Regression Models Week 4 - MPG Analysis

Keith Swaback, Apr 23 2017

Executive Summary

In this assignment we are asked to explore the relationship between a set of variables and a miles per gallon (MPG) outcome for a set of cars. With some exploratory analysis and some regression modeling, I seek to answer the questions: 1. "Is an automatic or manual transmission better for MPG?" 2. "Quantify the MPG difference between automatic and manual transmissions" We will use the mtcars data from package "datasets." Taking a look at the variables included in the datasets, we will then run some exploratory data analyses to identify which variables affect MPG most significantly.

```
library(datasets)
data(mtcars)
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.2
```

Question 1: Is automatic or manual transmission better for MPG?

Start by doign a quick boxplot showing the distribution of MPG data across the two groups, automatic (am = 0) and manual (am = 1):

```
boxplot(mpg~am, data = mtcars, main = "MPG by transmission type", xlab =
"Transmission Type 0 = auto, 1 = manual", ylab = "MPG")
```

(See Figure 1 in Appendix)

We see quickly that the median, 25th, and 75th quantiles for manual tranmsission vehicles are significantly higher than those for automatic tranmission vehicles. It seems likely that manual transmission cars are better for MPG. However it is possible that other variables are confounding this relationship - weight seems like an obvious candidate. Plot the relationship between MPG and weight, showing automatic transmission vehicles and manual transmission vehicles in diffferent colors.

```
g1 <- ggplot(mtcars, aes(x=wt, y=mpg, color=am))
g1 <- g1 + geom_point()
g1
```

(See Figure 2 in Appendix)

We can now see that manual transmission vehicles tend to weigh less; so weight is a confounding variable with transmission type. In fact, there is a strong correlation between vehicle weight and MPG. Now let's answer the question. We can set up a linear regression model that models mpg as a function of transmission type. We should also include other

regressors - especially weight - and see if, after this adjustment, the transmission type remains significant in predicting mpg.

```
fitMPG1 <- lm(mpg ~ am + 1, data = mtcars)
summary(fitMPG1)$coef

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04

fitMPG2 <- lm(mpg ~ am + wt + 1, data = mtcars)
summary(fitMPG2)$coef

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

If we model mpg as predicted by transmission type alone, we see that our expected mpg for automatic is ~17mpg, while our expected mpg for manual = 17 + 7.25 = 24.25mpg. However, after adjusting the model to include weight, we see that the transmission type drops to close to 0 (-.024). This means that holding weight constant and changing the transmission type does not have a significant effect on the mpg.

Therefore, manual transmission is better for MPG, but this relationship is mostly explained by the confounding of variable wt (weight).

Question 2: Quantify the MPG difference between automatic and manual transmissions

The goal here is to select a model that reasonably models the relationship between mpg and several variables in the dataset. Let's include a few more covariates in the regression and use ANOVA to determine whether the variables are significant:

```
fitMPG3 <- lm(mpg ~ am + wt + hp + 1, data = mtcars)
fitMPG4 <- lm(mpg ~ am + wt + hp + disp + 1, data = mtcars)
fitMPG5 <- lm(mpg ~ am + wt + hp + disp + cyl + 1, data = mtcars)
anova(fitMPG1, fitMPG2, fitMPG3, fitMPG4, fitMPG5)

## Analysis of Variance Table
##
## Model 1: mpg ~ am + 1
## Model 2: mpg ~ am + wt + 1
## Model 3: mpg ~ am + wt + hp + 1
## Model 4: mpg ~ am + wt + hp + disp + 1
## Model 5: mpg ~ am + wt + hp + disp + cyl + 1
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.5432 7.017e-09 ***
## 3     28 180.29  1     98.03 15.6250 0.0005286 ***
## 4     27 179.91  1      0.38  0.0611 0.8066730
```

```
## 5     26 163.12  1     16.79  2.6758 0.1139322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the wt (weight) and hp (horsepower) coefficient are sigifnicant (that is, we can say with 95% confidence that they are nonzero), but the cylinder and displacement coefficients are not.

Let's try one more thing and include an interaction term in our model between weight and transmission type. Check the ANOVA value and plot the residuals diagnostics. Then, review the model summary.

```
fitMPG6 <- lm(mpg ~ am + wt + hp + am*wt + 1, data = mtcars)
anova(fitMPG3, fitMPG6)

## Analysis of Variance Table
##
## Model 1: mpg ~ am + wt + hp + 1
## Model 2: mpg ~ am + wt + hp + am * wt + 1
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 180.29
## 2     27 146.84  1    33.446 6.1496 0.01968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(fitMPG6)

summary(fitMPG6)

##
## Call:
## lm(formula = mpg ~ am + wt + hp + am * wt + 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0639 -1.3315 -0.9347  1.2180  5.0822
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.947333   2.723411  11.363 8.55e-12 ***
## am          11.554813   4.023277   2.872  0.00784 **
## wt          -2.515586   0.844497  -2.979  0.00605 **
## hp          -0.026949   0.009796  -2.751  0.01048 *
## am:wt       -3.577910   1.442796  -2.480  0.01968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.332 on 27 degrees of freedom
## Multiple R-squared:  0.8696, Adjusted R-squared:  0.8503
## F-statistic: 45.01 on 4 and 27 DF,  p-value: 1.451e-11
```

(See Figure 3 in Appendix)

Our model fitMPG6 is a reasonable model to proceed with and explains about 85% of the variation in the dataset according to the adjusted R squared value. One inference we can make is that, all other variables being equal and unchanged, a vehicle with manual tranmission is expected to get about 11.5MPG more than an identical automatic transmission car (p = .01). By including the interaction term (significant with p = .05), we adjust the slope of the weight covariate depending on whether the vehicle is manual or automatic tranmission.

Looking at residuals it's clear that a couple points exert high leverage and have high residuals, therefore have unusually high effect on model fit. Let's identify these points:

```
dffits(fitMPG6)
```

```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710
##         -0.42527466         -0.24659491         -0.43614248
##      Hornet 4 Drive   Hornet Sportabout             Valiant
##          0.21210291          0.14449790         -0.18420096
##          Duster 360           Merc 240D            Merc 230
##         -0.26343393          0.69326813          0.37610673
##            Merc 280           Merc 280C           Merc 450SE
##          0.02666812         -0.14293578          0.06033118
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood
##          0.06500131         -0.15199579         -0.54420856
## Lincoln Continental   Chrysler Imperial            Fiat 128
##         -0.38343445          1.11784600          0.89683620
##         Honda Civic      Toyota Corolla       Toyota Corona
##         -0.21735923          0.89930286         -0.15403051
##    Dodge Challenger         AMC Javelin          Camaro Z28
##         -0.28200713         -0.35909397         -0.29173869
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2
##          0.28404536         -0.28433832         -0.14394379
##        Lotus Europa      Ford Pantera L        Ferrari Dino
##          0.06870801         -0.07670818         -0.18638708
##       Maserati Bora          Volvo 142E
##          1.99978530         -0.25516519
```

It's clear that the Maserati Bora datapoint exhibits high leverage and has a high residual - we may want to check the confirm the data values at this row.

Appendix

Figure 1

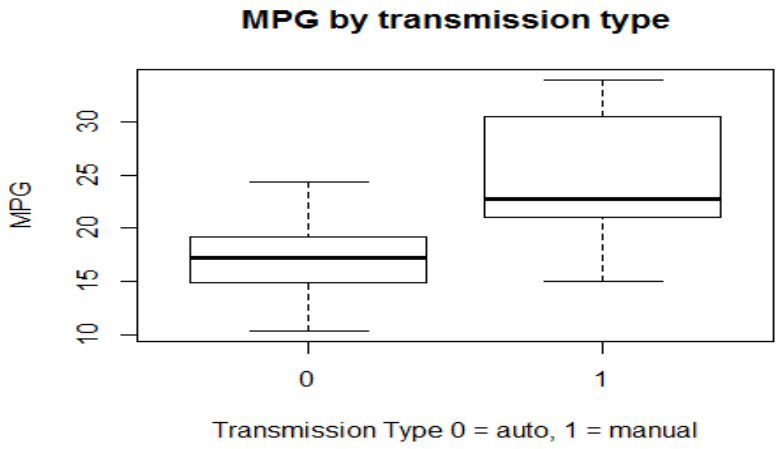


Figure 2

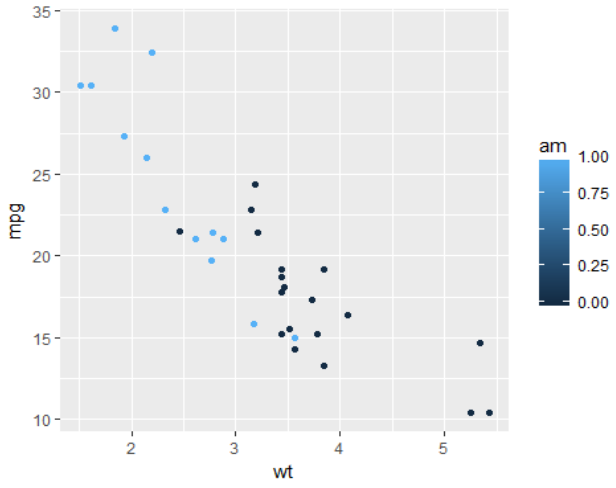
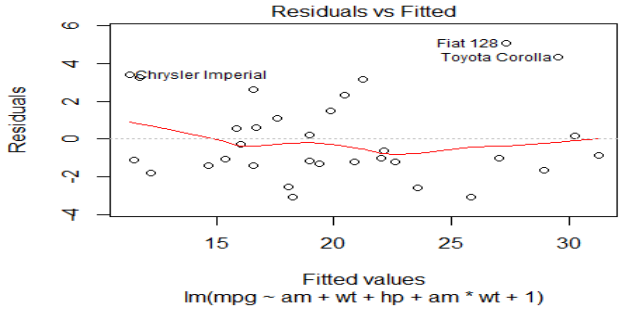
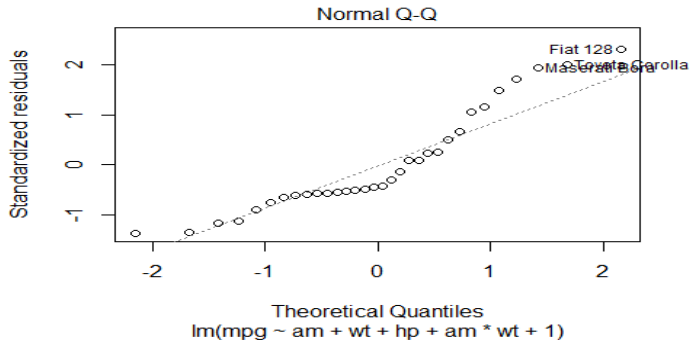
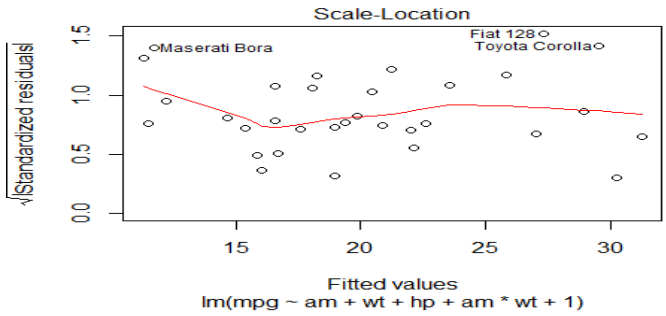
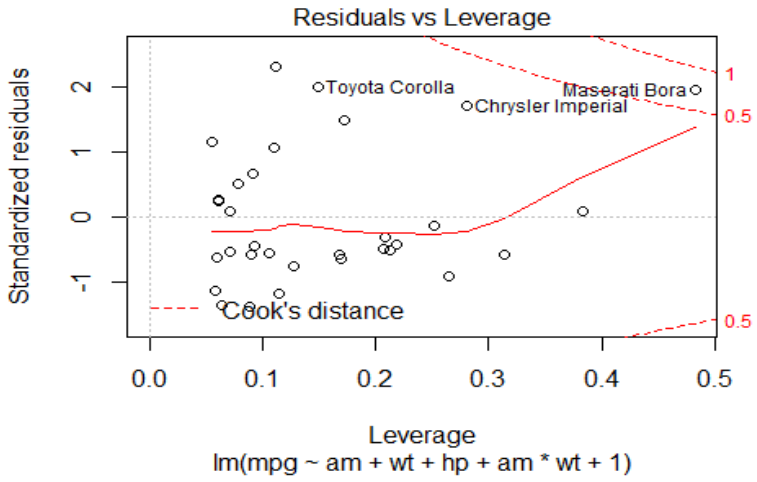


Figure 3




Normal Q-Q
Standardized residuals
Fiat 128
Toyota Corolla
Maserati Bora
-2
-1
0
1
2
Theoretical Quantiles
lm(mpg ~ am + wt + hp + am * wt + 1)


Scale-Location
√|Standardized residuals|
Maserati Bora
Fiat 128
Toyota Corolla
0.0
0.5
1.0
1.5
15
20
25
30
Fitted values
lm(mpg ~ am + wt + hp + am * wt + 1)


Residuals vs Leverage
Standardized residuals
Toyota Corolla
Maserati Bora
Chrysler Imperial
Cook's distance
1
0.5
0.5
-1
0
1
2
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
lm(mpg ~ am + wt + hp + am * wt + 1)