Homework 1

Chi-Yuan Fang

2021-02-25

Contents

Chapter 1

Textbook Exercises

Do the following problem sets from Stock & Watson (4th Edition).

2.6, 2.8, 2.14, 2.25, 3.2, 3.3, 3.4, 3.9, 3.13

1.1 Exercise 2.6

a.

$$E(Y) = 0 \cdot 0.12 + 1 \cdot 0.88 = 0.88 \tag{1.1}$$

b.

$$\begin{aligned} Pr(Y = 0) &= 1 - Pr(Y = 1) && (1.2)\\ &= 1 - E(Y) && (1.3)\\ &= 1 - 0.88 = 0.12 && (1.4) \end{aligned}$$

c. The conditional probabilities are

$$Pr(Y=0|X=0) = \frac{Pr(X=0,Y=0)}{Pr(X=0)} \tag{1.5}$$
$$= \frac{0.078}{0.751} = \frac{78}{751} \tag{1.6}$$
$$Pr(Y=1|X=0) = \frac{Pr(X=0,Y=1)}{Pr(X=0)} \tag{1.7}$$
$$= \frac{0.673}{0.751} = \frac{673}{751} \tag{1.8}$$
$$Pr(Y=0|X=1) = \frac{Pr(X=1,Y=0)}{Pr(X=1)} \tag{1.9}$$
$$= \frac{0.042}{0.249} = \frac{14}{83} \tag{1.10}$$
$$Pr(Y=1|X=1) = \frac{Pr(X=1,Y=1)}{Pr(X=1)} \tag{1.11}$$
$$= \frac{0.207}{0.249} = \frac{69}{83}. \tag{1.12}$$

Then, the conditional expectations are

$$E(Y|X=1) = 0 \cdot \frac{14}{83} + 1 \cdot \frac{69}{83} = \frac{69}{83} \approx 0.8313 \tag{1.13}$$
$$E(Y|X=0) = 0 \cdot \frac{78}{751} + 1 \cdot \frac{673}{751} = \frac{673}{751} \approx 0.8961. \tag{1.14}$$

d. Unemployment rate for college graduates is

$$Pr(Y=0|X=1) = 1 - E(Y|X=1) \tag{1.15}$$
$$= 1 - \frac{69}{83} \tag{1.16}$$
$$= \frac{14}{83} \approx 0.1687, \tag{1.17}$$

and unemployment rate for non-college graduates is

$$Pr(Y=0|X=0) = 1 - E(Y|X=0) \tag{1.18}$$
$$= 1 - \frac{673}{751} \tag{1.19}$$
$$= \frac{78}{751} \approx 0.1039. \tag{1.20}$$

e. Given a randomly selected member of unemployed population, the probability that the worker is a college graduate is

$$Pr(X=1|Y=0) = \frac{Pr(X=1,Y=0)}{Pr(Y=0)} \tag{1.21}$$
$$= \frac{0.042}{0.12} = 0.35, \tag{1.22}$$

and the probability that the worker is a non-college graduate is

$$Pr(X = 0|Y = 0) = \frac{Pr(X = 0, Y = 0)}{Pr(Y = 0)} \tag{1.23}$$

$$= \frac{0.078}{0.12} = 0.65. \tag{1.24}$$

f. Because

$$Pr(X = 0|Y = 0) = \frac{78}{751} \tag{1.25}$$

$$\neq Pr(X = 0) = 0.12, \tag{1.26}$$

educational achievement and employment status are not independent.

1.2 Exercise 2.8

We know

$$E(Y) = 4 \quad \text{and} \quad Var(Y) = \frac{1}{9}. \tag{1.27}$$

Then,

$$E(Z) = E\left[3(Y - 4)\right] \tag{1.28}$$

$$= 3E(Y - 4) \tag{1.29}$$

$$= 3E(Y) - 12 \tag{1.30}$$

$$= 3 \cdot 4 - 12 = 0 \tag{1.31}$$

and

$$Var(Z) = Var\left[3(Y - 4)\right] \tag{1.32}$$

$$= 3^2 \cdot Var(Y - 4) \tag{1.33}$$

$$= 9 \cdot Var(Y) \tag{1.34}$$

$$= 9 \cdot \frac{1}{9} = 1. \tag{1.35}$$

1.3 Exercise 2.14

By CLT, we have

$$\overline{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right) \overset{d}{=} N\left(50, \frac{21}{n}\right). \tag{1.36}$$

a.

$$Pr\left(\overline{Y} \leq 51\right) = Pr\left(\frac{\overline{Y} - 50}{\sqrt{21/50}} \leq \frac{51 - 50}{\sqrt{21/50}}\right) \approx 0.9386 \tag{1.37}$$

b.

$$Pr\left(\overline{Y} > 49\right) = Pr\left(\frac{\overline{Y} - 50}{\sqrt{21/150}} > \frac{49 - 50}{\sqrt{21/150}}\right) \approx 0.9962 \tag{1.38}$$

c.

$$Pr\left(50.5 \leq \overline{Y} < 51\right) = Pr\left(\frac{50.5 - 50}{\sqrt{21/45}} \leq \overline{Y} \leq \frac{51 - 50}{\sqrt{21/45}}\right) \approx 0.1605 \tag{1.39}$$

1.4 Exercise 2.25

a.

$$\sum_{i=1}^{n} ax_i = ax_1 + ax_2 + ... + ax_n \tag{1.40}$$

$$= a(x_1 + x_2 + ... + x_n) \tag{1.41}$$

$$= a\sum_{i=1}^{n} x_i \tag{1.42}$$

b.

$$\sum_{i=1}^{n}(x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + ... + (x_n + y_n) \tag{1.43}$$

$$= (x_1 + x_2 + ... + x_n) + (y_1 + y_2 + ... + y_n) \tag{1.44}$$

$$= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{1.45}$$

c.

$$\sum_{i=1}^{n} a = \underbrace{a + a + ... + a}_{n \text{ times}} \tag{1.46}$$

$$= n \times a \tag{1.47}$$

d.

$$\sum_{i=1}^{n}(a + bx_i + cy_i)^2 = \sum_{i=1}^{n}\left(a^2 + b^2x_i^2 + c^2y_i^2 + 2abx_i + 2acy_i + 2bcx_iy_i\right) \tag{1.48}$$

$$= na^2 + b^2\sum_{i=1}^{n} x_i^2 + c^2\sum_{i=1}^{n} y_i^2 + 2ab\sum_{i=1}^{n} x_i + 2ac\sum_{i=1}^{n} y_i + 2bc\sum_{i=1}^{n} x_iy_i \tag{1.49}$$

1.5 Exercise 3.2

a.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y} \tag{1.50}$$

b.

$$\begin{aligned} E(\hat{p}) &= E\left(\overline{Y}\right) && (1.51)\\ &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) && (1.52)\\ &= \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i\right) && (1.53)\\ &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) && (1.54)\\ &= \frac{1}{n}\cdot np = p && (1.55) \end{aligned}$$

c.

$$\begin{aligned} Var(\hat{p}) &= Var\left(\overline{Y}\right) && (1.56)\\ &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) && (1.57)\\ &= \frac{1}{n^2}Var\left(\sum_{i=1}^{n} Y_i\right) && (1.58)\\ &= \frac{1}{n^2}\left[nVar(Y_i) + 2\sum_{i\neq j}\underbrace{Cov(Y_i, Y_j)}_{=0}\right] && (1.59)\\ &= \frac{1}{n^2}\cdot np(1-p) = \frac{p(1-p)}{n} && (1.60) \end{aligned}$$

1.6 Exercise 3.3

a.

$$\hat{p} = \frac{270}{500} = 0.54 \tag{1.61}$$

b. The estimated variance of $\hat{p}$ is

$$\begin{aligned} Var(\hat{p}) &= \frac{\hat{p}(1-\hat{p})}{n} && (1.62)\\ &= \frac{0.54(1-0.54)}{500} = 0.0004968, && (1.63) \end{aligned}$$

and the standard error is

$$SE\left(\hat{p}\right) = \sqrt{Var\left(\hat{p}\right)} \tag{1.64}$$
$$\approx 0.0223. \tag{1.65}$$

c. The t-statistic is

$$t^* = \frac{\hat{p} - p_0}{SE\left(\hat{p}\right)} \tag{1.66}$$
$$= \frac{0.54 - 0.5}{0.0223} \tag{1.67}$$
$$= \frac{400}{223} \approx 1.7937. \tag{1.68}$$

Then,

$$p - value = 2\Phi\left(-|t^*|\right) \tag{1.69}$$
$$= 2\Phi\left(-\frac{400}{223}\right) \approx 0.0729. \tag{1.70}$$

d. The t-statistic is

$$t^* = \frac{\hat{p} - p_0}{SE\left(\hat{p}\right)} \tag{1.71}$$
$$= \frac{0.54 - 0.5}{0.0223} \tag{1.72}$$
$$= \frac{400}{223} \approx 1.7937. \tag{1.73}$$

Then,

$$p - value = 1 - \Phi\left(|t^*|\right) \tag{1.74}$$
$$= 1 - \Phi\left(\frac{400}{223}\right) \approx 0.0364. \tag{1.75}$$

e. Part (c) is a two-sided test and the p-value is the area in the tails of the standard normal distribution outside the $\pm$ (calculated t-statistic). Part (d) is a one-sided test and the p-value is the area under the standard normal distribution to the right of the calculated t-statistic.

f. For the test $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$ and $\alpha = 0.05$, we reject H_0 because p-value is less than the significance level. There is statistically significant evidence that the democratic candidate was ahead of the republican candidate at the time of conducting the poll.

1.7 Exercise 3.4

a.

$$\overline{p} \pm Z_{0.025} SE\left(\overline{p}\right) = 0.54 \pm 1.96 \cdot 0.0223 \approx [0.4963, 0.5837] \tag{1.76}$$

b.

$$\overline{p} \pm Z_{0.005} SE\left(\overline{p}\right) = 0.54 \pm 2.576 \cdot 0.0223 \approx [0.4826, 0.5974] \tag{1.77}$$

c. Mechanically, the interval in part (b) is wider because of a larger critical value. Substantively, a 99% confidence interval is wider than a 95% confidence level because a 99% confidence interval must contain the true value of p in 99% of all possible samples, while a 95% confidence interval must contain the true value of p in only 95% of all possible samples.

d. Because $0.5 \in C.I.$, we do not reject H_0 at 5% significance level.

1.8 Exercise 3.9

a. We know

$$E(Y) = 1000 \tag{1.78}$$
$$\sigma_Y = 100 \tag{1.79}$$
$$n = 50 \tag{1.80}$$

and

$$E\left(\overline{Y}\right) = 1000 \tag{1.81}$$
$$\sigma_{\overline{Y}} = \frac{100}{\sqrt{50}} = 10\sqrt{2}. \tag{1.82}$$

Then,

$$\text{size} = \Pr\left(\overline{Y} > 1100 | \mu = 1000\right) \tag{1.83}$$
$$= \Pr\left(\frac{\overline{Y} - 1000}{10\sqrt{2}} > \frac{1100 - 1000}{10\sqrt{2}}\right) \approx 0. \tag{1.84}$$

b. The probability of type 2 error is

$$\beta = \Pr\left(\overline{Y} | \mu = 1150\right) \tag{1.85}$$
$$= \Pr\left(\frac{\overline{Y} - 1150}{10\sqrt{2}} \leq \frac{1100 - 1150}{10\sqrt{2}}\right) \approx 0.0002. \tag{1.86}$$

Then, the power of the manager's testing is

$$1 - \beta \approx 0.9998. \tag{1.87}$$

c. For a test with size 1%, the rejection region for H_0 contains those values of the t-statistic exceeding $Z_{0.01}$. That is,

$$t^* = \frac{\overline{Y} - 1000}{10\sqrt{2}} > Z_{0.01} = 2.326. \tag{1.88}$$

Then,

$$\overline{Y} > 1000 + 10\sqrt{2} \cdot 2.326 \approx 1032.8946. \tag{1.89}$$

Thus, the manager should believe the inventor's claim if the sample mean life of the new product is greater than 1032.8946 hours if she wants the size of the test to be 1%.

1.9 Exercise 3.13

a.

$$\overline{Y} \pm Z_{0.05} \cdot \frac{s_Y}{\sqrt{n}} = 712.1 \pm 1.645 \cdot \frac{23.2}{\sqrt{400}} = [710.1918, 714.0082] \tag{1.90}$$

b.

- **Prepare**: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$, where μ_1 is average salary with small class, μ_2 is average salary with large class. Let the significance level be 0.05.
- **Calculate**:

$$t^* = \frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1.91}$$

$$= \frac{721.8 - 710.9}{\sqrt{\frac{24.4^2}{150} + \frac{20.6}{250}}} \approx 4.5790 \tag{1.92}$$

- **Conclude**: Because

$$t^* > Z_{0.025} = 1.96, \tag{1.93}$$

we reject H_0. There is statistically significant evidence that districts with smaller classes have higher average test scores.

Chapter 2

Computer exercises

Do the following problem set from Stock & Watson (4th Edition).

E3.2

2.1 E3.2

A consumer is given the chance to buy a baseball card for $1, but he declines the trade. If the consumer is now given the baseball card, will he be willing to sell it for $1? Standard consumer theory suggests yes, but behavioral economists have found that "ownership" tends to increase the value of goods to consumers. That is, the consumer may hold out for some amount more than $1 (for example, $1.20) when selling the card, even though he was willing to pay only some amount less than $1 (for example, $0.88) when buying it. Behavioral economists call this phenomenon the "endowment effect." John List investigated the endowment effect in a randomized experiment involving sports memorabilia traders at a sports-card show. Traders were randomly given one of two sports collectibles, say good A or good B, that had approximately equal market value. Those receiving good A were then given the option of trading good A for good B with the experimenter; those receiving good B were given the option of trading good B for good A with the experimenter. Data from the experiment and a detailed description can be found on the text website, http://www.pearsonglobaleditions.com, in the files **Sportscards and Sportscards_Description**.

a. i. Suppose that, absent any endowment effect, all the subjects prefer good A to good B. What fraction of the experiment's subjects would you expect to trade the good that

they were given for the other good? (Hint: Because of random assignment of the two treatments, approximately 50% of the subjects received good A, and 50% received good B.)

ii. Suppose that, absent any endowment effect, 50% of the subjects prefer good A to good B, and the other 50% prefer good B to good A. What fraction of the subjects would you expect to trade the good they were given for the other good?

iii. Suppose that, absent any endowment effect, X% of the subjects prefer good A to good B, and the other $(100-X)$% prefer good B to good A. Show that you would expect 50% of the subjects to trade the good they were given for the other good.

Solution

i. A person will trade if he/she received good A but prefers good B or he/she received good B and prefers good A. 50% received good A, of these 100% prefer good B; 50% receive good B, of these 0% prefer good A. Thus, the expected fraction traded is

$$0.5 \times 1 + 0.5 \times 0 = 0.5. \tag{2.1}$$

ii. A person will trade if he/she received good A but prefers good B or he/she received good B and prefers good A. 50% received good A, of these 50% prefer good B; 50% receive good B, of these 50% prefer good A. Thus, the expected fraction traded is

$$0.5 \times 0.5 + 0.5 \times 0.5 = 0.5. \tag{2.2}$$

iii. A person will trade if he/she received good A but prefers good B or he/she received good B and prefers good A. 50% received good A, of these $(100-X)$% prefer good B; 50% receive good B, of these X% prefer good A. Thus, the expected fraction traded is

$$0.5 \times (1-x) + 0.5x = 0.5 \tag{2.3}$$

where $x = \frac{X}{100}$.

b. Using the sports-card data, what fraction of the subjects traded the good they were given? Is the fraction significantly different from 50%? Is there evidence of an endowment effect? (Hint: Review Exercises 3.2 and 3.3.)

Solution

- **Prepare**

$H_0 : p = 0.5$ (no endowment effect) v.s. $H_A : p \neq 0.5$ (endowment effect), where p is the fraction of trades.

Let the significance level be 0.05.

- **Calculate**

```
# import data
library(readxl)
sportscards <- read_xlsx("sportscards/Sportscards.xlsx")

# the fraction of trades
fract_trade <- mean(sportscards$trade)
fract_trade
```

```
## [1] 0.3378378
```

```
# standard error of the fraction of trades
se_fract_trade <- sd(sportscards$trade)/sqrt(length(sportscards$trade))
se_fract_trade
```

```
## [1] 0.03901015
```

```
# test
t.test(sportscards$trade,
       alternative = c("two.sided"),
       mu = 0.5, # H0
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  One Sample t-test
##
## data:  sportscards$trade
## t = -4.1569, df = 147, p-value = 5.456e-05
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.2607447 0.4149310
## sample estimates:
## mean of x
## 0.3378378
```

- **Conclude**

The fraction of trades in the sample was 0.3378, with a standard error of 0.0390.

Because $p-value < 0.05$, we reject H_0. There is statistically significant evidence of an endowment effect.

c. Some have argued that the endowment effect may be present but that it is likely to disappear as traders gain more trading experience. Half of the experimental subjects were dealers, and

the other half were nondealers. Dealers have more experience than nondealers. Repeat (b) for dealers and nondealers. Is there a significant difference in their behavior? Is the evidence consistent with the hypothesis that the endowment effect disappears as traders gain more experience? (Hint: Review Exercise 3.15.)

Solution

i. Dealers:

- **Prepare**

$H_0 : p_1 = 0.5$ (no endowment effect) v.s. $H_A : p_1 \neq 0.5$ (endowment effect), where p_1 is the fraction of trades for dealers.

Let the significance level be 0.05.

- **Calculate**

```
# dealer
sportscards_dealer <- sportscards[sportscards$dealer == 1,]

# the fraction of trades for dealers
fract_trade_dealer <- mean(sportscards_dealer$trade)
fract_trade_dealer
```

```
## [1] 0.4459459
```

```
# standard error of the fraction of trades for dealers
se_fract_trade_dealer <- sd(sportscards_dealer$trade)/sqrt(length(sportscards_dealer$tr
se_fract_trade_dealer
```

```
## [1] 0.05817759
```

```
# test
t.test(sportscards_dealer$trade,
       alternative = c("two.sided"),
       mu = 0.5, # H0
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  One Sample t-test
##
## data:  sportscards_dealer$trade
## t = -0.92912, df = 73, p-value = 0.3559
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.3299982 0.5618937
## sample estimates:
## mean of x
```

```
## 0.4459459
```

- **Conclude**

The fraction of trades for dealers in the sample was 0.4459, with a standard error of 0.05818.

Because $p - value > 0.05$, we don't reject H_0. There is no evidence of an endowment effect.

ii. Nondealers:

- **Prepare**

$H_0 : p_2 = 0.5$ (no endowment effect) v.s. $H_A : p_2 \neq 0.5$ (endowment effect), where p_2 is the fraction of trades for nondealers.

Let the significance level be 0.05.

- **Calculate**

```
# nondealer
sportscards_nondealer <- sportscards[sportscards$dealer == 0,]

# the fraction of trades for nondealers
fract_trade_nondealer <- mean(sportscards_nondealer$trade)
fract_trade_nondealer
```

```
## [1] 0.2297297
```

```
# standard error of the fraction of trades for nondealers
se_fract_trade_nondealer <- sd(sportscards_nondealer$trade)/sqrt(length(sportscards_nondealer$tra
se_fract_trade_nondealer
```

```
## [1] 0.04923441
```

```
# test
t.test(sportscards_nondealer$trade,
       alternative = c("two.sided"),
       mu = 0.5, # H0
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  One Sample t-test
##
## data:  sportscards_nondealer$trade
## t = -5.4895, df = 73, p-value = 5.559e-07
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.1316057 0.3278538
## sample estimates:
## mean of x
```

```
## 0.2297297
```

- **Conclude**

The fraction of trades for nondealers in the sample was 0.2297, with a standard error of 0.0492.

Because $p-value < 0.05$, we reject H_0. There is statistically significant evidence of an endowment effect.

iii. Difference between dealers and nondealers:

- **Prepare**

$H_0 : p_1 - p_2 = 0$ v.s. $H_A : p_1 - p_2 \neq 0$, where p_1 is the fraction of trades for dealers, and p_2 is the fraction of trades for nondealers.

Let the significance level be 0.05.

- **Calculate**

```r
# difference
trade_diff <- sportscards_dealer$trade - sportscards_nondealer$trade

# the fraction of trades for nondealers
fract_trade_diff <- mean(trade_diff)
fract_trade_diff
```

```
## [1] 0.2162162
```

```r
# standard error of the fraction of trades for nondealers
se_fract_trade_diff <- sd(trade_diff)/sqrt(length(trade_diff))
se_fract_trade_diff
```

```
## [1] 0.07268651
```

```r
t.test(trade_diff,
       alternative = c("two.sided"),
       mu = 0.5, # H0
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  One Sample t-test
##
## data:  trade_diff
## t = -3.9042, df = 73, p-value = 0.0002088
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.07135221 0.36108022
## sample estimates:
## mean of x
## 0.2162162
```

- **Conclude**

Because $p - value < 0.05$, we reject H_0. There is statistically significant difference in the behavior of Traders and non-Traders.

Chapter 3

Introduction

3.1 TA Information

TA: Chi-Yuan Fang

TA sessions: Tuesday 1:20 – 3:10 PM (SS 501)

Email: r09323017@ntu.edu.tw

Office hours: Tuesday 3:20 - 4:10 PM or by appointments (SS 643)

Class group on Facebook: Statistics with Recitation (Fall 2020) https://www.facebook.com/groups/452292659024369/

Because screens are not clear in SS 501, I would provide live screen in the group.

3.2 TA Sessions Schedule

Week	TA Sessions	Quiz	Content	Remind
1	09/15: No class			
2	09/22: Class 1		Part 1: Introduction, Data Visualization	
3	09/29: Class 2		Part 1	10/07 Turn in HW1
4	10/06: Class 3		Part 2: Distributions (1)	10/07 Turn in HW1, 10/13 Quiz 1
5	10/13: Class 4	Quiz 1	Part 2	10/21 Turn in HW2

Week	TA Sessions	Quiz	Content	Remind
6	10/20: Class 5		Part 3: Distributions (2)	10/21 Turn in HW2, 10/27 Quiz 2
7	10/27: Class 6	Quiz 2	Part 3	11/04 Turn in HW3
8	11/03: Class 7		Part 4: Test	11/04 Turn in HW3, 11/10 Quiz 3
9	11/10: Class 8	Quiz 3	Part 4	**11/18 Midterm**
10	11/17: Class 9		Review and Q&A	**11/18 Midterm**, 11/25 Turn in HW4
11	11/24: Class 10		Part 5: Model (1) ANOVA	11/25 Turn in HW4, 12/01 Quiz 4
12	12/01: Class 11	Quiz 4	Part 5	12/09 Turn in HW5
13	12/08: Class 12		Part 6: Model (2) Regression	12/09 Turn in HW5, 12/15 Quiz 5
14	12/15: Class 13	Quiz 5	Part 6	12/23 Turn in HW6
15	12/22: Class 14		Review and Q&A	12/23 Turn in HW6, 12/29 Quiz 6
16	12/29: Class 15	Quiz 6	Review and Q&A	**01/13 Final Exam**
17	01/05: No class			**01/13 Final Exam**
18	01/12: No class			**01/13 Final Exam**

3.3 Reference

What is a good book on learning R with examples? https://www.quora.com/What-is-a-good-book-on-learning-R-with-examples

Chapter 4

Data Set: `possum`

4.1 Data Description

https://www.openintro.org/data/index.php?data=possum

4.1.1 Background

Data representing possums in Australia and New Guinea. This is a copy of the data set by the same name in the DAAG package, however, the data set included here includes fewer variables.

4.1.2 Variables

- pop - Population, either Vic (Victoria) or other (New South Wales or Queensland).
- sex - Gender, either m (male) or f (female).
- age - Age.
- head_l - Head length, in mm.
- skull_w - Skull width, in mm.
- total_l - Total length, in cm.
- tail_l - Tail length, in cm.

4.2 Input Data

4.2.1 csv File

```
# input data

#possum_csv <- read.csv('/Users/chi-yuan/Desktop/NTU ECON/Statistics/Stat/possum.csv')
```

Remark How to get a file path on a Mac?

1. Right-click the file
2. Click Get Info

4.2.2 Package

```
library(openintro)
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
#library(tidyverse)
#data(COL)

data(possum)
```

4.3 Data Cleaning

4.3.1 Remove Missing Values

```
# return a logical vector indicating which cases are complete, i.e., have no missing v
possum_new <- possum[complete.cases(possum),]
```

4.3.2 Delete Rows with Specific Condition(s)

For example, we want to delete $site = 2$ rows.

```
possum_new2 <- possum_new[possum_new$site != 2, ]
```

4.4 Correlation

Calculate the correlation coefficient between x (total_l) and y (head_l).

```
x <- possum_new$total_l
y <- possum_new$head_l

cor(x, y)
```

```
## [1] 0.6742892
```

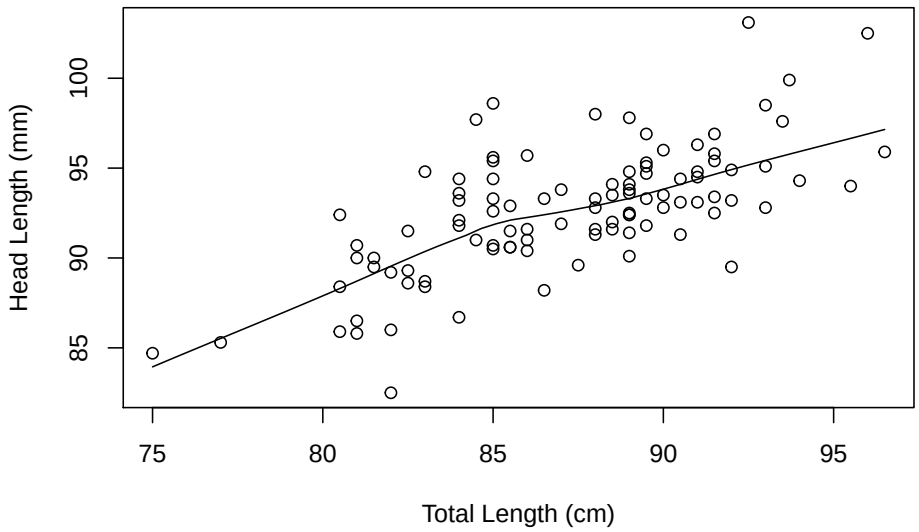
4.5 Graphical Analysis

4.5.1 Scatter Plot

Make a scatterplot for x (total_l) and y (head_l).

```
scatter.smooth(x = x, y = y,
               xlab = "Total Length (cm)",
               ylab = "Head Length (mm)",
               main = "Figure 8.6 (p.308)")
```

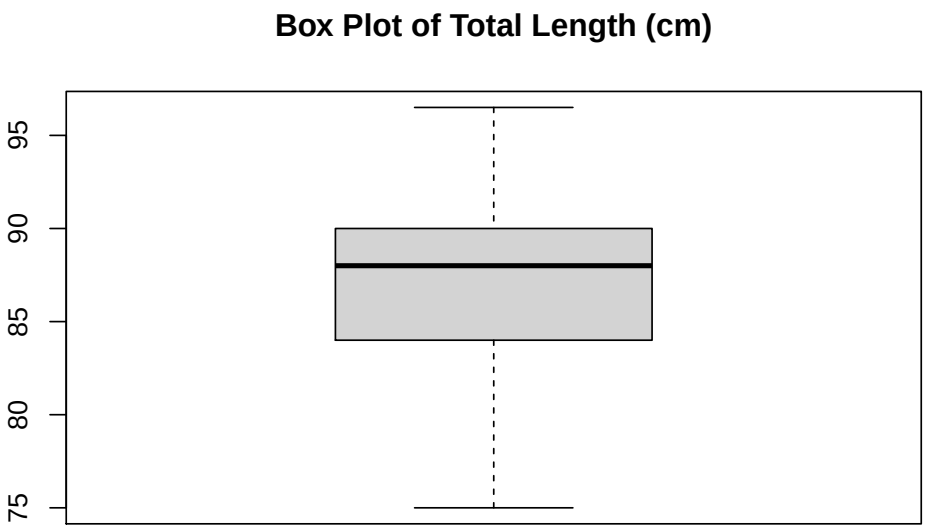
Figure 8.6 (p.308)



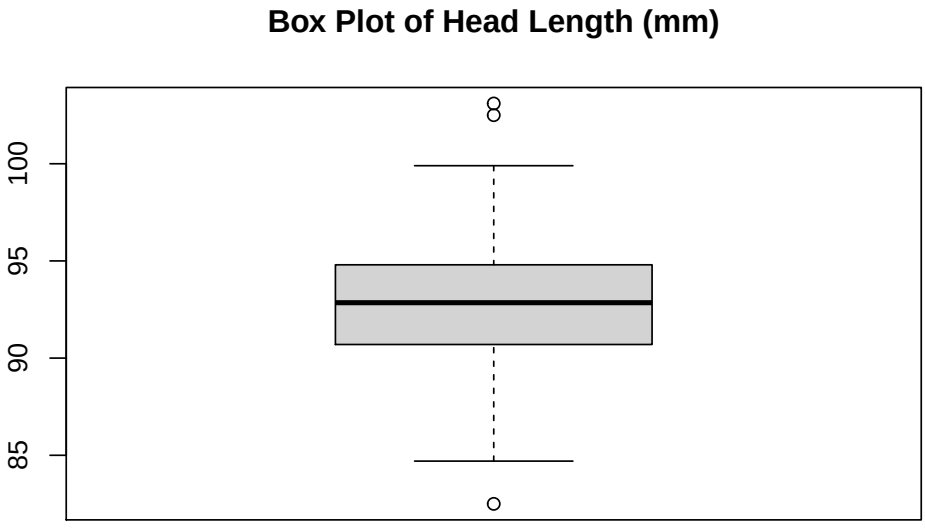
4.5.2 Box Plot: Check for Outliers

Make box plots for x (total_l) and y (head_l), respectively.

```
boxplot(x, main = "Box Plot of Total Length (cm)")
```



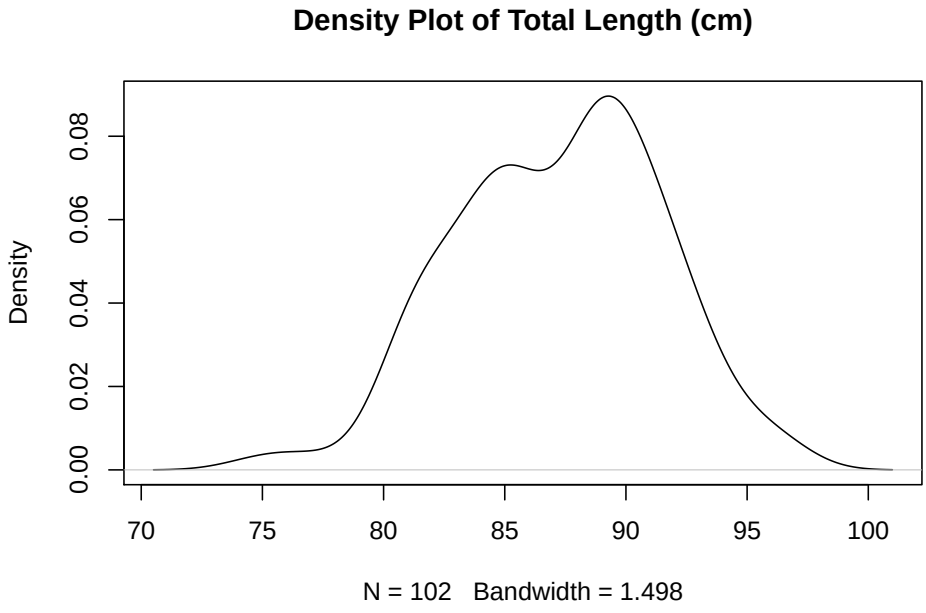

```
boxplot(y, main = "Box Plot of Head Length (mm)")
```



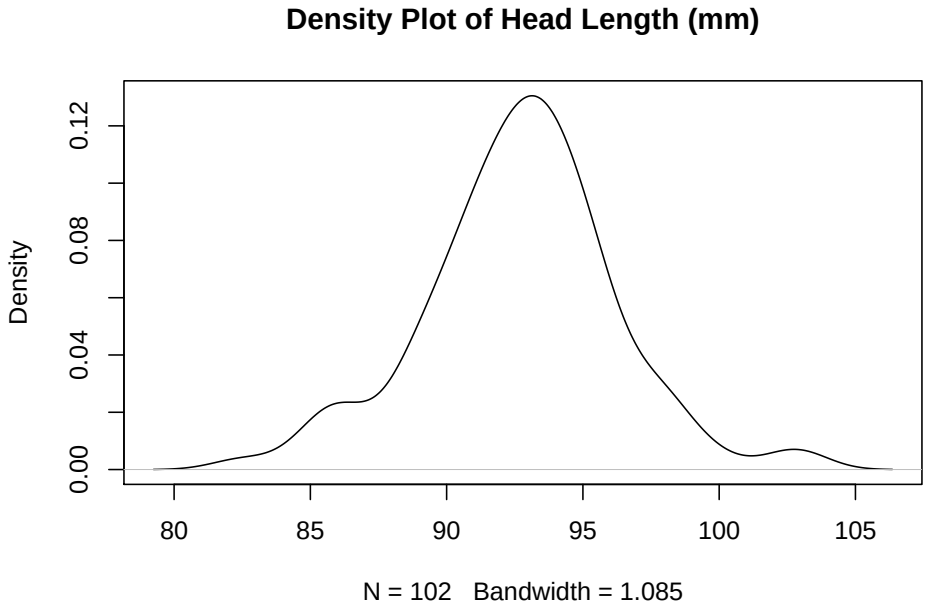
4.5.3 Density Plot: Check for Normality

Make density plots for x (total_l) and y (head_l), respectively.

```
plot(density(x), main = "Density Plot of Total Length (cm)")
```




```
plot(density(y), main = "Density Plot of Head Length (mm)")
```



4.6 Linear Regression

4.6.1 Model

Fit the least squares regression

$$head_l = \beta_0 + \beta_1 total_l + e. \tag{4.1}$$

```
# y ~ x
fit <- lm(head_l ~ total_l, data = possum_new)

summary(fit)
```

```
##
## Call:
## lm(formula = head_l ~ total_l, data = possum_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.226 -1.593 -0.326  1.303  7.424
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.25900    5.41959   7.982 2.49e-12 ***
## total_l      0.56667    0.06206   9.131 7.95e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.618 on 100 degrees of freedom
## Multiple R-squared:  0.4547, Adjusted R-squared:  0.4492
## F-statistic: 83.37 on 1 and 100 DF,  p-value: 7.946e-15
```

We have

$$\hat{\beta}_0 = 43.25900 \tag{4.2}$$

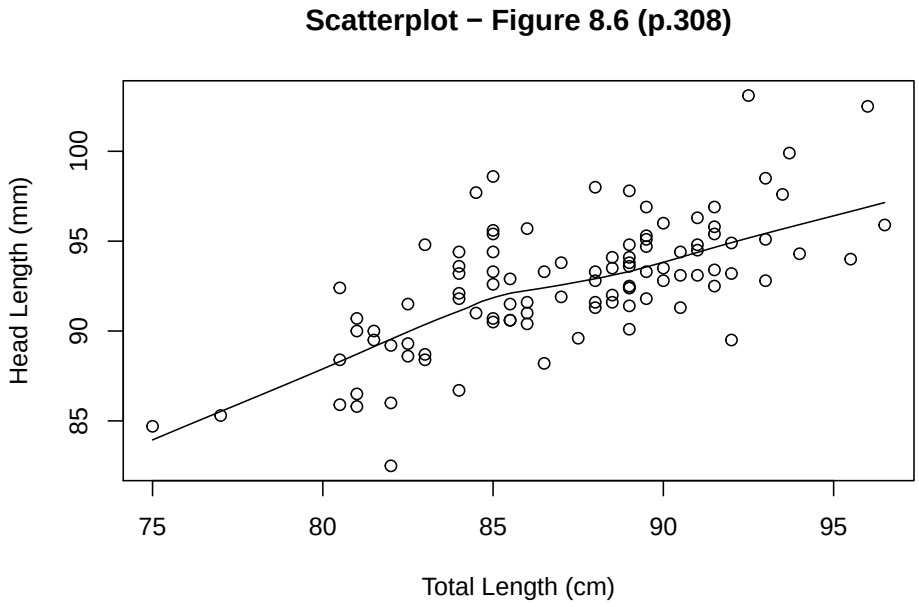
$$\hat{\beta}_1 = 0.56667. \tag{4.3}$$

4.6.2 Residual Analysis

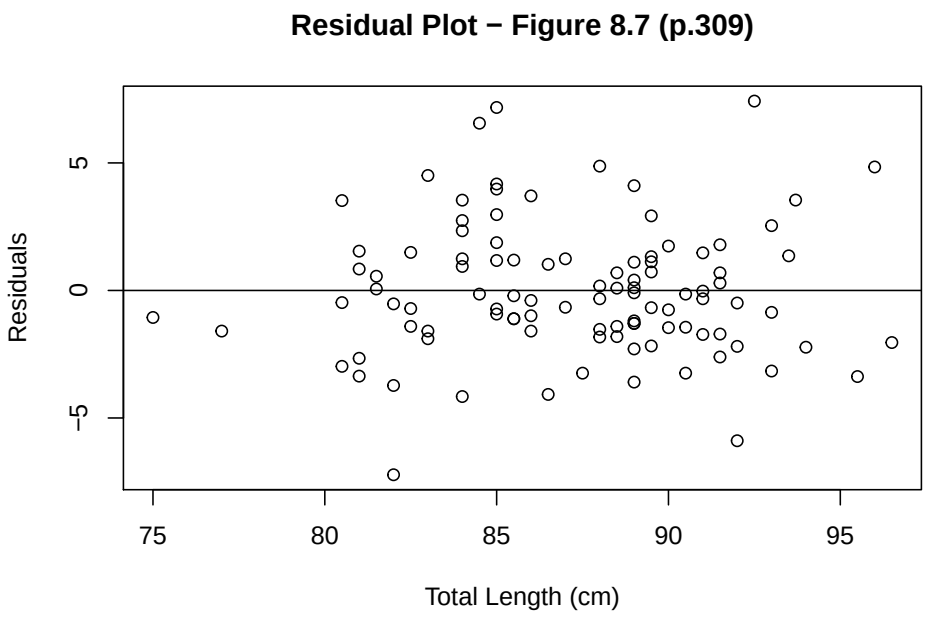
Make a scatterplot and a residual plot for regression. Discuss whether fitting a linear model would be appropriate.

```
# scatterplot
scatter.smooth(x = x, y = y,
```

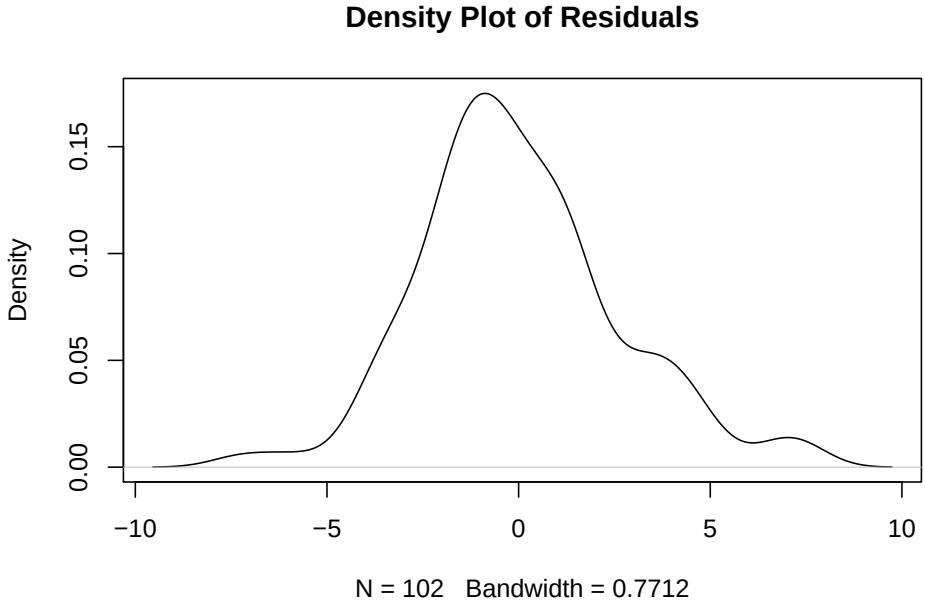
```
                xlab = "Total Length (cm)",
                ylab = "Head Length (mm)",
                main = "Scatterplot - Figure 8.6 (p.308)")
```




```
# residual plot
plot(x = possum_new$total_l, y = fit$residuals,
     xlab = "Total Length (cm)",
     ylab = "Residuals",
     main = "Residual Plot - Figure 8.7 (p.309)")
abline(h=0)
```




```
# density plot
plot(density(fit$residuals), main = "Density Plot of Residuals")
```



Check the following conditions:

- Linearity: linear trend, i.e., no patterns in residual plot. (valid)
- Normal residuals: no extremely large or small residuals. (valid)

- Constant variability: points around line dispersed in similar way. (valid)
- Independent observations: occurrence of one observation provides no information about occurence of the other. (valid)

Thus, fitting a linear model would be appropriate for this case.