Chapter 8: Linear regression

In Chapter 6, the normal distribution was used to model the distribution of a quantitative variable. Chapter 7 introduced the idea of a relationship between two quantitative variables, and this chapter introduces a model of the relationship between two quantitative variables.\(^1\) The linear regression model is a simple straight line.

- The method is only appropriate when the relationship is approximately linear, just as the normal model was appropriate only when the distribution was approximately normal.

- The figure to the right displays the carbon dioxide record from the European Project for Ice Coring in Antarctica, Dome C ice core, covering 0 to 800,000 BP. There are two major components of trend. One component is captured by the linear model and the second is not\(^2\).

- Recall that the equation of a straight line is \(y = mx + b\) where \(m\) is the slope (change in \(y\) per unit change in \(x\)) and \(b\) is the intercept (the point on the \(y\)-axis where the line and \(y\)-axis intersect). The linear model is the equation of a straight line, and thus has two parameters, the intercept and slope.\(^3\)

- In linear regression applications, it’s necessary to find values for the parameters that agree data.

The linear regression model: The linear regression model predicts a response variable (\(y\)) from the explanatory variable (\(x\)) according to the equation

\[
\hat{y} = b_0 + b_1 x
\]

where \(b_0\) is the intercept, \(b_1\) is the slope, and \(\hat{y}\) is a value of the response variable that is predicted by a value \(x\) of the explanatory variable. Every \(\hat{y}\) lies on the line described by the model.

---

\(^1\)The objective of modeling in this instance is to capture the principal features of a relationship

\(^2\)The trend captured by the linear model is the upward trend from 800,000 BP to the present. The ignored trend component is cyclical.

\(^3\)The normal model also had two parameters, \(\mu\) and \(\sigma\).
• $b_0$ is the predicted value of $y$ when $x = 0$.

• $b_1$ represents how much the predicted value of $y$ increases when $x$ increases by 1 unit.

• Prediction is subject to error (a prediction $\hat{y}$ is not likely to be exactly the same as the actual value $y$). The differences between actual and predicted values are key to understanding the utility of a linear regression model. Some notation:

\[
y = \text{the observed or actual response value},
\hat{y} = \text{the predicted or fitted value (spoken as y-hat)},
e = y - \hat{y} = \text{observed value} - \text{predicted value} = \text{residual or error}.
\]

• The values of $b_0$ and $b_1$ chosen to agree with the data are those values minimize the sum of the squared errors ($\sum e^2$). The method of computing these best estimates is called the least squares method.

**Example:** The figure to the right shows of winner’s average speeds (km/hour) of winning racers in the Tour de France plotted against year of the race. Is there evidence of systematic doping among winners?

The speed of the Tour de France winner can be estimated for 2012 after computing the coefficients of the model

\[
\text{winner's speed} = b_0 + b_1 \text{year}.
\]

The explanatory variable is $x = \text{year}$ and the response variable is $y = \text{winner's speed}$. Explanatory and response variables are assigned this way because \textit{year explains speed} instead of speed explaining year.

**The method of least squares:** The method of least squares finds $b_0$ and $b_1$ so that the residuals have the smallest possible sum of squares. Said another way, the least squares line is the line which minimizes the sum of the squared residuals (or errors)

\[
\sum e^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.
\]

61
Graphically, the least squares line is that line which minimizes the sum of the squared vertical deviations from the line.

Several formulas exist for the slope and intercept of the least squares line; in practice, linear regression is almost exclusively conducted using statistical software (eliminating hand computations).

The formulas

\[ b_1 = r \frac{s_y}{s_x}, \]
\[ b_0 = \bar{y} - b_1 \bar{x}. \]

are useful because they reveal two important properties of the least squares estimators.

**Property 1:** \( b_1 \) depends on the value of \( r \). \( b_1 \) has the same sign as \( r \).

\( b_0 \) has the same units as the response variable and \( b_1 \) has, as units, the response variable units per explanatory variable units.

The second property of \( b_0 \) and \( b_1 \) is revealed by substituting the expression for \( b_1 \) into the formula for \( b_0 \):

\[ y = \bar{y} - b_1 \bar{x} + b_1 x \]
\[ = \bar{y} + b_1(x - \bar{x}). \]

The prediction of \( y \) when at \( x = \bar{x} \) is

\[ = \bar{y} - b_1(\bar{x} - \bar{x}) \]
\[ = \bar{y}. \]

**Property 2:** the least squares regression line passes through the pair \((\bar{x}, \bar{y})\).

In the Tour de France example, year is the explanatory variable and winner’s average speed (kph) is the response variable. The necessary summary statistics are

\[ \bar{x} = 2005.5 \text{ (in years)} \]
\[ s_x = 3.606 \text{ years} \]
\[ \bar{y} = 40.21 \text{ kph} \]
\[ s_y = .7191 \text{ kph} \]
\[ r = -.4680. \]

Then,

\[ b_1 = r \frac{s_y}{s_x} \]
\[ = -.4680 \times \frac{.7191}{3.606} \]
\[ = -.09333 \text{ kph/year} \]

and

\[ b_0 = \bar{y} - b_1 \bar{x} \]
\[ = 40.21 - -.09333(2005.5) \]
\[ = 227.4 \text{ kph} \]
• The equation of the least squares line is

\[ \hat{y} = 227.4 - 0.09333x. \]

• In the context of the problem, the equation is

winner’s speed = 227.4 − 0.09333 × year.

• The predicted winner’s speed (kph) for 2012 is

\[
\begin{align*}
\text{winner’s speed} &= 227.4 - 0.09333 \times 2012 \\
&= 39.62 \text{ kph}.
\end{align*}
\]

Interpreting the slope and intercept

1. The slope coefficient of −0.09333 kph/year implies that the predicted winner’s speed decreases by 0.09333 kph for every additional year. Equivalently, the predicted winning race speed decreases by 0.9333 kph for every additional 10 years.

2. The intercept of 227.4 kph is the predicted winning average speed for the Tour de France in the year 0 AD. The intercept doesn’t have a sensible interpretation in this problem.\(^4\)

Remark: A linear regression equation should be used for prediction only using values of \(x\) that are within the range of the observed data. For example, the model predicts an average winning speed of 31.4 kph = 18.8 mph in the year 2100.

The connection between \(b_1\) and \(r\): Recall that

\[ b_1 = r \frac{s_y}{s_x}. \]

If \(b_1 \approx 0\), then either

1. \(r \approx 0\) and there is little association between response and explanatory variables, or

2. \(s_y \approx 0\) and there is little variation in \(y\).

In either case, the explanatory variable is not a good predictor of the response variable.

Residual Analysis A residual is the difference between an observed value and a model prediction:

\[ e = y - \hat{y} \]

\(^4\)It’s not uncommon that \(b_0\) lacks meaning in the context of the problem
Also,

\[ y = \hat{y} + e \]

Observed = Predicted + Residual

An observed value consists of two parts; the part explained by the model (predicted) plus the remaining part not explained by the model (residual).

The prediction and the residual for \( x = 2003 \), when the observed average winner’s speed was \( y = 41.7 \) kph, are computed as follows:

1. The prediction is

\[ \hat{y} = 227.4 - .09333 \times 2003 \]
\[ = 40.46 \text{ kph}. \]

2. The residual is

\[ e = y - \hat{y} \]
\[ = 41.7 - 40.46 \]
\[ = 1.24 \text{ kph}. \]

The sum of the residuals is zero whenever the method of least squares is used to compute the fitted model. In this sense, the fitted line passes through the middle of the cloud of data points.

If the residual is negative, then the observation is less than the prediction since

\[ y - \hat{y} < 0 \Rightarrow y < \hat{y}. \]

Where do points with negative residuals lie on the scatterplot relative to the least squares line?\(^5\)

What does a negative residual mean in terms of the prediction of \( y \)?\(^6\)

Using residuals to assess the adequacy of the regression model: The model must be considered a good (adequate) model before it is useful or interesting.

\(^5\)Below the fitted regression line.
\(^6\)\( y \) was over-predicted by the line.
Adequacy is assessed by plotting the residuals against the fitted values (a residual plot). The residual plot for the Tour de France data is shown above and right.

- If there is no pattern in the residual plot (i.e.: the residuals are scattered randomly), then the linear model form is appropriate.
- A pattern in the residuals, such as a curved pattern indicates that the linear model does not represent the relationship adequately. Some aspect of the relationship is unaccounted for by model.
- What does the residual plot for the Tour de France data reveal?

The advantage of the residual plot for assessing the adequacy of the model over just looking at the original scatterplot is that it’s easier to compare the magnitude of the residuals because linear trend has been removed. Further, the residual plot scales the $y$-axis to the range of the residuals, not to the range of the responses.

For comparison, the Tour de France regression model was re-fit without 2003. The Figure to the right shows the fitted models and the data. The 2003 observation pair affected the fitted model.

Using the residuals to assess model adequacy

The magnitude of the residuals reflect

- the strength of linear association between $x$ and $y$, and
- the accuracy of the model predictions.

The larger the residuals are, in magnitude, the worse the model performs in predicting the response variable.

Model assessment requires a simple alternative to linear regression against which the regression model may be compared. The alternative model predicts the response variable using

---

7The residuals appear to be randomly distributed about the horizontal line. There is one residual that is larger in magnitude than the rest associated with the year 2003.

8Magnitude and absolute value are equivalent.
the average response \( \bar{y} \). For example, without knowing year, a simple and logical prediction of the winning speed in 2012 would be the average winning speed \( \bar{y} \) of the 11 observations on winning time. If we write this prediction function as a model, then \( b_1 \) is fixed, or constrained to be equal to 0 (i.e., \( b_1 = 0 \)) and the (alternative) model is \( \hat{y} = b_0 \). The model is a baseline against which to compare linear regression model. Unless the linear regression model is substantially better than the model \( \hat{y} = b_0 \), there is little point to using the regression model.

The total variation measures the accuracy of the simple alternative model \( (b_0 = \hat{y}, b_1 = 0) \). Specifically, it measures the average squared distance from the \( y \)'s to \( \bar{y} \), the model estimate of each and every \( y \).

1. The total variation in the response values is

\[
SST = \sum_{i=1}^{n} (y_i - \bar{y})^2.
\]

2. The corresponding measure of fit for the linear regression model is the sum of the squared errors (or residuals):

\[
SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.
\]

A common term for SSE is the residual variation. SSE is the variation in the response variable that is unexplained by the regression model.

3. Regression variation (SSR) is the variation explained by the regression model. It’s the difference between total variation and unexplained variation. Specifically,

\[
SSR = SST - SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.
\]

Showing the equality of lines 2 and 3 requires some tedious algebra.

- Another view: the total variation SST in the response variable can be partitioned into two sources of variation according to

\[
SST = SSR + SSE
\]

where

1. SSR is the variation explained by the model.
2. SSE is the variation not explained by the model.

- The ratio of SSR to SST has an important role in regression as it is the proportion of the variation in the responses explained by the model.

- The ratio SSR/SST is equal to $r^2$, the square of the correlation coefficient. The standard definition and notation is

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}.$$ 

- For the Tour de France data, $R^2 = r^2 = (-.469)^2 = .219$. Thus, about 22% of the variation in average winning speed (kph) is explained by the regression of average winning speed (kph) on year. More tersely, 21.9% of the variation in average winning speed (kph) is explained by year.9

**Example**: The Hawaiian volcanoes were produced by the Hawaiian hot spot, presently beneath the Big Island of Hawaii. The Hawaiian volcanoes increase in age in a progression along the island chain away from the hot spot and from southeast to northwest. The Figure below and left shows the volcano ages (millions of years) and distances (km) from Kilauea along Hawaiian volcanic island chain (many of the volcanoes are sea mounts). The Figure below and right is a residual plot. Does it appear that a linear regression model adequately captures the principal features of the relationship between volcano ages and distances?10

---

9Explaining 22% of the variation is disappointing if the objective is to use the linear regression model for prediction.

10Yes, but there are two features of concern: 1. The residual plot clearly indicates that the relationship has a curvilinear component that is not captured by the linear model. 2. The intercept should be approximately 0, but the curvilinear relationship has apparently led to an incorrect value.
Pearson’s correlation between age and distance is $r = .991$, which indicates that there is a very strong, positive linear association between distance and age. Table 1, obtained from R, shows output from a regression of age on distance.

Table 1: Hawaiian hotspot data. Summary statistics: regression of age on distance.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$</td>
<td>229.830</td>
<td>52.554</td>
<td>4.373</td>
<td>.000115</td>
</tr>
<tr>
<td>$b_1$</td>
<td>75.687</td>
<td>1.732</td>
<td>43.700</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

The results reported in the context of the problem:

1. The least squares line (in the context of the problem) is
distance (km) = 229.8 + 75.7 × age (millions of years).

2. The slope implies that for each additional million years of age, distance from the hot spot is predicted to increase by 75.7 km.

3. Predict the distance from the hot spot for a volcano 30 million years old.
distance (km) = 229.8 + 75.7 × 30 = 2500.8 km.

4. Compute $R^2$ and interpret it in the context of the problem.
$R^2 = r^2 = .991^2 = .983$. 98.3% of the variation in distance is explained by age.