6 Linear Models

Concepts:

- Construct a Linear Model
- Gauge Accuracy of a Linear Model with Residuals
- Least Squares Linear Regression Line
- Interpret the Correlation Coefficient, r

(Section 2.5)

In many different fields of study, the collection and analysis of data is a very important process for examining trends and making decisions based on those trends. One way to analyze data is by creating a *mathematical model* (or equation) that most closely matches the data. The model can then be used to make predictions - interpolations or extrapolations - for points at which there is no data available. The accuracy of the model will affect the accuracy of the prediction.

In this section, we will learn to develop *linear* models and find ways to determine the accuracy of these models.

6.1 Constructing a Linear Model

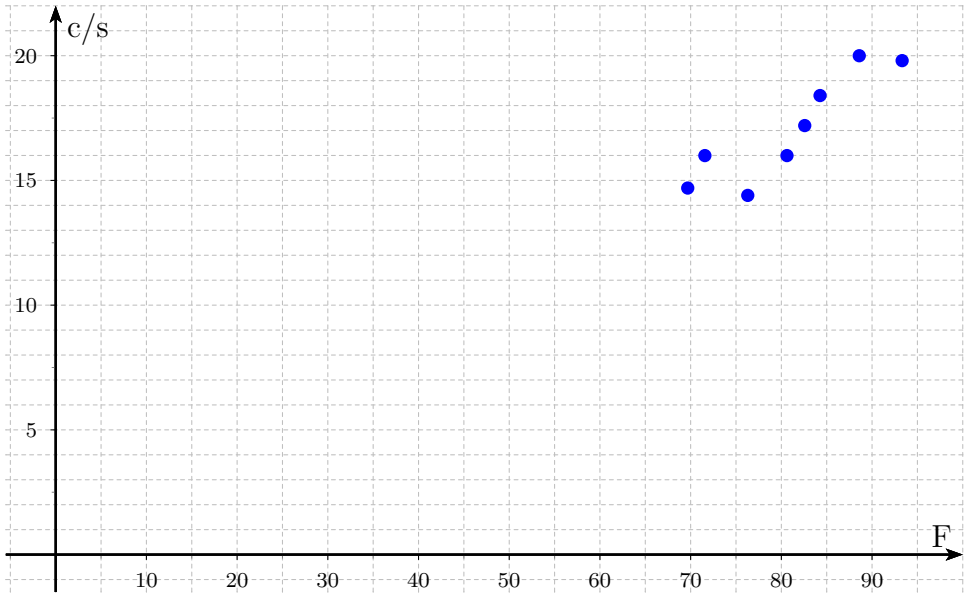
We will eventually develop methods for determining whether a model is an accurate representation of the data collected. However, we will begin with a very simple approach.

One of the easiest ways to construct a linear model is by using two of the existing data points, as we will see in our first example.

Example 6.1 (Data and Scatter Plots)
With the following data, we compare the temperature in degrees Fahrenheit to the number of chirps per second for the striped ground cricket. Reference: *The Song of Insects* by Dr.G.W. Pierce, Harvard College Press.

Some data points:

F^o	*chirps/sec*
69.7	14.7
71.6	16
76.3	14.4
80.6	16
82.6	17.2
84.3	18.4
88.6	20
93.3	19.8



We see from the scatter plot that the data is approximately linear, thus a line is a reasonable model for the data.

(a) Use the points (69.7, 14.7) and (93.3, 19.8) to find a linear model for the data.

(b) Use the points (76.3, 14.4) and (88.6, 20) to find a linear model for the data.

(c) Sketch the graphs for both of your models.

While both models appear to do a reasonable job of modeling the data, the previous example shows us that there are many linear models we can construct from the data values given.

Our next example will help us to determine which model is better by examining the errors in each of the models.

However, first we need to define what we mean by the error in a model.

Definition 6.2
Given a data point (x, d) and a corresponding model point (x, y), the error in the model for that specific value of x is the difference $d - y$. This difference is called a **residual**.

Thus, the *residual* in a linear model is the vertical distance from the data value to the linear model. When the data point lies above the linear model, the *residual* will be positive. When the data point lies below the linear model, the *residual* will be negative.

Because we want a model with the least amount of error, we will look at the overall error measured at each of the data points. This is done by calculating the sum of the ***squares*** of the residuals. For errors with an absolute value that is greater than one, the square of the error is larger. For errors with an absolute value that is smaller than one, the square of the error is smaller. Hence, the smaller the sum of the squares is, then the better the model is.

6.2 Gauging Accuracy by Using Residuals

Example 6.3
(Finding the Sum of Squares of Residuals) The two models for our cricket chirps from our first example are:

$$-5.1F^o + 23.6c = -8.55 \qquad\qquad -5.6F^o + 12.3c = -250.16$$

For each model, find the residuals, the squares of the residuals, and the sum of the squares of the residuals.

The following tables give the requisite information.

$-5.1F + 23.6c = -8.55$			
(F, d)	(F, c)	$d - c$	$(d - c)^2$
$(69.7, 14.7)$	$(69.7, 14.7)$	0.0	0.0
$(71.6, 16)$	$(71.6, 15.1)$	0.9	0.8
$(76.3, 14.4)$	$(76.3, 16.1)$	-1.7	3.0
$(80.6, 16)$	$(80.6, 17.1)$	-1.1	1.1
$(82.6, 17.2)$	$(82.6, 17.5)$	-0.3	0.1
$(84.3, 18.4)$	$(84.3, 17.9)$	0.5	0.3
$(88.6, 20)$	$(88.6, 18.8)$	1.2	1.5
$(93.3, 19.8)$	$(93.3, 19.8)$	0.0	0.0
			$\sum = 6.7$

$-5.6F + 12.3c = -250.16$			
(F, d)	(F, c)	$d - c$	$(d - c)^2$
$(69.7, 14.7)$	$(69.7, 11.4)$	3.3	10.9
$(71.6, 16)$	$(71.6, 12.3)$	3.7	14.0
$(76.3, 14.4)$	$(76.3, 14.4)$	0.0	0.0
$(80.6, 16)$	$(80.6, 16.4)$	-0.4	0.1
$(82.6, 17.2)$	$(82.6, 17.3)$	-0.1	0.0
$(84.3, 18.4)$	$(84.3, 18.0)$	0.4	0.1
$(88.6, 20)$	$(88.6, 20)$	0.0	0.0
$(93.3, 19.8)$	$(93.3, 22.1)$	-2.3	5.5
			$\sum = 30.6$

By comparing the sums of the squares of the residuals for each of our models, we see that the first model is a much better fit for the data, despite the fact that the majority of the data in the second model (5 of the 8 data points) were almost on the line.

6.3 Least Squares Linear Regression Line

Multivariate Calculus can show that among all the possible linear models you can find for a data set, there is one that is always the best.

Theorem 6.4 (Linear Regression Theorem)
For any set of data points, there is one and only one line for which the sum of the squares of the residuals is as small as possible. This line is called the **least squares regression line**.

Example 6.5 (Finding Regression Line)
With the data set from our first example, use technology to find the *least squares regression line*. Make a scatter plot and graph the regression line. This can be accomplished by entering the data points as two lists in the calculator's statistics editor with the STAT key.

Finding the Least Squares Regression Line

1. Entering data values in lists.
   (a) Hit STAT key.
   (b) Hit [ENTER] to EDIT lists. (lists are L_1, L_2, ...)
   (c) Enter the first values for each data point in L_1.
   (d) Use the RIGHT arrow to enter the second values in L_2.
   (e) Enter the second values for each data point in L_2.
2. Create a scatter plot of data points.
   (a) Hit [2nd] STAT PLOT. (STAT PLOT is above the [y=] key.)
   (b) Hit [ENTER] to open the STAT PLOT menu screen.
   (c) Hit [ENTER] again to turn ON Plot1
   (d) Use the DOWN arrow keys to select the Type, Lists, and Marks.
   (e) Select the [ZOOM] button to choose an appropriate viewing window.
   (f) Use the DOWN arrow keys to scroll to the "9:ZoomStat" window.
   (g) Hit [ENTER] to view a scatter plot of your data.
3. Obtaining the equation of the regression line.
   (a) Hit STAT key.
   (b) Use RIGHT arrow button to move to CALC menu.
   (c) Use DOWN arrow button to scroll to "4:LinReg (ax+b)".
   (d) Hit [ENTER].
   (e) Enter the list names and where the regression line is to be stored. (This can usually be done by simply hitting [ENTER] through the list).
   (f) Hit [ENTER] to get the coefficients for the equation of the linear regression line.

With the information provided here, we see that the linear equation that most closely matches our data is

$$c = .2393281598F^o - 2.293281598.$$

We can now use this model to make predictions for the number of chirps per second for the striped ground cricket at any Fahrenheit temperature. Predictions made with this model where $69.7 \leq F^o \leq 93.3$ are called *interpolations*, while predictions made where $F^o < 69.7$ or $93.3 < F^o$ are called *extrapolations*. *Interpolations* are typically more reliable than *extrapolations* as they will fall within (rather than outside of) the data values used to determine the equation for the least squares regression line.

6.4 Interpreting the Correlation Coefficient, r

In addition to the coefficients for the linear regression equation, the calculator (or other technology) will also provide another value r and its square r^2. The number r is called the **correlation coefficient**.

Definition 6.6
The **correlation coefficient** r is a statistical measure of how well the equation of the regression fits the data. This value is bounded as such, $-1 \leq r \leq 1$

If your calculator screen only displays the coefficients for the regression equation, follow these steps to turn on the additional diagnostics and reapply the steps for finding the equation. You will now see the additional diagnostic information.

Displaying Correlation Coefficient

1. Hit 2nd 0 key to access the catalog screen. Alpha-lock is automatically on.
2. Hit x^{-1} key to get to the "D" section of the catalog.
3. Scroll down to "DiagnosticOn" and hit ENTER to select the option.
4. Hit ENTER again to confirm it.
5. Rerun the linear regression equation process as described previously.
6. You should now see two additional statistics, r and r^2.

The closer that the absolute value of r is to 1 then the better the fit. If $r = 1$ or $r = -1$, then this indicates a perfect fit. Conversely, the closer that the absolute value of r is to 0 then to worse the fit, indicating there is no correlation and the data does not follow a linear model. See page 124 of the textbook for additional visual examples.

When $0 < r \leq 1$, this indicates a positive or *increasing* correlation.

When $-1 \leq r < 0$, this indicates a negative or *decreasing* correlation.