The syllabus for this four-hour exam is defined in the form of learning objectives, knowledge statements, and readings. It also includes various R packages and functions that candidates are expected to be familiar with.

LEARNING OBJECTIVES set forth, usually in broad terms, what the candidate should be able to do in actual practice. Included in these learning objectives are certain methodologies that may not be possible to perform from start to finish on an examination, but that the candidate would still be expected to explain conceptually if not demonstrate in the context of an examination.

KNOWLEDGE STATEMENTS identify some of the key terms, concepts, and methods that are associated with each learning objective. These knowledge statements are not intended to represent an exhaustive list of topics that may be tested, but rather are illustrative of the scope of each learning objective.

READINGS support the learning objectives. It is intended that the readings provide sufficient resources to allow the candidate to perform the learning objectives. Some readings are cited for more than one learning objective. Candidates are expected to use the readings cited in this *Syllabus* as their primary study materials.

Thus, the learning objectives, knowledge statements, and readings complement each other. The learning objectives define the behaviors, the knowledge statements illustrate more fully the intended scope of the learning objectives, and the readings provide the source material to achieve the learning objectives. Learning objectives should not be seen as independent units, but as building blocks for the understanding and integration of important competencies that the candidate will be able to demonstrate.

On a given examination, it is very possible that not every individual learning objective will be tested. Questions on a given learning objective may be drawn from any of the listed readings, or a combination of the readings. There may be no questions from one or more readings on a particular exam.

After each set of learning objectives, the references to the readings are provided in abbreviated form. Complete text references are provided at the end of this exam syllabus.

Items marked with a bold **OP** (Online Publication) are available at no charge and may be downloaded from the Internet at the links provided.

Prerequisites

- A working knowledge of R at an individual user level (not at a developer level). This includes the ability to write R functions and using the help() command. This knowledge may easily be gained by referencing one or more of the following:
  - Section 2.3 of ISL
  - http://faculty.chicagobooth.edu/richard.hahn/teaching/formulanotation.pdf
  - Roger Peng, R Programming for Data Science, See Chapter 14 on Control Structures and Chapter 15 on Functions. https://leanpub.com/rprogramming
  - https://cran.r-project.org/doc/manuals/R-intro.pdf is a good resource on the basics of R
  - http://thecasinstitute.org/wp-content/uploads/2016/09/Introduction_to_External_Readings_for_DS1.pdf *from the syllabus for the Data Concepts and Visualization exam*
- Basic linear regression functions in R. It may be helpful to consult chapter 4 of Fox and Weisberg to become familiar with these functions.
- A working knowledge of basic statistics. Needed concepts include hypothesis testing, confidence intervals, and basic linear regression. Some good sources for basic statistics, including confidence intervals and hypothesis testing, are chapters 1 and 4 of *An Introduction to Mathematical Statistics* by Hogg and Craig and Brian Caffo's *Statistical Inference for Data Science*. For linear regression models, if the student does not find the summary in chapter 3 of ESL to be review, it may be desirable to consult chapters 5 and 9 of Fox. An alternative source would be chapters 2-4 of Frees. It would also be helpful for the student's understanding to be familiar with concerns about multiple hypothesis testing as expressed, for example, in Regina Nuzzo's 2014 *Nature* article "Scientific Method: Statistical Errors" (https://www.nature.com/news/scientific-method-statistical-errors-1.14700).
- Candidates should have sufficient familiarity with the use of R's help facility to diagnose and resolve simple errors such as the names of function arguments, or values returned from functions. Candidates are expected to know that arguments to a function do not have to be named if they are provided in the same order expected by the function but must be named if the arguments are provided in a different order. If a function argument is named incorrectly, the function will likely result in an error. Remember that capitalization counts.
- Candidates should be aware of the default value of each argument in the function they use.
- In the exam several questions will be based in R. For these R based questions, candidates should not expect full credit for code which produces errors. Code which generates an error will be regarded as ambiguous with regard to the intent of the candidate. That is, it will not be clear to graders how much credit – if any at all – should be awarded for the question. Clear and ample comments within the code may help resolve ambiguities and could help a candidate earn partial credit when the code generates errors."

A. Planning a Modeling Project

Weight for Section A: 10-20 percent

The CRISP-DM paper lays out the steps of a typical data mining or predictive analytics project, while the Luca, et al. article discusses how to ensure your model is answering the business-relevant question, which is critical in the initial, business understanding, phase of a project. The Rubin and Waterman paper emphasizes the critical difference between observational and experimental data and points the way to more careful use of observational data when causal conclusions are required. The Thomke and Mazni article discusses how to monitor (and pilot) models to ensure they perform as they are supposed to. The concepts discussed here at a high level in Rubin and Waterman and Thomke and Mazni will find more detailed discussion in the experimental and quasi-experimental design sections of the syllabus. They should always be kept in mind even (especially!) when applying statistical or machine learning methods to observational data as they indicate ways in which models based on observational data can fail.

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
1. Identify the Business Problem	a. Apply the CRISP-DM framework to the process of planning out a predictive analytics project b. Understand the limitations of modeling, and apply this to determine how to leverage models effectively, and how to identify useful input data c. Understand the differences between working with observational data and experimental data, and why observational data can lead one astray d. Identify methods that can help the modeler make better use of observational data e. Understand and evaluate the role of business experiments in model implementation

READINGS

- Chapman
- Luca, et al.
- Rubin and Waterman
- Thomke and Manzi

B. Classical Models & Diagnostics

Weight for Section B: 25-35 percent

We begin with a brief section on the various types of data that are used in models and considerations that arise from certain sources of data (such as surveys) and from the frequent difficulty of missing information.

Interpretation of model diagnostics is critical in predictive analytics. As these diagnostics are very well developed for the classical statistical models, we start there. We then introduce the student to generalized linear models (including logistic regression, Poisson regression, and Tweedie regression), which are almost no more complex than linear regression but which find use in a much wider variety of situations. We also include material specific to adjustments that are often necessary in the case of logistic regression, especially when one of the classes of the response variable is rare. Finally, we close with a discussion of hierarchical models, which handle the common situation in which the regression assumption of independent observations is violated. Linear mixed models are introduced as the simplest case of hierarchical models and used as a link to what is called Bühlmann credibility in insurance, and used as a means to introduce the student to Vignette for R packages.

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
1. Types of Data, Missing and Incomplete Data	a. Describe types of data such as discrete and continuous data. Describe special issues that arise in data from surveys. b. Describe key patterns of missing data values, including censoring, truncation, missing-at-random, and missing-completely-at-random. c. Describe key underlying causes of missing data. Identify appropriate ways to deal with missing values in a given situation and identify the advantages and disadvantages of each.
READINGS	
• ESL, 2.2 • Fox, 15.5 • Gelman and Hill, Ch. 25 up to but not including 25.4 • Study Note on Truncation and Censoring	

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
2. **Linear Model Diagnostics**	a. Learning Objective: Interpret linear model output such as confidence intervals for parameter estimates and for predictions. Perform, interpret, and act upon standard diagnostics on linear models, including assessment and treatment -of outliers; -of appropriateness of model specification; -of nonnormal errors; -of nonlinear dependencies; -of heteroscedasticity; -of multicollinearity b. Understand and apply the hat matrix, hat values, residuals (raw, standardized, Studentized, and Pearson), and Cook's D to detect outliers and influential observations c. Apply residual plots, marginal model plots, and added variable plots to assess quality of fit and the impact of each predictor d. Use QQ plots to diagnose non-normal errors, e. Use F-tests, residual plots, component-plus-residual plots, and CERES plots to identify non-linear dependencies f. Use residual plots and spread-level plots to identify heteroscedasticity; determine when transformation of the target variable (possibly via Box Cox) is an appropriate remedy, and when weighted regression is appropriate. g. Identify collinearity via variance-inflation factors and generalized variance-inflation factors and discuss possible ways to deal with collinearity

READINGS

- ESL, Ch. 3 up to but not including 3.2.4
- Fox, Ch. 11 up to but not including 11.4.1; 11.6-11.8.3; Ch. 12 up to and including 12.5.1; results stated in exercises 12.3-12.4; Ch. 13 excluding 13.1.1
- Fox and Weisberg, Ch. 6 excluding the following four items: last subsection of 6.4.1, last subsection of 6.4.2, all of 6.5.2, all of 6.6

R Packages and Functions
• Functions in default packages : residuals, rstandard, rstudent, hatvalues, cooks.distance, dfbeta, dfbetas, lm.influence • in car package: residualPlots, marginalModelPlots, qqPlot, outlierTest, influenceIndexPlot, boxCox, powerTransform, crPlots, ceresPlots, spreadLevelPlot, vif, avPlots, boxCoxVariable • prerequisites from default R packages: lm and functions that manipulate lm objects (e.g., predict.lm, summary.lm, coef, effects, vcov)

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
3. Classical Models—Generalized Linear Models and Their Diagnostics	a. Understand the assumptions behind different forms of the Generalized Linear Model and be able to select the appropriate model b. Understand the relationship between mean and variance for various models within the GLM family c. Understand how to select the appropriate link function and distribution for the dependent variable. d. Understand the Tweedie as compound gamma-Poisson and also as the GLM with variance function a power law. e. Be able to describe the reason for a double GLM and two ways in which a double GLM might be fit. Be able to describe similarities and differences between a double GLM and a weighted GLM f. Use appropriate diagnostics to evaluate the fit of a GLM g. Describe the effect of non-canonical link function h. Define deviance and its relationship to a GLM

READINGS

- Allison
- Fox, Ch. 14.1, Ch. 15 up to but not including 15.5
- Fox and Weisberg, Ch. 5.10-5.11, Ch. 6.6
- Smyth and Jorgensen, sections 1-2, first paragraph of section 3

R PACKAGES AND FUNCTIONS

- glm, family in default packages
- glm.nb in MASS package
- logistf and summary.logistf in logistf package
- all R functions mentioned in section 6.6 of Fox-Weisberg

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
4. Hierarchical Models, including Linear Mixed Models, and Bühlmann Credibility	a. Describe sources of correlation in longitudinal data b. Describe REML and its rationale c. Use appropriate plots to do EDA of longitudinal data d. Build hierarchical models via the linear mixed-effects approach e. Apply Bühlmann credibility theory and describe the connection to linear mixed effects models

READINGS

- Fox, Ch. 23 up to but not including 23.9.2
- Gelman and Hill, Ch. 11-13
- Study Note on Credibility – Updated December 2018
- Vignette

R PACKAGES AND FUNCTIONS

- All functions mentioned in the Vignette for the lme4 package

C. Machine Learning Methods

Weight for Section C: 25-35 percent

We begin with sections that apply to all machine learning efforts, emphasizing the importance of out-of-sample data—both in a cross-validation context, to tune a model, and in a true holdout context, to validate a model. Machine learning models are sufficiently adaptive that in-sample ways of measuring goodness-of-fit are not reliable. Automated (as opposed to expert-driven) handling of non-linear dependencies and of interactions among variables, and model averaging approaches, are perhaps the most typical characteristics of machine learning methods. These are exemplified here by generalized additive models for non-linear effects and trees for interactions, with bagging, random forests, and boosting illustrating various model-averaging strategies. Finally, we conclude with a discussion of unsupervised learning, including applications.

It is not possible to cover all techniques in a single exam. Among techniques receiving little or no attention here that the student should be aware of, we should mention deep learning and Markov chain Monte Carlo algorithms. There are also areas of application, such as text analytics and image analytics, that typically require techniques not covered in this syllabus. This syllabus should provide a sufficiently deep introduction to the language and framework of machine learning to allow the student to learn additional techniques as needed from the relevant literature.

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
1. Validation Holdout vs Cross-Validation and Tuning Parameters	a. Explain and contrast holdout and Cross-Validation approaches and the best use of each b. For a given dataset and model, use cross-validation to estimate the accuracy of model predictions. c. Why might this estimate be inaccurate?
READINGS	
• ISL, Ch. 5 Intro, 5.1, 5.3.1-5.3.3, 2.2 Assessing Model Accuracy • Study Note on Validation and Holdout Data – Updated December 2018	
R PACKAGES AND FUNCTIONS	
• cv.glm function in boot package, sample function in default package	

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
2. **Evaluation: Goodness of Fit Metrics, Bootstrapping, Bias-Variance Tradeoff, and Presentation of Results**	a. Define and apply ROC curves, AUC, Lorenz curves, and Gini index b. Estimate variance of model estimates. c. Describe why your model may be biased. d. Describe how to build a model to minimize the expected mean squared error. e. What exhibits do you show for the holdout data f. What presentation material do you prepare and show

READINGS

- ESL, Ch. 7 up to but not including 7.8
- ISL, 5.2, 5.3.4
- Study Note on Validation and Holdout Data

R PACKAGES AND FUNCTIONS

- createDataPartition, defaultSummary, train, trainControl, resamples, resampleHist, resampleSummary in the caret package

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
3. Classification Models and Special Considerations	a. Describe and apply the ROC curve in evaluating a classification model b. Define and describe the Bayes error c. Apply linear regression, logistic regression, linear discriminant analysis, quadratic discriminant analysis, and nearest neighbors to fit classification models. Compare and contrast these methods as to when each might be preferable d. Fit a logistic regression by penalized maximum likelihood, and describe when that should be preferred to maximum likelihood e. Describe how unbalanced training datasets can influence classifiers and why that is a problem f. Identify algorithmic solutions to using unbalanced training sets, including various undersampling, oversampling, and cost-sensitive learning approaches g. Discuss the advantages and drawbacks of each

READINGS

- Ganganwar, up to section VI but excluding sections IV and V
- ISL, 2.2.3, Ch. 4 (including the Lab)
- Sokolova and Lapalme, up to but not including section 4

R PACKAGES AND FUNCTIONS

- glm in defaults packages
- kNN in class package
- lda, qda in MASS package
- logistf and summary.logistf in logistf package

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
4. Shrinkage and Feature Selection Methods	a. Apply forward stepwise selection b. Define "best subset" selection c. Define a shrinkage method and explain which penalty term corresponds to which method (ridge, lasso) d. Use shrinkage methods (lasso and ridge) to improve linear model predictions e. Select the tuning parameter for the penalty term. Comment on how this is done.

READINGS

- ESL, 3.3-3.4, 3.6

R PACKAGES AND FUNCTIONS

- glmnet, deviance.glmnet, cv.glmnet, plot.glmnet, plot.cv.glmnet, predict.glmnet, predict.cv.glmnet, print.glmnet in glmnet package

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
5. Non-Linear Effects and Additive Models	a. Be able to discuss several ways of capturing non-linear relationships in regressions and GLM models, including polynomials, step functions, splines, smoothing splines, and local regression b. Be able to build generalized additive models (GAM).
READINGS	
• ISL, Ch. 7	
R PACKAGES AND FUNCTIONS	
• smooth.spline and loess in default packages • gam in gam package	

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
6. Single Trees	a. Build regression and classification trees b. Use a tree to determine an estimate for an observation c. Discuss reasons for pruning and methods to prune d. Implement pruning
READINGS	
• ISL, Ch. 8 up to and including 8.1, 8.3.1, 8.3.2	
R PACKAGES AND FUNCTIONS	
• tree package, including prune.tree, plot.tree, and predict.tree	

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
7. Ensemble Methods, Random Forests, and Boosting	a. Be able to fit bagged tree models, boosted tree models, and random forests to data b. Be able to use each to get estimates for a new observation c. Discuss how each of these methods works, and what its pros and cons are

READINGS

- ISL, 8.2, 8.3.3, 8.3.4

R PACKAGES AND FUNCTIONS

- In randomForest: randomForest, predict, plot, summary, importance
- In gbm: gbm, summary, predict
- In tree: tree, plot, predict

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
8. Principal Components Analysis and Unsupervised Learning	a. Explain and apply principal components analysis b. Differentiate between supervised and unsupervised learning tasks c. Describe and apply principal components analysis in the context of dimension reduction d. Describe the choices involved in using k-means and hierarchical clustering and the implications thereof e. Interpret a dendrogram f. Summarize potential issues with using clustering and ways to mitigate them g. Cluster data using k-means and hierarchical clustering

READINGS

- ISL, 2.1.4, Ch. 10

R PACKAGES AND FUNCTIONS

- prcomp, kmeans, hclust, dist, cutree in default packages

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
9. Application Specific Methods: Association Models	a. Understand the basics of an association model as used in a marketbasket analysis b. Interpret rules metrics including: - Support - Confidence - Lift Association Model c. Identify the types of patterns that can be detected using simplified association rules d. Interpret and evaluate the support, confidence, and lift of a market basketanalysis
READINGS	
• ESL, 14.2.0-14.2.3, 14.2.7	

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
10. Application Specific Methods: Fraud Detection	a. Apply Pridit and Random Forests to fraud detection problems
READINGS	
• Frees, et al., Ch. 7 up to and including 7.12	
R PACKAGES AND FUNCTIONS	
• princomp in default packages; randomForest, predict, plot, summary, importance in randomForest	

D. Causal Inference

Weight for Section D: 10-20 percent

It is often important to estimate causal effects. For example, how many claims would be avoided by sending inspectors to a job site annually? It is notoriously tricky to do this with observational data—and indeed almost impossible to do this simply by including other confounding variables in a regression or GLM and interpreting the coefficient for the variable of interest as the size of a causal impact.

There are two main ways around this conundrum. One is to design randomized experiments, which are the gold standard for proving causation. Applications of this include A/B testing in on-line marketing and conjoint analysis in survey studies to support, for example, product design. Another is to interpret and analyze observational data in a new way, that focuses on what the results of an experiment would have looked like. While this can never be perfect, it is far preferable to the regression approach mentioned in the previous paragraph.

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
1. Experimental Design	a. Understand randomized experimental design, including factorial design, randomized block design, and covariance design. b. Understand the importance of assessing power in the design stage c. Understand internal validity and construct validity. d. Understand external validity is and what are the threats to it, and how it can be improved. e. Understand the Intention-to-Treat principle and apply it in the context of a business experiment f. Understand Simpson's paradox and explain how it can be misleading
READINGS	
• Newell • Oehlert, Ch. 1, 2.1-2.4, 3.1, 4.1 • Seltman, Ch. 8	

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
2. Experimental Methods	a. Understand choice models b. Given appropriate data, be able to fit a choice model c. Understand the use of conjoint analysis with survey data d. Given some survey data be able to apply a conjoint model e. Describe power and significance in A/B testing f. Summarize common mistakes and difficulties in A/B testing and techniques to address these g. Describe practical considerations in planning an A/B test h. Recognize alternative techniques to classical A/B testing

READINGS
• Chapman and Feit, 9.3.2-9.3.4, Ch. 13 except 13.4-13.5 • Vector • Bloggers • Miller

R PACKAGES AND FUNCTIONS
• mlogit package: mlogit and mlogit.data

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
3. Causal Inference from Observational Data	a. Understand coarsened exact matching (CEM), propensity scoring, and model-based methods for estimating causal effects, and explain the strengths and weaknesses of each b. Discuss the process for using propensity scores and CEM to estimate causal effects c. Distinguish causal effects from predictions d. Explain SATT (sample average treatment effect on the treated)

READINGS

- Iacus and Porro, Sections 1, 3.0, 3.2, 3.3, 3.5
- Stuart and Rubin, Ch. 11

R PACKAGES AND FUNCTIONS

- cem package. Functions att, cem, relax.cem

E. Ethical Considerations and Public Perception

Weight for Section E: 5-15 percent

Predictive models create a certain amount of fear in the public, partly because they are not fully understood, and partly because some of the goals in support of which models have been used. In this section we explore how to manage predictive model ethically, with a view to appropriate impacts on the public (Hancock and Wattenberg et al.) and also exhibit some of the different perspectives on these issues (Kleinberg et al., O'Neil). Whereas for testing for other portions of this syllabus we will try to avoid asking questions including the phrase "according to N", for the readings in this section, that will be considered fair, as these readings certainly represent a diversity of opinion that the student should gain an understanding of.

LEARNING OBJECTIVES	KNOWLEDGE STATEMENTS
1. Ethical Considerations and Public Perceptions	a. Describe ways models can promote fairness and ways in which they can perpetuate social inequalities b. Recognize the social implication of threshold selection models c. Identify situations where ethical care and oversight would be advisable d. Identify proxy variables and spurious correlations e. Describe the importance of feedback in model calibration f. Explain why it is possible to portray even the fairest model as biased g. Describe the key principles of the UK's ethical framework for predictive models h. Recognize concerns (whether fair or not) that are commonly expressed about predictive models in the press and in political discourse.

READINGS

- Hancock
- Kleinberg, et al.
- O'Neil, Ch. 6, 9
- Wattenberg, et al.

Complete Text References for Exam 3

Text references are alphabetized by the Abbreviation column.

Citation	Abbreviation	Learning Objective	Source
Allison, P., "Convergence Failures in Logistic Regression," SAS Global Forum 2008 proceedings (https://pdfs.semanticscholar.org/4f17/1322108dff719da6aa0d354d5f73c9c474de.pdf)	Allison	B.3	**OP**
Mount, J. "A clear picture of power and significance in A/B tests," R-bloggers, (http://www.win-vector.com/blog/2014/05/a-clear-picture-of-power-and-significance-in-ab-tests/)	Bloggers	D.2	**OP**
Chapman, P., et al., CRISP-DM: Step-by-Step Data Mining Guide, pp. 6-64 (https://www.the-modeling-agency.com/crisp-dm.pdf)	Chapman	A.1	**OP**
Chapman, P., and Feit, E., R for Marketing Research and Analytics, Springer, 2015, 9.3.2-9.3.4, Ch. 13 EXCLUDING 13.4-13.5	Chapman and Feit	D.2	**B - SK**
Hastie, T., et al., *The Elements of Statistical Learning*, 2nd ed., Chapter 2.2, Chapter 3 up to but not including 3.2.4, 3.3-3.4, 3.6, Chapter 7 up to but not including 7.8, 14.2.0-14.2.3, 14.2.7 (https://web.stanford.edu/~hastie/Papers/ESLII.pdf)	ESL	B.1, B.2, C.2, C.4, C.9	**OP**
Fox, J., *Applied Regression Analysis and Generalized Linear Models*, 3rd ed., Sage Publications, 2015: Chapter 11 up to but not including 11.4.1; 11.6-11.8.3; Chapter 12 up to and including 12.5.1, results stated in exercises 12.3-12.4; Chapter 13 EXCLUDING 13.1.1, section 14.1 (dichotomous), section 15.1-15.5, Chapter 23 up to but not including 23.9.2	Fox	B.1, B.2, B.3, B.4	**B - SK**
Fox, J., and Weisberg, S., *An R Companion to Applied Regression*, 2nd ed., Sage Publications, 2011. 5.10-5.11, All of Chapter 6 (pp. 285-327) EXCLUDING the last subsections of 6.4.1 (307-309), the last subsection of 6.4.2 (312-314), 6.5.2 (316)	Fox and Weisberg	B.2, B.3	**B - SK**
Frees, E., Meyers, G., and Derrig, R., eds., *Predictive Modeling Applications in Actuarial Science, Vol II: Case Studies in Insurance*, Cambridge University Press, 2016: Chapter 7 up to and including 7.12 (180-200)	Frees, et al.	C.10	**B - SK**

Citation	Abbreviation	Learning Objective	Source
Ganganwar, V., "An Overview of Classification Algorithms for Imbalanced Datasets", *International Journal of Emerging Technology and Advanced Engineering*, Volume 2, Issue 4, April, 2012, start – III, VI (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.413.3344&rep=rep1&type=pdf)	Ganganwar	C.3	**OP**
Gelman, A., and Hill, J., *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2007: Chapters 11-13, Chapter 25 (up to but not including section 25.4)	Gelman and Hill	B.1, B.4	**B - SK**
Hancock, M., "Data Science Ethics Framework", Cabinet Office, 2016 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/524298/Data_science_ethics_framework_v1.0_for_publication__1_.pdf)	Hancock	E.1	**OP**
Iacus, S., King, G., and Porro. G. "CEM: Software for Coarsened Exact Matching." *Journal of Statistical Software*, 2009, Vol. 30, issue i09: 2009. Copy at (http://j.mp/Te8KP5) -- Sections 1, 3.0, 3.2, 3.3, 3.5	Iacus and Porro	D.3	**OP**
Gareth, J., et al., *An Introduction to Statistical Learning with Applications in R*, Springer, 2017, Chapter 2.1.4, 2.2.3, Chapter 4 (including the Lab), Chapter 5 intro, 5.1, 5.2, 5.3.1-5.3.4, Chapter 7, Chapter 8 start, 8.1, 8.2, 8.3.1, 8.3.2. 8.3.3, 8.3.4, Chapter 10 (http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf)	ISL	C.1, C.2, C.3, C.5-C.8	**OP**
Kleinberg, J., Mullainathan, S., and Raghavan, M., "Inherent Trade-Offs in the Fair Determination of Risk Scores," Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017 (https://arxiv.org/pdf/1609.05807.pdf)	Kleinberg and Raghavan	E.1	**OP**
Luca, M., Kleinberg, J., and Mullainathan, S., "Algorithms Need Manages, Too", *Harvard Business Review*, Jan – Feb 2016 (https://hbr.org/2016/01/algorithms-need-managers-too)	Luca, et. al.	A.1	**OP**
Miller, E., "How Not To Run an A/B Test," Evanmiller.org (www.evanmiller.org/how-not-to-run-an-ab-test.html)	Miller	D.2	**OP**

Citation	Abbreviation	Learning Objective	Source
Newell, D., "Intention to Treat Analysis: Implications for Quantitative and Qualitative Research," *International Journal of Epidemiology*, 1992, Vol. 21, No. 5 (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.561.9039&rep=rep1&type=pdf)	Newell	D.1	**OP**
Oehlert, G., *A First Course in Design and Analysis of Experiments,* 2010: Chapters 1, 2.1-2.4, 3.1, 4.1 (http://users.stat.umn.edu/~gary/book/fcdae.pdf)	Oehlert	D.1	**OP**
O'Neil, C., *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Broadway Books, 2017: Chapters 6 and 9	O'Neil	E.1	**B - SK**
Rubin, D., and Waterman, R., "Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology", Statistical Science, 2006, Vol. 21, No. 2, 206-222 (https://arxiv.org/pdf/math/0609201.pdf)	Rubin and Waterman	A.1	**OP**
Seltman, H., "Threats to your Experiment," *Experimental Design for Behavioral and Social Sciences,* Chapter 8 (http://www.stat.cmu.edu/~hseltman/309/Book/chapter8.pdf)	Seltman	D.1	**OP**
Smyth, G., and Jorgensen, B., "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling", ASTIN BULLETIN, Vol. 32, No. 1, 2002, pp. 143-157, Section 1-2, first paragraph of section 3.1 (https://www.casact.org/library/astin/vol32no1/143.pdf)	Smyth and Jorgensen	B.3	**OP**
Sokolova, M. and Lapalme, G., "A systematic analysis of performance measures for classification tasks", Information Processing and management, 45 (2009) 427-437: start - III, VI (http://rali.iro.umontreal.ca/rali/sites/default/files/publis/SokolovaLapalme-JIPM09.pdf)	Sokolova and Lapalme	C.3	**OP**
Stuart, E., and Rubin, D., "Best Practices in Quasi-Experimental Designs: Matching Methods for Causal Inference", *Best Practices in Quantitative Methods,* Chapter 11 (https://www.corwin.com/sites/default/files/upm-binaries/18066_Chapter_11.pdf)	Stuart and Rubin	D.3	**OP**
"Study Note on Credibility", The CAS Institute (https://thecasinstitute.org/wp-content/uploads/2019/01/Exam-3-Study-Note-Credibility01162019.pdf)	Study Note on Credibility	B.4	**OP**

Citation	Abbreviation	Learning Objective	Source
”Study Note on Truncation and Censoring”, The CAS Institute (https://thecasinstitute.org/wp-content/uploads/2018/05/studynote-vF-050218.pdf)	Study Note on Truncation and Censoring	B.1	**OP**
“Study Note on Model Validation and Holdout Date”, The CAS Institute (https://thecasinstitute.org/wp-content/uploads/2019/01/Exam-3-Study-Note-Model-Validation-01162019.pdf)	Study Note on Validation and Holdout Data	C.1, C.2	**OP**
Thomke, S., and Manzi, J., "The Discipline of Business Experimentation", *Harvard Business Review*, Dec. 2014 (https://hbr.org/2014/12/the-discipline-of-business-experimentation)	Thomke and Manzi	A.1	**OP**
Mount, J., “Why does designing a simple A/B test seem so complicated?”, Win-Vector Blog, June 22, 2015 (http://www.win-vector.com/blog/2015/06/designing-ab-tests/)	Vector	D.2	**OP**
Bates, D., et al., "Fitting Linear Mixed Effects Models using lme4", Journal of Statistical Software, (https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf)	Vignette	B.4	**OP**
Wattenberg, M., Viégas, F., and Hardt, M., “Attacking Discrimination with Smarter Machine Learning” (https://research.google.com/bigpicture/attacking-discrimination-in-ml/)	Wattenberg and Hardt	E.1	**OP**

Source Key

B	Book—may be purchased from the publisher or reseller
OP	Online Publication
NEW	Indicates new or updated material
SK	Material included in the Study Kit