Modern Data Mining, HW 3

Group Member Wendy Deng Group Member Ruolan Li
Group Member Kira Nightingale

Due: 11:59Pm, 2/26, 2023

Contents

1 Overview

Multiple regression is one of the most popular methods used in statistics as well as in machine learning. We use linear models as a working model for its simplicity and interpretability. It is important that we use domain knowledge as much as we could to determine the form of the response as well as the function format for the factors. Then, when we have many possible features to be included in the working model it is inevitable that we need to choose a best possible model with a sensible criterion. `Cp`, `BIC` and regularizations such as LASSO are introduced. Be aware that if a model selection is done formally or informally, the inferences obtained with the final `lm()` fit may not be valid. Some adjustment will be needed. This last step is beyond the scope of this class. Check the current research line that Linda and collaborators are working on.

This homework consists of two parts: the first one is an exercise (you will feel it being a toy example after the covid case study) to get familiar with model selection skills such as, `Cp` and `BIC`. The main job is a rather involved case study about devastating covid19 pandemic. Please read through the case study first. This project is for sure a great one listed in your CV.

For covid case study, the major time and effort would be needed in EDA portion.

1.1 Objectives

- Model building process
- Methods
  - Model selection
    - All subsets
    - Forward/Backward
  - Regularization
    - LASSO (L1 penalty)
    - Ridge (L2 penalty)
    - Elastic net
- Understand the criteria
  - `Cp`
  - Testing Errors
  - `BIC`
  - `K fold Cross Validation`
  - `LASSO`
- Packages
  - `lm()`, `Anova`
  - `regsubsets()`
  - `glmnet()` & `cv.glmnet()`

2 Review materials

- Study lecture: Model selection
- Study lecture: Regularization
- Study lecture: Multiple regression

Review the code and concepts covered during lectures: multiple regression, model selection and penalized regression through elastic net.

3 Homework 2, Case study 3: Auto data set

If you haven't done this as part of the homework 2, please attach it here.

4 Case study 1: `ISLR::Auto` data

This will be the last part of the Auto data from ISLR. The original data contains 408 observations about cars. It has some similarity as the Cars data that we use in our lectures. To get the data, first install the package `ISLR`. The data set `Auto` should be loaded automatically. We use this case to go through methods learned so far.

Final modelling question: We want to explore the effects of each feature as best as possible.

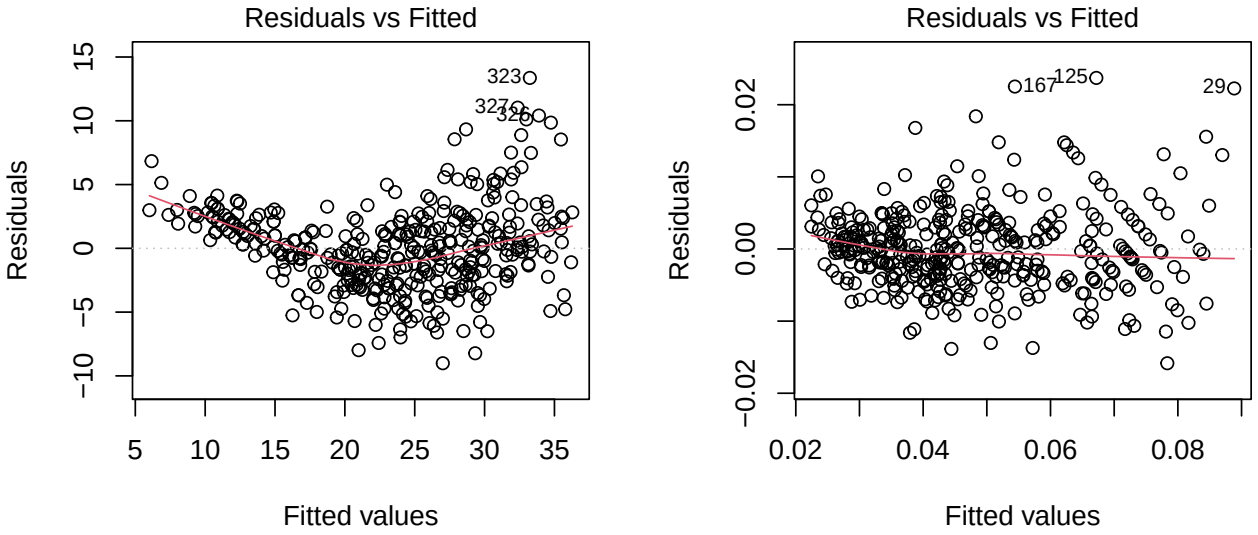
1) Preparing variables:

a) You may explore the possibility of variable transformations. We normally do not suggest to transform x for the purpose of interpretation. You may consider to transform y to either correct the violation of the linear model assumptions or if you feel a transformation of y makes more sense from some theory. In this case we suggest you to look into `GPM=1/MPG`. Compare residual plots of MPG or GPM as responses and see which one might yield a more satisfactory patterns.

```
library(dplyr)
#use MPG as response
Auto <- Auto %>%
  dplyr::select(-name) %>%
  mutate(origin = as.factor(origin)) #remove name column and convert origin as categorical variable
fit_mpg <- lm(mpg ~ ., data = Auto)


#use GPM as response
Auto_gpm <- Auto %>%
  mutate(gpm = 1/mpg) %>%
  dplyr::select(-mpg)
fit_gpm <- lm(gpm ~ ., data = Auto_gpm)

#residual plots
par(mfrow = c(1,2))
plot(fit_mpg, 1)
plot(fit_gpm, 1)
```



The residual plots indicate using GPM as response is more appropriate. The residual plot of GPM (on the right) is more close to a horizontal line, which indicates the relationship is more linear. Using GPM as response variable meets the regression assumptions. The residual plot of MPG suggests there may exist non-linear relationship.

In addition, can you provide some background knowledge to support the notion: it makes more sense to model `GPM`?

GPM measures fuel consumed per unit of distance, whereas MPG measures distance per unit of fuel consumed. GPM can reflect the actual difference in fuel consumption conditioning on same miles.

b) You may also explore by adding interactions and higher order terms. The model(s) should be as *parsimonious* (simple) as possible, unless the gain in accuracy is significant from your point of view.

```
fit_gpm1 <- lm(gpm ~ .+horsepower*origin, data = Auto_gpm)
anova(fit_gpm, fit_gpm1)
```

```
## Analysis of Variance Table
##
## Model 1: gpm ~ cylinders + displacement + horsepower + weight + acceleration +
##     year + origin
## Model 2: gpm ~ cylinders + displacement + horsepower + weight + acceleration +
##     year + origin + horsepower * origin
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1    383 0.0123
## 2    381 0.0122  2   7.2e-05 1.13   0.33
```

Because origin might affect horsepower, we checked interaction term of origin and horsepower. The p-value of F statistic is 0.3256, therefore, we fail to reject H_0, there is no significant difference between the model without interaction term and model with interaction term. That is to say, the interaction of horsepower and origin is not significant, and we do not need to include it in our model.

c) Use Mallow's C_p or BIC to select the model.

```r
library(leaps)
#model building
fit.exh <- regsubsets(gpm ~ ., data = Auto_gpm, nvmax = 10, method="exhaustive")
summary(fit.exh)
```

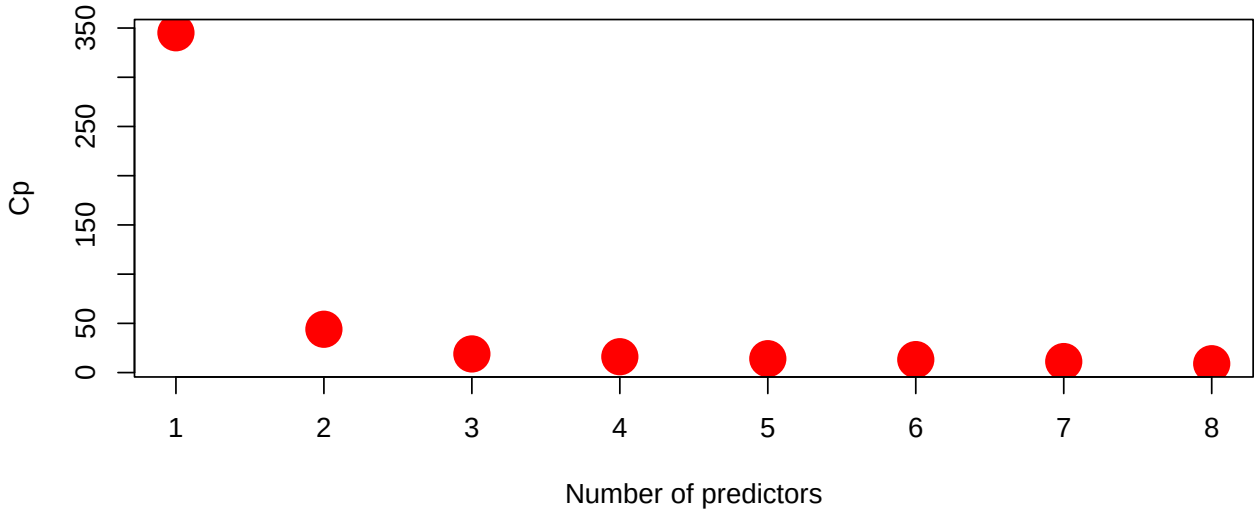
```
## Subset selection object
## Call: regsubsets.formula(gpm ~ ., data = Auto_gpm, nvmax = 10, method = "exhaustive")
## 8 Variables  (and intercept)
##              Forced in Forced out
## cylinders        FALSE      FALSE
## displacement     FALSE      FALSE
## horsepower       FALSE      FALSE
## weight           FALSE      FALSE
## acceleration     FALSE      FALSE
## year             FALSE      FALSE
## origin2          FALSE      FALSE
## origin3          FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          cylinders displacement horsepower weight acceleration year origin2
## 1  ( 1 ) " "       " "          " "        "*"    " "          " "  " "
## 2  ( 1 ) " "       " "          " "        "*"    " "          "*"  " "
## 3  ( 1 ) " "       " "          "*"        "*"    " "          "*"  " "
## 4  ( 1 ) " "       " "          "*"        "*"    " "          "*"  "*"
## 5  ( 1 ) " "       " "          "*"        "*"    "*"          "*"  "*"
## 6  ( 1 ) "*"       " "          "*"        "*"    "*"          "*"  "*"
## 7  ( 1 ) "*"       "*"          "*"        "*"    " "          "*"  "*"
## 8  ( 1 ) "*"       "*"          "*"        "*"    "*"          "*"  "*"
##          origin3
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

```r
#Plot the Cp values
f.e <- summary(fit.exh)
f.e$cp
```

```
## [1] 345.2  44.0  18.9  16.1  14.1  13.2  11.2   9.0
```

```r
plot(f.e$cp, xlab="Number of predictors",
     ylab="Cp", col="red", pch=16, cex =3,
     main = "Plot 1.1: Cp values")
```

Plot 1.1: Cp values



By using Mallow's C_p, we can see a model with all 8 variables has the smaller prediction error.

2) Describe the final model and its accuracy. Include diagnostic plots with particular focus on the model residuals.

```
#Final model
fit.exh.var <- f.e$which
final_var <- colnames(fit.exh.var)[fit.exh.var[5,]]
final_var
```

```
## [1] "(Intercept)"  "horsepower"   "weight"       "acceleration" "year"
## [6] "origin2"
```

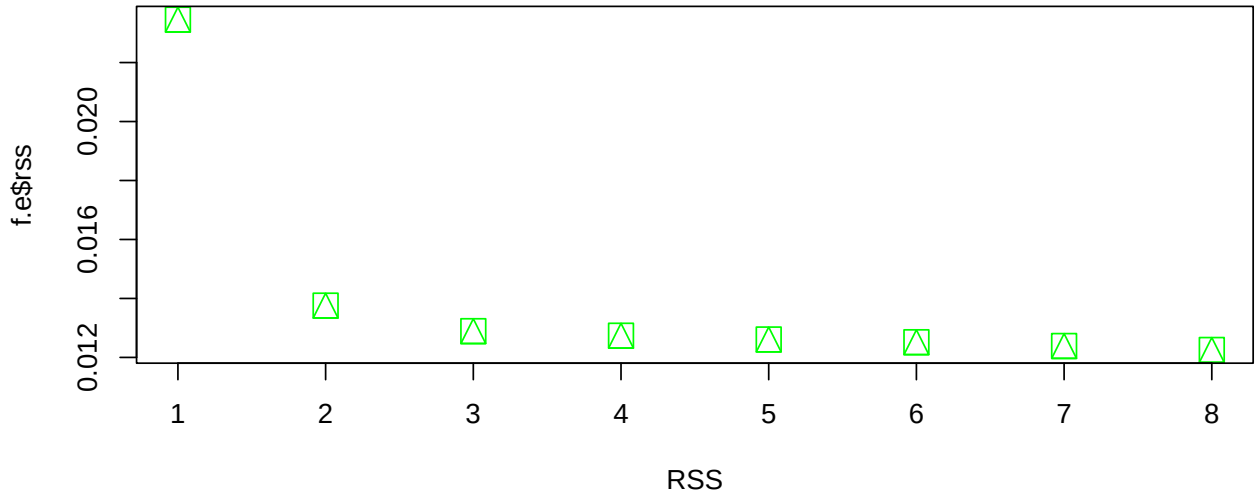
```
Auto_gpm1 <- Auto_gpm %>%
  mutate(origin2 = if_else(origin == 2, "1","0")) %>%
  dplyr::select(-origin) #if origin=2, assign it = 1, else = 0

#model summary
coef(fit.exh,5)
```

```
##  (Intercept)   horsepower       weight acceleration         year      origin2
##     9.66e-02     1.08e-04     1.16e-05     3.30e-04    -1.31e-03    -1.87e-03
```

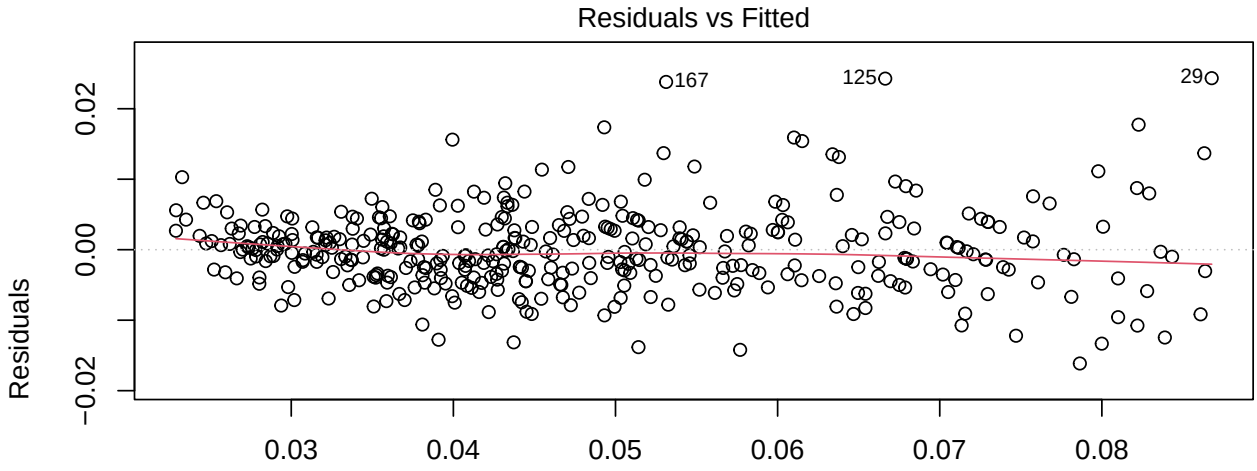
```
#diagnostic plots
plot(f.e$rss, xlab="RSS", pch = 14, col="green", cex=2, main = "Plot 1.2: RSS plot")
```

Plot 1.2: RSS plot



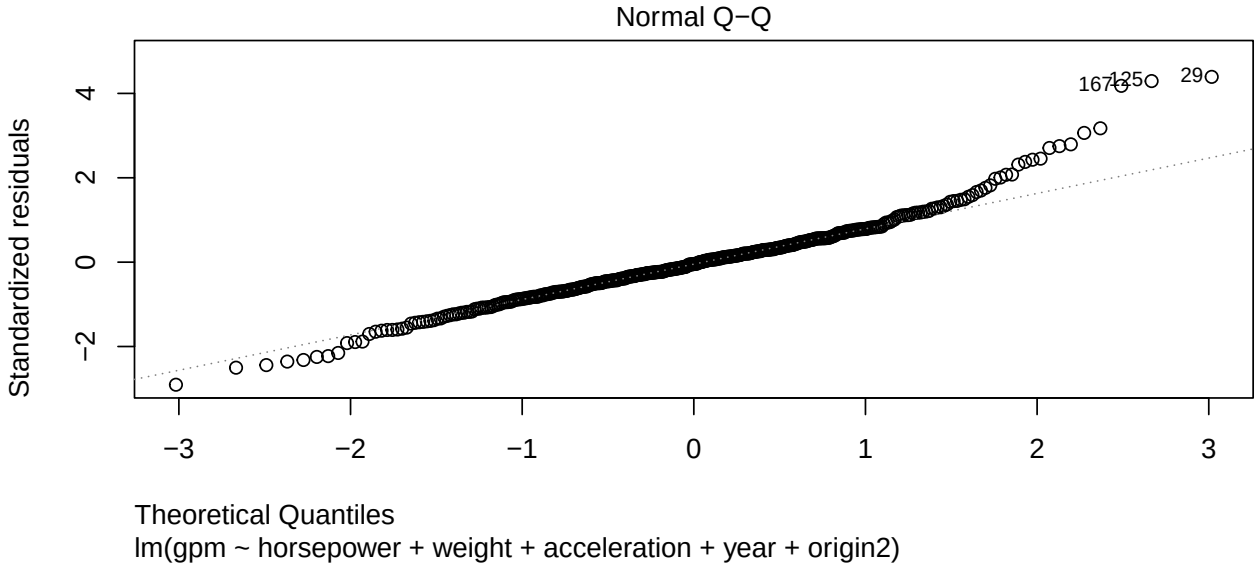

```
final_fit <- lm(gpm ~ horsepower+weight+acceleration+year+origin2, data = Auto_gpm1)
plot(final_fit,1, main = "Plot 1.3",adj=0)
```

Plot 1.3




```
plot(final_fit,2, main = "Plot 1.4",adj=0)
```

Plot 1.4



Since the C_p values are similar in scale, we choose final model with 5 variables. Small C_p value indicates more accurate model. The final model is:

$$gpm = 0.09662 + 0.0001081 * horsepower + 0.00001155 * weight + 0.00033 * acceleration - 0.001307 year - 0.00187 * origin2$$

- **From plot 1.2, we can see RSS is pretty small when selecting 5 variables.**
- **Plot 1.3 and 1.4 indicates the final model meets linearity and homoscedasticity assumption.**
  - Summarize the effects found.

```
summary(final_fit)
```

```
##
## Call:
## lm(formula = gpm ~ horsepower + weight + acceleration + year +
##     origin2, data = Auto_gpm1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.016142 -0.003487 -0.000249  0.002937  0.024335
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.66e-02   7.91e-03   12.22  < 2e-16 ***
## horsepower    1.08e-04   2.23e-05    4.85  1.8e-06 ***
## weight        1.16e-05   7.88e-07   14.66  < 2e-16 ***
## acceleration  3.30e-04   1.67e-04    1.98    0.048 *
## year         -1.31e-03   8.82e-05  -14.81  < 2e-16 ***
## origin21     -1.87e-03   8.13e-04   -2.30    0.022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00572 on 386 degrees of freedom
## Multiple R-squared:  0.884,  Adjusted R-squared:  0.882
## F-statistic:  586 on 5 and 386 DF,  p-value: <2e-16
```

- **R^2 is 0.8835 when selecting 5 variables, which suggests final model fits the data well.**
- **The p-values of all variables selected are significant. Interpretation of $\hat{\beta}_i$: take horsepower as an example, controlling for other variables, gpm will increase 0.0001081 if there is one unit increase in horsepower.**
  - Predict the mpg of a car that is: built in 1983, in the US, red, 180 inches long, 8 cylinders, 350 displacement, 260 as horsepower, and weighs 4,000 pounds. Give a 95% CI.

```
#car info
car_predict <- Auto_gpm1[1,]
car_predict$gpm = NA
car_predict$cylinders = 8
car_predict$displacement = 350
car_predict$horsepower = 260
car_predict$weight = 4000
car_predict$acceleration = mean(Auto_gpm1$acceleration) #no info about acceleration, use mean of accele
car_predict$year = 83
car_predict$origin2 = as.factor(0)

#predict
car_gpm <- predict(final_fit, car_predict, interval = "confidence", se.fit = TRUE) #predicted gpm
car_mpg <- 1/car_gpm$fit[1]
car_mpg_ci <- c(1/car_gpm$fit[3], 1/car_gpm$fit[2]) #reciprocal of gpm ci
car_mpg
```

```
## [1] 14.8
```

```
car_mpg_ci
```

```
## [1] 13.6 16.2
```

The predicted mpg for this car is 14.8. 95% CI for mpg is (13.6, 16.2).

- Any suggestions as to how to improve the quality of the study?

```
dim(Auto_gpm)
```

```
## [1] 392   8
```

The sample size of the study is 392. More observations can be collected if possible. Also, the study could include more variables, such as Type of Transmission.

5 Case study 2: COVID19

See a seperate file covid_case_study.Rmd for details.