# Key Concepts

<table>
<thead>
<tr>
<th>linear regression:</th>
<th>least square</th>
</tr>
</thead>
<tbody>
<tr>
<td>slope</td>
<td>homoscedasticity, heteroscedasticity</td>
</tr>
<tr>
<td>intercept</td>
<td>linear relationship, covariance,</td>
</tr>
<tr>
<td>residual</td>
<td>product-moment correlation, rank</td>
</tr>
<tr>
<td>error sum of squares or</td>
<td>correlation</td>
</tr>
<tr>
<td>residual sum of squares</td>
<td>multiple regression, stepwise regression,</td>
</tr>
<tr>
<td>sum of squares due to regression</td>
<td>regression diagnostics, multiple</td>
</tr>
<tr>
<td>mean squares</td>
<td>correlation coefficient, partial correlation</td>
</tr>
<tr>
<td>error mean squares</td>
<td>coefficient</td>
</tr>
<tr>
<td>regression (coefficient) of $x$ on $y$</td>
<td>regression toward the mean</td>
</tr>
</tbody>
</table>
SYMBOLS AND ABBREVIATIONS

\( b_0 \) sample intercept
\( b_1, b_2, \ldots \) sample regression coefficients
MS mean square
\( r \) correlation coefficient
\( R \) multiple correlation coefficient
\( R^2 \) proportion of the variability explained by a regression model
\( s_{xy} \) sample covariance of \( X \) and \( Y \) (estimate)
SS sum of squares
\( \hat{y} \) value of \( y \) predicted from a regression equation
\( e \) estimated error or residual from a regression model
\( \beta_0 \) population intercept
\( \beta_1, \beta_2, \ldots \) population regression coefficients
\( \varepsilon \) error or residual from a regression model (Greek letter epsilon)

SIMPLE LINEAR REGRESSION

In Chapter 9 we discussed categorical data and introduced the notion of response and predictor variables. We turn now to a discussion of relationships between response and predictor variables of the continuous type. To explain such a relationship we often search for mathematical models such as equations for straight lines, parabolas, or other mathematical functions. The anthropologist Sir Francis Galton (1822–1911) used the term regression in explaining the relationship between the heights of fathers and their sons. From a group of 1078 father–son pairs he developed the following model, in which the heights are in inches:

\[
\text{son's height} = 33.73 + 0.516 \text{ (father's height)}.
\]
If we substitute 74 inches into this equation for the father’s height, we arrive at about 72 inches for the son’s height (i.e. the son is not as tall as his father). On the other hand, if we substitute 65 inches for the father’s height, we find the son’s height to be 67 inches (i.e. the son is taller than his father). Galton concluded that although tall fathers tend to have tall sons and short fathers tend to have short sons, the son’s height tends to be closer to the average than his father’s height. Galton called this ‘regression toward the mean’. Although the techniques for modeling relationships among variables have taken on a much wider meaning, the term ‘regression’ has become entrenched in the statistical literature to describe this modeling. Thus, nowadays we speak of regression models, regression equations, and regression analysis without wishing to imply that there is anything ‘regressive’. As we shall see later, however, the phrase ‘regression toward the mean’ is still used in a sense analogous to what Galton meant by it.

**Figure 10.1** Graph of the equation $y = 3 + 2x$

We begin by discussing a relatively simple relationship between two variables, namely a straight-line relationship. The equation $y = 3 + 2x$ is the equation of the straight line shown in Figure 10.1. The equation $y = 2 - 3x$ is shown in Figure 10.2,

**Figure 10.2** Graph of the equation $y = 2 - 3x$
and the equation $y = 0.5x$ in Figure 10.3. In general, the equation of a straight line can be expressed in the form

$$y = \beta_0 + \beta_1 x$$

where $\beta_0$ and $\beta_1$ are specified constants. Any point on this line has an $x$-coordinate and a $y$-coordinate. When $x = 0$, $y = \beta_0$; so the parameter $\beta_0$ is the value of the $y$-coordinate where the line crosses the $y$-axis and is called the intercept. In Figure 10.1, the intercept is 3, in Figure 10.2 it is 2. When $\beta_0 = 0$, the line goes through the origin (Figure 10.3) and the intercept is 0. The parameter $\beta_1$ is the slope of the line and measures the amount of change in $y$ per unit increase in $x$. When $\beta_1$ is positive (as in Figures 10.1 and 10.3), the slope is upward and $x$ and $y$ increase together; when $\beta_1$ is negative (Figure 10.2), the slope is downward and $y$ decreases as $x$ increases. When $\beta_1$ is zero, $y$ is the same (it has value $\beta_0$) for all values of $x$ and the line is horizontal (i.e. there is no slope). The parameter $\beta_1$ is undefined for vertical lines but approaches infinity as the line approaches a vertical position.

So far in our discussion we have assumed that the relationship between $x$ and $y$ is explained exactly by a straight line; if we are given $x$ we can determine $y$ – and vice versa – for all values of $x$ and $y$. Now let us assume that the relationship between the two variables is not exact, because one of the variables is subject to random measurement errors. Let us call this random variable the response variable and denote it $Y$. The other variable $x$ is assumed to be measured without error; it is under the control of the investigator and we call it the predictor variable. This terminology is consistent with that of Chapter 9. In practice the predictor variable will also often be subject to random variability caused by errors of measurement, but we assume that this variability is negligible relative to that of the response variable. For example, suppose an investigator is interested in the rate at which a metabolite is consumed or produced by an enzyme reaction. A reaction mixture is set up from which aliquots are withdrawn at various intervals of time and the
concentration of the metabolite in each aliquot is determined. Whereas the time at which each aliquot is withdrawn can be accurately measured, the metabolite concentration, being determined by a rather complex assay procedure, is subject to appreciable measurement error. In this situation, time would be the predictor variable under the investigator’s control, and metabolite concentration would be the random response variable.

Since the relationship is not exact, we write

\[ Y = \beta_0 + \beta_1 x + \varepsilon \]

where \( \varepsilon \), called the error, is the amount by which the random variable \( Y \), for a given \( x \), lies above the line (or below the line, if it is negative). This is illustrated in Figure 10.4. In a practical situation, we would have a sample of pairs of numbers, \( x_1 \) and \( y_1 \), \( x_2 \) and \( y_2 \), and so on. Then, assuming a straight line is an appropriate model, we would try to find the line that best fits the data. In other words, we would try to find estimates \( b_0 \) and \( b_1 \) of the parameters \( \beta_0 \) and \( \beta_1 \), respectively. One approach that yields estimates with good properties is to take the line that minimizes the sum of the squared errors (i.e. that makes the sum of the squared vertical deviations from the fitted line as small as possible). These are called least-squares estimates. Thus, if we have a sample of pairs, we can denote a particular pair (the \( i \)th pair) \( x_i, y_i \), so that

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \]

**Figure 10.4** Graph of the regression equation \( y = 4 + 2x \) and errors for the points \((x = 2, y = 4)\), and \((x = 6, y = 17)\).
The $i$th error is
\[ \varepsilon_i = y_i - \beta_0 - \beta_1 x_i \]
as illustrated in Figure 10.4, and its square is
\[ \varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2. \]

Then the least-squares estimates $b_0$ and $b_1$ of $\beta_0$ and $\beta_1$, respectively, that minimize the sum of these squared deviations over all the sample values. The slope $\beta_1$ or its least-squares estimate $b_1$ is also called the regression of $y$ on $x$, or the regression coefficient of $y$ on $x$.

Notice that if the line provides a perfect fit to the data (i.e. all the points fall on the line), then $\varepsilon_i = 0$ for all $i$. Moreover, the poorer the fit, the greater the magnitudes, either positive or negative, of the $\varepsilon_i$. Now let us define the fitted line by $\hat{y}_i = b_0 + b_1 x_i$, and the estimated error, or residual, by
\[ e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i. \]

Then a special property of the line fitted by least squares is that the sum of $e_i$ over the whole sample is zero. If we sum the squared residuals $e_i^2$, we obtain a quantity called the error sum of squares, or residual sum of squares.

If the line is horizontal ($b_1 = 0$), as in Figure 10.5, the residual sum of squares is equal to the sum of squares about the sample mean. If the line is neither horizontal nor vertical, we have a situation such as that illustrated in Figure 10.6. The deviation of the $i$th observation from the sample mean, $y_i - \bar{y}$, has been partitioned into two components: a deviation from the regression line, $y_i - \hat{y}_i = e_i$, the estimated error or residual; and a deviation of the regression line from the mean, $\hat{y}_i - \bar{y}$, which we call the deviation due to regression (i.e. due to the straight-line model). If we square each of these three deviations ($y_i - \bar{y}$, $y_i - \hat{y}_i$, and $\hat{y}_i - \bar{y}$) and separately add them,
Figure 10.6  Graph of the regression equation $\hat{y} = b_0 + b_1 x_i$ showing how the difference between each $y_i$ and the mean $\bar{y}$ can be decomposed from the line $(y_i - \hat{y}_i)$ and a deviation of the line from the mean $(\hat{y}_i - \bar{y})$.

up over all the sample values, we obtain three sums of squares which satisfy the following relationship:

\[
\text{total sum of squared deviations from the mean} = \text{sum of squared deviations from the regression model} + \text{sum of squared deviations due to, or 'explained by', the regression model.}
\]

We often abbreviate this relationship by writing

\[
\text{SS}_T = \text{SS}_E + \text{SS}_R
\]

where $\text{SS}_T$ is the total sum of squares about the mean, $\text{SS}_E$ is the error, or residual, sum of squares, and $\text{SS}_R$ is the sum of squares due to regression.

These three sums of squares have $n - 1$, $n - 2$, and 1 d.f., respectively. If we divide the last two sums of squares by their respective degrees of freedom, we obtain quantities called mean squares: the error, or residual, mean square and the mean square due to regression. These mean squares are used to test for the significance of the regression, which in the case of a straight-line model is the same as testing whether the slope of the straight line is significantly different from zero. In the example discussed above, we may wish to test whether the metabolite concentration in the reaction mixture is in fact changing, or whether it is the same at all the different points in time. Thus we would test the null hypothesis $H_0: \beta_1 = 0$. Denote the error mean square $\text{MS}_E$ and the mean square due to the straight line regression model $\text{MS}_R$. Then, under $H_0$ and certain conditions that we shall specify, the ratio

\[
F = \frac{\text{MS}_R}{\text{MS}_E}
\]

follows the $F$-distribution with 1 and $n - 2$ d.f. (Note: as is always true of an $F$-statistic, the first number of degrees of freedom corresponds to the numerator, here $\text{MS}_R$, and the second to the denominator, here $\text{MS}_E$). As $\beta_1$ increases in magnitude, the numerator of the $F$-ratio will tend to have large values, and as $\beta_1$
approaches zero, it will tend toward zero. Thus, large values of \( F \) indicate departure from \( H_0 \), whether because \( \beta_1 \) is greater or less than zero. Thus if the observed value of \( F \) is greater than the 95th percentile of the \( F \)-distribution, we reject \( H_0 \) at the 5\% significance level for a two-sided test. Otherwise, there is no significant linear relationship between \( Y \) and \( x \).

The conditions necessary for this test to be valid are the following:

1. For a fixed value of \( x \), the corresponding \( Y \) must come from a normal distribution with mean \( \beta_0 + \beta_1 x \).
2. The \( Y \)'s must be independent.
3. The variance of \( Y \) must be the same at each value of \( x \). This is called \textit{homoscedasticity}; if the variance changes for different values of \( x \), we have \textit{heteroscedasticity}.

Furthermore, under these conditions, the least-squares estimates \( b_0 \) and \( b_1 \) are also the maximum likelihood estimates of \( \beta_0 \) and \( \beta_1 \).

The quantities \( b_0 \), \( b_1 \) and the mean squares can be automatically calculated on a computer or a pocket calculator. The statistical test is often summarized as shown in Table 10.1. Note that \( SS_T/(n - 1) \) is the usual estimator of the (total) variance of \( Y \) if we ignore the \( x \)-values. The estimator of the variance of \( Y \) about the regression line, however, is \( MSE \), the error mean square or the mean squared error. It estimates the error, or residual, variance not explained by the model.

### Table 10.1  Summary results for testing the hypothesis of zero slope (linear regression analysis)

<table>
<thead>
<tr>
<th>Source of Variability in ( Y )</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>( F )-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>( SS_R )</td>
<td>( MS_R )</td>
<td>( MS_R/MS_E )</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>( n - 2 )</td>
<td>( SSE )</td>
<td>( MSE )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td>( SST )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now \( b_1 \) is an estimate of \( \beta_1 \) and represents a particular value of an estimator which has a standard error that we shall denote \( SB_1 \). Some computer programs and pocket calculators give these quantities instead of (or in addition to) the quantities in Table 10.1. Then, under the same conditions, we can test \( H_0: \beta_1 = 0 \) by using the statistic

\[
t = \frac{b_1}{SB_1},
\]
which under $H_0$ comes from Student’s $t$-distribution with $n - 2$ d.f. In fact, the square of this $t$ is identical to the $F$-ratio defined earlier. So, in the case of simple linear regression, either an $F$-test or a $t$-test can be used to test the hypothesis that the slope of the regression line is equal to zero.

**THE STRAIGHT-LINE RELATIONSHIP WHEN THERE IS INHERENT VARIABILITY**

So far we have assumed that the only source of variability about the regression line is due to measurement error. But you will find that regression analysis is often used in the literature in situations in which it is known that the pairs of values $x$ and $y$, even if measured without error, do not all lie on a line. The reason for this is not so much because such an analysis is appropriate, but rather because it is a relatively simple method of analysis, easily generalized to the case in which there are multiple predictor variables (as we discuss later in this chapter), and readily available in many computer packages of statistical programs. For example, regression analysis might be used to study the relationship between triglyceride and cholesterol levels even though, however accurately we measure these variables, a large number of pairs of values will never fall on a straight line, but rather give rise to a scatter diagram similar to Figure 3.7. Regression analysis is not the appropriate statistical tool if, in this situation, we want to know how triglyceride levels and cholesterol levels are related in the population. It is, however, an appropriate tool if we wish to develop a prediction equation to estimate one from the other.

Let us call triglyceride level $x$ and cholesterol level $Y$. Using the data illustrated in Figure 3.7, it can be calculated that the estimated regression equation of $Y$ on $x$ is

$$\hat{y} = 162.277 + 0.217x$$

and the residual variance is estimated to be $776 \text{ (mg/dl)}^2$. This variance includes both measurement error and natural variability, so it is better to call it ‘residual’ variance rather than ‘error’ variance. Thus, for a population of persons who all have a measured triglyceride level equal to $x$, we estimate that the random variable $Y$, measured cholesterol level, has mean $162.277 + 0.217x \text{ mg/dl}$ and standard deviation $\sqrt{776} = 27.857 \text{ mg/dl}$. This is the way we use the results of regression analysis to predict the distribution of cholesterol level ($Y$) for any particular value of triglyceride level ($x$). But we must not use the same equation to predict triglyceride level from cholesterol level. If we solve the equation

$$\hat{y} = 162.277 + 0.217x$$

for \( x \), we obtain

\[
x = -747.820 + 4.608 \hat{y}.
\]

This is exactly the same line, the regression of \( Y \) on \( x \), although expressed differently. It is not the regression of the random variable \( X \) on \( y \). Such a regression can be estimated, and for these same data it turns out to be

\[
\hat{x} = -41.410 + 0.819y
\]

with residual variance 2930 (mg/dl)². This is the equation that should be used to predict triglyceride level from cholesterol level. Figure 10.7 is a repeat of Figure 3.7 with these two regression lines superimposed on the data. The regression of \( Y \) on \( x \) is obtained by minimizing the sum of the squared vertical deviations (deviations in \( Y \), that is, parallel to the \( y \)-axis) from the straight line and can be used to predict the distribution of \( Y \) for a fixed value of \( x \). The regression of \( X \) on \( y \), on the other hand, is the converse situation: it is obtained by minimizing the sum of the squared horizontal deviations (deviations in the random variable \( X \), that is, parallel to the \( x \)-axis), and it is the appropriate regression to use if we wish to predict the distribution of \( X \) for a fixed (controlled) value of \( y \). In this latter case, \( X \) is the response variable and \( y \) the predictor variable.

---

**Figure 10.7** Scatterplot of serum cholesterol versus triglyceride levels of 30 medical students with the two estimated regression lines.
It is clear that these are two different lines. Furthermore, the line that best describes the single, underlying linear relationship between the two variables falls somewhere between these two lines. It is beyond the scope of this book to discuss the various methods available for finding such a line, but you should be aware that such methods exist, and that they depend on knowing the relative accuracy with which the two variables $X$ and $Y$ are measured.

## CORRELATION

In Chapter 3 we defined variance as a measure of dispersion. The definition applies to a single random variable. In this section we introduce a more general concept of variability called covariance. Let us suppose we have a situation in which two random variables are observed for each study unit in a sample, and we are interested in measuring the strength of the association between the two random variables in the population. For example, without trying to estimate the straight-line relationship itself between cholesterol and triglyceride levels in male medical students, we might wish to estimate how closely the points in Figure 10.7 fit an underlying straight line. First we shall see how to estimate the covariance between the two variables.

*Covariance* is a measure of how two random variables vary together, either in a sample or in the population, when the values of the two random variables occur in pairs. To compute the covariance for a sample of values of two random variables, say $X$ and $Y$, with sample means $\bar{x}$ and $\bar{y}$, respectively, the following steps are taken:

1. For each pair of values $x_i$ and $y_i$, subtract $\bar{x}$ from $x_i$ and $\bar{y}$ from $y_i$ (i.e. compute the deviations $x_i - \bar{x}$ and $y_i - \bar{y}$).
2. Find the product of each pair of deviations (i.e. compute $(x_i - \bar{x}) \times (y_i - \bar{y})$).
3. Sum these products over the whole sample.
4. Divide the sum of these products by one less than the number of pairs in the sample.

Suppose, for example, we wish to compute the sample covariance for $X$, $Y$ from the following data:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>110</td>
</tr>
</tbody>
</table>
Note that the sample means are $\bar{x} = 30$ and $\bar{y} = 70$. We follow the steps outlined above.

1. Subtract $\bar{x}$ from $x_i$ and $\bar{y}$ from $y_i$:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x_i - \bar{x}$</th>
<th>$y_i - \bar{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$10 - 30 = -20$</td>
<td>$30 - 70 = -40$</td>
</tr>
<tr>
<td>2</td>
<td>$20 - 30 = -10$</td>
<td>$50 - 70 = -20$</td>
</tr>
<tr>
<td>3</td>
<td>$30 - 30 = 0$</td>
<td>$70 - 70 = 0$</td>
</tr>
<tr>
<td>4</td>
<td>$40 - 30 = 10$</td>
<td>$90 - 70 = 20$</td>
</tr>
<tr>
<td>5</td>
<td>$50 - 30 = 20$</td>
<td>$110 - 70 = 40$</td>
</tr>
</tbody>
</table>

2. Find the products

<table>
<thead>
<tr>
<th>$i$</th>
<th>$(x_i - \bar{x}) \times (y_i - \bar{y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$-20 \times (-40) = 800$</td>
</tr>
<tr>
<td>2</td>
<td>$-10 \times (-20) = 200$</td>
</tr>
<tr>
<td>3</td>
<td>$0 \times 0 = 0$</td>
</tr>
<tr>
<td>4</td>
<td>$10 \times 20 = 200$</td>
</tr>
<tr>
<td>5</td>
<td>$20 \times 40 = 800$</td>
</tr>
</tbody>
</table>

4. Divide by one less than the number of pairs:

$$s_{xy} = \text{sample covariance} = \frac{2000}{5 - 1} = 500.$$

Note in steps 1 and 2 that, if both members of a pair are below their respective means (as in the case for the first pair, $i = 1$), the contribution to the covariance is positive (+800 for the first pair). It is similarly positive when both members of the pair are above their respective means (+200 and +800 for $i = 4$ and 5, in the example). Thus, a positive covariance implies that $X$ and $Y$ tend to covary in such a manner that when one is either below or above its mean, so is the other. A negative covariance, on the other hand, would imply a tendency for one to be above its mean when the other is below its mean, and vice versa.

Now suppose $X$ is measured in pounds and $Y$ in inches. Then the covariance is in pound-inches, a mixture of units that is difficult to interpret. Recall that a variance
is measured in squared units and we take the square root of the variance to get back to the original units. Obviously this does not work for the covariance. Instead, we divide the covariance by the product of the estimated standard deviation of \( X \) and the estimated standard deviation of \( Y \), which we denote \( s_X \) and \( s_Y \), respectively. The result is a pure, dimensionless number (no units) that is commonly denoted \( r \) and called the \textit{correlation coefficient}, or \textit{Pearson's product-moment correlation coefficient},

\[
    r = \frac{s_{XY}}{s_X s_Y},
\]

where \( s_{XY} \) is the sample covariance of \( X \) and \( Y \), \( s_X \) is the sample standard deviation of \( X \), and \( s_Y \) the sample standard deviation of \( Y \). Thus, in the above example, \( s_X = \sqrt{250} \) and \( s_Y = \sqrt{1000} \), so

\[
    r = \frac{500}{\sqrt{250}\sqrt{1000}} = 1.
\]

In this example the correlation coefficient is +1. A scatterplot of the data indicates that all the points \((x_i, y_i)\) lie on a straight line with positive slope, as illustrated in Figure 10.8(a).

If all the points lie on a straight line with negative slope, as in Figure 10.8(b), the correlation coefficient is −1. These are the most extreme values possible: a correlation can only take on values between −1 and +1. Figures 10.8(a–h) illustrate a variety of possibilities, and it can be seen that the magnitude of the correlation measures how close the points are to a straight line. Remember that the correlation coefficient is a dimensionless number, and so does not depend on the units of measurement. In Figures 10.8(a–h), the scales have been chosen so that the numerical value of the sample variance of \( Y \) is about the same as that of \( X \) – you can see that in each figure the range of \( Y \) is the same as the range of \( X \). Now look at Figures 10.8(i, j). In each case the points appear to be close to a straight line, and you might therefore think that the correlation coefficient should be large in magnitude. If the scales are changed to make the range of \( Y \) the same as the range of \( X \), however, Figure 10.8(i) becomes identical to Figure 10.8(g), and Figure 10.8(j) becomes identical to Figure 10.8(h). Once the scales have been adjusted, it becomes obvious that the correlation coefficient is near zero in each of these two situations. Of course, if all the points are on a horizontal line or on a vertical line, it is impossible to adjust the scales so that the range is numerically the same for both variables. In such situations, as illustrated in Figures 10.8(k, l), the correlation coefficient is undefined.
Figure 10.8 Scatterplots illustrating how the correlation coefficient, $r$, is a measure of the linear association between two variables.
Note that the denominator in the correlation coefficient is the product of the sample standard deviations, which include both natural variability and measurement errors. Thus (unless the measurement errors in the two variables are themselves correlated), larger measurement errors automatically decrease the correlation coefficient. A small correlation between two variables can thus be due either to (1) little linear association between the two variables, or (2) large errors in their measurement. A correlation close to +1 or −1, on the other hand, implies that the measurement errors must be small relative to the sample standard deviations, and that the data points all lie close to a straight line. In fact, there is a close connection between the correlation coefficient and the estimated slope of the regression line. The estimated slope of the regression of $Y$ on $x$ is $r_{SY}/s_X$, and the estimated slope of the regression of $X$ on $y$ is $r_{SX}/s_Y$. The correlation coefficient is significantly different from zero if, and only if, the regression coefficients are significantly different from zero; and such a finding implies a dependency between the two variables. However, a correlation coefficient of zero does not imply two variables are independent (see Figure 10.8(h)) and, as we have seen before, a dependency between two variables does not necessarily imply a causal relationship between them.

**SPEARMAN’S RANK CORRELATION**

If we rank the $x$s from 1 to $n$ (from largest to smallest, or vice versa), and we rank the $y$s from 1 to $n$ in the same direction, and then compute the correlation coefficient between the pairs of ranks, we obtain the so-called *rank correlation coefficient*, or *Spearman’s rank correlation coefficient*. This correlation measures how closely the points can be fitted by a smooth, monotonic curve (i.e. a curve that is either always increasing or always decreasing). The rank correlations of the data in Figures 10.8(e, f) are +1, and −1, respectively. The curve that best fits the data in Figure 10.8(h), however, is not monotonic; it first increases and then decreases, and the rank correlation for these data is also approximately 0. Apart from being a measure of closeness to a monotonic curve, the rank correlation coefficient is less subject to influence by a few extreme values, and therefore sometimes gives a more reasonable index of association.

**MULTIPLE REGRESSION**

We have seen how a straight-line regression model can be fitted to data so that the variable $x$ ‘explains’ part of the variability in a random variable $Y$. A natural question to ask is, if one variable can account for part of the variability in $Y$, can more of
the variability be explained by further variables? This leads us to consider models such as

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \epsilon, \]

where \( x_1, x_2, \ldots, x_q \) are \( q \) distinct predictor variables. Just as before, we can partition the total sum of squared deviations from the mean of \( y \) as

\[ SS_T = SS_E + SS_R, \]

where \( SS_E \) is the sum of squared deviations from the regression model and \( SS_R \) is the sum of squared deviations due to, or ‘explained by’, the regression model. But now \( SS_E \) has \( n - q - 1 \) d.f. and \( SS_R \) has \( q \) d.f. Following the same line of reasoning as used in the case of simple linear regression, we can compute the quantities indicated in Table 10.2. Thus, \( F = MS_R/MS_E \) with \( q \) and \( n - q - 1 \) d.f. and provides a simultaneous test of the hypothesis that all the regression coefficients are equal to zero,

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0. \]

<table>
<thead>
<tr>
<th>Source of Y Variability in Y</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>( q )</td>
<td>( SS_R )</td>
<td>( MS_R )</td>
<td>( MS_R/MS_E )</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>( n - q - 1 )</td>
<td>( SS_E )</td>
<td>( MS_E )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td>( SS_T )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sum of squares for regression can be further partitioned into \( q \) terms, each with 1 d.f., so that each coefficient can be tested separately. Thus, the results of a multiple regression analysis with three predictor variables might look something like Table 10.3. The \( F \)-test provides an overall test of whether the coefficients \( \beta_1, \beta_2, \) and \( \beta_3 \) are simultaneously zero. The \( t \)-tests provide individual tests for each coefficient separately. These \( t \)-tests are to be interpreted only in light of the full model. Thus, if we drop \( x_2 \) from the model because the results suggest \( \beta_2 \) is not significantly different from zero, then we can fit a new model, say

\[ Y = \beta_0' + \beta_1' x_1 + \beta_3' x_3 + \epsilon, \]
Table 10.3  Summary results of regression analysis for the model
\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \]

<table>
<thead>
<tr>
<th>Source of Y Variability in Y</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression model</td>
<td>3</td>
<td>SS_R</td>
<td>MS_R</td>
<td>MS_R/MS_E</td>
</tr>
<tr>
<td>Error (residual)</td>
<td>n - 4</td>
<td>SS_E</td>
<td>MS_E</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n - 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error of Estimate</th>
<th>t</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>( b_0 )</td>
<td>( s_{\beta_0} )</td>
<td>( b_0/s_{\beta_0} )</td>
<td>( p_0 )</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>( b_1 )</td>
<td>( s_{\beta_1} )</td>
<td>( b_1/s_{\beta_1} )</td>
<td>( p_1 )</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>( b_2 )</td>
<td>( s_{\beta_2} )</td>
<td>( b_2/s_{\beta_2} )</td>
<td>( p_2 )</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>( b_3 )</td>
<td>( s_{\beta_3} )</td>
<td>( b_3/s_{\beta_3} )</td>
<td>( p_3 )</td>
</tr>
</tbody>
</table>

to make inferences about the coefficients of \( x_1 \) and \( x_3 \) with \( x_2 \) removed from the model. Note in particular that \( \beta_1 \) in the full model is not equal to \( \beta_1' \) in the reduced model. Similarly,

\[ \beta_3 \neq \beta_3'. \]

To study the effect of each predictor variable fully, it is necessary to perform a regression analysis for every possible combination of the predictor variables. In the example of Table 10.3, this would entail conducting a regression analysis for each of the following models:

1. \( Y \) regressed on \( x_1 \)
2. \( Y \) regressed on \( x_2 \)
3. \( Y \) regressed on \( x_3 \)
4. \( Y \) regressed on \( x_1 \) and \( x_2 \)
5. \( Y \) regressed on \( x_1 \) and \( x_3 \)
6. \( Y \) regressed on \( x_2 \) and \( x_3 \)
7. \( Y \) regressed on \( x_1, x_2, \) and \( x_3 \).

The larger the number of predictor variables, the greater the number of possible combinations, and so it is often not feasible to perform all possible regression analyses. For this reason a stepwise approach is often used, though it may not lead to the best subset of variables to keep in the model. There are two types of stepwise regression: forward and backward.
1. *Forward stepwise regression* first puts into the model the single predictor variable that explains most of the variability in \( Y \), and then successively at each step inserts the variable that explains most of the remaining (residual) variability in \( Y \). However, if at any step none of the remaining predictor variables explain a significant additional amount of variability in \( Y \), at a predetermined level of significance, the procedure is terminated.

2. *Backward stepwise regression* includes all the predictor variables in the model to begin with, and then successively at each step the variable that explains the least amount of variability in \( Y \) (in the presence of the other predictor variables) is dropped from the model. However, a variable is dropped only if, at a predetermined level of significance, its contribution to the variability of \( Y \) (in the presence of the other variables) is not significant.

Whatever method is used to select among a set of predictor variables in order to arrive at the ‘best’ subset to be included in a regression model, it must always be remembered that the final result is merely a *prediction* model, and not necessarily a model for the *causation* of variability in the response variable.

Suppose a multiple regression analysis is performed assuming the model

\[
Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon
\]

and, based on a sample of \( n \) study units, on each of which we have observations \((y, x_1, x_2, x_3)\), we find

\[
b_0 = 40, \quad b_1 = 5, \quad b_2 = 10, \quad b_3 = 7,
\]

as estimates of \( \beta_0, \beta_2, \beta_2, \) and \( \beta_3 \), respectively. The fitted regression model in this case is

\[
\hat{y} = 40 + 5x_1 + 10x_2 + 7x_3.
\]

For each of the \( n \) study units we can substitute \( x_1, x_2, \) and \( x_3 \) into the fitted model to obtain a value \( \hat{y} \). Suppose, for example, the observations on one of the study units were \((y, x_1, x_2, x_3) = (99, 4, 2, 3)\). On substituting the \( x \)s into the fitted model, we obtain

\[
\hat{y} = 40 + 5 \times 4 + 10 \times 2 + 7 \times 3 = 101.
\]

This procedure provides us with an estimate of the expected value of \( Y \) for the observed set of \( x \)s. We actually observed \( y = 99 \) for these \( x \)s; however, if we had
observed a second value of $Y$ for these same $x$s, that value would likely be some number other than 99. For each set of $x$s, our model assumes there is a distribution of $y$s corresponding to the random variable $Y$. Thus, $\hat{y}$ is an estimate of the mean value of $Y$ for that set of $x$s, and $y - \hat{y}$ ($99 - 101 = 2$, in our example) is the residual.

If we compute $y$ for each of the $n$ sample study units, and then compute the $n$ residuals $y - \hat{y}$, we can examine these residuals to investigate the adequacy of the model. In particular, we can obtain clues as to whether:

(i) the regression function is linear;
(ii) the residuals have constant variance;
(iii) the residuals are normally distributed;
(iv) the residuals are not independent;
(v) the model fits all but a few observations;
(vi) one or more predictor variables have been omitted from the model.

Methods for investigating model adequacy are called regression diagnostics. Regression diagnostics play an important role in statistical modeling because it is so easy to fit models with existing computer programs, whether or not those models are really appropriate. Use of good regression diagnostics will guard against blindly accepting misleading models.

Before leaving multiple regression, we should note that a special case involves fitting polynomial (curvilinear) models to data. We may have measured only one predictor variable $x$, but powers of $x$ are also included in the regression model as separate predictor variables. For example, we may fit such models as the quadratic model, or parabola,

$$Y = \beta_0' + \beta_1'x + \beta_2'x^2 + \varepsilon$$

and the cubic model

$$Y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \varepsilon.$$
overall linear association between the response variable $Y$ and the set of predictor variables $x_1, x_2, \ldots, x_q$ in the regression equation. In the special case where $q = 1$ (i.e. if there is only one predictor variable), the multiple correlation coefficient $R$ is simply equal to $r$, the correlation coefficient between $X$ and $Y$. In general, $R^2$ equals the ratio $SS_R/SS_T$, and so measures the proportion of the variability explained by the regression model. If we fit a model with three predictor variables and find $R^2 = 0.46$, and then fit a model that includes an additional, fourth predictor variable and find $R^2 = 0.72$, we would conclude that the fourth variable accounts for an additional 26% ($0.72 - 0.46$) of the variability in $Y$. The square of the multiple correlation coefficient, $R^2$, is often reported when regression analyses have been performed.

The partial correlation coefficient is a measure of the strength of linear association between two variables after controlling for one or more other variables. Suppose, for example, we are interested in the correlation between serum cholesterol and triglyceride values in a random sample of men aged 20–65. Now it is known that both cholesterol and triglyceride levels tend to increase with age, so the mere fact that the sample includes men from a wide age range would tend to cause the two variables to be correlated in the sample. If we wish to discount this effect, controlling for age (i.e. determine that part of the correlation that is over and above the correlation induced by a common age), we would calculate the partial correlation coefficient, ‘partialing out the effect of’ the variable age. The square of the partial correlation between the cholesterol and tryglyceride levels would then be the proportion of the variability in cholesterol level that is accounted for by the addition of triglyceride to a regression model that already includes age as a predictor variable. Similarly, it would also equal the proportion of the variability in triglyceride level that is accounted for by the addition of cholesterol to a regression model that already includes age as an predictor variable.

**REGRESSION TOWARD THE MEAN**

We conclude this chapter with a concept that, although it includes in its name the word ‘regression’ and is indeed related to the original idea behind regression, is distinct from modeling the distribution of a random variable in terms of one or more other variables. Consider the three highest triglyceride values among those listed in Table 3.1 for 30 medical students. They are (in mg/dl) 218, 225, and 287, with a mean of 243.3. Suppose we were to draw aliquots of blood from these three students on several subsequent days; should we expect the mean of the subsequent values to be 243.3? Alternatively, if we took later samples from those students with the three lowest values (45, 46, and 49), should we expect their average to remain the same (46.7)? The answer is that we should not: we should expect the mean of the
highest three to become smaller, and the mean of the lowest three to become larger, on subsequent determinations. This phenomenon is called regression toward the mean and occurs whenever we follow up a selected, as opposed to a complete, or random, sample. To understand why regression toward the mean occurs, think of each student's measured triglyceride value as being made up of two parts: a ‘true’ value (i.e. the mean of many, many determinations made on that student) and a random deviation from that true value; this latter could be due to measurement error and/or inherent variability in triglyceride value from day to day. When we select the three students with the highest triglyceride values based on a single measurement on each student, we tend to pick three that happen to have their highest random deviations, so that the mean of these three measurements (243.3 in our example) is most probably an overestimate of the mean of the three students’ ‘true’ values. Subsequent measures on these three students are equally likely to have positive or negative random deviations, so the subsequent mean will be expected to be the mean of their ‘true’ values, and therefore probably somewhat lower. Similarly, if we pick the lowest three students in a sample, these single measurements will usually be underestimates of their true values, because they were probably picked partly because they happened to have their lowest random deviations when they were selected. If we were to make subsequent observations on the whole sample of 30 students, however, or on a random sample of them, regression toward the mean would not be expected to occur.

It is important to distinguish between regression toward the mean and a treatment effect. If subjects with high cholesterol levels are given a potentially cholesterol-lowering drug, their mean cholesterol level would be expected to decrease on follow-up – even if the drug is ineffective – because of regression toward the mean. This illustrates the importance of having a control group taking a placebo, with subjects randomly assigned to the two groups. Regression toward the mean is then expected to occur equally in both groups, so that the true treatment effect can be estimated by comparing the groups.

Finally, something analogous to regression toward the mean tends to occur whenever multiple regression is used to select a set of variables that best explains, or predicts, a response variable. Given a sample of data, when a model is first fitted to a set of predictor variables all the estimated regression coefficients are unbiased. But if we now select the most significant predictors and only report these – or include only these in a new model that is fitted, using the same data set – we have automatically chosen predictors that best explain the response in this particular sample and the new estimates will be biased. In other words, we should expect estimates of these regression coefficients, when estimated from future samples taken from the same population, to be closer to zero and hence much less significant. This is one reason why many studies that first report a significant finding cannot be replicated by later investigators.
SUMMARY

1. In the equation of a straight line, \( y = \beta_0 + \beta_1 x \), \( \beta_0 \) is the intercept and \( \beta_1 \) is the slope.

2. In simple linear regression analysis, it is assumed that one variable (\( Y \)), the response variable, is subject to random fluctuations, whereas the other variable (\( x \)), the predictor variable, is under the investigator's control. Minimizing the sum of the squared deviations of \( n \) sample values of \( Y \) from a straight line leads to the least-squares estimates \( b_0 \) of \( \beta_0 \) and \( b_1 \) of \( \beta_1 \), and hence the prediction line \( \hat{y} = b_0 + b_1 x \). The sample residuals about this line sum to zero.

3. The total sum of squared deviations from the sample mean \( \bar{y} \) can be partitioned into two parts — that due to the regression model and that due to error, or the residual sum of squares. Dividing these by their respective degrees of freedom gives rise to mean squares.

4. Under certain conditions the estimates \( b_0 \) and \( b_1 \) are maximum likelihood estimates, and the ratio of the mean squares (that due to regression divided by that due to residual) can be compared to the \( F \)-distribution with 1 and \( n - 2 \) d.f. to test the hypothesis \( \beta_1 = 0 \). These conditions are: (a) for a fixed \( x \), \( Y \) must be normally distributed with mean \( \beta_0 + \beta_1 x \); (b) the \( Y \)s must be independent; and (c) there must be homoscedasticity — the variance of \( Y \) must be the same at each value of \( x \).

5. It is possible to determine a standard error for \( b_1 \) and, under the same conditions, \( b_1 \) divided by its standard error comes from a \( t \)-distribution with \( n - 2 \) d.f. In this situation, the \( F \)-test and the \( t \)-test are equivalent tests of the null hypothesis \( \beta_1 = 0 \).

6. The regression of \( Y \) on \( x \) can be used to predict the distribution of \( Y \) for a given value \( x \), and the regression of \( X \) on \( y \) can be used to predict the distribution of \( X \) for a given value \( y \). The line that best describes the underlying linear relationship between \( X \) and \( Y \) is somewhere between these two lines.

7. Covariance is a measure of how two random variables vary together. When divided by the product of the variables' standard deviations, the result is the (product-moment) correlation, a dimensionless number that measures the strength of the linear association between the two variables. If all the points lie on a straight line with positive slope, the correlation is \(+1\); if they all lie on a straight line with negative slope, the correlation is \(-1\). A nonzero correlation between two random variables does not necessarily imply a causal relationship between them. A correlation of \(0\) implies no straight-line association — but there may nevertheless be a curvilinear association.
8. The rank correlation (computed from the ranks of the observations) measures how closely the data points fit a monotonic curve. The rank correlation is not greatly influenced by a few outlying values.

9. Multiple regression is used to obtain a prediction equation for a response variable \( Y \) from a set of predictor variables \( x_1, x_2, \ldots \). The significance of the predictor variables can be jointly tested by an \( F \)-ratio, or singly tested by \( t \)-statistics. A stepwise analysis is often used to obtain the ‘best subset’ of the \( x \)-variables with which to predict the distribution of \( Y \), but theoretically we can only be sure of reaching the best subset by examining all possible subsets. The prediction equation obtained need not reflect any causal relationship between the response and predictor variables.

10. Regression diagnostics are used to investigate the adequacy of a regression model in describing a set of data. An examination of the residuals from a model gives clues as to whether the regression model is adequate, including whether the residuals are approximately normally distributed with constant variance.

11. The square of the multiple correlation coefficient is a measure of the proportion of the variability in a response variable that can be accounted for by a set of predictor variables. The partial correlation coefficient between two variables is a measure of their linear association after allowing for the variables that have been ‘partialed out’.

12. Whenever we make subsequent measurements on study units that have been selected for follow-up because they were extreme with respect to the variable being measured, we can expect regression toward the mean (i.e. study units with high initial values will tend to have lower values later, and study units with low initial values will tend to have higher values later). Similarly, if we select predictor variables that were most significant in the analysis of a particular sample, we can expect their effects to be closer to zero and less significant in a subsequent sample.

**FURTHER READING**

PROBLEMS

1. Suppose two variables under study are temperature in degrees Fahrenheit \((y)\) and temperature in degrees Celsius \((x)\). The ‘regression line’ for this situation is

\[ y = \frac{9}{5}x + 32. \]

Assuming there is no error in observing temperature, the correlation coefficient would be expected to be

A. \(\frac{9}{5}\)
B. \(\frac{5}{9}\)
C. \(-1\)
D. \(+1\)
E. 0

2. An investigator studies 50 pairs, of unlike-sex twins and reports that the regression of female birth weight \((y)\) on male birth weight \((x)\) is given by the following equation (all weights in grams):

\[ y = 1221 + 0.403x. \]

One can conclude from this that

A. the mean weight of twin brothers of girls who weigh 1000 g is predicted to be 1624 g
B. the mean weight of twin sisters of boys who weigh 1000 g is predicted to be 1624 g
C. the sample mean weight of the girls is 1221 g
D. the sample mean weight of the boys is 1221 g
E. the sample correlation between girl’s weight and boy’s weight is 0.403

3. In a regression analysis, the residuals of a fitted model can be used to investigate all the following except

A. the model fits all but a few observations
B. the error terms are normally distributed
C. the regression function is linear
D. the robustness of the rank sum test
E. one or more predictor variables have been omitted from the model
4. Which of the following plots might represent data for variables $X$ and $Y$ that have a correlation coefficient equal to 0.82?

A. a 
B. b 
C. c 
D. d 
E. a and b

5. The correlation coefficient for the data in the graph below would be expected to have a value that is

A. a positive number of magnitude approximately equal to one  
B. a negative number of magnitude approximately equal to one  
C. a positive number of magnitude approximately equal to zero  
D. a negative number of magnitude approximately equal to zero  
E. none of the above
6. The Pearson correlation coefficient between variables $A$ and $B$ is known to be $-0.50$, and the correlation between $B$ and $C$ is known to be $0.50$. What can we infer about the relationship between $A$ and $C$?

A. There is no association between $A$ and $C$.
B. $A$ and $C$ are independent.
C. $A$ and $C$ are negatively correlated.
D. $A$ and $C$ have a linear association.
E. The relationship cannot be determined from the information given.

7. It is reported that both Pearson’s product-moment correlation and Spearman’s rank correlation between two variables are zero. This implies

A. the two variables are independent
B. the two variables tend to follow a monotonic curve
C. there is no linear or monotonic association between the two variables
D. all of the above
E. none of the above

8. A data analyst is attempting to determine the adequacy of the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

for a particular set of data. The parameters of the model are

A. least-squares estimates
B. unbiased
C. $x_1$ and $x_2$
D. robust
E. $\beta_0$, $\beta_1$, and $\beta_2$

9. A study was conducted to investigate the relationship between the estriol level of pregnant women and subsequent height of their children at birth. A scatter diagram of the data suggested a linear relationship. Pearson’s product-moment correlation coefficient was computed and found to be $r = 0.411$. The researcher decided to re-express height in inches rather than centimeters and then recompute $r$. The recalculation should yield the following value of $r$:

A. $0.000$
B. $-0.411$
C. $0.411$
10. An investigator finds the following results for a regression analysis based on the model

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon. \]

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression model</td>
<td>3</td>
<td>120</td>
<td>40</td>
<td>1.43</td>
<td>0.25</td>
</tr>
<tr>
<td>Error (residual)</td>
<td>36</td>
<td>1008</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>1128</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assuming all assumptions that were made in the analysis are justified, an appropriate conclusion is

A. the mean of the outcome variable \( Y \) is not significantly different from zero
B. the intercept term \( \beta_0 \) is significantly different from zero
C. the parameters of the model are significantly different from zero
D. the predictor variables \( x_1, x_2, \) and \( x_3 \) do not account for a significant proportion of the variability in \( Y \)
E. none of the above

11. A forward stepwise regression analysis was performed according to the following models:

Step 1: \[ Y = \alpha_0 + \alpha_1 x_1 + \varepsilon \]

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression model</td>
<td>1</td>
<td>135</td>
<td>135</td>
<td>1.35</td>
<td>0.001</td>
</tr>
<tr>
<td>Error (residual)</td>
<td>28</td>
<td>280</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 2: \[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \]
The analysis was summarized as follows:

Step 1: \( R^2 = 32.5\% \)
Step 2: \( R^2 = 47.5\% \)

Assuming all assumptions made in the analysis are justified, an appropriate conclusion is

A. \( x_2 \) accounts for a significant amount of the variability in \( Y \) over and above that accounted for by \( x_1 \)
B. neither \( x_1 \) nor \( x_2 \) accounts for a significant amount of the variability in \( Y \)
C. the proportion of variability explained by the regression model containing both \( x_1 \) and \( x_2 \) is less than should be expected in a stepwise regression analysis
D. the residual sum of squares is too large for meaningful interpretation of the regression analysis
E. the \( F \)-ratio is too small for interpretation in step 2

12. A multiple regression analysis was performed assuming the model

\[
Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.
\]

The following results were obtained:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error of Estimate</th>
<th>( t )-Test</th>
<th>( p )-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>40</td>
<td>14.25</td>
<td>2.81</td>
<td>0.005</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>5</td>
<td>2.43</td>
<td>2.06</td>
<td>0.025</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>10</td>
<td>38.46</td>
<td>0.26</td>
<td>0.600</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>7</td>
<td>2.51</td>
<td>2.79</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Assuming all assumptions made in the analysis are justified, an appropriate conclusion is

A. none of the predictor variables considered belong in the model
B. all of the predictor variables considered belong in the model
C. $x_1$ and $x_2$ belong in the model, but $x_3$ does not
D. $x_1$ and $x_3$ belong in the model, but $x_2$ does not
E. $x_2$ and $x_3$ belong in the model, but $x_1$ does not

13. An investigator finds the following results for a regression analysis of data on 50 subjects based on the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>$F$</th>
<th>$p$-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression model</td>
<td>3</td>
<td>360</td>
<td>120</td>
<td>4.29</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Error (residual)</td>
<td>46</td>
<td>1288</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
<td>1648</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assuming all assumptions that were made in the analysis are justified, an appropriate conclusion is

A. the mean of the response variable $Y$ is not significantly different from zero
B. the intercept term $\beta_0$ is significantly different from zero
C. the parameters of the model are not significantly different from zero
D. the predictor variables $x_1$, $x_2$, and $x_3$ account for a significant proportion of the variability in $Y$
E. none of the above

14. Regression diagnostics are useful in determining the

A. coefficient of variation
B. adequacy of a fitted model
C. degrees of freedom in a two-way contingency table
D. percentiles of the $t$-distribution
E. parameters of a normal distribution

15. A group of men were examined during a routine screening for elevated blood pressure. Those men with the highest blood pressure – namely those with diastolic blood pressure higher than the 80th percentile for
the group – were re-examined at a follow-up examination 2 weeks later. It was found that the mean for the re-examined men had decreased by 8 mmHg at the follow-up examination. The most likely explanation for most of the decrease is

A. the men were more relaxed for the second examination
B. some of the men became concerned and sought treatment for their blood pressure
C. observer bias
D. the observers were better trained for the second examination
E. regression toward the mean