UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143 SEC 07

PROBABILITY & STATISTICAL DATA ANALYSIS

Project 2: Study of Cars Specification from Different Brands

Presentation video: https://youtu.be/lggrx2ul-uo

Lecturer: Dr Nor Azizah Ali

Group member :

Name	Matric Number
Yam Yuan Zhan	A21EC0146
Darren Leong Kah Xiang	A21EC0021
Lau Wen Zhen	A21EC0193
Brian Ting Chin Ing	A21EC0165

1.0 Introduction and background

Cars are quite significant in our everyday lives since we drive them to every location and they are more handy than other modes of transportation. However, fuel type, max rpm, engine size, and horsepower are all crucial. As a result, the primary goal of this study is to investigate and conduct research on the key data of a car, such as its engine size, horsepower peak rpm, and others, as well as to apply statistical analysis skills in the dataset to determine whether there is a relationship between the data using correlation and regression. During the inquiry and research phase, a few prospective variables are chosen and a series of test studies are performed.

The dataset regarding the potential variables is a secondary data source retrieved from the Kaggle website, this data was collected by Eleanor Xu who works as an analyst from Coursera, San Francisco, California, United States. This dataset is about the data of 205 sample cars and also related to their car companies, fuel types, car aspiration, number of doors, engine location, length, width, height, engine type, horsepower, price, and other specifications.

2.0 Dataset description

Source of dataset: https://www.kaggle.com/jingbinxu/sample-of-car-data

Population: Cars from different company

Sample: 205 Cars

Statistical Test Analysis for Data Selected

1) **Fuel Type and Horsepower**

- Objective: To test whether the mean of horsepower for cars that use gas fuel are greater than cars that use diesel.
- Test Analysis and Expected Outcome
  - Test Analysis: 2 Sample Hypothesis Testing
  - Expected Outcomes: Mean horsepower for car use gas fuel is larger than car use diesel.

2) **Wheel base and Curb Weight**

- Objective: To test the existence of a linear relationship between the size of wheelbase and curb weight.
- Test Analysis and Expected Outcomes
  - Test Analysis: Correlation
  - Expected Outcomes: There exists a strong relationship between the size of the wheelbase and the curb weight.

3) **Engine size and Peak RPM**

- Objective: To test whether the peak RPM of a car depends on the engine size used.
- Test Analysis and Expected Outcomes
    - Test Analysis: Regression
    - Expected Outcomes: The value of peak RPM depends on the engine size. The larger the engine size, the lower the peak RPM.

4) **Fuel System**

- Objective: To test whether there is difference between the observed frequency and the expected frequency of the fuel system used by cars.
- Test analysis and Expected outcomes
    - Test analysis: Goodness of Fit-Test
    - Expected Outcome: There is no difference between the observed frequency and the expected frequency, and it is a good fit to the assumed distribution.

5) **Drive Wheels and Number of Doors**

- Objectives: To test whether the variables, drive wheels and number of doors are related.
- Test analysis and Expected outcomes
    - Test analysis: Chi-Square Test of Independence
    - Expected Outcome: The drive wheels and number of doors are not related.

3.0 Data Analysis

3.1 2 Sample Hypothesis Test

In this 2-sample hypothesis test, we are using **horsepower** and **fuel type** to conduct the test and we will test whether the mean horsepower of a gas car is larger than the mean horsepower of a diesel car at the confidence level of 95% and we assume variance is unequal. From the data, count(n), mean($\overline{x}$), standard deviation(s) are calculated.

$n_1 = 185$	$n_2 = 20$
$\overline{X}_1 = 106.5676$	$\overline{X}_2 = 84.45$
$S_1 = 40.2073$	$S_2 = 25.95842$

**Note: group 1 is for gas cars, and group 2 is for diesel cars.

Hypothesis statement

$H0$: $\mu 1 = \mu 2$ $H1$: $\mu 1 > \mu 2$

where $\mu 1$ equals the mean horsepower of gas cars, and $\mu 2$ equals the mean horsepower of diesel cars.

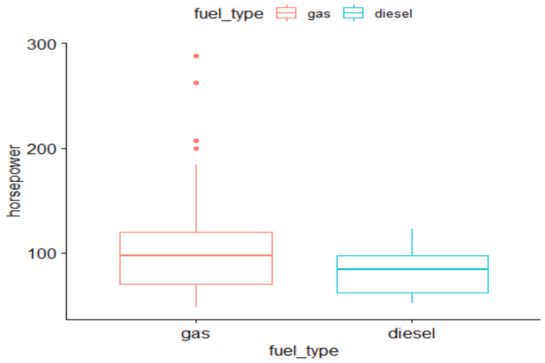
Given 95% confidence level, α = 0.05. The test statistics, t_0 from RStudio, test statistics, t_0 is 3.3955

Calculate the degrees of freedom:

By using RStudio, the degree of freedom, v is 29.9261Therefore, using α =0.05, we reject H_0 if $t_0 > t_{0.023,\ 29.9261} = 1.699$.

Conclusion

Since, $t_0 = 3.3955 > t_{0.023,\ 29.9261} = 1.699$ we reject the null hypothesis. There is sufficient evidence to conclude that the mean horsepower of gas cars is larger than the mean horsepower of diesel cars at the confidence level of 95%.

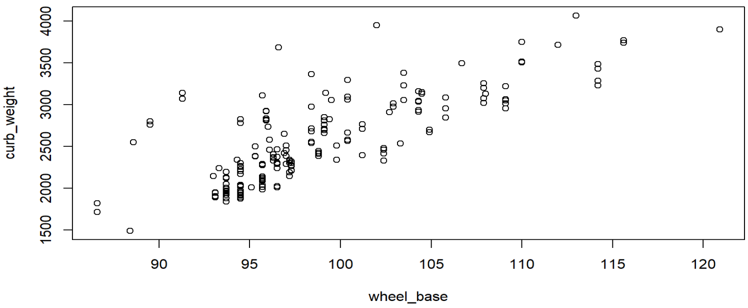


3.2 Correlation Analysis

In this analysis, we use the variables **wheelbase** and **curb weight** for this analysis. We wanted to measure the strength of the linear relationship between the variable wheelbase and curb weight. Since the measurement level of the two variables are ratios, we are using Pearson's Product-Moment Correlation Coefficient. The confidence level that we are using is 95% confidence level.

Calculate the sample correlation coefficient, r:

From the RStudio, the sample correlation coefficient is r = 0.7763863 which shows that the strength of the linear relationship is moderate because r is in the range of 0.5 to 0.8.



Based on the scatter plot graph that was visualised using RStudio, we can see that there is a positive linear correlation relationship between the wheel base and curb weight.

Hypothesis:
H0 : ρ = 0 (No linear correlation)
H1 : ρ ≠ 0 (linear correlation exists)

Test statistic:
From the RStudio, the test statistic, t = 17.55125

Critical value:
α = 0.05, $d.f$ = 205 − 2 = 203
Using the t Distribution table (two-tailed), the critical value is
$t_{\alpha/2=0.025,\ df=203}$ = ±1.960
So, if the test statistics > 1.960 or test statistics < -1.960, we reject H0.

Conclusion:
Since the test statistics t = 17.55125 > critical value $t+_{0.025,\ 203}$ =1.960, we reject the null hypothesis. There is sufficient evidence to conclude that there is a linear relationship between wheel base and curb weight at the confidence level of 95%.

3.3 Regression Analysis

In this analysis, we are using variables **engine size** and **peak rpm**, where we will test whether the value of peak rpm depends on the value of engine size, using engine size as the independent variable(x) and peak rpm as the dependent variable(y). Our regression model is a linear model. We assume that the changes in the values of peak rpm are affected by the changes in the values of engine size.

1. Estimated Regression Model
-b0 is the estimated average value of y(peak rpm)
-b1 is the estimated change in the average value of y(peak rpm) due to a one-unit change in x(engine size).

- From the RStudio w found that the value of b0 is 5480.152 and the value of b1 is -2.805273

$$\hat{yi}=5480.156-2.805273x$$

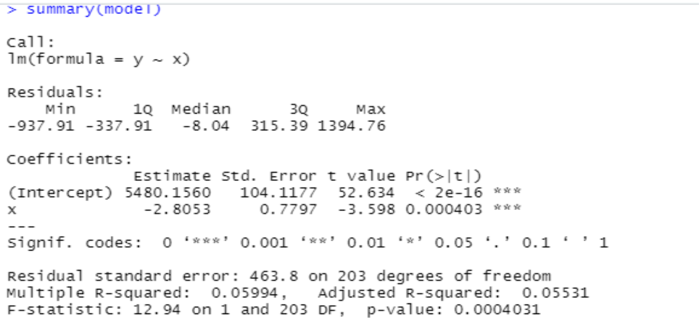
- From the equation, we can interpret and know that the b0= 5480.156 indicates that, for cars within the range of engine sizes observed, 5480.156 is the portion of peak rpm not explained by engine size. Whereas b1 =-2.805273 tells us that the average value of car peak rpm increases by -2.805273 on average.
- Find the SST, SSE and SSR:
  From RStudio, we found that SST is 46443976, SSE is 43659047, and SSR is 2783928.
- Coefficient of Determination, R2 = 0.06
- By using RStudio, we get Standard Error of Estimate, sƐ = 463.7555, and the standard deviation of regression is 0.7797138.

2. Inference about the Slope: t-Test

- Hypothesis Statement:
  H0: β1 = 0 (no linear relationship)
  H1: β1 ≠ 0 (linear relationship does exist)
- Find critical value, using α = 0.05, df = n-2 = 203
  since this is a two-tailed test, there are two critical values:
  Lower tail critical value -tα/2=0.025, df=203 = -1.960
  Upper tail critical value tα/2=0.025, df=203 = 1.960
  From RStudio, we also get p-value = 2.2e - 16.
  Hence, we reject H0 if test statistics > 1.960 / test statistics < -1.960.
  By using RStudio, we get the test statistic t = -3.597824.

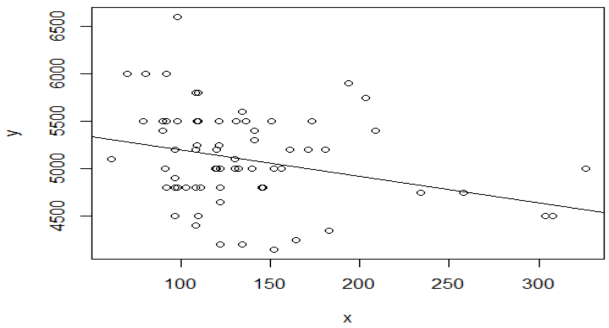
State the decision:
Since test statistics t = -3.597824 < lower tail critical value tα/2=0.025, df=203 = -1.960, we reject the null hypothesis. There is sufficient evidence that engine size affects peak rpm, at α = 0.05.

To perform linear regression in RStudio:



```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-937.91 -337.91   -8.04  315.39 1394.76 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5480.1560   104.1177  52.634  < 2e-16 ***
x             -2.8053     0.7797  -3.598 0.000403 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 463.8 on 203 degrees of freedom
Multiple R-squared:  0.05994,	Adjusted R-squared:  0.05531 
F-statistic: 12.94 on 1 and 203 DF,  p-value: 0.0004031
```

Based on our linear regression model, we can get b0 = 5480.156, b1 = -2.805273, $sb1$=0.7797138 , sƐ = 463.7555, df = 203, Coefficient of Determination, R2 = 0.06 and p-value = 2.2e – 16.
We use plot() function to plot graph and abline() to add linear model
y-peak rpm
X-engine size



3.4 Goodness of Fit Test

Variable used: **fuel system** (mpfi (Multi-point fuel injection), 2bbl (2 barrel), 1bbl (1 barrel), spfi (Sequential Port Fuel Injection), 4bbl (4 barrel), idi (Indirect fuel injection diesel engines))
Objective: Test whether there is a difference between the observed frequency and expected frequency of the fuel system of different cars at a confidence level of 95%.

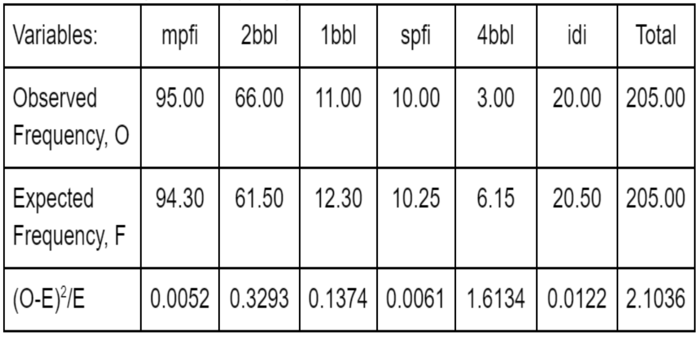
Our claims: $P_{mpfi} = 0.46, P_{2bbl} = 0.30, P_{1bbl} = 0.06, P_{spfi} = 0.05, P_{4bbl} = 0.03, P_{idi} = 0.10$

Statement of test hypothesis:

H0: P_{mpfi} = 0.46, P_{2bbl} = 0.30, P_{1bbl} = 0.06, P_{spfi} = 0.05, P_{4bbl} = 0.03, P_{idi} = 0.10

H1: There exists at least one of the proportions that are different from the claimed values.

Calculate the expected frequency:



Variables:	mpfi	2bbl	1bbl	spfi	4bbl	idi	Total
Observed Frequency, O	95.00	66.00	11.00	10.00	3.00	20.00	205.00
Expected Frequency, F	94.30	61.50	12.30	10.25	6.15	20.50	205.00
$(O-E)^2/E$	0.0052	0.3293	0.1374	0.0061	1.6134	0.0122	2.1036

```
1bbl 2bbl 4bbl  idi mpfi spfi 
  11   66    3   20   95   10 
```

Observed frequency from R studio.

From R Studio, we get the test statistic, X^2 = 2.1036 and critical value, $X^2_{5, 0.05}$ = 11.0705.

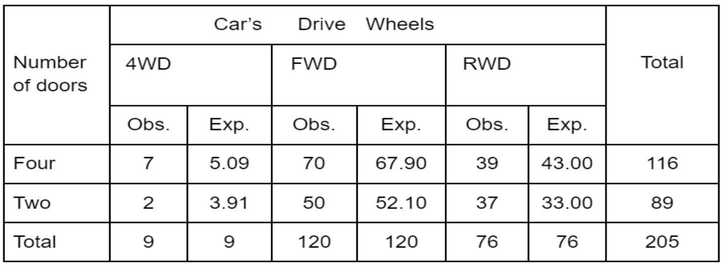
Conclusion: Since the X^2 value = 2.1036 is smaller than the critical values $X^2_{5, 0.05}$ = 11.0705, H_0 was not rejected. There is not sufficient evidence to reject the claim that the proportion is distributed with the proportion claimed.

3.5 Chi-Square Test of Independence

In the Chi-Square Test of Independence, we are using the car's **drive wheels** and the **number of doors** as the variables to conduct the test and we will test whether the car's drive wheels have a relationship with the number of doors using Two Way Contingency Table, at 95% confidence level.

State the hypothesis statement:
H_0: No relationship between variables.
H_1: Variables have a relationship.

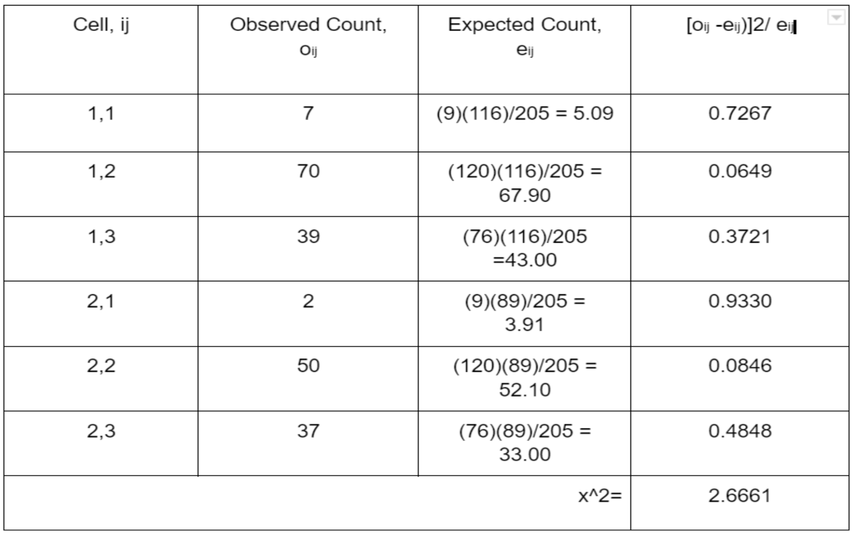
Find the critical value (refer to chi-square table):



Number of doors	Car's Drive Wheels						Total
	4WD		FWD		RWD		
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	
Four	7	5.09	70	67.90	39	43.00	116
Two	2	3.91	50	52.10	37	33.00	89
Total	9	9	120	120	76	76	205

α = 0.05
df = (3 - 1)(2 - 1) = 2
From RStudio, critical value of X^2 = 5.991

Calculate the test statistic value:
When we calculate test statistics manually, we get test statistic X^2 = 2.6661. Using RStudio, we get test statistic X^2 = 2.6537, with p-value = 0.2653



Cell, ij	Observed Count, o_{ij}	Expected Count, e_{ij}	[o_{ij} -e_{ij})]2/ e_{ij}
1,1	7	(9)(116)/205 = 5.09	0.7267
1,2	70	(120)(116)/205 = 67.90	0.0649
1,3	39	(76)(116)/205 =43.00	0.3721
2,1	2	(9)(89)/205 = 3.91	0.9330
2,2	50	(120)(89)/205 = 52.10	0.0846
2,3	37	(76)(89)/205 = 33.00	0.4848
		x^2=	2.6661

State the decision:
Since the test statistic value (X^2 = 2.6537) < critical value($X^2_{k=2,\ \alpha = 0.05}$ = 5.991), we fail to reject H_0. There is sufficient evidence to conclude that there is no relationship between the variable number of doors and drive wheels, at α = 0.05.

4.0 Conclusion

When choosing the dataset, we tried to choose the most complete dataset from the website called Kaggle. The dataset that we chose consists of 205 samples of cars. We used 5 types of statistical test analysis, which consist of 2 Sample Hypothesis Testing, Correlation Analysis, Regression Analysis, Goodness of Fit Test, and Chi-Square Test of Independence.

From the 2 Sample Hypothesis Testing, we found that gasoline burns faster than diesel, which enables it to create more horsepower. It matched our hypothesis testing.

From the Correlation Analysis, the strength of the linear relationship between wheelbase and curb weight is moderate. It means that the larger the wheelbase of a car, the more likely that the weight of the car will increase.

From the Regression Analysis. The scatter plot graph shows that there is a negative linear relationship between engine size and peak rpm, which means larger engine size causes lower peak RPM.

From the Goodness of Fit Test, we have learned that most cars use mpfi (Multi-point fuel injection) as it improves the fuel efficiency of a vehicle.

From the Chi-Square Test of Independence, we learned that the number of doors and drive wheels are completely not related because they are very different parts of a vehicle.

In conclusion, we learned a lot of things from this project such as using RStudio to perform various tests such as 2 sample hypothesis testing, correlation analysis, regression analysis, goodness of fit test, and chi-square test of independence.

5.0 Appendix

https://www.kaggle.com/jingbinxu/sample-of-car-data