CourseProject_Regression

Fei

Saturday, June 21, 2014

Executive summary

Through stepwise model selection through minimizing AIC, the output model turned out to be mpg ~ as.factor(cyl) + hp + wt + as.factor(am). However, performing ANOVA anlysis to compare the above model with and without the am variable showed that such two models are likely to be similiar. A Shapiro-Wilk test of normality over the residues the above model failed to reject the null hypothesis, validating the anova analysis. Diagnose analysis showed that this model is unlikely to be influenced by outliers. Therefore, basing on all the studies, the conclusions are: 1. with the given data, the manual and automatic transmission types do not significantly impact the MPG; 2. with the given data, holding all other variables constant, vehicle of manual transmission have 1.81 increase in MPG compared to vehicle of automatic transmission.

Use AIC (Akaike information criterion) for model selection. The AIC takes into account the number of parameters in the model, and the goodness of fit. So in theory, it not only rewards goodness of fit, but also encourages parsimony.

Backward, forward, and both direction stepwise regression gives the same model. Backward direction method is used here.

```
data('mtcars')
du<-capture.output(step(lm(mpg~ as.factor(cyl)+disp+hp+drat+wt+qsec
+as.factor(vs)+as.factor(gear)+as.factor(carb)
+as.factor(am),data=mtcars), direction='backward'))
```

Model generated: mpg ~ as.factor(cyl) + hp + wt + as.factor(am)

Coefficients are:

```
fit<-lm(formula = mpg ~ as.factor(cyl) + hp + wt + as.factor(am), data = mtcars)
summary(fit)$coefficients
```

```
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     33.70832    2.60489 12.9404 7.733e-13
## as.factor(cyl)6 -3.03134    1.40728 -2.1540 4.068e-02
## as.factor(cyl)8 -2.16368    2.28425 -0.9472 3.523e-01
## hp              -0.03211    0.01369 -2.3450 2.693e-02
## wt              -2.49683    0.88559 -2.8194 9.081e-03
## as.factor(am)1   1.80921    1.39630  1.2957 2.065e-01
```

However, AIC does not provide a test of the model. The reduction in the AIC score by inclusion or exclusion may not be significant. The anova test shows that the model with am and without am is not significantly different.

```
fit_wo<- lm (mpg ~ as.factor(cyl) + hp + wt,data=mtcars)
fit<-lm(formula = mpg ~ as.factor(cyl) + hp + wt + as.factor(am), data = mtcars)
anova(fit_wo,fit)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: mpg ~ as.factor(cyl) + hp + wt
## Model 2: mpg ~ as.factor(cyl) + hp + wt + as.factor(am)
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     27 161
## 2     26 151  1      9.75 1.68   0.21
```

P-value = 0.21, fail to reject the null. So this two models are likely to be similar.

ANOVA assumes normality of residues, so need to test the normality with Shapiro-Wilk test.

```
shapiro.test(fit$residuals) # Test normality
```

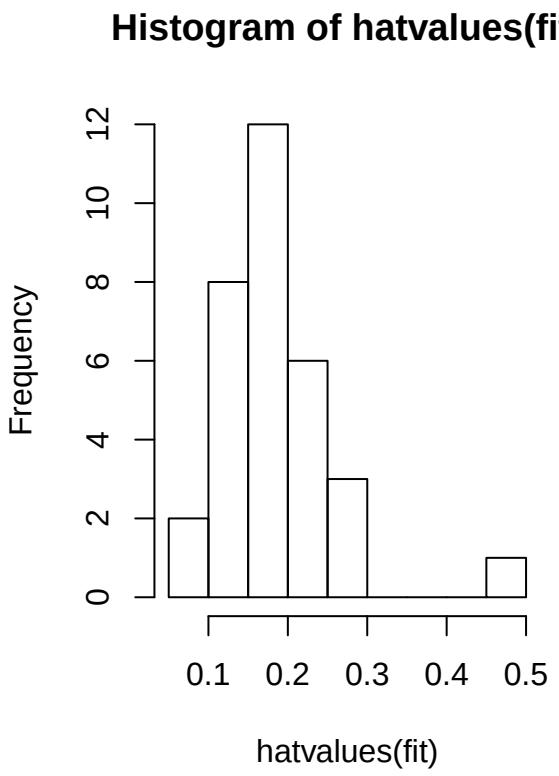
```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.9681, p-value = 0.4479
```

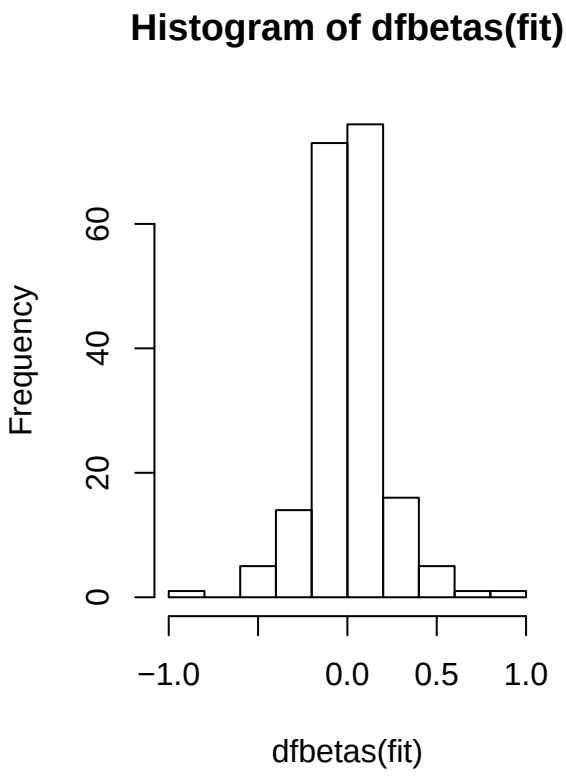
P-value = 0.45, fail to reject the null. So the residues are likely to follow a normal distribution. So our anova test is likely to be valid.

Diagnoses with R plot(fit) function, and influence measures on leverage and change in individiual coefficient with ith point deleted from the fitting model. All anlysis, codes and figures shown in appendex, show that the leverages and change in individual coefficients seem to be small. No obvious outlier effect is observed.

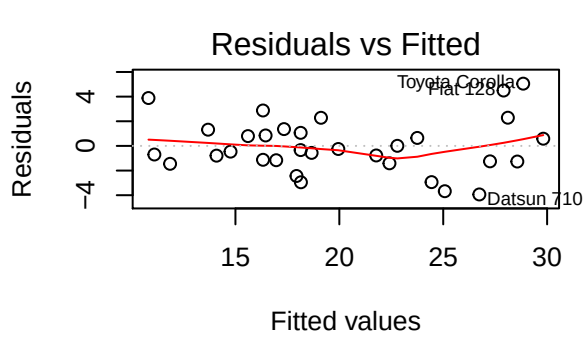
Appendix

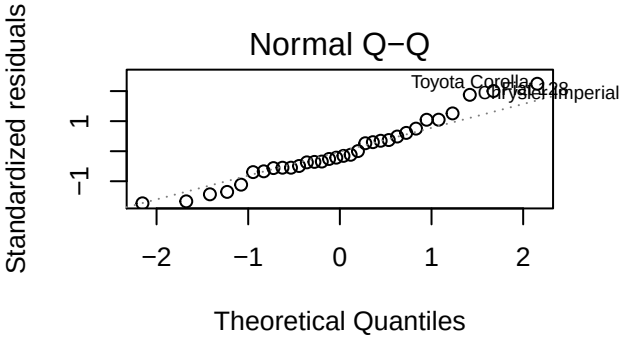
```
par(mfrow = c(1, 2))
hist(hatvalues(fit))
hist(dfbetas(fit))
```

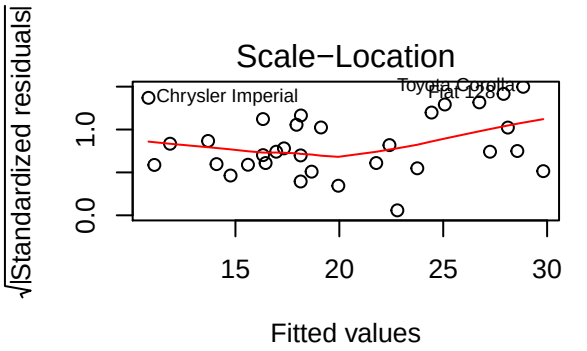


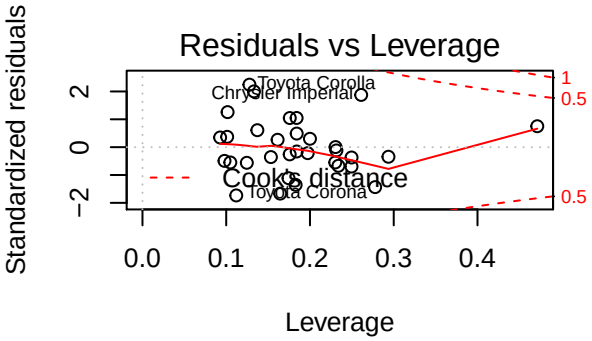



```
par(mfrow = c(2, 2))
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
15
20
25
30
−4
0
4


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
15
20
25
30
0.0
1.0


Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
0.0
0.1
0.2
0.3
0.4
−2
0
2