Statistical Linear Models

Assignment 1

Hanbin Liu 11912410

Problem 1

(a)

Data: $\mathcal{D} = \{(y_i, x_i), i = 1, ..., n\}$. The regression model is $y = \beta_0 + \beta_1 x + \epsilon$ with $\epsilon \overset{i.i.d}{\sim} (0, \sigma^2)$. The model for the data is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} (0, \sigma^2).$$

Some basic assumptions to the model are: 1) y_i and x_i have a linear relation; 2) $y_i^{'}s$(or $\epsilon_i^{'}s$) are independent, $i = 1, ..., n$; 3) $\mathrm{Var}(y_i) = \mathrm{Var}(\epsilon_i) = \sigma^2$; and $y_i \overset{i.n.d}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$ is an optional assumption.

If $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$, then we can apply MLE to all the unknown parameters. The derivation are as follows:

$$\begin{aligned}
&y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2),\ i = 1, ..., n \\
&p(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\} \triangleq \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{u_i^2}{2\sigma^2}\},\ u_i = y_i - \beta_0 - \beta_1 x_i.
\end{aligned}$$

Then,

$$\begin{aligned}
\log - \text{likelihood} = l(\beta_0, \beta_1, \sigma^2) &= \sum_{i=1}^n \log p(y_i) \\
&= \sum_{i=1}^n (\log \frac{1}{\sigma\sqrt{2\pi}} - \frac{u_i^2}{2\sigma^2}) \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n u_i^2.
\end{aligned}$$

$$0 = \frac{\partial l}{\partial \beta_0} = \frac{-1}{2\sigma^2} \sum_{i=1}^n 2u_i \times (-1) = \frac{1}{\sigma^2} \sum_{i=1}^n u_i \Rightarrow \sum_{i=1}^n u_i = 0 \qquad (*)$$

$$0 = \frac{\partial l}{\partial \beta_1} = \frac{-1}{2\sigma^2} \sum_{i=1}^n 2u_i \times (-x_i) = \frac{1}{\sigma^2} \sum_{i=1}^n x_i u_i \Rightarrow \sum_{i=1}^n x_i u_i = 0 \qquad (**)$$

From $(*)$, we have

$$0 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = n\bar{y} - n\beta_0 - n\beta_1 \bar{x}.$$

Solving the equation yields that $\beta_0 = \bar{y} - \beta_1 \bar{x}$.

From $(**)$, we have

$$\begin{aligned}
0 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i &= \sum_{i=1}^n x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 \\
&= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + \beta_1 n\bar{x} - \beta_1 \sum_{i=1}^n x_i^2.
\end{aligned}$$

Solving the equation yields that $\beta_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum i=1^n (x_i-\bar{x})^2}$. Therefore,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

For parameter σ^2, we have

$$0 = \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{2\pi\sigma^2} \times 2\pi + \frac{1}{2} \sum_{i=1}^n u_i^2 \cdot \frac{1}{\sigma^4}.$$

Thus,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n u_i^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

(b)

1)
LSE: $\epsilon_i \overset{i.i.d}{\sim} (0, \sigma^2)$, any distribution satisfies $E = 0$, Var $= \sigma^2$.
MLE: $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$, normal distribution.
2)
LSE: the estimator of σ^2 is $S^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$, unbiased.
MLE: the estimator of σ^2 is $\hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} = \frac{n-2}{n} S^2$, asymptotically unbiased.

$$E(\hat{\sigma}^2_{\text{MLE}}) = \frac{n-2}{n} E(S^2) \to E(S^2) = \sigma^2.$$

3)
If $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$, the maximum likelihood estimate of (β_0, β_1) is the same as the least squares estimate.

Problem 2

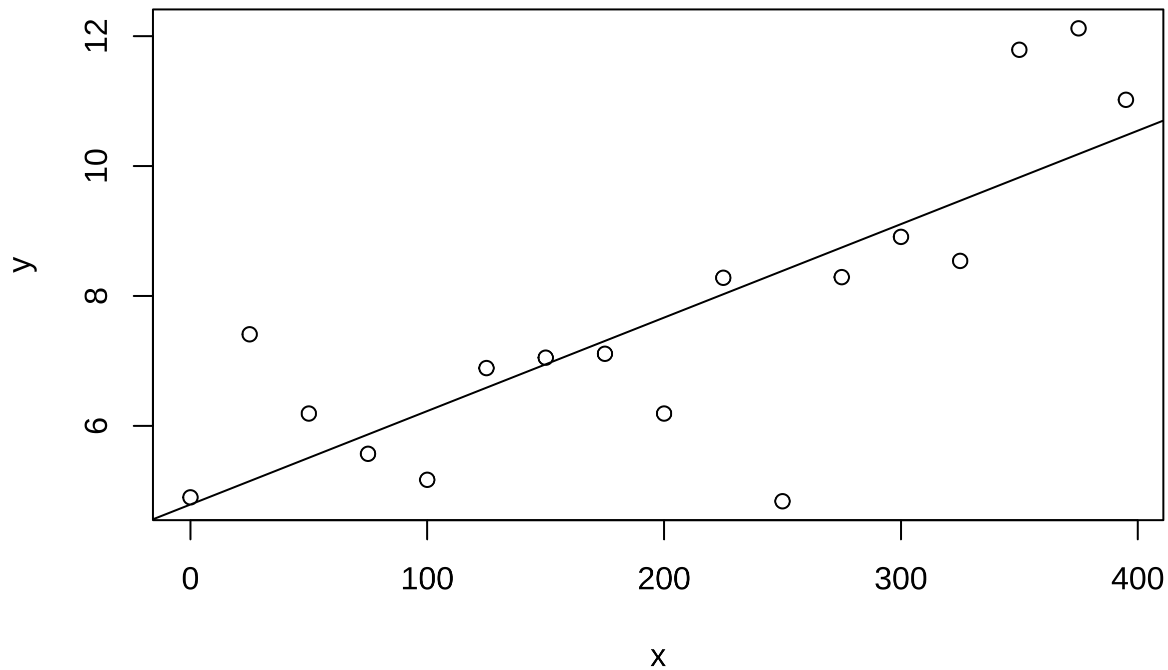
(a)&(b)

$\bar{x} = 199.7059$, $\bar{y} = 7.662941$, $S_{xx} = \sum_{i=1}^{17} (x_i - \bar{x})^2 = 253023.5$, $S_{xy} = \sum_{i=1}^{17} (x_i - \bar{x})(y_i - \bar{y}) = 3640.465$, $S_{yy} = \sum_{i=1}^{17} (y_i - \bar{y})^2 = 83.14615$. Therefore,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{3640.465}{253023.5} = 0.01438785,$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.662941 - 0.01438785 \times 199.7059 = 4.789603.$$

The least squares line is: $\hat{y} = 4.789603 + 0.01438785x$.

```
setwd("D:/  /Lesson/   /Data Sets")
mydata <- read.csv("DRILLROCK.csv")
x <- mydata[, 1]; y <- mydata[, 2]
plot(x, y)
model <- lm(y ~ x)
abline(model)
```



Comment on scatterplot: we can see some linear relation between x and y from the scatterplot.

(c)

Model for the data: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$. Assumptions: y_i and x_i have a linear relation; $y_i^{'}s$(or $\epsilon_i^{'}s$) are independent, $i = 1, ..., n$; $\text{Var}(y_i) = \text{Var}(\epsilon_i) = \sigma^2$; $y_i \overset{i.n.d}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$.

(d)

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

$\text{SSE} = \sum_{i=1}^{17} (y_i - \hat{y}_i)^2 = 30.76769$. $S^2 = \frac{\text{SSE}}{17-2} = \frac{30.76769}{15} = 2.051179$. Then

$$t = \frac{\hat{\beta}_1 - \beta_1}{S/S_{xx}^{\frac{1}{2}}} = \frac{0.01438785 - 0}{(2.051179/253023.5)^{\frac{1}{2}}} = 5.053294 > t_{0.025,15} = 2.131.$$

Hence, we should reject H_0.

(e)

$$t = \frac{\hat{\beta}_1 - \beta_1}{S/S_{xx}^{\frac{1}{2}}} \sim t(15).$$

$$95\% = \Pr\{ - t_{0.025,15} \leq \frac{\hat{\beta}_1 - \beta_1}{S/S_{xx}^{\frac{1}{2}}} \leq t_{0.025,15} \}.$$

C.I. of β_1:

$$-2.131 \leq \frac{0.01438785 - \beta_1}{(2.051179/253023.5)^{\frac{1}{2}}} \leq 2.131.$$

Therefore, the 95% C.I. of β_1 is

$$[0.00832042, 0.02045528].$$

(f)

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{30.76769}{83.14615} = 0.6299565.$$

Meaning: 63 percent of the variation of the time it takes to in the data is explained by the model(depth).

(g)

Regression prediction equation:

$$y_h = \hat{\beta}_0 + \hat{\beta}_1 x_h.$$

To construct the C.I.,

$$x_h = 6, \quad E(y_h) = \beta_0 + \beta_1 x_h.$$

Prediction: $\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h = 4.789603 + 0.01438785 \times 6 = 4.87593$.
The 95% C.I. of $E(y_h)$ is

$$\begin{aligned}
&\hat{y}_h \pm t_{0.025,15} \cdot S[\frac{1}{17} + \frac{(x_h - \bar{x})^2}{S_{xx}}]^{\frac{1}{2}} \\
=&4.87593 \pm 2.131 \cdot \sqrt{2.051179}[\frac{1}{17} + \frac{(6 - 199.7059)^2}{253023.5}]^{\frac{1}{2}} \\
=&4.87593 \pm 1.388974
\end{aligned}$$

Thus, the 95% C.I. of $E(y_h)$ is $(3.486956, 6.264904)$.

(h)

Predictive Interval of $y_{h,new} = \beta_0 + \beta_1 x_h + \epsilon_h$ at $x_h = 6$. Thus,

$$\begin{aligned}
&\hat{y}_h \pm t_{0.025,15} \cdot S[1 + \frac{1}{17} + \frac{(x_h - \bar{x})^2}{S_{xx}}]^{\frac{1}{2}} \\
=&4.87593 \pm 2.131 \cdot \sqrt{2.051179}[1 + \frac{1}{17} + \frac{(6 - 199.7059)^2}{253023.5}]^{\frac{1}{2}} \\
=&4.87593 \pm 3.353205
\end{aligned}$$

Thus, the 95% P.I. of $y_{h,new}$ is $(1.522725, 8.229135)$.

(i)

SSE = 30.76769, SST = 83.14615, then SSR = SST - SSE = 52.37846 and MSR = SSR/1 = SSR, MSE = SSE/15 = 2.051179. F = MSR/MSE = 25.53578. Thus, the ANOVA table is

	Sum of squares	Degree of freedom	Mean squares	F
Regression	52.37846	1	52.37836	25.53578
Error	30.76769	15	2.051179	
Total	83.14615	1		

Hypotheses:

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

F-test: F=$\frac{\text{MSR}}{\text{MSE}} \sim F_{1,15}$. And F $= 25.53578 > F_{0.01}(1, 15) = 8.683$. Therefore, we should reject H_0.