

26

Regression Analysis

Linear Regression
The Equation of a Line
The Linear Regression Equation
Standard Error of Estimate
The Multiple Regression Equation
A Walk Through a Computer Printout
A Multiple Regression Example

In Chapter 22, we discussed the use of **correlation** to measure the strength of association between two variables.

In this chapter we extend this concept to **regression analysis**, which allows us to predict the value of a variable from one or more others. **Linear regression** analyzes two variables – one **predicted** variable (called the criterion) and one **predictor** variable. **Multiple regression** analyzes **three or more variables** – one criterion and two or more predictor variables.

The mathematical computations for regression analysis are complex, but with the advent of the personal computer and the development of statistical packages, regression analysis is rapidly becoming the most popular statistical procedure -- particularly in the fields of psychology, sociology and education.

Dr. Martha Bessac studied predictor variables of **marital satisfaction** of 375 student couples at Southwestern Baptist Theological Seminary in 1986.¹ Aware of the increased stress on seminary marriages, including a rise in the number of divorces -- averaging twenty-four per year at that time² -- Dr. Bessac, as part of the Registrar's staff, wanted to determine **what factors might be contributing to this**.

She hypothesized, based on her literature search, the following variables as significant positive predictors of student marital satisfaction: sex (gender), age of husband, age of wife, seminary program of husband, number of semesters husband has been enrolled, number of hours husband enrolled in this semester, number of hours husband has completed towards degree, education level of husband, education level of wife, number of months married, number of children, child density, child spacing ratio, number of hours per week husband is employed, number of hours a week wife is employed, total income, number of hours per week husband engaged in church activities, and number of hours per week wife engaged in church activities.³

¹Martha Sue Bessac, "The Relationship of Marital Satisfaction to Selected Individual, Relational, and Institutional Variables of Student Couples at Southwestern Baptist Theological Seminary," (Fort Worth, Texas: Southwestern Baptist Theological Seminary, 1986)

²*Ibid.*, p. 19

³*Ibid.*, pp. 20-21

⁴*Ibid.*, pp. 41, 44

She found **four significant predictors** accounting for **9.6% of marital satisfaction** variability. These were **Months Married** ($t=-5.428$, $b= -0.054$), **Number of Hours Wife Works** ($t=-2.637$, $b= -0.183$), **Number of Hours Husband Works** ($t=-2.605$, $b= -0.094$), and **Income** ($t=-2.089$, $b= -0.158$). Further, the regression equation produced by the analysis was shown to be a **viable model** ($F=12.925$, $F_{cv}=2.39$).⁴

Notice that all the regression coefficients (b 's) are **negative**. As Months Married increased, marital satisfaction decreased. This is perhaps explained by considering two extreme groups of student couples: one group of newly-weds, in seminary-as-honeymoon mode, compared to older couples with teenage children, leaving behind "home, friends and family" for cramped quarters and hectic schedules.

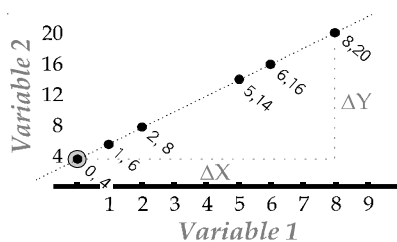
Increased hours of work for both husband and wife meant decreased marital satisfaction. Higher incomes, lower satisfaction. Number of children, degree plan, age, number of credit hours in the semester, hours engaged in church activities -- these and the other specified variables proved not significant.

Since only 9.6% of the variability of marital satisfaction ($Adj. R^2=0.096$) is accounted for by the four predictor variables, **90.4% of the variability of marital satisfaction was not accounted for**. This variability was either accounted for by unnamed variables, or by the unsystematic variation among the 375 couples. Still, the multiple regression procedure declared the model viable by posting a significant F-ratio of 12.925 ($F_{cv}=2.39$).⁵

Before introducing the concepts of regression, however, we need to review the fundamentals of linear equations upon which regression is built.

The Equation of a Line

Do you remember 'way back in high school math when you studied x- and y-coordinate systems? You would spend homework time plotting points on graph paper. You drew lines between points, and developed equations for the lines.



At left is such a graph, which shows a line drawn between the X,Y points of 0,4 and 8,20. This line is defined by a slope and a y-intercept point. The **slope** is defined by the ratio of the change in Y over the change in X ($\Delta Y/\Delta X$). Y-values change 16 points ($20-4=16$) and X values change 8 points ($8-0=8$), yielding a *slope of 16/8 = 2*.

The **y-intercept point** is the point where the line crosses the y-axis. It represents the value of Y when X=0. Here, when X=0, Y=4. Therefore the *y-intercept is 4*.

The equation of a line has the general form of $Y = mX + b$, where m represents the slope and b the y-intercept. The equation of the line in the graph at left is written as **$Y=2X+4$** .

Four other points are shown in the graph. The coordinates of any point on this line can be computed from the linear equation $Y=2X+4$, when given a value of X. The table below shows how each **Y-value** is computed from a **given X-value** and the equation:

X-value	$2X + 4$	Y-value	Coordinates
{0}	-----> $2 \times 0 + 4 =$	-----> {4}	{0,4}
1	$2 \times 1 + 4 =$	6	1,6
2	$2 \times 2 + 4 =$	8	2,8
5	$2 \times 5 + 4 =$	14	5,14
6	$2 \times 6 + 4 =$	16	6,16
8	$2 \times 8 + 4 =$	20	8,20

And if X=100?

100

2x100+4 =

204

100,204

The two elements of slope (m) and y-intercept (b) define a line.

These concepts of **slope and y-intercept** are used in computing a regression equation by which one variable can be predicted by the other.

Linear Regression

Several scattergrams are displayed on page 22-2. The first two scattergrams illustrate "perfect correlations" of +1.00 and -1.00 respectively. Remember that in both cases, the points fell along a straight line. The term "linear" (*lin-ee-er*) derives from the line represented by points in a scatterplot.

We can compute an equation for a line which fits any scatterplot. **Using this equation, we can predict one variable from another:** the stronger the relationship, the more closely the points cluster around a line, and the better the accuracy of the prediction.

Using a process called the **least squares method**, regression analysis produces a *best fit linear equation*. "Best fit?" In chapter 16, we learned that the *sum of deviations* about the mean equals **zero** ($\Sigma x=0$). It is also true that the *sum of squared deviations* about the mean (Σx^2) is a **minimum value**. That is, the sum of squares about the mean is smaller than it would be if computed about any other value. Looking at mean and Σx^2 another way, the **mean of a group of scores produces the smallest sum of squares**. It is a "least squares" measure of central tendency.

Just as the mean is the "best fit point" of a single group of scores, the **linear regression equation is the "best fit line" through a scatterplot of two groups of scores**. It is a "least squares" fit because -- just as $\Sigma x=0$ and $\Sigma x^2 =$ a minimum -- so deviations of scatterplot points about the computed line, called residuals (e), produce the values **$\Sigma e=0$ and $\Sigma e^2 =$ a minimum**. More on this a little later. Let's look at the regression equation.

The Linear Regression Equation

The general equation for the line is $Y = mX + b$. The equation used in linear regression is written like this:

$$Y' = a + bX$$

where **Y'** (pronounced "Y prime") is the *predicted value* of Y, **a** refers to the y-intercept point, and **b** refers to the *slope* of the regression line. Regression analysis produces values for a and b such that we can develop the best fit line through a scatterplot.

Computing a and b

Given a set of scores, how do we calculate the values of a and b? Here are the formulas we use:

$$b = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad a = \bar{Y} - b\bar{X}$$

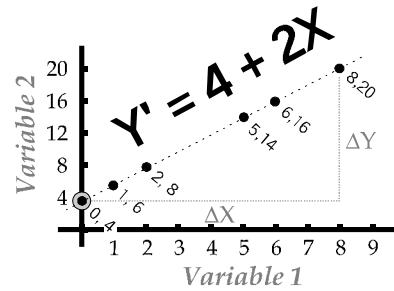
First, compute **b**. The elements of the formula bear a close resemblance to part of the Pearson's r correlation coefficient.

Second, use **b** and the means of X and Y to compute **a**.

Earlier, we computed values of Y from values of X using the equation $Y = 2X + 4$. Let's use those same values, and compute the equation components a and b. **If we do this right, we should get $a = 4$ and $b = 2$.**

The X- and Y- values below come from the computed coordinates at the bottom of 26-2.

XI	<u>X</u>	XY	<u>Y</u>	YI
0	0	0	4	16
1	1	6	6	36
4	2	16	8	64
25	5	70	14	196
<u>36</u>	<u>6</u>	<u>96</u>	<u>16</u>	<u>256</u>
66	14	188	48	568
ΣX^2	ΣX	ΣXY	ΣY	ΣY^2
	196			
	$(\Sigma X)^2$			
Means:	$14/5 = 2.8$		$48/5 = 9.6$	



First, compute **b**.

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{5(188) - (14)(48)}{5(66) - 196} = \frac{268}{234} = 2.0$$

Second, compute **a**.

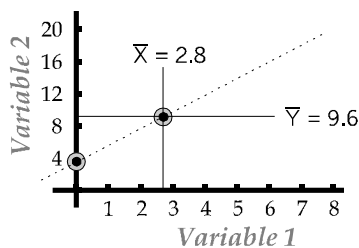
$$a = \bar{Y} - b\bar{X} = 9.6 - 2(2.8) = 4.0$$

Third, substitute the values of a and b into the equation $Y' = a + bX$, which results in

$$Y' = 2X + 4$$

This is the same equation we started with on page 26-2. While we may seem to be going around in circles, we have established the fundamentals of conducting regression analysis -- computing a linear equation from a set of matched scores.

Drawing the Regression Line on the Scatterplot

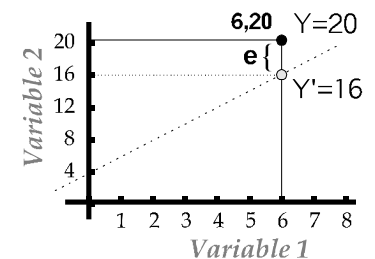


Once we have determined the regression equation, **how do we draw the regression line on the scatterplot diagram?** "Best fit" lines do not always *look* as if they fit the best, since "best fit" is a mathematical, rather than a visual, concept. In order to draw a line, we need two points. We already have one: the **y-intercept (a)**. This is the value of Y when X is 0. The second point is established by the **means of X and Y, which intersect on the regression line**. Establish these two points, and then draw a line through them using a ruler. This generates the "best fit" line through the data. Look at the diagram at right, which shows the two required points in shaded circles.

Errors of Prediction (e)

A *predicted value* of Y (Y'), computed from a regression equation, is identical to the actual Y *only when a perfect correlation exists* between the independent, or predictor, variable (X) and the dependent, or criterion, variable (Y). If the correlation is less than perfect, there will be differences between predicted scores (Y') and actual scores (Y). These **errors of prediction**, or **residuals**, and are defined as $e = Y - Y'$.

Look at the point (6,20) at right. This point does not fall on the regression line. The point labelled Y' indicates where it "should" fall. The vertical distance from the actual data point Y (6,20) to the predicted value on the regression line Y' (6,16) is the residual or error of prediction: $e = Y - Y' = 20 - 16 = 4$.



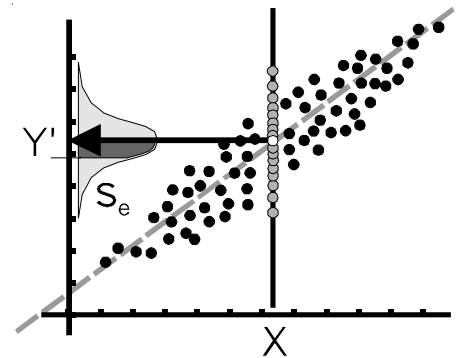
The **residual (e)** is comparable to the deviation (x) as discussed earlier. Residuals are deviations in two-dimensional space. **The sum of residuals about the regression line is equal to zero ($\Sigma e=0$)**, just as the sum of deviations about the mean equals zero ($\Sigma x=0$). When residuals are squared and summed, the result is the **sum of squared residuals (Σe^2)**. When the proper regression line has been computed, the value Σe^2 will be a **minimum**. Any other line drawn through the scattergram will produce a larger Σe^2 than the "best fit" regression line.

Standard Error of Estimate

When two variables have a perfect correlation ($r = 1.00$), then Y can be perfectly predicted from X. In most cases, however, the correlation between two variables is less than perfect. This means that prediction is less than perfect.

For each "X" value (vertical line X at right) there are many "Y" values centered around Y' (gray points on X vertical line).

If we plot data points around the predicted value (Y') on the Y-axis, we create a normal curve. The variability of the residuals can be plotted like the gray area in the normal curve at left. *The standard deviation of this Y-based normal curve is the standard deviation of the residuals, called the **standard error of estimate (s_e)**.* It is computed by the following formula:



$$s_e = \sqrt{\frac{\Sigma e^2}{n - 2}}$$

Compare this equation to the one for estimated population standard deviation (s) on page 16-9. The concepts are the same. The term $n - 2$ is used because two degrees of freedom are lost -- due to having two groups of scores.

Another way to compute the **standard error of estimate** is to use the correlation coefficient (r) as follows:

$$s_e = s_Y \sqrt{1 - r^2}$$

where s_Y is the standard deviation of the Y scores and r is the correlation between X and Y. **The larger the correlation (r) between X and Y, the smaller the term under the radical, and the smaller the standard error of estimate.** As r approaches 1.00, s_e ap-

proaches 0, which reflects a greater accuracy in prediction.

In this section we have reviewed the fundamentals of linear equations, the formulas for computing a and b for the equation $Y' = a + bX$, and the concepts of residuals and the standard error of estimate. But the real power of regression analysis for the complex studies of the social sciences is found in *multiple* regression analyses.

Multiple Linear Regression

Linear regression produces an equation which relates one predictor variable to one criterion variable. But real life problems are seldom so easily explained. *Multiple* regression analysis permits the study of several predictor variables for a given criterion.

Raw Score Regression Equation

The multiple regression equation is much like the simpler linear regression. Each additional predictor variable (X) has a regression coefficient (b) associated with it. The constant term is a . For k predictor variables, the general **raw score equation** is this:

$$Y' = a + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

One drawback to the raw score equation is that the size of b -coefficients is dependent on the scales of predictor variables. They **cannot, therefore, be compared directly** with each other to determine which predictor has the most influence on the criterion. If we could standardize both scores and coefficients, converting them to a single scale, we could make direct comparisons among the coefficients. This is exactly what the standardized regression equation does.

Standardized Score Regression Equation

The standardized equation has no constant term, uses *standardized regression coefficients* (beta, β) instead of b 's, and *z-scores* instead of raw scores (X). In this form, the regression equation looks like this:⁵

$$z'_Y = \beta_1z_{X1} + \beta_2z_{X2} + \dots + \beta_kz_{Xk}$$

Since β 's and z 's are standardized, **they can be directly compared to each other to determine the rank order of influence among the variables on the criterion.** We will see a little later how to use this important information.

Multiple Correlation Coefficient

Just as we can compute the correlation r between two variables, we can also compute the **multiple correlation coefficient, R** , the correlation between the criterion variable (Y') and all the predictor variables taken together.

Just as the coefficient of determination (r^2) tells us "proportion of variance accounted for" between two variables, so the **squared multiple R (R^2)** tells us *what proportion of variance is accounted for in Y' by all the predictor variables in the equation.*

The best way to see what multiple regression offers us is to work through an

⁵The numbers 1...2...3 are subscripts of the X 's, which in turn are subscripts of the z 's. But PageMaker 6.5 does not permit two levels of subscript. Therefore the subscript term " X_1 " is written as " $X1$."

⁶The data for this example is adapted from Howell, 416.

⁷Leland Wilkinson, *A System for Statistics*, Version 4, (Evanston, IL: SYSTAT, Inc., 1988)

actual analysis.⁶ The data analysis was done by the author using SYSTAT.

Multiple Regression Example

A number of years ago the Student Association of a large university published an evaluation of over 100 courses taught during the preceding semester. Students in each course completed a questionnaire in which they rated a number of different aspects of the course on a 5-point scale (1=failure, 5=exceptional). These variables were:

The overall quality of the courses:	OVERALL
The teaching skills of the instructor:	TEACH
The quality of tests and examinations:	EXAM
The instructor's perceived knowledge :	KNOW
The student's expected grade:	GRADE
The enrollment of the class:	ENROLL

The Data

The chart below displays 6 of the data sets collected from 50 courses. Each row is a single course. Scores are mean scores representing the entire class.

	OVERALL	TEACH	EXAM	KNOW	GRADE	ENROLL
1	3.4	3.8	3.8	4.5	3.5	21
2	2.9	2.8	3.2	3.8	3.2	50
3	2.6	2.2	1.9	3.9	2.8	800
4	3.8	3.5	3.5	4.1	3.3	221
49	4.0	4.2	4.0	4.4	4.1	18
50	3.5	3.4	3.9	4.4	3.3	90

The Correlation Matrix

Preliminary analysis of data sets such as these include computing Pearson r correlation coefficients on variables two at a time. These coefficients are efficiently displayed in a form called a *correlation matrix*, which shows all coefficients at one time, as shown below.

Correlation Matrix

N = 50	OVERALL	TEACH	EXAM	KNOW	GRADE
TEACH	0.804				
EXAM	0.596	0.720			
KNOW	0.682	0.526	0.451		
GRADE	0.301	0.469	0.610	0.224	
ENROLL	-0.240	-0.451	-0.558	-0.128	-0.337

With all coefficients shown, *any coefficient* between any two variables can be quickly found. Note the strong positive correlations between TEACH-OVERALL and ENROLL-EXAM above (0.804, -0.558). Notice also that ENROLL is negatively correlated with every other variable (As classes get bigger, student attitudes get more negative).

The Multiple Regression Equation

We can choose any variable of the six to serve as our criterion. However, since

one of them (OVERALL) reflects student attitude toward the course as a whole, it is the most appropriate variable to serve as criterion. TEACH, EXAM, KNOW, GRADE, and ENROLL are appropriate predictor variables.

How well do the predictor variables account for the variance in the criterion OVERALL? And what are the values of **a** (constant) and **b_x** based on the data? Here is our raw score regression model equation:

$$\text{OVERALL}' = \text{CONSTANT} + b_1\text{TEACH} + b_2\text{EXAM} + b_3\text{KNOW} + b_4\text{GRADE} + b_5\text{ENROLL}$$

The Essential Questions

We're looking for the values of CONSTANT (a) and the regression coefficients (b's). Beyond this, the multiple regression analysis provides information to answer three essential questions:

First, how much criterion variability is accounted for by the predictors?

Look for the **Adjusted Squared Multiple R value**.

Second, which predictor variables are significant?

Look for individual predictors whose **t-values are significant** (p<0.05).

Third, is the model-as-a-whole viable? Is it a good model?

Look for a **significant F-ratio** (p < 0.05).

Multiple Regression Printout

DEP VAR: OVERALL N: 50 MULTIPLE R: .869 SQUARED MULTIPLE R: .755
 ADJUSTED SQUARED MULTIPLE R: .728 STANDARD ERROR OF ESTIMATE: 0.320

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	-1.195	0.631	0.000	1.0000000	-1.893	0.065
TEACH	0.763	0.133	0.662	.4181886	5.742	0.000
EXAM	0.132	0.163	0.106	.3245736	0.811	0.422
KNOW	0.489	0.137	0.325	.6746330	3.581	0.001
GRADE	-0.184	0.165	-0.105	.6196885	-1.114	0.271
ENROLL	0.001	0.000	0.124	.6534450	1.347	0.185

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	13.934	5	2.787	27.184	0.000
RESIDUAL	4.511	44	0.103		

The above SYSTAT multiple regression printout of the data has *three distinct sections* which relate to the *three questions* stated above. These are delineated by dotted lines which are not normally seen in a printout. We will now take each section in turn.

Section One

DEP VAR: OVERALL N: 50 MULTIPLE R: .869 SQUARED MULTIPLE R: .755
 ADJUSTED SQUARED MULTIPLE R: .728 STANDARD ERROR OF ESTIMATE: 0.320

The first section of the regression printout, shown above, includes the elements defined below. The specific values for this example are displayed in brackets [].

DEP VAR: The dependent variable (criterion). [OVERALL]

N: Number of cases or subjects in the study. [50]

MULTIPLE R: Correlation between OVERALL and the predictors. [0.869]

SQUARED MULTIPLE R: Proportion of variance in OVERALL accounted for by predictors. [0.755]

ADJUSTED SQUARED MULTIPLE R: If you were to use a multiple regression equation with another set of data, the R^2 value from the second data set would be *smaller* than the R^2 produced by the original data set. This reduction in R^2 is called *shrinkage*. The adjustment depends on the number of subjects (N) and the number of variables (k) in the study. This is the “true” value of R^2 . [0.728]

STANDARD ERROR OF ESTIMATE: The standard deviation of the residuals. [0.320]

The answer to our first question is that the five predictor variables account for 72.8 percent of the variability of OVERALL (Adj.RI = 0.728).

Section Two

<u>VARIABLE</u>	<u>COEFFICIENT</u>	<u>STD ERROR</u>	<u>STD COEF</u>	<u>TOLERANCE</u>	<u>T</u>	<u>P(2 TAIL)</u>
CONSTANT	-1.195	0.631	0.000	1.0000000	-1.893	0.065
TEACH	0.763	0.133	0.662	.4181886	5.742	0.000
EXAM	0.132	0.163	0.106	.3245736	0.811	0.422
KNOW	0.489	0.137	0.325	.6746330	3.581	0.001
GRADE	-0.184	0.165	-0.105	.6196885	-1.114	0.271
ENROLL	0.001	0.000	0.124	.6534450	1.347	0.185
	b's	s_b	β's	multicollinearity	t=b/s_b	p(t)

The second section of the printout, shown above, details the **analysis of each predictor individually**. It is this part of the printout that provides the regression coefficients (b 's and β 's) as well as their significance tests.

VARIABLE: Heading for the variable names in the regression model. “Constant” is the value of OVERALL when all predictors equal zero.

COEFFICIENT: Heading for the values of the respective regression coefficients (the “ b 's”) of the regression equation. *Using these values, you can write the regression equation for OVERALL and the five predictors as follows:*

$$\text{OVERALL}' = -1.195 + 0.76\text{TEACH} + 0.13\text{EXAM} + 0.49\text{KNOW} - 0.18\text{GRADE} + 0.001\text{ENROLL}$$

\	\	\	\
predicted	Constant: the value	Regression coefficient:	Variable: Use
score for	of OVERALL when all	Multiply this by the	the raw scores
OVERALL	predictors = 0	value of VARIABLE	for these

Given the mean scores of the five predictors for any class, we can predict what that class' OVERALL score will be.

STD ERROR: Standard deviation of the regression coefficient (b). It is used in a t-test to determine whether the “ b ” is significant.

STD COEFFICIENT: Standardized regression coefficients, or beta weights (β). Betas are to b 's what z-scores are to X 's. While the b 's are used with raw scores in regression equations, as in **0.76TEACH** above, the betas are used with z-scores. The beta for TEACH equals 0.662. The proper term for TEACH in a standardized regression equation is **0.662z_{TEACH}**.

Because betas are standardized, they can be directly compared according to relative strength. The “ b ”s cannot be compared directly because they usually represent different score ranges. ENROLL ranges from a low of 7 to a high of 800, while the other scales range from 1 to 5. The standardization of the betas eliminates this problem of differing ranges, just as z-scores eliminates the problem of comparing raw scores with differing variabilities. *In our example, we see that TEACH is more than six times more influential than EXAM, and twice as influential as KNOW.*

TOLERANCE: The ideal condition in multiple regression analysis is for each predictor variable to be related to the criterion, but *not to other predictor variables*. Predictor variables are supposed to be independent of each other — but they rarely are. Tolerance values near zero (0) in this printout indicate that some

of the predictors are highly intercorrelated. This undesirable situation is called “multicollinearity.” Look for tolerance values near 1.0.

T: If you divide the value of each regression COEFFICIENT by its respective STD ERROR, you will get the values in this column. For example, the test for the “b” on the variable TEACH is equal to $0.763/0.133 = 5.742$. The t-test values are used to answer the question *Is this predictor significant?*

	T					
CONSTANT	-1.195	0.631	0.000	1.0000000	-1.893	0.065
>>> TEACH	0.763	0.133		=	5.742	0.000
EXAM	0.132	0.163	0.106	.3245736	0.811	0.422
KNOW	0.489	0.137	0.325	.6746330	3.581	0.001
GRADE	-0.184	0.165	-0.105	.6196885	-1.114	0.271
ENROLL	0.001	0.000	0.124	.6534450	1.347	0.185

P(2 TAIL): The probability of obtaining the computed t-value. For TEACH, p is very small – less than 0.001 – that we would get a t-value of 5.742 if $b_{teach}=0$. Therefore, we say that TEACH is a significant predictor. There is a 42.2% (0.422) chance of getting the t-value of 0.811 for EXAM by chance. This is not a significant predictor since $p>0.05$.

The answer to our second question is that TEACH and KNOW are significant predictors of OVERALL. EXAM, GRADE, and ENROLL are not.

Section Three

ANALYSIS OF VARIANCE					
SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	13.934	5	2.787	27.184	0.000
RESIDUAL	4.511	44	0.103		

The third section of the printout details the **analysis of the model as a whole**. Is the model, as represented by the regression equation being tested, a viable one?

SOURCE: There are two sources of variance in regression analysis. One is from the regression itself, and the other is from variance unaccounted for after the regression analysis. The predicted score (Y') seldom equals the criterion score (Y). There is always some error of estimate (e), such that $Y = Y' + e$. We can therefore divide the criterion scores into two parts: Y' (regression) and e (residual).

SUM-OF-SQUARES: This is the sum of squared deviations about the regression line. The total sum of squares is divided between “accounted for” REGRESSION (Y') and “unaccounted for” RESIDUALS (e).

Regression SS: This is the sum of squared deviations of Y' about the mean of Y : $\sum(Y' - \bar{Y})^2$

Residual sum of squares: This is e , which equals $\sum(Y - Y')^2$

DF: Degrees of freedom.

DF_{reg} equals the number of variables minus 1 [$df_{reg}=6-1=5$].

DF_{res} equals the number of subjects minus the number of variables [$df_{res}=50-6=44$].

MEAN-SQUARE: The mean-square terms are variances.

MS_{reg} equals SS_{reg} divided by df_{reg} .

MS_{res} equals SS_{res} divided by df_{res} .

F-RATIO: The F-ratio equals MS_{reg}/MS_{res} and is used to determine if the variance due to regression is enough greater than the variance due to residual noise to render the model “significant” (or “viable”).

P: The probability of the computed F-RATIO being this large by chance. Any time $p<0.05$, a significant model is indicated. In our case, a P of “0.000” is very small – less than 0.001 – and indicates a significant model.

The answer to our third question is that we do have a viable model. The F-ratio is significant.

Focus on the Significant Predictors

Notice in Section 2 that only two predictors, TEACH and KNOW, are significant. That is, only these two variables have t-test probabilities of 0.05 or less.

TEACH	0.763	0.133	0.662	.4181886	5.742	0.000
KNOW	0.489	0.137	0.325	.6746330	3.581	0.001

Since the other variables are not significant, let's analyze another model which includes only these two predictors.

$$\text{OVERALL}' = \text{CONSTANT} + b_1\text{TEACH} + b_2\text{KNOW}$$

DEP VAR: OVERALL N: 50 MULTIPLE R: .860 SQUARED MULTIPLE R: .739
ADJUSTED SQUARED MULTIPLE R: .728 STANDARD ERROR OF ESTIMATE: 0.320

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P (2 TAIL)
CONSTANT	-1.298	0.477	0.000	1.0000000	-2.720	0.009
TEACH	0.710	0.101	0.616	.7230389	7.021	0.000
KNOW	0.538	0.132	0.358	.7230389	4.082	0.000

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	13.627	2	6.814	66.467	0.000
RESIDUAL	4.818	47	0.103		

Study the printout above and the analysis below carefully:

First, did reducing the number of predictors from 5 to 2 *reduce the amount of variance accounted for* in OVERALL? The Adjusted R-Square is 0.728, exactly the same as we found with five predictors. We lost nothing here, which is good.

Second, did we *increase the standard error of estimate*? No, it is 0.320, the same as before. We gained nothing here, which is good.

Third, are the two *predictors significant*? Yes, both TEACH and KNOW show large t-test scores and very low probabilities (0.000 means $p < 0.001$, "very small"). This is good.

Fourth, is our *model more sound*? The F-Ratio is larger, showing a better ratio of regression to noise. Notice the values of the sum-of-squares are not much different from before, but the change in regression df from 5 to 2 produces a larger MEAN-SQUARE value.

In conclusion, this second model is better. For these 50 courses, we can account for nearly 73% of the students' ratings of courses by knowing their ratings of the instructor's TEACHing skills and their instructor's perceived KNOWledge of the subject. ENROLLment in the class, quality of EXAMs, and the students' anticipated GRADEs are not significant predictors of OVERALL quality.

Multiple Regression Equations

Our resulting raw score and standardized equations are:

$$\text{OVERALL}' = -1.298 + 0.71\text{TEACH} + 0.538\text{KNOW}$$

$$z'\text{OVERALL} = 0.616z\text{TEACH} + 0.358z\text{KNOW}$$

Summary

In this chapter you have been introduced to the world of regression analysis. You have seen how scattergrams of data can be reduced to a single predictor equation. You have learned how to compute the two variables in a linear regression line: a and b. You have learned how to read a computer printout from a multiple regression analysis.

Example

Dr. Dean Paret studied the relationship between “nuclear family health” and selected “family of origin” variables among 302 married subjects in 1991.⁸ The criterion (predicted) variable was “overall perceived nuclear family health” {FUNC}-- as measured by the Family Adaptability and Cohesion Evaluation Scale (FACES-R).⁹

Predictor variables related to family of origin. Autonomy {AUTON} measures an individual's sense of independence and self-reliance. It includes free expression, responsibility, mutual respect, openness and experiences of separation or loss.)¹⁰

The second predictor was Intimacy {INTIM}, which reflects close, familiar and usually affectionate or loving personal relationships without feeling threatened or overwhelmed. It includes expression of feelings, sensitivity and warmth, mutual trust, and the lack of undue stress in conflict situations.¹¹ Both AUTON and INTIM were measured by the Family-of-Origin Scale (FOS).¹²

Additional demographic variables were gathered by means of a questionnaire: educational level {EDUC}, degree program {DEGREE}, number of years in graduate school {YRS}, income level {SALARY}, sex of participant {SEX}, and whether or not the couple was a dual career family {DUAL}.¹³ Here is Dr. Paret's final printout:

MULTIPLE REGRESSION PRINTOUT¹⁴

DEP VAR:	FUNC	N:	302	MULTIPLE R:	.898	SQUARED MULTIPLE R:	0.806
ADJUSTED SQUARED MULTIPLE R:	.804	STANDARD ERROR OF ESTIMATE:					34.973
VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P (2 TAIL)	
CONSTANT	61.373	11.220	0.000	.	5.470	0.000	
AUTON	1.256	0.122	0.647	0.1656319	10.320	0.000	
EDUC	11.206	4.642	0.063	0.9462106	2.414	0.016	
INTIM	0.406	0.126	0.200	0.1700070	3.224	0.001	
YRS	7.588	1.973	1.107	0.8370002	3.847	0.000	
ANALYSIS OF VARIANCE							
	SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P	
	REGRESSION	1513896.178	4	378474.045	309.434	0.000	
	RESIDUAL	363265.557	297	1223.116			

⁸Dean Kevin Paret, “A Study of the Perceived Family of Origin Health as It Relates to the Current Nuclear Family in Selected Married Couples,” (Fort Worth, Texas: Southwestern Baptist Theological Seminary, 1991)

⁹*Ibid.*, p. 39

¹⁰*Ibid.*, p. 53

¹¹*Ibid.*, pp. 53-54

¹²*Ibid.*, p. 38

¹³*Ibid.*, p. 39

¹⁴*Ibid.*, Table 15, p. 149

Question One: How much **variability** of the subjects' family health (FUNC) was accounted for by **Family-Of-Origin autonomy and intimacy**, and the **demographic variables** of education level (EDUC: high school, college, graduate school) and years in seminary (YRS: 1, 2, 3, 4+)?

The Adjusted R^2 value (0.804) answers this question: **80.4%**. Less than 20% of the variability in current family health is unaccounted for. *This is a very strong finding.*

Question Two: Which **predictor variables are significant?** What is the order of influence of these variables on family health?

Since this is the fourth and final printout in a series, all nonsignificant predictors have been eliminated (DEGREE, DUAL, SALARY, SEX). **All of the variables listed above show $p(t)$ values less than 0.05.**

The rank order of influence is given by the **beta-values** under the heading "STD COEF." **Autonomy** (AUTON) has by far the **greatest influence** on family health (FUNC) with $\beta=0.647$. **Intimacy** (INTIM) is next with $\beta=0.200$, followed by **years enrolled in seminary** (YRS) with $\beta=1.107$, and finally **educational level** (EDUC) with $\beta=0.063$. The raw and standardized regression equations for Dr. Paret's study are

$$\text{FUNC}' = 61.373 + 1.256\text{AUTON} + 11.206\text{EDUC} + 0.406\text{INTIM} + 7.588\text{YRS}$$

$$z'_{\text{FUNC}} = 0.647z_{\text{AUTON}} + 0.063z_{\text{EDUC}} + 0.200z_{\text{INTIM}} + 0.107z_{\text{YRS}}$$

Question Three: Is this a **viable model?** Does it adequately predict family health among the 302 subjects?

The answer to this question is found in the ANOVA table and F-ratio. The F-ratio of 309.434 ($p<.001$) tells is **this is a very strong model.**

How important are the variables EDUC and YRS for the FUNC model? Dr. Paret dropped these out of his full model and produced the following:

MULTIPLE REGRESSION PRINTOUT¹⁵

DEP VAR: FUNC N: 302 MULTIPLE R: .890 SQUARED MULTIPLE R: 0.792
ADJUSTED SQUARED MULTIPLE R: .791 STANDARD ERROR OF ESTIMATE: 36.121

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P (2 TAIL)
CONSTANT	90.925	8.139	0.000	.	11.171	0.000
AUTON	1.350	0.124	0.696	0.1702486	10.887	0.000
INTIM	0.425	0.130	0.209	0.1702486	3.269	0.001

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	1487058.544	2	743529.272	569.888	0.000
RESIDUAL	390103.191	299	1304.693		

How much did Adj. R^2 change? The amount of variance-accounted-for dropped from 0.804 to 0.791, a change of -0.013, or a little over one percent. *This is good.* We did not lose much R^2 by dropping two of the four predictor variables.

Are AUTON and INTIM still significant predictors? **Yes** ($p<0.001$, $p=0.001$).

Did the model suffer from dropping EDUC and YRS? **No.** The F-ratio is larger than before, showing a **smaller, stronger model.**

¹⁵*ibid.*, p. 150

¹⁶*ibid.*, p. 101

Family patterns of relationship transfer from generation to generation. Healthy family relationships are rooted in the degree of autonomy and intimacy experienced in the family of origin. Likewise, dysfunctional family relationships are rooted in family of origin disfunction. These same patterns show up in seminary couples. Some enter the ministry to “help others” out of the need to help self because of a dysfunctional family background. Inability to establish healthy relationships in family has been found to transfer to the ministry: such ministers “have difficulty forming ministerial relationships in the pastorate.”¹⁶

The challenge to seminaries is to go far beyond “teaching students how to minister,” but requires also helping dysfunctional students break with past patterns and learn anew how to establish autonomous and appropriately intimate relationships with others. The health of our churches is at stake.

Vocabulary

adjusted squared multiple R	Multiple correlation coefficient after adjusted for shrinkage
correlation matrix	representation of multiple variables and their intercorrelations
criterion variable	predicted or dependent variable in multiple regression (Y')
linear equation	mathematical formula which describes a straight line ($Y = 2X + 3$)
linear regression	predicting one variable by another by best-fit line through scattergram
multicollinearity	degree of inter-correlation among predictor variables
multiple correlation coefficient	correlation between the criterion variable and all predictors together (R)
multiple linear regression	prediction of one variable by 2+ others
predictor variable	variable(s) used to estimate a criterion variable in regression analysis
regression sum of squares	sum of squared deviations between Y' and the mean of Y
regression coefficient	raw score correlates of criterion variable in regression (b)
residual sum of squares	sum of squared deviations between Y and Y' (e^2)
residual	difference between true Y value and the predicted value of Y ($e = Y - Y'$)
Shrinkage	Reduction in R^2 value when equation is applied to new data
slope	one of two determiners of a regression line: $m = (Y/X)$
squared multiple R	proportion of variance of Y accounted for by all predictors (R^2)
standard error of estimate	standard deviation of the residuals
standardized regr'n coefficient	standardized score correlates of criterion variable in regression (r)
tolerance	reflects the degree of multicollinearity among predictors
y-intercept	one of two determiners of a regression line: value of Y when $X=0$

Study Questions

1. Draw a set of axes. Label the X-axis (horizontal) from 0 to 10 and the Y-axis (vertical) from +4 to -1. Compute 10 values for Y when $X=1, 2, \dots, 10$ with the equation $Y = -0.5X + 4$. Plot the 10 points on your axes.
2. Define e . Show how it is calculated.
3. Work through the explanation of the first regression printout using the second printout on page 26-10. Identify and define each of the following elements:

a. Dep var	i. Tolerance
b. N	j. Multiple R
c. Squared multiple R	k. P(2 tail)
d. Adjusted squared multiple R	l. Regression sum-of-squares

- e. Standard error of estimate
 f. Coefficient
 g. Std error
 h. Std coef
- m. Residual df
 n. Regression Mean-square
 o. F-ratio
 p. P

4. A regression analysis was done on the data given below. Draw a scatterplot of the data. Compute a and b , and r , then draw the proper regression line on the scatterplot. Study the regression printout below and describe your findings. Include R , Adj. R -squared, coefficients, t -test values and probabilities, and the F -ratio. **The following data are scores from 15 students on Bible knowledge test scores (Y) and the number of semester hours of Bible in college (X).**

X 15 18 18 12 9 9 6 12 15 12 12 12 15 12 18
Y 23 27 30 19 18 21 17 21 27 29 25 22 26 25 24

MULTIPLE REGRESSION PRINTOUT
 DEP VAR: KNOW N: 15 MULTIPLE R: .728 SQUARED MULTIPLE R: .530
 ADJUSTED SQUARED MULTIPLE R: .494 STANDARD ERROR OF ESTIMATE: 2.792

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P (2 TAIL)
CONSTANT	13.066	2.845	0.000	1.0000000	4.593	0.001
HOURS	0.810	0.212	0.728	1.0000000	3.828	0.002

ANALYSIS OF VARIANCE						
SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P	
REGRESSION	114.259	1	114.259	14.657	0.002	
RESIDUAL	101.341	13	7.795			

Sample Test Questions

- Given the linear equation $Y = 5X - 10$, a value of $X=3$ would yield a Y of
 - +3
 - 10
 - +5
 - 5
- The two-dimensional "e" relates most closely to the one-dimensional
 - X
 - x
 - s
 - z
- Given the constant $(a) = 4$ and regression coefficient $(b) = 3$, the correct linear regression equation is
 - $Y' = 4X + 3$
 - $Y' = 4 + 3X$
 - $Y' = 3X + 4$
 - $Y' = a + 7X$
- The standardized regression coefficient is represented by the letter ____ and is used with _____ in the regression equation.
 - b , raw scores
 - b , z scores
 - β , raw scores
 - β , z scores