

Dealing with Data II Homework 4 Solutions

Joshua D. Ingram

2022-10-14

Reading from the textbook

a. Chapter 7: On the Normal Distribution and working with it

- Section 7.2

b. Chapter 12: More on the population regression model, and confidence intervals for the slope and prediction intervals

- Section 12.1: Review, but with the key definition of the Population Regression Model on page 636
- Section 12.2: Inference about the slope β (Hypothesis test and confidence interval)
- Section 12.3: skip!
- Section 12.3: Standardized Residuals, Confidence Interval for the mean response. Skip the last few page about the ANOVA table.

Reading the R Handout

a. Read Sections 5 and 6 of the R handout: “Scatterplots and Linear Regression” posted under Files -> R Handouts

Problems

1. Old Faithful (continued from last HW)

The Old Faithful dataset (available in R by typing `faithful` in the R prompt, e.g., `> faithful`) is a dataset about the duration of an eruption of that famous geyser (e.g., 3.6 minutes for the first observation in the dataset) and the waiting time to the next eruption (called “waiting”, e.g., 79 minutes for the first observation).

```
fit <- lm(waiting ~ eruptions, data = faithful)
summary(fit)

##
## Call:
## lm(formula = waiting ~ eruptions, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0796  -4.4831   0.2122   3.9246  15.9719
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.4744     1.1549   28.98  <2e-16 ***
## eruptions     10.7296     0.3148   34.09  <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.914 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic: 1162 on 1 and 270 DF, p-value: < 2.2e-16
```

h. Construct a 95% confidence interval for the slope and interpret it in context.

Answer:

```
# c.i. = b +/- margin of error
confint <- c(10.7296 - (1.96 * 0.3148), 10.7296 + (1.96 * 0.3148))
confint
```

```
## [1] 10.11259 11.34661
```

With 95% confidence, we predict that for every 1 minute increase in the eruption time, we expect the waiting time for the next eruption to increase by between 10.11259 and 11.34661 minutes, with a point estimate of a 10.7296 minute increase.

i. Construct a 95% confidence interval for the predicted (or mean) waiting time, for eruptions that last 4.5 minutes.

Answer:

```
predict(fit, newdata = data.frame(eruptions = 4.5), se = TRUE)

## $fit
##      1
## 81.75778
##
## $se.fit
## [1] 0.4796786
##
## $df
## [1] 270
##
## $residual.scale
## [1] 5.914009

confint <- c(81.75778 - 1.96 * 0.4796786, 81.75778 + 1.96 * 0.4796786)
confint
```

```
## [1] 80.81761 82.69795
```

With 95% confidence, we predict that the average waiting time between eruptions with an eruption time of 4.5 minutes is between 80.81761 and 82.69795 minutes, with a point estimate of 81.75778 minutes.

2. Textbok Exercise 12.22 - b.

Refer to the cell phone data set from Exercise 12.9 about various specs of cell phones, which is preloaded in the *Linear Regression* app. Treat the weight of the phone as the response and the capacity of the battery as the explanatory variable. Remove the outlier (phone no. 70).

- b. Confirm, using the output of the *Linear Regression* app or other software, that the 95% confidence interval for the population slope equals (0.028, 0.060). Interpret the interval and explain the correspondence with the result of the significance test in part a.

Instead of using the web app, download the data and find the p-value for the hypothesis test and the confidence interval in R.

Answer:

```
library(readr)
cell <- read_csv("https://img1.wsimg.com/blobby/go/bbca5dba-4947-4587-b40a-db346c01b1b3/downloads/Cellp

## Rows: 79 Columns: 13
## -- Column specification -----
## Delimiter: ","
## chr (5): Display, Resolution, Battery, Facebook, GPS
## dbl (8): Phone, Size(in), Talk_time(min), Storage(GB), Weight(g), Battery_Ca...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
# remove cell phone number 70 from the data
cell <- cell[-which(cell$Phone == 70),]
cell$capacity <- cell$`Battery_Capacity(mAh)`

fit <- lm(`Weight(g)` ~ capacity, data = cell)
summary(fit)

##
## Call:
## lm(formula = `Weight(g)` ~ capacity, data = cell)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34.010 -15.822  -2.981   9.580  61.604
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.658768  11.181471   5.962 7.34e-08 ***
## capacity     0.043568   0.008057   5.407 7.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.63 on 76 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2683
## F-statistic: 29.24 on 1 and 76 DF,  p-value: 7.123e-07
```

Our null and alternative hypotheses for our hypothesis test are as follows (with a significance level of $\alpha = 0.05$):

$$H_0 : \beta = 0$$

$$H_A : \beta \neq 0$$

From the `summary(fit)` function, we see we obtain a t-statistic of 5.407 for the slope, which corresponds to a p-value of 7.12e-07. This p-value is below our significance level, so we reject the null hypothesis and have evidence that $\beta \neq 0$. Our 95% confidence interval does not contain 0, which agrees with our hypothesis test, and is given by:

```
c(0.043568 - 1.96 * 0.008057 , 0.043568 + 1.96 * 0.008057 )
```

```
## [1] 0.02777628 0.05935972
```

With 95% confidence, we predict that for every 1 mAh increase in battery capacity, we expect that the weight of a phone to increase by between 0.0278 and 0.0594 grams, with a point estimate of 0.0436 grams.

3. Textbook Exercise 12.52 - a.

Refer to the cell phone specs data file on the book's website, which is also preloaded in the *Linear Regression* app. Regress $y = \text{weight}$ on $x = \text{capacity of battery}$, excluding the outlier (phone no. 70).

- a. Stating the necessary assumptions, find a 95% confidence interval for the mean weight of cell phones with a battery capacity of 1500 mAh. Interpret the interval.

This refers to the same dataset. Use R to obtain the confidence interval asked for in part a, and interpret.

Answer:

The necessary assumptions for making these inferences are the following:

1. Relationship in population satisfies the regression model (linear)
2. Data gathered randomly
3. Population y values at each x value have a normal distribution
4. Response has the same variance at each x value

```
predict(fit, newdata = data.frame(capacity = 1500), se = TRUE)
```

```
## $fit
##      1
## 132.0103
##
## $se.fit
## [1] 2.716695
##
## $df
## [1] 76
##
## $residual.scale
## [1] 21.6309
```

```
# confidence interval
c(132.0103 - 1.96 * 2.716695, 132.0103 + 1.96 * 2.716695)
```

```
## [1] 126.6856 137.3350
```

With 95% confidence, we predict that the average phone weight with a battery capacity of 1500 mAh is between 126.6856 and 137.335 grams, with a point estimate of 132.0103 grams.