# ISE-529 Predictive Analytics

Midterm Preparation

# Mid-Term

## Logistics

- Mid-term will be administered in class on Monday, July25
  - 4:00PM – 5:30PM
  - Can be taken in class or remotely
- Exam will be open-book/open-notes but must be done individually
  - Automated software will be used to identify identical responses
- There will be a one-hour lecture after completion of the mid-term (5:45 – 6:45PM)

# Mid-Term

## Format

- Mid-term will consist of short-answer questions to test your understanding of the theoretical concepts presented in module 1-4
  - Interpretation of model outputs and visualizations
  - Simple calculations that can be done manually of using Excel
- Exam will be a PowerPoint file
  - Similar to homework assignments, you will download the PowerPoint exam file, enter your responses, save it to a PDF file format and upload it to Gradescope

# Mid-Term

## Module 1/2 Sample Question Types

- Prediction vs inference - give examples and ask primary modeling objective
- Matrix notation - give actual vectors and ask for components by their symbols
- Bias/variance tradeoff
- KNN / calculate misclassification rates

# Problem 1

The training dataset for a simple KNN classification problem is given below:

| X1 | X2 | X3 | Y | Distance to (0,0,0) |
|---|---|---|---|---|
| -1 | 2 | 1 | B | $\sqrt{6} \approx 2.45$ |
| -2 | 1 | 2 | B | 3 |
| 2 | -3 | -3 | A | $\sqrt{22} \approx 4.7$ |
| 3 | 2 | 2 | A | $\sqrt{17} \approx 4.1$ |
| 0 | 0 | -2 | A | 2 |
| 0 | 3 | 2 | A | $\sqrt{13} \approx 3.6$ |

1A) Assume we are trying to use this dataset to make a prediction for Y when X1 = X2 = X3 = 0. Complete the “Distance to (0,0,0) Column” above using the Euclidian distance formula

# Problem 1

1B) What would your prediction be with K = 1? A

1C) What would your prediction be with K = 3? B

1D) If the true decision boundary is highly non-linear ("curvy") would you expect a higher or lower value for K to provide a better prediction? LOWER. Why? Lower values result in a closer fit to the data and a more complex model which would better fit a non-linear decision boundary. Higher values result in a "smoothed" boundary with fits best with a more linear decision boundary

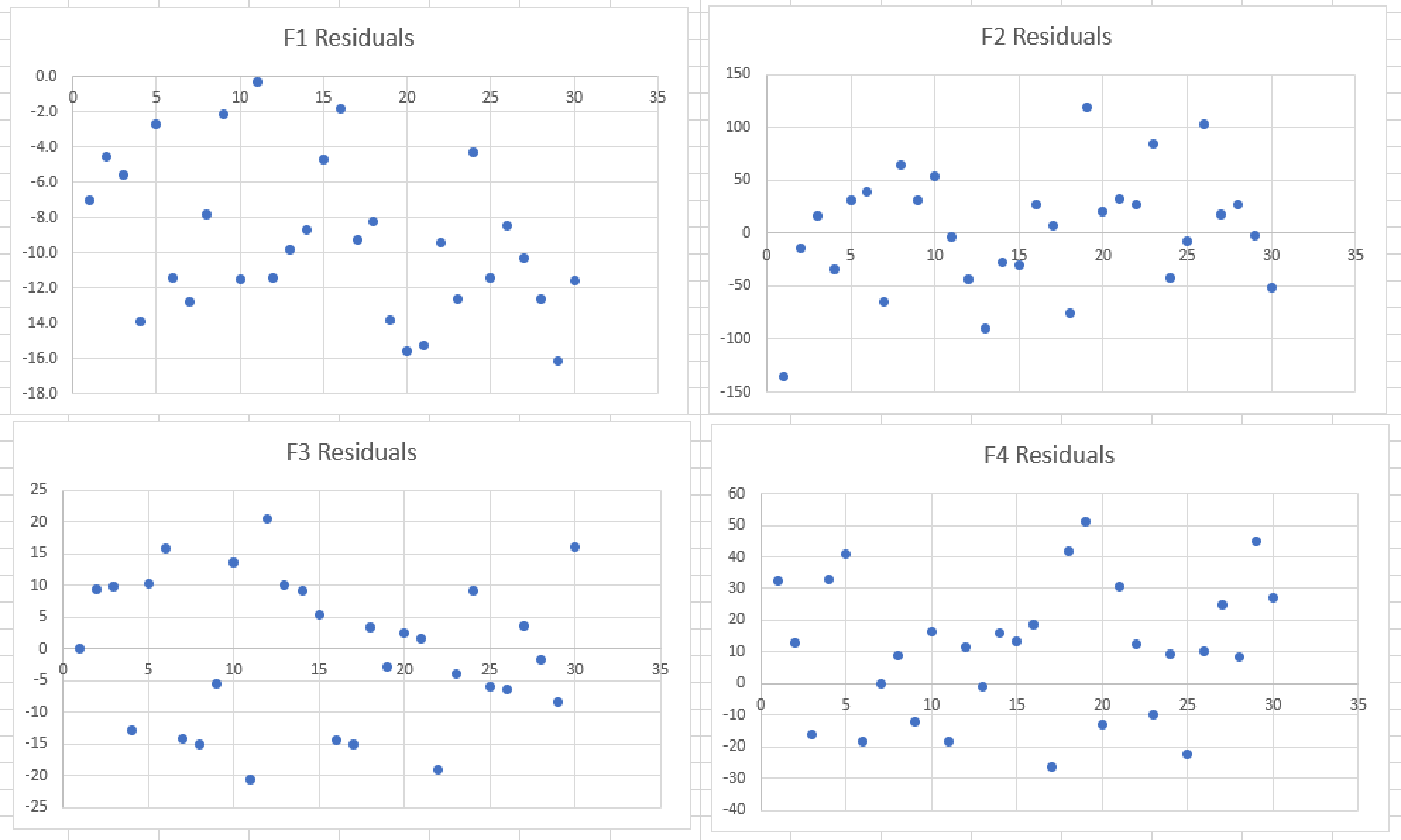
# Problem 2

You have created four regression models for a multidimensional datasets and plotted the test dataset residuals for each model on the following slides. Complete the following matrix by typing a  to assess each model in terms of its relative variance and bias:

| | Low Variance | High Variance |
|---|---|---|
| **Low Bias** | F3 | F2 |
| **High Bias** | F1 | F4 |

Of these four models, which would you select:  F3

# Problem 2


F1 Residuals
0.0
-2.0
-4.0
-6.0
-8.0
-10.0
-12.0
-14.0
-16.0
-18.0
0
5
10
15
20
25
30
35
F2 Residuals
150
100
50
0
-50
-100
-150
F3 Residuals
25
20
15
10
5
0
-5
-10
-15
-20
-25
F4 Residuals
60
50
40
30
20
10
0
-10
-20
-30
-40

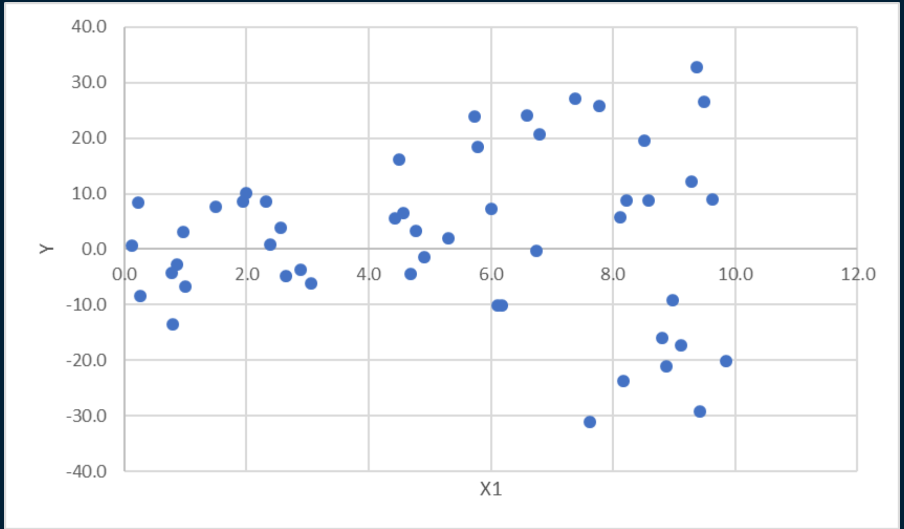
# Mid-Term

## Module 3 Sample Question Types

- Four assumptions of a linear model
- Write out equation from model coefficients
- Understanding and interpreting inference issues in the presence of multicollinearity
- Calculating various assessment statistics
- Interpreting p-values, confidence intervals, and F-statistics from a model result
- Interpreting and modeling interaction effects

# Problem 4

You are given a dataset with two inputs (X1 and X2) and one output (Y). X1 is a continuous attribute. X2 is a categorical attribute that has three possible values = "A", "B", or "C". The scatter plot below shows Y plotted as a function of X1:



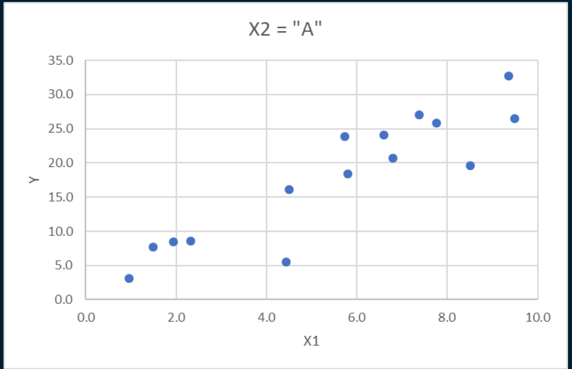
# Problem 4

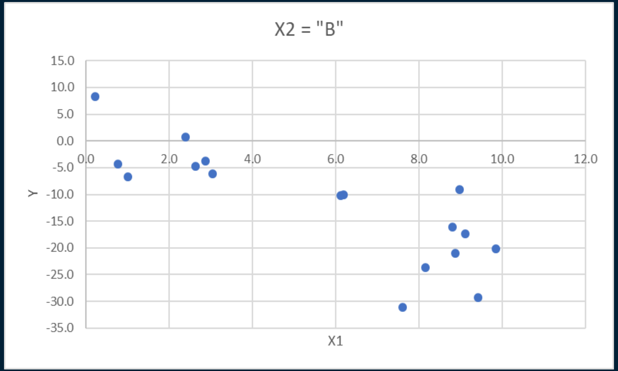
3A) How would you set up the equation for the linear regression?

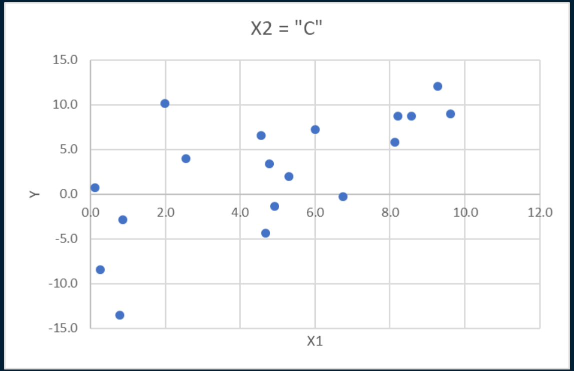
$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_{2a} + \beta_3 X_{2b} + \varepsilon}$$

# Problem 3

After tuning and evaluating your model, you are not happy with its performance. In attempt to understand what is going on, you plot scatter plots of Y plotted as a function of X1 for each of the three possible values of X2:



# Problem 3

3B) What appears to be going on with your data?

***There appears to be an interaction effect between X1 and X2***

3C) How would you modify your equation for the linear regression based on this?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_{2a} + \beta_3 X_{2b} + \beta_4 X_1 X_{2a} + \beta_5 X_1 X_{2b} + \varepsilon$$

Where $X_{2a}$=1 if $X_2 =$ "A" and $X_{2b}$=1 if $X_2 =$ "B" (and 0 otherwise)

# Problem 4

You are evaluating the following four candidate regression functions:

F1: $Y = 55{,}630.77 + 2591.30X_1 + 538.26X_2$

F2: $Y = 79{,}130.07 + 537.38X_1 + 736.2X_2 + 19.54{X_1}^2$

F3: $Y = 110{,}641.52 - 3174.0758X_1 + 561.356X_2 + 116.2{X_1}^2 - 0.64{X_1}^3$

F4: $Y = 128{,}431.04 - 3360.13X_1 - 351.65X_2 + 118.62{X_1}^2 - 0.65{X_1}^3 + 9.59{X_2}^2$

The various assessment metrics for the training and test partitions are given on the following page.

A) Which model would you select and why? F2 – has the lowest test MSE and highest test $R^2$

B) Do any of the candidate models exhibit possible overfitting? If yes, which ones? Yes – Models F3 and F4

# Problem 4

Training Partition

| Candidate regression function | Total Sum of Squares (TSS) | Residual Sum of Squares (RSS) | $R^2$ | Mean Squared Error |
|---|---|---|---|---|
| F1 | 261,238,232,586.45 | 58,463,155,045.00 | 0.78 | 2,338,526,201.80 |
| F2 | 261,238,232,586.45 | 52,260,913,256.85 | 0.80 | 2,090,436,530.27 |
| F3 | 261,238,232,586.45 | 48,838,883,468.24 | 0.81 | 1,953,555,338.73 |
| F4 | 261,238,232,586.45 | 47,699,709,571.19 | 0.82 | 1,907,988,382.85 |

Test Partition

| Candidate regression function | Total Sum of Squares (TSS) | Residual Sum of Squares (RSS) | $R^2$ | Mean Squared Error |
|---|---|---|---|---|
| F1 | 158,536,061,596.46 | 59,066,872,581.02 | 0.63 | 4,543,605,583.16 |
| F2 | 158,536,061,596.46 | 54,465,519,787.82 | 0.66 | 2,178,620,791.51 |
| F3 | 158,536,061,596.46 | 64,921,303,220.91 | 0.59 | 2,596,852,128.84 |
| F4 | 158,536,061,596.46 | 65,941,217,761.09 | 0.58 | 2,637,648,710.44 |

# Mid-Term

## Module 4 Sample Question Types

- Diagnosing residuals
  - Interpreting residuals plots
  - Recommending ways to resolve
- Manually calculating standard errors
- Bootstrap aggregation techniques

# Problem 2

The table below (and copied on the following page) shows a sample of 10 observations and then 10 bootstraps drawn from that sample. The mean, median, and sample standard deviation for the original sample and for each bootstrap is also provided. Use this information to complete the calculations on the following page. It may be helpful to copy this table into Excel.

| Sample | | | Bootstrap 1 | Bootstrap 2 | Bootstrap 3 | Bootstrap 4 | Bootstrap 5 | Bootstrap 6 | Bootstrap 7 | Bootstrap 8 | Bootstrap 9 | Bootstrap 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 75 | | 32 | 32 | 22 | 22 | 75 | 89 | 51 | 32 | 66 | 8 |
| 2 | 63 | | 32 | 32 | 47 | 22 | 51 | 8 | 89 | 63 | 75 | 63 |
| 3 | 32 | | 89 | 8 | 63 | 8 | 66 | 89 | 22 | 47 | 75 | 63 |
| 4 | 47 | | 89 | 89 | 8 | 66 | 51 | 51 | 89 | 89 | 51 | 22 |
| 5 | 51 | | 66 | 27 | 51 | 27 | 8 | 63 | 75 | 51 | 51 | 75 |
| 6 | 89 | | 75 | 32 | 27 | 66 | 63 | 63 | 63 | 47 | 27 | 89 |
| 7 | 66 | | 27 | 22 | 32 | 27 | 32 | 47 | 66 | 27 | 63 | 47 |
| 8 | 8 | | 63 | 75 | 32 | 75 | 89 | 32 | 22 | 75 | 89 | 75 |
| 9 | 22 | | 51 | 75 | 66 | 63 | 75 | 66 | 27 | 47 | 75 | 27 |
| 10 | 27 | | 89 | 51 | 32 | 22 | 32 | 32 | 89 | 51 | 27 | 47 |
| | | | | | | | | | | | | |
| Mean | 48 | | 61.3 | 44.3 | 38 | 39.8 | 54.2 | 54 | 59.3 | 52.9 | 59.9 | 51.6 |
| Median | 49 | | 64.5 | 32 | 32 | 27 | 57 | 57 | 64.5 | 49 | 64.5 | 55 |
| Std Dev | 25.7 | | 24.75 | 26.87 | 18.39 | 24.58 | 24.57 | 25.60 | 27.58 | 18.60 | 20.82 | 26.20 |

# Problem 2

| Sample | | | Bootstrap 1 | Bootstrap 2 | Bootstrap 3 | Bootstrap 4 | Bootstrap 5 | Bootstrap 6 | Bootstrap 7 | Bootstrap 8 | Bootstrap 9 | Bootstrap 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 75 | | 32 | 32 | 22 | 22 | 75 | 89 | 51 | 32 | 66 | 8 |
| 2 | 63 | | 32 | 32 | 47 | 22 | 51 | 8 | 89 | 63 | 75 | 63 |
| 3 | 32 | | 89 | 8 | 63 | 8 | 66 | 89 | 22 | 47 | 75 | 63 |
| 4 | 47 | | 89 | 89 | 8 | 66 | 51 | 51 | 89 | 89 | 51 | 22 |
| 5 | 51 | | 66 | 27 | 51 | 27 | 8 | 63 | 75 | 51 | 51 | 75 |
| 6 | 89 | | 75 | 32 | 27 | 66 | 63 | 63 | 63 | 47 | 27 | 89 |
| 7 | 66 | | 27 | 22 | 32 | 27 | 32 | 47 | 66 | 27 | 63 | 47 |
| 8 | 8 | | 63 | 75 | 32 | 75 | 89 | 32 | 22 | 75 | 89 | 75 |
| 9 | 22 | | 51 | 75 | 66 | 63 | 75 | 66 | 27 | 47 | 75 | 27 |
| 10 | 27 | | 89 | 51 | 32 | 22 | 32 | 32 | 89 | 51 | 27 | 47 |
| | | | | | | | | | | | | |
| Mean | 48 | | 61.3 | 44.3 | 38 | 39.8 | 54.2 | 54 | 59.3 | 52.9 | 59.9 | 51.6 |
| Median | 49 | | 64.5 | 32 | 32 | 27 | 57 | 57 | 64.5 | 49 | 64.5 | 55 |
| Std Dev | 25.7 | | 24.75 | 26.87 | 18.39 | 24.58 | 24.57 | 25.60 | 27.58 | 18.60 | 20.82 | 26.20 |

Standard Error for Mean, calculated by the following methods:

- Classical statistics: 8.13
- Bootstrap: 8.26

Standard Error for Median, calculated by using the bootstrap:

- 14.65

80% Confidence Interval for Mean (using the bootstrap):

- Lower Bound: 39.8
- Upper Bound: 59.9

80% Confidence Interval for Median:

- Lower Bound: 32
- Upper Bound: 64.5