620-270 Applied Statistics

Problem Set 4

Except for the first two questions, most problems in this set require MINITAB (or another package which performs regression) to complete all the questions.

4.1 Here is the height and weight of seven St. Kilda footballers:

| height (cm) | 172 | 175 | 181 | 187 | 190 | 191 | 203 |
| weight (kg) | 75  | 79  | 84  | 87  | 87  | 95  | 108 |

Do the following questions without using a statistical package:

(a) Draw a scatter plot of weight vs height.
(b) Guess the sample correlation coefficient r.
(c) Calculate \( \bar{x}, s_x, \bar{y}, s_y \) and the sample covariance \( s_{xy} \).
(d) Use these quantities to calculate \( r \).
(e) Use the quantities in (c) to find \( \hat{\beta} \), the estimated slope of the regression of weight on height. How would you roughly summarise the effect on weight of increasing a footballer’s height by 1 cm?
(f) Find \( \hat{a} \), the constant (or intercept) in the regression. Draw the fitted line on your plot.
(g) What is the predicted weight for a height of 185 cm?

4.2 Periodic measurements of salinity and water flow were taken in North Carolina’s Pamlico Sound, resulting in the following data (\( x = \) water flow, \( y = \) salinity):

<table>
<thead>
<tr>
<th>x</th>
<th>23</th>
<th>24</th>
<th>26</th>
<th>25</th>
<th>30</th>
<th>24</th>
<th>23</th>
<th>22</th>
<th>24</th>
<th>25</th>
<th>22</th>
<th>22</th>
<th>22</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>7.6</td>
<td>7.7</td>
<td>4.3</td>
<td>5.9</td>
<td>5.0</td>
<td>6.5</td>
<td>8.3</td>
<td>8.2</td>
<td>13.2</td>
<td>12.6</td>
<td>10.4</td>
<td>10.8</td>
<td>13.1</td>
<td>12.3</td>
</tr>
</tbody>
</table>

The following is part of a MINITAB output obtained using these data:

The regression equation is
\[
y = 30.8 - 0.910 \, x
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>30.816</td>
<td>7.025</td>
<td>4.39</td>
<td>0.001</td>
</tr>
<tr>
<td>x</td>
<td>-0.9105</td>
<td>0.2932</td>
<td>-3.10</td>
<td>0.008</td>
</tr>
</tbody>
</table>

\( s = 2.341 \quad R^2 = 42.6\% \quad R^2(\text{adj}) = 38.2\% \)

(a) Suppose that the population regression line is \( y = \alpha + \beta x \). Find estimates of \( \alpha \) and \( \beta \) and their standard errors.
(b) What are the assumptions made for inference on this linear regression model? Over what range of salinity levels would you expect your fitted model to apply?
(c) Obtain an estimate for average salinity when the water flow rate, \( x = 22 \), and a standard error for this estimate (you will need to calculate \( \bar{x} \) for this).
(d) Obtain an estimate of the error variance.
(e) Find the predicted values and residuals for the first two observations of the sample.
(f) Test the hypothesis that \( \beta = 0 \), at the 0.01 significance level.
4.3 A study was carried out on the effects of varying pressures on the density of cylindrical specimens made by dry pressing a ceramic compound. A mixture of Al₂O₃, polyvinyl alcohol and water was prepared, dried overnight, crushed and sieved to obtain 100 mesh size grains. These were pressed into cylinders at pressures ranging from 2000 psi to 10 000 psi and cylinder densities were calculated. The table below gives the data that were obtained: \( x \) denotes the pressure in thousands of psi, and \( y \) denotes the density in g/cm³.

\[
\begin{array}{cccccc}
2 & 4 & 6 & 8 & 10 \\
2.486 & 2.558 & 2.646 & 2.724 & 2.861 \\
2.479 & 2.570 & 2.657 & 2.774 & 2.879 \\
2.472 & 2.580 & 2.653 & 2.808 & 2.858 \\
\end{array}
\]

(a) Using MINITAB, draw a scatter plot of \( y \) vs \( x \).

(b) Find the least squares regression line.

(c) Comment on the statement that “the predicted density for zero pressure is 2.375 g/cm³”.

(d) Give an estimate of the increase in density that accompanies an increase of 2000 psi in the pressure, and specify the standard error of your estimate.

(e) Specify a predicted value for the density obtained when a pressure of 5000 psi is applied.

4.4 The following are the exam marks \( y \), test scores \( x_1 \) and number of classes missed \( x_2 \) of 12 students doing a particular subject:

\[
\begin{array}{ccccccccccccccccccccccccccc}
\text{ } & y & 85 & 74 & 76 & 90 & 85 & 87 & 94 & 98 & 81 & 91 & 76 & 74 \\
x_1 & 65 & 50 & 55 & 65 & 55 & 70 & 65 & 70 & 55 & 70 & 50 & 55 \\
x_2 & 1 & 7 & 5 & 2 & 6 & 3 & 2 & 5 & 4 & 3 & 1 & 4 \\
\end{array}
\]

(a) Specify the response variable and explanatory variables.

(b) Intuitively, which of the two explanatory variables do you think is more likely to be a good predictor of exam marks?

(c) Use multiple scatter plots for the data set to check your intuition and comment on the relationships between the variables.

(d) Obtain the correlation coefficients between the three variables and comment on them.

(e) Check that MINITAB gives the following multiple linear regression line:

\[
y = 27.5 + 0.922x_1 + 0.284x_2.
\]

Assuming this is a reasonable model, if a student had a test score of 70 and missed only one class, what would you predict her exam mark to be?

(f) For the first student in the table, calculate the fitted value, and then the residual for that observation.

(g) How do you account for the fact that the coefficient of \( x_2 \) in the regression line is positive, when the correlation between \( y \) and \( x_2 \) is negative? (run the regression without \( x_1 \) and see what happens). Would you include \( x_2 \) in the final model?

(h) Run the regression with only \( x_1 \) and recalculate the predicted exam mark for a student with test score 70.

(i) Verify that the value of \( R^2 \) is the square of the correlation coefficient found in (d).
4.5 An experiment was conducted to determine if the weight of an animal could be predicted after a given period of time on the basis of the initial weight of the animal and the amount of feed that was eaten. The following data, measured in kg, were recorded:

<table>
<thead>
<tr>
<th>final weight, $y$</th>
<th>95</th>
<th>77</th>
<th>80</th>
<th>100</th>
<th>97</th>
<th>70</th>
<th>50</th>
<th>80</th>
<th>92</th>
<th>84</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial weight, $x_1$</td>
<td>42</td>
<td>33</td>
<td>33</td>
<td>45</td>
<td>39</td>
<td>36</td>
<td>32</td>
<td>41</td>
<td>40</td>
<td>38</td>
</tr>
<tr>
<td>feed eaten, $x_2$</td>
<td>272</td>
<td>226</td>
<td>259</td>
<td>292</td>
<td>311</td>
<td>183</td>
<td>173</td>
<td>236</td>
<td>230</td>
<td>225</td>
</tr>
</tbody>
</table>

(a) Specify the response variable and explanatory variables.

(b) Use matrixplot to plot multiple scatter plots for the data set and comment on the relationships between the variables.

(c) Obtain the correlation coefficients between the three variables and comment on them. If you had to choose only one of the explanatory variables as a predictor, which would you choose? Why?

The regression command on MINITAB produces the following output for these data:

```
MTB > regr c1 2 c2 c3
```

The regression equation is

$y = -23.0 + 1.40 x_1 + 0.218 x_2$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-22.99</td>
<td>17.76</td>
<td>-1.29</td>
<td>0.237</td>
</tr>
<tr>
<td>$x_1$</td>
<td>1.3957</td>
<td>0.5825</td>
<td>2.40</td>
<td>0.048</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.21761</td>
<td>0.05777</td>
<td>3.77</td>
<td>0.007</td>
</tr>
</tbody>
</table>

$s = 6.051$ \hspace{1cm} R-sq = 87.3\% \hspace{1cm} R-sq(adj) = 83.7\%$

(d) The model being fitted is $\mu = \alpha + \beta_1 x_1 + \beta_2 x_2$. Obtain estimates and standard errors for $\alpha$, $\beta_1$, and $\beta_2$.

(e) Test the hypotheses $\beta_1 = 0$ and $\beta_2 = 0$ at the 5\% significance level.

(f) Obtain an estimate of the error standard deviation $\sigma$.

(g) What is number of degrees of freedom for the residual SS in this model? Write down the relation between the residual SS and the estimate obtained in (f).

(h) Assuming this is a reasonable model, if an animal has initial weight 40kg and has eaten 250kg of feed, what would you predict its final weight to be?

4.6 For Assignment 2

When anthropologists analyse human skeletal remains, an important piece of information is the stature (the height of the person when alive). Since skeletons are usually incomplete, inferences about stature are commonly based on statistical methods that utilise measurements on small bones. A research paper presented data to validate one such method. Consider the following representative data, where $x =$ metacarpal length (in mm) and $y =$ stature (in cm).

<table>
<thead>
<tr>
<th>$x$</th>
<th>45</th>
<th>51</th>
<th>39</th>
<th>41</th>
<th>52</th>
<th>48</th>
<th>49</th>
<th>46</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>171</td>
<td>178</td>
<td>157</td>
<td>163</td>
<td>183</td>
<td>172</td>
<td>183</td>
<td>172</td>
<td>175</td>
</tr>
</tbody>
</table>

(a) Specify the response variable and explanatory variable.

(b) Draw the scatter plot for the data set (either by hand or using MINITAB).

(c) Find the correlation coefficient between metacarpal length and stature, and comment on it.
(d) Fit the least-squares regression line and draw this line on your scatter plot.

(e) Assuming this is a reasonable model, what is the predicted stature of a person with metacarpal length of 50 mm?

The following are descriptive statistics for $x$ and $y$:

$x = 46.0; \ s_x = 4.44; \ \bar{y} = 172.7; \ s_y = 8.59; \ s_{xy} = 33.88.$

*Using these statistics*, and the residual mean square from the regression ($s^2 = 17.88$),

(f) Verify the correlation coefficient found in (c).

(g) Calculate the standard error of $\hat{\beta}$, the estimated slope of the regression line.

(h) Calculate the standard error of the predicted stature found in (e).

(i) For the two people with the largest metacarpal lengths and stature, calculate the fitted values and the residuals.