Solution week 11

1. A study was constructed to determine whether a linear relationship exists between the age \( x \) and the diameter \( y \) in inches at 1.5 meters level of chestnut trees. Ten randomly selected chestnut trees were examined and the data were recorded as follows:

<table>
<thead>
<tr>
<th>Age ((x))</th>
<th>Diameter ((y)) at 1.5 meters level</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
</tr>
<tr>
<td>8</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>2.3</td>
</tr>
<tr>
<td>10</td>
<td>3.2</td>
</tr>
<tr>
<td>12</td>
<td>4.9</td>
</tr>
<tr>
<td>13</td>
<td>3.7</td>
</tr>
<tr>
<td>14</td>
<td>4.5</td>
</tr>
<tr>
<td>17</td>
<td>5.6</td>
</tr>
<tr>
<td>19</td>
<td>6.5</td>
</tr>
</tbody>
</table>

\[
\sum_i x_i^2 = 1428, \quad \sum_i y_i^2 = 149.37, \quad \sum_i x_i y_i = 454.2.
\]

(a) Find a regression line fitted to the data.
(b) Test whether or not a linear relationship between the age and the diameter at 1.5 meters level of chestnut trees is significant (using t-test or F-test).
(c) Find an estimate of average diameter corresponding to age 30 and give a 95% prediction interval of this estimate.
(d) Find an estimate of the diameter for some special chestnut tree at age 30 and give a 95% prediction interval of this estimate.

Solution:

(a) We have that \( n = 10 \) and

\[
\bar{x} = 11, \quad \bar{y} = 3.33
\]

\[
S_{xx} = 1428 - 10 \times 11^2 = 218
\]

\[
S_{yy} = 149.37 - 10 \times 3.33^2 = 38.481
\]

\[
S_{xy} = 454.2 - 10 \times 11 \times 3.33 = 87.9.
\]
This implies that
\[ \hat{\beta} = \frac{S_{xy}}{S_{xx}} = 0.403211, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = -1.105321. \]

Hence the least squares line is
\[ \hat{y} = \hat{\alpha} + \hat{\beta} x = -1.105321 + 0.403211 x. \]

(b) The hypothesis to be tested is
\[ H : \beta = 0 \quad \text{vs} \quad H_A : \beta \neq 0. \]

Use the t-test. A observed value of the t-test with \( n = 8 \) is given by
\[ t_0 = \frac{\sqrt{8} S_{xy}}{\sqrt{S_{xx} S_{yy} - S_{xy}^2}} = 9.66. \]

The corresponding \( p \)-value is
\[ p = 2P(t_8 \geq 9.66) = 1.098279e - 05. \]

Conclusion: There are very strong evidence against the null hypothesis, that is, there is very strong evidence to support the claim that there is a relationship between the age and the diameter at 1.5 meters level of chestnut trees.

Instead, we may get the Regression ANOVA table
\[
\begin{array}{l|ccc}
\text{Source} & \text{df} & \text{SS} & \text{MS} & \text{F} \\
\hline
\text{Regression} & 1 & 35.44 & 35.44 & 93.23 \\
\text{Residuals} & 8 & 3.041 & 0.38 & \\
\text{Total} & 9 & 38.481 & & \\
\end{array}
\]

The corresponding \( p \)-value is
\[ p = P(F_{1,8} \geq 93.23) = 1.101921e - 05. \]

Conclusion: There are very strong evidence against the null hypothesis, that is, there is very strong evidence to support the claim that there is a relationship between the age and the diameter at 1.5 meters level of chestnut trees.

(c) Recall that the regression line is
\[ \hat{y} = \hat{\alpha} + \hat{\beta} x = -1.105321 + 0.403211 x. \]

Hence an estimator of average diameter corresponding to age 30 is given by
\[ \hat{y} = -1.105321 + 0.403211 \times 30 = 10.991. \]
A 95% prediction interval for the average diameter (at \( x_0 = 30 \)) is given by

\[
\left[ \hat{y} - t_{0.025} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, \ \hat{y} + t_{0.025} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right],
\]

where \( t_{0.025} \) satisfies that \( P(t_{n-2} \geq t_{0.025}) = 0.025 \).

Note that \( t_{0.025} = 2.306 \) (with \( n = 10 \)) and recall that \( \bar{x} = 11, \ S_{xx} = 218 \) and \( s^2 = 0.38 \) (see the Regression ANOVA table). We obtain \( s = 0.6165 \),

\[
\hat{y} - t_{0.025} s \sqrt{\frac{1}{10} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 9.1077
\]

\[
\hat{y} + t_{0.025} s \sqrt{\frac{1}{10} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 12.875
\]

This gives the 95% prediction interval of the average diameter

\[ [9.1077, \ 12.875]. \]

(d) According to the fitted line, the expected diameter of some special chestnut tree at age 30 still is

\[ \hat{y} = -1.105321 + 0.403211 \times 30 = 10.991. \]

With \( x_0 = 30 \),

\[
\hat{y} - t_{0.025} s \sqrt{1 + \frac{1}{10} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 8.631
\]

\[
\hat{y} + t_{0.025} s \sqrt{1 + \frac{1}{10} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 13.351
\]

This gives the 95% prediction interval of the some special chestnut tree at age 30:

\[ [8.631, \ 13.351]. \]

2. Let \( \hat{\alpha} \) and \( \hat{\beta} \) be the least-squares estimates in simple linear regression model:

\[ Y_i = \alpha + \beta x_i + \epsilon_i, \ \ i = 1, 2, \ldots, n \]

where \( \epsilon_i \) are iid random variables with zero mean and finite variance \( \sigma^2 \).

Show that

\[ \mathbf{E} \hat{\alpha} = \alpha, \ \ \mathbf{E} \hat{\beta} = \beta \ \ \text{and} \ \ Var(\hat{\beta}) = \sigma^2/S_{xx}. \]

Hint: \( S_{xx} = \sum_i (x_i - \bar{x})x_i \) and

\[ Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) \]

provided \( X \) and \( Y \) are independent.
Solution: Under the given model, we have that

\[ EY_i = \alpha + \beta x_i, \quad Y_i - EY_i = \epsilon_i, \quad i = 1, 2, \ldots, n. \]

\[ \hat{\beta} = \sum_i (x_i - \bar{x})Y_i/S_{xx}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}. \]

These, together with the properties of expectations, imply that

\[ \mathbb{E}\hat{\beta} = \sum_i (x_i - \bar{x})EY_i/S_{xx} \]

\[ = \frac{1}{S_{xx}} \sum_i (x_i - \bar{x})(\alpha + \beta x_i) \]

\[ = \frac{\alpha}{S_{xx}} \sum_i (x_i - \bar{x}) + \frac{\beta}{S_{xx}} \sum_i (x_i - \bar{x})x_i \]

\[ = 0 + \frac{\beta}{S_{xx}} S_{xx} = \beta, \]

\[ E\bar{Y} = \frac{1}{n} \sum_i EY_i = \frac{1}{n} \sum_i (\alpha + \beta x_i) = \alpha + \beta \bar{x}, \]

\[ \mathbb{E}\hat{\alpha} = \mathbb{E}\bar{Y} - \bar{x} \mathbb{E}\hat{\beta} = \alpha + \beta \bar{x} - \beta \bar{x} = \alpha, \]

\[ \text{Var}(\hat{\beta}) = \mathbb{E}\left( \hat{\beta} - \mathbb{E}\hat{\beta}\right)^2 \]

\[ = \mathbb{E}\left( \frac{1}{S_{xx}} \sum_i (x_i - \bar{x})(Y_i - EY_i) \right)^2 \]

\[ = \frac{1}{S_{xx}^2} \mathbb{E}\left( \sum_i (x_i - \bar{x})\epsilon_i \right)^2 \]

\[ = \frac{1}{S_{xx}^2} \mathbb{E}\left( \sum_{i \neq j} (x_i - \bar{x})(x_j - \bar{x})\epsilon_i\epsilon_j + \sum_i (x_i - \bar{x})^2 \epsilon_i^2 \right) \]

\[ = \frac{1}{S_{xx}^2} \left( \sum_{i \neq j} (x_i - \bar{x})(x_j - \bar{x})\mathbb{E}\epsilon_i\epsilon_j + \sum_i (x_i - \bar{x})^2 \mathbb{E}\epsilon_i^2 \right) \]

\[ = \frac{1}{S_{xx}^2} \left( 0 + \sigma^2 S_{xx} \right) = \frac{\sigma^2}{S_{xx}}, \]

where we have used the results: \( \mathbb{E}\epsilon_i\epsilon_j = 0 \) for \( i \neq j \) (because of independence) and \( \text{var}(\epsilon_j) = \mathbb{E}\epsilon_j^2 = \sigma^2 \).
1. Consider the data frame `fuel.frame`, which has information on makes of cars taken from the April 1990 issue of Consumer Reports.

(a) Set a 2 by 2 graphic window and prepare your data for analysis by typing `attach(fuel.frame)`.

(b) Create two vectors \( x \) and \( y \) whose elements correspond to Weight (in pounds) and Fuel (in gallons per 100 miles) on the data `fuel.frame` respectively.

(c) Plot the Fuel against Weight with Weight on the \( x \)-axis and comment on the plot.

(d) Find the regression line for Fuel on Weight using `lm` and draw this line on the first plot (Hint: use `abline`).

(e) Test whether the weight has an influence on the fuel in terms of \( p \)-value.

(f) Plot the residuals against the fitted values of fuel and draw in the line \( y = 0 \).

(g) Obtain a boxplot of residuals and a qq-plot of the residuals.

(h) Comment on the plots in (f) and (g), and explain whether the assumptions required in (e) are probably satisfied.

Solution:

```r
# (a)
par(mfrow=c(2,2))
attach(fuel.frame)

# (b)
x<- fuel.frame[,1]  # give the Weight
y<- fuel.frame[,4]  # give the Fuel

# (c)
plot(x,y, xlab="Weight", ylab="Fuel")

# The scatterplot shows clearly that fuel used in gallons per 100 miles
# increases as weight of a car increases, and a linear model might be
# suitable to describe the
# relationship between fuel and weight.

# (d)
```
Call:
  lm(formula = y ~ x)

Coefficients:
  (Intercept)   x
  0.3914324  0.00131638

# The regression line is given by
# y = 0.3914324 + 0.00131638 x

abline(lm(y~x))

# (e)

# The hypothesis to be test is
# \beta=0 vs \beta\not= 0

# first calculate Sxx, Syy and Sxy

> c(mean(x),mean(y))
  [1] 2900.833333  4.210033

> Sxx<-sum((x-mean(x))*(x-mean(x)))
> Sxy<-sum((x-mean(x))*(y-mean(y)))
> Syy<-sum((y-mean(y))*(y-mean(y)))

> c(Sxx, Sxy, Syy)
  [1] 1.450711e+07  1.909687e+04  3.385687e+01

> n<-length(x)
> t0<-sqrt(n-2)*Sxy/sqrt(Sxx*Syy-Sxy^2)
  # give a observed value of t-statistic

> p<-2*(1-pt(t0,n-2))  # give p-value of the test
> p
  [1] 0

# Conclusion: There are very strong evidence that
# the weight of a car has an influence on the fuel.
# (f)

fitted<-y- lsfit(x,y)$res
# give the fitted value
plot(fitted, lsfit(x,y)$res)
# give the plot: residual against the fitted value

abline(0)
# combine the line y=0 into the plot

# (g)

boxplot(lsfit(x,y)$res)
qqnorm(lsfit(x,y)$res)
qqline(lsfit(x,y)$res)

# (h)

# The plot of the residuals against the fitted value has no apparent pattern,
# which implies that a linear regression model is appropriate to describe
# the relationship between the weight and the fuel.

# The boxplot of the residuals looks symmetric and the qq-normal plot
# of the residuals looks quite linear
# Hence it is reasonable to believe the data satisfies the assumptions
# required for the test in (e). The assumptions required for the test are
# that the errors are iid normal random variables with zero mean and finite variance.