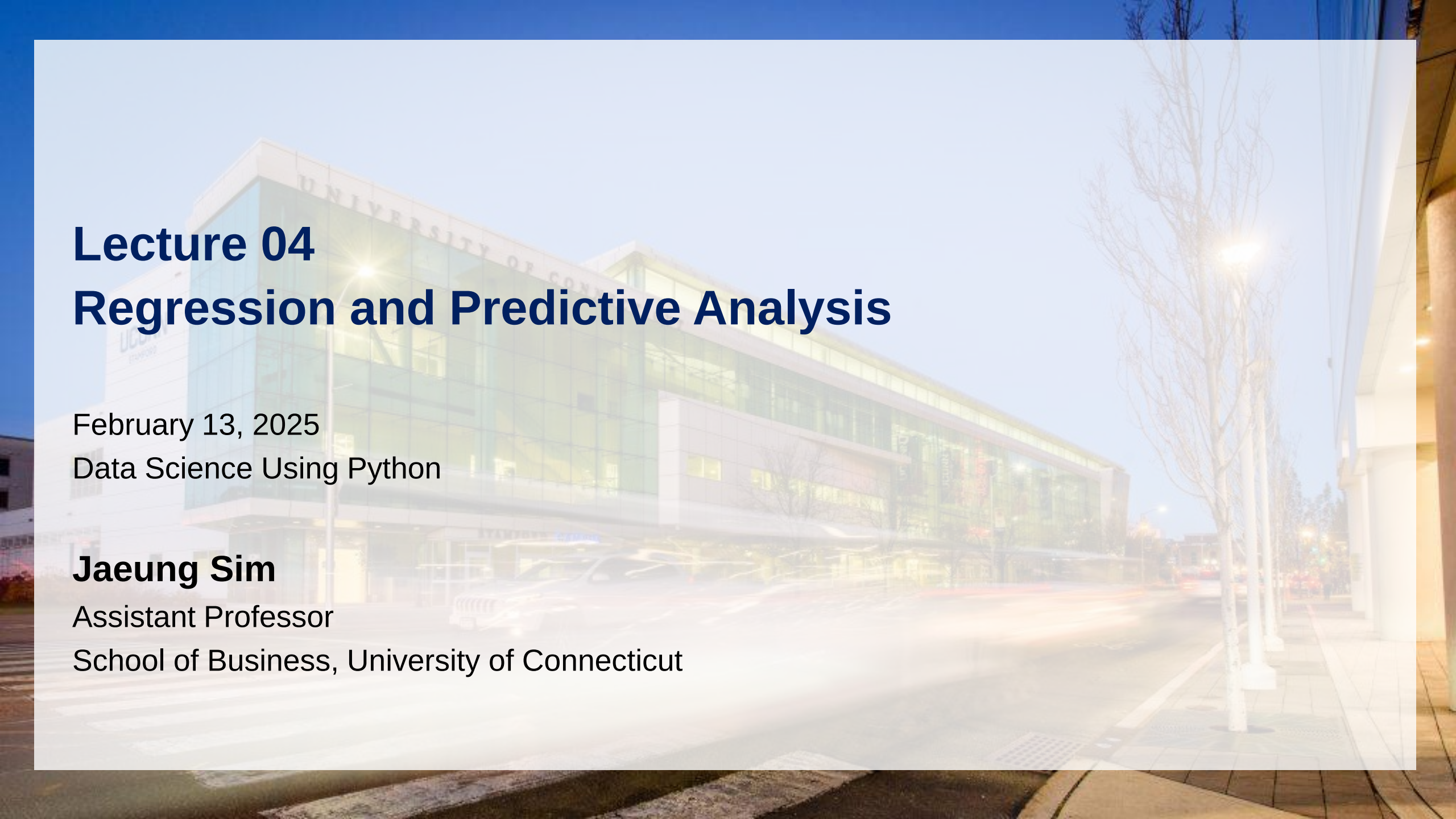


Lecture 04
Regression and Predictive Analysis

February 13, 2025

Data Science Using Python

Jaeung Sim

Assistant Professor

School of Business, University of Connecticut

Agenda

- **Python Hands-on for Lecture 02**
- **Basic Regression Models**
- **Advanced Models**
- **Data Partition and Evaluation**
- **Announcements**

Python Hands-on for Lecture 02

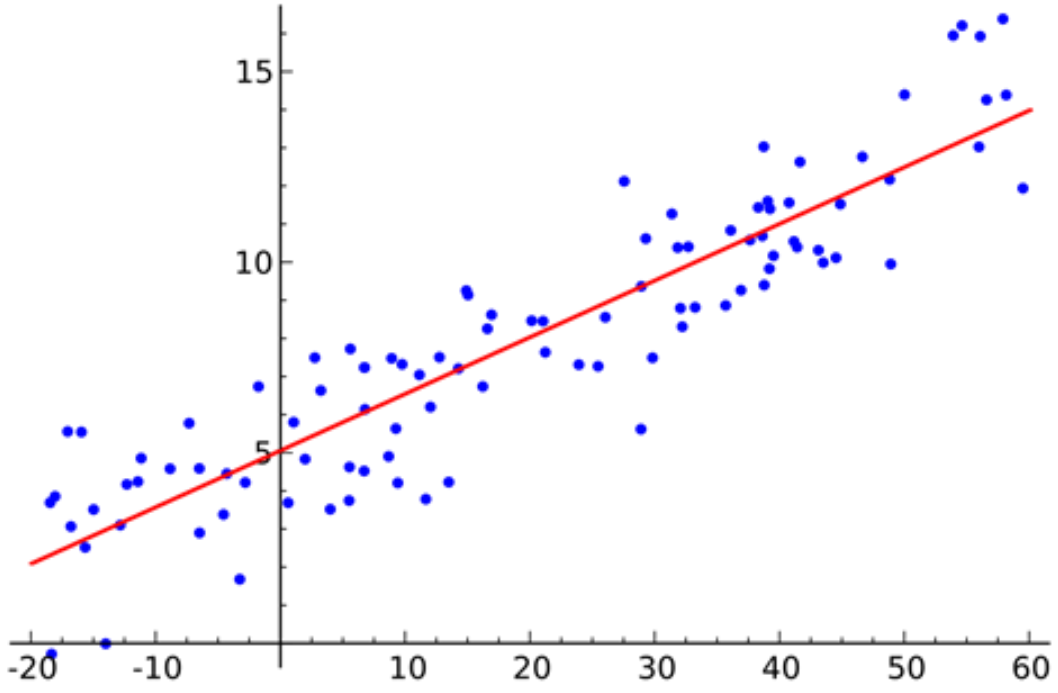
Pandas and Visualization

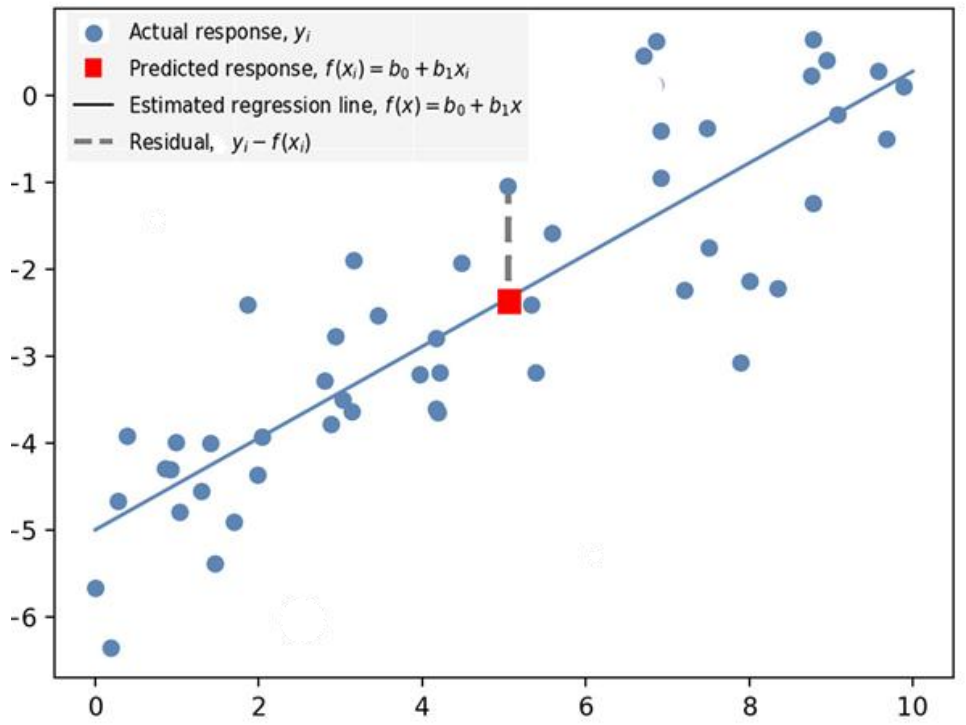
Basic Regression Models

Linear Regression, Logistic Regression, Poisson Regression

Linear Regression

- **Verbal Description**
  - Linear regression assumes linear relationships between a dependent and independent variables and attempts to fit a linear equation to observed data.





Linear Regression

- **Mathematical Form**

For observation $i \in \{1, \dots, n\}$, we can express the relationship between y_i and $x_{1i}, x_{2i}, \dots, x_{pi}$ as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i,$$

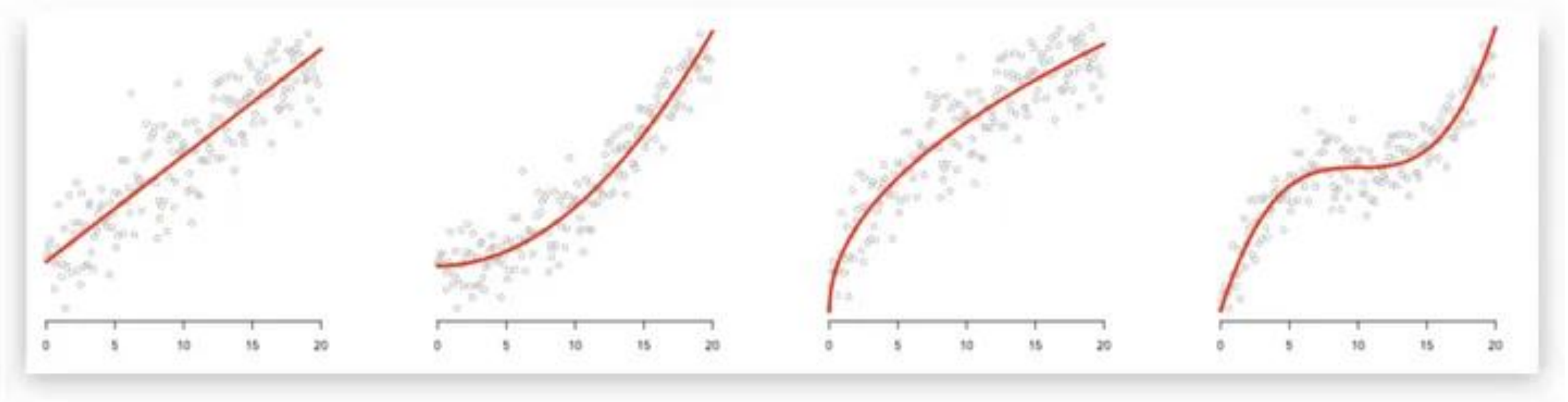
where ε_i is an error term. By defining a few vectors, we can simplify the equation in matrix notation as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$

Linear Regression

- **What does a linear relationship actually mean?**



Linear Regression

- **What does a linear relationship actually mean?**

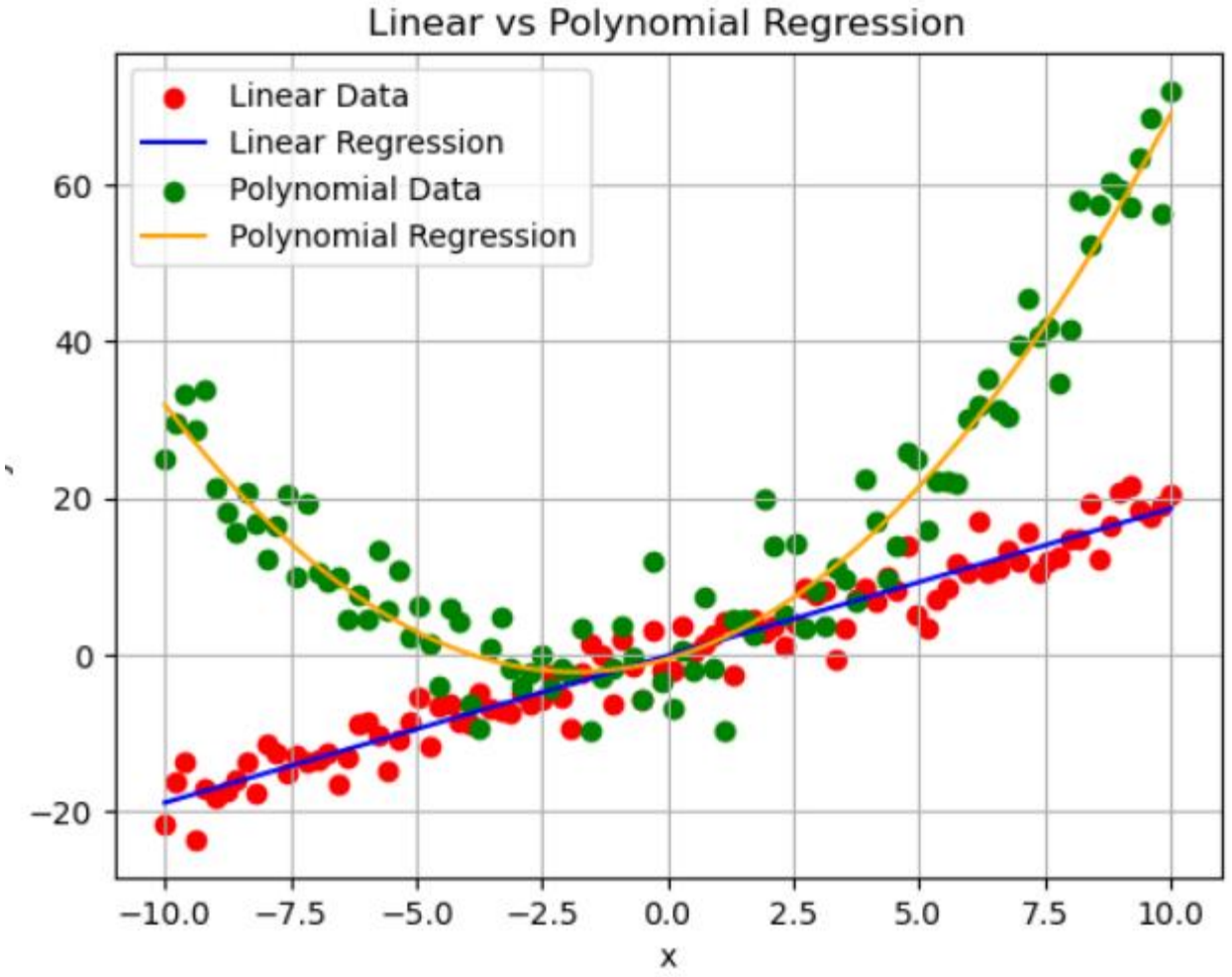
  If some function of x (i.e., $f(x)$) has a linear relationship with a function y (i.e., $g(y)$), then the relationship between x and y can be estimated with a linear regression model.

  For example, the following relationships can be fitted to this model:

  - Polynomial relationships: $y = a_2 x^2, y = a_3 x^3, \ldots, y = a_n x^n$
  - Multiplicative relationships: $y = e^{a_1 x_1 + a_2 x_2} \Leftrightarrow ln(y) = a_1 x_1 + a_2 x_2$

Linear Regression

- **What does a linear relationship actually mean?**



Logistic Regression

- **Verbal Description**
  - The logistic regression models the probability of an event taking place by estimating its log-odds as a linear function of independent variables.
- **Mathematical Form**

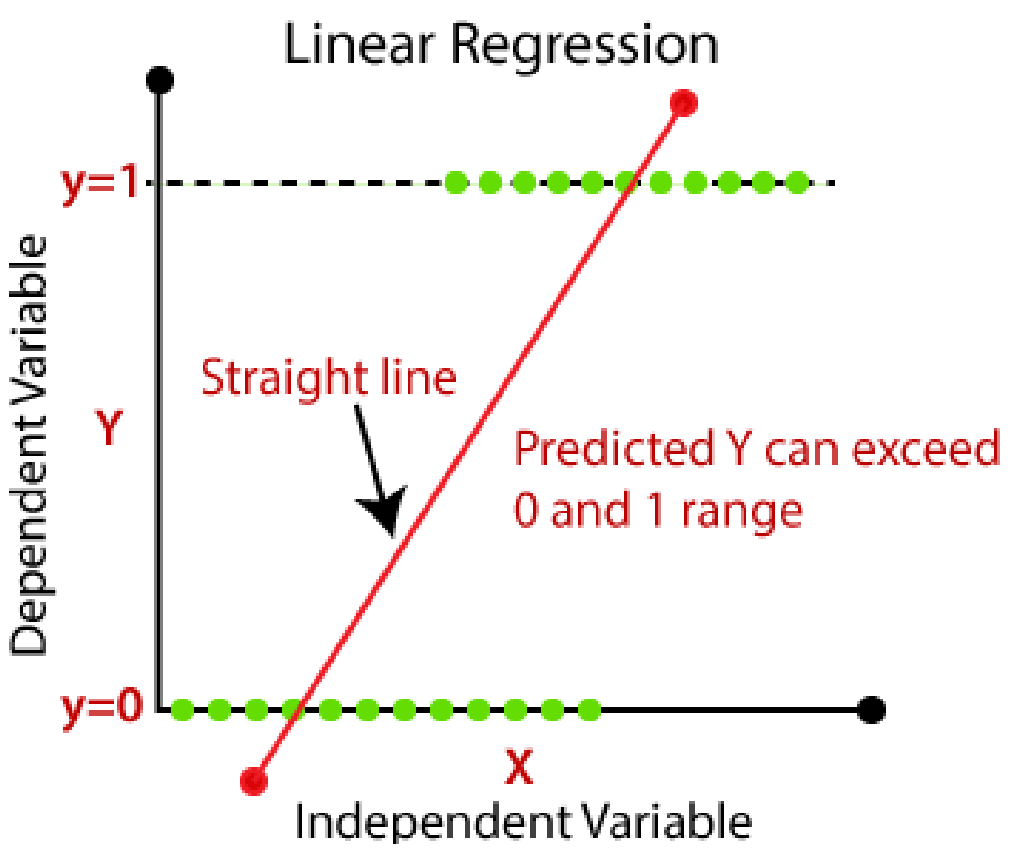
For observation $i \in \{1, \ldots, n\}$, we can express the relationship between $P(y_i = 1)$ (hereafter, P) and $x_{1i}, x_{2i}, \ldots, x_{pi}$ as:

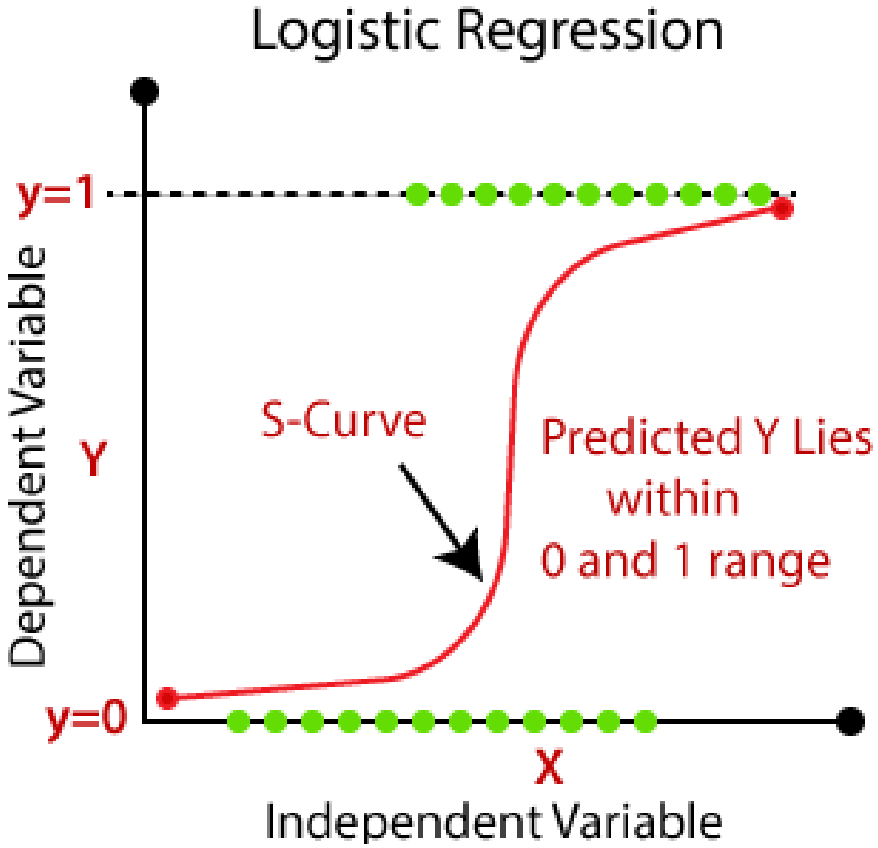
$$P = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_p x_{pi}}} \Leftrightarrow ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}.$$

where $\frac{P}{1-P}$ is the odds of P and $ln(\frac{P}{1-P})$ is the logit function.

Logistic Regression

- **Linear Regression vs. Logistic Regression**





Logistic Regression

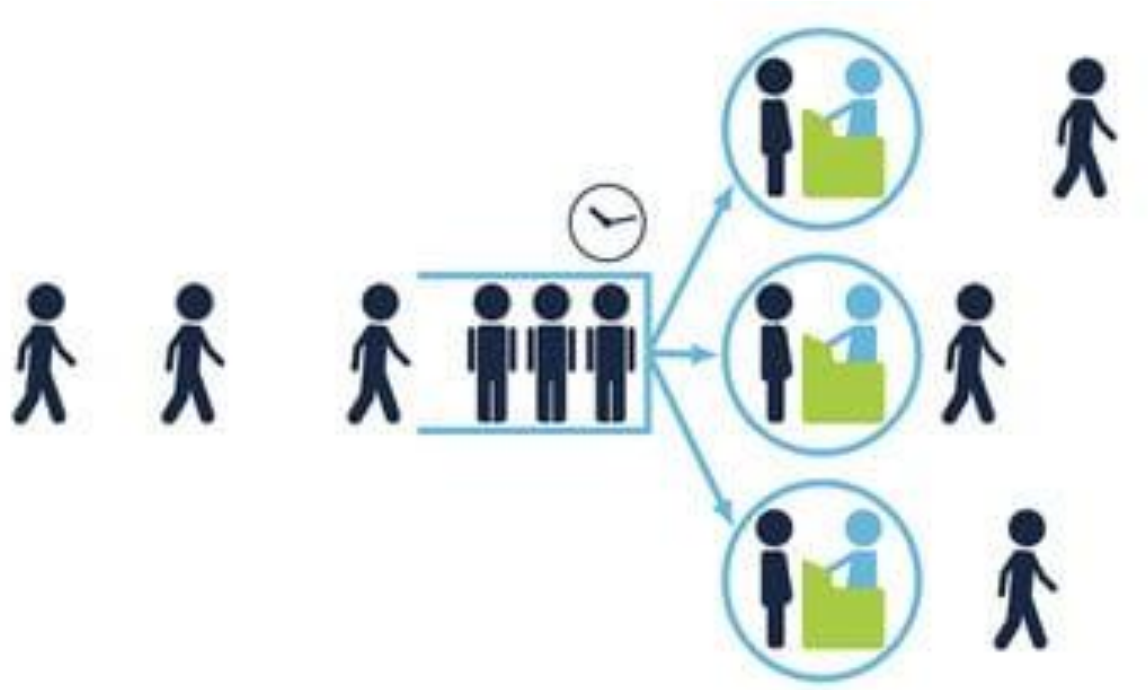
- **Linear Regression vs. Logistic Regression**
  - Consider a dependent variable y having either 0 or 1 as its value.
  - **Linear regression**
    - Assumes a linear relationship between x and $P(y = 1)$ in percentage points (%p).
    - Generates unbiased estimates with heteroskedasticity-robust standard errors.
    - Predicted values might appear outside [0, 1].
  - **Logistic regression**
    - Assumes a linear relationship between $\ln\left(\frac{P(y=1)}{1-P(y=1)}\right)$ and x.
    - Odds and x have an exponential relationship, providing an odds-based interpretation (i.e., odds are N times larger).
    - When P is sufficiently small, you can approximate it to N times larger probability.
    - Predicted values appear within (0, 1).

Logistic Regression

- **Linear Regression vs. Logistic Regression**
  - Why can’t you model the logit function with linear regression?
    - You see the linear relationship between a logit function of $P(y_i)$ and covariates as:
    - $logit(P) = ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$
    - But the thing is, you only observe whether an event occurred or not, instead of its probability. Since the observed values are either 0 or 1, the logit function cannot be defined (i.e., $ln(0)$ and $ln(\infty)$).

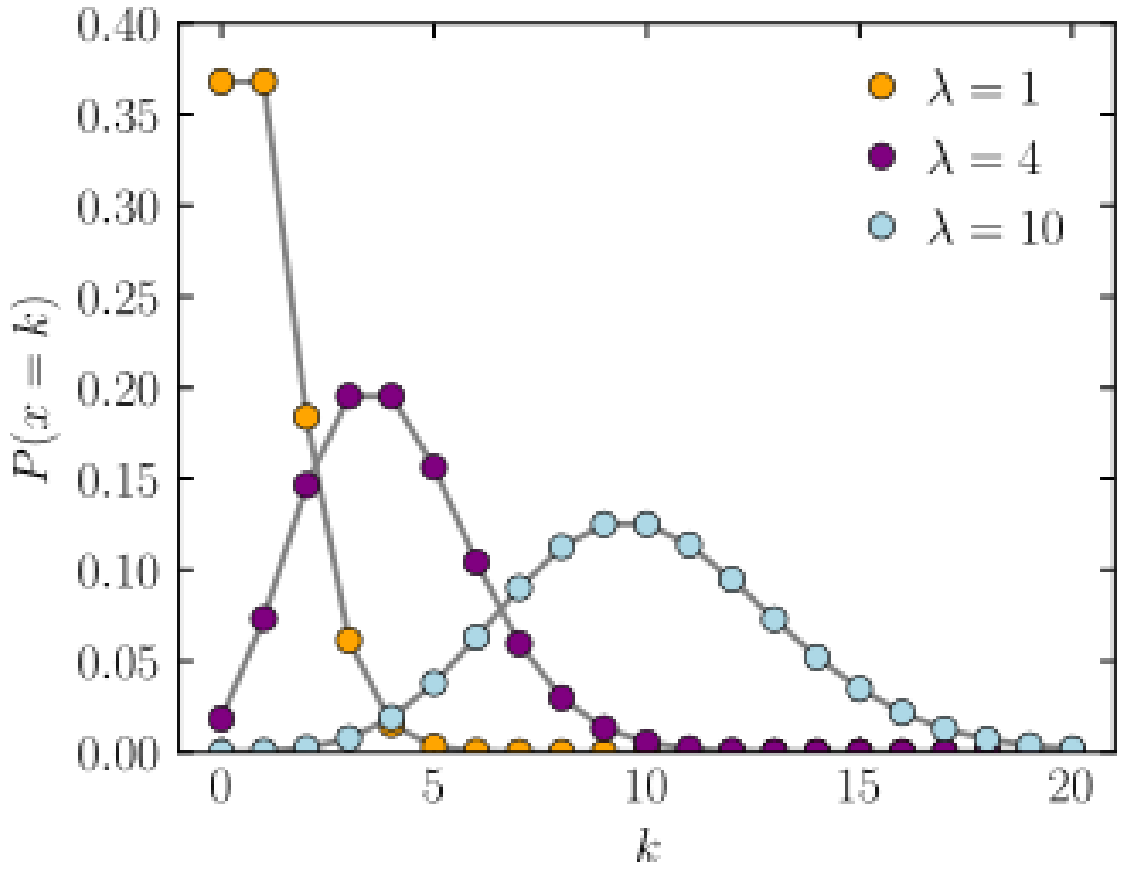
Poisson Regression

- **Verbal Description**
  - Poisson regression is a generalized linear model form of regression analysis used to model **count data**, which assumes that a dependent variable follows a Poisson distribution.



Poisson Regression

- **Mathematical Form**
  - **Poisson Distribution**



Poisson Regression

- **Mathematical Form**
  - **Poisson Distribution**

    A discrete random variable X has the following probability mass function with parameter $\lambda > 0$:

    $$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

    where k is the number of occurrences ($k = 0, 1, 2, \dots$).

    In the Poission distribution, $\lambda = E(X) = Var(X)$.

Poisson Regression

- **Mathematical Form**
  - **Regression**

For observation $i \in \{1, \ldots, n\}$, we can express the relationship between $P(y_i = 1)$ (hereafter, P) and $x_{1i}, x_{2i}, \ldots, x_{pi}$ as:

$$ln(E(y_i|x_{1i}, x_{2i}, \ldots, x_{pi})) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}.$$

Define $\lambda \equiv E(y_i|x_{1i}, x_{2i}, \ldots, x_{pi}) = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}}$, then we can express the Poisson distribution's probability mass function as:

$$P(y|x_{1i}, x_{2i}, \ldots, x_{pi}; \beta_0, \beta_1, \ldots, \beta_p) = \frac{\lambda^y}{y!} e^{-\lambda}.$$

Poisson Regression

- **What is the good of Poisson regression?**
  - In handling a count dependent variable, Poisson regression can be more appropriate than linear regression.
  - It is well known to model the arrival of service requests effectively.

- **What should we be cautious about Poisson regression?**
  - Poisson regression may not provide plausible interpretations on continuous variables.
  - Also, the equal mean and variance assumption is very strong, calling for more advanced approaches, such as negative binomial regression and zero-inflated Poisson regression.

Advanced Models

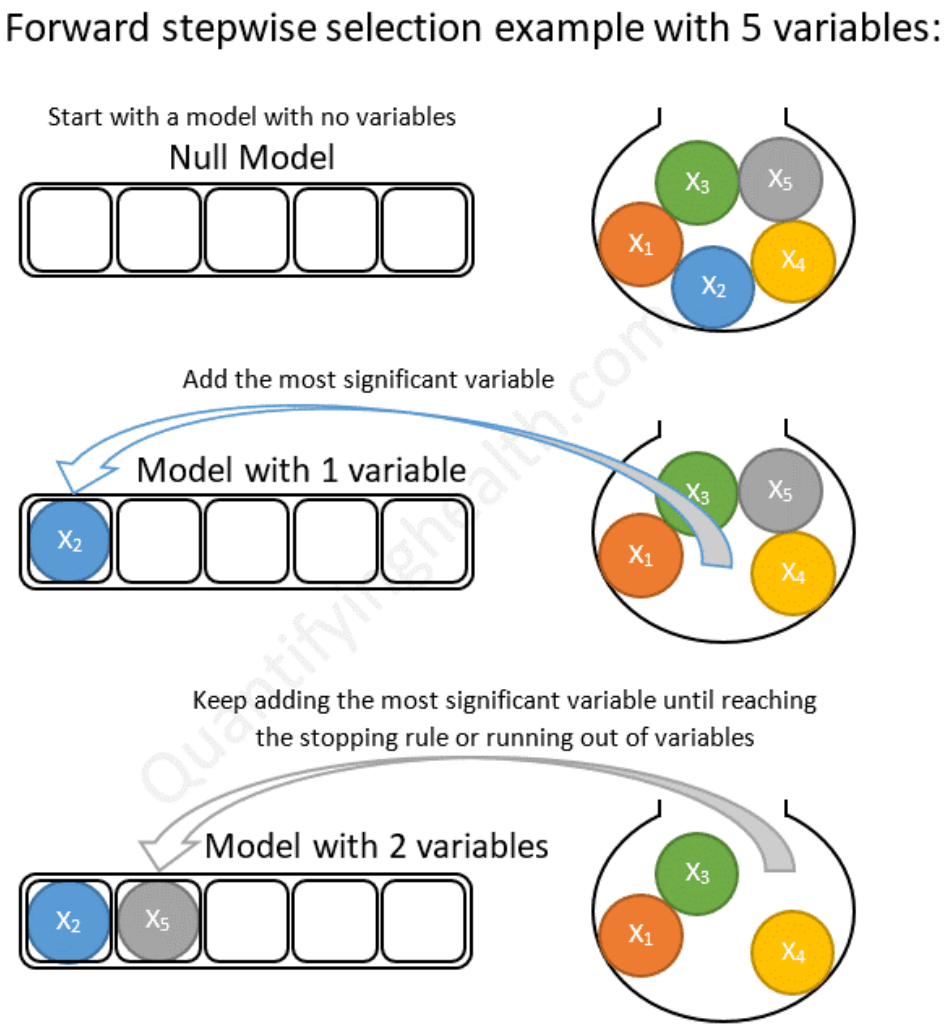
Stepwise Regression, Multinomial Logistic Regression, Negative Binomial Regression

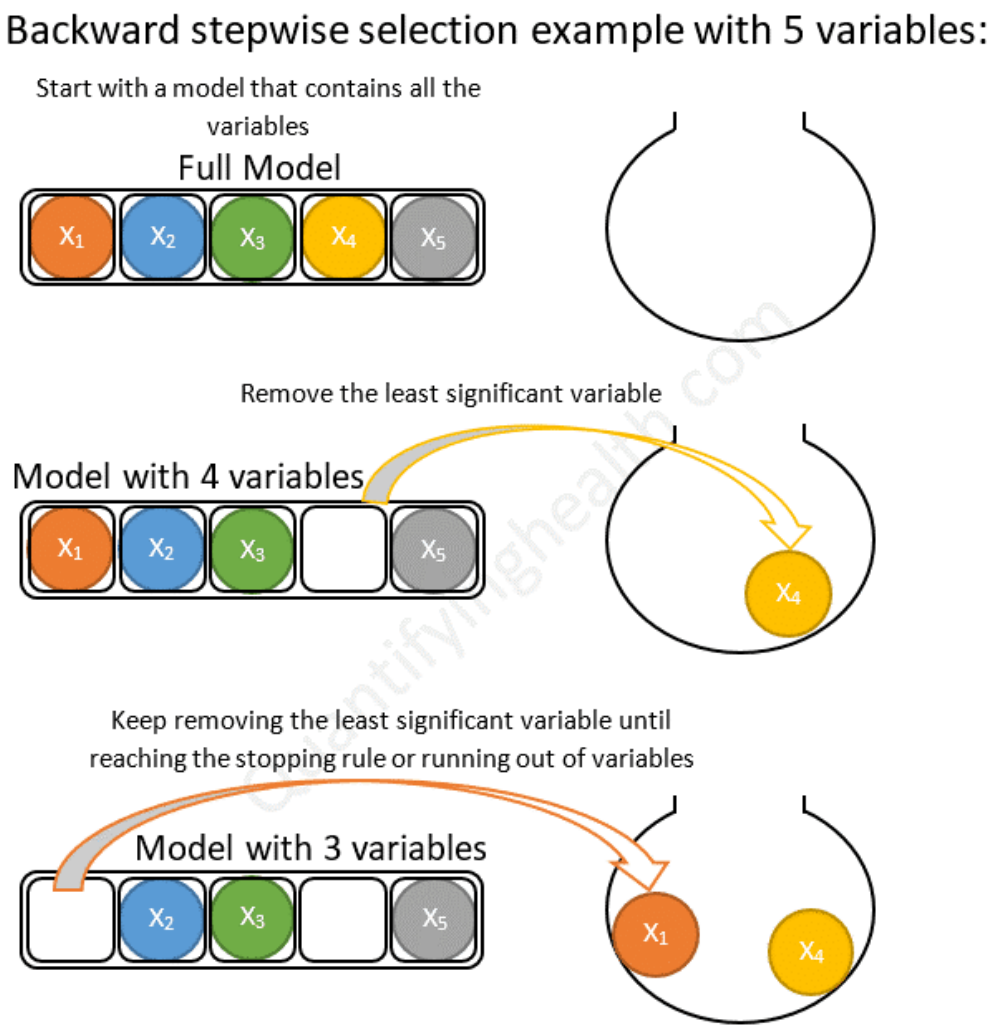
Stepwise Regression

- **Verbal Description**
  - Stepwise regression is a method of fitting a regression model by iteratively adding or removing variables. It is used to build a model that is accurate and parsimonious, meaning that it has the smallest number of variables that can explain the data.
- **Implementation Process**
  - Forward Selection
    - In forward selection, the algorithm starts with an empty model and iteratively adds variables to the model until no further improvement is made.
  - Backward Elimination
    - In backward elimination, the algorithm starts with a model that includes all variables and iteratively removes variables until no further improvement is made.

Stepwise Regression

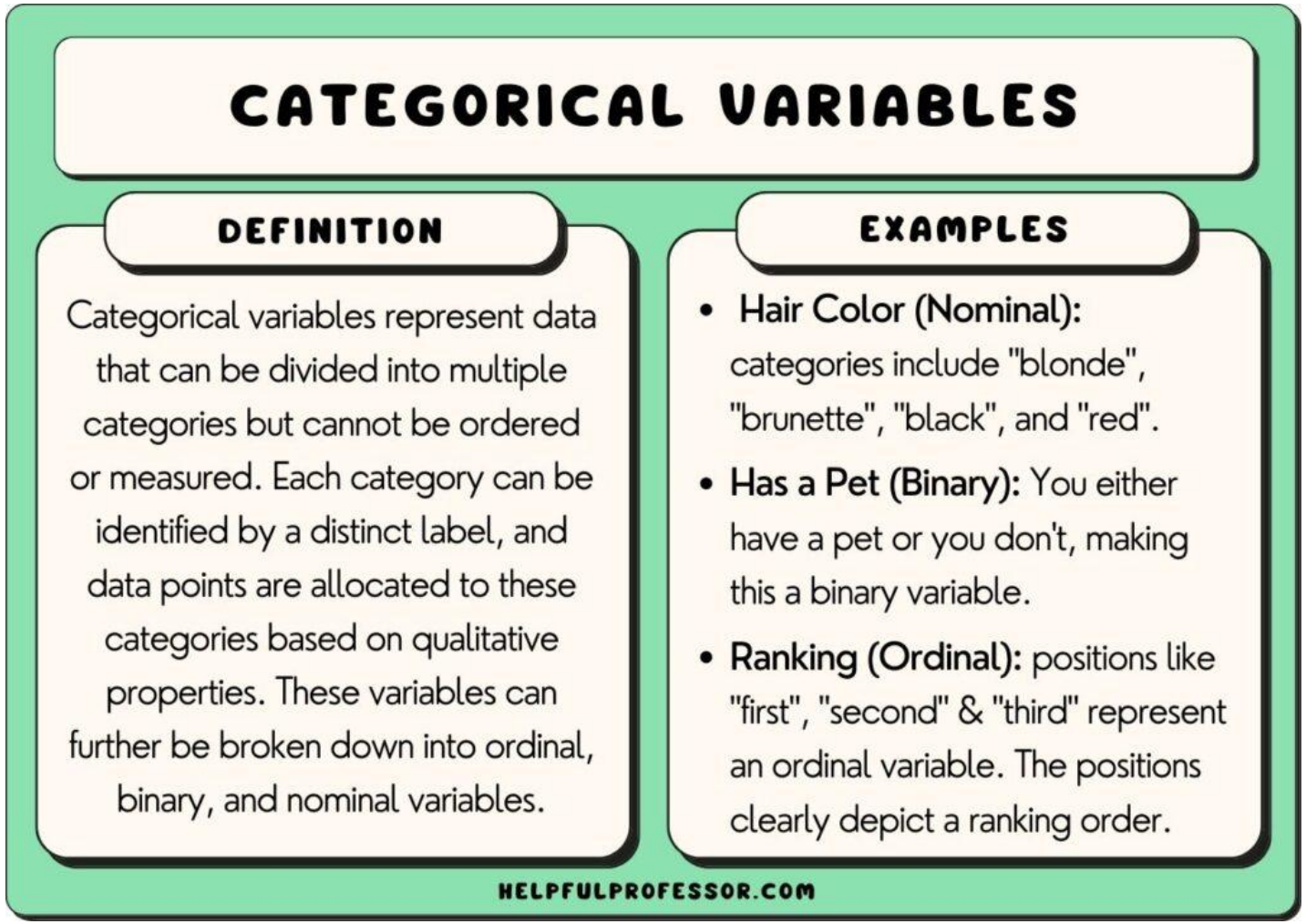
- **Implementation Process**





Multinomial Logistic Regression

- **Nominal (or categorical) outcome variables**



Multinomial Logistic Regression

- **Reminder: Binary Logistic Regression**
  - **Mathematical Form**

For observation $i \in \{1, \ldots, n\}$, we can express the relationship between $P(y_i = 1)$ (hereafter, P) and $x_{1i}, x_{2i}, \ldots, x_{pi}$ as:

$$P = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_p x_{pi}}} \Leftrightarrow ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}.$$

where $\frac{P}{1-P}$ is the odds of P and $ln(\frac{P}{1-P})$ is the logit function.

- To simplify, let's define $\mathbf{X\beta} \equiv \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$
- Then, you can express $P = \frac{e^{\mathbf{X\beta}}}{1+e^{\mathbf{X\beta}}}$

Multinomial Logistic Regression

- **Deriving the Multinomial Model**

For multinomial logistic regression, all you need to do is to consider the probabilities of other categories like this:

$$
\begin{aligned}
ln\frac{P(y_i=1)}{P(y_i=K)} &= \boldsymbol{X_i\beta_1}. \\
ln\frac{P(y_i=2)}{P(y_i=K)} &= \boldsymbol{X_i\beta_2}. \\
&\vdots \\
ln\frac{P(y_i=K-1)}{P(y_i=K)} &= \boldsymbol{X_i\beta_{K-1}}
\end{aligned}
$$

In other words, you consider the event $y_i = K$ the baseline and quantify how much the covariates $\boldsymbol{X_i}$ increases the logit $P(y_i = k)$. We can express this equation as follows:

$$
P(Y_i = k) = \frac{e^{\boldsymbol{X_i\beta_k}}}{1+\sum_{j=1}^{K-1} e^{\boldsymbol{X_i\beta_j}}}.
$$

Ordered Logistic Regression

- **Definition**
  - Ordered logistic regression (also called ordinal logistic regression) is a statistical model used for predicting an ordinal dependent variable, meaning a categorical variable where the categories have a meaningful order but **unequal or unknown spacing** between them.
- **Example of Ordinal Outcomes**
  - Survey Ratings: ("Poor", "Fair", "Good", "Very Good", "Excellent")
  - Educational Levels: ("High School", "Bachelor's", "Master's", "PhD")
  - Customer Satisfaction: ("Dissatisfied", "Neutral", "Satisfied")
- **Functional Form**
  - $P(Y \leq j|X) = \frac{1}{1+e^{-(\alpha_j - X\beta)}}$

Negative Binomial Regression

- **Verbal Description**
  - Negative binomial regression is for modeling count variables, usually for over-dispersed count outcome variables. Specifically, when the mean of the count is lesser than the variance of the count, then Negative binomial regression is used to test for connections between confounding and predictor variables on a count outcome variable.
- **Mathematical Form**

The probability mass function of the negative binomial distribution is as follows:

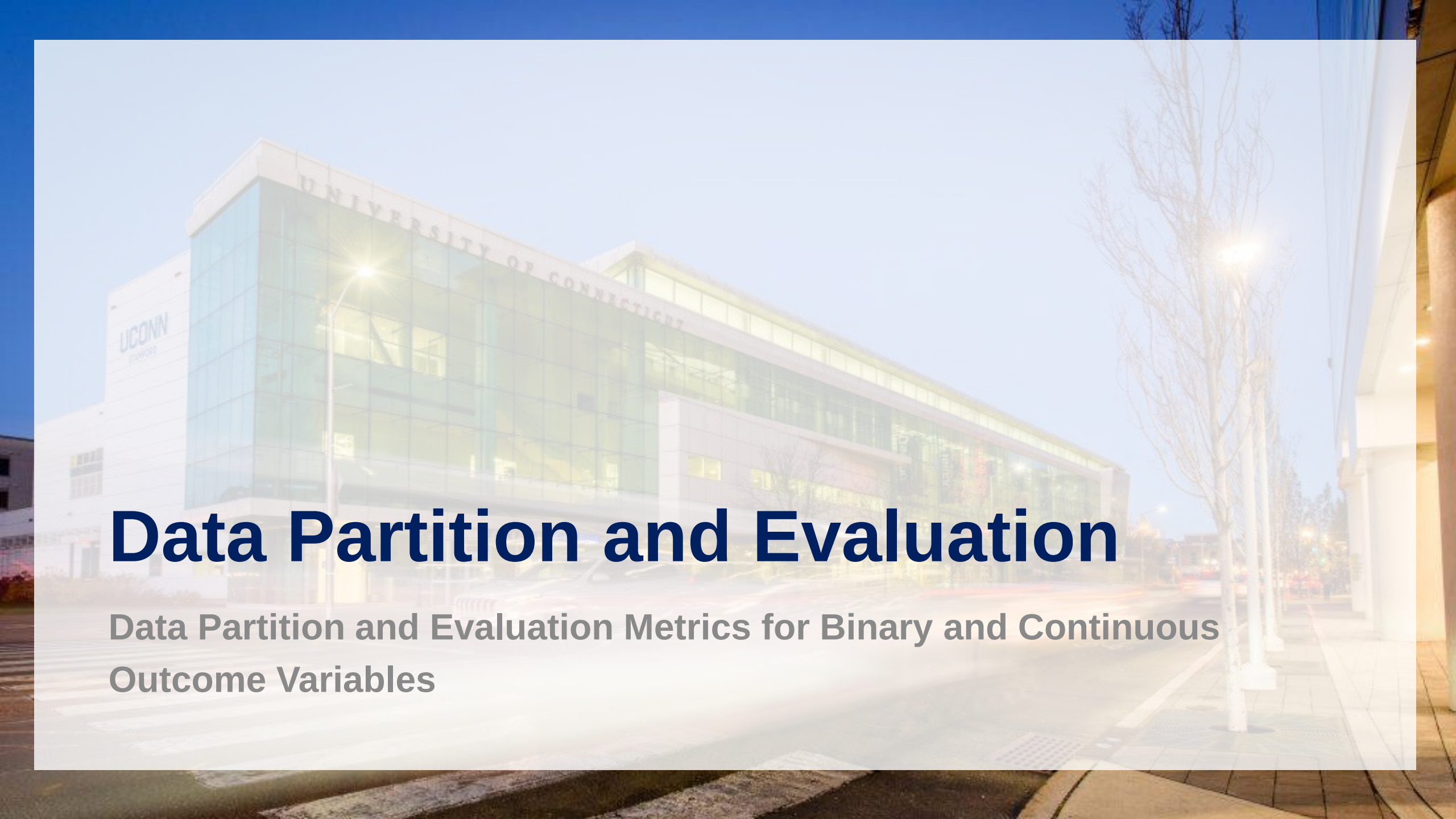
$$f(k;r,p) \equiv P(X=k) = \binom{k+r-1}{k}(1-p)^k p^r,$$

and its mean and variance are obtained as $\frac{r(1-p)}{p}$ and $\frac{r(1-p)}{p^2}$, respectively.

Given that $p < 1$ and $\frac{r(1-p)}{p} < \frac{r(1-p)}{p^2}$, this model can address the over-dispersion problem of Poisson regression.

Negative Binomial Regression

- **Poisson vs. Negative Binomial**
  - Poisson regression assumes that the mean and the variance of the outcome variable are equal; that is, $\lambda = \mathrm{E}[\mathrm{Y}] = \mathrm{Var}[\mathrm{Y}]$.
  - Negative binomial model incorporates an additional term to account for higher variance.
  - Negative binomial regression performs better than Poisson regression only if your outcome variable is over-dispersed because:
    - It uses additional parameters, perhaps some of them can be redundant.
    - For every $p < 1$, variance is determined to be larger than mean.
  - Some statistical packages provide diagnostic statistics for additional parameters.



Data Partition and Evaluation

Data Partition and Evaluation Metrics for Binary and Continuous Outcome Variables

Data Partition and Cross-validation

- **Motivation**
  - A model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data (i.e., **overfitting**).
- **Data Partitioning**
  - **Training Set:** This is the largest portion of the data, used to train the model. The model learns to make predictions from this dataset.
  - **Validation Set (Optional):** It serves as an interim test set to tune the model's hyperparameters. It helps in the model selection process without touching the test set.
  - **Test Set:** This portion of the data is held back and used only after the model has been trained and hyperparameters have been chosen. It serves to evaluate the model's final performance, simulating how the model would perform on unseen data.

Data Partition and Cross-validation

- **Cross-validation**
  - Involves systematically creating and evaluating multiple models on different subsets of the dataset and aims to mitigate the risk of the model's performance being dependent on the way the data was split.
- **k-fold cross-validation**
  - **Divide the data into k subsets (or folds):** For example, with 5-fold cross-validation, the data is divided into 5 subsets.
  - **Iteratively train and evaluate k models:** In each iteration, a different fold is held back as the test set, and the remaining k-1 folds are used for training. This process repeats k times, with each fold used exactly once as the test set.
  - **Aggregate the results:** The performance measure (e.g., accuracy) from each of the k models is averaged to get a single performance metric.

Confusion Matrix and Evaluation Metrics

- **Confusion Matrix**

		Predicted	
		Negative **(N)** **-**	Positive **(P)** **+**
Actual	Negative **-**	True Negative **(TN)**	**False Positive (FP)** **Type I Error**
	Positive **+**	**False Negative (FN)** **Type II Error**	True Positive **(TP)**

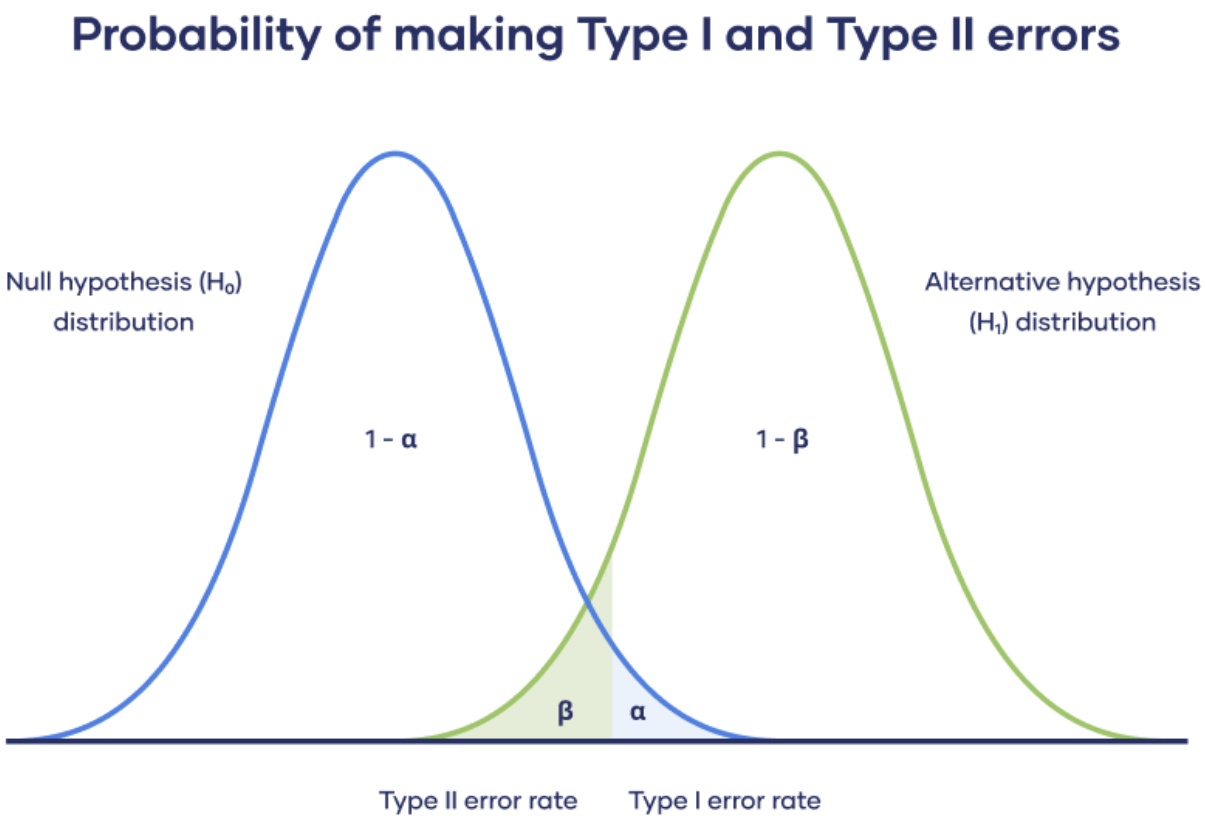
Confusion Matrix and Evaluation Metrics

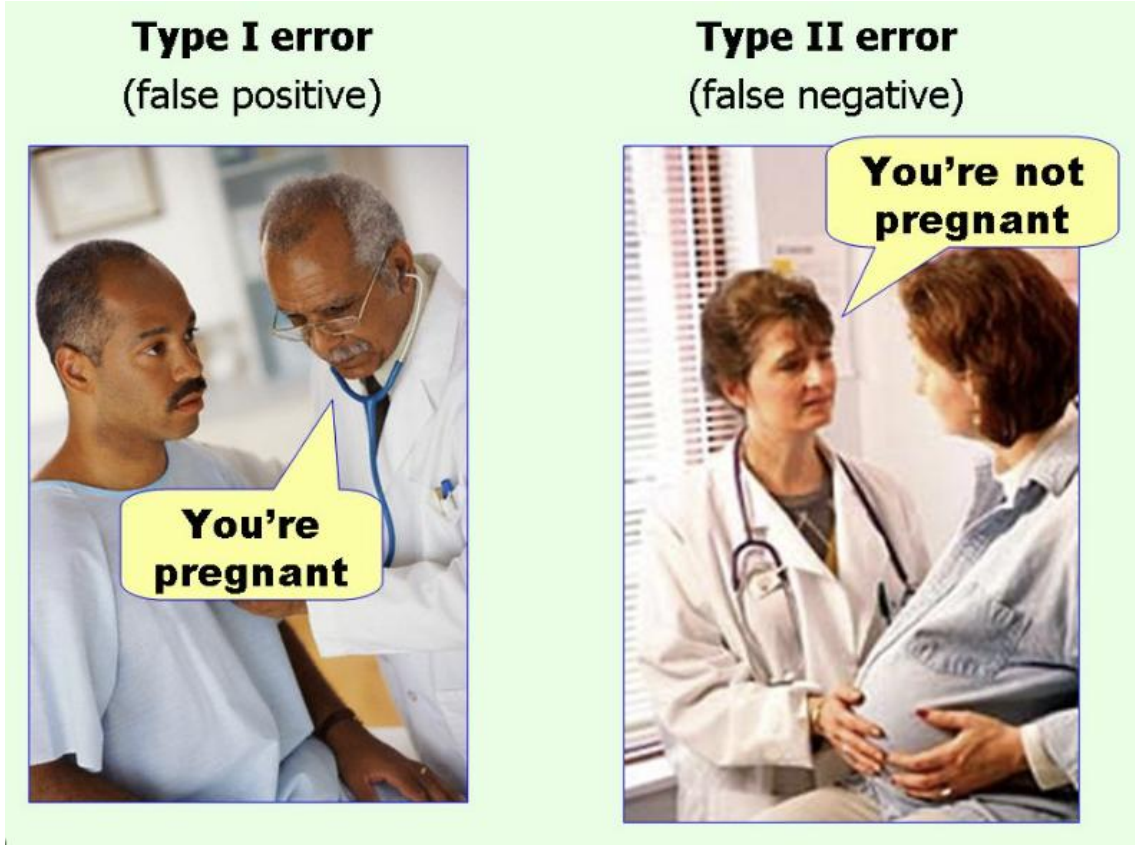
- **Confusion Matrix**
  - **True Positives (TP):** The count of instances that were positive and correctly predicted as positive by the model.
  - **False Positives (FP):** The count of instances that were negative but incorrectly predicted as positive (Type I error).
  - **True Negatives (TN):** The count of instances that were negative and correctly predicted as negative by the model.
  - **False Negatives (FN):** The count of instances that were positive but incorrectly predicted as negative by the model (Type II error).

	Negative **(N)** -	Positive **(P)** +
Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

Confusion Matrix and Evaluation Metric

- **Confusion Matrix**





Confusion Matrix and Evaluation Metric

- **Classification Measures**
  - **Accuracy** $(= \frac{TP+TN}{P+N})$

    > Accuracy simply measures how often the classifier makes the correct prediction. It's the ratio between the number of correct predictions and the total number of predictions. The accuracy metric is not suited for imbalanced classes.

  - **Precision** $(= \frac{TP}{PP})$

    > It is a measure of correctness that is achieved in true prediction. In simple words, it tells us how many predictions are actually positive out of all the total positive predicted.

Confusion Matrix and Evaluation Metric

- **Classification Measures**
  - **Recall** $(= \frac{TP}{P})$

    > It is a measure of actual observations which are predicted correctly, i.e. how many observations of positive class are actually predicted as positive. It is also known as sensitivity. Recall is a valid choice of evaluation metric when we want to capture as many positives as possible.

  - **F1-Score** $(= 2\frac{Precision \times Recall}{Precision + Recall})$

    > The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. We use harmonic mean because it is not sensitive to extremely large values, unlike simple averages.
    >
    > F1 score sort of maintains a balance between the precision and recall for your classifier. If your precision is low, the F1 is low and if the recall is low again your F1 score is low.

Continuous Outcomes and Evaluation Metrics

- **R-squared score, the coefficient of determination**

> R-squared (or R^2) represents the proportion of variance (of y) that has been explained by the independent variables in the model. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance.
>
> Formally, the estimated R^2 is defined as:
>
> $$R^2(y, \hat{y}) = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}$$
>
> where $\bar{y} = \frac{1}{n}\Sigma_{i=1}^{n} y_i$ and $\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 = \Sigma_{i=1}^{n}\varepsilon_i^2$.

Continuous Outcomes and Evaluation Metrics

- **Adjusted R-squared**
  - A modified version of R-squared that incorporates the number of predictors.
  - Unlike R-squared, adjusted R-squared can be negative.
  - Adjusted R-squared is always less than or equal to R-squared.

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$\text{Adjusted } R^2 = 1 - \frac{SS_{residuals} / (n - K)}{SS_{total} / (n - 1)}$$

Continuous Outcomes and Evaluation Metrics

- **Mean squared error (MSE)**

> Mean squared error (MSE) is a risk metric corresponding to the expected value of the squared (quadratic) error or loss. It is formally defined as:
>
> $$MAE(y, \hat{y}) = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}.$$

- **Mean absolute error (MAE)**

> Mean absolute error (MAE) is a risk metric corresponding to the expected value of the absolute error loss. It is formally defined as:
>
> $$MAE(y, \hat{y}) = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} = \frac{\sum_{i=1}^{n} |\varepsilon_i|}{n}.$$

Continuous Outcomes and Evaluation Metrics

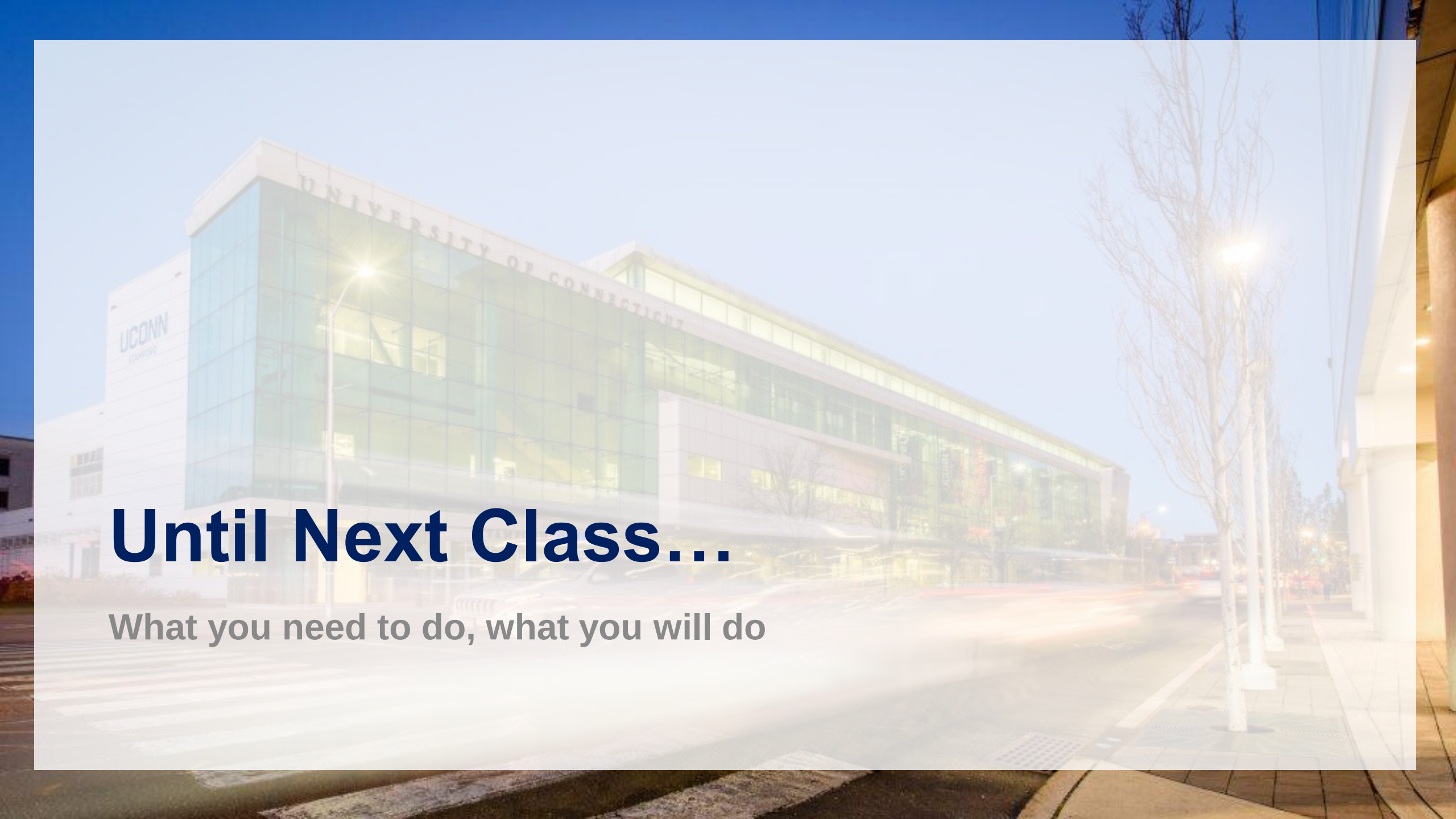
- **Akaike information criterion (AIC)**

> The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting.
>
> Let k be the number of estimated parameters in the model and $\hat{L}$ be the maximized value of the likelihood function for the model. Then the AIC value of the model is the following:
>
> $$AIC = 2k - 2ln(\hat{L}).$$
>
> Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

Continuous Outcomes and Evaluation Metrics

- **Bayesian information criterion (BIC)**

> The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred.
>
> The BIC is formally defined as:
>
> $$BIC = kln(n) - 2ln(\hat{L}),$$
>
> where n is the number of observations; the other notations are identically defined as the AIC's formula.
>
> The formula for BIC is similar to the formula for AIC, but with a different penalty for the number of parameters. With AIC the penalty is $2k$, whereas with BIC the penalty is $kln(n)$.



Until Next Class…

What you need to do, what you will do

One week to our next meeting

- **Feb 20 (Thu): Next Class**
  - **Python Hands-on for Concepts in Lecture 04**
  - **Q&A for Data Collection Methods**
  - **Deep Learning: Basics (1 of 2)**
- **Feb 21 (Fri): Hands-on Assignment #1 Due**
- **Feb 27 (Thu): Two Weeks Later**
  - **Deep Learning: Basics (2 of 2)**
  - **Short Session for Mid-term Preparation**
- **Mar 6 (Thu): Mid-term Exam**

Jaeung Sim
Assistant Professor
School of Business, University of Connecticut
jaeung.sim@uconn.edu