Statistics and Data Science: A Modeling Approach

CourseKata Goals and Learning Objectives

High-level Performance Objective

At the end of the course students will be able to:
Generate research questions that could be answered with data--whether collected by students or provided to them--and will engage in the cycle of data analysis: exploring variation, modeling variation, evaluating their models, and communicating the results of what they have learned from data analysis.

To achieve this high-level objective, students will master a wide range of specific learning objectives, as shown in the table below.

How to Read the Learning Goals and Objectives Table

What students will learn in CourseKata is defined by a hierarchical system of high-level goals, sub-goals, and learning objectives. High-level goals are shown in bold and numbered 1, 2 and 3. Each high-level goal is organized into two levels of sub-goals shown in bold and numbered 1.1, 1.2, etc., or 1.1.1, 1.1.2, etc.

The meaning of each sub-goal is then made clear by specific learning objectives, which are identified by codes combining letters and numbers: CM1, CM2, etc. The letters represent important words in the sub-goal: for example, CM = Concepts of Measurement.

Many objectives require effective use of R code to achieve them. This is indicated by (R) at the end of the objective. Numbers in parentheses—e. g., (2, 4)—show the textbook chapters that address the objective.

ID	Learning Goals and Objectives
1. UNDERSTAND DATA	
1.1. Understand where data come from	
CM	**1.1.1. Understand and apply concepts of measurement**
CM1	Quantify variation in data by assigning numbers or categories to attributes (2)
CM2	Differentiate quantitative and categorical levels of measurement (2)
CM3	Explain the inherent issue of measurement error (2)

CM4	Recognize the limitations of measurement, including measurement bias (2)
SM	**1.1.2. Understand sampling methodology**
SM1	Explain that samples are the result of a sampling process and a data generating process (DGP) (2, 3)
SM2	Define independent, random sampling and explain the consequences of violating independent, random sampling (2)
SM3	Explain and apply the definition of sampling variation and consider it a possible explanation for observed group differences (2, 4)
SM4	Recognize that samples (and even random samples) are not perfectly representative of the entire population (2)
SM5	Recognize that samples are studied in order to find out about the population and the DGP (2)
SM6	Recognize the limitations of sampling, including sampling bias (2)
RD	**1.1.3. Understand elements of research design**
RD1	Distinguish outcome variables versus explanatory variables (2,3, 4)
RD2	Differentiate observational and experimental studies and recognize the limitations of each (4)
RD3	Differentiate random sampling and random assignment and explain the advantages of each (4)
1.2. Understand the organization of data	
DF	**1.2.1. Understand data frames**
DF1	Interpret the structure of data: rows (i.e. observations) and columns (i.e. variables) (R) (2)
DF2	Differentiate and interpret variables versus values in a data set (2)
DF3	Differentiate and interpret levels of variables (2)
MD	**1.2.2. Manipulate data in a data frame**
MD1	Write code to perform basic R commands (R) (1 - 12)
MD2	Create summary variables (R) (2)

MD3	Aggregate variables across rows to create a new data frame (R) (2)
MD4	Recode quantitative variables into categorical variables e.g., using ntile() (R) (2)
MD5	Identify and decide how to handle missing data (R) (2)
MD6	Filter, organize, and manipulate data in a data frame; interpret the results (e.g. arrange, select; R) (2)
1.3. Understand the purpose of data	
DQ	**1.3.1. Imagine data that could answer a question**
DQ1	Generate questions that could be answered with a given data set (2)
QD	**1.3.2. Generate questions you could ask of data**
QD1	Evaluate appropriateness of data to specific questions and purposes (2)
2. EXPLORE VARIATION	
2.1. Understand sources of variation	
SV	**2.1.1. Categorize sources of variation**
SV1	Recognize that variation can be divided into explained and unexplained; unexplained variation into real (i.e. a product of the system) or induced; induced variation into measurement error, sampling error, or mistakes (4)
EV	**2.1.2. Understand what it means to explain variation**
EV1	Explain and apply an intuitive definition of ‘explaining variation’ (4)
EV2	Identify within and between group variations in a graph (4)
EV3	Apply the definition of explaining variation to graphs (e.g. scatterplots, faceted histograms) (4)
CCC	**2.1.3. Understand correlation, causation, and confounding**
CCC1	Recognize that explaining variation does not always mean the relationship is causal (4)
CCC2	Identify and apply the concepts of confounding variables and the problem of directionality (4, 7)

CCC3	Differentiate experimental from observational designs; understand which can lead to causal conclusions (4, 7, 8, 11)
CCC4	Recognize that randomness can cause an apparent relationship in data (4)
2.2. Describe distributions of data	
CD	**2.2.1. Understand concept of distribution**
CD1	Define a distribution as the pattern of variation in a variable or combination of variables (3)
CD2	Describe a distribution using shape, center, and spread, and reason about the possible causes of a distribution's characteristics (3)
CD3	Reason about measures of central tendency (5)
DGP	**2.2.2. Understand that distributions of data are generated by DGP, which is usually unknown**
DGP1	Recognize that populations are the long-run result of a host of causal processes we call the DGP; (3)
DGP2	Simulate samples of data from a DGP; anticipate what the data may look like ® (3, 4, 9, 10)
DGP3	Generate hypotheses about the DGP based on distributions of data; recognize limitations (3, 4)
DGP4	Recognize the role of randomness and the law of large numbers in interpreting distributions of data (3)
RUD	**2.2.3. Represent univariate distributions in tables and graphs**
RUD1	Create frequency tables and relative frequency tables (e.g., tally) (R) (3)
RUD2	Create and interpret box plots, histograms, and bar graphs to visualize univariate data (R) (3)
RUD3	Choose the appropriate visualization to represent a given variable (3)
RUD4	Create and interpret a five-number summary (i.e., favstats()) (R) (3)
RUD5	Find and interpret the interquartile range (IQR) (4)
RUD6	Identify outliers and decide how to handle them (3)
IUD	**2.2.4. Interpret tables and graphs of univariate distributions**

IUD1	Interpret the axes of a histogram: identify that frequency is represented on the y axis; variable is represented on the x axis (3)
IUD2	Interpret bar graphs and understand why shape, center, and spread are not meaningful characteristics for these visualizations (3)
IUD3	Recognize and explain the differences and similarities between a frequency histogram and a relative frequency histogram (3)
2.3. Describe relationships in data	
RBR	**2.3.1. Represent bivariate relationships in tables and graphs**
RBR1	Create two-way contingency tables (e.g. using tally function) with both frequencies and proportions (4) (R)
RBR2	Create scatterplots, faceted histograms, box plots, jitter plots to visualize bivariate data (4) (R)
RBR3	Choose the appropriate visualization to represent a bivariate relationship (4)
IBR	**2.3.2. Interpret tables and graphs of bivariate relationships**
IBR1	Interpret bivariate relationships presented in tables, scatter plots, faceted histograms (4)
IBR2	Visually evaluate the strength of relationships in scatter plots, faceted histograms, box plots (8)
WE	**2.3.3. Represent relationships in word equations**
WE1	Create path diagrams to represent bivariate relationships (4)
WE2	Write word equations to represent relationships between explanatory and outcome variables (4)
3. MODEL VARIATION	
3.1. Understand data = model + error	
SM	**3.1.1. Understand concept of statistical models**
SM1	Map DATA = MODEL + ERROR in different contexts (4, 5, 7, 8)
SM2	Define and use the concept of a statistical model as a function that produces a predicted score for each observation (5, 7, 8)
SM3	Differentiate between the empty model and a more complex model (5, 7, 8)

SM4	Identify the most appropriate model to use based on how variables are measured (e.g., group versus regression models) (5, 7, 8)
SM5	Distinguish between a model of data and a model of DGP (5, 6, 7, 8)
CE	**3.1.2. Understand concept of error**
CE1	Visually intuit which models or distributions show more error (4, 5, 7)
CE2	Define, calculate, and reason about residuals from models (5, 6, 7, 8)
CE3	Recognize that residuals aggregate to measures of error (7)
CE4	Recognize that error from the empty model is all of the unexplained variation (6)
PE	**3.1.3. Understand partitioning error**
PE1	Recognize that sum of squares from the empty model (SS total) can be partitioned into error and model sums of squares (6)
PE2	Recognize that SS total is determined by the outcome variable and so total variation for an outcome variable will be the same regardless of explanatory variables included in the model (7,8)
PE3	Interpret visual representations of SS error, SS model, and SS total from empty, group, and regression models (8)
TV	**3.1.4. Understand how transforming variables can help build models**
TV1	Use difference scores to model paired samples (7)
TV2	Use standardized variables (e.g., z-scores) to compare the strengths of linear relationships between different sets of variables (6, 8)
TV3	Calculate and interpret Pearson's R, and estimate correlation coefficients from scatterplots (8)
TV4	Link transformed variables to visualizations (e.g., scatterplots, linear representations of models) (6, 8)
3.2 Specify models	
GLM	**3.2.1. Write models using GLM notation**
GLM1	Use and interpret GLM notation to represent models of data, the DGP, and sampling distributions (5, 7, 8)
GLM2	Flexibly use GLM notation to represent group and regression models in different contexts (7, 8)

GLM3	Differentiate between variables and parameters in GLM equations (5, 7, 8)
GLM4	Use dummy codes to model multi-group data in GLM (7)
IPE	**3.2.2. Interpret parameter estimates in context**
IPE1	Map GLM components to contexts, graph, and word equations (5, 7, 8)
IPE2	Interpret GLM components computationally (5,7,8)
3.3. Fit models	
FM	**3.3.1. Understand what it means to fit a model**
FM1	Recognize that fitting a model means to calculate best fitting parameter estimates (5,7,8)
FM2	Identify that the mean acts as a balancing point for (error) residuals in a distribution (5)
FM3	Recognize that the best fitting model minimizes the appropriate measure of error (in this course, sum of squares) (6, 7, 8)
FM4	Understand the concept of “overfitting” a model (7)
EP	**3.3.2. Estimate parameters**
EP1	Differentiate between parameters and statistics; explain why a statistic is our best estimate of a (usually) unknowable parameter (5, 7, 8)
EP2	Estimate parameters for empty models, group models, and regression models (i.e., fit models) (5, 7, 8) (R)
EP3	Represent and interpret visual representations of models (e.g. vline() (5, 7, 8)
IE	**3.3.3. Interpret estimates**
IE1	Interpret the output of lm() for empty, group and regression models (R) (5,7,8)
IE2	Interpret parameter estimates for group and regression models in context (8)
IE3	Write best fitting model based on lm() output (5,7,8)
3.4. Assess model fit	

UR	**3.4.1. Use residuals to assess model fit**
UR1	Recognize that the empty model is made up of all unexplained variation (5)
UR2	Recognize that a residual from the empty model can be partitioned into error and model (7, 8)
UR3	Identify and interpret residuals on a graph of data for empty, group, and regression models (6, 7, 8)
UR4	Plot and interpret distributions of residuals (R) (7, 8)
UR5	Generate and analyze predictions and residuals from a model (R) (6, 7, 8)
AE	**3.4.2. Quantify aggregate error around a model**
AE1	Quantify aggregate error as Sum of Absolute Deviations (SAD), Sum of Squares (SS), variance, and standard deviation, and interpret these measures (R) (6)
AE2	Explain how sums of squares are constructed from residuals (6)
AE3	Recognize the pros and cons of different methods of quantifying error (6)
AE4	Calculate and interpret z-scores within a distribution or across distributions (6)
CF	**3.4.3. Compare the fit of two models**
CF1	Use SS, PRE, and F statistic to compare models (7, 8)
CF2	Calculate (e.g., find ANOVA tables; R) and interpret PRE and F statistic (R) (7, 8)
CF3	Interpret a t distribution and relate the critical t to critical z (10)
CF4	Interpret PRE similarly across ANOVA models and regression models: recognize that these two measures express the same thing (8)
ES	**3.4.4. Understand and calculate different measures of effect size**
ES1	Describe and determine appropriate measures of effect size such as difference of means, PRE, Cohen's d (7, 8)
3.5 Use models to generate predictions and probabilities	
ND	**3.4.5. Use the normal distribution to model variation**

ND1	Explain why the normal distribution may be used to model variation in the DGP/population (6)
ND2	Identify the features of a normal distribution and use data to construct best fitting normal model of variation (e.g., defined by mean and standard deviation) (6)
PE	**3.4.6. Make point predictions based on parameter estimates**
PE1	Use parameter estimates in functions to make point predictions from the empty, group, and regression models (R) (6)
PD	**3.4.7. Make likelihood predictions based on probability distributions**
PD1	Explain the concept of probability distribution as a model of a random DGP (6, 9)
PD2	Explain the relationship between a discrete (e.g. data, simulated normal distributions) and continuous probability distribution (6)
PD3	Use the empirical rule to estimate the likelihood of scores under a normal distribution (6)
PD4	Use the distribution of data to predict likelihood of future observations falling within a specified range (6)
PD5	Use the normal distribution to estimate likelihood of future observations falling within a specified range (R) (6)
4. EVALUATE MODELS	
4.1. Model the distributions of estimates	
CSD	**4.1.1. Understand the concept of sampling distribution**
CSD1	Explain the concept of a sampling distribution (9)
CSD2	Explain the problem a sampling distribution solves, i.e. that it helps to make sense of a particular estimate given sampling variation (9)
CSD3	Interpret sampling distributions in specific contexts (9)
CSD4	Recognize that sampling distributions can depict the random distribution of any statistic (not just the sample mean) from some DGP (9)
CSD5	Recognize that sampling distributions are imaginary (9)
FSD	**4.1.2. Understand features of sampling distributions**
FSD1	Describe the shape, center, and spread of a sampling distribution of estimates, including SDoM, b0, b1, F, and PRE (9, 10, 11)

FSD2	Explain how the shape/center/spread of population and sample size are related to sampling distributions, and how the population's features will affect shape/center/spread of sampling distributions (9)
FSD3	Define and calculate standard error (9)
FSD4	Identify standard error visually (9)
SDR	**4.1.3. Construct sampling distributions (R)**
SDR1	Construct sampling distributions by resampling (bootstrapping), simulation, randomization, and using mathematical probability distributions (R) (9, 10, 11)
SDR2	Explain the relationship between a discrete (e.g. data, simulated normal distributions) and continuous probability distribution (6)
SDR3	Be able to predict the center and shape of a sampling distribution depending on the method used to construct it (9, 10, 11)
ISD	**4.1.4. Interpret sampling distributions**
ISD1	Calculate and evaluate the likelihood of selecting a random sample with a certain sample statistic, (e.g., a t distribution) (R) (9)
ISD2	Identify what questions can and cannot be addressed using sampling distributions; i.e. questions about individual observations require a population distribution whereas questions about samples require a sampling distribution (9)
4.2. Use confidence intervals to compare models	
CI	**4.2.1. Understand confidence intervals**
CI1	Recognize that just as a fixed DGP could produce a range of estimates, a single estimate could be produced by a range of DGPs (9)
CI2	Recognize that a confidence interval represents a range of parameter values that could, with some degree of likelihood, have produced your estimate (10)
CI3	Define margin of error and be able to find it (10)
CI4	Recognize how standard error relates to confidence intervals (10)
CI5	Recognize that the upper and lower bounds of a confidence interval represent the highest and lowest parameter values beyond which our sample would have been unlikely (10)
CCI	**4.2.2. Construct a confidence interval around an estimate (R)**
CCI1	Construct confidence intervals using simulation, bootstrapping, and the mathematical probability distribution (i.e., z, t) for various estimates (e.g., b0, b1) (10) (R)

CCI2	Explain how level of confidence, sample size, and variance impact the size of a confidence interval (10)
ICI	**4.2.3. Interpret a confidence interval**
ICI1	Interpret confidence intervals in regression and grouping models for various estimates (e.g., b0, b1) (10)
ICI2	Explain why you might reject the empty model when a confidence interval for b1 contains 0 (10)
ICI3	Interpret a confidence interval correctly (10)
4.3. Compare models using the F distribution	
F	**4.3.1. Understand F**
F1	Define F as a measure of the strength of a relationship *per parameter* used in a model; and as a ratio of variation explained to variation unexplained (11)
F2	Use simulation to explore which statistics can be modeled with the F distribution. (R) (11)
F3	Interpret an F distribution, and distinguish the F statistic from the F distribution (11)
F4	Recognize the that the shape of the F distribution depends on the degrees of freedom for model and for error (11)
MC	**4.3.2. Conduct model comparison**
MC1	Conduct F tests for ANOVA and regression models in some context for some purpose (7, 8)
MC2	Identify the region that corresponds to p-value in a sampling distribution of F or PRE (11)
MC3	Define p-value as the probability of obtaining an F or PRE statistic as extreme or more extreme than the one observed assuming that the empty model is true (11)
MC4	Explain how the numbers in an ANOVA table are calculated and what they mean (7, 8)
IMC	**4.3.3. Interpret results of model comparison**
IMC1	Determine which statistics can be used to compare two group or three group models versus an empty model (11)
IMC2	Recognize the limitations of the F test in a three group model (11)
IMC3	Explain the relationship between an F test and a t test (11)

IMC4	Recognize and explain the need for simple effects tests (11)
IMC5	Use the results of ANOVA tables to make predictions about confidence intervals (11)
IMC6	Explain the problem of simultaneous comparisons and how it is addressed by the bonferroni correction (11)
IMC7	Explain the inherent risk of statisticians' p hacking; calculate likelihood of p-hacking (11)
IMC8	Explain and recognize Type I error and why its probability of occurring is never zero (11)
IMC9	Explain the use of an alpha level and how it relates to type I and type II error (11)
IMC10	Explain and recognize Type II error (11)
IMC11	Use the F statistic to compare models; intuit about effect size based on size of F ratio (7)
IMC12	Interpret a p-value and link to the assumed Data Generating Process (11)