Final Project: Analysis of Automobile MPG

Xihao Wu & Anhui Shi

PSTAT 126

Section Time

Tuesday & Thursday 5:00 - 5:50 pm

# Introduction

The data we used in this project comes from a sample of 398 automobile specifications. Data for each of automobile contains 9 attributes in varied data types. Dataset includes continuous variables: *mpg*, *displacement*, *horsepower*, *weight* and *acceleration*. The Dataset also includes discrete variables: *cylinders* (2 to 8), *model year* (1970 to 1982) and *origin* (1 being originated in America, 2 being in Europe and 3 being in Japan). The dataset at last contains each *car’s name* in string for further reference. Our research questions centered at automobile mpg. We assume the response variable is normally distributed with independent observations. Hence, under the linear regression model, we had located 3 explanatory variables (*weight, model year, origin*) best suited for the regression model to answer our questions of interest.

# Questions of Interest

1. What is the goodness of fit of the final model when mpg as response given different origins?
2. What is the predicted average mpg of latest model American cars with average weight in the dataset with model above?

# Regression Method

For research question 1, we would first use stepwise regression and best subset regression to select predictors when setting *mpg* as response and preserving *origin* as one of the predictors. Then we perform the residual analysis and check the four LINE conditions. If necessary, we conduct transformation based on which assumption the model has violated. We would also look for influential points and refit the model if they exit. Then, we conduct hypothesis test to explore if interaction terms are required. At last, we calculate the R2 value of the final model to find out the goodness of fit.

For research question 2, after we had settled down on the final model, we locate the data entry with correct specifications and calculated the 95% confidence interval for our prediction.

# Regression Analysis, Results and Interpretation

*Goodness of Fit (R2)*

Before conducting any analysis, we first corrected variable type of each attribute. We found out that the attribute *horsepower* was in string. Hence, we had converted it to double and deleted data with missing entries. We also decided that the variable *origin* should be categorical as we are interested in response under different categories not the effect of the category on response. Hence, we converted it to factor for the convenience of computation.

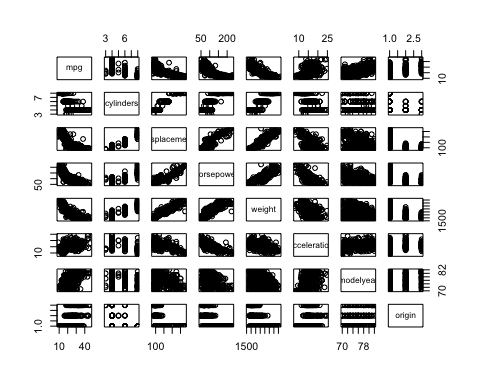
On a larger picture, we performed preliminary analysis on the data to determine relationships between variables, shown in Figure 1. The top five scatterplot shows that the

Figure 1. Scatter Plot Matrix

variable *displacement, horsepower and weight* all exhibit strong negative relationship with *mpg*; *origin* shows a slight positive relationship with *mpg;* *model year* however shows a strong positive relationship with the response. On the other hand, the relationship between *mpg* and *cylinders* or *acceleration* is rather unclear. Under closer look, *displacement*, *horsepower* and *weight* also show a strong positive relationship with one another on the scatterplot. This could infer that these three variables could be interchangeable and only one is sufficient for the model.

To begin the variable selection process, we only input the explanatory variable excluding the categorical one, *origin*, since we need to preserve this variable to answer the question regarding car’s origin. The stepwise regression under AIC selected *weight* and *model* *year* as final predictors1. Best subset regression under both adjusted R2 and Mallow’s Cp reached the same model as well2. This supports our preliminary analysis before. Thus, the predictors we selected through variable selection are *weight* (continuous), *model year* (continuous), and *origin* (categorical) with regression function: (origin2 = 1 indicating the data is European car, 0 otherwise; origin3 = 1 indicating the data is Japanese car, 0 otherwise)

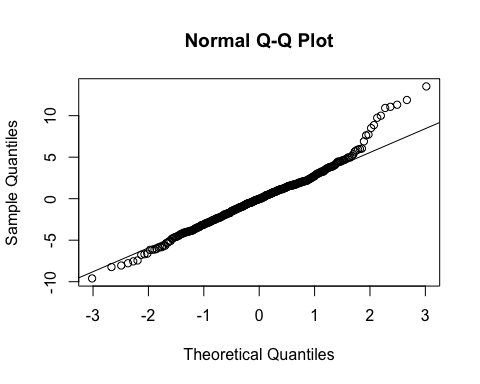
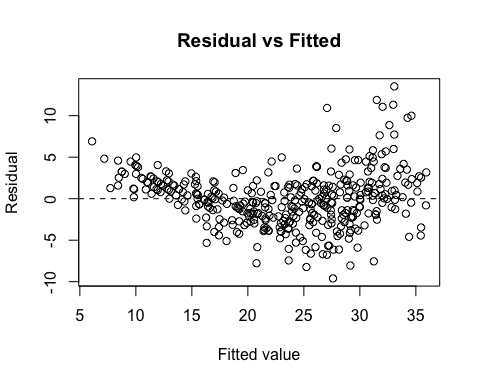
The residual analysis with above regression function, however, showed that the model violated both equal variance assumption (see figure. 2) and normality assumption (see figure. 3). 

Figure 2. Residual Analysis Figure 3. QQ plot

We could see a clear fan out pattern on Residual vs. Fitted and a heavy tails on QQ plot. The Shapiro-Wilk Test verified our observation with a p-value very close to zero3, thus we can reject the null hypothesis that the model has normality.

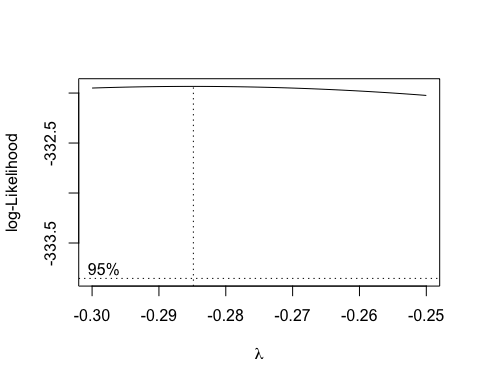
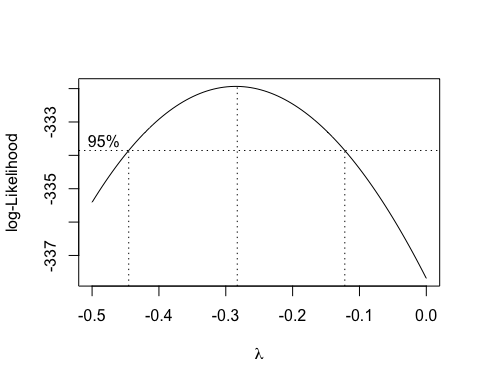
Above analysis calls for transformation on response. Boxcox gave us the right lambda value to transform mpg (see figure. 4 and figure. 5). 

Figure 4. Boxcox Figure 5. Boxcox approximation

We then decided that the lambda should be -0.285. The Residual vs. Fitted plot did look better as the spread of data points are evenly distributed along the average line (see figure. 6).

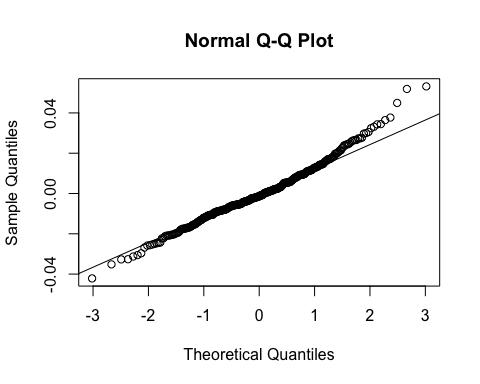
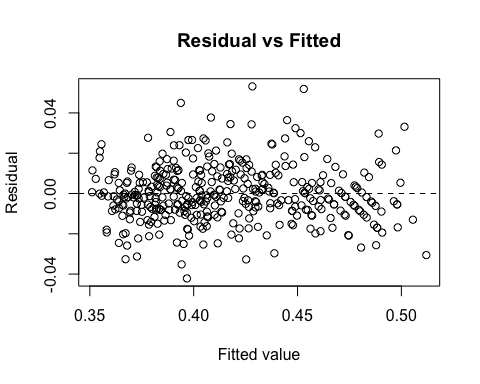


Figure 6. Transformed Residual Analysis Figure 7. Transformed QQ plot

However, normality assumption was still not met as the transformed QQ plot still had heavy tails and the p-value given by Shapiro-Wilk test was still close to zero although significantly improved4.

To fix the normality assumption, we checked possible influential points and outliers. Using studentized deleted residuals, we found out that 4 data points could be classified as outliers5, although all the data cannot be deemed influential using Cook’s distance6. We tried to delete the outliers and refit the model to see if the normality condition was improved. However, the largest p-value we could obtain is still not enough to prove normality7.

To further explore the relationship between the response and predictors, we conduct hypothesis test to look for possible interaction terms with transformed response. We defined null hypothesis as a reduced model with no interaction term with regression function:

and defined alternative hypothesis as a full model with all the possible interaction term with regression function:

We want to test if at least of one the slope parameter for the interaction term is not zero. However, the partial F-test suggested that we don’t have enough evidence to reject the null hypothesis as the p-value calculated in significantly larger than the alpha value. Thus, we conclude that the final model for mpg would be:

under linear regression. However, we cannot conclude that the data points are normality distributed at each predictor using this model.

Thus, we can now calculate the goodness of fit (R2) value, which is 0.882310. This means that the final model explains 88.23% of the data, which suggests that our model is valid and sound.

*95% Confidence Interval for Mean Response*

The 95% confidence interval calculated under the model for average mpg of American cars with average weight for the latest model is between 26.6 and 25.211. This means that we are 95% percent confidence that the expected mpg for American cars with average weight in year 1982 would be between 26.6 and 25.2.

# Conclusion

The predictors that best answer the questions of interest is its *weight*, *model year* and its *origin*. However, the response does not distribute normally. The outliers in the data set contribute the violation but deleting them doesn’t solve the problem. The final model is certainly not the best model for the response as we arbitrarily select *origin* as one of the predictors since it’s in the questions of interest. The model will be different if we include *origin* under variable selection process.

Looking at the summary table, we can conclude that European cars have higher mpg on average than American cars. Since we had transformed the response, the order of data would reverse after boxcox transformation. This difference under transformation is denoted by , meaning that the of American car is smaller than the of European cars by . Japanese cars also have higher mpg on average than American cars with the difference and European cars have higher mpg on average than Japanese cars with the difference under transformation. These values are significant with respect to the scale after transformation and we can then conclude that there is a significant difference in mpg among these three regions.

Appendix

1Stepwise regression under AIC

## Start: AIC=1611.93  
## mpg ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + weight 1 16497.8 7321.2 1151.5  
## + displacement 1 15440.2 8378.8 1204.4  
## + horsepower 1 14433.1 9385.9 1248.9  
## + cylinders 1 14403.1 9415.9 1250.1  
## + modelyear 1 8027.7 15791.3 1452.8  
## + acceleration 1 4268.5 19550.5 1536.5  
## <none> 23819.0 1611.9  
##   
## Step: AIC=1151.49  
## mpg ~ weight  
##   
## Df Sum of Sq RSS AIC  
## + modelyear 1 2752.3 4569.0 968.66  
## + horsepower 1 327.4 6993.8 1135.56  
## + acceleration 1 168.3 7152.9 1144.37  
## + displacement 1 150.9 7170.3 1145.33  
## + cylinders 1 115.1 7206.1 1147.28  
## <none> 7321.2 1151.49  
## - weight 1 16497.8 23819.0 1611.93  
##   
## Step: AIC=968.66  
## mpg ~ weight + modelyear  
##   
## Df Sum of Sq RSS AIC  
## <none> 4569.0 968.66  
## + acceleration 1 10.5 4558.5 969.77  
## + cylinders 1 5.0 4564.0 970.24  
## + horsepower 1 3.3 4565.7 970.38  
## + displacement 1 0.0 4568.9 970.66  
## - modelyear 1 2752.3 7321.2 1151.49  
## - weight 1 11222.4 15791.3 1452.81

2Best subset regression under adjusted R­­­2 and Mallow’s Cp

## Call:  
## lm(formula = mpg ~ weight + modelyear, data = auto\_mpg[, -8])  
##   
## Coefficients:  
## (Intercept) weight modelyear   
## -14.347253 -0.006632 0.757318

## (Intercept) cylinders displacement horsepower weight acceleration  
## 1 TRUE FALSE FALSE FALSE TRUE FALSE  
## 2 TRUE FALSE FALSE FALSE TRUE FALSE  
## 3 TRUE FALSE FALSE FALSE TRUE TRUE  
## 4 TRUE FALSE TRUE FALSE TRUE TRUE  
## 5 TRUE TRUE TRUE FALSE TRUE TRUE  
## 6 TRUE TRUE TRUE TRUE TRUE TRUE  
## modelyear  
## 1 FALSE  
## 2 TRUE  
## 3 TRUE  
## 4 TRUE  
## 5 TRUE  
## 6 TRUE

## [1] 0.6918423 0.8071941 0.8071393 0.8067872 0.8067841 0.8062826

## [1] 232.396144 1.169751 2.284200 3.992019 5.000800 7.000000

3Shapiro-Wilk Test before transformation

##   
## Shapiro-Wilk normality test  
##   
## data: e.2.selected  
## W = 0.97513, p-value = 2.951e-06

4Shapiro-Wilk Test after Boxcox

## Shapiro-Wilk normality test  
##   
## data: e.2.trans  
## W = 0.98356, p-value = 0.0001951

5Studentized deleted residuals for the largest 5 data points

## 165 124 111 382 274   
## 3.882467 3.788329 3.286674 3.067614 2.748952

6Cook’s Distance for the largest 3 data points

## 111 274 276   
## 0.03684560 0.03020079 0.02938542

7Shapiro-Wilk test with deleted datasets

##   
## Shapiro-Wilk normality test  
##   
## data: e.deleted  
## W = 0.99213, p-value = 0.0378

8Hypothesis test checking for interaction

## Analysis of Variance Table  
##   
## Model 1: mpg^(-0.285) ~ weight + modelyear + origin  
## Model 2: mpg^(-0.285) ~ weight + modelyear + origin + weight \* modelyear +   
## weight \* origin + modelyear \* origin  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 387 0.075545   
## 2 382 0.074799 5 0.00074532 0.7613 0.5782

9Summary table of the final table for R2 value

##   
## Call:  
## lm(formula = mpg^(-0.285) ~ weight + modelyear + origin, data = auto\_mpg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.042143 -0.008286 -0.001333 0.008135 0.053164   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.975e-01 1.682e-02 35.526 < 2e-16 \*\*\*  
## weight 3.491e-05 1.088e-06 32.084 < 2e-16 \*\*\*  
## modelyear -3.731e-03 2.038e-04 -18.313 < 2e-16 \*\*\*  
## origin2 -7.769e-03 2.168e-03 -3.583 0.000383 \*\*\*  
## origin3 -5.412e-03 2.172e-03 -2.492 0.013123 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01397 on 387 degrees of freedom  
## Multiple R-squared: 0.8823, Adjusted R-squared: 0.8811   
## F-statistic: 725.5 on 4 and 387 DF, p-value: < 2.2e-16

11Confidence Interval

## fit lwr upr  
## 1 25.92349 26.63265 25.23841

Code

#Load the data set, delete the data that we don't want  
library(readr)  
auto\_mpg <- read\_table2("auto-mpg.txt", col\_names = FALSE)

auto\_mpg<-auto\_mpg[,-c(9,10,11)]  
names(auto\_mpg)<-c('mpg','cylinders','displacement','horsepower','weight','acceleration','modelyear','origin')  
auto\_mpg$horsepower <- as.numeric(auto\_mpg$horsepower)

auto\_mpg<-na.omit(auto\_mpg)  
auto\_mpg$origin<-as.factor(auto\_mpg$origin)

#Pairwise scatterplot   
pairs(auto\_mpg)

library(faraway)  
  
mod.1 = lm(mpg ~ 1, data = auto\_mpg[,-8])  
mod.1.upper = lm(mpg ~ ., data = auto\_mpg[,-8])  
  
step(mod.1, scope = list(lower = mod.1, upper = mod.1.upper))

#Variable selection  
library(leaps)  
mod.2 = regsubsets(auto\_mpg[,-c(1,8)], auto\_mpg$mpg)  
summary(mod.2)$which

summary(mod.2)$adjr2

summary(mod.2)$cp

#Residual Analysis  
mod.2.selected = lm(mpg ~ weight + modelyear + origin , data = auto\_mpg)  
yhat.2.selected= fitted(mod.2.selected)  
e.2.selected = auto\_mpg$mpg - yhat.2.selected  
plot(yhat.2.selected, e.2.selected, xlab = 'Fitted value', ylab = 'Residual', main = 'Residual vs Fitted')  
abline(h = 0, lty = 2)

#QQ-plot  
qqnorm(e.2.selected)  
qqline(e.2.selected)

#Shapiro-Wilk Test  
shapiro.test(e.2.selected)

#boxcox  
library(MASS)  
boxcox(mod.2.selected)

boxcox(mod.2.selected, lambda = seq(-0.5,0, length = 10))

boxcox(mod.2.selected, lambda = seq(-0.3,-0.25, length = 10))

#Residual Analysis after boxcox transformation  
mod.2.trans = lm(mpg^(-0.285) ~ weight + modelyear + origin , data = auto\_mpg)  
yhat.2.trans= fitted(mod.2.trans)  
e.2.trans = (auto\_mpg$mpg)^(-0.285) - yhat.2.trans  
plot(yhat.2.trans, e.2.trans, xlab = 'Fitted value', ylab = 'Residual', main = 'Residual vs Fitted')  
abline(h = 0, lty = 2)

#QQ-plot  
qqnorm(e.2.trans)  
qqline(e.2.trans)

#Shapiro-Wilk Test after boxcox transformation  
shapiro.test(e.2.trans)

summary(mod.2.trans)

#studentized deleted residuals  
sort(abs(rstudent(mod.2.trans)), decreasing = TRUE )[c(1,2,3,4,5)]

#Cook's distance  
sort(cooks.distance(mod.2.trans), decreasing =TRUE)[c(1,2,3)]

#Residual Analysis with deleted data   
mod.deleted= lm(mpg^(-0.285) ~ weight + modelyear + origin , data = auto\_mpg[-c(165,124,111), ] )  
yhat.deleted= fitted(mod.deleted)  
e.deleted = (auto\_mpg[-c(165,124,111), ]$mpg)^(-0.285) - yhat.deleted  
  
#QQ-plot with deleted data  
qqnorm(e.deleted)  
qqline(e.deleted)

#Shapiro-Wilk Test with deleted data  
shapiro.test(e.deleted)

#Find interaction  
mod.3.reduced = lm(mpg^(-0.285) ~ weight + modelyear + origin , data = auto\_mpg)  
mod.3.full =lm(mpg^(-0.285) ~ weight + modelyear + origin + weight\*modelyear + weight\*origin + modelyear\*origin , data = auto\_mpg)  
anova(mod.3.reduced, mod.3.full)

summary(mod.3.full)

#Final model and R^2 value  
mod.final = lm(mpg^(-0.285) ~ weight + modelyear + origin , data = auto\_mpg )  
summary(mod.final)

#Confidence Interval  
  
new\_ave = data.frame(weight = mean(auto\_mpg$weight), modelyear = max(auto\_mpg$modelyear), origin = as.factor(1) )  
  
ci\_ave = predict(mod.final , newdata = new\_ave, interval = 'confidence', level = 0.95)  
ci\_ave^(-1/0.285)