STAR 513: HW 2

Yvette Uwineza

Total points: 24
Questions are worth **2 pts** each, except where noted.
See Canvas calendar for due date.

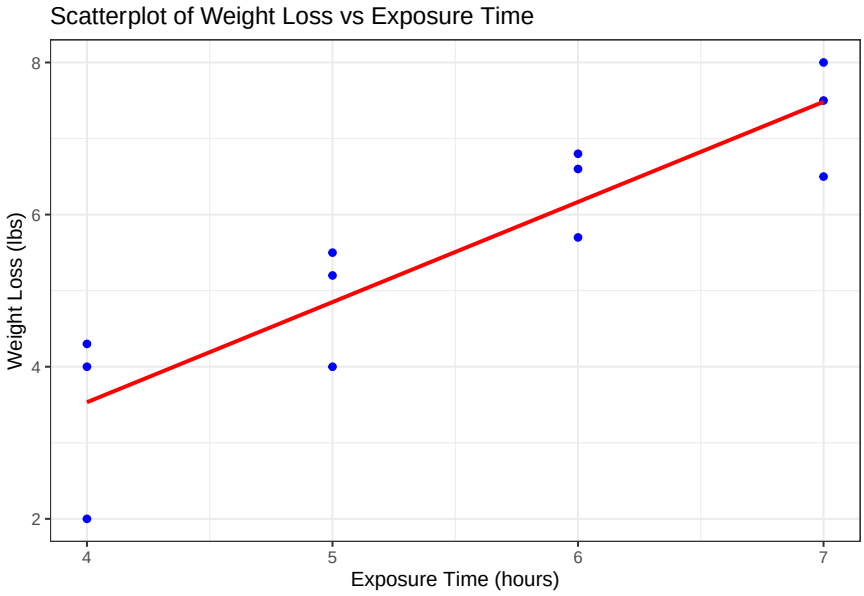
Homework should be submitted as a pdf, doc or docx file via Canvas.
Use of R markdown HW template is strongly encouraged.
Add or delete code chunks as needed.
Knit frequently to avoid last minute problems!
Your submitted assignment should be neatly formatted and organized.

Ott & Longnecker Example 11.32: A chemist is interested in the association between weight loss in lbs (y) versus the exposure time in hours (x) for a particular compound. The data includes n = 12 observations. The data ex11-32.csv is available from Canvas.

This assignment is very similar to Lec02_examples: Simple Linear Regression!

Q1 (4 pts)

For this question, please use the ggplot2 package (available through tidyverse). You may need to install this package if it is the first time you have used it. Create a scatterplot of the data with fitted regression line overlaid. Your plot should include axis labels that include the units for each variable.



Q2

Fit an appropriate regression model and show the summary() output.

```
##
## Call:
## lm(formula = WeightLoss ~ ExposureTime, data = ex11_32)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5333 -0.5625  0.3917  0.5458  0.7667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.7333     1.1652  -1.488    0.168
## ExposureTime   1.3167     0.2076   6.342 8.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8041 on 10 degrees of freedom
## Multiple R-squared:  0.8009, Adjusted R-squared:  0.781
## F-statistic: 40.22 on 1 and 10 DF,  p-value: 8.437e-05
```

Q3

For this question, please use the tidy() function from the broom package. You may need to install this package if it is the first time you have used it. From the model you fit in the previous question, present "tidy" results.

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     -1.73     1.17      -1.49 0.168
## 2 ExposureTime     1.32     0.208      6.34 0.0000844
```

Q4 (4 pts)

Provide a detailed interpretation of the estimated **slope** in context of this research study. Your interpretation should include appropriate units and the numeric value for the estimated slope.

The slope estimated above is 1.317. It can be interpreted as follow: For one hour of additional exposure time, the expected weight loss is increased by 1.317 lbs.

Q5

Consider the p-value corresponding to ExposureTime. State the null hypothesis using standard greek letter notation and subscripting. Hint: See the end of the Lec1_notes for LaTex code examples.

Question 5: Null Hypothesis in LaTeX

$H_0 : \beta_1 = 0$ No linear relationship between exposure time and weight loss

$H_A : \beta_1 \neq 0$ There is a linear relationship between exposure time and weight loss.

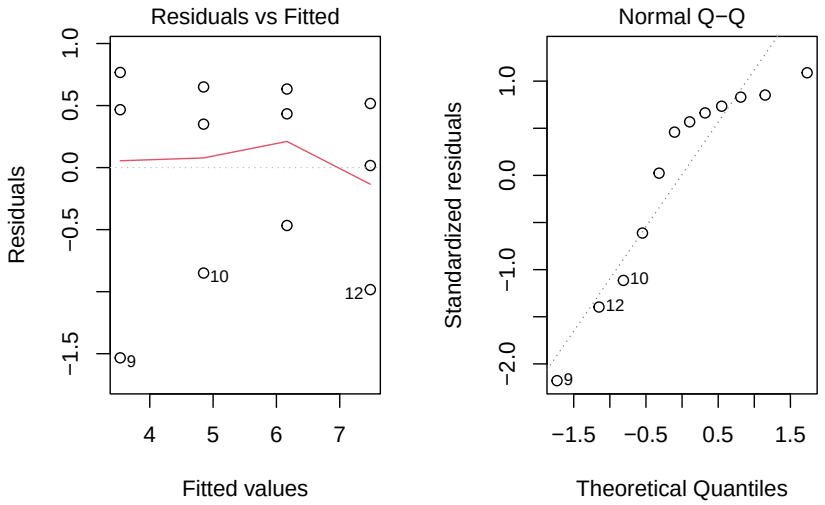
Q6

Do we have evidence (at the alpha = 0.05 level) of a linear association between weight loss and exposure time? Is the association positive or negative? Justify your response using an appropriate p-value.

At 0.05 significance level, the pvalue is less than 0.05, so we reject the null hypothesis that there is no linear relationship between exposure time and weight loss. Looking at the slope we got earlier which is positive 1.317, we can say that exposure time increases so does the weight loss as well.

Q7

Create the plots of (1) residuals vs fitted values and (2) qqplot of residuals.



Q8 (4 pts)

The four assumptions of simple linear regression are listed below. For each assumption, state a graph that can be used to check the assumption. If an assumption cannot be checked graphically, write "Cannot be checked graphically". You do NOT need to evaluate the assumptions for this question.

Independence: We can not check it graphically Equal variance: We can use the residuals vs fitted values plots Normality of Residuals: QQ plots of residuals can be used. Linearity: We can use the residuals vs fitted values plots

---

Q9

Use model.matrix() to examine the design or model matrix (but you do not need to include it in your assignment).
How many rows are there? How does the number of rows relate to the number of observations (n)?
How many columns are there? How does the number of columns relate to the number of model coefficients/parameters/"betas"?

Number of rows = 12 = the number of rows represents the number of observations which are 12 in this case.

Number of cols = 2 = the number of columns represent two coefficients, the intercept and the slope.

```
## Number of rows: 12
```

```
## Number of columns: 2
```

Appendix

```
#Retain this code chunk!!!
library(knitr)
library(tidyverse)
library(ggplot2)
library(broom)

knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)



# loading the data
```

```r
ex11_32 <- read_csv("Homework_1/ex11-32.csv")


#Q1

ggplot(ex11_32, aes(x = ExposureTime, y = WeightLoss)) +
  geom_point(color = "blue") +  # Scatter points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Regression line
  labs(
    title = "Scatterplot of Weight Loss vs Exposure Time",
    x = "Exposure Time (hours)",
    y = "Weight Loss (lbs)"
  ) +
  theme_bw()

#Q2

# fitting a model
model <- lm(WeightLoss ~ ExposureTime, data = ex11_32)

summary(model)


#Q3

results <- tidy(model)


print(results)
#Q7

par(mfrow = c(1,2))

plot(model, which = 1)

plot(model, which = 2)

par(mfrow = c(1,1))


#Q9

# model mattrix
Value <- model.matrix(model)


num_rows <- nrow(Value)
num_cols <- ncol(Value)


cat("Number of rows:", num_rows, "\n")
cat("Number of columns:", num_cols, "\n")
```