Homework #2 (Solutions)
STAT-UB.0003: Regression and Forecasting Models

1. MBS, Ex. 7.34.

Solution:

(a) The null and alternative hypotheses are

\[ H_0 : \mu = 85 \]
\[ H_a : \mu \neq 85, \]

where \( \mu \) is the true mean Mach rating score of all purchasing managers.

(b) We reject the null hypothesis if the absolute value of the test statistic is larger than \( z_{0.05} = 1.645 \). (It is also acceptable to give \( t_{0.05,121} \approx t_{0.05,120} = 1.658 \).)

(c) The test statistic is

\[ t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{(99.6) - (85)}{(12.6)/\sqrt{122}} \]
\[ = 12.80 \]

(d) Since \(|t| > 12.80\), we reject the null hypothesis.

2. MBS, Ex. 7.55, parts (a)-(e).

Solution:

(a) The null and alternative hypotheses are \( H_0 : \mu = 2 \) and \( H_a : \mu \neq 2 \), where \( \mu \) is the mean surface roughness of coated interior pipe.

(b) The test statistic (from the Minitab output) is \( t = -1.02 \).

(c) We reject \( H_0 \) if \(|t| > t_{0.025,19} = 2.093.\) You get full credit for this problem if you say \(|t| > z_{0.025} = 1.96 \) or \(|t| > 2 \).

(d) We do not reject \( H_0 \).

(e) The \( p \)-value is \( p = 0.322 \). If the true mean surface roughness were 2 microns, then there would be a 32.2% chance of seeing data at least as extreme as observed. In other words, the data is consistent with the null hypothesis (it would be very typical if \( H_0 \) were true).
3. MBS, Ex. 11.7. Give one example where a probabilistic model is preferable, and one example where a deterministic model is preferable.

Solution:
Probabilistic models are generally preferable because they allow for approximate rather than exact linear relationships; they allow for deviations around the linear trend. One example where a deterministic relationship is more appropriate is a taxi fare:

\[
\text{(total fare)} = (\text{flag drop fee}) + \beta_1 (\text{number of miles driven})
\]

One example where a probabilistic model is more appropriate is a model relating a movie’s box office gross to its advertising budget:

\[
\text{(box office gross)} = \beta_0 + \beta_1 (\text{advertising budget}) + \epsilon;
\]

there will not be an exact relationship between these quantities, so we need a probabilistic model.
(Many other examples of deterministic and probabilistic linear models are valid.)

4. MBS, Ex. 11.9.

Solution: No, it does not. The random error will cause a deviation above or below the regression line (the mean).

5. MBS, Ex. 11.10.

Solution:
(a) Here is the completed table:

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( x_i^2 )</th>
<th>( x_i y_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2</td>
<td>49</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>36</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
<td>18</td>
</tr>
</tbody>
</table>

Totals \( \sum x_i = 24 \) \( \sum y_i = 31 \) \( \sum x_i^2 = 116 \) \( \sum x_i y_i = 80 \)
(b) 
\[ SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \]
\[ = (80) - \frac{(24)(31)}{7} \]
\[ = -26.28571. \]

(c) 
\[ SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \]
\[ = (116) - \frac{(24)^2}{7} \]
\[ = 33.71429 \]

(d) 
\[ \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \]
\[ = \frac{-26.28571}{33.71429} \]
\[ = -0.77966 \]

(e) 
\[ \bar{x} = \frac{\sum x_i}{n} \]
\[ = \frac{24}{7} \]
\[ = 3.42857, \]
\[ \bar{y} = \frac{\sum y_i}{n} \]
\[ = \frac{31}{7} \]
\[ = 4.42857 \]

(f) 
\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 - \bar{x} \]
\[ = (4.42857) - (-0.77966)(3.42857) \]
\[ = 7.10169 \]

(g) The least squares line is 
\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 + x \]
\[ = 7.10169 - 0.77966 x \]
6. MBS, Ex. 11.15. Here is the Minitab output from fitting the model described in the problem; use this output instead of the SPSS output given in the textbook:

Model Summary

<table>
<thead>
<tr>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.9119</td>
<td>92.06%</td>
<td>91.89%</td>
<td>91.44%</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-97.4</td>
<td>26.7</td>
<td>-3.65</td>
<td>0.001</td>
<td>1.00</td>
</tr>
<tr>
<td>MATH2001</td>
<td>1.1882</td>
<td>0.0499</td>
<td>23.83</td>
<td>0.000</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Regression Equation

\[ \text{MATH2011} = -97.4 + 1.1882 \text{MATH2001} \]

Solution:

(a) \( E(Y | x) = \beta_0 + \beta_1 x \). (The book writes this as \( E(y) = \beta_0 + \beta_1 x \)). You get full credit for a deterministic model \( y = \beta_0 + \beta_1 x \), but you should understand why a probabilistic model is better.

(b) \( \hat{y} = -97.4 + 1.1882x \)

(c) No practical interpretation; it doesn’t make sense to have an SAT score of 0.

(d) For every additional point on a state’s average 2001 Math SAT score, a state’s expected average Math 2011 SAT score increases by 1.1882 points. The interpretation is valid for the range of the \( x \) values in the data (between 474 and 603; this information is not given in the problem, but can be gotten by looking at the raw data).