STAT3032 HW4

Mingming Xu

2019/3/25

Problem 1

```
diamonds =
    read.csv("https://raw.githubusercontent.com/tidyverse/ggplot2/master/data-raw/diamonds.csv")
diamonds = diamonds[,-5]
diamonds = diamonds[-46477, ]
n = nrow(diamonds)

set.seed(3032)
sub = sample(1:n, 500, replace=FALSE)
diamonds = diamonds[sub,]
```

(a) If we fit this model: log(price) ~ log (carat) + cut + color + clarity + table + x + y + z, we need to estimeate 23 coefficients. carat,table,x,y,z are quantitative varaibles with 5 coefficients;cut is dummy variable with 5 categories, with 4 coefficients;color is dummy variable with 7 categories, with 6 coefficients;clarity is dummy variable with 8 categories, with 7 coefficients. Hence, 1+4+6+7+1+1+1+1+1=23.

(b)

Because there are 8 predictors in the population model,the maximum number of subsets for a stepsize is 1+p(p+1)/2,which is 37.

(c)

```
mod1=lm(log(price)~log(carat)+cut+color+clarity+table+x+y+z, data=diamonds)
backAIC = step(mod1, direction="backward", data=diamonds)
```

```
## Start:  AIC=-1981.8
## log(price) ~ log(carat) + cut + color + clarity + table + x +
##     y + z
##
##              Df Sum of Sq     RSS     AIC
## - z           1    0.0006  8.6632 -1983.8
## - table       1    0.0007  8.6633 -1983.8
## - y           1    0.0015  8.6641 -1983.7
## - x           1    0.0125  8.6751 -1983.1
## <none>                     8.6626 -1981.8
## - cut         4    0.2710  8.9336 -1974.4
## - log(carat)  1    3.9507 12.6133 -1795.9
## - color       6    5.6632 14.3258 -1742.3
## - clarity     7   15.2787 23.9413 -1487.5
##
## Step:  AIC=-1983.76
## log(price) ~ log(carat) + cut + color + clarity + table + x +
##     y
##
##              Df Sum of Sq     RSS     AIC
## - table       1    0.0003  8.6635 -1985.7
## - y           1    0.0016  8.6648 -1985.7
```

```
## - x           1    0.0147  8.6779 -1984.9
## <none>                     8.6632 -1983.8
## - cut         4    0.3051  8.9683 -1974.5
## - log(carat)  1    4.8327 13.4959 -1764.1
## - color       6    5.7652 14.4284 -1740.7
## - clarity     7   15.2924 23.9556 -1489.2
##
## Step:  AIC=-1985.74
## log(price) ~ log(carat) + cut + color + clarity + x + y
##
##              Df Sum of Sq     RSS     AIC
## - y           1    0.0016  8.6651 -1987.7
## - x           1    0.0150  8.6785 -1986.9
## <none>                     8.6635 -1985.7
## - cut         4    0.3390  9.0025 -1974.5
## - log(carat)  1    4.9136 13.5771 -1763.1
## - color       6    5.8078 14.4713 -1741.2
## - clarity     7   15.3379 24.0014 -1490.2
##
## Step:  AIC=-1987.66
## log(price) ~ log(carat) + cut + color + clarity + x
##
##              Df Sum of Sq     RSS     AIC
## <none>                     8.6651 -1987.7
## - x           1    0.2196  8.8847 -1977.1
## - cut         4    0.3459  9.0109 -1976.1
## - log(carat)  1    5.1461 13.8111 -1756.6
## - color       6    5.8115 14.4765 -1743.0
## - clarity     7   15.6376 24.3026 -1486.0
```

The predictor variables included in the best model are carat , cut,color , clarity and x.

(d)

```
backBIC = step(mod1, direction="backward", data=diamonds,k=log(n))
```

```
## Start:  AIC=-1777.2
## log(price) ~ log(carat) + cut + color + clarity + table + x +
##     y + z
##
##              Df Sum of Sq     RSS     AIC
## - cut         4    0.2710  8.9336 -1805.4
## - z           1    0.0006  8.6632 -1788.1
## - table       1    0.0007  8.6633 -1788.0
## - y           1    0.0015  8.6641 -1788.0
## - x           1    0.0125  8.6751 -1787.4
## <none>                     8.6626 -1777.2
## - log(carat)  1    3.9507 12.6133 -1600.2
## - color       6    5.6632 14.3258 -1591.0
## - clarity     7   15.2787 23.9413 -1345.2
##
## Step:  AIC=-1805.38
## log(price) ~ log(carat) + color + clarity + table + x + y + z
##
##              Df Sum of Sq     RSS     AIC
## - y           1    0.0057  8.9393 -1816.0
```

```
## - z           1    0.0347  8.9683 -1814.3
## - x           1    0.0406  8.9741 -1814.0
## - table       1    0.0572  8.9908 -1813.1
## <none>                     8.9336 -1805.4
## - log(carat)  1    3.8973 12.8308 -1635.3
## - color       6    5.5898 14.5234 -1627.8
## - clarity     7   15.6465 24.5800 -1375.6
##
## Step:  AIC=-1815.96
## log(price) ~ log(carat) + color + clarity + table + x + z
##
##              Df Sum of Sq     RSS     AIC
## - z           1    0.0343  8.9736 -1824.9
## - table       1    0.0609  9.0002 -1823.5
## <none>                     8.9393 -1816.0
## - x           1    0.3892  9.3285 -1805.5
## - log(carat)  1    4.0517 12.9910 -1640.0
## - color       6    5.6034 14.5427 -1638.0
## - clarity     7   16.0840 25.0232 -1377.5
##
## Step:  AIC=-1824.94
## log(price) ~ log(carat) + color + clarity + table + x
##
##              Df Sum of Sq     RSS     AIC
## - table       1    0.0373  9.0109 -1833.8
## <none>                     8.9736 -1824.9
## - x           1    0.3895  9.3631 -1814.6
## - color       6    5.8465 14.8200 -1639.5
## - log(carat)  1    4.8999 13.8734 -1618.0
## - clarity     7   16.3050 25.2786 -1383.4
##
## Step:  AIC=-1833.76
## log(price) ~ log(carat) + color + clarity + x
##
##              Df Sum of Sq     RSS     AIC
## <none>                     9.0109 -1833.8
## - x           1    0.3681  9.3790 -1824.6
## - color       6    5.8092 14.8201 -1650.4
## - log(carat)  1    5.0202 14.0311 -1623.2
## - clarity     7   16.2951 25.3060 -1393.7
```

The predictor variables included in the best model are carat , color , clarity and x.

(e)

```
bestmod=lm(log(price) ~ log(carat) + color + clarity + x,data = diamonds)
summary(bestmod)
```

```
##
## Call:
## lm(formula = log(price) ~ log(carat) + color + clarity + x, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59079 -0.08810  0.00805  0.08573  0.39242
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.67696    0.30417  21.952  < 2e-16 ***
## log(carat)   1.46801    0.08940  16.421  < 2e-16 ***
## colorE      -0.01267    0.02495  -0.508    0.612
## colorF      -0.01961    0.02443  -0.803    0.423
## colorG      -0.10483    0.02479  -4.229 2.81e-05 ***
## colorH      -0.20547    0.02530  -8.123 3.82e-15 ***
## colorI      -0.29101    0.02888 -10.077  < 2e-16 ***
## colorJ      -0.44603    0.04042 -11.034  < 2e-16 ***
## clarityIF    1.00979    0.06341  15.924  < 2e-16 ***
## claritySI1   0.51486    0.05115  10.065  < 2e-16 ***
## claritySI2   0.33344    0.05195   6.419 3.28e-10 ***
## clarityVS1   0.72695    0.05264  13.810  < 2e-16 ***
## clarityVS2   0.65429    0.05148  12.709  < 2e-16 ***
## clarityVVS1  0.94263    0.05679  16.598  < 2e-16 ***
## clarityVVS2  0.88228    0.05415  16.294  < 2e-16 ***
## x            0.20581    0.04629   4.446 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1364 on 484 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:  0.982
## F-statistic:  1812 on 15 and 484 DF,  p-value: < 2.2e-16
```

(i)The estimated coefficient of log(carat) is 1.46801, which means that on average, log(carat) increasing one unit, holding other variables, the log(price) increases 1.46801.

(ii)The estimated coefficient of colorJ is -0.44603, which means that on average,when the color is J,holding other variables, the log(price) decreases 0.44603.

(f)

```
new.dat=data.frame(carat=1,cut="Ideal",color="G",clarity="VS1",x=5.7,y=5.7,z=3.5,table=57)
predict(bestmod,newdata = new.dat,interval = "confidence",level = 0.95)
```

```
##        fit      lwr      upr
## 1 8.472201 8.389143 8.555259
```

```
exp (predict(bestmod,newdata = new.dat,interval = "confidence",level = 0.95))
```

```
##        fit      lwr      upr
## 1 4780.025 4399.048 5193.996
```

Hence, the confidence interval is (4399.048,5193.996).

on avearge, we have 95% confident to say that the price will be between 4394.813 and 5192.666.

Problem 2

Two major concerns

1. " The author found that “seven of these (predictor) variables had a statistically significant impact on attendance revenue” " “.It is a reduced model,not versus the full model with 12 varaibles and testing nested model. We cannot explain if it is better to add extra predictors of the bigger model.
2. In the text, it refers to it had a t-statistic significant at least at the 10% level.If we decrease the significant level to less than 10%, we cannot sure that if seven variables will keep this sighnificant.

Problem 3

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
?Rateprof
n = nrow(Rateprof)
```

The simplest model : quality ~ 1

The most complicated model :quality ~ 1 + gender+ numYears + numRaters + numCourses + pepper + dept+ helpfulness + clarity + easiness + raterInterest

(a)

(B) will have the highest R^2,because R^2 usually increases as we add more regressors in the model. The model (B) has the most regressors.

(b) We have 10 predictors in the most complicated model. Thus,the number of possible models is 2^{10},1024.

(c)

```
mod2_1=lm(quality  ~  1 + gender+ numYears + numRaters + numCourses + pepper + dept+ helpfulness + clari
backAIC = step(mod2_1, direction="backward", data=Rateprof)
```

```
## Start:  AIC=-2180.38
## quality ~ 1 + gender + numYears + numRaters + numCourses + pepper +
##     dept + helpfulness + clarity + easiness + raterInterest
##
##                 Df Sum of Sq    RSS     AIC
## - dept          47    0.0253 0.7187 -2261.3
## - pepper         1    0.0000 0.6934 -2182.4
## - gender         1    0.0000 0.6934 -2182.4
## - raterInterest  1    0.0009 0.6943 -2181.9
## - numYears       1    0.0024 0.6957 -2181.1
## - numCourses     1    0.0029 0.6963 -2180.8
## - easiness       1    0.0036 0.6970 -2180.5
## <none>                       0.6934 -2180.4
## - numRaters      1    0.0165 0.7099 -2173.8
## - clarity        1    7.1520 7.8454 -1294.4
## - helpfulness    1    8.3857 9.0791 -1241.0
##
## Step:  AIC=-2261.28
## quality ~ gender + numYears + numRaters + numCourses + pepper +
##     helpfulness + clarity + easiness + raterInterest
##
##                 Df Sum of Sq     RSS     AIC
## - pepper         1    0.0001  0.7187 -2263.2
## - raterInterest  1    0.0001  0.7187 -2263.2
## - gender         1    0.0006  0.7193 -2263.0
## - numCourses     1    0.0024  0.7210 -2262.1
## - numYears       1    0.0024  0.7211 -2262.1
## <none>                       0.7187 -2261.3
```

```
## - easiness       1    0.0059  0.7245 -2260.3
## - numRaters      1    0.0136  0.7323 -2256.4
## - clarity        1    8.4893  9.2080 -1329.8
## - helpfulness    1   10.5279 11.2466 -1256.6
##
## Step:  AIC=-2263.25
## quality ~ gender + numYears + numRaters + numCourses + helpfulness +
##     clarity + easiness + raterInterest
##
##                 Df Sum of Sq     RSS     AIC
## - raterInterest  1    0.0001  0.7188 -2265.2
## - gender         1    0.0006  0.7193 -2264.9
## - numCourses     1    0.0023  0.7210 -2264.1
## - numYears       1    0.0023  0.7211 -2264.1
## <none>                        0.7187 -2263.2
## - easiness       1    0.0059  0.7246 -2262.3
## - numRaters      1    0.0136  0.7323 -2258.4
## - clarity        1    8.5775  9.2962 -1328.3
## - helpfulness    1   10.5611 11.2798 -1257.5
##
## Step:  AIC=-2265.2
## quality ~ gender + numYears + numRaters + numCourses + helpfulness +
##     clarity + easiness
##
##               Df Sum of Sq     RSS     AIC
## - gender       1    0.0006  0.7194 -2266.9
## - numYears     1    0.0023  0.7211 -2266.0
## - numCourses   1    0.0024  0.7212 -2266.0
## <none>                      0.7188 -2265.2
## - easiness     1    0.0060  0.7249 -2264.1
## - numRaters    1    0.0136  0.7324 -2260.4
## - clarity      1    8.6811  9.3999 -1326.3
## - helpfulness  1   10.7026 11.4214 -1255.0
##
## Step:  AIC=-2266.92
## quality ~ numYears + numRaters + numCourses + helpfulness + clarity +
##     easiness
##
##               Df Sum of Sq     RSS     AIC
## - numCourses   1    0.0023  0.7217 -2267.8
## - numYears     1    0.0026  0.7220 -2267.6
## <none>                      0.7194 -2266.9
## - easiness     1    0.0057  0.7251 -2266.0
## - numRaters    1    0.0136  0.7330 -2262.1
## - clarity      1    8.6805  9.3999 -1328.3
## - helpfulness  1   10.7527 11.4720 -1255.4
##
## Step:  AIC=-2267.76
## quality ~ numYears + numRaters + helpfulness + clarity + easiness
##
##               Df Sum of Sq     RSS     AIC
## - numYears     1    0.0019  0.7235 -2268.8
## <none>                      0.7217 -2267.8
## - easiness     1    0.0058  0.7275 -2266.8
```

```
## - numRaters    1    0.0163  0.7380 -2261.6
## - clarity      1    8.7828  9.5044 -1326.2
## - helpfulness  1   10.7845 11.5062 -1256.3
##
## Step:  AIC=-2268.82
## quality ~ numRaters + helpfulness + clarity + easiness
##
##               Df Sum of Sq     RSS     AIC
## <none>                      0.7235 -2268.8
## - easiness     1    0.0051  0.7286 -2268.2
## - numRaters    1    0.0145  0.7380 -2263.6
## - clarity      1    8.8197  9.5432 -1326.7
## - helpfulness  1   10.7834 11.5069 -1258.2
```

(i)The final model selsected is quality ~ numRaters + helpfulness + clarity + easiness.

```
mod2_2=lm(quality ~ numRaters + helpfulness + clarity + easiness, data = Rateprof)
summary(mod2_2)
```

```
##
## Call:
## lm(formula = quality ~ numRaters + helpfulness + clarity + easiness,
##     data = Rateprof)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75221 -0.00729  0.00271  0.01206  0.15629
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0147493  0.0124380  -1.186  0.23647
## numRaters   -0.0003745  0.0001394  -2.686  0.00756 **
## helpfulness  0.5354934  0.0073004  73.352  < 2e-16 ***
## clarity      0.4646970  0.0070050  66.338  < 2e-16 ***
## easiness     0.0058524  0.0036630   1.598  0.11098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04477 on 361 degrees of freedom
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9971
## F-statistic: 3.184e+04 on 4 and 361 DF,  p-value: < 2.2e-16
```

The fitted model is $\hat{quality} = -0.0147493 - 0.0003745 * numRaters + 0.5354934 * helpfulness + 0.4646970 * clarity + 0.0058524 * easiness$

(ii) 1+10+9+8+7+6+5+4=50.Totally,in this backward elimination process, R fits 50 models.

(iii)An example: quality ~ 1+dept + numYears + numRaters + numCourses + pepper + helpfulness + clarity + easiness + raterInterest.

(d)

```
backBIC = step(mod2_1, direction="backward", data=Rateprof,k=log(n))
```

```
## Start:  AIC=-1957.93
## quality ~ 1 + gender + numYears + numRaters + numCourses + pepper +
##     dept + helpfulness + clarity + easiness + raterInterest
##
```

```
##                 Df Sum of Sq    RSS     AIC
## - dept          47    0.0253 0.7187 -2222.2
## - pepper         1    0.0000 0.6934 -1963.8
## - gender         1    0.0000 0.6934 -1963.8
## - raterInterest  1    0.0009 0.6943 -1963.4
## - numYears       1    0.0024 0.6957 -1962.6
## - numCourses     1    0.0029 0.6963 -1962.3
## - easiness       1    0.0036 0.6970 -1961.9
## <none>                       0.6934 -1957.9
## - numRaters      1    0.0165 0.7099 -1955.2
## - clarity        1    7.1520 7.8454 -1075.9
## - helpfulness    1    8.3857 9.0791 -1022.4
##
## Step:  AIC=-2222.25
## quality ~ gender + numYears + numRaters + numCourses + pepper +
##     helpfulness + clarity + easiness + raterInterest
##
##                 Df Sum of Sq     RSS     AIC
## - pepper         1    0.0001  0.7187 -2228.1
## - raterInterest  1    0.0001  0.7187 -2228.1
## - gender         1    0.0006  0.7193 -2227.8
## - numCourses     1    0.0024  0.7210 -2227.0
## - numYears       1    0.0024  0.7211 -2226.9
## - easiness       1    0.0059  0.7245 -2225.2
## <none>                       0.7187 -2222.2
## - numRaters      1    0.0136  0.7323 -2221.3
## - clarity        1    8.4893  9.2080 -1294.7
## - helpfulness    1   10.5279 11.2466 -1221.5
##
## Step:  AIC=-2228.13
## quality ~ gender + numYears + numRaters + numCourses + helpfulness +
##     clarity + easiness + raterInterest
##
##                 Df Sum of Sq     RSS     AIC
## - raterInterest  1    0.0001  0.7188 -2234.0
## - gender         1    0.0006  0.7193 -2233.7
## - numCourses     1    0.0023  0.7210 -2232.8
## - numYears       1    0.0023  0.7211 -2232.8
## - easiness       1    0.0059  0.7246 -2231.0
## <none>                       0.7187 -2228.1
## - numRaters      1    0.0136  0.7323 -2227.2
## - clarity        1    8.5775  9.2962 -1297.1
## - helpfulness    1   10.5611 11.2798 -1226.3
##
## Step:  AIC=-2233.98
## quality ~ gender + numYears + numRaters + numCourses + helpfulness +
##     clarity + easiness
##
##               Df Sum of Sq     RSS     AIC
## - gender       1    0.0006  0.7194 -2239.6
## - numYears     1    0.0023  0.7211 -2238.7
## - numCourses   1    0.0024  0.7212 -2238.7
## - easiness     1    0.0060  0.7249 -2236.8
## <none>                     0.7188 -2234.0
```

```
## - numRaters    1    0.0136  0.7324 -2233.1
## - clarity      1    8.6811  9.3999 -1299.0
## - helpfulness  1   10.7026 11.4214 -1227.7
##
## Step:  AIC=-2239.6
## quality ~ numYears + numRaters + numCourses + helpfulness + clarity +
##     easiness
##
##               Df Sum of Sq     RSS     AIC
## - numCourses   1    0.0023  0.7217 -2244.3
## - numYears     1    0.0026  0.7220 -2244.2
## - easiness     1    0.0057  0.7251 -2242.6
## <none>                      0.7194 -2239.6
## - numRaters    1    0.0136  0.7330 -2238.6
## - clarity      1    8.6805  9.3999 -1304.8
## - helpfulness  1   10.7527 11.4720 -1231.9
##
## Step:  AIC=-2244.34
## quality ~ numYears + numRaters + helpfulness + clarity + easiness
##
##               Df Sum of Sq     RSS     AIC
## - numYears     1    0.0019  0.7235 -2249.3
## - easiness     1    0.0058  0.7275 -2247.3
## <none>                      0.7217 -2244.3
## - numRaters    1    0.0163  0.7380 -2242.1
## - clarity      1    8.7828  9.5044 -1306.7
## - helpfulness  1   10.7845 11.5062 -1236.8
##
## Step:  AIC=-2249.3
## quality ~ numRaters + helpfulness + clarity + easiness
##
##               Df Sum of Sq     RSS     AIC
## - easiness     1    0.0051  0.7286 -2252.6
## <none>                      0.7235 -2249.3
## - numRaters    1    0.0145  0.7380 -2248.0
## - clarity      1    8.8197  9.5432 -1311.1
## - helpfulness  1   10.7834 11.5069 -1242.6
##
## Step:  AIC=-2252.63
## quality ~ numRaters + helpfulness + clarity
##
##               Df Sum of Sq     RSS     AIC
## <none>                      0.7286 -2252.6
## - numRaters    1    0.0159  0.7445 -2250.6
## - clarity      1    8.8658  9.5945 -1315.1
## - helpfulness  1   11.3950 12.1237 -1229.4
```

(i)The final model selsected is quality ~ numRaters + helpfulness + clarity

```
mod2_3=lm(quality ~numRaters + helpfulness + clarity,data = Rateprof)
summary(mod2_3)
```

```
##
## Call:
## lm(formula = quality ~ numRaters + helpfulness + clarity, data = Rateprof)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75556 -0.00600  0.00187  0.01161  0.15943
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0069440  0.0114630  -0.606  0.54505
## numRaters   -0.0003915  0.0001393  -2.810  0.00522 **
## helpfulness  0.5379656  0.0071498  75.242  < 2e-16 ***
## clarity      0.4652794  0.0070105  66.369  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04486 on 362 degrees of freedom
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9971
## F-statistic: 4.228e+04 on 3 and 362 DF,  p-value: < 2.2e-16
```

The fitted model is $\hat{quality} = -0.0069440 - 0.0003915 * numRaters + 0.5379656 * helpfulness + 0.4652794 * clarity$. This model is not the same as the one selected by the backward elimination process with AIC in Part (c).

(ii)

Because in the formula, BIC has log(n) more than AIC. BIC has the heavy penelty term.

```
simplemod= lm(quality~ 1, data=Rateprof)
addto = list(lower = ~1,upper = ~ gender+ numYears + numRaters + numCourses + pepper + dept+ helpfulness
forwardAIC = step(simplemod, scope=addto,direction="forward", data=Rateprof)
```

```
## Start:  AIC=-128.82
## quality ~ 1
##
##                 Df Sum of Sq     RSS      AIC
## + helpfulness    1   246.388   9.620 -1327.79
## + clarity        1   243.848  12.161 -1242.02
## + easiness       1    81.758 174.251  -267.62
## + raterInterest  1    56.713 199.295  -218.47
## + pepper         1    35.915 220.093  -182.14
## + numCourses     1     1.903 254.105  -129.55
## + numRaters      1     1.872 254.137  -129.50
## <none>                       256.008  -128.82
## + gender         1     0.530 255.479  -127.58
## + numYears       1     0.325 255.683  -127.28
## + dept          47    31.682 224.327   -83.17
##
## Step:  AIC=-1327.79
## quality ~ helpfulness
##
##                 Df Sum of Sq    RSS     AIC
## + clarity        1    8.8756 0.7445 -2262.3
## + numCourses     1    0.1048 9.5153 -1329.8
## + raterInterest  1    0.0879 9.5323 -1329.2
## + pepper         1    0.0694 9.5508 -1328.4
## + easiness       1    0.0566 9.5635 -1328.0
## <none>                       9.6201 -1327.8
```

```
## + numRaters      1    0.0257 9.5945 -1326.8
## + numYears       1    0.0107 9.6095 -1326.2
## + gender         1    0.0014 9.6187 -1325.8
## + dept          47    1.5766 8.0436 -1299.3
##
## Step:  AIC=-2262.34
## quality ~ helpfulness + clarity
##
##                 Df Sum of Sq     RSS     AIC
## + numRaters      1 0.0158925 0.72862 -2268.2
## + easiness       1 0.0065470 0.73797 -2263.6
## + numCourses     1 0.0053016 0.73921 -2263.0
## <none>                       0.74452 -2262.3
## + numYears       1 0.0001256 0.74439 -2260.4
## + gender         1 0.0000823 0.74443 -2260.4
## + pepper         1 0.0000291 0.74449 -2260.4
## + raterInterest  1 0.0000111 0.74450 -2260.3
## + dept          47 0.0215126 0.72300 -2179.1
##
## Step:  AIC=-2268.24
## quality ~ helpfulness + clarity + numRaters
##
##                 Df Sum of Sq     RSS     AIC
## + easiness       1 0.0051161 0.72351 -2268.8
## <none>                       0.72862 -2268.2
## + numCourses     1 0.0017586 0.72686 -2267.1
## + numYears       1 0.0011402 0.72748 -2266.8
## + raterInterest  1 0.0003144 0.72831 -2266.4
## + gender         1 0.0002463 0.72838 -2266.4
## + pepper         1 0.0000060 0.72862 -2266.2
## + dept          47 0.0252781 0.70334 -2187.2
##
## Step:  AIC=-2268.82
## quality ~ helpfulness + clarity + numRaters + easiness
##
##                 Df Sum of Sq     RSS     AIC
## <none>                       0.72351 -2268.8
## + numYears       1 0.0018542 0.72165 -2267.8
## + numCourses     1 0.0015227 0.72198 -2267.6
## + gender         1 0.0006559 0.72285 -2267.2
## + raterInterest  1 0.0001264 0.72338 -2266.9
## + pepper         1 0.0000331 0.72347 -2266.8
## + dept          47 0.0240624 0.69944 -2187.2
```

(a) The final model selsected is quality ~ helpfulness + clarity + numRaters + easiness.

```
mod2_4=lm(quality ~ helpfulness + clarity + numRaters + easiness, data = Rateprof)
summary(mod2_4)
```

```
##
## Call:
## lm(formula = quality ~ helpfulness + clarity + numRaters + easiness,
##     data = Rateprof)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.75221 -0.00729  0.00271  0.01206  0.15629
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0147493  0.0124380  -1.186  0.23647
## helpfulness  0.5354934  0.0073004  73.352  < 2e-16 ***
## clarity      0.4646970  0.0070050  66.338  < 2e-16 ***
## numRaters   -0.0003745  0.0001394  -2.686  0.00756 **
## easiness     0.0058524  0.0036630   1.598  0.11098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04477 on 361 degrees of freedom
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9971
## F-statistic: 3.184e+04 on 4 and 361 DF,  p-value: < 2.2e-16
```

The fitted model is $\hat{quality} = -0.0147493 + 0.5354934 * helpfulness + 0.4646970 * clarity - 0.0003745 * numRaters + 0.0058524 * easiness$

(b) In the part (c),the final model selsected is quality ~ numRaters + helpfulness + clarity + easiness. So,the result is consistent with that of part (c).

(c)In the final model of backward elimination, AIC is -2268.82.In the final model of forward elimination AIC=-2268.82. So, they are consistent. Because, they select the same model.

Problem 4

Case 1

```
set.seed(3032)
e = rnorm(100, 0, 2)
x1 = rnorm(100)
x2 = x1 + 2
y = x1 + 0.1*x2 + e
dat1 = data.frame(x1 = x1, x2 = x2, y = y)
```

(a)

```
cor(x1,x2)
```

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

(b)

```
mod3_1=lm(y~1+x1+x2,data = dat1)
summary(mod3_1)
```

```
##
## Call:
## lm(formula = y ~ 1 + x1 + x2, data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2848 -1.5666 -0.0297  1.3564  6.1789
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1890     0.2106   0.897    0.372
```

```
## x1            1.0262     0.2203   4.658    1e-05 ***
## x2                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.098 on 98 degrees of freedom
## Multiple R-squared:  0.1813, Adjusted R-squared:  0.1729
## F-statistic:  21.7 on 1 and 98 DF,  p-value: 1.004e-05
```

Because X2=x1+2,x1 and x2 have a linear relationship. Hence, R does not give an estimator for the slope of x2.

For the fitted model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x1 + \hat{\beta}_2 * x2$. But x2=x1+2,the model could be written as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x1 + \hat{\beta}_2 * (x1 + 2) = \hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 + \hat{\beta}_2) * x1 + 2 * \hat{\beta}_2$.

Case 2

```
  set.seed(3032)
  e = rnorm(100, 0, 2)
  x1 = rnorm(100)
  x2 = x1 + 2 + rnorm(100, 0, 0.5)
  y = x1 + 0.1*x2 + e
  dat2 = data.frame(x1 = x1, x2 = x2, y = y)
```

(a)

```
cor(x1,x2)
```

```
## [1] 0.8825145
```

(b)

```
mod3_2=lm(y~1+x1+x2,data = dat2)
summary(mod3_2)
```

```
##
## Call:
## lm(formula = y ~ 1 + x1 + x2, data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2883 -1.6227  0.0646  1.2512  6.1781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6213     0.9028  -0.688    0.493
## x1            0.6380     0.4697   1.358    0.177
## x2            0.4068     0.4413   0.922    0.359
##
## Residual standard error: 2.104 on 97 degrees of freedom
## Multiple R-squared:  0.1861, Adjusted R-squared:  0.1693
## F-statistic: 11.09 on 2 and 97 DF,  p-value: 4.599e-05
```

The fitted model is $\hat{y} = -0.6213 + 0.6380 * x1 + 0.4068 * x2$, the estimates of the coefficients $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ are not consistent with the true parameters $(\beta_0 = 0, \beta_1 = 1, \beta_2 = 0.1)$.

Case 3

```
 set.seed(999)
 e = rnorm(100, 0, 2)
```

```
 x1 = rnorm(100)
x2 = x1 + 2 + rnorm(100, 0, 0.5)
 y = x1 + 0.1*x2 + e
  dat2 = data.frame(x1 = x1, x2 = x2, y = y)
```

(a)

```
cor(x1,x2)
```

```
## [1] 0.8636606
```

(b)

```
mod3_3=lm(y~1+x1+x2,data = dat2)
summary(mod3_3)
```

```
##
## Call:
## lm(formula = y ~ 1 + x1 + x2, data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4818 -1.6162  0.1636  1.2418  5.0283
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8181     0.6942   1.178   0.2415
## x1            0.9794     0.4028   2.431   0.0169 *
## x2           -0.4531     0.3458  -1.310   0.1932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.809 on 97 degrees of freedom
## Multiple R-squared:  0.07939,    Adjusted R-squared:  0.06041
## F-statistic: 4.182 on 2 and 97 DF,  p-value: 0.0181
```

The fitted model is $\hat{y} = 0.8181 + 0.9794 * x1 - 0.4531 * x2$,the estimates of the coefficients $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ are not similiar to the estimates in Case2.Because the values of correaltion between x1 and x2 are diffrent.