Regression Models: Course Project

Impact of Transmission Type on Automobile Mileage (USA, 1974)

by David Modjeska

Executive Summary

This report analyzes the relationship between automobile transmission type and gas mileage. The goal was to identify which transmission type (automatic or manual) was associated with higher mileage, as well as the difference in expected mileage for each transmission type. To perform this analysis, we used data from the US magazine *Motor Trend* on aspects of automobile design and performance for 1973-74 automobile models. Analysis determined that manual transmissions had an expected higher mileage than did automatic transmissions. Moreover, the expected difference in mileage was approximately 11.57 miles per gallon.

Reading and cleaning the cars data

To perform this analysis, I used data from the US magazine *Motor Trend* on aspects of automobile design and performance for 1973-74 automobile models. Conveniently, this data is available as the R "mtcars" dataset that is built into R. After loading this small dataset into R, the only necessary cleaning was to rename the 'am' variable to "man" for legibility, and to indicate the "man" and "cyl" variables as categorical.

Exploratory data analysis

The dataset contains records for 32 vehicles. The variables of key interest were fuel consumption (in MPG), number of cylinders (4, 6, or 8), weight (in 1000 pounds), and transmission type (automatic or manual).

Through pairwise plotting and correlation analyses, we see that mileage correlates highly with displacement (-0.8476) and weight (-0.8677). Also, a boxplot shows a relationship between weight and number of cylinders. Cylinders are associated with displacement and horsepower, which may be redundant in our analysis. Finally, transmission is not clearly associated with other variables. (Please see the Appendix for these plots.)

Plotting just the key variables, we see patterns. First, transmission type seems to be associated with mileage, with manual transmissions having higher mileage. Second, cylinders seem to be associated with mileage, with more cylinders showing lower mileage. Third, weight seems to show an inverse relationship with mileage. Finally, manual transmissions seem to be associated with lighter cars. (These plots are in the Appendix.)

Fitting Multiple Models and Selection Strategy

Two questions motivated this analysis: Is an automatic or manual transmission better for MPG? How can we quantify the MPG difference between automatic and manual transmissions?

So my overall strategy for fitting models was to begin with the main variables of interest, transmission type and mileage. Then I added in related explanatory variables of interest, such as weight, number of cylinders, horsepower, and displacement, on a trial basis. After each addition of a new model variable in this nested style, I compared the models using ANOVA. Where an added variable showed a significant improvement, I retained it. The retained explanatory variables were weight, number of cylinders, and transmission type.

For completeness, I examined all potential interactions among the explanatory variables. The interaction between weight and transmission type was found to be significant, so it was added to the model. (Since the heaviest vehicles tend to have automatic transmissions, this interaction makes sense, informally speaking.)

Because of the two categorical variables (number of cylinders and transmission type) in the model, I removed the intercept term when using R's "lm()" function. To seek a parsimonious model, I set an alpha level of 0.01 for ANOVA comparisons between nested models.

Final model and interpreting coefficients

A summary of the final model is as follows. (A plot of this model can be found in the Appendix.)

```
##      cyl4      cyl6      cyl8        wt      man1    wt:man1
## 29.774836 27.065059 24.998726 -2.398713 11.568790 -4.067981
```

With an automatic transmission, the expected mileage for a car with a hypothetical weight of 0 pounds is 29.7748357 (for 4 cylinders), 27.0650588 (for 6 cylinders), and 24.9987258 (for 8 cylinders). With a manual transmission, the expected mileage for a car with a hypothetical weight of 0 pounds is 41.3436258 (for 4 cylinders), 38.6338488 (for 6 cylinders), and 36.5675159 (for 8 cylinders).

With an automatic transmission, for each additional 1000 pounds of weight in the vehicle, the expected mileage decreases by 2.3987132. With a manual transmission, for each additional 1000 pounds of weight in the vehicle, the expected mileage decreases by 6.4666938.

Residual Plot and Diagnostics

Before stepping back to answer the questions of interest, let's perform some residual plotting and diagnostics. Looking at the scaled and unscaled residuals plotted against the fitted values, the values look reasonable. There are no obvious patterns in the plot. Looking at a quantile-quantile plot for the residuals, they appear to have an approximately normal distribution. Finally, the plot of residuals by leverage looks reasonable, without any obvious patterns in the data. (These diagnostic plots can be found in the Appendix.)

In the plotting above, two vehicles have more extreme values: the Toyota Corolla and the Fiat 128. Analysis with R's "hatvars" function shows that these points lack exceptional leverage. R's "dfbetas" gives a similar result for influence. With R's "dffits" function, though, these two points have the strongest influence. The data seem reputable, so it's possible that these vehicles just have distinguishing mileage characteristics.

Answers to Questions of Interest

We are now in a position to answer the questions that motivated this analysis:

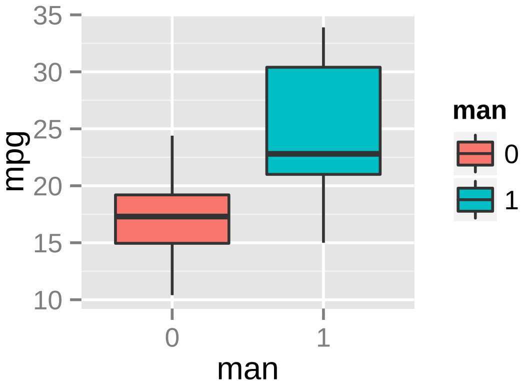
- A *manual* transmission is better for MPG
- The MPG difference between automatic and manual transmissions is approximately 11.57 MPG.

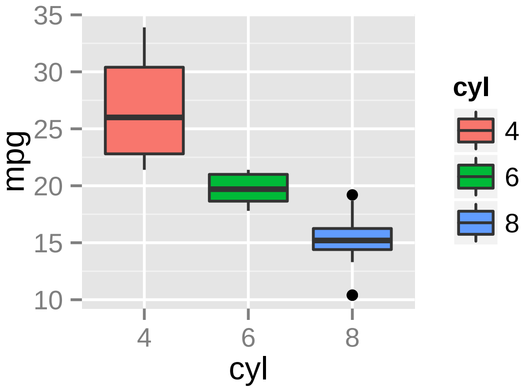
Inference and Uncertainty

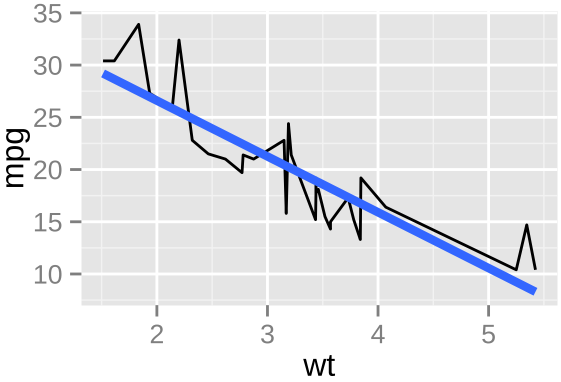
Let's now test the alternate hypothesis that manual transmissions *do* exhibit higher mileage than automatic transmissions. Our null hypothesis is this: there is no difference in expected mileage between transmission types. This is a right-tailed test. Let's set a conservative alpha level of 0.01 for this test. Using R's "summary()" function with the model above, the P-value for the difference in expected mileage with different transmissions is 0.0088538. This value is less than alpha, so we conlude our test by rejecting the null hypothesis. At the 0.01 level of significance, it appears that cars with manual transmissions *do* have higher mileage.

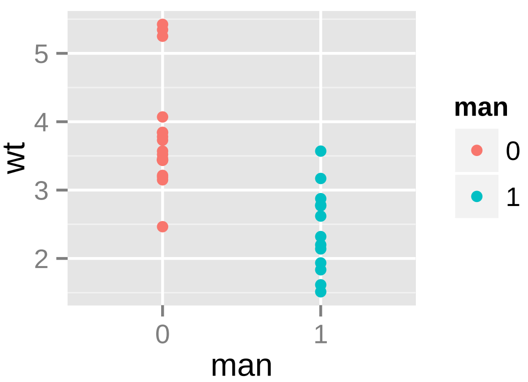
It is useful to create a confidence interval for the MPG difference discussed above. Using R's "summary()" and df.residual() functions with our model: the expected MPG difference (11.57), the standard error(4.0878), and the degrees of freedom (26). If we use a 95% confidence interval, a T-distribution gives a margin of error of 8.4026. So we can have 95% confidence that the true MPG difference lies between 3.17 and 19.97.

Appendix

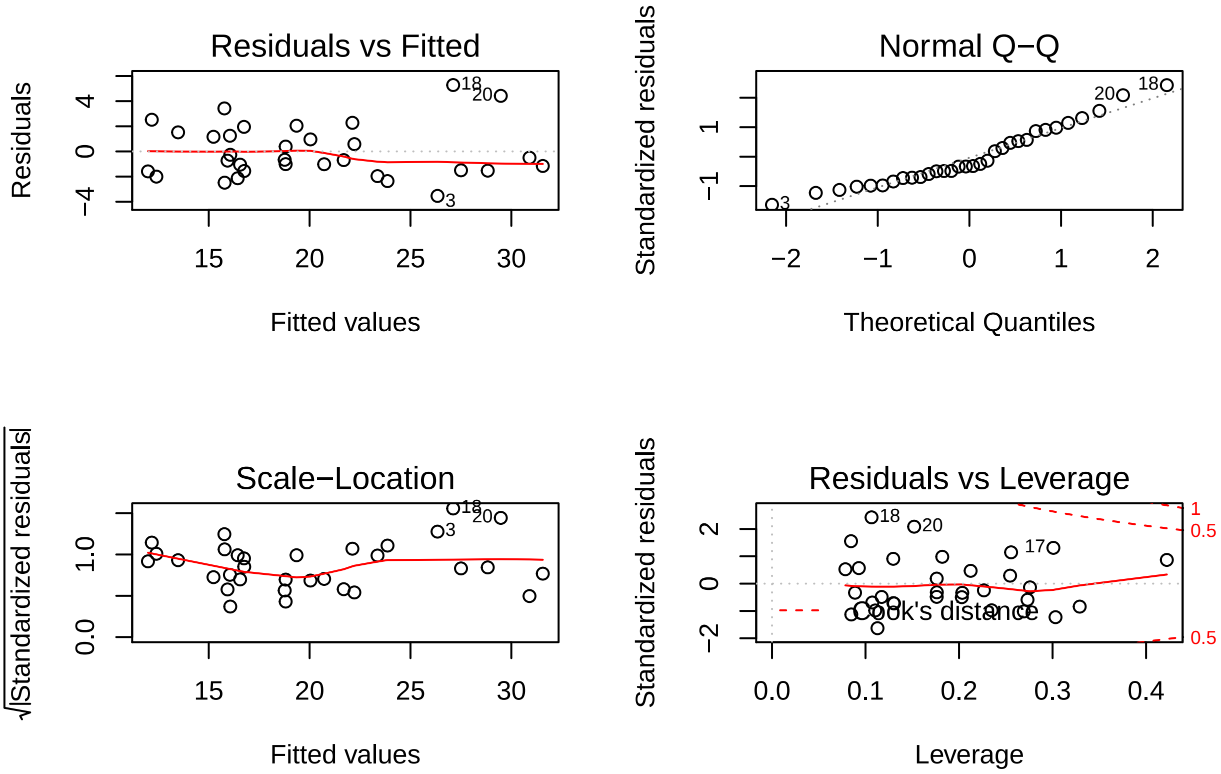

35
30
25
20
15
10
mpg
0
1
man
man
0
1

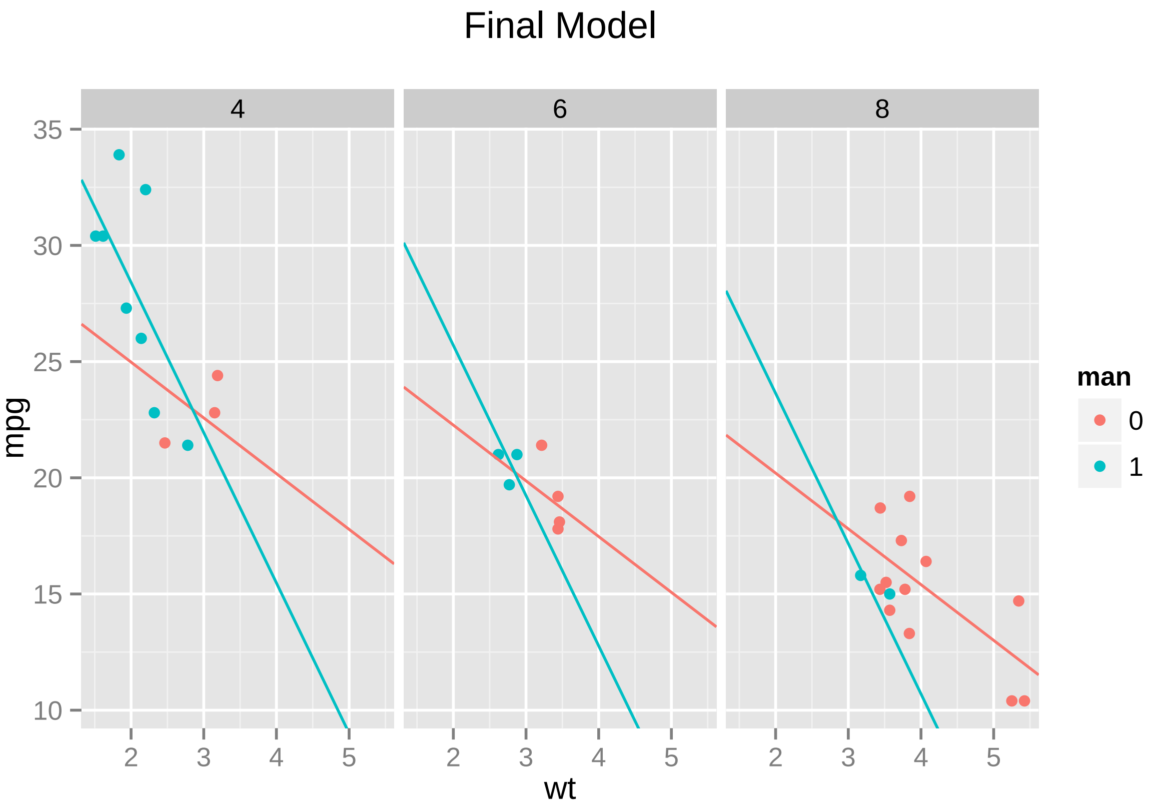

35
30
25
20
15
10
mpg
4
6
8
cyl
cyl
4
6
8


35
30
25
20
15
10
mpg
2
3
4
5
wt


5
4
3
2
wt
0
1
man
man
0
1

.


Residuals vs Fitted
Residuals
4
0
−4
15
20
25
30
Fitted values
18
20
3
Normal Q−Q
Standardized residuals
1
−1
−2
−1
0
1
2
Theoretical Quantiles
20
18
3
Scale−Location
√|Standardized residuals|
1.0
0.0
15
20
25
30
Fitted values
18
20
3
Residuals vs Leverage
Standardized residuals
2
0
−2
0.0
0.1
0.2
0.3
0.4
Leverage
18
20
17
Cook's distance
1
0.5
0.5



R Code

```
# read and clean data
library(datasets); library(ggplot2); library(GGally); library(dplyr)
data(mtcars)
mtcars1 <- mtcars %>% rename(man = am) %>% mutate(man = factor(man)) %>%    mutate(cyl = factor(cyl))

# explore data --------------------

str(mtcars1); summary(mtcars1);
pairs_plot <- ggpairs(mtcars1, params=list(size=3)) + theme(axis.text=element_text(size=3))

mpg_histo <- ggplot(mtcars1, aes(x = mpg)) +
    geom_histogram(aes(y = ..density..), binwidth = 2.5, colour = "white", fill = "gray50")
man_box <- ggplot(mtcars1, aes(x = man, y = mpg, fill = man)) + geom_boxplot()
cyl_box <- ggplot(mtcars1, aes(x = cyl, y = mpg, fill = cyl)) + geom_boxplot()
wt_line <- ggplot(mtcars1, aes(x = wt, y = mpg)) + geom_line() +
    geom_smooth(method = "lm", se = FALSE, lwd = 1.5)
man_wt_plot <- ggplot(mtcars1, aes(x = man, y = wt, color = man)) + geom_point()

# fit and test multiple models ------------------

fitM <- lm(data = mtcars1, mpg ~ man - 1); summary(fitM)

fit5 <- lm(data = mtcars1, mpg ~ wt + man - 1); summary(fit5)
anova(fitM, fit5) # difference is significant
```

```
fit6 <- lm(data = mtcars1, mpg ~ cyl + man - 1); summary(fit6)
anova(fitM, fit6) # difference is significant

fit7 <- lm(data = mtcars1, mpg ~ wt + cyl + man - 1); summary(fit7)
anova(fit5, fit7); anova(fit6, fit7) # differences are significant

fit8 <- lm(data = mtcars1, mpg ~ wt + disp + cyl + man - 1); summary(fit8)
anova(fit7, fit8) # difference not significant, cyl and disp highly correlated?

fit9 <- lm(data = mtcars1, mpg ~ wt + hp + cyl + man - 1); summary(fit9)
anova(fit7, fit9) # difference not significant, cyl and hp highly correlated?

fit10 <- lm(data = mtcars1, mpg ~ cyl + wt + man + wt:man - 1); summary(fit10)
anova(fit7, fit10) # difference is significant
# plot(fit10) # inspect model fit visually (later in .Rmd file)
hatvalues(fit10) # diagnose point leverages
dffits(fit10) # diagnose point influences on intercepts
dfbetas(fit10) # diagnose point influences on coefficients

fit11 <- lm(data = mtcars1, mpg ~ wt*cyl + man - 1); summary(fit11)
anova(fit7, fit11) # difference not significant

fit12 <- lm(data = mtcars1, mpg ~ wt + cyl*man - 1); summary(fit12)
anova(fit7, fit12) # difference not significant

# plot best model --------------------
coeff <- fit10$coeff
model_plot <- ggplot(mtcars1, aes(x = wt, y = mpg, color = man)) + geom_point() +
    geom_abline(data = subset(mtcars1, cyl =='4' & man == '0'), aes(color = man),
                intercept = coeff[1], slope = coeff[4]) +
    geom_abline(data = subset(mtcars1, cyl =='4' & man == '1'), aes(color = man),
                intercept = coeff[1] + coeff[5], slope = coeff[4] + coeff[6]) +
    geom_abline(data = subset(mtcars1, cyl =='6' & man == '0'), aes(color = man),
                intercept = coeff[2], slope = coeff[4]) +
    geom_abline(data = subset(mtcars1, cyl =='6' & man == '1'), aes(color = man),
                intercept = coeff[2] + coeff[5], slope = coeff[4] + coeff[6]) +
    geom_abline(data = subset(mtcars1, cyl =='8' & man == '0'), aes(color = man),
                intercept = coeff[3], slope = coeff[4]) +
    geom_abline(data = subset(mtcars1, cyl =='8' & man == '1'), aes(color = man),
                intercept = coeff[3] + coeff[5], slope = coeff[4] + coeff[6]) +
    facet_wrap(~ cyl, ncol = 3) + ggtitle("Final Model\n")

# calculate 95% confidence interval for best model -------------
man_mpg_diff <- fit10$coeff[5]
standard_error <- summary(fit10)$coefficients[5 , 2]
df <- df.residual(fit10)
error_margin <- standard_error * qt(0.975, df)
confint <- c(man_mpg_diff - error_margin, man_mpg_diff + error_margin)
```

R markdown file on GitHub

https://github.com/dmodjeska/RegMods_Project/blob/master/RegMods_Project_Modjeska.Rmd