ST5202_Tut4

Name:Zhu Xu
User ID:E0337988
Matriculation ID:A0191344H

9/April/2019

8.15:

a)

```
copier = cbind(read.table(
  "/Users/xuzhu/Desktop/Notes/Sem2/ST5202-Applied_Regression_Analysis/Tut/copier_maintenance.txt"),
  read.table(
    "/Users/xuzhu/Desktop/Notes/Sem2/ST5202-Applied_Regression_Analysis/Tut/copier_maintenance_additiona
colnames(copier) = c("Y","X1","X2")
fit.lm = lm(Y ~ X1 + X2, data=copier)
summary(fit.lm)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = copier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5390  -4.2515   0.5995   6.5995  14.9330
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9225     3.0997  -0.298    0.767
## X1           15.0461     0.4900  30.706   <2e-16 ***
## X2            0.7587     2.7799   0.273    0.786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.011 on 42 degrees of freedom
## Multiple R-squared:  0.9576, Adjusted R-squared:  0.9556
## F-statistic: 473.9 on 2 and 42 DF,  p-value: < 2.2e-16
```

$b_0 = -0.9225$ is the intercept of the model, $b_1 = 15.0461$ means if one new copier is serviced the number of minutes will increase 15.0461, $b_2 = 0.7587$ means if the one of the copiers changes from large to small, the total time will increase 0.7587 minutes.

b)

The fitted model is

$$\hat{Y} = -0.9225 + 15.0461X_1 + 0.7587X_2$$

c)

```r
b_2 = summary(fit.lm)$coefficients[ 3, 1]
sd_2 = summary(fit.lm)$coefficients[ 3, 2]
paste0("(",round(b_2 - qt(0.975,42)*sd_2,2),
       ",",round(b_2 + qt(0.975,42)*sd_2,2),")")
```

```
## [1] "(-4.85,6.37)"
```

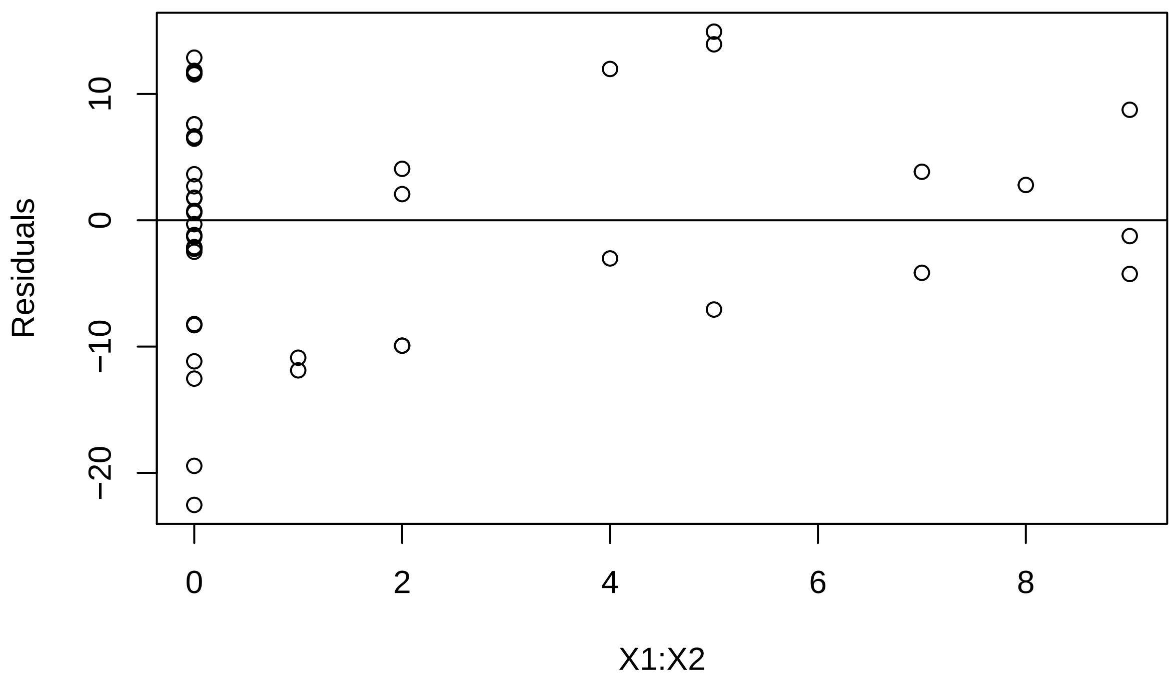
The 95% confidence interval is $(-4.85, 6.37)$.

d)

Since if it is not concluded, the fitted result may be undervalued.

e)

```r
par(mfrow=c(1,1))
plot(y=fit.lm$residuals,x=copier$X1*copier$X2,xlab="X1:X2",
                        ylab="Residuals")
abline(0,0)
```



8.19:

a)

```r
fit.lm1 <- lm(Y ~ X1 + X2 + X1:X2, data=copier)
summary(fit.lm1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1:X2, data = copier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2072  -6.7887  -0.1708   7.1504  14.7441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8131     3.6468   0.771   0.4449
## X1           14.3394     0.6146  23.333   <2e-16 ***
## X2           -8.1412     5.5801  -1.459   0.1522
## X1:X2         1.7774     0.9746   1.824   0.0755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.771 on 41 degrees of freedom
## Multiple R-squared:  0.9608, Adjusted R-squared:  0.9579
## F-statistic: 334.6 on 3 and 41 DF,  p-value: < 2.2e-16
```

The fitted model is

$$\hat{(Y)} = 2.8131 + 14.3994X_1 - 8.1412X_2 + 1.7774X_1X_2$$

b)

```r
anova(fit.lm,fit.lm1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 + X2 + X1:X2
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     42 3410.3
## 2     41 3154.4  1    255.89 3.326 0.07549 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

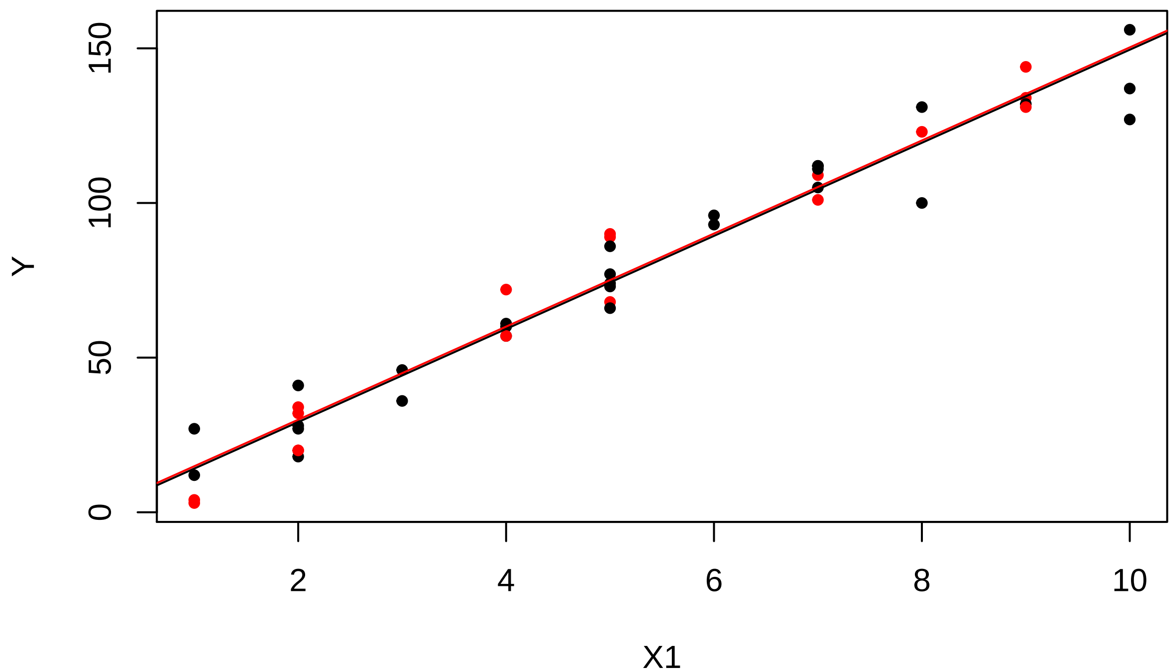
$H_0 : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \qquad H_1 : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

The F-statistic is 3.326, which is larger than $F(0.9, 41)$, thus we reject H_0

Q1:

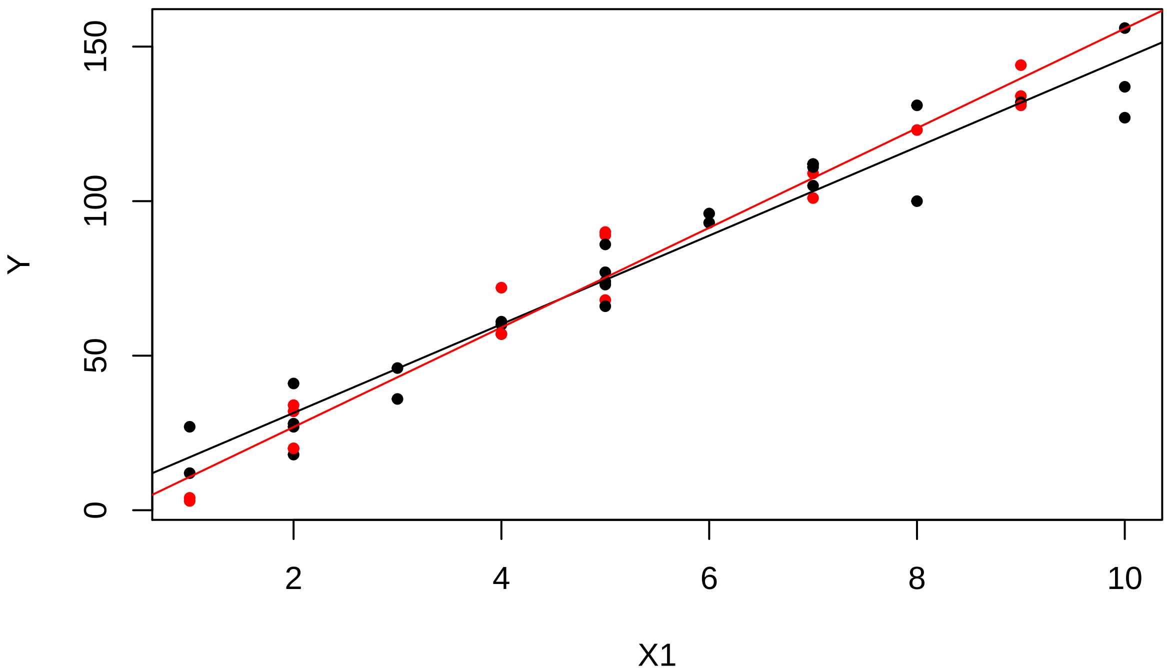
a)

```r
 plot(y=copier$Y, x=copier$X1,
      xlab="X1", ylab="Y",
      col=as.factor(copier$X2), pch=20)
abline(coefficients(fit.lm)[1], coefficients(fit.lm)[2])
abline(coefficients(fit.lm)[1]+coefficients(fit.lm)[3],
       coefficients(fit.lm)[2],col="red")
```



b)

```
plot(y=copier$Y,x=copier$X1,
     xlab="X1",ylab="Y",
     col=as.factor(copier$X2),pch=20)
abline(coefficients(fit.lm1)[1], coefficients(fit.lm1)[2])
abline(coefficients(fit.lm1)[1]+coefficients(fit.lm1)[3],
       coefficients(fit.lm1)[2]+coefficients(fit.lm1)[4],
       col="red")
```



8.21:

a)

For hard hat: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

For bump cap: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_3 X_3$

For none: $E\{Y\} = \beta_0 + \beta_1 X_1$

b)

(1) $H_0 : \beta_3 < 0, \quad H_a : \beta_3 \geq 0$

(2) $H_0 : \beta_2 = \beta_3, \quad H_a : \beta_2 \neq \beta_3$

9.10:

```r
jobdata = read.table(
  "/Users/xuzhu/Desktop/Notes/Sem2/ST5202-Applied_Regression_Analysis/Tut/job_proficiency.txt")
colnames(jobdata) = c("X1",'X2','X3','X4','Y')
fit.lm <- lm(Y~.,data=jobdata)
summary(fit.lm)
```

```
##
## Call:
## lm(formula = Y ~ ., data = jobdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7283 -2.6769 -0.7255  3.1096  9.3900
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 100.88142   30.03086   3.359  0.00312 **
## X1            0.84060    0.21337   3.940  0.00081 ***
## X2           -0.19182    0.09142  -2.098  0.04879 *
## X3           -0.04574    0.07258  -0.630  0.53570
## X4           -0.58529    0.40537  -1.444  0.16427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.212 on 20 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7617
## F-statistic: 20.17 on 4 and 20 DF,  p-value: 8.612e-07
```

X_3 and X_4 are not significant, we should not keep them.

9.11:

a)

```r
library('MuMIn')
options(na.action = "na.fail")
combinations = dredge(fit.lm,extra="adjR^2")
```

```
## Fixed term is "(Intercept)"
```

```r
print(combinations)
```

```
## Global model call: lm(formula = Y ~ ., data = jobdata)
## ---
## Model selection table
##      (Intrc)     X1       X2        X3       X4 adjR^2 df  logLik  AICc
## 2    50.6200 0.4779                                0.7562  3 -76.539 160.2
## 4    55.2600 0.5242 -0.08623                       0.7760  4 -75.484 161.0
## 12   98.5000 0.8234 -0.18620           -0.60010 0.7979  5 -74.204 161.6
## 6    52.5300 0.4907          -0.029010             0.7579  4 -76.454 162.9
## 10   49.7900 0.4725                     0.01323 0.7563  4 -76.538 163.1
## 8    59.1700 0.5521 -0.09544 -0.051830             0.7811  5 -75.198 163.6
## 16 100.9000 0.8406 -0.19180 -0.045740 -0.58530 0.8018  6 -73.959 164.6
## 14   50.3900 0.4769          -0.030790  0.03577 0.7581  5 -76.445 166.0
## 9    -0.3569                            0.94280 0.6119  3 -82.342 171.8
## 11   -6.5630         0.10080            0.90110 0.6474  4 -81.144 172.3
## 13   -0.3216                 -0.007777  0.95070 0.6120  4 -82.338 174.7
## 15   -6.5250         0.10080 -0.009144  0.91030 0.6476  5 -81.139 175.4
## 5    68.5400                  0.244900             0.1575  3 -92.023 191.2
## 7    54.7700         0.15210  0.226700             0.2402  4 -90.731 191.5
## 3    76.9200         0.17180                       0.1068  3 -92.753 192.6
## 1    94.6800                                       0.0000  2 -94.163 192.9
##    delta weight
## 2   0.00  0.327
## 4   0.75  0.225
## 12  1.34  0.167
## 6   2.69  0.085
## 10  2.85  0.078
## 8   3.33  0.062
## 16  4.36  0.037
## 14  5.83  0.018
## 9  11.61  0.001
## 11 12.07  0.001
## 13 14.46  0.000
## 15 15.21  0.000
```

```
## 5  30.97  0.000
## 7  31.24  0.000
## 3  32.43  0.000
## 1  32.65  0.000
## Models ranked by AICc(x)
```

So the 4 best regression models are:

$$E\{Y\}_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$
$$E\{Y\}_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4$$
$$E\{Y\}_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$
$$E\{Y\}_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

b)

R-squard, AIC, BIC…

Q2:

a)

```r
combinations = dredge(fit.lm,rank='AIC')
```

```
## Fixed term is "(Intercept)"
```

```r
print(combinations)
```

```
## Global model call: lm(formula = Y ~ ., data = jobdata)
## ---
## Model selection table
##      (Intrc)     X1       X2        X3       X4 df  logLik   AIC delta
## 12   98.5000 0.8234 -0.18620           -0.60010  5 -74.204 158.4  0.00
## 4    55.2600 0.5242 -0.08623                     4 -75.484 159.0  0.56
## 2    50.6200 0.4779                              3 -76.539 159.1  0.67
## 16  100.9000 0.8406 -0.19180 -0.045740 -0.58530  6 -73.959 159.9  1.51
## 8    59.1700 0.5521 -0.09544 -0.051830           5 -75.198 160.4  1.99
## 6    52.5300 0.4907          -0.029010           4 -76.454 160.9  2.50
## 10   49.7900 0.4725                     0.01323  4 -76.538 161.1  2.67
## 14   50.3900 0.4769          -0.030790  0.03577  5 -76.445 162.9  4.48
## 11   -6.5630         0.10080            0.90110  4 -81.144 170.3 11.88
## 9    -0.3569                            0.94280  3 -82.342 170.7 12.28
## 15   -6.5250         0.10080 -0.009144  0.91030  5 -81.139 172.3 13.87
## 13   -0.3216                 -0.007777  0.95070  4 -82.338 172.7 14.27
## 7    54.7700         0.15210  0.226700           4 -90.731 189.5 31.05
## 5    68.5400                  0.244900           3 -92.023 190.0 31.64
## 3    76.9200         0.17180                     3 -92.753 191.5 33.10
```

```
## 1   94.6800                                    2 -94.163 192.3 33.92
##     weight
## 12  0.252
## 4   0.190
## 2   0.180
## 16  0.118
## 8   0.093
## 6   0.072
## 10  0.066
## 14  0.027
## 11  0.001
## 9   0.001
## 15  0.000
## 13  0.000
## 7   0.000
## 5   0.000
## 3   0.000
## 1   0.000
## Models ranked by AIC(x)
```

The best 4 models according to AIC are models with varibles(1,2,4), (1,2), (1), (1,2,3,4)

b)

```r
combinations <- dredge(fit.lm,rank='BIC')
```

```
## Fixed term is "(Intercept)"
```

```r
print(combinations)
```

```
## Global model call: lm(formula = Y ~ ., data = jobdata)
## ---
## Model selection table
##      (Intrc)     X1       X2        X3       X4 df  logLik   BIC delta
## 2    50.6200 0.4779                               3 -76.539 162.7  0.00
## 4    55.2600 0.5242 -0.08623                      4 -75.484 163.8  1.11
## 12   98.5000 0.8234 -0.18620           -0.60010 5 -74.204 164.5  1.77
## 6    52.5300 0.4907          -0.029010          4 -76.454 165.8  3.05
## 10   49.7900 0.4725                     0.01323 4 -76.538 166.0  3.22
## 8    59.1700 0.5521 -0.09544 -0.051830          5 -75.198 166.5  3.75
## 16  100.9000 0.8406 -0.19180 -0.045740 -0.58530 6 -73.959 167.2  4.49
## 14   50.3900 0.4769          -0.030790  0.03577 5 -76.445 169.0  6.25
## 9    -0.3569                            0.94280 3 -82.342 174.3 11.61
## 11   -6.5630         0.10080            0.90110 4 -81.144 175.2 12.43
## 13   -0.3216                 -0.007777  0.95070 4 -82.338 177.6 14.82
```

```
## 15   -6.5250         0.10080 -0.009144  0.91030  5 -81.139 178.4 15.64
## 5    68.5400                  0.244900           3 -92.023 193.7 30.97
## 7    54.7700         0.15210  0.226700           4 -90.731 194.3 31.60
## 1    94.6800                                     2 -94.163 194.8 32.03
## 3    76.9200         0.17180                     3 -92.753 195.2 32.43
##     weight
## 2    0.368
## 4    0.212
## 12   0.152
## 6    0.080
## 10   0.074
## 8    0.056
## 16   0.039
## 14   0.016
## 9    0.001
## 11   0.001
## 13   0.000
## 15   0.000
## 5    0.000
## 7    0.000
## 1    0.000
## 3    0.000
## Models ranked by BIC(x)
```

Model with variables(1,2,4), (1), (1,2), (1,2,3,4) according to BIC.

9.18:

```
fit.lm = lm(Y ~ 1, data =jobdata)
anova(lm(Y ~ X1, data =jobdata),fit.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     23  667.9
## 2     24 2735.4 -1   -2067.5 71.198 1.699e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(Y ~ X2, data =jobdata),fit.lm)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: Y ~ X2
## Model 2: Y ~ 1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 2443.5
## 2     24 2735.4 -1    -291.9 2.7475  0.111
```

```r
anova(lm(Y ~ X3, data =jobdata),fit.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X3
## Model 2: Y ~ 1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 2304.9
## 2     24 2735.4 -1    -430.5 4.2958 0.0496 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(Y ~ X4, data =jobdata),fit.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X4
## Model 2: Y ~ 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     23 1062.5
## 2     24 2735.4 -1     -1673 36.215 3.887e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fit.lm =update(fit.lm, .~. +X1)
anova(fit.lm,lm(Y ~ X1+X2, data =jobdata))
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 667.90
## 2     22 613.84  1    54.064 1.9377 0.1778
```

```r
anova(fit.lm,lm(Y ~ X1+X3, data =jobdata))
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
```

```
## Model 2: Y ~ X1 + X3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 667.90
## 2     22 663.36  1    4.5473 0.1508 0.7015
```

```
anova(fit.lm,lm(Y ~ X1+X4, data =jobdata))
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 667.90
## 2     22 667.84  1  0.064314 0.0021 0.9637
```

The model shoud be $E\{Y\} = \beta_0 + \beta_1 X_1$

b)

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Q3:

a)

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$$

b)

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$$

c)

Comparing AIC of simple linear regression model and simple population model, finding all 4 models can decrease the AIC of simple population model. We choose the minimum one as best one, thus we add X_3, X_1, X_4 into our model.

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$$

d)

We drop the variable X_2.

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$$