1 Inferences using the Least-Squares coefficients

Recall that if your data follows the model

\[ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \]

then the Least-Squares estimators of \( \beta_0 \), \( \beta_1 \) are given by

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}
\]

You learned that

\[
E(\hat{\beta}_1) = \beta_1, \quad E(\hat{\beta}_0) = \beta_0
\]

\[
\text{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \text{Var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),
\]

and that the estimation of the standard deviation is based on the residuals:

\[
s = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n - 2}} = \sqrt{\frac{(1 - r^2) \sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 2}}
\]

which yield the standard errors of the coefficients:

\[
s_{\hat{\beta}_1} = s \sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}, \quad s_{\hat{\beta}_0} = s \sqrt{\frac{1}{n + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}
\]

The estimated \( y_i \) is given by \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \). It has a (sample) standard deviation given by

\[
s_{\hat{y}_i} = s \sqrt{\frac{1}{n + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}
\]

which leads to the \((1 - \alpha)100\%\) confidence interval for \( \hat{y}_i \)

\[
\hat{y}_i \pm t_{n-2, \alpha/2}s_Y
\]

2 Exercise 8.1.7, page 332

The article “Evaluation of the Expansion Attained to Date by Concrete Affected by Alkali-Silica Reaction. Part III: Application to Existing Structures” (M. Bérubé, N. Smaoui, et al., 2005) reports measurements of expansion for several concrete bridges in the area of Quebec City. Following are measurements of horizontal and vertical expansion (in parts per hundred thousand) for several locations on the Père-Lelièvre bridge.

<table>
<thead>
<tr>
<th>Horizontal (x)</th>
<th>20</th>
<th>15</th>
<th>43</th>
<th>5</th>
<th>18</th>
<th>24</th>
<th>32</th>
<th>10</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical (y)</td>
<td>58</td>
<td>58</td>
<td>55</td>
<td>80</td>
<td>58</td>
<td>68</td>
<td>57</td>
<td>69</td>
<td>63</td>
</tr>
</tbody>
</table>

(a) Compute the least-squares line for predicting vertical expansion from horizontal expansion.

(b) Compute 95% confidence intervals for \( \beta_0 \) and \( \beta_1 \).
3 Checking Assumptions

For things to work in inference for linear models, the following must be true:

1. The errors $\varepsilon_1, \ldots, \varepsilon_n$ should be independent.
2. The errors $\varepsilon_1, \ldots, \varepsilon_n$ must have mean 0.
3. The errors $\varepsilon_1, \ldots, \varepsilon_n$ should have the same variance, denoted by $\sigma^2$.
4. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are normally distributed.

4 Exercise 8.2.2

The processing of raw coal involves "washing," in which coal ash (nonorganic, incombustible material) is removed. The article “Quantifying Sampling Precision for Coal Ash Using Gy’s Discrete Model of the Fundamental Error” (1989) provides data relating the percentage of ash to the volume of a coal particle. The average percentage of ash for six volumes of coal particles was measured. The data are as follows:

<table>
<thead>
<tr>
<th>Volume (x cm$^3$)</th>
<th>0.01</th>
<th>0.06</th>
<th>0.58</th>
<th>2.24</th>
<th>15.55</th>
<th>276.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent ash (y)</td>
<td>3.32</td>
<td>4.05</td>
<td>5.69</td>
<td>7.06</td>
<td>8.17</td>
<td>9.36</td>
</tr>
</tbody>
</table>

(a) Compute the least-squares line for predicting percent ash (y) from volume (x). Plot the data, and the residuals versus the fitted values. Does the linear model seem appropriate? Explain.

(b) Compute the least-squares line for predicting percent ash from ln(volume). Plot the data, and the residuals versus the fitted values. Does the linear model seem appropriate? Explain.

(c) Compute the least-squares line for predicting percent ash from $\sqrt{\text{volume}}$. Plot the data, and the residuals versus the fitted values. Does the linear model seem appropriate? Explain.

(d) Using the most appropriate model, predict the percent ash for particles with a volume of 50 cm$^3$.

(e) Using the most appropriate model, construct a 95% confidence interval for the mean percent ash for particles with a volume of 50 cm$^3$.

(f) There’s no item (f) in the book, but there’s a typo in it. The units for x should be cm$^3$. Would your prediction be appropriate if you used x in cm$^3$ for the model construction, and then tried to predict a value for $x = 50m^3$? What would the predicted percent ash value be in this case?