MA 214 Lecture 3.

- Multiple Regression.
  - Choose proper variables.
    - Issue with "overlapping"
  - Computational intensive expensed.

- So far: Continuous variables.
- How to incorporate categorical covariates in a regression model?

Example: Salaries of the recent graduates.

- Continuous: Salary

  Years of Experience

- Categorical: Gender.

Indicator variable.

Define an indicator / Dummy variable:

$$X_2 = \begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female} \end{cases}$$

Thus if the fitted model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Then implies:

$$\hat{Y} = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1 & \text{if Male} \\ \beta_0 + \beta_1 X_1 & \text{if Female.} \end{cases}$$

- Why don't we just go through two subsets of data. like "Male" and "Female"?
  - Some will have large amounts of categorical variables.
    - Therefore it will be extremely inefficient.

- Convert categorical variables to "indicator variable".

- Some case will have more than 2 "values"
  - Major?

Sci/Math (SM), Social Science (SS), Humanities (HU)

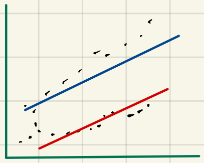
- We need to two indicators.

	X_1	X_2
SM	1	0
SS	0	1
HU	0	0

Thus the fitted model is: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

$$\hat{Y} = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1 & \text{if SM} \\ (\beta_0 + \beta_3) + \beta_1 X_1 & \text{if SS} \\ \beta_0 + \beta_1 X_1 & \text{if HU} \end{cases}$$

- The order of founding the indicator variable does not affect the interpretation.



- Two lines are parallel line.
- It is kind of like we combine two seperate models into one.

- If we have k values (categorical)
- One less for indicators.

Interaction Term.

- The effect of one covariate on the response variable may depend on the value of another covariate.

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

If X_2 is an indicator variable (say, 1 if Male and 0 if Female), then the model implies.

$$\hat{Y} = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 & \text{if Male} \\ \beta_0 + \beta_1 X_1 & \text{if Female.} \end{cases}$$

First-order model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Second-order model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

- In JMP, choose "cross" after select all the covariates.
  - Result: R^2 increase significantly.

Non-Linear Regression.

- Multiple types of them.

Polynomial Regression

- PR of order 2, second order model with one covariate.

$$y_i = \beta_0 + \underbrace{\beta_1 (x_i - \bar{x})}_{x_1} + \underbrace{\beta_2 (x_i - \bar{x})^2}_{x_2} + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

This can be viewed as a multiple regression model with two covariates.

$$x_1 = (x_i - \bar{x}) \text{ and } x_2 = (x - \bar{x})^2$$

- When some variables are not significant, we need to carry out variable selection.