Regression Models Course Project

Erik Johnson

11/9/2017

Synopsis

We analyzed the mtcars dataset from the R data package to determine whether automatic or manual transmission cars are better for miles per gallon (mpg). We hypothesized that manual cars would have on average better gas mileage than automatic cars. However, we found no significant mpg difference between automatic and manual when controlling for weight and horsepower. Indeed, we found that weight and horsepower are the most significant predictors of mpg.

Exploratory Data Analysis

We start by loading the mtcars dataset from the R data package. We add a new column, amf, that converts the transmission type (am) to a factor.

```
library(ggplot2); library(dplyr);
data(mtcars)
mtcars$amf <- factor(mtcars$am, levels=c(0,1), labels=c("automatic", "manual"))
```

Next, we plot mpg vs transmission type (see Figure 1 in appendix). We clearly see that manual transmission cars on average get better mpg than automatic cars (24.39 vs 17.15, respectively), and there is no overlap at the 95% confidence level.

However, the picture gets murkier if we consider other variables. We see in Figure 2 that manual transmission cars tend to cluster in the lower horsepower (hp) range, and lower hp cars tend to get lower mpg. Additionally, we see in Figure 3 that automatic cars tend to be heavier than manuals, and heavier cars tend to get lower mpg.

This tells us that we need to take confounding variables into account when selecting our model.

Model Selection

We start with a linear fit of miles per gallon (mpg) as the outcome and transmission type (amf) as a factor predictor, mpg ~ amf (fit1). Next, we try adding additional confounding variables. As we saw before, horsepower (hp) and weight (wt) are possible candidates. We create a second fit, fit2, by adding hp as a predictor to the model (mpg ~ amf + hp). Then, we create a third fit, fit3, by adding wt as a third predictor (mpg ~ amf + hp + wt).

```
fit1 <- lm(mpg ~ amf, data=mtcars)
fit2 <- lm(mpg ~ amf + hp, data=mtcars)
fit3 <- lm(mpg ~ amf + hp + wt, data=mtcars)
```

Using anova, we find that fit2 is better than fit1, and fit3 is better than fit2 ($\Pr(>F) < 0.05$ for each).

```
fitAnova <- anova(fit1, fit2, fit3); fitAnova
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ amf
## Model 2: mpg ~ amf + hp
```

```
## Model 3: mpg ~ amf + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 73.841 2.445e-09 ***
## 3     28 180.29  1     65.15 10.118  0.003574 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We also find that including any other predictors results in an anova Pr(>F)-value above 0.05, so we conclude that fit3 is the best linear model to use.

Residuals

We investigate the residuals of our model (fit3). In Figure 4, we plot Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage. We find that the residuals don't exhibit any discernable pattern and the errors seem to be normally distributed.

In Figure 5, we also plot histograms of the dfbetas and hatvalues to show the leverage and power of the points in the data, respectively. We see that the betas and hatvalues are mostly clustered near zero with no major outliers, showing that there aren't any high-leverage or high-power points skewing the data.

We conclude that our model is a good fit for the data.

Conclusion

We finish by looking at the summary of our model.

```
fit3coeffs <- summary(fit3)$coefficients; fit3coeffs
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## amfmanual    2.08371013 1.376420152  1.513862 1.412682e-01
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
```

```
fit3confint <- confint(fit3); fit3confint
```

```
##                    2.5 %      97.5 %
## (Intercept) 28.58963286 39.41611738
## amfmanual   -0.73575874  4.90317900
## hp          -0.05715454 -0.01780291
## wt          -4.73232353 -1.02482730
```

We see that the beta value for going from an automatic transmission to a manual transmission while keeping other variables constant predicts a 2.08 increase in mpg. However, the p-value for this coefficient is $0.14127 > 0.05$, indicating that we can't reject the null hypothesis that it's equal to zero. Indeed, looking at the 95% confidence interval for that coefficient, we see that it contains zero.

On the other hand, the coefficients for hp and wt both have $p < 0.05$, and both predict a decrease in mpg for every unit increase in hp and wt.

We therefore conclude that hp and wt are significant predictors of mpg. On the other hand, there is no significant difference between manual and automatic transmission cars in predicting mpg, when hp and weight are held constant.

Appendix

```
fig1 <- ggplot(data=mtcars, mapping=aes(x = amf, y = mpg)) +
    geom_boxplot() +
    labs(x = "Transmission Type", y = "Miles Per Gallon",
         title = "Miles Per Gallon vs Transmission Type")
print(fig1)
```

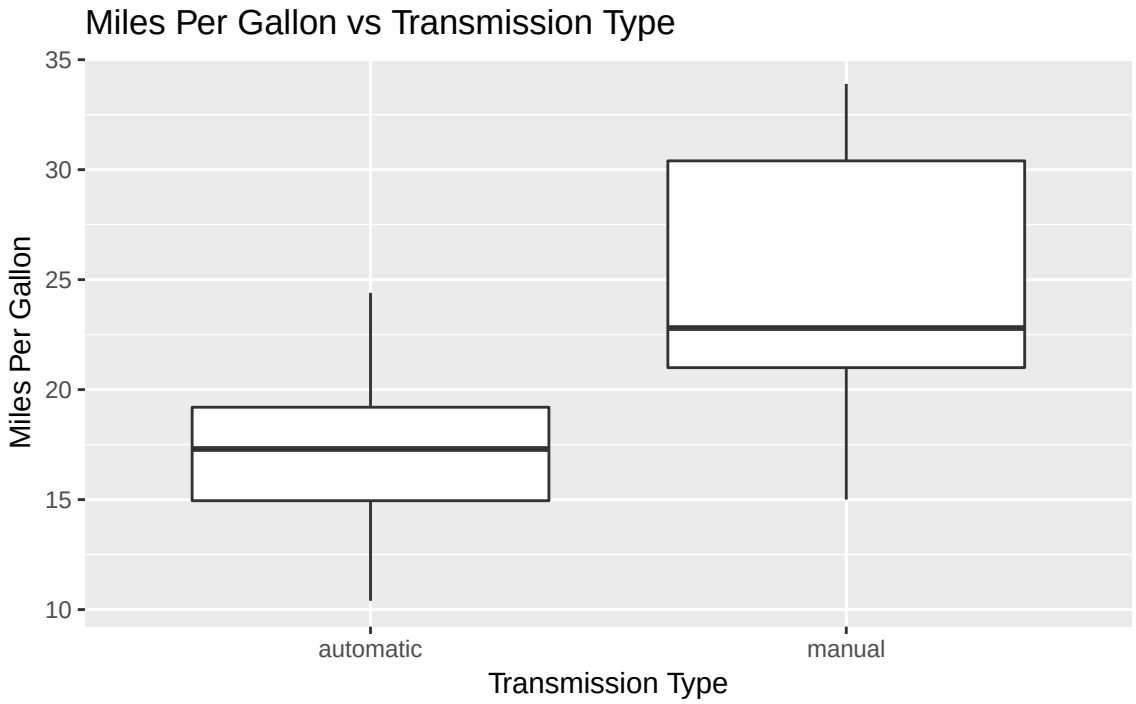

Figure 1: We see that manual vehicles appear to get better mpg than automatic vehicles.

```
fig2 <- ggplot(data=mtcars, mapping=aes(x = hp, y = mpg, color = amf)) +
    geom_point() +
    labs(x = "Horsepower", y = "Miles Per Gallon",
        title = "Miles Per Gallon vs Horsepower", color = "Transmission Type")
print(fig2)
```

```
fig3 <- ggplot(data=mtcars, mapping=aes(x = wt, y = mpg, color = amf)) +
    geom_point() +
    labs(x = "Weight (1000 lbs)", y = "Miles Per Gallon",
        title = "Miles Per Gallon vs Weight", color = "Transmission Type")
print(fig3)
```

```
par(mfrow = c(2, 2)); plot(fit3)
```

```
par(mfrow = c(1, 2)); hist(dfbetas(fit3)[,2]); hist(hatvalues(fit3))
```

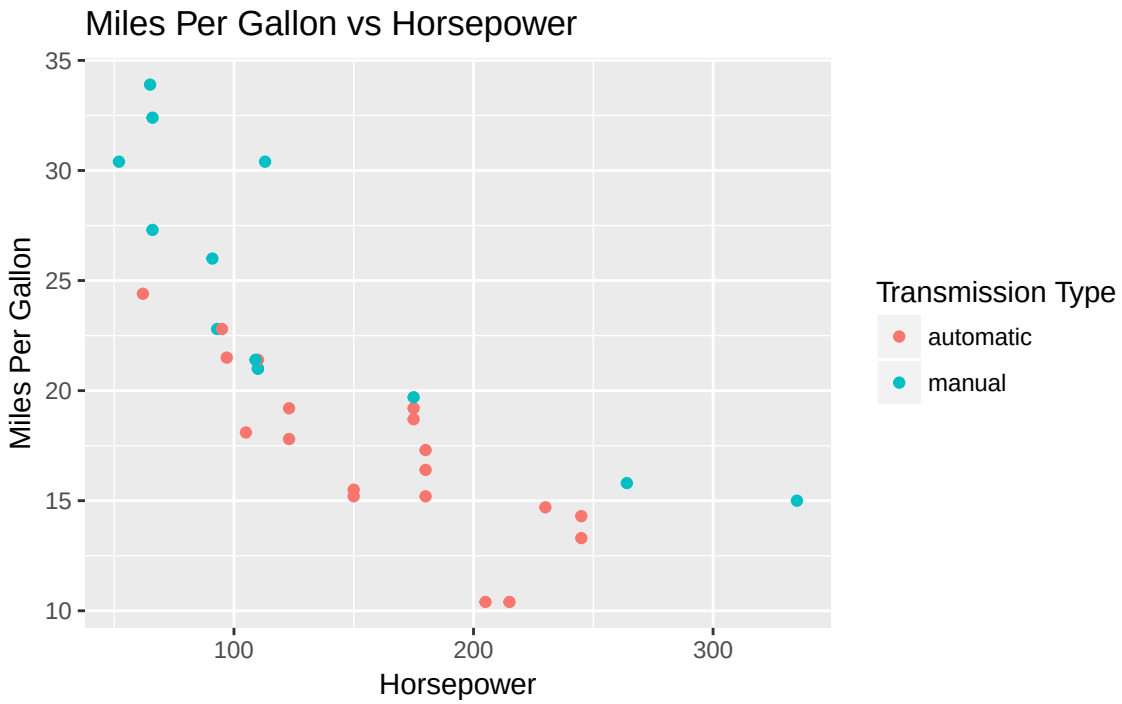

Figure 2: We see that manual vehicles tend to have lower horsepower than automatic vehicles, and lower horsepower vehicles tend to get better mpg. Therefore, hp is a possible confounder.

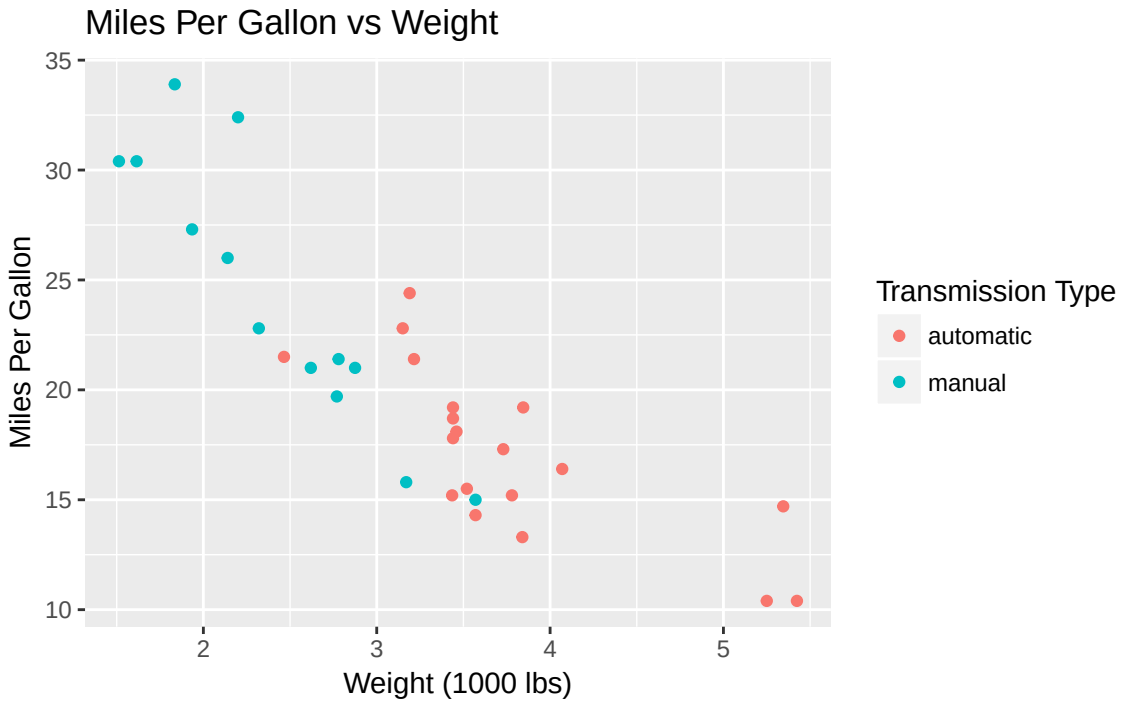

Figure 3: We see that automatic vehicles tend to be heavier than manuals, and heavier vehicles get worse mpg. Therefore, weight is a possible confounder.

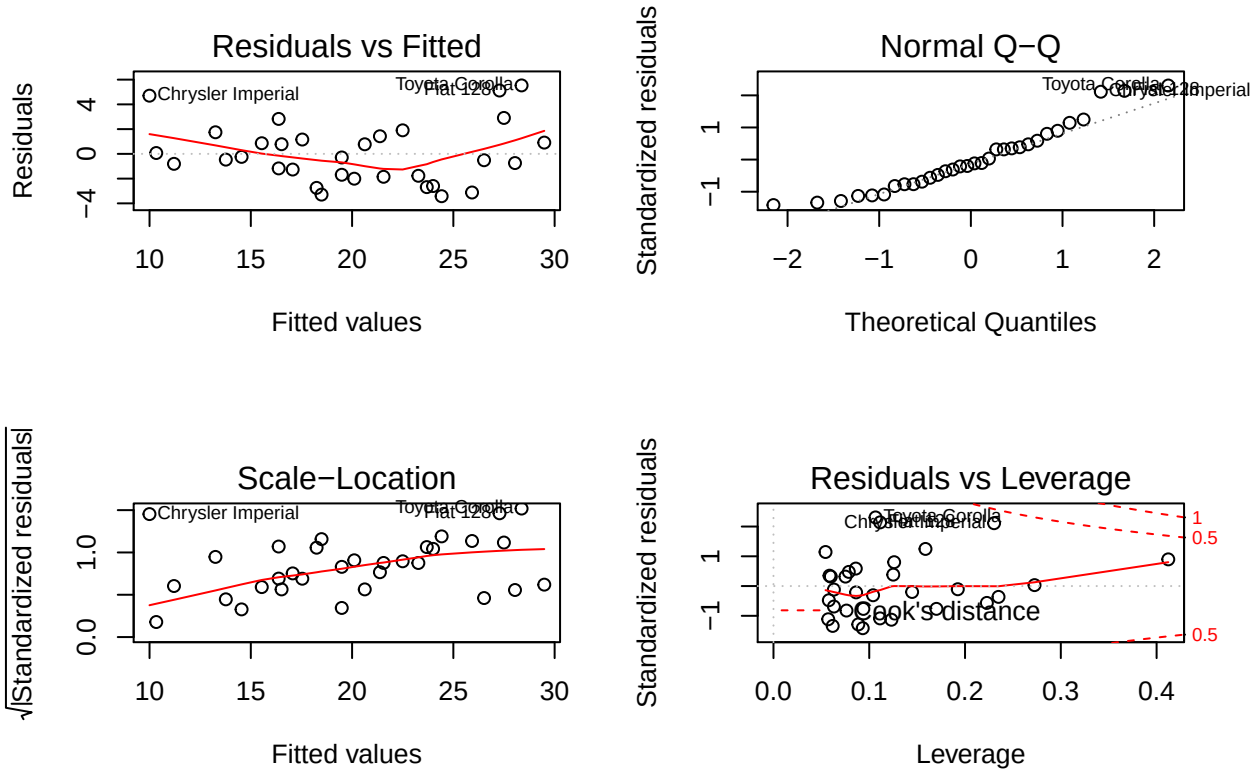

Figure 4: Residual plots showing that the errors are normally distributed and don't have discernable patterns.

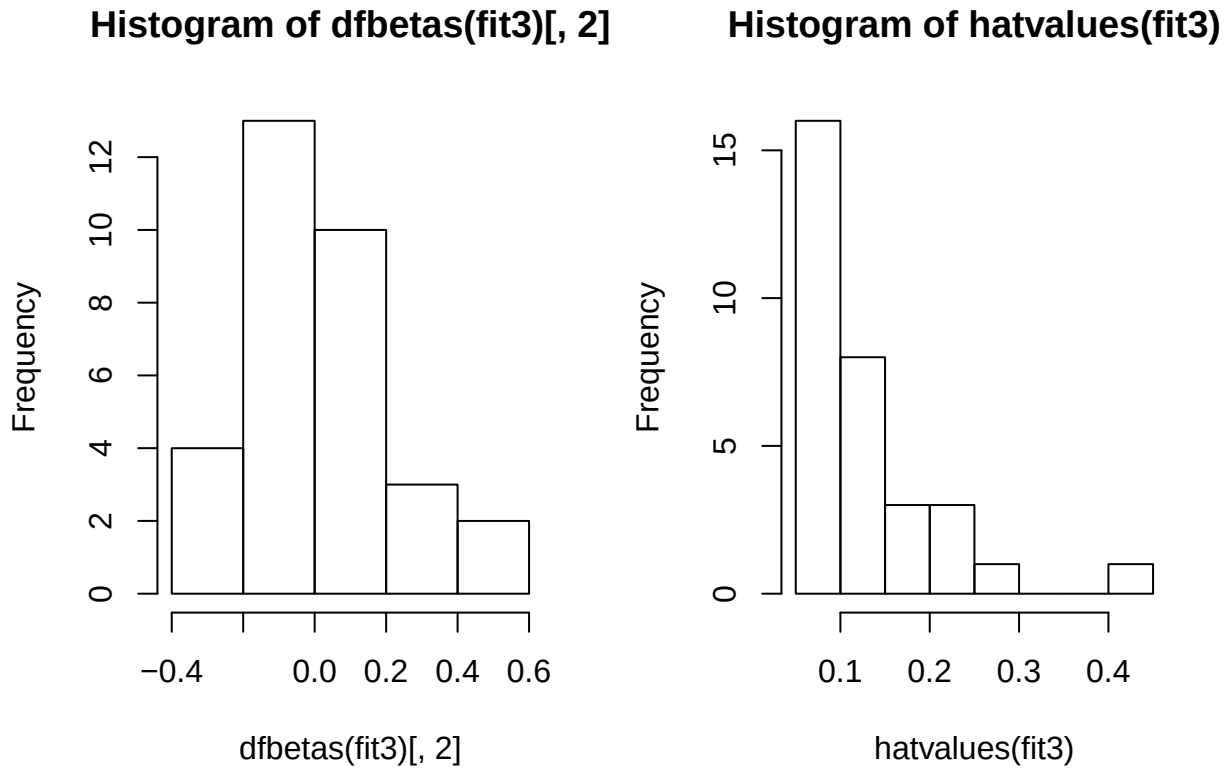

Figure 5: Histogram plots of dfbetas and residual hat values showing that there aren't any outliers with high leverage or power.