Regression Models Project (Two Page Condensed Version)- Jeff Sheremata

Executive Summary Motor Trend, is interested in the following two questions: 1. Is an automatic or manual transmission better for MPG? 2. What is the MPG difference between automatic and manual transmissions? The approach taken was to use the mtcars dataset, and create a linear regression model with MPG as the predicted variable, with transmission and other variables as predictors. A linear regression model (Rˆ2 =0.866) determined that when the other model variables (hp, cyl, and wt) were constant, manual trasmissions improved MPG by an average of 1.8 MPH compared to automatic transmissions.

Step 1 mtcars data is loaded and some variables are transformed into factors

```
data(mtcars)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"));
mtcars$carb <- factor(mtcars$carb); mtcars$cyl<- factor(mtcars$cyl);
mtcars$gear <-factor(mtcars$gear); mtcars$vs<- factor(mtcars$vs)
```

Step 2 Exploratory Data Analysis A pairs plot of all of the variables in this study is presented in the Appendix in Figure 1. By visual interpretation, most of the variables apear to impact MPG.

Step 3 Statistical Inference

```
options(digits=4); ttest<-t.test(mpg ~ am, data = mtcars)
```

In mtcars, the mean MPGs of the cars with manual and automatic transmissions are 17.1474 and 24.3923 MPG. A boxplot of the automatic vs. manual transmission MPG data is presented in Figure 2 of the Appendix. The p-value of the t-test is 0.0014. We reject the null hypothesis and conclude that the automatic and manual cars are from different populations.

Step 4 Regression Model Development The step function is used to compose a linear predictive model of MPG variables. Step uses both forward and backward selection to identify the variables that most significantly impact the ouput variable (in our case MPG). Step requires an initial model. For an initial model, I supplied a linear model (initialmodel) that had all of the variables as inputs. For statistical comparison basemodel is defined as the model that uses only transmission type (am) as a predictor.

```
initialmodel <- lm(mpg ~ ., data = mtcars)
basemodel <- lm(mpg ~ am, data = mtcars)
stepmodel <- step(initialmodel, direction = "both")
supermodel1 <- lm(mpg ~ cyl + hp+wt*am, data = mtcars)
supermodel2 <- lm(mpg ~ cyl + hp*wt+am, data = mtcars)
```

The only predictor variable in the basemodel is am. The basemodel DF of freedom is 30. The adjusted Rˆ2 is 0.3598 the residual standard error is 4.902. The Rˆ2 value indicates that only aproximately 36% of the variance of MPG can be explained by the linear model.

```
summary(stepmodel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.939 -1.256 -0.401  1.125  5.051
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7083     2.6049   12.94  7.7e-13 ***
## cyl6         -3.0313     1.4073   -2.15   0.0407 *
## cyl8         -2.1637     2.2843   -0.95   0.3523
## hp           -0.0321     0.0137   -2.35   0.0269 *
## wt           -2.4968     0.8856   -2.82   0.0091 **
## amManual      1.8092     1.3963    1.30   0.2065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

The variables that were selected for the stepmodel are: cyl, hp, wt, and am. The stepmodel DF is 26. The adjusted Rˆ2 is 0.8659 the residual standard error is 2.4101. This is lower than the residual standard error of the base model. The Rˆ2 value for the stepmodel indicates that the model can explain 86.6% of the variance of MPG. Thus, the stepmodel resulted in both a lower residual standard error and a higher Rˆ2. supermodel1 adds a wt-am interaction term. The adjusted Rˆ2 is 0.8841 the residual standard error is 2.2845. Relative to the stepmodel, supermodel1 has both an improved Rˆ2 and residual standard error. The model can explain 88.4% of the MPG variance. supermodel2 is conceptual and designed to see if there is interaction between hp and weight. The adjusted Rˆ2 is 0.888 the residual standard error is 2.2456. Including the nonlinear interaction term improves both Rˆ2 and residual standard error. This indicates that there might be a nonlinear interaction between horsepower and wt that should be addressed in a future nonlinear model. The model interpretation is beyond the scope of this course.

```
anova(basemodel,stepmodel,supermodel1)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
## Model 3: mpg ~ cyl + hp + wt * am
##   Res.Df RSS Df Sum of Sq     F  Pr(>F)
## 1     30 721
## 2     26 151  4       570 27.30 8.4e-09 ***
## 3     25 130  1        21  3.94   0.058 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The signifficant p-value of the ANOVA analysis for the stepmodel relative to the basemodel (that only has am as a predictor), the model with cyl + hp + wt is signifficant different than the base model. We conclude that adding cyl + hp + wt to the base model signifficantly improves the model accuracy. However, the p-value of the ANOVA anlaysis for supermodel1 realative to the stepmodel is not signifficant. There is not signifficant statistical evidence that the supermodel1 with the wt-am interaction is different than the stepmodel.

The stepmodel residual plots are in Appendix Figure 3. There are no trends or biases in a plot of residuals vs fitted values, and they appear randomly scattered. The Normal QQ plot indicates that the residuals are normally distributed. The Scale-Location plot indicates a random scatering of standarized residuals, indicating a constant variance. The Residuals vs Leverage plot indicates that there are no outliers beyond the 0.5 bands. We conclude that the residuals of the stepmodel are normally distributed and homosekedastic.

Conclusions We conclude: 1. t.test and linear regression indicate that an a manual transmission is beter for MPG 2. The selected linear stepmodel(Rˆ2 =0.866), when hp, cyl, and wt are constant, indicates a manual transmission improves MPG with an average increase of 1.8 MPG compared to an automatic transmission.

Appendix

Figure 1. Pair plots of the variables in mtcars

```r
pairs(mpg ~ ., data = mtcars)
```

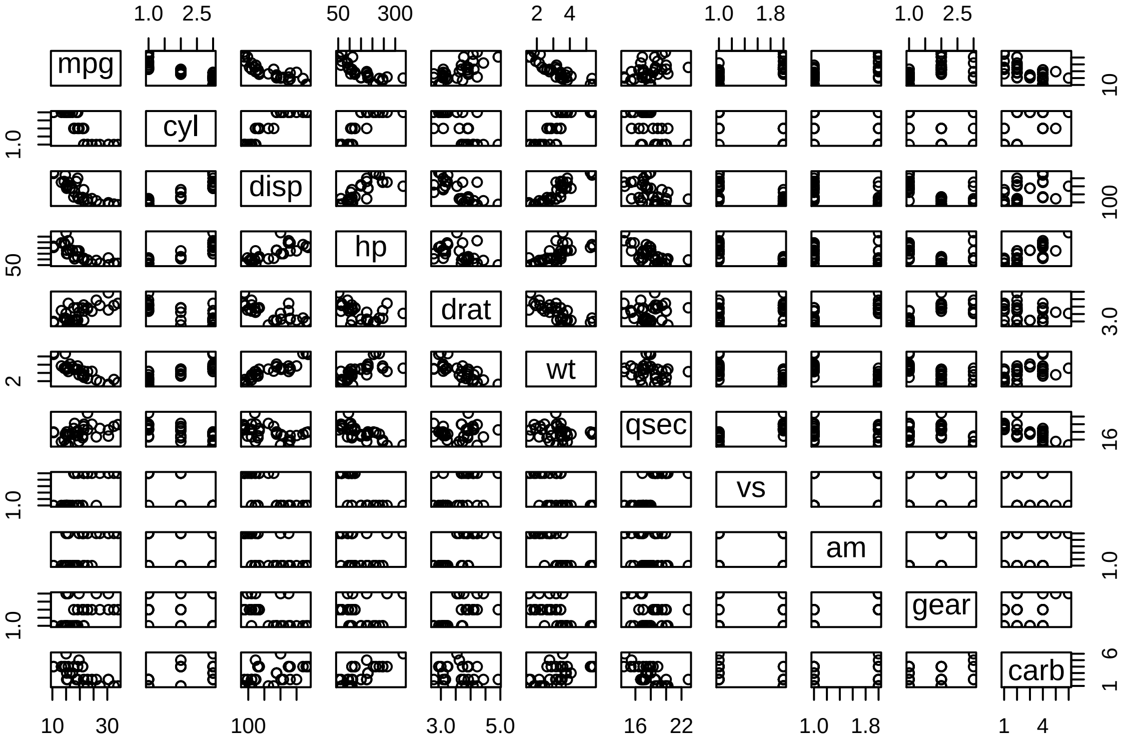


Figure 2. Boxplot of MPG by transmission type

```
boxplot(mpg ~ am, data = mtcars, ylab = "MPG", xlab = "Transmission Type")
```

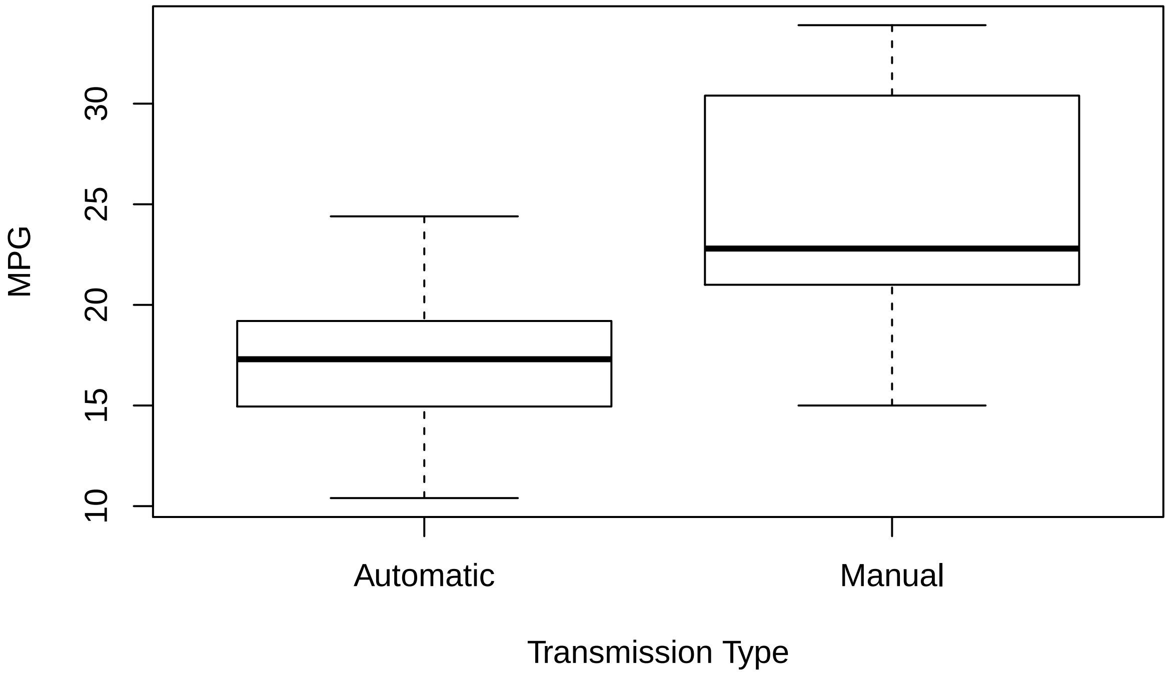
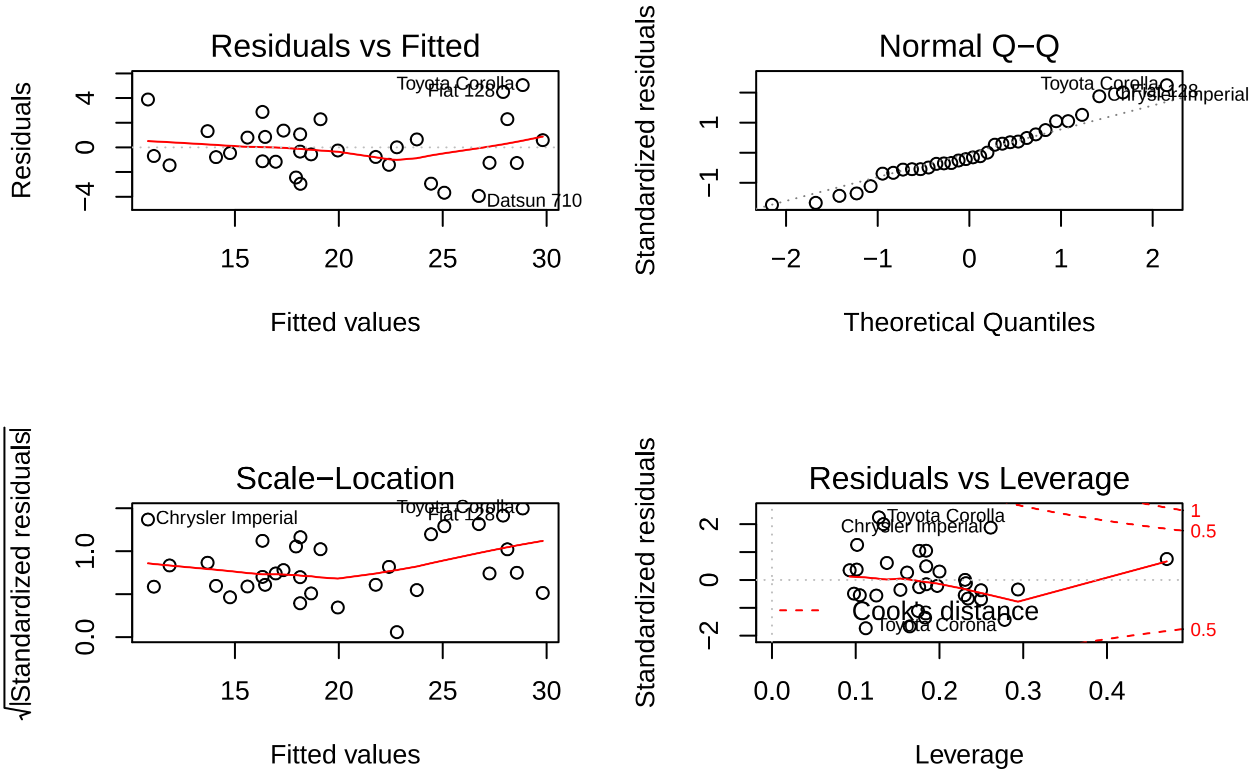


Figure 3. Residual analysis of the stepmodel

```r
par(mfrow=c(2, 2))
plot(stepmodel)
```


Residuals vs Fitted
Residuals
Fitted values
−4
0
4
15
20
25
30
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5