502_HW3

Shengbo Jin

1/29/2022

Q1

```
load("Homework 3 Data.Rdata")
attach(mac)
head(mac)
```

```
##     BigMac    Bread WorkHrs VacDays    BusFare  Service   TeachSal TeachTax
## 1 3.433987 2.197225    1714    31.9  0.2390169 5.634790 3.08190997     28.2
## 2 3.496508 2.197225    1792    23.5 -1.3093333 5.135798 2.24070969     14.8
## 3 4.584967 3.135494    2152    17.4 -2.4079456 4.605170 0.78845736      4.3
## 4 4.875197 3.295837    2052    30.6 -2.4079456 4.248495 0.09531018     11.7
## 5 3.433987 2.484907    1708    24.6  0.1043600 5.521461 3.12236492     38.2
## 6 4.653960 3.258097    1971    16.2 -1.4271164 5.347108 0.83290912     17.0
##     EngSal EngTax
## 1 3.790985   44.1
## 2 2.965273   23.7
## 3 2.734368   20.3
## 4 1.547563   37.6
## 5 3.908015   50.7
## 6 2.708050   18.5
```

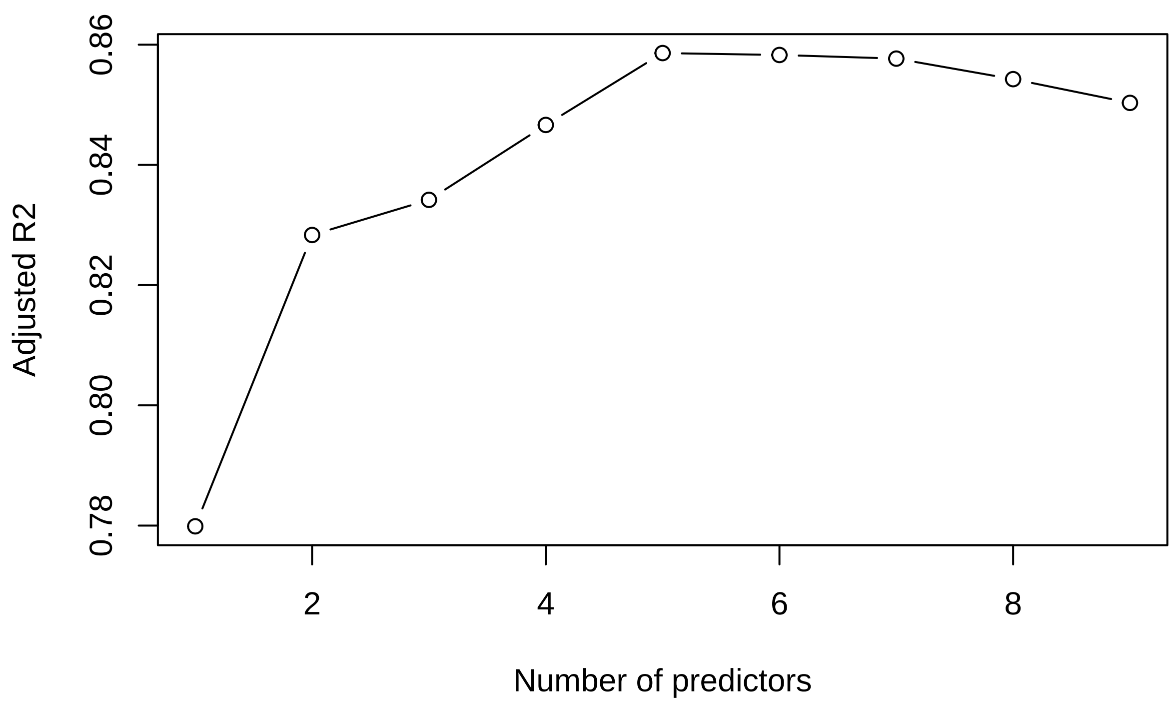
(a)

```
library(leaps)
best.subsets <- regsubsets(BigMac ~ ., data = data.frame(BigMac, Bread,
                           WorkHrs, VacDays, BusFare, Service, TeachSal,
                           TeachTax, EngSal, EngTax), nvmax = 9)
(b <-summary(best.subsets))
```

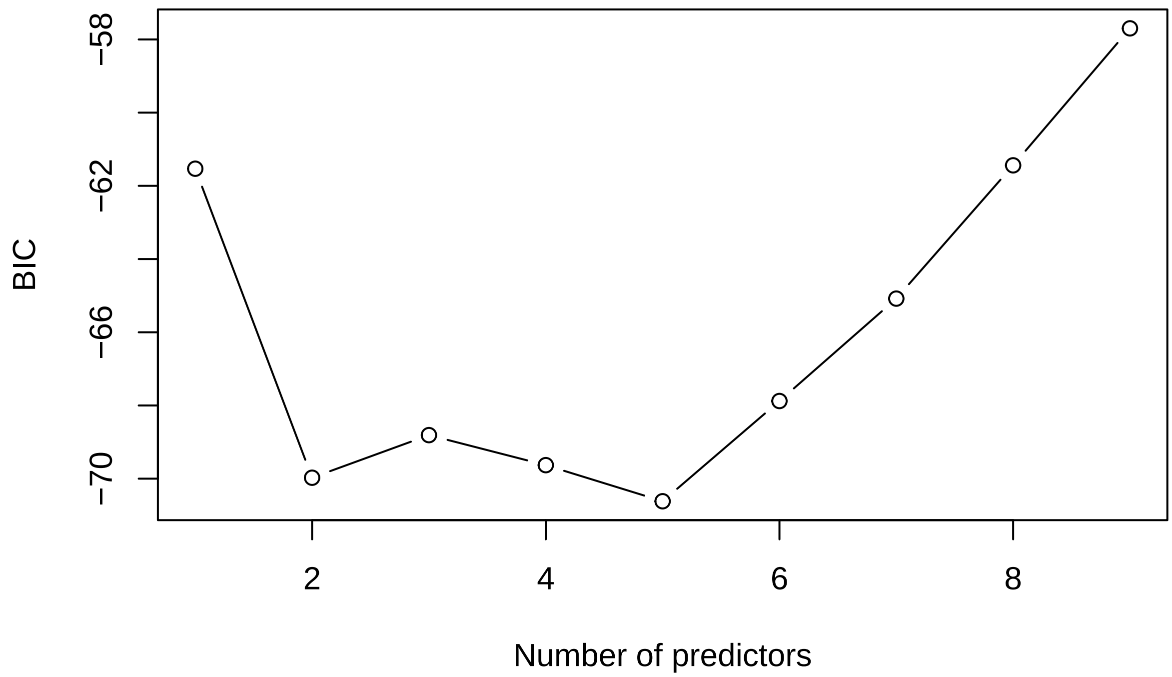
```
## Subset selection object
## Call: regsubsets.formula(BigMac ~ ., data = data.frame(BigMac, Bread,
##     WorkHrs, VacDays, BusFare, Service, TeachSal, TeachTax, EngSal,
##     EngTax), nvmax = 9)
## 9 Variables  (and intercept)
##          Forced in Forced out
## Bread        FALSE      FALSE
## WorkHrs      FALSE      FALSE
## VacDays      FALSE      FALSE
```

```
## BusFare      FALSE      FALSE
## Service      FALSE      FALSE
## TeachSal     FALSE      FALSE
## TeachTax     FALSE      FALSE
## EngSal       FALSE      FALSE
## EngTax       FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          Bread WorkHrs VacDays BusFare Service TeachSal TeachTax EngSal EngTax
## 1  ( 1 ) " "   " "     " "     " "     " "     "*"      " "      " "    " "
## 2  ( 1 ) " "   " "     " "     " "     " "     "*"      "*"      " "    " "
## 3  ( 1 ) " "   " "     " "     "*"     " "     "*"      "*"      " "    " "
## 4  ( 1 ) " "   " "     " "     "*"     " "     "*"      "*"      "*"    " "
## 5  ( 1 ) " "   " "     " "     "*"     "*"     "*"      "*"      "*"    " "
## 6  ( 1 ) "*"   " "     " "     "*"     "*"     "*"      "*"      "*"    " "
## 7  ( 1 ) "*"   " "     "*"     "*"     "*"     "*"      "*"      "*"    " "
## 8  ( 1 ) "*"   " "     "*"     "*"     "*"     "*"      "*"      "*"    "*"
## 9  ( 1 ) "*"   "*"     "*"     "*"     "*"     "*"      "*"      "*"    "*"
```

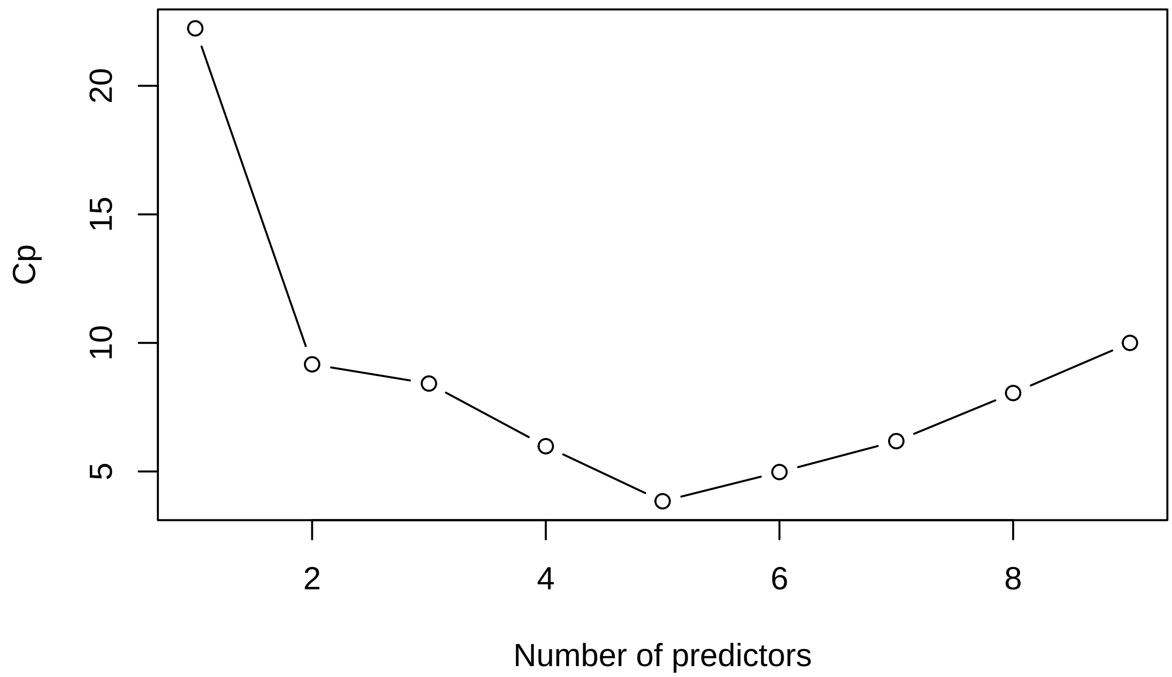
```
plot(1:9, b$adjr2, type="b", xlab="Number of predictors", ylab="Adjusted R2")
```




```
plot(1:9, b$bic, type="b", xlab="Number of predictors", ylab="BIC")
```




```
plot(1:9, b$cp, type="b",xlab="Number of predictors", ylab="Cp")
```



Under adjusted R2, BIC and Mallow's Cp criteria, the best model is the 5 predictor model with predictors BusFare, Service, TeachSal, TeachTax and EngSal.

(b)

For each fixed k, the best model is the same regardless of which criteria is used, and is equivalent to minimizing the residual sum of squares.

```
extractAIC(lm(BigMac~TeachSal))
```

```
## [1]    2.0000 -102.2656
```

```
extractAIC(lm(BigMac~TeachSal+TeachTax))
```

```
## [1]    3.0000 -112.5141
```

```
extractAIC(lm(BigMac~BusFare+TeachSal+TeachTax))
```

```
## [1]    4.000 -113.157
```

```
extractAIC(lm(BigMac~BusFare+TeachSal+TeachTax+EngSal))
```

```
## [1]    5.0000 -115.7838
```

```r
extractAIC(lm(BigMac~BusFare+Service+TeachSal+TeachTax+EngSal))
```

```
## [1]    6.0000 -118.5766
```

```r
extractAIC(lm(BigMac~Bread+BusFare+Service+TeachSal+TeachTax+EngSal))
```

```
## [1]    7.000 -117.645
```

```r
extractAIC(lm(BigMac~Bread+VacDays+BusFare+Service+TeachSal+TeachTax+EngSal))
```

```
## [1]    8.0000 -116.6536
```

```r
extractAIC(lm(BigMac~Bread+VacDays+BusFare+Service+TeachSal+TeachTax+EngSal
              +EngTax))
```

```
## [1]    9.000 -114.819
```

```r
extractAIC(lm(BigMac~Bread+WorkHrs+VacDays+BusFare+Service+TeachSal+TeachTax
              +EngSal+EngTax))
```

```
## [1]   10.0000 -112.8837
```

Under AIC criterion, the best model is 5 predictors model with predictors BusFare, Service, TeachSal, TeachTax and EngSal.

(c)

```r
best.subsets<- regsubsets(BigMac ~ ., data = data.frame(BigMac, Bread, WorkHrs,
                          VacDays, BusFare, Service, TeachSal, TeachTax,
                          EngSal, EngTax), nvmax = 1)
(b <-summary(best.subsets))
```

```
## Subset selection object
## Call: regsubsets.formula(BigMac ~ ., data = data.frame(BigMac, Bread,
##     WorkHrs, VacDays, BusFare, Service, TeachSal, TeachTax, EngSal,
##     EngTax), nvmax = 1)
## 9 Variables  (and intercept)
##          Forced in Forced out
## Bread        FALSE      FALSE
## WorkHrs      FALSE      FALSE
## VacDays      FALSE      FALSE
## BusFare      FALSE      FALSE
## Service      FALSE      FALSE
## TeachSal     FALSE      FALSE
## TeachTax     FALSE      FALSE
## EngSal       FALSE      FALSE
## EngTax       FALSE      FALSE
## 1 subsets of each size up to 1
## Selection Algorithm: exhaustive
##          Bread WorkHrs VacDays BusFare Service TeachSal TeachTax EngSal EngTax
## 1  ( 1 ) " "   " "     " "     " "     " "     "*"      " "      " "    " "
```

TeachSal is best for modeling BigMac.

(d)

```
library(MASS)
stepAIC(lm(BigMac~Bread+WorkHrs+VacDays+BusFare+Service+TeachSal+TeachTax
           +EngSal+EngTax), direction="both")
```

```
## Start:  AIC=-112.88
## BigMac ~ Bread + WorkHrs + VacDays + BusFare + Service + TeachSal +
##     TeachTax + EngSal + EngTax
##
##            Df Sum of Sq    RSS     AIC
## - WorkHrs   1   0.00338 2.3517 -114.82
## - EngTax    1   0.00952 2.3578 -114.70
## - VacDays   1   0.03102 2.3793 -114.29
## - Bread     1   0.07643 2.4247 -113.44
## <none>                  2.3483 -112.88
## - Service   1   0.11722 2.4655 -112.69
## - EngSal    1   0.12029 2.4686 -112.64
## - TeachSal  1   0.40949 2.7578 -107.65
## - BusFare   1   0.43648 2.7848 -107.21
## - TeachTax  1   0.44598 2.7943 -107.06
##
## Step:  AIC=-114.82
## BigMac ~ Bread + VacDays + BusFare + Service + TeachSal + TeachTax +
##     EngSal + EngTax
##
##            Df Sum of Sq    RSS     AIC
## - EngTax    1   0.00866 2.3603 -116.65
## - VacDays   1   0.06175 2.4134 -115.65
## - Bread     1   0.07306 2.4247 -115.44
## <none>                  2.3517 -114.82
## - EngSal    1   0.11886 2.4705 -114.60
## - Service   1   0.12548 2.4771 -114.48
## + WorkHrs   1   0.00338 2.3483 -112.88
## - TeachSal  1   0.43810 2.7898 -109.13
## - TeachTax  1   0.44468 2.7963 -109.03
## - BusFare   1   0.47540 2.8271 -108.53
##
## Step:  AIC=-116.65
## BigMac ~ Bread + VacDays + BusFare + Service + TeachSal + TeachTax +
##     EngSal
##
##            Df Sum of Sq    RSS     AIC
## - VacDays   1   0.05350 2.4138 -117.64
## - Bread     1   0.06968 2.4300 -117.34
## <none>                  2.3603 -116.65
## - Service   1   0.18570 2.5460 -115.25
## + EngTax    1   0.00866 2.3517 -114.82
## + WorkHrs   1   0.00252 2.3578 -114.70
## - EngSal    1   0.24017 2.6005 -114.29
```

```
## - TeachSal  1   0.46307 2.8234 -110.59
## - BusFare   1   0.48161 2.8419 -110.30
## - TeachTax  1   1.07226 3.4326 -101.80
##
## Step:  AIC=-117.64
## BigMac ~ Bread + BusFare + Service + TeachSal + TeachTax + EngSal
##
##            Df Sum of Sq    RSS     AIC
## - Bread     1   0.05800 2.4718 -118.58
## <none>                  2.4138 -117.64
## + VacDays   1   0.05350 2.3603 -116.65
## + WorkHrs   1   0.03051 2.3833 -116.22
## - EngSal    1   0.22043 2.6343 -115.71
## - Service   1   0.22252 2.6364 -115.68
## + EngTax    1   0.00042 2.4134 -115.65
## - BusFare   1   0.44110 2.8549 -112.09
## - TeachSal  1   0.59846 3.0123 -109.68
## - TeachTax  1   1.17596 3.5898 -101.78
##
## Step:  AIC=-118.58
## BigMac ~ BusFare + Service + TeachSal + TeachTax + EngSal
##
##            Df Sum of Sq    RSS     AIC
## <none>                  2.4718 -118.58
## + Bread     1   0.05800 2.4138 -117.64
## + VacDays   1   0.04182 2.4300 -117.34
## + WorkHrs   1   0.01277 2.4591 -116.81
## + EngTax    1   0.00011 2.4717 -116.58
## - Service   1   0.27780 2.7496 -115.78
## - EngSal    1   0.33272 2.8046 -114.89
## - BusFare   1   0.42209 2.8939 -113.48
## - TeachSal  1   0.68882 3.1607 -109.52
## - TeachTax  1   1.22783 3.6997 -102.43

##
## Call:
## lm(formula = BigMac ~ BusFare + Service + TeachSal + TeachTax +
##     EngSal)
##
## Coefficients:
## (Intercept)      BusFare      Service     TeachSal     TeachTax       EngSal
##     3.39209     -0.23172      0.29650     -0.38618      0.02451     -0.28257
```

Using stepwise selection, the best model in terms of AIC is BigMac ~ BusFare + Service + TeachSal + TeachTax + EngSal.

(e)

```r
library(faraway)
fit1 <- lm(BigMac~Bread+WorkHrs+VacDays+BusFare+Service+TeachSal+TeachTax
           +EngSal+EngTax)
fit2 <- lm(BigMac~BusFare+Service+TeachSal+TeachTax+EngSal)
```

```
vif(fit1)
```

```
##      Bread   WorkHrs   VacDays   BusFare   Service  TeachSal  TeachTax    EngSal
##   2.674683  2.719598  1.831019  5.615845  3.449345 17.926595  7.702699 14.643748
##     EngTax
##   6.873528
```

```
vif(fit2)
```

```
##    BusFare   Service  TeachSal  TeachTax    EngSal
##   4.649925  2.641037 13.085365  2.426280  8.607902
```

In the full set of predictors, TeachSal, TeachTax, EngSal and EngTax are clearly collinear as their VIF is very large. The model selection has reduced it with smaller VIF of TeachSal, TeachTax and EngSal.

Q2

(a)

```
set.seed(100)
n <- 100
x1 <- rnorm(n)
x2 <- rnorm(n)
x3 <- x1+x2+rnorm(n, sd=0.001)
eps <- rnorm(n, sd=0.1)
y <- 3*x1+3*x2+eps
summary(lm(y~x1+x2+x3))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26057 -0.06113 -0.00340  0.06866  0.34316
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.008296   0.010877  -0.763    0.447
## x1           13.686681  10.939966   1.251    0.214
## x2           13.688708  10.935505   1.252    0.214
## x3          -10.680584  10.939157  -0.976    0.331
##
## Residual standard error: 0.1087 on 96 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.9991
## F-statistic: 3.676e+04 on 3 and 96 DF,  p-value: < 2.2e-16
```

For X1, p-value is 0.214 which is larger than 0.05, we fail to reject H0. The coefficient of X1 should be 0, which is less than the true value 3; For X2, p-value is 0.214 which is larger than 0.05, we fail to reject H0.

The coefficient of X2 should be 0, which is less than the true value 3; For X3, p-value is 0.331 which is larger than 0.05, we fail to reject H0. The coefficient of X3 should be 0, which is equal to the true value 0. To some extent, the t-test shows Y has no relationship with X1 and X2. But, in reality, X1 and X2 make up Y. The multicollinearity greatly decreases the reliability of the test results.

(b)

```r
set.seed(100)
n <- 10000
x1 <- rnorm(n)
x2 <- rnorm(n)
x3 <- x1+x2+rnorm(n, sd=0.001)
eps <- rnorm(n, sd=0.1)
y <- 3*x1+3*x2+eps
summary(lm(y~x1+x2+x3))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36643 -0.06626 -0.00119  0.06689  0.43715
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0003147  0.0009927  -0.317 0.751275
## x1           3.2430077  0.9846944   3.293 0.000993 ***
## x2           3.2423925  0.9846962   3.293 0.000995 ***
## x3          -0.2416992  0.9846958  -0.245 0.806109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09926 on 9996 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9994
## F-statistic: 6.035e+06 on 3 and 9996 DF,  p-value: < 2.2e-16
```

Basically Correct. From the test results, we can see the p-value becomes 0.000993, 0.000995, 0.806109. Hence the coefficient of X1 and X2 are statistically significant and close to the true value. And X3 is not statistically significant, which means X3 should be 0. The test result matches better with our real model. Thus, increasing the sample size can address multicollinearity.

(a) Based on the assumption, ε_0 is uncorrelated with ε_i for $i=1,\dots,n$.
Thus, ε_0 is uncorrelated with $\hat{Y}_0^{(0)}$ and $\hat{Y}_0^{(1)}$

$$E[(Y_0-\hat{Y}_0^{(0)})^2] = Var(Y_0-\hat{Y}_0^{(0)}) + (E[Y_0-\hat{Y}_0^{(0)}])^2$$
$$= Var(\beta_0+\beta_1 X_0+\varepsilon_0-\hat{Y}_0^{(0)}) + (bias(\hat{Y}_0^{(0)}))^2$$
$$= Var(\varepsilon_0-\hat{Y}_0^{(0)}) + (bias(\hat{Y}_0^{(0)}))^2$$
$$= Var(\varepsilon_0) + Var(\hat{Y}_0^{(0)}) + (bias(\hat{Y}_0^{(0)}))^2$$
$$= \sigma^2 + Var(\hat{Y}_0^{(0)}) + (bias(\hat{Y}_0^{(0)}))^2$$

$$E[(Y_0-Y_0^{(1)})^2] = Var(Y_0-\hat{Y}_0^{(1)}) + (E[Y_0-\hat{Y}_0^{(1)}])^2$$
$$= Var(\beta_0+\beta_1 X_0+\varepsilon_0-\hat{Y}_0^{(1)}) + (bias(\hat{Y}_0^{(1)}))^2$$
$$= Var(\varepsilon_0) + Var(\hat{Y}_0^{(1)}) + (bias(\hat{Y}_0^{(1)}))^2$$
$$= \sigma^2 + Var(\hat{Y}_0^{(1)}) + (bias(\hat{Y}_0^{(1)}))^2$$

(b)
$$Var(\hat{Y}_0^{(0)}) = Var(\hat{\beta}_0^{(0)}) = Var(\bar{Y})$$
$$= Var\left(\frac{\sum_{i=1}^{n}\varepsilon_i}{n}\right)$$
$$= \frac{\sigma^2}{n} \leq \sigma^2\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{(n-1)S_X^2}\right) = Var(\hat{Y}_0^{(1)})$$

$$bias(\hat{Y}_0^{(0)}) = E[\hat{Y}_0^{(0)}] - E[Y_0]$$
$$= E[\hat{\beta}_0^{(0)}] - E[\beta_0+\beta_1 X_0+\varepsilon_0]$$
$$= E[\bar{Y}] - (\beta_0+\beta_1 X_0)$$
$$= \beta_0+\beta_1\bar{X}-\beta_0-\beta_1 X_0$$
$$= \beta_1(\bar{X}-X_0)$$
$$(bias(\hat{Y}_0^{(0)}))^2 = \beta_1^2(\bar{X}-X_0)^2$$

$$\text{bias}(\hat{Y}_0^{(1)}) = E[\hat{Y}_0^{(1)}] - E[Y_0]$$

$$= E[\hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x_0] - (\beta_0 + \beta_1 x_0)$$

$$= (\beta_0 + \beta_1 x_0) - (\beta_0 + \beta_1 x_0)$$

$$= 0$$

$$(\text{bias}(\hat{Y}_0^{(1)}))^2 = 0 \le \beta_1^2 (\bar{x} - x_0)^2 = (\text{bias}(\hat{Y}_0^{(0)}))^2$$

$$E[(Y_0 - \hat{Y}_0^{(1)})^2] = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1) S_x^2}\right)$$

$$E[(Y_0 - \hat{Y}_0^{(0)})^2] = \sigma^2 + \frac{\sigma^2}{n} + \beta_1^2 (\bar{x} - x_0)^2$$

$$E[(Y_0 - \hat{Y}_0^{(1)})^2] - E[(Y_0 - \hat{Y}_0^{(0)})^2] = \sigma^2 \frac{(x_0 - \bar{x})^2}{(n-1) S_x^2} - \beta_1^2 (\bar{x} - x_0)^2$$

if $\frac{\sigma^2}{(n-1) S_x^2} > \beta_1^2$

Then Model$_1$ is worse than the underfit model