Readme

Statistical Analysis with R

Alberto Valdez

October 13, 2022

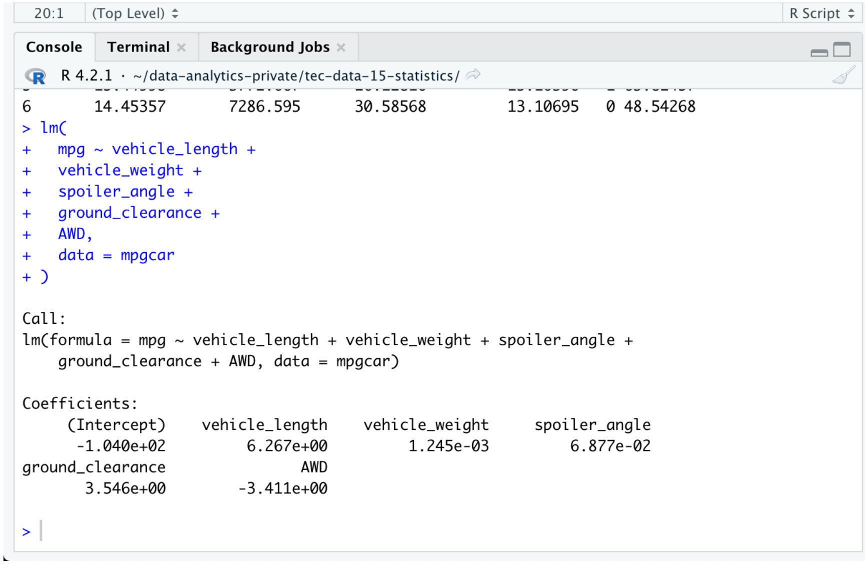
Contents

1 Statistical Analysis with R

In this analysis we will perform multiple tests for finding flaws in the production line of the MechaCar for the CarsRUs company. Then we will propose more tests for the future.

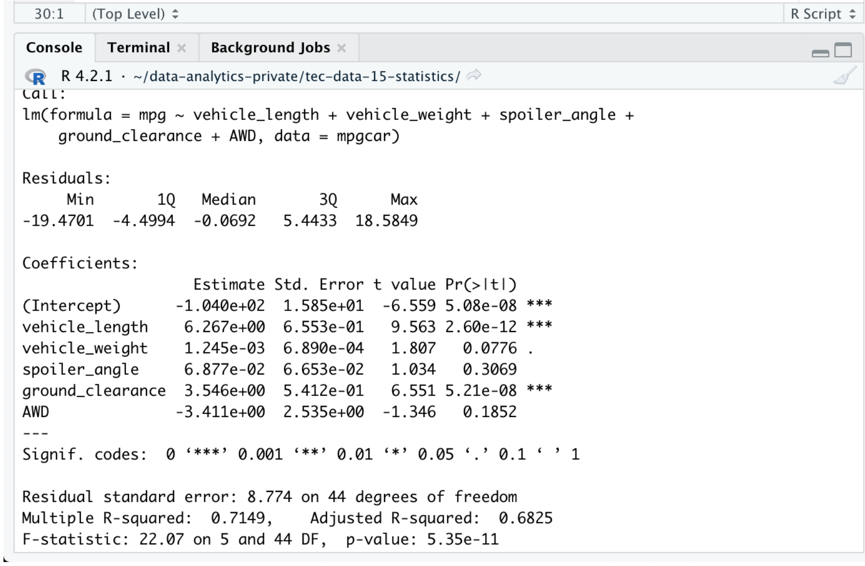
1.1 Linear Regression to Predict MPG

The first linear regression gives us the following results.



```
6       14.45357       7286.595      30.58568         13.10695    0 48.54268
> lm(
+   mpg ~ vehicle_length +
+   vehicle_weight +
+   spoiler_angle +
+   ground_clearance +
+   AWD,
+   data = mpgcar
+ )

Call:
lm(formula = mpg ~ vehicle_length + vehicle_weight + spoiler_angle +
    ground_clearance + AWD, data = mpgcar)

Coefficients:
     (Intercept)    vehicle_length    vehicle_weight     spoiler_angle
      -1.040e+02         6.267e+00         1.245e-03         6.877e-02
ground_clearance               AWD
       3.546e+00        -3.411e+00

>
```

When we call the summary we get our **p-value** and **multiple r-square**.



```
Call:
lm(formula = mpg ~ vehicle_length + vehicle_weight + spoiler_angle +
    ground_clearance + AWD, data = mpgcar)

Residuals:
     Min       1Q   Median       3Q      Max
-19.4701  -4.4994  -0.0692   5.4433  18.5849

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

Residual standard error: 8.774 on 44 degrees of freedom
Multiple R-squared:  0.7149,	Adjusted R-squared:  0.6825
F-statistic: 22.07 on 5 and 44 DF,  p-value: 5.35e-11
```

1. The coefficients that provide a non-random amount of variance are: vehicle_length and ground_clearance (as well as the Intercept) because their **Pr(>|t|)** value is very small.

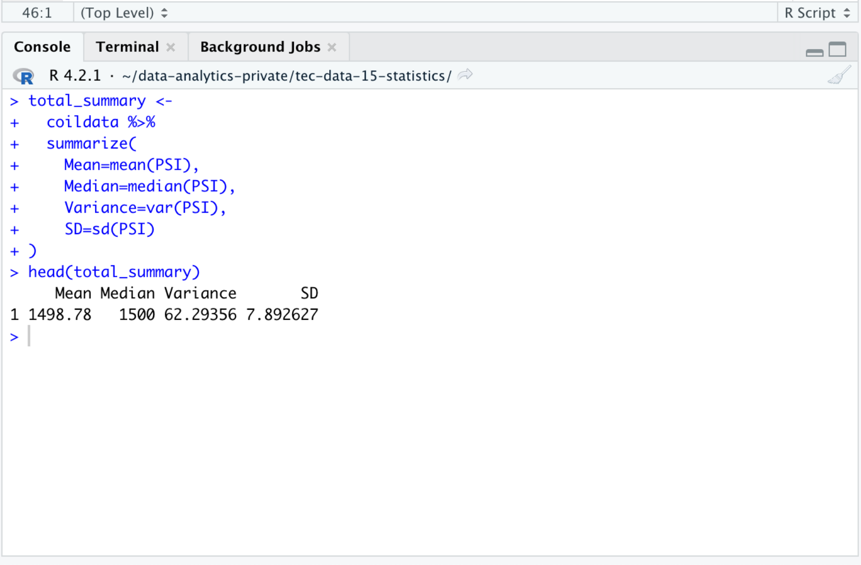
2. The slope is not considered zero because the **r-squared** value is 0.7149,

which means the relationship is linear. If we were to plot a line, it would be almost completely perpendicular to both axis.

3. This model is effective on showing us which variables have the most impact (greater correlation) on the **Milles per Gallon** and which do not affect it as much, so now we can make decisions based on those predictions.

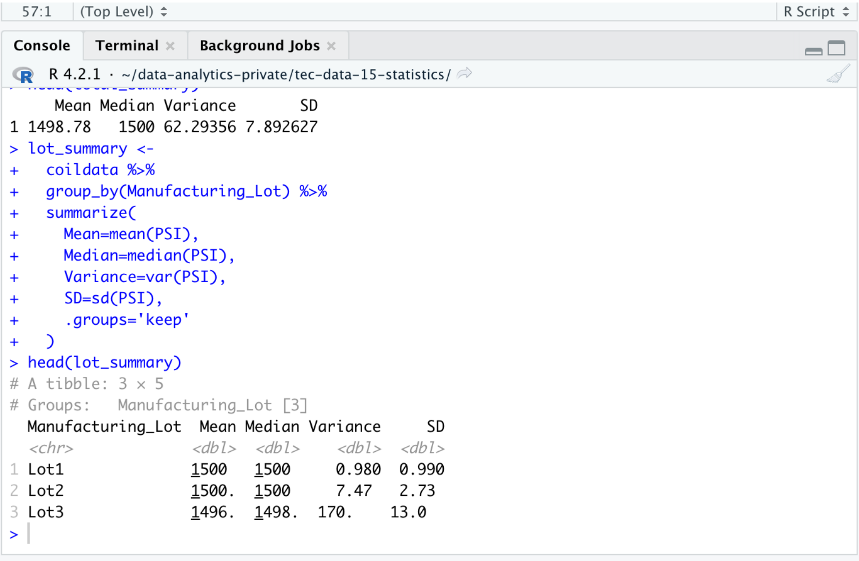
1.2 Summary Statistics on Suspension Coils

These are both summaries on the suspension coils.



```
> total_summary <-
+   coildata %>%
+   summarize(
+     Mean=mean(PSI),
+     Median=median(PSI),
+     Variance=var(PSI),
+     SD=sd(PSI)
+ )
> head(total_summary)
     Mean Median Variance       SD
1 1498.78   1500 62.29356 7.892627
> 
```

We can see that the total variance on the suspension coils is under 100 pounds per square inch, as per the design specifications. However, when we look at the lots **individually**, we can see that there is a problem with `Lot3`.



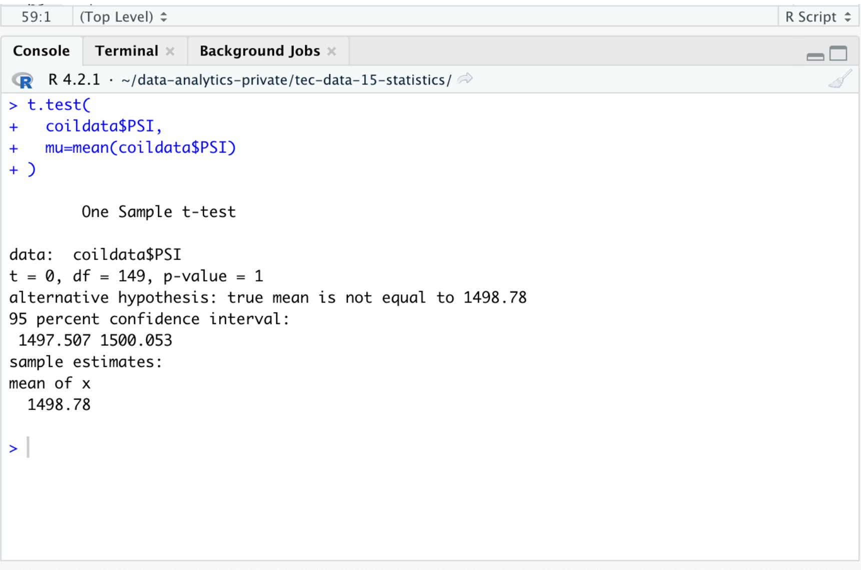
```
     Mean Median Variance       SD
1 1498.78   1500 62.29356 7.892627
> lot_summary <-
+   coildata %>%
+   group_by(Manufacturing_Lot) %>%
+   summarize(
+     Mean=mean(PSI),
+     Median=median(PSI),
+     Variance=var(PSI),
+     SD=sd(PSI),
+     .groups='keep'
+   )
> head(lot_summary)
# A tibble: 3 × 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
>
```

Lot3 doesn't comply with the variance specification, as it's value is 170, much higher than the required 100.

1.3 T-Tests on Suspension Coils

The following tests are made to assert that the mean of each sample is apoximates the population mean. We can verify the measure in our summary table that we have already calculated.

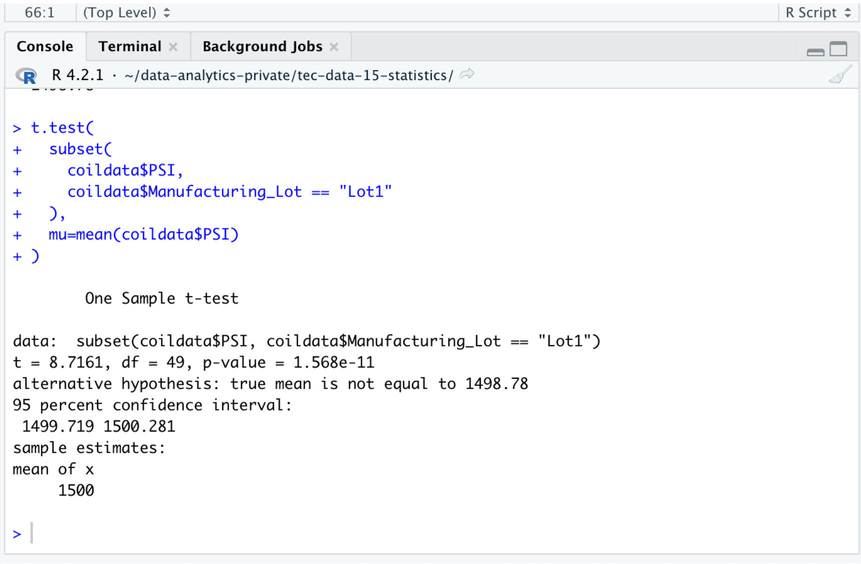
If we compare the mean of the population against itself, we will get a **p-value** of 1 because the ratio is perfect.



```
> t.test(
+   coildata$PSI,
+   mu=mean(coildata$PSI)
+ )

	One Sample t-test

data:  coildata$PSI
t = 0, df = 149, p-value = 1
alternative hypothesis: true mean is not equal to 1498.78
95 percent confidence interval:
 1497.507 1500.053
sample estimates:
mean of x
  1498.78

>
```

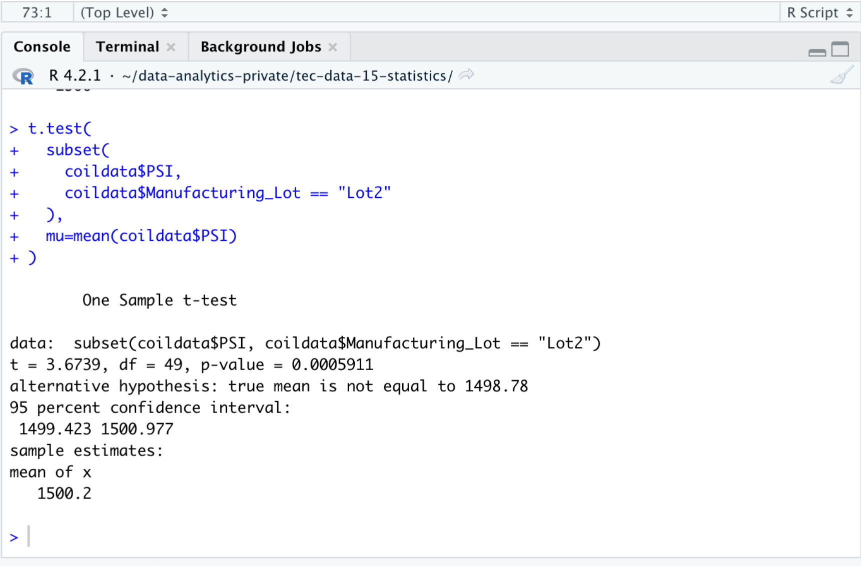
However, when doing **one-sample-t-test** for each of the lots, we get a more revealing picture.

Lot1 T-Test



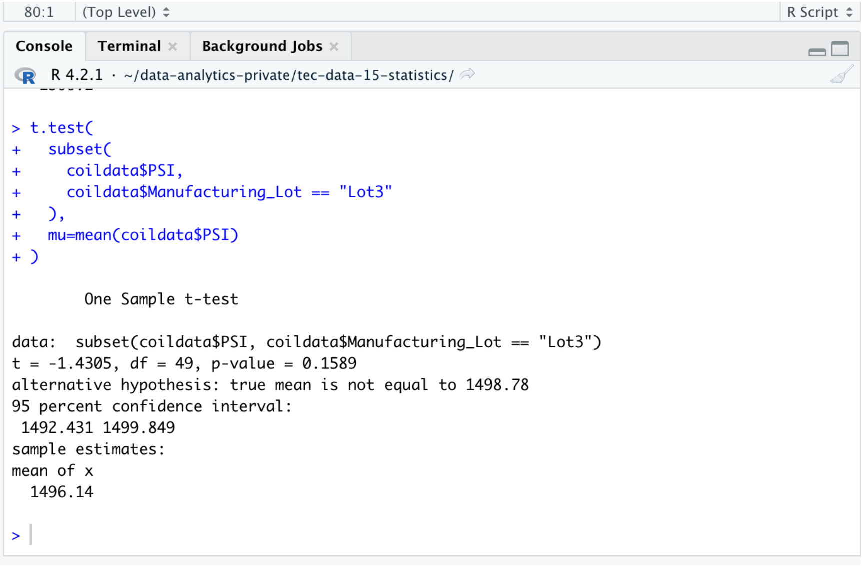
```
> t.test(
+   subset(
+     coildata$PSI,
+     coildata$Manufacturing_Lot == "Lot1"
+   ),
+   mu=mean(coildata$PSI)
+ )

	One Sample t-test

data:  subset(coildata$PSI, coildata$Manufacturing_Lot == "Lot1")
t = 8.7161, df = 49, p-value = 1.568e-11
alternative hypothesis: true mean is not equal to 1498.78
95 percent confidence interval:
 1499.719 1500.281
sample estimates:
mean of x
     1500

>
```

Lot2 T-Test



```
> t.test(
+   subset(
+     coildata$PSI,
+     coildata$Manufacturing_Lot == "Lot2"
+   ),
+   mu=mean(coildata$PSI)
+ )

        One Sample t-test

data:  subset(coildata$PSI, coildata$Manufacturing_Lot == "Lot2")
t = 3.6739, df = 49, p-value = 0.0005911
alternative hypothesis: true mean is not equal to 1498.78
95 percent confidence interval:
 1499.423 1500.977
sample estimates:
mean of x
   1500.2

>
```

Lot3 T-Test



```
> t.test(
+   subset(
+     coildata$PSI,
+     coildata$Manufacturing_Lot == "Lot3"
+   ),
+   mu=mean(coildata$PSI)
+ )

        One Sample t-test

data:  subset(coildata$PSI, coildata$Manufacturing_Lot == "Lot3")
t = -1.4305, df = 49, p-value = 0.1589
alternative hypothesis: true mean is not equal to 1498.78
95 percent confidence interval:
 1492.431 1499.849
sample estimates:
mean of x
  1496.14

>
```

We can see that the **p-value** on `Lot1` and `Lot2` are within the confidence interval, but the value for `Lot3` which is equal to `0.1589` is not, which means this sample deviates from the mean of the population more than we would like it to be.

If our alternative hypothesis is the following:

> If we measure the mean of the Lot3, it should not deviate from the mean of the population considerably.

Then we can't validate it, so the null hypothesis would still stand.

> The mean of the Lot3 will deviate from the population mean considerably when measured.

1.4 Study Design: MechaCar vs Competition

We want to test the MechaCar product at a larger scale so we should develop a set of tests to help us obtain insight and make predictions about the performance of the product in the market. This means that we have to compare against competitors and include on the consumer as a source for measurements.

We can start with a test to measure safety rating, which would consist of a controlled environment to perform tests on both the MechaCar and its competitors. The null hypothesis will be the following:

> Given a set of terrains and crash scenarios, the MechaCar won't protect the passengers of the vehicle better than its competitors.

And the alternative hypothesis will be the following:

> If we measure damages to test to dummy mannequins and interior of the MechaCar in different crash situations, it will perform better than its competitors in the majority of the tests.

The measurements we can perform are speed of the vehicles, which will be continuous data, then ordinal data can be severity of the damage in mannequins and interior of the car. We can also measure the PSI on the tires as numerical data.

Given that the number of tests can be scarce, we can do a multiple linear regression on the severity of the interior damage v.s. all the other variables.

For example, given a few numerical variables and a few ordinal variables, we can find a relationship of how much each numerical data affects it. Then we can compare the MechaCar data to other vehicles and look for differences on variance and how consistent each sample is against the population.

```
library(tidyverse)
```

```
# mock data
data <- tibble(severity = 1:5, psi = 31:35, speed = 60 + 20*log(severity))
```

severity	psi	speed
1	31	60
2	32	73.8629436111989
3	33	81.9722457733622
4	34	87.7258872223978
5	35	92.188758248682

The previous data is just a mock to demonstrate how we could perform the tests with R and then use linear regression to find correlation and variance.

```
summary(lm(severity ~ psi + speed, data=data))


Call:
lm(formula = severity ~ psi + speed, data = data)

Residuals:
         1          2          3          4          5 
-2.408e-16  8.565e-16 -6.017e-16 -4.030e-16  3.890e-16 

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)    
(Intercept) -3.000e+01  2.777e-14 -1.080e+15   <2e-16 ***
psi          1.000e+00  1.180e-15  8.473e+14   <2e-16 ***
speed        5.456e-16  1.468e-16  3.716e+00   0.0654 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.566e-16 on 2 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:      1 
F-statistic: 6.815e+30 on 2 and 2 DF,  p-value: < 2.2e-16

Warning message:
In summary.lm(lm(severity ~ psi + speed, data = data)) :
  essentially perfect fit: summary may be unreliable
```

The test will gain relevance if we measure more variables of the environment and the state of the cars as well as including the competition.

The null hypothesis will be negated if the MechaCar performs better and there is less interior damage than its competitors in a majority of the scenarios.