9.10 Model I: simple linear regression: only one $y$ for each $x$

In this section we consider the calculations in two boxes. The first (Box 9.4) describes how to calculate a regression line which allows you to model the association or make predictions. In Box 9.5 the significance of this association is tested. Both calculations are illustrated in relation to Example 9.5.

Example 9.5 Heavy metal contamination of soil under electricity pylons

An undergraduate investigated heavy metal tolerance in plants growing under electricity pylons. As part of her study, she recorded the concentration of zinc in soil samples taken at regular intervals moving away from the pylons (Table 9.11). One of the objectives of her investigation was to see whether there was an association between the distance from the pylon and the concentration of zinc in the soil.

Table 9.11 Zinc concentrations in soil at specific distances from an electricity pylon

<table>
<thead>
<tr>
<th>Distance from pylon (m)</th>
<th>Zinc concentration ($\mu g$ Zn/g soil)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>648</td>
</tr>
<tr>
<td>1.5</td>
<td>610</td>
</tr>
<tr>
<td>2.0</td>
<td>534</td>
</tr>
<tr>
<td>2.5</td>
<td>500</td>
</tr>
<tr>
<td>3.0</td>
<td>472</td>
</tr>
</tbody>
</table>

BOX 9.4 How to carry out a Model I simple linear regression (one $y$ value for each $x$ value): drawing a regression line

1. How to work out the regression coefficient $b$
   
   i. Variable $x$, the distance from the pylon (m), is under the investigators control and is the independent variable. Variable $y$ is the concentration of zinc ($\mu g$ Zn g$^{-1}$ g soil) in the soil. The independent variable is that plotted on the $x$ axis. The dependent variable is plotted on the $y$ axis.
   
   ii. Calculate the following:
   
   $\Sigma x$ Add all $x$ values together.
   
   $\Sigma x = 1.0 + 1.5 + 2.0 + 2.5 + 3.0 = 10$
   
   $(\Sigma x)^2$ Square the $\Sigma x$ value.
   
   $(\Sigma x)^2 = (10)^2 = 100$
   
   $\Sigma x^2$ Square all the $x$ values and sum.
   
   $\Sigma x^2 = 1.0^2 + 1.5^2 + 2.0^2 + 2.5^2 + 3.0^2 = 10.0 + 2.25 + 4.0 + 6.25 + 9.0 = 22.5$
   
   (continued)
**RESEARCH METHODS FOR THE BIOSCIENCES**

\[ \sum y \text{ Add all the } y \text{ values together.} \]
\[ \sum y = 648 + 610 + 534 + 500 + 472 = 2764 \]
\[ \sum xy \text{ Multiply each } x \text{ by its corresponding } y. \text{ Add all these } xy \text{ values together.} \]
\[ \sum xy = (1.0 \times 648) + (1.5 \times 610) + (2.0 \times 534) + (2.5 \times 500) + (3.0 \times 472) \]
\[ = 648 + 915 + 1068 + 1250 + 1416 = 5297 \]

Table 9.12 summarizes these calculations.

**Table 9.12** Calculating a Model I simple linear regression (one y value for each x value) using the data from Example 9.5 Heavy metal contamination of soil under electricity pylons

<table>
<thead>
<tr>
<th>Distance from pylon (m) (x)</th>
<th>Zinc concentration (µg Zn/g soil) (y)</th>
<th>xy</th>
<th>x × y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>648</td>
<td></td>
<td>648</td>
</tr>
<tr>
<td>1.5</td>
<td>610</td>
<td></td>
<td>915</td>
</tr>
<tr>
<td>2.0</td>
<td>534</td>
<td></td>
<td>1068</td>
</tr>
<tr>
<td>2.5</td>
<td>500</td>
<td></td>
<td>1250</td>
</tr>
<tr>
<td>3.0</td>
<td>472</td>
<td></td>
<td>1416</td>
</tr>
<tr>
<td>n = 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Σx = 10</td>
<td>Σy = 2764</td>
<td></td>
<td>Σ(xy) = 5297</td>
</tr>
<tr>
<td>(Σx)^2 = 100</td>
<td>(Σy)^2 = 7639696</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Σ(x^2) = 22.5</td>
<td>Σ(y^2) = 1549944</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x̄ = 2.0</td>
<td>ȳ = 552.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
iii. b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{(5 \times 5297) - (10 \times 2764)}{(5 \times 22.5) - 100} = \frac{26485 - 27640}{112.5 - 100} = \frac{-1155}{12.5} = -92.4
\]

2. How to work out a

i. Examine your data and record the means for each variable. 
   \[ x̄ = 2.0, \; ȳ = 552.8 \]

ii. \( a \) can be calculated using the general formula for this straight line, so that 
   \[ a = \bar{y} - b \bar{x} = 552.8 - (-92.4 \times 2) = 552.8 + 184.8 = 737.6 \]

3. What is the regression equation?

For a Model 1 regression the general equation of the line is 
\[ y = a + bx = 737.6 + (-92.4)x \]
\[ y = 737.6 - 92.4x \]

4. How to draw the line

The regression line indicates the apparent trend in your data. If you want to draw this line on your scatter plot you will need to calculate three pairs of data points as follows:
The second step of the simple linear regression is to see if there is a statistically significant association that is described by this line. Box 9.5 shows you this part of the calculation.

Take any $x$ value from your observed data and place this and the values for $a$ and $b$ into the general equation $y = a + bx$. The only unknown value will be $y$. Work out the $y$ value and plot the $x$ you selected and this $y$ value on your graph. Repeat for at least two more points and draw a line through them within the range of your data set. This is the regression line and should be labelled with the regression equation.

When $x = 1m$
\[ y = 737.6 - (92.4 \times 1) \]
\[ = 645.2 \ \mu g \ \text{Zn g}^{-1} \ \text{soil} \]

When $x = 2m$
\[ y = 737.6 - (92.4 \times 2) \]
\[ = 552.8 \ \mu g \ \text{Zn g}^{-1} \ \text{soil} \]

When $x = 3m$
\[ y = 737.6 - (92.4 \times 3) \]
\[ = 460.4 \ \mu g \ \text{Zn g}^{-1} \ \text{soil} \]

FIGURE 9.13 illustrates the scatter plot with the regression line added. But you need to confirm that this is a statistically significant association (see Box 9.5).

The second step of the simple linear regression is to see if there is a statistically significant association that is described by this line. Box 9.5 shows you this part of the calculation.

**BOX 9.5 How to carry out a Model I simple linear regression (one $y$ value for each $x$ value): testing the significance of the association**

1. **Hypotheses to be tested**
   - $H_0$: There is no linear association between the distance from the pylons (m) and concentrations of zinc in the soil ($\mu g \text{ Zn g}^{-1} \text{ soil}$).
   - $H_1$: There is a linear association between the distance from the pylons (m) and concentrations of zinc in the soil ($\mu g \text{ Zn g}^{-1} \text{ soil}$).

2. **Have the criteria for using this test been met?**
   Sort of (9.10.2ii).

3. **How to work out $t_{calculated}$**
   i. First calculate the basic terms $SS(x)$, $SS(y)$ and $SP(xy)$. (These calculations are also summarized in Table 9.12.)
   \[
   SS(x) = \frac{\sum (x^2) - (\sum x)^2}{n}
   \]
   \[
   \sum x = 1.0 + 1.5 + 2.0 + 2.5 + 3.0 = 10
   \]
   \[
   (\sum x)^2 = 10^2 = 100
   \]
   \[
   (\sum x^2) = \frac{102}{5} = \frac{100}{5} = 20
   \]
   \[
   \sum (x^2) = 1.0^2 + 1.5^2 + 2.0^2 + 2.5^2 + 3.0^2 = 1.0 + 2.25 + 4.0 + 6.25 + 9.0 = 22.5
   \]

   (continued)
RESEARCH METHODS FOR THE BIOSCIENCES

\[
SS(x) = \Sigma \left( x^2 \right) - \left( \frac{\Sigma x}{n} \right)^2 = 22.5 - 20.0 = 2.5
\]

\[
SS(y) = \Sigma \left( y^2 \right) - \left( \frac{\Sigma y}{n} \right)^2
\]

\[
\Sigma y = 648 + 610 + 534 + 500 + 472 = 2764
\]

\[
(\Sigma y)^2 = 7639696
\]

\[
\frac{(\Sigma y)^2}{n} = \frac{7639696}{5} = 1527939.2
\]

\[
\Sigma (y^2) = 648^2 + 610^2 + 534^2 + 500^2 + 472^2 = 419904 + 372100 + 285156 + 250000 + 222784 = 1549444
\]

\[
SS(y) = \Sigma \left( y^2 \right) - \left( \frac{\Sigma y}{n} \right)^2 = 1549444 - 1527939.2 = 22004.8
\]

\[
SS(xy) = \Sigma (xy) - \left( \frac{\Sigma x}{n} \right) \left( \frac{\Sigma y}{n} \right)
\]

\[
\Sigma xy = (1.0 \times 648) + (1.5 \times 610) + (2.0 \times 534) + (2.5 \times 500) + (3.0 \times 472) = 648 + 915 + 1068 + 1250 + 1416 = 5297
\]

\[
ss(xy) = \Sigma xy - \left( \frac{\Sigma x}{n} \right) \left( \frac{\Sigma y}{n} \right) = 5297 - \left( \frac{10 \times 2764}{5} \right) = 5297 - \frac{27640}{5} = 5297 - 5528 = -231.0
\]

ii. Calculate the residual variance \((s_r^2)\).

\[
s_r^2 = \frac{1}{n-2} \left[ SS(y) - \left( \frac{SP(xy)}{SS(x)} \right)^2 \right] = \frac{1}{5-2} \left[ 22004.8 - \left( \frac{-231.0}{2.5} \right)^2 \right]
\]

\[
= \frac{1}{3} \left[ 22004.8 - \frac{53361}{2.5} \right] = \frac{1}{3} \left[ 660.4 \right] = 220.1333
\]

iii. Calculate the standard error of \(b (SE(b))\).

\[
SE(b) = \frac{s_r^2}{SS(x)} = \sqrt{\frac{220.1333}{2.5}} = \sqrt{88.053333} = 9.38367
\]

iv. Calculate the test statistic, \(t_{calculated}\).

\[
t_{calculated} = \frac{b}{SE(b)} = \frac{-92.4}{9.38367} = -9.84689
\]

4. **How to find \(t_{critical}\)**

Use a t table for a two-tailed test where \(p = 0.05\) (Appendix d, Table D6). The degrees of freedom \((\nu)\) are \(n - 2\). When \(\nu = 5 - 2 = 3\) and at \(p = 0.05\), \(t_{critical} = 3.182\).

5. **The rule**

If the absolute value of \(t_{calculated}\) is greater than the critical value of \(t\) then you may reject \(H_0\). As \(t_{calculated} (9.847)\) is more than \(t_{critical} (3.182)\) at \(p = 0.05\), you may reject the null hypothesis. In fact, at \(p = 0.01\), \(t_{critical} = 5.841\), and at \(p = 0.001\), \(t_{critical} = 12.941\), so you may reject the null hypothesis at \(p = 0.01\) but not at \(p = 0.001\).

6. **What does this mean in real terms?**

There is a highly significant \((0.01 > p > 0.001)\) negative linear association between the distance from the pylons (m) and concentrations of zinc in the soil (\(\mu g Zn/g soil\)) described by \(y = 737.6 - 92.4x\).