# 8.3: Introduction to Simple Linear Regression

**Learning Objectives**

- Define independent and dependent variables
- Differentiate between observed and predicted values of the dependent variable
- Compare different linear models using the sum of squared errors
- Motivate the line of best fit and provide means for its computation
- Develop the coefficient of determination
- Conduct linear regression analysis by checking the reasonability of using a linear model, computing $R^2$, and finding the line of best fit
- Nuance the predictive and interpretive power of linear regression

Section 8.3 Excel File: (contains all of the data sets for this section)

## Review and Preview

In studying bivariate quantitative data, we try to determine whether there is an association between two particular variables or not. If there is an association, a relationship between the variables, we would like to describe the relationship. We are interested in what happens to one variable as the other variable changes. We have discussed several ways to build this understanding: constructing scatter plots, classifying associations, and determining correlation. While more advanced textbooks address nonlinear correlation, we restricted ourselves to linear correlation. Linear correlation assesses the strength of an underlying linear relationship between the two variables of interest. If there is a linear relationship, it seems appropriate to think that there is a linear function that models the relationship. Knowledge of such a function would deepen our understanding of the relationship and allow us to extrapolate regarding values that were not explicitly measured in our collection of the data. In essence, such a function would enable us to make predictions about cases that were not explicitly studied. One of the fundamental motivations of statistical inquiry is to understand the world better so that we may better predict what will happen and act accordingly. This section develops the ideas of constructing a linear function that is the best fit for the data at hand.

When there is a linear relationship between two variables $x$ and $y$, we expect there to be constants $m$ and $b$ such that $y = mx + b$. In this formulation, we refer to $y$ (the vertical variable) as the **dependent variable** and $x$ (the horizontal variable) as the **independent variable**. Recall that $m$ is the slope of the line, the amount of change in $y$ if $x$ is increased by 1, and $b$ is the $y$-intercept, the $y$ value of the line when the $x$ value is $0$. Since there is a relationship between the two variables, one variable changes as the other changes; that is, the slope of the line is defined and not $0$; we have $m \neq 0$.

Remember that our study of correlation did not depend on the ordering of the variables. If there is a linear relationship between $x$ and $y$, there is a linear relationship between $y$ and $x$. In which case, we would expect another set of constants say $M$ and $B$ such that $x = My + B$. We would call $x$ the dependent variable and $y$ the independent variable. When studying associations, we do not assume causal relationships; do not let the terminology influence your thought in this regard.

When we collect data to understand the relationship, we expect the data to have some natural variation from the equation due to measurement error, natural variation, and the random noise that occurs in reality. As such, when we use collected data to construct a linear function, we are approximating the values of $m$ and $b$. Throughout this section, we will use the hat symbol to indicate approximation values. Thus $m$ is approximated by $\hat{m}$ and $b$ is approximated by $\hat{b}$. We will use these values to approximate $y$ values which we denote $\hat{y}$ using the following equation.

$$\hat{y} = \hat{m}x + \hat{b}$$

Several fundamental questions arise. How do we pick the values of $\hat{m}$ and $\hat{b}$? How do we know how well we did in picking the values of $\hat{m}$ and $\hat{b}$? Is there an optimal choice of values for $\hat{m}$ and $\hat{b}$? Even if we have the best line, it will not be perfect unless all the points are on the same line. Given that there is some error, how well does the line fit the data? These are questions we will begin to answer. We call the process of finding and evaluating these lines **regression analysis**.

## Modeling Using Linear Functions

When studying bivariate quantitative data, we do not expect, even when there is a linear relationship, that all of the data points fall precisely on the same line. As such, even once we decide upon a linear function to model the data, it is impossible for the function to match the data perfectly. There will necessarily be some error between some of the observed values and the predicted values associated with them. To visualize this consider the following data set along with a scatter plot visualizing it.

Table 8.3.1: Paired values of variables $x$ and $y$

| Observation | $x$ | $y$ |
|---|---|---|
| 1 | 7 | 5 |
| 2 | 2 | 4 |
| 3 | 6 | 6 |
| 4 | 3 | 3 |
| 5 | 1 | 2 |

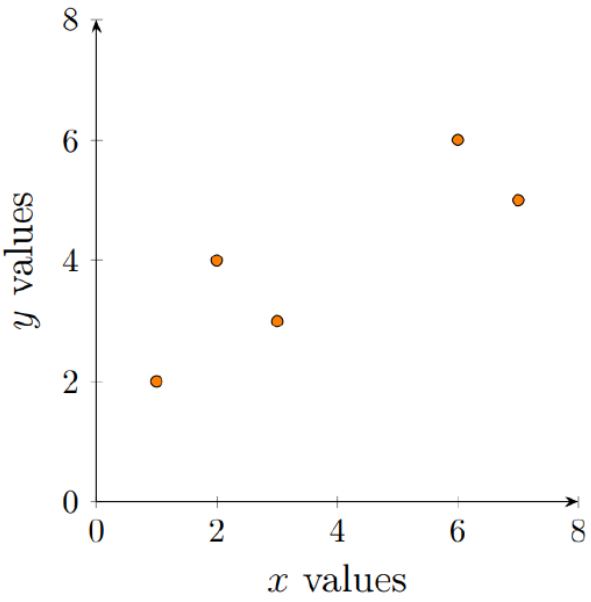


Figure 8.3.1: Scatter plot of $x$ and $y$

Notice that we cannot draw a line that goes through every point of the scatter plot. We could fairly easily plot a line through 2 of the points or even 3 of the points, but we cannot go through all 5 points; we cannot avoid the presence of error in our model! Suppose that we decided upon the linear function $\hat{y} = 0.5x + 2$ to model this particular data set and then plotted it below in dark blue. Notice that the line does not go through any of the observed values plotted on the scatter plot. The observed values remain in orange, and the predicted values are colored in a light blue.

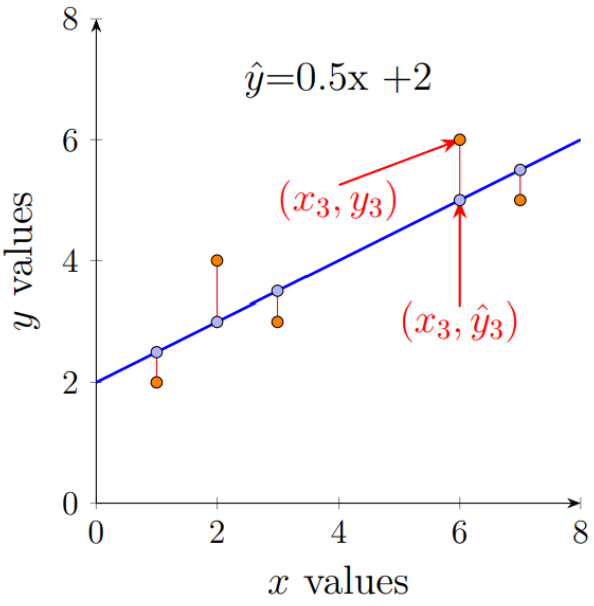


Figure 8.3.2: Scatter plot with linear model

The scatter plot labels two particular coordinate pairs on the scatter plot: $(x_3, y_3) = (6, 6)$ and $(x_3, \hat{y}_3) = (6, 5)$. The former is the observed pair; while, the latter is the predicted pair. Notice how the $x$-coordinates are the same. The predicted value of $y_3$ is $\hat{y}_3$. We compute $\hat{y}_3$ using the equation of the line $\hat{y}_3 = 0.5x_3 + 2 = 0.5 * 6 + 2 = 5$. We call the difference between $y_3$ and $\hat{y}_3$ the error at $x_3$ which we denote $e_3$. In general, we define the error at any observed $x$ value $x_i$ as follows.

$$e_i = y_i - \hat{y}_i$$

**? Text Exercise 8.3.1**

1. Using the linear function $\hat{y} = 0.5x + 2$, predict the value of $y$ when $x = 8$.

**Answer**

We can predict the value of the variable $y$ be evaluating the linear function at the indicated $x$ value. We have a predicted value of $0.5 \cdot 8 + 2 = 6$.

2. Using the linear function $\hat{y} = 0.5x + 2$, compute the error at each of the 5 collected $x$ values.

Table 8.3.2: Values for variables $x$ and $y$

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i = y - \hat{y}$ |
|---|---|---|---|---|
| 1 | 7 | 5 | | |
| 2 | 2 | 4 | | |
| 3 | 6 | 6 | | |
| 4 | 3 | 3 | | |
| 5 | 1 | 2 | | |

**Answer**

The error is computed by taking the difference between the $y$ value and the $\hat{y}$ value at a given $x$ value. We must determine each $\hat{y}$ value. We do so in the table that follows.

Table 8.3.3 Computation of predicted values and errors

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i = y - \hat{y}$ |
|---|---|---|---|---|
| 1 | 7 | 5 | $0.5 \cdot 7 + 2 = 5.5$ | $5 - 5.5 = -0.5$ |
| 2 | 2 | 4 | $0.5 \cdot 2 + 2 = 3$ | $4 - 3 = 1$ |
| 3 | 6 | 6 | $0.5 \cdot 6 + 2 = 5$ | $6 - 5 = 1$ |
| 4 | 3 | 3 | $0.5 \cdot 3 + 2 = 3.5$ | $3 - 3.5 = -0.5$ |
| 5 | 1 | 2 | $0.5 \cdot 1 + 2 = 2.5$ | $2 - 2.5 = -0.5$ |

Notice that when the observed value is above the linear function modeling the data that the error is positive and when the observed value is below the line the error is negative. With this basis, we begin to develop the process of determining the best fitting line.

## The Line of Best Fit

When we gather information about the world around us, we collect a lot of information. In order to understand best, we try to incorporate as much of the data as we can in our analyses and considerations. We do not collect samples from many people and then only use the results from a handful. Each observation provides important information; we do not want to exclude information without due cause. So, how do we decide upon a line to model our data, when no line perfectly predicts our data in practice? We want to use all of the data in the construction of the line, but how do we achieve such a goal when a line is uniquely determined given two points or a point and a slope?

The answer resides in considering error. We could assess the quality of a line by looking at the error across all observed values, but how are we to assess the totality of the error? If we were sum or average the errors, we would get cancelation between positive and negative errors. Indeed, if we modeled the data from above with the constant function $\hat{y} = 4$, the sum of the errors and hence the average would be $0$, but the function would only go through one point and would indicate that there is no relationship between the variables. This does not match with the reality of the data. It is natural to desire that the measure of the totality of error is $0$ only when the model perfectly fits the data. We, therefore, must expand our considerations.

Hopefully, we remember a similar discussion surrounding how to measure the dispersion of a data set. We went through several possibilities until we settled on our definition of variance which involved summing the deviations from the mean squared. We will utilize a similar sort of methodology without much motivation; we will consider the **sum of the squared errors** (**SSE**) as our measure of the totality of the error present in the model. In setting this as our measure, we have that there exists a unique line that minimizes the SSE. We can thus find a line and assert that it is one and only best line. We refer to this line as the line of best fit or the least-squares regression line.

### ? Text Exercise 8.3.2

1. Compute the SSE using the function $\hat{y} = 0.5x + 2$ on the same data set as before, which we reproduce below.

Table 8.3.4: Values for variables $x$ and $y$

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i = y - \hat{y}$ | $e_i^2 = (y - \hat{y})^2$ |
|---|---|---|---|---|---|
| 1 | 7 | 5 | 5.5 | $5 - 5.5 = -0.5$ | |
| 2 | 2 | 4 | 3 | $4 - 3 = 1$ | |
| 3 | 6 | 6 | 5 | $6 - 5 = 1$ | |
| 4 | 3 | 3 | 3.5 | $3 - 3.5 = -0.5$ | |
| 5 | 1 | 2 | 2.5 | $2 - 2.5 = -0.5$ | |

**Answer**

We have computed the error for each of the observed $x$ values in a previous text exercise. All that is left to do is square each of the errors and then add them together.

Table 8.3.5 Computation of predicted values, errors, and squared errors

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i = y - \hat{y}$ | $e_i^2 = (y - \hat{y})^2$ |
|---|---|---|---|---|---|
| 1 | 7 | 5 | 5.5 | $5 - 5.5 = -0.5$ | $(-0.5)^2 = 0.25$ |
| 2 | 2 | 4 | 3 | $4 - 3 = 1$ | $1^2 = 1$ |
| 3 | 6 | 6 | 5 | $6 - 5 = 1$ | $1^2 = 1$ |
| 4 | 3 | 3 | 3.5 | $3 - 3.5 = -0.5$ | $(-0.5)^2 = 0.25$ |
| 5 | 1 | 2 | 2.5 | $2 - 2.5 = -0.5$ | $(-0.5)^2 = 0.25$ |

The sum of the squared errors is thus SSE $= 0.25 + 1 + 1 + 0.25 + 0.25 = 2.75$.

2. We can show that the function $\hat{y} = 0.5x + 2$ is not the best fitting line by finding another line that has a smaller SSE. One of the properties of the line of best fit is that it goes through the point $(\bar{x}, \bar{y})$. Let us keep the same slope but adjust the $y$-intercept so that our function goes through the point $(\bar{x}, \bar{y})$ and then compute the SSE.

Table 8.3.6: Values for variables $x$ and $y$

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i^2 = (y - \hat{y})^2$ |
|---|---|---|---|---|
| 1 | 7 | 5 | | |
| 2 | 2 | 4 | | |
| 3 | 6 | 6 | | |
| 4 | 3 | 3 | | |
| 5 | 1 | 2 | | |

**Answer**

We are going to model our data with the function $\hat{y} = 0.5x + b$ so that our function passes through the point $(\bar{x}, \bar{y})$. In order to compute $b$, we must compute $\bar{x}$ and $\bar{y}$, which we find to be $\bar{x} = 3.8$ and $\bar{y} = 4$. Knowing the slope and a point through which our line passes is enough to determine the value of $b$. We plug the $x$- and $y$-coordinates into the function and solve for $b$.

$$\begin{aligned} 4 &= 0.5 \cdot 3.8 + b \\ 4 &= 1.9 + b \\ 4 - 1.9 &= b \\ 2.1 &= b \end{aligned}$$

Thus producing the function $\hat{y} = 0.5x + 2.1$ as our linear model for the data set. From this we can produce a table of the approximated values at each observed $x$ value and then compute each of the squared errors.

Table 8.3.7 Computation of predicted values and squared errors

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i^2 = (y - \hat{y})^2$ |
|---|---|---|---|---|
| 1 | 7 | 5 | 5.6 | $(-0.6)^2 = 0.36$ |
| 2 | 2 | 4 | 3.1 | $0.9^2 = 0.81$ |
| 3 | 6 | 6 | 5.1 | $0.9^2 = 0.81$ |
| 4 | 3 | 3 | 3.6 | $(-0.6)^2 = 0.36$ |
| 5 | 1 | 2 | 2.6 | $(-0.6)^2 = 0.36$ |

The sum of the squared errors is thus SSE $= 0.36 + 0.81 + 0.81 + 0.36 + 0.36 = 2.7$. Notice how the individual errors increased in magnitude for some values but decreased for others when comparing the error values using the previous model with these error values. Also, note that the sum of the errors is 0, again emphasizing the inadequacy of using the sum or average of the errors as a measure of the totality of error. We finally notice that the SSE is smaller with this new model than with the old. We thus say that this model is better than the previous model. The question is, have we found the best model yet?

3. Using technology, we determined that the line of best fit for the data set in question is given by the equation $\hat{y} = .5224x + 2.0149$ where the values of $\hat{m}$ and $\hat{b}$ are rounded to four decimal places. Compute the SSE.

Table 8.3.8: Values for variables $x$ and $y$

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i^2 = (y - \hat{y})^2$ |
|---|---|---|---|---|
| 1 | 7 | 5 | | |
| 2 | 2 | 4 | | |
| 3 | 6 | 6 | | |
| 4 | 3 | 3 | | |
| 5 | 1 | 2 | | |

**Answer**

We have computed the error for each of the observed $x$ values in a previous text exercise. All that is left to do is square each of the errors and then add them together.

Table 8.3.9 Computation of predicted values and squared errors

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i^2 = (y - \hat{y})^2$ |
|---|---|---|---|---|
| 1 | 7 | 5 | 5.6717 | $(-0.6717)^2 \approx 0.4512$ |
| 2 | 2 | 4 | 3.0597 | $0.9403^2 \approx 0.8842$ |
| 3 | 6 | 6 | 5.1493 | $0.8507^2 \approx 0.7237$ |
| 4 | 3 | 3 | 3.5821 | $(-0.5821)^2 \approx 0.3388$ |
| 5 | 1 | 2 | 2.5373 | $(-0.5373)^2 \approx 0.2887$ |

The sum of the squared errors is thus SSE $\approx 2.6866$. We again note that this is the smallest of the sum of squared errors that we have yet to see for this data set. Indeed, this is the smallest attainable value for any linear function! No matter what slope and $y$-intercept picked, the sum of the squared errors will be larger than this value. Readers with a background in calculus or linear algebra are encouraged to read or work out the details for why this is!

Establishing the uniqueness and computational formulas for the line of best fit requires mathematics beyond the scope of this course. The mathematics does, however, produce very elegant results regarding the slope and $y$-intercept for the line of best fit, the line that minimizes the sum of the squared errors. We provide the formulas without proof.

$$\hat{m} = r\frac{s_y}{s_x} \qquad \hat{b} = \bar{y} - r\frac{s_y}{s_x}\bar{x} = \bar{y} - \hat{m}\bar{x}$$

**? Text Exercise 8.3.3**

Using the coefficients for the line of best fit provided above, show that every line of best fit, regardless of the data set, goes through the point $(\bar{x}, \bar{y})$. That is, show that when you substitute $\bar{x}$ in the formula for the line of best fit, the value returned is $\bar{y}$.

**Answer**

The line of best fit is given by the formula $\hat{y} = \hat{m}x + \hat{b}$. We were given the formulas for $\hat{m}$ and $\hat{b}$. We have $\hat{y} = r\frac{s_y}{s_x}x + \bar{y} - r\frac{s_y}{s_x}\bar{x}$. When we substitute $\bar{x}$ into the variable $x$, we obtain the following.

$$\begin{aligned}\hat{y} &= r\frac{s_y}{s_x}\bar{x} + \bar{y} - r\frac{s_y}{s_x}\bar{x} \\ &= \bar{y}\end{aligned}$$

In addition to the fact that the point $(\bar{x}, \bar{y})$ always falls on the line of best fit, we have the sum and the average of all the errors is always 0 (when we do not round the numbers). We will not provide a proof of such a fact here, but it is a fact worth noting.

Just as with the computation of the correlation coefficient, we generally rely on technology to compute the slope and $y$-intercept for the line of best fit. We provide you with the function in Excel that returns the desired information as an array with the slope in the left cell and the $y$-intercept in the right cell. The function name is $\mathrm{LINEST}$ and takes four arguments: the first is the array of $y$ values (values of the dependent variable); the second is the array of $x$ values (values of the independent variable); the third is set of $\mathrm{TRUE}$ or 1; and the fourth is set to $\mathrm{FALSE}$ or simply 0. Further information is available using the fourth argument but will not be utilized in this course.

## Assessing the Line of Best Fit

We have now established that we can find the line of best fit, but another consideration must be made. Just because something is the best does not necessarily mean it is good. Of all the lines that could be used to model the data, we can find the best one, but does this best line actually fit the data well? This is the question we seek to answer and seems closely related to the correlation coefficient. Since the correlation coefficient measures the strength of an apparent linear relationship, we would expect that the closer $|r|$ is to 1, the better our line of best fit will model the data. This intuition is correct and will be confirmed as we approach the problem from a different direction.

When we are studying bivariate quantitative data (variables $x$ and $y$) we are interested in how one variable changes as the other changes. With this, we may ask how much of the change in one variable can be attributed to the change of the other variable? Inherently, this question requires the development of some method or model which can measure the amount of change in the dependent variable which can be attributed to the model. When making such a measurement, the interest lies in the proportion of the change in one variable that can be attributed to the model, not the raw amount of variation that can be attributed. This allows the measure to be compared across data sets composed of data with vastly different magnitudes and makes the measure value independent of the units of the measurement. A high percentage indicates that the model fits well. Most of the change in $y$ can be explained as due to the change in the $x$ variable. If the percentage is low, the model does not fit well. The majority of the change in $y$ is not understood as due to changes in $x$ under the model.

In order to continue, we must decide how to measure the change in the $y$ variable; this is really a question of dispersion. In general, the more the $y$ variable changes, the greater the spread of the $y$ variable data. Our most commonly used measure of dispersion has been standard deviation, but as we have seen throughout our bonus discussions, the real statistical power lies not in standard deviation but in variance. Recall that the variance is closely related to the average of square deviations from the mean, but we are not interested in a typical value, rather, we want the total change in the $y$ variable. As such, we define our measure of change in $y$ to be the **total variation** of $y$ which we can compute with the following for data sets with $n$ observations, note the similarity to the definition of variance.

$$\text{Total Variation} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

We are interested in computing the percentage of the total variation of $y$ that is explained by using the line of best fit to model the data; we call this percentage the **coefficient of determination** and denote it using the symbol $R^2$. To determine the coefficient of determination, we must be able to compute the explained variation in our model. Either the variation is explained or it is not explained. As such, we know that the total variation is equal to the sum of the unexplained variation and the explained variation. The disparity between predicted values and observed values is the source of the unexplained variation. At this point, we recognize that the $\mathrm{SSE}$ is the unexplained variation. Recall the meaning of the sum of the squared errors and think of the formula that would compute it in general.

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

With these ideas in hand we are ready to derive a computational formula for the coefficient of determination which measures the percent of variation in a data set that is accounted for by using the line of best fit to model the data.

$$R^2 = \frac{\text{Total Variation} - \text{SSE}}{\text{Total Variation}} = 1 - \frac{\text{SSE}}{\text{Total Variation}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Note that we dropped the indices on the summations in our formula to ease, but remember that for each of the summations, we are summing through all observations, 1 through $n$. The equivalency of the last computational formula is not obvious but can derived using the definitions of the line of best fit and the correlation coefficient, we encourage the reader to verify the simplification. In simplifying this last formulation, we develop the connection between the correlation coefficient and the coefficient of determination (a connection possibly already surmised given the symbolic representations); namely, that in squaring the correlation coefficient, we obtain the coefficient of determination.

**Optional Derivation Connecting Correlation Coefficient and Coefficient of Determination for the Mathematically Inclined**

$$\begin{aligned}
R^2 &= \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{\sum \left( r\frac{s_y}{s_x} x_i + \bar{y} - r\frac{s_y}{s_x}\bar{x} - \bar{y} \right)^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{\sum \left( r\frac{s_y}{s_x} x_i - r\frac{s_y}{s_x}\bar{x} \right)^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{\sum \left( r\frac{s_y}{s_x} (x_i - \bar{x}) \right)^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{\left( r\frac{s_y}{s_x} \right)^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} \\
&= \frac{r^2 \frac{s_y^2}{s_x^2} (n-1) s_x^2}{(n-1) s_y^2} \\
&= \frac{r^2 \frac{\cancel{s_y^2}}{\cancel{s_x^2}} \cancel{(n-1)} \cancel{s_x^2}}{\cancel{(n-1)} \cancel{s_y^2}} = r^2
\end{aligned}$$

## Text Exercise 8.3.4

1. Using technology, confirm that the line of best fit matches the formula provided in the earlier text exercise when rounded to 4 decimal places. Using the slope and $y$-intercept of the line of best fit provided by technology without rounding, determine the SSE, the Total Variation,and $R^2$ by applying the formulas derived thus far in this section. Provide an interpretative sentence explaining the meaning of the coefficient of determination in light of the computed value. Do not round answers in intermediate steps.

Table 8.3.10: Values for variables $x$ and $y$

| Observation | $x$ | $y$ |
|---|---|---|
| 1 | 7 | 5 |
| 2 | 2 | 4 |
| 3 | 6 | 6 |
| 4 | 3 | 3 |
| 5 | 1 | 2 |

**Answer**

Using the LINEST function in Excel, we confirm the accuracy of the previous text exercise rounded to 4 decimal places with a computed slope of $0.52238806\ldots$ and a $y$-intercept of $2.014925373\ldots$. We provide a table with values rounded to 6 decimals for checking.

Table 8.3.11: Computation of predicted values, squared errors, and variations

| Observation | $x$ | $y$ | $\hat{y}$ | $e_i^2 = (y-\hat{y})^2$ | $(y_i-\bar{y})^2$ |
|---|---|---|---|---|---|
| 1 | 7 | 5 | 5.671642 | $(-0.671642)^2 \approx 0.451113$ | $(5-4)^2 = 1$ |
| 2 | 2 | 4 | 3.059702 | $0.940299^2 \approx 0.884161$ | $(4-4)^2 = 0$ |
| 3 | 6 | 6 | 5.149254 | $0.850746^2 \approx 0.723769$ | $(6-4)^2 = 4$ |
| 4 | 3 | 3 | 3.582090 | $(-0.582090)^2 \approx 0.338828$ | $(3-4)^2 = 1$ |
| 5 | 1 | 2 | 2.537313 | $(-0.537313)^2 \approx 0.288716$ | $(2-4)^2 = 4$ |

The sum of the squared errors is thus $\text{SSE} \approx 2.686567$, the smallest possible value for this particular data set. The $\text{Total Variation}$, the summation of the last column is 10. We can compute the coefficient of determination $R^2 \approx 1 - \frac{2.686567}{10} \approx 1 - 0.268657 \approx 0.731343$. Approximately $73.1\%$ of the variation present in the $y$ variable is accounted for using the linear function $\hat{y} = 0.522388x + 2.014925$. This indicates that the linear model fits the data to a certain degree, but there is a decent amount of random variation, error, or noise present.

2. Using the results of the previous part of this text exercise and technology, confirm that the square of the correlation coefficient is equal to the coefficient of determination. Compute the correlation coefficient using technology.

**Answer**

We computed in the previous part that $R^2 \approx 0.731343$. Using the $\text{CORREL}$ function in excel, we compute that $r \approx 0.855186$ and note that $0.855186^2 \approx 0.731343$ which is the value that we computed for the coefficient of determination.

The coefficient of determination $R^2$ can be computed directly using the Excel function $\text{RSQ}$. The function takes two arrays of numbers, similar to the $\text{LINEST}$ function, the first array consists of the known $y$-values (dependent variable) and the second array consists of the known $x$-values (independent variable).

## Simple Linear Regression: Predictions and Interpretations

We have yet to conduct simple linear regression outside of a purely mathematical context. Having developed the concepts, we now address the application of these ideas and provide insight to their interpretations. Let us return to a data set that we have started to analyze, the ages of the bride and groom on their wedding day. Using a scatter plot of the data, we have already determined that a linear model would be appropriate. Let us determine the line of best fit and assess how well the model fits the data. In our previous considerations, we had the groom's age on the horizontal axis, the axis traditionally associated with the independent variable. For continuity of presentation, we will continue in this vein of thought. When evaluating the linear model we will be predicting the age of the bride based on the input of a groom's age. If we want to predict the age of a groom based on a particular age of a bride, we

will either have to solve for the age of the groom or conduct the linear regression analysis with the variables switched. Either option is fairly straightforward and will produce the same predictions.

### Text Exercise 8.3.5

1. Letting the age of the bride on her wedding day be the dependent variable, determine the line of the best fit and the coefficient of determination for the data set. Explain the results of the linear regression in the context of the problem.

Table 8.3.12: Ages of bride and groom on wedding day

| Married Couple | Groom's Age (years) | Bride's Age (years) |
|---|---|---|
| 1 | 20 | 21 |
| 2 | 26 | 20 |
| 3 | 32 | 34 |
| 4 | 30 | 30 |
| 5 | 21 | 22 |
| 6 | 29 | 28 |
| 7 | 26 | 25 |
| 8 | 34 | 34 |
| 9 | 29 | 28 |
| 10 | 55 | 50 |
| 11 | 30 | 26 |
| 12 | 43 | 39 |
| 13 | 30 | 29 |
| 14 | 24 | 22 |
| 15 | 20 | 19 |

**Answer**

Using Excel, we have the line of best fit given by $\hat{y} = 0.878562x + 2.168364$ with a coefficient of determination equal to $0.938862$. The computed coefficient of determination indicates that almost $94\%$ of the variation in the age of a bride on her wedding day can be accounted by modeling the relationship between the ages of the bride and groom with the function $\hat{y} = 0.878562x + 2.168364$, where $x$ is the age of the groom. This is a fairly high percentage which indicates that the model is a good fit. The positive slope indicates a positive association.

2. The $y$-intercept of a function is the value of the dependent variable when the independent variable is equal to $0$. Within the context of our problem $x = 0$ corresponds to the groom's age being $0$. The $y$-intercept is about $2.17$, indicating that the bride would be just a little older than $2$. Explain why, contextually, these considerations do not make any sense. What does this say about our model? What does this say about linear regression models in general?

**Answer**

Infants and toddlers do not get married. Adults get married. It is remiss to try to use the line of best fit to model the relationship where the relationship does not exist. There is a mathematical domain for our function and there is a contextual domain for our relation. If we are trying to understand the reality around us, the contextual domain must be at the forefront of our minds. We do not want to extend our model where the relationship ceases or beyond where our data permits us to engage. As such, we would not want to use our model for any ages less than 16 or 18 years of age for either the bride or the groom as those are the ages commonly set as the minimum ages for which marriage is legal. This does not say anything

negative about our model or models in general; we must be cognizant of when it is appropriate to use the models. Contextual clues are a big help. We will develop more nuance as we progress through this section.

3. Using the model constructed in part 1, predict the age of the bride when the groom is 27 years old.

**Answer**

We have that the line of best fit is given by $\hat{y} = 0.878562x + 2.168364$. We are being asked to predict the age of the bride when the groom is 27 years old. This is equivalent to evaluating the function when the $x$ variable is 27. We predict the bride's age to be $0.878562 \cdot 27 + 2.168364 = 25.88955$ years old; the bride will be just shy of 26 years old, when the groom is 27 years old.

4. Using the model constructed in part 1, predict the age of the groom when the bride is 32 years old.

**Answer**

We have again that the line of best fit is given by $\hat{y} = 0.878562x + 2.168364$. We are asked to predict the age of the groom when the bride is 32 years old. This is not equivalent to evaluating the function when the $x$ variable is 32 because the $x$ variable corresponds to the age of the groom, not the bride. We predict the groom's age by solving for $x$ when our linear equation equals 32.

$$\begin{aligned} 32 &= 0.878562x + 2.168364 \\ 32 - 2.168364 &= 0.878562x \\ \frac{32 - 2.168364}{0.878562} &= x \\ 33.955055 &\approx x \end{aligned}$$

We predict that the groom will be about just shy of 34 when the bride is 32 years old.

5. Using the model constructed in part 1, when does the model predict that the bride and groom will be exactly the same age? Does this seem like an appropriate use of the model?

**Answer**

For the last time, we have that the line of best fit is given by $\hat{y} = 0.878562x + 2.168364$. We want to find when the model predicts the two ages to be the same, i.e. $\hat{y} = x$. To do so, we replace $\hat{y}$ with $x$ in the equation and solve.

$$\begin{aligned} x &= 0.878562x + 2.168364 \\ x - 0.878562x &= 2.168364 \\ x(1 - 0.878562) &= 2.168364 \\ x &= \frac{2.168364}{1 - 0.878562} \approx 17.855796 \end{aligned}$$

We predict the bride and groom to be the same age when they are both just shy of 18 on their wedding day.

In general, once a person is about 2 years of age, the primary focus is on the number of years. As such, our interest might be more of when the model predicts that both the bride and the groom would be in the same year of life. This would seem to be a more appropriate question given the context of the model; although, it is a much harder question to solve for ages of 17.5 and 17.8 would constitute solutions as well as what we just found.

## ? Text Exercise 8.3.6

1. When ordering custom clothing or preparing to rent formal wear, many measurements are taken to ensure that the clothes fit well. Two common measurements are height and the length from the center of the back between the scapulae to the tips of the fingers when the arm is fully extended to the side. Let us refer to this latter measurement as an individual's radius.

We asked a random selection of 50 online elementary statistics students to obtain these measurements for themselves in centimeters and report their findings to analyze as a group. We provide the data in the attached Excel file. Examine the data to check that a linear model is appropriate. If not, explain. If so, find the line of best fit and coefficient of determination using the radius as the independent variable.

**Answer**

We first create a scatter plot to check if a linear relationship is reasonable. We provide two scatter plots with different scaling.

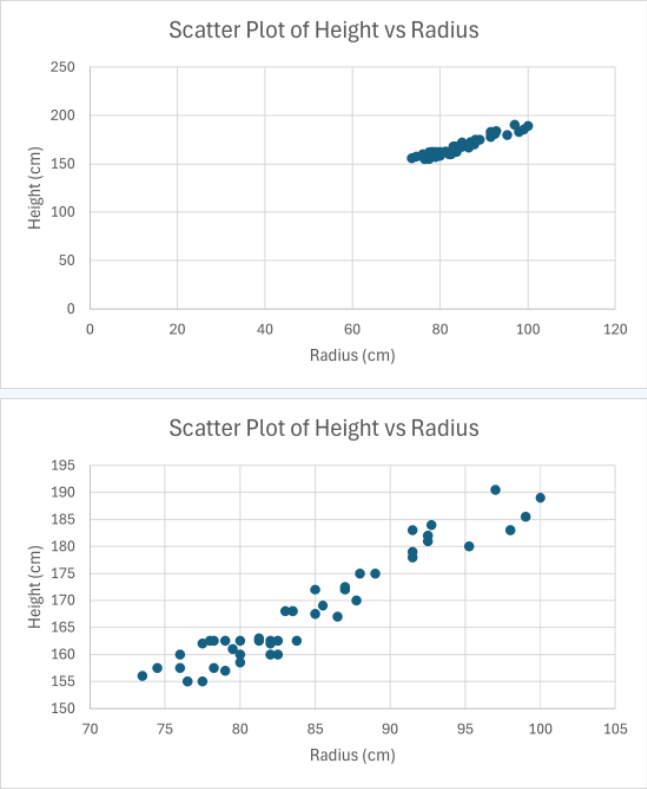


Figure 8.3.3 Scatter plots of height and radius using different scales and ranges

The scatter plot on the left includes the origin $(0, 0)$ while the other scatter plot does not. Both indicate a fairly linear relationship. So we proceed with a linear regression analysis. The coefficient of determination is $0.9212$with the linear model defined by $\hat{y} = 1.381x + 51.3125$.

In the previous text exercise, we determined the line of best fit and saw that the line fit fairly well. A little more than $92\%$ of the variation in the height variable was attributed to the difference values of the radius variable through our linear model. We have a nice model to help us understand the relationship between the height and radius of individuals. The possible values of an individual's radius go beyond those collected in our sample. This is one of the reasons that we desired a model; so, that we could estimate values for points where we did not have any data collected. As such, we might be tempted to estimate the height of an individual with a radius of 40 centimeters.

**? Text Exercise 8.3.7**

Using the line of best fit found in the previous text exercise, estimate the height of an individual with a radius of 40 centimeters. Consider the validity of such an estimation.

**Answer**

We would estimate that individual's height to be $1.381 \cdot 40 + 51.3125 = 106.5525$centimeters. Both the radius and the height values are within the contextual domain of our variables, but can we use the model in such a way? The predicted

height is about three and a half feet tall; a rather short person. In reality, most likely a child. This begs the question: should we use data from a sample of elementary statistics students who are fully formed adults to make predictions about a child? Hopefully, at this stage in our development of statistics, we would be inclined to say no. We would not think that a sample of adults would be representative of children without some significant argumentation explaining why they are fundamentally the same. Our intuition would naturally be that the body structure of children is different than the body structure of adults. We do not want to overgeneralize our results beyond that which we have actually studied. In practice, we must consider both the contextual domain and the extent to which our sample is representative. In general, we do not want to utilize our model too far beyond the values seen in our collected data. Do you want to predict the height of an individual with a radius of $90$ centimeters? Go right ahead! But, if you want to predict the height of an individual with a radius of $15,$ best go collect data from individuals around that size.

We conclude this section with one last interpretative guideline. The slope of a linear function describes the rate of change of the function. If the value of $x$ increases by $1$, the value of $y$ changes in value equal to the slope. In the case of our last text exercise, when we increase the radius by one centimeter, the predicted $y$ value increases by $1.381$ centimeters. Are we to interpret this to indicate that if an individual had a radius of $75$ centimeters and height of $150$ centimeters and then grew to $76$ centimeters, the individual's height would be $151.381$ centimeters? Unfortunately, the answer is no. We built the model by using data from $50$ individuals. The model predicts the typical relationship between the variables; it does not predict the individual change, nor does it predict the changes in a perfect way. We must temper ourselves from concluding more than we can. We can expect that as individuals increase in radius by $1$ centimeter, the average gain in height is going to be close to $1.381$ centimeters, but we cannot make such a claim on the individual level.

This is, in fact, a theme pertaining to the entirety of this textbook. Statistics seeks to understand trends in large groups, and it is almost always inappropriate to use information about a group to infer facts about an individual. If we say one group is shorter than another group on average, that does not necessarily mean that every individual in the first group is shorter than every individual in the second group. If we say that $80\%$ of some group has a particular disease, that does not necessarily mean that each individual in that group has an $80\%$ chance of obtaining that disease. If we say that a hypothesis or model predicts a group to have certain parameters, that says nothing about a specific individual in that group. Many issues in modern society arise from people misunderstanding this. People often use facts about a group to inform their thoughts about individuals (such as with stereotyping). People also often ignore facts about groups because of facts they know about an individual; for example, Bob smoked his whole life and lived to be $100$. We hope that this text has helped the reader understand how to properly understand facts about groups and why such understanding can be useful.

---