Stat 331 Applied Linear Models – Assignment 1
Solution

1(a) 4 points

$$\begin{aligned}
cov(y_i, \hat{\beta}_0 + x_i\hat{\beta}_1) &= cov(y_i, \bar{y} - \bar{x}\hat{\beta}_1) + cov(y_i, x_i\hat{\beta}_1) \\
&= cov(y_i, \frac{1}{n}\sum y_j) - \bar{x}cov(y_i, \hat{\beta}_1) + x_i cov(y_i, \hat{\beta}_1) \\
&= cov(y_i, \frac{1}{n}y_i) + (x_i - \bar{x})cov(y_i, \sum c_j y_j) \\
&= \frac{1}{n}\sigma^2 + (x_i - \bar{x})cov(y_i, c_i y_i) \\
&= \frac{1}{n}\sigma^2 + (x_i - \bar{x})c_i\sigma^2 \\
&= \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]\sigma^2,
\end{aligned}$$

where $c_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}$ as introduced in class. The third step is because by assumption (iii), $cov(y_i, y_j) = 0$ for $i \neq j$

1(b) 4 points

$$\begin{aligned}
E(r_i) &= E(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= E(y_i) - E(\hat{\beta}_0) - x_i E(\hat{\beta}_1) \\
&= \beta_0 + x_i\beta_1 + 0 - \beta_0 - x_i\beta_1 = 0.
\end{aligned}$$

The second last step is because by Assumption (i), $E(\epsilon_i) = 0$, and by what we proved in class, $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$.

$$\begin{aligned}
V(r_i) &= V(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1) \\
&= V(y_i) + V(\hat{\beta}_0 + x_i\hat{\beta}_1) - 2cov(y_i, \hat{\beta}_0 + x_i\hat{\beta}_1) \\
&= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] - 2cov(y_i, \hat{\beta}_0 + x_i\hat{\beta}_1),
\end{aligned}$$

where the second term in the last step is proved in class (in a very similar manner).

Hence

$$V(r_i) = \sigma^2 - \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] = \sigma^2[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}].$$

1(c) 2 points

$$\begin{aligned}
E(\sum r_i^2) &= \sum[V(r_i) - (E(r_i))^2] \\
&= \sigma^2\left[n - 1 - \frac{S_{xx}}{S_{xx}}\right] \\
&= \sigma^2(n - 2).
\end{aligned}$$

2(a) 2 points

We are to minimize

$$S(\beta) = \sum (y_i - \beta x_i)^2.$$

Take partial derivative with respect to β,

$$\frac{\partial S(\beta)}{\partial \beta} = -2 \sum x_i (y_i - \beta x_i) = 0.$$

Therefore,

$$\begin{aligned} \sum x_i (y_i - \beta x_i) = 0, \\ \sum x_i y_i = \beta \sum x_i^2, \\ \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}. \end{aligned}$$

2(b) 4 points

$$\begin{aligned} E(\hat{\beta}) &= E(\frac{\sum x_i y_i}{\sum x_i^2}) \\ &= \frac{\sum x_i (\beta x_i)}{\sum x_i^2} \\ &= \beta. \end{aligned}$$

$$\begin{aligned} V(\hat{\beta}) &= \frac{\sum V(x_i y_i)}{(\sum x_i^2)^2} \\ &= \frac{\sum x_i^2 \sigma^2}{(\sum x_i^2)^2} \\ &= \frac{\sigma^2}{\sum x_i^2}. \end{aligned}$$

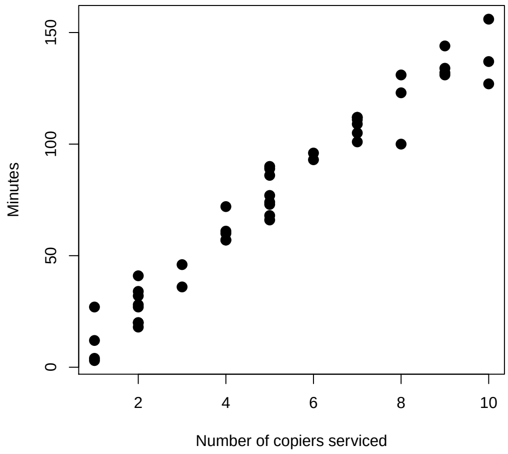
2(c) 4 points

(i) $\sum r_i x_i = 0$ can be obtained directly from step (a) in this question.

(ii) Not necessarily true. Note that $\sum r_i = \sum y_i - \hat{\beta} \sum x_i$. One can easily construct an example such that $\sum y_i \neq 0$ and $\sum x_i = 0$, in which case $\sum r_i = \sum y_i \neq 0$.

The code is provided after the solutions.

3(a) 2 points



It looks OK, as (1) the linear relationship assumption looks good, and (2) the variance of the random error looks consistent for all x values.

3(b) 2 points

$\hat{y} = -0.58 + 15.035x$.

3(c) 1 point

15.035.

3(d) 2 points

$(14.061, 16.009)$

3(e) 4 points

$H_0 : \beta_1 = 0$ vs. $H_\alpha : \beta_1 \neq 0$.
T statistic: 31.12
p-value: very close to 0.
Conclusion: reject the null hypothesis.

3(f) 2 points

Yes. First, notice that the confidence level 0.95 in (d) plus the significance level 0.05 in (e) equals 1, thus we can use the confidence interval in (d) for hypothesis testing in (e). As $0 \notin (14.061, 16.009)$, we should reject the null hypothesis, which is consistent with the p-value approach in (e).

3(g) 3 points

74.596
(56.421,92.771)

```
setwd("") ## set your own work directory

copier = read.table("Copier.txt",header=T)

x = copier$Serviced

y = copier$Minutes

### (a)

plot(x,y, main="", xlab="Number of copiers serviced",
    ylab="Minutes", pch=19,cex=1.4)

### (b)

n=length(x)

b1 = (sum(x*y)-n*mean(x)*mean(y))/(sum(x^2)-n*(mean(x))^2)

b0 = mean(y)-b1*mean(x)

print(paste("y=",round(b1, digits = 3),
    "x",round(b0, digits = 3),sep=""))

### (c)

b1

### (d)

yhat = b0+b1*x

resid = y-yhat

s2 = sum(resid^2)/(n-2)

sxx = sum(x^2) - n*(mean(x))^2

se.b1 = sqrt(s2/sxx)

t_star = qt((1-(1-0.95)/2),df=n-2)

lower = round(b1-t_star*se.b1,digits=3)
upper = round(b1+t_star*se.b1,digits=3)

print(paste("(",lower,",",upper,")",sep=""))

### (e)
```

```
tstats = b1/se.b1

pvalue = 2*(1-pt(tstats, n-2))

pvalue ## p-value approach

tstats > t_star ## critical value approach

### (f)

### (g)

yp = b1*5+b0

yp

se.yp = sqrt( (1+1/n+(5-mean(x))^2/sxx)*s2)

lower = round(yp-t_star*se.yp,digits=3)
upper = round(yp+t_star*se.yp,digits=3)

print(paste("(",lower,",",upper,")",sep=""))
```