Course Context and Content

D208 | Predictive Modeling

- Predictive Modeling (using Regression) is a statistical method that helps us to analyze and understand the relationship between two or more variables of interest.
- The process helps to understand which factors are important, which factors can be ignored, and how they are influencing each other.
- Therefore, your goal in this course is to:
  - Multiple Linear Regression
  - Logistic Regression
- In this course, you will either use Python or R to conduct regression analysis.
- Your competence will be demonstrated by the successful completion of the performance assessment.

Common Vocabulary

Term	Definition
Target (Dependent) variable	Attribute selected as Y in a regression equation [aka response attribute, response variable, and effect].
Explanatory (Independent) variables	Attributes selected as X1, X2, …, in a regression equation [aka predictor attributes, predictor variables, explanatory variables, and cause].
Multiple Linear regression	A statistical tool used to model the relationship between a Continuous responses (dependent) variable and one or more continuous and/or categorical explanatory (independent) variables.
Logistic Regression	Statistical Tool to understand the relationship between the Categorical responses (dependent) variable and one or more continuous and/or categorical explanatory (independent) variables to estimate probabilities.
Continuous variable	A type of data that has an infinite number of values between any two values (e.g., temperature). Weight, temperature and length are all examples of continuous data.
Categorical variable	Categorical Data, sometimes called qualitative data, are data whose values describe some characteristic or category. There are two types of categorical data, namely; the nominal and ordinal data.

THE BIG PICTURE FOR D208

Getting Started with D208 (Part I)

Step 1: Research Question
a. Design research question (target variable)
b. Select variables for your model (explanatory variables)

Step 2. Data Preparation and Manipulation
a. Clean the data (check for missing values, duplicates, outliers, etc.)
b. Explore the data (Create visualizations)
c. Wrangle the data (Re-expression of categorical variables)

Step 3: Construct the Regression Model
a. Create initial model
b. Create reduced model
c. Create residual plots (Task 1) / Confusion Matrix (Task II)

Getting Started with D208 (Part II)

Step 4: Model Evaluation
a. Evaluate the initial model
b. Evaluate the reduced model
c. Compare both models (initial and reduced)

Step 5: Discussion and Interpretation of Analysis
a. Create a regression equation.
b. Discuss the coefficients in the regression equation.
c. Discuss Statistical and Practical Significance of the reduced model
d. Discuss limitations of analysis
e. Recommended Course of Action

- **Note: Many of the requirements in D208 are a culmination of D206 and D207.**
- **Note: Before you begin, know the assumptions of Regression (MLR and Logistic)**
- **Note: Consider first, executing the steps above (Steps 1-3) and then writing responses to the PA requirement (including Step 4).**

Step 1: Research Question

Step 1: Research Question

A. Design a Research Question

- The research questions must be relatively broad, and not too narrow.
- **For example, "Does A, B and C cause D?" is too narrow. A much better approach is to ask, "What causes D?“**
- When designing your research question, you are selecting your target variable (Y):

  Task 1: Any logical Continuous Variable
  Task 2: Any logical Categorical Variable

B. Select explanatory variables

- Cast a wide net (i.e., include a good number of explanatory variables) to ensure you find the independent variables that do influence your dependent variable. There is no magic number –but 10 is better than 5 and 15 is better than 10, etc.).
- This requirement uses the word "all" and sometimes causes confusion for students. It does not mean "all predictor variables in the dataset". It means the predictor variables that you think are needed to answer your research question.

NOTE: For Task 1, must have at least one categorical explanatory (predicting) variable for your initial model.

Step 2:
Data Preparation and Manipulation

Step 2: Data Preparation and Manipulations

A. Data Cleaning (D206)

- Detect and Treat Nulls (if any)
- Detect and Treat Outliers (if any)

B. Data Exploration (D207)

- Descriptive Statistics (mean, median, mode)
- Univariate Analysis (visualization) – **every explanatory variables**
- Bivariate Analysis (visualization) – **every explanatory variables**

C. Data Wrangling (D206)

- Re-expression of Categorical Variables
- Logarithmic Transformations **(not required; optional)**

Data Wrangling: *Re-expression of Categorical Variables*

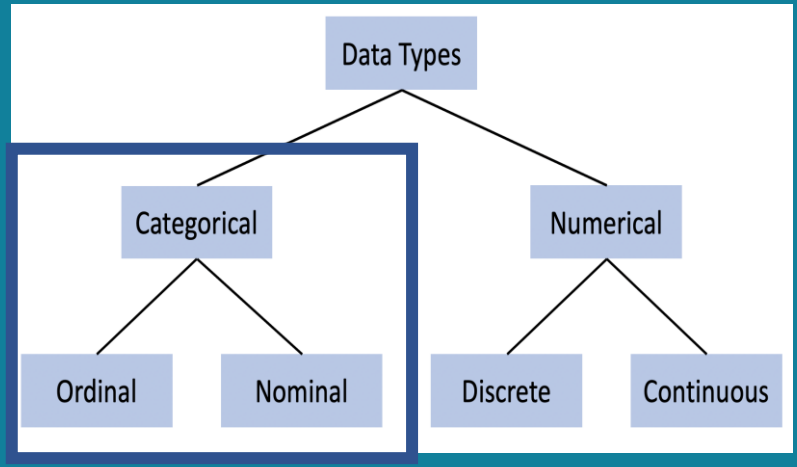
- Most statistical methods/machine learning algorithms mining work exclusively with numeric data.

- This means that if your data contains categorical data, you must encode it to numbers do perform statistical modeling.

- **Categorical Data**, sometimes called qualitative data, are data whose values describe some characteristic or category.

- There are two types of categorical data, namely:

  - **Ordinal Categorical Data** → Inherent Order
  - **Nominal Categorical Data** → Names only; no inherent order



Ordinal Data

- Economic status (poor, middle income, wealthy)
- Course grades (A+, A-, B+, B-, C)
- Education level (Elementary, High School, College, Graduate, Post-graduate)
- Likert scales (Very satisfied, satisfied, neutral, dissatisfied, very dissatisfied)
- Military ranks (Colonel, Brigadier General, Major General, Lieutenant General)
- Age (child, teenager, young adult, middle-aged, retiree)

Note: Some binary variables can be arranged as ordinal categorical variables (Yes/No, True/False)

Method: Ordinal Encoder Technique

- Python: Replace Function
- R Studio: Revalue/Replace Function

Nominal Data

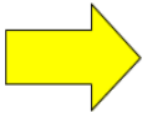
What is your Gender	Your Marital Status
Female	Single
Male	Married
Non-Binary	Divorced

Method: One-hot encoding

- One hot encoding is the most widespread approach, and it works very well unless your categorical variable takes on a large number of values.
- In this approach, for each category of a feature/variable, we create a new column (sometimes called a dummy variables).
- The new column contain a binary encoding (0 or 1) to denote whether a particular row belongs to this category.
- Very common method for nominal data (especially with more than two levels)
- NOTE: Use k-1 (number of columns -1) when adding the variables (columns) to your regression model. This will mitigate multicollinearity.

Interpretation of Dummy Variables https://www.unifyingdatascience.org/html/interpreting_indicator_vars.html

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

- The values in the original data are *Red*, *Yellow* and *Green*.
- We create a separate column for each possible value.
- Wherever the original value was *Red*, we put a 1 in the *Red* column.

Step 3: Construct a Regression Model

Step 3:

Construct the Regression Model and other relevant plots

A. Create the Initial Model

- Task 1: Multiple Linear Regression Model
- Task 2: Logistic Regression Model

B. Create the Reduce Model

- Reduce Model (using a feature selection technique) which will remove explanatory (independent) variables that have little or no influence on the dependent variable.

C. Residual Plots / Confusion Matrix

- Create a residuals plot (Task 1)
- Create a confusion matrix (Task 2)

A. Create an Initial Regression Model

- Create a gross model with all explanatory variables you have selected.
- Use the “Kitchen Sink” approach.
- Include a y-intercept in your regression model.

B. Create a Reduced Regression Model

Create a reduced model (**reducing the number of input variables).**

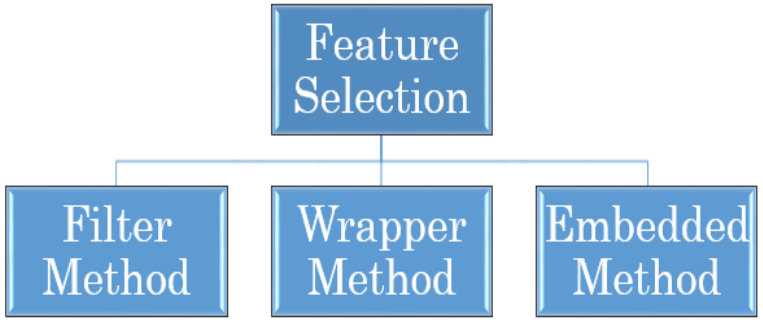
Technique: Variable (Feature) selection

Feature Selection Methods

Feature selection can be done in multiple ways but there three broad categories:

Wrapper Methods:

- Feed the features (variables) for your model and based on the model performance you add/remove the features.
- This is an iterative process and could use computation power (depending on the size of the dataset)



Feature (Variable) Selection Methods

Wrapper Methods

Note: These methods do not remove multicollinearity. Therefore, you must still check for multicollinearity.

SOME COMMON EXAMPLES OF WRAPPER METHODS:

- **Backward Stepwise Elimination:** In backward elimination, we start with all the features and removes the least significant feature (based on p-value) at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features (e.g., no variables “features” greater than .05).

- **Recursive Feature Elimination:** Aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It uses accuracy metric to rank the feature according to their importance.

python

OLS Regression Results

Dep. Variable:	Y	**R-squared:**	0.732
Model:	OLS	**Adj. R-squared:**	0.730
Method:	Least Squares	**F-statistic:**	387.9
Date:	Thu, 15 Aug 2019	**Prob (F-statistic):**	2.96e-200
Time:	18:42:03	**Log-Likelihood:**	-147.90
No. Observations:	716	**AIC:**	307.8
Df Residuals:	710	**BIC:**	335.2
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.6957	0.046	-15.250	0.000	-0.785	-0.606
X1	0.1814	0.009	19.220	0.000	0.163	0.200
X2	0.1845	0.009	19.983	0.000	0.166	0.203
X3	0.1702	0.009	18.200	0.000	0.152	0.189
X4	0.1913	0.009	20.961	0.000	0.173	0.209
X5	0.1954	0.009	21.547	0.000	0.178	0.213

Omnibus:	5.569	**Durbin-Watson:**	1.845
Prob(Omnibus):	0.062	**Jarque-Bera (JB):**	4.274
Skew:	-0.069	**Prob(JB):**	0.118
Kurtosis:	2.648	**Cond. No.**	20.7

R

```
Call:
lm(formula = heart.disease ~ biking + smoking, data = heart.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1789 -0.4463  0.0362  0.4422  1.9331

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.984658   0.080137  186.99   <2e-16 ***
biking      -0.200133   0.001366 -146.53   <2e-16 ***
smoking      0.178334   0.003539   50.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.654 on 495 degrees of freedom
Multiple R-squared:  0.9796,    Adjusted R-squared:  0.9795
F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

C. Create Residual Plots (Task I)

Task 1: Residual Plot

- Residuals are obtained by subtracting the **observed (actual) responses** from the **predicted responses**.
- Residuals are used to are used to visually check that your data is a) a good fit for the model and b) does your data/model meet the assumptions of normality and homoscedasticity.

a. **A scatterplot of the residuals versus the predicted(fitted) values of Y. (Note: Residuals are along the Y axis and predicted(fitted values) are along the X-axis.**
This is helpful to visually check for the assumption of homoscedasticity.

a. **Q-Q plot or histogram of the residuals**
This is helpful to visually check for the assumption normality of residuals.

b. **Calculate the residual standard error**
Used to measures the standard deviation of the residuals in a regression model. It can also determine how well a regression model fits a dataset. The smaller the residual standard error, the better a regression model fits a dataset.

Resources: Python	Resources: R
• https://www.statology.org/residuals/)	• https://www.statology.org/residual-plot-r/ • https://www.statology.org/how-to-interpret-residual-standard-error/#:~:text=The%20residual%20standard%20error%20is,residuals%20in%20a%20regression%20model.

THE BIG PICTURE FOR D208

"Getting Started with D208" (Part I)

Step 1: Research Question

a. Design research question (target variable)
b. Select variables for your model (explanatory variables)

Step 2. Data Preparation and Manipulation

a. Clean the data (check for missing values, duplicates, outliers, etc.)
b. Explore the data (Create visualizations)
c. Wrangle the data (Re-expression of categorical variables)

Step 3: Construct the Regression Model

a. Create initial model
b. Create reduced model
c. Create residual plots (Task 1) / Confusion Matrix (Task II)

"Getting Started with D208" (Part II)

Step 4: Model Evaluation

a. Evaluate the initial model
b. Evaluate the reduced model
c. Compare both models (initial and reduced)

Step 5: Discussion and Interpretation of Analysis

a. Create a regression equation.
b. Discuss the coefficients in the regression equation.
c. Discuss Statistical and Practical Significance of the reduced model
d. Discuss limitations of analysis
e. Recommended Course of Action

- **Note: Many of the requirements in D208 are a culmination of D206 and D207.**
- **Note: Before you begin, know the assumptions of Regression (MLR and Logistic)**
- **Note: Consider first, executing the steps above (Steps 1-3) and then writing responses to the PA requirement (including Step 4).**

The Performance Assessment

Overview of the Performance Assessment

1. Written Report (Addressing ALL the requirements)

- Word documents highly recommended.
- Use headers (helps with reviewing the assessment).
- Professional Communication: APA Format, References and Free from Grammatical Errors].
- Include Visualizations

2. Cleaned "Prepared" Dataset

- Extract the cleaned data from the R or Python environment.

3. Panopto Video

- a demonstration of the functionality of the code used for the analysis
- an identification of the version of the programming environment
- a comparison of the **two** multiple regression models you used in your analysis
- an interpretation of the coefficients.

4. Submit a copy of all your code (very highly recommended) several components of your code will be required in the submission of your report.

Note: .ipynb and R files are accepted.

Performance Assessment Requirement Area	Helpful Tips and Reminders
Part I: Research Question	A1. Provide a research question that would be relevant to a real-world organizational situation. Remember, you need to select the appropriate target variable. A2. Define the objectives or goals of the data analysis.
Part II: Method Justification	B1. Summarize all the assumptions related to (Task I: Multiple Linear Regression and Task II: Logistic Regression). B2. Describe the programming language you have decided to use and why are you using this tool for analysis. B3. Explain why multiple regression an appropriate technique is to analyze the research question summarized in Part I.
Part III: Data Preparation	C1. Describe your data preparation and the data manipulations goals. Remember, your goals are high level as discussed with the Getting Started with D208 Part I. C2. As a customary step of exploratory data analysis, perform summary stats on the variables that you have selected for your initial model. Provide the output of your summary statistics and a discussion of the output. Assume that the reader of your assessment is unaware of summary stats (mean, median, mode, etc.) C3. Explain the steps used to manipulate and prepare the data for the analysis. These steps should reflect the goals referenced in C1. Also, be certain to include the including the annotated code. C4. Generate univariate and bivariate visualizations for all variables used in your initial model (at minimum). Include the target variable in your bivariate visualizations. Remember for Task I, you must have at least one categorical variable as a predicting variable. C5. Provide a copy of the cleaned and manipulated dataset. This can be exported in a .csv file.
Part IV: Model, Comparison and Evaluation	D1. Construct an initial regression model from all predictor variables you have selected. Ensure that at least one categorical variable is in your initial model. D2. State and justify the statistically based variable selection procedure and a model evaluation metric that you will use to reduce the model. Ensure that your justification is in alignment with the research question. D3. Provide the output of the reduced regression model.