Lab 3: Multiple Linear Regression

Reese Quillian

2022-10-10

```
# Loading packages and files

traindir <- "C:/Users/Student/OneDrive - University of Virginia/Documents/SYS4021/In Class/Data/Train Da
sourcedir <-"C:/Users/Student/OneDrive - University of Virginia/Documents/SYS4021/Lab 3"

# load data
setwd(sourcedir)
source("AccidentInput.R")

# load libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.1.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(devtools)
```

```
## Warning: package 'devtools' was built under R version 4.1.3

## Loading required package: usethis

## Warning: package 'usethis' was built under R version 4.1.3
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.3

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
acts <- file.inputl(traindir)

totacts <- combine.data(acts)
```

Part 2

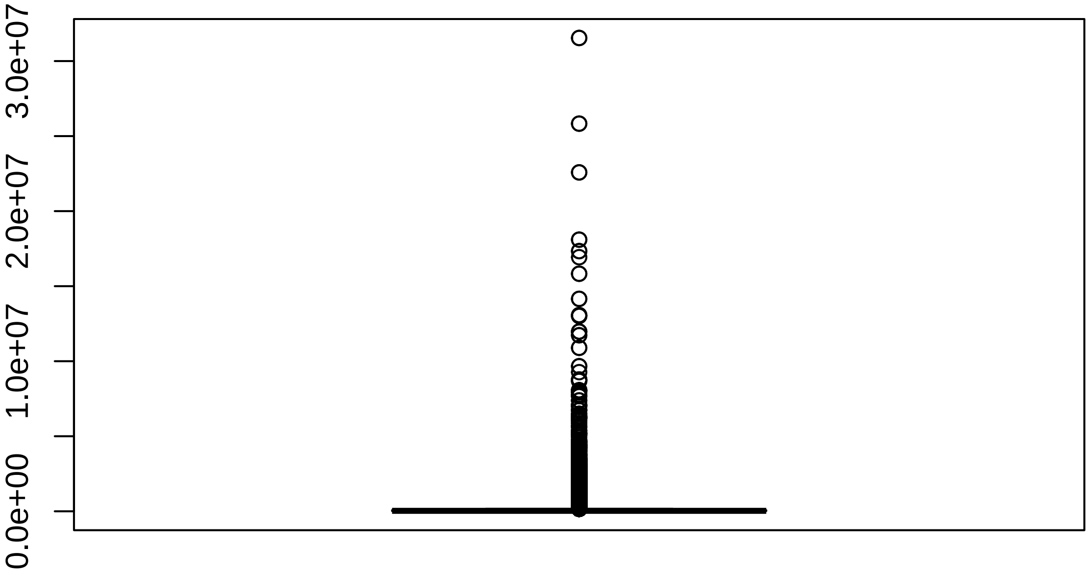
For this part you should use a data frame that has all rail accident data from year 2001 to 2022.

1. Create a new data frame which includes only accidents above the upper whisker for ACCDMG.
2. Remove 9/11 for your extreme accidents data frame.
3. Remove the duplicated reports from the new data frame with these extreme accidents.
4. For the all of the questions in this lab use this new "de-duplicated" data frame.

In all the following questions, use the significance level of 0.05. Use 2 digits to the right of the decimal for any numeric response.

```
# Build a data frame with only extreme accidents for ACCDMG

dmgbox <-boxplot(totacts$ACCDMG)
```



```r
# accidents above upper whisker
xdmg <- totacts[totacts$ACCDMG > dmgbox$stats[5],]

#remove 9/11
xdmg <- xdmg[-183,]

## Remove duplicates from xdmg and call new data frame xdmgnd
xdmgnd <- xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY", "TIMEHR", "TIMEMIN")])),]
```

Question 22 of 41 Build a model accdmg.lm1, ACCDMG~TEMP + TRNSPD + CARS + HEADEND1.

In the model accdmg.lm1, the value of adjusted R2 is:

```r
accdmg.lm1<-lm(ACCDMG~TEMP+TRNSPD+CARS+HEADEND1, data=xdmgnd)
summary(accdmg.lm1)
```

```
##
## Call:
## lm(formula = ACCDMG ~ TEMP + TRNSPD + CARS + HEADEND1, data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2334474  -361168  -203644    63205 31048030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 479947.6    45017.8  10.661  < 2e-16 ***
## TEMP           153.9      583.5   0.264    0.792
## TRNSPD       18017.4      786.7  22.902  < 2e-16 ***
## CARS          4057.1     1028.0   3.947 8.00e-05 ***
## HEADEND1    -69702.0    10323.6  -6.752 1.57e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1205000 on 7596 degrees of freedom
## Multiple R-squared:  0.06693,    Adjusted R-squared:  0.06644
## F-statistic: 136.2 on 4 and 7596 DF,  p-value: < 2.2e-16
```

Question 23 of 41 Which of the following is true about accdmg.lm1?

A. accdmg.lm1 is significant overall.
B. All of the predictors in accdmg.lm1 are significant at the 0.05 level.
C. None of the predictors in accdmg.lm1 are significant at the 0.05 level. D. All of the predictors except 1 in accdmg.lm1 are significant at the 0.05 level.

```r
# A and D are true
```

Question 24 of 41 Which interpretation(s) of accdmg.lm1 is(are) correct?

A. CARS is statistically significant in explaining the variance of total accident damage *true B. If the train speed (TRNSPD) increases, there is a statistically significant increase in accident damage.* true C. If temperature increases, there is a statistically significant decrease in accident damage. D. This model allows us to examine the relationship between train speed and accident damage while controlling for the effects of temperature, number of cars carrying hazmat, and number of head end locomotives.

```
coef(accdmg.lm1)
```

```
## (Intercept)        TEMP      TRNSPD        CARS    HEADEND1
## 479947.6141    153.9491  18017.3925   4057.1264 -69701.9690
```

Question 25 of 41 Answer the following questions about the estimated parameters for the linear model accdmg.lm1:

The intercept, B0, is:

The coefficient for TRNSPD, BTRNSPD is:

```
# B0 = 479947.6141
# BTRSNPD = 18017.3925
```

Question 26 of 41 The model accdmg.lm1 has an F-statistic of ____________________ on ____________________________ and _________________________ degrees of freedom.

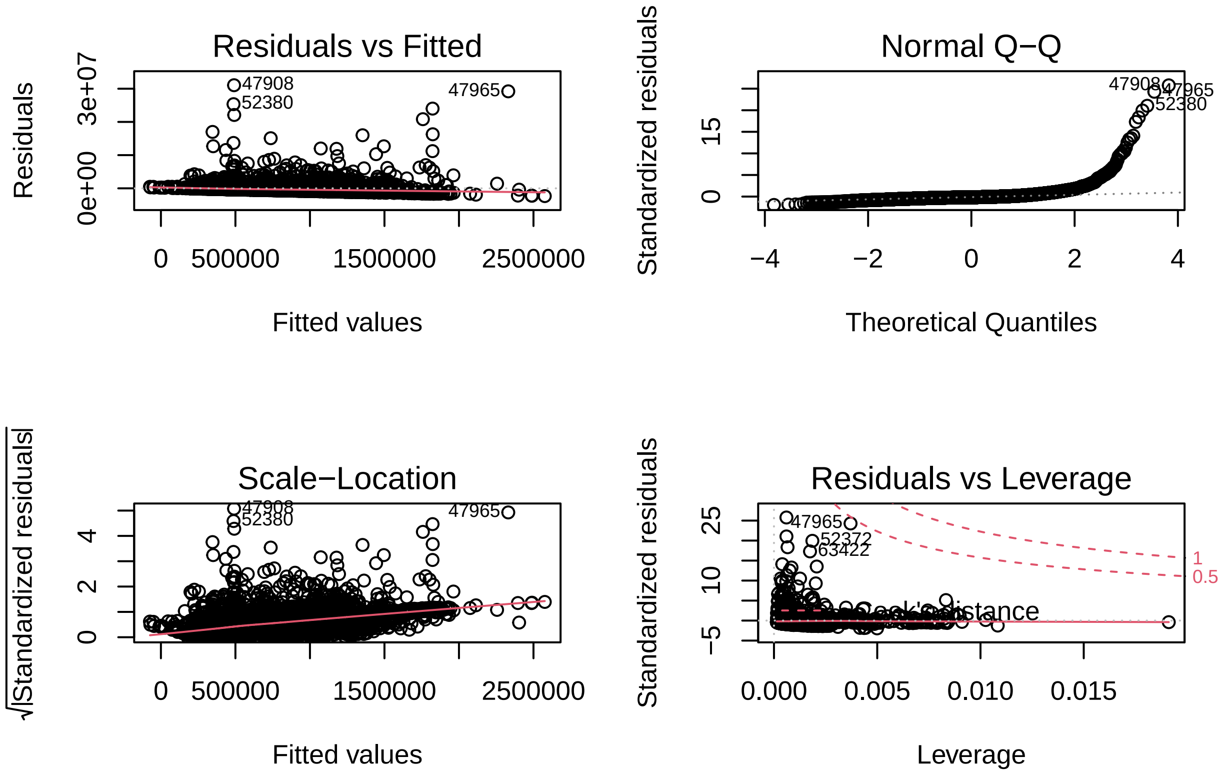
```
# F-statistic: 136.2 on 4 and 7596 DF,  p-value: < 2.2e-16
```

Question 27 of 41 Create diagnostic plots for this new model accdmg.lm1.

There are (is) _________ point(s) with a Cook's distance greater than 0.4.

A. 0 B. 1 C. 2 D. 3

```
par(mfrow = c(2, 2))
plot(accdmg.lm1)
```



```
cooks <- cooks.distance(accdmg.lm1)
cooks[(cooks > 0.4)]
```

```
##    47965
## 0.439021
```

Question 28 of 41 Which of the following statements is true about the observation with the highest Cook's distance?

A. The accident is of type derailment. B. This accident has Cook's distance greater than 1.0
C. This accident had over 50 injuries. D. This accident costs more than $30 million.

Question 29 of 41 Based on the Residual vs Fitted diagnostic plot for accdmg.lm1, which of the following statements are true? A. The model accdmg.lm1 shows no violation of the regression assumptions.
B. The variance of the error term is constant and has a mean of 0. C. There are no outliers
D. The error term has a non constant variance.

Question 30 of 41 Based on your diagnostic plots for accdmg.lm1, the residuals do not follow a normal distribution indicating the potential need for a transformation of the response.

A. True B. False

Question 31 of 41 Use the boxcox function (load library(MASS) first) with model accdmg.lm1. The optimal lambda to normalize ACCDMG is ________________.

Question 32 of 41 Use the boxcox function with model accdmg.lm1. The maximum y value (log-likelihood) from the boxcox function is ____________________.

Question 33 of 41 Transform the response variable ACCDMG as suggested by the Box-Cox plot. Build a new model with the transformed response variable and the same predictors as in model accdmg.lm1. Call this new model accdmg.lm1.trans.

Which of the following statements is true?

A. accdmg.lm1 is better based on the diagnostics plots.
B. accdmg.lm1.trans is better based on the diagnostics plots.
C. In the transformed model, there are 3 points in the Cook's distance greater than 0.5.
D. In the transformed model, there are no points in the Cook's distance greater than 0.02.

Question 34 of 41 Consider the WEATHER variable. Suppose that we want to determine if accident damage (ACCDMG) that occurred in foggy weather is the same as the damages that occur in rainy weather. We can use linear regression models to test this hypothesis. Which of the following statements is correct?

A. We can use any weather level as the base case. B. We don't have to convert the WEATHER variable into a categorical variable before building the linear model. C. We can use the default coding in R to code as.factor(WEATHER) by dummy variables. D. We can use foggy weather or rainy weather as the base case.

Question 35 of 41 Build a model accdmg.lm2, to predict accident damage in terms of type of accident. response variable: ACCDMG predictor: TYPE (with thirteen values)

_________________________ dummy variables are used to code TYPE

Question 36 of 41 In the model accdmg.lm2, the default base case (without recoding) in R is: A. Derailment B. Head on collision C. Rear end collision D. Side collision E. Raking collision F. Other

Question 37 of 41 Create a new categorical variable, Derail using 1 dummy variable- 1 level denoting accidents of type derailment and 0 denoting all other types of accidents that are not of type derailment.

Build a new model accdmg.lm3 to compare derailments to all other types of accidents (your model should have only 1 predictor). Test the following hypothesis: "Derailments do not increase the severity of accident damage."

Which of the follow statements is true?

A. Derailments cost significantly more than other types of accidents. B. Derailments cost significantly less than other types of accidents. C. There is not a significant difference in the cost of derailments and other types of accidents.

Question 38 of 41 Build a stepwise model from the full main effects + interaction model with Derail, TRNSPD, TONS, CARS, and HEADEND1. Call the main effects + interaction model accdmg.lm4 and the stepwise model accdmg.lm4.step. Which of the following statements is true about the stepwise model?

A. Derailments cost significantly more than other types of accidents.
B. High speed derailments cost significantly more that other types of accidents
C. High weight derailments cost significantly more than other types of accidents. D. There are 10 parameters significant at the $P < 0.05$ level.

Question 39 of 41 Perform a partial F test between accdmg.lm4 and accdmg.lm4.step. Which model do you choose?

A. accdmg.lm4 B. accdmg.lm4.step C. They are the same. D. None of the above

Question 40 of 41 Which of the following statements is true about the models accdmg.lm4 and accdmg.lm4.step?

A. accdmg.lm4 is the best based on AIC.
B. The BIC is the same for accdmg.lm4 and accdmg.lm4.step. C. accdmg.lm4.step is the best model according to AIC. D. none of the above

Question 41 of 41 Plot the following interactions with log(ACCDMG) using the median as the cutpoint for TRNSPD and 1 as the cutpoint for CARS.

TRNSPD, CARS TRNSPD, Freight TRNSPD, Derail Note, you'll have to create a binary variable for Freight like you did for Derail. Which of the following statements are true?

A. The interaction plot of TRNSPD and CARS with log(ACCDMG) shows there are greater damages at high speeds when more hazard cars derail.
B. High speed freight train accidents have less damages than high speed non-freight accidents. C. For the interaction plot of TRNSPD and Derail with ACCDMG, the lines have different non-zero slopes, indicating a main effect from TRNSPD and an interaction between TRNSPD and Derail. D. High speed derailments result in more damages on average than low speed derailments.