TOPIC3 In Class Problems

PROBLEM 10

Use the CREDIT.CSV data.

(a) Use stepwise regression to identify potentially important predictor variables (make sure to specify which are qualitative).

(b) Manually build a model in which these potentially important variables have two-way interactions.

(c) Use stepwise regression to find a final model based on this interaction model in which all predictors contribute.

ANSWER TO PROBLEM 10

(a) Use stepwise regression to identify potentially important predictor variables (make sure to specify which are qualitative).
(b) Manually build a model in which these potentially important variables have two-way interactions.
(c) Use stepwise regression to find a final model based on this interaction model in which all predictors contribute.
.

```
fullmodel<-lm(Balance~Income+Limit+Rating+Cards+Age+Education+factor(Gender)+factor(Ethnicity)
+factor(Married)+factor(Student),data=credit)
stepmod=ols_step_both_p(fullmodel,pent = 0.1, prem = 0.3, details=TRUE)
summary(stepmod$model)
```

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + "
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-170.00  -77.85  -11.84   56.87  313.52

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -493.73419   24.82476 -19.889  < 2e-16 ***
Rating               1.09119    0.48480   2.251   0.0250 *
Income              -7.79508    0.23342 -33.395  < 2e-16 ***
factor(Student)Yes 425.60994   16.50956  25.780  < 2e-16 ***
Limit                0.19369    0.03238   5.981 4.98e-09 ***
Cards               18.21190    4.31865   4.217 3.08e-05 ***
Age                 -0.62406    0.29182  -2.139   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.61 on 393 degrees of freedom
Multiple R-squared:  0.9547,    Adjusted R-squared:  0.954
F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

ANSWER TO PROBLEM 10

(a) Use stepwise regression to identify potentially important predictor variables (make sure to specify which are qualitative).

(b) Manually build a model in which these potentially important variables have two-way interactions.

(c) Use stepwise regression to find a final model based on this interaction model in which all predictors contribute.

```
bestmodel2<-lm(Balance~Income+Limit+Rating+Cards+Age+factor(Student)+Income*Rating
+Income*factor(Student)+Limit*Rating+Limit*factor(Student),data=credit)
summary(bestmodel2)
```

```
Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
    factor(Student) + Income * Rating + Income * factor(Student) +
    Limit * Rating + Limit * factor(Student), data = credit)

Residuals:
     Min       1Q   Median       3Q      Max
-231.817  -41.097    7.283   38.913  153.038

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.945e+02  2.160e+01  -9.006  < 2e-16 ***
Income                     -1.837e+00  5.235e-01  -3.508 0.000504 ***
Limit                       1.079e-01  2.158e-02   5.000 8.70e-07 ***
Rating                     -3.121e-01  3.200e-01  -0.976 0.329914
Cards                       1.832e+01  2.786e+00   6.575 1.57e-10 ***
Age                        -7.660e-01  1.886e-01  -4.063 5.87e-05 ***
factor(Student)Yes          1.555e+02  2.634e+01   5.905 7.68e-09 ***
Income:Rating              -1.694e-02  1.187e-03 -14.272  < 2e-16 ***
Income:factor(Student)Yes  -1.784e+00  4.460e-01  -4.001 7.55e-05 ***
Limit:Rating                3.373e-04  1.711e-05  19.710  < 2e-16 ***
Limit:factor(Student)Yes    7.868e-02  7.666e-03  10.264  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.6 on 389 degrees of freedom
Multiple R-squared:  0.9813,    Adjusted R-squared:  0.9809
F-statistic:  2046 on 10 and 389 DF,  p-value: < 2.2e-16
```

ANSWER TO PROBLEM 10

(a) Use stepwise regression to identify potentially important predictor variables (make sure to specify which are qualitative).
(b) Manually build a model in which these potentially important variables have two-way interactions.
(c) Use stepwise regression to find a final model based on this interaction model in which all predictors contribute.
.

```
bestmodel1<-lm(Balance~Income+Limit+Rating+Cards+Age+factor(Student)+Rating*Limit
+Rating*Income+factor(Student)*Income+factor(Student)+Limit+Rating*Age+Income*Age
,data=credit) ## This includes all
summary(bestmodel1)
```

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Residuals:
     Min       1Q   Median       3Q      Max
-189.484  -41.039    7.709   37.960  161.701

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -2.712e+02  3.225e+01  -8.409 8.12e-16 ***
Income                      -2.818e+00  6.643e-01  -4.243 2.77e-05 ***
factor(Student)Yes           2.115e+02  4.253e+01   4.973 9.95e-07 ***
Age                          5.513e-01  4.794e-01   1.150 0.250820
Cards                        1.757e+01  2.764e+00   6.356 5.86e-10 ***
Rating                       2.009e-01  3.525e-01   0.570 0.568976
Limit                        9.471e-02  2.233e-02   4.242 2.78e-05 ***
Rating:factor(Student)Yes -1.317e+00  8.839e-01  -1.490 0.137125
Rating:Limit                 3.425e-04  1.707e-05  20.068  < 2e-16 ***
Rating:Income               -1.721e-02  1.189e-03 -14.475  < 2e-16 ***
factor(Student)Yes:Income -1.637e+00  4.452e-01  -3.678 0.000268 ***
factor(Student)Yes:Limit     1.642e-01  5.945e-02   2.762 0.006020 **
Rating:Age                  -6.274e-03  1.953e-03  -3.212 0.001429 **
Income:Age                   1.945e-02  8.694e-03   2.237 0.025863 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.85 on 386 degrees of freedom
Multiple R-squared:  0.9819,    Adjusted R-squared:  0.9813
F-statistic:  1613 on 13 and 386 DF,  p-value: < 2.2e-16
```

PROBLEM 11

Use the CREDIT.CSV data.

(a) Use backwards regression to identify potentially important predictor variables (make sure to specify which are qualitative).

(b) In what ways is it different than additive-only stepwise model?

Use all variables (including factors) – additive model

ols_step_backward_p(_____)

ANSWER TO PROBLEM 11

(a) Use backwards regression to identify potentially important predictor variables (make sure to specify which are qualitative).

(b) In what ways is it different than the additive-only stepwise model

```
fullmodel<-lm(Balance~Income+Limit+Rating+Cards+Age+Education+factor(Gender)+factor(Ethnicity)+
factor(Married)+factor(Student),data=credit)
backmodel=ols_step_backward_p(fullmodel,prem = 0.3,details=TRUE)
summary(backmodel$model)
```

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + "
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-174.30  -77.35  -12.01   55.99  308.38

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -488.61587   25.28900 -19.321  < 2e-16 ***
Income                -7.80363    0.23352 -33.417  < 2e-16 ***
Limit                  0.19362    0.03238   5.980 5.02e-09 ***
Rating                 1.09405    0.48474   2.257   0.0246 *
Cards                 18.10917    4.31910   4.193 3.41e-05 ***
Age                   -0.62065    0.29179  -2.127   0.0340 *
factor(Gender)Female  -10.45315    9.88956  -1.057   0.2912
factor(Student)Yes    426.58126   16.53266  25.802  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 98.6 on 392 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.954
F-statistic:  1183 on 7 and 392 DF,  p-value: < 2.2e-16
```

ANSWER TO PROBLEM 11

(a) Use backwards regression to identify potentially important predictor variables (make sure to specify which are qualitative).

(b) In what ways is it different than the additive-only stepwise model

.

Backwards regression

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + "
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-170.00  -77.85  -11.84   56.87  313.52

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -493.73419   24.82476 -19.889  < 2e-16 ***
Rating                1.09119    0.48480   2.251   0.0250 *
Income               -7.79508    0.23342 -33.395  < 2e-16 ***
factor(Student)Yes  425.60994   16.50956  25.780  < 2e-16 ***
Limit                 0.19369    0.03238   5.981 4.98e-09 ***
Cards                18.21190    4.31865   4.217 3.08e-05 ***
Age                  -0.62406    0.29182  -2.139   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.61 on 393 degrees of freedom
Multiple R-squared:  0.9547,    Adjusted R-squared:  0.954
F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

Stepwise regression

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + "
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-174.30  -77.35  -12.01   55.99  308.38

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -488.61587   25.28900 -19.321  < 2e-16 ***
Income               -7.80363    0.23352 -33.417  < 2e-16 ***
Limit                 0.19362    0.03238   5.980 5.02e-09 ***
Rating                1.09405    0.48474   2.257   0.0246 *
Cards                18.10917    4.31910   4.193 3.41e-05 ***
Age                  -0.62065    0.29179  -2.127   0.0340 *
factor(Gender)Female -10.45315    9.88956  -1.057   0.2912
factor(Student)Yes   426.58126   16.53266  25.802  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.6 on 392 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.954
F-statistic:  1183 on 7 and 392 DF,  p-value: < 2.2e-16
```

PROBLEM 12

Use the CREDIT.CSV data.

(a) Use forward regression to identify potentially important predictor variables (make sure to specify which are qualitative).

(b) In what ways is it different than the others?

ANSWER TO PROBLEM 12

(a) Use forward regression to identify potentially important predictor variables (make sure to specify which are qualitative).

(b) In what ways is it different than the others?

Forward regression

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-170.00  -77.85  -11.84   56.87  313.52

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -493.73419   24.82476 -19.889  < 2e-16 ***
Rating                1.09119    0.48480   2.251   0.0250 *
Income               -7.79508    0.23342 -33.395  < 2e-16 ***
factor(Student)Yes  425.60994   16.50956  25.780  < 2e-16 ***
Limit                 0.19369    0.03238   5.981 4.98e-09 ***
Cards                18.21190    4.31865   4.217 3.08e-05 ***
Age                  -0.62406    0.29182  -2.139   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.61 on 393 degrees of freedom
Multiple R-squared:  0.9547,    Adjusted R-squared:  0.954
F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

Backwards regression

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + "
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-170.00  -77.85  -11.84   56.87  313.52

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -493.73419   24.82476 -19.889  < 2e-16 ***
Rating                1.09119    0.48480   2.251   0.0250 *
Income               -7.79508    0.23342 -33.395  < 2e-16 ***
factor(Student)Yes  425.60994   16.50956  25.780  < 2e-16 ***
Limit                 0.19369    0.03238   5.981 4.98e-09 ***
Cards                18.21190    4.31865   4.217 3.08e-05 ***
Age                  -0.62406    0.29182  -2.139   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.61 on 393 degrees of freedom
Multiple R-squared:  0.9547,    Adjusted R-squared:  0.954
F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

Stepwise regression

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + "
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-174.30  -77.35  -12.01   55.99  308.38

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -488.61587   25.28900 -19.321  < 2e-16 ***
Income               -7.80363    0.23352 -33.417  < 2e-16 ***
Limit                 0.19362    0.03238   5.980 5.02e-09 ***
Rating                1.09405    0.48474   2.257   0.0246 *
Cards                18.10917    4.31910   4.193 3.41e-05 ***
Age                  -0.62065    0.29179  -2.127   0.0340 *
factor(Gender)Female -10.45315    9.88956  -1.057   0.2912
factor(Student)Yes   426.58126   16.53266  25.802  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.6 on 392 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.954
F-statistic:  1183 on 7 and 392 DF,  p-value: < 2.2e-16
```

PROBLEM 13

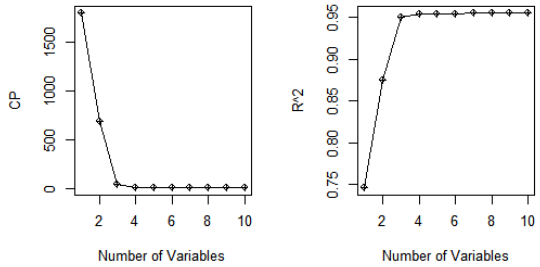
Use the CREDIT.CSV data.

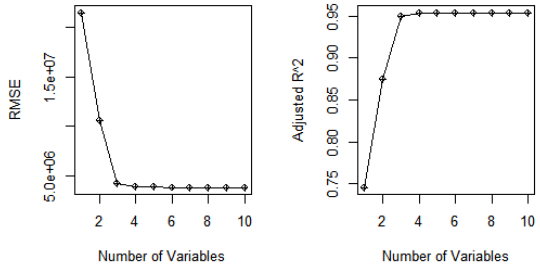
(a) Use all subsets to identify best model. Use: `leaps::regsubsets()`

(b) Is it the same as with the other methods?

ANSWER TO PROBLEM 13

```
1 subsets of each size up to 10
Selection Algorithm: exhaustive
         Income Limit Rating Cards Age Education factor(Gender)Female factor(Ethnicity)Asian factor(Ethnicity)Caucasian factor(Married)Yes factor(Student)Yes
1  ( 1 )  " "    " "   "*"    " "   " " " "       " "                 " "                    " "                      " "               " "
2  ( 1 )  "*"    " "   "*"    " "   " " " "       " "                 " "                    " "                      " "               "*"
3  ( 1 )  "*"    " "   "*"    " "   " " " "       " "                 " "                    " "                      " "               "*"
4  ( 1 )  "*"    "*"   " "    "*"   " " " "       " "                 " "                    " "                      " "               "*"
5  ( 1 )  "*"    "*"   "*"    "*"   " " " "       " "                 " "                    " "                      " "               "*"
6  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       " "                 " "                    " "                      " "               "*"
7  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"                 " "                    " "                      " "               "*"
8  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"                 "*"                    " "                      " "               "*"
9  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"                 "*"                    " "                      "*"               "*"
10  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       "*"                 "*"                    "*"                      "*"               "*"
```






```
        rsquare          cp     RMSE AdjustedR
[1,] 0.7458484 1800.308406 21435122 0.7452098
[2,] 0.8751179  685.196514 10532541 0.8744888
[3,] 0.9498788   41.133867  4227219 0.9494991
[4,] 0.9535800   11.148910  3915058 0.9531099
[5,] 0.9541606    8.131573  3866091 0.9535789
[6,] 0.9546879    5.574883  3821620 0.9539961
[7,] 0.9548167    6.462042  3810759 0.9540098
[8,] 0.9548880    7.845931  3804746 0.9539649
[9,] 0.9549636    9.192355  3798367 0.9539243
[10,] 0.9550468  10.472883  3791345 0.9538912
```

ANSWER TO PROBLEM 13

All subsets

```
Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
    factor(Gender) + factor(Student), data = credit)

Residuals:
    Min      1Q  Median      3Q     Max
-174.30  -77.35  -12.01   55.99  308.38

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -488.61587   25.28900 -19.321  < 2e-16 ***
Income                 -7.80363    0.23352 -33.417  < 2e-16 ***
Limit                   0.19362    0.03238   5.980 5.02e-09 ***
Rating                  1.09405    0.48474   2.257   0.0246 *
Cards                  18.10917    4.31910   4.193 3.41e-05 ***
Age                    -0.62065    0.29179  -2.127   0.0340 *
factor(Gender)Female  -10.45315    9.88956  -1.057   0.2912
factor(Student)Yes    426.58126   16.53266  25.802  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.6 on 392 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.954
F-statistic:  1183 on 7 and 392 DF,  p-value: < 2.2e-16
```

Forward regression

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-170.00  -77.85  -11.84   56.87  313.52

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -493.73419   24.82476 -19.889  < 2e-16 ***
Rating                1.09119    0.48480   2.251   0.0250 *
Income               -7.79508    0.23342 -33.395  < 2e-16 ***
factor(Student)Yes  425.60994   16.50956  25.780  < 2e-16 ***
Limit                 0.19369    0.03238   5.981 4.98e-09 ***
Cards                18.21190    4.31865   4.217 3.08e-05 ***
Age                  -0.62406    0.29182  -2.139   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.61 on 393 degrees of freedom
Multiple R-squared:  0.9547,    Adjusted R-squared:  0.954
F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

Backwards regression

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + "
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-170.00  -77.85  -11.84   56.87  313.52

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -493.73419   24.82476 -19.889  < 2e-16 ***
Rating                1.09119    0.48480   2.251   0.0250 *
Income               -7.79508    0.23342 -33.395  < 2e-16 ***
factor(Student)Yes  425.60994   16.50956  25.780  < 2e-16 ***
Limit                 0.19369    0.03238   5.981 4.98e-09 ***
Cards                18.21190    4.31865   4.217 3.08e-05 ***
Age                  -0.62406    0.29182  -2.139   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.61 on 393 degrees of freedom
Multiple R-squared:  0.9547,    Adjusted R-squared:  0.954
F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

Stepwise regression

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + "
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-174.30  -77.35  -12.01   55.99  308.38

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -488.61587   25.28900 -19.321  < 2e-16 ***
Income                 -7.80363    0.23352 -33.417  < 2e-16 ***
Limit                   0.19362    0.03238   5.980 5.02e-09 ***
Rating                  1.09405    0.48474   2.257   0.0246 *
Cards                  18.10917    4.31910   4.193 3.41e-05 ***
Age                    -0.62065    0.29179  -2.127   0.0340 *
factor(Gender)Female  -10.45315    9.88956  -1.057   0.2912
factor(Student)Yes    426.58126   16.53266  25.802  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.6 on 392 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.954
F-statistic:  1183 on 7 and 392 DF,  p-value: < 2.2e-16
```