11. MULTIPLE REGRESSION

Engineers operate at the interface between science and society.

– Dean Gordon Brown

Up to this point, we’ve been looking at regression models in which a single regressor variable is used to predict the response variable. Often, the response of a system depends on more than just one variable. For these systems, we can postulate a multiple-regression model

\[ y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \ldots + \hat{a}_k x_k + e \]

where each partial regression coefficient \( \hat{a}_i \) \( (i \neq 0) \) describes the expected change in the response variable \( y \) brought about by a unit change in the regressor variable \( x_i \) while all the other regressor variables are held constant. Of course this can only be achieved if the \( k \) regressor variables are independent of one another (i.e., the value of one variable is not influenced by the values of the others). This is one of the fundamental underlying assumptions of multiple regression.

As in our simple regression model, \( e \) is a normally-distributed random error with a mean value of zero and an unknown but constant variance \( \sigma^2 \). Our goal is still to find the regression model that minimizes \( e \) for the entire sample set and we accomplish that by deriving an equation for the sum of the squares of the residuals:

\[ S_r = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} \left( y_i - \hat{a}_0 - \hat{a}_1 x_{1i} - \hat{a}_2 x_{2i} - \ldots - \hat{a}_k x_{ki} \right)^2 \]

then setting the partial derivatives of the equation to zero:

\[ \frac{\partial S_r}{\partial \hat{a}_0} = 0, \quad \frac{\partial S_r}{\partial \hat{a}_1} = 0, \quad \frac{\partial S_r}{\partial \hat{a}_2} = 0, \quad \ldots, \quad \frac{\partial S_r}{\partial \hat{a}_k} = 0 \]

and solving the resulting system of simultaneous linear equations for \( \hat{a}_0, \hat{a}_1, \ldots, \hat{a}_n \).

With more than two variables, it becomes quite cumbersome to calculate the regression model coefficients by hand. Therefore, we’ll use Excel to do all of our calculations and concentrate on understanding the results that Excel provides. This is important because we can’t plot the data and visually assess the fit of the regression model as we did in simple regression. Instead, we have to rely on our understanding of various regression statistics (such as \( r^2 \) and \( s_{yi} \)) to ascertain how well our model describes the behavior of the system.

Example

A local bottling company wants to fine-tune the routing and scheduling of its soda-machine route salesmen in order to minimize its labor costs. To do this, the company needs to know how long the route salesman (driver) will remain at each site. Ignoring, for now, the driving time between soda machine installations, we can postulate that two things determine how long it takes the driver to restock the machines at a given location. The first is the distance the driver must walk to reach the vending machines, which we’ll express in units of feet. The second is the amount of soda needed to restock the machines, which we’ll express in terms of cases of soda.
We can safely assume that the number of sodas needed to restock the machines has nothing to do with the distance the driver must walk to reach the machines, so we can consider cases of soda \( (x_1) \) and distance \( (x_2) \) to be independent regressor variables. If there was some sort of correlation between the two, we could not use multiple regression to model the delivery time.

Below are 24 delivery time observations that were obtained by randomly selecting 24 different vending machine locations around the city:

<table>
<thead>
<tr>
<th>Vending Location Number</th>
<th>Delivery Time (min)</th>
<th>Number of Cases Stocked</th>
<th>Walking Distance (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>8</td>
<td>110</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>11</td>
<td>120</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>10</td>
<td>550</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>8</td>
<td>295</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>4</td>
<td>200</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>2</td>
<td>375</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>24</td>
<td>9</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>27</td>
<td>8</td>
<td>300</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>4</td>
<td>410</td>
</tr>
<tr>
<td>12</td>
<td>37</td>
<td>11</td>
<td>400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vending Location Number</th>
<th>Delivery Time (min)</th>
<th>Number of Cases Stocked</th>
<th>Walking Distance (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>42</td>
<td>12</td>
<td>500</td>
</tr>
<tr>
<td>14</td>
<td>12</td>
<td>2</td>
<td>360</td>
</tr>
<tr>
<td>15</td>
<td>18</td>
<td>4</td>
<td>400</td>
</tr>
<tr>
<td>16</td>
<td>69</td>
<td>20</td>
<td>600</td>
</tr>
<tr>
<td>17</td>
<td>10</td>
<td>1</td>
<td>585</td>
</tr>
<tr>
<td>18</td>
<td>35</td>
<td>10</td>
<td>540</td>
</tr>
<tr>
<td>19</td>
<td>47</td>
<td>15</td>
<td>250</td>
</tr>
<tr>
<td>20</td>
<td>45</td>
<td>15</td>
<td>290</td>
</tr>
<tr>
<td>21</td>
<td>54</td>
<td>16</td>
<td>510</td>
</tr>
<tr>
<td>22</td>
<td>57</td>
<td>17</td>
<td>590</td>
</tr>
<tr>
<td>23</td>
<td>22</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>24</td>
<td>21</td>
<td>5</td>
<td>400</td>
</tr>
</tbody>
</table>

The SLOPE and INTERCEPT functions in Excel can only be used for simple regression models (i.e., those with a single regressor variable). For multiple regression, we have to use a function called LINEST. For now, we’ll use the function in its simplest form:

\[
\text{LINEST(known}_y\text{'s, known}_x\text{'s, , )}
\]

where known_y’s is the block of cells containing the response data and known_x’s is the block of cells containing the regressor data. We’ll discuss the purpose of the two empty fields later.

This is known as an array function because it returns an array of data rather than a single value. The array of data contains the regression model intercept and the \( k \) partial regression coefficients. This means you have to simultaneously enter the function in \( k+1 \) cells. To do this, select \( k+1 \) adjacent cells in a row (they cannot be in a column), type the function in the formula bar at the top of the spreadsheet, and simultaneously hit [Ctrl], [Shift], and [Enter] to enter the formula into the \( k+1 \) spreadsheet cells.
For our soft-drink problem, the result looks like this:

\[
\begin{array}{ccc}
\hat{a}_2 & \hat{a}_1 & \hat{a}_0 \\
0.013 & 2.785 & 1.487 \\
\end{array}
\]

Note that the order of the coefficients is exactly the reverse of their order in the model:

\[
y = 1.487 + 2.785 x_1 + 0.013 x_2 + e
\]

Now that we have our model, we can estimate \(\hat{y}\) and compute a residual for each \((x_1, x_2)\) pair in the table. From these we can determine the measures of variation \(SSR\), \(SSE\), and \(SST\):

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Number} & \text{\(y_i - \bar{y}\)} & \text{\(y_i - \hat{y}_i\)} & \text{\(\hat{y}_i - \bar{y}\)} \\
\hline
1 & -19 & -2 & -21 \\
2 & -5 & 2 & -3 \\
3 & 3 & 2 & 5 \\
4 & 6 & 2 & 8 \\
5 & -4 & 3 & -1 \\
6 & -12 & -1 & -13 \\
7 & -15 & -2 & -17 \\
8 & -19 & -2 & -21 \\
9 & -5 & 4 & -1 \\
10 & -2 & 1 & -1 \\
11 & -12 & 1 & -11 \\
12 & 8 & 0 & 8 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Number} & \text{\(y_i - \bar{y}\)} & \text{\(y_i - \hat{y}_i\)} & \text{\(\hat{y}_i - \bar{y}\)} \\
\hline
13 & 13 & -1 & 12 \\
14 & -17 & 0 & -17 \\
15 & -11 & 0 & -11 \\
16 & 40 & -4 & 36 \\
17 & -19 & 2 & -17 \\
18 & 6 & 1 & 7 \\
19 & 18 & 0 & 18 \\
20 & 16 & 2 & 18 \\
21 & 25 & -1 & 24 \\
22 & 28 & -1 & 27 \\
23 & -7 & -2 & -9 \\
24 & -8 & 0 & -8 \\
\hline
\end{array}
\]

\[
\begin{align*}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = 6089 \\
SSR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 6007 \\
SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 82
\end{align*}
\]
From these values, we can compute the coefficient of multiple determination

\[ R^2 = \frac{SSR}{SST} = \]

and the standard error of the \( y \) estimate

\[ s_{yx} = \sqrt{\frac{SSE}{n-k-1}} = \]

Note that \( SSE \) has \( n – 3 \) degrees of freedom because there are three model coefficients (including the intercept) that are derived from the data. Note, too, that we use the upper case \( R \) to distinguish the coefficient of multiple determination from the coefficient of determination, even though they share the same definition.

The addition of more regressor variables decreases the number of degrees of freedom of the \( SSE \). This makes it difficult to fairly compare two or more multiple regression models that predict the same response using different numbers of regressor variables. To “level the playing field,” some researchers advocate using an adjusted \( R^2 \):

\[ R^2_{adj} = 1 - \left[ (1 - R^2) \frac{n-1}{n-k-1} \right] = \]

**Residual Analysis**

As with simple regression, we really should look at the residuals to see if there are any patterns that suggest a lack of normality, homoscedasticity or independence. To check for normality, we can either plot a histogram of the residuals:

The plot suggests that the residuals are normally distributed about the regression line.

To check for independence, we can plot the residuals in the same order we collected the data to see if there is any correlation between adjacent residuals:
There doesn’t appear to be any correlation between adjacent observations, so we can assume that the observations are independent of one another.

To check for a lack of homoscedasticity and randomness, we should plot the model residuals as a function of each regressor variable and of the models predictions themselves.

Starting with the distance from the truck to the vending machines and back,

we see that the variability of the data is fairly constant over all values of \( x_2 \) (which is the definition of homoscedasticity). There doesn’t appear to be any trends in the residuals that might suggest a nonlinear relationship between delivery time and the delivery distance, so we should be pretty happy with our linear function of delivery distance.

Next, we’ll plot the model residuals as a function of the delivery volume:
Here we see that the largest underestimation corresponds to the largest number of cases, even though the model doesn’t consistently underestimate delivery times when the number of cases is large. This might suggest that there is something different about that one observation. Perhaps the driver stopped to use the restroom. Maybe the driver couldn’t fit 20 cases of soda on his hand truck and had to make a second trip. Since our current model assumes a single trip to the vending machines, it would naturally underestimate the delivery time for multiple trips.

Finally, we’ll plot the residuals as a function of the predicted delivery time itself:

Here, the same data point stands out noticeably from the rest. Note, however, that we know

\[ s_{y|x} = 1.98 \approx 2 \text{ minutes} \]

so 97.73% of the residuals will fall within a range of \([-4,+4]\) minutes. This data point may stand out visually but, statistically, it is not particularly out of line with the rest of the residuals. In fact, the worst overestimation of delivery time is also 4 minutes. So this may be a valid data point that stands out simply because there are so few data points in its immediate vicinity.
Hypothesis Testing

Once we’ve completed the residual analysis and convinced ourselves that we haven’t violated any of the underlying assumptions of linear regression, we can determine whether or not there is a significant relationship between the dependent variable and the regressor variables. We’ll do this using the same $F$-test we used for simple linear regression.

The null and alternative hypotheses are as follows:

- $H_0: a_1 = a_2 = 0$ (There is no linear relationship)
- $H_a: \text{At least one } a_i \neq 0$ (There is at least one linear relationship)

The appropriate test statistic is:

\[ T.S.: \quad F = \frac{MSR}{MSE} \]

which should follow an $F$ distribution with $k$ and $n - k - 1$ degrees of freedom. We’ll have to reject the null hypothesis if the ratio of the variances is large, so this is a one-tailed test and the rejection criterion is given by

\[ R.R.: \quad F > F_{\alpha,k,n-k-1} \]

We have all of the data we need to compute our test statistic:
We can also use a *t*-test to independently test the significance of each regressor variable $x_i$ for predicting the response variable $y$. In this case, we are actually testing for the significance of *adding* one particular variable to the regression model given that the other variables have already been included in the model. In other words, we’re testing for the *contribution* of each regressor variable to the overall prediction.

**Example**

For example, to test the significance of delivery volume ($x_1$) for predicting delivery time *given that delivery distance has already been accounted for*, our null and alternate hypotheses are

$$
H_0: \quad a_1 = 0 \quad \text{(delivery volume does not contribute anything to the regression)}
$$

$$
H_a: \quad a_1 \neq 0 \quad \text{(delivery volume does contribute something to the regression)}
$$

The appropriate test statistic is:

$$
T.S.: \quad t = \frac{\hat{a}_1}{s_{\hat{y}|x}/\sqrt{SSX_1}} \quad \text{where} \quad SSX_1 = \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2
$$

This test statistic should follow a *t* distribution with $n - k - 1$ degrees of freedom. We’ll have to reject the null hypothesis if $\hat{a}_1$ is appreciably greater than or less than zero, so this is a two-tailed test and the rejection criterion is given by

$$
R.R.: \quad |t| > t_{\alpha/2,n-k-1}
$$

From our earlier regression analysis (and a quick calculation of the variance of the $x_1$ data):

$$
\hat{a}_1 = 2.785 \quad s_{\hat{y}|x} = 1.98 \quad SSX_1 = 680
$$
To test the significance of delivery distance ($x_2$) for predicting delivery time given that delivery volume has already been accounted for, our null and alternate hypotheses are:

$$H_0: \ a_2 = 0 \quad \text{(delivery distance does not contribute anything to the regression)}$$

$$H_a: \ a_2 \neq 0 \quad \text{(delivery distance does contribute something to the regression)}$$

The appropriate test statistic is:

$$T.S.: \quad t = \frac{\hat{a}_2}{\frac{\hat{\sigma}_{y|x}}{\sqrt{SSX_2}}} \quad \text{where} \quad SSX_2 = \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2$$

which should also follow a $t$ distribution with $n - k - 1$ degrees of freedom:

$$R.R.: \quad |t| > t_{\alpha/2, n-k-1}$$

From our earlier regression analysis (and a quick calculation of the variance of the $x_2$ data):

$$\hat{a}_2 = 0.013 \quad \hat{\sigma}_{y|x} = 1.98 \quad SSX_2 = 764,341$$
A Closer Look at LINEST

Now we can take a closer look at the LINEST function in Excel. We said earlier that, in its simplest form, it was an array function that returns a row of numbers that represent the regression model coefficients. In reality, LINEST returns an array of numbers that covers nearly everything we’ve examined here. To get these additional statistics, two things must happen:

1. You must simultaneously enter the LINEST function into a block of cells 5 rows deep and \( k + 1 \) columns wide (where \( k \) is the number of regressor variables)

2. You must make the last argument in the LINEST argument list TRUE (which tells LINEST to return the statistics along with the regression coefficients)

\[
\text{LINEST(known_y’s, known_x’s, \text{TRUE})}
\]

The organization of the block of cells returned by LINEST is as follows:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \hat{a}_k )</td>
<td>( \hat{a}_{k-1} )</td>
<td>( \ldots )</td>
<td>( \hat{a}_2 )</td>
<td>( \hat{a}_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( s_k )</td>
<td>( s_{k-1} )</td>
<td>( \ldots )</td>
<td>( s_2 )</td>
<td>( s_1 )</td>
</tr>
<tr>
<td>3</td>
<td>( r^2 )</td>
<td>( s_{y</td>
<td>x} )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>4</td>
<td>( F )</td>
<td>( d_f )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>5</td>
<td>( SSR )</td>
<td>( SSE )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
</tbody>
</table>

The first row contains the regression model coefficients, as before. The second row contains the standard error of each regression coefficient:

\[
s_i = \frac{s_{y|x}}{\sqrt{SSX_i}}
\]

We haven’t mentioned these by name, but we’ve used them to compute the \( t \)-test statistics for determining the significance of the contribution of each regressor:

\[
t = \frac{\hat{a}_i}{s_{y|x}/\sqrt{SSX_i}} = \frac{\hat{a}_i}{s_i}
\]

Even the intercept, \( \hat{a}_0 \), has a standard error associated with it. This standard error can be used to construct a confidence interval about the intercept

\[
\text{C.I.} = \hat{a}_0 \pm \left( t_{\alpha/2,n-k-1} \right) (s_0)
\]

or it can be used to test whether the intercept is really equal to zero:
If we fail to reject the null hypothesis, the regression model may really be

\[ y = \hat{a}_1 x_1 + \hat{a}_2 x_2 + \ldots + \hat{a}_k x_k + e \]

The third row contains just two values: the coefficient of determination, \( r^2 \), and the standard error of the \( y \) estimate, \( s_{y|x} \).

The fourth row also contains just two values: the value of the \( F \)-test statistic used to test the overall significance of the regression and the number of degrees of freedom \( n - k - 1 \) associated with the residuals.

The last row contains the regression sum of squares, \( SSR \), and the error sum of squares, \( SSE \), from which you can calculate the total sum of squares as

\[ SST = SSR + SSE \]

These are the variance measures used to compute the statistics \( r^2 \), \( s_{y|x} \), \( F \), and others.

**Example**

For our soda delivery problem, the **LINEST** function returns the following array of values:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.013</td>
<td>2.785</td>
<td>1.487</td>
</tr>
<tr>
<td>2</td>
<td>0.00243</td>
<td>0.08145</td>
<td>0.94563</td>
</tr>
<tr>
<td>3</td>
<td>0.98651</td>
<td>1.97797</td>
<td>#N/A</td>
</tr>
<tr>
<td>4</td>
<td>767.714</td>
<td>21</td>
<td>#N/A</td>
</tr>
<tr>
<td>5</td>
<td>6007.17</td>
<td>82.16</td>
<td>#N/A</td>
</tr>
</tbody>
</table>

Without looking back at the calculations we’ve already done, we can determine from these cells alone that the regression model is given by

\[ y = 1.487 + 2.785 x_1 + 0.013 x_2 + e \]

The overall \( r^2 \) is a very good 0.98651, which means our model explains 98.65% of the variation in the delivery times from one site to the next.
The standard error of the $y$ estimate is 1.97797, which tells us that the model residuals are scattered about the regression line with a standard deviation of just under 2 minutes. We can also infer from this statistic that 97.73% of the residuals fall within the range $[-4,+4]$ if they are approximately normally distributed.

The $F$-test statistic on the overall significance of the regression is 767.714. To determine if we should reject the null hypothesis that there is no statistically significant relationship between the delivery time and the two regressor variables (delivery volume and delivery distance) we must determine if

$$F > F_{\alpha,2,21}$$

where $\alpha$ is the desired significance level, 2 is the number of degrees of freedom of SSR, and 21 is the number of degrees of freedom of SSE. If we chose a 0.05 level of significance,

$$F_{\alpha,2,21} = 3.47$$

so we reject the null hypothesis and conclude that there is a statistically significant relationship between the delivery time, delivery volume, and delivery distance.

To determine which of the regressor variables contributes significantly to the regression, we can perform a $t$-test on the slopes. The mean value of $\hat{a}_1$ is 2.785 and its standard error is 0.08145, so the $t$-test statistic for the delivery volume is

$$t = \frac{\hat{a}_1}{s_1} = \frac{2.785}{0.08145} = 34.193$$

Likewise, the mean value of $\hat{a}_2$ is 0.013 and its standard error is 0.00243, so the $t$-test statistic for the delivery distance is

$$t = \frac{\hat{a}_2}{s_2} = \frac{0.013}{0.00243} = 5.35$$

Both should also follow a $t$ distribution with 21 degrees of freedom so, at the 0.05 significance level, both should be greater than

$$t_{0.025,21} = 2.08$$

if we are to reject the hypotheses that one or the other contributes nothing to the regression. In this case, we would reject both hypotheses.

To determine if the intercept is actually zero, we can perform a $t$-test on the intercept. The mean value of $\hat{a}_0$ is 1.487 and its standard error is 0.94563, so the $t$-test statistic for the intercept equal to zero is

$$t = \frac{\hat{a}_0}{s_0} = \frac{1.487}{0.94563} = 1.572$$
This, too, should follow a $t$ distribution with 21 degrees of freedom so, at the 0.05 significance level, we fail to reject the null hypothesis (that the intercept is equal to zero) and conclude that

There is insufficient evidence to reject the hypothesis that the intercept is zero.

Note that we haven’t accepted anything. Recall that

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$$

$$\beta = P(\text{type II error}) = P(\text{accept } H_0 | H_0 \text{ is false})$$

Recall, too, that the power of the test is given by

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 | H_0 \text{ is false})$$

Since we don’t know the power of our test (we know $\alpha$ but not $\beta$) we can’t definitively conclude that the intercept is zero, we can only conclude that there’s not enough evidence to show that it’s not zero.

More on Excel’s Regression Tool

If we model the soda delivery times using the regression tool from Excel’s Data Analysis Toolpak, we get the results shown below:
By now, you should recognize the “Multiple R” and “Adjusted R Square” entries. The former is the coefficient of multiple determination $R^2$ and the latter is the adjusted coefficient $R_{adj}^2$.

The ANOVA table located immediately below the regression statistics shows $SSR$, $SST$, and $SSE$ in the column denoted “SS” and their corresponding degrees of freedom in the column with the heading “df”. Dividing $SSR$ and $SSE$ by their respective degrees of freedom provides the values of $MSR$ and $MSE$ shown in the column marked “MS”. The ratio of $MSR$ to $MSE$ is the $F$-test statistic shown in the next column to the right. That $F$-test statistic corresponds to the $P$-value shown as “Significance F”. As before, the $P$-value represents the probability that there is no linear relationship between the response variable and the $k$ regressor variables.

The table immediately below the ANOVA table shows the regression coefficients $\hat{a}_j$ along with their associated standard errors,

$$s_j = \frac{s_{y|x}}{\sqrt{SSX_i}}$$

and the $t$-test statistics needed to determine the significance of each variable’s contribution to the overall regression:

$$t = \frac{\hat{a}_i}{s_{y|x}} = \frac{\hat{a}_i}{s_i}$$

The final columns of that table show the 95% confidence limits for those regression coefficients:

$$C.I. = \hat{a}_i \pm s_i t_{0.025, n-k-1}$$

which are interpreted as follows: In repeated sampling from the population, 95% of the regression coefficients will fall within the limits shown.

Collinearity

At the start of our soda delivery example, we stated that we could safely assume that the number of sodas needed to restock the machines has nothing to do with the distance the driver must walk to reach the machines. That allowed us to consider delivery volume ($x_1$) and delivery distance ($x_2$) to be independent regressor variables.

Sometimes, two or more of the regressor variables are highly correlated with each other. This may be because the two variables are actually related to each other or, more often, that the two variables are both related to a third, unidentified, variable.

In these cases, it is hard to determine the contribution of each regressor variable to the prediction because two collinear variables (two variables that are linearly related to each other) do not bring twice as much information to bear on the problem. This makes the regression statistics somewhat hazy because you can’t test the significance of one regressor independent of the other.

One way to measure collinearity is to build a regression model to predict each regressor variable as a function of the remaining variables. The coefficient of multiple determination ($R^2$) should be reasonably close to zero if the variables are truly independent.