Intermediate Regression Analysis With SPSS

James M. Clark
Department of Psychology
University of Winnipeg

TABLE OF CONTENTS

PREFACE

Unit 1 reviews basic aspects of statistics, including univariate correlation and regression (i.e., single predictor). Unit 2 introduces multiple regression (i.e., multiple predictors) with two predictors. Unit 3 extends regression analysis to multiple predictors and various applications (e.g., nonlinear relationships, categorical predictors). A companion manuscript on Analysis of Variance (ANOVA) demonstrates the equivalence of regression and ANOVA analyses despite their seeming differences. In essence, ANOVA involves applications of regression (sometimes labeled General Linear Model) to certain research scenarios that allow for alternative but equivalent calculations to those used for regression. This point is made briefly in the present chapters where relevant.

Regression as presented in this manuscript involves only basic mathematical operations. Understanding material like statistics requires practice and repetition, and also benefits from exposure to alternative conceptualization of important features, such as the unique contribution of individual predictors in a multiple regression. Although, conceptual repetition can seem confusing and redundant, it results in a deeper understanding of the analyses.

With respect to SPSS, I focus on syntax, which is to be recommended over a menu approach. It provides a record of the analyses, makes it easy to correct and rerun analyses, allows creation of simulations to generate data, and gives access to some procedures not available by menu (e.g., MANOVA).

Thanks to several colleagues who have contributed to my own understanding of regression and to the many students over the years who tolerated my sometimes “casual” lectures on this material. Errors or suggestions? Please e-mail j.clark@uwinnipeg.ca. Thanks ... Jim

CHAPTER 1 - SINGLE SAMPLE STATISTICS

Data analysis begins with a collection of scores, called a sample. The sample generally comes from a large population of scores, but for this chapter the population consists of 51 students who rated the statement "I look forward with great pleasure to Intermediate Statistics" on a 7-point scale (1 = Strongly Disagree, 7 = Strongly Agree). A random sample of 9 observations was selected from this population of 51 ratings.

Descriptive Statistics

Descriptive statistics describe characteristics of a sample of scores, such as where along the dimension of agreement the set of scores tends to fall (central location) and how spread out the sample scores are on this dimension (variability). These descriptive statistics are used to make inferences about the population of scores from which they come.

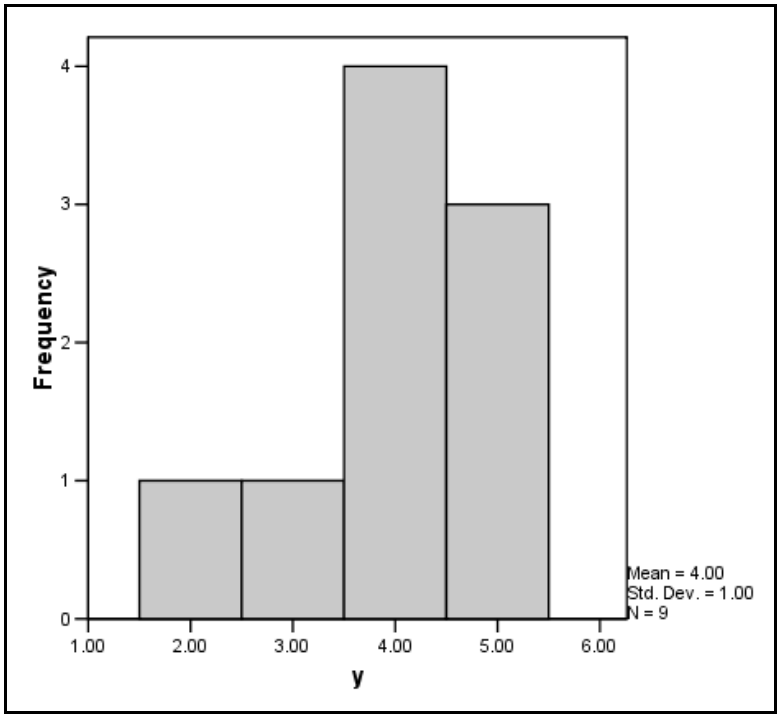


Figure 1-1. Frequency Distribution (Histogram) of Scores.

The following commands are typed in SPSS's syntax window (see supplementary handout on SPSS) and create an SPSS data file. Bolded text contains SPSS commands in capital letters and user text in lowercase letters. Lines beginning with * are comments and ignored by SPSS. Non-bolded regular text contains SPSS printouts from the commands that precede the output, and italicized text contains notes added to the output.

```
*Sample of 9 observations.
DATA LIST FREE / y.
BEGIN DATA
2 4 3 5 5 4 4 4 5
END DATA.

*Listing (includes later calculations).
LIST.
```

y	mean	ydev	ydev2
2.00	4.00	-2.00	4.00
4.00	4.00	.00	.00
3.00	4.00	-1.00	1.00
5.00	4.00	1.00	1.00
5.00	4.00	1.00	1.00
4.00	4.00	.00	.00
4.00	4.00	.00	.00
4.00	4.00	.00	.00
5.00	4.00	1.00	1.00

A good first step in data analysis is to represent the results visually. For this sample of scores, a frequency distribution is appropriate. The following command produced the graph in Figure 1-1 and the table below. The horizontal axis represents the rating scale and the vertical axis the frequency of each value.

```
FREQUENCIES y /HISTOGRAM.
```

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	1	11.1	11.1	11.1
	3.00	1	11.1	11.1	22.2
	4.00	4	44.4	44.4	66.7
	5.00	3	33.3	33.3	100.0
	Total	9	100.0	100.0	

A frequency distribution has two qualities, central location and variability. The central location, middle, or average of the distribution appears to be around 4. With respect to variability, scores range from 2 to 5. Consider first central location or central tendency. There are various measures of average, but the most common is the mean: the sum of the scores or total divided by the number of scores. In summation notation (see supplementary handout on summation notarion), $\bar{y} = \sum y/n$, where $\sum y$ (sum of y or sigma y) is the total and n is the number of scores. SPSS can perform such calculations even for very many scores. Here is the command to get the sum of the 9 scores.

```
DESCRIPTIVES y /STATISTICS = SUM.
```

	N	Sum	
y	9	36.00	$= \sum y$

Given a sum of 36.0 and n = 9 scores, the mean is $\bar{y} = 36/9 = 4.0$ (see formula in Box 1-1), exactly where the center appeared to be in the histogram. The first COMPUTE below adds the mean to the file (see the second column headed *mean* in the earlier listing). The second COMPUTE subtracts the mean from each of the scores, producing *ydev* in column three, and the third COMPUTE squares these deviations producing *ydev2* in column 4. These calculations can be used to demonstrate why the mean is a good measure of the average or central tendency for a distribution.

$$\bar{y} = \frac{\sum y}{n}$$

Box 1-1.

```
COMPUTE mean = 36 / 9.          = ∑y/n = ȳ
COMPUTE ydev  = y - mean.       = y−ȳ
COMPUTE ydev2 = ydev**2.        = (y−ȳ)²
```

One reason the mean is a good measure of the average is that it is the point of balance of the set of scores. The distance from the mean to all scores above the mean is exactly the same as the distance to all scores below. In terms of summation notation, $\sum(y-\bar{y}) = 0$. Summing the values in the *ydev* column gives a total of 0. The following command instructs SPSS to compute the sum of *ydev*. The result is 0, as expected.

```
DESCRIPTIVES ydev ydev2 /STATISTICS = SUM.
```

	N	Sum	
ydev	9	.00	$= \sum(y-\bar{y})$
ydev2	9	8.00	$= \sum(y-\bar{y})^2$

A second way to think about an average is how close it is to all scores. Because deviations from the mean always sum to 0, they do not measure how close the mean is to all 9 scores. Squaring the deviations removes the positive and negative signs and squared deviations can be summed. The further the scores are from the mean, the larger the sum of squared deviations. In summation notation, the sum of squared deviations about the mean is $\sum(y-\bar{y})^2 = 8.0$ as calculated above by SPSS or by summing the *ydev2* scores in the earlier listing. This quantity is abbreviated SS for sum of squares and is used often in statistics.

SS provides a second reason why the mean is a good measure of central tendency; namely, the sum of squared deviations about the mean is a minimum. That is, the sum of squared deviations about the mean is smaller than the sum of squared deviations from any other value. For example, subtracting 4.5 from each score and squaring those deviations gives a value greater than SS = 8.0. The mean comes as close as possible to all scores in terms of squared deviations or squared distances. Figure 1-2 shows $\sum(y-k)^2$ for different values of k. The sum of deviations about k is a minimum, 8.0, when k = 4.0, which is the mean. The mean is the point closest to all the data points in terms of squared deviations.

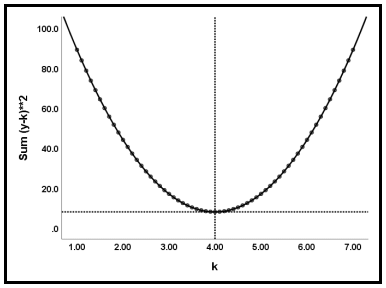


Figure 1-2.

A final reason the sample mean is a good measure of average is that the expected value (EV) of all possible sample means is the average for the population from which the sample comes; that is, $EV(\bar{y}) = \mu$, the population mean (represented by the Greek letter mu). A later section shows this more concretely.

The second property of the sample to quantify is its variability. SS measures variability because the further the scores are from the mean, the larger SS will be (see Box 1-2). When all scores are identical (no variability), then the scores equal $\bar{y}$ and SS equals 0. The value for our sample, SS = 8.0, measures the amount of variability in the 9 scores. A major goal of statistics is to determine how much other variables can predict or explain of SS.

$$SS_y = \sum (y-\bar{y})^2$$
$$s^2 = \frac{SS_y}{n-1} \qquad s = \sqrt{s^2}$$
Box 1-2.

One limitation of SS is that it depends on the number of scores and not just the extent of variability in the scores. If we doubled the number of scores by duplicating the original 9 scores, the extent of variability has not increased, but SS will. We need to calculate something like the "average" variability in the scores. The initial impulse might be to divide SS by the number of scores, but this would only be correct IF SS was based on the entire population of scores, whereas our SS is based on a sample rather than the population. The

reason it would be inappropriate to divide by n can be thought of in two ways.

One way is to remember that the mean used the sum of the scores, 36.0. Given this total, not all scores are "free to vary" because the scores must always sum to 36.0. For example, if we did not know the first score in the sample (the 2 in the listing), then the scores would sum to 34.0. But the sum has to be 36.0 to produce a mean of 4.0, so the first (missing) score must be 2. The same logic applies if any other score is removed, but not if two or more scores are removed. In statistical terms, only n-1 scores are free to vary. This quantity is called the degrees of freedom or df for short. To calculate an "average" variability, SS is divided by df = n - 1 = 9 - 1 = 8. This statistic is called the variance, $s^2 = SS/df = 8.0/8 = 1.0$, for this sample.

A second way to appreciate why we must divide SS by a number less than n is because the sample SS is a minimum; that is, $\sum(y-\bar{y})^2 = 8.0$ is a minimum. This is an issue because researchers are actually interested in the variability in the population, rather than the sample. In the population, $SS = \sum(y-\mu)^2$, where μ is the population mean not the sample mean. But μ generally will differ from $\bar{y}$ except rarely when the two are exactly the same. This means that $\sum(y-\bar{y})^2$ will generally be less than $\sum(y-\mu)^2$ and too small an estimate of the population variability or variance, σ^2 (represented by lowercase Greek letter sigma). To adjust for a sample SS that is too small, SS is divided by a number smaller than n, namely n-1. This will give a better estimate (Expected Value) of the population variance; that is, $EV(s^2) = \sigma^2$, as shown later.

One issue with s^2 is that it is the "average" squared deviation from the mean. People normally think in terms of actual units of distance, not squared distances. To eliminate the squaring, the square root of the variance produces another measure of variability, the standard deviation, $s = \sqrt{s^2} = \sqrt{1.0} = 1.0$ in the present sample. Normally, s^2 and s will rarely be equal.

SPSS can calculate these descriptive statistics directly. Note below that the full SPSS command words DESCRIPTIVE and STATISTICS are not required. SPSS only needs enough characters to uniquely identify the specific command. As well, SPSS calculates default statistics if /STATISTIC = is omitted.

```
DESCR y /STAT = MEAN VARIANCE STDDEV.

                N Mean   Std. Deviation Variance
 y              9 4.0000 1.00000        1.000
```

Although SPSS can calculate most of the quantities needed for analyses, calculating the quantities by hand helps to understand the concepts and interpret the SPSS output. Therefore, practice the calculations so that they become automatic.

Sampling Distributions

The sample of 9 scores was selected from a population of 51 scores. The population mean, variance, and standard deviation are: $\mu = 3.82353$, $\sigma^2 = 2.8904$, $\sigma = 1.7001$. Normally researchers do not know these population values, but must estimate them from a sample.

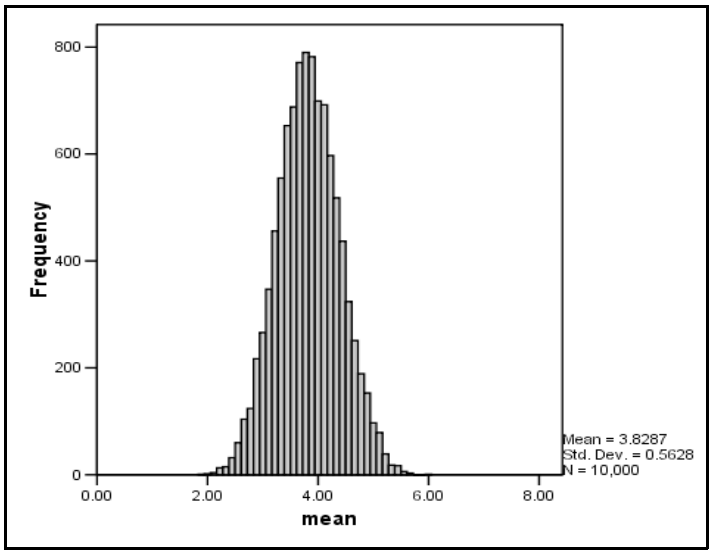


Figure 1-4. Sampling distribution of $\bar{y}$s.

Our sample was one of many possible samples that could be selected from this population. To illustrate, 10,000 samples of n = 9 observations were randomly selected to produce the 10,000 sample means, $\bar{y}$s, plotted in Figure 1-4. A frequency or probability distribution for a sample statistic, $\bar{y}$ here, is called a Sampling Distribution. Observe that the mean of the sample means, 3.8287 = EV($\bar{y}$), is very close to the population mean, despite variability across samples. We previously noted that EV($\bar{y}$) is μ; the expected value is the mean of all possible $\bar{y}$s.

Figure 1-5 shows the sampling distribution of the variances for 10,000 samples. The mean of the sample variances, 2.9029, is close to the population variance, $\sigma^2 = 2.8904$. Each sample variance was SS divided by df = n - 1. If SS had been divided by n, the mean of the variances would be 2.5804, much smaller than the population variance. SS/n gives a value too small because $SS_y = \Sigma(y - \bar{y})^2$, whereas the desired SS is $\Sigma(y - \mu)^2$. Because the sum of squared deviations about the sample mean is a minimum, $\Sigma(y - \bar{y})^2$ will always be less than or (rarely) equal to $\Sigma(y - \mu)^2$. That is, sample SSs are too small.

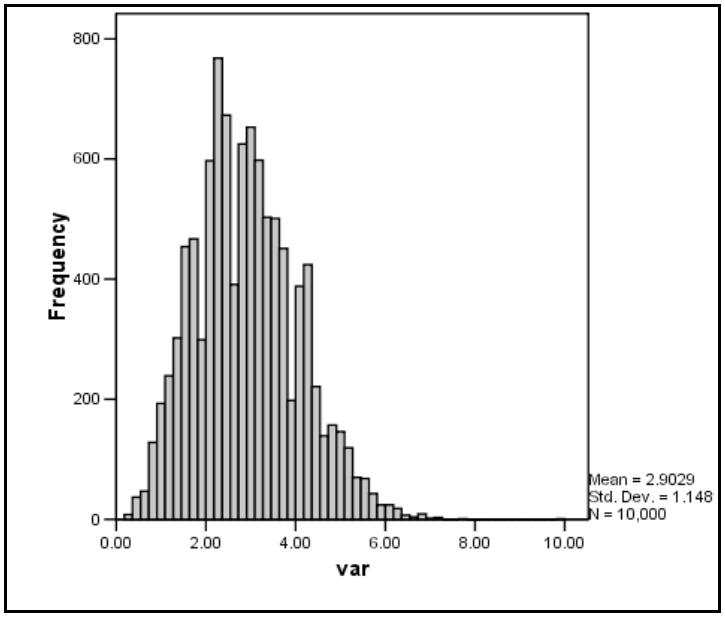


Figure 1-5. Sampling distribution of variances (s^2).

In fact the mean of the sample SSs is 23.2232, whereas the mean of $\Sigma(y - \mu)^2$ is 26.0738. To adjust for SS being too small, SS is divided by n-1 to produce a better estimate of σ^2. Observe that dividing 26.0738, the squared deviations about the population mean rather than the sample mean, by 9 (i.e., n) produces a value of 2.8971, very close to the population variance.

$$\mu_{\bar{y}} = \mu_y$$
$$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}}$$
Box 1-3.

A second important aspect of the Sampling Distribution of $\bar{y}$ is the variability in the $\bar{y}$

from sample to sample: $s_{\bar{y}} = .5628$ in Figure 1-4. Normally, we do not have 10,000 samples to calculate $s_{\bar{y}}$ for the sample means, but the Central Limit Theorem (CLT) states that $s_{\bar{y}}$ will be $\sigma/\sqrt{n}$ (see Box 1-3). In the present case where we know σ, we expect $s_{\bar{y}} = 1.7001/\sqrt{9} = .5667$, very close to $s_{\bar{y}}$ for the simulation. The standard deviation for a sample statistic is also called its Standard Error (SE).

So if μ_y and σ_y are known for the population, which is rare but true in the present case, then $\mu_{\bar{y}} = \mu_y$, and $\sigma_{\bar{y}} = \sigma_y/\sqrt{n}$. The CLT also states that the distribution of $\bar{y}$ will be normal, which means that a z-score could be calculated for observed $\bar{y}$s using $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$; that is, $z = (\bar{y} - \mu_{\bar{y}})/\sigma_{\bar{y}}$.

Appendix 1-1 presents some samples of data with which to practice calculation of the descriptive statistics described here. They focus on the calculations only, but you want those to be as automatic as possible so that they can be done quickly and do not disrupt the cognitive processes required to understand and explain what they mean. They also provide an opportunity to learn about and practice working with your calculator (e.g., memory, brackets, how it performs successive operations). Again, you want the ability to do operations to be fluent, certainly for tests, but even for class activities and assignments. You can also enter the data into SPSS to become more familiar with SPSS commands and output. When working with larger sets of data, use SPSS to do the necessary calculations to put into formula {e.g., $\sum y$, $\sum(y-\bar{y})^2$}.

APPENDIX 1-1
DESCRIPTIVE STATISTICS EXERCISE SHEET

$$\text{MEAN} = \bar{y} = \frac{\sum y}{n} \qquad SS = \sum (y-\bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = (n-1)s^2 \qquad \text{VAR} = s^2 = \frac{SS}{n-1} \qquad SD = s = \sqrt{s^2}$$

SAMP	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	MEAN	SS	VAR	SD
1 -	14	13	13	8	9								11.400	29.200	7.300	2.702
2 -	11	8	7	7	6	11	6	7	4				7.444	42.222	5.278	2.297
3 -	8	8	8	7	6								7.400	3.200	.800	.894
4 -	23	24	31	23	22	27	19	24	29				24.667	110.000	13.750	3.708
5 -	7	6	7	6	10								7.200	10.800	2.700	1.643
6 -	14	19	21	19	21	20	18	22	19	19	20		19.273	44.182	4.418	2.102
7 -	3	10	10	9	6	7							7.500	37.500	7.500	2.739
8 -	19	19	17	20	18	20	22	15	18	23	22	9	18.500	155.000	14.091	3.754
9 -	21	25	23	16	27	21	18	13	27	21	19		21.000	194.000	19.400	4.405
10 -	3	5	3	3	5	6	4	5					4.250	9.500	1.357	1.165
11 -	22	25	17	18	18	21	23	17	29	19	17		20.545	152.727	15.273	3.908
12 -	14	10	16	12	11	17	15	13	16	18			14.200	63.600	7.067	2.658
13 -	13	18	23	18	14								17.200	62.800	15.700	3.962
14 -	5	4	4	5	5	5	6	2	5	5			4.600	10.400	1.156	1.075
15 -	10	5	8	9	9	11	6	9	8	8	11		8.545	34.727	3.473	1.864
16 -	16	13	18	8	18	18							15.167	80.833	16.167	4.021
17 -	19	21	17	17	15	19	20						18.286	25.429	4.238	2.059
18 -	18	19	15	13	18	19	12	18					16.500	54.000	7.714	2.777
19 -	11	13	15	9	15	14	16	13					13.250	37.500	5.357	2.315
20 -	22	27	11	20	13								18.600	173.200	43.300	6.580
21 -	6	6	6	5	5	7	6	8					6.125	6.875	.982	.991
22 -	7	7	11	12	9	8	10	8					9.000	24.000	3.429	1.852
23 -	5	4	6	5	6	5	6						5.286	3.429	.571	.756
24 -	14	18	18	16	16	15	20	15	17				16.556	28.222	3.528	1.878
25 -	13	16	8	18	15	16	17	12	18	21	18	18	15.833	131.667	11.970	3.460
26 -	16	16	18	13	15	17	20	18	16	13	17		16.273	44.182	4.418	2.102
27 -	11	10	10	9	12	12							10.667	7.333	1.467	1.211
28 -	30	23	26	19	21	24	32	29	22	26	25	33	25.833	213.667	19.424	4.407
29 -	16	16	24	23	26								21.000	88.000	22.000	4.690
30 -	5	9	10	7	9	7	8	6	11				8.000	30.000	3.750	1.936
31 -	8	12	7	9	9	11	9	7	10	8	7	13	9.167	43.667	3.970	1.992
32 -	10	11	15	15									12.750	20.750	6.917	2.630
33 -	21	18	15	14	11	11							15.000	78.000	15.600	3.950
34 -	18	12	19	20	19	21	19	14	10				16.889	120.889	15.111	3.887
35 -	5	6	7	7	5	5	6	6	7				6.000	6.000	.750	.866
36 -	18	17	17	18									17.500	1.000	.333	.577
37 -	6	7	6	6	7	8	7	7	6	7			6.700	4.100	.456	.675
38 -	18	19	19	26	22	24	21	22	10				20.111	166.889	20.861	4.567
39 -	7	7	8	7	8	8	9						7.714	3.429	.571	.756
40 -	20	21	21	24	22								21.600	9.200	2.300	1.517
41 -	8	10	8	11	9	7	9	6	6	8	11		8.455	30.727	3.073	1.753
42 -	8	7	8	6	7	5							6.833	6.833	1.367	1.169
43 -	7	7	9	8	7	7	7						7.429	3.714	.619	.787
44 -	8	13	13	12	11	15	12						12.000	28.000	4.667	2.160
45 -	21	21	14	16	12	18	18	19	18	22	14	17	17.500	105.000	9.545	3.090
46 -	4	4	5	6	4	4	5	6	4	5	4	6	4.750	8.250	.750	.866
47 -	18	26	19	22	15	14	20	25	25	25	25		21.273	188.182	18.818	4.338
48 -	11	12	11	11	9	11	11						10.857	4.857	.810	.900
49 -	29	27	26	25	26	14	24	21	25	16	29		23.818	241.636	24.164	4.916
50 -	17	16	17	24									18.500	41.000	13.667	3.697

CHAPTER 2 - HYPOTHESIS TESTING

The preceding lecture showed the relationship between samples and populations. The population was known in our example, but researchers generally want to make an inference about an unknown population based on a sample. This is called hypothesis testing and involves inferences about some population value (e.g., population mean, difference between population means, correlation in population). Consider an hypothesis about a population mean. Suppose researchers want to test a theory that predicts a certain population (e.g., university graduates) has an IQ higher than the general population (i.e.,$\mu = 100$).

The first step might seem unusual at first. Start with a statistical hypothesis that university students do *not* have a higher IQ. This is called the Null Hypothesis: H_0 $\mu = 100$. The prediction is called the Alternative Hypothesis and reflects the prediction: H_a $\mu > 100$. Researchers select a sample of university students and calculate their IQ: $\bar{y} = 107$, for example. A statistical test is applied to decide what is the probability that this sample came from the H_0 population (i.e., people with an average IQ = 100). If the probability is lower than a value that researchers have selected, called Alpha (α), then they reject H_0 $\mu = 100$ and accept H_a $\mu > 100$. But if the probability of $\bar{y}$ coming from the H_0 population is greater than alpha, they do not (i.e., fail to) reject H_0 and accept H_a $\mu > 100$, which is what the theory predicted.

Box 2-1 demonstrates four possible outcomes given the actual (but unknown) state of affairs and the decision made by the researchers. Two of the four cells represent correct decisions. One is if the researchers reject a false H_0, and the other is if they do not reject a true H_0. The other two cells represent errors. Rejecting a true H_0 is one kind of error, called a Type I error. Failing to reject a false H_0 is another error, called a Type II error. This course focuses almost entirely on statistics related to Type I errors. Type II errors are also very important, but the statistics involved benefit from a good appreciation of analyses and Type I errors.

	TRUE (UNKNOWN) STATE	
DECISION	**H_0 True**	**H_0 False**
Reject H_0	Type I Error	Correct
Do Not Reject H_0	Correct	Type II Error

Box 2-1. Hypothesis Testing Framework.

By selecting different values for alpha, researchers control the probability of a Type I error. If they choose $\alpha = .05$, then we expect to reject a true H_0 5% of the time or 5 out of 100 times. If researchers are not willing to take that much risk of a Type I error, they could use $\alpha = .01$. Then the probability of a Type I error is 1 out of 100 times. If willing to take more of a risk than .05, researchers could use $\alpha = .10$ and accept a probability of rejecting a true H_0 as 10 out of 100 times. Note that choosing a very small value for α increases the probability of a Type II error. By making it less likely to Reject H_0, researchers increase the decision Do Not Reject H_0, which increases the probability of a Type II error.

A major challenge in this hypothesis testing scenario is how to calculate the probability of the

observed outcome if H_0 is true so that the observed probability can be compared to α. Researchers calculate the probability based on the sampling distribution of the relevant test statistic for the sample, $\bar{y}$ for μ in our example, or other sample statistics for hypotheses about different population values.

Hypothesis Test for Single Population Mean ($\mu=\mu_0$)

Researchers generally do not know what μ is, but they can hypothesize a value denoted as μ_0 and then determine whether $\bar{y}$ is farther from the hypothesized value than expected. The hypothesized value is called the Null Hypothesis, H_0. Inferential statistics calculate the probability of the sample statistic if the H_0 is true. For tests of a single μ, the null specifies a specific value, μ_0 in Box 2-2.

$$H_0: \mu = \mu_0$$
$$H_a: \mu < \mu_0 \text{ or } \mu > \mu_0 \text{ or } \mu \neq \mu_0$$
$$z = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{or} \quad t = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Box 2-2.

Because μ = 3.82353 and σ = 1.7001 for our population, we can illustrate the logic of hypothesis testing using the z distribution. For each of 10,000 samples we calculate a z score using the following command. Note that it uses μ and σ for the population. The distribution of the resulting zs appears in Figure 2-1.

```
COMPUTE z = (mean-3.82353)/[1.7001/SQRT(9)].
```

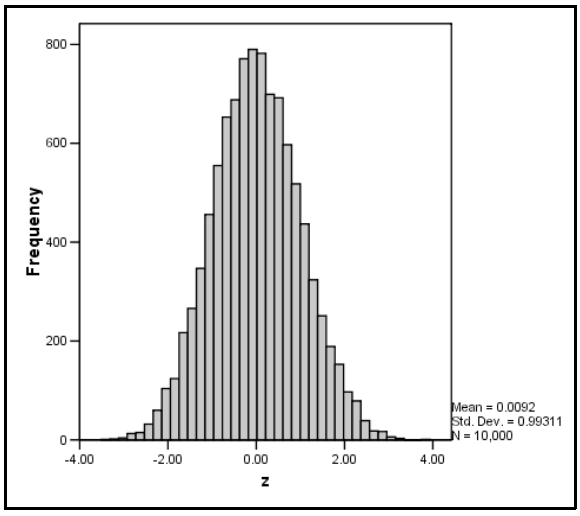


Figure 2-1. Distribution of z Statistic.

As expected for a normal distribution, $\bar{z} \approx 0$ and $s_Z \approx 1$. The probability distribution for normal z scores states that about 5% of zs should be less than or equal to -1.96 or greater than or equal to +1.96. So to reject H_0: μ = 3.82353 only 5% of the time when it is true (i.e., use α = .05), researchers reject the H_0 if z for $\bar{y}$ is less than or equal to -1.96 or greater than or equal to +1.96. In fact, in the 10,000 samples, 4.9% of the zs were less than or equal to -1.96 or greater than or equal to +1.96, close to the 5.0% value expected theoretically when the H_0 is true.

The z distribution cannot be used in most studies because researchers usually do not know what σ is. Instead they only have s for a single sample, which means they cannot calculate a z-statistic. Instead, a different sampling distribution must be used, such as the t or F distribution. When σ is unknown, s (*std* in the simulated data) is used in the denominator of the t statistic (see Box 2-2).

```
COMPUTE t = (mean - 3.82353)/(std/SQRT(9)).
```

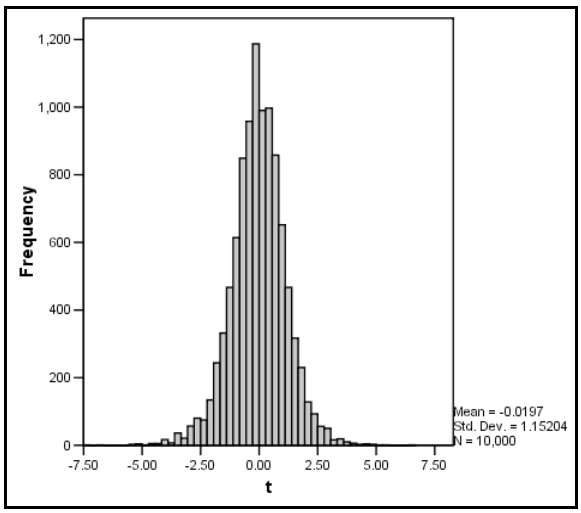


Figure 2-2. Distribution of t Statistic.

The sampling distribution of the 10,000 ts is shown in Figure 2-2. The mean t is still very close to 0, as expected given the correct population mean; however, there is more variability in the t statistic than in the z statistic, $s_t = 1.152$ rather than 1.0. This occurs because s varies from sample to sample, and an s smaller than σ will produce a t larger than z for that sample.

Given the greater variability, the percent of ts less than or equal to -1.96 OR greater than or equal to +1.96 is more than 5%. Using ±1.96 as a cut-off, 8.3% (p = .083) of the ts would lead researchers to reject H_0 even when it is true. This would be higher than expected if the desired probability of a Type I Error (i.e., rejecting a true null hypothesis) was .05.

The correct probability uses a critical value from the t distribution rather than the z distribution. The distribution of t varies with its df, n-1 here (from s). From the table for t, $t_{.025} = 2.306$ for df = n - 1 = 8. If H_0 is true, then $p(t \geq 2.306) = .025$. Since the t distribution is symmetrical, $p(t \leq -2.306) = .025$ as well. Therefore, p(t ≤ -2.306 OR t ≥ +2.306 IF H_0 true) = .05 = p(Reject H_0 if it is true) = p(Type I Error) = α (alpha). In the 10,000 samples, 5.3% of the ts fall outside these limits, close to 5.0%.

A second, equivalent test about the population mean uses the F distribution. The F statistic is basically a ratio of two variances, $F = s_1^2 / s_2^2$. F can test different hypotheses by generating a numerator variance that is sensitive to deviations from the null hypothesis and a denominator variance that represents random variation or noise. If F is too large (it can never be negative), then the null hypothesis is rejected.

$$F = \frac{n \times (\bar{y} - \mu_0)^2}{s_y^2}$$

Box 2-3.

The denominator variance in the present case is simply the sample variance s_y^2 (called *var* in the simulation). This represents random or error variation about the sample mean. For the numerator variance, square the deviation of the sample mean from the hypothesized value, and multiply by n, the number of observations (see Box 2-3). That is, $s^2_{Numerator} = n \times (\bar{y} - \mu_0)^2$, where μ_0 is the hypothesized value. Both the numerator and denominator will be positive, and so will F. The greater the distance between $\bar{y}$ and μ_0, the greater the variance in the numerator and the greater the value for F. For the SPSS simulation, the calculation is:

```
COMPUTE f = (9*(mean-3.82353)**2)/var.
```

The distribution of the 10,000 Fs is shown in Figure 2-3. All F values are positive, and the distribution is highly skewed (i.e., not symmetrical, but with a long tail to the right). One sample produced

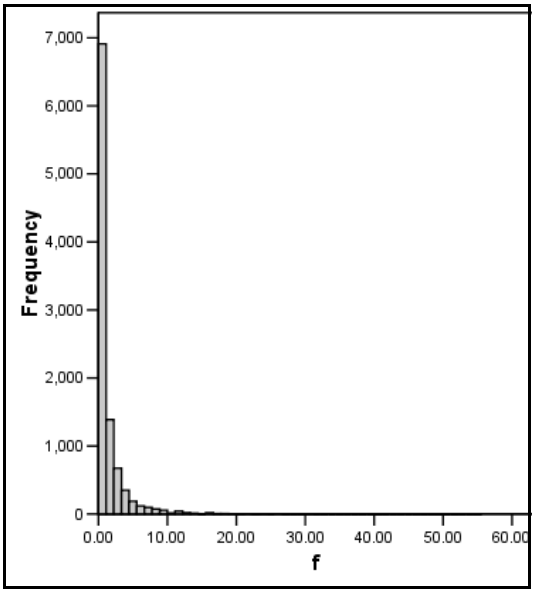


Figure 2-3. Distribution of 10,000 Fs.

F = 55.56, an extremely large value. The critical value for F depends on df for the numerator and df for the denominator, 1 and 9-1 here. The F table gives F = 5.32 as the critical value for $\alpha = .05$ with df = 1 for the numerator and df = 8 for the denominator. In fact, 5.3% of the 10,000 Fs were greater than or equal to that value, close to what is expected when the H_0 is true and the same percentage as for the t distribution.

The t and F tests are equivalent when $df_{Numerator}$ for F is 1, as for the single sample test for μ. Specifically, the observed value t equals the square root of the observed F (conversely, $F = t^2$). Similarly, the critical value for t, 2.306, equals the square root of the critical value for F, 5.32 (conversely, $5.32 = 2.306^2$). The two tests lead to the same conclusion because if $t_{Observed} > t_\alpha$, then $F_{Observed} > F_\alpha$ (i.e., $t^2_{Obs} > t_\alpha^2$). One way to think about this relationship is that the F distribution sometimes equals the t distribution "folded" over its middle value of 0. That is, the negative side of the t distribution overlaps the positive side, hence $F = t^2$.

The preceding demonstrations examined the distribution of statistics when the null hypothesis is true. If the null hypothesis is false, then the probability of extreme values for z, t, or F is greater than when the null is true. For example, if we hypothesized H_0 $\mu = 4.5$, then the command,

```
COMPUTE t = (mean - 4.5)/(std/SQRT(9))
```

produces more extreme values because the sampling distribution of t is no longer centered at H_0. In fact, 25.7% of these ts fall outside the critical values of -2.306 and +2.306, and would correctly reject the null. For the other 74.3% of samples, H_0 is not rejected, giving a Type II Error (i.e., fail to reject false null hypothesis).

SPSS can perform these tests about the mean, as shown below for the sample used last class. Assume the hypothesis that students on average would give the question a rating of 5 or greater. The value of 5 becomes our null hypothesis. The sample data is entered the same way as before.

```
DATA LIST FREE / y.
BEGIN DATA
2 4 3 5 5 4 4 4 5
END DATA.
```

Below is SPSS's TTEST command for testing an hypothesis about a single mean. The observed t value is -3.00, which falls in the rejection region given the critical value of t determined earlier, $t_\alpha = \pm 2.306$. Equivalently, the observed probability of .017 in the Sig. column is less than $\alpha = .05$, which means the null hypothesis that $\mu = 5.0$ can be rejected. The significance (Sig.) indicates how likely our observed $\bar{y}$ is given $\mu = 5.0$. Namely, $p(t_{obs} \leq -3.00$ OR $t_{obs} \geq +3.00$ IF Ho is true) = .017.

```
TTEST /TESTVALUE = 5 /VARIABLE = y.

   N Mean   Std. Deviation Std. Error Mean
 y 9 4.0000 1.00000        .33333

   Test Value = 5
   t         df Sig.       Mean
                (2-tailed) Difference
 y -3.000    8  .017       -1.00000
```

Calculations (see Box 2-2): $t = (4.0 - 5.0) / (1.0/\sqrt{9}) = -1.0 / .3333 = -3.00$, $df = n - 1 = 8$

Figure 2-4 shows the relationship between $t_{Observed}$ falling in the rejection region (i.e., greater than $t_{Critical}$) and $p_{Observed}$ being less than α. The values are shown just for the right side (tail) of the t distribution; the left tail (-2.306) would be equivalent. The $p_{Observed}$ value .0085 = .017/2 because .017 includes both tails of the distribution. As illustrated, whenever $t_{Observed}$ is greater than $t_{Critical}$, $p_{Observed}$ will be less than α. In the present case, $t_{Observed} = +3.00 > t_{Critical} = +2.306$ and half of $p_{Observed} = .0085 < \alpha/2$. In terms of both tails, $p_{Observed} = .017 < \alpha = .05$.

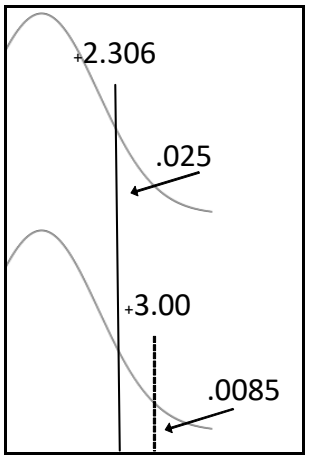


Figure 2-4.

Using SPSS to calculate the significance of the F statistic for this test requires a little "trick." GLM and other ANOVA commands can test the significance of an observed mean relative to a hypothesized value of 0, but not to some non-zero value (e.g., 5.0). To get around this limitation, subtract 5.0 from the scores and test whether the new scores differ from 0, as shown below. To determine whether to reject the null hypothesis that $\mu = 5.0$, compare F = 9.0 to the critical value of 5.32 or the observed significance of .017 to $\alpha = .05$. Using either approach, the null hypothesis is rejected. The equivalence of the t and F tests appears below.

```
COMPUTE yminus5 = y - 5.
GLM yminus5 /PRINT = DESCRIPTIVES.

 Mean    Std. Deviation N
 -1.0000 1.00000        9

 Source          Type III Sum of Squares df Mean Square F     Sig.
 Corrected Model .000(a)                 0  .           .     .

 Intercept       9.000                   1  9.000       9.000 .017
 Error           8.000                   8  1.000

 Total           17.000                  9
 Corrected Total 8.000                   8
```

Calculations (see Box 2-3): $F = \{9 \times (4.0-5.0)^2\}/s^2 = 9.0/1.0 = 9.0$, $df = 1, n - 1$

Equivalence of F & t: both p values (Sig) = .017; $t^2 = -3.0^2 = 9.0 = F$, OR $\sqrt{F} = \sqrt{9} = 3.0 = t$

Hypothesis Testing for Relationships: Difference Between Two Independent Means

Researchers develop theories and hypotheses to identify independent variables that can explain variation (i.e., SS) in a dependent variable of interest. Independent variables are also called predictors, and dependent variables can also be called criterion or outcome variables. Statistically, independent variables can be naturally occurring (e.g., gender, age) or experimental (e.g., treatment vs control, time given to study list of words). Independent variables can also be categorical (e.g., gender, treatment vs control) or numerical (e.g., age, time to study list of words). The critical factor in deciding about the appropriate statistical analysis of relationships between independent and dependent variables is whether the independent variable is categorical or numerical, not whether it is experimental or not.

For categorical predictors, researchers test whether scores for two or more samples of observations come from the same population (i.e., no difference between population means) or from different populations (i.e., different population means). In some studies observations in the two samples are related to one another (e.g., pre-test versus post-test scores from the same subjects). In other studies, scores are obtained from two samples that are unrelated or independent (e.g., control versus treatment groups to which people were randomly assigned). Here t and F are used to test the difference between two means for independent samples.

Two samples of five observations each were drawn from the population of 51 ratings (i.e., the two samples come from populations with the same mean; $\mu = 3.82353$). Data for the two samples are shown below, along with calculations required for the independent groups t-test (see Box 2-4).

The independent groups t-test tests whether the sample means $\bar{y}_1$ and $\bar{y}_2$ are sufficiently different to reject the H_0: $\mu_1 = \mu_2$ or its equivalent $\mu_1 - \mu_2 = 0$. The numerator is how far from 0 is the difference between sample means. The denominator is the standard error of the difference between means and tests whether the numerator difference is further than expected by chance from the hypothesized value of 0; for example, in less than 5% of random samples if $\alpha = .05$. Box 2-4 shows the formula to calculate the pooled variance, s_p^2, as the sum of the SSs for the two groups in the numerator and the sum of the dfs in the denominator. The rationale for this is that s_p^2 is a better estimate of σ^2 than the separate sample variances. This value and the ns are used to calculate $SE_{\bar{y}1-\bar{y}2}$, the denominator for t in Box 2-4.

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 > \mu_2 \quad \text{or} \quad \mu_1 < \mu_2 \quad \text{or} \quad \mu_1 \neq \mu_2$$

$$s^2_{Pooled} = \frac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)}$$

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad df = n_1 + n_2 - 2$$

Box 2-4. Independent t-test.

	G1		G2
	2		4
	4		4
	3		4
	5		5
	5		6
$\bar{y}_1$	3.80	$\bar{y}_2$	4.60
SS_1	6.80	SS_2	3.20

$s_p^2 = (SS_1 + SS_2) / (df_1 + df_2) = (6.80 + 3.20) / [(5 - 1)+(5 - 1)]$
$= 10.0 / 8 = 1.25$

$SE = SQRT\{1.25(1/5+1/5)\} = .7071$

$t_{Obs} = (3.80 - 4.60) / .7071 = -0.80 / .7071 = -1.13$

$df = (5-1) + (5 - 1) = 5 + 5 - 2 = 8$

SPSS requires two variables, one to indicate the group (1 or 2) and the second for y, the dependent variable. The following commands enter the data and perform an independent groups t-test.

```
DATA LIST FREE / group y.
BEGIN DATA
1 2   1 4   1 3   1 5   1 5         2 4   2 4   2 4   2 5   2 6
END DATA.

TTEST /GROUP = group /VARIABLE = y.
```

	group	N	Mean	Std. Deviation	Std. Error Mean
y	1.00	5	3.8000	1.30384	.58310
	2.00	5	4.6000	.89443	.40000

$SS_1 = (n_1-1)*s_1^2 = (5-1)*1.30384^2 = 6.80$

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Std. Error Difference**
y	Equal variances assumed	1.024	.341	**-1.131**	**8**	**.291**	**-.80000**	**.70711**
	Equal variances not assumed			-1.131	7.082	.295	-.80000	.70711

The initial output reports descriptive statistics for the samples and provides everything required for the independent t-test, namely sample means and standard deviations that can be converted to SSs. The relevant results in the test section are in bold: observed t, df, mean difference (numerator), and standard error (denominator), and significance. The significance indicates that the probability t is ≤ -1.131 or ≥ +1.131 is .291 > .05 if H_0 is true. Also, $t_{Observed}$ is not ≥ 2.306 or ≤ -2.306, $t_{Critical}$, $df = n_1 + n_2 - 1 = 8$. Therefore we do not reject the H_0. Two elements in the output are ignored here: the F and Sig. for

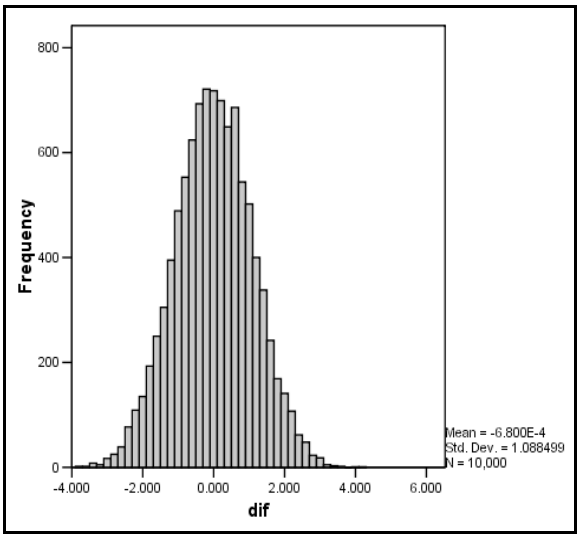


Figure 2-5. Distribution of $\bar{y}_1 - \bar{y}_2$.

Levene's test for the Equality of Variances and the t test in the second row for Equal variances not assumed. Both have to do with the assumption that it is reasonable to aggregate the separate variances to calculate the pooled variance.

Figure 2-5 shows the sampling distribution of *dif* = $\bar{y}_1 - \bar{y}_2$ for 10,000 samples selected from the population. The mean *dif* is 0, as expected, but the sample means vary, with $s_{dif} = 1.0885$, close to the expected value for the variability in the difference between means when σ is known:

$\sigma_{dif} = \text{SQRT}\{\sigma^2(1/n_1+1/n_2)\} = \text{SQRT}\{2.9482(1/5+1/5)\}$

$= 1.086$

Because σ is generally unknown, it must be estimated from the sample standard deviations or variances. The formula in Box 2-4 produces s_p^2, a pooled estimate of the population variance, which is used to calculate $SE_{\bar{y}1 - \bar{y}2}$ and $t_{Observed}$.

Figure 2-5 shows the sampling distribution of $t_{Observed}$ for the 10,000 samples. The mean t is 0, as expected, but s_t is greater than 1.0, which is what the standard deviation would be if we had calculated a z statistic using σ^2 instead of s_p^2. With df = 5 + 5 - 2 = 8, a nondirectional test, and α = .05 (i.e., a = .05/2 = .025), $t_{Critical} = \pm 2.306$. We do not reject H_0: $\mu_1 = \mu_2$ for our sample with t = -1.13. In fact, only 5.2% of the 10,000 *t*s are significant, very close to the theoretically expected percentage of Type I errors; that is, differences between means large enough to reject H_0 even though H_0 is true.

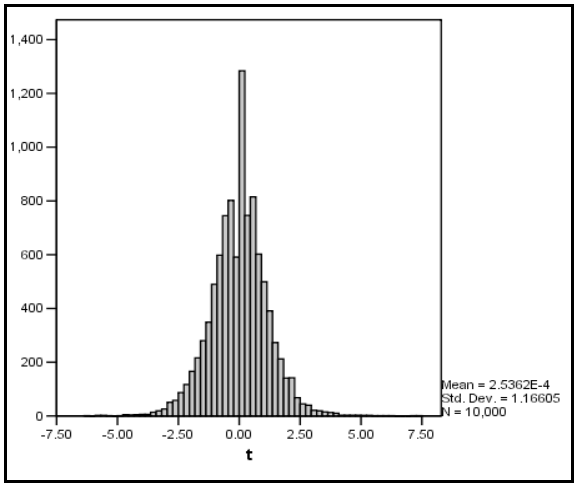


Figure 2-5. Sampling Distribution of t.

Independent Groups F test (Analysis of Variance)

The two means can also be compared by an *F* test (i.e., Analysis of Variance or ANOVA), with a numerator variance to represent the difference between means, and a denominator variance to represent random or error variation (see Box 2-5). The denominator is s_p^2, as calculated earlier. The numerator is based on the deviation of sample means from the overall (Grand) mean; $\bar{y}_G$ is the grand mean averaged across groups. The number of groups is k (2 here) and j is an index for the groups; $\bar{y}_j$, for example, represents $\bar{y}$ for each group; j = 1 and 2 for two groups.

$$SS_{Numerator} = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}_G)^2$$
$$= n_1 \times (\bar{y}_1 - \bar{y}_G)^2 + n_2 \times (\bar{y}_2 - \bar{y}_G)^2$$
$$F = \frac{\frac{SS_{Numerator}}{k-1}}{s_p^2} \quad df = k-1, N-k$$

Box 2-5.

Group (j)	$\bar{y}_j$		$\bar{y}_j - \bar{y}_{Grand}$		n_j
j=1	$\bar{y}_1$ 3.80	$\bar{y}_1 - \bar{y}_G$	−0.40	n_1	5
j=2	$\bar{y}_2$ 4.60	$\bar{y}_2 - \bar{y}_G$	+0.40	n_2	5
	$\bar{y}_{Grand}$ 4.20			N	10

$SS_{Num} = n_j\Sigma(\bar{y}_j - \bar{y}_G)^2 = 5\times(-0.40^2 + 0.40^2) = 5\times.32 = 1.60$ df = 2−1 = 1

$s^2_{Num} = SS_{Num}/df = 1.60/(2-1) = 1.60$

$F_{Observed} = s^2_{Num}/s_p^2 = (5\times.32)/1.25 = 1.6/1.25 = 1.28$ ($\sqrt{1.28} = 1.13 = t_{Obs}$)

df = k−1, N−k (k = # groups and N = total # subjects across all groups)

= 2 −1, 10 − 2 = 1, 8 $F_{Critical} = 5.32$ ($= 2.306^2 = t_{Critical}^2$)

The F-test is equivalent to a t-test when there are two groups (i.e., $df_{Numerator} = 2 - 1 = 1$) and leads to the same conclusion. However, F can be used with more than two groups. Several SPSS commands perform ANOVA, including GLM and MANOVA.

```
GLM y BY group /PRINT = DESCR.
```

group	Mean	Std. Deviation	N
1.00	3.8000	1.30384	5
2.00	4.6000	.89443	5
Total	4.2000	1.13529	10

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1.600(a)	1	1.600	1.280	.291
Intercept	176.400	1	176.400	141.120	.000
group	**1.600**	**1**	**1.600**	**1.280**	**.291**
Error	**10.000**	**8**	**1.250**		
Total	188.000	10			
Corrected Total	11.600	9			

```
MANOVA y BY group(1 2) /PRINT = CELL.
```

Cell Means and Standard Deviations

FACTOR	CODE	Mean	Std. Dev.	N
group	1	3.800	1.304	5
group	2	4.600	.894	5
For entire sample		4.200	1.135	10

Source of Variation	SS	DF	MS	F	Sig of F
WITHIN CELLS	**10.00**	**8**	**1.25**		
group	**1.60**	**1**	**1.60**	**1.28**	**.291**
(Model)	1.60	1	1.60	1.28	.291
(Total)	11.60	9	1.29		

R-Squared = .138 Adjusted R-Squared = .030

There is one important aspect of the analysis that merits emphasis. Not that SS_{Total} (Corrected Total in GLM) is 11.60, which is the sum of SS_{Group} and SS_{Error}, also referred to as $SS_{Between}$ and SS_{Within} in ANOVA. The value of 11.60 is the total variability in the 10 scores, which can be calculated from the standard

deviation for all 10 scores. SS_{Total} is partitioned (divided) into variability between groups and variability within groups.

$$SS_{Group} + SS_{Error} = SS_{Total} = (10-1)1.13529^2 = 11.60 = 1.60 + 10.00$$

General Schema for Hypothesis Testing

Given several examples of statistical tests, now is a good time to develop more specifically a general template for hypothesis testing. The Null Hypothesis involves a parameter or statistic for the population from which samples are theoretically selected (e.g., μ, $\mu_1 - \mu_2$, σ^2, ρ, ...), where ρ (rho) represents the population correlation coefficient covered in the next chapter.

$$H_0: \mu = \mu_0, \mu_1 = \mu_2, \rho = 0, ...$$

The Alternative Hypothesis is an expectation about the parameter if the null hypothesis is false. This may be a general hypothesis that H_0 is false, or a more specific prediction based on past research or theory.

$$H_a: \mu \neq \mu_0, \mu > \mu_0, \mu_1 \neq \mu_2, \mu_1 < \mu_2, \rho > 0, ...$$

To decide between the Null and Alternative Hypotheses, researchers calculate a sample statistic that estimates the population parameter and a measure of the variability expected in the sample statistic (i.e., the standard deviation or standard error of the sampling distribution for the statistic). These values are used to compute an inferential statistic (e.g., t or F) to determine the probability that the observed statistic or a more extreme value would occur if the Null Hypothesis was true. For example, $p(t \geq t_{Observed})$ IF H_0 true.

If the observed outcome is too unexpected if the null hypothesis is true (i.e., it's probability is low), then the H_0 may be rejected and the H_a accepted. A standard value for "too unexpected" is $\alpha = .05$, although sometimes smaller or larger values are used for α. The value alpha, α, represents the probability of a Type I Error = p(Reject H_0 IF H_0 true). The H_0 is rejected and H_a accepted if $p_{Observed} \leq \alpha$, or if the observed test statistic is more extreme than the critical value corresponding to α; for example, $p(t_{Observed} \leq -t_{Critical}$ OR $t_{Observed} \geq +t_{Critical})$ IF H_0 true.

A final detail concerns the alternative hypothesis and what outcomes lead researchers to reject the null and accept the alternative hypothesis. A nondirectional H_a means researchers have no prediction about the direction of difference. For nondirectional t tests (e.g., $H_a: \mu \neq \mu_0$), α is divided between both tails of the t distribution (top image in Figure 2-6). The critical value of t is found in the column for $\alpha/2$ because the table only represents one end of the

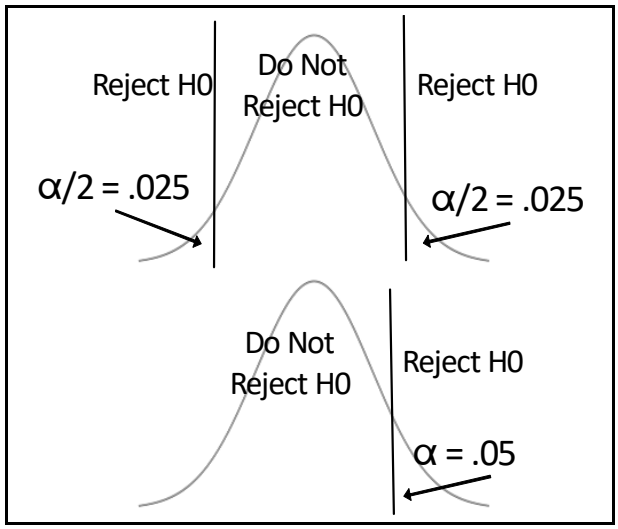


Figure 2-6. Nondirectional & Directional Hypothesis Testing.

distribution. The equivalent significance reported by SPSS equals alpha, the sum of the probabilities from both ends of the distribution. The F test is normally nondirectional. Recall that $F = t^2$, which translates to the negative half of the t distribution being folded over (i.e., it becomes positive) so that both ends of the t distribution are in the right tail of the F distribution. Non-directional tests are often referred to as two-tailed (see SPSS t-test output), which is accurate for t but *not* for F because both tails of the t distribution are in one tail (the positive tail) of F.

For directional t tests (e.g., H_a: $\mu > \mu_0$), all of α is in one tail of the distribution (bottom image in Figure 2-6). The critical value of t in the table is in the column equal to α and the nondirectional significance reported by SPSS must be divided by two to obtain the one-tailed probability. Determining critical values of F for directional tests can be confusing but the significance reported by SPSS is divided by two. In essence to obtain critical values of F for directional tests from a table, keep in mind that both ends of the t distribution are in the positive tail of the F distribution. It also helps to keep in mind that $F = t^2$.

The preceding tests allow researchers to determine whether there is a relationship between a categorical predictor variable and a numerical dependent variable. Researchers often want to examine relationships for numerical predictors as well. The appropriate tests are described next. Appendix 2-1 presents sample data to practice some of the tests.

APPENDIX 2-1
PRACTICE DATA FOR INDEPENDENT GROUPS T-TEST AND F-TEST

The data below can be used to practice the various calculations for the preceding tests. The columns are:

- EFF difference between population means: 0, 1, or 2 units
- NJ number of observations per group
- M# mean for group 1 or 2
- SD# standard deviation for group 1 or 2
- VARP pooled variance
- SE standard error of difference between means
- t observed t statistic; $F = t^2$ to practice F calculation
- p probability of t as extreme as observed t or more extreme
- TCRIT critical value of t given df; $F_{critical} = t^2_{Critical}$
- SIG 1 if observed t significant, 0 if not significant

You can also calculate F from the following data. It should equal t^2.

EFF	NJ	M1	SD1	M2	SD2	VARP	SE	t	p	tcrit	SIG
2	9	6.000	2.062	4.000	1.581	3.375	.87	2.309	.035	2.120	1
2	18	6.556	1.617	3.722	1.742	2.825	.56	5.057	.000	2.032	1
0	6	3.500	1.517	4.000	2.191	3.550	1.09	-.460	.656	2.228	0
2	20	5.350	1.631	3.200	1.795	2.941	.54	3.965	.000	2.024	1
2	13	5.692	1.750	4.692	1.932	3.397	.72	1.383	.179	2.064	0
2	22	6.364	1.840	3.909	1.743	3.212	.54	4.542	.000	2.018	1
0	10	4.400	1.174	2.900	1.370	1.628	.57	2.629	.017	2.101	1
1	12	5.583	1.564	3.000	1.859	2.951	.70	3.684	.001	2.074	1
1	9	5.000	2.236	3.889	1.269	3.306	.86	1.296	.213	2.120	0
2	15	5.333	1.447	3.867	1.727	2.538	.58	2.521	.018	2.048	1
0	15	4.067	1.624	3.933	1.223	2.067	.52	.254	.801	2.048	0
2	6	5.167	1.472	3.500	1.761	2.633	.94	1.779	.106	2.228	0
2	6	5.833	1.472	4.000	1.789	2.683	.95	1.938	.081	2.228	0
1	22	4.045	1.618	4.182	1.967	3.244	.54	-.251	.803	2.018	0
2	16	5.563	2.128	3.125	1.784	3.856	.69	3.511	.001	2.042	1
0	15	2.867	1.552	4.333	1.877	2.967	.63	-2.332	.027	2.048	1
0	5	4.800	1.643	3.800	2.168	3.700	1.22	.822	.435	2.306	0
1	18	4.778	1.592	3.778	2.016	3.301	.61	1.651	.108	2.032	0
1	10	4.100	1.197	4.400	1.430	1.739	.59	-.509	.617	2.101	0
2	10	5.700	1.160	3.300	1.567	1.900	.62	3.893	.001	2.101	1
1	17	4.765	1.985	5.000	1.323	2.846	.58	-.407	.687	2.037	0
2	13	5.692	1.653	3.692	1.316	2.231	.59	3.414	.002	2.064	1
0	11	3.636	1.567	3.636	2.248	3.755	.83	.000	1.000	2.086	0
1	20	4.900	1.410	3.750	1.552	2.199	.47	2.453	.019	2.024	1
1	15	5.133	1.506	3.867	2.100	3.338	.67	1.899	.068	2.048	0
2	20	5.600	1.465	3.850	1.927	2.930	.54	3.233	.003	2.024	1
1	19	5.211	1.903	3.421	1.774	3.383	.60	2.999	.005	2.028	1
2	17	6.059	1.638	4.118	1.536	2.522	.54	3.564	.001	2.037	1
1	7	6.000	1.826	3.714	1.380	2.619	.87	2.642	.021	2.179	1
0	9	4.667	1.658	3.778	1.716	2.847	.80	1.117	.280	2.120	0
2	12	6.167	1.193	3.667	1.723	2.197	.61	4.131	.000	2.074	1
1	23	5.304	2.032	3.522	2.020	4.105	.60	2.984	.005	2.015	1
1	9	4.222	1.093	4.444	1.333	1.486	.57	-.387	.704	2.120	0
1	13	5.077	1.706	4.000	1.472	2.538	.62	1.723	.098	2.064	0
0	9	3.444	1.424	3.889	1.269	1.819	.64	-.699	.495	2.120	0
1	12	5.083	1.165	3.167	1.801	2.299	.62	3.096	.005	2.074	1
0	9	4.111	1.900	3.556	1.130	2.444	.74	.754	.462	2.120	0
2	12	6.083	1.311	4.917	1.621	2.174	.60	1.938	.066	2.074	0
0	6	4.333	2.338	3.667	1.366	3.667	1.11	.603	.560	2.228	0
2	22	6.091	1.509	3.909	1.875	2.896	.51	4.252	.000	2.018	1
0	16	4.188	1.424	3.438	1.632	2.346	.54	1.385	.176	2.042	0
0	5	4.200	2.168	3.000	1.581	3.600	1.20	1.000	.347	2.306	0
0	17	3.588	1.622	3.529	1.505	2.449	.54	.110	.913	2.037	0
0	11	4.455	2.252	4.182	1.401	3.518	.80	.341	.737	2.086	0
1	23	4.565	1.805	3.435	1.903	3.439	.55	2.067	.045	2.015	1
0	5	3.400	1.140	4.400	1.517	1.800	.85	-1.179	.272	2.306	0
1	21	5.190	1.940	4.095	1.300	2.726	.51	2.149	.038	2.021	1
0	20	4.000	1.835	3.250	1.618	2.993	.55	1.371	.178	2.024	0

CHAPTER 3 - NUMERICAL PREDICTORS

The independent groups t-test concerns differences between means; that is, a relationship between a categorical predictor (group) and a numerical dependent variable. But researchers also study numerical predictors to determine whether scores on a numerical predictor X are associated with an increase or a decrease in scores on Y. That is, do the variables correlate rather than do their means differ. This analysis requires pairs of numerical scores corresponding to X and Y. Variables can correlate whether or not there are differences between means or it would be meaningless to compare means (e.g., IQ as a predictor of GPA).

The following commands enter 9 pairs of X, Y scores to produce the first two columns in the listing below and construct the graph of the relationship between x and y shown in Figure 3-1. Menu commands for the graph are: Graph | Legacy Dialogs | Scatter/Dot | Simple Scatter | Define | y → y axis | x → x axis | OK. A double-click on the resulting graph opens the Chart Editor, which can be used to modify the default graph.

```
DATA LIST FREE /x y .
BEGIN DATA
7  5    3  2    4  3    3  5    4  3    6  7    2  3    5  6    2  2
END DATA.

GRAPH SCATTER(BIVARIATE) x WITH y.
```

The vertical solid line in Figure 3-1 is $\bar{x} = 4.0$ and the horizontal solid line is $\bar{y} = 4.0$. The means create four quadrants. The cross-product $(x-\bar{x})\times(y-\bar{y})$ is positive for observations in the lower-left or upper-right quadrants, and negative for the upper-left or lower-right quadrants. Therefore, the sum of cross-products (SCP) will be positive when most CPs are positive, negative when most CPs are negative, and about 0 when CPs are scattered across all four quadrants. SCP is used to calculate correlation and regression statistics. The following SPSS commands compute SS_x, SS_y, and SCP.

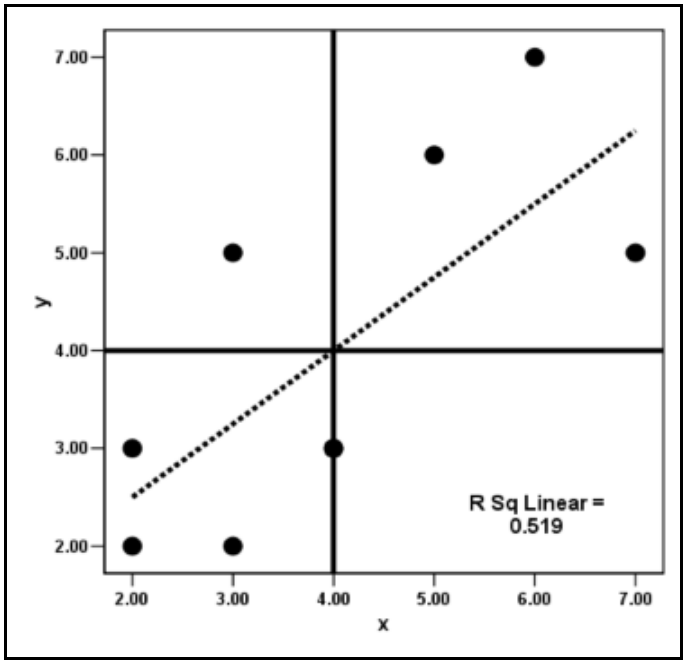


Figure 3-1. Scattergram of y as a function of x.

```
COMPUTE xdev = x - 4.0.        (x-x̄)
COMPUTE ydev = y - 4.0.        (y-ȳ)
COMPUTE xdev2 = xdev**2.       (x-x̄)²
COMPUTE ydev2 = ydev**2.       (y-ȳ)²
COMPUTE cp = xdev*ydev.        (x-x̄) (y-ȳ)
```

```
LIST x y xdev ydev cp.
     x       y    xdev    ydev      cp
  7.00    5.00    3.00    1.00    3.00
  3.00    2.00   -1.00   -2.00    2.00
  4.00    3.00     .00   -1.00     .00
  3.00    5.00   -1.00    1.00   -1.00
  4.00    3.00     .00   -1.00     .00
  6.00    7.00    2.00    3.00    6.00
  2.00    3.00   -2.00   -1.00    2.00
  5.00    6.00    1.00    2.00    2.00
  2.00    2.00   -2.00   -2.00    4.00

 DESCR xdev2 ydev2 cp /STAT = SUM.
                 N Sum
 xdev2           9 24.00      = SS_x
 ydev2           9 26.00      = SS_y
 cp              9 18.00      = SCP = ∑(x−x̄) (y−ȳ)
```

$$r = \frac{SCP}{\sqrt{SS_x \times SS_y}}$$

Box 3-1.

These quantities are used to calculate the correlation coefficient r, as shown in Box 3-1. The value of r varies between -1 and +1; that is, $-1 \le r \le +1$. For the present data, r = 18.0/SQRT{24.0×26.0} = .7206. The following command calculates r and its significance, described shortly. Figure 3-1 reported $r^2 = .519 = .7206^2$.

```
CORRELATION x y.
                     x     y
 x Pearson           1     .721      r = 18.0/SQRT{24×26}
   Sig. (2-tailed)   .     .029      Sig < .05, therefore reject H0 (test shown later)
```

$$b_1 = \frac{SCP}{SS_x} \qquad b_0 = \bar{y} - b_1 \times \bar{x}$$

Box 3-2.

The relationship between X and Y can also be conceptualized in terms of the dashed regression line shown in Figure 3-1. Points on the line are predicted values for Y given different values for X and the linear relationship between X and Y. The best-fit line minimizes the difference between predicted and observed values. The formula for a line is determined by its slope and intercept. The slope is the amount of change in Y per unit change in X, and the intercept is the predicted value when X = 0. Box 3-2 shows the formula for the slope and intercept of the best-fit line; the slope uses SCP in the numerator, like the formula for r.

b_1 = 18.0 / 24.0 = .75 **b_0 = 4.0 − .75 × 4.0 = 1.0**

The best-fit regression line is: $\hat{y} = b_0 + b_1 x = 1.0 + .75x$, where $\hat{y}$ is the predicted value (i.e., points on the line for each subject). This is the formula for the dashed line in Figure 3-1. The following SPSS commands produce the best-fit regression line and other statistics. The intercept and slope appear in the Unstandardized Coefficients column. The SAVE command adds predicted $\hat{y}$s (*prd*) and residual scores (*res*) to the dataset, *res* being the deviation of observed from predicted scores (i.e., $y - \hat{y}$). These appear in columns three and four in the listing below.

```
REGRESSION DEPENDENT = y /ENTER x /SAVE PRED(prd) RESID(res).

 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .721(a) .519     .451            1.33631

 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 13.500         1  13.500      7.560 .029(a)
       Residual   12.500         7  1.786
       Total      26.000         8

 Model            Unstandardized             Standardized    t     Sig.
                  Coefficients               Coefficients
                  B               Std. Error Beta
 1     (Constant) 1.000           1.179                      .849  .424
       x           .750            .273       .721           2.750 .029

Residuals Statistics(a)
                 Mean    Std. Deviation N
 Predicted Value 4.0000  1.29904        9       SS_Prd = (9-1)1.29904² = 13.50
 Residual         .0000  1.25000        9       SS_Res = (9-1)1.25000² = 12.50

LIST x, y, prd, res.
     x      y         prd         res
  7.00   5.00     6.25000    -1.25000          ŷ₁ = 1.0 + .75×7.0 = 6.25
  3.00   2.00     3.25000    -1.25000          Y₁ - ŷ₁' = 5.0 - 6.25 = -1.25
  4.00   3.00     4.00000    -1.00000
  3.00   5.00     3.25000     1.75000
  4.00   3.00     4.00000    -1.00000
  6.00   7.00     5.50000     1.50000
  2.00   3.00     2.50000      .50000
  5.00   6.00     4.75000     1.25000
  2.00   2.00     2.50000     -.50000
```

The predicted scores $\hat{y}$ are obtained by entering values for X into the prediction equation, and the residual scores by subtracting predicted from observed scores (i.e., $y - \hat{y}$). These operations are shown above for the first subject. Given predicted and residual scores it is possible to calculate SSs from the standard deviations shown in the Residuals Statistics section of the output. As shown above, $SS_{Predicted} = 13.50 = \Sigma(\hat{y}-\bar{y})^2$ and $SS_{Residual} = 12.50 = \Sigma(y-\hat{y})^2$. $SS_{Predicted}$ is also referred to as $SS_{Regression}$, and $SS_{Residual}$ as $SS_{Error.}$ These quantities appear in the Sum of Squares column in the output.

Notice that $SS_y = 26.0 = SS_{Total}$ has been partitioned (divided) into what can be predicted using X, $SS_{Regression} = 13.5$, and what cannot be predicted, $SS_{Residual} = 12.5$. Figure 3-2 represents the partitioning of SS_{Total} as a Venn diagram. The circle Y represents SS_{Total} and the overlap represents how much of the total is predicted by X. As a proportion, the predictor X accounts for $SS_{\hat{y}}/SS_y = 13.5/26.0 = .519$ of the total variability in y. This quantity is $r^2 = .721^2 = .520$. R and R^2 appear in the output above and reflect the strength of the relationship between X and Y in terms of r (between -1 and +1) or, more precisely, r^2, the proportion of variability in Y

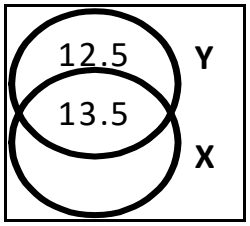


Figure 3-2.

predicted by X.

The result of the present analysis is a best-fit regression line. The idea of best-fit is analogous to the mean being the best measure of central tendency. Recall that the sum of the deviations of scores about the mean equals 0; that is, $\sum(y-\bar{y}) = 0$. In regression, the sum of the deviations of observed scores about the predicted values equals 0; that is, $\Sigma(y-\hat{y}) = 0$. Note in the regression analysis that the mean residual score is 0 because the sum of the residual scores is 0. Deviations above and below the line balance. A second way to think of the mean as a good measure of average is that $SS_y = \Sigma(y-\bar{y})^2$ = a minimum. No other value comes as close to all the scores in terms of squared deviations. Similarly, the best-fit regression line minimizes $SS_{Residual}$; that is, $\Sigma(y-\hat{y})^2$ = a minimum. No other line comes as close to the observed scores in terms of squared deviations. The claim of best-fit is specific to a straight line fit. It is possible that an equation for a curve could do a better job predicting the scores. Nonlinear regression is examined in a later chapter.

Significance of Regression and Correlation

In addition to the strength of the relationship, researchers are interested in its significance because a strong relationship can be not significant (i.e., can occur by chance) and a weak relationship can be significant. For a single predictor X, there are several equivalent tests that the correlation in the population, denoted by the Greek letter rho (ρ), is 0 (see Box 3-3). The various tests reflect different relationships when there is more than one predictor as demonstrated in later chapters.

$$H_0: \rho = 0$$
$$H_a: \rho \neq 0 \text{ or } \rho > 0 \text{ or } \rho < 0$$
$$t_r = \frac{r - \rho_0}{\sqrt{\frac{1-r^2}{n-2}}} \quad df = n-2$$

Box 3-3.

First, the correlation coefficient r can be tested for significance using a t-test. The formula is shown in Box 3-3 and the calculations follow. For the present sample, the null hypothesis that the population correlation coefficient (ρ) equals 0 is rejected.

$t = (r-0)/SQRT\{(1-r^2)/(n-2)\} = .721/SQRT\{(1-.721^2)/(9-2)\} = .721/.2619 = 2.75$

$df = n - 2 = 9 - 2 = 7$

Nondirectional, $\alpha = .05$, $t_{Critical} = \pm 2.365$

$t_{Observed} > +t_{Critical}$ Reject H_0, Accept H_a

The correlation coefficient r can also be tested for significance using an F test based on $SS_{Regression}$ and $SS_{Residual}$ or on r^2 and $(1-r^2)$, as shown in Box 3-4, where p = the number of predictors, 1 in the present case. Calculations appear below and the result appears on the Regression line of the output.

$$F = \frac{MS_{Regression}}{MS_{Residual}} = \frac{\frac{SS_{Regression}}{p}}{\frac{SS_{Residual}}{n-p-1}} = \frac{\frac{r^2}{p}}{\frac{1-r^2}{n-p-1}}$$

Box 3-4.

$F = (r^2/p)/\{(1-r^2)/(n-p-1)\} = (.721^2/1)/\{(1 - .721^2)/(9-1-1)\}$

$= .5198/.0686 = 7.578 = 2.75^2 = t_r^2$

$df = 1, 7$ $F_{Critical} = 5.59 = t^2_{Critical}$ Reject H_0: $\rho^2 = 0$, Accept H_a

Third, the regression coefficient b_1 can be tested for significance using the t-test shown in Box 3-5. The result appears on the Coefficient line for X.

$t_{b1} = (b_1-0)/SQRT\{MS_{Res}/SS_x\} = .750/SQRT\{1.786/24\}$

$= .750/.2728 = 2.75 = t_r$

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0 \text{ or } \beta_1 > 0 \text{ or } \beta_1 < 0$

$$t_{b_1} = \frac{b_1 - \beta_1}{\sqrt{\frac{MSE}{SS_x}}}$$

Box 3-5.

Regression can also be conceptualized in terms of relationships among the original X and Y variables and the new variables, $\hat{y}$ and $y - \hat{y}$. The following SPSS commands produce the rs among the four variables and the relationships are represented in Figure 3-3. Let's examine the six correlations. The r between x and $\hat{y}$ is 1 and between x and $y - \hat{y}$ is 0 because any variability in y related to x goes to $\hat{y}$, the points on the best fit line, and none goes to $y - \hat{y}$. By the same reasoning, the r between $\hat{y}$ and $y - \hat{y}$ is also 0. Because x and $\hat{y}$ correlated perfectly, the r between y and $\hat{y}$ equals the r between x and y, our original correlation coefficient. The variability in y has been divided between $\hat{y}$ and $y - \hat{y}$; the proportion accounted for by $\hat{y}$ equals $r^2 = .721^2 = .52$, and the proportion accounted for by $y - \hat{y}$ equals $1- r^2 = 1 - .52 = .48$. The square root of $1 - r^2 = .693$, as shown below. Observe that $.721^2 + .693^2 = 1$ because $\hat{y}$ and $y - \hat{y}$ account for all of the variability in y given $y = \hat{y} + (y - \hat{y})$. The fact that the correlation between x and $y - \hat{y}$ is 0 will be important for understanding aspects of multiple regression. In general, residual scores from a regression will be uncorrelated with any predictors in the equation.

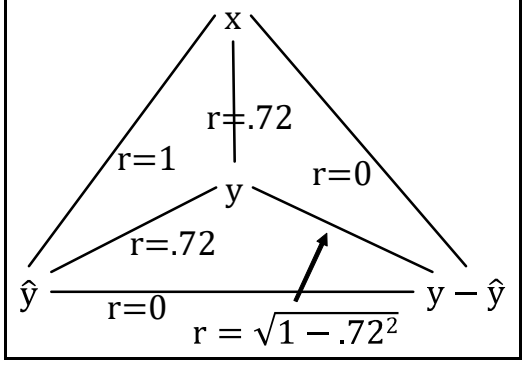


Figure 3-3. Partition of y.

```
VARIABLE LABEL prd '' res ''.
CORRELATION y x prd res /STATISTICS.
```

	Mean	Std. Deviation	N	
y	4.0000	1.80278	9	
x	4.0000	1.73205	9	
prd	4.0000000	1.29903811	9	$M_{Pred} = 4.0 = \bar{y}$
res	.0000000	1.25000000	9	$M_{Res} = 0.0$

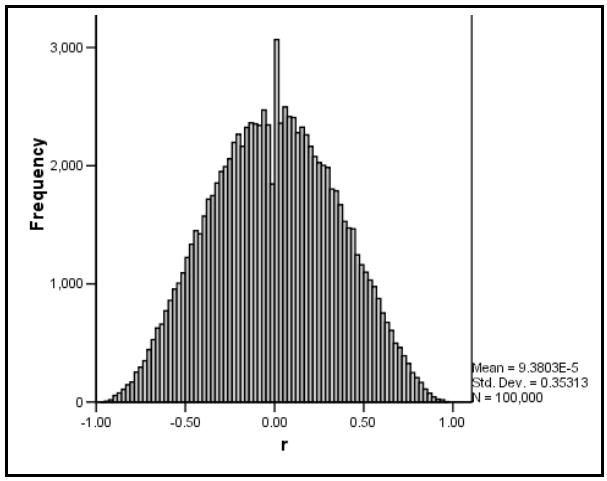


Figure 3-4. Sampling distribution of r.

	y	x	prd
x	.721		
prd	.721	1.000	
res	.693	.000	.000

Sampling Distributions for Test Statistics

The logic of the three tests (t_r, t_{b1}, F) can be illustrated with a simulation. Figure 3-4 shows the frequency distribution of 100,000 *r*s calculated for our samples. As expected since the population correlation coefficient ρ is 0, the mean of the *r*s is .00009 ≈ 0. That is, EV(r) = ρ. But there is variability about this value. Indeed r is occasionally close to -1 and +1, indicating a perfect relationship between X and Y. To determine whether a given *r* is significant, we calculate a *t* test.

Figure 3-5 shows the distribution of the 100,000 *t*s calculated using the formula for either *r* or b_1. Again as expected, the mean of all the *t*s is equal to 0, but there is variability about this value. It turns out that exactly 5% of the *t*s were ≤-2.365 OR ≥+2.365, as expected given H_0 is true. Our sample was one of the 5% that (wrongly) produced a significant *t*. That is, our conclusion was a Type I error; rejecting a true H_0. A simulation of F would show the same result.

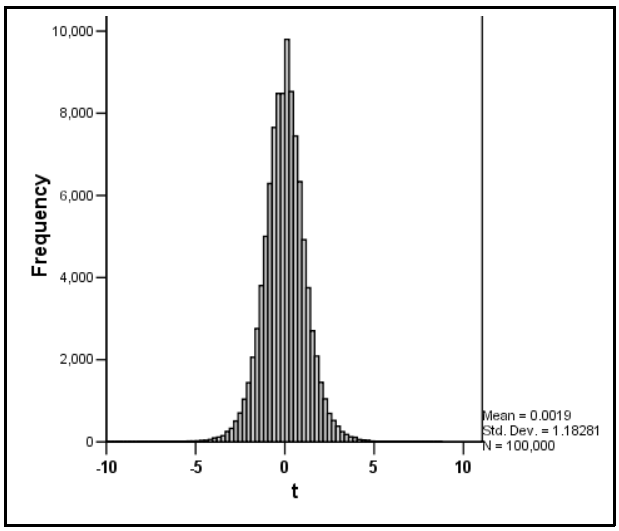


Figure 3-5. Sampling distribution of t.

Dependent t-test for Difference Between Means

The independent *t*-test compared means from two unrelated samples. That is, scores in the two groups are not expected to correlate with one another. A dependent or paired-difference *t*-test compares two means when scores in the two sets are expected to correlate, either because observations come from the same subjects (e.g., pre versus post treatment), or because of a pre-existing relationship between subjects (e.g., twins, animals from same litter, matched subjects). The proper test is a modified single sample t-test of difference scores obtained by subtracting individual scores in one group from those in the other group (see Box 3-6). If the null hypothesis is true, the mean of the difference scores should equal 0.

Pre	Post	D=Post−Pre
7	7	0
3	4	+1
4	5	+1
3	7	+4
4	5	+1
6	9	+3
2	5	+3
5	8	+3
2	4	+2
		ΣD = 18.0

H_0: $\mu_D = 0$

H_a: $\mu_D \neq 0$ or $\mu_D > 0$ or $\mu_D < 0$

$t = (M_D - 0) / (s_D/\sqrt{n_D})$

$= (2.0 - 0) / (1.3228/\sqrt{9})$

$= 2.0 / .4410$

$= 4.535$ $\quad df = n_D - 1 = 8$

$M_D = 18/9 = 2.00$ $\quad s_D = 1.3228$

$$t_{Dependent} = \frac{\bar{D} - 0}{\frac{s_D}{\sqrt{n_D}}}$$

Box 3-6.

```
DATA LIST FREE / pre post.
BEGIN DATA
 7 7    3 4    4 5    3 7    4 5    6 9    2 5    5 8    2 4
END DATA.

COMPUTE diff = post - pre.
TTEST /TESTVALUE = 0 /VARI = diff.
```

```
      N Mean  Std. Deviation Std. Error Mean
 diff 9 2.000 1.32288        .44096

      Test Value = 0
      t         df Sig.       Mean
                   (2-tailed) Difference
 diff 4.536      8  .002      2.0000
```

SPSS can do the paired-difference test without computing difference scores. The modified TTEST command is:

```
TTEST PAIR pre post.
```

```
           Mean   N Std. Deviation Std. Error Mean

 Pair Pre  4.0000 9 1.73205        .57735
      Post 6.0000 9 1.80278        .60093

                N Correlation Sig.
 Pair 1 y1 & y2 9 .721        .029

                Mean          Std.      Std. Error  t      df Sig.
                              Deviation Mean                  (2-tailed)
 Pair 1 Pre-Post 2.0000       1.32288   .44096      4.536  8  .002
```

The expectation for the dependent or paired-difference *t* is that scores for the two samples will correlate. The standard deviation of the difference scores depends on r, as well as on the original variability in the two sets of scores. Specifically,

$$s_D^2 = s_1^2 + s_2^2 - \mathbf{2 \times r_{12}} \times s_1 \times s_2$$

$$= 1.73205^2 + 1.80278^2 \mathbf{- 2 \times .721} \times 1.73205 \times 1.80278$$

$$= 1.7474$$

$$s_D = SQRT\{1.7474\} = 1.322$$

To the extent that the scores do correlate, variability in the difference scores will be less than the variability in the original scores, resulting in a smaller denominator for the dependent t than for the independent t. In the present case:

$SE_D = .44096 \quad < \quad SE_{\bar{y}1-\bar{y}2} = SQRT\{s_p^2(1/n_1+1/n_2)\} = .8333$

Since the numerators for the two *t*s are identical because $\bar{y}_D = \bar{y}_1 - \bar{y}_2$, the dependent *t* will *generally* be larger than the independent *t*. At the same time, the df for the dependent *t* will be $n_D - 1$, smaller than $df = n_1 + n_2 - 2$ for the independent *t*. A smaller df means that the critical value will be somewhat larger for the dependent *t*. The gain from the smaller SE is expected to offset the loss of degrees of freedom, especially when n is large. The independent groups analyses from earlier are repeated below. As expected given the SE is larger, the independent t is smaller, and its p value larger.

```
*Independent Scores Analyses (t and F).
DATA LIST FREE / grp y.
BEGIN DATA
 1 7   1 3   1 4   1 3   1 4   1 6   1 2   1 5   1 2
 2 7   2 4   2 5   2 7   2 5   2 9   2 5   2 8   2 4
END DATA.

TTEST /GROUP = grp /VARI = y.

 y 1.00 9 4.0000 1.73205        .57735
   2.00 9 6.0000 1.80278        .60093

                 t-test for Equality of Means
                 t              df     Sig.       Mean       Std. Error
                                       (2-tailed) Difference Difference
 y Equal variances-2.400          16     .029      -3.76659    .83333

MANOVA y BY grp(1 2).
 Source of Variation          SS       DF         MS          F  Sig of F
 WITHIN CELLS              50.00       16       3.13
 grp                       18.00        1      18.00       5.76       .029
 (Total)                   68.00       17       2.94
```

Regression and Difference Between Independent Means

It might appear that tests for the difference between means and for regression are unrelated, but they are actually equivalent in some cases, such as the independent groups design when $df_{Numerator} = 2 - 1 = 1$. The following regression analysis demonstrates this. The predictor is a categorical variable *grp* that represents the two groups (*grp* = 1 or 2) and the dependent variable is Y. Observe the many parallels between the regression output and the independent groups t-test and corresponding ANOVA. Later material on ANOVA for more complex designs (e.g., three or more groups) will explore this equivalence in greater depth.

```
REGRE /DEP = y /ENTER grp /SAVE PRED(prd2) RESI(res2).
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.514(a)	.265	.219	1.76777

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18.000	1	18.000	**5.760**	**.029**(a)
	Residual	50.000	16	3.125		
	Total	68.000	17			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.000	1.318		1.518	.149
	grp	2.000	.833	.514	**2.400**	**.029**

Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4.0000	6.0000	5.0000	1.02899	18
Residual	-2.00000	3.00000	.00000	1.71499	18

```
VARIABLE LABEL prd2 '' res2 ''.
LIST.
```

grp	y	prd2	res2
1.00	7.00	4.00000	3.00000
1.00	3.00	4.00000	-1.00000
1.00	4.00	4.00000	.00000
1.00	3.00	4.00000	-1.00000
1.00	4.00	4.00000	.00000
1.00	6.00	4.00000	2.00000
1.00	2.00	4.00000	-2.00000
1.00	5.00	4.00000	1.00000
1.00	2.00	4.00000	-2.00000
2.00	7.00	6.00000	1.00000
2.00	4.00	6.00000	-2.00000
2.00	5.00	6.00000	-1.00000
2.00	7.00	6.00000	1.00000
2.00	5.00	6.00000	-1.00000
2.00	9.00	6.00000	3.00000
2.00	5.00	6.00000	-1.00000
2.00	8.00	6.00000	2.00000
2.00	4.00	6.00000	-2.00000

Appendix 3-1 presents data with which to practice correlation and regression calculations. Appendix 3-2 presents some research scenarios to practice deciding about the appropriate test for different designs that have been discussed.

APPENDIX 3-1

CORRELATION & REGRESSION WITH SINGLE PREDICTOR

CALCULATION EXERCISE

Each line below shows statistics from 20 samples (s = 1 to 20) of x and y scores. N is the number of observations in a sample. Use the formula sheet and the data below to calculate the correlation, the best fit equation, and the strength and significance of the linear relationship. The raw data for each sample is shown next and can be pasted into SPSS to do the analyses corresponding to your calculations. Omit s, n, and the periods when entering the data. Here are the commands for sample 5.

```
DATA LIST FREE / x y.
BEGIN DATA
49 53 48 54 57 67 62 64 64 51 62 62
END DATA.
```

S	N	Mx	My	SDx	SDy	SCP	r	b0	b1	t	SDreg	SDres	Fcrit	F	Sig
1	9	60.111	71.778	14.5640	12.5576	243.2222	.166	63.162	.143	.446	2.0875	12.3829	5.59	.199	.669
2	11	58.364	59.455	10.2983	7.3398	−51.8182	−.069	62.306	−.049	−.206	.5626	8.1868	5.12	.042	.841
3	7	67.714	62.000	14.5569	12.2610	−63.0000	−.059	65.355	−.050	−.132	.6247	10.6000	6.61	.017	.900
4	7	57.000	63.286	9.3986	10.1442	−213.0000	−.372	86.193	−.402	−.897	3.2711	8.1534	6.61	.805	.411
5	6	57.000	58.500	6.9857	6.6558	77.0000	.331	40.512	.316	.702	1.7428	4.9649	7.71	.493	.521
6	9	53.222	63.556	10.8947	5.8119	375.8889	.742	42.487	.396	2.929	4.3128	3.8959	5.59	8.578	.022
7	7	49.571	64.714	12.6736	16.5400	994.1429	.790	13.578	1.032	2.885	11.3222	8.7742	6.61	8.326	.034
8	10	51.900	68.400	14.8283	7.3212	359.4000	.368	58.974	.182	1.119	2.8564	7.2209	5.32	1.252	.296
9	6	51.833	74.167	5.6362	13.8912	90.1667	.230	44.742	.568	.473	2.5295	10.6867	7.71	.224	.661
10	12	57.083	67.750	11.7509	8.1589	90.2500	.086	64.358	.059	.272	.8187	9.5321	4.96	.074	.791
11	11	57.273	65.727	12.0340	9.0121	35.8182	.033	64.311	.025	.099	.3328	10.0704	5.12	.010	.923
12	6	57.333	69.167	9.9130	11.5658	394.6667	.688	23.113	.803	1.899	6.2950	6.6315	7.71	3.604	.130
13	6	52.000	63.833	9.3381	16.3758	−56.0000	−.073	70.512	−.128	−.147	.9482	12.9114	7.71	.022	.890
14	11	53.545	66.818	9.5117	11.3386	319.0909	.296	47.933	.353	.929	3.7507	12.1094	5.12	.863	.377
15	12	57.167	71.083	10.6757	8.9388	162.8333	.155	63.658	.130	.497	1.6260	10.3548	4.96	.247	.630
16	12	49.667	61.667	12.6874	16.9670	1566.6667	.662	17.722	.885	2.790	13.1633	14.9185	4.96	7.785	.019
17	10	54.900	59.500	7.9366	15.6223	−116.5000	−.104	70.782	−.206	−.297	1.7299	16.4794	5.32	.088	.774
18	8	58.125	67.750	11.5194	8.6644	58.2500	.083	64.105	.063	.205	.6757	8.0766	5.99	.042	.844
19	12	52.333	63.000	16.0189	8.6655	592.0000	.388	52.024	.210	1.330	3.9396	9.3664	4.96	1.769	.213
20	12	55.083	62.000	12.9857	11.2412	119.0000	.074	58.466	.064	.235	.9769	13.1452	4.96	.055	.819

s	n	x1	y1	x2	y2	x3	y3	x4	y4	x5	y5	x6	y6	x7	y7	x8	y8	x9	y9	x10	y10	x11	y11	x12	y12
1	9	41	66	74	45	65	77	59	74	45	63	77	79	70	81	70	88	40	73	.	.	.	.	.	.
2	11	42	67	67	60	59	65	62	60	61	65	63	55	54	52	77	59	41	60	60	68	56	43	.	.
3	7	66	65	73	54	89	57	40	63	67	49	67	59	72	87	.	.	.	.	.	.	.	.	.	.
4	7	50	52	64	52	72	64	54	77	48	72	63	56	48	70	.	.	.	.	.	.	.	.	.	.
5	6	49	53	48	54	57	67	62	64	64	51	62	62	.	.	.	.	.	.	.	.	.	.	.	.
6	9	51	64	61	64	48	68	63	63	28	49	60	63	60	68	50	66	58	67	.	.	.	.	.	.
7	7	58	70	39	70	63	77	39	37	35	47	47	69	66	83	.	.	.	.	.	.	.	.	.	.
8	10	71	73	41	55	37	64	66	82	73	66	30	75	43	64	57	70	55	69	46	66	.	.	.	.
9	6	58	63	49	94	47	69	54	77	58	85	45	57	.	.	.	.	.	.	.	.	.	.	.	.
10	12	60	67	47	66	63	68	71	67	72	86	43	68	38	65	53	74	44	71	66	50	68	68	60	63
11	11	68	76	33	68	67	79	64	65	53	76	74	52	60	63	41	55	59	64	59	56	52	69	.	.
12	6	65	78	45	66	62	66	48	49	70	80	54	76	.	.	.	.	.	.	.	.	.	.	.	.
13	6	68	60	51	95	51	65	43	59	56	47	43	57	.	.	.	.	.	.	.	.	.	.	.	.
14	11	55	73	34	55	60	53	59	61	55	77	44	49	63	79	60	61	45	81	65	74	49	72	.	.
15	12	51	73	59	59	72	79	53	57	38	80	67	72	50	69	62	81	63	85	62	69	41	64	68	65
16	12	38	41	59	79	31	58	67	91	51	63	39	66	45	52	68	69	32	33	52	77	57	44	57	67
17	10	61	89	51	55	53	60	51	33	60	45	37	74	66	49	55	65	55	59	60	66	.	.	.	.
18	8	65	64	70	66	55	66	41	82	55	63	74	80	60	64	45	57	.	.	.	.	.	.	.	.
19	12	55	77	54	55	72	76	58	69	48	64	65	55	24	67	45	62	58	60	25	48	48	57	76	66
20	12	61	53	42	49	44	77	62	63	43	45	43	82	60	60	64	58	71	61	34	60	65	60	72	76

APPENDIX 3-2

WHICH TEST IS APPROPRIATE?

Given this review of several statistical tests, we can practice making decisions about what tests are appropriate for different studies. Here are some brief descriptions of studies. Decide whether they require a single sample test about means, an independent groups t-test, a paired-difference t-test, or a correlation and regression analysis.

1. Educational researchers wanted to identify who would struggle in school. They administered an IQ test to 100 students entering grade 10 and recorded their GPA at the end of the year.

2. To determine whether a novel training program was effective for new employees, human resource researchers assigned 15 new employees to the current training program, and 15 employees to the novel program. Work performance was measured over the first two weeks on the job.

3. A science practitioner conducts a pilot study on a new treatment for anxiety. Anxiety was measured in 24 high anxiety people after the two-week therapy program. Norms for the test indicate that high anxiety people score 45.0 on the test without any treatment.

4. To test the hypothesis that old people will perform better on a memory test with practice, cognitive researchers gave 36 people (average age of 75) two memory lists to learn. Each list contained 24 words. List one was presented on day one, followed by 10 sessions of practice learning lists of words. List two was presented on day two to see if performance differed from list one.

CHAPTER 4 - INTRODUCTION TO MULTIPLE REGRESSION

Most (all?) psychological phenomena depend on multiple factors that cause or are otherwise related to human behaviour and experience. Analyses that include only a single predictor do not accurately reflect this complexity. In a non-experimental study, researchers cannot know whether any relationship with the outcome variable represents the influence of a predictor or is due to other correlated variables. Teasing apart such confounds requires multiple regression analyses that include more than one predictor and control statistically for confounding variables. Multiple regression (MR) is also a way to potentially test whether the relationship between an outcome variable and a predictor is mediated by some other variable, sometimes represented as X ➡ M ➡ Y, where M is the variable hypothesized to mediate the relationship between X and Y. In addition, MR can test for interactions when the relationship between X and Y (that is, X ➡ Y) is influenced by a third variable. And inclusion of a second predictor that is unrelated to X can remove variability from $SS_{Residual}$, which gives a smaller denominator for hypothesis testing and measures of effect size.

The Multiple Regression Equation

Consider a study in which school psychologists hypothesized that more intelligent students use better memory strategies than less intelligent students when learning academic material. To test this hypothesis, nine students were administered an intelligence test, studied a chapter, and then were tested for use of memory strategies while studying and for recall of material from the chapter. Scores for intelligence (*int*), memory strategies (*mem*), and recall (*rec*) are entered into SPSS and then listed in the first three columns of the following table. The last two columns show results generated by later analyses.

```
DATA LIST FREE / int mem rec.
BEGIN DATA
96  25 47 111  27 51 118  23 41 112  21 37  88  14 25  92  12 30 112  25 41  89  18 32 106  23 46
END DATA.

LIST.
```

int	mem	rec	prdr.im	resr.im
96	25	47	45.72339	1.27661
111	27	51	48.32103	2.67897
118	23	41	41.65077	−.65077
112	21	37	38.69435	−1.69435
88	14	25	28.46649	−3.46649
92	12	30	25.11143	4.88857
112	25	41	45.08556	−4.08556
89	18	32	34.81783	−2.81783
106	23	46	42.12914	3.87086

$\hat{y}_1 = 9.605 - .04 \times 96 + 1.598 \times 25 = 45.72$

$y_1 - \hat{y}_1 = 47 - 45.72339 = 1.27661$

Simple correlations show that the two predictors are related to recall, but they also correlate with one another. The 3-D plot in Figure 4-1 shows the inter-relationships. The left-right horizontal axis represents Intelligence, the front-back horizontal axis is Memory Strategies, and the vertical axis is Recall. The following command created the graph and could be generated by menu.

```
GRAPH /SCATTERPLOT(XYZ)=mem WITH rec WITH int.
```

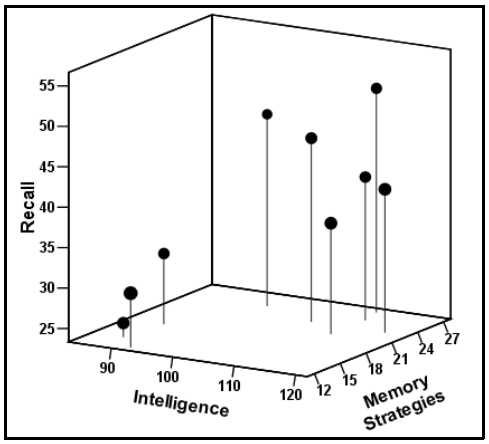


Figure 4-1. 3-D Plot

The lines from observed recall scores to the floor are called spikes and show that higher intelligence people used better strategies than lower intelligence people; that is *int* and *mem* are confounded. The relationship between recall and memory is confounded by intelligence, and the relationship between recall and intelligence is confounded by memory. Imagine looking down at the floor from above. Note that scores cluster in the lower-left and upper-right quadrants, indicating a positive correlation. The three relationships correspond to the following correlations.

```
CORR rec int mem /STAT.
```

```
       Mean   Std. Deviation   N
rec   38.89     8.623          9
int  102.67    11.456          9
mem   20.89     5.183          9

            rec   int
int         .630
mem         .923  .711
```

The correlation of .711 between the two predictors complicates interpretation of their relationship with recall. MR can determine the combined and unique contribution of the predictors, which serves to test the research hypothesis.

MR extends regression to designs in which multiple predictors are used to generate predicted scores for the dependent variable. In the present case, both intelligence and memory strategy would be included in a single regression equation. Box 4-1 shows the form of the equation, the calculation of the regression coefficient for X1 (i.e., $b_{y1.2}$ or simply b_1), and b_0. The coefficient $b_{y2.1}$ for X2 would involve a rearrangement of the correlation coefficients used to calculate $b_{y1.2}$.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

$$b_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2} \times \frac{s_y}{s_1}$$

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

Box 4-1. Multiple Regression Formula.

The equation in Box 4-1 defines a plane (a rigid surface with linear edges) that has a left-right tilt (slope), a front-back tilt (slope), and an elevation. In simple regression with a single predictor, b_{y1} was the tilt of a line and b_0 was the elevation of a line that minimized $SS_{Residual}$, deviations of observed from predicted values. In multiple regression with two predictors, the regression coefficients and elevation similarly minimize deviations of observed values from predicted values, the points on the plane when values for X1 and X2 are inserted in the equation.

The critical part of the MR coefficient formula is: r_{y1}-$r_{y2}r_{12}$. In the present example, all three correlations are positive. A positive value, $r_{y2}\times r_{12}$, will be subtracted from the positive r_{y1}, resulting in a value for $b_{y1.2}$ that will be less positive and perhaps even negative. The notation $b_{y1.2}$ indicates the slope for X1 when X2 is held constant (i.e., controlled or removed statistically). The ultimate value for $b_{y1.2}$ will depend on the magnitude and direction of the three correlations. In a study where r_{y1} and r_{y2} are positively related to y but the two predictors are *negatively* correlated, subtraction of a negative value, $r_{y2}\times r_{12}$, from r_{y1} will result in $b_{y1.2}$ becoming more positive, even if r_{y1} was 0 or negative. In this case, the relation of each predictor with the criterion variable is masked or hidden by their negative correlation with one another. One example would be if people low in *int* (reduces recall) studied more (increases recall) than people high in *int* (increases recall but is masked by studying less). Simple correlations of recall with *int* and study time would be weak, but the relationships become more positive in an MR analysis with both *int* and study time as predictors. Consider what happens in MR with other values for r_{y1}, r_{y2}, and r_{12}.

Here are the SPSS commands and output for predicting the dependent variable *rec* with *inc* and *mem* as predictors. The only difference in the command from the single predictor is the inclusion of both predictors after /ENTER. The /SAVE option is explained shortly.

```
REGRESS /DEPENDENT = rec /ENTER int mem /SAVE PRED(prdr.im) RES(resr.im).
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.923**(b)	**.853**	.804	3.822

R^2 = 507.258/594.889 = .853
$1-R^2$ = 87.631/594.889 = .147
$\surd$.147 = .384

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	**507.258**	2	253.629	17.366	.003(b)
	Residual	**87.631**	6	14.605		
	Total	**594.889**	8			

F = ($.923^2$/2)/{(1−$.923^2$)/(9−2−1)} = 17.26
SS_{Total} = (9−1)8.523^2 = 594.85

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	**9.605**	12.979		.740	.487
	int	**−.040**	.168	−.053	−.238	.820
	mem	**1.598**	.371	.960	4.311	.005

Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	25.11	48.32	38.89	**7.963**	9
Residual	−4.086	4.889	.000	**3.310**	9

SS_{Reg} = (9−1)7.963^2 = 507.27
SS_{Res} = (9−1)3.310^2 = 87.64

Box 4-2 shows equations to calculate the best-fit regression equation along with calculations for the regression coefficient (slope) for *int* controlling for *rec*. Analogous calculations for *mem* produce a coefficient of 1.598. Box 4-2 also shows the calculation of the intercept, $b_0 = 9.615$, which ensures that the prediction plane goes through the intersection of the three means and minimizes $SS_{Residual}$. Observed values for the intercept and regression coefficients are in the *Unstandardized Coefficients* column of the regression output. The best-fit regression equation is: $\hat{y} = 9.605 - .040 \times int + 1.598 \times mem$. The slopes for *int* and *mem* adjust or control for the correlation between predictors, as shown later. For now, we focus on the overall relationship between *rec* and the two predictors.

$$
\begin{aligned}
b_{ri.m} &= \frac{r_{ri} - r_{rm} \times r_{im}}{1 - r_{im}^2} \times \frac{s_r}{s_i} \\
&= \frac{.630 - .923 \times .711}{1 - .711^2} \times \frac{8.623}{11.456} \\
&= -.040 \\
b_0 &= \bar{y}_r - b_i \times \bar{x}_i - b_m \times \bar{x}_m \\
&= 38.89 - -.040 \times 102.67 - 1.598 \times 20.89 \\
&= 9.615
\end{aligned}
$$

Box 4-2. Regression Equation.

The equation can be used to calculate predicted recall scores for each participant based on their *int* and *mem* scores. The results are shown in the initial LIST as *prdr.im*. The equation defines a two dimensional plane with a tilt on the left-right axis and a separate tilt on the front-back axis, as shown in Figure 4-2. Although predicted scores could be plotted for just observed values of *int* and *mem*, the graph would reflect the equation poorly because there are few values on the predictors and they are confounded. Instead, a wide range of possible predictor values and their combinations is created and used to generate the best-fit plane in Figure 4-2 (see Appendix 4-1).

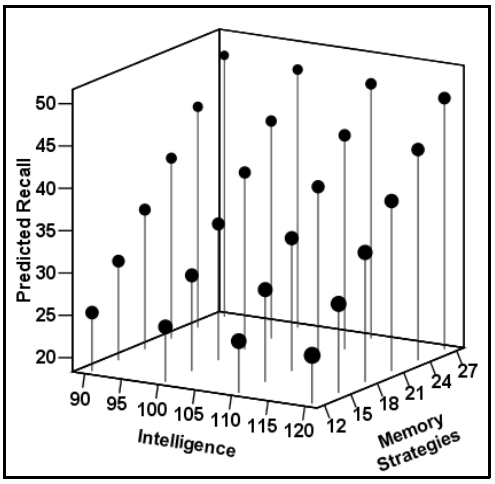


Figure 4-2. Best fit regression plane.

The MR equation, specifically $b_{rm.i} = 1.598$, and the front-back axis in the graph show a positive relationship between recall and memory strategies controlling for intelligence. But when memory strategies are controlled, the effect of intelligence disappears and even reverses slightly, $b_{ri.m} = -.04$, as seen in the slight downward trend in Figure 4-2. This is consistent with the hypothesis that memory strategies mediate the relationship between intelligence and recall; that is, the inferred model is INT ➡ MEM ➡ REC. Be cautious about this conclusion, however, because other possible models could account for the results.

The actual predicted (*prdr.im*) and residual (*resr.im*) scores generated by SPSS using the /SAVE option on the REGRESSION command are shown in the initial table of data, and descriptive statistics for *prdr.im* and *resr.im* appear in the *Residuals Statistics* section of the regression output. The mean of the residual scores is 0, indicating that deviations above the plane (i.e., positive values for $y - \hat{y}$) balance perfectly

deviations below the plane (i.e., negative values for $y - \hat{y}$). The plane goes through the middle of the data points. Also the mean for the predicted scores is equal to the mean for the recall scores (i.e., $\hat{y} = \bar{y}$), as was observed for predicted scores from the single predictor regression equation.

Sum of Squares Regression & Residual

Standard deviations for predicted and residual scores shown to the right of the *Residuals Statistics* output are used to calculate $SS_{Reg} = (n-1)s_{\hat{y}}^2 = 507.258$ and $SS_{Res} = (n-1)s_{(y-\hat{y})}^2 = 87.631$, which sum to the total variability in recall scores, $SS_{Total} = 594.889$. These values appear in the ANOVA Summary table. Our best-fit regression plane partitions SS_{Total} into SS_{Reg} and SS_{Res}. It is the best-fit plane because it goes through the centre of the scores and SS_{Res} is a minimum. No other values for the coefficients and intercept can produce as small a sum of squared deviations of observed from predicted values.

MR can be conceptualized in terms of Venn diagrams, as in Figure 4-3. The circle labelled Y represents the total variability or SS_y for the criterion variable (*rec* in our example). Circles labelled X1 and X2 represent the predictors, *int* and *mem*. Overlap of circles indicates shared variability. The overlap of predictors with Y divides SS_y into four regions, labelled a, b, c, and d. Area a represents variability in Y that cannot be predicted by X1, X2, or their combined effect; this is SS_{Res} in the MR analysis with both predictors. Area b+c+d represents variability predicted by X1, X2, or both; this is SS_{Reg} in the MR analysis. Area b+c represents what X1 alone can predict, that is, SS_{Reg} when only X1 is in the equation (i.e., r_{y1}^2). Area c+d represents what X2 by itself can predict, SS_{Reg} when X2 is in the equation alone (i.e., r_{y2}^2).

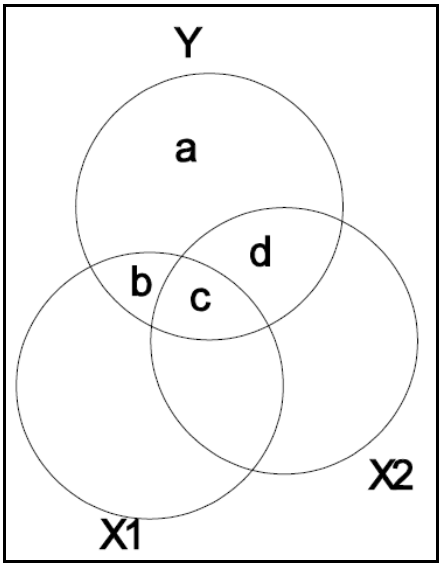


Figure 4-3. Venn Diagram Representation of MR Equation

Calculations of b+c+d, SS_{Reg}, and a, $SS_{Residual}$, are shown by the regression output: $SS_{Reg} = (9-1)7.963^2 = 507.27$ = b+c+d, and $SS_{Res} = (9-1)3.310^2 = 87.64$ = a. Their sum is $SS_{Total} = SS_y = (9-1)8.523^2 = 594.85$ = a+b+c+d, the total variability in recall. SS_y has been partitioned into what can be predicted by intelligence and memory strategies and what cannot be predicted by the two predictors; that is, $SS_{Total} = SS_{Reg} + SS_{Res}$, as was observed for a single predictor. Additional predictors (e.g., X3, X4, ...) could account for some residual variability, but the overall equation would still produce just two values, SS_{Reg} and SS_{Res}, that would sum to SS_{Total}.

Strength and Significance of the Overall Regression

As with a single predictor, SS_{Reg} and SS_{Res} can be used to calculate the proportion of variability in recall that can and cannot be predicted collectively by *int* and *mem*. The relevant calculations (division by SS_{Total}) are shown in the Model Summary section of the SPSS output. The square roots of these proportions produce rs that represent the variability in recall that can and cannot be predicted by both predictors (see Box 4-3).

$$R^2_{y.12} = \frac{SS_{\hat{y}}}{SS_y} \quad 1-R^2_{y.12} = \frac{SS_{y-\hat{y}}}{SS_y}$$

$$H_0: \rho^2_{y.12} = 0 \quad H_a: \rho^2_{y.12} > 0$$

$$F_{Overall} = \frac{\frac{SS_{Regression}}{p}}{\frac{SS_{Residual}}{n-p-1}} = \frac{\frac{R^2_{y.12}}{p}}{\frac{1-R^2_{y.12}}{n-p-1}}$$

Box 4-3. Strength and Significance

The overall relation between recall and both predictors can be tested for significance using the *F* test shown in Box 4-3 and in the ANOVA section of the output. The *df* for the numerator SS_{Reg} is $p = 2$, where p = the number of predictors, and the *df* for the denominator SS_{Res} is $n - p - 1 = 9 - 2 - 1 = 6$. Dividing SS by *df* gives Mean Squares (variances) for the numerator and denominator, which form an *F* ratio to test the H_0: $\rho^2_{r.im} = 0$. *F* can also be calculated using *R*, as shown in Box 4-3 and the regression output. Here we reject H_0. There is no corresponding *t* test for *F* when $df_{Numerator} > 1$.

As for a single predictor, several features of the overall multiple regression analysis can be demonstrated by correlating original variables and new variables created by the regression (see Figure 4-4 and rs below). Predicted and residual scores correlate 0 because they represent independent sources of variability; that is, SS that can (b+c+d) and SS that cannot be predicted (a) are mutually exclusive; they do not overlap. As well, residual scores correlate 0 with the two predictors used to generate the predicted scores because any variability in recall related to *int* or *mem* is in predicted recall scores. The correlation between observed and predicted recall scores is equal to the multiple R calculated earlier and the correlation between observed and residual recall scores represents variability not predicted by *int* and *mem*. Predicted and residual scores account for all the variability in recall, that is, $.923^2 + .384^2 = 1.0$.

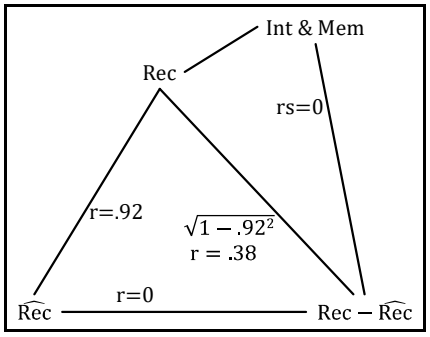


Figure 4-4. Relationships among original and derived variables.

```
VARIABLE LABEL prdr.im '' resr.im ''.
CORR rec int mem prdr.im resr.im /STAT.
                 rec   int   mem   prdr.im
 int            .630
 mem            .923   .711
 prdr.im        .923   .682  .999
 resr.im        .384   .000  .000  .000
```

Analyses to this point have concerned the collective prediction of recall from the combination of *int*

and *mem* scores, although the prediction plane and the regression coefficients suggest recall increased as a function of *mem* but not *int*. To be more specific about the strength and significance of the unique contribution of each predictor, the first step is to calculate an SS that represents unique contribution (i.e., variability in recall predicted *only* by one predictor and not the other).

Sum of Squares Unique

To determine an SS that represents the unique contribution, consider the Venn diagrams in Figure 4-3. The area b+c+d represents $SS_{Regression}$ with both predictors, which can be represented as $SS_{\hat{y}.12}$, the variability in y predicted by a combination of X1 and X2. Area b+c represents the variability in y that can be predicted by X1 alone, represented by $SS_{\hat{y}.1}$. Area c+d represents the variability in y that can be predicted by X2 alone, $SS_{\hat{y}.2}$. The unique contribution of each predictor is overlap with Y that is *not* shared with the other predictor: b for X1 and d for X2. As shown in Box 4.4., the unique contribution for X1 can be calculated by subtraction: b = (b+c+d) - (c+d) = $SS_{\hat{y}.12}$ - $SS_{\hat{y}.2}$ = $SS_{\hat{y}1.2}$, the SS in y that can be predicted by X1 controlling for X2. $SS_{\hat{y}2.1}$ can be calculated in a similar manner.

b+c+d	=	$SS_{\hat{y}.12}$
c+d	=	$SS_{\hat{y}.2}$
b	=	$SS_{\hat{y}1.2}$

Box 4.4. Unique contribution of X1.

The SSs required to calculate $SS_{\hat{y}.12}$ can be obtained from two REGRESSIONS, one for X2 alone ($SS_{\hat{y}.2}$) and the other for X1 and X2 ($SS_{\hat{y}.12}$ - $SS_{\hat{y}.2}$). However, SPSS can produce the needed quantities from a single regression because successive ENTER commands *add* predictors to those already in the equation. Calculation of $SS_{\hat{y}1.2}$ requires a regression with just X2 in it and a regression with both X1 and X2 included. In the present study, calculation of $SS_{\hat{r}i.m}$ requires a regression with just *mem* to obtain $SS_{\hat{r}.m}$ and a regression with both *int* and *mem* to obtain $SS_{\hat{r}.im}$. The following commands provide both from a single regression. Model 1 is the regression with just *mem* and Model 2 contains both *int* and *mem* because /ENTER int adds *int* to the predictor(s) already in the equation, in this case *mem*.

```
REGRESS/DEP = rec /ENTER mem /ENTER int.

 Model R       R      Adjusted Std.Error
               Square R Square the Estimate
 1     .923(a) .851   .830
 2     .923(b) .853   .804     3.822

 Model            Sum of Squares df Mean Square F      Sig.
 1     Regression 506.432        1  506.432     40.076 .000(a)
       Residual   88.457         7  12.637
       Total      594.889        8

 2     Regression 507.258        2  253.629     17.366 .003(b)     SSChamge = 507.258 - 506.432
       Residual   87.631         6  14.605                                  = 0.826
       Total      594.889        8
```

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.821	5.202		1.311	.231
	mem	1.535	.242	.923	6.331	.000
2	(Constant)	9.605	12.979		.740	.487
	mem	1.598	.371	.960	4.311	.005
	int	−.040	.168	−.053	−.238	.820

Calculation of $SS_{\hat{i}.m}$ is obtained from SS_{Reg} for Model 2 (b+c+d) minus SS_{Reg} for Model 1 (c+d) = $SS_{\hat{i}.im}$ - $SS_{\hat{i}.m}$= 507.258 - 506.432 = 0.826. This quantity is the change in SS_{Reg} from Model 1 to Model 2. Consistent with earlier impressions, *int* uniquely predicts very little of recall once its correlation with *mem* is eliminated (i.e., subtracted, controlled, kept constant).

A related way to conceptualize this process is to consider what would happen if the overlap between X1 and X2 was removed, ignoring for the moment Y. Eliminating the overlap would remove areas c and d in the Venn diagrams and leave only area b, the unique contribution of X1. The correlation between Y and the variability in X1 that is unique (i.e., independent of X2) represents the unique contribution of X1 (see Figure 4-5). This approach will be examined more fully later, after considering the significance and strength of the unique contribution of each predictor.

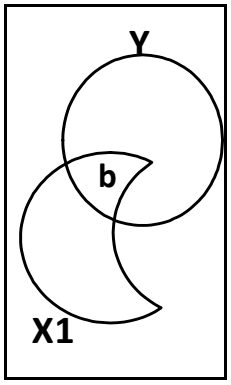


Figure 4-5. Unique X1

One final observation. Observe that the final output for Model 2 (i.e., after *int* was added to *mem* so both predictors are in the equation) is identical to the output with both predictors entered at the same time. The final equation and associated statistics do not differ when predictors are entered separately or in a different order. That is, REGRESS /DEP = *rec* /ENTER *int* /ENTER *mem* will produce the same final equation (i.e., Model 2) as observed here.

APPENDIX 4-1

To plot the best-fit plane with two predictors (intelligence and memory strategies in this example), first create a wider range of values for the two predictors and then generate the plane. Given the equation, the following commands and some chart-editing produced the best-fit plane (i.e., predicted values) shown in Figure 4-2 and reproduced in Figure 4-6. The range of values for i and m were chosen from Figure 4-1, a plot of the actual data, and r is the predicted recall score computed from the best-fit equation obtained from the MR analysis.

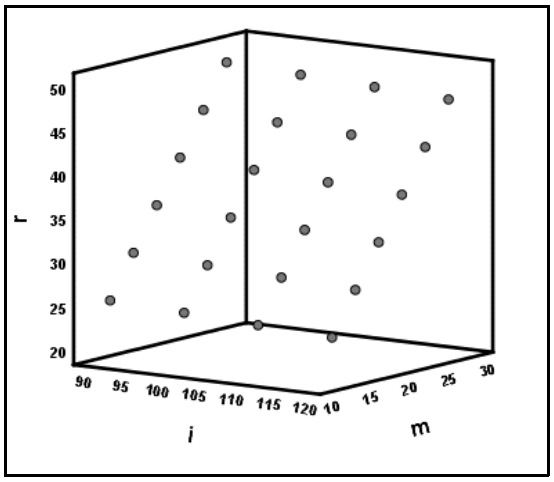


Figure 4-6. Plot of Best-Fit Plane.

```
INPUT PROGRAM.
LOOP i = 90 TO 120 BY 10.
LEAVE i.
LOOP m = 12 TO 27 BY 3.
END CASE.
END LOOP.
END LOOP.
END FILE.
END INPUT PROGRAM.

COMPUTE r = 9.605 -.04*i + 1.598*m.

GRAPH  /SCATTERPLOT(XYZ) = m WITH r WITH I.
```

To better appreciate what the program does, Figure 4-6 plots the relationship between the artificial values generated for i (intelligence) and m (memory strategies). Whereas the original observed values for *int* and *mem* were highly correlated, the artificial values *i* and *m* do not correlate and cover the full range of possible values. This generates predicted values that better illustrate the best-fit plane.

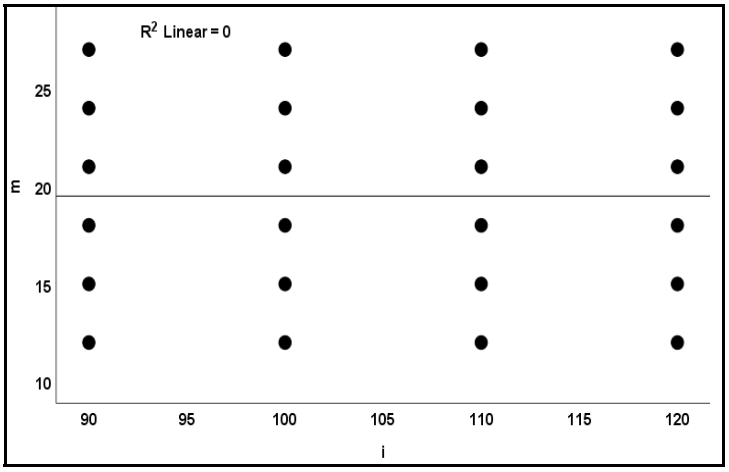


Figure 4-6. Plot of Artificial Predictors Generated by INPUT PROGRAM.

CHAPTER 5 - STRENGTH & SIGNIFICANCE OF UNIQUE CONTRIBUTION

Chapter 4 showed that intelligence (*int*) and memory strategies (*mem*) together accounted for a significant and substantial amount of variability in recall scores (*rec*). As well, an SS_{Unique} or SS_{Change} for each predictor was calculated by subtraction: $SS_{\hat{y}i.m}$ and $SS_{\hat{y}m.i}$. In general, $SS_{\hat{y}1.2} = SS_{\hat{y}.12} - SS_{\hat{y}.2}$. This chapter uses SS_{Unique} to calculate a measure of the strength of the unique contribution of each predictor and an F test of the significance of the unique contribution of each predictor controlling for other predictors in the equation. An equivalent t-test for the regression coefficient is also presented.

Part Correlation and the Strength of the Unique Contribution

As shown previously, a common way to measure the strength of a predictor is a correlation coefficient squared, r^2, which indicates the proportion of variability explained by the predictor. Calculation of r^2 involves dividing $SS_{Predicted}$ by SS_y, the total variability in the predictor. For the unique contribution of predictor X1 in MR, the strength is the SS_{Change} calculated earlier divided by SS_y, that is, $SS_{\hat{y}1.2}$ / SS_y (see Box 5-1). This is called a part correlation, represented as $r_{y(1.2)}$. The rationale for the parentheses in the notation is explained later. The following regression provides the SSs required to calculate $SS_{\hat{y}i.m}$. As well, some new commands produce the additional statistics. These printouts contain more precise values that were obtained from SPSS for several quantities because the default values were too coarse given the tiny unique contribution of *int*. Some results may not be perfectly equal.

$$r^2_{y(1.2)} = \frac{SS_{\hat{y}.12} - SS_{\hat{y}.2}}{SS_y} = R^2_{y.12} - R^2_{y.2}$$

Box 5-1. Part r Formula

```
REGRESS /STAT = DEFAULT CHANGE ZPP /DEP = rec /ENTER mem /ENTER int.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.923(a)	***.851306***	.830	3.555	.851	40.076	1	7	.000
2	.923(b)	***.852693***	.804	3.822	**.001388**	.057	1	6	.820

$r^2_{r(i.m)}$ = .825424/594.889
= .001388 = $R^2_{r.im} - R^2_{r.m} \approx .852693 - .851306$

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	***506.432322***	1	506.432	40.076	.000(a)
	Residual	88.457	7	12.637		
	Total	594.889	8			
2	Regression	***507.257746***	2	253.629	17.366	.003(b)
	Residual	87.631	6	14.605		
	Total	594.889	8			

$SS_{\hat{y}i.m}$ = 507.257746 – 506.432322
= .825424

Model		Unstandardized Coefficients: B	Std. Error	Standardized Coefficients: Beta	t	Sig.	Correlations: Zero-order	Partial	Part
1	(Constant)	6.821	5.202		1.311	.231			
	mem	1.535	.242	.923	6.331	.000	.923	.923	.923
2	(Constant)	9.605	12.979		.740	.487			
	mem	1.598	.371	.960	4.311	.005	.923	.869	.675
	int	–.040	.168	–.053	–.238	.820	.630	–.097	**–.037**

$r_{r(i.m)}$ = √.001387
= –.03725

As before, *mem* is entered into the equation first and *int* is added second. This gives the unique

contribution of *int*. The first step is to calculate $SS_{\hat{y}i.m}$ for *int* added to *mem*. This quantity can be referred to simply as SS_{Change}, although that label is vague about which predictor it is for. The part r^2 is SS_{Change} divided by SS_{Total} = .825424/597.889 = .001387 for *int* controlling for *mem*.

The *CHANGE* option on the *REGRESSION* command produces R Square Change in the Change Statistics section of the output. For Model 2, R^2_{Change} = .001388, the part r^2 for *int*, which was added to the equation at step 2. As shown above, R^2_{Change} can also be calculated as the increase in R^2 from Model 1 to Model 2; that is, $R^2_{r(i.m)} = R^2_{r.im} - R^2_{r,m}$.

The part *r*, $r_{r(i.m)} = \sqrt{}.001387 = -.037$, appears on the regression line for *int*. The additional output for each predictor was produced by the ZPP option in the REGRESSION command. ZPP refers to zero, partial, and part correlation. Zero is the simple correlation between the dependent variable and a predictor, that is, $r_{r.i}$ for *int* and $r_{r.m}$ for *mem*. The part *r* is negative because of the negative coefficient for *int*.

Box 5-2 shows the relationship between part r^2 and the difference in R^2; specifically, $r^2_{y(1.2)} = R^2_{y.12} - R^2_{y.2}$. These calculations for the part r^2 for *int* are shown in the preceding regression below the change statistics.

$$\frac{SS_{\hat{y}_{1.2}}}{SS_y} = \frac{SS_{\hat{y}.12} - SS_{\hat{y}.2}}{SS_y} = \frac{R^2_{y.12} \times SS_y - R^2_{y.2} \times SS_y}{SS_y} = \frac{(R^2_{y.12} - R^2_{y.2}) \times SS_y}{SS_y} = R^2_{y.12} - R^2_{y.2}$$

Box 5-2. Alternative calculation for part r.

It is telling to contrast the part *r* for *int*, -.037, with the simple *r* of .630 (see the Zero-order column generated by the ZPP option). A large positive correlation of .630 has become tiny and even slightly negative, $r_{r'i.m}$ = -.037 once the correlation between *int* and *mem* is controlled. The original simple correlation was positive and large because people high on *int* tended to be high on *mem*, and *mem* was related to the criterion variable *rec*. Controlling for *mem* (i.e., removing c in the Venn diagram) eliminates the relation between *int* and *rec*. The analysis below can be used to calculate the part correlation for *mem*, $r_{r(m.i)}$ = .675, which is less than r_{rm} = .923, but still substantial.

```
REGRESS /STAT = DEFAU CHANGE ZPP /DEP = rec /ENTER int /ENTER mem.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.630(a)	.397	.310	7.162	.397	4.599	1	7	.069
2	.923(b)	.853	.804	3.822	.456	18.581	1	6	.005

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	235.878	1	235.878	4.599	.069(a)
	Residual	359.011	7	51.287		
	Total	594.889	8			
2	Regression	507.258	2	253.629	17.366	.003(b)
	Residual	87.631	6	14.605		
	Total	594.889	8			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-9.772	22.815		-.428	.681			
	int	.474	.221	.630	2.145	.069	.630	.630	.630
2	(Constant)	9.605	12.979		.740	.487			
	int	-.040	.168	-.053	-.238	.820	.630	-.097	-.037
	mem	1.598	.371	.960	4.311	.005	.923	.869	.675

The contribution of *mem* remains moderately strong even with *int* in the equation, $r^2_{r(m.i)} = .456$ and $r_{r(m.i)} = .675$. However, the strength of *mem* is considerably reduced when *int* is controlled statistically (i.e., .675 is weaker than .923, the simple correlation between *rec* and *mem*). The reduction is even more notable in terms of r^2: $.675^2 = .456$ versus $.923^2 = .852$. The decrease occurs because some, but not all, of what *mem* predicts when alone was due to variability in *mem* shared with *int* (i.e., area c). In theory, this could be viewed as consistent with the mediation hypothesis: INT ➡ MEM ➡ REC. Although less plausible, however, the results are also consistent with MEM ➡ INT ➡ REC and other underlying models. In any case, some variation in *mem* remains related to recall when *int* is controlled.

Note that the unique contributions of the two predictors does *not* add up to the total variability accounted for when both predictors are in the equation; that is, $r^2_{r(i.m)} + r^2_{r(m.i)} = -.037^2 + .675^2 = .457 < .853 = R^2_{r.im}$. This inequality could also be demonstrated in terms of SSs. The other 40% or so of variability in *rec* that is predictable from *int* and *mem* is due to variability shared by the two predictors (i.e., area c in the Venn diagrams) and cannot be allocated to a specific predictor.

Significance of SS_{Change}

Although the part r reflects the strength of the unique contribution of a predictor, it does not tell us about its significance; that is, whether it could have occurred by chance. There are two equivalent tests of significance for the unique contribution of a single predictor, an *F* test and a t-test. The *F* tests the significance of the increase in SS_{Reg} when a predictor is added to other predictors in the equation (see Box 5.3). Model 1 of the regression output shows that *mem* by itself accounts for $SS_{Reg} = 506.432$ units of variability. When *int* was added second, SS_{Reg} became 507.258, an increase of just .826 units of variability (i.e, 507.258 - 506.432 = .826). This SS_{Change} has $df = 1$ because just one predictor was added to the equation, which gives $MS_{Change} = .826/1 = .826$. Dividing MS_{Change} by MS_{Res} from the regression analysis with both predictors tests the significance of the unique contribution of *int*.

$H_0: \rho_{y1.2} = 0$

$H_a: \rho_{y1.2} \neq 0$ or $H_0: \rho_{y1.2} > 0$ or $H_0: \rho_{y1.2} < 0$

$SS_{Change} = SS_{\hat{y}1.2} = SS_{\hat{y}.12} - SS_{\hat{y}.2}$

$$F_{Change} = \frac{\frac{SS_{\hat{y}1.2}}{1}}{MS_{Residual}}$$

Box 5-3

For the unique contribution of *int*, $F = 0.826/ 14.605 = .057$. The denominator is MS_{Res} for both predictors because the test is for the unique contribution of *int* with *mem* controlled (i.e., with *mem* also in the

equation). The *F* value appears in the Change Statistics section of the MR analysis and is not significant given its *p* value of .820. Here are relevant sections of the earlier regression. The probability of an F of .057 or larger if the H_0 is true is .820; that is, p(F ≥.057 IF H_0 true) = .820, clearly greater than .05. The H_0 is not rejected.

```
REGRESS /STAT = DEFAU CHANGE ZPP /DEP = rec /ENTER mem /ENTER int.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.923(a)	.851306	.830	3.555	.851	40.076	1	7	.000
2	.923(b)	.852693	.804	3.822	.001388	**.057**	**1**	**6**	**.820**

Model		Sum of Squares	df	Mean Square	F	Sig.
...						
2	Regression	507.257746	2	253.629	17.366	.003(b)
	Residual	87.631	6	14.605		
	Total	594.889	8			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
...									
2	(Constant)	9.605	12.979		.740	.487			
	mem	1.598	.371	.960	4.311	.005	.923	.869	.675
	int	-.040	.168	-.053	-.238	.820	.630	-.097	-.037

There are several ways to conceptualize the *F* test and the equivalent t-test presented in the next section. Specifically, what do the tests measure and what happened to reduce *int* from marginal significance alone to nonsignificance when *mem* is controlled? We noted earlier that the simple correlation between *rec* and *int* was confounded because *int* was correlated with *mem*. MR shows that controlling for or eliminating the confounding has reduced the contribution of *int*. In terms of the graphs of the data and the best-fit plane, rather than comparing front-left observations (i.e., low on *int* and *mem*) with back-right observations (i.e., high on *int* and *mem*), the MR determines what the change in *rec* would be if *int* varied independently of *mem*. The graph of the plane in Chapter 4 (Figure 4-1) showed that the slope for *int* is virtually flat (i.e., close to 0).

A second way to conceptualize the *F* and *t* for the unique contribution is in terms of the Venn diagram shown previously and repeated in Figure 5-1. The circle labelled Y represents the variability in the criterion variable (*rec* in our example). The circles labelled X1 and X2 represent our predictors, *int* and *mem*. Area b represents the unique contribution of X1 (*int*), that part of Y that overlaps with X1

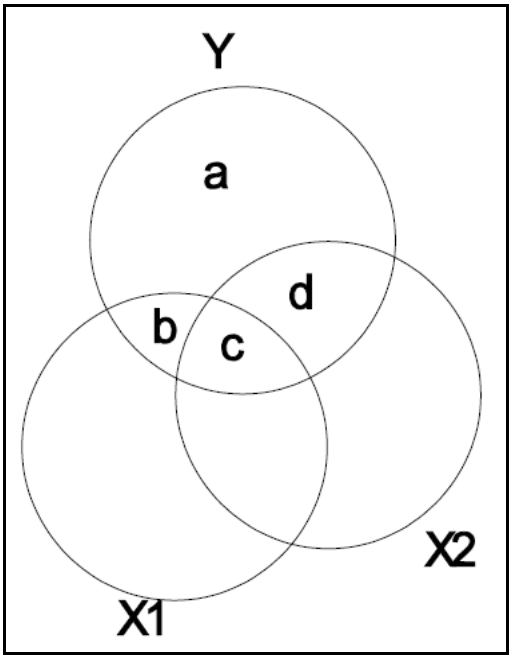


Figure 5-1. Venn Diagram Representation of Multiple Regression

but is independent of X2. In essence, the significance for each predictor in a multiple regression equation represents the significance of the increase in SS_{Reg} when that predictor is added to predictors already in the regression analysis. If a third predictor X3 was added and overlapped with an area of Y that did not overlap with X1 and X2, the additional variability accounted for would be unique to X3 and could be used to calculate a part r and to test the significance of the unique contribution of X3. And so on, with multiple predictors. The generalized version for p predictors is: $SS_{\hat{y}1.23...p} = SS_{\hat{y}.123...p} - SS_{\hat{y}.23...p}$ as discussed in chapter 7.

Significance of Regression Coefficients

Because *F* involves df = 1 for the numerator, there is an equivalent t. For *int*, the t-test can be considered a test of significance of the part r, $r_{r(i.m)}$, or of the regression coefficient, $b_{ri.m}$. Both are equivalent and we will just do the significance of the regression coefficient since that is how SPSS reports it. Specifically, the t-test determines whether the regression coefficient for each predictor differs significantly from 0 when other predictors are in the equation. The regression coefficient in the population is represented by the Greek letter beta; that is, $\beta_{y1.2}$ is the population coefficient for X1 controlling for X2. A regression coefficient that equals 0 uniquely explains none of the variability in the dependent variable when the other predictor is controlled.

Box 5-4 shows the relevant formula. First, calculate a Standard Error for the observed regression coefficient to reflect how much variability is expected from sample to sample if the null hypothesis is true. Note that the denominator for SE is variability in X1 that is independent of X2. $SE_{by1.2}$ can be used to compute a *t* to test the significance of the regression coefficient. The calculations appear after the following regression output. Earlier calculations using *F* to test significance are included for comparison.

$$H_0: \beta_{y1.2} = 0$$
$$H_0: \beta_{y1.2} \neq 0 \text{ or } H_0: \beta_{y1.2} > 0 \text{ or } H_0: \beta_{y1.2} < 0$$
$$t = \frac{b_{y1.2} - 0}{SE_{b_{y1.2}}}$$
$$SE_{b_{y1.2}} = \sqrt{\frac{MS_{Residual}}{SS_1(1 - r_{12}^2)}}$$

Box 5-4. T-test for regression coefficient.

```
REGRESS /STAT = DEFAU CHANGE /DEP = rec /ENTER mem /ENTER int.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.923(a)	.851	.830	3.555	.851	40.076	1	7	.000
2	.923(b)	.853	.804	3.822	.001	**.057**	1	6	.820

F_{change} = 0.826/14.605 = .057

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	506.432	1	506.432	40.076	.000(a)
	Residual	88.457	7	12.637		
	Total	594.889	8			
2	Regression	507.258	2	253.629	17.366	.003(b)
	Residual	87.631	6	14.605		
	Total	594.889	8			

SS_{Chamge} = 507.258 – 506.432 = 0.826

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.821	5.202		1.311	.231
	mem	1.535	.242	.923	6.331	.000
2	(Constant)	9.605	12.979		.740	.487
	mem	1.598	.371	.960	4.311	.005
	int	**-.040**	**.168**	-.053	**-.238**	**.820**

$SE_{bi} = SQRT\{MSE/((1-r_{im}^2)*SS_i)\} = SQRT\{14.605/((1-.711^2)*1050)\} = .168$

$t_{int} = -.04/.168 = -.238$ $df = n - p - 1 = 6$ DNR H_0: $\beta_{ri.m} = 0$

As expected, the observed *p* value for *t*, .820, equals the *p* for *F*. Also, $F = t^2 = -.238^2 = .057$. *F* and *t* tests are equivalent because the *F* test numerator has *df* = 1. It tests the significance of one additional predictor in the equation. With a *p* value of .820, the slope is clearly not significant, even though the simple relationship between recall and intelligence was close to significance, *p* = .069.

The preceding SPSS analysis produced change statistics for the unique contribution of *int* because *mem* was entered first and then *int* added. A regression with *int* entered first and *mem* second was presented earlier. Use that analysis to determine and explain the unique contribution of *mem*. Perform calculations to test the significance of the unique contribution of *mem* and relate the results to the printout. Observe that the t-tests for the regression coefficients are the same in both analyses irrespective of the order predictors are entered. It is not necessary to obtain the Change statistics to determine the significance or the strength of the unique contribution for each predictor, although that is a useful way to conceptualize the test.

One important observation about SS_{Change} is that the increase in SS_{Reg} is the same as the decrease in SS_{Res}. This must be true given SS_{Total} is identical in both regressions. In terms of Figure 5.1, the unique contribution of a predictor is variability in the dependent variable that was error with only the other predictor in the equation. Below is the relevant section of the previous analysis with *mem* alone in model 1 and *int* added in model 2. SS_{Change} obtained by subtracting SS_{Res} (model 1 minus model 2) is the same as that obtained by subtracting SS_{Reg} (model 2 minus model 1). That is, $SS_{Change} = 507.258 - 506.432 = 88.457 - 87.631 = .826$. This alternative way to view SS_{Change} will help understand a second measure of the strength of the unique contribution of a predictor, as covered in Chapter 6.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	506.432	1	506.432	40.076	.000(a)
	Residual	88.457	7	12.637		
	Total	594.889	8			
2	Regression	507.258	2	253.629	17.366	.003(b)
	Residual	87.631	6	14.605		
	Total	594.889	8			

Sampling Distributions for Significance Tests

The following distributions are based on 100,000 samples of X1, X2, and Y for 24 participants. In

the population, all correlations are 0. That is, $\rho_{y1} = 0$, $\rho_{y2} = 0$, and $\rho_{12} = 0$. In a Venn diagram, there would be three, non-overlapping circles representing X1, X2, and Y.

Figure 5-2 shows the distribution of the 100,000 multiple correlation coefficients, $R_{y.12}$. In the population, $\rho_{y.12} = 0$, but because a multiple R cannot be negative, only positive values are produced.

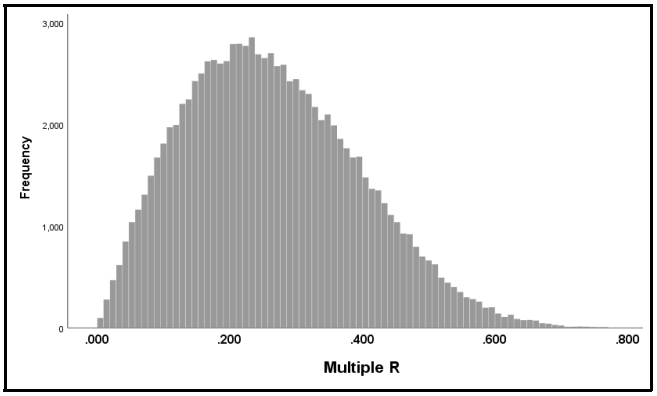


Figure 5-2. Multiple R.

The statistical question is how large $R_{y.12}$ must be to reject H_0: $\rho_{y.12} = 0$. To answer that question, F(2, 21) is calculated for each sample. The 100,000 Fs appear in Figure 5-3. Although not shown, $p_{Observed}$ was also calculated for each F. If H_0 is true, as it is here, 5% of those ps should be less than or equal to .05.

The dashed vertical line in Figure 5-3 is $F_{Critical} = 3.47$ for df = p, n-p-1 = 2, 21. There are two equivalent ways to determine how many samples produced a significant $F_{Observed}$. We can count the number of ps less than .05 or the number of Fs greater than 3.47. The following commands do this, then count the number of significant values for each. As expected, they agree and are close to the expected proportion of .05.

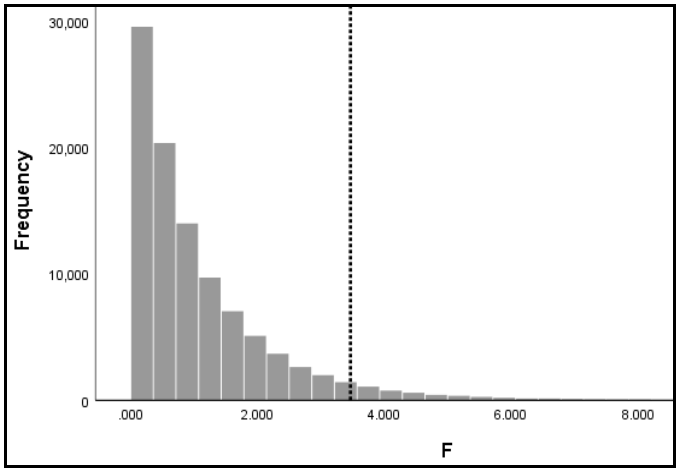


Figure 5-3. Distribution of Fs for 100,000 Samples.

```
COMPUTE sigf = f GE 3.4668.
COMPUTE sigp = p LE .05.
FREQ sigp sigf.
```

Sigf	Frequency	Percent
0	95091	95.1
1	4909	4.9

Sigp	Frequency	Percent
0	95091	95.1
1	4909	4.9

F tests the significance of the relationship between Y and both X1 and X2. It does not measure the significance of the unique contribution of X1 or X2. A t-test for the significance of each regression coefficient provides that information. Here the null hypothesis is H_0: $\beta_{Y1.2} = 0$ or equivalently $\rho_{Y(1.2)} = 0$.

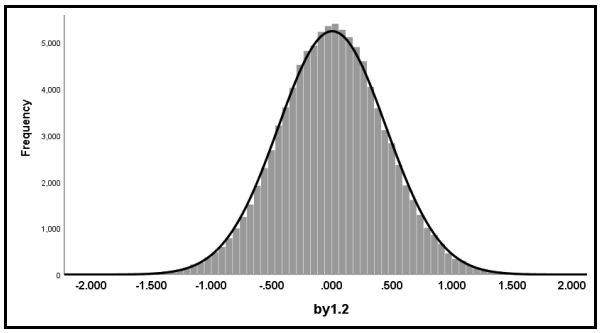


Figure 5-4. Distribution of Regression Coefficient.

Figure 5-4 shows the distribution of the regression coefficient for X1 controlling for X2. It can take on positive or

negative values and has an expected value of 0 in the present case. A distribution for $r_{Y(1.2)}$ would also have an expected value of 0. However, by chance $b_{y1.2}$ can deviate from 0, sometimes by a substantial amount. To determine whether it deviates enough to reject H_o: $\beta_{Y1.2} = 0$, a t-test is calculated or an F test for SS_{Change}.

Figure 5-5 shows the distribution of ts for the 100,000 samples and the critical values for a non-directional test. A p value was also calculated for each $t_{Observed}$ and the number of samples with significant results calculate from t and p. Again, the results confirm those expected when there is no relationship among the three variables.

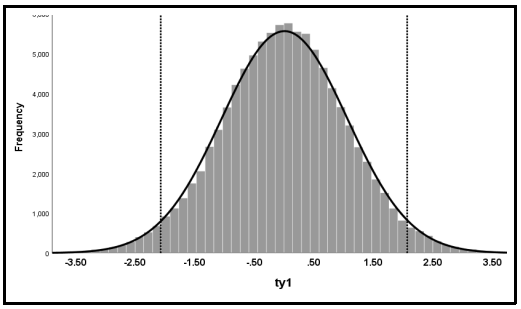


Figure 5-5. T for Regression Coefficient.

```
COMPUTE sigp = py1 LE .05.
COMPUTE sigt = ty1 LE -2.0796 OR ty1 GE 2.0796.
FREQ sigp sigt.
```

Sigp	Frequency	Percent
.000	94966	95.0
1.000	5034	5.0

Sigt	Frequency	Percent
.000	94966	95.0
1.000	5034	5.0

Given this introduction to multiple regression with two predictors, let's examine mediation analysis, one common application of multiple regression as mentioned at the start of chapter 4. The goal of mediation analyses is to determine whether the relationship between a predictor X and a dependent variable Y is due to a mediator variable. The hypothesized underlying model is X ➡ M ➡ Y. If this model is correct, the statistical relationship between X and Y (i.e., significance, strength) should weaken and perhaps even disappear when Y is regressed on both X and M. Appendix 5-2 describes the process more fully and demonstrates, *most importantly*, that mediation analyses can be consistent with other underlying models and must be considered thoughtfully

APPENDIX 5-1: PRACTICE SAMPLES

Below, i1, i2, ..., i9 represent the 9 intelligence scores; m1, m2, ..., m9 represent the 9 memory strategy scores; and r1, r2, ..., r9 represent the 9 recall scores. Descriptive statistics, simple rs, and MR results are shown for each sample in the next two listings. Note the correspondence between the results for Sample 10 below and analyses discussed in previous pages.

SAMPLE	i1	m1	r1	i2	m2	r2	i3	m3	r3	i4	m4	r4	i5	m5	r5	i6	m6	r6	i7	m7	r7	i8	m8	r8	i9	m9	r9
1	96	22	55	109	25	42	90	21	27	123	25	46	109	23	48	94	12	29	94	22	39	116	26	37	83	14	37
2	99	23	42	74	11	35	111	24	44	85	23	42	108	21	39	105	18	31	96	19	45	98	21	52	93	23	37
3	77	14	41	80	17	41	98	22	39	95	18	51	96	24	42	98	20	43	97	19	33	112	18	37	118	23	57
4	117	30	53	104	19	33	115	21	38	76	17	44	83	21	36	84	17	37	74	15	30	93	19	45	91	21	38
5	105	21	43	103	17	38	114	24	45	93	18	32	79	17	34	95	21	43	94	18	37	83	8	31	110	25	43
6	85	17	41	114	20	40	118	23	42	81	18	42	87	17	41	79	17	36	87	14	39	88	13	37	96	22	44
7	90	20	31	89	26	42	108	24	45	105	23	46	75	17	41	104	17	38	106	22	42	74	17	46	96	16	33
8	76	17	35	85	16	40	115	25	39	75	16	40	133	30	57	119	27	45	98	22	53	78	15	33	96	12	37
9	103	23	36	87	19	35	80	23	42	106	22	44	111	23	38	100	22	42	98	20	46	117	29	54	103	19	43
10	96	25	47	111	27	51	118	23	41	112	21	37	88	14	25	92	12	30	112	25	41	89	18	32	106	23	46

SAMPLE	MNi	SDi	MNm	SDm	MNy	SDy	Rri	Rrm	Rim	Byi	Bym	B0
1	101.5556	13.2393	21.1111	4.9103	40.0000	8.9303	.3944	.4760	.7450	.0602	.7449	18.1621
2	96.5556	11.5878	20.3333	4.0311	40.7778	6.1599	.1578	.4312	.5869	-.0773	.7894	32.1898
3	96.7778	13.0459	19.4444	3.1667	42.6667	7.2457	.3416	.2851	.6199	.1487	.2725	22.9745
4	93.0000	15.8745	20.0000	4.3012	39.3333	6.9642	.4342	.7136	.7543	-.1059	1.4502	20.1775
5	97.3333	11.7580	18.7778	4.9944	38.4444	5.2941	.8226	.8410	.7677	.1941	.5406	9.3992
6	92.7778	14.0337	17.8889	3.3333	40.2222	2.5386	.3559	.6976	.6835	-.0410	.6493	32.4129
7	94.1111	13.0714	20.2222	3.6667	40.4444	5.4569	.0220	.4130	.4219	-.0773	.7310	32.9399
8	97.2222	21.0225	20.0000	6.2048	42.1111	8.1155	.6988	.7447	.8701	.0808	.7359	19.5401
9	100.5556	11.3700	22.2222	3.0322	42.2222	5.7615	.4558	.5978	.5289	.0982	.9411	11.4310
10	102.6667	11.4564	20.8889	5.1828	38.8889	8.6233	.6297	.9227	.7109	-.0399	1.5978	9.6053

SAMPLE	SSREG	SSRES	R2	F	SIG	Ti	SIGi	Tm	SIGm
1	146.8463	491.1537	.2302	.8969	.4562	.1662	.8735	.7628	.4745
2	60.6585	242.8971	.1998	.7492	.5123	-.3223	.7581	1.1453	.2957
3	52.6822	367.3178	.1254	.4303	.6689	.5503	.6020	.2448	.8148
4	207.3191	180.6809	.5343	3.4423	.1010	-.5689	.5901	2.1108	.0793
5	175.6881	48.5341	.7835	10.8597	.0101	1.4545	.1960	1.7206	.1361
6	26.4994	25.0562	.5140	3.1728	.1148	-.5816	.5820	2.1867	.0714
7	47.3552	190.8670	.1988	.7443	.5143	-.4596	.6620	1.2186	.2687
8	297.8104	229.0785	.5652	3.9001	.0822	.3831	.7149	1.0301	.3427
9	102.1020	163.4536	.3845	1.8740	.2332	.5136	.6259	1.3124	.2373
10	507.2577	87.6311	.8527	17.3657	.0032	-.2377	.8200	4.3106	.0050

Here are SPSS commands to enter the data for sample 10.

```
DATA LIST FREE / int mem rec.
BEGIN DATA
 96 25 47  111 27 51  118 23 41  112 21 37   88 14 25
 92 12 30  112 25 41   89 18 32  106 23 46
END DATA.
```

APPENDIX 5-2: CAUTIONS ABOUT MEDIATION ANALYSIS

Mediation analysis determines whether the relationship between a predictor X and a criterion Y is due to a mediator M. The general model is shown in Figure 1, along with standardized regression coefficients from four regression analyses: X➜Y, X&M➜Y, X➜M, M➜Y. The hypothesis that M is a mediator of X➜Y can be tested in two equivalent ways. One way to determine whether the path from X to M to Y is significant is to test whether the product of β_{YM} times β_{MX} differs from 0. The second is to determine whether $\beta_{YX.M}$ is less than β_{YX}. The second approach is used below although we do not actually test the significance. The appropriate test is one that we have not learned.

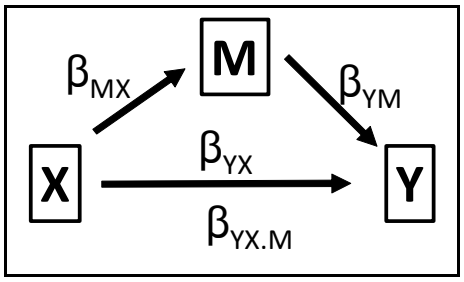


Figure 13. Mediation Model.

The following demonstrations show that a reduction in the relationship between X and Y is ambiguous because alternative models to X ➜ M ➜ Y produce the same statistical result. The simulated data are generated using the models shown before the SPSS commands and also in the bolded lines of the syntax. Remaining lines stay the same. The four models all produce a decrease in the relationship between X and Y when M enters the regression. In the results, note the decrease in significance for X from Model 1 to Model 2 and the part r for X being much less than the simple r for X, as bolded in the ZPP section of Model 2.

Model 1: X ➜ M ➜ Y

```
SET SEED = 7654321.
INPUT PROGRAM.
LOOP o = 1 TO 1000.
COMPUTE x = RV.NORM(0,1).
COMPUTE m = x*.7071 + RV.NORM(0,1)*.7071.
COMPUTE y = m*.7071 + RV.NORM(0,1)*.7071.
END CASE.
END LOOP.
END FILE.
END INPUT PROGRAM.

REGRESS /STAT = DEFAU ZPP /DEP = y /ENTER x /ENTER m.
```

Model		Unstandardized Coefficients		t	Sig.	Correlations		
		B	Std. Error			Zero-order	Partial	Part
1	(Constant)	-0.006	0.027	-.233	.816			
	x	.504	0.027	**18.737**	**.000**	.510	.510	0.51
2	(Constant)	.004	0.022	.185	.854			
	x	.010	0.032	**.324**	**.746**	**.510**	.010	**0.007**
	m	.691	0.032	21.757	.000	.706	.567	0.488

Model 2: Y ➜ M ➜ X

```
. . .
COMPUTE y = RV.NORM(0,1).
COMPUTE m = y*.7071 + RV.NORM(0,1)*.7071.
COMPUTE x = m*.7071 + RV.NORM(0,1)*.7071.
. . .
```

Model		Unstandardized Coefficients		t	Sig.	Correlations		
		B	Std. Error			Zero-order	Partial	Part
1	(Constant)	.012	.027	.430	.667			
	x	.516	.028	**18.737**	**.000**	.510	.510	.510
2	(Constant)	.016	.022	.736	.462			
	x	.010	.032	**.324**	**.746**	**0.51**	0.01	**.007**
	m	.709	.031	22.673	.000	.715	0.583	.502

Model 3: M ➜ X & Y

```
. . .
COMPUTE m = RV.NORM(0,1).
COMPUTE x = m*.7071 + RV.NORM(0,1)*.7071.
COMPUTE y = m*.7071 + RV.NORM(0,1)*.7071.
. . .
```

Model		Unstandardized Coefficients		t	Sig.	Correlations		
		B	Std. Error			Zero-order	Partial	Part
1	(Constant)	0.016	.027	.580	.562			
	x	0.498	.027	**18.240**	**.000**	.500	.500	0.5
2	(Constant)	0.004	.022	.185	.854			
	x	-0.016	.032	**-.509**	**.611**	**0.5**	-0.016	**-0.011**
	m	0.717	0.032	22.608	0	0.71	0.582	0.504

Model 4: z ➜ X&M, M ➜ Y

```
. . .
COMPUTE z = rv.norm(0,1).
COMPUTE x = RV.NORM(0,1)*.7071 + z*.7071.
COMPUTE m = RV.NORM(0,1)*.7071 + z*.7071.
COMPUTE y = RV.NORM(0,1)*.7071 + m*.7071.
...
```

Model		Unstandardized Coefficients		t	Sig.	Correlations		
		B	Std. Error			Zero-order	Partial	Part
1	(Constant)	-.070	.145	-.486	0.629			
	x	0.515	.153	3.369	0.002	0.445	0.445	0.445
2	(Constant)	-.001	.100	-.010	.992			
	x	.093	.120	**0.771**	**0.444**	**0.445**	0.114	**0.07**
	m	.707	.098	7.244	.000	.790	.734	0.657

CHAPTER 6 - MORE ON UNIQUE CONTRIBUTION

Chapter 6 examines three topics related to the strength of the unique contribution of predictors: an alternative conceptualization of part rs, standardized regression coefficients, and the partial r, a less common and more ambiguous alternative to the part r.

Unique Contribution and Residual Predictors

Part *r* is equal to SS_{Change} for a predictor divided by the *total* variability in the criterion variable y. A closely related way to conceptualize part *r*s is in terms of a residual predictor; that is, a predictor with variability shared with *other* predictors removed. This approach generalizes to MR with more than two predictors better than Venn diagrams.

Recall that a residual dependent variable is uncorrelated with the predictor(s) used to generate predicted and residual scores; that is, $r_{(y\text{-}\hat{y})x} = 0$. Given this fact, the variability in one predictor that is independent of other predictors in the equation can be determined by regressing one predictor on the other predictors. For two predictors, if X2 is used to predict X1 the X1 residuals will be completely independent of X2. Below is the standard regression analysis for recall (*rec*) as a function of intelligence (*int*) and memory strategies (*mem*). Previous calculations for the part r for *int* are shown to compare with the following analyses.

```
REGRESS /STAT = DEFAU CHANGE ZPP /DEP = rec /ENTER mem /ENTER int.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.923(a)	***.851306***	.830	3.555	.851	40.076	1	7	.000
2	.923(b)	***.852693***	.804	3.822	**.001388**	.057	1	6	.820

$r^2_{r(i.m)}$ = .825424/594.889

= .001388 = $R_{r.im}^2 - R_{r.m}^2 \approx .852693 - .851306$

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	***506.432322***	1	506.432	40.076	.000(a)
	Residual	88.457	7	12.637		
	Total	594.889	8			
2	Regression	***507.257746***	2	253.629	17.366	.003(b)
	Residual	87.631	6	14.605		
	Total	594.889	8			

$SS_{i.m}$ = 507.257746 − 506.432322 = .825424

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Correlations: Zero-order	Partial	Part
1	(Constant)	6.821	5.202		1.311	.231			
	mem	1.535	.242	.923	6.331	.000	.923	.923	.923
2	(Constant)	9.605	12.979		.740	.487			
	mem	1.598	.371	.960	4.311	.005	.923	.869	.675
	int	−.040	.168	−.053	−.238	.820	.630	−.097	**−.037**

$r_{y(i.m)} = \sqrt{.001387} = -.03725$

The Venn diagrams in Figures 6-1 and 6-2 show one way to visualize how residual scores capture the unique contribution of predictors. The diagram expands on previous versions. As before, area a is the residual variability and area b+c+d is the explained variability with both predictors. Area b+c is what X1 by itself predicts and c+d is what X2 by itself predicts. By subtraction, we obtained b, the unique contribution of X1, and d, the unique contribution of X2. Figure 6-1 labels some additional areas, e, f, and g. Note in particular that c+f represents the overlap between the two predictors. Ignore Y for now and focus on X1 and X2.

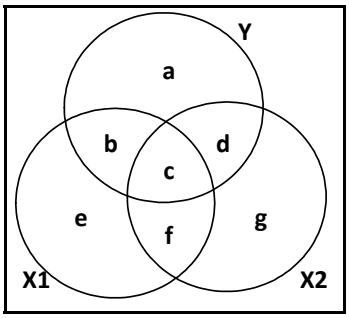


Figure 6-1.

To get the variability in X1 that is independent of X2 (i.e., unique to X1), c+f must be subtracted from the total variability in X1. Using X1 as the dependent variable and X2 as predictor, SS_{Res} for X1 (i.e., res1.2) equals the variability in X1 that does not overlap with X2 (i.e., is uncorrelated with X2). This corresponds to area b+e in Figure 6-1. Therefore, the overlap between Y and the residual X1 predictor now corresponds to b, namely $SS_{\hat{y}1.2}$, and the part r for X1, $r_{y(1.2)}$. Figure 6-2 illustrates these operations. In the top-left Venn diagram, the overlap between X1 and Y contains both b and c, the unique and shared variability, respectively. The lower-right Venn diagram shows that by removing the overlap between X1 and X2 (i.e., c+f), the residual X1 variable and Y share only area b, the unique contribution of X1. The same strategy could be used to obtain d in Figure 6-1, $SS_{\hat{y}2.1}$, and $r_{y(2.1)}$, the unique contribution of X2. See the analyses in Appendix 6-2.

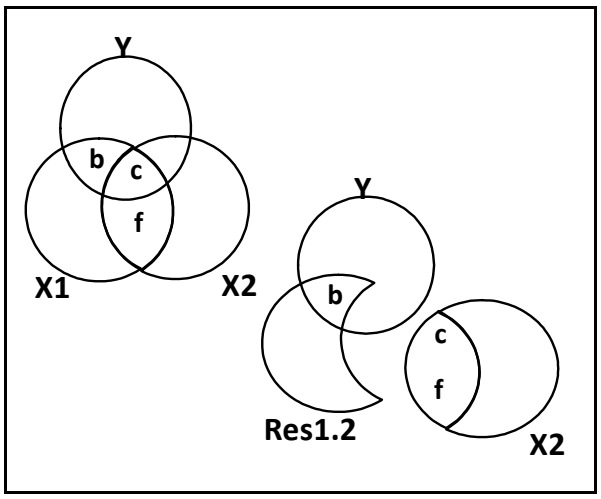


Figure 6-2.

This way to conceptualize the unique contribution of any predictor generalizes to multiple predictors, whereas the Venn diagram approach becomes overly complex. In terms

of residual predictors, removing overlap with all other predictors leaves only the portion of variability in X1 that is unique and independent of the other predictors. The relationship between this residual predictor is what X1 can predict in Y beyond what all other predictors can account for, producing $SS_{\hat{y}1.23...p}$, and $r_{y(1.23...p)}$, the relationship between y and X1 controlling for X2, X3, ..., Xp, where p is the number of predictors.

Calculating the part *r* from a residual predictor for *int* involves several steps summarized in Figure 6-3. Regress *int* on *mem* (i.e., *int* is the dependent variable and *mem* the predictor) and save residual scores from this regression, $i\text{-}\hat{i}$ in the Figure and *resi.m* in the following regression. This residual predictor is unique in that it correlates 0 with *mem*. Finally, correlating *resi.m* with *rec* or regressing *rec* on *resi.m*, produces the same results as earlier calculations for $r_{r(i.m)}$, the part *r* for *int*.

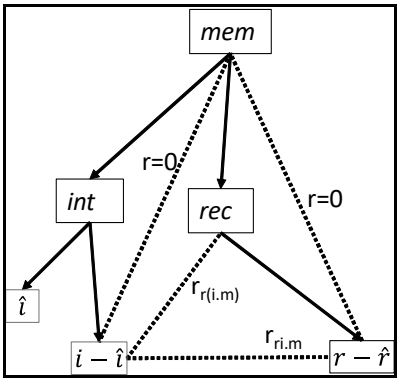


Figure 6-3. Part & Partial r as Residual Predictors

The following regression shows the relevant analyses. Note that *mem* accounts for 50.5% of the variability in *int*. This shared variability is removed to obtain variability unique to *int* and ultimately the unique contribution of *int* to the prediction of *rec*. Note that $SS_{Res} = SS_1(1\text{-}r^2_{12})$, the quantity in the denominator of SE for the t-test reported earlier.

```
REGRESS /DEP = int /ENTER mem /SAVE RESI(resi.m).

 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .711(a) .505     .435            8.614           1-R² = .495

 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 530.594        1  530.594     7.151 .032(a)
       Residual   519.406        7  74.201
       Total      1050.000       8

 Model            Unstandardized             Standardized    t     Sig.
                  Coefficients               Coefficients
                  B               Std. Error Beta
 1     (Constant) 69.843          12.606                     5.540 .001
       mem        1.571           .588       .711            2.674 .032

Residuals Statistics(a)
                 Minimum Maximum Mean   Std. Deviation N
 Predicted Value 88.70   112.27  102.67 8.144          9
 Residual        -13.127 12.016  .000   8.058          9 unique variation in int
```

Correlating *resi.m* with *mem* shows that the new predictor, *resi.m*, is independent of the other predictor, *mem*. When *resi.m* is correlated with *rec*, the part r of -.037 is obtained. This part r shows that *resi.m*, the variability in *int* independent of *mem*, accounts uniquely for $-.037^2 = .0014$ (.14%) of the *total* variability in y. These correlations appear in the following analysis.

```
VARIABLE LABEL resi.m ''.
CORR rec mem resi.m.
                          rec   mem   resi.m
 mem    Pearson           .923  1     .000
 resi.m Pearson           -.037 .000  1
```

Regressing *rec* on *resi.m* also shows that this approach is equivalent to previous calculations for the part r. SS_{Reg} equals SS_{Change} from earlier work and R^2, the simple *r*, and the *part r* all equal earlier calculations of part r. Relevant to the distinction between part and partial rs (discussed later), note that the denominator for SS_{Change} is the *total* variability in *rec*, that is, SS_{Recall}.

```
REGRESS /STAT = DEFA ZPP /DEP = rec /ENTER resi.m.
 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .037(a) .001     -.141           9.212

 Model            Sum of Squares df Mean Square F    Sig.
 1     Regression .825           1  .825        .010 .924(a)
       Residual   594.063        7  84.866
       Total      594.889        8

 Model           Unstandardized            Standardized    t      Sig. Correlations
                 Coefficients              Coefficients
                 B             Std. Error Beta                      Zero-order Partial Part
 1     (Constant) 38.889        3.071                    12.664 .000
       resi.m     -.040         .404       -.037         -.099  .924 -.037      -.037   -.037
```

Although residual predictors work well for the strength of the unique contribution, observe that the significance in the preceding analysis is *not* correct. In removing X2 from X1, overlap with areas c and d was also removed. So the error above represents a+c+d rather than area a as in the multiple regression. The denominator is too large and produces F and t statistics that are smaller, as shown by the fact that Sig. above equals .924 versus .820 in the earlier multiple regression analyses.

Standardized Regression Coefficients

Although not a measure of strength in the same sense as correlation coefficients, the magnitude of the change in y as reflected in the regression coefficients, can also be used to compare the contribution of different predictors. Which predictor produces the largest change in the dependent variable? But unstandardized coefficients are generally less than ideal for such a comparison because the amount of change in y depends not only on the relationship between y and the predictor, but also on the variation in predictors. Specifically, a predictor with a large range of values could have a greater impact on y even if its regression coefficient was smaller than a predictor with a small range of values. For example, a predictor with values ranging from 1 to 5 and a slope of 5.0 produces predicted scores ranging from 5 to 25, whereas a predictor with values ranging from 5 to 100 and a smaller slope of 1.0 produces predicted scores from 5 to 100, a much larger change in y.

The solution is to use standardized regression coefficients based on predictor and criterion variables with equivalent variability (i.e., s = 1.0). The problem with unstandardized coefficients and the solution is illustrated below. The descriptive statistics show that s for *int* is much larger than s for *mem*. This difference needs to be eliminated to compare fairly the contribution of the two predictors. Calculating z or normalized scores for the predictor and criterion variables (i.e., subtracting the mean from the scores and dividing by the standard deviation) produces $\bar{z} = 0$, $s_z = 1$. The following analyses illustrate the process. Note that default regression analyses also produce standardized regression coefficients.

```
DESCR rec mem int.
                N Minimum Maximum Mean   Std. Deviation
 rec            9 25      51      38.89  8.623
 mem            9 12      27      20.89  5.183
 int            9 88      118     102.67 11.456

COMPUTE zrec = (rec -  38.8889)/8.6233.                    z = (y - M)/SD
COMPUTE zmem = (mem -  20.8889)/5.1828.
COMPUTE zint = (int - 102.6667)/11.4564.

DESCR zrec zmem zint.

                N Minimum Maximum Mean     Std. Deviation
 zrec           9 -1.6106 1.4045  -.000001 .9999987       All Ms = 0
 zmem           9 -1.7151 1.1791  -.000002 .9999943       All SDs = 1
 zint           9 -1.2802 1.3384  -.000003 1.0000034
```

```
REGRESS /DEP = rec /ENTER mem int.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.923(a)	.853	.804	3.822

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	507.258	2	253.629	17.366	.003(a)
	Residual	87.631	6	14.605		
	Total	594.889	8			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.605	12.979		.740	.487
	mem	1.598	.371	**.960**	4.311	.005
	int	-.040	.168	**-.053**	-.238	.820

```
REGRESS /DEP = zrec /ENTER zmem zint.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.923(a)	.853	.804	.4431800

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.822	2	3.411	17.366	.003(a)
	Residual	1.178	6	.196		
	Total	8.000	8			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.16E-007	.148		.000	1.000
	zmem	**.960**	.223	.960	4.311	.005
	zint	**-.053**	.223	-.053	-.238	.820

Standardized coefficients are interpreted in terms of standard deviations. Specifically, a standardized coefficient indicates how much y changes in standard deviation units given a one standard deviation change in the predictor. Given a one standard deviation change in *int*, *rec* will decrease by .053 SDs. Given a one standard deviation change in *mem*, *rec* will increase by .960 SDs. These values are still far apart, but closer than were the unstandardized coefficients, -.040 for *int* and 1.598 for *mem*. Adjusting for different variability in the predictors has changed somewhat their relative magnitude. The change could be even more substantial if the predictor with the larger unstandardized coefficient had the smaller standardized coefficient, depending on the s for the respective predictors.

Partial Correlation Coefficient

A second coefficient representing the unique contribution of a

$$r^2_{y1.2} = \frac{SS_{\hat{y}.12} - SS_{\hat{y}.2}}{SS_y - SS_{\hat{y}.2}}$$

Box 6-1. Partial r

predictor is the partial correlation coefficient (see Box 6-1). Partial r is used less frequently and must be interpreted with caution. Recall that SS_{Change} can be calculated from SS_{Res} in Models 1 and 2, rather than SS_{Reg} because the increase in SS_{Reg} from Model 1 to Model 2 must be due to variability moving from SS_{Res} in Model 1 since SS_{Total} remains the same. For the unique contribution of *int*, $SS_{\hat{r}.im} = SS_{Res1} - SS_{Res2} = 88.457 - 87.631 = .826$, the same quantity calculated earlier in terms of SS_{Reg}. In terms of the Venn diagram, b = (a+b) - a. So SS_{Change} equals both the increase in SS_{Reg} when one predictor is added to the other, and the decrease in SS_{Res} when one predictor is added.

The partial correlation coefficient represents the percentage reduction in SS_{Res}. The partial correlation coefficient $r_{y1.2}$ (note absence of parentheses), is SS_{Change} divided by SS_{Res1} and can best be conceptualized in terms of SS_{Change} as the difference in SS_{Res}. In essence, $r^2_{y1.2}$ reflects the proportion of residual variability from Model 1 that is now accounted for by the predictor added in Model 2. That is, $r^2_{y1.2} = SS_{Change} / SS_{Res1} = (SS_{\hat{y}.12} - SS_{\hat{y}.2}) / (SS_y - SS_{\hat{y}.2})$. Note that for the partial correlation, $SS_{\hat{y}.2}$ is subtracted from both the numerator and denominator. For the part correlation, $SS_{\hat{y}.2}$ was subtracted only from the numerator. The partial correlation can be calculated from the same analyses used for the part correlation. Here is relevant output.

```
REGRESS /STAT = DEFAU CHANGE ZPP /DEP = rec /ENTER mem /ENTER int.
...
 Model            Sum of Squares df Mean Square F      Sig.
 1     Regression 506.432        1  506.432     40.076 .000(a)
       Residual   88.457         7  12.637
       Total      594.889        8

 2     Regression 507.258        2  253.629     17.366 .003(b)       SS(r̂.im) = 88.457 - 87.631
       Residual   87.631         6  14.605                                  = .826
       Total      594.889        8

 Model           Unstandardized            Standardized    t     Sig. Correlations
                 Coefficients              Coefficients
                 B              Std. Error Beta                      Zero-order Partial Part
...
 2     (Constant) 9.605          12.979                    .740  .487
       mem        1.598          .371       .960           4.311 .005 .923       .869    .675
       int        -.040          .168       -.053          -.238 .820 .630       -.097   -.037
                                                  r(ri.m)² = .826 / 88.457 = .0093
                                                  r(ri.m) = √.0093 = .0966
```

The partial r for *int* shows that intelligence accounts for .93% of the 88.457 units of variability in recall not already accounted for by mem, whereas the part (or semi-partial) *r*

indicates that intelligence accounts for .14% of the total variability in recall. The partial correlation appears in the ZPP portion of the regression output. The entry for *int*, -.097, agrees with the preceding calculations. The negative sign is added because the slope for *int* is negative. The partial *r* for *mem* would be obtained in a similar way. Partial *r*s must be interpreted with caution. The partial *r* for *int* is larger than the part *r* (although still modest because SS_{Change} was so small) not because its contribution to prediction of recall was stronger but because the contribution of *mem* alone was strong, which removed much variability in y from the denominator and inflated the apparent strength of *int*.

In terms of the Venn diagram in Figure 6-1, the part correlation is b/(a+b+c+d) and the partial is b/(a+b). Any difference between part and partial rs is due to the smaller denominator for the partial since the numerators are the same. In terms of residual predictors in Figure 6-3, the partial correlation is the correlation between the residual *int* and a residual *rec* variable obtained by regressing *rec* on *mem*, as in the following regression. Note below that the variability in residual scores for *rec* (*resr.m*) with variability predicted by *mem* partialled out, i.e., $SS_{resr.m} = 88.457 = SS_r - SS_{\hat{r}.m}$, the value used previously as the denominator for the partial *r* for *int*.

```
REGRESS /DEP = rec /ENTER mem /SAVE RESI(resr.m).
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.923(a)	.851	.830	3.555

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	506.432	1	506.432	40.076	.000(a)
	Residual	**88.457**	7	12.637		
	Total	594.889	8			

Correlating *resr.m* with *resi.m* produces -.097, the partial r for *int*. The partial r is greater than the part r because *resi.m* can predict a larger proportion of the residual variability in *rec* (i.e., *resr.m*) than of the total variability in *rec*. The correlation matrix also shows that *resr.m* is uncorrelated with *mem*, the predictor used to generate the residual scores, and the subsequent regression shows the relationship of this approach to the SSs used earlier to calculate the partial correlation.

```
CORR rec mem resi.m resr.m.
```

		rec	mem	resi.m	
resi.m	Pearson	**−.037**	.000		*part r*
resr.m	Pearson	.386	.000	**−.097**	*partial r*

```
REGRESS /STAT = DEFA ZPP /DEP = resr.m /ENTER resi.m.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.097(a)**	**.009**	−.132	3.53818241

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	**.825**	1	.825	.066	.805(a)
	Residual	87.631	7	12.519		
	Total	**88.457**	8			

This completes our introduction to MR based on two predictors. The material generalizes readily to the overall strength and significance of more than two predictors and the strength and significance of the unique contribution of individual predictors when there are more than two predictors. New material generally concerns how to apply MR to specific situations, such as nonlinear relationships and interactions.

APPENDIX 6-1

The following analyses compute residual scores that reproduce the part and partial rs for *mem* controlling for *int*. Work through the commands and output.

```
REGRESS /DEP = mem /ENTER int /SAVE RESI(resm.i).

 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .711(a) .505     .435            3.897

 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 108.589        1  108.589     7.151 .032(a)
       Residual   106.300        7  15.186
       Total      214.889        8
...
Residuals Statistics(a)
                 Minimum Maximum Mean  Std. Deviation N
 Predicted Value 16.17   25.82   20.89 3.684          9
 Residual        -5.459  6.255   .000  3.645          9

REGRESS /STAT = DEFA ZPP /DEP = rec /ENTER resm.i.

 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .675(a) .456     .378            6.798

 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 271.380        1  271.380     5.872 .046(a)
       Residual   323.509        7  46.216
       Total      594.889        8

 Model           Unstandardized            Standardized    t      Sig. Correlations
                 Coefficients              Coefficients
                 B             Std. Error  Beta                      Zero-order Partial Part
 1     (Constant) 38.889        2.266                    17.161 .000
       resm.i     1.598         .659        .675         2.423  .046 .675       .675    .675

REGRESS /DEP = rec /ENTER int /SAVE RESI(resr.i).
 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .630(a) .397     .310            7.162
 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 235.878        1  235.878     4.599 .069(a)
       Residual   359.011        7  51.287
       Total      594.889        8
...
Residuals Statistics(a)
                 Minimum Maximum Mean  Std. Deviation N
 Predicted Value 31.94   46.16   38.89 5.430          9
 Residual        -6.937  11.271  .000  6.699          9

VARIABLE LABEL resm.i ''  resr.i ''.
CORR rec mem resm.i resr.i.
                       rec  mem  resm.i
 mem    Pearson        .923
 resm.i Pearson        .675 .703
 resr.i Pearson        .777 .611 .869
```

```
REGRESS /STAT = DEFA ZPP /DEP = resr.i /ENTER resm.i.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.869(a)	.756	.721	3.53818241

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	271.380	1	271.380	21.678	.002(a)
	Residual	87.631	7	12.519		
	Total	359.011	8			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	1.11E-015	1.179		.000	1.000			
	resm.i	1.598	.343	.869	4.656	.002	.869	.869	.869

CHAPTER 7 - MULTIPLE PREDICTORS & AUTOMATED SELECTION

Let p equal the number of predictors in a study. Previous material covered p = 2, but most of the conceptualization generalizes to p ≥ 2. Researchers want to examine the overall relationship between the dependent variable and all predictors collectively, and also examine the unique contribution of individual predictors. The general form of the equation is: $\hat{y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_pX_p$, where the ellipses (...) represent all predictors between the third and the last. The first example includes three predictors.

Researchers for Child and Family Services needed to determine whether involvement with family services had a beneficial effect on children, specifically whether CFS involvement led children to be less likely to engage in delinquent activities as adolescents. Previous research failed to obtain a significant correlation between degree of involvement with family services and later delinquency, suggesting that services were not effective. To further examine this relationship, the researchers obtained data for 15 adolescents on their delinquent activities as adolescents (*deli*), how deprived their home environment was of normal parental control (*home*), how impulsive the adolescents were as children (*impu*), and their involvement with family services (*serv*). The MR equation will be: $\hat{d} = b_0 + b_s \times s + b_i \times i + b_h \times h$.

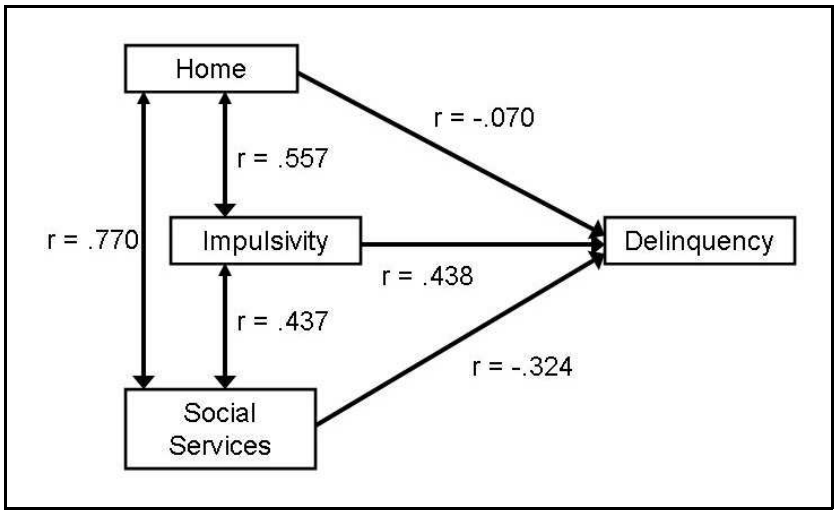


Figure 7-1. Relations among the four variables.

The results of the study appear below, followed by some descriptive statistics.

S	DELI	IMPU	SERV	HOME
1	20	44	38	27
2	24	42	39	25
3	21	52	41	28
4	25	60	41	36
5	22	35	24	21
6	20	51	44	29
7	37	54	29	22
8	22	45	35	21
9	26	67	45	34
10	23	55	33	25
11	29	52	32	29
12	20	45	37	23
13	31	57	32	23
14	22	32	28	22
15	23	43	45	33

```
CORR /VARI deli TO home /STAT.
```

```
      Mean     Std. Deviation N
 deli 24.33    4.761          15
 impu 48.93    9.331          15
 soci 36.20    6.483          15
 home 26.53    4.882          15

                      deli  impu serv
 impu Pearson         .438
      Sig. (2-tailed) .103

 serv Pearson         -.324 .437
      Sig. (2-tailed) .239  .103

 home Pearson         -.070 .557 .770
      Sig. (2-tailed) .805  .031 .001
```

As found in previous research, the simple correlation shows little evidence that involvement with social services reduced delinquency, r_{ds} = -.324, p = .239. The relationship is in the expected direction (negative), but is weak and far from significant. Impulsivity is also unrelated to delinquency, r_{di} = .438, p = .103, as is home environment, r_{dh} = -.070, p = .805. The relationships are complicated, however, because the three predictors correlate positively with one another. The various relationships and simple rs appear in Figure 7-1. A multiple regression analysis with all three predictors follows. We consider first the overall relationship between delinquency and all three predictors, then the unique contribution of each predictor.

Overall Relationship between Delinquency and Three Predictors

When there are more than two predictors, the complexity of the relationships among the predictors and the dependent variable requires more sophisticated mathematics to calculate the best-fit regression coefficients and are not considered in this course. Think of all the direct and indirect pathways in Figure 7-1 that must be "controlled" for. Given the regression equation from SPSS, however, the steps for the overall relationship are identical to analyses with two predictors. The equation generates predicted and residual scores, which are used to calculate statistics that reflect the significance and strength of the overall relationship. The Unstandardized Coefficients give the following equation:

$$\hat{d} = 23.275 + .357 \times i + .062 \times h - .498 \times s.$$

```
REGRE /DEP = deli /ENTER impu home serv  /SAVE PRED(prdd.ihs) RESI(resd.ihs).
```

```
 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .722(a) .522     .391            3.71535          R² = 165.491/317.333 = .522

 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 165.491        3  55.164      3.996 .038(a) MS_Reg = 165.491/3  = 55.164
       Residual   151.842        11 13.804                    MS_Res = 151.842/11 = 13.804
       Total      317.333        14                           F = 55.164/13.804 = 3.996
```

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	**23.275**	6.452		3.607	.004
	impu	**.357**	.128	.699	2.783	.018
	home	**.062**	.346	.063	.178	.862
	serv	**−.498**	.240	−.678	−2.073	.062

Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N	
Predicted Value	18.2317	29.4480	24.3333	**3.43814**	15	$SS_{Reg}=(15-1)*3.43814^2=165.491$
Residual	−4.99704	7.55203	.00000	**3.29330**	15	$SS_{Res}=(15-1)*3.29330^2=151.842$

The best-fit regression equation can generate predicted and residual scores for each participant. These are listed below. Standard deviations for for predicted and residual scores can be used to calculate SS_{Reg} and SS_{Res}, as shown to the right of the Residuals Statistics printout. These agree with the SSs in the ANOVA section. Note that despite having more than two predictors, the analysis partitions SS_{Total} into just two components, what can be predicted and what cannot be predicted. That is, $SS_{Total} = SS_{Reg} + SS_{Res}$. This is true no matter how many predictors there are.

$$H_0: \rho^2_{y.123...p} = 0$$
$$H_a: \rho^2_{y.123...p} \neq 0$$
$$F = \frac{\frac{MS_{Reg}}{p}}{\frac{MS_{Res}}{n-p-1}}$$
$$df = p, n-p-1$$

Box 7-1.

The SSs can be used to calculate $R^2_{d.his}$ and F, which reflect the strength and significance of the overall relationship. These statistics require no new formula. Collectively, the three predictors account for 52.2% of the total variability in delinquency, and the F test in Box 7-1 warrants rejection of the null hypothesis of no overall relationship in the population between delinquency and the three predictors; that is, we reject H_0: $\rho_{d.his} = 0$. As with two predictors, this test is nondirectional; only individual predictors can be positive or negative.

LIST.

				y	$\hat{y}$	$y-\hat{y}$
s	deli	impu	serv	home	prdd.ihs	resd.ihs
1.00	20.00	44.00	38.00	27.00	21.70558	−1.70558
2.00	24.00	42.00	39.00	25.00	20.37075	3.62925
3.00	21.00	52.00	41.00	28.00	23.12667	−2.12667
4.00	25.00	60.00	41.00	36.00	26.47348	−1.47348
5.00	22.00	35.00	24.00	21.00	25.09908	−3.09908
6.00	20.00	51.00	44.00	29.00	21.33717	−1.33717
7.00	37.00	54.00	29.00	22.00	29.44797	7.55203
8.00	22.00	45.00	35.00	21.00	23.18698	−1.18698
9.00	26.00	67.00	45.00	34.00	26.85485	−.85485
10.00	23.00	55.00	33.00	25.00	27.99704	−4.99704
11.00	29.00	52.00	32.00	29.00	27.67145	1.32855
12.00	20.00	45.00	37.00	23.00	22.31396	−2.31396
13.00	31.00	57.00	32.00	23.00	29.08540	1.91460
14.00	22.00	32.00	28.00	22.00	22.09798	−.09798
15.00	23.00	43.00	45.00	33.00	18.23165	4.76835

The correlation matrix below demonstrates the same relationships as regression with one and two predictors. Residual scores correlate 0 with predicted scores and with all three predictors. The correlation between the original delinquency score and the predicted

$$r_{\hat{y}(y-\hat{y})} = 0$$
$$r_{x(y-\hat{y})} = 0$$
$$r_{y\hat{y}} = R_{y.12...p}$$
$$r_{y(y-\hat{y})} = \sqrt{1-R^2}$$

Box 7-2.

scores is the multiple R, and the correlation between the delinquency score and the residual score is the square root of 1 - R^2. These relationships reflect the partitioning of SS_{Total} into what can and cannot be predicted by *impu*, *serv*, and *home*, and appear in Box 7-2.

```
VARI LABEL prdd.ihs '' resd.ihs ''.
CORR deli TO resd.ihs /STAT.
```

```
          Mean       Std. Deviation N
deli      24.3333    4.76095        15          ȳ=24.3333
impu      48.9333    9.33095        15
serv      36.2000    6.48294        15
home      26.5333    4.88243        15
prdd.ihs 24.3333333 3.43814203     15          ŷ = ȳ
resd.ihs .0000000   3.29330321     15          Σ(y-ŷ)=0

            deli  impu  serv  home  prdd.ihs
impu         .438
serv        -.324  .437
home        -.070  .557  .770
prdd.ihs     .722  .606 -.449 -.096
resd.ihs     .692  .000  .000  .000  .000
```

Unique Contribution

The preceding analyses examined the strength and significance of the overall relationship between delinquency (*deli*) and the three predictors: impulsivity (*impu*), home environment (*home*), and degree of involvement with social services (*serv*). Additional analyses concern the strength and significance of the unique contribution of each predictor, below the unique contribution of *serv* controlling for *impu* and *home*. The formula in Box 7-3 measure the strength and significance of the unique contribution of a predictor controlling for any number of other predictors.

The approach extends that used with two predictors. To calculate SS_{Change} for three predictors, the two control predictors are entered first and then the target predictor is added, *serv* in our example. The increase or change in SS_{Reg} reflects the unique contribution of the added predictor *serv*. SS_{Change} is used to calculate a part r^2 that reflects the strength of the unique contribution and an F_{Change} to test the significance of the unique contribution.

$$SS_{Change} = SS_{\hat{y}1.23\ldots p} = SS_{\hat{y}.123\ldots p} - SS_{\hat{y}.23\ldots p}$$

$$r^2_{y(1.23\ldots p)} = \frac{SS_{\hat{y}1.23\ldots p}}{SS_y} = R^2_{y.123\ldots p} - R^2_{y.23\ldots p}$$

$$MS_{Change} = \frac{SS_{Change}}{1} \qquad F_{y(1.23\ldots p)} = \frac{MS_{Change}}{MS_{Residual}}$$

Box 7-3. Change Statistics.

```
REGRE /STAT = DEFAU ZPP CHANGE /DEP = deli /ENTER impu home /ENTER serv.
```

```
Model R       R      Adjusted Std. Error of Change Statistics
              Square R Square the Estimate  R Square Change F Change df1 df2 Sig. F Change
1     .578(a) .335   .224     4.19496       .335            3.016    2   12  .087
2     .722(b) .522   .391     3.71535       .187            4.298    1   11  .062
```

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	**106.161**	2	53.081	3.016	.087(a)
	Residual	211.172	12	17.598		
	Total	317.333	14			
2	Regression	**165.491**	3	55.164	3.996	.038(b)
	Residual	151.842	11	13.804		
	Total	317.333	14			

SS_{Change}=165.491−106.161=59.33

$r^2_{d(s.hi)}$=59.33/317.333 = .187
=.522−.335 = .187

F_{Change} =(59.33/1)/13.804
= 4.298

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	18.843	6.873		2.741	.018			
	impu	.353	.145	.692	2.438	.031	.438	.576	.574
	home	−.444	.277	−.455	−1.605	.135	−.070	−.420	−.378
2	(Constant)	23.275	6.452		3.607	.004			
	impu	.357	.128	.699	2.783	.018	.438	.643	.580
	home	.062	.346	.063	.178	.862	−.070	.054	.037
	serv	**−.498**	**.240**	−.678	**−2.073**	**.062**	−.324	−.530	**−.432**

The part r is shown in the ZPP section, $r_{d(s.hi)} = \sqrt{.187} = (-).432$, with the negative sign obtained from the regression coefficient for *serv* (i.e., -.498). When SS_{Change} is based on the addition of a single predictor, as here, a test of the significance of the regression coefficient is equivalent to F_{Change}. The formula for the standard error (SE) must be adjusted for additional predictors, as shown in Box 7-4. The denominator for SE represents the variability in a predictor (SS_{serv} in the present example) that is independent of other predictors in the equation. The previous version with two predictors used $SS_1 \times (1-r^2_{12})$, but with more than two predictors a multiple R^2 is required to remove the variability shared with all other predictors. The following regression performs this operation to give $R^2_{s.hi} = .594$. The /STAT = R option limits the output as we only need $R^2_{s.hi}$. The SS for *serv* can be calculated from its standard deviation, shown earlier, $SS_{serv} = (15-1) \times 6.42894^2 = 578.638$. The final calculations for SE are in Box 7-4.

$$SE_{b_1} = \sqrt{\frac{MS_{Residual}}{SS_1 \times (1 - R^2_{1.23...p})}} = \sqrt{\frac{13.804}{578.638(1-.594)}} = .242$$

Box 7-4.

This gives, $t = -.498/.240 = -2.075$ and $t^2 = 2.073^2 = 4.297 = F_{Change}$. As well, $p_{b1} = .062$ is identical to p for F_{Change}. The two tests are equivalent because F_{Change} tests the significance of a single predictor; that is, the $df_{Numerator}$ for F is 1.

```
REGRE /STAT = R /DEP = serv /ENTER impu home /SAVE RESI(ress.ih).
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.770(a)	**.594**	.526	4.46381

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	30.4774	45.8917	36.2000	4.99495	15
Residual	−6.72673	5.28101	.00000	4.13268	15

Because the regression includes /SAVE RESI(ress.ih), the preceding regression also gives a residual

serv predictor, which can be used to calculate $r_{d(s.hi)}$, the correlation between the dependent variable *deli* and the variability in *serv* that is independent of *home* and *imp*, as shown in the following correlation matrix.

```
VARIABLE LABEL ress.ih ''.
CORR ress.ih WITH deli impu home.

                          deli  impu  home
 ress.ih Pearson          -.432 .000  .000
```

The partial correlation coefficient can be calculated using a slight modification of the procedures with two predictors. Conceptually, SS_{Change} is the decrease in SS_{Res} when the last predictor is added and can be used to calculate the percentage reduction in SS_{Res} from Model 1 to Model 2, which gives the results shown in Box 7-5. The value for partial r appears in the ZPP output. In terms of residual variables, the following regression computes the variability in *deli* that is independent of *home* and *impu*. The correlation of the residual *serv* and residual *deli* scores gives $r_{ds.hi}$.

$$SS_{Change} = 211.172 - 151.842 = 59.33$$
$$r^2_{ds.hi} = \frac{59.33}{211.172} = .281$$
$$r_{ds.hi} = \sqrt{.281} = (-).530$$

Box 7-5. Partial r

```
REGRE /STAT = R /DEP = deli /ENTER impu home /SAVE RESI(resd.ih).

 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .578(a) .335     .224            4.19496

                 Minimum  Maximum Mean    Std. Deviation N
 Predicted Value 19.3694  28.7482 24.3333 2.75372        15
 Residual        -4.51361 8.86663 .00000  3.88378        15       SS resd.ih = 211.172
```

```
VARIABLE LABEL resd.ih ''.
CORR deli impu home serv ress.ih resd.ih.

                          deli  impu  home  serv  ress.ih resd.ih
 impu    Pearson          .438
 home    Pearson          -.070 .557
 serv    Pearson          -.324 .437  .770
 ress.ih Pearson          -.432 .000  .000  .637                  r d{s.ih}
 resd.ih Pearson          .816  .000  .000  -.338 -.530           r ds.ih
```

The next two regressions demonstrate the difference between the part and partial rs. The part r^2 represents the proportion of the total variability in *deli* explained uniquely by *serv* and the partial r^2 represents the proportion of the variability not explained by the other predictors now explained uniquely by *serv*.

```
REGRE /DEP = deli /ENTER ress.ih.
 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .432(a) .187     .124            4.45493                  R = r d(s.ih)
```

Model		Sum of Squares	df	Mean Square	F	Sig.	
1	Regression	**59.330**	1	59.330	2.989	.107(a)	$SS_{Regression} = SS_{d's.ih}$
	Residual	258.003	13	19.846			
	Total	**317.333**	14				SS_{Deli}

```
REGRE /DEP = resd.ih /ENTER ress.ih.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	**.530**(a)	.281	.226	3.41762225	$R = r_{ds.ih}$

Model		Sum of Squares	df	Mean Square	F	Sig.	
1	Regression	**59.330**	1	59.330	5.080	.042(a)	
	Residual	151.842	13	11.680			
	Total	**211.172**	14				$SS_{Deli} - SS_{d'.ih}$

Visual Representations of Multiple Predictors

Regression with multiple predictors cannot be visualized as readily as with two predictors. Figure 7-2 shows a Venn diagram for the present study. Remember that it is symbolic and areas will not correspond to actual degrees of overlap among the variables.

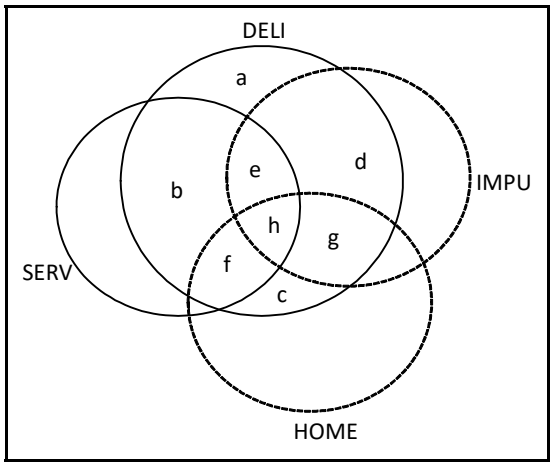


Figure 7-2. Venn Diagram.

Together, all three predictors account for the following portions of SS_d: b+c+d+e+f+g+h = $SS_{d.his}$. Portion b cannot be accounted for by *home* or *impu*, and represents the unique contribution of *serv*. To obtain b we remove the variability already predicted by *impu* and *home*: $SS_{d.hi}$ = c+d+e+f+g+h. This gives $SS_{d\ s.hi} = SS_{d.his} - SS_{d.hi}$.

The part r^2 is area b over the total variability in *deli*, that is, all the lettered areas. The partial r^2 is area b over a+b, that is, the variability in *deli* not predicted by *home* and *impu* alone.

Figure 7-3 shows a representation of the unique contribution and part r in terms of a residual *serv* predictor. A regression with *serv* as a dependent variable and *home* and *impu* as predictors creates a residual *serv* that is independent of the other predictors ($Res_{s.hi}$). The solid lines represent the regression. The correlations of the residual *serv* with *home* and *impu* are both 0, while the residual's correlation with *deli* gives $r_{d(s.hi)}$, the part r.

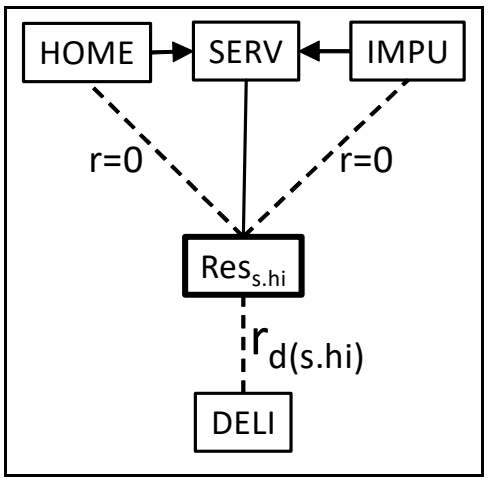


Figure 7-3. Residual Predictor

Although Venn diagrams and residual predictors, especially the former, become extremely complex, they serve as metaphors for the unique contribution of a predictor. No matter how many other

predictors (circles) there are overlapping with the dependent variable, adding another predictor (circle) to the Venn diagram could account for variability not already accounted for by predictors already in the equation. And the total of all areas covered represents the overall relationship between the dependent variable and the predictors.

With respect to a residual predictor, one predictor can be regressed on all other predictors to create a residual predictor that is independent of all the other predictors (i.e., rs = 0). The correlation between the residual and the dependent variable will be the part r for that predictor, that is $r_{y(1.23...p)}$.

Given such statistics, researchers must make sense of relationships among the variables. It appears that delinquency can be predicted by both impulsivity (positive relationship) and involvement with social services (negative relationship), but home deprivation makes no direct contribution. Home deprivation, however, appears to increase both impulsivity and involvement with social services, which explains their positive correlation and masking due to the opposite direction of their influence on delinquency. Below is the two-predictor regression analysis, without *home*, as well as plots of actual and predicted scores.

```
REGR /DEP = deli /ENTER impu serv.
 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .721(a) .520     .440            3.56231

 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 165.053        2  82.526      6.503 .012(a)
       Residual   152.281        12 12.690
       Total      317.333        14

 Model            Unstandardized             Standardized    t      Sig.
                  B               Std. Error Beta
 1     (Constant) 23.385          6.157                      3.798  .003
       impu       .366            .113       .716            3.222  .007
       serv       -.468           .163       -.637           -2.866 .014
```

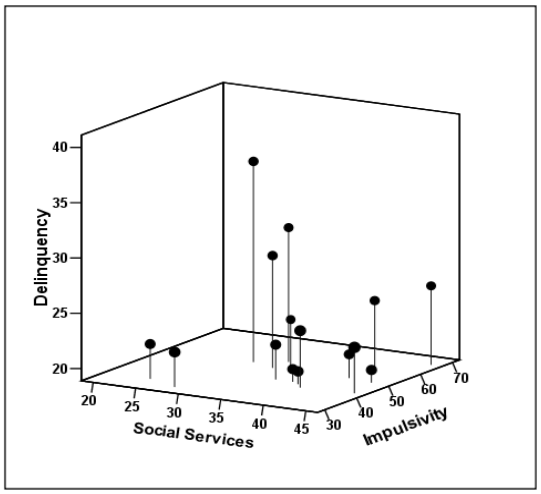


Figure 7-4. Graph of observed values.

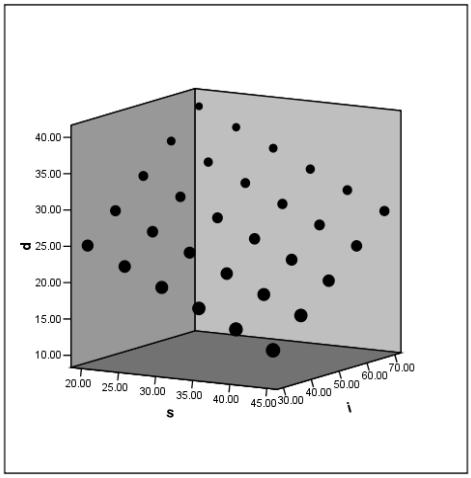


Figure 7-5. Graph of two-predictor equation.

Automated Procedures for Entering Predictors

Sometimes researchers only include significant predictors in the final equation as in the previous analysis without *home*. A predictor may not improve prediction or it could reduce the unique contribution of other predictors. Note in the preceding analysis that *serv* became more significant than it was with *home* in the equation. SPSS and other statistical packages have automated procedures for selecting equations that exclude nonsignificant predictors. Such automated procedures must be used cautiously because they consider only statistical criteria and not what the variables represent or hypotheses about the underlying structure. The three procedures that we consider are FORWARD, BACKWARD, and STEPWISE. In general, the procedures are based on whether the unique contribution of a predictor is significant; that is, whether the t-test for the regression coefficient would be significant if it was in the equation with other predictors.

FORWARD regression enters variables one at a time IF their unique contribution is the most significant of all predictors not yet in the equation AND their *p* value if entered is less than PIN (default .05, but can be modified). Here is the correlation matrix for the delinquency dataset. No predictor is entered using the default value for PIN because no predictor has a significance of .05 or smaller. However, *impu* and *serv* are entered if PIN is set higher than .103, the *p* for *impu*, which is the predictor closest to being entered. Because SPSS will choose what predictors to enter or exclude, all variables must be listed on the /VARIABLE option, the dependent variable is identified with /DEPENDENT, and remaining variables are treated as potential predictors.

```
CORR /VARI deli TO home.
       deli  impu  serv
 impu  .438
       .103

 serv -.324  .437
       .239  .103

 home -.070  .557  .770
       .805  .031  .001
```

```
REGRE /VARIABLE = impu serv home deli /STAT = DEFAU ZPP /DEP = deli /FORWARD.

Variables Entered/Removed(a)
 a Dependent Variable: deli
```

```
REGRE /VARI = impu serv home deli /STAT = DEFAU ZPP /CRITERIA = PIN(.11)
 /DEP = deli /FORWARD.

Variables Entered/Removed(a)
 Model Variables       Variables  Method
       Entered         Removed
 1     impu            .          Forward (Criterion: Probability-of-F-to-enter <= .110)
 2     serv            .          Forward (Criterion: Probability-of-F-to-enter <= .110)
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.438(a)	.192	.130	4.44184
2	.721(b)	.520	.440	3.56231

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	60.845	1	60.845	3.084	.103(a)
	Residual	256.489	13	19.730		
	Total	317.333	14			
2	Regression	165.053	2	82.526	6.503	.012(b)
	Residual	152.281	12	12.690		
	Total	317.333	14			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	13.401	6.330		2.117	.054			
	impu	.223	.127	.438	1.756	.103	.438	.438	.438
2	(Constant)	23.385	6.157		3.798	.003			
	impu	.366	.113	.716	3.222	.007	.438	.681	.644
	serv	−.468	.163	−.637	−2.866	.014	−.324	−.637	−.573

Excluded Variables(c)

Model		Beta In	t	**Sig.**	Partial Correlation	Collinearity Statistics Tolerance
1	serv	−.637(a)	−2.866	**.014**	−.637	.809
	home	−.455(a)	−1.605	**.135**	−.420	.689
2	home	.063(b)	.178	**.862**	.054	.346

Once *impu* is in the equation, SPSS considers the revised significance of the remaining predictors to find which is most significant and below PIN = .11 when added to *impu*. Relevant statistics are shown in the Excluded Variables section of the output. If *serv* was entered second, its *p* value would be .014. If *home* was entered second, its *p* value would be .135. SPSS enters *serv* as it has a *p* less than .11 and is more significant than *home*. Note that the statistics for *serv* once in the main regression with *impu* are the same as in the Excluded Variables section. Once *serv* and *impu* are in the equation, *home* would not contribute further to prediction of delinquency if it was added to the equation given its p value would be *p* = .862.

BACKWARD regression works in reverse. First, all predictors are entered into the equation and then predictors are removed one at a time IF the *p* value for that predictor is the least significant of all predictors in the equation AND the *p* value is greater than POUT (by default .10, but can be changed). The backward results follow. In Model 1 with all three predictors, *home* has the weakest unique relationship and *p* = .862 is greater than .10. It is removed, after which the *p* values for both *impu* and *serv* are less than .10 and remain in the final equation.

```
REGRE /VARI = impu serv home deli /STAT = DEFAU ZPP /DEP = deli /BACKWARD.

Variables Entered/Removed(b)
 Model Variables       Variables Method
       Entered         Removed
 1     home, impu,serv .         Enter
 2     .               home      Backward (criterion: Probability of F-to-remove >= .100).

 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
 1     .722(a) .522     .391            3.71535
 2     .721(b) .520     .440            3.56231

 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 165.491        3  55.164      3.996 .038(a)
       Residual   151.842        11 13.804
       Total      317.333        14

 2     Regression 165.053        2  82.526      6.503 .012(b)
       Residual   152.281        12 12.690
       Total      317.333        14

 Model           Unstandardized            Standardized    t      Sig. Correlations
                 Coefficients              Coefficients
                 B              Std. Error Beta                      Zero-order Partial Part
 1     (Constant) 23.275         6.452                    3.607  .004
       impu       .357           .128      .699           2.783  .018 .438       .643    .580
       serv       -.498          .240      -.678          -2.073 .062 -.324      -.530   -.432
       home       .062           .346      .063           .178   .862 -.070      .054    .037

 2     (Constant) 23.385         6.157                    3.798  .003
       impu       .366           .113      .716           3.222  .007 .438       .681    .644
       serv       -.468          .163      -.637          -2.866 .014 -.324      -.637   -.573

...
```

To understand these automated procedures, remember that the relevant *p* values are revised for each predictor every time a new equation is created. The relevant *p* value for *serv* in the FORWARD procedure, for example, was .014, its *p* value when *impu* is already in the equation, not .239, its *p* value alone. Similarly, the relevant *p* value for *serv* in the BACKWARD procedure was again .014, its *p* value when only *impu* and *serv* predictors were in the equation, not .062, its *p* value when all three predictors were in the equation.

A third procedure, STEPWISE, combines features of the FORWARD and BACKWARD methods. STEPWISE begins by entering predictors one at a time using the PIN criterion. But after each new predictor has been added, it reviews all of the predictors in the equation to see if any of the *p* values now exceed POUT and should be removed. Here the STEPWISE procedure leads to the same results as FORWARD because both predictors remain significant using PIN = .11. There are occasions, however, when entry of subsequent predictors decreases the p value for earlier predictors to above POUT. In terms of a Venn diagram, predictors added later could overlap with what was earlier the unique contribution of a predictor.

A limitation of automated procedures is that they do not consider all possible equations. Because variables are entered or removed successively, only some of the 2^p possible equations (including one with 0 predictors) are considered. Sometimes it is useful to examine all possible equations to determine the "best"

equation, as shown below for the delinquency study (# equations = $2^p - 1 = 2^3 - 1 = 7$, plus a null equation). In the present case, the equation with just *impu* and *serv* appears satisfactory. In other studies, however, there may be a "better" (i.e., stronger or more interpretable) combination of predictors.

```
 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
REGRE /STAT = R /DEP = deli /ENTER impu.
 1     .438(a) .192     .130            4.44184
REGRE /STAT = R /DEP = deli /ENTER serv.
 1     .324(a) .105     .036            4.67417
REGRE /STAT = R /DEP = deli /ENTER home.
 1     .070(a) .005     -.072           4.92867
REGRE /STAT = R /DEP = deli /ENTER impu serv.
 1     .721(a) .520     .440            3.56231
REGRE /STAT = R /DEP = deli /ENTER impu home.
 1     .578(a) .335     .224            4.19496
REGRE /STAT = R /DEP = deli /ENTER serv home.
 1     .430(a) .185     .049            4.64344
REGRE /STAT = R /DEP = deli /ENTER impu serv home.
 1     .722(a) .522     .391            3.71535
```

Multiple Confounded Predictors

Understanding multiple predictors in a regression can be a challenge, but at the same time benefit our understanding of psychological phenomena. For example, a predictor can appear to be very strong on its own, but become quite weak when analyzed with other predictors. The challenge is to appreciate why this happens. It can be understood, for example, in terms of residual predictors or Venn diagrams as presented previously. It can be nicely illustrated by an old crime dataset.

Crime rates vary markedly across the USA states. Forensic psychologists examined a data set that included: crime rate (*crime*), percentage of the population living in metropolitan areas (*pm*), percentage white (*pw*), percentage high school graduates (*hs*), percentage living in poverty (*pv*), and percentage single parent families (*sp*). Preliminary analyses indicated that 2 of the 51 states should be excluded as outliers; they had extreme values on *crime*, *pw*, or both.

Researchers hypothesized that the simple correlation of *crime* with *pw* was extremely misleading because it was confounded with so many measures of societal disadvantage. Also the direction of any relationship was ambiguous since white Americans could move out of high crime areas. The *hs* statistic is a good measure of education level in previous generations when high school graduation was less universal than today. The predictors correlate highly with *crime* and with one another; notably *pw* correlates with all four other predictors, especially *sp*.

```
CORR crime pw pm pv hs sp /STAT /MISS = LISTWISE.
```

	Mean	Std. Deviation
crime	572.90	295.603
pw	86.0653	9.12696
pm	66.5755	21.86913
pv	14.1388	4.24135
hs	76.2082	5.66196
sp	11.1510	1.46132

	crime	pw	pm	pv	hs
pw	-.684				
	.000				
pm	.610	-.293			
	.000	.041			
pv	.350	-.434	-.148		
	.014	.002	.310		
hs	-.287	.508	.008	-.773	
	.046	.000	.956	.000	
sp	.639	-.686	.171	.407	-.222
	.000	.000	.239	.004	.125

On its own, *pw* is a is a strong predictor and indeed would be the first predictor entered in a regression using the FORWARD option. However, a regression with all predictors reveals that the relationship between *crime* and *pw* becomes quite weak when other predictors are controlled, as indicated by its part r squared: $r^2_{cw.mvsp} = -.184^2 = .03$. The significance has also been much reduced; by itself, the significance is .000 in the correlation matrix above versus .020 in the regression alone, which benefits from a very small error term given predictors collectively account for 75% of the variability in crime rates.

```
REGRESS /STAT = DEFAU ZPP /DEP = crime /ENTER pw pm pv hs sp.
```

Model	R	R Square
1	.867	.751

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3151554.002	5	630310.800	25.993	.000
	Residual	1042736.488	43	24249.686		
	Total	4194290.490	48			

		Unstandardized Coefficients		Standardized Coefficients			Correlations	
Model		B	Std. Error	Beta	t	Sig.	Zero-order	Part
1	(Constant)	-599.783	656.310		-.914	.366		
	pw	-10.125	4.193	-.313	-2.415	.020	-.684	-.184
	pm	7.090	1.137	.525	6.234	.000	.610	.474
	pv	23.113	9.719	.332	2.378	.022	.350	.181
	hs	9.265	7.425	.177	1.248	.219	-.287	.095
	sp	48.353	24.191	.239	1.999	.052	.639	.152

This pattern can be interpreted in terms of residual predictors. Most of what *pw* alone accounts for is due to its considerable overlap with other predictors, as shown in the correlation matrix and the following regression; *pm*, *pv*, *hs*, and *sp* account for 65.5% of the variability in *pw*.

```
REGRESS /DEP = pw /ENTER pm pv hs sp /SAVE RESID(respw).
```

Model	R	R Square
1	.809	.655

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2619.093	4	654.773	20.886	.000
	Residual	1379.378	44	31.349		
	Total	3998.471	48			

		Unstandardized Coefficients			
Model		B	Std. Error	t	Sig.
1	(Constant)	62.605	21.628	2.895	.006
	pm	-.067	.040	-1.684	.099
	pv	.443	.343	1.292	.203
	hs	.855	.234	3.654	.001
	sp	-3.900	.641	-6.083	.000

	Mean	Std. Deviation	N
Predicted Value	86.0653	7.38677	49
Residual	.00000	5.36069	49

The overlap with other predictors is removed from the residual *pw* variable (*respw*) and reveals the weak unique relationship between *crime* and *pw*, as shown by the part r for *pw* in the ZPP output above and the correlation matrix below. Note as well that the residual predictor correlates 0 with the four other predictors (i.e., the residual variability is unique to *pw*). Although the correlation is not significant in the correlation matrix, recall that the test uses a denominator that includes variability in *crime* that would be accounted for by the other predictors in the multiple regression. That is, the error is inflated.

```
VARI LABEL respw ''.
CORR respw WITH crime pm pv hs sp.
```

	crime	pm	pv	hs	sp
respw	-.184	.000	.000	.000	.000
	.207	1.000	1.000	1.000	1.000

Venn diagrams can also be used to conceptualize the difference between simple and part rs; that is, why the strongest predictor alone might become the weakest when other predictors are included in the regression. But Venn diagrams are not easy to represent when there are multiple predictors. Figure 7-6 demonstrates the principle for three of the crime predictors: *hs*, *pw*, and *pv*. The oval with a heavy solid line represents the dependent variable *crime* and the oval with a light solid line represents *pw*. There is a high degree of overlap between the two (areas c, e, f, and h). But one or both of the other predictors represented by the ovals with dashed lines overlap with most of this area (i.e., areas e, f, and h), leaving only c as the unique contribution of *pw*.

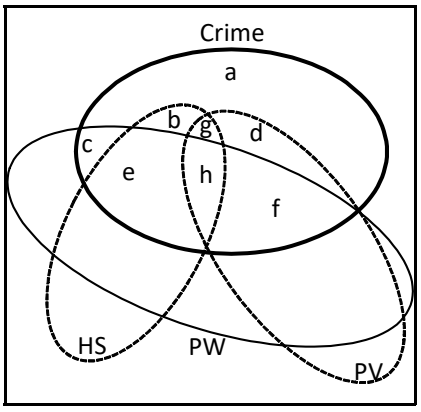


Figure 7-6.

The critical insight is to appreciate that the area represented by c, e, f, and h can be larger than the areas for the simple correlation of *crime* with *hs* (b, e, g, h) or *pv* (d, f, g, h). But as well, area c representing the unique contribution of *pw* can be weaker than the corresponding areas b for *hs* and d for *pv*. These relationships are summarized in Box 7-6.

	Alone	Unique
HS	b e g h	b
PW	**c e f h**	**c**
PV	d f g h	d

Box 7-6.

Although weak, the remaining relationship of *crime* with *pw* (i.e., c in Figure 9-12 and -.184 as a part r) might be explained by other factors not included in the regression, including some associated directly with ethnicity. For example, non-whites may experience discrimination or economic factors prevent non-whites moving as easily as whites to low crime communities.

Outcomes like that illustrated here for ethnicity can occur when a predictor captures much of the variability in other predictors that is related to the dependent variable. A socioeconomic status variable (SES), for example, might be a strong predictor by itself but not when education and occupation are included in a regression because SES is a measure determined in large part by education and occupation (or income). Predictors like ethnicity and SES, can be thought of as proxy variables for the collective effects of other predictors. Analyses with such predictors can be misleading. In an extreme case, for example, including SES as a predictor might lead to the wrong conclusion that education and occupation are unrelated to the dependent variable.

Given how difficult it can be to tease apart the unique effect of predictors given their complex correlations with one another and the dependent variable, alternative sophisticated statistical procedures are sometimes more appropriate, such as factor analysis or structural equation modelling (SEM). These procedures can identify and capture overlap of predictor and dependent variables. In essence, they identify a possible hypothetical factor that underlies shared variability but is not explicitly measured.

CHAPTER 8 - CATEGORICAL PREDICTORS

In addition to examining relationships between numerical dependent variables and numerical predictors, MR can accommodate categorical predictors that involve membership in two or more groups or categories, such as gender (male or female), religious affiliation (Protestant, Catholic, Jew, Muslim, ...), academic major (humanities, social sciences, ...), or different treatment conditions in an experiment (experimental versus control group). Sometimes the categories are ordered (e.g., young, middle-aged, and old adults), but often there is no meaningful ordering and numbers assigned to groups are arbitrary (e.g., male = 1 and female = 2; Protestant = 1, Catholic = 2, ...). In essence the numbers serve as labels rather than amounts.

The way to accommodate such categorical variables is to use k - 1 predictors, where k equals the number of groups and each predictor has two or more values that collectively define the groups. For two groups, a single predictor with two values (1 or 2, 0 or 1, -1 or +1, or whatever) is sufficient. For three groups, two predictors are required, and so on. Here we consider two groups, starting with the example of delinquency as a function of involvement with social services, but treating involvement with social services as a categorical variable with just two levels (e.g., none vs. some, or little vs. a lot, ...).

Categorical Predictor Only

Delinquency scores were obtained for 18 adolescents who had low or high levels of involvement with social services (9 in each group). One test for the difference between delinquency means of the two groups is the independent t-test presented earlier.

```
TTEST /GROUP = serv2 /VARI = deli.

      serv2 N Mean    Std. Deviation Std. Error Mean
 deli 1     9 25.2222 3.45607        1.15202
      2     9 22.7778 4.20648        1.40216

Independent Samples Test
                 Levene's Test for      t-test for Equality of Means
                 Equality of
                 Variances
                    F              Sig.   t         df     Sig.       Mean       Std. Error
                                                       (2-tailed) Difference Difference
 deli Equal variances .049          .827   1.347     16     .197       2.44444    1.81472
```

Calculations for the independent *t*-test are shown in Box 8-1. The H_0: $\mu_1 = \mu_2$ cannot be rejected, although the results "approach significance" by a directional test with $p = .197/2 = .0985$.

$$s_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(9-1)3.45607^2 + (9-1)4.20648^2}{9+9-2} = \frac{237.111}{16} = 14.8194$$

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{25.222 - 22.778}{\sqrt{14.8194(\frac{1}{9} + \frac{1}{9})}} = \frac{2.4444}{1.81472} = 1.347$$

Box 8-1. Independent Groups T-Test.

The significance can also be obtained by analysis of variance using an *F* test equivalent to *t*. Formula are summarized in Box 8-2 and calculations shown below.

```
GLM deli BY serv2 /PRINT = DESCR.
```

```
serv2 Mean    Std. Deviation N      ȳj−ȳG       SSȳ    = 9×(1.2222² + −1.2222²)
1     25.2222 3.45607        9      1.2222             = 9×2.9875 = 26.888
2     22.7778 4.20648        9     −1.2222      MSNum  = 26.888/(2−1) = 26.888
Total 24.0000 3.94074        18                 MSDen  = sp² = 14.819

Source          Type III SS    df Mean Square F       Sig.
Intercept       10368.000       1  10368.000   699.621 .000
serv2           26.889          1  26.889      1.814   .197      F = t²     pF = pt
Error           237.111         16 14.819
                                          SSError =(9−1)3.45607² + (9−1)4.20648² =237.111
Corrected Total 264.000         17        SSTotal = (18−1)3.94074² = 264.000
                                          SSTreatment = 264.000 − 237.111   =  26.889
```

The preceding analyses are equivalent to a regression of delinquency scores on a categorical predictor that represents different levels of involvement with social services. The regression analysis below shows that the regression coefficient equals the difference between means, predicted scores are the group means, and so on. The values of *serv2* are recoded to 0 and 1, which helps to interpret the regression results and shows that the significance test does not depend on the specific values used to "label" the groups. As *serv2* increases by 1 unit from 0 to 1, *deli* decreases by 2.444 units. If the group labels (0, 1) were reversed, the coefficient would be positive, but the magnitude would remain the same.

$$SS_{\bar{y}} = \sum n_j \times (\bar{y}_j - \bar{y}_G)^2$$

$$MS_{\bar{y}} = \frac{SS_{\bar{y}}}{k-1}$$

$$F = \frac{MS_{\bar{y}}}{s_p^2}$$

Box 8-2. F-test

```
RECODE serv2 (1 = 0) (2 = 1).
REGRESS /DEP = deli /ENTER serv2 /SAVE PRED(prdd.s2) RES(resd.s2).
```

```
Model R       R Square
1     .319(a) .102

Model           Sum of Squares df Mean Square F     Sig.
1     Regression 26.889         1  26.889      1.814 .197(a)      = ANOVA F test
      Residual   237.111        16 14.819                    MSRes = sp²
      Total      264.000        17

Model           Unstandardized            t      Sig.
                Coefficients
                B              Std. Error
1     (Constant) 25.222 =ȳ0     1.283      19.656 .000
      serv2      −2.444 =ȳ1−ȳ0  1.815      −1.347 .197   = t-test
```

```
LIST serv2 deli prdd.s2 resd.s2.
```

serv2	deli	prdd.s2	resd.s2
0	24.00	25.22222	-1.22222
0	29.00	25.22222	3.77778
0	24.00	25.22222	-1.22222
0	29.00	25.22222	3.77778
0	23.00	25.22222	-2.22222
0	23.00	25.22222	-2.22222
0	28.00	25.22222	2.77778
0	19.00	25.22222	-6.22222
0	28.00	25.22222	2.77778
1	24.00	22.77778	1.22222
1	16.00	22.77778	-6.77778
1	23.00	22.77778	.22222
1	26.00	22.77778	3.22222
1	22.00	22.77778	-.77778
1	18.00	22.77778	-4.77778
1	30.00	22.77778	7.22222
1	25.00	22.77778	2.22222
1	21.00	22.77778	-1.77778

```
GRAPH /SCATTERPLOT(BIVAR)=serv2 WITH deli.
```

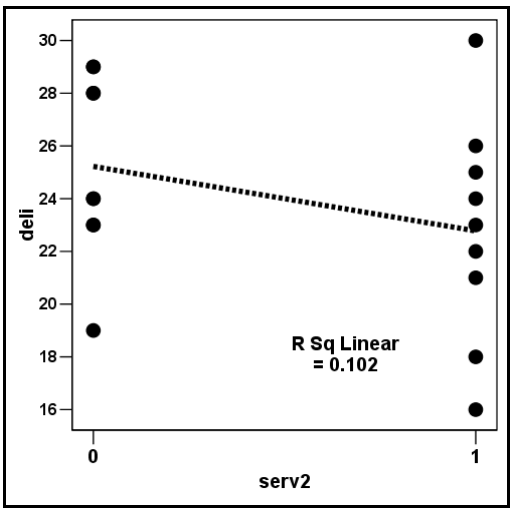


Figure 8-1.

The graph in Figure 8-1 shows that a single predictor accommodates the difference between two groups because a straight line can always go through two points, in this case the two means. In addition to the equivalent tests above, t for just one predictor can be calculated using the formula for the significance of a correlation coefficient, as shown in Box 8-3.

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.319}{\sqrt{\frac{1 - .319^2}{18 - 2}}} = \frac{.319}{.2369} = 1.346$$

Box 8-3.

The situation is more complex with more than two groups (i.e., when k > 2), which requires k - 1 predictors to ensure that the regression equation generates cell means as the predicted values. More than two groups also rules out an independent *t*-test and requires an *F*-test that can accommodate any number of groups (although the number of groups is generally modest for categorical predictors). The extension to more than two groups is covered in analysis of variance.

Categorical with a Numerical Predictor

Although the null hypothesis was not rejected in the preceding analyses, this is a non-experimental study. Therefore, confounded variables may mask the hypothesized benefits of involvement with social services. This problem was solved with multiple regression in earlier analyses of social services as a numerical predictor by including the confounding variable (e.g., impulsivity) as another predictor. The following analysis shows that the social service group (*serv2* = 1) does indeed have a higher mean impulsivity score, *M* = 51.78 for *serv2* = 1 and *M* = 42.22 for *serv2* = 0. If impulsivity correlates with delinquency, then previous analyses do not represent the unique contribution of involvement with social services independent of differences in impulsivity. MR can control for confounding of categorical and numerical predictors. Other examples of categorical and numerical predictors are shown in Appendix 8-1.

```
GLM impu BY serv2 /PRINT = DESCR.
```

```
 serv2 Mean    Std. Deviation N
 0     42.2222 6.41829        9
 1     51.7778 9.54958        9
Total 47.0000 9.29896        18
```

$SS_{Total} = (18-1)9.29896^2 = 1470.00$

$SS_{Error} = (9-1)6.41829^2 + (9-1)\ 9.54958^3 = 1059.11$

$SS_{Treatment} = 1470.00 - 1059.11 = 410.89$

```
 Source          Type III Sum of Squares df Mean Square F       Sig.
 Intercept       39762.000               1  39762.000   600.685 .000
 serv2           410.889                 1  410.889     6.207   .024
 Error           1059.111                16 66.194
 Corrected Total 1470.000                17
a R Squared = .280 (Adjusted R Squared = .234)
```

Reject H_0

Specifically, confounding is controlled statistically by regressing the dependent variable on both the confounding variable and the k - 1 indicator variables that define the categorical predictor. The analyses are shown below for *serv2*. The categorical predictor was added to impulsivity so that F_{Change} represents the effect of the categorical predictor controlling for impulsivity. For k = 2, we can also use the *t*-test for the categorical variable regression coefficient. Compare the $p = .026$ below to $p = .197$ obtained from the t and ANOVA for *serv2* alone. Researchers would now reject the null hypothesis, which can be stated in terms of the part correlation or the regression coefficient in the population.

```
REGRE /STAT = DEFAU CHANGE ZPP /DEP = deli /ENTER impu /ENTER serv2
 /SAVE PRED(prdd.is2).
```

```
 Model R       R      Adjusted Std. Error of Change Statistics
               Square R Square the Estimate  R Square Change F Change df1 df2 Sig.
 1     .234(a) .055   -.004    3.94889       .055            .930     1   16  .349
 2     .572(b) .327   .238     3.44067       .272            6.076    1   15  .026
```

```
 Model            Sum of Squares df Mean Square F     Sig.
 1     Regression 14.501          1  14.501      .930  .349(a)
       Residual   249.499         16 15.594
       Total      264.000         17

 2     Regression 86.426          2  43.213      3.650 .051(b)
       Residual   177.574         15 11.838
       Total      264.000         17
```

$SS_{Change} = 86.426-14.501 = 71.925$

```
 Model           Unstandardized            Standardized    t      Sig. Correlations
                 Coefficients              Coefficients
                 B              Std. Error Beta                        Zero-order Partial Part
 1   (Constant)  19.332         4.929                     3.922  .001
     impu        .099           .103       .234           .964   .349 .234       .234    .234

 2   (Constant)  15.211         4.609                     3.300  .005
     impu        .237           .106       .559           2.243  .040 .234       .501    .475
     serv2       -4.710         1.911      -.615          -2.465 .026 -.319      -.537   -.522
```

When impulsivity is controlled, *serv2* becomes significant, $p = .026$, in contrast to the nonsignificant effect with impulsivity alone as a predictor, $p = .197$. The strength has also increased, as reflected in $r_{d(s.i)} = -.522$ versus the simple $r_{d.s} = -.319$. There are several reasons the significance and strength increased. First, with impulsivity in the equation, the numerator for the t-test (i.e., the regression coefficient) for *serv2* is -

4.710 versus -2.444 in the previous analysis; the social service predictor has become a stronger predictor of delinquency. The SS_{Change} = 71.925 for the numerator of F_{Change} is also larger than SS_{Reg} = 26.889 in the earlier analysis with *serv2* alone. The adjustment for differences between groups in impulsivity is responsible for these increases, as demonstrated shortly. The larger value for SS_{Change} also explains why the part r for *serv2* is stronger than the simple r; that is, $r_{d(s.i)}$ = -.522 versus $r_{d.s}$ = -.319.

Additional predictors can also reduce the error terms (i.e., the denominators) for F and t. For example, MS_{Res} = 11.838 in the above analysis, versus 14.819 earlier. Even numerical predictors that do not correct for confounding can still improve the sensitivity of analyses by removing variability from the error terms for the statistical tests. A test of the significance of a categorical predictor controlling for a numerical variable is called Analysis of Covariance (ANCOVA). The numerical predictor is a covariate (impulsivity here). MANOVA and GLM can also do ANCOVA.

The complicated graph in Figure 8-2 shows visually the effects of the preceding analysis. The open circles and dashed lines are for *serv2* = 0 (low involvement with social services) and the filled circles and solid lines are for *serv2* = 1 (high involvement).

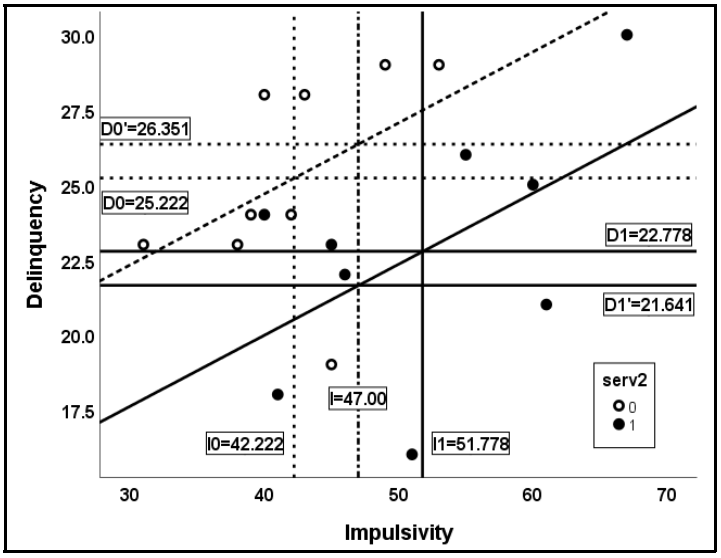


Figure 8-2. Categorical and Numerical Predictors.

The vertical lines show the mean impulsivity scores; the dashed line on the left is for *serv2* = 0 ($\bar{y}_{Imp}$ = 42.222) and the solid line on the right is for *serv2* = 1 ($\bar{y}_{Imp}$ = 51.778). The higher impulsivity scores for the social services group complicates interpretation of the fact that mean delinquency scores did not differ significantly in the earlier tests when impulsivity was not controlled. The multiple regression, on the other hand, indicates the difference between the group means if both groups had obtained the same average impulsivity scores, represented by the middle vertical line ($\bar{y}_{Imp}$ = 47.00).

The sloped lines are the best fit regression lines for the two groups using the multiple regression equation from above. The calculations shown in Box 8-4 demonstrate how the mean difference changes. The regression equation is used to predict delinquency values for each group, first using the observed impulsivity means for the individual groups, 42.222 and 51.778, and then using a common mean, 47.000.

$\hat{d} = 15.211 - 4.715 \times s + .237 \times i$

for $s = 0 \,\&\, i = 42.222$, $\hat{d} = 25.22 = \bar{d}_0$

for $s = 1 \,\&\, i = 51.778$, $\hat{d} = 22.77 = \bar{d}_1$

for $s = 0 \,\&\, i = 47.000$, $\hat{d} = 26.35$

for $s = 1 \,\&\, i = 47.000$, $\hat{d} = 21.64$

Box 8-4.

The difference between the unadjusted means is 25.222 - 22.778 = 2.444, while the difference between means adjusted for the difference in impulsivity is 26.351 - 21.641 = 4.710. These values are the regression coefficients for *serv2* alone and *serv2* with *impu*, respectively. The horizontal lines in Figure 8-2 are means for delinquency, dashed for *serv2* = 0 unadjusted for impulsivity (bottom dashed line) and

adjusted. The solid horizontal lines are means for *serv2* = 1, top for unadjusted and bottom adjusted for impulsivity differences.

To equate on impulsivity, the *serv2* = 1 group shifts down from its higher impulsivity score and the *serv2* = 0 group shifts up from its lower impulsivity score up. Note that the shift is up or down *along* the regression lines and *not* simply a shift directly left or right. Shifts follow the regression lines because impulsivity correlates with delinquency. The difference between *serv2* = 0 and 1 becomes larger, as just calculated. See Appendix 8-2 for instructions to create Figure 8-2.

Although it helps to think about the adjusted difference between means as representing the outcome if there was no difference on impulsivity, the fact that the lines are parallel (i.e., $b_{di.s}$ = .237) means that the difference on delinquency will be the same (i.e., 4.710) anywhere along the line (e.g., at the intercepts). If the lines have different slopes, however, then the difference will vary depending on the value of impulsivity. The possibility that slopes differ (i.e., there is an interaction between the two predictors) is considered next.

Categorical Predictors and Interaction

The preceding analysis assumed that a single regression coefficient or slope for impulsivity ($b_{di.s}$ = .237) could be used for both groups; that is, the regression lines were expected to be parallel. It is conceivable, however, that the regression coefficient for impulsivity differs as a function of involvement with social services. When the relationship between two variables depends on the levels of a third variable, then the variables are said to interact. In the present example, the relationship between delinquency and impulsivity may differ for low and high levels of involvement with social services.

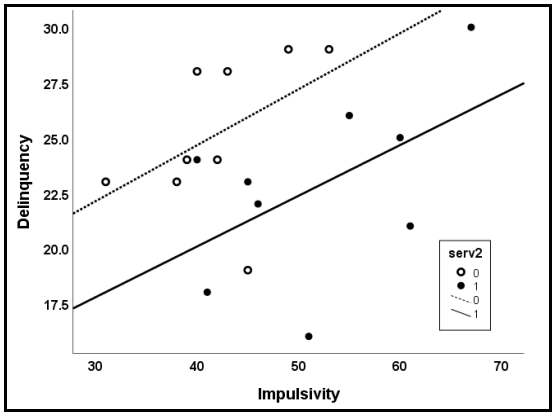


Figure 8-3. Categorical & Numerical Interaction.

Figure 8-3 graphs separate lines with different slopes, that is, a separate regression equation for each group. These lines were created in the Chart Editor and appear to deviate very little from parallel lines; that is, they have similar slopes that likely do not differ significantly. However, statistical tests are required to determine the strength and significance of any deviation from parallel lines. The first step in understanding the procedure is to find the regression coefficients from separate regressions for the two groups.

The separate regressions can be obtained using SPSS's SPLIT FILE command. After a SPLIT FILE command, SPSS procedures are computed separately for the groups specified on the SPLIT FILE command. To use SPLIT FILE, the dataset must be sorted on the variable used to define groups; for example, in the present study all *serv2* = 0 and *serve2* = 1 cases would be clustered together. If data is not already sorted properly, SORT CASES BY can be used to sort cases in the appropriate order.

```
SPLIT FILE by serv2.
REGRE /DEP = deli /ENTER impu.
 serv2 Model R       R Square Adjusted R      Std. Error of
                              Square          the Estimate
 0     1     .471(a) .222     .111            3.25951
 1     1     .521(a) .272     .168            3.83729

 serv2 Model            Sum of Squares df Mean Square F     Sig.
 0     1     Regression 21.185         1  21.185      1.994 .201(a)
             Residual   74.371         7  10.624
             Total      95.556         8

 1     1     Regression 38.482         1  38.482      2.613 .150(a)
             Residual   103.073        7  14.725
             Total      141.556        8

 serv2 Model            Unstandardized             Standardized    t     Sig.
                        Coefficients               Coefficients
                        B               Std. Error Beta
 0     1     (Constant) 14.517          7.659                     1.896 .100
             impu       .254            .180       .471           1.412 .201

 1     1     (Constant) 10.886          7.466                     1.458 .188
             impu       .230            .142       .521           1.617 .150
SPLIT FILE OFF.
```

The graph and the separate regressions demonstrate that the two lines are almost but not quite parallel. The slope for group 1 is slightly steeper than the slope for group 2; specifically, $b_{d.i}$ = .254 for *serv2* = 0 and .230 for *serv2* = 1, a difference of .024 units. The graph and regressions demonstrate that the intercepts for the two lines also differ (14.517 versus 10.886). Unlike parallel lines, this difference is difficult to interpret because it represents the difference between the lines when impulsivity equals 0. The difference between the lines will vary with impulsivity when the lines have different slopes.

The graph and preceding analyses do not tell whether the difference between slopes is significant, although the difference is certainly small. Testing the significance of the difference in slopes requires that we incorporate the difference between slopes into a single equation to test whether it is significant. Here each slope was clearly not significant, but even if one were significant and the other not, we could still not conclude that they differed significantly from one another. For example, two slopes could be quite close to one another with one significantly different from 0 and the other not, depending on where the slopes fall relative to the value required to reject H_0.

To test the significance of the difference between slopes with only two groups, a third predictor is created to represent the interaction, specifically the product of the first two predictors. In the present example, we would multiple *serv2* times *impu* to create a third predictor, called *s2ximp* below. The regression coefficient for this new predictor will reflect the difference between the regression coefficients, the .024 calculated above.

```
COMPUTE s2ximp = serv2*impu.
REGRE /DEP = deli /ENTER serv2 impu s2ximp.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.573(a)	.328	.184	3.56014

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	86.556	3	28.852	2.276	.125(a)
	Residual	177.444	14	12.675		
	Total	264.000	17			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	14.517	8.365		1.735	.105
	serv2	−3.631	10.861	−.474	−.334	.743
	impu	.254	.196	.598	1.293	.217
	s2ximp	**−.024**	.236	−.166	**−.101**	**.921**

That the new predictor *s2ximp* represents the difference between the two slopes is shown by its regression coefficient (-.024). Here the difference in slopes is not even close to significant, $p = .921$ for the interaction term, indicating that the null hypothesis of no interaction (i.e., no difference between the two slopes for impulsivity) cannot be rejected. Although less obvious, the single equation from this analysis incorporates both regressions obtained earlier. Substituting the values for *serv2* (0 or 1, depending on the group) into the equation gives the equations in Box 8-5. The multiple regression equation with the interaction term incorporates both earlier equations.

$$\hat{d} = 14.517 - 3.631 \times s + .254 \times i - .024 \times si$$
$$= 14.517 - 3.631 \times s + (.254 - .024 \times s) \times i$$
$$\text{for } s = 0, \quad \hat{d} = 14.517 - 3.631 \times 0 + (.254 - .024 \times 0) \times i$$
$$= 14.517 + .254i$$
$$\text{for } s = 1, \quad \hat{d} = 14.517 - 3.631 \times 1 + (.254 - .024 \times 1) \times i$$
$$= 10.886 + .230i$$

Box 8-5.

Summary

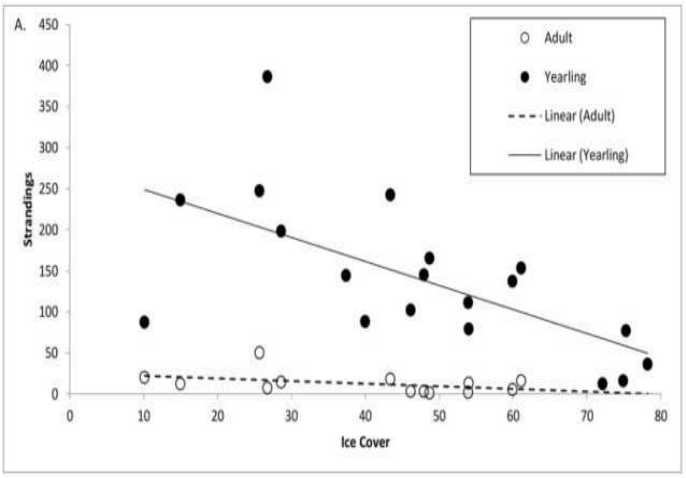
Given a categorical predictor, three analyses are possible and the corresponding equations are illustrated in Box 8-6. If $b_{x.cn} = 0$, then equation three reduces to equation two and if $b_{n.c} = 0$, the second equation reduces to the first. Graphs of some interactions between categorical and numerical predictors are shown in Appendix 8-1. Such interactions can be important for theoretical or applied reasons.

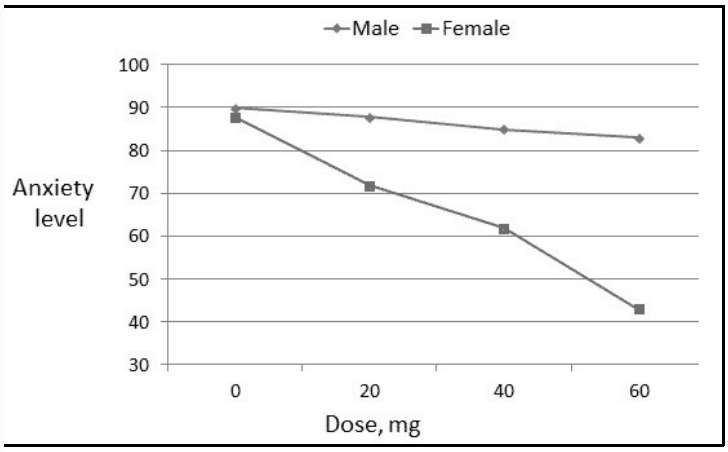
$$\hat{Y} = b_0 + b_c \times C$$
$$\hat{Y} = b_0 + b_{c.n} \times C + b_{n.c} \times N$$
$$\hat{Y} = b_0 + b_{c.nx} \times C + b_{n.cx} \times N + b_{x.cn} \times X$$
$$C = \text{categorical}, \; N = \text{numerical}, \; X = C \times N$$

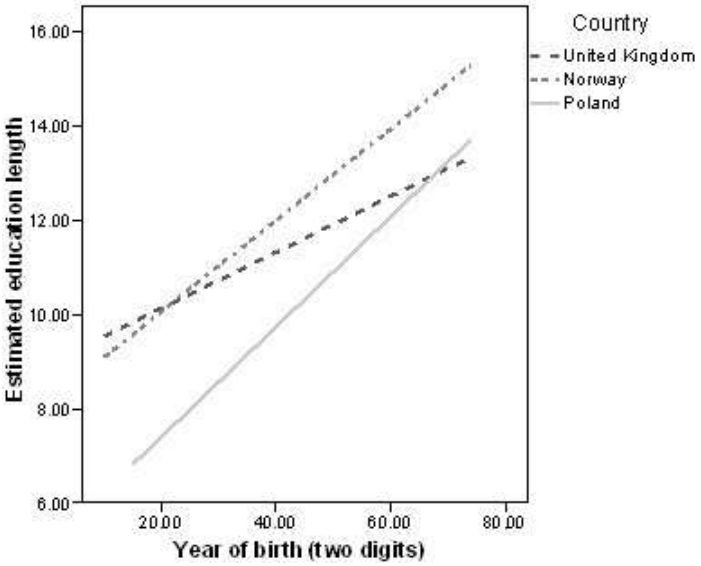
Box 8-6.

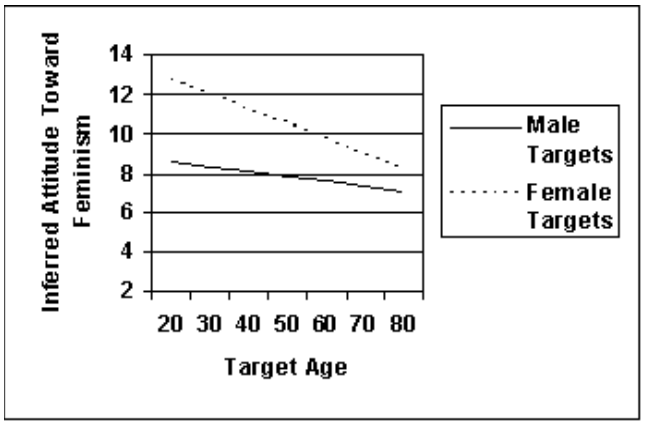
Appendix 8-1

Examples of Categorical and Numerical Predictors









Appendix 8-2
Creating the Graph in Figure 8-2

First, create the basic graph via the following menu steps: Graph | Legacy | Scatterplot | Simple | Define. This brings up the Simple Scatterplot menu shown in Figure 1. The Y Axis, X Axis, and Set Markers by have been added. Set Markers is where the categorical variable goes, *serv2* in our example. Y is the dependent variable *deli* and X is the numerical predictor *impu*. Click Ok to create the basic graph.

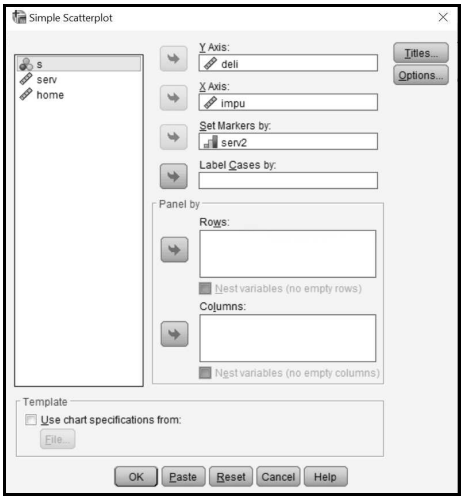


Figure 1. Simple Scatterplot Menu

Next, double click on the graph to open the Graph Editor shown in Figure 2. The left image in Figure 2 shows the pull-down menu for Options. Figure 8.2 requires that we add vertical lines (X Axis Reference Line), horizontal lines (Y Axis Referent Line), and an equation for each group derived from our regression analysis (Reference Line from Equation). These can be added by selecting the relevant option or by clicking on the corresponding symbol on the overall Chart Editor menu shown in the right image in Figure 2.

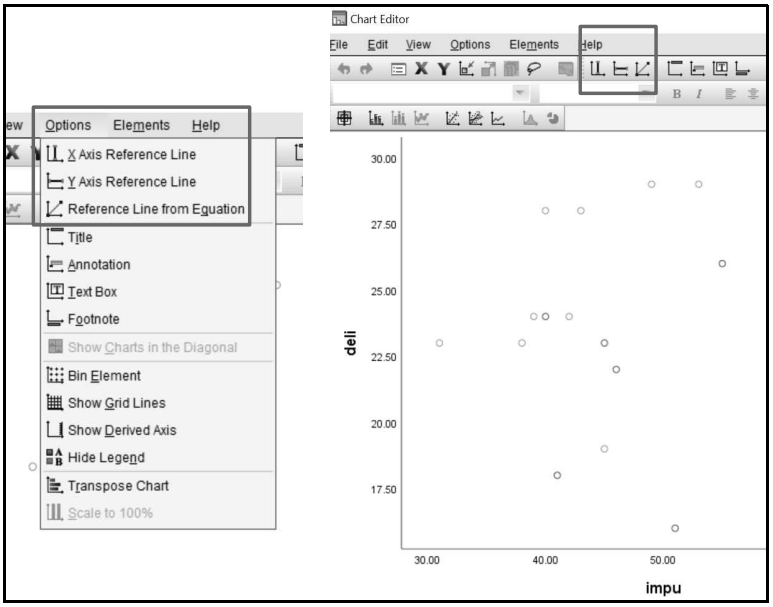


Figure 2. Chart Editor

Now add and modify the elements shown in Figure 8-2. Figure 3 shows the outcome of selecting the X Axis Reference Line option. SPSS inserts a line at some point it deems reasonable, but the actual location can be set in the Properties menu box labelled Position. Here it contains the mean *impu* score for *serv2* = 1.

Using the Y Axis Reference Line option, the mean *deli* score for *serv2* = 1 can be added and is shown in Figure 4.

The final requirement for *serv2* = 1 is the equation derived from: $\hat{d}=15.211-4.715\times s+.237\times i$. For s = 1, the equation becomes: $\hat{d}=15.211-4.715\times 1+.237\times i=10.496-.237\times i$. Selecting Reference Line from Equation brings up Figure 4. SPSS has

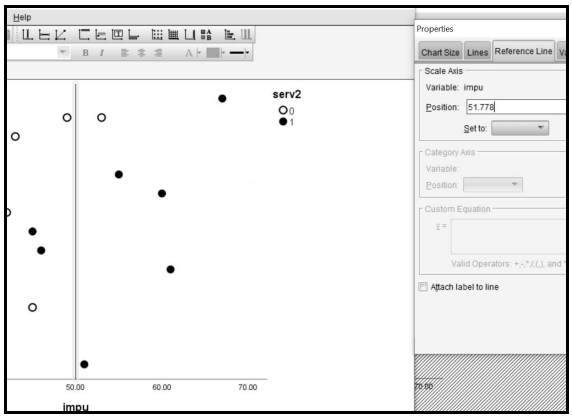


Figure 3. Vertical Line Added.

added a line from an equation that appeared initially in the Custom Equation box. But now the above equation has been added. When applied, the line in the figure will move and, as expected for a regression line, it will pass exactly through the intersection of our vertical and horizontal lines (i.e., the means for X and Y).

The same steps would be followed for *serv2* = 0: vertical line at its mean *impu*, horizontal line at its mean *deli*, and the equation from above for s = 0, which would be 15.211 + .237i, because 4.715×0 = 0.

The remaining lines in Figure 8.2 are a vertical line at the overall mean *impu* score for both groups and horizontal lines where that vertical line intersects the equations. These intersection points represent mean *deli* scores adjusted for impulsivity.

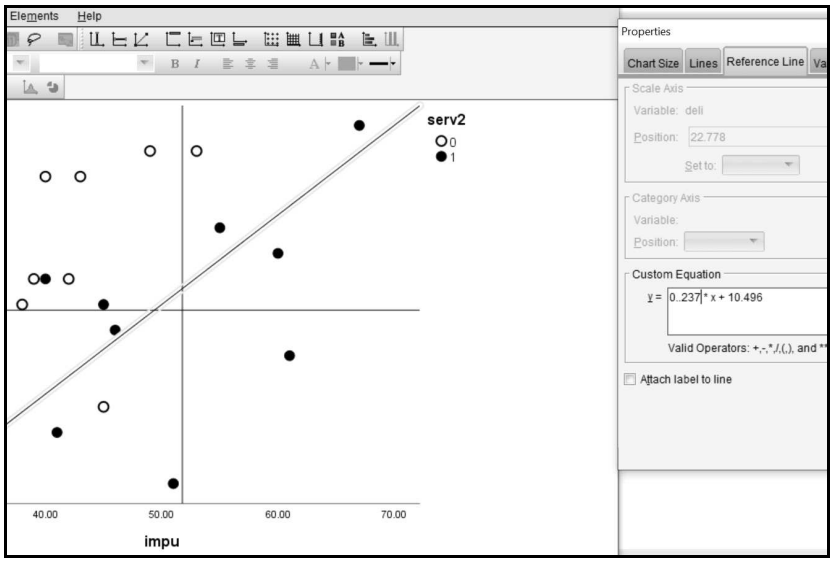


Figure 4. Reference Line from Equation Option

Chapter 9 - Nonlinear Regression

Not all relationships between predictor and criterion variables are linear. For a linear relationship, a unit change on X produces exactly the same amount of change on Y no matter where X falls along the scale. For example, the effect on Y of a 10 point difference on X would be the same whether the comparison was 110 vs 100 or 210 vs 200. This defines a straight line (i.e., an equation with a single slope across the entire range of X). In multiple regression, the two regression coefficients represent linear effects for both dimensions giving a rigid, flat plane of predicted values.

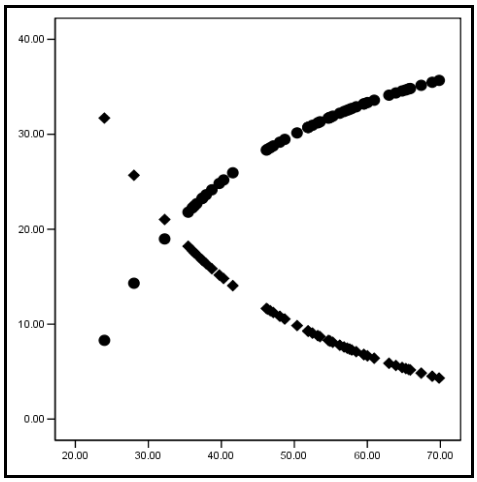


Figure 9-1. Decelerating nonlinear relationships.

But sometimes the change in Y varies (i.e., grows steeper or less steep) across values of X. In studies of forgetting, for example, the amount of information lost may be greater early in the retention interval and less later in the retention interval. This pattern is represented by the descending curve (diamonds) in Figure 9-1. Note that the loss between X = 20 and X = 30 (about 10 unit decrease on Y) is much greater than the loss between X = 60 and X = 70 (less than 5 unit decrease on Y). The increasing curve (circles) shows a pattern in which more positive change occurs early and less positive change occurs later along the values of X. A study of learning, for example, may show more learning across the first 10 trials than across the last 10 trials.

In other situations, the amount of change (increase or decrease) may be greater at higher values of X than at lower values of X. Idealized patterns of this sort are shown in Figure 9-2. The decreasing curve (circles) shows a negative relationship that becomes even more negative. This might occur, for example, in a study of fatigue over time; there may be little deterioration initially, and more marked loss in performance later. The increasing curve (diamonds) shows a positive relationship that becomes even more positive. In learning to solve insight problems, for example, people may not improve very much initially, but eventually improve more rapidly with greater experience.

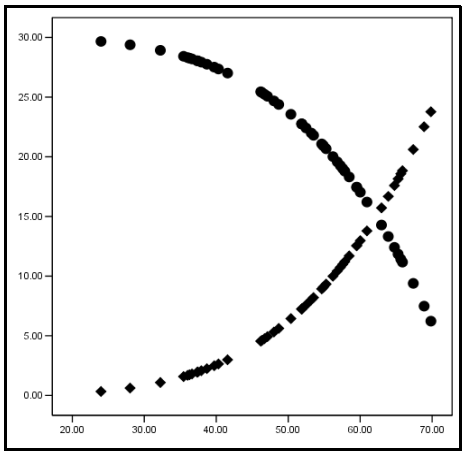


Figure 9-2. Accelerating nonlinear relationships.

Such nonlinear relationships occur in many areas of psychology. See Appendix 9-1 for some examples. There are several ways to accommodate nonlinear relationships; we consider two main approaches: polynomial regression and transformations. An interaction approach is discussed briefly, as well as more sophisticated methods to analyze nonlinear relationships that avoid some problems with the preceding approaches.

Identifying Nonlinear Relationships with Graphs

To illustrate analysis of nonlinear relationships, consider the old USA dataset on crime rates by state that illustrates not only nonlinear relationships but also how misleading simple correlations can be (see discussion in Appendix 7-2). One of several predictors was percentage of the population that was white (abbreviated as *pw* or *pctwhite*). The basic relationship is shown in Figure 9-3. The linear relation shown by the solid line is quite good, $r^2 = .468$, although additional analyses in Appendix 7-2 revealed that the linear relationship between the two variables became very weak when other factors were controlled. Moreover, some part of the relationship could reflect aspects of the criminal justice system that are biased against minorities. But here we examine how to capture the non-linear relationship.

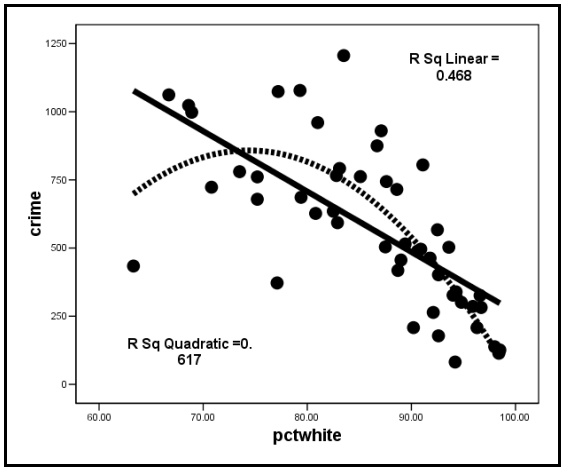


Figure 9-3. Nonlinear relationship.

One visual cue for nonlinearity is that deviations of observed values from predicted values vary systematically across values of *pctwhite*. Specifically, residuals for low and high values of *pctwhite* tend to be negative (i.e., the linear equation over-predicts those observations), whereas residuals for moderate values of *pctwhite* tend to be positive (i.e., the linear equation under-predicts). For purely linear relationships, deviations above and below the best-fit straight line tend to be evenly distributed across the range of X. We return to the dashed line of fit in a moment.

Although plots of Y against X often reveal the nonlinear nature of relationships, a stronger visual "test" plots residual Y scores from a linear regression of Y on X (i.e., REGRESS /DEP = crime /ENTER pctwhite /SAVE RESI(resc.w). Residual crime scores are shown in Figure 9.4 as a function of the predictor. As shown before, r = 0 for the relationship between the residual and X. A dashed line has been inserted at 0, the mean of residual scores, and makes the pattern of positive and negative residuals more obvious. Specifically, negative residuals are more common at the extremes of *pctwhite* and positive residuals more common in the middle. Adding a nonlinear prediction in the chart editor would strengthen this impression. It would be an inverted U-shape.

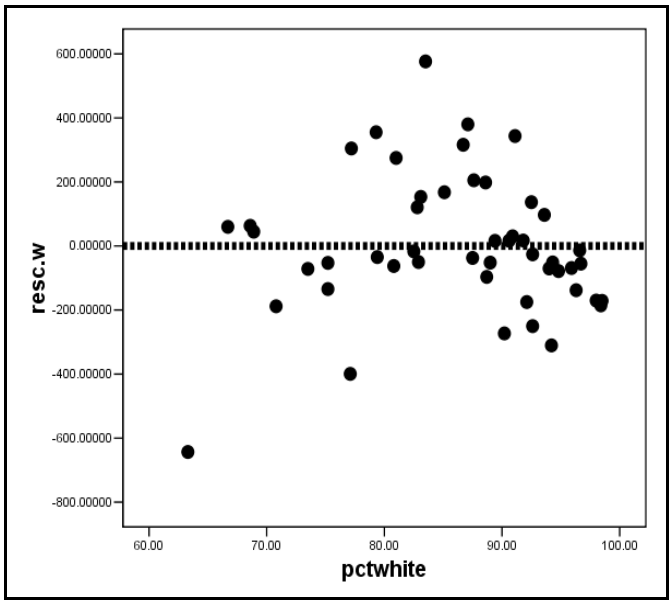


Figure 9.4.

Polynomial Regression

One approach to the analysis of nonlinear relationships is polynomial regression. In polynomial regression, the predictor X is used to generate X^2 (and sometimes X^3, X^4, and so on) values. The new X^2

predictor is included with X in a multiple regression. This allows for a nonlinear relationship; the dashed line in Figure 9-3 is a polynomial fit, called quadratic because it includes X and X^2, but not higher powers of X.

The quadratic regression for the crime study is shown below. Including *pctwhite2* (i.e., X^2) along with *pctwhite* (i.e., X) markedly increases SS_{Reg} and R^2, resulting in a substantial part $r^2 = .149$ and a highly significant unique effect for the *pctwhite2* predictor, $F = 17.959$ or $t = -4.238$, $p = .000$. In short, *pcwhite2* demonstrates a substantial and significant unique contribution to the prediction of crime rates, over above the linear effect represented by *pctwhite*. If the relationship was purely linear, the statistical results for the unique contribution of *pctwhite2* would be minimal.

```
COMPUTE pctwhite2 = pctwhite**2.
REGRE /STAT = DEFAU CHANGE /DEP = crime /ENTER pctwhite /ENTER pctwhite2.
```

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.684(a)	.468	.457	217.875	.468	41.358	1	47	.000
2	.786(b)	.617	.601	186.770	**.149**	**17.959**	**1**	**46**	**.000**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1963225.985	1	1963225.985	41.358	.000(a)
	Residual	2231064.505	47	47469.458		
	Total	4194290.490	48			
2	Regression	2589676.343	2	1294838.172	37.120	.000(b)
	Residual	1604614.147	46	34882.916		
	Total	4194290.490	48			

Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.
1	(Constant)	2479.966	298.172		8.317	.000
	pctwhite	-22.158	3.446	-.684	-6.431	.000
2	(Constant)	-6352.548	2099.853		-3.025	.004
	pctwhite	193.942	51.079	5.988	3.797	.000
	pctwhite2	**-1.304**	**.308**	**-6.683**	**-4.238**	**.000**

Essentially the fit is better because the quadratic equation in model 2 generates predicted scores that follow a curve rather than a straight line. The resulting curve is the dashed line shown in Figure 9-3 along with the best-fit straight line. The curve was created by SPSS in the chart editor, but could also have been generated by saving the predicted values from the best-fit equation in model 2. This analysis reveals that the crime rate stays relatively flat until *pctwhite* reaches 85% or so. Then it starts to decline. Such a pattern probably involves a different explanation than a simple linear relationship, as discussed later.

Transformations of Predictor

A second approach to nonlinear relationships is to transform the predictor. In essence the predictor can be transformed (and/or the criterion variable) to make the relationship more linear. This involves stretching out or compressing the predictor or criterion variable. We focus on the predictor because transforming the dependent variable can be problematic for multiple predictors given the specific transformation required for the dependent variable could vary across predictors. But there is only one dependent variable and it cannot be transformed in multiple ways in the same analysis.

Figure 9-5 shows a hypothetical curvilinear relationship between X and Y. Deviations of observed from the linear predicted values are positive for low and high values of X, and negative for intermediate values of X. Although $R^2 = .958$ for the linear fit is excellent, $R^2 = .994$ for the quadratic is even better. Transforming X would improve the linear relationship. The data in Figure 9-5 improves if we compress X. Compressing or stretching the predictor is done by raising the predictor to some power. Powers less than 1 compress X and powers greater than 1 stretch it out.

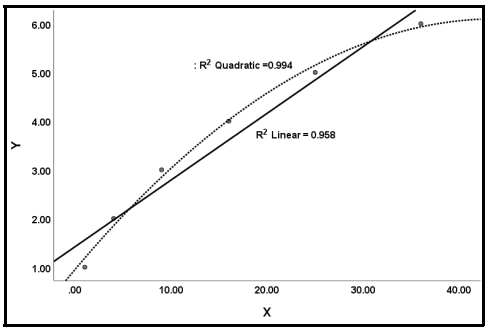


Figure 9-5.

To illustrate what is meant by compression and stretching, consider the numbers 1, 4, and 9. On this scale, the distance between 9 and 4 (5 units) is greater than the distance between 4 and 1 (3 units). But if we compress the scale by raising the numbers to a power less than 1 (.5 = square root), we obtain $1^{.5} = 1$, $4^{.5} = 2$, and $9^{.5} = 3$ with equal distances between the previous differences. The upper end has been compressed. Values less than .5 (e.g., ~0 = logarithm, -1 = reciprocal) would compress the upper values even further. Figure 9-6 shows the relationship between Y and the square root of X (i.e., $X^{.5}$). The linear relationship is now perfect, not surprising since Y equals the square root of X.

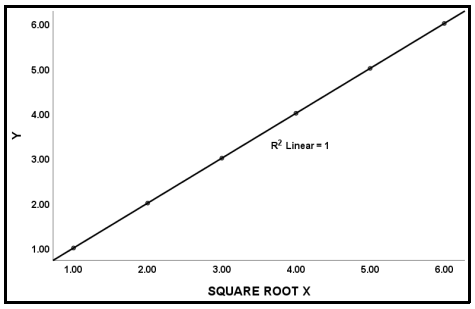


Figure 9-6.

Consider next starting with the values 1, 2, and 3. Raising these numbers to a power greater than 1 (e.g., squaring them) stretches out the upper end so that what were equal distances become larger for the upper values; that is, $1^2 = 1$, $2^2 = 4$, and $3^2 = 9$. The equal difference between 1-2 (1 unit) and 2-3 (1 unit) on the original scale is now 1-4 (3 units) and 4-9 (5 units). X is more spread out.

Figure 9-7 shows a relationship that would benefit from an expansion of X. The curve is steeper at higher values of X than lower values. Again, the quadratic improves on the linear, even though the linear equation is itself very good. Expansion requires a power greater than 1. Figure 9-8 shows the plot for Y as a function of X^2. The fit is perfect, in this case because Y equals X^2.

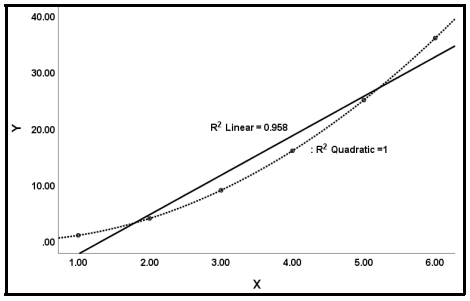


Figure 9-7.

Examine the graphs in Figures 9-1 and 9-2, to determine why the data in Figure 9-1 results in a better fit if X is compressed, whereas the data in Figure 9-2 benefits from X being stretched out.

The *crime* and *pctwhite* relationship requires that *pctwhite* be stretched out. The following analyses examine the effect of different transformations on R^2 (recall that $R^2 = .617$ for the quadratic regression and $R^2 = .468$ for the linear regression). With the transformation

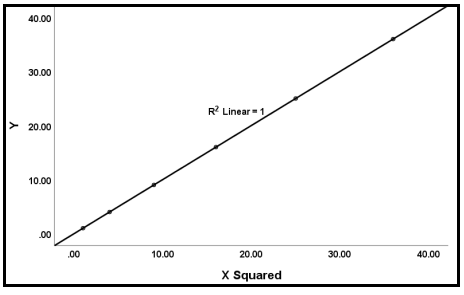


Figure 9-8.

approach, ONLY the transformed variable is used as a predictor, and not the original X (as was done in quadratic regression). Note below that transformations less than one produce a poorer fit than the linear, while transformations greater than one produce a better fit, although never as good as the quadratic, at least up to the fourth power. One issue with higher powers is making sense of the transformed predictor. The best that can be done in many cases is to simply think of it as expanding or contracting the predictor. SPSS's CURVE ESTIMATION procedure described later can also perform and analyze some transformations.

```
COMPUTE wreciprocal = pctwhite**-1.
COMPUTE wlogarithm = lg10(pctwhite).
COMPUTE wsquareroot = pctwhite**.5.
COMPUTE w2 = pctwhite**2.
COMPUTE w3 = pctwhite**3.
COMPUTE w4 = pctwhite**4.
```

```
 Model R       R Square Adjusted R      Std. Error of
                        Square          the Estimate
REGRE /STAT = R /DEP = crime /ENTER wreciprocal.
 1     .632(a) .400     .387            231.425
REGRE /STAT = R /DEP = crime /ENTER wlogarithm.
 1     .660(a) .435     .423            224.490
REGRE /STAT = R /DEP = crime /ENTER wsquareroot.
 1     .672(a) .452     .440            221.131
REGRE /STAT = R /DEP = crime /ENTER pctwhite.
 1     .684(a) .468     .457            217.875        <<<<< Original predictor
REGRE /STAT = R /DEP = crime /ENTER w2.
 1     .705(a) .498     .487            211.755
REGRE /STAT = R /DEP = crime /ENTER w3.
 1     .723(a) .523     .513            206.272
REGRE /STAT = R /DEP = crime /ENTER w4.
 1     .738(a) .545     .535            201.524
```

One challenge students sometimes have with transformations is the incorrect belief that there is something "sacred" about the numerical values for measures and they cannot be changed. That is, isn't there something fixed about IQ scores, measures of depression, or a person's age that makes it inappropriate to stretch or compress the scales? But psychologically the answer is no. Consider age. Is it really the case that the "aging" process is necessarily the same between 40 and 50 and between 70 and 80 (i.e., 10 years of "aging" in each case). No, because 10 years might represent greater physiological or other changes that result in more loss of cognitive functioning between 70 and 80 than between 40 and 50. That defines a nonlinear relationship. But given this, it would be perfectly sensible to transform the age variable so that the interval between 70 and 80 was larger than the interval between 40 and 50. Squaring the numbers produces an interval of 1500 between 70^2 and 80^2 versus an interval of 900 between 40^2 and 50^2.

Another useful example is the relationship between happiness and gross domestic product (GDP) of countries. Certainly an increase in GDP from $1,000 a year to $2,000 a year is going to have more impact on well-being than an increase from $50,000 a year to $51,000 a year. Again that defines a nonlinear relationship.

Consider also performance on a cognitive task (e.g., arithmetic) and three people who took 10, 20, and 30 minutes to solve 60 problems. It would appear that the difference between persons 1 and 2 (i.e., 10

minutes) is the same as the difference between persons 2 and 3 (i.e., 10 minutes). But what if we instead measure problems-per-minute (ppm); now we have scores of 60/10 = 6 ppm for person 1, 60/20 = 3 ppm for person 2, and 60/30 = 2 ppm for person 3. Now the difference in scores between persons 1 and 2 (i.e., 3 ppms) is greater than the difference in scores between persons 2 and 3 (1 ppm). Clearly nothing indicates that one of these measures is inherently better than the other despite the fact they are not linear equivalents. Similar logic applies to many variables. Apply this same reasoning to the graphs of nonlinear relationships or transformed predictors shown in Appendix 9-1.

Supplementary Material on Nonlinear Relationships

The following material demonstrates some limitations of the preceding analyses and alternative approaches to nonlinear relationships. However, you are not responsible for learning these analyses. One limitation of the procedures described so far is that quadratic equations modify the direction of change only once, which means that eventually the curve begins to arc in a direction that makes little sense. The quadratic equation in Figure 9-9 curves down as *pctwhite* becomes increasingly small. While there may be situations in which such a reversal is appropriate, it is more likely in the present case that the fit should flatten out rather than start to decrease again. A limitation of the transformation approach is that some transformations may not be meaningful. The log of X, for example, might indeed correspond to a rational transformation. But the logic of raising X to the power of four might be difficult to grasp. We consider two approaches to nonlinear relationships that avoid these problems.

Interaction and Nonlinear Relationships

One alternative approach to nonlinear relationships is to fit two or more straight lines to the data by entering interactions between levels of the predictor (e.g., low vs high) and the values. Each line would have a different slope, allowing for a better fit to nonlinear data and often demonstrating something important about the relationship. Although there are ways to determine statistically the optimal break-point, the following regression uses *pctwhite* = 85 based on the graph in Figure 9.3. The graph in Figure 9-9 shows the results corresponding to the following regression analysis. Up to *pctwhite* = 85, there is little change in crime rate, r^2 = .0006. Beyond 85, there is a marked decline, r^2 = .62.

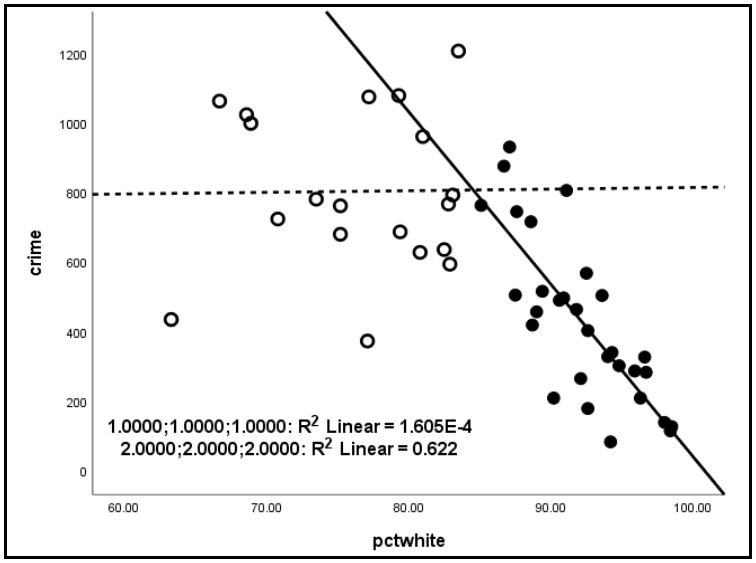


Figure 9-9.

The first step in the analysis is to create a categorical variable that represents low and high values for *pctwhite*. The following commands create *w2*, which equals 0 for low values and 1 for high values. The values of 0 and 1 work well for the later regression, but for the graph we need to use values of 1 and 2 as

GRAPH does not recognize 0 as a distinct group when plotting separate lines for two groups. RECODE produces a new variable *w2b* with values of 1 and 2 for the low and high *pctwhite* groups, respectively. The GRAPH command then produces the graph in Figure 9-9 and the Chart Editor adds the separate regression lines. The line is quite flat for low scores (*w2b* = 1) and strongly negative for the high scores (*w2b* = 2).

```
COMPUTE w2 = 0.
IF pctwhite > 85 w2 = 1.
RECODE w2 (0=1) (1=2) INTO w2b.

GRAPH /SCATTERPLOT(BIVAR)=pctwhite WITH crime BY w2b.
```

The regression analysis corresponding to Figure 9-9 requires an interaction term as in earlier analyses for categorical and numerical predictors. The predictor *wxw2* below represents the interaction. The R^2 = .628 in the regression is higher than R^2 = .468 from the simple linear regression. Also, the interaction term is highly significant confirming that the two regression coefficients represented in Figure 9-9 are significantly different.

```
COMPUTE wxw2 = pctwhite*w2.
REGRE /DEP = crime /ENTER w2 pctwhite wxw2.
```

Model	R	R Square
1	.793	0.628

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2634626.129	3	878208.71	25.338	.000
	Residual	1559664.36	45	34659.208		
	Total	4194290.49	48			

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	766.843	539.277		1.422	0.162
	w2	4232.248	1009.095	7.048	4.194	.000
	pctwhite	0.468	7.035	.014	.066	0.947
	wxw2	-50.063	11.618	-7.701	-4.309	0

Although not shown, separate regressions resulted in a regression coefficient of +.47 for the low scores and -49.59 for the high scores. This gives a difference between the slopes of -49.59 - +.47 = -50.06, the coefficient for *wxw2*.

One benefit of a better fit for the data is that it can reveal important features of the data. In the present case, for example, Figure 9-9 shows that there is no relationship between *pctwhite* and crime until we reach states that are almost entirely white. It is *not* the case that in general more white residents translates into less crime in a linear way, and there would be many features that are unique to states with largely white populations.

Figure 9-10 shows one strong candidate, the difference between largely urban and largely rural states, with *pctwhite* being especially high in rural states. Note that the pattern here is very similar to the pattern in Figure 9-9, relatively flat up to about 85% *pctwhite* and then a decline above that. Crime rates might be lower in rural states, if only because opportunities for many crimes (e.g., shop-lifting) could be fewer.

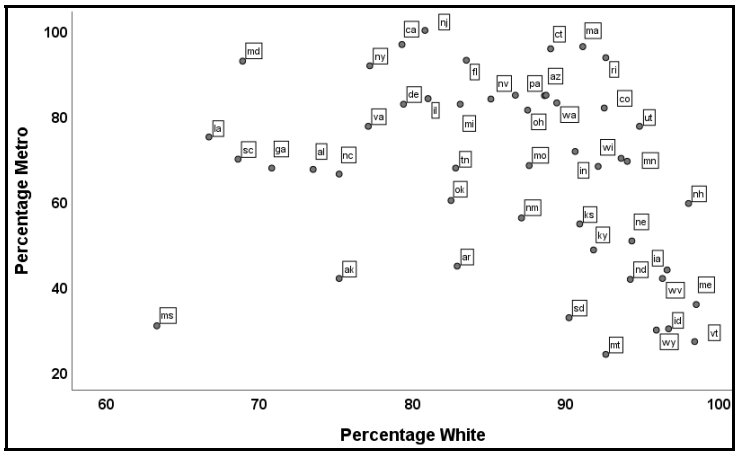


Figure 9-10.

Nonlinear Regression

There are statistical tools more powerful than the preceding approaches to fit nonlinear equations, including NLR in SPSS. Like most such programs, NLR requires the researcher to specify the general form of a nonlinear model for the data and then NLR finds the values (parameters) for the model that give the best fit to the data. Below are commands that calculate a nonlinear equation to fit the crime rate data. The form of the equation is: $y = a + b*c^{pctwhite}$ and NLR estimates the best values for a, b, and c. These values are bolded in the output. Raising a value to the power of *pctwhite* (or in general X) produces a nonlinear prediction.

```
MODEL PROGRAM  a=10 b=.5 c=1.01.
COMPUTE  PRED_=a + b*c**pctwhite.
NLR crime  /PRED PRED_   /CRITERIA SSCONVERGENCE 1E-8 PCON 1E-8.
```

Parameter	Estimate	Std. Error
a	**894.121**	97.601
b	**-0.031**	.108
c	**1.109**	.039

Source	Sum of Squares	df	Mean Squares
Regression	18583488.537	3	6194496.179
Residual	1693193.463	46	36808.554
Uncorrected Total	20276682.000	49	
Corrected Total	4194290.49	48	

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .596.

The parameters for the model generate predicted crime rates, as shown below, although NLR can also generate predicted values. The GRAPH command creates a single plot containing the observed data (*crime*) and the predicted values (*crimenl*) as a function of *pctwhite*. The resulting plot is Figure 9-11.

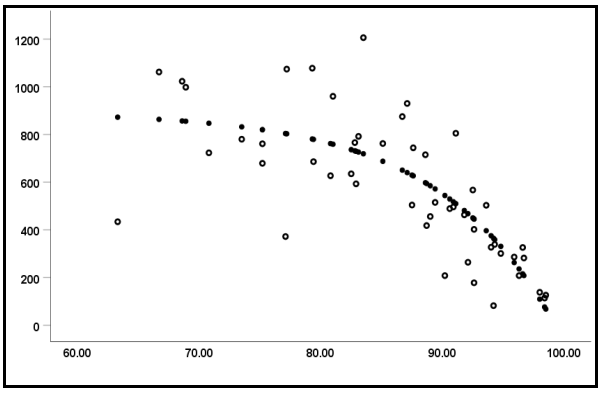


Figure 9-11.

```
COMPUTE crimenl = 894.121 + -.031*1.109**pctwhite.
GRAPH  /SCATTERPLOT(OVERLAY)= pctwhite pctwhite
     WITH crimenl crime (PAIR).
```

The open circles are the actual data. The filled circles are the predicted values and follow the data quite closely. In contrast to the quadratic equation, the predicted values stay flat at low values for *pctwhite*, rather than decreasing. Parameter a = 894 determines where the equation levels off, and is called the asymptote. The other parameters determine whether there is an increase or decrease and how rapid any nonlinear change is. Correlating the observed crime scores and the predicted crime scores produces $r = .772$ and $r^2 = .596$ as shown above in the NLR output. Recall that $R^2 = .617$ for the quadratic regression. Here we have a similarly strong relationship with a single predictor. This "elegant" statistical analysis just needs an equally elegant theoretical explanation. Appendix 9-1 show some examples of nonlinear relationships.

APPENDIX 9-1: EXAMPLES OF NONLINEAR RELATIONSHIPS