Regression Models Project

Dylan Peters

October 22, 2015

Executive Summary

This is an analysis of the *mtcars* dataset to determine what the relationship is between MPG and a car's attributes. We use R and a simple linear regression model to determine:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Exploratory Data Analysis

A quick look at the data shows that the transmission variable (am) has a big impact on MPG. See Figure 1 for a boxplot of the MPG by am (automatic = 0, manual = 1). The mean MPG for automatic transmission is 17.1473684 and for manual it is 24.3923077.

Initial Model

When we do a linear regression on the two variables, the coefficients reflect the difference in mean MPG between the two values of am.

```
fit1 <- lm(mpg ~ am, data=mtcars)
summary(fit1)$coefficients
```

```
##              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
```

See Figure 2 for a plot of the model with a line to show the regression model. Also see Figure 3 for a comparison of the residuals of this model against the predicted values (there are only two predicted values since the variable only has two values).

Further analysis

However, when we look at the am variable along with other variables, the picture changes. When we combine it with the wt variable (vehicle weight in 1000's of pounds), the relationship basically vanishes.

```
fit2 <- lm(formula = mpg ~ am + wt, data = mtcars)
summary(fit2)$coefficients
```

```
##                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
```

Now the coefficient for am is only -0.0236152, which is much smaller than the standard error for the variable in the model (1.5456453). Also, the sum of the squared residuals is 720.8965992 for the first model and 278.3196972 for the second model, indicating that the second model is a much better fit.

We can examine the influence of am vs wt by looking at the analysis of variance to determine which has the greatest impact on the outcome:

```
fitwt <- lm(mpg ~ wt, data=mtcars)
tab1 <- anova(fit1, fit2); tab2 <- anova(fitwt, fit2)
tab1[2,6]
```

```
## [1] 1.867415e-07
```

```
tab2[2,6]
```

```
## [1] 0.9879146
```

The first table shows that the second model (am+wt) greatly improves over the first model (am), with a very small p-value. When we compare a model with just weight to the second model, the p-value indicates that there is very little difference.

To get a clearer picture, we evaluate the data based on all the variables, to isolate any correlations.

```
fitall <- lm(formula = mpg ~ ., data = mtcars)
summary(fitall)$coefficients
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

This is the best model so far, with a sum of the squared residuals is 147.49443. Figure 4 shows the predicted values against the residuals, and the residuals are in a much narrower range. Figure 5 shows the actual MPG against the predicted values, with a line with slope one to show that the values correlate.

Now we can see that wt is the biggest factor in MPG, with am being the second biggest. However, the effects are not very strong. The wt variable has a p-value of 0.0632522, which is not quite significant enough to satisfy a $p < 0.05$ test. The p-value for the am variable is worse: 0.2339897. So while these variables are the most influential, their influence is not very strong.

Conclusion

The variables that affect MPG are all correlated with each other, making it difficult to isolate a single best one. At best, we can say that a manual transmission vehicle might have a slight benefit over an automatic transmission vehicle.

Appendix

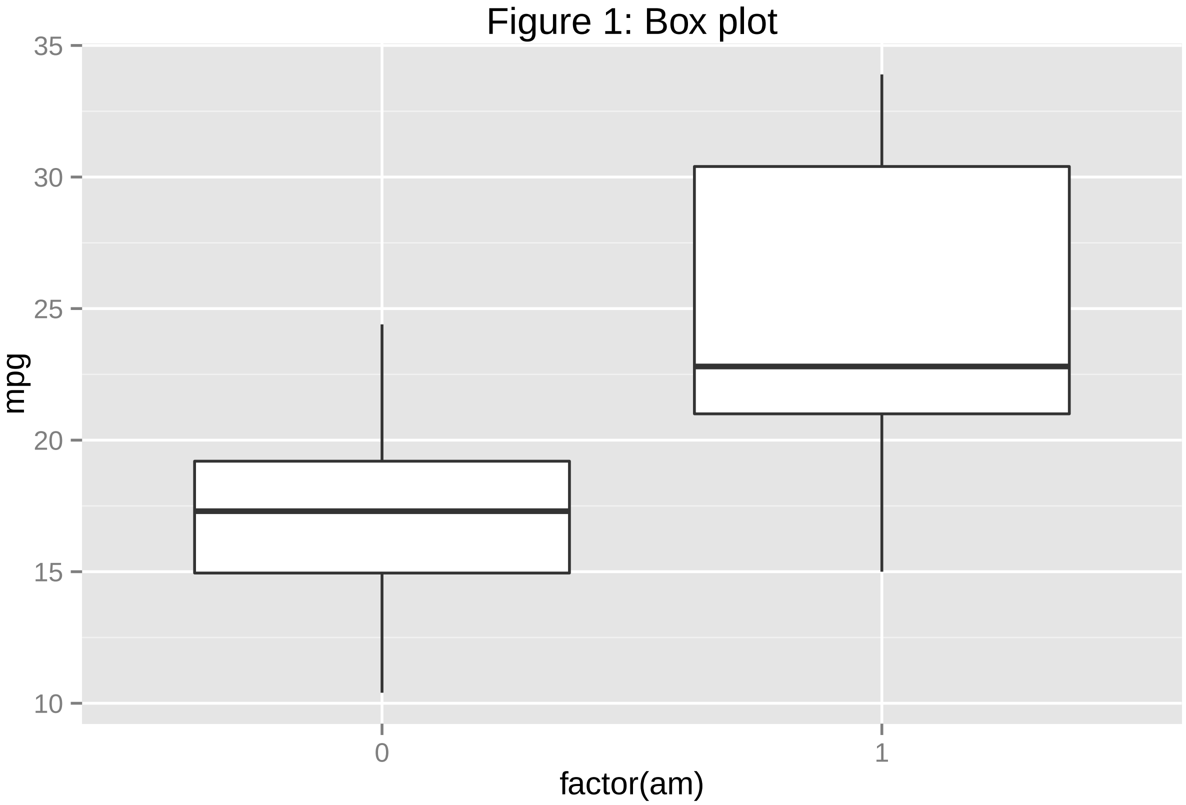


Figure 1: Box plot
35
30
25
20
15
10
mpg
0
1
factor(am)

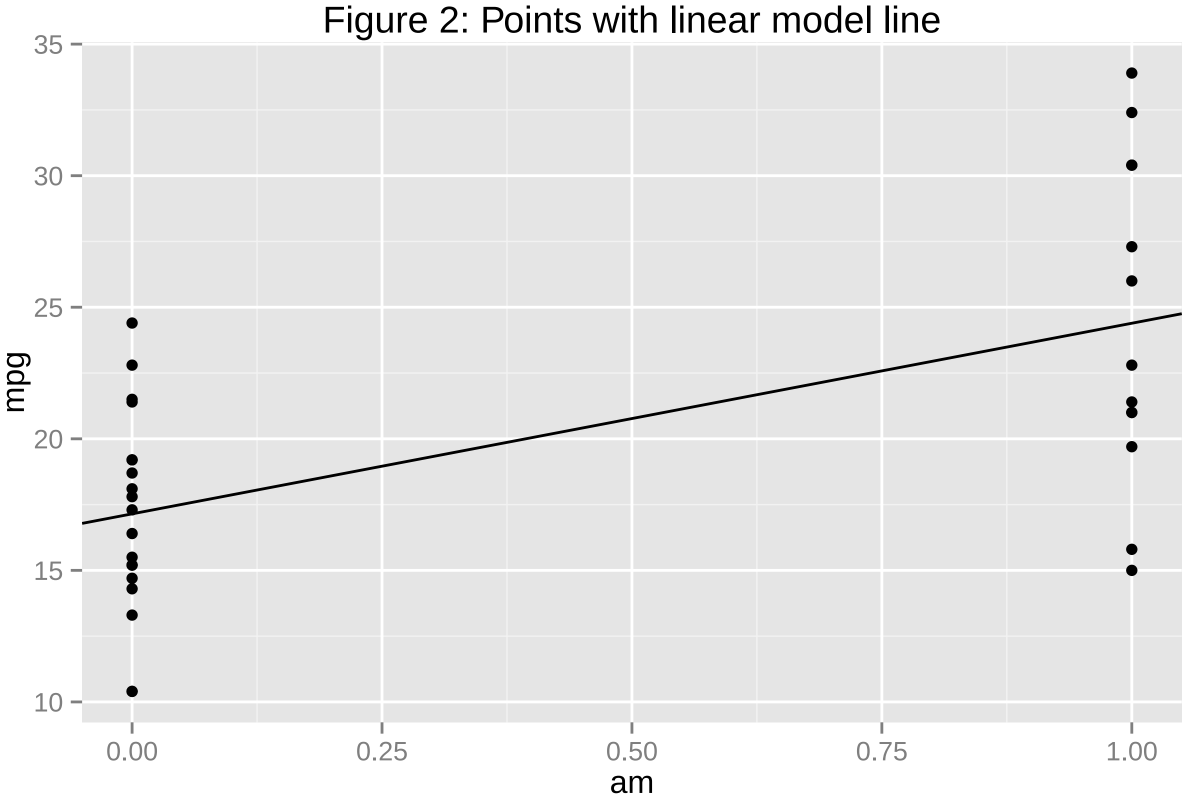


Figure 2: Points with linear model line
35
30
25
20
15
10
mpg
0.00
0.25
0.50
0.75
1.00
am

Figure 3: Model 1 Predicted fit vs residuals

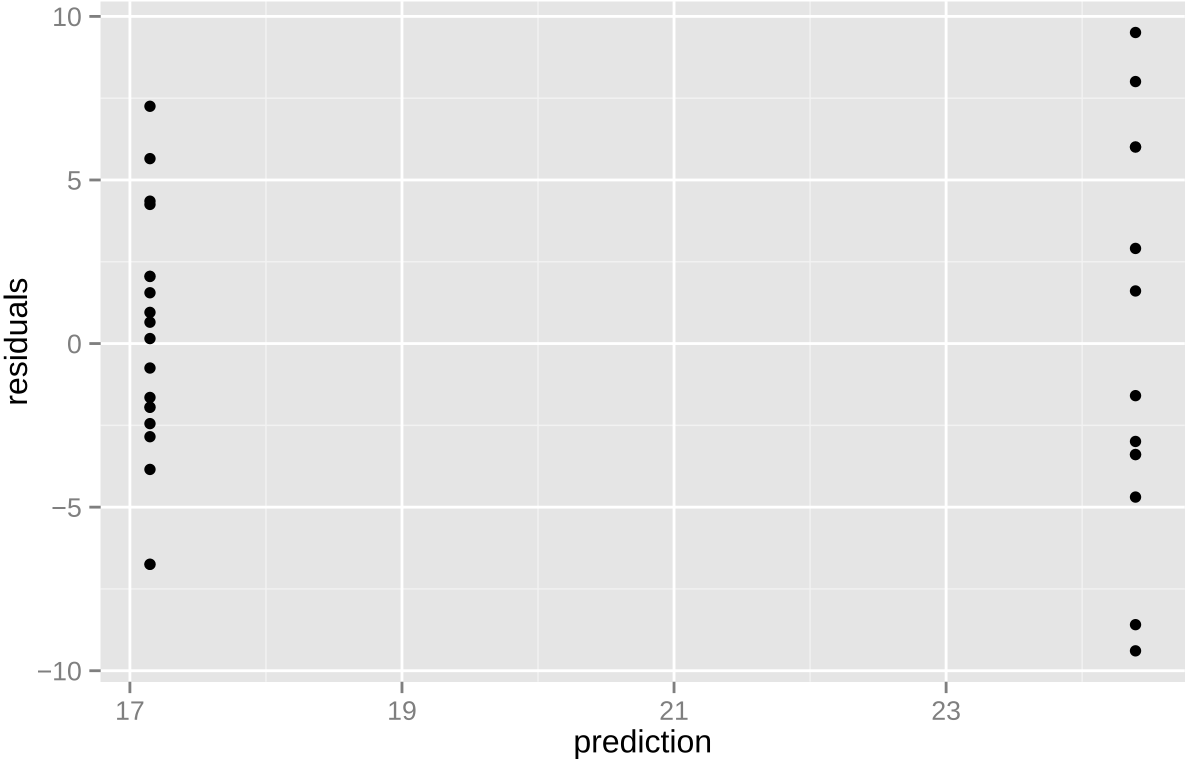


Figure 4: Final Predicted fit vs residuals

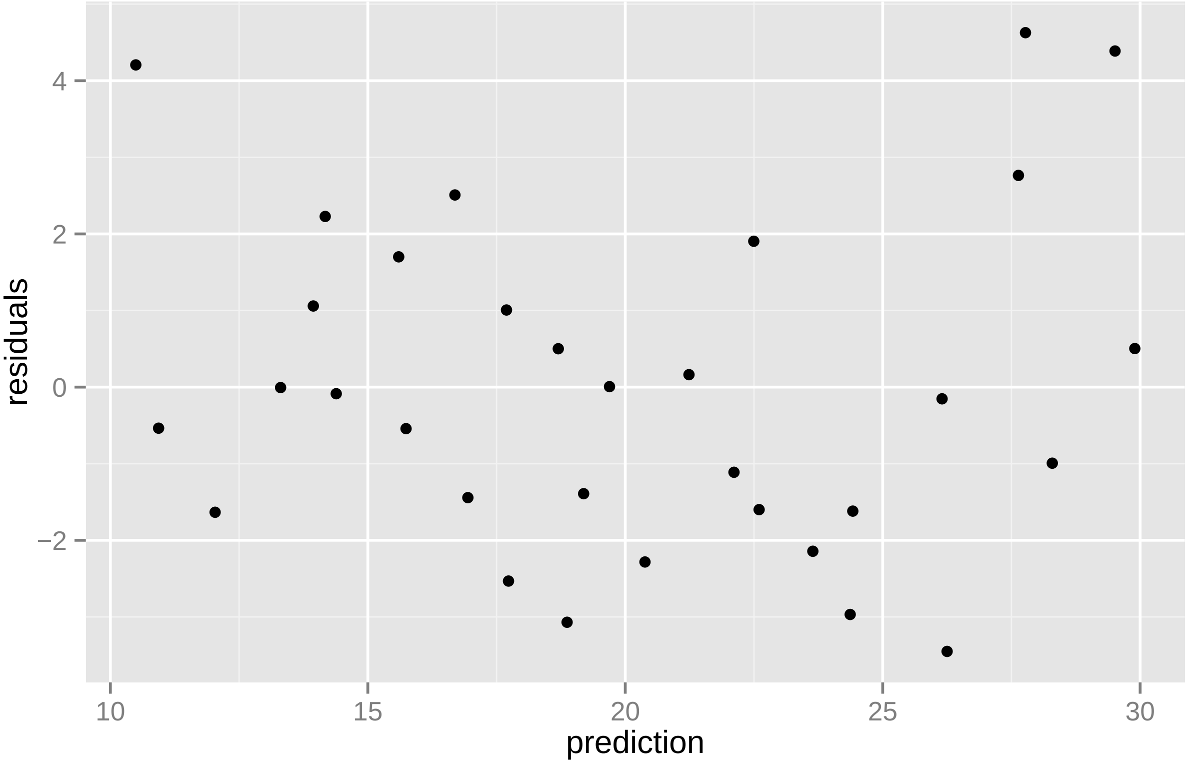
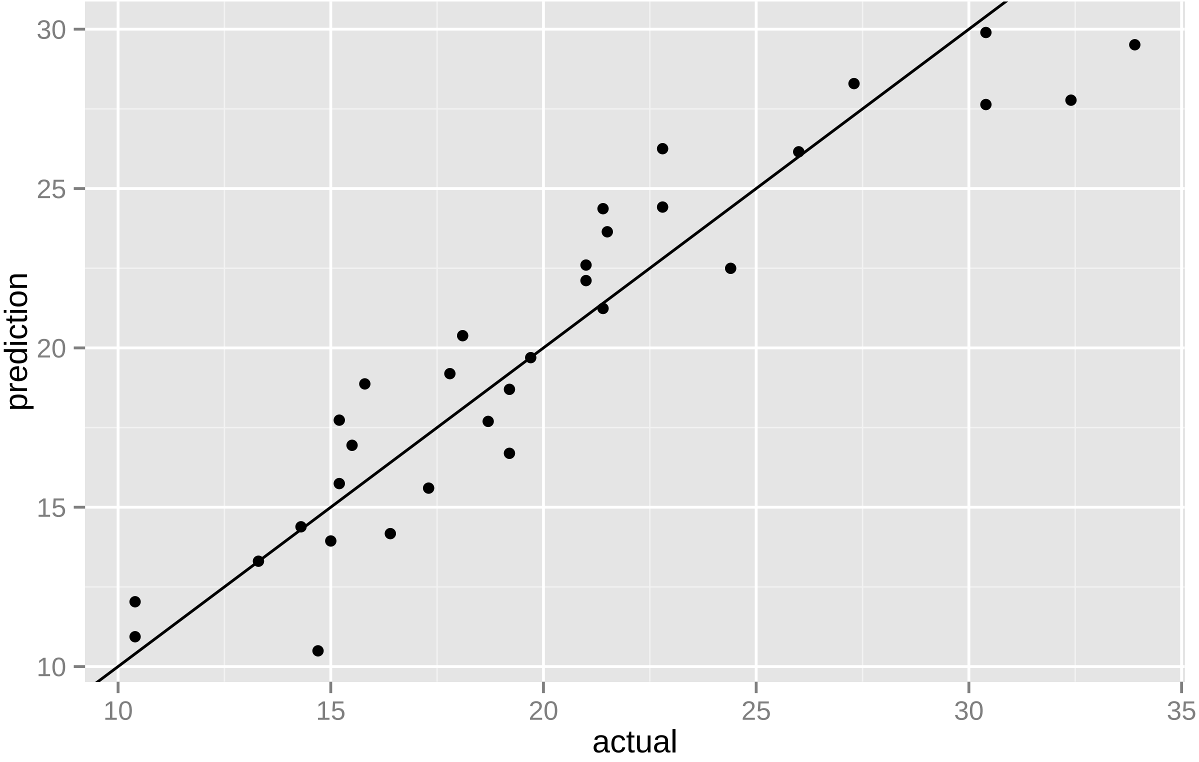


Figure 5: Final Actual value vs Predicted fit


prediction
actual