Homework # 1
Challenge 15- R
July 22, 2021

Kari Hodge

Contents

1 Introduction

AutosRUs' newest prototype, the MechaCar, is suffering from production troubles that are blocking the manufacturing team's progress. The following analysis will be performed:

1. Perform multiple linear regression analysis to identify which variables in the dataset predict the mpg of MechaCar prototypes.
2. Collect summary statistics on the pounds per square inch (PSI) of the suspension coils from the manufacturing lots.
3. Run t-tests to determine if the manufacturing lots are statistically different from the mean population.
4. Design a statistical study to compare vehicle performance of the MechaCar vehicles against vehicles from other manufacturers. For each statistical analysis, you'll write a summary interpretation of the findings.

```
# read in both data sets
mecha<-read.csv(file='/Users/karihodge/Dropbox/UT_Data_Camp/Class_Folder/R-AutosRUs/MechaCa
head(mecha)

##   vehicle_length vehicle_weight spoiler_angle
```

```
## 1       14.69710       6407.946      48.78998
## 2       12.53421       5182.081      90.00000
## 3       20.00000       8337.981      78.63232
## 4       13.42849       9419.671      55.93903
## 5       15.44998       3772.667      26.12816
## 6       14.45357       7286.595      30.58568
##   ground_clearance AWD      mpg
## 1         14.64098   1 49.04918
## 2         14.36668   1 36.76606
## 3         12.25371   0 80.00000
## 4         12.98936   1 18.94149
## 5         15.10396   1 63.82457
## 6         13.10695   0 48.54268
```

```
susp<-read.csv(file='/Users/karihodge/Dropbox/UT_Data_Camp/Class_Folder/R-AutosRUs/Suspensi
head(susp)
```

```
##   VehicleID Manufacturing_Lot  PSI
## 1    V40858              Lot1 1499
## 2    V40607              Lot1 1500
## 3    V31443              Lot1 1500
## 4     V6004              Lot1 1500
## 5     V7000              Lot1 1501
## 6    V17344              Lot1 1501
```

2 Deliverable 1

Deliverable 1: Linear Regression to Predict MPG

The MechaCar datavset contains mpg test results for 50 prototype MechaCars. The MechaCar prototypes were produced using multiple design specifications to identify ideal vehicle performance. Multiple metrics, such as vehicle length, vehicle weight, spoiler angle, drivetrain, and ground clearance, were collected for each vehicle.

2.1 Test Normality

Evaluate Skewness and Kurtosis

```
library(psych)
describe(mecha)
```

```
##                  vars  n    mean      sd  median trimmed
## vehicle_length      1 50   15.02    2.03   14.60   14.83
## vehicle_weight      2 50 6154.15 1846.09 5928.69 6122.17
## spoiler_angle       3 50   57.12   19.56   58.55   58.06
## ground_clearance    4 50   12.71    2.53   12.98   12.71
## AWD                 5 50    0.50    0.51    0.50    0.50
##                    mad  min   max range  skew kurtosis
```

```
## vehicle_length      1.78   12    20     8  0.76    -0.28
## vehicle_weight   1802.33 2000 10000  8000  0.13    -0.52
## spoiler_angle      15.75    0    90    90 -0.53     0.04
## ground_clearance    2.55    6    18    12 -0.16    -0.29
## AWD                 0.74    0     1     1  0.00    -2.04
##                      se
## vehicle_length     0.29
## vehicle_weight   261.08
## spoiler_angle      2.77
## ground_clearance   0.36
## AWD                0.07
##  [ reached 'max' / getOption("max.print") -- omitted 1 rows ]
```

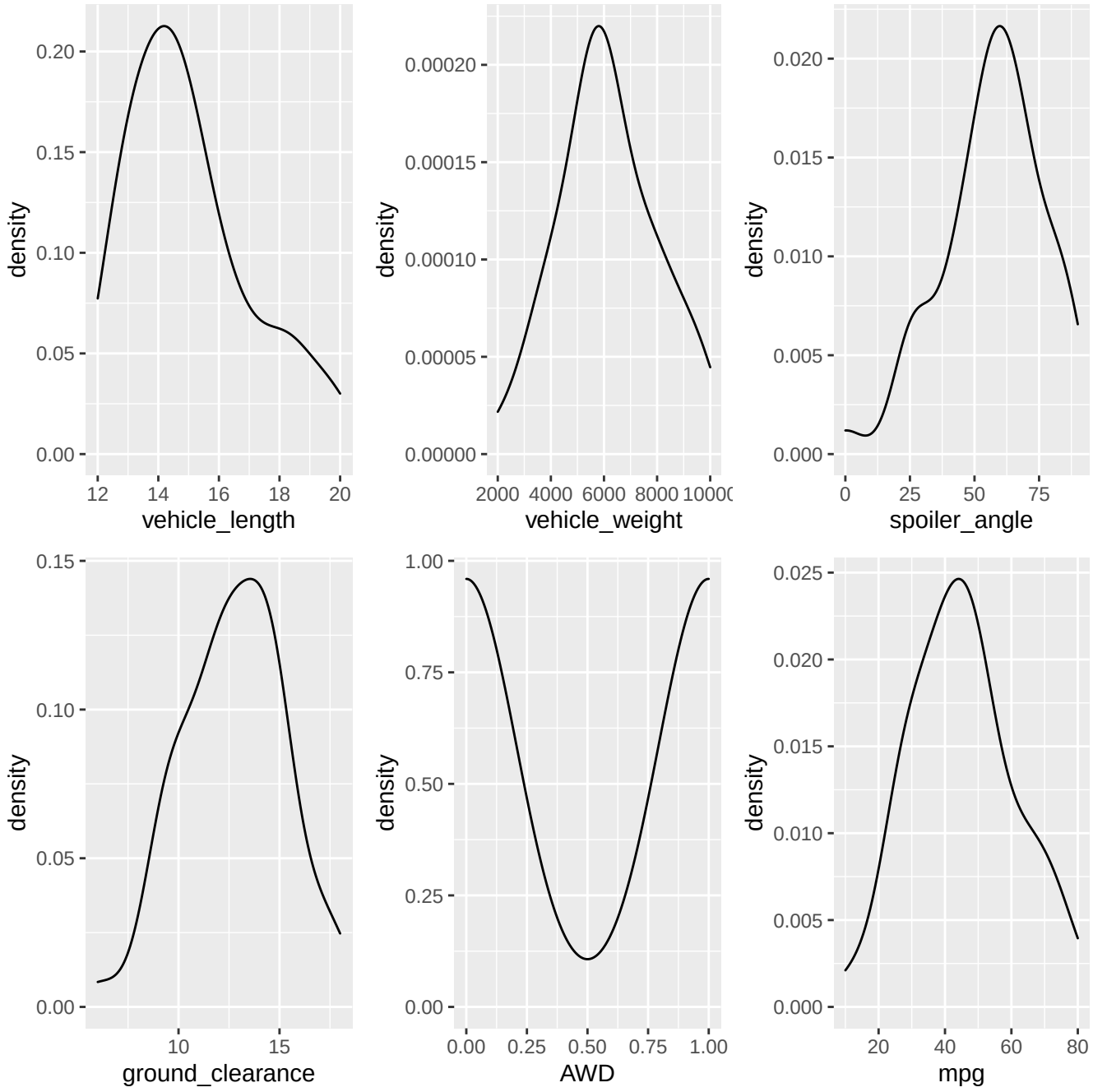
The descriptive statistics summary provides evidence that the variables are not normally distributed. Skewness values should be near 0 and kurtosis near 3. Only vehicle length and AWD meet the skewness criteria and non of the variables meet kurtosis.

```
# Test of Normality
# # plot the distribution using the geom_density() function
# put plots together
library(ggplot2)

##
## Attaching package:  'ggplot2'

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

library(gridExtra)
# create plots and name them
plot1<-ggplot(mecha,aes(x=vehicle_length)) + geom_density()
plot2<-ggplot(mecha,aes(x=vehicle_weight)) + geom_density()
plot3<-ggplot(mecha,aes(x=spoiler_angle)) + geom_density()
plot4<-ggplot(mecha,aes(x=ground_clearance)) + geom_density()
plot5<-ggplot(mecha,aes(x=AWD)) + geom_density()
plot6<-ggplot(mecha,aes(x=mpg)) + geom_density()
# arrange plots
grid.arrange(plot1, plot2, plot3, plot4,plot5, plot6,  nrow = 2, ncol = 3)
```



According to the plots the variables are not normal. AWD is not continuous.

Shapiro-Wilk test of Normality

```
vl<-shapiro.test(mecha$vehicle_length)
vw<-shapiro.test(mecha$vehicle_weight)
sa<-shapiro.test(mecha$spoiler_angle)
gc<-shapiro.test(mecha$ground_clearance)
awd<-shapiro.test(mecha$AWD)
mpg<-shapiro.test(mecha$mpg)

# create table with p-values in latex syntax
```

Table 1: Shaprio Test of Normality

Variable	P Value	Normality
vehicle length	0.01	No
vehicle weight	0.82	Yes
spoiler angle	0.28	Yes
ground clearance	0.84	Yes
AWD	0.00	No
mpg	0.79	Yes

The Shapiro-Wilk test of normality hypothesis states that the distribution is normal. Vehicle weight, spoiler angle, ground-clearance, and mpg provide evidence that they are distributed normally with p-vales greater than 0.05.

However, based on all of the evidence the variable are not normally distributed and a non-parametric model should be used.

Correlation Matrix

```
# using the used_cars data set
# select numeric columns and convert to matrix
mecha_matrix <- as.matrix(mecha[,c("vehicle_length","vehicle_weight","spoiler_angle", "grou
#used_matrix
cor(mecha_matrix)
```

```
##                  vehicle_length vehicle_weight
## vehicle_length       1.00000000    -0.12271790
## vehicle_weight      -0.12271790     1.00000000
## spoiler_angle        0.02577114    -0.11307851
## ground_clearance    -0.31663112     0.08511338
## AWD                  0.08565668    -0.03698098
## mpg                  0.60947984     0.09068314
##                  spoiler_angle ground_clearance         AWD
## vehicle_length      0.02577114      -0.31663112  0.08565668
## vehicle_weight     -0.11307851       0.08511338 -0.03698098
## spoiler_angle       1.00000000      -0.21112057 -0.09120266
## ground_clearance   -0.21112057       1.00000000 -0.15214456
## AWD                -0.09120266      -0.15214456  1.00000000
## mpg                -0.02083999       0.32874886 -0.14166977
##                           mpg
## vehicle_length     0.60947984
## vehicle_weight     0.09068314
## spoiler_angle     -0.02083999
## ground_clearance   0.32874886
## AWD               -0.14166977
## mpg                1.00000000
```

Vehicle length and ground clearance or the only variables that are somewhat correlated to mpg.

The estimates from the regression model may be suspect.

2.2 Non-Parametnric Multiple Regression

A Generalized additive model will be used in place of a multiple linear regression model as the GAM model is more flexible with a variety of variables.

```
library(mgcv)

## Loading required package:  nlme

## This is mgcv 1.8-36.  For overview type 'help("mgcv-package")'.

model.g = gam(mpg ~ vehicle_length + vehicle_weight + spoiler_angle + ground_clearance + AW
              data = mecha,
              family=gaussian())

summary(model.g)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## mpg ~ vehicle_length + vehicle_weight + spoiler_angle + ground_clearance +
##     AWD
##
## Parametric coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.040e+02  1.585e+01  -6.559 5.08e-08 ***
## vehicle_length    6.267e+00  6.553e-01   9.563 2.60e-12 ***
## vehicle_weight    1.245e-03  6.890e-04   1.807   0.0776 .
## spoiler_angle     6.877e-02  6.653e-02   1.034   0.3069
## ground_clearance  3.546e+00  5.412e-01   6.551 5.21e-08 ***
## AWD              -3.411e+00  2.535e+00  -1.346   0.1852
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.683   Deviance explained = 71.5%
## GCV = 87.489  Scale est. = 76.99     n = 50
```

THe p-value is 5.77 x $10^{-}6$, $and is much smaller than the 0.05 a priori significance level. There is significant evidence$
$values for this model vehicle weight, spoiler angle and AWD are not significant predictors in the variance of mpg. Veh$
$00 units. Holding all other variable constant for a one unit increase in mpg ground clearance increases 3.546e +$
$00 units. The intercept is statistically significant, it means that the intercept term explains a significant amount of var$
$squared is 0.68 meaning that the model explains 68\% of the variability in mpg.$

3 Deliverable 2

Deliverable 2: Summary Statistics on Suspension Coils

The MechaCar Suspension Coi data set contains the results from multiple production lots. In this data set, the weight capacities of multiple suspension coils were tested to determine if the manufacturing process is consistent across production lots. Analysis will include:

1. The suspension coil's PSI continuous variable across all manufacturing lots.
2. The following PSI metrics for each lot: mean, median, variance, and standard deviation.

```
head(susp)
```

```
##   VehicleID Manufacturing_Lot  PSI
## 1    V40858              Lot1 1499
## 2    V40607              Lot1 1500
## 3    V31443              Lot1 1500
## 4     V6004              Lot1 1500
## 5     V7000              Lot1 1501
## 6    V17344              Lot1 1501
```

```
library(tidyverse)
```

```
## -- Attaching packages ------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.2    v dplyr  1.0.7
## v tidyr   1.1.3    v stringr 1.4.0
## v readr   1.4.0    v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ---------------------- tidyverse_conflicts() --
## x ggplot2::%+%()    masks psych::%+%()
## x ggplot2::alpha()  masks psych::alpha()
## x dplyr::collapse() masks nlme::collapse()
## x dplyr::combine()  masks gridExtra::combine()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
```

```
# summarize PSI
total_summary<-summarise(susp, Mean=mean(PSI), Median=median(PSI), Variance=sd(PSI)^2, SD=s
total_summary
```

```
##      Mean Median Variance       SD
## 1 1498.78   1500 62.29356 7.892627
```

```
# summarize PSI by lot

lot_summary<- susp %>% group_by(Manufacturing_Lot) %>% summarise(Mean=mean(PSI), Median=med
lot_summary
```

```
## # A tibble: 3 x 5
## # Groups:   Manufacturing_Lot [3]
```

```
##   Manufacturing_Lot  Mean Median Variance     SD
##   <chr>             <dbl>  <dbl>    <dbl>  <dbl>
## 1 Lot1              1500   1500     0.980  0.990
## 2 Lot2              1500.  1500     7.47   2.73
## 3 Lot3              1496.  1498.  170.    13.0
```

The design specifications for the MechaCar suspension coils dictate that the variance of the suspension coils must not exceed 100 pounds per square inch. Lot 3 variance is 170 and is higher than the maximum allowable pounds per square inch. Lot 1 and 2 are with in acceptable ranges.

4 Deliverable 3

Deliverable 3: T-Test on Suspension Coils

```
# determine if the PSI across all manufacturing lots is statistically different from the
describe(susp$PSI)
```

```
##    vars   n    mean   sd median trimmed  mad  min  max
## X1    1 150 1498.78 7.89   1500 1499.66 1.48 1452 1542
##    range  skew kurtosis   se
## X1    90 -1.69    17.63 0.64
```

```
t.test(susp$PSI, mu=1500)
```

```
##
##  One Sample t-test
##
## data:  susp$PSI
## t = -1.8931, df = 149, p-value = 0.06028
## alternative hypothesis: true mean is not equal to 1500
## 95 percent confidence interval:
##  1497.507 1500.053
## sample estimates:
## mean of x
##   1498.78
```

```
# non-parametric t-test for comparison
wilcox.test(susp$PSI,1500)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  susp$PSI and 1500
## W = 68.5, p-value = 0.8894
## alternative hypothesis: true location shift is not equal to 0
```

The Variable PSI is not nomrally distributed and a non parametric t-test should be used.

The null hypothesis for this t-test is that the sample mean for PSI is statistically similar to the population mean. According the p-value, we do not have sufficient evidence to reject this hypothesis

because the p-value is greater than than the Type I error rate of 5%. There is 95% confidence that the population mean of PSI is between 1497.51 and 1500.05.

Complimentary to the t-test the wilcox non parametric equivilant to the t-test conclusion is the same. There is not have sufficient evidence to reject this hypothesis because the p-value is greater than than the Type I error rate of 5%.

```
# determine if the PSI for each manufacturing lot is statistically different from the pop
lo1<-susp[susp$Manufacturing_Lot== "Lot1",]
t.test(lo1$PSI, mu=1500)

##
##  One Sample t-test
##
## data:  lo1$PSI
## t = 0, df = 49, p-value = 1
## alternative hypothesis: true mean is not equal to 1500
## 95 percent confidence interval:
##  1499.719 1500.281
## sample estimates:
## mean of x
##      1500

lo2<-susp[susp$Manufacturing_Lot== "Lot2",]
t.test(lo2$PSI, mu=1500)

##
##  One Sample t-test
##
## data:  lo2$PSI
## t = 0.51745, df = 49, p-value = 0.6072
## alternative hypothesis: true mean is not equal to 1500
## 95 percent confidence interval:
##  1499.423 1500.977
## sample estimates:
## mean of x
##    1500.2

lo3<-susp[susp$Manufacturing_Lot== "Lot3",]
t.test(lo3$PSI, mu=1500)

##
##  One Sample t-test
##
## data:  lo3$PSI
## t = -2.0916, df = 49, p-value = 0.04168
## alternative hypothesis: true mean is not equal to 1500
## 95 percent confidence interval:
##  1492.431 1499.849
## sample estimates:
```

```
## mean of x
##   1496.14
```

```
# determine if PSI by lot is statistically different a kin to Anova and doing it this way
t.test(formula= PSI~Manufacturing_Lot, data= susp, subset= Manufacturing_Lot %in% c('Lot1',
```

```
##
##  Welch Two Sample t-test
##
## data:  PSI by Manufacturing_Lot
## t = -0.48653, df = 61.635, p-value = 0.6283
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0218165  0.6218165
## sample estimates:
## mean in group Lot1 mean in group Lot2
##             1500.0             1500.2
```

```
t.test(formula= PSI~Manufacturing_Lot, data= susp, subset= Manufacturing_Lot %in% c('Lot1',
```

```
##
##  Welch Two Sample t-test
##
## data:  PSI by Manufacturing_Lot
## t = 2.0856, df = 49.564, p-value = 0.04218
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1418273 7.5781727
## sample estimates:
## mean in group Lot1 mean in group Lot3
##            1500.00            1496.14
```

```
t.test(formula= PSI~Manufacturing_Lot, data= susp, subset= Manufacturing_Lot %in% c('Lot2',
```

```
##
##  Welch Two Sample t-test
##
## data:  PSI by Manufacturing_Lot
## t = 2.1533, df = 53.29, p-value = 0.03584
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.278647 7.841353
## sample estimates:
## mean in group Lot2 mean in group Lot3
##            1500.20            1496.14
```

The null hypothesis for this t-test is that the sample mean for PSI is statistically similar to the population mean. According the p-value for Lot1 and Lot2, we do not have sufficient evidence to reject this hypothesis because the p-value is greater than than the Type I error rate of 5%. There is 95% confidence that the population mean of PSI for Lot 1 is between 1499.719 and 1500.281.

Table 2: Shaprio Test of Normality

Manufacturing Lot Mean Difference	Mean Confidence Interval	P Value
Lot1	1500.00	1
0	1499.719, 1500.281	
Lot2	1500.20	0.61
0.2	1499.42, 1500.98	
Lot3	1496.14	0.04
-4.06	1492.43, 1499.85	

There is 95% confidence that the population mean of PSI for Lot 2 is between 1499.42 and 1500.98. Notice that 1500 is in the confidence interval for these two lots. There is 95% confidence that the population mean of PSI for Lot 3 is between 1492.43 and 1499.85.Notice that 1500 is not in the confidence interval for Lot 3.

5 Deliverable 4

Deliverable 4: Design a Study Comparing the MechaCar to the Competition

This study design will compare performance of the MechaCar vehicles against performance of vehicles from other manufacturers. A google search of how to compare cars turned up cars.com research comparison site website (https://www.cars.com/research/compare/). According to cars site the following perfromace features are used for comparison: MSRP, MPG, drivetrain, safety features. J.D. Power NADA guide compares two additional variable engine and transmission.

Based on the google search the variables that will be included in this study include: consumer ratings, car class, MSRP, MPG, drivetrain, engine, transmission, safety, and consumer ratings.

The first step will be to evaluate the data for data type and normality. The second step will be to create a correlation matrix to evaluate the relationship of the variables. Next Consumer ratings will be used as a dependent variable in a categorical multiple regression model to estimate how much variability is explained in consumer ratings by all of the other variables. Using the output of the regression variables that contribute the most to the variability in consumer ratings will be used in the ANOVA analysis to compare Mechacar vehicles by car class to other vehicles in that class. ANOVA will be used as there are multiple categorical variables being used compare Mechacar to other vehicles. In the ANOVA analysis the null hypothesis is that there is no difference between Mechacar and other vehicles of the same car class on mpg, msrp, drivetrain, engine, transmission, and safety.

6 References

https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

https://dplyr.tidyverse.org/

https://tidyr.tidyverse.org/reference/pivot$_l$*onger.html*

https://ggplot2.tidyverse.org/reference/ggsave.html

http://www.cookbook-r.com/Graphs/Axes$_{(}ggplot2)/$

https://stackoverflow.com/questions/1330989/rotating-and-spacing-axis-labels-in-ggplot2

https://ggplot2.tidyverse.org/reference/geom$_{b}oxplot.htmlaesthetics$

https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

http://www.statsoft.com/Textbook/Power-Analysis

https://www.statisticssolutions.com/statistical-power-analysis/

https://onezero.blog/combining-multiple-ggplot2-plots-for-scientific-publications/

https://rcompanion.org/handbook/F$_1$2.*html*