Statistical Modelling with Data

May 23 – June 02, 2023

Instructor: Qing (Leah) Li, Ph.D. Candidate at Cumming School of Medicine

qing.li2@uclagary.ca

Thank you Dr. Thuntida Ngamkham for contributing the contents

Thank you Dr. Qingrun Zhang and Dr. Quan Long for contributing some slides

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression procedures
  - Lecture 5: Model selection: Forward and Backward selection procedures
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, and equal variance assumptions.
  - Lecture 7: Multiple regression diagnostics: verify normality assumptions and identify multicollinearity and outliers.
  - Lecture 8: Multiple regression diagnostics: data transformation
- Topic 5: Transfer learning
  - Lecture 9: Transfer-learning (Bonus): standing on the shoulders of giants.

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression procedures
  - Lecture 5: Model selection: Forward and Backward selection procedures
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, and equal variance assumptions.
  - Lecture 7: Multiple regression diagnostics: verify normality assumptions and identify multicollinearity and outliers.
  - Lecture 8: Multiple regression diagnostics: data transformation
- Topic 5: Transfer learning
  - Lecture 9: Transfer-learning (Bonus): standing on the shoulders of giants.

Statistical Modelling with Data

Learning Outcomes: At the end of the course, participants will be able to

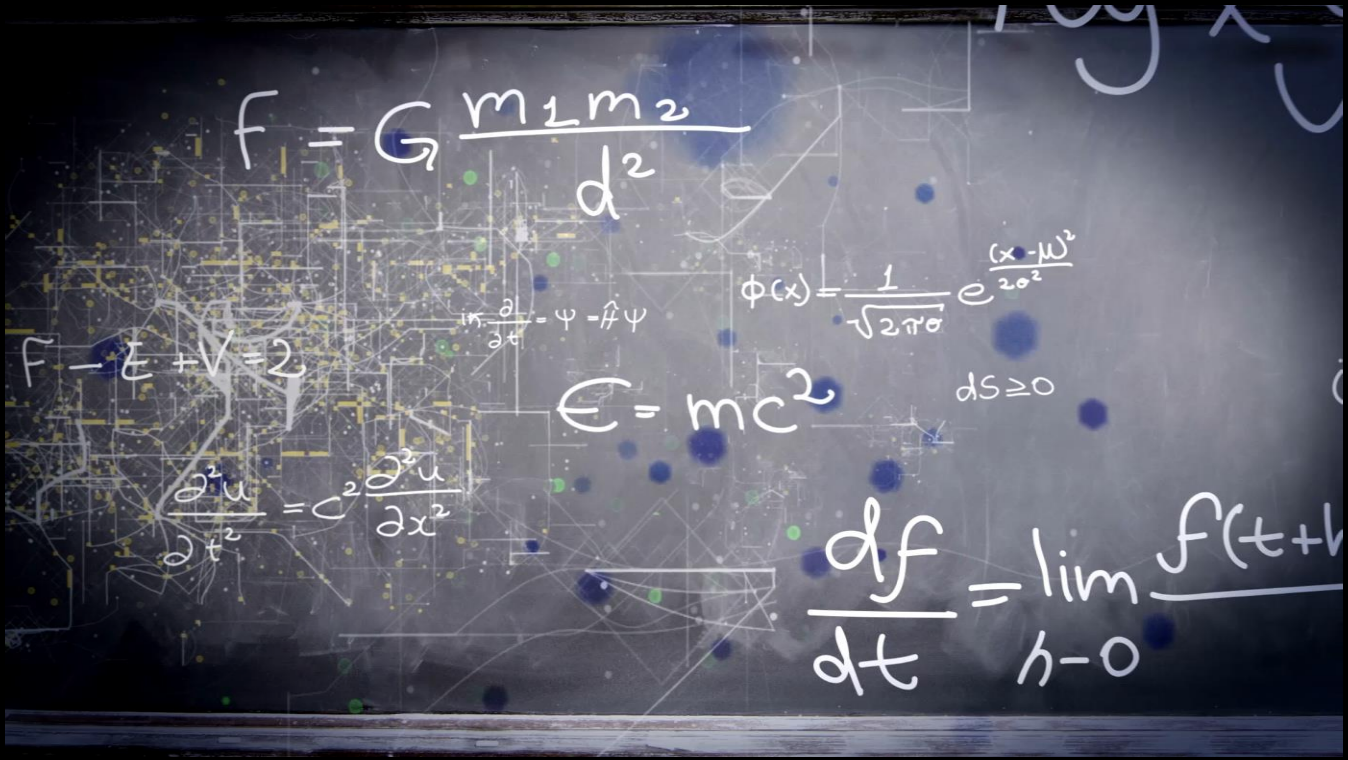
1. Model the multiple linear relationships between a response variable (Y) and all explanatory variables (both categorical and numerical variables) with interaction terms. Interpret model parameter estimates, construct confidence intervals for regression coefficients, evaluate model fits, and visualize correlations between a response variable (Y) and all explanatory variables (X) by graphs (scatter plot, residual plot) to assess model validity.
2. Predict the response variable at a certain level of the explanatory variables once the fit model exists.
3. Implement R-software and analyze statistical results for biomedical and other data.

- **Evaluations**
  1. Assignments will be posted on Slack (our communication tool with students).
  2. Students must attend 70% (6/9) of the sessions in order to receive the certificate and are encouraged to work on the assignments progressively throughout the course as the relevant material is covered.

Statistical Modelling with Data

- Supportive materials
  - Lectures slides (2023)
  - R code scripts (2023)
  - PDF (dated 2022)
  - Two Assignments (dated 2022)
- Slack channels
  - Recoding videos
  - Exercises
  - Course-documents

Lecture 3: Multivariate linear regression Interaction effect and A Quadratic (Second-Order) Model with Quantitative Predictors



Quick recap of lecture 2

- Statistics:
  - Interactions: x1:x2 or (x1+x2)^2 or x1*x2
  - Dummy coding: the number of dummy variable = the number of levels -1
  - Interpretation of coefficients
- Code:
  - lm(y ~ x1+x2+(x1+x2)^2)
  - lm(y ~ factor(x1))

Gender	X1
Male	1
Female	0

	x1	x2
Assistant	0	0
Associate	1	0
Full	0	1

Interaction Effect in Multiple Regression with both Quantitative and Qualitative (Dummy) Variable models

In previous topics, we considered Multiple Regression models for both quantitative and qualitative variables. We also discussed an interaction in Multiple Regression for quantitative variables. However, the concept of interactions applies just as well to qualitative variables, or to a combination of quantitative and qualitative variables. In fact, an interaction between a qualitative variable and a quantitative variable has a particularly nice interpretation.

Consider the Credit data set example and suppose that we wish to predict balance using the income (quantitative) and student (qualitative) variables. In the absence of an interaction term, the model takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon$$

$$balance_i = \beta_0 + \beta_1 Income_i + \begin{cases} \beta_2 & \text{if } i^{th}\text{person is a student} \\ 0 & \text{if } i^{th}\text{person is not a student} \end{cases}$$

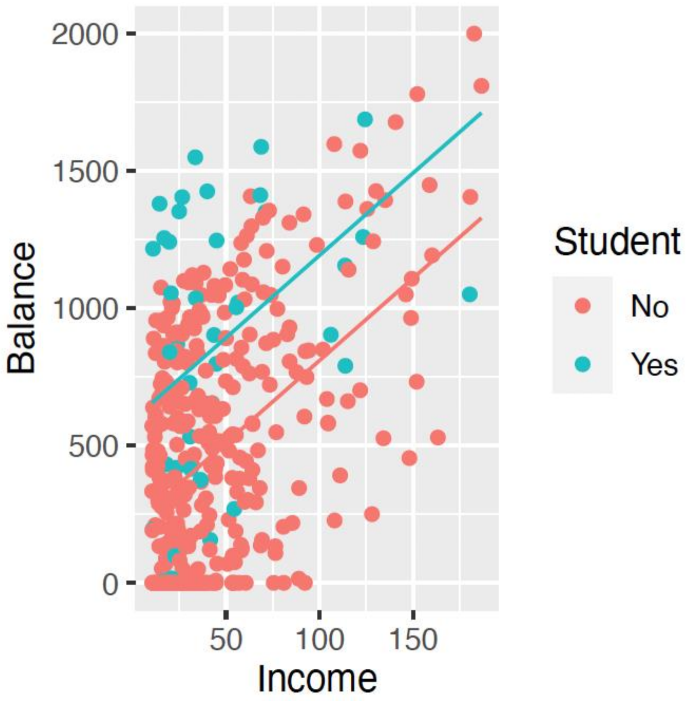
$$balance_i = \beta_1 Income_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i^{th}\text{person is a student} \\ \beta_0 & \text{if } i^{th}\text{person is not a student} \end{cases}$$

Interaction Effect in Multiple Regression with both Quantitative and Qualitative (Dummy) Variable models

```
> credit=read.csv("credit.csv",header = TRUE)
> mixmodel<-lm(Balance~Income+factor(Student), data=credit)
> summary(mixmodel)

Call:
lm(formula = Balance ~ Income + factor(Student), data = credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-762.37 -331.38  -45.04  323.60  818.28 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        211.1430    32.4572   6.505 2.34e-10 ***
Income               5.9843     0.5566  10.751  < 2e-16 ***
factor(Student)Yes 382.6705    65.3108   5.859 9.78e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 391.8 on 397 degrees of freedom
Multiple R-squared:  0.2775,	Adjusted R-squared:  0.2738 
F-statistic: 76.22 on 2 and 397 DF,  p-value: < 2.2e-16
```

$$balance_i = 5.9843 Income_i + \begin{cases} 211.1430 + 382.6705 = 593.8135 & \text{if } i^{th} \text{person is a student} \\ 211.1430 & \text{if } i^{th} \text{person is not a student} \end{cases}$$

$$balance_i = \begin{cases} 593.8135 + 5.9843 Income_i & \text{if } i^{th} \text{ person is a student} \\ 211.1430 + 5.9843 Income_i & \text{if } i^{th} \text{ person is not a student} \end{cases}$$



The fact that the lines are parallel means that the average effect on balance of a one-unit increase in income does not depend on whether or not the individual is a student.
This represents a potentially serious limitation of the model, since in fact a change in income may have a very different effect on the credit card balance.

Interaction Effect in Multiple Regression with both Quantitative and Qualitative (Dummy) Variable models

```
> credit=read.csv("credit.csv",header = TRUE)
> mixmodel<- lm(Balance~Income+factor(Student)+Income*factor(Student),data=credit)
> summary(mixmodel)

Call:
lm(formula = Balance ~ Income + factor(Student) + Income * factor(Student),
    data = credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-773.39 -325.70  -41.13  321.65  814.04 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               200.6232    33.6984   5.953 5.79e-09 ***
Income                      6.2182     0.5921  10.502  < 2e-16 ***
factor(Student)Yes        476.6758   104.3512   4.568 6.59e-06 ***
Income:factor(Student)Yes  -1.9992     1.7313  -1.155    0.249    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 391.6 on 396 degrees of freedom
Multiple R-squared:  0.2799,    Adjusted R-squared:  0.2744 
F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

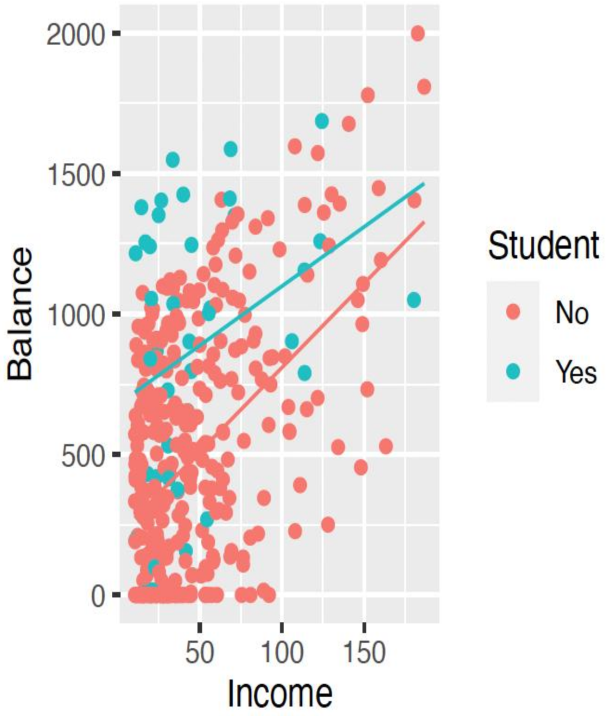
Interaction Effect in Multiple Regression with both Quantitative and Qualitative (Dummy) Variable models

$$Y_i = 200.6232 + 6.2182X_{i1} + 476.6758X_{i2} - 1.9992X_{i1}X_{i2} + \epsilon$$

$$balance_i = 200.6232 + 6.2182 Income_i + \begin{cases} 476.6758 - 1.9992 Income_i & \text{if } i^{th}\text{person is a student} \\ 0 & \text{if } i^{th}\text{person is not a student} \end{cases}$$

$$\widehat{balance}_i = \begin{cases} (200.6232 + 476.67582) + (6.2182 - 1.9992) Income_i & \text{if } i^{th} \text{ person is a student} \\ 200.6232 + 6.2182 Income_i & \text{if } i^{th} \text{ person is not a student} \end{cases}$$

$$\widehat{balance}_i = \begin{cases} 677.29902 + 4.219 Income_i & \text{if } i^{th} \text{ person is a student} \\ 200.6232 + 6.2182 Income_i & \text{if } i^{th} \text{ person is not a student} \end{cases}$$



Disregard the p-value for the interaction term, we have two different regression lines for the students and the non-students. But now those regression lines have different intercepts, $\beta_0 + \beta_2$ versus β_1, as well as different slopes, $\beta_1 + \beta_3$ versus β_1. This allows for the possibility that changes in income may affect the credit card balances of students and non-students differently. The output shows the estimated relationships between income and balance for students and non-students in the model. We note that the slope for students (4.219) is lower than the slope for non-students (6.218). This suggests that increases in income are associated with smaller increases in credit card balance among students as compared to non-students.

One Health at UCalgary

UNIVERSITY OF CALGARY

In class Practice Problem 8

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

1. Build a full additive model with only significant predictors
2. Build interaction model with predictors from 1
3. Remove non-significant interactions and rerun the model
4. Interpret the final model

Hints:

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

Lecture_3.R

In class Practice Problem 8

1. Build a full additive model with only significant predictors

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

Lecture_3.R

In class Practice Problem 8

2. Build interacting model with predictors from 1

```
> model2 = lm(formula = Balance ~ (Income+Limit+Rating+Cards+Age+factor(Student))^2, data=credit)
> summary(model2)

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
```

Lecture_3.R

In class Practice Problem 8

3. Remove non-significant interactions and rerun the model

```
> model3=lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+           +Income*Age+Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student),data=credit)
> summary(model3)

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

Residuals:
     Min       1Q   Median       3Q      Max
-216.057  -40.976    7.601   39.380  152.057

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -2.035e+02  2.525e+01  -8.058 9.64e-15 ***
Income                    -1.683e+00  5.696e-01  -2.955 0.003316 **
Limit                      1.084e-01  2.161e-02   5.017 8.00e-07 ***
Rating                    -3.136e-01  3.202e-01  -0.980 0.327918
Cards                      1.822e+01  2.792e+00   6.525 2.13e-10 ***
Age                       -5.975e-01  3.096e-01  -1.930 0.054395 .
factor(Student)Yes         1.554e+02  2.636e+01   5.896 8.13e-09 ***
Income:Age                -3.532e-03  5.144e-03  -0.687 0.492784
Income:Rating             -1.683e-02  1.199e-03 -14.041  < 2e-16 ***
Income:factor(Student)Yes -1.759e+00  4.478e-01  -3.928 0.000101 ***
Limit:Rating               3.363e-04  1.718e-05  19.575  < 2e-16 ***
Limit:factor(Student)Yes   7.852e-02  7.675e-03  10.230  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.64 on 388 degrees of freedom
Multiple R-squared:  0.9814,    Adjusted R-squared:  0.9808
F-statistic:  1858 on 11 and 388 DF,  p-value: < 2.2e-16
```

Lecture_3.R

In class Practice Problem 8

3. Remove non-significant interactions and rerun the model

```
> model4=lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+           +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student),data=credit)
> summary(model4)

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
```

Lecture_3.R

In class Practice Problem 8

4. Interpret the final model

$$\hat{y} = -0.01945 - 1.837Income + 0.1079Limit - 0.3121Rating + 10.832Cards - 0.766Age + 155.5Student - 0.01694Income \times Rating - 1.784Income \times Student + 0.0003373Limit \times Rating + 0.07868Limit \times Student$$

What is the effect of income on final credit balance if not a student?

$$-1.837Income + 155.5 \times 0 - 0.01694Income \times Rating - 1.784Income \times 0$$
$$= -1.837Income - 0.01694Income \times Rating$$
$$= -(1.837 + 0.01694Rating) \times Income$$

If not a student, with income increase, credit balance decreases. The person is likely to spend more.

What is the effect of income on final credit balance if a student?

$$-1.837Income + 155.5 \times 1 - 0.01694Income \times Rating - 1.784Income \times 1$$
$$= -(1.837 + 0.01694Rating - 1.784) \times Income + 155.5$$

If a student, with income increase, credit balance likely to increase decreases. The student is likely to spend less.

Lecture_3.R

In class Practice Problem 8

Dr. Thuntida Ngamkham's approach

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

Leah's approach:

1. Start with an interaction model with all predictors
2. Remove non-significant interactions
3. Rerun model to ensure all predictors are significant
4. Iterate step 3 until done.

Lecture_3.R

One Health at UCalgary

UNIVERSITY OF CALGARY

Practice Problem 8

```
> ###Leah's approach
> #Step1: Build an interaction model with all predictors
> model_inter = lm(formula = Balance ~ (Income+Limit+Rating+Cards+Age+Education
+              +factor(Gender)+factor(Ethnicity)+factor(Married)+factor(Student))^2, data=credit)
> summary(model_inter)

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

Residuals:
     Min       1Q   Median       3Q      Max
-155.561  -40.024    4.531   40.274  149.757

Coefficients:
                                      Estimate Std. E
(Intercept)                         -3.174e+02  1.068
Income                              -1.697e-01  1.264
Limit                                1.837e-02  1.345
Rating                               9.731e-01  2.005
Cards                                1.049e+01  2.110
Age                                  1.356e+00  1.240
Education                            1.883e+00  6.207
factor(Gender)Female                -5.120e+01  4.851
factor(Ethnicity)Asian               8.756e+01  6.985
factor(Ethnicity)Caucasian           3.215e+01  6.187
factor(Married)Yes                   4.418e+01  5.085
factor(Student)Yes                   1.854e+02  1.076
Income:Limit                         9.216e-04  6.515
Income:Rating                       -3.099e-02  9.560
Income:Cards                        -1.726e-01  1.352
Income:Age                           2.667e-02  9.653
Income:Education                    -1.583e-01  5.265
Income:factor(Gender)Female         -9.414e-01  3.400
Income:factor(Ethnicity)Asian        6.959e-01  4.744
Income:factor(Ethnicity)Caucasian    9.808e-01  4.149
Income:factor(Married)Yes           -1.746e-01  3.454
Income:factor(Student)Yes           -1.850e+00  5.514
Limit:Rating                         3.510e-04  1.850
```

```
Limit:Cards                             6.023e-04  1.257e-02   0.048 0.961810
Limit:Age                               1.318e-03  1.420e-03   0.928 0.353969
Limit:Education                        -6.330e-03  7.710e-03  -0.821 0.412184
Limit:factor(Gender)Female              5.305e-02  4.479e-02   1.184 0.237078
Limit:factor(Ethnicity)Asian            2.658e-02  6.557e-02   0.405 0.685461
Limit:factor(Ethnicity)Caucasian        4.695e-02  5.355e-02   0.877 0.381195
Limit:factor(Married)Yes               -3.005e-02  4.633e-02  -0.649 0.517037
Limit:factor(Student)Yes                2.103e-01  7.703e-02   2.730 0.006674 **
Rating:Cards                            2.416e-02  1.861e-01   0.130 0.896741
Rating:Age                             -2.765e-02  2.127e-02  -1.300 0.194501
Rating:Education                        1.171e-01  1.168e-01   1.002 0.316978
Rating:factor(Gender)Female            -5.896e-01  6.719e-01  -0.877 0.380847
Rating:factor(Ethnicity)Asian          -5.097e-01  9.797e-01  -0.520 0.603237
Rating:factor(Ethnicity)Caucasian      -8.542e-01  8.015e-01  -1.066 0.287310
Rating:factor(Married)Yes               4.231e-01  6.940e-01   0.610 0.542530
Rating:factor(Student)Yes              -1.927e+00  1.152e+00  -1.673 0.095275 .
Cards:Age                               1.322e-01  1.865e-01   0.709 0.478916
Cards:Education                        -1.133e+00  9.315e-01  -1.216 0.224880
Cards:factor(Gender)Female              1.201e+01  6.014e+00   1.997 0.046621 *
Cards:factor(Ethnicity)Asian            2.302e-01  9.303e+00   0.025 0.980273
Cards:factor(Ethnicity)Caucasian        7.929e+00  7.862e+00   1.008 0.313952
Cards:factor(Married)Yes               -1.953e+00  6.520e+00  -0.300 0.764718
Cards:factor(Student)Yes                1.024e+01  1.045e+01   0.980 0.327718
Age:Education                          -5.001e-02  6.500e-02  -0.769 0.442215
Age:factor(Gender)Female                5.275e-01  4.047e-01   1.304 0.193292
Age:factor(Ethnicity)Asian              2.694e-01  5.584e-01   0.483 0.629753
Age:factor(Ethnicity)Caucasian         -2.452e-01  4.662e-01  -0.526 0.599256
Age:factor(Married)Yes                  2.979e-01  4.171e-01   0.714 0.475527
Age:factor(Student)Yes                  2.481e-01  8.618e-01   0.288 0.773596
Education:factor(Gender)Female         -1.212e+00  2.181e+00  -0.556 0.578771
Education:factor(Ethnicity)Asian       -4.307e+00  3.146e+00  -1.369 0.171942
Education:factor(Ethnicity)Caucasian    1.612e+00  2.721e+00   0.592 0.554028
Educati
Educati
factor(
factor(
factor(
factor(
factor(
```

```
factor(Gender)Female:factor(Married)Yes          -3.218e+00  1.441e+01  -0.223 0.823423
factor(Gender)Female:factor(Student)Yes          -2.927e+00  2.515e+01  -0.116 0.907401
factor(Ethnicity)Asian:factor(Married)Yes        -8.929e+00  1.979e+01  -0.451 0.652194
factor(Ethnicity)Caucasian:factor(Married)Yes    -2.771e+01  1.666e+01  -1.664 0.097124
factor(Ethnicity)Asian:factor(Student)Yes         1.045e+01  3.091e+01   0.338 0.735434
factor(Ethnicity)Caucasian:factor(Student)Yes     7.257e-01  2.931e+01   0.025 0.980263
factor(Married)Yes:factor(Student)Yes            -1.371e+00  2.383e+01  -0.058 0.954143
```

Lecture_3.R

In class Practice Problem 8

```
> #Step2: Select significant predictors, remove non-significant predictors
> coefficeints_model_inter = data.frame(summary(model_inter)[4])
> sig_coefficeints_model_inter = coefficeints_model_inter[coefficeints_model_inter$coefficients.Pr...t.. < 0.05,]
> sig_coefficeints_model_inter
                                  coefficients.Estimate coefficients.Std..Error coefficients.t.value coefficients.Pr...t..
(Intercept)                               -3.173700e+02           1.068183e+02            -2.971119          3.182266e-03
Income:Rating                             -3.099275e-02           9.560304e-03            -3.241816          1.307449e-03
Income:Age                                 2.667199e-02           9.652560e-03             2.763204          6.040905e-03
Income:Education                          -1.583361e-01           5.265257e-02            -3.007187          2.836756e-03
Income:factor(Gender)Female               -9.413676e-01           3.399607e-01            -2.769049          5.936021e-03
Income:factor(Ethnicity)Caucasian          9.808143e-01           4.148621e-01             2.364194          1.864107e-02
Income:factor(Student)Yes                 -1.849814e+00           5.514055e-01            -3.354725          8.858117e-04
Limit:Rating                               3.509598e-04           1.850477e-05            18.965906          5.896138e-55
Limit:factor(Student)Yes                   2.102837e-01           7.703386e-02             2.729757          6.674036e-03
Cards:factor(Gender)Female                 1.201146e+01           6.014331e+00             1.997140          4.662094e-02
```

```
> #Step3: Build refined model 1 with significant interaction predictors
> model_inter_refine1 = lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+Education
+                                             +factor(Gender)+factor(Ethnicity)+factor(Student)+
+                              Income:Rating+ Income:Age+ Income:Education+ Income:factor(Gender)+
+                              Income:factor(Ethnicity)+Income:factor(Student)+Limit:Rating+
+                              Limit:factor(Student)+Cards:factor(Gender), data=credit)
> summary(model_inter_refine1)
```

Lecture_3.R

In class Practice Problem 8

```
> #Step 4: Select significant interaction predictors for refined model 1
> coefficeints_model_inter_refine1=data.frame(summary(model_inter_refine1)[4])
> si_coefficients_model_inter_refine1 = coefficeints_model_inter_refine1[coefficeints_model_inter_refine1$coefficients.Pr...t.. < 0.05,]
> si_coefficients_model_inter_refine1
                         coefficients.Estimate coefficients.Std..Error coefficients.t.value coefficients.Pr...t..
(Intercept)                      -1.901534e+02            3.618799e+01            -5.254601          2.481193e-07
Limit                             1.118835e-01            2.156661e-02             5.187811          3.473609e-07
Cards                             1.463119e+01            3.560129e+00             4.109736          4.858097e-05
Age                              -6.331507e-01            3.142269e-01            -2.014947          4.461588e-02
factor(Student)Yes                1.467450e+02            2.646051e+01             5.545811          5.497256e-08
Income:Rating                    -1.676704e-02            1.218299e-03           -13.762663          3.106820e-35
Income:Education                 -6.434522e-02            2.996033e-02            -2.147680          3.237272e-02
Income:factor(Student)Yes        -1.840636e+00            4.593894e-01            -4.006701          7.413385e-05
Limit:Rating                      3.405165e-04            1.751980e-05            19.436102          7.102362e-59
Limit:factor(Student)Yes          8.168285e-02            7.723380e-03            10.576049          4.431681e-23
```

```
> #Step 5: Build refined model 2 with significant interaction predictors from model 1
> model_inter_refine2 = lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)+Education+
+                               Income:Rating+Income:Education+Income:factor(Student)+Limit:Rating+
+                               Limit:factor(Student), data=credit)
> summary(model_inter_refine2)
```

Lecture_3.R

In class Practice Problem 8

```
> #Step 6: Select significant interaction predictors for refined model 2
> coefficeints_model_inter_refine2=data.frame(summary(model_inter_refine2)[4])
> si_coefficients_model_inter_refine2 = coefficeints_model_inter_refine2[coefficeints_model_inter_refine2$coefficients.Pr...t.. < 0.05,]
> si_coefficients_model_inter_refine2
                         coefficients.Estimate coefficients.Std..Error coefficients.t.value coefficients.Pr...t..
(Intercept)                      -1.900214e+02            3.103170e+01            -6.123463          2.250449e-09
Limit                             1.107595e-01            2.150280e-02             5.150934          4.137290e-07
Cards                             1.850824e+01            2.769060e+00             6.683941          8.135102e-11
Age                              -7.571316e-01            1.872836e-01            -4.042700          6.377404e-05
factor(Student)Yes                1.519892e+02            2.620767e+01             5.799418          1.383826e-08
Income:Rating                    -1.713310e-02            1.191675e-03           -14.377328          7.423761e-38
Income:factor(Student)Yes        -1.861321e+00            4.441825e-01            -4.190441          3.452314e-05
Limit:Rating                      3.421202e-04            1.726112e-05            19.820276          7.486054e-61
Limit:factor(Student)Yes          8.070936e-02            7.656363e-03            10.541475          5.215057e-23
```

```
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
```

```
> #Step 7: Build refined model 3 with significant interaction predictors from model 2
> model_inter_refine3 = lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)+
+                              Income:Rating+Income:factor(Student)+Limit:Rating+
+                              Limit:factor(Student), data=credit)
> summary(model_inter_refine3)
```

Lecture_3.R

In class Practice Problem 8

Dr. Thuntida Ngamkham's final model

```
> model4=lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+           +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student),data=credit)
> summary(model4)

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
```

Leah's final model

```
> #Step 7: Build refined model 3 with significant interaction predictors from model 2
> model_inter_refine3 = lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)+
+                          Income:Rating+Income:factor(Student)+Limit:Rating+
+                          Limit:factor(Student), data=credit)
> summary(model_inter_refine3)

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
```

Lecture_3.R

In class Practice Problem 8

Dr. Thuntida Ngamkham's approach

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

Leah's approach:

1. Start with an interaction model with all predictors
2. Remove non-significant interactions
3. Rerun model to ensure all predictors are significant
4. Iterate step 3 until done.

Pros:	Cons
• Involve a small number of interaction predictors to keep model simple.	• Risk of missing some interaction predictors.

Either way works, and they lead to the same results for problem 8!

Pros:	Cons
• No risk of missing any interaction predictors.	• Got a really long list of coefficients. Not very eyes-friendly.

Lecture_3.R

A Quadratic (Second Order) Model with Quantitative predictors

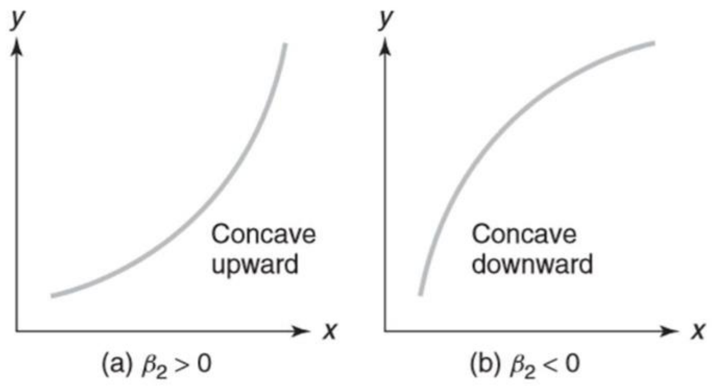
All of the models discussed in the previous sections proposed straight-line relationships between *E*(*y*) and each of the independent variables in the model. In this slide, we consider a model that allows for curvature in the relationship. This model is a second-order model because it will include an X^2 term. Here, we consider a model that includes only one independent variable X_1. The form of this model, called the *quadratic model*, is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \in$$

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_1^2$$

How to interpret the regression coefficients?
How to let R know we are putting higher order term?

Interpretation of the regression coefficients



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \in$$

Differentiate with respect to X_1

$$Y' = \beta_1 + \beta_2 X_1$$

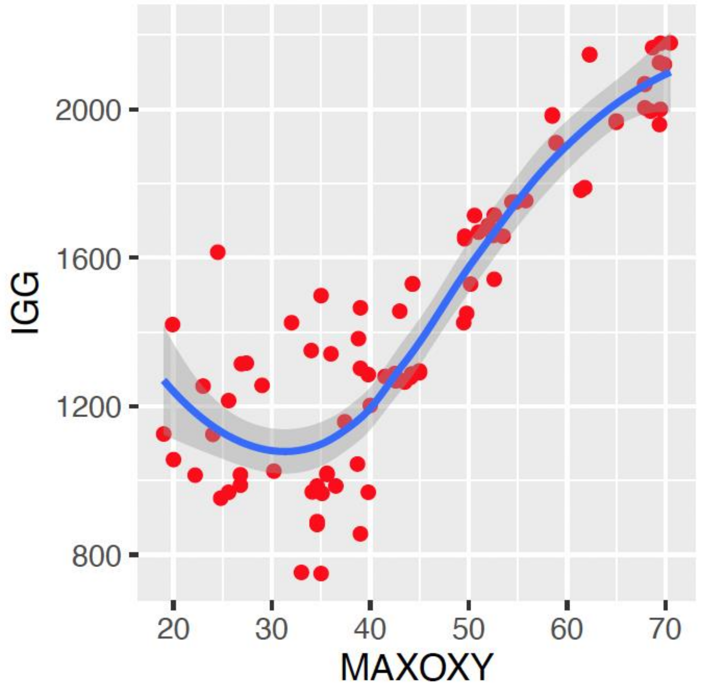
$\widehat{\beta_0}$ can be meaningfully interpreted only if the range of the independent variable includes zero-that is, if X_1 = 0 is included in the sampled range of X_1.

$\widehat{\beta_1}$ no longer represents a slope in the presence of the quadratic term X_1^2. The estimated coefficient of the first-order term X_1 will not, in general, have a meaningful interpretation in the quadratic model.

The sign of the coefficients, $\widehat{\beta_2}$ is the indicator of whether the curve is concave downward (mound-shaped) or concave upward (bowl-shaped). A negative $\widehat{\beta_2}$ implies downward concavity, as in this example, and a positive $\widehat{\beta_2}$ implies upward concavity.

Example: obviously nonlinear

Example A physiologist wants to investigate the impact of exercise on the human immune system. The physiologist theorizes that the amount of immunoglobulin Y in blood (called IgG, an indicator of long-term immunity, milligrams) is related to the maximal oxygen uptake x (a measure of aerobic fittness level, milliliters per kilogram). The data file is provided in **AEROBIC.CSV** file. Construct a scatterplot for the data. Is there evidence to support the use of a quadratic model? What is the best model to fit the data.

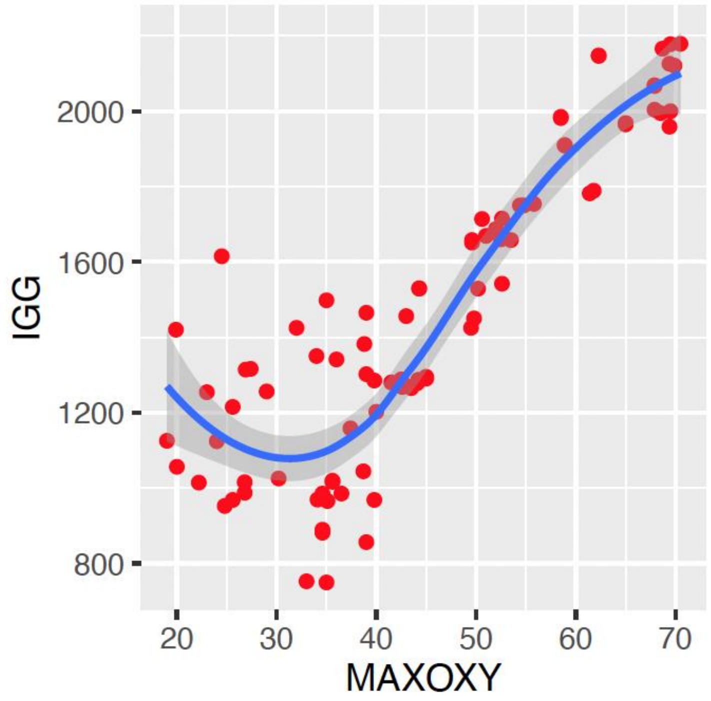


Higher-order models

```
quadmodel=lm(IGG~MAXOXY+I(MAXOXY^2),data=aerobicdata)
summary(quadmodel)
```

```
cubemodel=lm(IGG~MAXOXY+I(MAXOXY^2)+I(MAXOXY^3),data=aerobicdata)
summary(cubemodel)
```

```
forthmodel=lm(IGG~MAXOXY+I(MAXOXY^2)+I(MAXOXY^3)+I(MAXOXY^4),data=aerobicdata)
summary(forthmodel)# should stop at cubemodel because all variables are not significant.
```



Higher-order models

```
> simplemodel=lm(IGG~MAXOXY,data=aerobicdata)
> summary(simplemodel)

Call:
lm(formula = IGG ~ MAXOXY, data = aerobicdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-478.11 -127.30   28.04  116.38  636.34 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  398.954     69.561   5.735 1.38e-07 ***
MAXOXY        23.662      1.468  16.120  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 201.4 on 87 degrees of freedom
Multiple R-squared:  0.7492,	Adjusted R-squared:  0.7463 
F-statistic: 259.8 on 1 and 87 DF,  p-value: < 2.2e-16
```

```
> quadmodel=lm(IGG~MAXOXY+I(MAXOXY^2),data=aerobicdata)
> summary(quadmodel)

Call:
lm(formula = IGG ~ MAXOXY + I(MAXOXY^2), data = aerobicdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-439.91  -86.43  -30.15  139.15  517.61 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1270.41137  186.19900   6.823 1.18e-09 ***
MAXOXY       -18.10744    8.52049  -2.125   0.0364 *  
I(MAXOXY^2)    0.45082    0.09088   4.960 3.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 178.6 on 86 degrees of freedom
Multiple R-squared:  0.805,	Adjusted R-squared:  0.8004 
F-statistic: 177.5 on 2 and 86 DF,  p-value: < 2.2e-16
```

I(X^2) :add quadratic term to the model

```
> cubemodel=lm(IGG~MAXOXY+I(MAXOXY^2)+I(MAXOXY^3),data=aerobicdata)
> summary(cubemodel)

Call:
lm(formula = IGG ~ MAXOXY + I(MAXOXY^2) + I(MAXOXY^3), data = aerobicdata)

Residuals:
   Min     1Q Median     3Q    Max 
-356.7 -100.1  -12.5  103.6  496.1 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.502e+03  5.015e+02   6.982 6.03e-10 ***
MAXOXY      -1.902e+02  3.727e+01  -5.103 2.01e-06 ***
I(MAXOXY^2)  4.527e+00  8.680e-01   5.216 1.27e-06 ***
I(MAXOXY^3) -2.999e-02  6.357e-03  -4.717 9.29e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 159.9 on 85 degrees of freedom
Multiple R-squared:  0.8454,	Adjusted R-squared:   0.84 
F-statistic:   155 on 3 and 85 DF,  p-value: < 2.2e-16
```

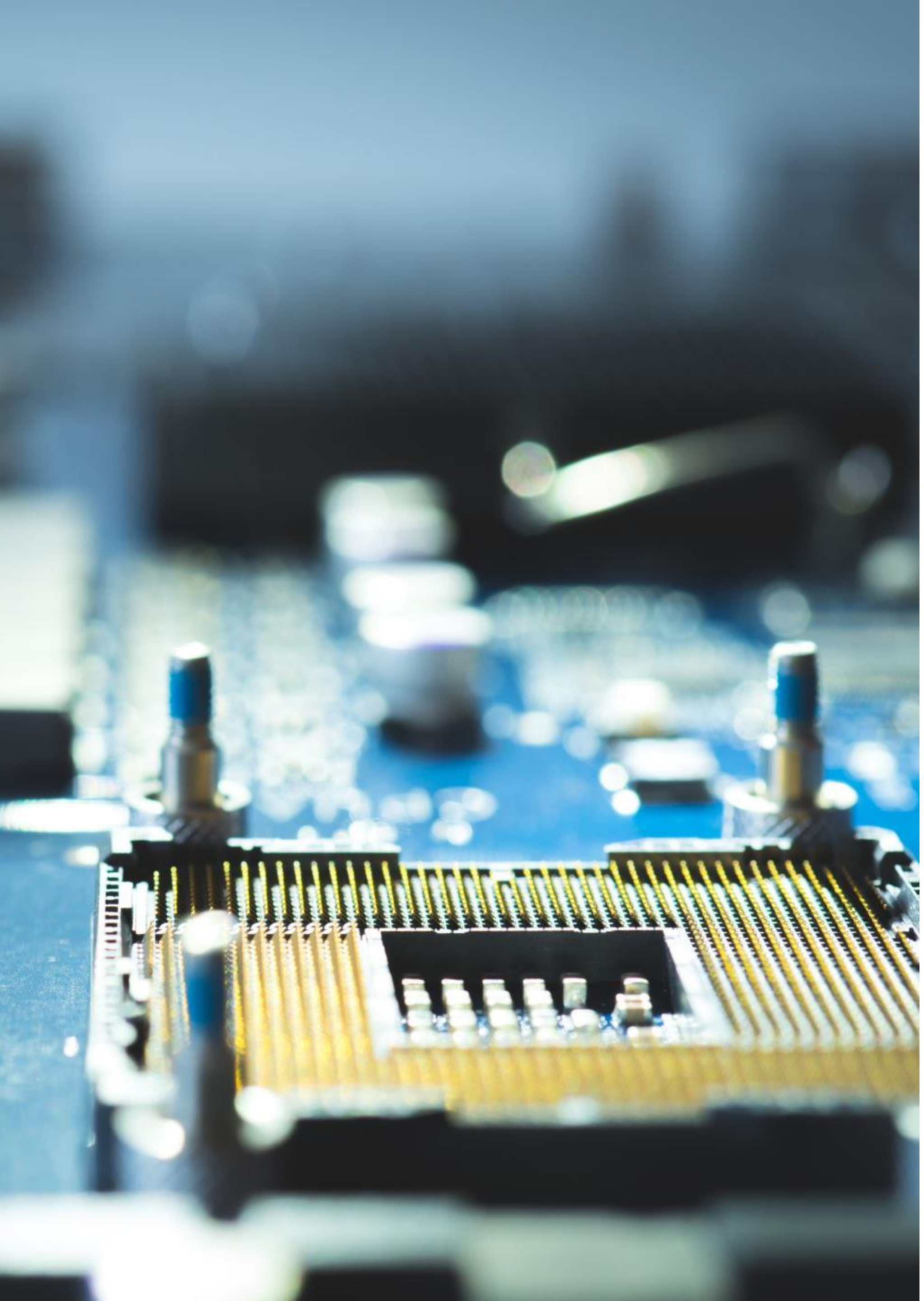
```
> forthmodel=lm(IGG~MAXOXY+I(MAXOXY^2)+I(MAXOXY^3)+I(MAXOXY^4),data=aerobicdata)
> summary(forthmodel)# should stop at cubemodel because all variables are not significant.

Call:
lm(formula = IGG ~ MAXOXY + I(MAXOXY^2) + I(MAXOXY^3) + I(MAXOXY^4), 
    data = aerobicdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-362.89 -104.07   -8.92   98.60  481.75 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.012e+03  1.596e+03   1.261    0.211
MAXOXY      -3.370e+01  1.635e+02  -0.206    0.837
I(MAXOXY^2) -1.255e+00  5.947e+00  -0.211    0.833
I(MAXOXY^3)  5.979e-02  9.156e-02   0.653    0.516
I(MAXOXY^4) -4.976e-04  5.063e-04  -0.983    0.328

Residual standard error: 160 on 84 degrees of freedom
Multiple R-squared:  0.8472,	Adjusted R-squared:  0.8399 
F-statistic: 116.4 on 4 and 84 DF,  p-value: < 2.2e-16
```

Higher-order models

- From the output, considering the scatterplot between Y and $X1$, we found that the best model to fit the data is

$$\hat{Y} = 3502 \; - 190.2X_1 + 4.527X_1^2 - 299.9X_1^3$$

- Moreover, R2 adj = 0.84 and RMSE=159.9,with the lowest RMSE and highest R2 adj among four models. We can conclude that the cube model fits the data better than the simple linear regression model.
- Note! Model interpretations are not meaningful outside the range of the independent variable. Although the model appears to support the data. To make a prediction for Y , value of X should be inside the range of the independent variable. Otherwise, the prediction will not be meaningful.



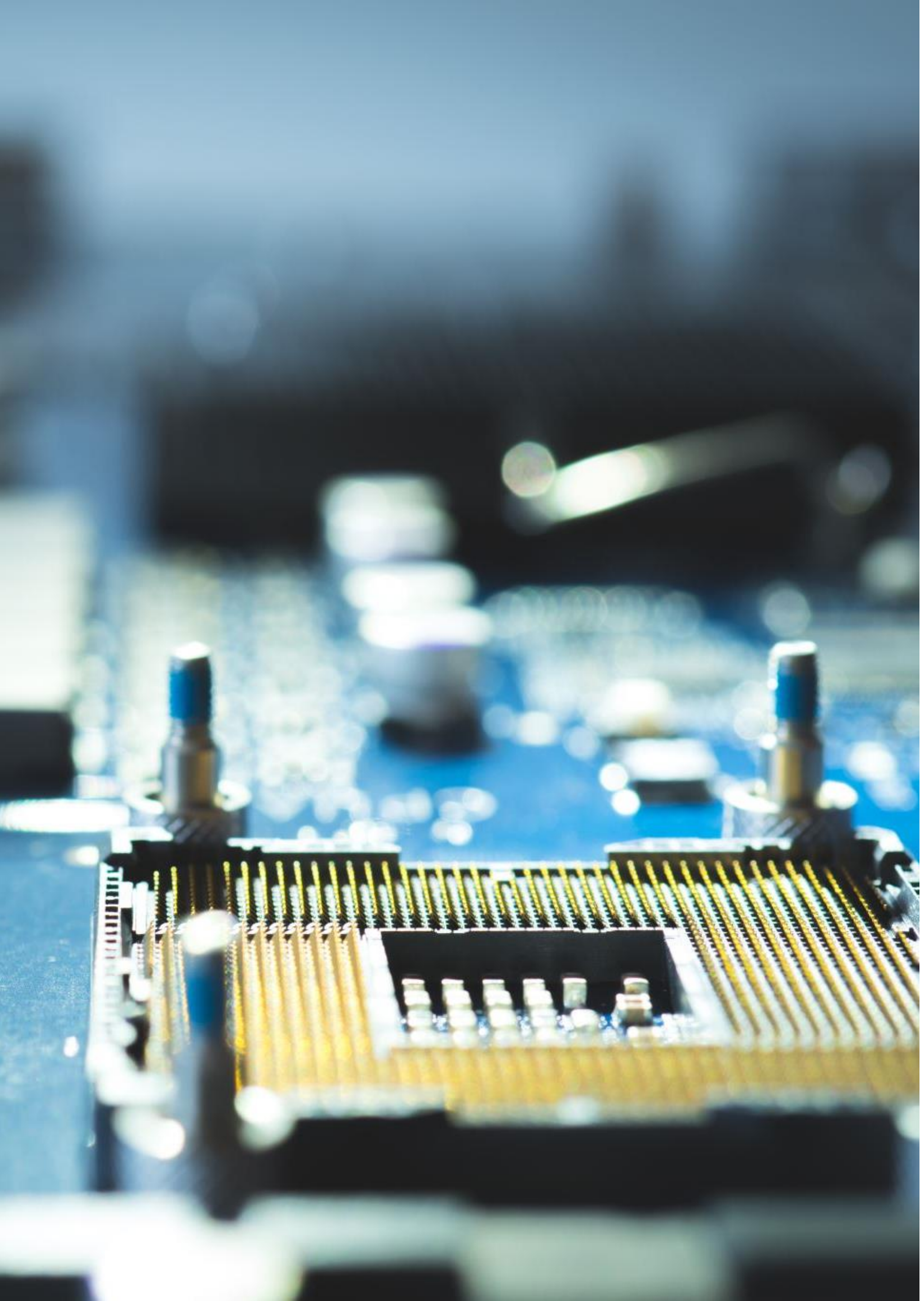
In class Practice Problem 9

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.

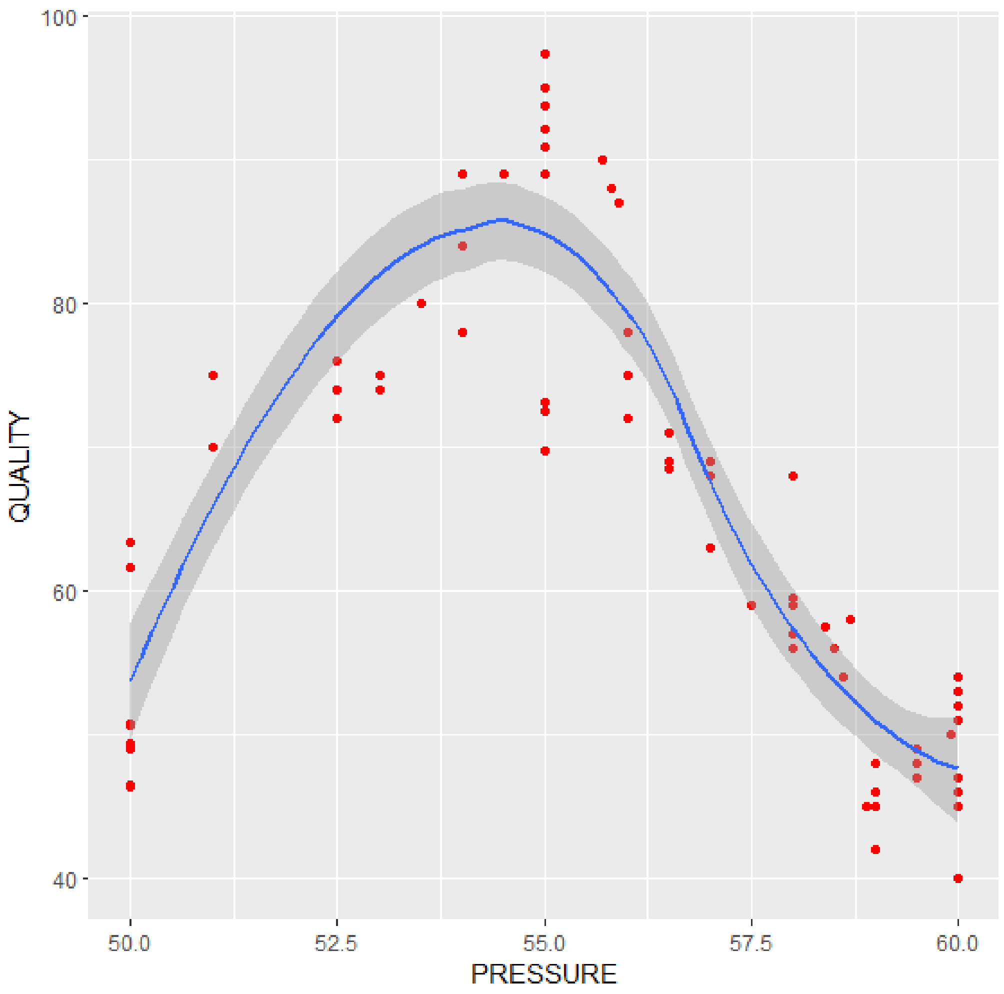
Fit a second-order model to the data and sketch the scatterplot. The data are provided in PRODQUAL.csv file

Which order would you select?



In class Practice Problem 9

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



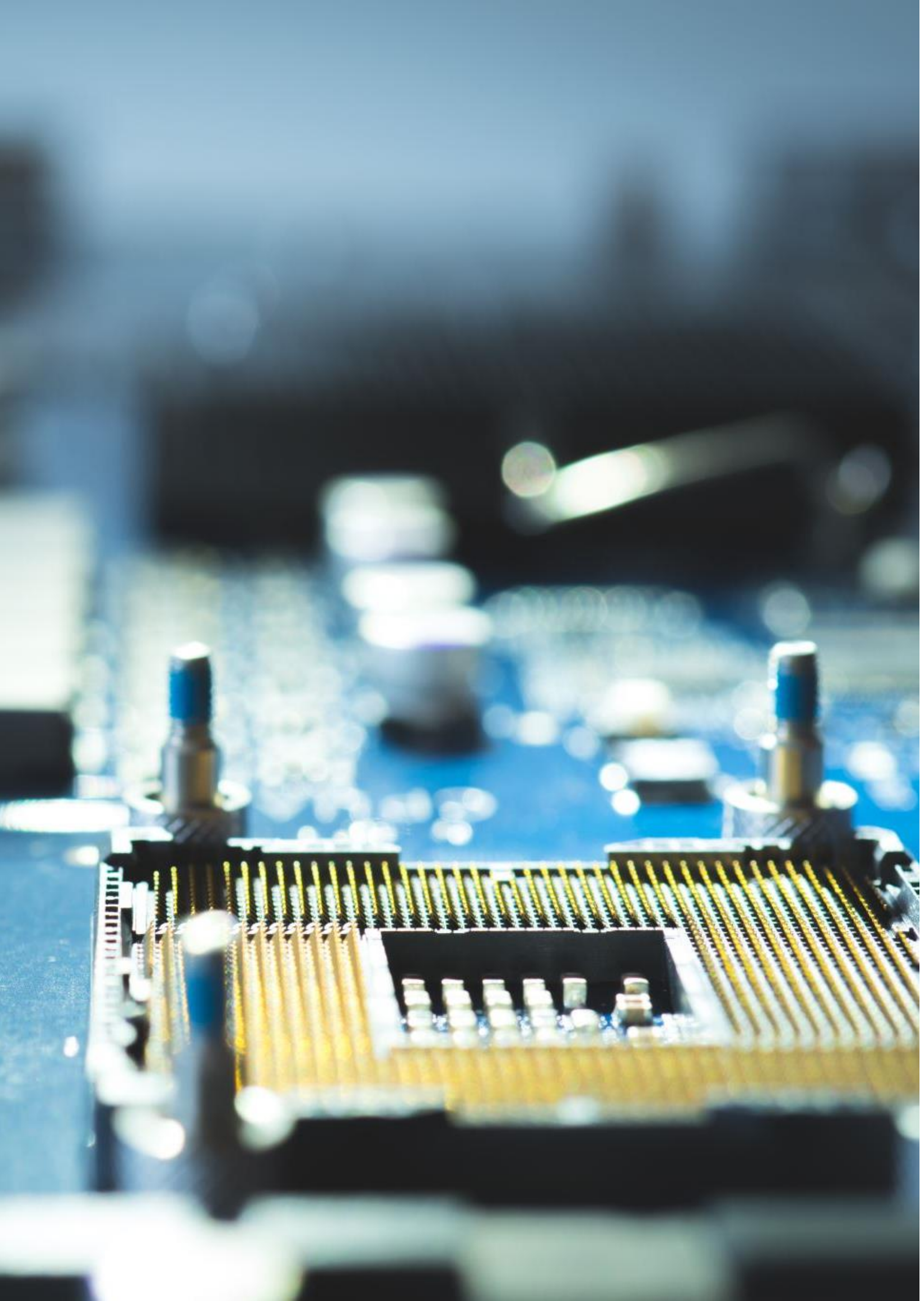
In class Practice Problem 9

```
> model1=lm(formula = QUALITY ~ PRESSURE, data=quality)
> summary(model1)

Call:
lm(formula = QUALITY ~ PRESSURE, data = quality)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.441 -10.698  -2.543   7.108  30.735 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 167.5999    30.3011   5.531 4.57e-07 ***
PRESSURE     -1.8352     0.5403  -3.397   0.0011 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.07 on 74 degrees of freedom
Multiple R-squared:  0.1349,	Adjusted R-squared:  0.1232 
F-statistic: 11.54 on 1 and 74 DF,  p-value: 0.0011
```

```
> model2=lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2), data=quality)
> summary(model2)

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
```

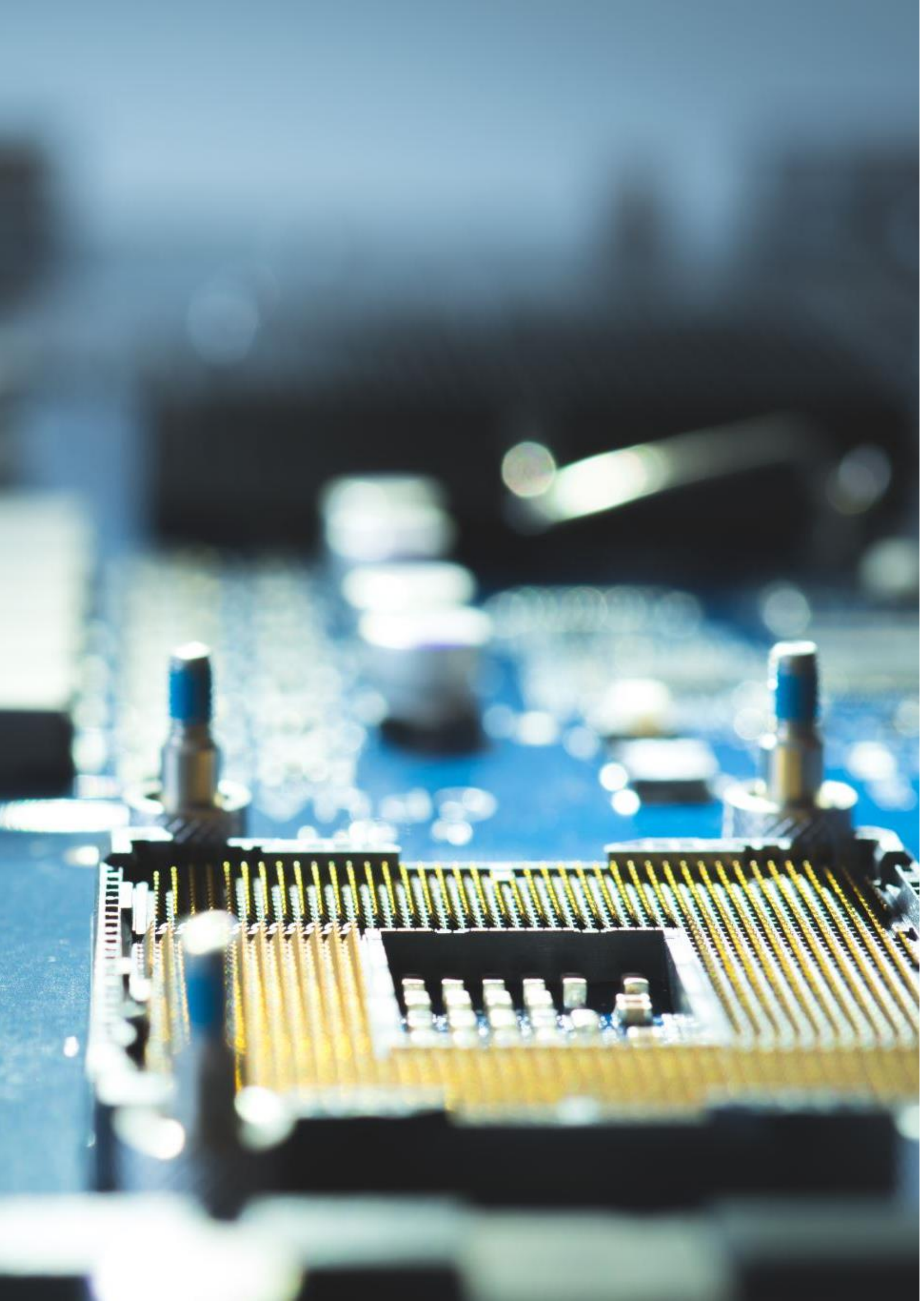
```
> model3=lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2) + I(PRESSURE^3), data=quality)
> summary(model3)

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
```

```
> model4=lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2) + I(PRESSURE^3)+I(PRESSURE^4), data=quality)
> summary(model4)

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
```



In class Practice Problem 9

```
> model5=lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2) + I(PRESSURE^3)+I(PRESSURE^4)+I(PRESSURE^5), data=
quality)
> summary(model5)

Call:
lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2) + I(PRESSURE^3) +
    I(PRESSURE^4) + I(PRESSURE^5), data = quality)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9191  -4.9140  -0.6831   4.3809  12.6809

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.647e+06  3.020e+06  -1.208    0.231
PRESSURE       3.401e+05  2.746e+05   1.239    0.220
I(PRESSURE^2) -1.268e+04  9.976e+03  -1.271    0.208
I(PRESSURE^3)  2.361e+02  1.810e+02   1.304    0.197
I(PRESSURE^4) -2.196e+00  1.641e+00  -1.338    0.185
I(PRESSURE^5)  8.162e-03  5.945e-03   1.373    0.174

Residual standard error: 6.306 on 70 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.8465
F-statistic: 83.74 on 5 and 70 DF,  p-value: < 2.2e-16

> model6=lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2) + I(PRESSURE^3)+I(PRESSURE^4)+I(PRESSURE^5)+I(PRESSURE^
6), data=quality)
> summary(model6)

Call:
lm(formula = QUALITY ~ PRESSURE + I(PRESSURE^2) + I(PRESSURE^3) +
    I(PRESSURE^4) + I(PRESSURE^5) + I(PRESSURE^6), data = quality)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9191  -4.9140  -0.6831   4.3809  12.6809

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.647e+06  3.020e+06  -1.208    0.231
PRESSURE       3.401e+05  2.746e+05   1.239    0.220
I(PRESSURE^2) -1.268e+04  9.976e+03  -1.271    0.208
I(PRESSURE^3)  2.361e+02  1.810e+02   1.304    0.197
I(PRESSURE^4) -2.196e+00  1.641e+00  -1.338    0.185
I(PRESSURE^5)  8.162e-03  5.945e-03   1.373    0.174
I(PRESSURE^6)         NA         NA      NA       NA

Residual standard error: 6.306 on 70 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.8465
F-statistic: 83.74 on 5 and 70 DF,  p-value: < 2.2e-16
```



In class Practice Problem 9

- Adjusted R2: model 5 > model 4 > model 3> model 2 > model 1.
- RMSE: model 5 < model 4 < model 3 < model 2 < model 1.
- Which order /model should we choose?

Too many predictors? Overfitting?

Coffee break

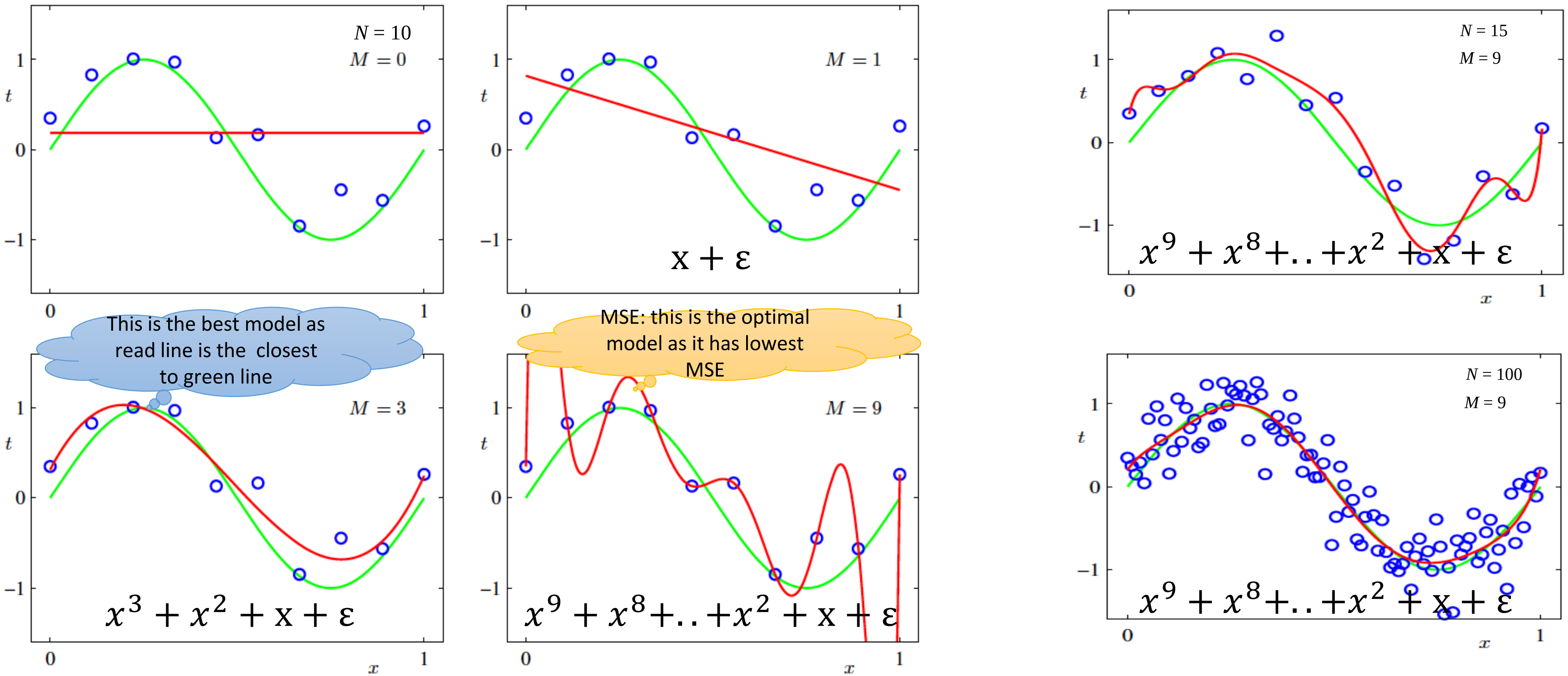
Which model should we choose?

$Y = \sin(x) + \varepsilon$

$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_1^2$

Least square method

SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

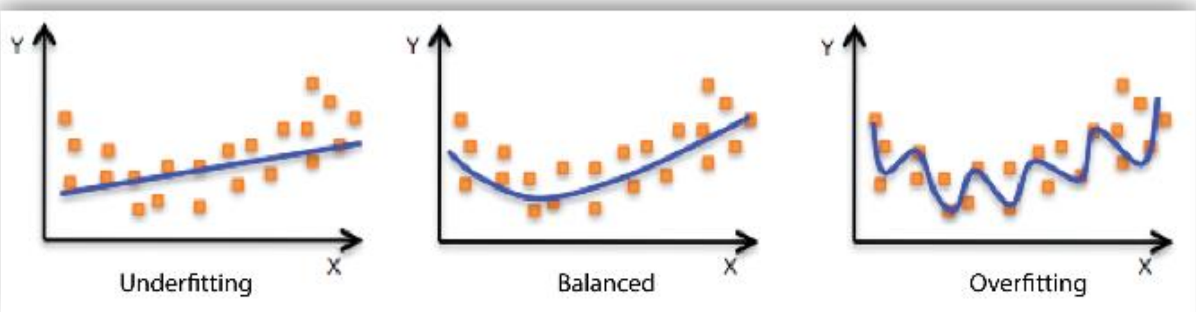


M = degree of polynomial

N = sample size

Bishop 2006, Pattern Recognition and Machine Learning

Which model should we choose?



- One can make the model complicated enough so that the MSE is very small.
- Overfitting: a scenario in data science where model is too closely or exactly to a particular set of data and may therefore fail to fit to additional data or predict future observations reliably.
- Underfitting: another scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data.
- We want to avoid overfitting and underfitting and we want have a balanced model.

Cross validation

Least square method
SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

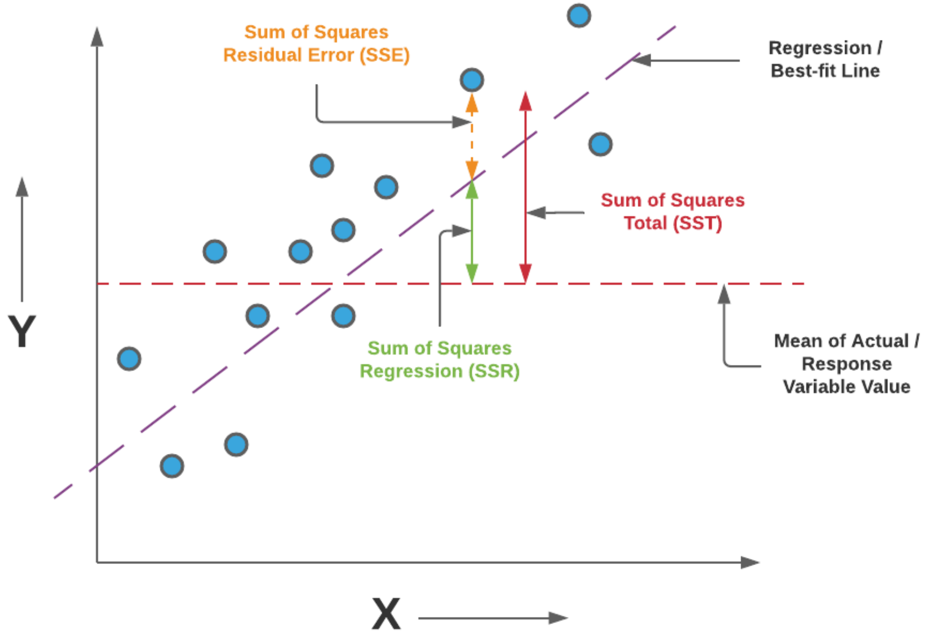
Least square method
SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

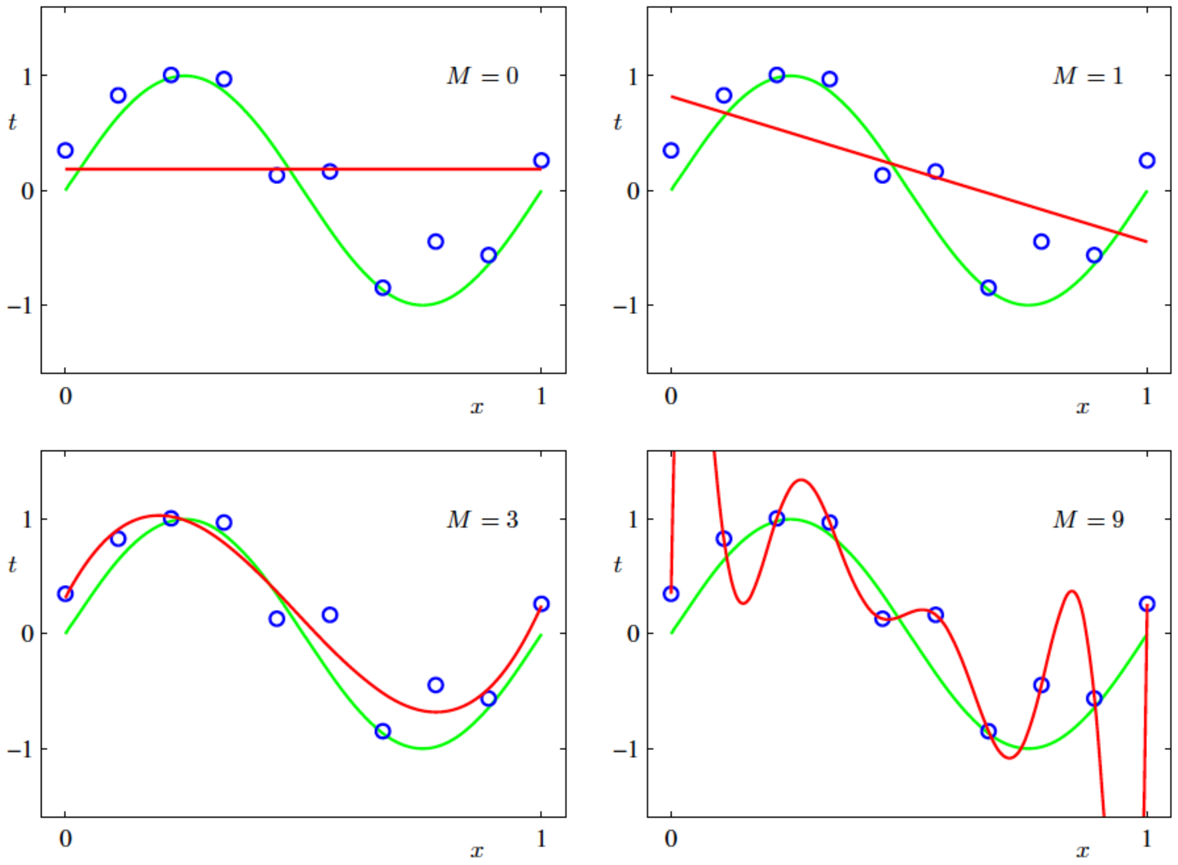
Least square method
SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Least square method
SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.
- The procedure has a single parameter called **k** that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.
- Cross-validation is primarily used in applied machine learning to estimate the performance of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.
- It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model performance than other methods, such as a simple train/test split.

Which model should we choose?

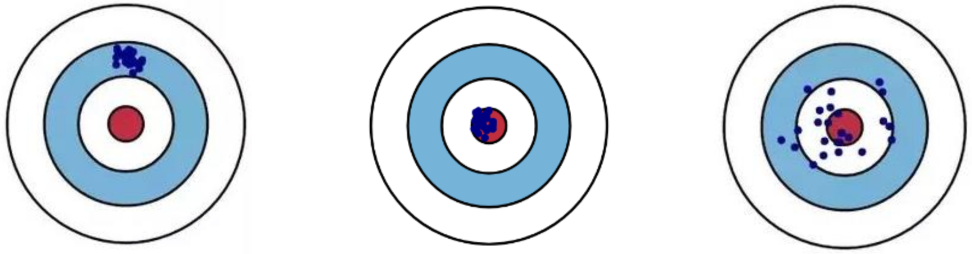




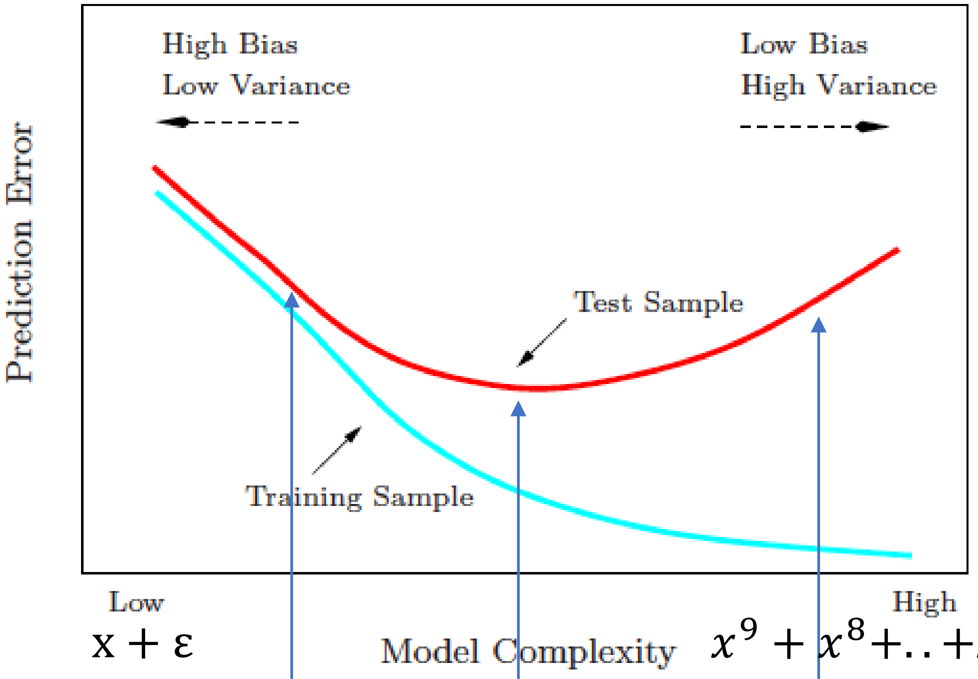
Least square method

SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Which model should we choose?



$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



$\mathrm{x} + \varepsilon$

$x^9 + x^8 + .. + x^2 + \mathrm{x} + \varepsilon$

Underfitting | Balanced | Overfitting

Cross validation (DAAG)

```
> out1<-CVlm(data=credit, m=mk, seed=20230525,
+            form.lm = formula(Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                              +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student)))


fold 1
Observations in test set: 40
                    6         10        21        39         46        64        69        83        94       103       120       137      138
Predicted   1134.52127 1384.40784  189.4167 499.53293 1021.06446  234.1633 801.97737 515.98595 860.14302 1694.5837 -122.7813  179.3011 255.7569
cvpred      1134.38532 1389.77662  191.3352 503.28371 1022.05598  235.8375 802.34248 517.63806 862.80783 1697.3476 -123.1055  179.5909 258.2165
Balance     1151.00000 1350.00000   89.0000 531.00000  997.00000  133.0000 822.00000 503.00000 937.00000 1587.0000    0.0000   75.0000 187.0000
CV residual   16.61468  -39.77662 -102.3352  27.71629  -25.05598 -102.8375  19.65752 -14.63806  74.19217 -110.3476  123.1055 -104.5909 -71.2165
                 143       146        175       178        208       211       243       257       258       269       280       288      289
Predicted   640.95954 589.55576 1664.54527 391.40720 1156.94835 181.32173 102.18309 -89.19359  67.19496 -128.4911 296.14722 -30.11823 824.6291
cvpred      641.35407 591.52828 1664.69487 394.03729 1160.59347 182.62262 109.13149 -89.38667  68.46046 -128.8340 303.67641 -29.98238 826.9321
Balance     669.00000 642.00000 1573.00000 384.00000 1216.00000  95.00000  16.00000   0.00000   0.00000    0.0000 269.00000   0.00000 863.0000
CV residual  27.64593  50.47172  -91.69487 -10.03729   55.40653 -87.62262 -93.13149  89.38667 -68.46046  128.8340 -34.67641  29.98238  36.0679
                 292       295       301       302       320       323       330       334       341        354       364       368        387
Predicted   335.10208  2.975958 558.01605 254.26787  105.2528 315.46725 793.96161 243.59172 362.44330 432.697214 594.61479 267.38983 370.408605
cvpred      337.23163  4.164024 559.74472 256.62579  107.1743 318.07979 795.20617 244.88758 365.28727 434.299534 594.34959 268.52042 372.672499
Balance     309.00000  0.000000 580.00000 172.00000    0.0000 265.00000 846.00000 182.00000 320.00000 425.000000 578.00000 216.00000 371.000000
CV residual -28.23163 -4.164024  20.25528 -84.62579 -107.1743 -53.07979  50.79383 -62.88758 -45.28727  -9.299534 -16.34959 -52.52042  -1.672499
                 393
Predicted    30.18593
cvpred       32.44740
Balance       0.00000
CV residual -32.44740

Sum of squares = 173112.4    Mean square = 4327.81    n = 40
```

- Package “DAAG”: Data Analysis And Graphing. The ‘DAAG’ package contains three functions for k – fold cross validation; the ‘cv.lm’ function is used for simple linear regression models, the ‘CVlm’ function is used for multiple linear regression models, and the ‘CVbinary’ function is used for logistic regression models. The k –fold method randomly removes k – folds for the testing set and models the remaining (training set) data.
- R command: library(DAAG); CVlm(data, form.lm, m=3)
- The input data frame is returned, with additional columns Predicted (Predicted values using all observations) and cvpred (cross-validation predictions). The cross-validation residual sum of squares (ss) and degrees of freedom (df) are returned as attributes of the data frame.
- Here, at the bottom of the output we get the cross validation **residual sums of squares** (Overall MS); which is a corrected measure of prediction error averaged across all folds. The function also produces a plot of each fold’s predicted values against the actual outcome variable (y); with each fold a different color.

In class Practice Problem 8+

```
> library(DAAG)
>
> model_inter_refine3=lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                         +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student)
+                         ,data=credit)
> summary(model_inter_refine3)

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
```

Model	Adjusted R2	RMSE
Model_inter_refine3	0.9809	63.6

In class Practice Problem 8+

```
> model_inter_high_order1=lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                            +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student)
+                            +I(Income^2)
+                            ,data=credit)
> summary(model_inter_high_order1)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-203.523  -38.565    6.857   37.878  123.752 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -1.633e+02  1.925e+01  -8.486 4.56e-16 ***
Income                     1.403e+00  5.522e-01   2.541 0.011437 *  
Limit                      6.481e-02  1.943e-02   3.336 0.000933 ***
Rating                    -4.319e-01  2.820e-01  -1.532 0.126438    
Cards                      1.814e+01  2.453e+00   7.393 8.94e-13 ***
Age                       -7.455e-01  1.661e-01  -4.489 9.43e-06 ***
factor(Student)Yes         1.564e+02  2.320e+01   6.743 5.66e-11 ***
I(Income^2)                5.716e-02  5.363e-03  10.659  < 2e-16 ***
Income:Rating             -4.134e-02  2.516e-03 -16.428  < 2e-16 ***
Income:factor(Student)Yes -2.327e+00  3.960e-01  -5.876 9.04e-09 ***
Limit:Rating               5.227e-04  2.302e-05  22.710  < 2e-16 ***
Limit:factor(Student)Yes   8.310e-02  6.764e-03  12.286  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56 on 388 degrees of freedom
Multiple R-squared:  0.9856,	Adjusted R-squared:  0.9852 
F-statistic:  2409 on 11 and 388 DF,  p-value: < 2.2e-16
```

Model	Adjusted R2	RMSE
Model_inter_refine3	0.9809	63.6
Model_inter_high_order1	0.9852	56

In class Practice Problem 8+

```
> model_inter_high_order2=lm(formula = Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                            +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student)
+                            +I(Income^2)+I(Rating^2)+I(Cards^2)+I(Age^2)
+                            ,data=credit)
> summary(model_inter_high_order2)

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

Residuals:
     Min       1Q   Median       3Q      Max
-169.522  -39.994    6.786   38.310  129.475

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.324e+02  3.896e+01  -5.966 5.51e-09 ***
Income                    1.325e+00  5.562e-01   2.382  0.01770 *
Limit                    -4.416e-02  3.958e-02  -1.116  0.26523
Rating                    1.362e+00  6.445e-01   2.114  0.03517 *
Cards                     6.870e+00  7.527e+00   0.913  0.36194
Age                       2.733e-01  1.103e+00   0.248  0.80446
factor(Student)Yes        1.527e+02  2.295e+01   6.655 9.76e-11 ***
I(Income^2)               5.623e-02  5.368e-03  10.474  < 2e-16 ***
I(Rating^2)              -4.396e-03  1.453e-03  -3.026  0.00265 **
I(Cards^2)                1.555e+00  9.976e-01   1.559  0.11981
I(Age^2)                 -9.187e-03  9.810e-03  -0.936  0.34962
Income:Rating            -4.091e-02  2.534e-03 -16.143  < 2e-16 ***
Income:factor(Student)Yes -2.229e+00  3.927e-01  -5.676 2.72e-08 ***
Limit:Rating              8.155e-04  9.769e-05   8.348 1.26e-15 ***
Limit:factor(Student)Yes  8.303e-02  6.695e-03  12.403  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.35 on 385 degrees of freedom
Multiple R-squared:  0.986,	Adjusted R-squared:  0.9855
F-statistic:  1939 on 14 and 385 DF,  p-value: < 2.2e-16
```

Model	Adjusted R2	RMSE
Model_inter_refine3	0.9809	63.6
Model_inter_high_order1	0.9852	56
Model_inter_high_order2	0.9855	55.35

Is there any overfitting for the last model?

Cross validation with CVlm() function!

In class Practice Problem 8+

```
> out1<-CVlm(data=credit, m=mk, seed=20230525,
+            form.lm = formula(Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                              +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student)))
```

```
fold 10
Observations in test set: 40
                    25        28        40        59         70        79         80        93         96        98        99        141       153       164
Predicted     1.962058 457.65709 355.82829 337.75830 1061.73262 366.63122  14.12024  119.2585 -103.28773 243.18944 384.39904 1457.91009 206.48467  111.1840
cvpred        8.743194 453.36958 355.56673 344.85299 1061.34702 357.19056  19.81941  120.7995  -94.98682 246.19752 388.02545 1453.51163 202.11619  119.9058
Balance       0.000000 467.00000 344.00000 333.00000 1084.00000 391.00000   0.00000    0.0000    0.00000 155.00000 375.00000 1425.00000 156.00000    0.0000
CV residual  -8.743194  13.63042 -11.56673 -11.85299   22.65298  33.80944 -19.81941 -120.7995   94.98682 -91.19752 -13.02545  -28.51163 -46.11619 -119.9058
                   166       168       186       190       217       218        226       228       234       238       244       254        274       278
Predicted   573.306453 -12.00866 436.14914 206.73536  160.8509 878.05289  994.79381 499.08996  79.05650 466.75094 817.75517 266.98637 1201.45591 503.74114
cvpred      574.156052 -15.38844 428.83052 205.88421  163.8298 879.07664  988.72668 495.72799  86.29129 459.00244 813.11219 262.02215 1199.17784 497.67929
Balance     570.000000   0.00000 450.00000 126.00000   52.0000 955.00000 1075.00000 482.00000   0.00000 443.00000 856.00000 218.00000 1255.00000 531.00000
CV residual  -4.156052  15.38844  21.16948 -79.88421 -111.8298  75.92336   86.27332 -13.72799 -86.29129 -16.00244  42.88781 -44.02215   55.82216  33.32071
                    284       290        296       315       324       357        370       372       385      390       395       397
Predicted   886.0090455 463.97964 -100.03935 1140.3608 2230.8172 938.31918 1258.78208 -61.77125 -36.84096 752.4319 700.30244 460.57652
cvpred      890.1367546 461.87366  -93.76674 1150.7876 2298.0696 932.97393 1264.32352 -61.34712 -27.41481 747.1448 692.14919 463.21554
Balance     890.0000000 485.00000    0.00000 1140.0000 1999.0000 962.00000 1208.00000   0.00000   0.00000 806.0000 734.00000 480.00000
CV residual  -0.1367546  23.12634   93.76674  -10.7876 -299.0696  29.02607  -56.32352  61.34712  27.41481  58.8552  41.85081  16.78446

Sum of squares = 213608.8    Mean square = 5340.22    n = 40

Overall (Sum over all 40 folds)
      ms
4378.823
```

Model	Test MSE	Overall mean squared error
Model_inter_refine3		4378.823
Model_inter_high_order1		3320.602
Model_inter_high_order2		3330.393

Cross validation with CVlm() function!

In class Practice Problem 8+

```
> out2<-CVlm(data=credit, m=mk, seed=20230525,
+            form.lm = formula(Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                              +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student)
+                              +I(Income^2)))
```

```
fold 10
Observations in test set: 40
                   25       28           40         59         70        79        80        93        96        98         99        141       153       164
Predicted   -9.252072 435.4248 3.429202e+02 336.455768 1068.15207 429.65957  1.076396  96.56571 -88.48731 218.59290 366.253949 1467.92996 230.47997  93.39682
cvpred      -3.284561 431.7985 3.439905e+02 340.495728 1067.90712 420.21383  5.682573  97.52032 -81.30322 221.82232 370.288551 1465.27941 229.35166 101.72216
Balance      0.000000 467.0000 3.440000e+02 333.000000 1084.00000 391.00000  0.000000   0.00000   0.00000 155.00000 375.000000 1425.00000 156.00000   0.00000
CV residual  3.284561  35.2015 9.482638e-03  -7.495728   16.09288 -29.21383 -5.682573 -97.52032  81.30322 -66.82232   4.711449  -40.27941 -73.35166 -101.72216
                  166       168       186       190       217        218         226       228       234       238      244       254        274       278
Predicted  565.588668  -8.00710 415.75904 196.45399 140.75479 952.893587 1075.187069 455.74465  61.96973 432.55959 803.2079 269.28798 1234.15148 480.65454
cvpred     565.402262 -11.39794 409.62211 194.58702 143.95463 952.406848 1069.956471 452.26781  68.36292 427.67356 799.1088 265.30367 1236.03625 476.17323
Balance    570.000000   0.00000 450.00000 126.00000  52.00000 955.000000 1075.000000 482.00000   0.00000 443.00000 856.0000 218.00000 1255.00000 531.00000
CV residual  4.597738  11.39794  40.37789 -68.58702 -91.95463   2.593152    5.043529  29.73219 -68.36292  15.32644  56.8912 -47.30367   18.96375  54.82677
                 284       290       296        315       324        357        370       372       385       390      395       397
Predicted  873.22777 469.60890 -75.93885 1110.66480 2202.5229 961.367200 1243.06145 -25.03509 -44.39880 724.87071 677.1383 459.04520
cvpred     875.39454 467.16337 -70.70722 1119.18179 2261.3795 956.816081 1246.87027 -24.39788 -37.68827 720.50497 672.3150 462.44357
Balance    890.00000 485.00000   0.00000 1140.00000 1999.0000 962.000000 1208.00000   0.00000   0.00000 806.00000 734.0000 480.00000
CV residual 14.60546  17.83663  70.70722   20.81821 -262.3795   5.183919  -38.87027  24.39788  37.68827  85.49503  61.6850  17.55643

Sum of squares = 159798.1    Mean square = 3994.95    n = 40

Overall (Sum over all 40 folds)
      ms
3320.602
```

Model	Test MSE	Overall mean squared error
Model_inter_refine3		4378.823
Model_inter_high_order1		3320.602
Model_inter_high_order2		3330.393

Cross validation with CVlm() function!

In class Practice Problem 8+

```
> out3<-CVlm(data=credit, m=mk, seed=20230525,
+            form.lm = formula(Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                               +Income*Rating+Income*factor(Student)+Limit*Rating+Limit*factor(Student)
+                               +I(Income^2)+I(Rating^2)+I(Cards^2)+I(Age^2)))
```

```
fold 10
Observations in test set: 40
                   25        28         40        59         70        79        80        93         96        98         99        141      153       164
Predicted   -8.433411 436.23058 339.250784 341.89871 1061.97240 433.91626 -2.689230 107.4459 -59.29174 215.63257 370.330291 1456.47921 223.5515  97.18885
cvpred      -5.920932 430.35797 339.505034 343.16932 1059.36362 421.34817  0.472763 102.5062 -62.09733 220.75699 371.921941 1461.33861 233.7753 108.88477
Balance      0.000000 467.00000 344.000000 333.00000 1084.00000 391.00000  0.000000   0.0000   0.00000 155.00000 375.000000 1425.00000 156.0000   0.00000
CV residual  5.920932  36.64203   4.494966 -10.16932   24.63638 -30.34817 -0.472763 -102.5062 62.09733 -65.75699   3.078059  -36.33861 -77.7753 -108.88477
                  166       168       186       190       217        218         226       228       234       238       244      254        274       278
Predicted   564.741526 -14.75389 415.12651 189.81114 135.37869 951.611801 1067.933816 451.16684  57.30127 431.65577 821.79714 274.0248 1239.04489 482.21358
cvpred      562.560665 -15.82321 413.26064 196.42005 137.19324 949.044603 1070.832509 445.99317  64.62524 427.90158 807.86676 266.3535 1239.54615 474.02935
Balance     570.000000   0.00000 450.00000 126.00000  52.00000 955.000000 1075.000000 482.00000   0.00000 443.00000 856.00000 218.0000 1255.00000 531.00000
CV residual   7.439335  15.82321  36.73936 -70.42005 -85.19324   5.955397    4.167491  36.00683 -64.62524  15.09842  48.13324 -48.3535   15.45385  56.97065
                  284       290       296        315       324        357        370       372       385       390       395       397
Predicted   860.75032 465.76103 -57.5651 1118.69949 2163.4724 958.796114 1221.13597 -37.70804 -34.00446 733.77369 676.96460 460.77306
cvpred      866.47445 468.86588 -58.8099 1121.31659 2245.4291 955.808316 1225.75797 -31.87942 -28.74142 724.12654 671.93019 462.14406
Balance     890.00000 485.00000   0.0000 1140.00000 1999.0000 962.000000 1208.00000   0.00000   0.00000 806.00000 734.00000 480.00000
CV residual  23.52555  16.13412  58.8099   18.68341 -246.4291   6.191684  -17.75797  31.87942  28.74142  81.87346  62.06981  17.85594

Sum of squares = 147136.9    Mean square = 3678.42    n = 40

Overall (Sum over all 40 folds)
      ms
3330.393
```

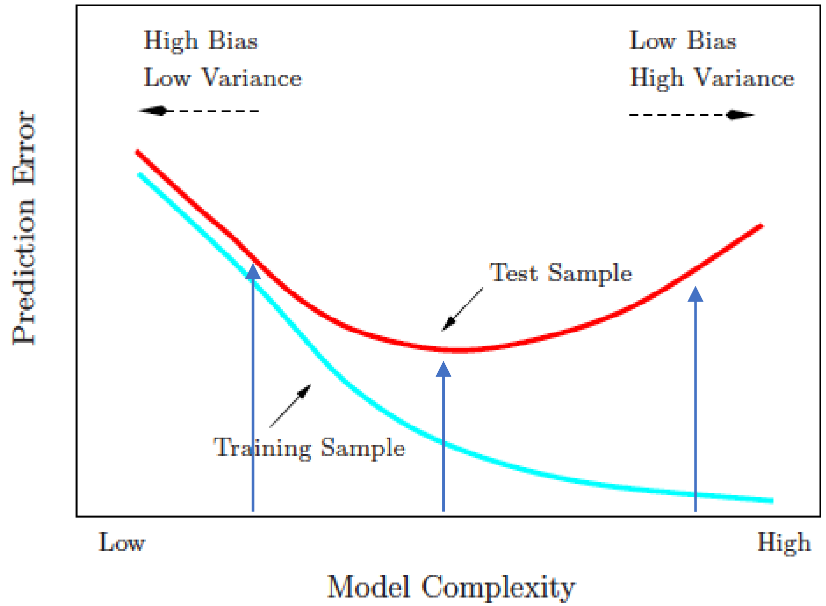
Model	Test MSE	Overall mean squared error
Model_inter_refine3		4378.823
Model_inter_high_order1		3320.602
Model_inter_high_order2		3330.393

In class Practice Problem 8+

```
cv_error1<-mean((out1$cvpred-out1$Balance)^2)

cv_error2<-mean((out2$cvpred-out2$Balance)^2)

cv_error3<-mean((out3$cvpred-out3$Balance)^2)

print(paste(cv_error1, cv_error2, cv_error3))
```

Model	Test MSE	Overall mean squared error
Model_inter_refine3	4378.822	4378.823
Model_inter_high_order1	3320.60	3320.602
Model_inter_high_order2	3330.393	3330.393

In class Practice Problem 8+

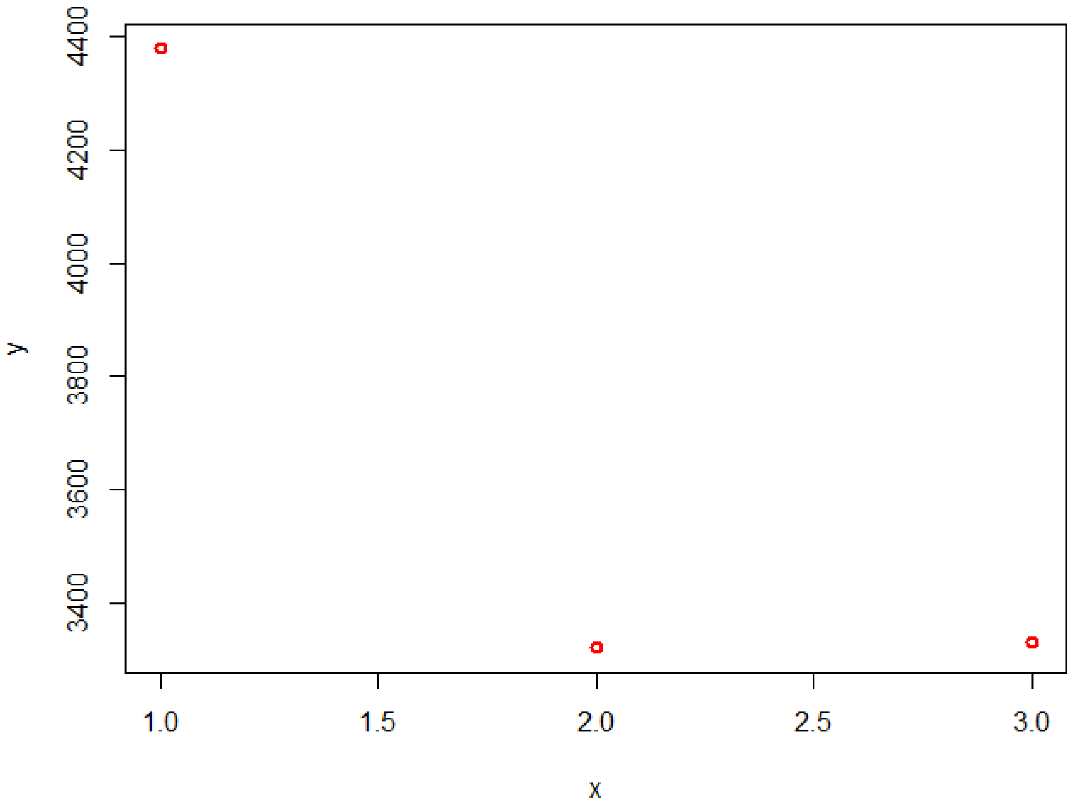


Underfitting Just-right Overfitting

Model	Adjusted R2	RMSE
Model_inter_refine3	0.9809	63.6
Model_inter_high_order1	0.9852	56
Model_inter_high_order2	0.9855	55.35

Model	Test MSE	Overall mean squared error
Model_inter_refine3	4378.822	4378.823
Model_inter_high_order1	3320.60	3320.602
Model_inter_high_order2	3330.393	3330.393

Leah: Oops, the last model is overfitting, and the second model seems just right



Model selection

- More about model selection? See you tomorrow at Next lecture

Dr. Thuntida Ngamkham's approach

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

Leah's approach:

1. Start with an interaction model with all predictors
2. Remove non-significant interactions
3. Rerun model to ensure all predictors are significant
4. Iterate step 3 until done.

Take away messages

- Statistics:
  - Interaction Effect in Multiple Regression with both Quantitative and Qualitative (Dummy) Variable models
  - Two different approaches but result in the same optimal model
  - A Quadratic (Second Order) Model with Quantitative predictors
  - Cross validation to avoid overfitting model
- Code:
  - lm(y ~ x1+x2+(x1+x2)^2 + I(X1^2)+I(X2^2))
  - CVlm()

Thank you

- Questions OR Comments?
- Slack channel: section2-course-documents
- Email: qing.li2@uclagary.ca