Regression Models – Final Project

Darrell Gerber

5/17/2021

Executive Summary

The mtcars data set was analyzed to determine if the use of an automatic transmission impacts fuel efficiency, and, if so, by how much.

The average fuel-efficiency of cars with manual transmissions is 7.3 mpg better than those with automatic transmissions. An ANOVA analysis is performed to determine the best predictors in mtcars to include in a linear regression model. It was found that *cyl*, *hp*, and *wt* were important to include along with *am*, our predictor of interest. However, the lack of predictive significance for *am* in the model indicates unmodeled confounding factors that interfere with our attempts to determine the relationship between transmission-type and fuel-efficiency using the mtcars data set. It is likely that manual transmissions have higher fuel efficiency than automatic transmissions, but we need a better data set to properly quantify the difference.

Note to reviewers: All code, figures, and tables are included in the Appendix.

Exploratory Analysis

The distribution of fuel efficiency for cars with and without automatic transmission is shown in Figure 1. The average fuel efficiency for vehicles with manual transmission is clearly higher than that for the vehicles with automatic transmission. The average fuel-efficiency for cars with automatic transmissions is 17.1 mpg. The distribution of *mpg* for automatic transmissions appears compact and roughly normally distributed. The average fuel-efficiency for cars with manual transmissions is 24.4 mpg. The distribution of *mpg* for manual transmissions is dispersed and shows a slight two-lobed form.

Modeling

A series of linear models are created where each model incrementally adds an additional parameter from the mtcars data set. We can compare these nested models with ANOVA to determine which terms are necessary to include in the model.

The P-values in Table 1 test the likelihood that all of the added variables are zero. In other words, if the P-value is high for a model, the term added in that model is likely not necessary. The P-values for Model 2, Model 4 and Model 6 are below 5%, so we can conclude that the following terms are likely necessary to include in our model:

- *am* – Transmission type. The predictor of interest.
- *cyl* – Number of cylinders
- *hp* – Gross horsepower
- *wt* – Weight (1000 lbs)

$$mpg_i = \beta_0 + \beta_1 * am_i + \beta_2 * cyl_i + \beta_3 * hp_i + \beta_4 * wt_i + \epsilon_i$$

The final model is a good fit with an overall P-value that is nearly zero (Table 2). Comparing the residuals of the final selected model versus a linear model using transmission type (Figure 2) shows a decrease in residuals by adding additional predictors further indicating an improved fit by including *cyl*, *hp*, and *wt*.

However, the effect of transmission-type on the fuel efficiency appears to be minimal (about 1.5 additional mpg for a manual transmission versus an automatic transmission). The P-value for the *am* coefficient is very high, though, indicating there is likely over 30% chance the result is due to random and/or unmodeled factors.

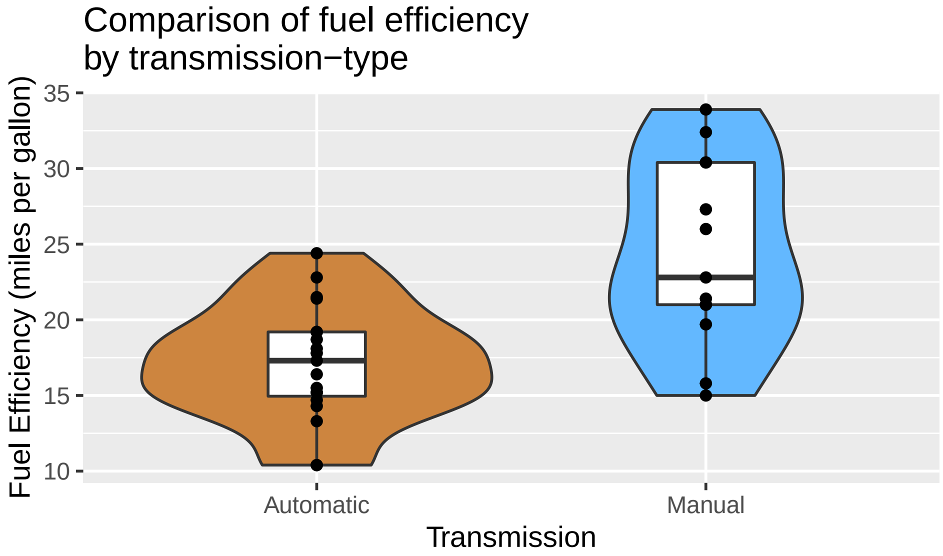
Conclusion

The simplest linear model (using transmission-type as the only predictor) indicates a positive relationship between the use of an manual transmission and fuel-efficiency. However, an ANOVA analysis indicates the need to include additional predictors in the model (cylinders, horsepower, and weight). The more complex model showed no significant effect on fuel-efficiency from transmission-type.

The lack of predictive significance in the complex model indicates unmodeled confounding factors that interfere with our attempts to determine the relationship between transmission-type and fuel-efficiency using the mtcars data set. A simple change in test design to remove the effect of confounding factors is A/B tests with the same model of car with and without an automatic transmission. It is likely that manual transmissions have higher fuel efficiency than automatic transmissions, but we need a better data set to properly quantify the difference.

Appendix

```
library(ggplot2)
library(gridExtra)
g1 <- ggplot(mtcars, aes(x=factor(am), y=mpg, fill=factor(am)) )
g1 <- g1 + geom_violin()
g1 <- g1 + geom_boxplot(width=.25, fill="white")
g1 <- g1 + geom_point()
g1 <- g1 + scale_fill_manual(values=c("tan3","steelblue1"), guide=FALSE)
g1 <- g1 + labs(title = "Comparison of fuel efficiency\nby transmission-type",
              x = "Transmission", y = "Fuel Efficiency (miles per gallon)")
g1 <- g1 + scale_x_discrete(labels = c("Automatic", "Manual"))
g1
```



```r
meanManual <- round(mean(mtcars[mtcars$am==1,]$mpg),1)
meanAuto <-  round(mean(mtcars[mtcars$am==0,]$mpg),1)
```

Figure 1: Comparison of the fuel-efficiency of cars in the mtcars data set based on the transmission-type (manual versus automatic). The average fuel-efficiency for cars with automatic transmissions is 17.1. The distribution of mpg for automatic transmissions appears compact and roughly normally distributed. The average fuel-efficiency for cars with manual transmissions is 24.4. The distribution of mpg for manual transmissions is dispersed and shows a slight two-lobed form.

```r
library(ggplot2)
library(gridExtra)
fit1 <- lm(mpg ~ am, data=mtcars)
fit2 <- lm(mpg ~ am + cyl, data=mtcars)
fit3 <- lm(mpg ~ am + cyl + disp, data=mtcars)
fit4 <- lm(mpg ~ am + cyl + disp + hp, data=mtcars)
fit5 <- lm(mpg ~ am + cyl + disp + hp + drat, data=mtcars)
fit6 <- lm(mpg ~ am + cyl + disp + hp + drat + wt, data=mtcars)
fit7 <- lm(mpg ~ am + cyl + disp + hp + drat + wt + qsec, data=mtcars)
fit8 <- lm(mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear, data=mtcars)
fitall <- lm(mpg ~ ., data=mtcars)
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fitall)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + drat
## Model 6: mpg ~ am + cyl + disp + hp + drat + wt
## Model 7: mpg ~ am + cyl + disp + hp + drat + wt + qsec
## Model 8: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear
## Model 9: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 64.0039 8.231e-08 ***
## 3     28 252.08  1     19.28  2.7452   0.11241
## 4     27 216.37  1     35.71  5.0849   0.03493 *
## 5     26 214.50  1      1.87  0.2663   0.61121
## 6     25 162.43  1     52.06  7.4127   0.01275 *
## 7     24 149.09  1     13.34  1.8999   0.18260
## 8     22 147.90  2      1.19  0.0846   0.91917
## 9     21 147.49  1      0.41  0.0579   0.81218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 1: Output from the ANOVA analysis comparing sequentially nested linear models of the output mpg (fuel-efficiency) with predictors from the mtcars data set. Low P-values indicate a high likelihood that the added predictor is a necessary addition to the model. Only Models 2, 4, and 6 have P-values below a 5% threshold indicating that *cyl*, *hp*, and *wt* are important predictors of fuel-efficiency.

```r
fitFinal <- lm(mpg ~ am + cyl + hp + wt, data=mtcars)
summary(fitFinal)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## am           1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## hp          -0.02495    0.01365  -1.828   0.0786 .
## wt          -2.60648    0.91984  -2.834   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

Table 2: A summary of the linear model $mpg_i = \beta_0 + \beta_1 * am_i + \beta_2 * cyl_i + \beta_3 * hp_i + \beta_4 * wt_i + \epsilon_i$ applied to the *mtcars* data set. The model P-value is nearly zero indicating that it is highly unlikely that all coefficients are zero. However, the predictor of interest, *am* has a high P-value and the coefficient is barely more than one standard deviation away from zero ($tvalue = 1.026$).

```
fitAM <- lm(mpg ~ am, data=mtcars)
residAM <- data.frame(Residual = resid(fitAM))
gam <- ggplot(residAM )
gam <- gam + geom_point(aes(x=rownames(residAM), y=Residual))
gam <- gam + ylim( min(residAM$Residual), max(residAM$Residual))
gam <- gam + theme(axis.text.x = element_text(angle=30, hjust=1,
                                              vjust=1, size=rel(0.5)))
gam <- gam + labs(title = expression(Residuals %->% mpg[i] == beta[0] + beta[1]*am[i] + epsilon[i]),
                 x = "Car", y = "Residual")

residFinal <- data.frame(Residual = resid(fitFinal))
gf <- ggplot(residFinal )
gf <- gf + geom_point(aes(x=rownames(residFinal), y=Residual))
gf <- gf + ylim( min(residAM$Residual), max(residAM$Residual))
gf <- gf + theme(axis.text.x = element_text(angle=30, hjust=1,
                                            vjust=1, size=rel(0.5)))
gf <- gf + labs(title = expression(Residuals %->% mpg[i] == beta[0] + beta[1]*am[i] + beta[2]*cyl[i] +
                 x = "Car", y = "Residual")

grid.arrange(gf, gam, nrow=2)
```

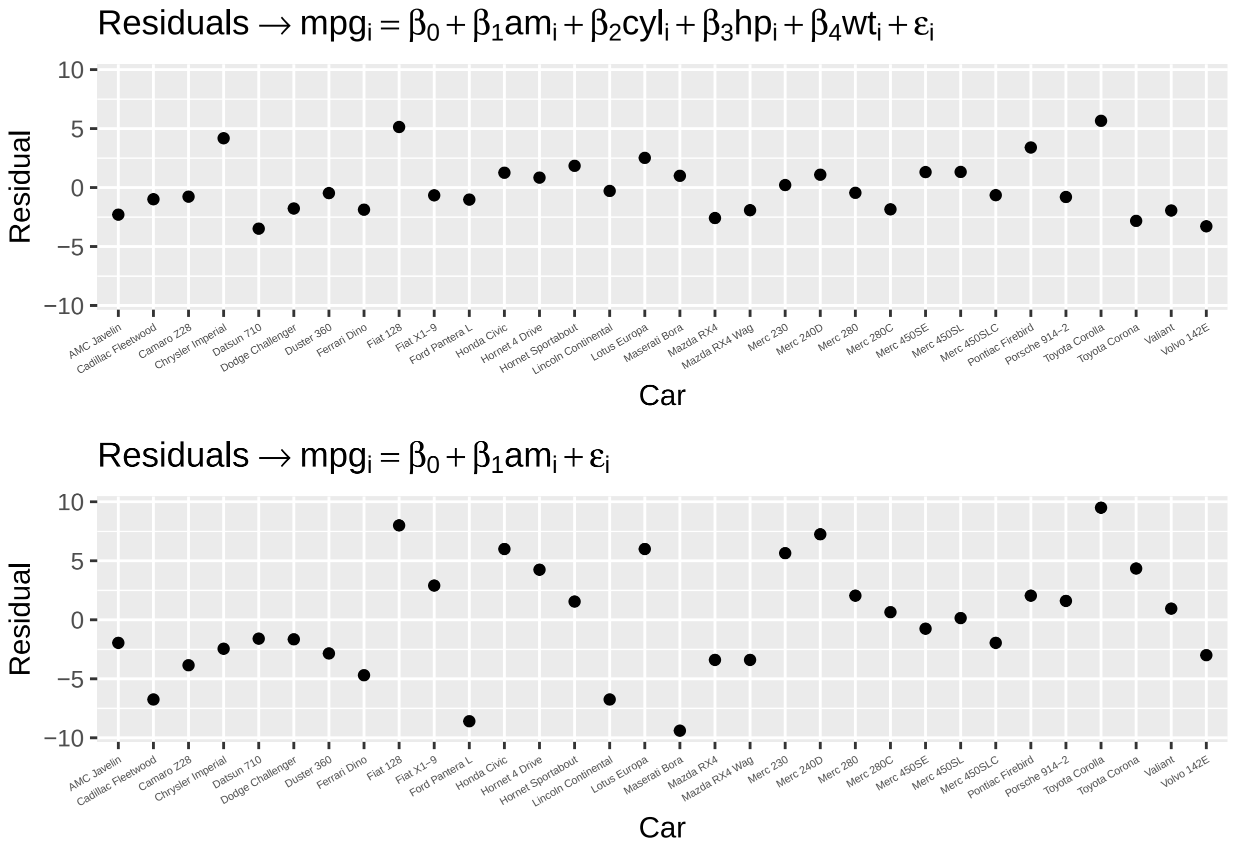


Figure 2: Plotting the residuals for the final linear model used ($mpg_i = \beta_0 + \beta_1 * am_i + \beta_2 * cyl_i + \beta_3 * hp_i + \beta_4 * wt_i + \epsilon_i$) compared to a simple linear model containing only the transmission type as a predictor ($mpg_i = \beta_0 + \beta_1 * am_i + \epsilon_i$). The more complex model shows a narrowing of residuals indicating a better fit.