



12/10/2014

IDS 570 Statistics for Management

Project Report

Submitted By -

Adwaith Gangireddy Varun

agangi2@uic.edu

Mounica Sirineni

msirin2@uic.edu

Nishanth Reddy Konkala

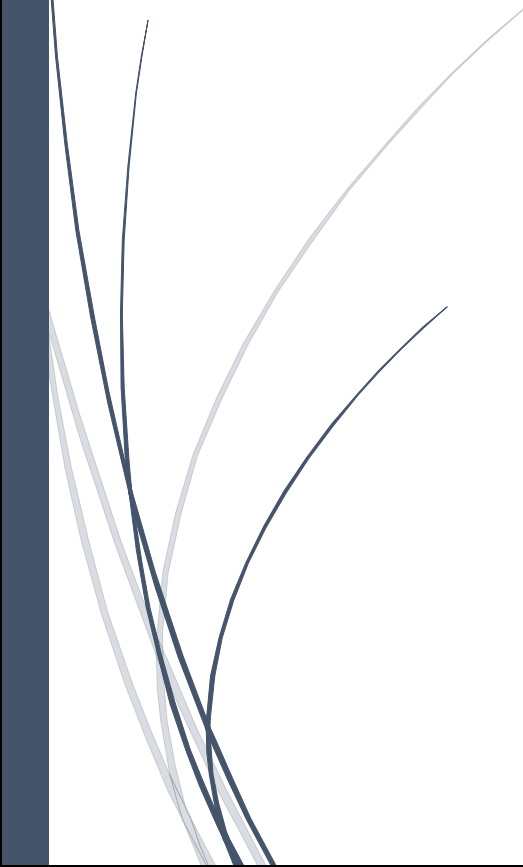
nkonka2@uic.edu

Sanjeeta Behera

sbeher2@uic.edu

Sreekanth Reddy

schint9@uic.edu



Contents

Purpose	2
Data Dictionary	2
Basics	3
Quantiles	3
Box plots comparing the different types of cars	3
Histogram	4
Means Plot	5
Stem plot of Selling Price	6
Quantile Plots	7
t-test Analysis	7
ANOVA test	8
Multiple Linear Regression	9
Residual Plots	10
Adjusted Multiple Linear Regression	11
Prediction	11
Appendix	12

Purpose

To compare if the selling price of the car is based on the type and prediction is made on the basis of the type of the car and its manufacturer, desired miles per gallon for the customer. The different types of the cars are – compact, large, midsize, small and sporty.

Data Dictionary

The dataset Car_Sales_Project_Final is about the different types of cars manufactured by different manufacturers and the attributes of the cars.

It consists of 78 rows and 15 columns

A - Manufacturer

B - Model

C – Type (Compact, Midsize, Large, Small, Sporty)

D – Selling Price of the car

E – Miles per gallon

F – Airbags Standard (standard 0 = none, 1 = driver only, 2 = driver & passenger)

G - Drive train type (0 = rear wheel drive 1 = front wheel drive 2 = all-wheel drive)

H – No of Cylinders

I – Engine Size (liters)

J – Horse Power

K – RPM (Revolutions per minute at maximum horsepower)

L – Engine revolutions (per mile)

M – Manual transmission available (0 = No, 1 = Yes)

N – Fuel tank capacity (gallons)

O – Domestic (0 = non-U.S. manufacturer 1 = U.S. manufacturer)

Basics

Mean of Selling Price = 18.03

Median of Selling Price= 15.90

Minimum Selling Price= 7.40

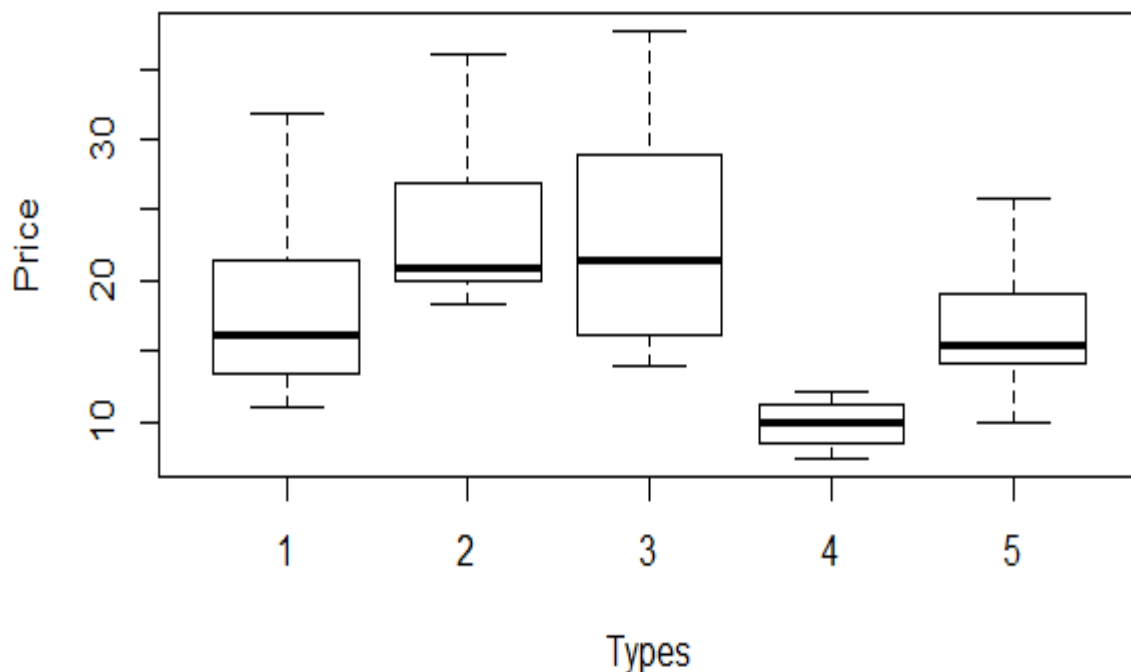
Maximum Selling Price= 37.70

Quantiles

0%	25%	50%	75%	100%
7.40	11.45	15.90	22.40	37.70

Boxplots comparing the different types of car

Boxplot comparing different Types



The box plot here is used to compare the selling price of the different types of the cars. It shows that the Selling Price's median of Car type 3 (Midsize) is higher than the others. On seeing the plot diagram, one can notice that there are no outliers in the data.

Mean of Selling Price according to Car type

1	2	3	4	5
18.21250	24.30000	23.62632	9.88000	16.75000

Standard deviation of Selling Price according to Car type

1	2	3	4	5
6.686890	6.337507	7.878652	1.483098	4.521766

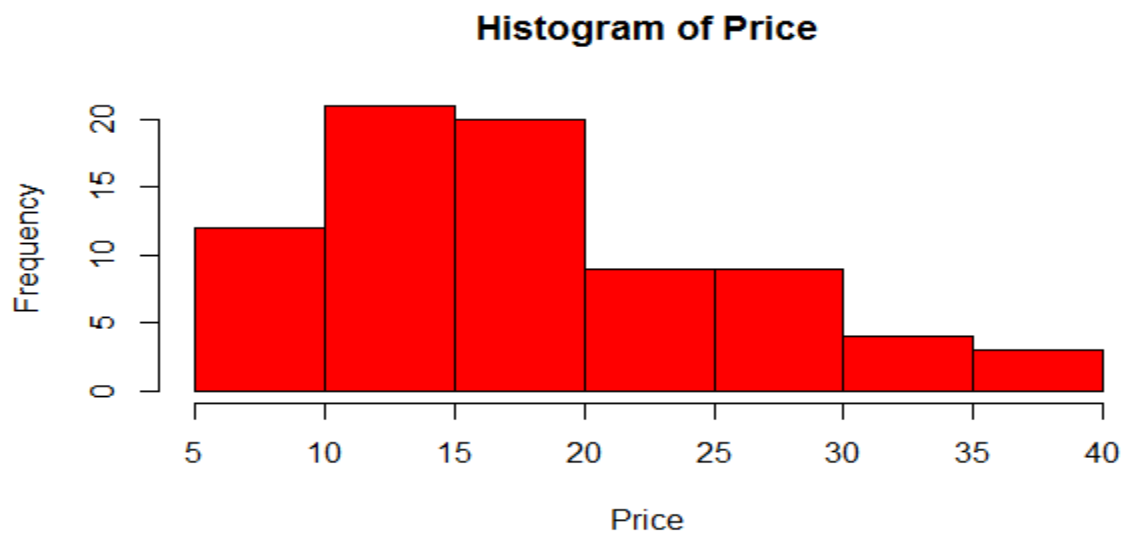
Length of Selling Price according to Car type

1	2	3	4	5
16	11	19	20	12

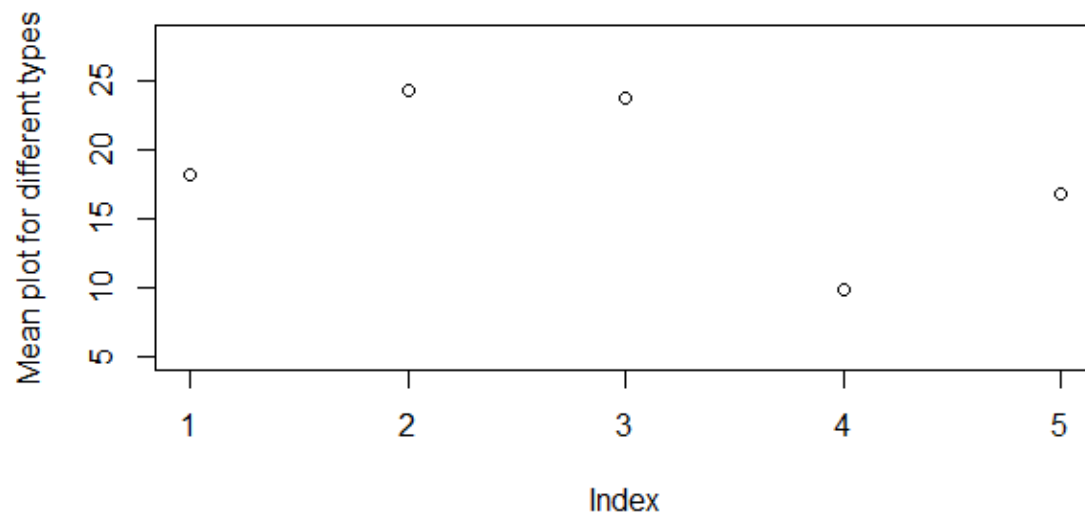
Variance of Selling Price = 62.7599

Standard deviation of Selling Price = 7.922119

[Histogram](#)



Means Plot



The histogram for the selling price of the car is right skewed with no outliers. The means plot shows that the mean is highest for the second type of the car (Large) and is lowest for the fourth type of the car (Small). The mean selling price of the car based on the 5 types is shown above.

Stem plot of Selling Price

The decimal point is at the |

6 | 4

8 | 034460128

10 | 001391133468

12 | 1253459

14 | 01491677899

16 | 3557

18 | 24458358

20 | 027895

22 | 737

24 | 48

26 | 137

28 | 0715

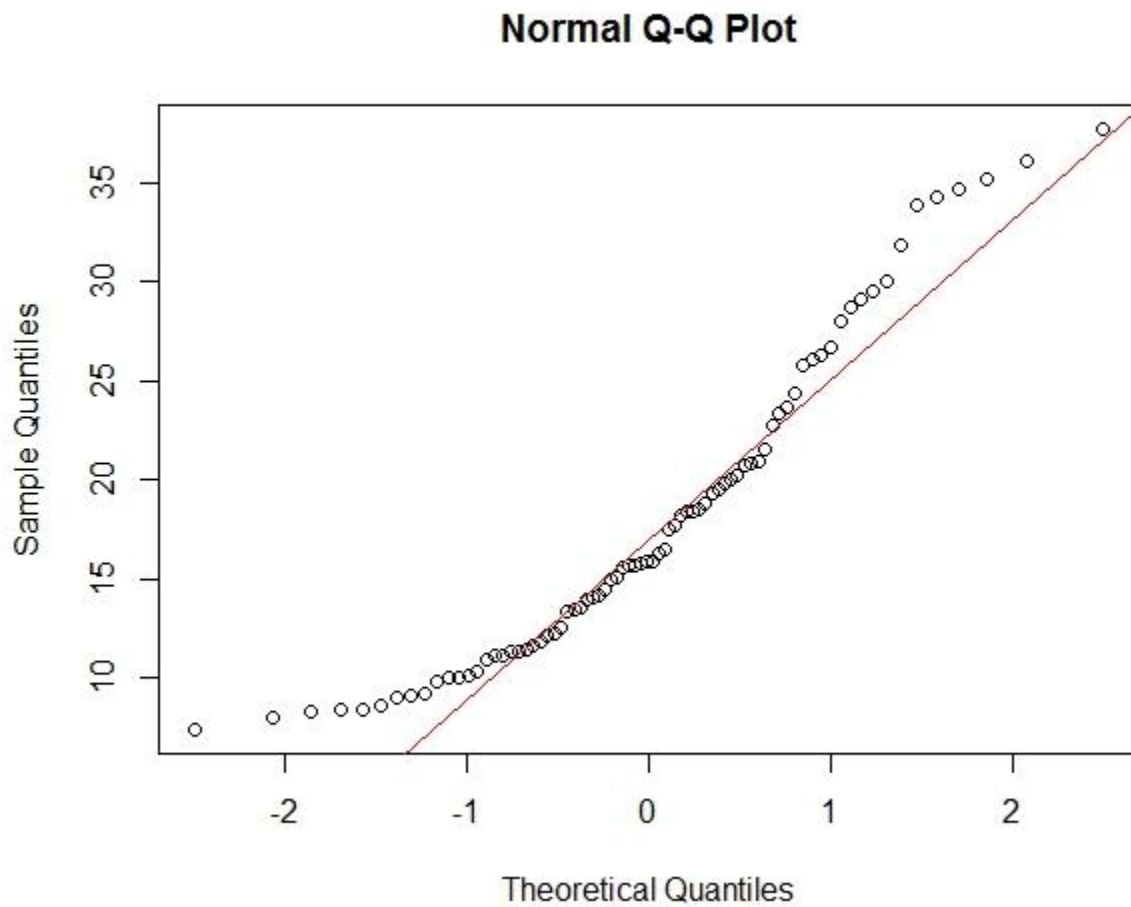
30 | 09

32 | 9

34 | 372

36 | 17

The stem plots show that the data is right skewed and the data doesn't have outliers. It is similar to the histogram which has been plotted.



The quantile plot shows that the values are not deviated from the line hence we can conclude that the data is almost symmetric.

t-test Analysis

μ_1 = Mean selling price of car type 1.

μ_4 = Mean selling price of car type 4.

Null Hypothesis:

$$H_0 : \mu_1 = \mu_4$$

Alternate Hypothesis:

$$H_a : \mu_1 \neq \mu_4$$

95 percent confidence interval:

(0.1538496, 10.6036504)

$t = 2.1094$, $df = 27.8$, $p\text{-value} = 0.04405$

Since $p\text{-value} < 0.05$, the null hypothesis is rejected. We have sufficient evidence that the means of car type 1 and 4 are not equal.

ANOVA Test

Consider the following,

μ_1 = Mean selling price of car type 1.

μ_2 = Mean selling price of car type 2.

μ_3 = Mean selling price of car type 3.

μ_4 = Mean selling price of car type 4.

μ_5 = Mean selling price of car type 5.

Null Hypothesis:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

Alternate hypothesis:

H_a : All the means are not equal.

The results of the ANOVA test on the data give us the following results –

```
> summary(test)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Type	1	262	261.93	4.355	0.0402 *
Residuals	76	4571	60.14		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The P-value is 0.0402; which is less than 0.05 (5% significance level) and thus we have sufficient evidence to reject the null hypothesis H_0 .

Multiple Linear Regression

```
Residuals:
    Min       1Q   Median       3Q      Max
-9.497 -4.252 -1.546  3.671 15.150

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   41.25240    3.02572   13.634 < 2e-16 ***
Manufacturer  -0.05790    0.07892   -0.734   0.465
Type          -0.66513    0.52541   -1.266   0.210
Miles.per.gallon -0.87325    0.12905   -6.767 2.67e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.931 on 74 degrees of freedom
Multiple R-squared:  0.4613,    Adjusted R-squared:  0.4394
F-statistic: 21.12 on 3 and 74 DF,  p-value: 5.49e-10
```

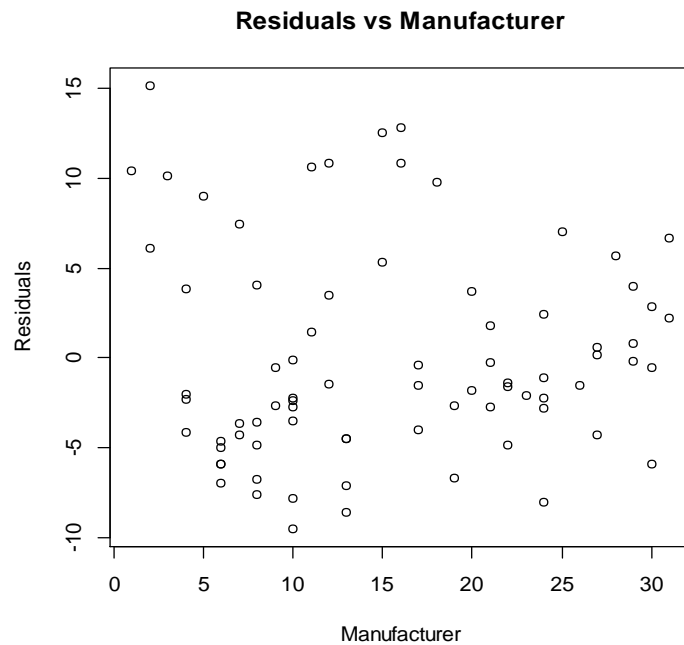
'Manufacturer type', 'Car type', 'Miles per gallon' is used as the variables in the Multiple linear regression. By using the above mentioned variables an equation is derived to determine the selling price.

Selling.Price= 41.25240 – 0.05790*Manufacturer – 0.66513*Type – 0.87325*Miles.per.gallon

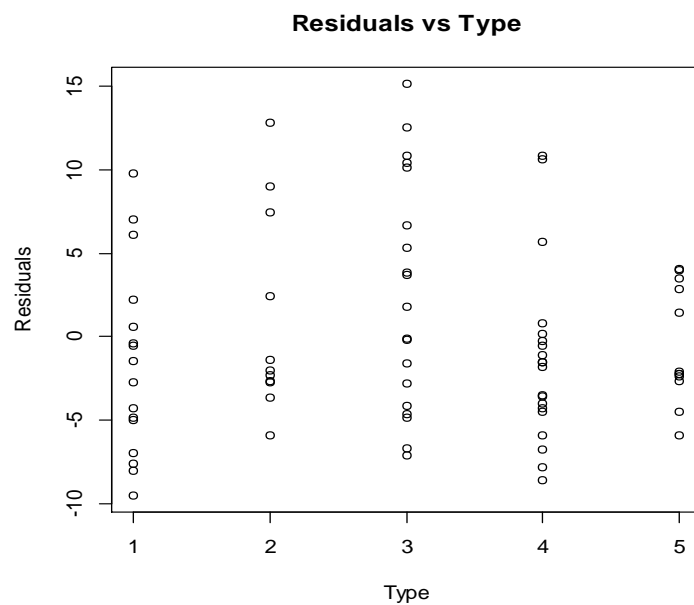
From the above equation one can infer that by decreasing the Manufacturer type by 0.05790, the selling price increases by 1 unit. Similarly if the car type value decreases by 0.66513 units and the Miles per gallon units decreases by 0.87325 units the selling price increase by 1 unit.

Multiple R-squared value for this is 0.4613.

Residual Plots



For the Residual plot of manufacturer type, the data doesn't follow a pattern so we use the squared value of manufacturer type in the next linear regression model.



For the Residual plot of car type, the data follows a pattern that is linear, so we use the normal value of the car type in the next linear regression model.

Adjusted Multiple Linear Regression

```
Residuals:
    Min       1Q   Median       3Q      Max
-8.747 -3.782 -1.272  2.978 14.362

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   43.75183    3.35502   13.041 < 2e-16 ***
Manufacturer  -0.62863    0.35554   -1.768  0.0812 .
Type          -0.59611    0.52114   -1.144  0.2564
Miles.per.gallon -0.84295    0.12891   -6.539 7.36e-09 ***
ManufacturerSq  0.01722    0.01047    1.645  0.1042
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.864 on 73 degrees of freedom
Multiple R-squared:  0.4806,    Adjusted R-squared:  0.4521
F-statistic: 16.88 on 4 and 73 DF,  p-value: 7.68e-10
```

'Manufacturer type', 'Car type', 'Miles per gallon' and 'ManufacturerSq type' is used as the variables in the Multiple linear regression. By using the above mentioned variables an equation is derived to determine the selling price.

Selling.Price = 43.75183 – 0.62863*Manufacturer – 0.59611*Type – 0.84295*Miles.per.gallon + 0.01722*ManufacturerSq

From the above equation one can infer that by decreasing the Manufacturer type by 0.05790, the selling price increases by 1 unit. Similarly if the car type value decreases by 0.66513 units and the Miles per gallon units decreases by 0.87325 units and if the ManufactureSq type increases by 0.01722 units the selling price increase by 1 unit.

Multiple R-squared for this is 0.4806.

Prediction

We will be estimating the Selling Price for the following group: Manufacturer = 2, Type = 1, Miles per gallon=20.

The **Predicted Selling Price** = $43.75183 - 0.62863*2 - 0.59611*1 - 0.84295*20 + 0.01722*4 = 28.156$

Predicted value of Selling Price of this group is **28.156**.

Appendix

Basics Code:

```
data <- read.csv("Car_Sales_Project_Final.csv", header=T);
data
attach(data)
Selling.Price<- data[,4]
Selling.Price
t1 <- Selling.Price[1:16]
t2 <- Selling.Price[17:27]
t3 <- Selling.Price[28:46]
t4 <- Selling.Price[47:66]
t5 <- Selling.Price[67:78]
mean(Selling.Price)
median(Selling.Price)
min(Selling.Price)
max(Selling.Price)
quantile(Selling.Price, probs = c(0,25,50,75,100)/100, type=1)
boxplot(Selling.Price~Type, main="Boxplot comparing different Types", xlab="Types", ylab="Price")
tapply(Selling.Price, Type, mean)
plot(tapply(Selling.Price, Type, mean), ylab="Mean plot for different types", ylim=c(5,28))
tapply(Selling.Price, Type, sd)
tapply(Selling.Price, Type, length)
var(Selling.Price)
sd(Selling.Price)
hist(Selling.Price, main="Histogram of Price", xlab="Price", col="Red")
stem(Selling.Price, scale=3)
qqnorm(Selling.Price);
qqline(Selling.Price, col="red")
t.test(t1,t4)
```

ANOVA Code:

```
data <- read.csv ("Car_Sales_Project_Final.csv", header=T);
attach (data);
test <- aov (Selling.Price~Type);
summary (test)
```

Linear Regression Code:

```
regression.lm <- lm(Mid.range.Price ~ Manufacturer + Type + City.mpg);

summary(regression.lm)
```

```
regression.res = resid(regression.lm)

plot(Manufacturer, regression.res, ylab="Residuals", xlab="Manufacturer", main="Residuals vs
Manufacturer")

plot(Type, regression.res, ylab="Residuals", xlab="Type", main="Residuals vs Type")

ManufacturerSq <- (Manufacturer)^2;

regression2 <- lm(Mid.range.Price ~ Manufacturer + Type + City.mpg + ManufacturerSq);

summary(regression2)
```