Suppose we know that one variable is a function of the other, i.e., we know there is a causal relation. For example, we know that Y depends on X. We would like to see exactly how this dependence works, so we would be able to predict the amount that Y would change for a given amount of change in X. That is, we would like to have a nice, simple linear equation, \( Y = a + bX \). Regression analysis is a methodology for establishing such a relationship by estimating the parameters \( a \) and \( b \) from the data. Regression may also be used in some cases where causal relations are not established, and only a correlation is known. For example, if \( X \) is the score on a college entrance examination, and \( Y \) is the grade-point average for the first year of college, it would be nice to be able to predict a student's performance as a freshman, as measured by grade point, from the college entrance examination score.

In linear regression, we assume that the values of \( X \) are not random but are selected by the researcher. For each of these selected values of \( X \), there is a normal distribution of possible \( Y \) values.

**Assumptions of regression**

1. Values of the independent variable \( X \) are fixed and determined by the experimenter.
2. The variable \( X \) is measured without error (since its values are set by the experimenter), so all the error is in \( Y \).
3. For each \( X_i \), there is a subpopulation of \( Y \) values.
4. The means of the subpopulations of \( Y \) all lie on the same straight line.
5. The variances of the subpopulations of \( Y \) are all equal.
6. The \( Y \) values are independent of each other.

The first assumption is frequently ignored, since it isn't essential for fitting a regression line. However, in designing a study, it should be used. If we have control of the \( X \) values, we pick nice, evenly spaced \( X \) values over as wide a range as feasible. This will, in general, produce results that we will have more confidence in than those obtained by simply fitting a regression line to any set of \( X, Y \) values.

These assumptions are clarified by Figures 11.1 through 11.4 and the following discussion.

For each possible value of \( X \), there are many possible values of \( Y \). In fact, we assume an infinite number of normally distributed possible values of \( Y \) for each \( X \) value. For example, suppose we are trying to express body weight (\( Y \)) as a function of height (\( X \)). Let's pick one specific \( X \) value, \( X = 60 \) inches. Suppose we select a random sample of 100 people who are exactly 60 inches tall and weigh them. Will all of them have the same weight? No.
There will be a distribution of weights, $Y_i$ clustered around some mean $\bar{Y}_{X=60}$ (which is the mean of $Y$ for $X = 60$ inches), as in Figure 11.1. If the sample size is large enough, this distribution will appear normal and $\bar{Y}_{X=60}$ will be a good estimate of the true mean.

![Figure 11.1. Normally distributed subpopulation of $Y$ values for $X = 60$.](image)

We now take another $X$ value, $X = 65$ inches, and repeat the process. We now have a distribution of weight ($Y$) values of people who are 65 inches tall, with $\bar{Y}_{X=65}$, as shown in Figure 11.2.

If we take another $X$ value, $X = 70$ inches, and repeat the process, we have a third distribution, with $\bar{Y}_{X=70}$, as shown in Figure 11.3.

![Figure 11.3. Normally distributed subpopulation of $Y$ values for $X = 70$.](image)

Thus, if we summarize all these results in one figure, we see that we could draw a line through the means of all these different $Y$ values corresponding to each $X$ value and that each distribution of $Y$ values is centered around its mean, as in Figure 11.4.

![Figure 11.4. Regression line showing normally distributed subpopulation of $Y$ values for each value of $X$.](image)
Note that each subpopulation distribution has the same shape, indicating that
the amount of variability in each is equal. (Of course, if these were the
entire subpopulations the $\bar{Y}$'s should be replaced by $\mu$'s.)

We will use the following notation for expressing the regression
equation:

$$ Y = b_0 + b_1 X $$

$b_0$ = the $Y$ intercept, where the line hits the $Y$ axis

$b_1$ = the slope of the line, the amount of change in $Y$ for a unit change
in $X$

These terms are illustrated in Figure 11.5. If we assign specific values, say
$b_0 = 1$ and $b_1 = 1.5$, the equation is $Y = 1 + 1.5X$, which is shown in Figure
11.6.

![Figure 11.5. Regression line $Y = b_0 + b_1 X$, showing $b_0$ as $Y$
intercept and $b_1$ as slope.]

![Figure 11.6. Regression line $Y = 1 + 1.5X$ with $b_0 = 1$
and $b_1 = 1.5$.]

**Notation and calculations**

We use the same preliminary calculations for regression that we did for
correlation. We first calculate

- $\sum X_i$ and $\bar{X}$
- $\sum Y_i$ and $\bar{Y}$
- $\sum X_i Y_i$
- $\sum X_i^2$
- $\sum Y_i^2$
Then we calculate

\[ SS_X = \sum x_i^2 - \left( \frac{\sum x_i}{n} \right)^2 \]

\[ SS_Y = \sum y_i^2 - \left( \frac{\sum y_i}{n} \right)^2 \]

\[ SS_{XY} = \sum x_i y_i - \left( \frac{\sum x_i \sum y_i}{n} \right) \]

We now use the following notation and new calculations:

\[ Y = b_0 + b_1 X = \text{derived regression equation} \]

\[ b_0 = \text{Y intercept} \]

\[ b_1 = \text{slope of the regression line} \]

To calculate \( b_1 \) and \( b_0 \), we use

\[ b_1 = \frac{SS_{XY}}{SS_X} \]

\[ b_0 = \bar{Y} - b_1 \bar{X} \]

To predict the value of \( Y \) corresponding to a particular value of \( X \), we use

\[ Y_{p,i} = b_0 + b_1 X_i \]

where

\[ Y_{p,i} = \text{predicted value of } Y \text{ corresponding to } X_i \]

The error or deviation from the predicted value is

\[ \epsilon_i = Y_i - Y_{p,i} \]

where

\[ \epsilon_i = \text{error, deviation from regression, deviation from predicted value} \]

\[ Y_i = \text{observed value corresponding to } X_i \]

**Example**

In a study of the effect of diazepam on physiologically measured stress, the stress response was measured both before and after injection of a placebo for the control group or diazepam for the experimental group (R. Rafoth and L.J. Peterson, J. Oral Surg., 33, 189-191, 1975). The data for the control group are shown in Table III.1. We would like to be able to predict \( Y \) from \( X \).
We want to derive a linear regression equation $Y = b_0 + b_1X$ from the data.

We first calculate $\sum X_i Y_i$, $\sum X_i^2$ and $\sum Y_i^2$ (in columns D, E, and F of the table). Then we calculate $SS_X$, $SS_Y$, and $SS_{XY}$.

Table 11.1. Stress responses of control subjects before (X) and after (Y) injection with placebo.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>Y</td>
<td>$X_i Y_i$</td>
<td>$X_i^2$</td>
<td>$Y_i^2$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>44</td>
<td>30</td>
<td>44 x 30 = 1320</td>
<td>44$^2$ = 1936</td>
<td>30$^2$ = 900</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>49</td>
<td>50 x 49 = 2450</td>
<td>50$^2$ = 2500</td>
<td>49$^2$ = 2401</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>30</td>
<td>30 x 30 = 900</td>
<td>30$^2$ = 900</td>
<td>30$^2$ = 900</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>26</td>
<td>33 x 26 = 858</td>
<td>33$^2$ = 1089</td>
<td>26$^2$ = 676</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>97</td>
<td>80</td>
<td>97 x 80 = 7760</td>
<td>97$^2$ = 9409</td>
<td>80$^2$ = 6400</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>109</td>
<td>107</td>
<td>109 x 107 = 11663</td>
<td>109$^2$ = 11881</td>
<td>107$^2$ = 11449</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>53</td>
<td>46</td>
<td>53 x 46 = 2438</td>
<td>53$^2$ = 2809</td>
<td>46$^2$ = 2116</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>60</td>
<td>40</td>
<td>60 x 40 = 2400</td>
<td>60$^2$ = 3600</td>
<td>40$^2$ = 1600</td>
<td></td>
</tr>
</tbody>
</table>

\[ \sum X_i = 476 \quad \sum Y_i = 408 \quad \sum X_i Y_i = 29789 \quad \sum X_i^2 = 34124 \quad \sum Y_i^2 = 26442 \]

\[ \bar{X} = 59.5 \quad \bar{Y} = 51.0 \]

\[ SS_X = \sum X_i^2 - \left( \frac{\sum X_i}{n} \right)^2 = 34124 - \left( \frac{476}{8} \right)^2 = 34124 - 28322 = 5802 \]

\[ SS_Y = \sum Y_i^2 - \left( \frac{\sum Y_i}{n} \right)^2 = 26442 - \left( \frac{408}{8} \right)^2 = 26442 - 20808 = 5634 \]

\[ SS_{XY} = \sum X_i Y_i - \left( \frac{\sum X_i \sum Y_i}{n} \right) = 29789 - \left( \frac{476 \times 408}{8} \right) = 29789 - 24276 = 5513 \]

We now calculate the regression equation terms.

The slope is \( b_1 = \frac{SS_{XY}}{SS_X} = \frac{5513}{5802} = 0.9502 \)

The Y intercept is \( b_0 = \bar{Y} - b_1 \bar{X} = 51.0 - 0.9502(59.5) = 51.0 - 56.537 = -5.537 \)
Yes, this can happen in regression. Even though we know a negative stress response isn't possible, these equations are giving us the straight line that best fits the data points. In this case it makes sense intuitively, because the Y values tend to be lower than the X values, so when X = 0, a negative Y seems reasonable (even if we can't measure it).

Our regression equation is

\[ Y = b_0 + b_1X = -5.537 + 0.9502X \]

We can use this equation to predict the value of Y for a specific value of X, or to estimate the mean value of Y for this specific value of X. The predicted value of Y is the same as the estimated value of the mean of Y for this X; we make the distinction because the error in predicting a single value is different from the error in estimating the mean.

For example, for \( X_2 = 50 \), we measured \( Y_2 = 49 \). What is the Y value that our equation would predict for \( X_2 = 50 \)? We will use \( Y_i \) for the measured value and \( Y_{p,i} \) for the predicted value of \( X_i \).

\[
Y_{p,2} = -5.537 + 0.9502X_2 \\
= -5.537 + 0.9502(50) \\
= 41.97
\]

The deviations from regression are expressed as \( \varepsilon_i = Y_i - Y_{p,i} \). Thus, we see where the deviations from regression (the \( \varepsilon_i \)) come from. The \( Y_2 = 49 \) deviates from the value of \( Y_{p,2} \) predicted by regression by

\[ \varepsilon_2 = Y_2 - Y_{p,2} = 49 - 41.97 = 7.03 \]

Another example. For \( X_8 = 60 \), we measured \( Y_8 = 40 \). What would our equation predict for \( X_8 = 60 \)?

\[
Y_{p,8} = -5.537 + 0.9502X_8 \\
= -5.537 + 0.9502(60) \\
= 51.475
\]

The error or deviation from regression in this case is

\[ \varepsilon_8 = Y_8 - Y_{p,8} = 40 - 51.475 = -11.475 \]

For \( X_6 = 109 \), we measured \( Y_6 = 107 \).

\[
Y_{p,6} = -5.537 + 0.9502X_6 \\
= -5.537 + 0.9502(109) \\
= 98.035 \\
\varepsilon_6 = Y_6 - Y_{p,6} = 107 - 98.035 = 8.965
\]
Once we have derived the regression equation, we generally want to draw the line corresponding to the equation, giving a graphic representation of the relationship. To draw the regression line, we first note that since
\[ b_0 = \bar{Y} - b_1 \bar{X}, \]
the line must pass through the point \((\bar{Y}, \bar{X})\). We also know that the regression line must pass through the Y intercept, i.e., the point \((Y = -5.537, X = 0)\). Now we can draw a straight line through these two points and we have a graphic display of our derived regression line, as shown in Figure 11.7. Two of the deviations from regression, \(\varepsilon_1\) and \(\varepsilon_6\), are shown in this figure.

The derived regression line is the straight line that best fits the data based on the criterion that the sum of squares of all the deviations \((\varepsilon_i = Y_i - \hat{Y}_i)\) of observed values from the predicted values on the line is a minimum. This is called the method of least squares, or least squares regression.
Of course, once we have derived the regression equation, we can use it to predict values of \( Y \) for other values of \( X \) not in the original data, with two precautions: (1) The error in the prediction increases the further we get from the mean, so predicted values of \( Y \) for values of \( X \) near \( \bar{X} \) will be more accurate; (2) Since the lowest \( X \) value observed was 30 and the highest \( X \) value observed was 109, we can only feel confident about predicting \( Y \) for values of \( X \) between and including 30 and 109. Extrapolating beyond this interval is dangerous and should not be done, since it is based on the assumption that the relationship continues to be linear outside this interval. (Yes, estimating \( b_0 \), the \( Y \) intercept, is equivalent to extrapolating beyond this interval, but this one exception is acceptable because the equation in terms of \( b_0 \) is the most convenient version to use.)

**Testing hypotheses about regression**

There are three kinds of hypothesis tests we can make in linear regression: (1) We can test hypotheses about \( \beta_0 \), the true intercept for the population of \( X, Y \) values, where \( \hat{b}_0 \) is our estimate of \( \beta_0 \); (2) We can test hypotheses about \( \beta_1 \), the true slope for the population of \( X, Y \) values, where \( \hat{b}_1 \) is our estimate of \( \beta_1 \); (3) We can test the hypothesis that \( X \) and \( Y \) are linearly related. We use the \( t \) test for the first two and the \( F \) test (explained later) for the third one.

**A. Testing hypotheses about \( \beta_0 \), the \( Y \) intercept**

1. Specify \( H_0: \beta_0 = 0 \)
   
   \( H_0: \beta_0 \neq 0 \) 2 tailed test
   
   or
   
   \( H_0: \beta_0 > 0 \) 1 tailed test
   
   \( H_0: \beta_0 < 0 \)

2. Determine degrees of freedom. In regression, \( df = n - 2 \).
   
   Here, \( n = 8 \), so \( df = 8 - 2 = 6 \).

3. Select significance level. Let \( \alpha = .05 \).

4. Determine critical value of \( t \).
   
   For \( df = 6 \), \( \alpha = .05 \), 2 tailed test, \( t_{CR} = \pm 2.447 \)

5. Calculate test statistic. This is done in several steps.

   a. Calculate the sum of squares for error, \( SSE \)

   \[
   SSE = SS_y - \hat{b}_1 SS_{xy}
   = 5634 - 0.9502(5513)
   = 5634 - 5238.4526
   = 395.5474
   \]

   (Don't round off here, since this is an intermediate value to be used in further calculations.)
b. Calculate the variance, $s^2$

$$s^2 = \frac{SSE}{n-2} = \frac{395.5474}{6} = 65.9246$$

c. Calculate the standard error of $b_0$, $s_{b_0}$

$$s_{b_0} = s^2 \left( \frac{1 + \left( \frac{\bar{X}^2}{SS_X} \right)}{n} \right)$$

$$= 65.92 \left( \frac{1 + \left( \frac{59.5}{5802} \right)^2}{8} \right)$$

$$= 65.92 \left[ 0.125 + 0.610 \right]$$

$$= 48.4512$$

Then the standard error of $b_0$ is

$$s_{b_0} = \sqrt{s_{b_0}^2} = 6.9607$$

d. Calculate $t$,

$$t = \frac{b_0 - \beta_0}{s_{b_0}}$$

$$= \frac{-5.537 - 0}{6.9607}$$

(since $\beta_0 = 0$ by $H_0$)

$$= -0.795$$

6. Compare calculated and critical values

$$t_{calc} = -0.795$$

$$t_{cr} = \pm 2.447$$

7. Conclusion. Accept $H_0$: $\beta_0 = 0$, since $t_{calc} = -0.795$ falls in the region of acceptance. Thus, the intercept is not significantly different from zero.

B. Testing hypotheses about $\beta_1$, the slope

First 4 steps are same as above. Let's use

$H_0$: $\beta_1 = 0$ (says slope is 0, or $X$ and $Y$ are unrelated)

$H_a$: $\beta_1 \neq 0$ (says slope is not 0, $X$ and $Y$ are related, 2 tailed)

$\alpha = .05$, df = 6, $t_{cr} = \pm 2.447$. 
5. Calculate the test statistic. Again, it's done in several steps.
   a. Calculate SSE. (This one is same as above.)
      
      \[ \text{SSE} = SS_Y - b_1SS_{XY} = 395.5474 \]
   
   b. Calculate \( S^2 \). (This is also same as above.)
      
      \[ S^2 = \frac{\text{SSE}}{n-2} = 65.9246 \]
   
   c. Calculate standard error of \( b_1 \). (Different than above, also easier.)
      
      \[ S_{b_1} = \frac{S^2}{SS_X} = \frac{65.9246}{5802} = 0.01136 \]
      
      Then
      
      \[ S_{b_1} = \sqrt{S_{b_1}^2} = 0.1066 \]
   
   d. Calculate \( t \)
      
      \[ t = \frac{b_1 - \beta_1}{S_{b_1}} \]
      
      \[ = \frac{0.9502 - 0}{0.1066} \quad \text{(since } H_0: \beta_1 = 0) \]
      
      \[ = 8.91 \]

6. Compare calculated and critical values

   \( t_{\text{calc}} = 8.91 \)
   
   \( t_{\text{cr}} = \pm 2.447 \)

7. Conclusion. Reject \( H_0: \beta_1 = 0 \), since 8.91 > 2.447. Thus, the slope is significantly different from zero, implying that \( X \) and \( Y \) are related.

C. Testing hypotheses about the linear relation between \( X \) and \( Y \)

We now want to consider a somewhat different way of testing hypotheses that \( X \) and \( Y \) are linearly related. We want to consider our regression equation/line in terms of how much of the variability it can explain, or how closely the line fits the observed set of \( Y \) values. In regression, we're assuming all the variability is in \( Y \). This variability
can be partitioned into different sources, as follows. First, we calculate $\overline{Y}$ and then calculate the total sum of squares of deviations about the mean (which is just $SS_y$),

$$SS_y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \frac{n}{n} \sum_{i=1}^{n} Y_i - \left( \frac{\sum_{i=1}^{n} Y_i}{n} \right)^2$$

Thus, $SS_y$ accounts for all the variability in the $Y$ variable. We now partition this variability into 2 components -- the amount that can be explained or accounted for by our derived regression equation and the amount that cannot be explained by the regression equation. Remember, in correlation, $X$ and $Y$ were said to be perfectly correlated if all the data points were on the straight line. If this situation were true in regression, intuitively we see that the regression equation would explain all the variability in $Y$, since it could predict every value of $Y$ from $X$ values. However, when some observed $Y$ values are not on the line, we say these are deviations from the values predicted by regression, or that this amount of variability is unexplained by regression. We thus break each deviation $Y_i - \overline{Y}$ into 2 other deviations as shown in Figure 11.8. First, $Y_i - Y_{p,i}$ is the deviation of the observed $Y_i$ value corresponding to $X_i$ from the value $Y_{p,i}$ predicted by our regression equation. Second, each $Y$ value on the regression line is a predicted or explained value, whose deviation from the mean $\overline{Y}$ can be expressed as $Y_{p,i} - \overline{Y}$. Thus, we have expressed the total deviation as

$$Y_i - \overline{Y} = (Y_i - Y_{p,i}) + (Y_{p,i} - \overline{Y})$$

![Figure 11.8. Illustration of deviations due to regression ($Y_{p,i} - \overline{Y}$) and unexplained deviations ($Y_i - Y_{p,i}$).]
If we square each of these deviations and add them up for all values of \( Y \), we get

\[
\frac{\sum_{i=1}^{n} (Y_{i} - \bar{Y})^2}{n} = \frac{\sum_{i=1}^{n} (Y_{i} - Y_{p,i})^2}{n} + \frac{\sum_{i=1}^{n} (Y_{p,i} - \bar{Y})^2}{n}
\]

where \( \frac{\sum_{i=1}^{n} (Y_{i} - \bar{Y})^2}{n} \) = total sum of squares = \( SS_Y \)

\( \frac{\sum_{i=1}^{n} (Y_{i} - Y_{p,i})^2}{n} \) = unexplained or residual sum of squares = \( SSE \)

\( \frac{\sum_{i=1}^{n} (Y_{p,i} - \bar{Y})^2}{n} \) = explained or regression sum of squares = \( SS_{reg} \)

\( SS_Y = SSE + SS_{reg} \)

If all the data points were very close to the regression line, it is obvious that \( SS_{reg} \) would be relatively large, while \( SSE \) would be relatively small, i.e., each \( Y_{i} - Y_{p,i} \) would be small since each \( Y_{i} \) would be near its predicted value on the line, so most of the variability in \( Y \) would be accounted for by the line itself, or \( SS_{reg} \). Thus the ratio of \( SS_{reg} \) to \( SSE \) is given a special name,

\( r^2 = \frac{SS_{reg}}{SSE} = \) coefficient of determination.

If we had calculated the correlation coefficient \( r \), we could square it to get the coefficient of determination, since we're talking about the same \( r \). In other words, the higher the correlation, the larger \( r^2 \) will be, and thus, the greater the amount of variability accounted for by the regression equation.

After all this digression, back to the hypothesis test. We divide each of these 3 sums of squares by its appropriate degrees of freedom to form a variance. Thus, the variance for all the data is

\[
S_Y^2 = \frac{SS_Y}{n-1} = \frac{\sum_{i=1}^{n} (Y_{i} - \bar{Y})^2}{n-1}
\]

The error variance or residual variance is

\[
S^2 = \frac{SSE}{n-2}
\]

The variance due to regression is

\[
S_{reg}^2 = \frac{SS_{reg}}{1}
\]
There is only 1 degree of freedom due to the regression.

We use two of these variances, which are also called mean squares, as the basis of our test of whether X and Y are linearly related. The data are summarized as in Table 11.3.

Table 11.3. Summary table for regression.

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source of</td>
<td>Sum of</td>
<td>Degrees of</td>
<td>Mean Square</td>
<td></td>
</tr>
<tr>
<td>Variation</td>
<td>Squares</td>
<td>Freedom</td>
<td>(Variance)</td>
<td>F</td>
</tr>
<tr>
<td>Linear regress</td>
<td>SSreg</td>
<td>1</td>
<td>MSreg = $\frac{SS_{reg}}{1}$</td>
<td>$F = \frac{MS_{reg}}{MS_{res}}$</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>SSE</td>
<td>n-2</td>
<td>MSres = $\frac{SSE}{n-2}$</td>
<td>with df(_1)=1, df(_2)=n-2</td>
</tr>
<tr>
<td>Total</td>
<td>SS(_y)</td>
<td>n-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Column I lists the 3 sources of variation, column II lists their respective sums of squares, column III lists their respective degrees of freedom, and column IV shows the mean squares (variances) for the regression and residual (error) terms. Finally, in column V, a ratio of these two variances is formed. This ratio is given a special name— the F test (named after Sir R.A. Fisher, who developed it).

$$F = \frac{MS_{reg}}{MS_{res}} = \frac{\text{variability explained by regression}}{\text{variability unexplained by regression}}$$

From inspection of the F ratio, it is obvious that the more variability accounted for by regression, the larger the calculated value of F, the test statistic. Thus, $H_0$ will say X and Y are not linearly related. If we get a calculated F larger than the critical value of F, we reject $H_0$. Since each variance in F has a different number of degrees of freedom, we have to use both in determining the critical value of F.

Now let's go back and use our example to do this third hypothesis test. However, instead of using the definition formulas on page 8, we use the equivalent versions below:

$$SS_{Y} = \sum Y_i^2 - (\bar{Y})^2_n = \text{total sum of squares}$$

$$SSE = SS_{Y} - b_1SS_{XY} = \text{residual (error) sum of squares}$$

$$SS_{reg} = b_1SS_{xy} = \text{regression sum of squares}$$
We calculated SS\(_Y\) = 5634, \(b_1 = 0.9502\), and SS\(_{XY}\) = 5513. So we now have

\[
\begin{align*}
SS_Y &= 5634 \\
SSE &= SS_Y - b_1 SS_{XY} = 5634 - 0.9502(5513) = 395.5474 \\
SS_{reg} &= b_1 SS_{XY} = 0.9502(5513) = 5238.4526
\end{align*}
\]

We enter the appropriate values in the summary table, Table 11.4, and calculate \(F\).

Table 11.4. Summary table for regression example.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square (Variance)</th>
<th>(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regress</td>
<td>SS(_{reg}) = 5238.4526</td>
<td>1</td>
<td>MS(<em>{reg}) = \frac{SS</em>{reg}}{1}</td>
<td></td>
</tr>
<tr>
<td>Residual (error)</td>
<td>SSE = 395.5474</td>
<td>n-2=6</td>
<td>MS(<em>{res}) = \frac{SS</em>{res}}{n-2} = \frac{395.5474}{6}</td>
<td>\frac{MS_{reg}}{MS_{res}} = \frac{5238.4526}{65.9246} = 79.46</td>
</tr>
<tr>
<td>Total</td>
<td>SS(_Y) = 5634</td>
<td>n-1=7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The critical value of \(F\) for \(df_1 = 1\) and \(df_2 = 6\) at the 0.05 significance level is 5.99, and our calculated value of 79.46 is much larger, so we reject \(H_0\) and conclude that \(X\) and \(Y\) are linearly related. We will defer the discussion of how to look up critical values of \(F\) until we consider analysis of variance.

Several measurements of \(Y\) for each \(X\)

Suppose that for \(X = 60\) inches, we have recorded the weights of several people. For \(X = 65\) inches, we have recorded the weights of several more people, for \(X = 70\) inches, we have recorded several more weights, etc. Thus, we have several \(Y_i\) values for each \(X_i\). Is it better to use all the \(Y_i\) values or to just use the means of the \(Y\) values corresponding to each \(X_i\) in fitting the regression line? For example, if the \(Y\) values (weights) corresponding to \(X_i = 60\) inches are 140 lb., 100 lb., and 120 lb., do we use all 3 of these or do we use \(\overline{Y_i} = 120\) lb.? It is perfectly valid to do either. In general, it is best to use all the \(Y_i\) values since the degrees of freedom will be larger, which will result in smaller critical values in hypothesis testing, so you will be more likely to detect a significant difference. However, for a quick and dirty estimate, the calculations are much easier using the means of the \(Y_i\) corresponding to each \(X_i\). Just calculate the \(\overline{Y_i}\) for each \(X_i\) and treat these as the \(Y\) values in all the formulas.