

Regression Models Course Project: Data Analysis on Motor Trend

Cheng-Han Yu

August 19, 2015

Executive Summary

We examine `mtcars` data set to answer the following questions:

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

We first summarize the relationships between variables, and then fit a linear regression model that has the smallest BIC and the largest adjusted R^2 , followed by residual analysis and diagnostics. Whether with or without interaction, our model tells us a manual transmission is better for MPG, and no-pattern residual plots are indications for good model fitting.

Exploratory Data Analysis

The `mtcars` data set is an R data frame with 32 observations on 11 variables. Figure 1 in Appendix gives us a general picture of the variables including their histogram, scatter plots and correlation between variables. Marginally manual transmission seems to have higher MPG than automatic transmission.

Regression Models and Subset Selection

We first consider two naive models, the model including all predictors (`fit.full`) and the one with variable `am` only (`fit.am`).

```
fit.full <- lm(mpg ~ ., data = mtcars); round(summary(fit.full)$coef[, 4][-1], 2) # p-value

## cyl6 cyl18 disp hp drat wt qsec vs1 am1 gear4 gear5 carb2
## 0.40 0.96 0.28 0.09 0.64 0.09 0.70 0.51 0.71 0.77 0.51 0.68
## carb3 carb4 carb6 carb8
## 0.50 0.81 0.49 0.40

fit.am <- lm(mpg ~ am, data = mtcars); summary(fit.am)$coef[2, ]

## Estimate Std. Error t value Pr(>|t|)
## 7.2449392713 1.7644216316 4.1061269831 0.0002850207
```

In the full model, all coefficients are not significant at 5% significance level, although it has large adjusted $R^2 = 0.78$. Fitting many correlated variables results in *multicollinearity* and *overfitting* with inflated estimated standard error. On the contrary, the coefficients of the `am`-only model are significantly different from zero, saying that on average, a manual transmitted car has 7.245 MPG higher than an automatic transmitted car. However, the model has small adjusted $R^2 = 0.34$, implying small explanatory power for MPG.

To do variable selection, we use forward and backward stepwise selection with two criteria AIC ($-2 \log L(M) + 2k$) and BIC ($-2 \log L(M) + k \log(n)$), where $L(M)$ is the maximum of the likelihood function of model M and k is the number of parameters in M and n is the number of observations.

Forward stepwise selection starts with a intercept-only model, and then adds predictors to the model, one at the time, until all of the predictors are in the model. At each step the variable that gives the greatest *additional* improvement to the fit is added to the model. Backward method, on the other hand, begins with the full model, and then removes the least useful covariate, one at the time.

There are four models `for.aic`, `for.bic`, `back.aic` and `back.bic` to be considered, each of which is the best model with the smallest AIC or BIC.

```
for.aic <- step(lm(mpg ~ 1, data = mtcars), direction = "forward",
                  scope = formula(fit.full), k = 2, trace = 0) # forward AIC
for.bic <- step(lm(mpg ~ 1, data=mtcars), direction = "forward",
                  scope = formula(fit.full), k = log(32), trace = 0) # forward BIC
back.aic <- step(fit.full, direction = "backward", k = 2, trace = 0) # backward AIC
back.bic <- step(fit.full, direction = "backward", k = log(32), trace = 0) # backward BIC

# back.aicRsq back.bicRsq  for.aicRsq  for.bicRsq
#  0.8335561   0.8335561   0.8263446   0.8185189
```

```
##             Estimate    Pr(>|t|) 
## (Intercept) 9.617781 1.779152e-01
## wt          -3.916504 6.952711e-06
## qsec         1.225886 2.161737e-04
## am1          2.935837 4.671551e-02
```

Since the model `back.bic` has the largest adjusted $R^2 = 0.834$, the model including `wt`, `qsec`, and `am` has the most explanatory power for `mpg`. Under this model, a manual transmission car, on average, has 2.936 miles per gallon more than an automatic transmission car, holding values of weights and 1/4 mile time constant.

We then fit four possible interaction models `fit.int`, `fit.int.aq`, `fit.int.aw` and `fit.int.wq` to check if any interaction is needed to be in the model.

```
fit.int <- summary(lm(mpg ~ wt * qsec * am, data = mtcars))
fit.int.aq <- summary(lm(mpg ~ wt + qsec * am, data = mtcars))
fit.int.aw <- summary(lm(mpg ~ qsec + wt * am, data = mtcars))
fit.int.wq <- summary(lm(mpg ~ am + qsec * wt, data = mtcars))

##     int_Rsq int.aq_Rsq int.aw_Rsq int.wq_Rsq
##  0.8759496  0.8531624  0.8804219  0.8347545
```

Since model `fit.int.aw` has the largest adjusted $R^2 = 0.88$, the model `mpg = 9.723 + (1.017)qsec + (-2.937)wt + (14.079)am + (-4.141)wt*am` is our final model. When `am = 0`, the slope of `wt` is -2.937 and the intercept is 9.723. When `am = 1`, the slope of `wt` is -7.078 and the intercept is 23.802. In term of uncertainty, the 95% confidence interval for the coefficients are shown below.

```
fit <- lm(mpg ~ qsec + wt * am, data = mtcars)
t(confint(fit))

##             (Intercept)      qsec        wt        am1      wt:am1
## 2.5 %    -2.380779  0.4998811 -4.303102  7.030875 -6.597032
## 97.5 %   21.826884  1.5340661 -1.569960 21.127981 -1.685721
```

Residual Diagnostics

Some plots for residual diagnostics are shown in Figure 2. There is no particular pattern in residuals vs fitted, scale-location, and residuals vs leverage plots. For QQ-plot, it seems that the residual is a little bit right skewed, but it still can be seen as normal from Shapiro-Wilk normality test.

We use $2 * k/n$ as a threshold for hat values, and there are four high leverage points, but according to `dfbeta()`, they are not so influential to our model.

```
shapiro.test(fit$res)

## 
##  Shapiro-Wilk normality test
##
## data: fit$res
## W = 0.9444, p-value = 0.1001

round(hatvalues(fit)[hatvalues(fit) > 2*5/32], 2) # high leverage

##          Merc 230 Lincoln Continental      Lotus Europa
##             0.35           0.32            0.33
##      Maserati Bora
##             0.37

round(dfbeta(fit)[which(hatvalues(fit) > 2*5/32), ], 2) # check influence

##                (Intercept)  qsec      wt     am1 wt:am1
## Merc 230          1.53 -0.09  0.01  0.46 -0.19
## Lincoln Continental 1.87 -0.03 -0.37 -1.14  0.30
## Lotus Europa       0.12 -0.01  0.00  0.10 -0.04
## Maserati Bora     0.38 -0.02 -0.01 -1.37  0.64
```

In sum, our model fit the data quite well. Although there are some high leverage points, they does not affect the model much. We may use this model to do inference and prediction as long as we pay attention to those data points with careful explanation.

Appendix

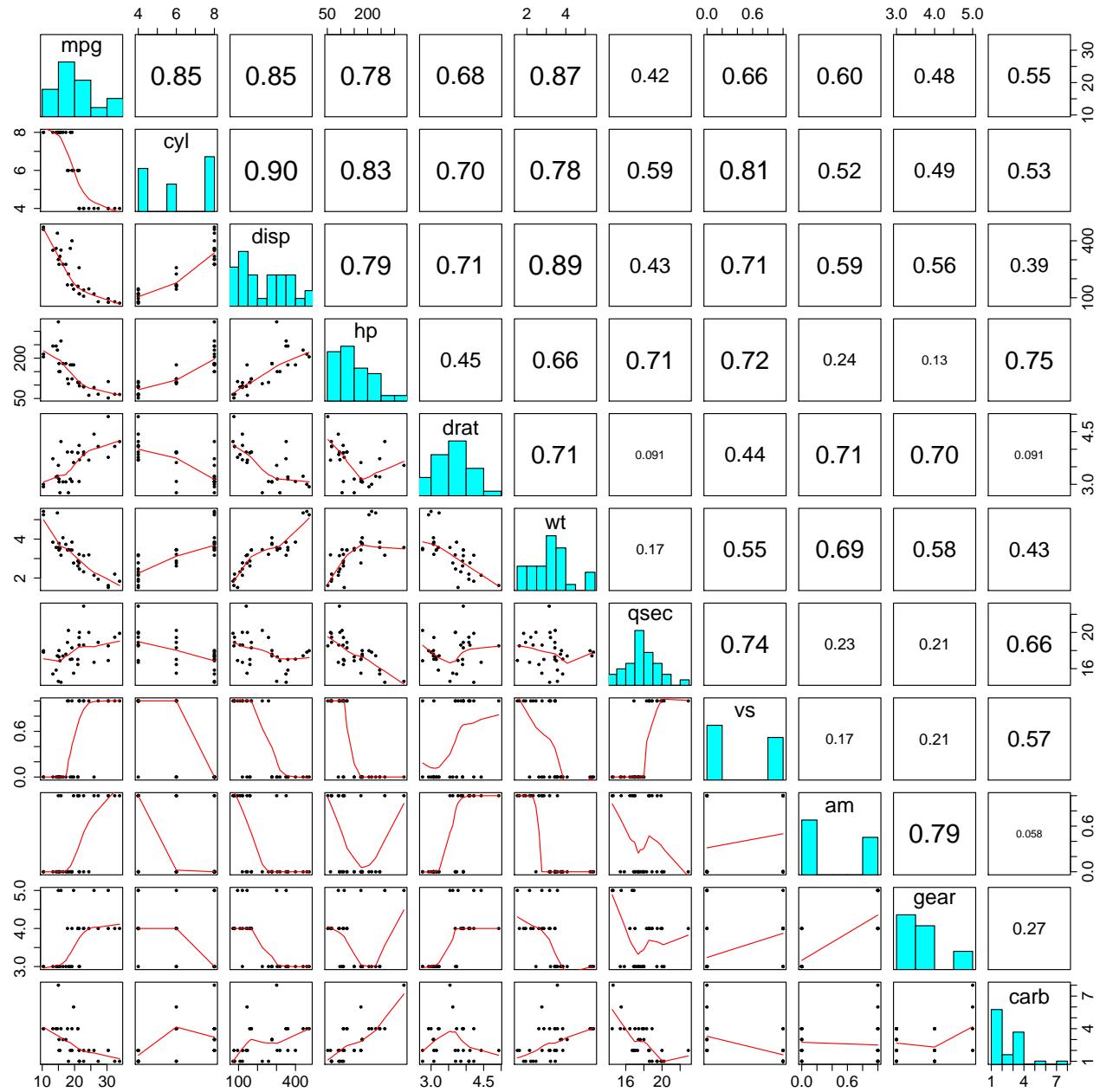


Figure 1: Descriptive Summary of Variables

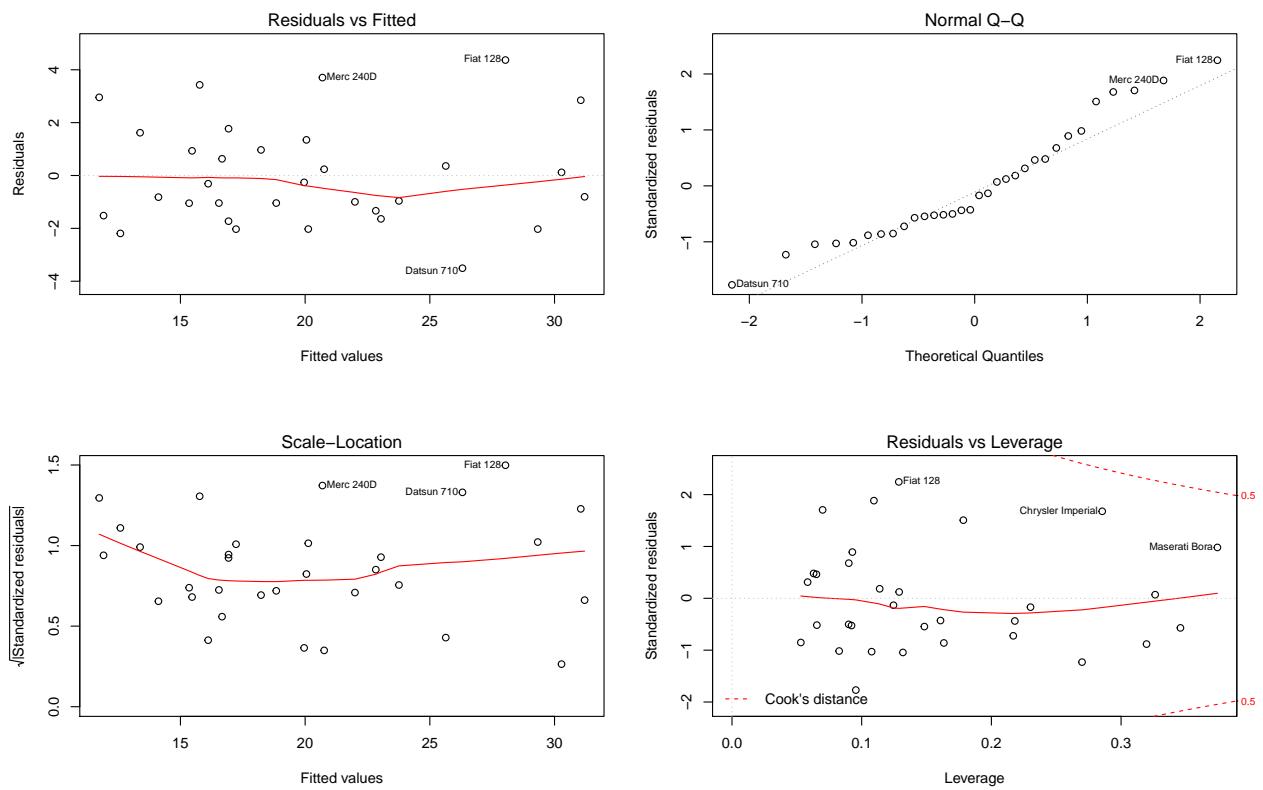


Figure 2: Residual Diagnostics