Regression Models Course – Solution to Project 1

Dr. Guy Cohen

August 19, 2015

Effect of Transmission Type on Fuel Economy in 1973-74 Cars

Executive Summary

We look at the data of the *Motor Trend* Car Road Tests which was extracted from the 1974 *Motor Trend* US magazine and which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). We perform an exploratory analysis of the data. We use forward selection to build a multiple linear regression predictive model for mpg using the other variables. The best model has 4 parameters and has an adjusted R-squared of 0.84. From this model we conclude, with 90% confidence, that manual transmission is associated with better fuel economy. For the cars in the dataset and keeping the other variables in the model constant, choosing a car with manual, rather than automatic, transmission improves fuel economy by 2.26 miles per gallon (90% confidence interval of [0.04,4.49]).

Exploratory Data Analysis

The data of the *Motor Trend* Car Road Tests was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). The dataset consists of 32 observations on 11 variables: mpg - Miles/(US) gallon, cyl - Number of cylinders, disp - Displacement (cu.in.), hp - Gross horsepower, drat - Rear axle ratio, wt - Weight (1000 lb), qsec - 1/4 mile time, vs - V/S, am - Transmission (0 = automatic, 1 = manual), gear - Number of forward gears, carb - Number of carburetors.

We first convert the factor variables to factor type.

```
data(mtcars)
mtcars2 <- mtcars
mtcars2[,c("cyl","vs","am","gear","carb")] <-
    as.data.frame(lapply(mtcars2[,c("cyl","vs","am","gear","carb")],as.factor))
```

Second, we plot boxplots of continuous variables and barplots for categorical ones (Figs. 1-11 in App.). Of the numerical variables, none seem to be excessively skewed. In the weight variable we note 2 outliers. Since the dataset presumably has no errors, we leave those there. From the categorical variables, the number of forward gears and the number of carburetors seem to be imbalanced, so caution should be used when employing those as predictors.

We plot a pairs plot to see how the variables relate to each other (Fig. 12 in App.). From the pairs plot it is notable that weight and displacement have a correlation coefficient of +0.888 . MPG is highly correlated with displacement and with weight (correlation coefficients of -0.848 and -0.868 respectively). Thus, when predicting MPG we should take either displacement, weight, a linear combination of both or none.

Detailed Solution

We build the best model to predict MPG and assume that the data are randomized, so that the unexplained variance is composed of random Gaussian noise and of contributions from missing important predictors with values that are randomly distributed. We confirm this assumption later by the residual plot. **Our strategy for feature selection is "forward selection"** by which one starts with the null model with no predictors. **We examine many models**. At each stage we add the new feature that decreases MSE the most.

```
library(leaps)
regfit.fwd=regsubsets(mpg ~ ., data = mtcars2, nvmax = 12, method="forward")
summary(regfit.fwd)$outmat[1:9,1:13]
```

```
##          cyl6 cyl8 disp hp  drat wt  qsec vs1 am1 gear4 gear5 carb2 carb3
## 1  ( 1 ) " "  " "  " "  " " " " "*" " "  " " " " " " " "   " "   " "   " "
## 2  ( 1 ) " "  " "  " "  "*" " " "*" " "  " " " " " " " "   " "   " "   " "
## 3  ( 1 ) "*"  " "  " "  "*" " " "*" " "  " " " " " " " "   " "   " "   " "
## 4  ( 1 ) "*"  " "  " "  "*" " " "*" " "  " " "*" " " " "   " "   " "   " "
## 5  ( 1 ) "*"  " "  " "  "*" " " "*" " "  "*" "*" " " " "   " "   " "   " "
## 6  ( 1 ) "*"  " "  " "  "*" " " "*" " "  "*" "*" " " " "   "*"   " "   " "
## 7  ( 1 ) "*"  " "  " "  "*" " " "*" "*"  "*" "*" " " " "   "*"   " "   " "
## 8  ( 1 ) "*"  "*"  " "  "*" " " "*" "*"  "*" "*" " " " "   "*"   " "   " "
## 9  ( 1 ) "*"  "*"  " "  "*" "*"  "*" "*"  "*" "*" " " " "   "*"   " "   " "
```

We plot a graph of adjusted R-squared vs. number of predictors (Fig. 13 in App.). We see the best performance is obtained with 5 predictors. However, since the value for 4 predictors is within the confidence interval of the value for 5 predictors, we choose to have 4 predictors, which from the summary of the forward selection procedure are: dummy variable for cylinders equal to 6, horsepower, weight and whether transmission is manual (cyl6, hp, wt and am1).

We fit a linear model using the parameters found.

```
fit <- lm(mpg ~ I(cyl == "6") + hp + wt + I(am == "1"), data = mtcars2)
coef(summary(fit))
```

```
##                      Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)       34.34835479 2.510930248 13.679534 1.171231e-13
## I(cyl == "6")TRUE -2.15170905 1.055362386 -2.038834 5.136745e-02
## hp                -0.04158974 0.009326389 -4.459362 1.301187e-04
## wt                -2.67523820 0.863677640 -3.097496 4.518122e-03
## I(am == "1")TRUE   2.26599959 1.307887711  1.732564 9.458327e-02
```

We plot residual plots (Last 4 Figs. in App.). From the residual plots we see the assumptions of multiple linear regression are satisfied. There is no pattern in the graph of the residuals vs. the fitted value and the residuals are approximately normal.

From the adjusted R-squared plot we see that R-squared of the fit is is 0.84. From the fit summary, we see the coefficient of `I(am == 1)` is 2.266. Thus, the effect of the transmission type is to add 2.266 miles per gallon to the mpg for manual transmission and 0 for automatic transmission. The standard error is 1.307. The p-value is 0.09458>0.05, so with 95% confidence we fail to reject the null hypothesis that transmission type is associated with change in fuel economy. However, since the p-value is less than 0.1, with 90% confidence, we reject this null hypothesis.

To summarize, we answer the original questions. **Question**: Is an automatic or manual transmission better for MPG? Answer: With 90% confidence, manual transmission is better for MPG for the cars in the dataset. **Question**: Quantify the MPG difference between automatic and manual transmissions. Answer: The expected addition to MPG due to having manual transmission and not automatic transmission, keeping horsepower, weight and whether number of cylinders is 6, constant, is 2.26 ± 1.31 miles per gallon. The 90% confidence interval for this value is $(0.04, 4.49)$.

```
summary(fit)$coef[5,1]+qt(.95,27)*c(-1,1)*summary(fit)$coef[5,2]
```

```
## [1] 0.03828957 4.49370962
```