A REVIEW OF CORRELATION AND REGRESSION

We begin by going back to the basics of the regression model in this chapter. Our goal is to build up the components of the standard multiple regression model step-by-step first, before introducing major variations of that model in the chapters ahead.

To introduce this model, we unpeel the layers of the regression framework to get down to essential core concepts that must be understood first, such as widely used but often undefined terms like "association," "independence," "controlling for," and "effect." In doing this, we also take positions on some of the foundational assumptions of the regression framework.

If this chapter is purely review, it could be skipped. But we encourage readers to start here, if only to see how we discuss these starting points.

1.1 ASSOCIATION IN A BIVARIATE TABLE

This section asks an important question: What does it mean to say that variables are "related," or associated? To investigate these terms, we look at an example from a study by Radelet (1981), assessing which individuals convicted of murder received the death penalty in Florida in 1976–1977. This is a widely used example, because it is in part so clear, and in part also sobering.

To look at the possibility of an association between the race of the victim and the likelihood of a death penalty verdict, we produce a cross-tabulation of these variables. This cross-tabulation shows the joint values of two variables: whether the victim was Black or White (variable name **victim)** and whether the death penalty (variable name **penalty)** was given in each case. We consider the race of the victim an independent variable, possibly influencing the likelihood of the death penalty—the dependent variable in this example. Notice that the column variable in the Table 1.1 is the independent variable (race of the victim), and the row variable is the dependent variable (death penalty or not).

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Here are the frequencies for this two-way table:

TABLE 1.1 ⬢ A CROSS-TABULATION OF THE RACE OF THE VICTIM AND THE USE OF THE DEATH PENALTY IN 326 HOMICIDE CASES

		Victim		
		Black	White	Total
Penalty				
no	Frequency	106	184	290
yes	Frequency	6	30	36
Total	Frequency	112	214	326

There are 326 cases considered in this table. The table shows the marginal frequencies for the race of the victim (bottom row) and the marginal frequencies for the death penalty (right-hand column) overall. These are the same as the one-way frequencies. The cell frequencies show the number of cases for each possible joint value of the two variables.

Of course, all we have here are the raw frequencies. We need to make the pattern of frequencies more interpretable—to show whether there is a relationship between the two variables and what kind of relationship it is.

Table 1.1 is also hard to interpret in this form, mainly because the row and column marginal totals differ from each other. We want to see whether the tendency to receive the death penalty was related to the race of the victim. The most straightforward way to see this is to calculate the percentages in each category of the independent variable—that is, the percentages in each column. If those percentages are the same, there is no association. If they vary, there is some association.

Here is the same table with the column percentages shown:

TABLE 1.2 ⬢ DEATH PENALTY OUTCOME BY RACE OF THE VICTIM, WITH PERCENTAGES

		Victim		
		Black	White	Total
Penalty				
no	Frequency	106	184	290
	Col Pct	94.64	85.98	
yes	Frequency	6	30	36
	Col Pct	5.36	14.02	
Total	Frequency	112	214	326

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

It is now easy to see that there is a greater tendency to receive the death penalty when the victim is White, compared to cases when the victim is Black. It looks like the race of the victim may be ***associated*** with the chances of the death penalty. We can interpret the table this way: If the victim is Black, only about 5% of convicted murderers receive the death penalty, but when the victim is White, the percentage who receive the death penalty increases to about 14%.

We can refer back to probability theory to state exactly what is meant by "independence" versus "association." A relationship means that there is some form of association, or *dependence*, between the two variables. In probability theory terms, this means that the probability of Y changes with the categories of X. In this specific example, this means that the probability of the death penalty changes with the race of the victim. Notice that this probability is in fact shown in Table 1.2, since it is just the proportion in each category of victim race that receive the death penalty. So, the probability of the death penalty in cases where the victim is Black is .0536, and the probability of the death penalty in cases where the victim is White is .1402.

1.1.1 Probability Rules for Defining Independence

A formal definition of independence of variables allows us to detect departures from independence. If two "events" X and Y are independent, this means

$$Pr(Y/X) = Pr(Y)$$

In words, the probability of Y, given a level of X ($Pr\ (Y/X)$), is the same as the probability of Y overall. It would help to translate this statement into the world of variables: The probability of a given specific category of Y occurring, given a specific category of X, is the same as the overall (marginal) probability of Y. In other words, knowing the category of X tells us nothing about the probability of Y.

Notice in Table 1.2 the overall probability of the death penalty can be found from the marginal frequency in that row divided by the total number of cases—that is, 36 / 326 = .11. But the probabilities across the two categories of victim race vary around this overall probability quite a bit (.05 to .14).

Under the assumption of independence, the probability of any combination of X and Y values is

$$Pr(X_i \text{ and } Y_j) = Pr(X_i) \cdot Pr(Y_j)$$

In words, the probability of the joint occurrence of a particular value of X and a particular value of Y is equal to the ***overall*** probability of the ith value of X times the overall probability of the jth value of Y. What this rule says, in more specific terms, is this: If X and Y are independent, then the probability of any joint value of X and Y can be inferred from the overall probabilities of each. The probability rule above says that we can find the specific probability of any combined values of X_i and Y_j – if independent – by taking successive proportions: The proportion of people with the value Y_j as a proportion of the people with the value X_i.

But, if Y depends on X, we know that the probability of specific categories of Y will differ across categories of X. From the definition of conditional probability

$$Pr(Y_j/X_i) = \frac{Pr(X_i \text{ and } Y_j)}{Pr(X_i)},$$

$$\text{we can derive } Pr(X_i \text{ and } Y_j) = Pr(X_i) \cdot Pr(Y_j/X_i)$$

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

If this is the case, then, using the rule for independent probabilities will result in observed probabilities in the table that depart from the actual probabilities. The more that observed probabilities depart from probabilities expected under the assumption of independence, the greater the evidence of an association, or in other words, a dependence between the variables.

In Table 1.2, we can use the marginal frequencies from the table to find the overall probabilities of each category of X and Y:

$$Pr\ (Death\ Penalty = No) = 290 / 326 = .89$$
$$Pr\ (Death\ Penalty = Yes) = 36 / 326 = .11$$
$$Pr\ (Victim = Black) = 112 / 326 = .34$$
$$Pr\ (Victim = White) = 214 / 326 = .66$$

So, for example, if the race of the victim and the likelihood of the death penalty are independent, then the probability of the death penalty when the victim is Black would be the following:

$$Pr(X_1) \cdot Pr(Y_2) = .34 \times .11 = .037$$

However, the actual probability of the death penalty given a Black victim is

$$Pr(X_1 \text{ and } Y_2) = 6 / 326 = .018$$

The actual probability is one half the probability implied by independence.

As another example: under the assumption of independence, the probability of the death penalty given a White victim is:

$$Pr(X_2) \cdot Pr(Y_2) = .66 \cdot .11 = .073$$

However the actual probability is:

$$Pr(X_2 \text{ and } Y_2) = 30 / 326 = .092$$

which is somewhat higher.

If we set up a definition of what is implied by independence in every cell in the table, we could test the overall departure from independence to assess the existence of an association between variables. Of course, this only begins to define the many characteristics implied by the term association: There are also considerations of direction, strength, and form of the association—all issues we will eventually deal with.

We have seen how association is defined in tables, for nominal or ordinal categorized variables. Next we look at the association between variables that are interval or ratio level.

1.2 CORRELATION AS A MEASURE OF ASSOCIATION

We will use data from the United States to investigate the relationship between the mobility rate and the divorce rate in nine geographical regions. Assume that the divorce rate is measured as the number of persons per 1000 population getting a divorce or annulment. The mobility rate is the percentage of people living in a different house than five years earlier. This example is restricted to an N of 9 observations (for 9 regions), so that we can more easily see how the association works.

Do not copy, post, or distribute

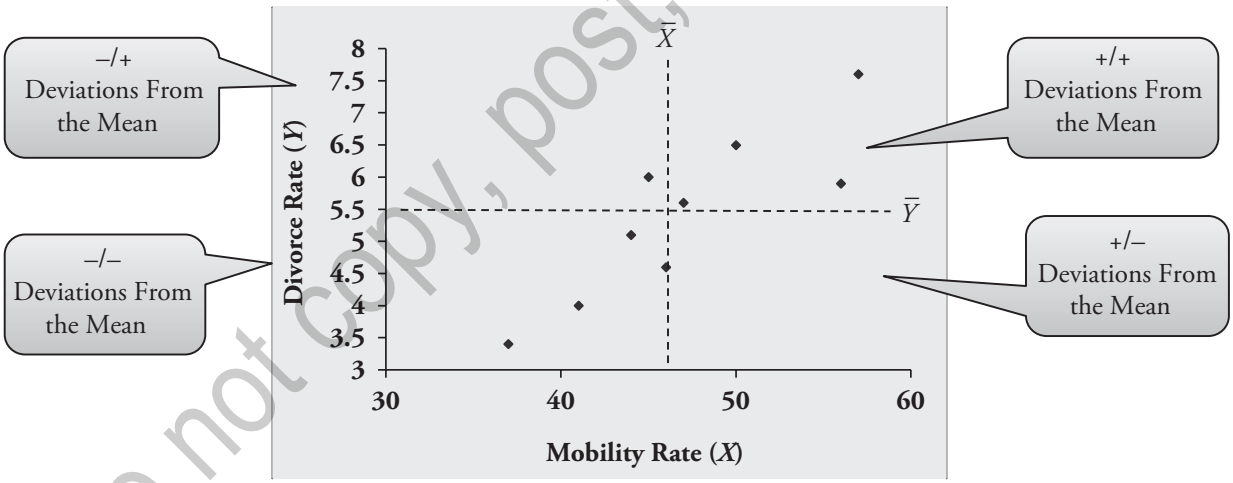
Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The data are shown in Table 1.3.

TABLE 1.3 ⬢ MOBILITY AND DIVORCE RATES FOR NINE REGIONS OF THE UNITED STATES, 1960

Region	Mobility Rate	Divorce Rate
New England	41	4.0
Middle Atlantic	37	3.4
E. North Central	44	5.1
W. North Central	46	4.6
South Atlantic	47	5.6
East South Central	44	6.0
West South Central	50	6.5
Mountain	57	7.6
Pacific	56	5.9

FIGURE 1.1 ⬢ SCATTERPLOT OF THE MOBILITY RATE AND DIVORCE RATE IN NINE REGIONS OF THE UNITED STATES



Assume you are interested in the ***amount*** of association between these two variables. To see visually the possibility of an association, construct a scatterplot showing the joint data points for all nine regions on values of mobility and divorce (see Figure 1.1).

The scatterplot shows the two variables are definitely "associated," or to use a synonymous term, related. Higher mobility rates are associated with higher divorce rates, *on average*. We say "on average" because there will always be exceptions, deviations from this tendency. But the issue here is to capture the pattern, not each exception.

The question is, how closely related are the mobility rate and the divorce rate? Is there a way to state the level of association on a common scale so that it can be compared across different pairs

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

of variables or situations? How can we develop a measure of association for these data that is sensitive to the direction and strength of the association, assuming linearity? (Note: We need to assume something about form of the association at the outset.)

1.2.1 Developing a Measure of Association for Interval/Ratio Data

We have added the mean of X and Y to the scatterplot, thus dividing the plot into quadrants. From the data, we calculate that

$$\bar{X} = 46.89 \qquad \bar{Y} = 5.41$$

$$s_X = 6.56 \qquad s_Y = 1.29$$

$$s_X^2 = 43.111 \qquad s_Y^2 = 1.674$$

This shows the mean ($\bar{X}, \bar{Y}$), the standard deviation (s_X, s_Y), and the variance (s_X^2, s_Y^2) of X and Y in the scatterplot. Notice almost all of the points fall in the lower left and upper right quadrants of the scatterplot.

1.2.1.1 Developing a Measure

Version 1. The Sum of the Cross-Products

Use deviation scores on X and Y, so that ***both*** are positive points in the upper right quadrant, both are negative for points in lower left quadrant (with combined signs either +/+ or -/-), and the combined signs would be (+/ -) and (- /+) in the upper left and lower right quadrants. Using that information, you can sum these individual deviations from the mean, as in

$$\sum_{i=1}^{N} (X_i - \bar{X}) \cdot (Y_i - \bar{Y})$$

This is called the ***sum of the cross-products.*** The summation operator here ($\sum$) denotes summing across all units (i) in a sample, from 1 to N. Unless applied differently, we will drop this notation to simplify the presentation.

This measure is more positive the more points that fall in lower left and upper right quadrants (each with positive cross-products) compared to the upper left and lower right quadrants (each with negative cross-products). Thus, this measure is also more negative the more points in upper left and lower right quadrants relative to the other two. If points occur equally in all quadrants, they will cancel out when summed, since half will be positive and half will be negative. Thus, this measure is sensitive to the direction and strength of the association.

But there is a problem: The sum depends on sample size. You can increase its size just by increasing the size of the sample. This leads to the next necessary component of a useful measure of association: the covariance.

Version 2. The Covariance

Divide the sum of the cross-products by $N - 1$ to get

$$\frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{N - 1} = \text{the covariance, a } \textbf{\textit{very}} \text{ important associational statistic.}$$

The covariance is like the average amount of association per individual. (Note: N –1 is used to take into account that there are $N - 1$ degrees of freedom, since the last observation on X and Y can be deduced from the other $N - 1$ observations, because the sum of deviations around the mean is by definition zero.)

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

There is still a problem, however: X and Y have very different size units. The sum will change depending on the size of the units of each variable. If you multiply each divorce score by 10, the sum will change accordingly. This is fine if one wants the association to reflect the units of the variables, but if one wants the association to reflect how ***relative*** differences on one variable are related to the same relative differences on the other, or in other words, if one wants to judge the ***strength*** of the association, the two variables must be in the same units. This leads us to the final component defining the correlation.

Version 3. The Correlation

The correlation "equalizes" the units of the two variables by dividing each variable's deviation score by the size of its standard deviation. This puts the two variables in a standardized metric, equal to a Z score on each variable:

$$r_{xy} = \frac{\Sigma\left(\frac{(X_i - \bar{X})}{s_x} \cdot \frac{(Y_i - \bar{Y})}{s_y}\right)}{N-1} = \Sigma \frac{Z_x \cdot Z_y}{N-1}$$

In this version—the actual correlation—deviations on X and Y are measured relative to the standard deviation of each, thus removing differences in the size of the units. Scores on both variables are Z scores: They stand for the number of standard deviations above or below the mean a given X or Y score is.

Theoretically, this value varies between -1 (a perfect negative correlation) to +1 (a perfect positive correlation), with 0 standing for no correlation—in other words, no association. In the latter situation, you would see equally scattered points in all quadrants of the plot. As you can see in the plot, the correlation here is very high.

The formula for the correlation is often stated in an equivalent form. If you rearrange the formula above, by inverting the divisor and multiplying, you get

$$r_{xy} = \Sigma\left(\frac{(X_i - \bar{X})(Y_i - \bar{Y})}{s_x \cdot s_y} \cdot \frac{1}{N-1}\right) = \Sigma\left(\frac{(X_i - \bar{X})(Y_i - \bar{Y})}{N-1} \cdot \frac{1}{s_x \cdot s_y}\right)$$
$$= \Sigma \frac{\frac{(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}}{s_x \cdot s_y} = \frac{s_{xy}}{s_x \cdot s_y}$$

Note the shorthand notation used for the covariance in the numerator: s_{xy}. When the correlation between mobility and the divorce rate is calculated, it is .854. (You should be able to calculate the sum of the cross-products here as 58.10.)

1.2.2 Factors Affecting the Size of *r*

1. The more dissimilar the distributions of X and Y and/or the more skewed their distributions, the lower the possible value of r.
2. Unreliability—random error components—in the measurement of X or Y, so that observed scores imperfectly reflect true X or Y values, will introduce noise into r and lower it.
3. When the range of values on X or Y is restricted, r will be lower (because larger deviations have a larger influence on the numerator than denominator of r). This is usually the result of problems in sampling: An incomplete sampling frame, a restricted population, or sample selection bias.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

4. Outliers, or unusual values of *X* and *Y,* will have a major influence on *r*.
5. Curvilinear relationships will be underestimated by *r,* unless *X* and/or *Y* are transformed to reflect the nonlinearity.

1.3 BIVARIATE REGRESSION THEORY

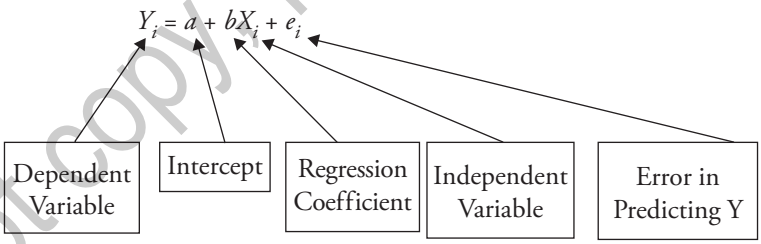
While the correlation is a ***symmetric*** measure of overall linear association between *X* and *Y*, the regression coefficient is an ***asymmetric*** measure of the effect of *X on Y*—that is, how much of a change in *Y* results from a given change in *X*. There is a very important change in language that signals a change in intentions: Here we talk about *X having an effect* on *Y*.

Some believe this language is not appropriate because it seems to say there is a causal connection: *X* causally precedes *Y*. In our approach, we emphasize the fact that the regression model is inherently an asymmetric model and use the term "effect" to signify a causal *claim, not causal proof.* We believe this language is most consistent with the structure of the model and its intentions. Here, association is not enough because it implies simply a relationship, without regard to the direction of causation.

Consider again the scatterplot in Figure 1.1 for the relationship between the mobility rate (*X*) and the divorce rate (*Y*). The question now is, what is the line drawn through these points that maximizes our ability to predict *Y* from *X*? That is, how much do we know about the regional rate of divorce (*Y*) as a result of knowing the mobility rate (*X*)?

1.3.1 The Regression Model

The bivariate regression model expresses each observation's score on *Y* as a function of a combination of components:



The equation says the individual score on *Y* is composed of:

- A constant *a*, called the intercept, defined as the point where the regression line crosses the *Y* axis and thus the predicted value of *Y* when $X = 0$
- The observation's score on *X* times *b*, the regression coefficient, defined as ***the number of units of change in*** **Y** ***resulting from a one unit increase in*** **X**
- An error term (e_i), reflecting the degree to which scores of *Y* are not predicted by *X*.

For any case, the predicted value of *Y*, $\hat{Y}$, is: $\hat{Y}_i = a + bX_i$

and the prediction error is: $Y_i - \hat{Y}_i = e_i$

In other words, this is the distance on *Y* from the predicted point on the regression line to the actual Y_i score at a given level of *X*. $\hat{Y}$ is the predicted score for anyone with the same *X* value.

Do not copy, post, or distribute

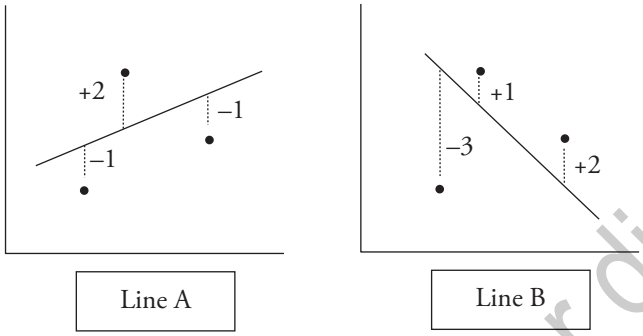
Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

1.3.2 The Least Squares Criterion

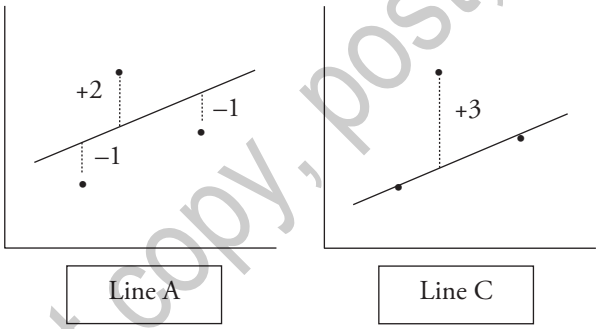
Perhaps the most important element of the regression framework is the development of a criterion telling us where to draw the line showing how we can maximize the prediction of *Y* from *X* and thereby minimize error in prediction. Again we develop this concept in stages (using a classic example in Wonnacott and Wonnacott, 1979):

Version 1. Minimize $\sum(Y_i - \hat{Y}_i)$

Problem: + and - errors cancel out; so this doesn't work as a criterion. Note below that this sum is zero in both cases, but line B is a worse fit to the "sense" in the scatter of points.



Version 2. Minimize $\sum\left|Y_i - \hat{Y}_i\right|$ (the absolute value of the error)



Note here that Line C is a better fit by this criterion but a less reasonable line because it ignores one data point while maximizing the fit of the other two.

Version 3. Minimize $\sum(Y_i - \hat{Y}_i)^2$ (**The Least Squares Criterion**)

This is called the "sum of squared errors," or "sum of squares error." Note that it solves the canceling problem with Version 1 and takes all data into account by definition, unlike Version 2.

1.3.2.1 Deriving the Value of *b* That Minimizes the Sum of Squares Error

We want to find a value of *b* that results in the least error in the prediction of *Y*. This is a complicated problem, but it can be derived mathematically.

Begin with the fitted model: $\hat{Y}_i = a + bX_i$

First, we need to express the intercept, *a*, in terms of the other components of the model. This will show us that the intercept depends on *b*, as well the means of *X* and *Y*.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

First, from the complete regression model:

$$Y_i = a + bX_i + e_i$$

Divide by N: $$\frac{Y_i}{N} = \frac{a}{N} + \frac{bX_i}{N} + \frac{e_i}{N}$$

Sum: $$\frac{\sum Y_i}{N} = \frac{\sum a}{N} + \frac{b\sum X_i}{N} + \frac{\sum e_i}{N}$$

The sum of error is zero, by assumption, the sum of constants is N times the constant, and the sum of a constant times a variable is the constant times the sum of the variable, so . . .

$$\bar{Y} = \frac{N \cdot a}{N} + b\bar{X}$$
$$\bar{Y} = a + b\bar{X}$$
$$a = \bar{Y} - b\bar{X}$$

This shows the intercept is the mean of Y minus b times the mean of X. It is also the ***bivariate formula for the intercept.***

Now substitute for a in the prediction equation:

$$\hat{Y}_i = (\bar{Y} - b\bar{X}) + bX_i$$

Collecting terms: $$\hat{Y}_i = \bar{Y} + b(X_i - \bar{X})$$

Now expand the least squares criterion, substituting for $\hat{Y}_i$ with the preceding equation:

$$\sum\left(Y_i - \hat{Y}\right)^2 = \sum(Y_i - (\bar{Y} + b(X_i - \bar{X})))^2$$
$$= \sum((Y_i - \bar{Y}) - b(X_i - \bar{X}))^2$$

We can expand this expression using the rule for expansion of a difference squared: $(a-b)^2 = a^2 - 2ab + b^2$, applied to the equation above turns into

$$(a \quad - \quad b)^2 \quad = \quad a^2 \quad - (2 \quad a \quad \cdot \quad b) \quad + \quad b^2$$

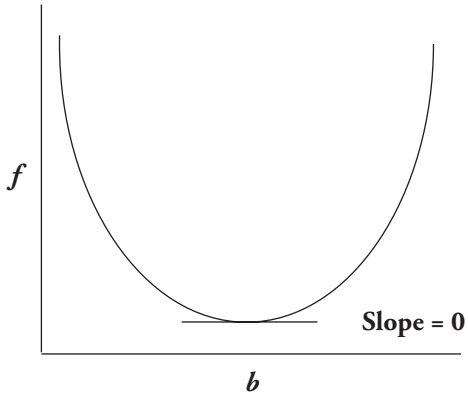
$$\sum((Y_i - \bar{Y}) - b(X_i - \bar{X}))^2 = \sum((Y_i - \bar{Y})^2 - 2b(X_i - \bar{X}) \cdot (Y_i - \bar{Y}) + b^2(X_i - \bar{X})^2)$$
$$= \sum(Y_i - \bar{Y})^2 - 2b\sum(X_i - \bar{X}) \cdot (Y_i - \bar{Y}) + b^2 \sum(X_i - \bar{X})^2$$

This can be arranged to show it is a quadratic function of the form $a + bx + cx^2$, but you have to read the equation realizing that the b are the variables in the equation above, and the constants a, b, and c from the quadratic formula are given constants involving the sum of squared deviations in Y ($\sum(Y_i - \bar{Y})^2 = a$), twice the sum of the cross-products ($2\sum(X_i - \bar{X}) \cdot (Y_i - \bar{Y}) = b$), and the sum of squared deviations in X ($\sum(X_i - \bar{X})^2 = c$).

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

FIGURE 1.2 ⬢ THE FIRST DERIVATIVE AS A FUNCTION OF VALUES OF *b*



Using this, we rearrange the right-hand side of the equation as follows:

$$\sum\left(Y_i - \hat{Y}\right)^2 = \sum(Y_i - \bar{Y})^2 - 2\sum(X_i - \bar{X}) \cdot (Y_i - \bar{Y}) \cdot b + \sum(X_i - \bar{X})^2 \cdot b^2$$

Figure 1.2 shows values of this function as we change values of *b*. To find the formula that always ensures minimizing the error in predicting *Y*, we find the *first derivative*, representing the rate of change in *Y* resulting from the smallest possible change in *b* on the right side of the equation. This rate of change is the slope of the tangent to the curve at each value of *b*. We want to take as our value of *b* the value that produces the minimum value of this function. That minimum is ensured if we choose the value that produces a first derivative of 0.

The derivative is (using rules discussed in a later chapter)

$$\frac{d(f)}{d(b)} = 0 - 2\sum(X_i - \bar{X}) \cdot (Y_i - \bar{Y}) + 2\sum(X_i - \bar{X})^2 \cdot b$$

As noted above, we want to evaluate this expression when the derivative, the slope of the function, is zero. When we do this and solve for *b*, we derive this expression:

$$-2\sum(X_i - \bar{X}) \cdot (Y_i - \bar{Y}) + 2\sum(X_i - \bar{X})^2 \cdot b = 0$$

$$2\sum(X_i - \bar{X})^2 \cdot b = 2\sum(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$$

$$b = \frac{2\sum(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{2\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

This is in fact the formula for the bivariate regression coefficient *b*.

This can written shorthand as $b = \frac{s_{xy}}{s_x^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)}$

which is produced by dividing both the numerator and denominator of the derived least squares formula by $N - 1$. This formula shows that changes in *Y* associated with changes in *X* (the numerator) are measured in terms of the size of units of *X* (captured by the variance of *X*).

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

You can compare this formula to the previous formula for the correlation. It is very similar in the numerator, but only the variance of X occurs in the denominator here. That reflects two facts about the regression coefficient: (1) It is asymmetric, and (2) it is expressed in units change in Y ***per*** unit change in X.

When we calculate b using this formula, we get

$$b = \frac{58.10/8}{43.111} = .168$$

And the intercept is

$$a = 5.41 - (.168) \cdot 46.89 = -2.47$$

So the prediction equation here is

$$\hat{Y} = -2.47 + .168X$$

Note the intercept is *negative*. How can this be? In this case, easily. This happens because the value $X = 0$ is far beyond the *observed* boundaries for values of X. In other words, a is a "meaningless" number in this case. However, it is necessary for the correct prediction of Y values.

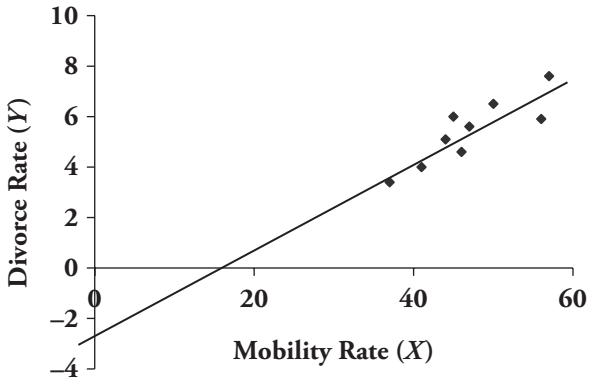
The earlier scatterplot for mobility and divorce is reproduced in Figure 1.3, with the addition of a fitted regression line. In this case, the extension of the fitted line back to the Y axis results in the intercept as a negative number. This is not a real value of Y, but it is the correct baseline for predicting values of Y occurring in its actual range.

1.3.2.2 Interpretation of *b*

The interpretation of b follows from the definition of the first derivative, noting that in this application the smallest possible change in X is 1 unit of X. Two rules of derivatives are used here: One that states the derivative of a constant is zero, and the other that states that the derivative of a constant times a variable equals the constant:

$$\frac{dy}{dx} = \frac{d(a + bX_i)}{dX_i} = \frac{d(a)}{dX_i} + \frac{d(bX_i)}{dX_i} = 0 + b = b$$

FIGURE 1.3 ⬢ THE FITTED REGRESSION LINE



Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

That is, b is the amount of change in Y for a one unit change in X—*always*! Parenthetically, the derivative can be used to figure out the effect of X in nonlinear and non-additive equations as well because it has a general interpretation that applies to all regression equations.

1.3.3 Unstandardized versus Standardized Coefficients

The coefficient in the preceding equation is an *unstandardized (aka metric) regression coefficient*; this means it is expressed in terms of the raw units of the X and Y variables. You can also express this coefficient in *standardized* terms—that is, where you equalize the units of the two variables by expressing the effect of X on Y in terms of the standard deviations of each variable.

Standardizing b involves a simple transformation:

$$\beta = b \cdot \frac{s_x}{s_y} \text{ where } \beta \text{ is the } \textbf{\textit{standardized}} \text{ regression coefficient.}$$

Note what this involves when worked out by substituting for b:

$$\beta = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x}{s_y} = \frac{s_{xy}}{s_x \cdot s_y}, \text{ which is in fact the same as } r_{xy}, \text{ the correlation.}$$

The equivalence between r and β holds ***only in the bivariate case***.

1.3.3.1 Interpretation of β

β = the number of standard deviation unit changes in Y due to a one standard deviation (SD) unit increase in X.

To see how this works in words, suppose

$b = .92$, $s_x = 2$, $s_y = 3$.

Then, A 1 SD unit change in X = 2 raw units of X.
Since b =.92, we know that a 1 raw unit increase in X leads to a .92 raw unit increase in Y.
Therefore, a 1 SD—that is, 2 unit change in X, leads to a 2 × .92 = 1.84 raw unit increase in Y.
Also, a 1 SD unit change in Y = 3 raw units of Y.
So, in terms of SD units of Y, the 1 SD change in X leads to a 1.84/3 = .61 SD change in Y.

$\beta = r = .61$ in this example. In words, a 1 standard deviation increase in X will increase Y by about .6 standard deviations.

1.4 PARTITIONING OF VARIANCE IN BIVARIATE REGRESSION

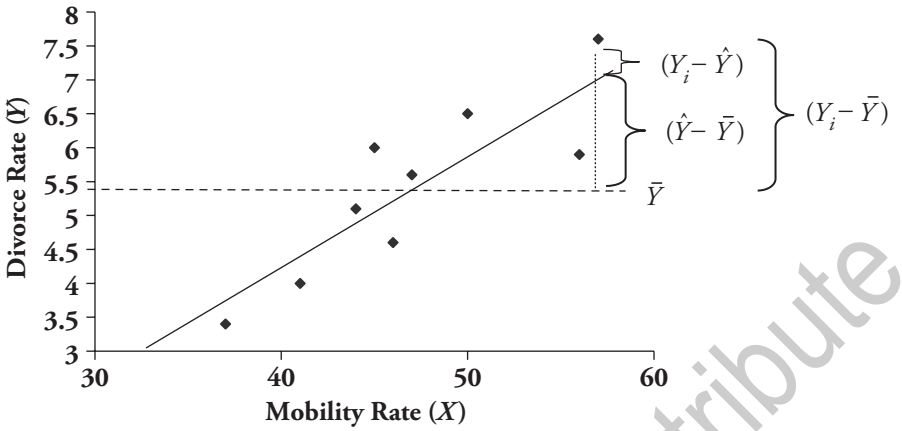
The regression model conceptualizes the explanation of individual Y values as partial, allowing for true indeterminacy. The baseline for comparison in evaluating a regression is no explanation at all, sometimes referred to as the "null model." In this case, our "best guess" about each person's Y score is the mean of Y, since that would minimize the error overall.

Partitioning of variance refers to the division of the overall variance on Y into two parts: (a) explained by X (the regression) and (b) error. The regression line in Figure 1.4 helps conceptualize how this partitioning works.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

FIGURE 1.4 ⬢ PARTITIONING OF THE TOTAL VARIANCE IN *Y*



The total variability of individual Y values around the mean of Y can be expressed as

$$\Sigma(Y_i - \bar{Y})^2 \quad \textit{the total sum of squares.}$$

This total sum of squares can be partitioned into two components:

(1) The sum of squared deviations of predicted Y values $(\hat{Y})$ around the mean $(\bar{Y})$, telling us how much the regression line helps in accounting for individual Y values:

$$\Sigma(\hat{Y} - \bar{Y})^2 \quad \textit{the sum of squares regression}$$

and (2) the sum of squared deviations of actual Y values around the regression line, telling us the degree to which the regression line is not predicting individual Y values:

$$\Sigma(Y_i - \hat{Y})^2 \quad \textit{the sum of squares error.}$$

So

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y})^2 + \Sigma(\hat{Y} - \bar{Y})^2 \tag{1}$$

SS total = SS error + SS regression.

Using the proof in the box on the next page, this equation can be shown to be equal to this:

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y})^2 + \beta^2 \cdot \Sigma(Y_i - \bar{Y})^2 \tag{2}$$

Divide by $\Sigma(Y_i - \bar{Y})^2$ to show this result as proportions adding to 1, where 1 is the total variance:

$$\frac{\Sigma(Y_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2} = \frac{\Sigma(Y_i - \hat{Y})^2}{\Sigma(Y_i - \bar{Y})^2} + \beta^2 \cdot \frac{\Sigma(Y_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

$$1 = \frac{\Sigma(Y_i - \hat{Y})^2}{\Sigma(Y_i - \bar{Y})^2} + \beta^2$$

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

In words:

1 = proportion of total variance due to error + proportion of variance due to regression.

This shows that β^2, which is also R^2, is the proportion of explained variance, which is also

$$\beta^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

This implies that $1 - \beta^2$ is the proportion of error variance.

Proof

Looking at the last term in equation (1), we can show that:

$$\sum(\hat{Y} - \bar{Y})^2 = b^2 \sum(X_i - \bar{X})^2$$

$$\hat{Y}_i = (\bar{Y} - b\bar{X}) + bX_i \text{ (substituting for } a)$$

$$\hat{Y}_i - \bar{Y} = -b\bar{X} + bX_i$$

$$\hat{Y}_i - \bar{Y} = b(X_i - \bar{X})$$

$$\sum(\hat{Y} - \bar{Y})^2 = b^2 \sum(X_i - \bar{X})^2$$

Substituting the result into (1):

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y})^2 + b^2 \sum(X_i - \bar{X})^2$$

Substituting for b^2 using: $b = \beta \frac{s_y}{s_x}$:

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y})^2 + \left(\beta^2 \frac{s_y^2}{s_x^2}\right) \sum(X_i - \bar{X})^2$$

Substituting for the variance of X and Y, and inverting the divisor:

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y})^2 +$$

$$\beta^2 \cdot \frac{\sum(Y_i - \bar{Y})^2}{N-1} \cdot \frac{N-1}{\sum(X_i - \bar{X})^2} \cdot \sum(X_i - \bar{X})^2$$

Canceling results in equation (2):

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y})^2 + \beta^2 \cdot \sum(Y_i - \bar{Y})^2$$

1.5 BIVARIATE REGRESSION EXAMPLE

We can run a bivariate regression example using PROC REG in SAS, a general procedure for running bivariate and multiple regressions. This program also allows you to estimate descriptive statistics and correlations as well as test specific hypotheses about the variables in the model.

In this example, we consider the impact of education on job income, separately for each parent in the Toronto Study of Intact Families. This is a study of 888 husband–wife families in Toronto, with at least one child aged 9 to 16.These are two separate bivariate regressions, but later, we specify the issue differently, by using the concept of an *interaction,* which would allow us to directly estimate the differential impact of education on income by gender of the parent in the same model.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The first results are for the mother. In Table 1.4, we show all of the requested output from PROC REG in SAS, because the different elements of the output are common to many regression programs.

TABLE 1.4 ⬢ BIVARIATE REGRESSION OUTPUT IN SAS

Number of Observations Read	888
Number of Observations Used	607
Number of Observations with Missing Values	281

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation	Label
Intercept	607.00000	1.00000	607.00000	0	0	Intercept
momeduc	8764.00000	14.43822	131290	7.84395	2.80071	mother's education in years
mjobinc	20145	33.18758	897499	377.78793	19.43677	mother's job income

Correlation

Variable	Label	momeduc	mjobinc
momeduc	mother's education in years	1.0000	0.3655
mjobinc	mother's job income	0.3655	1.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > F
Model	1	30592	30592	93.31	<.0001
Error	605	198348	327.84744		
Corrected Total	606	228939			

Root MSE	18.10656	R-Square	0.1336
Dependent Mean	33.18758	Adj R-Sq	0.1322
Coeff Var	54.55823		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	−3.44034	3.86237	−0.89	0.3734	0
momeduc	mother's education in years	1	2.53687	0.26262	9.66	<.0001	0.36555

Do not copy, post, or distribute

1.5.0.1 Observations Used in the Analysis

At the top of the output, you can see that 607 of the 888 observations are used in the analysis. This is because of missing values, most of which occur for job income because some mothers don't work outside of the home.

1.5.0.2 Descriptive Statistics

The descriptive statistics show that the mean level of education in this sample of mothers is 14.44 years, and the average income is $33,187.58. The value shown is income in thousands because that is

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

the way the variable was coded. So to get to dollars, you multiple the value by 1000. This is a pretty low level of income, but there is likely to be a significant sub-sample of part-time workers here.

1.5.0.3 Correlation

Note the correlation between the variables in the analysis is .3655—moderately high, as one would expect.

1.5.0.4 Partitioning of Variance

The "Analysis of Variance" table that partitions the variance shows the sum of squares regression (explained by the model), the sum of squares error, and the "corrected" total sum of squares. This sum is corrected for the 1 degree of freedom due to the one independent variable in the model. If you divide the model by the total sum of squares, you get

$$\frac{SS_{model}}{SS_{total}} = \frac{30592}{228939} = .1336$$

Note that this is the R^2 of the model, printed below in the next table in the output. This means that mother's education explains about 13% of the total variance in mother's job incomes in this sample. Considering this is only one variable, this is not a small amount.

1.5.0.5 Regression Results

The results of the bivariate regression equation are shown in the Parameter Estimates table. The intercept is -3.44, and the regression coefficient b is 2.537. This coefficient can be interpreted this way: Each year of education increases job income by $2,537 on average among these women.

The prediction equation would look like this:

$$\widehat{Mjobinc} = -3.44 + 2.537 \cdot Momeduc$$

Notice here that the intercept value is again negative. This is caused by the fact that there are no real zero values for mother's education in this sample. In fact, the lowest year of education reported is six years. As a result, when the line is extended back to 0, it goes into the negative on the *Y* axis.

The Parameter Estimates table also shows the standardized estimate for the effect of mother's education on her job income. This coefficient is .3655. As discussed above, it should be exactly the same as the correlation in the bivariate case. The interpretation is that a 1 standard deviation increase in mother's education, which we see from the results is 2.8 years, increases job income by .3655 standard deviations. Given that the standard deviation in mother's job income is about 19.44, we could also say that a 2.8-year increase in mother's education leads to, on average, an increase in job income of .3655 x 19.44 = 7.1053, or just over seven thousand dollars.

The same model was run among the husbands, to compare the impact of education on job income among the husbands. The output is shown in Table 1.5.

TABLE 1.5 ⬢ BIVARIATE RESULTS FOR THE HUSBANDS

Number of Observations Read	**888**
Number of Observations Used	**750**
Number of Observations with Missing Values	**138**

(*Continued*)

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

TABLE 1.5 ⬢ (Continued)

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation	Label
Intercept	750.00000	1.00000	750.00000	0	0	Intercept
dadeduc	11402	15.20267	188058	19.64913	4.43273	father's education in years
fjobinc	45217	60.28979	4769362	2727.92760	52.22957	father's job income

Correlation

Variable	Label	dadeduc	fjobinc
dadeduc	father's education in years	1.0000	0.1796
fjobinc	father's job income	0.1796	1.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > F
Model	1	65876	65876	24.92	<.0001
Error	748	1977342	2643.50542		
Corrected Total	749	2043218			

Root MSE	51.41503	R-Square	0.0322
Dependent Mean	60.28979	Adj R-Sq	0.0309
Coeff Var	85.27983		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	28.12580	6.71109	4.19	<.0001	0
dadeduc	father's education in years	1	2.11568	0.42382	4.99	<.0001	0.17956

There are some notable differences in these results. The impact of education among the husbands on their income appears to be ***smaller***, not larger, as you might expect. The b is 2.1157, compared to the 2.537 among the wives. This means that for every year of education the husbands' income increases on average by $2,116. Even though this is smaller, it is also misleading. Notice the differences in the average incomes of husbands and wives here. The husband's average income is $60,289, almost twice the income of the wives. Because of the fact that the husbands are starting at a higher level on average or have been working longer in the labor force, the impact of their education does not "need" to be as high. If you already have an advantage, the meritocratic impact of education may be—ironically—weaker. Following from this, note also that the R^2 and the correlation are also both much lower.

Of course, this example raises many questions, and there are many possibilities to consider in interpreting this relationship. That is why we use multiple regression: both to control for, that is, take into account, alternative explanations of the difference here and to study the role of mediating variables (intervening variables between education and income), which help to explain the effect of education.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

1.5.1 Bivariate Regression Example in STATA

In Table 1.6, we present comparable results in STATA for our first regression model, predicting mother's job income from her years of education. The following model uses the *reg* procedure in STATA (note: all commands in STATA must be stated in lower case).

The first part of the output presents the Analysis of Variance table. Next to this is an overview of the model fit. Again, we see an R^2 statistic of .1322, duplicating the result in SAS.

The latter part of the output presents the relevant unstandardized coefficient (*b*, in the column labeled "Coef"), followed by its associated standard error (2.63). Dividing these two numbers, we get the *t*-statistic (9.66) and its noted significance ($P > |t|$, .000). The *beta* option in STATA produces the standardized regression coefficient. Note that one key difference in the STATA regression output relative to SAS is the placement and labeling of the *y*-intercept. You will find this value in the last row of the regression output ("_cons") rather than the first.

TABLE 1.6 ⬢ BIVARIATE REGRESSION OUTPUT IN STATA

```
      Source |       SS       df       MS              Number of obs =     607
-------------+------------------------------           F(  1, 606)   =   93.31
       Model |      30592        1      30592           Prob > F      =  0.0000
    Residual |     198348      605   327.847443         R-squared     =  0.1336
-------------+------------------------------           Adj R-squared =  0.1322
       Total |     228939      606   377.787129         Root MSE      = 18.1066

------------------------------------------------------------------------------
      mjobinc |      Coef.   Std. Err.      t    P>|t|       Beta
--------------+---------------------------------------------------------------
      momeduc |    2.53687    .26262      9.66   0.000     .3655542
        _cons |   -3.44034   3.86237      -.89   0.373            .
------------------------------------------------------------------------------
```

1.6 ASSUMPTIONS OF THE REGRESSION MODEL

There are a number of assumptions involved in the regression model, although some of them are not the final word on what is possible—because modifications of the model often solve the problem. We state the main assumptions here: A few of these are crucial:

- *No autocorrelation of errors:* $cov(e_i e_j) = 0$

  Y observations are independent of each other—that is, they do not have common systematic components. In other words, this means that different errors are uncorrelated. This assumption is sometimes violated when the same observations are followed through time or when sampling is clustered. Usually, however, observations are sampled independently.

- *Homoscedasticity:* $Var(e_i / X) = E(e_i^2) = \sigma^2$

  In words, the variance of the errors is the same at all levels of *X* and does not depend on the level on *X*. When this assumption is violated, modifications of the regression can be used to address the problem, resulting in *weighted least squares*.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

- *Independence of independent variables and errors*: $E(Xe_i) = cov(Xe_i) = 0$

  Independent variables are uncorrelated with factors in the "true" error term. If this assumption is violated, the regression model is basically misspecified. This assumption amounts to saying either that all relevant explanatory variables for *Y* are included in the model so that what is left involves only random factors or that the excluded factors are uncorrelated with *X*.

- *Linearity*

  The relationship between *X* and *Y* is linear in the population. This assumption can be modified because there are many transformations of nonlinear relations which can be "fit" into a linear model.

- *No measurement error in X*

  The independent variable is measured without error. There are consequences when there is significant measurement error, but the consequences to estimates are minimal when measurement error is minimal to modest.

- *Normality of errors*

  This is a very important assumption because it is so often misunderstood. The assumption of the model is that the ***errors*** are normally distributed around the regression line. ***The assumption is not that* Y *is normal.*** This assumption is not necessary for unbiased estimation of *b*; it is necessary for correct application of significance tests. However, the central limit theorem applies, so that even when *Y* is skewed, suggesting errors may be skewed, an *N* of 100 or more will often result in a sufficiently normal sampling distribution for testing.

1.7 MULTIPLE REGRESSION

What happens when there is more than one independent variable in a regression? What does it mean to "control for" other variables or "partition" their effects or "hold constant" other variables? These are widely used synonymous terms, but one rarely sees a detailed discussion of what exactly is going on when these terms are invoked in an analysis. In general, the intention in considering more than one variable is to derive an estimate of the effect of each variable that is purged of any confounding (overlap) with the effects of all other correlated independent variables. How is this done?

Excluding the individual subscripts for variables to simplify, the general form of the multiple regression equation is

$$X_3 = a + b_{31}X_1 + b_{32}X_2 + e_i$$

Note the changes in the notation used here: The dependent variable does not have to be *Y*, it can be anything. Here X_3 is the dependent variable. It is helpful to distinguish the variables by using differently numbered subscripts, but you could use anything to stand for the variables: letters, acronyms, and variable names are all acceptable. Because there are now multiple independent variables, the regression coefficients also have to be distinguished. It is customary to order the subscripts for the coefficients with the dependent variable number first, then the independent variable number second.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

In standardized form, where the intercept is by definition zero, the equation would look like this, using small *x*'s to stand for the standardized variables.

$$x_3 = \beta_{31}x_1 + \beta_{32}x_2 + e_i$$

The problem in multiple regression is that X_1 and X_2 are usually correlated—that is, confounded. Therefore, the usual way of estimating b in the bivariate case would include a portion of the effect of the other independent variable. To get a better estimate of the effect of each X, the influence of the other variable must be removed.

1.7.1 Covariance Equations

The regression equation above can be used to derive ***covariance equations,*** which in turn can be used to actually solve for the coefficients, as well as form the basis of interpretation of results in causal models (in Chapter 6).

To develop covariance equations to solve the coefficients, first replace a by solving for it. This is done using a multiple regression extension of the formula for the intercept developed in the last section on bivariate regression:

$$a = \bar{X}_3 - b_{31}\bar{X}_1 - b_{32}\bar{X}_2$$

Then we have

$$X_3 = (\bar{X}_3 - b_{31}\bar{X}_1 - b_{32}\bar{X}_2) + b_{31}X_1 + b_{32}X_2 + e$$

Rearranging and factoring out the coefficients leads to an equation where the variables are expressed in deviation form:

$$X_3 - \bar{X}_3 = -b_{31}\bar{X}_1 - b_{32}\bar{X}_2 + b_{31}X_1 + b_{32}X_2 + e$$
$$X_3 - \bar{X}_3 = b_{31}(X_1 - \bar{X}_1) + b_{32}(X_2 - \bar{X}_2) + e$$

Now multiply each side of this equation by each independent variable in deviation form in turn:

$$(X_1 - \bar{X}_1)(X_3 - \bar{X}_3) = b_{31}(X_1 - \bar{X}_1)(X_1 - \bar{X}_1) + b_{32}(X_1 - \bar{X}_1)(X_2 - \bar{X}_2) + (X_1 - \bar{X}_1)e$$

$$(X_2 - \bar{X}_2)(X_3 - \bar{X}_3) = b_{31}(X_1 - \bar{X}_1)(X_2 - \bar{X}_2) + b_{32}(X_2 - \bar{X}_2)(X_2 - \bar{X}_2) + (X_2 - \bar{X}_2)e$$

Then sum and divide both sides of both equations by N - 1:

$$\sum \frac{(X_1 - \bar{X}_1)(X_3 - \bar{X}_3)}{N-1} = b_{31}\sum \frac{(X_1 - \bar{X}_1)(X_1 - \bar{X}_1)}{N-1} + b_{32}\sum \frac{(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{N-1} + \sum \frac{(X_1 - \bar{X}_1)e}{N-1}$$

$$\sum \frac{(X_2 - \bar{X}_2)(X_3 - \bar{X}_3)}{N-1} = b_{31}\sum \frac{(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{N-1} + b_{32}\sum \frac{(X_2 - \bar{X}_2)(X_2 - \bar{X}_2)}{N-1} + \sum \frac{(X_2 - \bar{X}_2)e}{N-1}$$

What do we have? Note the form of the term on the left-hand side of the equality—*it is the formula for the covariance between each independent variable and the dependent variable.* The covariance between the two independent variables also occurs on the right, as well as the variances of

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

X_1 and X_2, resulting from terms where the deviation score on an independent variable is multiplied by itself, and therefore squared.

Given that *X*s do not covary with *e* by assumption, this produces the following covariance equations, using more efficient notation with *"s"* for the covariances and variances:

$$s_{13} = b_{31} \cdot s^2_{X_1} + b_{32} \cdot s_{12}$$
$$s_{23} = b_{31} \cdot s_{12} + b_{32} \cdot s^2_{X_2}$$

Note that this is two equations in two unknowns. This means we could use these equations to solve for the two coefficients.

If you look at the covariance equations, you can see the difference between bivariate and multiple regression.

The bivariate coefficient for X_1 is

$$b_{31} = \frac{s_{13}}{s^2_{X_1}}$$

Using the first covariance equation and solving for b_{31}, the effect of X_1 in the multiple regression is now

$$s_{13} = b_{31} \cdot s^2_{X_1} + b_{32} \cdot s_{12}$$
$$s_{13} - b_{32} \cdot s_{12} = b_{31} \cdot s^2_{X_1}$$
$$b_{31} = \frac{s_{13} - b_{32} s_{12}}{s^2_{X_1}}$$

In effect, the covariance between X_1 and X_3 needs to have the portion removed that is due to the covariance between the two independent variables X_1 and X_2 and the fact that X_2 also has an effect on the dependent variable. This amount is $b_{32}s_{12}$, which is the covariance between X_1 and X_2 times the effect of X_2 on X_3.

To provide an example, we have the following variables from the National Survey of Families and Households (NSFH), a widely used national longitudinal study of households in the United States, with three waves. Here we focus on a sample of married people at Wave 2 followed through to Wave 3:

X_3 = depression score at Wave 3 of the NSFH (Dep_3 below).

X_1 = a dummy variable for getting a divorce between Waves 2 and 3. A dummy variable (explained later) is just a 1/0 variable comparing two groups. Here it equals 1 if the person got divorced and 0 if they stayed married (Div_{23} below).

X_2 = depression score at Wave 2 of the NSFH (Dep_2 below).

We rewrite the regression equation specifically to show the variables involved:

$$Dep3_3 = a + b_{31} Div23_1 + b_{32} Dep2_2 + e$$

The subscript here gives the variable a number for reference in what follows. You could use any statistical software to derive the covariances and variances of these variables. Here we used PROC CORR in SAS to get the covariances, and PROC UNIVARIATE to get basic descriptive statistics, including the standard deviation. The results of doing this are shown in Table 1.7.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

TABLE 1.7 ⬢ BASIC DESCRIPTIVE STATISTICS FOR THE DIVORCE MODEL

Covariances	Standard Deviations	Means
$s_{13} = .085$	$s_1 = .099$	$\bar{X}_1 = .01$
$s_{12} = .096$	$s_2 = 15.43$	$\bar{X}_2 = 13.23$
$s_{23} = 102.57$	$s_3 = 15.19$	$\bar{X}_3 = 13.06$

We ran a bivariate regression of depression at Wave 3 on getting a divorce between Waves 2 and 3, to establish a baseline. That value was

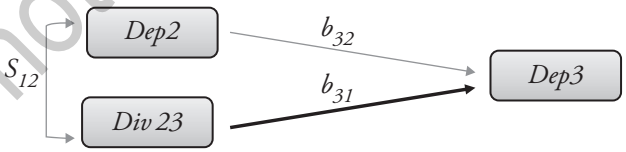
$$b_{31} = \frac{s_{13}}{s^2_{X_1}} = 8.67$$

However, in the multiple regression controlling for prior depression at Wave 2, the effect of *Dep2* (X_2) on *Dep3* (X_3) and its overlap (covariance) with *Div23* (X_1) is removed first—that is, it is "controlled." This is important because the effect of *Dep2* on *Dep3* represents the continuity in depression across waves—that is, the lack of change. Thus, in this equation, the effect of *Div23* is the effect of divorce on change in depression across waves. Also, note, previous depression could lead to a higher risk of divorce, so the causal direction could be wrong unless it is controlled here.

Here the effect of X_1 on X_3 is: $b_{31} = \frac{s_{13} - b_{32}s_{12}}{s^2_{X_1}}$

The amount to be removed from the effect of *Div23* is shown in the graphic in Figure 1.5 in paths with grey arrows, and the net effect of *Div23* after removing the overlap with *Dep2* is the Black arrow.

FIGURE 1.5 ⬢ A MODEL FOR A TWO VARIABLE MULTIPLE REGRESSION



1.7.1.1 Solving the Equation

Using the given information about the variances and covariances, it is possible to solve for each unknown, because we then have two covariance equations in two unknowns. For example, we can solve for b_{31} (though we do not show the derivation of this formula):

$$b_{31} = \frac{s_{13} \cdot s^2_{X_2} - s_{12} \cdot s_{23}}{s^2_{X_1} s^2_{X_2} - s^2_{12}}$$

$$b_{31} = 4.83$$

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Substituting this value in the covariance equation for s_{23} leads to

$$b_{32} = .432$$

Solving for *a,*

$$a = \bar{X}_3 - b_{31}\bar{X}_1 - b_{32}\bar{X}_2 = 13.06 - (4.83 \cdot .01) - (.432 \cdot 13.23) = 7.23$$

The overall prediction equation i**s**

$$\widehat{Dep3}_3 = 7.23 + 4.83 Div23_1 + .432 Dep2_2$$

Here we use the concept of "symptom days" to interpret results. That's because the depression questions asked how many days a week each symptom occurred. Because there are 12 symptoms, there are potentially 7 x 12 = 84 symptom days a week that could occur.

1.7.1.2 The Equation Interpreted

1. A divorce between Waves 2 and 3 leads to 4.83 more depression symptom days per week at Wave 3, over and above depression at Wave 2.
2. Each symptom day of depression at Wave 2 leads to .432 symptoms days of depression at Wave 3.
3. Note, importantly, that the effect of divorce is 4.83 / 8.67 = 56%, just over half, of its original size. This reflects the confounding and suggests that prior depression *is* also related to the risk of divorce over time.

You can calculate exactly how much the effect of divorce has been reduced by the presence of prior depression in the equation, as follows:

$$\frac{b_1 - b_2}{b_1} = \frac{8.67 - 4.83}{8.67} = .44$$

That is, controlling for prior depression explains about 44% of the original association.

The results of the regression from the SAS output are shown in Table. 1.8.

TABLE 1.8 ⬢ RESULTS FOR THE MULTIPLE REGRESSION

Number of Observations Read	4600
Number of Observations Used	4247
Number of Observations with Missing Values	353

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	189842	94921	518.58	<.0001
Error	4244	776819	183.03924		
Corrected Total	4246	966660			

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Root MSE	13.52920	R-Square	0.1964
Dependent Mean	12.97078	Adj R-Sq	0.1960
Coeff Var	104.30520		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.22571	0.27373	26.40	<.0001
div23	1	4.83030	2.05421	2.35	0.0187
dep2	1	0.43160	0.01353	31.91	<.0001

You can see in the final table the regression coefficients under "Parameter Estimates." Note there is also a *t* test for each coefficient and a significance level. This significance level is "two-tailed," so if you hypothesize a direction to the effect, as for divorce, the probability here could be "one-tailed," which means you divide the printed probability by two (.0187/2 = .0094).

1.7.2 Tests for Multiple Regression

There are three basic tests used in multiple regression:

- ***Significance of the whole equation***: The alternative hypothesis is that at least one of the independent variables has a significant effect on the dependent variable, against the null hypothesis that *no* independent variable has a significant effect.

$$F = \frac{R^2 / k}{\left(1 - R^2\right) / (N - k - 1)} \text{ with } k \text{ and } (N - k - 1) \text{ df}$$

Where k = number of independent variables in the equation and N = sample size.

- ***Individual variables in the equation*:** Tests of individual variables in the equation amount to a 1 degree of freedom test of the difference in the R^2 with the variable in the model versus the R^2 when the variable is not in the model.

The null hypothesis is that X has no effect on Y.

$$F = \frac{(R_k^2 - R_{(k-1)}^2) / 1}{(1 - R_k^2) / (N - k - 1)} \text{ with 1 and } (N - k - 1) \text{ df}$$

Where R_k^2 = the R^2 with all variables in the equation.

$R_{(k-1)}^2$ = the R^2 with the variable to be tested removed from the equation.

This test is equivalent to the test for* b *printed by most programs.

- ***Group of variables added to an equation:*** This test compares the R^2 in a model with variables added to a baseline model to the R^2 of the baseline model, to test collectively for the significance of the effect of the group of variables added.

This test is a comparison of nested models, in which Model A (the smaller model) is nested in Model B (the larger model), and thus all variables in A are contained in B, but B has additional variables.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The null hypothesis is that none of the new variables added has a significant effect on *Y*.

$$F = \frac{(R_B^2 - R_A^2)/(k_B - k_A)}{(1 - R_B^2)/(N - k_B - 1)} \qquad \text{with } k_B - k_A \text{ and } N - k_B - 1 \text{ df}$$

Where R_B^2 = the R^2 from the larger model (B),

R_A^2 = the R^2 from the smaller model (A),

and k_B and k_A are the number of independent variables in B and A respectively.

1.7.3 Nested Models

One model (A) is nested in another more complex model (B) when A occurs completely as a subset of B, and B has additional variables. You could, for example, add a group of variables to the depression equation to test a single idea or hypothesis. One version of this occurs when you want to study differences across groups. Another occurs when you want to test interactions (discussed in the next chapter).

Suppose you are concerned that the effect of divorce in the previous example is confounded with ethno-racial differences in rates of divorce, and these group differences represent basic differences in status that are reflected in differences in depression. If we added groups to this equation representing race/ethnicity, we could control for this possibility as an alternative explanation.

The equations being compared here are

Model 1: $Dep3_3 = a + b_{31}Div23_1 + b_{32}Dep2_2 + e_i$

Model 2: $Dep3_3 = a + b_{31}Div23_1 + b_{32}Dep2_2 + b_{33}Black + b_{34}Hispanic + b_{35}Asian + e_i$

Note that Model 1 is contained in Model 2 and thus is "nested" in Model 2. We can therefore compare the R^2 across these models. Here we added three groups—Black, Hispanic, and Asian—from the NSFH data to the equation for the effect of divorce, with the reference comparison group non-Hispanic Whites. Each group variable is coded 1 versus 0, and the reference group is the group left out of this coding. This makes that group the baseline for comparison. The relevant output is reproduced in Table 1.9.

TABLE 1.9 ● ADDING GROUP DIFFERENCES TO THE MULTIPLE REGRESSION

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	***F* Value**	**Pr > F**
Model	5	196220	39244	216.02	<.0001
Error	4241	770440	181.66465		
Corrected Total	4246	966660			

Root MSE	13.47830	R-Square	0.2030
Dependent Mean	12.97078	Adj R-Sq	0.2020
Coeff Var	103.91281		

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	1	6.79018	0.28282	24.01	<.0001
div23	1	4.67567	2.04693	2.28	0.0224
dep2	1	0.42115	0.01360	30.98	<.0001
black	1	3.38370	0.60390	5.60	<.0001
hispanic	1	2.48630	1.12877	2.20	0.0277
asian	1	2.83244	2.81983	1.00	0.3152

We can see in these results that the effect of divorce is still significant and therefore independent of group differences. The net effect of divorce here (4.67) is very close to the prior model, suggesting that ethno-racial differences in divorce do *not* account for the effect of divorce in this model.

If we want to test for the effect of "race/ethnicity" here, we would need to compare the R^2 from this model to the previous model for divorce and prior depression only. This comparison isolates the effect of "race/ethnicity" and asks whether the independent partialled effect of race/ethnicity is significant over and above divorce and prior depression.

Substituting the values from the output into the *F*-test,

$$F = \frac{(.2030 - .1964)/(5-2)}{(1-.2030)/(4247-5-1)} = 11.71$$

with 3 and 4241 df. Any *F* table will verify that this is significant beyond the .0001 level. Race/ethnicity does make a difference here, but its effect is independent of the relevance of divorce.

1.8 A MULTIPLE REGRESSION EXAMPLE: THE GENDER PAY GAP

When you hear about the gender pay gap, it is often stated in bivariate terms: It is the overall difference between the average (or median) pay of men and women in the labor force. For example, you may have heard that women make something like 76 cents on the male dollar earned.

The question is how much of that pay gap is due to "natural" or expected differences in pay due to other causes. These other causes include factors that are generally rewarded with higher pay—such as education, experience, and performance—or are due to differences in economic sectors, occupations, or regional economies that have an imbalance of men and women. The concept of "equal pay for equal work" is not as easy to specify as it sounds.

This is admittedly a tricky question and thus also a good example. The entire issue is subtler than it appears, if women choose occupations that provide more "flex time," for example, which generally have lower pay, but feel constrained to choose those occupations. This possibility suggests larger causes of unequal pay are at work, limiting the sense of choice and access women have.

As an example of interpreting multiple regression, we present results from the National Survey of Families and Households, using data from Wave 2 in 1992 to1994. We would expect

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

a substantial pay difference to exist at that point in history. To keep things as straightforward as possible, we make some simplifying assumptions, relative to the large number of alternative debates surrounding this issue.

Our dependent variable here is the wages earned per hour of work reported. This is one prevalent approach in this literature because it already takes into account gender differences in hours worked overall.

The standard approach to this issue adjusts—controls—for differences in human capital, captured by level of education and total years worked in the labor force. There may also be other confounders, if, for example, women are overrepresented in groups that are also at a pay disadvantage, but the cause is not gender per se. This happens in our example in the case of race: 19% of our employed females are Black, but only 14% of the employed males are Black.

Here is the result of a bivariate regression of wages per hour worked on gender, coded here as a two-category variable, with 1 = female, and 0 = male (see Table 1.10). This coding allows us to see the average difference in dollar income per hour directly.

TABLE 1.10 ⬢ BIVARIATE MODEL FOR GENDER PAY GAP

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	15.67842	0.36708	42.71	<.0001
female	1	−4.15118	0.49088	−8.46	<.0001

This result says that women make on average $4.15 less than men per hour worked. The results are not stated proportionally. Because of the way "female" is coded, we know that men make on average $15.68 dollars per hour worked. This is because the intercept is the value of Y when independent variables = 0. We can make the result proportional by expressing the difference in income this way:

$$\textit{female-to-male}\ \text{income ratio} = \frac{15.678 - 4.1511}{15.678} = .73$$

This says that in this equation, women make about 73 cents per dollar earned by a man. Given the year is around 1992 to 1994, this figure—broadly—makes sense.

This result does not control for qualifications, tenure (experience), and performance. The last is difficult to capture in most data, but the first two are represented by level of education and experience in the labor force. When you control for education and the total number of years worked in the labor force, this produces the results shown in Table 1.11.

TABLE 1.11 ⬢ THE GENDER PAY GAP CONTROLLING FOR EDUCATION AND SENIORITY

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	1	−9.28590	1.42357	−6.52	<.0001
female	1	−2.84527	0.48218	−5.90	<.0001
education2	1	1.50659	0.09479	15.89	<.0001
Totalworkyearswgt	1	0.26889	0.02406	11.18	<.0001

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Here we see that each variable affects income controlling for the others. Each year of education (*education2)* increases average hourly income by about $1.51 per hour. The effect of labor force experience is captured by the variable "*totalworkyearswgt*." This variable is the total number of accumulated work years, weighted by whether the job was full-time or part-time. Here each weighted work year increases hourly income by about $.27, or 27 cents. In both cases, results for human capital predictors are very significant and reflect the usually observed influence of these standard predictors of pay.

You can also see here that the net effect of "female" is now -2.8453—that is, the net difference with men is now $2.85 per hour less than men. This net difference with males is considerably smaller. You cannot figure out the net female to male income ratio here, however, because now the intercept has a different interpretation. It is in fact still the predicted value of *Y* when all independent variables = 0, but now those variables include education and total years worked. This means the intercept is the predicted hourly income for men with zero years of education and zero years worked—in other words, it is not interpretable as is. The intercept is negative, mainly due to the fact that no one in the actual sample has zero years of education and reports zero years worked.

You can adjust for this problem and make the intercept interpretable again, by "centering" the control variables here. That means subtracting the *male* mean from each raw score, like a deviation score. We subtract the male mean, assuming we want to interpret the male / female difference as if female workers had the same level of education and time in the labor force as male workers.

Male (female = 0) and female (female = 1) means on the control variables are shown in Table 1.12.

TABLE 1.12 ⬢ MALE VERSUS FEMALE MEANS ON CONTROL VARIABLES

female	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	2118	education2	2118	13.5193579	2.6023414	0	20.0000000
		Totalworkyearswgt	2118	17.0931500	10.3754841	0.0833333	53.1666667
1	2687	education2	2687	13.2690733	2.4021010	0	20.0000000
		Totalworkyearswgt	2687	13.6387855	9.3584510	0.0833333	48.2500000

One thing that is very clear from this table is that women work significantly fewer years than men, despite the fact that we observe in other results that they are slightly older than the men in this sample. This is presumably due to more time in home work and childcare roles. However, it is also much less clear that there are any differences in level of education by gender.

When the education and total years worked are centered on the male mean, you get the results show in Table 1.13.

TABLE 1.13 ⬢ RE-CENTERING THE VARIABLES FOR COMPARISON TO THE BIVARIATE DIFFERENCE

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	1	15.67842	0.35447	44.23	<.0001
female	1	−2.84527	0.48218	−5.90	<.0001
educcenter	1	1.50659	0.09479	15.89	<.0001
totalyearscenter	1	0.26889	0.02406	11.18	<.0001

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Note here that the intercept is essentially equal to the original intercept, and the coefficients for each variable are exactly the same—as expected. Using this result, we will compare net differences between women and men relative to that overall male mean income. Now the adjusted female to male ratio is

$$female\text{-}to\text{-}male \text{ income ratio } = \frac{15.678 - 2.8453}{15.678} = .82$$

You can see that the net difference is smaller. Now women make 82 cents per dollar earned by men, *net of other factors representing human capital differences.*

Finally, we add race to this example, because we know that women are overrepresented among Blacks, and race is a distinct source of income discrimination. In other words, we do not want to attribute a pay difference to gender, when it is in fact due to race.

We added two variables to the equation representing race: "Black" is a comparison of Blacks] to non-Hispanic Whites (1 vs. 0), and Hispanic is a comparison of Hispanics to non-Hispanic Whites (also 1 vs. 0).

When we add these variables to the regression, using the centering approach, we get the results shown in Table 1.14.

TABLE 1.14 ⬢ CONTROLLING FOR RACE DIFFERENCES

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	1	15.67842	0.35439	44.24	<.0001
female	1	−2.78185	0.48313	−5.76	<.0001
educcenter	1	1.47579	0.09691	15.23	<.0001
totalyearscenter	1	0.27117	0.02409	11.26	<.0001
blackcenter	1	−1.30546	0.64290	−2.03	0.0424
hispcenter	1	−0.67357	0.99173	−0.68	0.4971

The net difference between women and men is now –$2.78 dollars per hour of work. Computing this as a ratio of female to male income,

$$female\text{-}to\text{-}male \text{ income ratio } = \frac{15.678 - 2.7818}{15.678} = .82$$

The ratio is about the same: Even though Blacks receive a lower per hour income compared to Whites and there is a slight difference in proportion female among Blacks versus Whites, it does not further account for the gender pay ratio.

A caveat: There are other, more comprehensive controls we could use here, discussed widely in this literature. And of course, historical change would suggest we use more recent data as well. But the principle of accounting for the female to male difference has been demonstrated in this example, even considering this restricted set of other factors determining pay.

There is an important complication to consider in this example that will be discussed thoroughly in Chapter 6. The issue is whether standard human capital differences are really controls here, as opposed to mediators that transmit the effect of gender to pay differences—in other words, actual consequences of gender that are part of the overall gender difference in pay, not an alternative explanation.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The last model can be duplicated using the *reg* command in STATA. Again, we present results just to create a cross-walk between SAS and STATA, this time with a multiple regression example (Table 1.15).

TABLE 1.15 ⬢ REGRESSION RESULTS FOR THE GENDER WAGE GAP IN STATA

```
      Source |       SS           df       MS      Number of obs   =     4805
-------------+----------------------------------   F( 5,  4799)    =    86.21
       Model |  114657.249         5  22931.4499   Prob > F        =   0.0000
    Residual |  1276521.15      4799  265.997322   R-squared       =   0.0824
-------------+----------------------------------   Adj R-squared   =   0.0815
       Total |  1391178.4       4804  289.58751    Root MSE        =   16.309

------------------------------------------------------------------------------
  wagesperhr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -2.78185   .4831318    -5.76   0.000    -3.72901    -1.83469
  educcenter |   1.475792   .0969146    15.23   0.000     1.285795    1.665789
totalyearsc~r|   .2711658   .0240869    11.26   0.000     .2239445    .3183871
  blackcenter|  -1.305463   .6429024    -2.03   0.042    -2.565847   -.0450801
   hispcenter|  -.6735712    .991734    -0.68   0.497    -2.617824    1.270682
        _cons |   15.67842   .3543853    44.24   0.000     14.98366    16.37317
------------------------------------------------------------------------------
```

1.9 DUMMY VARIABLES

The variables in the examples in the previous section are not all "continuous" variables that vary from zero to the highest value in a sample. In those examples, divorce is a "dummy variable," equal to 1 if there was a divorce, and equal to 0 if not. "Female" is also a dummy variable, = 1 for females, and 0 for males. This kind of variable (aka an indicator variable) just stands for whether you are in a certain group or not, whether you experienced an event or not, or whether you are in a certain category or not. It is straightforward to interpret results because there are only two groups involved—divorced versus not divorced, women versus men.

A trickier case occurs when you want to assess the differences among a ***set*** of groups, considered as a set of independent variables. In this case, you have to use a set of dummy variables to represent differences among groups. This happens for variables like marital status, race, ethnicity, religion, region, and so forth. This too came up in the previous section when we added race to the model to demonstrate nested models.

There are certain rules in interpreting these variables that you need to be aware of. For one thing, there is a left-out reference group that is the comparison point. For example, to study marital status, you may have five groups in total: married, divorced or separated, widowed, never married, and cohabiting. You choose a reference group for comparison, usually a standard norm, and compare other groups to this group. In this case, you could make married the reference group.

1.9.1 How Do Dummy Variables Work in Regression?

Interpreting *b* is *always* the same in general: The amount of change in *Y* resulting from a 1-unit change in *X*. For dummy variables, that one unit represents two groups, and as a result, because

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

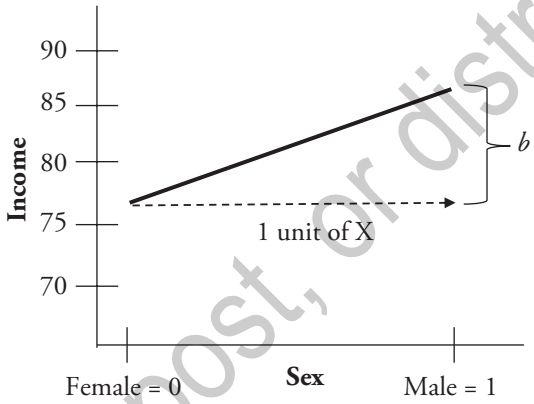
of the "Least Squares Criterion" discussed in the earlier section, the *b* for a dummy variable is *also* the mean difference between the two groups on *Y*.

In this context, *a,* as usual, is the predicted value of *Y* when $X = 0$. Thus it is the mean of *Y* for the group coded 0 on the dummy variable.

1.9.1.1 Example 1: Gender—Two Categories

Imagine you are studying gender differences in starting salaries at universities at the assistant professor level. The graph in Figure 1.6 shows the difference as a regression slope comparing males (= 1) to females (= 0), based on imagined data circa 2010.

FIGURE 1.6 ⬢ DIFFERENCES IN STARTING SALARY OF MALE VERSUS FEMALE ASSISTANT PROFESSORS



Given that there are two values of *X*, the regression line will pass **exactly** through the mean values on *Y* within each of the two groups (the conditional *Y* means).

So b = the difference between the means in the two groups.

a = the intercept, the mean of *Y* in the group coded = 0.

The equation: $\hat{Y} = 77 + 9X$

Interpreted:

When $X = 0$ (women): $\hat{Y} = 77 + 9(0) = 77$

When $X = 1$ (men): $\hat{Y} = 77 + 9(1) = 86$

The regression coefficient, $b = 9$, expresses the difference between the mean incomes in the two groups. So, the mean income for women is $77,000, and the mean income for men is $86,000.

1.9.1.2 Example 2: Divorce

In the previous section, we considered the effect of divorce on depression. Divorce is a dummy variable with two groups, coded in the previous example to be = 1 for divorced and = 0 if still married.

The output from this regression is shown in Table 1.16.
From these results, we could write out the equation as

$$\widehat{Dep3} = 12.97 + 8.57 \cdot Div23$$

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

TABLE 1.16 ⬢ THE BIVARIATE EFFECT OF DIVORCE ON DEPRESSION

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > F
Model	1	3344.10531	3344.10531	14.53	0.0001
Error	4589	1056268	230.17386		
Corrected Total	4590	1059612			

Root MSE	15.17148	R-Square	0.0032
Dependent Mean	13.06394	Adj R-Sq	0.0029
Coeff Var	116.13247		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	1	12.97808	0.22504	57.67	<.0001
div23	1	8.56935	2.24820	3.81	<.0001

This means that those not divorced had a mean level of depression at Wave 3 of 12.97 on this scale and that the divorced had an average level 8.57 points higher than that—12.97 + 8.57 = 21.54. We note the estimate here is slightly different than the by-hand calculation, which is subject to rounding error.

1.9.2 Dummy Variables with Multiple Categories

Many categorical variables have more than two categories: *ethnicity, marital status, religion, region, employment status, family type,* to name just a few examples.

In Table 1.17, assume there are four measured marital statuses. Note in this scheme that you will see variables for divorced or separated, widowed, and never married but not for married. Each variables is coded = 1 to stand for that group and 0 for ***all other groups***. The only group with 0 on all three dummy variables is the married. Each dummy variable can be interpreted as the mean difference on *Y* of that group versus the reference group, in this case, the married. You cannot create a separate dummy variable for the married because they are already uniquely defined in this coding:

TABLE 1.17 ⬢ CREATING DUMMY VARIABLES FOR THE GROUPS WITH MARRIED AS THE REFERENCE

	Variable Names		
Groups	Divsep	Widow	Nevermarr
Divorced / Separated	1	0	0
Widowed	0	1	0
Never Married	0	0	1
Married	0	0	0

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

If you try to create a separate dummy variable for all four groups here, you end up with one variable that is perfectly determined by the scores on three others and thus is perfectly "collinear," a term referring to the fact that the last dummy variable considered is completely determined by the values of the other three and thus is not separable from the other three.

Imagine you did create separate dummy variables for each marital status, as illustrated in Table 1.18.

TABLE 1.18 ⬢ CREATING DUMMY VARIABLES FOR ALL GROUPS

Groups	Variable Names			
	Divsep	Widow	Nevermarr	Married
Divorced / Separated	1	0	0	0
Widowed	0	1	0	0
Never Married	0	0	1	0
Married	0	0	0	1

In this table, ***Married*** = 1 – *Divsep* – *Widow* – *Nevermarr*.

You can do the calculation as follows:

For the divorced/separated, Married = 1 - 1 - 0 - 0 = 0
For the widowed, Married = 1 - 0 - 1 - 0 = 0
For the never married, Married = 1 - 0 - 0 - 1 = 0
For the married, Married = 1 - 0 - 0 - 0 = 1

In other words, the values on the married dummy variable are already defined by the combined information on the first three dummy variables. Basically, *you don't need the fourth dummy variable here.*

The reference group you choose is always left out of the group of dummy variables constructed. This means that each dummy variable will be the mean difference on *Y* between the group defined by that dummy variable and the reference group. If you want to know about mean differences *among* those groups, you can still extract that information from the regression results, since the difference in the coefficients of any two variables is equal to the difference in the means in those groups.

The ultimate lesson is this: You only need *k* - 1 variables to represent differences among *k* groups. You must choose a reference group, often a standard reference representing an extreme or what is expected, and then create dummy variables showing the differences between the other groups and this reference group.

1.9.3 Interpreting Results for Dummy Variables with Multiple Categories

This example uses the National Survey of Families and Households (NSFH) Wave 2 data to assess marital status differences in "close and trusting relations," a general scale derived from the Ryff Well-Being scales (Ryff, 1989). This scale is constructed as a 6-point scale, from 0 to 5, where 0 represents no trusting or close relations and 5 would represent strong agreement that you have close and trusting relationships with others.

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.
Do not copy, post, or distribute

Marital status at Wave 2 also considered whether the respondent was cohabiting. For all nonmarried statuses, this trumped divorce, widowhood, or never married status. Thus, everyone in the dummy variables for divorce / separation, widowhood, and never married is *not* living with a partner.

The results are shown in Table 1.19.

TABLE 1.19 ⬢ MARITAL STATUS DIFFERENCES IN TRUSTING RELATIONSHIPS

Number of Observations Read	4600
Number of Observations Used	4242
Number of Observations with Missing Values	358

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > F
Model	4	179.01967	44.75492	34.62	<.0001
Error	4237	5477.06122	1.29267		
Corrected Total	4241	5656.08089			

Root MSE	1.13696	R-Square	0.0317
Dependent Mean	3.64785	Adj R-Sq	0.0307
Coeff Var	31.16793		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	1	3.79430	0.02179	174.10	<.0001
divsep2	1	−0.50006	0.04591	−10.89	<.0001
widow2	1	−0.26325	0.06239	−4.22	<.0001
nevermar2	1	−0.37252	0.07140	−5.22	<.0001
cohab2	1	0.08167	0.09772	0.84	0.4033

Looking at the results for the dummy variables here, you should first see that the intercept value of 3.79 *is* the mean on *Y* among the married, the reference group. You can also see that the divorced, widowed, and never married all experience lower levels of trusting relationships compared to the married ($p < .0001$), but cohabitors are not significantly different from the married on this outcome. The basic dividing line in the results is having a live-in partner.

You can derive the means in these groups straightforwardly, by calculating each mean as $a + b_k$. For the divorced, for example, the mean is 3.79 + (–.500) = 3.29.

What you don't know from these results is whether the groups differ from each other. This is useful information in interpreting the results, in order to get a sense of an overall pattern and the essential differences among groups.

You can in fact run regression programs in SAS or STATA with specific options or additional statements to test the differences among the groups in the equation. The test you set up is a difference of means between two groups. You need to see first what the mean of each group is made up of and then create a test of a null hypothesis of no difference across groups.

For example, the mean among the divorced/separated in this equation is

$a + b_1$

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The mean among the widowed in this equation is

$$a + b_2$$

So the hypothesis of no difference between these groups is

$$H_o: (a + b_1) - (a + b_2) = 0$$

When you perform the subtraction above, the intercepts cancel, resulting in

$$H_o: b_1 - b_2 = 0$$

So you can set up a difference of means test for pairs of groups by just subtracting the two coefficients. This is possible because each is a difference from the same baseline. An example of how that works is this: If the mean in the reference group is 10, the mean in Group 1 is 15, and the mean in Group 2 is 12, then the difference between Group 1 and the reference is 5, and the difference between Group 2 and the reference is 2. So the difference between has to be 15 – 12 = 5 – 2 = 3. Results for *F*-tests of mean differences for all six pairs of groups here are shown in Table 1.20, using SAS output. This output names the test using a short-form of the comparison—for example, "divvwid" is divorced versus widowed.

TABLE 1.20 ⬢ POST-HOC TESTS FOR DIFFERENCES AMONG GROUPS

1.

Test divvwid Results for Dependent Variable closetrust2				
Source	DF	Mean Square	*F* Value	Pr > F
Numerator	1	14.51279	11.23	0.0008
Denominator	4237	1.29267		

2.

Test divvnmar Results for Dependent Variable closetrust2				
Source	DF	Mean Square	*F* Value	Pr > F
Numerator	1	3.50184	2.71	0.0999
Denominator	4237	1.29267		

3.

Test divvcoh Results for Dependent Variable closetrust2				
Source	DF	Mean Square	*F* Value	Pr > F
Numerator	1	31.76358	1.50	0.2214
Denominator	4237	1.29267		

4.

Test widvnmar Results for Dependent Variable closetrust2				
Source	DF	Mean Square	*F* Value	Pr > F
Numerator	1	1.93369	1.50	0.2214
Denominator	4237	1.29267		

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

5.

Test widvcoh Results for Dependent Variable closetrust2				
Source	**DF**	**Mean Square**	***F* Value**	**Pr > F**
Numerator	1	11.03322	8.54	0.0035
Denominator	4237	1.29267		

6.

Test nmarvcoh Results for Dependent Variable closetrust2				
Source	**DF**	**Mean Square**	***F* Value**	**Pr > F**
Numerator	1	15.76829	12.20	0.0005
Denominator	4237	1.29267		

The first test shows that the divorced (−.50) are significantly lower on trusting relationships compared to the widowed (−.26). However, they are not significantly different from the never married (Test 2). As would be expected given the nonsignificant difference between cohabitors and the married, the divorced are also lower in trusting relations compared to cohabitors (Test 3). The widowed are not different from the never married (Test 4, note that the never married are between the divorced and the widowed in trusting relations) and also lower than cohabitors (Test 5), as are the never married (Test 6).

The picture that emerges is this: Groups that have no live-in partner express lower levels of trusting relationships in their life. But there is a further distinction among those without a partner, specifically when previous relationships have ended naturally or "successfully," as in the death of a spouse, relative to those who have been divorced. This puts the widowed in-between the other nonmarried groups and the partnered groups.

1.9.4 One-Way Analysis of Variance

The one-way analysis of variance refers to a simple test for any group differences in the equation. This is equivalent to the test in multiple regression for "any" effect of at least one independent variable.

The test is the significance of the R^2 in the equation: if the R^2 is significant, this could only occur because there are significant differences among the groups somewhere in the equation.

The *F*-test is structured as follows:

$$F = \frac{R^2 / k}{(1 - R^2) / N - k - 1}$$

testing the null hypothesis for k groups in the equation

$$H_0: b_1 = \ldots = b_k = 0$$

against an alternative that at least one group differs from one other

$$H_A: b_1 \neq \ldots \neq b_k \neq 0$$

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

This test is printed in the earlier regression output in the "Analysis of Variance" table. Here the *F* value is 34.62, with 4 degrees of freedom, significant at the .0001 level. Thus the one-way analysis of variance overall test suggests there are group differences here.

In fact, this test should be conducted first ***before*** you conduct any tests comparing specific pairs of groups. We do not take the time or space here to review the problem of cumulative Type I error, mainly because this is a large topic, also more suited to experimental data. However, we do advocate *only* comparing groups as suggested by specific hypotheses you are testing, instead of all groups.

Concluding Words

This chapter travels a considerable distance in relatively few pages. Consistent with the idea that this book is more for second courses, this chapter is intended as a review, a tune-up, more than an initial introduction. However, we did include important detail at a number of points, which will be useful as reference points for the material in the chapters ahead.

Our discussion started with a formal definition of an association between variables using probability theory. This was illustrated first using a cross-tabulation of two categorical variables. We introduced the correlation as a measure of association, in steps, to demonstrate the structure of a measure of association for more continuous variables.

The regression model was introduced in steps as well, starting with the structure of the bivariate regression model and adding to that model for the rest of the chapter. There are a number of important introductory concepts here: The Least Squares Criterion for where to fit the line through a scatter of points, unstandardized versus standardized coefficients, and the partitioning of variance. *All* of these concepts will be invoked as we move forward.

Multiple regression was introduced, initially, as a way of accounting for confounding—overlap—between different independent variables. In this demonstration, we see what it means to "control for," or " account for," or "adjust for," confounding among variables. Our example using the gender pay gap illustrates a very important feature of control variables that will come up in later chapters: For a control variable to be relevant, it must be related to *both* the focal independent variable (gender) and the outcome (pay). Usually we are trying to test the focal association at issue, so the most threatening control variables are also those whose patterns of association are consistent with the overall focal association. For example, women had fewer total years of work, and fewer years was related to lower pay. As a result, total years working partially accounted for the bivariate gender difference.

Finally, we introduced dummy variables as a special kind of variable in regression, designed to handle variables that designate membership in different groups.

At this point, all we have is the standard linear ordinary least squares regression model. The structure of it is additive—meaning that different variables push or pull *Y* up or down, but independently—and it is linear—meaning that independent variables have the *same* effect across all levels—and it is restricted to outcomes that are quasi- to completely continuous. ***All of these assumptions will be modified in the chapters ahead.*** As we will eventually see, the regression model is quite flexible, allowing a broader range of representation of the predominant ideas and styles of thinking in the social sciences than is possible using only the additive, linear model.

Generalizations of Regression: A Graphical Road Map

Here is a graphical overview of the variations of the basic multiple regression model we will consider in the chapters ahead.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The linear additive model . . . the starting point

$$Y = a + b_1X_1 + b_2X_2 + e$$

Continuous | Linear | Additive | Normal

Including conditional multiplicative effects . . . interactions

$$Y = a + b_1X_1 + b_2X_2 + b_3(X_1 \cdot X_2) + e$$

Not Additive

Including nonlinearity . . . multiple forms possible

$$Y = a + b_1X_1 + b_2X_2 + b_3X_1^2 + e$$

$$Y = a + b_1X_1 + b_2 \ln(X_2) + e$$

Not Linear

Including categorical outcomes . . . logistic

$$\text{ln odds } Y = a + b_1X_1 + b_2X_2 + e$$

Categorical: 1 / 0

Where Odds = $\dfrac{\Pr(Y = 1)}{1 - \Pr(Y = 1)}$

Including nonnormal outcomes and errors (Poisson)

$$\ln Y = a + b_1X_1 + b_2X_2 + e$$

Nonnormal

Do not copy, post, or distribute

Practice Questions

1. Bivariate Regression and Correlation

The output below from SAS shows the descriptive statistics and covariance for two variables from the 2015 Canadian General Social Survey (Statistics Canada, 2017):

- **overallhealth** is an index of overall physical and mental health, varying from 0 to 30.
- **hhincome** is the household's income in thousands of dollars.

Table 1.A shows the means and standard deviations of each variable.

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Table 1.A Means and Standard Deviations

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
overallhealth	22832	15.58368	2.94471	355807	0	30.00000	Overall health: Index of health, mental health, and well-being
hhincome	22832	72.91137	46.10670	1664713	0	150.00000	Household income

Table 1.B shows the variances and covariance of the two variables. The variances are in the diagonal (the cells in Row 1, Column 1, and Row 2, Column 2), and the covariance is the off-diagonal. Notice that the covariance is the same in Cells 1,2 and 2,1, as it should be.

Table 1.B Variances and Covariance

Covariance Matrix, DF = 22831		
	overallhealth	hhincome
overallhealth	8.671341	24.800833
hhincome	24.800833	2125.827490

a. Using the information in these tables, calculate the bivariate regression equation showing the effect of household income (the independent variable) on overall health (the dependent variable), including the intercept *a* and the bivariate regression coefficient *b*.

b. Transform the unstandardized *b* you have calculated to a standardized β.

2. Dummy Variables

Results of a dummy variable regression in SAS are shown in Table 1.C. The results show differences in pride in Canada by household living arrangements (from the definition provided in Q5): "(*canadaproud*), a scale of pride in Canada that varies from 20 to 100. The living arrangements variable classified people into five categories, and these were used to develop dummy variables for four of these categories. ***The reference group is living alone***.

There are four dummy variables in the regression:

- **livespouse**: = 1 if living with spouse only; 0 otherwise
- **livenuclearfam**: = 1 if living with spouse and children; 0 otherwise
- **liveparent**: = 1 if living with parents; 0 otherwise
- **liveother**: = 1 if other arrangement; 0 otherwise

Table 1.C Regression Results in SAS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > F
Model	4	44832	11208	71.87	<.0001
Error	27402	4273432	155.95330		
Corrected Total	27406	4318264			

Root MSE	12.48813	R-Square	0.0104
Dependent Mean	76.24155	Adj R-Sq	0.0102
Coeff Var	16.37969		

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	Intercept	1	74.87337	0.15609	479.68	<.0001
livespouse	Live with Spouse Only	1	0.48689	0.21190	2.30	0.0216
livenuclearfam	Live with Spouse and Children	1	2.15710	0.20704	10.42	<.0001
liveparent	Live with one or two parents	1	3.78327	0.25753	14.69	<.0001
liveother	Live with others	1	1.22804	0.38425	3.20	0.0014

a. What is the result of the one-way analysis of variance test for any differences among these groups? (No calculations necessary; available in output).
b. What is the mean level of pride in Canada of people who live with a spouse and children?
c. Calculate the mean *difference* in pride in Canada between people who live with a spouse and children and people who live with a spouse only.

3. *Multiple Regression with Dummy Variables*

This is output from a regression of depression at Wave 2 of the National Survey of Families and Households (NSFH) on four variables: education, age, a sex dummy variable, and welfare status in childhood, also a dummy, standing for whether the respondent's parents were on welfare when they were growing up.

Table 1.D Regression of Depression on Education, Age, Sex, and Welfare Status in Childhood

Number of Observations Read	4600
Number of Observations Used	4252
Number of Observations with Missing Values	348

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > F
Model	4	61418	15354	68.55	<.0001
Error	4247	951342	224.00330		
Corrected Total	4251	1012760			

Root MSE	14.96674	R-Square	0.0606
Dependent Mean	13.22474	Adj R-Sq	0.0598
Coeff Var	113.17232		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	1	28.73657	1.72143	16.69	<.0001
education2	1	−0.91151	0.08780	−10.38	<.0001
age2	1	−0.12839	0.01951	−6.58	<.0001
female	1	3.92017	0.48401	8.10	<.0001
welfare	1	4.35591	0.82259	5.30	<.0001

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The effect of every variable here is significant. Answer these questions:

a. What is the total difference in depression between a female who grew up in a family on welfare versus a male who did not?
b. What is the difference between a male who grew up on welfare and a female who did not?
c. Work out the difference between a person with 12 years of education who is 40 years old and someone with 16 years of education who is 60 years old.

4. *Multiple Regression—with Dummy Variables*

The multiple regression results below are from the National Longitudinal Survey of Youth in the United States. The dependent variable here is ***level of education,*** measured in **years.** In this sample, it varies from 0 to 20 years. The *N* (sample size) in this regression is 4,737.

The independent variables in the output below are

- **momed**: The respondent's mother's education, in years.
- **stablepov100**: A dummy variable = 1 if the person lived in a household in adolescence consistently below the poverty line; 0 otherwise.
- **unstablepov100**: A dummy variable = 1 if the person lived in a household in adolescence that was sometimes below, sometimes above the poverty line; 0 otherwise.
- **earlyMHcv**: This is an index of emotional and behavioral problems in early adolescence. It varies from 0 problems to 15 problems.
- **asvab**: This is the person's percentile rank on a national achievement test given in early high school. Here it is measured in 10% increases, so it varies from 0 to 10. The mean is 4.5, reflecting the 45th percentile on this test.
- **female**: A dummy variable = 1 if female and = 0 if male.
- **Black**: A dummy variable = 1 if the respondent identifies as African American and = 0 otherwise.
- **Hispanic**: A dummy variable = 1 if the respondent identifies as Hispanic and = 0 otherwise.

The reference group for the poverty dummy variables is "no poverty," and the reference group for the two race dummy variables is "non-Hispanic White."

Results are shown in the following table from SAS: Two models are shown: The second model adds the effects of poverty background to the first model (i.e., stablepov100 and unstablepov100).

Table 1.E Without Poverty

Root MSE	**2.10799**	**R-Square**	**0.3401**
Dependent Mean	**12.93878**	**Adj R-Sq**	**0.3393**
Coeff Var	**16.29201**		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	**1**	**9.60066**	**0.14919**	**64.35**	**<.0001**
momed	**1**	**0.07750**	**0.00934**	**8.29**	**<.0001**
earlyMHcv	**1**	**−0.04787**	**0.01253**	**−3.82**	**0.0001**
asvab	**1**	**0.46205**	**0.01201**	**38.49**	**<.0001**
female	**1**	**0.58104**	**0.06245**	**9.30**	**<.0001**
black	**1**	**0.14741**	**0.08077**	**1.83**	**0.0681**
Hispanic	**1**	**−0.03119**	**0.08933**	**−0.35**	**0.7270**

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Table 1.F With Poverty

Root MSE	**2.09231**	**R-Square**	**0.3501**
Dependent Mean	**12.93878**	**Adj R-Sq**	**0.3490**
Coeff Var	**16.17086**		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	**1**	**9.84612**	**0.15212**	**64.73**	**<.0001**
momed	**1**	**0.07136**	**0.00936**	**7.62**	**<.0001**
stablepov100	**1**	**−2.04608**	**0.41475**	**−4.93**	**<.0001**
unstablepov100	**1**	**−0.60992**	**0.08387**	**−7.27**	**<.0001**
earlyMHcv	**1**	**−0.04560**	**0.01244**	**−3.66**	**0.0003**
asvab	**1**	**0.44264**	**0.01214**	**36.47**	**<.0001**
female	**1**	**0.59656**	**0.06203**	**9.62**	**<.0001**
black	**1**	**0.22988**	**0.08082**	**2.84**	**0.0045**
Hispanic	**1**	**0.03816**	**0.08905**	**0.43**	**0.6683**

a. Write out the *F*-test you would use to test the overall effect of childhood poverty in Model 2. Plug the relevant values into this formula, and do the calculation. Is the effect of poverty significant?
b. Using Model 2, what is the mean level of education among White males who did not grow up in poverty, had no mental health problems in adolescence, had a mother who graduated high school (12 years), and scored at the 50th percentile on the Asvab test (asvab = 5)?
c. Using Model 2 again, what is the predicted difference in education between Black females and White males? Note, you only need to state the difference, not the actual levels of each, so you can ignore other variables not involved in this comparison.

5. Multiple Regression

The multiple regression results in Table 1.G are from the Canadian General Social Survey in 2015 (Statistics Canada, 2017), restricted to the provinces of Quebec and Ontario.

The dependent variable here is pride in Canada ***(canadaproud)***, a scale of pride in Canada that varies from 20 to 100.

The independent variables in the output are

- **educyrs**: Education in years
- **female**: A dummy variable = 1 for female and 0 for male
- **provqc**: A dummy variable = 1 if the respondent lived in Quebec and = 0 if Ontario
- **employed**: A dummy variable = 1 if working now and = 0 otherwise
- **fedvote**: A dummy variable = 1 if the person voted in the last federal election and = 0 if they did not vote
- **freqnews**: Frequency of following the news in the last week, in days
- **nrcivilgroups**: Number of civil participation groups the respondent belongs to

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

TABLE 1.G ⬢ PREDICTING PRIDE IN CANADA IN A MULTIPLE REGRESSION

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |t|
Intercept	Intercept	1	80.86589	0.66115	122.31	<.0001
educyrs	Years of Education of Respondent	1	−0.13912	0.04740	−2.93	0.0033
female	Female =1, Male=0	1	−0.80933	0.22926	−3.53	0.0004
provqc	Province of residence: Quebec	1	−10.21617	0.23884	−42.77	<.0001
employed	Employed last week?	1	−0.57010	0.24150	−2.36	0.0183
fedvote	Voted in last federal election?	1	−0.52063	0.26723	−1.95	0.0514
freqnews	Frequency of following news and current affairs - Week	1	0.22482	0.04460	5.04	<.0001
nrcivilgroups	Civil Society Participation - Number of Groups - past 12 months	1	−0.22978	0.06828	−3.37	0.0008

a. According to the results, does education increase or decrease pride in Canada or have no effect?
b. If you follow the news every day (7 days a week), how much would this increase pride in Canada, according to the results?
c. Given a respondent who is female, lives in Quebec, and does not work, what is the total ***difference*** in pride in Canada compared to an employed male living in Ontario?
d. According to the results, who has the greater pride in Canada, a working person who voted in the last election or a nonworking person who did not vote?

6. Multiple Regression

The multiple regression results below are from the Canadian Quality of Life Panel Survey conducted between 1977 and 1981. The variables in the output are following.

The dependent variable is
LQ16: life satisfaction (from 1 to 11)

The independent variables are

- **NCHILD:** = Number of children living in the household
- **ONTARIO:** = 1 if from Ontario, 0 otherwise (dummy)
- **FEMALE:** = 1 if female; 0 if male (dummy)
- **HLTHPROB:** = 1 if person reports a chronic health problem, 0 otherwise (dummy)
- **Q158:** Income, measured in $10,000 increments
- **EMPLOY77:** = 1 if working, 0 if not working (dummy)

The output is from SPSS but is essentially similar to SAS output:

- *B* = unstandardized coefficient (*b*);
- BETA = the standardized coefficient (β);
- CONSTANT = the intercept (a).
- SIG T = two-tailed significance of B.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Regression Output

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	SE B	BETA	T	SIG T
NCHILD	−.099524	.036070	−.078042	−2.759	.0059
ONTARIO	−.277151	.100515	−.075744	−2.757	.0059
FEMALE	.113122	.105134	.031733	1.076	.2821
HLTHPROB	−.379606	.111195	−.094136	−3.414	.0007
Q158 ***(income)***	.056744	.014720	.113645	3.855	.0001
EMPLOY77	.256317	.111073	.072146	2.308	.0212
(CONSTANT)	8.396110	.174691		48.063	.0000

a. Only one variable here does not have a significant effect on life satisfaction. What variable is that?

b. Do children in the household increase or decrease life satisfaction?

c. How much would a $20,000 increase in income increase life satisfaction, according to these results?

d. What is the ***difference*** in life satisfaction between an employed person in Quebec and an unemployed person in Ontario?

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.