TOPIC2 In Class Problems

In-class Practice Problem 4

From the condominium problem:

(a) Create a model with all interaction terms (based on the best model from last class)

(b) Are interaction terms appropriate? Conduct a test to find out.

ANSWER PROBLEM 4

```
Call:
lm(formula = listprice ~ livingarea + floors + baths, data = condo)

Residuals:
    Min      1Q  Median      3Q     Max
-11.796  -1.483   1.077   2.903  11.892

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.590      7.501   2.078 0.061888 .
livingarea    65.192      6.446  10.114  6.6e-07 ***
floors       -14.925      5.465  -2.731 0.019533 *
baths         28.381      5.715   4.966 0.000425 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.622 on 11 degrees of freedom
Multiple R-squared:  0.9706,    Adjusted R-squared:  0.9625
F-statistic: 120.9 on 3 and 11 DF,  p-value: 1.059e-08
```

```
Call:
lm(formula = listprice ~ livingarea + floors + baths + livingarea:baths +
    livingarea:floors + floors:baths, data = condo)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2701  -2.0116  -0.4466   3.8389   6.2406

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         40.502     77.021   0.526    0.613
livingarea          27.339     44.761   0.611    0.558
floors             -19.319    109.400  -0.177    0.864
baths               28.625     24.946   1.147    0.284
livingarea:baths    10.209     31.157   0.328    0.752
livingarea:floors   13.143     25.315   0.519    0.618
floors:baths        -8.062     37.341  -0.216    0.834

Residual standard error: 6.611 on 8 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9626
F-statistic: 61.13 on 6 and 8 DF,  p-value: 3.008e-06
Analysis of Variance Table
```

```
Model 1: listprice ~ livingarea + floors + baths +
livingarea:baths +
    livingarea:floors + floors:baths
Model 2: listprice ~ livingarea + floors + baths
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 349.67
2     11 482.39 -3   -132.72 1.0121 0.4364
```

In-class Practice Problem 5

sales.csv

Data on last year's sale (Y in 100,000s dollars) in 40 sales districts are given in the sales.csv file. This file also contains:

- promotional expenditures (X_1: in 1,000s dollars),
- the number of active accounts (X_2),
- the number of competing brands (X_3) and
- the district potential (X_4, coded) for each of the district (OMIT THIS VARIABLE FOR NOW)

1. Find the best fit additive to predict sales using some or all of the variables X_1, X_2, X_3 only.
2. Find the best fit model with interaction terms (if needed) using some or all of the variables X_1, X_2, X_3
3. Which model would you choose? Explain.
4. Once you obtain the best fit model, interpret the regression coefficient for X_3 (Hint: it will interact with another variable).

X_3 → involve an interaction

ANSWER PROBLEM 5 (1 of 3)

PART 1

```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = sale)

Residuals:
     Min       1Q   Median       3Q      Max
-106.803   -6.726   -1.967    7.072   81.964

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 162.2269    31.0376   5.227 7.50e-06 ***
X1            2.0192     2.5763   0.784    0.438
X2            3.4568     0.3426  10.088 4.91e-12 ***
X3          -19.4589     1.8054 -10.778 8.08e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.35 on 36 degrees of freedom
Multiple R-squared:  0.9175,    Adjusted R-squared:  0.9106
F-statistic: 133.4 on 3 and 36 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Y ~ X2 + X3, data = sale)

Residuals:
     Min       1Q   Median       3Q      Max
-109.096   -5.888   -3.440    8.780   83.982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 172.4595    28.0109   6.157 3.85e-07 ***
X2            3.5011     0.3362  10.414 1.50e-12 ***
X3          -19.7308     1.7625 -11.195 1.94e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.22 on 37 degrees of freedom
Multiple R-squared:  0.9161,    Adjusted R-squared:  0.9115
F-statistic: 201.9 on 2 and 37 DF,  p-value: < 2.2e-16
Analysis of Variance Table

Model 1: Y ~ X2 + X3
Model 2: Y ~ X1 + X2 + X3
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     37 23532
2     36 23137  1    394.79 0.6143 0.4383
```

ANSWER PROBLEM 5 (2 of 3)

PART 2

```
Call:
lm(formula = Y ~ X2 + X3 + X2 : X3, data = sale)

Residuals:
    Min      1Q  Median      3Q     Max
-98.788  -6.804  -1.861   6.225  58.055

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.3191    62.5599   0.309    0.7592
X2            6.0809     1.0084   6.030 6.33e-07 ***
X3           -2.9261     6.4576  -0.453    0.6532
X2:X3        -0.2903     0.1079  -2.689    0.0108 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.33 on 36 degrees of freedom
Multiple R-squared:  0.9301,    Adjusted R-squared:  0.9243
F-statistic: 159.7 on 3 and 36 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Y ~ (X1 + X2 + X3)^2, data = sale)

Residuals:
    Min      1Q  Median      3Q     Max
-93.253  -9.208   0.852   6.606  51.455

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.0421    91.7901  -0.240  0.81171
X1           11.9200    14.7995   0.805  0.42633
X2            4.8325     1.6708   2.892  0.00672 **
X3            7.0945     7.8501   0.904  0.37268
X1:X2         0.1193     0.2169   0.550  0.58607
X1:X3        -2.1302     0.9540  -2.233  0.03246 *
X2:X3        -0.2495     0.1168  -2.136  0.04021 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.59 on 33 degrees of freedom
Multiple R-squared:   0.94, Adjusted R-squared:  0.929
F-statistic: 86.09 on 6 and 33 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X2 : X3 + X1 : X3, data = sale)

Residuals:
    Min      1Q  Median      3Q     Max
-93.180  -9.362   0.929   7.712  53.205

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -55.0832    68.6781  -0.802   0.4281
X1           18.4248     8.8028   2.093   0.0439 *
X2            5.5102     1.1166   4.935 2.09e-05 ***
X3            6.9378     7.7640   0.894   0.3778
X2:X3        -0.2524     0.1155  -2.185   0.0358 *
X1:X3        -2.1387     0.9441  -2.265   0.0300 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.35 on 34 degrees of freedom
Multiple R-squared:  0.9394,    Adjusted R-squared:  0.9305
F-statistic: 105.4 on 5 and 34 DF,  p-value: < 2.2e-16
```

ANSWER PROBLEM 5 (3 of 3)

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X2 : X3 + X1 : X3, data = sale)

Residuals:
    Min      1Q  Median      3Q     Max
-93.180  -9.362   0.929   7.712  53.205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -55.0832    68.6781  -0.802   0.4281
X1           18.4248     8.8028   2.093   0.0439 *
X2            5.5102     1.1166   4.935 2.09e-05 ***
X3            6.9378     7.7640   0.894   0.3778
X2:X3        -0.2524     0.1155  -2.185   0.0358 *
X1:X3        -2.1387     0.9441  -2.265   0.0300 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.35 on 34 degrees of freedom
Multiple R-squared:  0.9394,    Adjusted R-squared:  0.9305
F-statistic: 105.4 on 5 and 34 DF,  p-value: < 2.2e-16
```

1. Which model would you choose? Explain.
2. Once you obtain the best fit model, interpret the regression coefficient for X_3 (Hint: it will interact with another variable).

Inclass Practice Problem 6

Suppose that we wish to investigate differences in credit card balance between marital status. Based on the Married variable, we can create a dummy variable which 0 is NO and 1 is Yes.

(a) Create a simple linear regression model to predict the credit card balance by using the Married variable.

(b) How much is the average credit card debt for an unmarried person.

(c) What is the difference in debt between a married and single person.

Ignore the individual t-test output

ANSWER PROBLEM 6

```
Call:
lm(formula = Balance ~ factor(Married), data = credit)

Residuals:
    Min      1Q  Median      3Q     Max
-523.29 -451.03  -60.12  345.06 1481.06

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          523.290     36.974  14.153   <2e-16 ***
factor(Married)Yes    -5.347     47.244  -0.113     0.91
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.3 on 398 degrees of freedom
Multiple R-squared:  3.219e-05, Adjusted R-squared:  -0.00248
F-statistic: 0.01281 on 1 and 398 DF,  p-value: 0.9099
```

	number	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
1	1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	3	104.593	7075	514	4	71	11	Male	No	No	Asian	580
4	4	148.924	9504	681	3	36	11	Female	No	No	Asian	964
5	5	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331
6	6	80.180	8047	569	4	77	10	Male	No	No	Caucasian	1151

$\hat{\beta}_0 + \hat{\beta}_1 x_1$

$\hat{y} = 523.29 - 5.347\,[\text{MARRIED}]$

$\hat{\beta}_0 = 523.29 \rightarrow$ Average debt for unmarried. ($x_1 = 0$)

$\hat{\beta}_1 = -5.35 \rightarrow$ Difference (between levels)

$\hat{\beta}_0 + \hat{\beta}_1 = 523.29 + (-5.35)$

$= 517.94$ average debt for married.

In-class Practice Problem 7

There is always a certain curiosity and controversy surrounding professors' salaries and whether they are overpaid or not paid enough. A university would like to study the effects of ranks and departments on salaries. 30 observations were randomly chosen from 3 different departments. The data are provided in the salary.csv data file. Dept= Department (1=Family Studies, 2=Biology, 3=Business)

Instead of the rank variable, practice how to interpret the dept variable.

STATE THE AVERAGE SALARY OF A PERSON FROM EACH DEPT

ANSWER PROBLEM 7

```
Call:
lm(formula = salary ~ factor(dept), data = salary)

Residuals:
    Min      1Q  Median      3Q     Max
-12.250  -6.838  -3.925   4.662  30.000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    42.250      2.788  15.154 1.01e-14 ***
factor(dept)2   7.750      4.408   1.758  0.09008 .
factor(dept)3  12.350      4.135   2.986  0.00594 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.658 on 27 degrees of freedom
Multiple R-squared:  0.2543,    Adjusted R-squared:  0.199
F-statistic: 4.603 on 2 and 27 DF,  p-value: 0.01905
```

$$\hat{salary} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

(0 or 1) (0 or 1)

$$\hat{salary} = \begin{cases} \hat{\beta}_0 & \text{FAMILY STUDIES} \quad (0,0) \\ \hat{\beta}_0 + \hat{\beta}_1 & \text{BIOLOGY} \quad (1,0) \\ \hat{\beta}_0 + \hat{\beta}_2 & \text{BUSINESS} \quad (0,1) \end{cases}$$

(β_1, β_2)

Average

FAMILY STUDIES

$\$\hat{\beta}_0 \times 1000 = \$42,250$

BIOLOGY

$\$(\hat{\beta}_0 + \hat{\beta}_1) \times 1000 = \$42,250 + 7,750$

BUSINESS

$\$(\hat{\beta}_0 + \hat{\beta}_2) \times 1000 = 42,250 + 12350$

Inclass Practice Problem 8

From the credit card example, use the lm() function to perform the best fit model. How would you interpret the regression coeffients (if possible)? Would you recommend this model for predictive purpose?

STEPS
1. Build additive model
2. Determine significant predictors
3. Build interaction model with significant predictors
4. Remove non-significant interactions
5. Rerun model to ensure all predictors significant
6. Iterate at step 5 until done.

(a) Build full additive model with only significant predictors

(b) Build interacting model with predictors from (a)

(c) Remove non-significant interactions and rerun model

ANSWER PROBLEM 8

(a) Build full additive model with only significant predictors
(b) Build interacting model with predictors from (a)
(c) Remove non-significant interactions and rerun model
(d) Interpret the final model

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

Residuals:
    Min      1Q  Median      3Q     Max
-161.64  -77.70  -13.49   53.98  318.20

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -479.20787   35.77394 -13.395  < 2e-16 ***
Income                     -7.80310    0.23423 -33.314  < 2e-16 ***
Limit                       0.19091    0.03278   5.824 1.21e-08 ***
Rating                      1.13653    0.49089   2.315   0.0211 *
Cards                      17.72448    4.34103   4.083 5.40e-05 ***
Age                        -0.61391    0.29399  -2.088   0.0374 *
Education                  -1.09886    1.59795  -0.688   0.4921
factor(Gender)Female      -10.65325    9.91400  -1.075   0.2832
factor(Ethnicity)Asian     16.80418   14.11906   1.190   0.2347
factor(Ethnicity)Caucasian 10.10703   12.20992   0.828   0.4083
factor(Married)Yes         -8.53390   10.36287  -0.824   0.4107
factor(Student)Yes        425.74736   16.72258  25.459  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 98.79 on 388 degrees of freedom
Multiple R-squared:  0.9551,    Adjusted R-squared:  0.9538
F-statistic: 750.3 on 11 and 388 DF,  p-value: < 2.2e-16
```

ANSWER PROBLEM 8

(a) Build full additive model with only significant predictors
(b) Build interacting model with predictors from (a)
(c) Remove non-significant interactions and rerun model
(d) Interpret the final model

```
Call:
lm(formula = Balance ~ (Income + Limit + Rating + Cards + Age +
    factor(Student))^2, data = credit)

Residuals:
     Min       1Q   Median       3Q      Max
-166.579  -40.014    8.191   38.844  163.054

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -2.923e+02  4.966e+01  -5.886 8.72e-09 ***
Income                        -1.907e+00  8.011e-01  -2.381  0.01777 *
Limit                          3.230e-03  8.354e-02   0.039  0.96918
Rating                         1.446e+00  1.252e+00   1.154  0.24912
Cards                          8.495e+00  1.426e+01   0.596  0.55182
Age                            9.420e-01  7.315e-01   1.288  0.19862
factor(Student)Yes             1.909e+02  6.589e+01   2.898  0.00398 **
Income:Limit                   6.667e-04  5.931e-04   1.124  0.26168
Income:Rating                 -2.708e-02  8.703e-03  -3.112  0.00200 **
Income:Cards                  -1.755e-01  1.247e-01  -1.407  0.16021
Income:Age                     1.878e-02  8.833e-03   2.126  0.03414 *
Income:factor(Student)Yes     -1.565e+00  4.769e-01  -3.282  0.00113 **
Limit:Rating                   3.420e-04  1.751e-05  19.536  < 2e-16 ***
Limit:Cards                    3.130e-03  1.168e-02   0.268  0.78883
Limit:Age                      8.277e-04  1.281e-03   0.646  0.51860
Limit:factor(Student)Yes       2.075e-01  6.806e-02   3.048  0.00247 **
Rating:Cards                  -4.870e-03  1.734e-01  -0.028  0.97761
Rating:Age                    -1.869e-02  1.919e-02  -0.974  0.33075
Rating:factor(Student)Yes     -1.966e+00  1.019e+00  -1.929  0.05447 .
Cards:Age                      3.773e-02  1.748e-01   0.216  0.82920
Cards:factor(Student)Yes       1.073e+01  9.452e+00   1.136  0.25678
Age:factor(Student)Yes         2.499e-01  7.669e-01   0.326  0.74475
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.94 on 378 degrees of freedom
Multiple R-squared:  0.9822,    Adjusted R-squared:  0.9813
F-statistic: 995.8 on 21 and 378 DF,  p-value: < 2.2e-16
```

ANSWER PROBLEM 8

(a) Build full additive model with only significant predictors
(b) Build interacting model with predictors from (a)
(c) Remove non-significant interactions and rerun model
(d) Interpret the final model

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

Residuals:
     Min       1Q   Median       3Q      Max
-218.008  -42.145    7.003   39.734  147.616

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -2.058e+02  2.531e+01  -8.130 5.86e-15 ***
Income                        -1.684e+00  5.693e-01  -2.958 0.003291 **
Limit                          1.008e-01  2.251e-02   4.478 9.91e-06 ***
Rating                        -2.018e-01  3.334e-01  -0.605 0.545365
Cards                          1.810e+01  2.792e+00   6.482 2.76e-10 ***
Age                           -6.310e-01  3.107e-01  -2.031 0.042970 *
factor(Student)Yes             1.957e+02  4.276e+01   4.578 6.34e-06 ***
Income:Age                    -3.194e-03  5.149e-03  -0.620 0.535339
Income:Rating                 -1.687e-02  1.198e-03 -14.078  < 2e-16 ***
Income:factor(Student)Yes     -1.703e+00  4.500e-01  -3.784 0.000179 ***
Limit:Rating                   3.367e-04  1.717e-05  19.605  < 2e-16 ***
Limit:factor(Student)Yes       1.498e-01  5.999e-02   2.497 0.012955 *
Rating:factor(Student)Yes     -1.067e+00  8.910e-01  -1.198 0.231808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.6 on 387 degrees of freedom
Multiple R-squared:  0.9814,    Adjusted R-squared:  0.9809
F-statistic:  1705 on 12 and 387 DF,  p-value: < 2.2e-16
```

ANSWER PROBLEM 8

(a) Build full additive model with only significant predictors
(b) Build interacting model with predictors from (a)
(c) Remove non-significant interactions and rerun model
(d) Interpret the final model

```
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
```

keep rtin dk to hierarchical principle

ANSWER PROBLEM 8

(a) Build full additive model with only significant predictors
(b) Build interacting model with predictors from (a)
(c) Remove non-significant interactions and rerun model
(d) Interpret the final model

```
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
```

TRY THIS ON YOUR OWN

$$\hat{y} = b_0 + b_1 Income + b_2 Limit + b_3 Rating + b_4 Cards + b_5 Age + b_6 Student$$
$$+ b_7 Income * Rating + b_8 Income * Student + b_9 Limit * Rating + b_{10} Limit * Student$$

$\hat{credit}$

$$\hat{y} = -0.01945 - 1.837 Income + 0.1079 Limit - 0.3121 Rating + 10.832 Cards$$
$$-0.7660 Age + 155.5 Student - 0.01694 Income * Rating$$
$$-1.784 Income * Student + 0.0003373 Limit * Rating + 0.07868 Limit * Student$$

ANSWER PROBLEM 8

(a) Build full additive model with only significant predictors
(b) Build interacting model with predictors from (a)
(c) Remove non-significant interactions and rerun model
(d) Interpret the final model

```
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
```

TRY THIS ON YOUR OWN

$$\hat{y} = b_0 + b_1 Income + b_2 Limit + b_3 Rating + b_4 Cards + b_5 Age + b_6 Student$$
$$+ b_7 Income * Rating + b_8 Income * Student + b_9 Limit * Rating + b_{10} Limit * Student$$

$$\hat{y} = -0.01945 - 1.837 Income + 0.1079 Limit - 0.3121 Rating + 10.832 Cards$$
$$-0.7660 Age + 155.5 Student - 0.01694 Income * Rating$$
$$-1.784 Income * Student + 0.0003373 Limit * Rating + 0.07868 Limit * Student$$

ANSWER PROBLEM 8

(a) Build full additive model with only significant predictors
(b) Build interacting model with predictors from (a)
(c) Remove non-significant interactions and rerun model
(d) Interpret the final model

This is our final model → let's interpret it.

$$\hat{y} = -0.01945 - 1.837Income + 0.1079Limit - 0.3121Rating + 10.832Cards$$
$$-0.7660Age + 155.5Student - 0.01694Income * Rating$$
$$-1.784Income * Student + 0.0003373Limit * Rating + 0.07868Limit * Student$$

student = 0

WHAT IS THE EFFECT ON INCOME IF NOT A STUDENT

=0

$-1.837[INCOME] - 0.01694[INCOME][RATING] - 1.784[INCOME][STUDENT]$

Factor out θ to make easier to explain

$(-1.837 - 0.01694[RATING])[INCOME] - 0$

$(-1.837 + 0.01694[RATING])$

IS THIS AN INCREASE OR DECREASE?

→ decrease

In-class Practice Problem 9

Suppose you wanted to model the quality, y, of a product as a function of the pressure pounds per square inch (psi), at which it is produced. Four inspectors independently assign a quality score between 0 and 100 to each product, and then the quality, y, is calculated by averaging the four scores. An experiment is conducted by varying temperature in F. The data are provided in **PRODQUAL.csv** file

Fit a higher-order model to the data and sketch the scatterplot.

Which order would you select?

QUALITY ~ PRESSURE

Find the highest order model that we could use.

ANSWER PROBLEM 9

```
Call:
lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2), data = quality)

Residuals:
    Min      1Q  Median      3Q     Max
-12.136  -6.234  -2.852   7.660  16.410

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.791e+03  2.857e+02  -13.27   <2e-16 ***
PRESSURE       1.423e+02  1.039e+01   13.70   <2e-16 ***
I(PRESSURE^2) -1.307e+00  9.418e-02  -13.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.956 on 73 degrees of freedom
Multiple R-squared:  0.7622,    Adjusted R-squared:  0.7557
F-statistic:   117 on 2 and 73 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2) + I(PRESSURE^3),
    data = quality)

Residuals:
    Min      1Q  Median      3Q     Max
-12.430  -5.536  -0.779   5.710  15.170

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.083e+04  6.089e+03  -5.064 3.04e-06 ***
PRESSURE       1.623e+03  3.332e+02   4.871 6.38e-06 ***
I(PRESSURE^2) -2.827e+01  6.065e+00  -4.661 1.41e-05 ***
I(PRESSURE^3)  1.633e-01  3.672e-02   4.446 3.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.097 on 72 degrees of freedom
Multiple R-squared:  0.8134,    Adjusted R-squared:  0.8056
F-statistic: 104.6 on 3 and 72 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2) + I(PRESSURE^3) +
    I(PRESSURE^4), data = quality)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3715  -4.4458  -0.7475   3.9742  13.2232

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.958e+05  1.208e+05   4.106 0.000106 ***
PRESSURE      -3.669e+04  8.780e+03  -4.178 8.24e-05 ***
I(PRESSURE^2)  1.015e+03  2.391e+02   4.246 6.48e-05 ***
I(PRESSURE^3) -1.245e+01  2.890e+00  -4.309 5.18e-05 ***
I(PRESSURE^4)  5.710e-02  1.308e-02   4.366 4.22e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.345 on 71 degrees of freedom
Multiple R-squared:  0.8529,    Adjusted R-squared:  0.8446
F-statistic: 102.9 on 4 and 71 DF,  p-value: < 2.2e-16
```

```
ggplot(data=quality) +
  aes(x=PRESSURE, y=QUALITY) +
  geom_point(color='red') +
  geom_smooth()
```