# CSC465HW01

*Kaiyuan Hu*

*2017/2/12*

## Question 1

### a)

Adding a constant $c$ in each response $Y$ means we can rewite the original model as $Y + C = \beta_0 + \beta_1 X + C$, which is $Y + C = (\beta_0 + C) + \beta_1 X$. So, the new $\beta_1$ is equal to the previous one and th new $\beta_0$ is equal to the original $\beta_0 + c$.

### b)

For the new model:

$$\hat{Y} + C = \hat{\beta}_0^{'} + \hat{\beta}_1^{'} X_{new}$$
$$\hat{Y} = \hat{\beta}_0^{'} + \hat{\beta}_1{'} X_{new} - C$$
$$\hat{\beta}_1{'} X_{new} - C = \hat{\beta}_1 X$$
$$\hat{\beta}_1{'} X_{new} = \hat{\beta}_1 X + C$$
$$X_{new} = (\hat{\beta}_1 X + C)/\hat{\beta}_1{'}$$

---

## Question 2

### a)

$$trace(AB^T) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij} b_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji} a_{ij} = trace(B^T A)$$

Then:

$$trace(H) = trace(X(X^T X)^{-1} X^T) = trace(X^T X (X^T X)^{-1}) = trace(I_{q*q}) = q$$

### b)

#### I)

$$H * H = X(X^T X)^{-1} X^T * X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} * I * X^T = H$$

So matrix $H$ is idempotent.

#### II)

We assume $H \neq HH$, which means there exist $n$ such that $H^n y$ is the closet point in $S_q$ closet to $y$. This statement is contradict to the fact that $Hy$ is the closet point in $S_q$ closet to $y$. So, $H = HH$.

## c)

### I)

$$\mathbf{H}^{'} = \begin{bmatrix} 1/n & \cdots & 1/n \\ \cdots & \cdots & \cdots \\ 1/n & \cdots & 1/n \end{bmatrix}_{n*n}$$

$trace(H^{'}) = 1$

### II)

$$\mathbf{H}^{'} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 & \cdots \\ 1/3 & 1/3 & 1/3 & 0 & 0 & \cdots \\ 0 & 1/3 & 1/3 & 1/3 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & \cdots & 1/2 & 1/2 \end{bmatrix}$$

$trace(H^{'}) = 1 + (n-2)/3$

### III)

$$\mathbf{H}^{'} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 1 \end{bmatrix}_{n*n}$$

$trace(H^{'}) = n$

## d)

When $n = 2$, the order of complexity is $I = II < III$ When $n > 2$, the order of complexity is $I < II < III$

The degree of freedom of $I$ is 1, so the model complexity is low. For model $II$, as the complexity increase the $MSE$ will decrease. For model $III$, the $MSE$ is zero, this model is really complexity.

---

# Question 3

```r
library(MASS)
```

## a)

```r
modela = lm(bwt~age, data = birthwt)
```

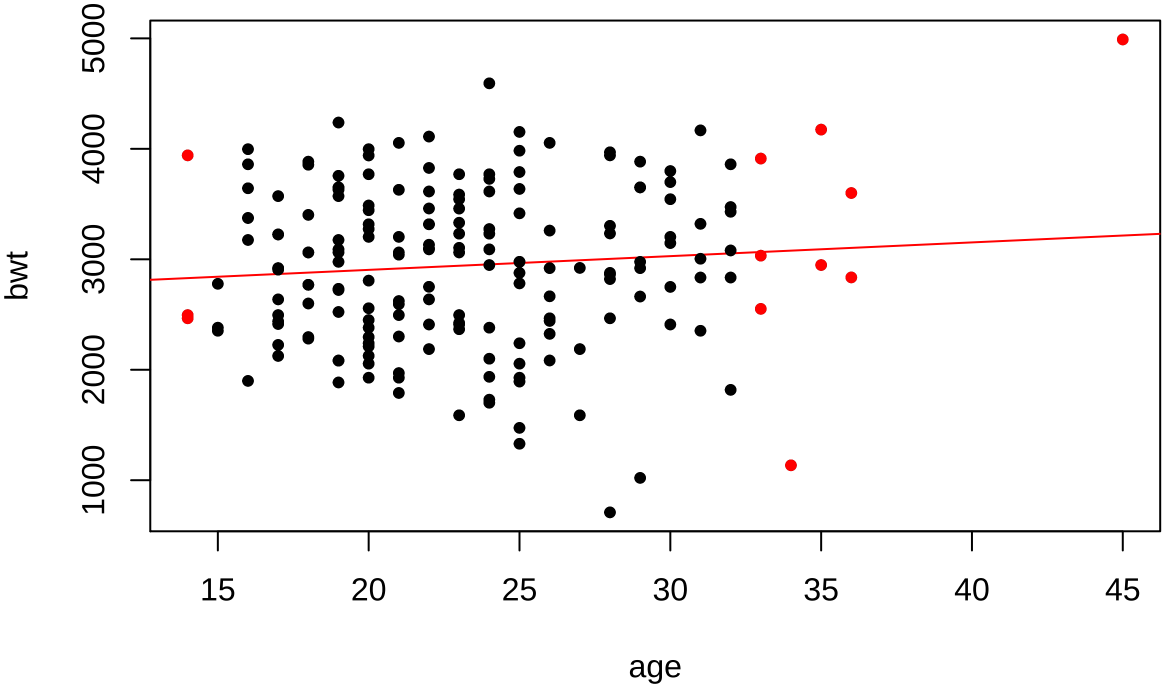
### b)

```
inf_measure = influence.measures(modela)
```

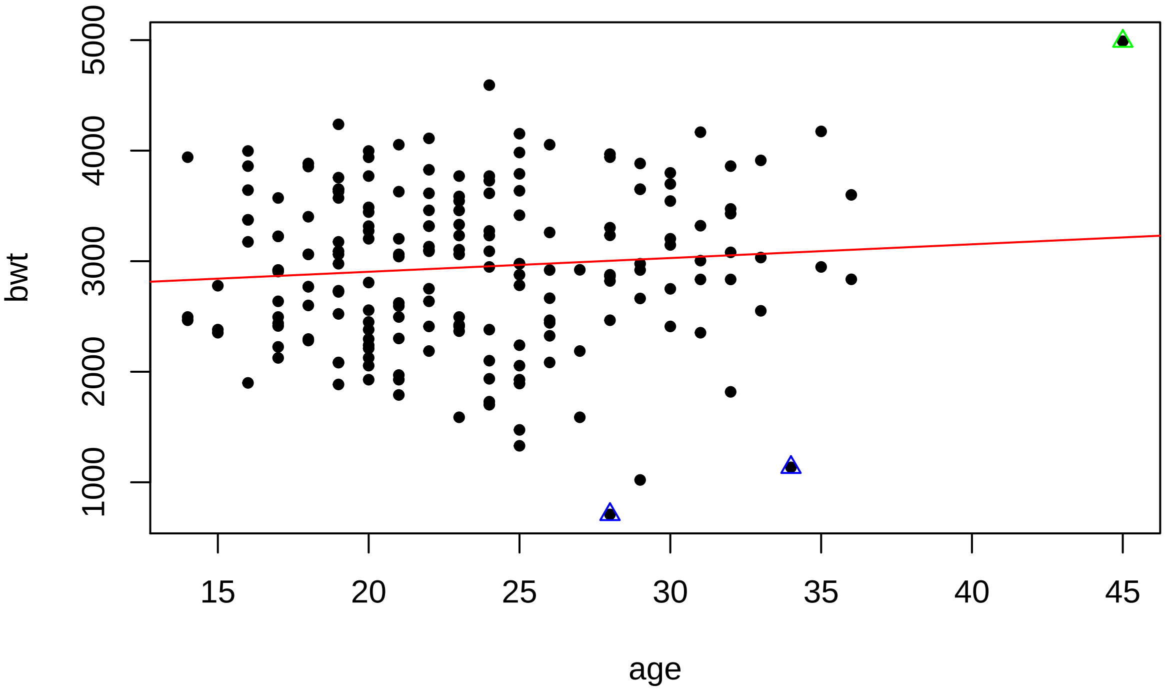
### c)

#### I&II&III&IV&V)

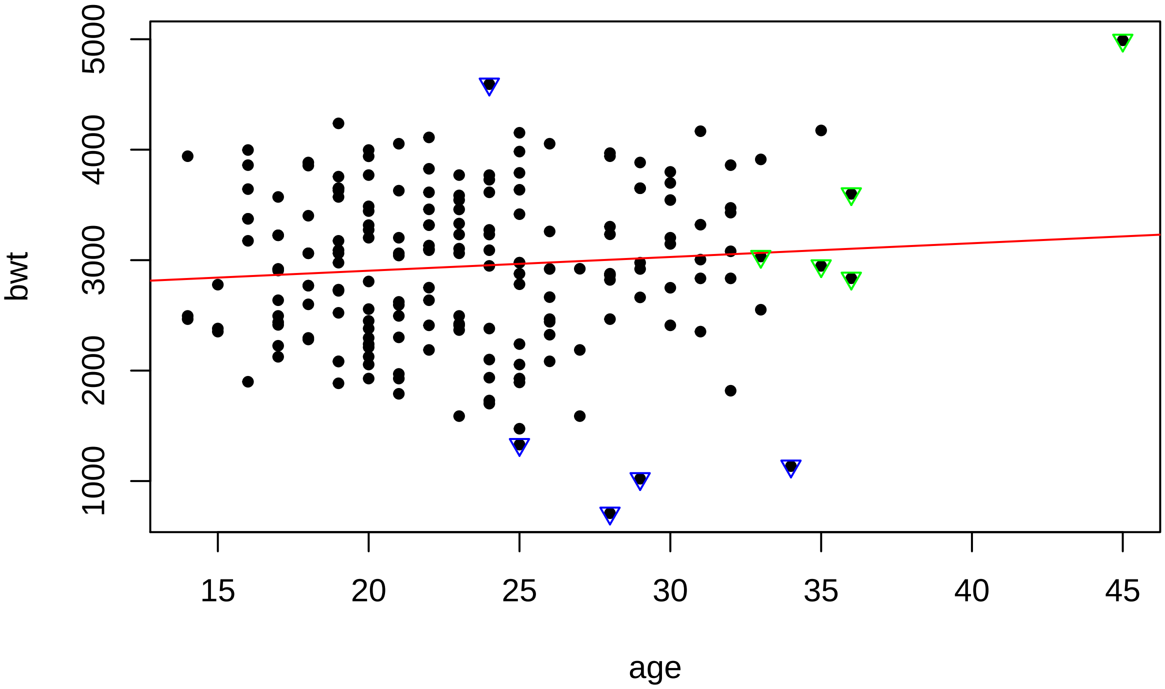
```
plot(birthwt$bwt~birthwt$age, pch = 20, xlab = 'age', ylab = 'bwt') # I
abline(modela, col = 'red') # II
flag = inf_measure$infmat
points(birthwt$age, birthwt$bwt, col = ifelse(flag[,6] > 4/dim(birthwt)[1], 'red', 'black'), pch = 20)
```



```
plot(birthwt$bwt~birthwt$age, pch = 20, xlab = 'age', ylab = 'bwt') # I
abline(modela, col = 'red')
points(birthwt$age[inf_measure$is.inf[,'dffit'] & inf_measure$infmat[,'dffit'] < 0], birthwt$bwt[inf_mea
points(birthwt$age[inf_measure$is.inf[,'dffit'] & inf_measure$infmat[,'dffit'] >= 0], birthwt$bwt[inf_me
```



```
plot(birthwt$bwt~birthwt$age, pch = 20, xlab = 'age', ylab = 'bwt') # I
abline(modela, col = 'red')# IV
points(birthwt$age[inf_measure$is.inf[,'cov.r'] & inf_measure$infmat[,'cov.r'] < 1], birthwt$bwt[inf_mea
points(birthwt$age[inf_measure$is.inf[,'cov.r'] & inf_measure$infmat[,'cov.r'] >= 1], birthwt$bwt[inf_me
```



### d)

It seems the fitted line distinguishes high and low *DFFITSi*. High *DFFITSi* value usually above the fitted line and low *DFFITSi* value usually below the fitted line.

### e)

No. As we can see in the figure, no flag criterion is dominates. We need them all to detect outliers.

### f)

As we can see in the figure, If the bwt value is higher than 4000 or lower than 2000 or age large than 35, observations will have high covariance ratio.

If a value is too far away from most other values, remove this point, the standard error of $\beta$ will decrease.

### g)

$X_{,,}$ yield lower varience, since $(X_{,,}^T X_{,,})^{-1} < (X_{,}^T X_{,})^{-1}$.

### h)

The large sample variance will cause small beta varience.

### i)

No. Coveriance ratio it alone can not tell which point is really a outlier, since coveriance ratio only compute the distance between points. However, far distance does not necessaily means it is a outlier.

---

## Question 4

### a)

```
newbirthwt = sub <- subset(birthwt,!inf_measure$is.inf[,'dffit'])
```

### b)

```
summary.table = function(fit){
    Nonsmoke.Intercept = summary(fit)$coef[1,1]
    Nonsmoke.slope = 0
    Smoke.Intercept = summary(fit)$coef[2,1] + Nonsmoke.Intercept
    Smoke.slope = 0

    if(i == 1){
      Nonsmoke.slope = summary(fit)$coef[2,1]
      Smoke.Intercept = NA
      Smoke.slope = NA
    }

    if(i == 2){
      Smoke.Intercept = Nonsmoke.Intercept + summary(fit)$coef[2,1]
    }

    if(i == 3){
      Smoke.Intercept = Smoke.slope = summary(fit)$coef[3,1]
    }

    if(i == 4){
      Smoke.slope = summary(fit)$coef[3,1]
    }

    if(i == 5){
      Nonsmoke.slope = summary(fit)$coef[3,1]
    }

    if(i == 6){
      Nonsmoke.slope = summary(fit)$coef[3,1]
      Smoke.slope = Nonsmoke.slope + summary(fit)$coef[4,1]
    }

    Radj = summary(fit)$adj.r.squared
    Numerator.d.f = summary(fit)$f[2]
    Denominator.d.f = summary(fit)$f[3]
    Fstat = summary(fit)$f[1]
    Pv = 1 - pf(Fstat,Numerator.d.f,Denominator.d.f)

    sumname = data.frame( Nonsmoke.Intercept, Nonsmoke.slope, Smoke.Intercept, Smoke.slope,
                          Radj, Numerator.d.f, Denominator.d.f, Fstat, Pv)
    sum = c( Nonsmoke.Intercept, Nonsmoke.slope, Smoke.Intercept, Smoke.slope, Radj,
             Numerator.d.f,Denominator.d.f, Fstat, Pv)

    names(sum) = names(sumname)

    # print(summary(fit))
  return(sum)
}

newbirthwt$nonsmoke <- ifelse(newbirthwt$smoke == 1, 0, 1)
newbirthwt$smoke <- as.numeric(newbirthwt$smoke)
newbirthwt$nonsmoke<- as.numeric(newbirthwt$nonsmoke)
bwt <- newbirthwt$bwt
smoke <- newbirthwt$smoke
nonsmoke <- newbirthwt$nonsmoke
age <- newbirthwt$age

formula.list = list(
  bwt ~ age,
  bwt ~ smoke,
  bwt ~ smoke + age,
  bwt ~ smoke + I(smoke*age),
  bwt ~ smoke + I(nonsmoke*age),
  bwt ~ smoke + age + I(smoke*age)
)
```

```
table = NULL
for (i in 1:6)  table = rbind(table, summary.table(lm(formula.list[[i]])))
row.names(table) = formula.list
table
```

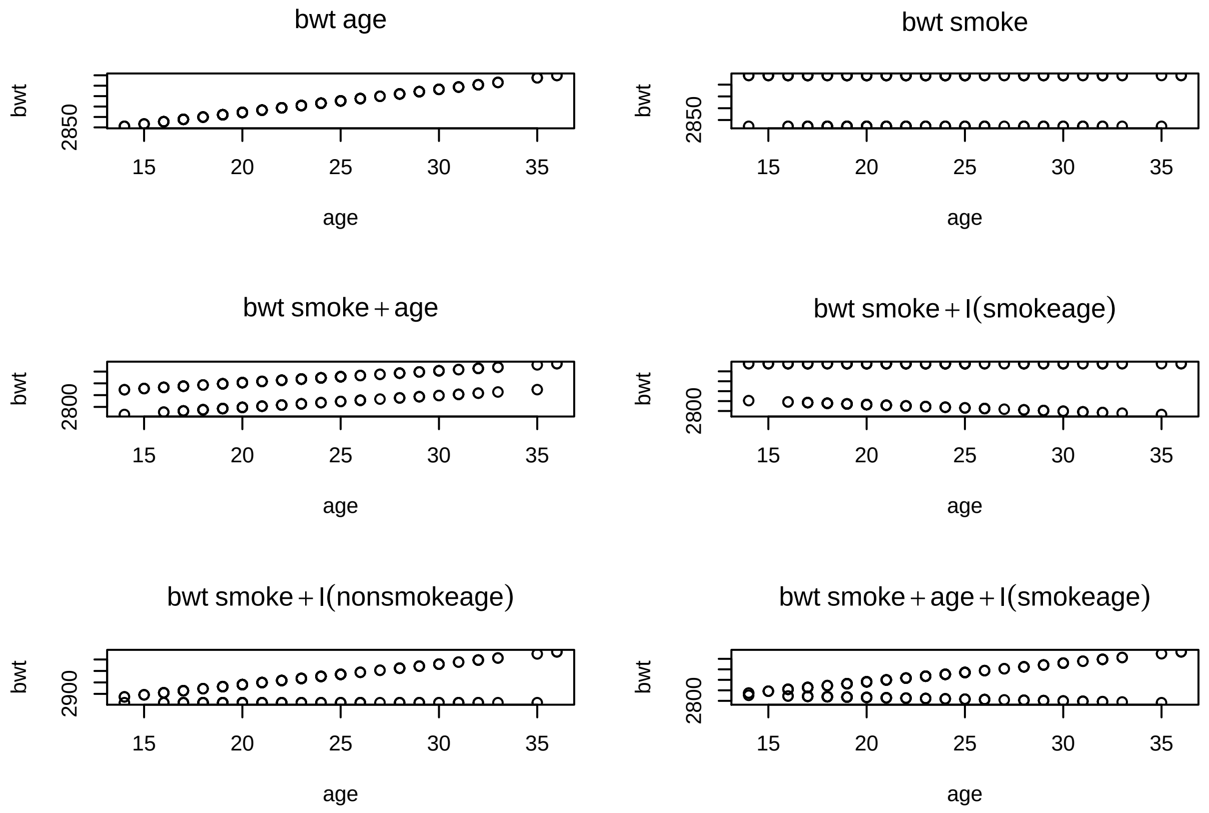
```
##                                    Nonsmoke.Intercept Nonsmoke.slope
## bwt ~ age                                    2700.092       11.07953
## bwt ~ smoke                                  3038.728        0.00000
## bwt ~ smoke + age                            2805.090        0.00000
## bwt ~ smoke + I(smoke * age)                 3038.728        0.00000
## bwt ~ smoke + I(nonsmoke * age)              2624.257       17.83681
## bwt ~ smoke + age + I(smoke * age)           2624.257       17.83681
##                                    Smoke.Intercept Smoke.slope        Radj
## bwt ~ age                                       NA          NA 0.001137205
## bwt ~ smoke                             2823.30556    0.000000 0.018099026
## bwt ~ smoke + age                         10.05465   10.054649 0.018161743
## bwt ~ smoke + I(smoke * age)            2898.99408   -3.331035 0.012952481
## bwt ~ smoke + I(nonsmoke * age)         2823.30556    0.000000 0.023536039
## bwt ~ smoke + age + I(smoke * age)      2898.99408   -3.331035 0.018391090
##                                    Numerator.d.f Denominator.d.f    Fstat
## bwt ~ age                                      1             184 1.210623
## bwt ~ smoke                                    1             184 4.410038
## bwt ~ smoke + age                              2             183 2.711037
## bwt ~ smoke + I(smoke * age)                   2             183 2.213827
## bwt ~ smoke + I(nonsmoke * age)                2             183 3.229559
## bwt ~ smoke + age + I(smoke * age)             3             182 2.155366
##                                            Pv
## bwt ~ age                          0.27264626
## bwt ~ smoke                        0.03709128
## bwt ~ smoke + age                  0.06913803
## bwt ~ smoke + I(smoke * age)       0.11220034
## bwt ~ smoke + I(nonsmoke * age)    0.04184134
## bwt ~ smoke + age + I(smoke * age) 0.09488082
```

### c)

```
par(mfrow=c(3,2))
newbirthwt$smoke <- as.numeric(newbirthwt$smoke)
newbirthwt$nonsmoke <- ifelse(newbirthwt$smoke == 1, 0, 1)
newbirthwt$smoke <- as.numeric(newbirthwt$smoke)
newbirthwt$nonsmoke<- as.numeric(newbirthwt$nonsmoke)
bwt <- newbirthwt$bwt
smoke <- newbirthwt$smoke
nonsmoke <- newbirthwt$nonsmoke
age <- newbirthwt$age

formula.list = list(
  bwt ~ age,
  bwt ~ smoke,
  bwt ~ smoke + age,
  bwt ~ smoke + I(smoke*age),
  bwt ~ smoke + I(nonsmoke*age),
  bwt ~ smoke + age + I(smoke*age)
)
data = list(
  data.frame(age),
  data.frame(smoke),
  data.frame(smoke, age),
  data.frame(smoke, I(smoke*age)),
  data.frame(smoke, I(nonsmoke*age)),
  data.frame(smoke, age, I(smoke*age))
  )


for(i in 1:6){
  plot(predict(lm(formula.list[[i]]), newdata = data[i])~age,
       ylab = "bwt", xlab = "age",main= formula.list[[i]])}
```



## d)

The model *bwt ~ smoke + I(nonsmoke * age)* has the highest $R^2_{adj}$ value, which is 0.023..

```
summary(lm(bwt ~ smoke + I(nonsmoke * age)))
```

```
##
## Call:
## lm(formula = bwt ~ smoke + I(nonsmoke * age))
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -2120.52  -441.77    66.33   509.03  1540.66 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        2624.26     298.17   8.801    1e-15 ***
## smoke               199.05     308.74   0.645    0.520    
## I(nonsmoke * age)    17.84      12.54   1.423    0.156    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 679.6 on 183 degrees of freedom
## Multiple R-squared:  0.03409,    Adjusted R-squared:  0.02354 
## F-statistic:  3.23 on 2 and 183 DF,  p-value: 0.04184
```

For nonsmoker the model is $Y = 2624.25 + 7.84 * X$. For smoker the model is $Y = 2823.30 + 0 * X$.

## e)

Model 2 is the submodel of model5 has the largest number of model degree of freedom.

```
M2 = lm(bwt ~ smoke)
M5 = lm(bwt ~ smoke + I(nonsmoke * age))
anova(M2,M5)
```

```
## Analysis of Variance Table
## 
## Model 1: bwt ~ smoke
## Model 2: bwt ~ smoke + I(nonsmoke * age)
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    184 85444096                           
## 2    183 84509173  1    934922 2.0245 0.1565
```

As we can the p-value in the anvoa, the result seems not yield the same conclusion as the $R^2_{adj}$ ranking.