What does transmission is most economic: automatic or manual?

Héderson Pereira dos Santos

Summary Executive

This article shows how identify the best transmission of a car (automatic or manual) compared to gasoline consumption (represented by miles per gallon – mpg). It uses statistical concepts of regression linear to make these calculations. It presents two experiments, one uses linear regression model and the other uses multivariable regression model. The data of experiments are from R dataset mtcars. The first model is formed by a predicting variable (mpg attribute) and a dummy variable (binary variable) and the second model forms a relation between the weight and the horsepower to predicting which type of transmission is most economical.

Experiment with linear regression model

Figure 1 shows that the mean of gasoline consumption to automatic transmissions is less than to manual transmissions. Then, we have the following hypothesis: it appears that automatic cars have lower miles per gallon, and therefore a better fuel efficiency, than manual cars. But, it is possible that this is happening because the cars with manual transmissions of the sample are not economic cars. We have to use statistical test to check it.

To test if the miles per gallon depend on whether it's an automatic or manual transmission in the mtcars dataset, it uses Student's T-test. Figure 2 shows the result of this test made in R. We note that the p-value equal to 0.001374 and the confidence interval between -11.280 and -3.210. The mean to group 0 is 17.147 mpg and to group 1 is 24.392 mpg. Note that p-value is a low value and the 95% confidence interval describes how much lower the miles per gallon is in manual cars than it is in automatic car and the true difference is between 3.2 and 11.28.

To test if the automatic or manual transmission affects the gasoline consumption, it uses correlation test. Note in Figure 3 that the correlation between the kind of transmission and the miles per gallon may be weak. The correlation test in Figure 4 shows that there is a positive correlation and the coefficient correlation is equal to 0.5998324. This value varies from -1 to 1, where 1 represents a perfectly linear positive relationship, and -1 represents a perfectly linear negative relationship. 0 represents that the two are not correlated. The p-value is equal to 0.000285. P-value is the probability that this data would appear to be this strongly correlated by chance alone. Then, the correlation between mpg and am attributes is weak because the coefficient correlation is near 0 and p-value indicates a low probability to exist a strong correlation between them.

Consider predicting y (as mpg attribute) at a value of x (as am attribute). Figure 5 shows the summary coefficients of this function prediction. Estimate of the y-intercept is 17.147368 and the slope coefficient is 7.244939. This shows that there's a positive relationship, where increasing the am attribute value increases the miles per gallon. The stand error value represents the amount of uncertainty in our estimate of the slope. It is 1.764422. P-Value is the same of the t-test shows in the Figure 2, it´s 1.133983e-15.

Figure 6 shows that the confidence interval between 3.64151 and 10.84837. It´s the uncertainty in the fit and has 95% confidence interval for the expected mpg at am attribute. Figure 7 shows R-squared value. It is 36%. One way to check if the adjusted model is appropriate is to look the result of the determination coefficient (R-squared). This coefficient measures how much the dependent variable is explained by the model. Higher the value of R-squared better the dependent variable is explained by the model. Note in the Figure 7 and in the Figure 8 that the most predictions has residuals standard errors equal to 4.902 and that the distance of predictions are so far of center.

It concludes that attributes mpg and am not are goods variables to determine alone the economy of cars. The positive coefficient correlations attributes is weak because the coefficient correlation is near 0 and p-value indicates a low probability to exist a strong correlation between them. R-squared value is a low rate to

dependent variable explains the model. It´s very difficult make a regression model with a variable that has binary values.

Experiment with multivariable regression model

It presents a multivariable regression model with the aim of obtaining a more correct prediction of which type of transmission has a lower consumption of gasoline. It uses the attributes weight (wt) and horsepower (hp). It creates two models, one to automatic transmission cars and other to manual transmission car.

Figure 9 and Figure 10 shows the plot graphs of attributes wt and hp. Note that there may be a negative regression between wt and mpg and between hp and mpg. Figure 11 shows the coefficients values of regressions models to automatic and manual cars. Note that according to these values, when increases the weight and the horse power than decreases the gasoline consumption because the coefficients of the linear terms of these attributes are all negative. The multivariable regression model to automatic car is y = 30.703 – 1.856x1 – 0.041x2. The multivariable regression model to manual car is y = 44.443 – 7.625x1 – 0.013x2. Let y = mpg, x1 = wt and x2 = hp.

Note in the Figure 11 that – to automatic car - R-squared value is 77% and the Adjusted R-Square value is 74%. To manual car, R-squared value is 84% and the Adjusted R-Square value is 80%. It is better than the model with only the am attribute. P-value is 8.52 x 10-6 to automatic car and 0.0001153 to manual car

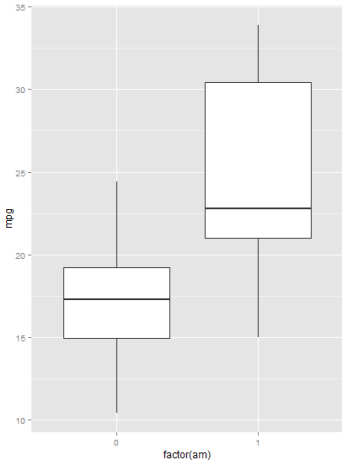
. To assess the suitability (quality) of the fit, it is necessary to verify if the hypotheses of normality and constant variance (homoscedasticity) of the residuals are satisfied.

Using the R commands in the Figure 15, it have done the graphs in the Figures 12, 13 and 14 to automatic cars and it have done the graphs in the Figure 16, 17 e 18 to manual cars.

Note in the Figure 13 and in the Figure 17 that the residuals of this model has normal distribution because, the plots are close to the line identity. Residual has constant variance (homoscedasticity) like shows in the Figure 14 and 18. To automatic cars the residual are between -3 and 3 and to manual cars the residuals are between -2 and 2.

It concludes that a multivariable regression model may be more effective than a linear regression model. In these models presented, the automatic car is more economical than the manual car. Automatic car consumes 30.703 mpg when weight and horsepower variable is zero (this makes no sense because the area is always positive). Miles per gallon reduces in 1.856% when the weight is greater than 1 and the others variables maintains the values. It reduces in 0.041 when the horsepower is greater than 1 and the others variables maintains the values. Manual cars consumes 44.443 mpg when weight and horsepower variable is zero (this value is greater than automatic car value) and reduces in 7.625% when the weight is greater than 1 and reduces 0.013% when the horsepower is greater than 1.

Appendix

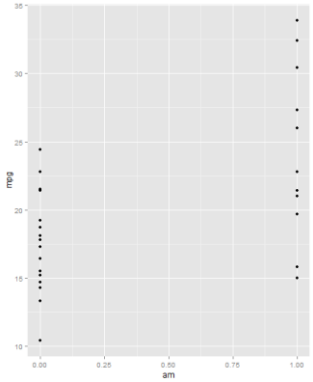



```
> library(ggplot2)
> ggplot(mtcars, aes(x=factor(am), y=mpg)) + geom_boxplot()
```

Figure 1 –Comparison of gasoline consumption (attribute mpg) between automatic and manual transmissions.

```
> 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 0 mean in group 1
       17.14737        24.39231
```

Figure 2 – T test results for the mpg attributes and am. It´s was made in R




```
> ggplot ( mtcars , aes ( x = am , y = mpg ))  + geom_point ()
```

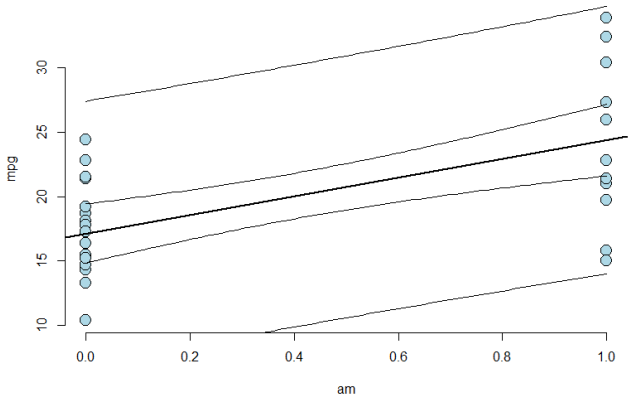
Figure 3 - Plot graph of gasoline consumption by automatic or manual transmission

```
> cor.test ( mtcars $ mpg , mtcars $ am )

 Pearson's product-moment correlation

data:  mtcars$mpg and mtcars$am
t = 4.1061, df = 30, p-value = 0.000285
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3175583 0.7844520
sample estimates:
      cor
0.5998324
```

Figure 4 - Correlation test between mpg and am attributes

```
> y <- mtcars$mpg; x <- mtcars$am; n <- length(y)
> fit <- lm(y ~ x);
> summary(fit)$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 17.147368   1.124603 15.247492 1.133983e-15
x            7.244939   1.764422  4.106127 2.850207e-04
```

Figure 5 – Values of coefficients of regression function lm(mpg~am)




```
> newdata <- data.frame(x = xvals)
> p1 <- predict(fit, newdata, interval = ("confidence"))
> p2 <- predict(fit, newdata, interval = ("prediction"))
> plot(x, y, frame=FALSE,xlab="am",ylab="mpg",pch=21,col="black", bg="lightblue", cex=2)
> abline(fit, lwd = 2)
> lines(xvals, p1[,2]); lines(xvals, p1[,3])
> lines(xvals, p2[,2]); lines(xvals, p2[,3])
> sumCoef <- summary(fit)$coefficients
> sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[2, 2]
[1]  3.64151 10.84837
```

Figure 6 – Confidence Interval of regression function lm(mpg~am)

```
> summary.lm(fit)

Call:
lm(formula = y ~ x)

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 ***
x              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
```

Figure 7 Residual Standard error and Multiple R-squared values

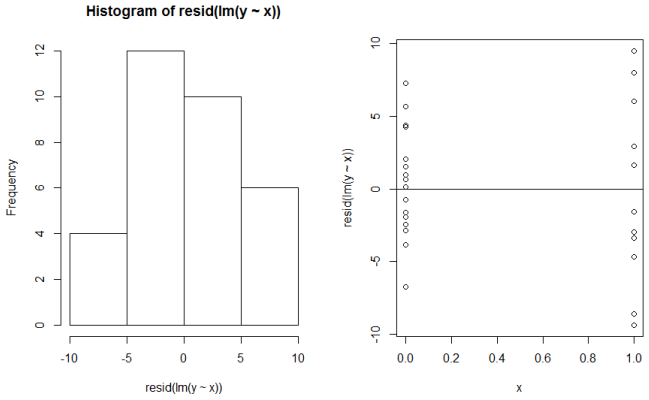


Figure 8 – Histogram of residuals of regression function lm(x~y) and graph plot of residuals of regression function lm(x~y).

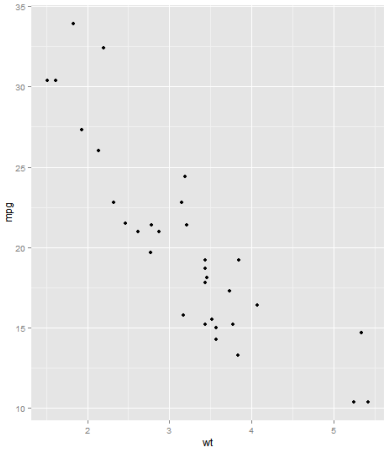


Figure 9 - Plot graph of gasoline consumption by weight

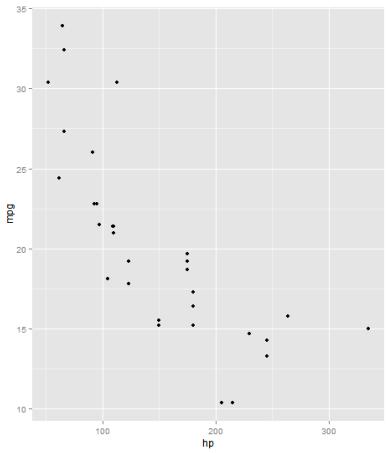


Figure 10 - Plot graph of gasoline consumption by hp

```
> fit_auto<-lm(mtcars$mpg[mtcars$am==0]~mtcars$wt[mtcars$am==0]+mtcars$hp[mtcars$am==0])
> summary(fit_auto)

Call:
lm(formula = mtcars$mpg[mtcars$am == 0] ~ mtcars$wt[mtcars$am ==
    0] + mtcars$hp[mtcars$am == 0])

Residuals:
    Min      1Q  Median      3Q     Max
-2.9873 -1.4590 -0.0835  1.3561  3.3331

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             30.70393    2.29396  13.385 4.16e-10
mtcars$wt[mtcars$am == 0] -1.85591    0.81043  -2.290  0.03594
mtcars$hp[mtcars$am == 0] -0.04094    0.01169  -3.503  0.00294

(Intercept)             ***
mtcars$wt[mtcars$am == 0] *
mtcars$hp[mtcars$am == 0] **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.96 on 16 degrees of freedom
Multiple R-squared:  0.7676, Adjusted R-squared:  0.7385
F-statistic: 26.42 on 2 and 16 DF,  p-value: 8.515e-06
```

```
> fit_manual<-lm(mtcars$mpg[mtcars$am==1]~mtcars$wt[mtcars$am==1]+mtcars$hp[mtcars$am==1])
> summary(fit_manual)

Call:
lm(formula = mtcars$mpg[mtcars$am == 1] ~ mtcars$wt[mtcars$am ==
    1] + mtcars$hp[mtcars$am == 1])

Residuals:
    Min      1Q  Median      3Q     Max
-2.7313 -1.3217 -1.0014 -0.0759  5.5987

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             44.44393    3.89917  11.398 4.73e-07
mtcars$wt[mtcars$am == 1] -7.62486    2.19997  -3.466  0.00606
mtcars$hp[mtcars$am == 1] -0.01315    0.01615  -0.814  0.43436

(Intercept)             ***
mtcars$wt[mtcars$am == 1] **
mtcars$hp[mtcars$am == 1]
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.728 on 10 degrees of freedom
Multiple R-squared:  0.8369, Adjusted R-squared:  0.8043
F-statistic: 25.66 on 2 and 10 DF,  p-value: 0.0001153
```

Figure 11 – Values of coefficients and residuals of models automatic and manual.

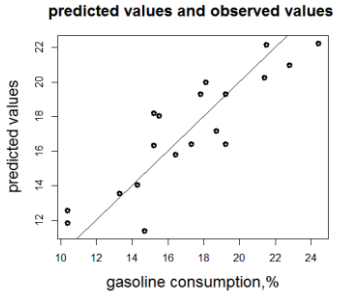


Figure 12 – Predicted values and observed values by mph to automatic cars. Plot graph are the predicted values and line graph are the observed values.

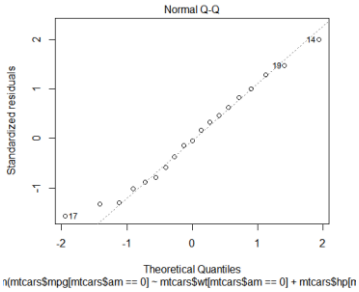


Figure 13 - Normal probability plot of the residuals to automatic cars.

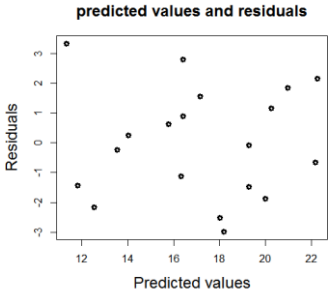


Figure 14 – Residuals by predicted values to automatic cars.

```
> a<-fitted(fit)
> t<-resid(fit)
> plot(y,a,main="predicted values and observed values ",ylab="predicted values",xlab="gasoline consumption,%",pch=1,lwd=3,
cex.lab=1.5, cex.main=1.5)
> abline(0,1)
> plot(fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> plot(a,t,main="predicted values and residuals",ylab="Residuals",xlab="Predicted values",pch=1,lwd=3,cex.lab=1.5,cex.main
=1.5)
> abline(0,1)
```

Figure 15 – R commands to create the graphs in the Figures 12 to 14 and 16 to 18. Fit argument in the first and second line must be changed to fit_auto or fit_manual model.

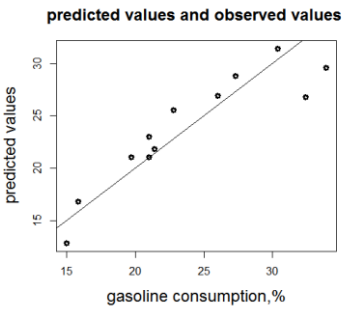


Figure 16 – Predicted values and observed values by mph to manual cars. Plot graph are the predicted values and line graph are the observed values.

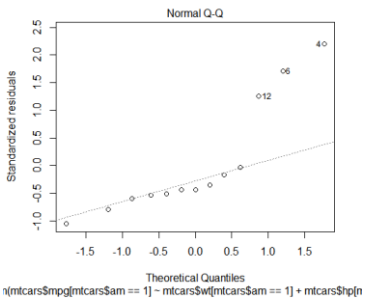


Figure 17 - Normal probability plot of the residuals to automatic cars.

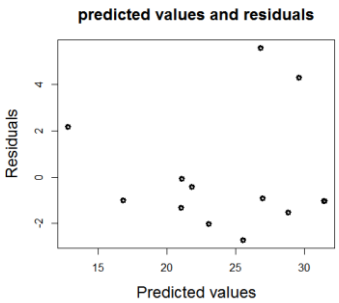


Figure 18 – Residuals by predicted values to automatic cars.