Homework for 4/14

1. Consider the linear model

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \]

where \( \varepsilon_i \)'s i.i.d \( \mathcal{N}(0, \sigma^2) \) random variables. Show that the RSS (Residual Sum of Squares) satisfies

\[
\text{RSS} = \left[\sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \right] - \frac{n \left[ \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right) \right]^2}{n^2 \sum x_i^2 - n \left( \sum x_i \right)^2}
\]

\( \text{(Hint: use the facts that } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \text{ and } \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y}.) \)

First we have

\[
\text{RSS} = \sum \left[ (y_i - \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 \right] = \sum \left[ (y_i - \bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i \right]^2
\]

\[
= \sum \left[ (y_i - \bar{y})^2 - 2\hat{\beta}_1 (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2 (x_i - \bar{x})^2 \right]
\]

\[
= \sum (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2
\]

\[
= \sum (y_i - \bar{y})^2 - 2 \left( \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right) \sum (y_i - \bar{y})(x_i - \bar{x})
\]

\[
+ \left( \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right)^2 \sum (x_i - \bar{x})^2
\]

\[
= \sum (y_i - \bar{y})^2 - \frac{\left[ \sum (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum (x_i - \bar{x})^2}
\]

\[
= \sum y_i^2 - n \bar{y}^2 - \frac{\left[ \sum x_i y_i - n \bar{x} \bar{y} \right]^2}{\sum x_i^2 - n \bar{x}^2}
\]

\[
= \left[ \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \right] - \frac{1}{n^2} \left[ \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right) \right]^2
\]

\[
= \left[ \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \right] - \frac{n \left[ \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right) \right]^2}{n^2 \sum x_i^2 - n \left( \sum x_i \right)^2}. \]
2. During oil drilling operations, components of the drilling assembly may suffer from sulfide stress cracking. The article “Composition Optimization of High-Strength Steels for Sulfide Cracking Resistance Improvement” (Corrosion Science, 2009: 2878-2884) reported on a study in which the composition of a standard grade of steel was analyzed. The following data on $y = \text{threshold stress (\% SMYS)}$ and $x = \text{yield strength (MPa)}$ was read from a graph in the article (which also included the equation of the least squares line).

<table>
<thead>
<tr>
<th>$x$</th>
<th>635</th>
<th>644</th>
<th>711</th>
<th>708</th>
<th>836</th>
<th>820</th>
<th>810</th>
<th>870</th>
<th>856</th>
<th>923</th>
<th>878</th>
<th>937</th>
<th>948</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>100</td>
<td>93</td>
<td>88</td>
<td>84</td>
<td>77</td>
<td>75</td>
<td>74</td>
<td>63</td>
<td>57</td>
<td>55</td>
<td>47</td>
<td>43</td>
<td>38</td>
</tr>
</tbody>
</table>

\[ \sum x_i = 10576, \quad \sum y_i = 894, \quad \sum x_i^2 = 8741264, \]
\[ \sum y_i^2 = 66224, \quad \sum x_i y_i = 703192. \]

a. Find the least square linear model.
b. Compute the estimated standard deviation.
c. Calculate a confidence interval using confidence level 95\% for the expected change in stress associated with a 1 MPa increase in strength.

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \]

\[ \hat{\beta}_0 = \frac{\left( \sum_{i=1}^{n} x_i^2 \right) \left( \sum_{i=1}^{n} y_i \right) - \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} x_i y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2} \]
\[ = \frac{8741264 \cdot 894 - 10576 \cdot 703192}{13 \cdot 8741264 - 10576^2} \]
\[ = 211.655, \]

\[ \hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2} \]
\[ = \frac{13 \cdot 703192 - 10576 \cdot 894}{13 \cdot 8741264 - 10576^2} \]
\[ = -0.176. \]

Therefore, the least square linear model is
\[ \hat{y} = 211.655 - 0.176x. \]
b. According to problem 1, the residual sum of squares is given by

\[
\text{RSS} = \left[ \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \right] - \frac{n \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right)}{n^2 \sum x_i^2 - n \left( \sum x_i \right)^2}^2
\]

\[
= 66224 - \frac{1}{13} \cdot 894^2 \frac{(13 \cdot 703192 - 10576 \cdot 894)^2}{13^2 \cdot 8741264 - 13 \cdot 10576^2}
\]

\[
= 509.505.
\]

Thus the estimated standard deviation is

\[
s = \sqrt{\frac{\text{RSS}}{n - 2}} = \sqrt{\frac{509.505}{13 - 2}} = 6.806.
\]

c. The slope \( \beta_1 \) describes the expected change in stress associated with a 1 MPa increase in strength. The estimated standard deviation for \( \hat{\beta}_1 \) is

\[
s_{\hat{\beta}_1} = \sqrt{\frac{n s^2}{n \sum x_i^2 - \left( \sum x_i \right)^2}} = \sqrt{\frac{13 \cdot 6.806^2}{13 \cdot 8741264 - 10576^2}} = 0.0184.
\]

Thus a 95% confidence interval for \( \beta_1 \) is

\[
(\hat{\beta}_1 - t_{0.975,11} \cdot s_{\hat{\beta}_1}, \hat{\beta}_1 - t_{0.975,11} \cdot s_{\hat{\beta}_1})
\]

\[
= (-0.176 - 2.201 \cdot 0.0184, -0.176 + 2.201 \cdot 0.0184)
\]

\[
= (-0.216, -0.136).
\]
3. How does lateral acceleration—side forces experienced in turns that are largely under driver control—affect nausea as perceived by bus passengers? The article “Motion Sickness in Public Road Transport: The Effect of Driver, Route, and Vehicle” (Ergonomics, 1999: 1646-1664) reported data on $x =$ motion sickness dose (calculated in accordance with a British standard for evaluating similar motion at sea) and $y =$ reported nausea (%). Relevant summary quantities are

$$n = 17, \quad \sum x_i = 222.1, \quad \sum y_i = 193, \quad \sum x_i^2 = 3056.69, \quad \sum y_i^2 = 2975, \quad \sum x_i y_i = 2759.6.$$  

Values of dose in the sample ranged from 6.0 to 17.6.

a. Assuming that the simple linear regression model is valid for relating these two variables (this is supported by the raw data), calculate an estimate of the slope parameter that conveys information about the precision and reliability of estimation.

b. Does it appear that there is a useful linear relationship between these two variables (that is, the slope is not 0) at $\alpha = 0.05$?

a. The estimate for the slope is

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right)}{n \sum x_i^2 - \left( \sum x_i \right)^2} = \frac{17 \cdot 2759.6 - 222.1 \cdot 193}{17 \cdot 3056.69 - 222.1^2} = 1.536.$$  

Since

$$\text{RSS} = \left[ \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \right] - \frac{n \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right)}{n^2 \sum x_i^2 - n \left( \sum x_i \right)^2}^2 = 418.138.$$  

Thus the estimated standard deviation is

$$s = \sqrt{\frac{\text{RSS}}{n - 2}} = \sqrt{\frac{418.138}{17 - 2}} = 5.28.$$
and the estimated standard deviation for \( \hat{\beta}_1 \) is

\[
s_{\hat{\beta}_1} = \sqrt{\frac{\frac{n s^2}{n \sum x_i^2 - (\sum x_i)^2}}{17 \cdot 5.28^2}} = \frac{\sqrt{17 \cdot 5.28^2}}{17 \cdot 3056.69 - 222.12} = 0.424.
\]

Therefore, a 100(1 - \( \alpha \))% confidence interval for \( \beta \) is given by

\[
(\hat{\beta}_1 - t_{1-\alpha/2,n-2} \cdot s_{\hat{\beta}_1}, \hat{\beta}_1 + t_{1-\alpha/2,n-2} \cdot s_{\hat{\beta}_1})
\]

\[
= (1.536 - t_{1-\alpha/2,15} \cdot 0.424, 1.536 + t_{1-\alpha/2,15} \cdot 0.424).
\]

For example, a 95% \( (t_{0.975,15}) \) confidence interval for \( \beta_1 \) is

\[
(1.536 - 2.131 \cdot 0.424, 1.536 + 2.131 \cdot 0.424) = (0.63, 2.44).
\]

b. The hypotheses are

\[
H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0.
\]

The test statistic is

\[
T = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}},
\]

and the rejection region is \( R = \{|T| > t_{1-\alpha/2,n-2}\} \). Therefore,

\[
T = \frac{1.536}{0.424} = 3.623.
\]

Since \( t_{0.975,15} = 2.131 \), we have \( T > t_{0.975,15} \). Therefore, we should reject \( H_0 \). In other words, it appears that there is a useful linear relationship between these two variables.
Homework for 4/16

1. Let \( \{Z_t, t = 0, \pm 1, \pm 2, \ldots \} \) be a sequence of independent random variables, each with mean \( \mathbb{E}[Z_t] = 0 \) and variance \( \text{Var}[Z_t] = \sigma^2 \). Let \( a, b, \) and \( c \) be real constants. Compute the mean and autocovariance function of the following processes, and identify the stationary processes. 

(Hint: Section 4.3 may be helpful. Especially, Theorem A, Corollary A, and Corollary B.)

a. \( X_t = a + bZ_t + cZ_{t-2} \).

b. \( X_t = Z_1 \cos(ct) + Z_2 \sin(ct) \).

c. \( X_t = Z_t \cos(ct) + Z_{t-1} \sin(ct) \).

d. \( X_t = a + bZ_0 \).

e. \( X_t = Z_0 \cos(ct) \).

e. \( X_t = Z_t Z_{t-1} \).

\( a. \) \( X_t = a + bZ_t + cZ_{t-2} \).

\[ \mathbb{E}[X_t] = \mathbb{E}[a + bZ_t + cZ_{t-2}] = a + b\mathbb{E}[Z_t] + c\mathbb{E}[Z_{t-2}] = a. \]

\[ \text{Cov}[X_{t+h}, X_t] = \text{Cov}[a + bZ_{t+h} + cZ_{t+h-2}, a + bZ_t + cZ_{t-2}] \]
\[ = b^2 \text{Cov}[Z_{t+h}, Z_t] + bc \text{Cov}[Z_{t+h}, Z_{t-2}] \]
\[ + bc \text{Cov}[Z_t, Z_{t+h-2}] + c^2 \text{Cov}[Z_{t+h-2}, Z_{t-2}] \]
\[ = \begin{cases} (b^2 + c^2)\sigma^2, & h = 0 \\ b c \sigma^2, & h = \pm 2 \\ 0, & \text{otherwise} \end{cases} \]

Thus \( \{X_t\} \) is stationary.

b. \( X_t = Z_1 \cos(ct) + Z_2 \sin(ct) \).

\[ \mathbb{E}[X_t] = \mathbb{E}[Z_1 \cos(ct) + Z_2 \sin(ct)] = \mathbb{E}[Z_1] \cos(ct) + \mathbb{E}[Z_2] \sin(ct) = 0. \]

\[ \text{Cov}[X_{t+h}, X_t] \]
\[ = \text{Cov}[Z_1 \cos(c(t + h)) + Z_2 \sin(c(t + h)), Z_1 \cos(ct) + Z_2 \sin(ct)] \]
\[ = \text{Var}[Z_1] \cos(c(t + h)) \cos(ct) + \text{Cov}[Z_1, Z_2] \cos(c(t + h)) \sin(ct) \]
\[ + \text{Cov}[Z_1, Z_2] \sin(c(t + h)) \cos(ct) + \text{Var}[Z_2] \sin(c(t + h)) \sin(ct) \]
\[ = \sigma^2 \cos(c(t + h)) \cos(ct) + \sigma^2 \sin(c(t + h)) \sin(ct) \]
\[ = \sigma^2 \cos(ch). \]

Thus \( \{X_t\} \) is stationary.
c. $X_t = Z_t \cos(ct) + Z_{t-1} \sin(ct)$.

$\mathbb{E}[X_t] = \mathbb{E}[Z_t \cos(ct) + Z_{t-1} \sin(ct)] = \mathbb{E}[Z_t] \cos(ct) + \mathbb{E}[Z_{t-1}] \sin(ct) = 0$.

$\text{Cov}[X_{t+h}, X_t]$

$= \text{Cov}[Z_{t+h} \cos(c(t+h)) + Z_{t+h-1} \sin(c(t+h)), Z_t \cos(ct) + Z_{t-1} \sin(ct)]$

$= \text{Cov}[Z_{t+h}, Z_t] \cos(c(t+h)) \cos(ct)$

$+ \text{Cov}[Z_{t+h}, Z_{t-1}] \cos(c(t+h)) \sin(ct)$

$+ \text{Cov}[Z_{t+h-1}, Z_t] \sin(c(t+h)) \cos(ct)$

$+ \text{Cov}[Z_{t+h-1}, Z_{t-1}] \sin(c(t+h)) \sin(ct)$

$= \begin{cases} 
\sigma^2, & h = 0 \\
\sigma^2 \sin(ct + c) \cos(ct), & h = 1 \\
\sigma^2 \sin(ct) \cos(ct + c), & h = -1 \\
0, & \text{otherwise} 
\end{cases}$

Thus $\{X_t\}$ is not stationary.

d. $X_t = a + bZ_0$.

$\mathbb{E}[X_t] = \mathbb{E}[a + bZ_0] = a + b\mathbb{E}[Z_0] = a$.

$\text{Cov}[X_{t+h}, X_t] = \text{Cov}[a + bZ_0, a + bZ_0]$

$= \text{Var}[a + bZ_0] = b^2 \text{Var}[Z_0]$

$= b^2 \sigma^2$.

Thus $\{X_t\}$ is stationary.

e. $X_t = Z_0 \cos(ct)$.

$\mathbb{E}[X_t] = \mathbb{E}[Z_0 \cos(ct)] = \mathbb{E}[Z_0] \cos(ct) = 0$.

$\text{Cov}[X_{t+h}, X_t] = \text{Cov}[Z_0 \cos(c(t+h)), Z_0 \cos(ct)]$

$= \cos(c(t+h)) \cos(ct) \text{Var}[Z_0]$

$= \cos(c(t+h)) \cos(ct) \sigma^2$.

Thus $\{X_t\}$ is not stationary.

e. $X_t = Z_t Z_{t-1}$.

$\mathbb{E}[X_t] = \mathbb{E}[Z_t Z_{t-1}] = \mathbb{E}[Z_t] \mathbb{E}[Z_{t-1}] = 0$.

$\text{Cov}[X_{t+h}, X_t] = \text{Cov}[Z_t Z_{t+h-1}, Z_t Z_{t-1}]$

$= \mathbb{E}[Z_t Z_{t+h-1} Z_t Z_{t-1}] - \mathbb{E}[Z_t Z_{t+h-1}] \mathbb{E}[Z_t Z_{t-1}]$

$= \mathbb{E}[Z_t Z_{t+h-1} Z_t Z_{t-1}]$

$= \begin{cases} 
\sigma^4, & h = 0 \\
0, & h \neq 0 
\end{cases}$

Thus $\{X_t\}$ is stationary.