Regression Models - Project

Geoff Williams

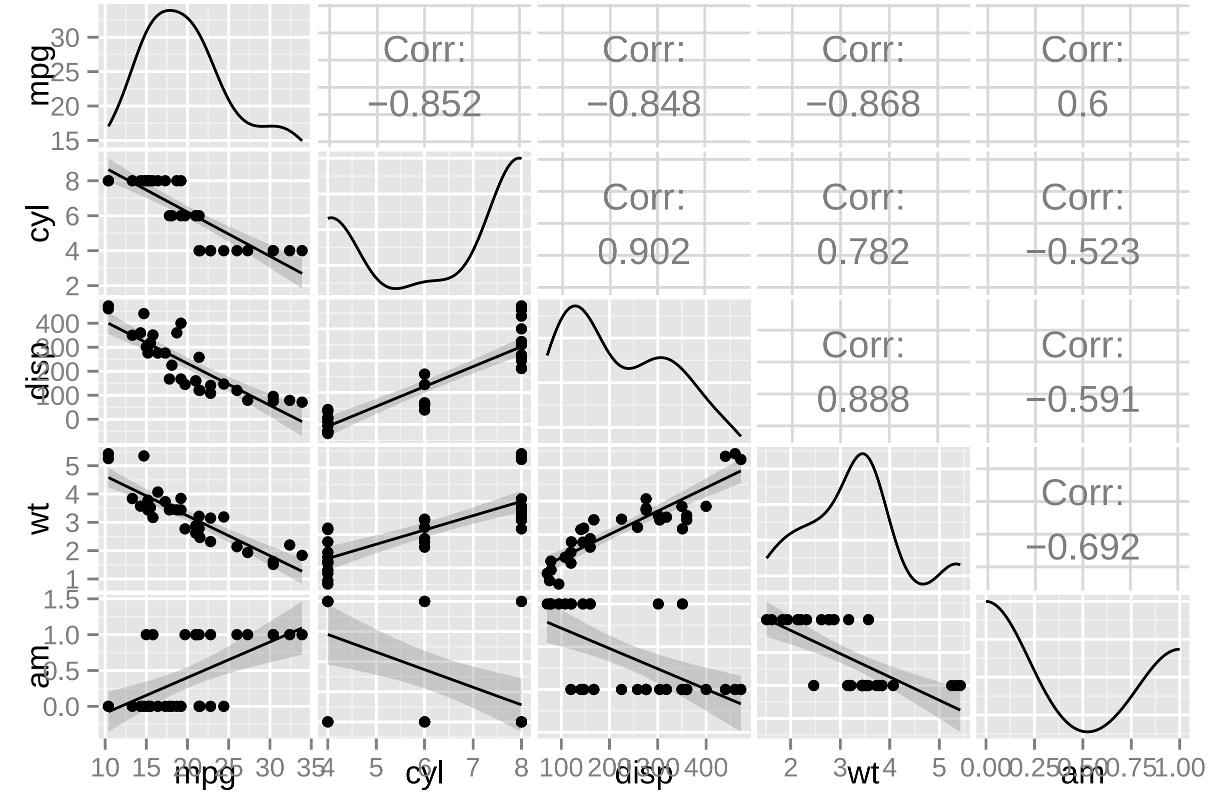
Monday, September 21, 2015

Executive Summary

This project was specifically to look at whether automatic vs manual transmission has any statistically valid effect on fuel economy and if so, to what degree. From the data in the mtcars set, it was seen that while there is correlation between economy (in mpg) and factors such as cylinders and weight, there didn't appear any direct correlation with transmission types. While there was some relationship it could be more attributed to the fact that larger cars with more cylinders which had lower economy also generally had automatic transmissions.

Exploratory

If we first explore the economy with the probable main contributors with fuel usage (weight, cylinders, transmission and displacement)



We can see that direct correlation between mpg and “am”" (transmission type) is relatively low compared to the other factors such as cylinders and weight which are quite a bit higher

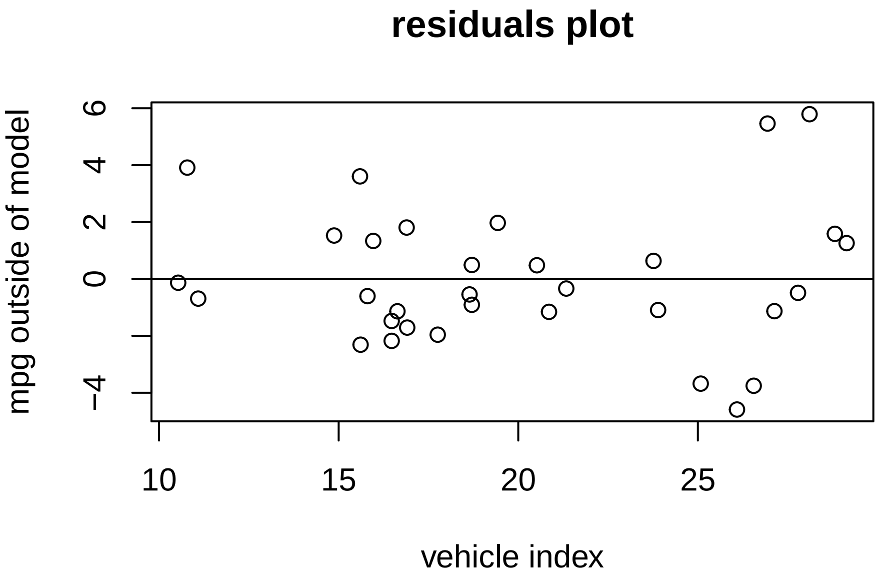
Model Selection

So, if we proceed to look at four models for mpg correlation, one with just tranmission (Table 1), another with transmission and weight (Table 2) , the next with the addition of cylinders (Table 3) and one final one with the addition of displacement (Table 4), we see something interesting. When we look at the results in table 1, there initally does seem to be a fit with a P value significantly less than 0.05. However when we include the weight and cylinders in the next two models (Table 2 and 3), the P value becomes very high (close to 1) suggesting that the transmission factor ("am") when cylinder and or weight is introduced becomes uncorrelated. We also find that the P-Value for diplacement (table 4) is very high indicating it also doesn't add extra information to the model and is most likely also related to the other two factors. To test this we used a nested model test (using anova - Table 5 and using Weight as the benchmark) and observed, based on the P-value, that the weight parameter definetely seemed relevant while the cylinders parameter while not as much was still seem relevent to include based on a 5% P-value threshold

So the result selected model was mpg = B0 + B1*weight + B2*cylinders The overall fit for this model gave as an R-Squared of 0.82 and residual standard error of around 2.5 compared to nearly (table 5) when using using the transmission type

Residuals and diagnostics

To check this model we looked at how the residuals looked in reference to the model



This tends to indicate that thee model is reasonably close with no apprent residual pattern and only a few vehicles falling outside of more than 2 mpg outside the esimtate.

Lastly we look at the HatValues (Table 6), we see that most values are wihtin 0.1 and at most 0.25 indicating that no vehicle adds substantial leverage to the overall model

Conclusion

So from the model fit it seems that weight is the major factor influencing the mpg values with some influence from the number of clyinders. Not all variation is explained however and would mostly likely be due to individual engineering variances as indicated by the hat values

Appendix - Tables and figures

Table 1 - Model for mpg = B0 + B1*transmission

```r
fit1<-lm(mpg~factor(am),data=mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

Table 2 - Model for mpg = B0 + B1*transmission + B2*Weight

```r
fit2<-lm(mpg~wt+factor(am),data=mtcars)
summary(fit2)$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
## factor(am)1 -0.02361522  1.5456453 -0.01527855 9.879146e-01
```

Table 3 - Model for mpg = B0 + B1*transmission + B2*Weight + B3*Cylinders

```r
fit3<-lm(mpg~wt+factor(cyl)+factor(am), data=mtcars)
summary(fit3)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  33.7535920  2.8134831 11.9970836 2.495549e-12
## wt           -3.1495978  0.9080495 -3.4685309 1.770987e-03
## factor(cyl)6 -4.2573185  1.4112394 -3.0167231 5.514697e-03
## factor(cyl)8 -6.0791189  1.6837131 -3.6105432 1.227964e-03
## factor(am)1   0.1501031  1.3002231  0.1154441 9.089474e-01
```

Table 4 - Model for mpg = B0 + B1*transmission + B2*Weight + B3*Cylinders +B4*Displacement

```r
fit4<-lm(mpg~wt+factor(cyl)+factor(am)+disp, data=mtcars)
summary(fit3)$coef
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  33.7535920  2.8134831 11.9970836 2.495549e-12
## wt           -3.1495978  0.9080495 -3.4685309 1.770987e-03
## factor(cyl)6 -4.2573185  1.4112394 -3.0167231 5.514697e-03
## factor(cyl)8 -6.0791189  1.6837131 -3.6105432 1.227964e-03
## factor(am)1   0.1501031  1.3002231  0.1154441 9.089474e-01
```

Table 5 - Nested Model selection

```r
# Using Weight as the benchmark as it had the highest correlation
fit1<-lm(mpg~wt,data=mtcars)
fit2<-lm(mpg~wt+factor(am),data=mtcars)

anova(fit1, fit2 ,fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + factor(am)
## Model 3: mpg ~ wt + factor(cyl) + factor(am)
## Model 4: mpg ~ wt + factor(cyl) + factor(am) + disp
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 278.32
## 2     29 278.32  1     0.002 0.0003 0.985897
## 3     27 182.97  2    95.351 6.7784 0.004273 **
## 4     26 182.87  1     0.099 0.0141 0.906470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 6 - Selected Model -> mpg = B0 + B1*Weight + B2*Cylinders

```r
fit5<-lm(mpg~wt+factor(cyl), data=mtcars)
summary(fit5)
```

```
##
## Call:
## lm(formula = mpg ~ wt + factor(cyl), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     33.9908     1.8878  18.006  < 2e-16 ***
## wt              -3.2056     0.7539  -4.252 0.000213 ***
## factor(cyl)6    -4.2556     1.3861  -3.070 0.004718 ** 
## factor(cyl)8    -6.0709     1.6523  -3.674 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82 
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

Table 7 - Hat values diagnostic

```
hatvalues(fit5)
```

```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710 
##          0.16434300          0.14795437          0.09101121 
##      Hornet 4 Drive   Hornet Sportabout             Valiant 
##          0.14368963          0.09861466          0.15307634 
##          Duster 360           Merc 240D            Merc 230 
##          0.08744400          0.16199592          0.15584604 
##            Merc 280           Merc 280C          Merc 450SE 
##          0.15191887          0.15191887          0.07186417 
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood 
##          0.07772924          0.07560618          0.20743395 
## Lincoln Continental   Chrysler Imperial            Fiat 128 
##          0.24790608          0.22887836          0.09154798 
##         Honda Civic      Toyota Corolla       Toyota Corona 
##          0.13001858          0.10857020          0.09370304 
##    Dodge Challenger         AMC Javelin          Camaro Z28 
##          0.09139267          0.09910298          0.07363228 
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2 
##          0.07349604          0.10160282          0.09275526 
##        Lotus Europa      Ford Pantera L        Ferrari Dino 
##          0.14281810          0.13120415          0.15333341 
##       Maserati Bora          Volvo 142E 
##          0.08744400          0.11214758
```

Plot Code

Code for Exploratory plot

g<-ggpairs(mtcars[,c(1,2,6,9)],lower=list(continuous="smooth"),params=c(method="loess"))

Code for residuals plot

fit<-lm(mpg~wt+factor(cyl), data=mtcars) plot(predict(fit),resid(fit),xlab="vehicle index",ylab="mpg outside of model",main="residuals plot") abline(h=0)