

**Modeling Data for Inference**

**Course Number:** DATA-142  
**Duration:** 5 days

**Overview**

This Modeling Data for Inference course teaches attendees how to use Python to perform causal inference on observational data. Participants learn how to work with inferential models, missing data, and experimental design.

**Prerequisites**

Attendees must have a solid foundation in Python programming for descriptive analytics.

**Materials**

All Data Modeling training students receive comprehensive courseware.

**Software Needed on Each Student PC**

* Windows, Mac, or Linux
* A current version of Anaconda for Python 3.x, or a comparable Python installation with the necessary libraries (Accelebrate can provide a list)

**Objectives**

* Perform causal inference in observational data using Python
* Perform and interpret null hypothesis testing in Python
* Implement generalized linear models in statsmodels
* Understand missing data
* Impute missing data
* Generate accurate power calculations
* Implement non-parametric methods to test hypotheses.
* Use causal inference frameworks to identify causal effects from observational data

**Outline**

* Introduction
* GLMs with Python using Stats Models
  + Applying Statistical Models for Analysis in Python: The A/B test
    - Explanation of statsmodels library of functions
    - Inferential and descriptive statistics refresher
    - Implementing A/B tests
* Modeling Continuous Data (Linear models)
  + Formulation of the simple linear model
  + Application of the intercept only, null model
    - Binary predictor
    - Interpreting results
    - Categorical predictor
    - Continuous predictor
    - Polynomial expansions
    - Multiple linear regression
    - Spline models
    - Interaction terms
    - Picking the “best” model
    - Discussion of confounding, interaction terms, and model building approaches
  + Modeling Binary Data (Logistic models)
    - Discussion of the generalized linear model
    - The Logit link function
    - Binomial distribution
    - Intercept only model
    - Back transformation of coefficients
    - Simple predictor
    - Multiple predictors
    - Odds ratio interpretations
    - Generating a scoring data set
    - Predicting from the model with new data
  + Modeling Count Outcomes
    - How are count outcomes different?
    - Poisson models
    - Over dispersed modeling options
    - Log link functions
    - Using offsets to model rates / uneven follow-up
* Power Analyses/Study Design
  + Understanding and estimating statistical power
  + Type 1 and type 2 errors
  + Using existing power estimators
  + Simulating power through the data-generating process
* Non-Parametric Analysis Methods
  + Using bootstrapping/permutation tests
    - Bootstrapping versus depending on asymptotic behavior to estimate confidence intervals
    - How different/stable are my results?
    - resampling a data set
    - bias-corrected bootstrap interval
    - Extending the bootstrap function to calculate more statistics
    - Permutation tests for p-values
* Missing data
  + Quantifying
  + Visualizing missing data
  + MAR,MCAR,MNAR
  + Sensitivity analysis
  + Imputation
    - MICE/trees pre-processing
* Time to Event (Survival) Analysis
  + Visualizing Hazards Across Time
  + Understanding the Log Rank Test
  + Cox Proportional Hazards Modeling
    - Understanding and interpreting the Hazard Ratio
    - Model diagnostics and assumptions
    - Implementing Time Varying Covariates
  + Parametric Survival Models
    - Weibull Model
    - Exponential Model
    - Predicting Failure Times
* Causal Inference: The Potential Outcomes Framework
  + Defining treatment effects (ATT, ATE)
  + Identifying populations of interest
  + Defining your causal hypothesis
  + Understanding the counterfactual
  + Establishing the causal diagram for your problem
  + Different methods for conditioning on variables:
    - Propensity Scores
    - Direct regression adjustment
    - G-computation formulas
  + Instrumental variable analysis
* Conclusion