Homework #4

Dustin Leatherman

10/12/2019

3.18

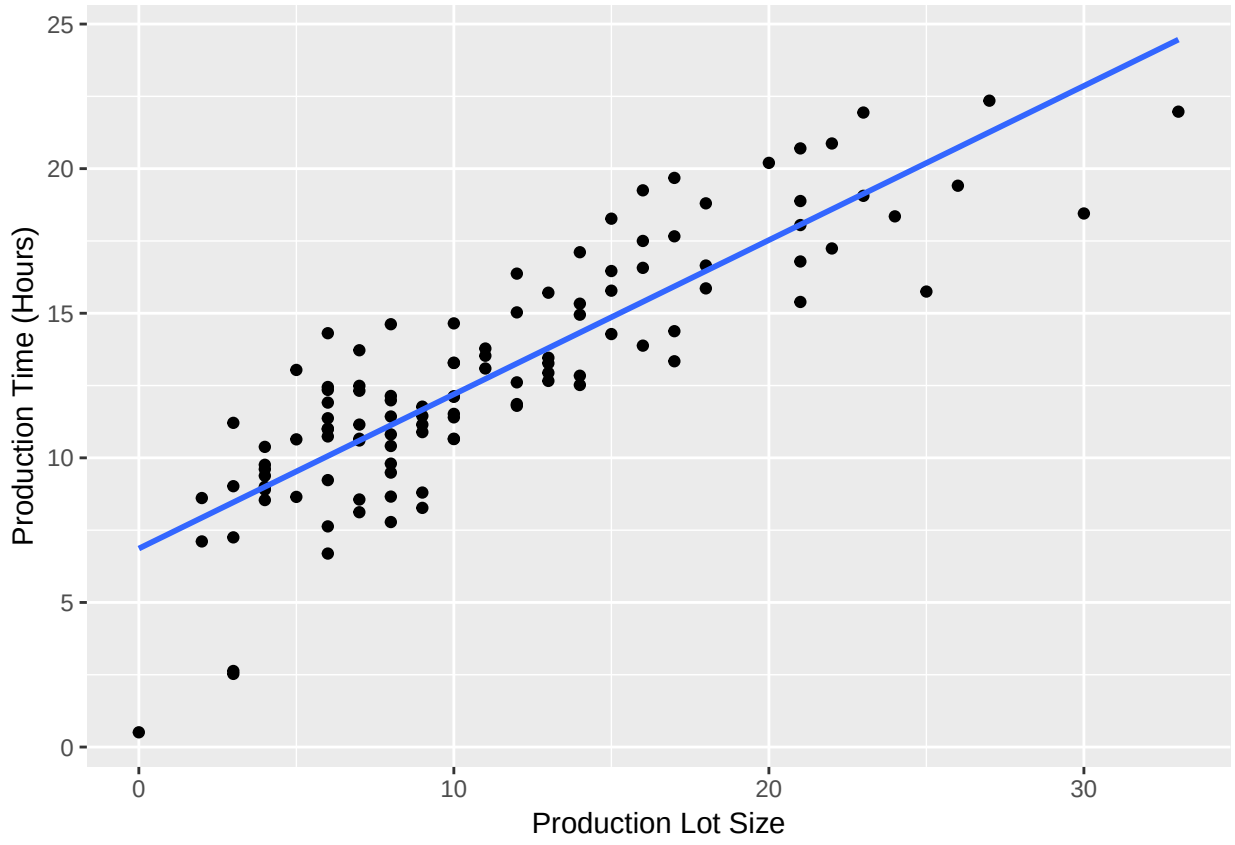
In a manufacuring study, the production times for 111 recent production runs were obtained.

a

Prepare a scatterplot of the data. Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate? Why?

```
production <- read.csv("~/Downloads/ProductionTime.csv")

production %>%
  qplot(LotSize, Hours, data = ., xlab = "Production Lot Size", ylab = "Production Time (Hours)") +
  geom_smooth(method = lm, se = F)
```



There appears to be a positive correlation between Production Lot Size and Production Time but it doesn't appear to be completely linear. There is some curvature in the graph with increased distance from the

regression line as Lot Size increases. This indicates that a transformation on the X would be appropriate in order to meet the linearity assumption.

b

Use the transformation $X' = \sqrt{X}$ and obtain the estimated linear regression function for the the transformed data.

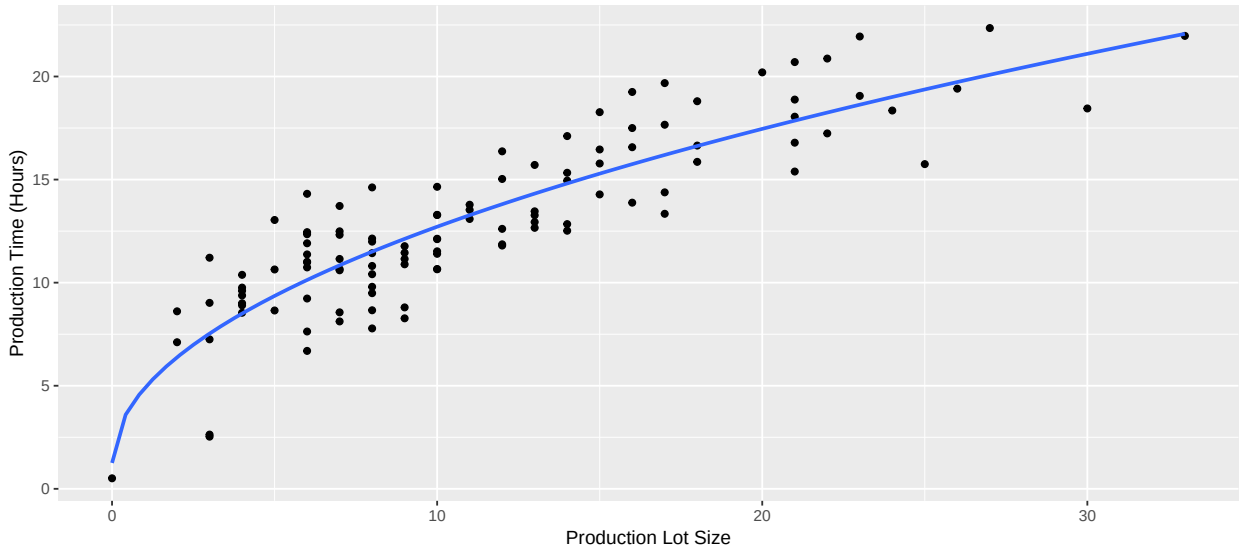
```
production.model1 <- lm(Hours ~ sqrt(LotSize), data = production)
```

$$\hat{Y} = 1.2547 + 3.6235\sqrt{LotSize}$$

c

Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

```
production %>%
  qplot(LotSize, Hours, data = ., xlab = "Production Lot Size", ylab = "Production Time (Hours)") +
  geom_smooth(method = lm, formula = y ~ sqrt(x), se = F)
```

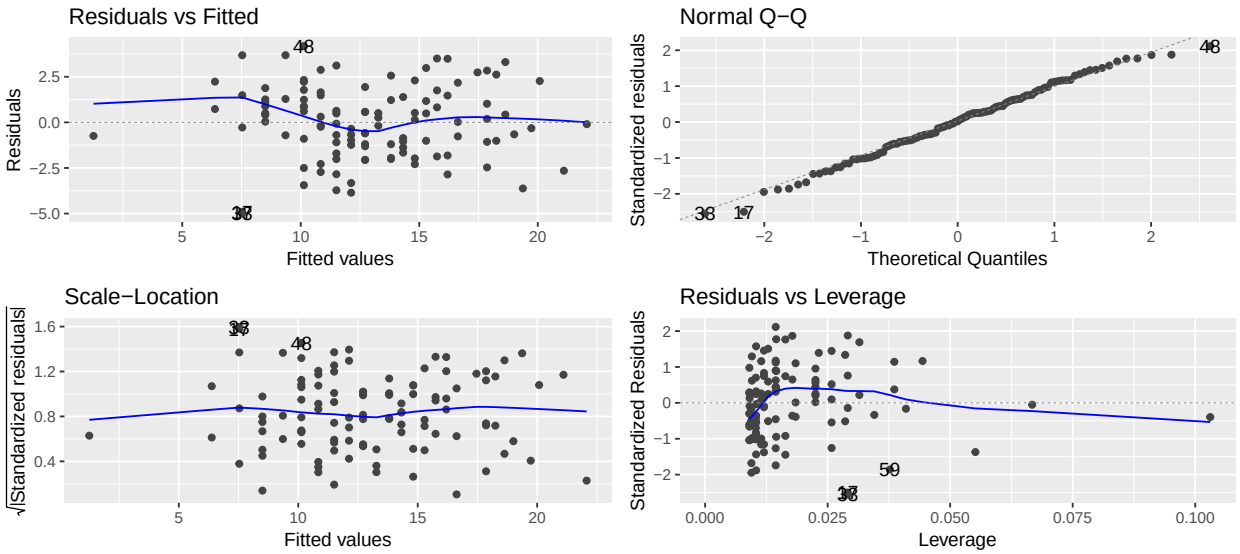


The regression line fits much better than the untransformed model. It is a good fit for the data.

d

Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

```
autoplot(production.model1)
```



The Residuals plot show a larger spread for fitted values less than 10 compared to the rest of the fitted values. Based on the plots, one could justify constant variance but the Brown-Forsyth or Bruesch-Pagan test should be used to confirm.

The normal probability plot shows the residuals falling on or near the line indicating the residuals are normally distributed.

e

Express the estimated regression function in the original units

$$\hat{Y} = 1.2547 + 3.6235\sqrt{LotSize}$$

3.20

If the error terms in a regression model are independent $N(0, \sigma^2)$, what can be said about the error terms after transformation $X' = \frac{1}{X}$ is used? Is the situation the same after transformation $Y' = \frac{1}{Y}$ is used?

Transforming X nor Y does not affect the independence of the error terms since transforming the variables would not affect the relationship between the two variables, it would only scale it.

4.9

Refer to Plastic Hardness

a

Management wishes to obtain interval estimates of the mean hardness when the elapsed time is 20, 30, and 40 hour, respectively. Calculate the desired confidence intervals, using Bonferroni

procedure and a 90% family confidence coefficient. What is the meaning of the family confidence coefficient here?

```
plastic <- readxl::read_excel("~/Downloads/PlasticHardness.xlsx")
plastic.model1 <- lm(Hardness ~ Hours, data = plastic)
newData <- data.frame(Hours=c(20,30,40))
myAlpha <- 0.1
g <- count(newData) %>% as.integer
myLevel <- 1 - myAlpha / g
predict(plastic.model1, newdata = newData, interval = "confidence", level = myLevel) %>%
  as_tibble %>%
  bind_cols(newData) %>%
  select("Hours","lwr","upr") %>% kable
```

Hours	lwr	upr
20	206.7277	211.8473
30	227.6762	231.5863
40	246.7824	253.1676

With 90% confidence, the average hardness for 20, 30, and 40 hours will all fall within the aforementioned intervals for any given random sample.

b

Is the Bonferroni procedure employed in (a) the most efficient one that could be employed here? Explain.

The Bonferroni procedure tends to work better for smaller sets of predictors whereas the Scheffe works better on a larger set of predictors. The Bonferroni procedure is the most efficient in this case.

c

The next two test items will be measured after 30 and 40 hours of elapsed time, respecitvely. Predict the hardness for each of these two items, using the most efficient procedure and a 90% family confidence coefficient

```
newData <- data.frame(Hours=c(30,40))
myAlpha <- 0.1
g <- count(newData) %>% as.integer
myLevel <- 1 - myAlpha / g
predict(plastic.model1, newdata = newData, interval = "predict", level = myLevel) %>%
  as_tibble %>%
  bind_cols(newData) %>%
  select("Hours", "fit","lwr","upr") %>% kable
```

Hours	fit	lwr	upr
30	229.6312	222.4710	236.7915
40	249.9750	242.4562	257.4938

It is estimated that after 30 and 40 hours of elapsed time, the test items will have a hardness of 229.6312 and 249.975 respectively. With 90% confidence, given 30 and 40 hours of elapsed time, the hardness will be between 222.471 and 236.7915 after 30 hours and between 242.4562 and 257.4938 after 40 hours.

4.20

Refer to the Plastic Hardness problem. The measurement of a new test item showed 238 Brinell units of Hardness.

a

Obtain a 99% confidence interval for the elapsed time before the hardness was measured. Interpret your confidence interval.

```
calibrate(plastic.model1, 238, interval = "Wald", level = 0.99)
```

```
## estimate    lower    upper       se
## 34.11367 29.16920 39.05815  1.66098
```

With 99% confidence, a test item with 238 Brinell units of Hardness has an associated elapsed time between 29.1692 and 39.05815 Hours.

b

Is criterion (4.33) as to the appropriateness of the approximate confidence interval met here?

```
b1 <- plastic.model1$coefficients[2] %>% unname

(qt(0.995, 16 - 2)^2 * anova(plastic.model1)$'Mean Sq'[2]) / (b1 * sum((plastic$Hours - mean(plastic$Hou
```

```
## [1] 0.03559246
```

$\frac{(t_{1-\frac{\alpha}{2},n-2})^2 MSE}{b_1^2 \Sigma (X_i - \bar{X})^2} = 0.0356$

This value is less than 0.1 so the confidence interval is appropriate.

4.19

Refer to the GPA dataset. A new student earned a GPA of 3.4 in the freshman year.

a

Obtain a 90% Confidence Interval for the student's ACT test score. Interpret your confidence interval.

```
gpa <- readxl::read_excel("~/Downloads/GradePointAverage.xlsx")
gpa.model <- lm(GPA ~ ACT, data = gpa)
calibrate(gpa.model, 3.4, interval = "Wald", level = 0.90)
```

```
##  estimate     lower     upper        se
## 33.119904  6.013148 60.226660 16.350355
```

With 90% confidence, a freshman student's GPA of 3.4 is associated with an ACT score between 6.013 and 36 points.

b

Is criterion (4.33) as to the appropriateness of the approximate confidence interval met here?

```
b1 <- gpa.model$coefficients[2] %>% unname

(qt(0.95, (count(gpa) %>% as.integer) - 2)^2 * anova(gpa.model)$'Mean Sq'[2]) / (b1 * sum((gpa$ACT - mea
```

```
## [1] 0.01154922
```

$\frac{(t_{1-\frac{\alpha}{2},n-2})^2 MSE}{b_1^2 \Sigma (X_i - \bar{X})^2} = 0.0115$

The value is below 0.1 so this confidence interval is appropriate.

4.22

Derive an extension of the Bonferroni inequality (4.2a) for the case of three statements, each with statement confidence coefficient $1 - \alpha$

$$b_0 \pm t(1 - \alpha/2; n - 2)s(b_0)$$
$$b_1 \pm t(1 - \alpha/2; n - 2)s(b_1)$$
$$b_2 \pm t(1 - \alpha/2; n - 2)s(b_2)$$

Let $P(A)$ represent the event that the first confidence interval does not cover b_0. Let $P(B)$ represent the event that the first confidence interval does not cover b_1. Let $P(C)$ represent the event that the first confidence interval does not cover b_2.

Thus:

$$\begin{aligned} P(A) =& \alpha \\ P(B) =& \alpha \\ P(C) =& \alpha \end{aligned}$$

$$\begin{aligned} P(A \cup B \cup C) =& P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \\ P(\bar{A} \cap \bar{B} \cap \bar{C}) =& 1 - P(A \cup B \cup C) \\ =& 1 - P(A) - P(B) - P(C) + P(A \cap B) + P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) \end{aligned}$$

An inequality is introduced because of the following assumption: $P(A \cap B) \geq 0, P(A \cap C) \geq 0, P(B \cap C) \geq 0, P(B \cap C) - P(A \cap B \cap C) \geq 0$

Thus,

$P(\bar{A} \cap \bar{B} \cap \bar{C}) \geq 1 - P(A) - P(B) - P(C) = 1 - 3\alpha$

5.1

For the matrices, obtain 1. A + B 2. A - B 3. AC 4. AB' 5. B'A

1.

$$\begin{bmatrix} 2 & 7 \\ 3 & 10 \\ 5 & 13 \end{bmatrix} (3 \times 2)$$

2.
$$\begin{bmatrix} 0 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} (3 \times 2)$$

3.
$$\begin{bmatrix} 23 & 24 & 1 \\ 36 & 40 & 2 \\ 49 & 56 & 3 \end{bmatrix} (3 \times 3)$$

4.
$$\begin{bmatrix} 7 & 8 & 13 \\ 12 & 14 & 22 \\ 17 & 20 & 31 \end{bmatrix} (3 \times 3)$$

5.
$$\begin{bmatrix} 26 & 76 \\ 9 & 26 \end{bmatrix} (2 \times 2)$$

5.15

Consider the simultaneous equations:

$$\begin{aligned} 5y_1 + 2y_2 &= 8 \\ 23y_1 + 7_y 2 &= 28 \end{aligned}$$

a

Write these equations in Matrix Notation

$$\begin{bmatrix} 5 & 2 & 8 \\ 23 & 7 & 28 \end{bmatrix}$$

b

Using matrix methods, find the solutions for y_1 and y_2

$$\begin{bmatrix} 5 & 2 & 8 \\ 23 & 7 & 28 \end{bmatrix} \underset{\frac{1}{5}R_1}{\rightarrow} \begin{bmatrix} 1 & \frac{2}{5} & \frac{8}{5} \\ 23 & 7 & 28 \end{bmatrix} \underset{-23R_2+R_1}{\rightarrow} \begin{bmatrix} 1 & \frac{2}{5} & \frac{8}{5} \\ 0 & \frac{-11}{5} & \frac{-44}{5} \end{bmatrix} \underset{\frac{-5}{11}R_2}{\rightarrow} \begin{bmatrix} 1 & \frac{2}{5} & \frac{8}{5} \\ 0 & 1 & 4 \end{bmatrix} \underset{\frac{-2}{5}R_2+R_1}{\rightarrow} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 4 \end{bmatrix}$$

$$y_1 = 0, y_2 = 4$$

5.17

Consider the following functions of the random variables Y_1, Y_2, Y_3

$$\begin{aligned} W_1 &= Y_1 + Y_2 + Y_3 \\ W_2 &= Y_1 - Y_2 \\ W_3 &= Y_1 - Y_2 - Y_3 \end{aligned}$$

a

State the above in matrix notation

$$W = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \vec{Y}$$

b

Find the expectation of the random vector W

$$E(\vec{W}) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \end{bmatrix} E(\vec{Y})$$

c

Find the variance-covariance matrix of W

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \sigma^2(\vec{Y}) \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 0 & 1 \end{bmatrix}$$

5.5

The data below show, for a consumer finance company operating in six cities, the number of competing loan companies operating in the city (X) and the number per thousand of the comapny's loans made in that city that are currently delinquent (Y) Assume that first-order regression model is applicable. Using matrix methods, find (1) $Y'Y$, (2) $X'X$, (3) $X'Y$

$$\begin{aligned} \vec{X} &= \begin{bmatrix} 4 & 1 & 2 & 3 & 3 & 4 \end{bmatrix} \\ \vec{Y} &= \begin{bmatrix} 16 & 5 & 10 & 15 & 13 & 22 \end{bmatrix} \end{aligned}$$

1.

$$Y'Y = \begin{bmatrix}16 & 5 & 10 & 15 & 13 & 22\end{bmatrix} \begin{bmatrix}16\\5\\10\\15\\13\\22\end{bmatrix} = 1259$$

2.

$$X'X = \begin{bmatrix}4 & 1 & 2 & 3 & 3 & 4\end{bmatrix} \begin{bmatrix}4\\1\\2\\3\\3\\4\end{bmatrix} = 55$$

3.

$$\begin{bmatrix}4 & 1 & 2 & 3 & 3 & 4\end{bmatrix} \begin{bmatrix}16\\5\\10\\15\\13\\22\end{bmatrix} = 261$$

5.13

$$\begin{aligned} X'X =& 55 \\ 55^{-1} =& 0.01818 \end{aligned}$$

Regression through the origin (again)

Let $Y_i = \beta X_i + \epsilon_i$ where $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2$

a

Write the model as $Y = X\beta + \epsilon$ defining each matrix/vector

$$\begin{aligned} \underset{(n\times 1)}{\vec{Y}} &= \begin{bmatrix}Y_1\\...\\Y_n\end{bmatrix} \\ \underset{(n\times 2)}{X} &= \begin{bmatrix}1 & X_1\\... & ...\\1 & X_n\end{bmatrix} \\ \underset{(2\times 1)}{\vec{\beta}} &= \begin{bmatrix}0\\\beta_1\end{bmatrix} \\ \underset{(n\times 1)}{\vec{\epsilon}} &= \begin{bmatrix}\epsilon_1\\...\\\epsilon_n\end{bmatrix} \end{aligned}$$

b

Show that $\hat{\beta} = (X'X)^{-1}X'Y = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$

$$X'Y = \begin{bmatrix} X_1 & ... & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ ... \\ Y_n \end{bmatrix} = X_1Y_1 + ... + X_nY_n = \sum_{i=1}^{n} X_iY_i$$

$$X'X = \begin{bmatrix} X_1 & ... & X_n \end{bmatrix} \begin{bmatrix} X_1 \\ ... \\ X_n \end{bmatrix} = X_1^2 + ... + X_n^2 = \sum_{i=1}^{n} X_i^2$$

$$(\sum_{i=1}^{n} X_i^2)^{-1} = \frac{1}{\sum_{i=1}^{n} X_i^2} \times \sum_{i=1}^{n} X_iY_i = \frac{\sum_{i=1}^{n} X_iY_i}{\sum_{i=1}^{n} X_i^2}$$

c

Show $var(\hat{\beta}) = \sigma^2(X'X)^{-1} = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$

Per above, $(X'X)^{-1} = \frac{1}{\sum_{i=1}^{n} X_i^2}$. Thus $\sigma^2(X'X)^{-1} = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$