***EXECUTIVE SUMMARY***

*Dataset 🡪 mtcars.*

*Motor Trend,*  a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

* *“Is an automatic or manual transmission better for MPG”*
* *"Quantify the MPG difference between automatic and manual transmissions"*

#### Objective🡪Using simple linear regression analysis, we determine that there is a signficant difference between the mean MPG for automatic and manual transmission cars.

data(mtcars)

head(mtcars)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  |  |  |  |  |   mpg cyl disp hp drat wt qsec vs am gear carb |
| Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4 |
| Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4 |
| Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1 |
| Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1 |
| Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2 |
| Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1 |

mtcars$vs <- as.factor(mtcars$vs)

mtcars$am <- as.factor(mtcars$am)

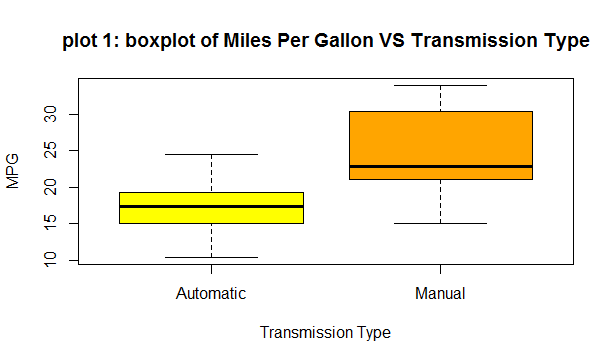
summary(mtcars$mpg)

|  |
| --- |
|  |

|  |
| --- |
| Min. 1st Qu. Median Mean 3rd Qu. Max. |
| 10.40 15.43 19.20 20.09 22.80 33.90 |

mtcars$am <- factor(mtcars$am,labels=c("Automatic","Manual"))

boxplot(mpg ~ am, data = mtcars, col = (c("yellow","orange")), ylab = "MPG", xlab = "Transmission Type",main="plot 1: boxplot of Miles Per Gallon VS Transmission Type")



fit <- lm(mpg ~ am, data = mtcars)

summary(fit)

SIMPLE LINEAR REGRESSION MODEL

Call:

lm(formula = mpg ~ am, data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-9.3923 -3.0923 -0.2974 3.2439 9.5077

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 17.147 1.125 15.247 1.13e-15 \*\*\*

am1 7.245 1.764 4.106 0.000285 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.902 on 30 degrees of freedom

Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385

F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

The p-value of 0.000285 is small and the CI does not include zero, so we can reject null in favor of the alternative hypothesis that there is a significant difference in MPG between the two groups at 95% Confidence Interval.

Furthermore, Adjusted R squared value is only 0.3385 which means that only 33.85% of the regression variance can be explained by our model. We will have to consider more predictor variables to see if they played a larger role in the model.

bestfit <- lm(mpg ~., data = mtcars)

summary(bestfit)

MULTIVARIABLE LINEAR REGRESSION

Call:

lm(formula = mpg ~ ., data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-3.4506 -1.6044 -0.1196 1.2193 4.6271

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.30337 18.71788 0.657 0.5181

cyl -0.11144 1.04502 -0.107 0.9161

disp 0.01334 0.01786 0.747 0.4635

hp -0.02148 0.02177 -0.987 0.3350

drat 0.78711 1.63537 0.481 0.6353

wt -3.71530 1.89441 -1.961 0.0633 .

qsec 0.82104 0.73084 1.123 0.2739

vs1 0.31776 2.10451 0.151 0.8814

am1 2.52023 2.05665 1.225 0.2340

gear 0.65541 1.49326 0.439 0.6652

carb -0.19942 0.82875 -0.241 0.8122

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom

Multiple R-squared: 0.869, Adjusted R-squared: 0.8066

F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07

.

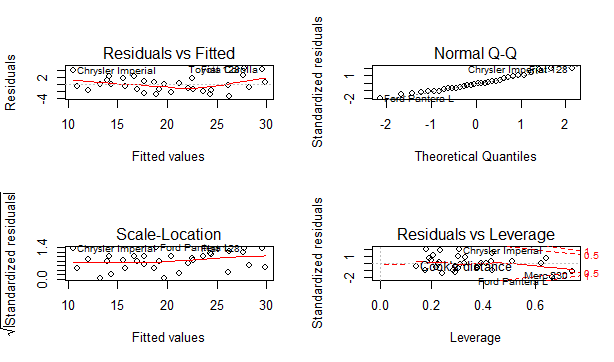
anova(fit, bestfit)

|  |
| --- |
| Analysis of Variance Table |
|  |
| Model 1: mpg ~ am |
| Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb |
| Res.Df RSS Df Sum of Sq F Pr(>F) |
| 1 30 720.90 |
| 2 21 147.49 9 573.4 9.0711 1.779e-05 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
| Based on the p-value of 3.745e-09, we can now reject the H0 and states that our Multivariable Model is significantly different than the Simple Linear Regression |

par(mfrow = c(2,2))

plot(bestfit)

## Model Residuals and Diagnostics



t.test(mpg ~ am, data = mtcars)

Welch Two Sample t-test

data: mpg by am

t = -3.7671, df = 18.332, p-value = 0.001374

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-11.280194 -3.209684

sample estimates:

mean in group Automatic mean in group Manual

17.14737 24.39231

P value<0.05

#### 🡪  we perform a t-test on the two subsets of mpg data: manual and automatic transmission assuming that the transmission data has a normal distribution and tests the null hypothesis that they come from the same distribution. Based on the t-test results, we reject the null hypothesis that the mpg distributions for manual and automatic transmissions are the same.