Statistics 206

Homework 7 (Solution)

Due : Nov. 18, 2022, 11:59PM

Instructions:

- You should upload homeworkX files on canvas (under "Assignments/hwX") before its due date.
- Your homework may be prepared by a word processor (e.g., Latex) or through handwriting.
- For handwritten homework, you should either scan or take photos of your homework: Please make sure the pages are clearly numbered and are in order and the scans/photos are complete and clear; Check before submitting.
- Please name the files following the format: "FirstName-LastName-HwX". If there are several files, you can use "-Questions1-5", "-Questions6", etc., to distinguish them. E.g., "Jie-Peng-Hw1-Questions1-5.pdf", "Jie-Peng-Hw1-Questions6.rmd".
- Your name should be clearly shown on the submitted files: By putting on your name, you also acknowledge that you are the person who did and prepared the submitted homework.
- **Optional Problems** are more advanced and are not counted towards the grade.
- Showing/sharing/uploading homework or solutions outside of this class is prohibited.

1. Tell true or false of the following statements.

   (a) To quantify a qualitative variable with three classes C_1, C_2, C_3, we need the following dummy variables:

$$X_1 = \begin{cases} 1 & if \quad C_1 \\ 0 & if \quad otherwise \end{cases} \quad X_2 = \begin{cases} 1 & if \quad C_2 \\ 0 & if \quad otherwise \end{cases} \quad X_3 = \begin{cases} 1 & if \quad C_3 \\ 0 & if \quad otherwise \end{cases}$$

ANS. FALSE. We only need X_1 and X_2. C_3 is represented by both $X_1 = X_2 = 0$. Indeed, $X_1 + X_2 + X_3 \equiv 1$, so three of them are in perfect intercorrelation. If all three are included in a model, the LS estimators will not be defined.

   (b) Polynomial regression models with higher than the third power terms are preferred since they provide better approximations to the regression relation.

ANS. FALSE. Polynomial regression models with higher-order powers could be highly variable and hard to generalize.

(c) In interaction regression models, the effect of one variable depends on the value of another variable with which it appears together in a cross-product term.

ANS. TRUE.

(d) With a qualitative variable, the best way is to fit separate regression models under each of its classes.

ANS. FALSE. This usually would not be as efficient as fitting one regression model using indicator variables due to loss of degrees of freedom (since each class will have a smaller sample size and more parameters are being fitted).

2. **(Cars) Exploratory Data Analysis**. You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

   (a) Conduct a visual inspection of the data in "Cars.csv" and then read the data into R.

   (b) Are there missing values? If so, replace missing values by "NA".

   (c) Check the variable types. Which variables do you think should be treated as quantitative and which should be treated as qualitative/categorical? Fix the problems that you have identified (if any).

   (d) Draw histogram for each quantitative variable. Comment on their distributions.

   (e) Draw scatter plot matrix among quantitative variables with the lower panel showing correlation coefficients. Comment on their relationships.

   (f) Draw pie chart (with class percentage) for each categorical variable.

   (g) Draw side-by-side box plots for "mpg" with respect to each categorical variable. What do you observe?

3. **(Cars Cont'd) Regression with Categorical Variables**. In this question, we consider models for "mpg" using "cylinders", "horsepower", and "weight" as predictors, where "cylinders" should be treated as a categorical variable. You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

   (a) Decide on whether you'd like to make any transformation of the "mpg".

   (b) Fit a first-order model with the (transformed) variables. Conduct model diagnostics. Does this model appear to be adequate?

   (c) Derive the regression function for cars with 4 cylinders.

   (d) Fit a model including interactions between "cylinders" and "horsepower", and, "cylinders" and "weight". Derive the regression function for cars with 4 cylinders.

   (e) Compare the two models using the function `anova()`. What do you find?

(f) Construct a 95% prediction interval of "mpg" for a car with 4 cylinders, 100 horsepower and 3000 pounds under these two models. What do you observe?

4. **(Optional problem). Regression coefficients as partial coefficients.** Let $X = (X_1, X_2)$ where $X_1 \in \mathbb{R}^{n\times s}, X_2 \in \mathbb{R}^{n\times t}$. Write the LS fitted regression coefficients as $\hat{\beta} = \begin{pmatrix} \hat{\beta}^{(1)} \\ \hat{\beta}^{(2)} \end{pmatrix}$. Show that:

(a) The LS fitted regression coefficients of X_2 is

$$\hat{\beta}^{(2)} = (\tilde{X}_2^T\tilde{X}_2)^{-1}\tilde{X}_2^T Y = (\tilde{X}_2^T\tilde{X}_2)^{-1}\tilde{X}_2^T(Y - \hat{Y}(X_1)), \quad \tilde{X}_2 = X_2 - \hat{X}_2(X_1),$$

i.e., $\hat{\beta}^{(2)}$ **is the LS fitted regression coefficients by regressing** Y **(or** $Y - \hat{Y}(X_1)$**) onto** $X_2 - \hat{X}_2(X_1)$**.** Such coefficients are called **partial coefficients.**

Proof. Since $X_2 - \hat{X}_2(X_1)$ is orthogonal to $span\{X_1\}$, and $\hat{Y}(X_1)$ is in the space of $span\{X_1\}$, we can show that $\tilde{X}_2^T\hat{Y}(X_1) = 0$. This tells us that $(\tilde{X}_2^T\tilde{X}_2)^{-1}\tilde{X}_2^T Y = (\tilde{X}_2^T\tilde{X}_2)^{-1}\tilde{X}_2^T(Y - \hat{Y}(X_1))$.
Now, to show $\hat{\beta}^{(2)} = (\tilde{X}_2^T\tilde{X}_2)^{-1}\tilde{X}_2^T Y$. We know

$$X^T X = \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix}.$$

Denote

$$(X^T X)^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where $B_{22} = (X_2^T X_2 - X_2^T H(X_1) X_2)^{-1}$, and $B_{21} = -B_{22} X_2^T X_1 (X_1^T X_1)^{-1}$. (Remember the **Inverse of a partitioned matrix** from hw6. Suppose A is a $(p+q)\times(p+q)$ square matrix $(p, q \geq 1)$:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} is a $p \times p$ square matrix and A_{22} is a $q \times q$ square matrix. Suppose A_{11} and A_{22} are invertible. Then A is invertible and

$$A^{-1} = \begin{bmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} A_{12}A_{22}^{-1} \\ -\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} A_{21}A_{11}^{-1} & \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{bmatrix}.)$$

Since $\hat{\beta} = (X^T X)^{-1} X^T Y$, we have

$$\begin{aligned} LHS &= B_{21}X_1^T Y + B_{22}X_2^T Y \\ &= B_{22}(X_2^T - X_2^T X_1 (X_1^T X_1)^{-1} X_1^T) Y \\ &= B_{22} X_2^T (I - H(X_1)) Y \\ &= (\tilde{X}_2^T\tilde{X}_2)^{-1}\tilde{X}_2^T Y \\ &= RHS. \end{aligned}$$

□

(b) If $X_1 \perp X_2$ (i.e., the columns of X_1 and the columns of X_2 are orthogonal), then

$$\hat{\beta}^{(2)} = (X_2^T X_2)^{-1} X_2^T Y, \quad if, \quad X_1 \perp X_2,$$

i.e., the LS fitted regression coefficients by regressing Y onto X_2 alone.

Proof. $X_1 \perp X_2$ indicates that $X_1^T X_2 = 0$. Thus,

$$(X^T X)^{-1} = \begin{pmatrix} (X_1^T X_1)^{-1} & 0 \\ 0 & (X_2^T X_2)^{-1} \end{pmatrix}.$$

Plug it into the expression of $\hat{\beta}$, we can obtain

$$\hat{\beta}^{(2)} = (X_2^T X_2)^{-1} X_2^T Y.$$

□

5. **(Optional problem). Simultaneous confidence bands of the regression function.** Under the Normal error model, derive the simultaneous confidence bands of the regression function by the following steps.

(a) Show that

$$\frac{(\hat{\beta} - \beta)^T (X^T X)(\hat{\beta} - \beta)}{MSE} \sim pF_{p,n-p}.$$

Proof. Under the normal error model, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$. Thus,

$$(X^T X)^{1/2} (\hat{\beta} - \beta) \sim \mathcal{N}(0, \sigma^2 I),$$

and

$$(\hat{\beta} - \beta)^T (X^T X)(\hat{\beta} - \beta) \sim \sigma^2 \chi_p^2.$$

From the lecture notes, we know that $pMSE \sim \sigma^2 \chi_{n-p}^2$. Since

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}, (I - H)Y) &= \mathrm{Cov}(HY, (I - H)Y) \\ &= 0, \end{aligned}$$

and $\hat{\beta}$ and $(I - H)Y$ are jointly normal, $\hat{\beta}$ and $(I - H)Y$ are independent. Thus,

$$\frac{(\hat{\beta} - \beta)^T (X^T X)(\hat{\beta} - \beta)}{MSE} \sim pF_{p,n-p}.$$

□

(b) Show that for a constant $C \geq 0$, $|x^T \beta - x^T \hat{\beta}| \leq \sqrt{C x^T (X^T X)^{-1} x}$ for all $x \in \mathbb{R}^p$ if and only if $(\hat{\beta} - \beta)^T (X^T X)(\hat{\beta} - \beta) \leq C$.

Proof. Let $y = (X^TX)^{-1/2}x$. Using Cauchy-Schwarz inequality, we have

$$\begin{aligned}|x^T\beta - x^T\hat{\beta}|^2 &= |y^T(X^TX)^{1/2}(\hat{\beta}-\beta)|^2 \\ &\le (y^Ty)(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta) \\ &= x^T(X^TX)^{-1}x(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta).\end{aligned}$$

Since the equality can be achieved, this suggests the equivalency. □

(c) Show that the $(1-\alpha)$ simultaneous confidence bands for the regression function, $x^T\beta, x \in \mathbb{R}^p$, are:

$$x^T\hat{\beta} \pm \sqrt{pF(1-\alpha;p,n-p))}\sqrt{MSEx^T(X^TX)^{-1}x}, \quad x \in \mathbb{R}^p,$$

i.e.,

$$P(x^T\beta \in x^T\hat{\beta} \pm \sqrt{pF(1-\alpha;p,n-p))}\sqrt{MSEx^T(X^TX)^{-1}x}, \text{ for all } x \in \mathbb{R}^p) = 1-\alpha.$$

Proof. To show the probability equals $1-\alpha$.

$$\begin{aligned}LHS = P(&|x^T\beta \\ &- x^T\hat{\beta}| \le \sqrt{pF(1-\alpha;p,n-p))}\sqrt{MSEx^T(X^TX)^{-1}x}, \quad \text{for all } x) \\ &= P((\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta) \le pF(1-\alpha;p,n-p)MSE) \quad \text{Using (b)} \\ &= P(\frac{(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta)}{MSE} \le pF(1-\alpha;p,n-p)) \\ &= P(F_{p,n-p} \le F(1-\alpha;p,n-p)) \quad \text{Using (a)} \\ &= 1-\alpha.\end{aligned}$$

□

HW7_Question23

Wookyeong Song (most of them from Yan-Yu Chen)

2022/11/18

2. (Cars) Exploratory Data Analysis

You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

(a) Conduct a visual inspection of the data in "Cars.csv" and then read the data into R.

```
cars <- read.csv('Cars.csv', header=TRUE, na.strings = "?")
summary(cars)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.0   1st Qu.: 75.0   1st Qu.:2223
##  Median :23.00   Median :4.000   Median :146.0   Median : 93.5   Median :2800
##  Mean   :23.52   Mean   :5.458   Mean   :193.5   Mean   :104.5   Mean   :2970
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0   3rd Qu.:126.0   3rd Qu.:3609
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##                                                  NA's   :5
##   acceleration    country.code
##  Min.   : 8.00   Min.   :1.000
##  1st Qu.:13.80   1st Qu.:1.000
##  Median :15.50   Median :1.000
##  Mean   :15.56   Mean   :1.574
##  3rd Qu.:17.10   3rd Qu.:2.000
##  Max.   :24.80   Max.   :3.000
##
```

(b) Are there missing values? If so, replace missing values by `NA`.

There are 5 missing values in `horsepower`. They have been replaced in (a).

(c) Check the variable types. Which variables do you think should be treated as quantitative and which should be treated as qualitative/categorical? Fix the problems that you have identified (if any).

```
sapply(cars, class)
```

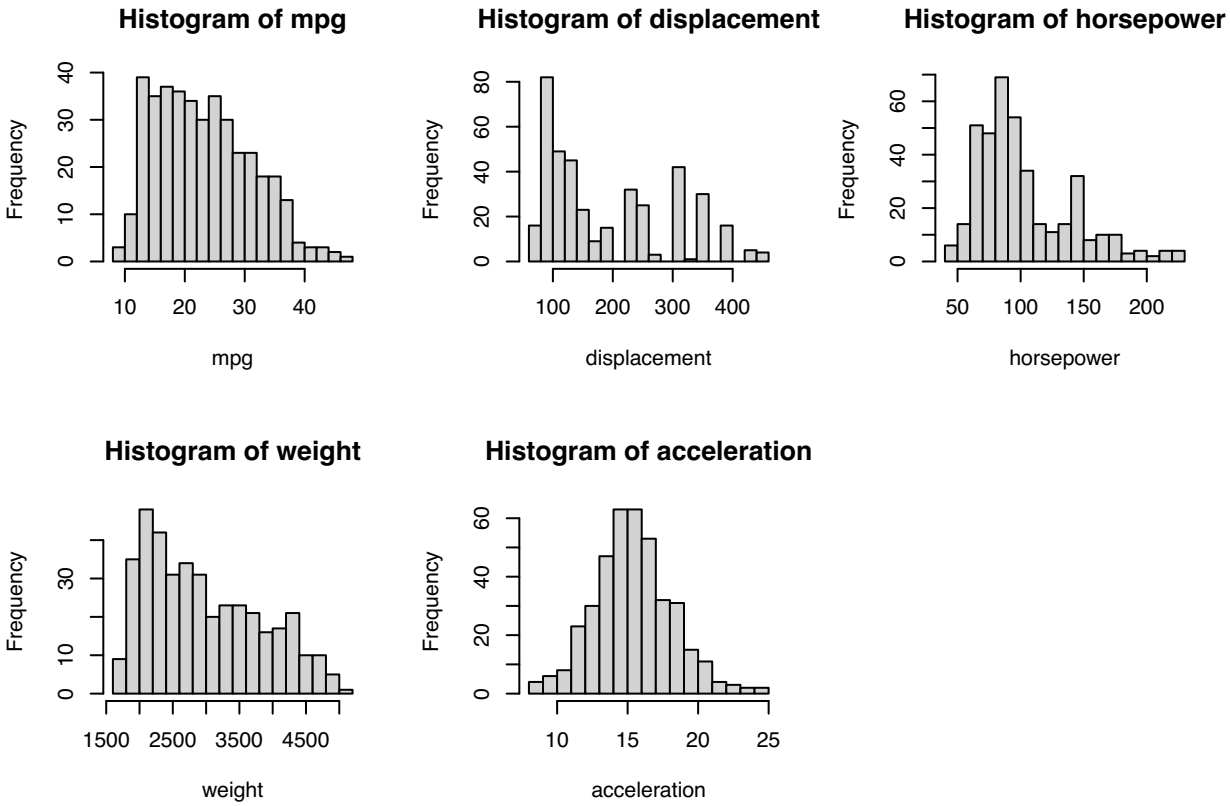
```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    "numeric"    "integer"    "numeric"    "integer"    "integer"    "numeric"
## country.code
##    "integer"
```

`country.code` only takes value 1, 2, 3. `cylinders` only takes value 3,4,5,6,8. They should be treated as categorical variables. The others should be treated as quantitative.

```
cars$cylinders<-as.factor(cars$cylinders)
cars$country.code<-as.factor(cars$country.code)
```

(d) Draw histogram for each quantitative variable. Comment on their distributions.

```
par(mfrow=c(2,3))
with(cars,{
  hist(mpg, breaks=15)
  hist(displacement, breaks=15)
  hist(horsepower, breaks=15)
  hist(weight, breaks=15)
  hist(acceleration, breaks=15)
})
par(mfrow=c(1,1))
```

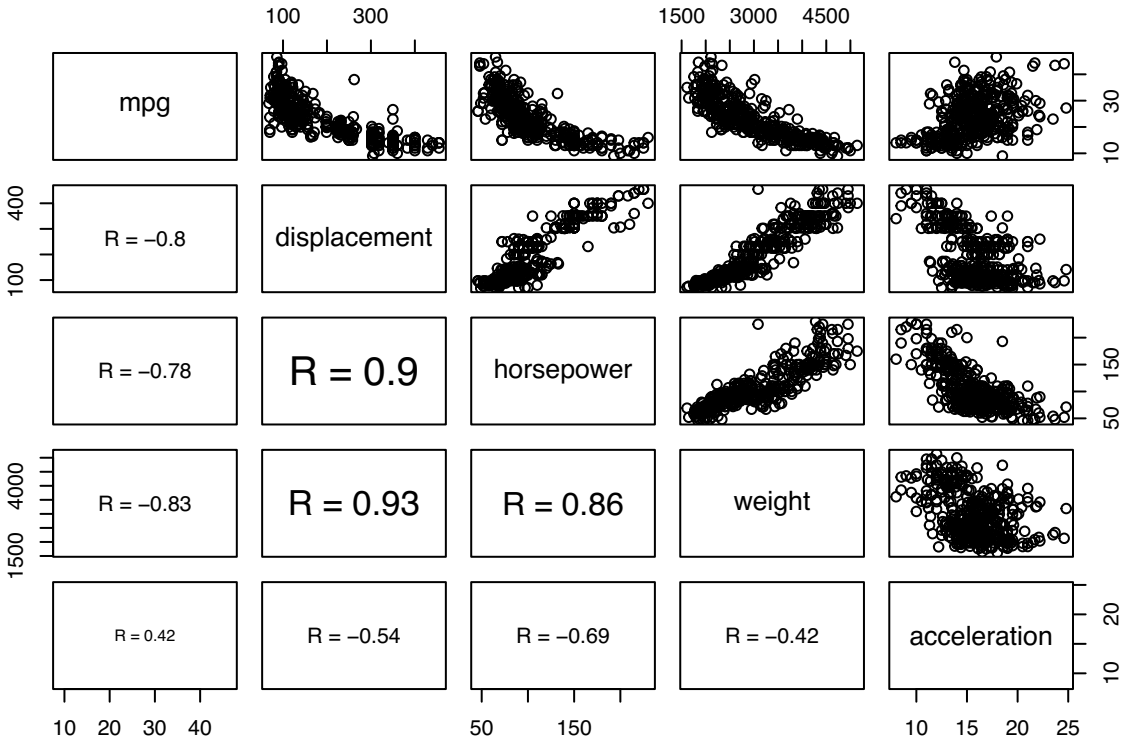


The histogram of `mpg` is right-skewed. `displacement` and `horsepower` have similar bimodal shapes. `Weight` is also right-skewed.

(e) Draw scatter plot matrix among quantitative variables with the lower panel showing correlation coefficients. Comment on their relationships.

```
panel.cor <- function(x, y){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y,use="na.or.complete"), digits=2)
  txt <- paste0("R = ", r)
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(cars[,c("mpg","displacement","horsepower","weight","acceleration")], lower.panel = panel.cor)
```



Comment: `displacement`, `horsepower`, and `weight` are strongly correlated with each other. `mpg` has a negative correlation with `displacement`, `horsepower`, and `weight`, while a positive correlation with `acceleration`. However, their relationships seem slightly non-linear.

(f) Draw pie chart (with class percentage) for each categorical variable.

```
n = nrow(cars)
par(mfrow=c(1,2))
lbls=names(table(cars$cylinders))
pct=round(100*table(cars$cylinders)/n)
lab=paste("#",lbls,": ", pct, "%", sep='')
lab
```

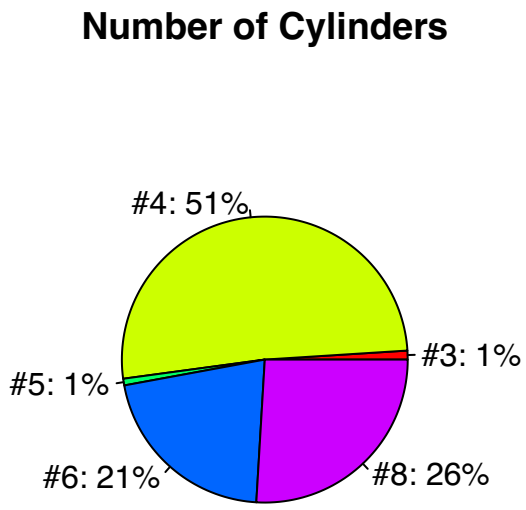
```
## [1] "#3: 1%"  "#4: 51%" "#5: 1%"  "#6: 21%" "#8: 26%"
```

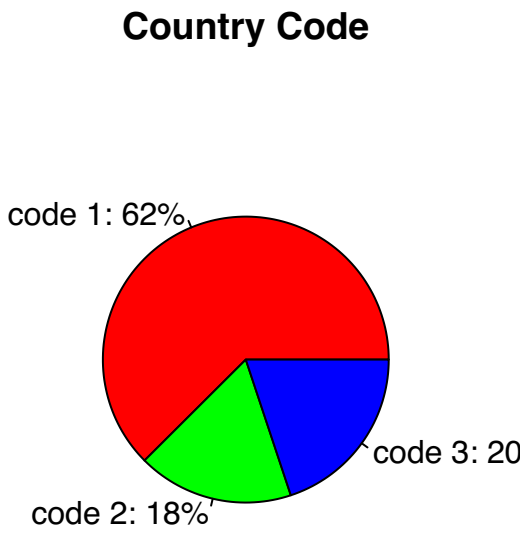
```
pie(table(cars$cylinders),labels=lab,col=rainbow(length(lab)),
    main='Number of Cylinders')


lbls=names(table(cars$country.code))
pct=round(100*table(cars$country.code)/n)
lab=paste("code ", lbls,": ", pct, "%", sep='')
lab
```

```
## [1] "code 1: 62%" "code 2: 18%" "code 3: 20%"
```

```
pie(table(cars$country.code),labels=lab,col=rainbow(length(lab)),
    main='Country Code')
```



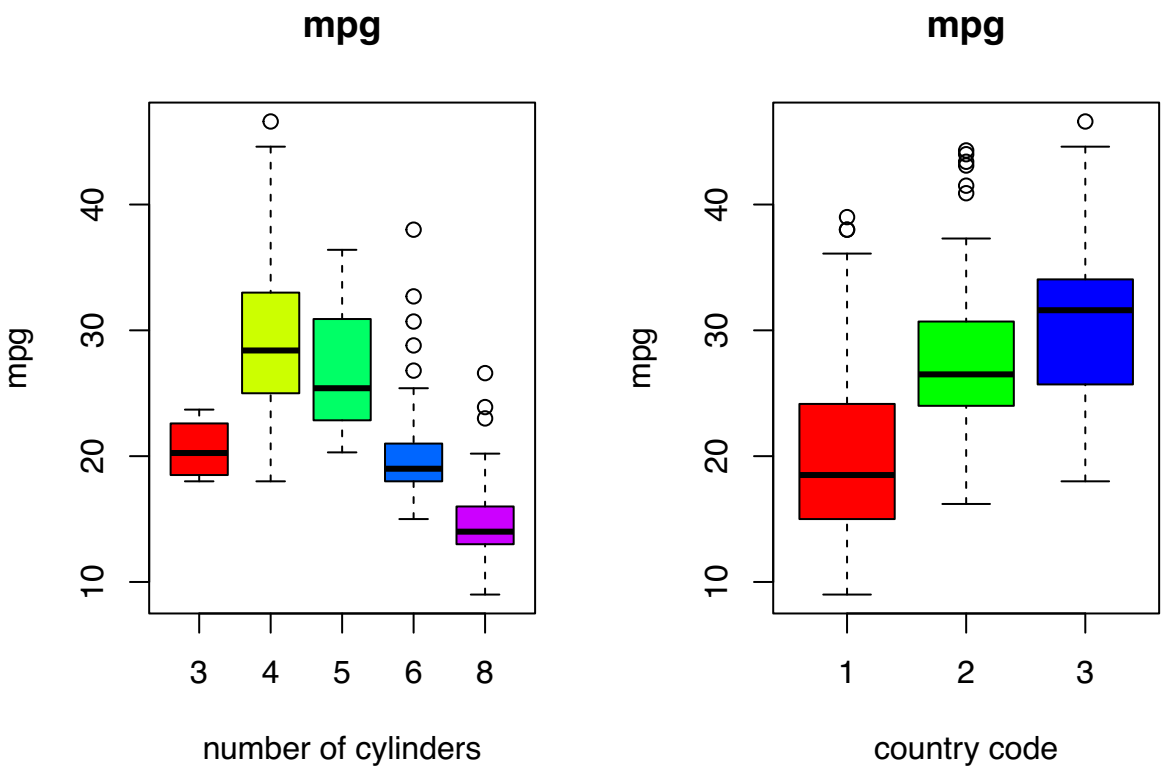



```
par(mfrow=c(1,1))
```

(g) Draw side-by-side box plots for "mpg" with respect to each categorical variable. What do you observe?

```
par(mfrow=c(1,2))
boxplot(cars$mpg~cars$cylinders,main='mpg',
        xlab='number of cylinders',ylab='mpg',col=rainbow(length((levels(cars$cylinders)))))

boxplot(cars$mpg~cars$country.code,main='mpg',
        xlab='country code',ylab='mpg',col=rainbow(length((levels(cars$country.code)))))
```




```
par(mfrow=c(1,1))
```

Cars with 3, 6, and 8 cylinders tend to have lower mpg compared to others. Country code 3 appears to have higher mpg, followed by 2 followed by 1.

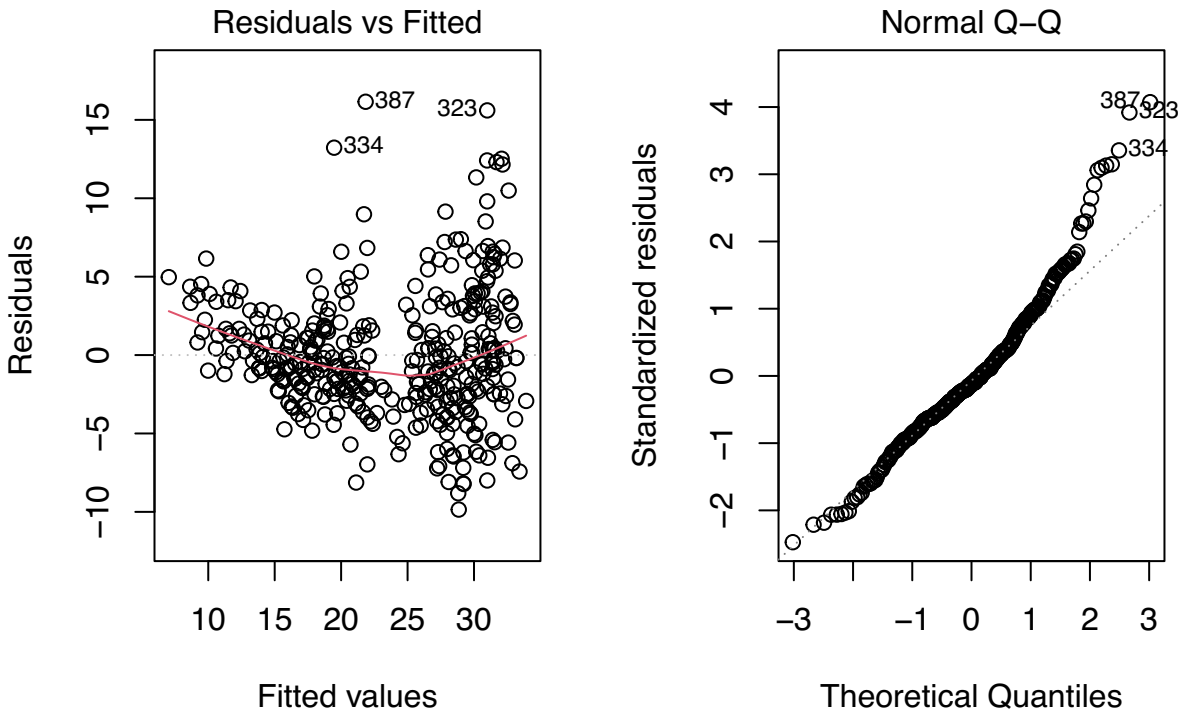
3. (Cars Cont'd) Regression with Categorical Variables

In this question, we consider models for `mpg` using `cylinders`, `horsepower`, and `weight` as predictors, where `cylinders` should be treated as a categorical variable. You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

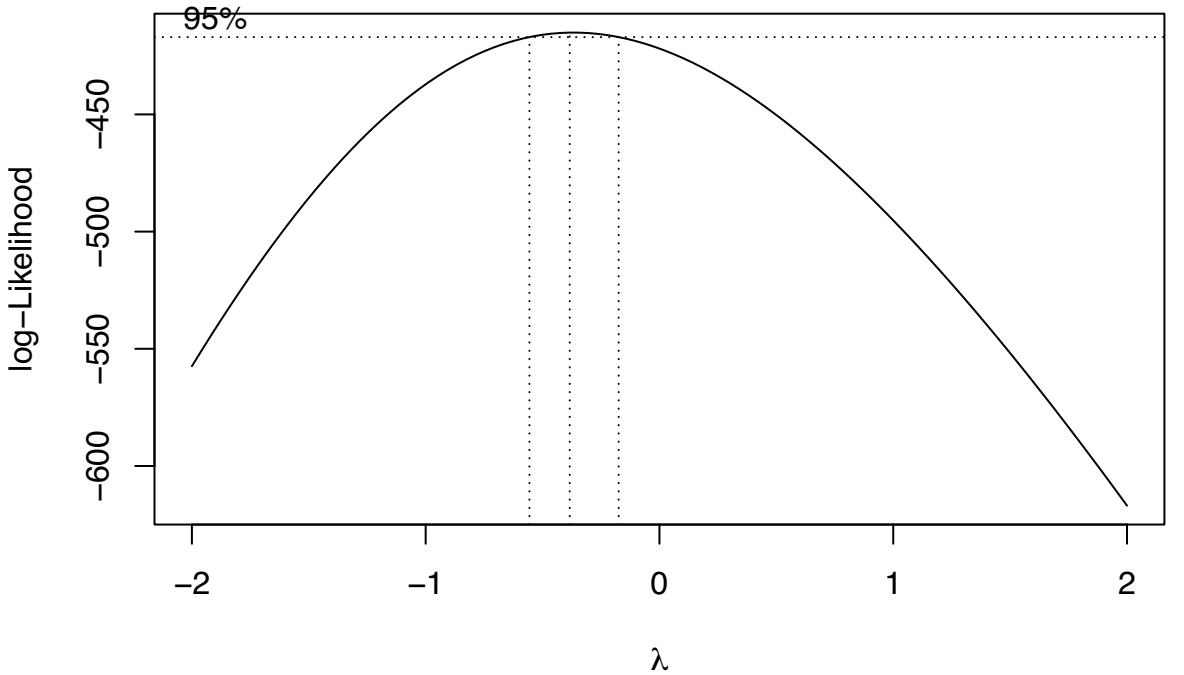
(a) Decide on whether you'd like to make any transformation of the `mpg`.

```
attach(cars)
fit1<-lm(mpg~cylinders+horsepower+weight)

par(mfrow=c(1,2))
plot(fit1, which=1:2)
```



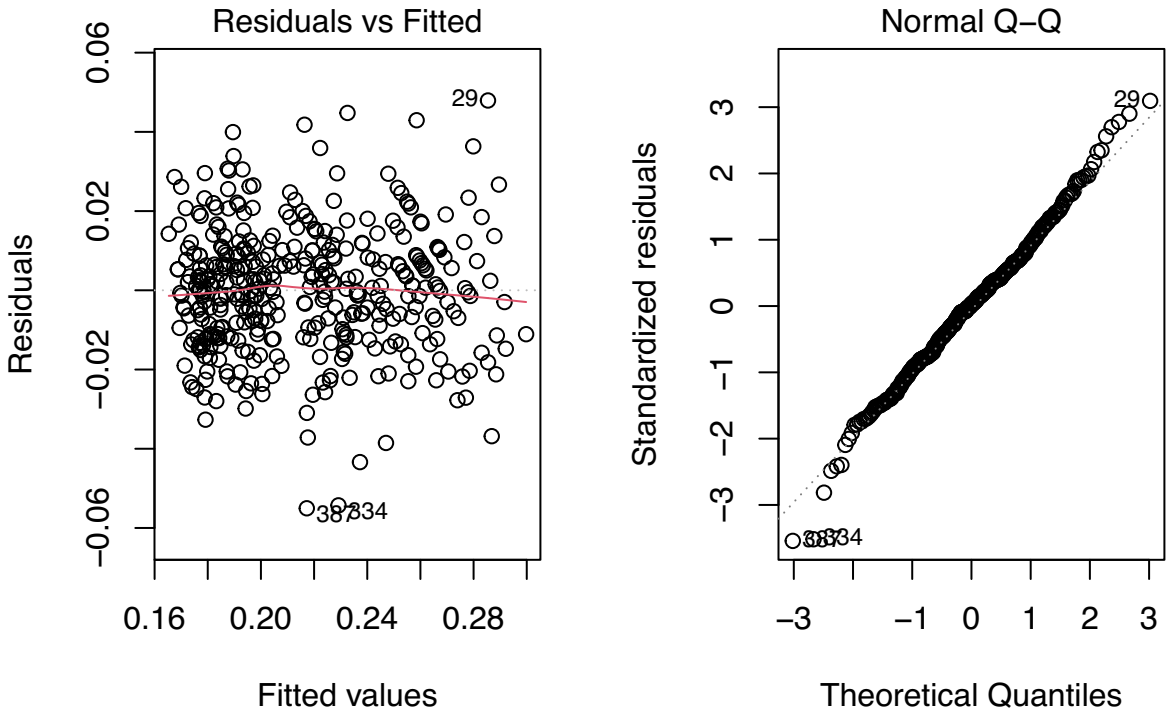

```
par(mfrow=c(1,1))
MASS::boxcox(fit1)
```



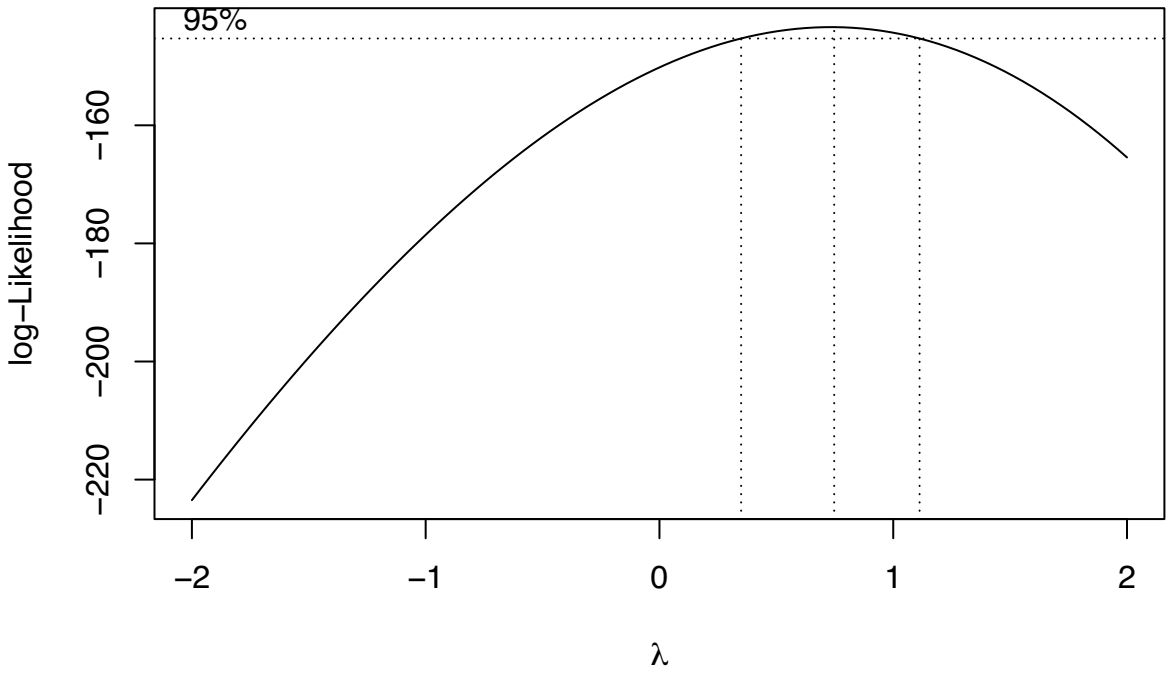
The Box-Cox likelihood suggests a power transformation with power -0.5.

(b) Fit a first-order model with the (transformed) variables. Conduct model diagnostics. Does this model appear to be adequate?

```
fit2<-lm(mpg^{-0.5}~cylinders+horsepower+weight)
par(mfrow=c(1,2))
plot(fit2, which=1:2)
```




```
par(mfrow=c(1,1))
MASS::boxcox(fit2)
```




```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg^{
##     -0.5
## } ~ cylinders + horsepower + weight)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055028 -0.011058  0.000403  0.009268  0.047923
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.402e-01  9.556e-03  14.667  < 2e-16 ***
## cylinders4  -2.588e-02  7.952e-03  -3.254  0.00124 **
## cylinders5  -3.620e-02  1.215e-02  -2.979  0.00307 **
## cylinders6  -1.323e-02  8.231e-03  -1.607  0.10882
## cylinders8  -1.448e-02  8.818e-03  -1.642  0.10132
## horsepower   3.016e-04  4.703e-05   6.413  4.2e-10 ***
## weight       2.146e-05  2.427e-06   8.842  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01564 on 385 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.8245, Adjusted R-squared:  0.8218
## F-statistic: 301.5 on 6 and 385 DF,  p-value: < 2.2e-16
```

By examining the residual plots and the R^2, this model appears to be reasonable. In particular, $R^2 = 0.8245$ is much larger than in the original model (0.7077).

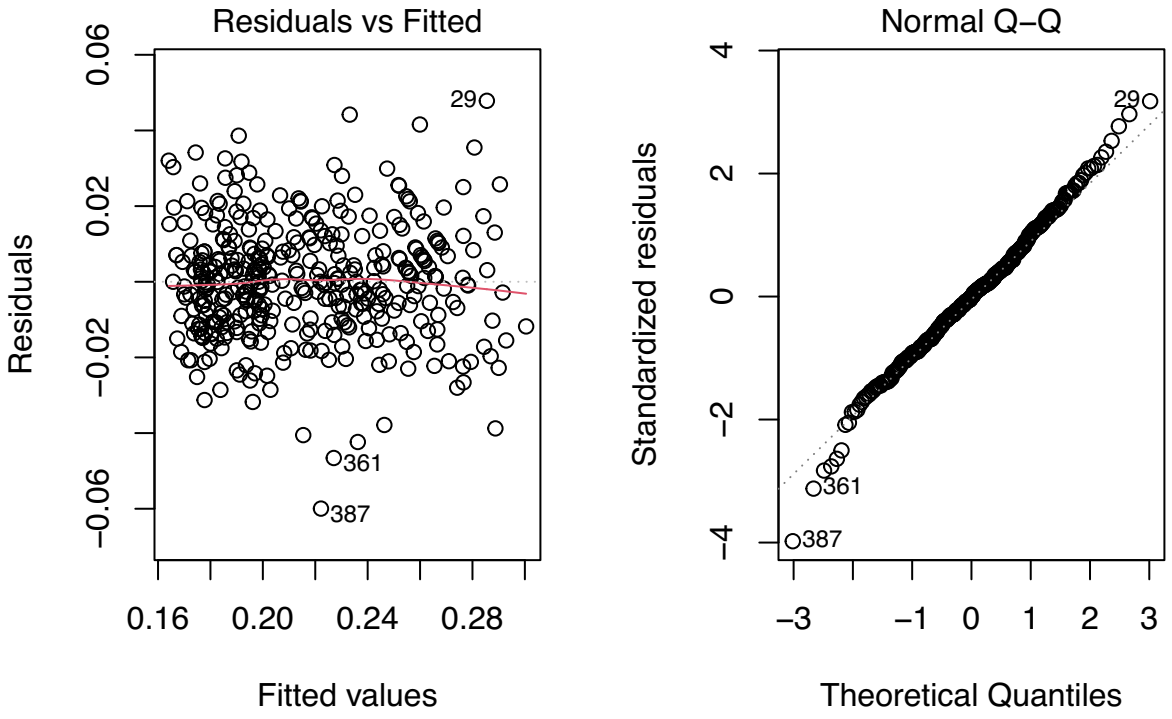
(c) Derive the regression function for cars with 4 cylinders.

For cars with 4 cylinders, $\texttt{mpg} = 0.1143 + 3.0158 \times 10^{-4}$ `horsepower` $+ 2.146 \times 10^{-5}$ `weight`

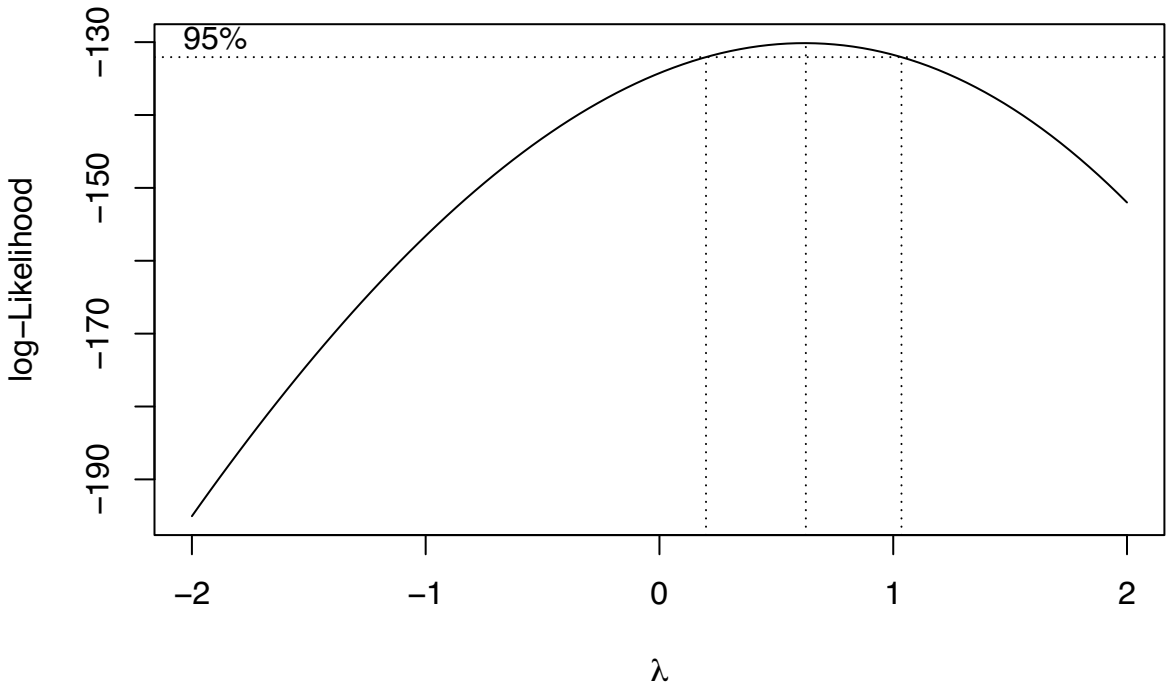
(d) Fit a model including interactions between `cylinders` and `horsepower`, and, `cylinders` and `weight`. Derive the regression function for cars with 4 cylinders.

```
fit3<-lm(mpg^{-0.5}~cylinders+horsepower+weight+cylinders:horsepower+cylinders:weight)
par(mfrow=c(1,2))
plot(fit3, which=1:2)
```

```
## Warning: not plotting observations with leverage one:
##   111, 273, 296, 326
```




```
par(mfrow=c(1,1))
MASS::boxcox(fit3)
```




```
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg^{
##     -0.5
## } ~ cylinders + horsepower + weight + cylinders:horsepower +
##     cylinders:weight)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.059962 -0.009924  0.000000  0.008757  0.047836
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.054183   3.706702   0.554    0.580
## cylinders4              -1.940515   3.706709  -0.524    0.601
## cylinders5              -2.081293   3.708930  -0.561    0.575
## cylinders6              -1.911858   3.706753  -0.516    0.606
## cylinders8              -1.926868   3.706729  -0.520    0.603
## horsepower              -0.092746   0.196876  -0.471    0.638
## weight                   0.003074   0.006602   0.466    0.642
## cylinders4:horsepower    0.093250   0.196876   0.474    0.636
## cylinders5:horsepower    0.094400   0.196877   0.479    0.632
## cylinders6:horsepower    0.092667   0.196876   0.471    0.638
## cylinders8:horsepower    0.093078   0.196876   0.473    0.637
## cylinders4:weight       -0.003059   0.006602  -0.463    0.643
## cylinders5:weight       -0.003046   0.006602  -0.461    0.645
## cylinders6:weight       -0.003045   0.006602  -0.461    0.645
## cylinders8:weight       -0.003054   0.006602  -0.463    0.644
##
```

```
## Residual standard error: 0.01531 on 377 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.8354, Adjusted R-squared:  0.8293
## F-statistic: 136.7 on 14 and 377 DF,  p-value: < 2.2e-16
```

For cars with 4 cylinders, `mpg` $= 0.1137 + 5.0379 \times 10^{-4}$ `horsepower` $+ 1.485 \times 10^{-5}$ `weight`

(e) Compare the two models using the function `anova()`. What do you find?

```
anova(fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg^{
##     -0.5
## } ~ cylinders + horsepower + weight
## Model 2: mpg^{
##     -0.5
## } ~ cylinders + horsepower + weight + cylinders:horsepower +
##     cylinders:weight
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1    385 0.094230
## 2    377 0.088374  8 0.0058556 3.1225 0.00198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The overall effects of the interaction terms are significant.

(f) Construct a 95% prediction interval of `mpg` for a car with 4 cylinders, 100 horsepower and 3000 pounds under these two models. What do you observe?

```
newX=data.frame("cylinders"="4", "horsepower"=100, "weight"=3000)
1/predict(fit2, newX, interval="prediction")^2
```

```
##        fit      lwr      upr
## 1 22.93608 31.62339 17.39251
```

```
1/predict(fit3, newX, interval="prediction")^2
```

```
##        fit      lwr      upr
## 1 22.97944 31.52113 17.49192
```

Predictions given by the two models are close. This is what we expected since the prediction models for 4-cylinder car in (b) and (d) are similar.