

MechaCar_Statistical_Analysis

Deliverable 1

Purpose:

- ✓ To identify which variables in the dataset predict the mpg of MechaCar prototypes using multiple linear regression model.

Materials:

Dataset: MechaCar_mpg.csv

Software: RStudio and R-programing

Model: Multiple linear regression

Assumptions

1. The relationship between X and the mean of Y is linear
2. The variance of residual is the same for any value of X.
3. The observations are independent of each other.
4. For any fixed value of X, Y is normally distributed.

Hypothesis

Null hypothesis: all variables do not have any impact on MPG.

Alternate hypothesis: all or some of the variables have impact on MPG.

Results

Multiple Linear Regression Model to Predict MPG

```
R 4.1.2 ~ /UC Bootcamp/Projects/MechaCar_Statistical_Analysis/R-analysis/
> lm(mpg ~ vehicle_length + vehicle_weight + spoiler_angle + ground_clearance + AWD, data = mech_car)

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

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

> summary(lm(mpg ~ vehicle_length + vehicle_weight + spoiler_angle + ground_clearance + AWD, data = mech_car))

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

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
```

Model p-value = 5.35e-11 which is < 0.05. And at the same time the p-value for Vehicle length and Ground clearance is also less than 0.05.

Conclusion

Based on the results, vehicle length and ground clearance are statistically ($p < 0.05$) different or have a significant impact on mpg. Therefore, the two variables can predict MPG.

Which variables/coefficients provided a non-random amount of variance to the mpg values in the dataset?

According to the analysis, vehicle length and ground clearance have provided a non-random amount of variance to the mpg values in the dataset. Because vehicle length and ground clearance have a significant impact on MPG since their p-values are less than 0.05 (significance level). But, Vehicle weight, spoiler angle and AWD have provided a random amount of variance.

Is the slope of the linear model considered to be zero? Why or why not?

The slope of the linear model is not considered to be zero, because the overall linear model shows that the p-value ($5.35e-11$) is less than 0.05. This shows there is relationship between x and y axis. Additionally, the R-squared = 0.7149; i.e. 71.5% of the variability of our mpg variable is explained using this linear model.

Does this linear model predict mpg of MechaCar Car prototypes effectively? Why or why not?

The R-squared of the model indicated that 71.5% of the variability of our mpg is explained by this model. But for better prediction, further modelling analysis is necessary to access other factors or adding data for maximizing the R-squared value.

Deliverable 2

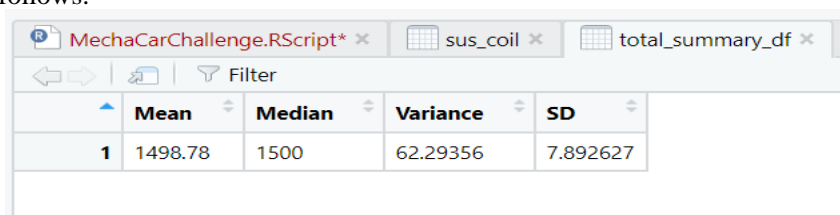
Purpose

- To get the summary statistics of Suspension Coils.

Results

Summary Statistics on Suspension Coils

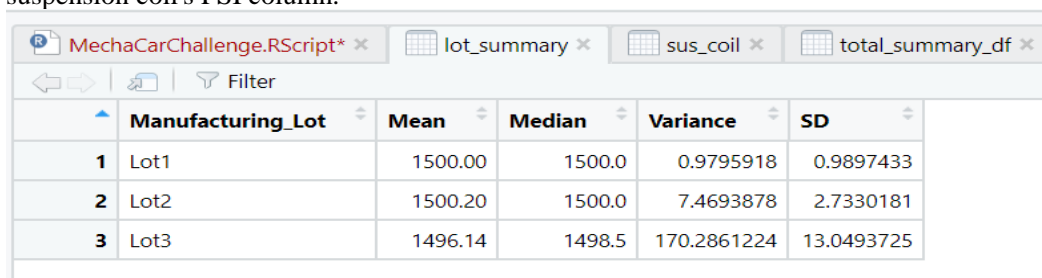
The total summary; mean, median, variance and standard deviation of the suspension coils (PSI) are as follows: -



The screenshot shows the RStudio interface with three tabs: 'MechaCarChallenge.RScript*', 'sus_coil', and 'total_summary_df'. The 'total_summary_df' tab is active, displaying a table with summary statistics for the 'sus_coil' data. The table has columns for 'Mean', 'Median', 'Variance', and 'SD'. There is one row of data.

	Mean	Median	Variance	SD
1	1498.78	1500	62.29356	7.892627

Lot summary to group each manufacturing lot by the mean, median, variance, and standard deviation of the suspension coil's PSI column.



The screenshot shows the RStudio interface with four tabs: 'MechaCarChallenge.RScript*', 'lot_summary', 'sus_coil', and 'total_summary_df'. The 'lot_summary' tab is active, displaying a table with summary statistics for the 'lot_summary' data. The table has columns for 'Manufacturing_Lot', 'Mean', 'Median', 'Variance', and 'SD'. There are three rows of data.

	Manufacturing_Lot	Mean	Median	Variance	SD
1	Lot1	1500.00	1500.0	0.9795918	0.9897433
2	Lot2	1500.20	1500.0	7.4693878	2.7330181
3	Lot3	1496.14	1498.5	170.2861224	13.0493725

Based on the total summary, the suspension coil for all manufacturing lots is 62 pounds per square inch. Hence the current manufacturing data meet design specification. At the individual level, Lot 1 and Lot 2 meet the design specification at 0.98 and 7.47 pounds per square inch, respectively. Whereas lot 3 (170.3 PSI) exceeding the design specification and fail to meet the specification.

Deliverable 3

T-Test on Suspension Coils

Hypothesis

Null hypothesis: there is no statistical difference between the manufacturing lots mean and population mean of PSI.

Alternate hypothesis: there is a statistical difference between the manufacturing lots mean and population mean of PSI.

The t-test below compares all manufacturing lots against mean PSI of the population

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

```
One Sample t-test
```

```
data: sus_coil$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
```

The three t-tests compare each manufacturing lot against mean PSI of the population as follows:-

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

One Sample t-test

data: sus_coil$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

>
> ## t.test for each lot
>
> t.test(subset(sus_coil, Manufacturing_Lot=="Lot1")$PSI, mu=1500)

One Sample t-test

data: subset(sus_coil, Manufacturing_Lot == "Lot1")$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

> t.test(subset(sus_coil, Manufacturing_Lot=="Lot2")$PSI, mu=1500)

One Sample t-test

data: subset(sus_coil, Manufacturing_Lot == "Lot2")$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

> t.test(subset(sus_coil, Manufacturing_Lot=="Lot3")$PSI, mu=1500)

One Sample t-test

data: subset(sus_coil, Manufacturing_Lot == "Lot3")$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
```

Results and conclusion

According to the first t-test analysis, the p-value (0.06) is greater than the significance level, therefore, there is no statistical difference between the manufacturing lots and population mean of PSI. However, the t-test analysis for individual lots shows that the p-value for Lot 1 and 2 is greater than significance level, whereas the p-value for Lot 3 is less than (0.04) the significance level. Therefore, there is statistical difference between the mean of Lot 3 and population mean of PSI.

Deliverable 4

Study Design: MechaCar Vs Competition

Mostly consumers would like to consider cost, city or highway fuel efficiency and horse power to purchase a car.

Metric to test

Cost of car is a dependent variable.

To find the variable that mainly affects the cost of a car.

Hypothesis

Null hypothesis: all variables from city or highway fuel efficiency and horse power does not have impact on the cost of car.

Alternate hypothesis: all or some of the variable have impact on the cost of car.

Statistical test

The best method to test the hypothesis is multiple linear regression model. Because we have 4 variables and the dependent variable, and all independent variables have continuous data types. Additionally, this model can easily access the impact of fuel efficient and horse power on the cost of car.

Data

We need the following data to run the model; fuel efficiency both from city and highway, horse power and the cost of a car. Regarding sample size, as a rule of thumb we need 10 observations per variable, with approximately 50 total observations.