Statistics 13

Instructor: Nicolas Christou

Exam 1
29 October 2014

Name: SOLUTIONS

UCLA ID: ______________________

Section: ______

Problem 1  (20 points)
Answer the following questions:

a. Consider the simple linear regression model \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \). Using data on \( y \) and \( x \) we find the least squares estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) and fit the best line \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \). Is it possible that \( \operatorname{var}(y) < \sigma^2_{\epsilon} \)? Note: \( \operatorname{var}(y) \) is the variance of the \( y \) values and \( \sigma^2_{\epsilon} \) is the variability around the fitted line. If this is possible can you draw (approximately) the corresponding scatterplot of \( y \) on \( x \)?

![Yes, very weak association](image)

b. Consider the simple regression model \( y_i = \gamma_0 + \beta_1 x_i + \epsilon_i \). If \( \beta_0 = 4 \) what is the least squares estimate of \( \beta_1 \)? Note: Use the procedure of least squares we discussed in class, but here the intercept is known!

\[
\begin{align*}
\sum_{i=1}^{n} y_i &= 4 + \beta_1 \sum_{i=1}^{n} x_i + \epsilon_i, \\
\frac{d}{dx} \sum_{i=1}^{n} y_i &= -2 \sum_{i=1}^{n} (y_i - 4 - \beta_1 x_i) x_i = 0, \\
\sum_{i=1}^{n} x_i y_i - 4 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\beta_1 &= \frac{\sum_{i=1}^{n} x_i y_i - 4 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}
\end{align*}
\]

c. Consider the simple linear regression model \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \). One of the goals of regression is to use the fitted regression line to make predictions for \( y \) when a new value of \( x \) is given. If a new value of \( x \) is below \( \bar{x} \) is it true that the predicted value of \( y \) will be always below \( \bar{y} \)? Please explain your answer!

![Graphs showing relationship between y and x](image)
Problem 2  (20 points)
The data in the table below shows the depth of a stream and the rate of its flow.

<table>
<thead>
<tr>
<th>Flow Rate</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.636</td>
<td>0.34</td>
</tr>
<tr>
<td>0.319</td>
<td>0.29</td>
</tr>
<tr>
<td>0.734</td>
<td>0.28</td>
</tr>
<tr>
<td>1.327</td>
<td>0.42</td>
</tr>
<tr>
<td>0.487</td>
<td>0.29</td>
</tr>
<tr>
<td>0.924</td>
<td>0.41</td>
</tr>
<tr>
<td>7.350</td>
<td>0.76</td>
</tr>
<tr>
<td>5.800</td>
<td>0.73</td>
</tr>
<tr>
<td>1.979</td>
<td>0.46</td>
</tr>
<tr>
<td>1.124</td>
<td>0.40</td>
</tr>
</tbody>
</table>

(a)  

(b)  

(c)  

(d)  

Plot (a) shows the scatterplot of Rate against Depth.  
Plot (b) shows the scatterplot of Residuals against Depth from the simple regression of Rate on Depth.  
Plot (c) shows the scatterplot of Log(Rate) against Log(Depth).  
Plot (d) shows the scatterplot of Residuals against Log(Depth) from the simple regression of Log(Rate) on Log(Depth).

Answer the following questions:

a. Write the regression model equation based on plot (a).
   \[ \hat{\text{Rate}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Depth} + \varepsilon_i \]

b. Write the fitted line equation based on plot (c).
   \[ \log \hat{(\text{Rate})} = \hat{\beta}_0 + \hat{\beta}_1 \log (\text{Depth}) \]

c. Explain why it was necessary to transform the data using logarithms.
   *Non-linear, take logs to linearize relationship. It show also on plot (b)*
Problem 3  (20 points)

Part A:
Data on cadmium ($x$) and cobalt ($y$) of soil at a certain area gave the following information:

\[
\begin{align*}
\sum_{i=1}^{6} x_i &= 9.545, \\
\sum_{i=1}^{6} y_i &= 61.668, \\
\sum_{i=1}^{6} x_i^2 &= 15.78468, \\
\sum_{i=1}^{6} y_i^2 &= 719.9573, \\
\sum_{i=1}^{6} x_i y_i &= 96.43722.
\end{align*}
\]

It was discovered that the last two $x$ values were interchanged. The incorrect table is shown below:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.145</td>
<td>16.320</td>
</tr>
<tr>
<td>1.565</td>
<td>3.508</td>
</tr>
</tbody>
</table>

\[\sum_{i=1}^{6} x_i^2 \cdot \sum_{i=1}^{6} y_i - (\sum_{i=1}^{6} x_i \cdot \sum_{i=1}^{6} y_i) \Rightarrow \hat{\beta}_1 = 6.1895\]

Part B:
Data on cadmium ($x$) and cobalt ($y$) of soil at a certain area gave the following information:

\[
\begin{align*}
\sum_{i=1}^{6} x_i &= 11.908, \\
\sum_{i=1}^{6} y_i &= 69.305, \\
\sum_{i=1}^{6} x_i^2 &= 26.77971, \\
\sum_{i=1}^{6} y_i^2 &= 708.9622, \\
\sum_{i=1}^{6} x_i y_i &= 101.8183.
\end{align*}
\]

It was discovered that a pair of $x$ and $y$ was interchanged. The incorrect table is shown below:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.508</td>
<td>1.145</td>
</tr>
</tbody>
</table>

\[\sum_{i=1}^{6} x_i^2 \cdot \sum_{i=1}^{6} y_i - (\sum_{i=1}^{6} x_i \cdot \sum_{i=1}^{6} y_i) \Rightarrow \hat{\beta}_1 = 6.1895\]

SAME DATA WERE USED FOR BOTH QUESTIONS.
Problem 4  (20 points)
Answer the following questions:

a. Plot the ecdf of the following numbers: 1, 13, 9, 8, 10, 8.

```
\begin{align*}
  x & \quad F(x) \\
  1 & \quad 0.1667 \\
  8 & \quad 0.5 \\
  9 & \quad 0.6667 \\
  10 & \quad 0.8333 \\
  13 & \quad 1
\end{align*}
```

b. Consider the body fat data (see handout “Introduction to R”). We use here the variables \( y, x_6, x_9, x_{10} \). The variance covariance matrix and the sample means of the four variables are shown next:

```
# Variance covariance matrix:
\begin{align*}
  y & \quad x_6 & \quad x_9 & \quad x_{10} \\
  y & 70.036 & 9.980 & 37.483 & 24.587 \\
  x_6 & 9.980 & 5.909 & 12.799 & 8.879 \\
  x_9 & 37.483 & 12.799 & 51.324 & 33.715 \\
  x_{10} & 24.587 & 8.879 & 33.715 & 27.562
\end{align*}
```

```
# Sample means:
\begin{align*}
  y & \quad x_6 & \quad x_9 & \quad x_{10} \\
\end{align*}
```

Find the fitted line of the regression of \( y \) on \( x_9 \). Show all your work.

\[
\hat{\beta}_9 = \frac{\text{Cov}(y, x_9)}{\text{Var}(x_9)} = \frac{3.7483}{51.324} \Rightarrow \hat{\beta}_9 = 0.073
\]

\[
\hat{\beta}_c = \bar{y} - \hat{\beta}_9 \bar{x}_9 = 19.151 - 0.73 \left( 99.905 \right) \Rightarrow \hat{\beta}_0 = 73.78
\]

\[
\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_9 x_9
\]

c. Refer to part (b). Compute the correlation coefficient between \( y \) and \( x_{10} \).

\[
\text{Cov}(y, x_{10}) = \frac{\text{Cov}(y, x_{10})}{\text{Var}(y) \text{Var}(x_{10})} = \frac{24.587}{127.562} \approx 0.5596
\]

d. Refer to part (b). Consider the regression of \( y \) on \( x_6 \). Compute the value of the regression sum of squares.

\[
SSR = \hat{\beta}^2 \sum (x_i - \bar{x}_6)^2 = 1.69 \sum (1477.75) \Rightarrow SSR = 4219.74
\]

\[
\hat{\beta}_6 = \frac{\text{Cov}(y, x_6)}{\text{Var}(x_6)} \quad \text{Var}(x_6) = \frac{\sum (x_i - \bar{x}_6)^2}{n-1}
\]

\[
= \frac{9.780}{5.909} = 1.69 \quad \sum (x_i - \bar{x}_6)^2 = (n-1) \sum y_i - \left( \bar{y} \right) \left( \bar{x}_6 \right) = 1477.23
\]
Problem 5  (20 points)

Normal temperature (degrees Fahrenheit) readings (x) and heart rates (beats per minute) of 65 males and 65 females were used in this problem.

a. Consider the data for males: The OLS estimate of $\beta_1$ is $\hat{\beta}_1 = 1.66$. A permutation test was run to test the hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$ of the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Under the null hypothesis the permutation test produced the following histogram. What is your conclusion?

$b_0 \sqrt{n} \approx t(\alpha/2) \Rightarrow H_0.$

a. Consider the data for females: The OLS estimate of $\beta_1$ is $\hat{\beta}_1 = 3.13$. A permutation test was run to test the hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$ of the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Under the null hypothesis the permutation test produced the following histogram. What is your conclusion?

Two sided test

$p-value \approx \frac{2 \times 10^{-6}}{10000} = 0.000002 \Rightarrow H_0.$