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 conceptuallyif 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.

©2021, The CAS Institute, All Rights Reserved
The CAS Institute, 4350 North Fairfax Drive, Suite 250, Arlington, VA 22203
TheCASInstitute.org

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. Some of this knowledge is provided in the CSPA Exam 2 syllabus. Sources for additional background and a refresher on R are provided in the first two (2) parts of Section A of the Exam 3 syllabus.
- Basic linear regression functions in R. It may be helpful to consult chapter 3 of ISL and Chapter 6 of MASS 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 of MASS and Chapter 2 of CAS.
- 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.

©2021, The CAS Institute, All Rights Reserved
The CAS Institute, 4350 North Fairfax Drive, Suite 250, Arlington, VA 22203
TheCASInstitute.org

A. Classical Models, Diagnostics and Their Application

Weight for Section A: 40-60 percent

We begin with an introduction to programming and data manipulation in R. Note that some of these topics are also covered on DS1 - CSPA Exam 2. This prepares the candidate for understanding and using models.

Interpretation of model diagnostics is critical in predictive analytics. As these diagnostics are very well developed for classical statistical models, we start there. We introduce the student to generalized linear models, which are used a wider variety of situations than linear regressions. We introduce topics in model validation and selection and discuss how they are applied in two key casualty insurance applications – ratemaking and reserving. More specifically, we include the use of R and the application of multiple linear regression models and generalized linear models on general insurance pricing, claim reserving, and loss triangles using insurance data examples.

Topics	Learning objectives	Reference
A1. Basic concepts of the R language	a. Introducing R environment. b. Introducing R language and building personal R functions. c. Identifying and changing the class of each R object. d. Inputting an outputting different type of objects with various format. e. Creating simple graphics.	ISL Ch. 2. CAS Ch. 1.1-1.3, 2.1, 2.3**. This is Optional Background for later chapters.** MASS Ch.1, 2, 3 and 4.
A2. Understanding and manipulating data	a. Applying univariate statistics to a dataset. b. Handling missing data. c. Merging different datasets.	ISL Ch. 2.3 CAS Ch. 1.2-1.3 **(Background material for later chapters.)** MASS Ch. 2.3, 5.1-5.7
A3. Linear regression	a. Understanding the assumptions of the linear regression model. b. Applying modeling diagnostics to select the most useful linear model. c. Understanding interaction terms.	ISL Ch. 3.1-3.3, 3.4, 3.6 CAS Ch. 2.4.2, 2.4.3 MASS Ch. 6.1-6.6

©2021, The CAS Institute, All Rights Reserved
The CAS Institute, 4350 North Fairfax Drive, Suite 250, Arlington, VA 22203
TheCASInstitute.org

A4. Generalized linear models	a. Understanding the assumptions of a GLM. b. Understanding link functions and their role in GLMs. c. Understanding the over-dispersion in binomial and Poisson GLMs.	MASS Ch. 7.1-7.5 ISL Ch.4.1-4.4, 4.6
A5. Model selection in classic models	a. Understanding and applying stepwise forward and backward variable selection based on the AIC and BIC to select the optimal sub-model. b. Applying the F-statistic/likelihood ratio test in model selection of nested models	CAS Ch. 4.2.3-4.2.4 MASS Ch. 7.2, 7.4 ISL Ch. 6.1
A6. General Insurance pricing modeling	a. Understanding claim count and claim severity related to premium setting. b. Applying GLMs to model claim frequency. c. Applying GLMs to model claim severity. d. Applying double GLMs to model premium. e. Applying Tweedie GLMs to model aggregate losses.	CAS Ch. 14.1-14.8
A7. Claim reserving and IBNR modeling	a. Understanding the use of loss triangles in reserving. b. Applying the Mack chain ladder method to model loss triangles. c. Applying GLMs to model a loss triangle. d. Using bootstrapping to estimate the variability of loss reserves.	CAS Ch. 16.1-16.4, 16.6, 16.7

©2021, The CAS Institute, All Rights Reserved
The CAS Institute, 4350 North Fairfax Drive, Suite 250, Arlington, VA 22203
TheCASInstitute.org

B. Machine Learning Models and Their Application

Weight for Section B: 40-60 percent

We begin with a background on statistical learning followed by classification methods. We follow that 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 both non-linear effects and interactions. We cover bagging, random forests, and boosting illustrate various model- averaging strategies. Finally, we conclude with a discussion of unsupervised learning, including applications. All the insurance application examples in section A (Classical Models) can be addressed with all the machine learning models in section B. In practice, it is not unusual to compare the results of both types of approaches to determine which provides the best model for the given business application.

Topics	Learning objectives	Reference
B1. Statistical learning	a. Understanding the tradeoff between prediction accuracy and model interpretability and bias/variance tradeoff. b. Assessing model accuracy (i.e., MSE, error rate, training error)	ISL Ch. 2.1-2.2
B2. Classification method and its application	a. Applying multiple logistic regression in prediction. b. Applying LDR, QDA and KNN in prediction. c. Comparing logistic regression, LDA, QDA and KNN prediction results	ISL Ch. 4.1-4.7 MASS Ch. 12.1-12.2

©2021, The CAS Institute, All Rights Reserved
The CAS Institute, 4350 North Fairfax Drive, Suite 250, Arlington, VA 22203
TheCASInstitute.org

B3. Resampling method and model selection.	a. Understanding bootstrapping, leave-one-out cross validation, k-fold cross validation and test sample. b. Use LRT to select the best smoothing parameters and predictor variables. c. Use complexity parameters and cross-validation to select optimal sub-models.	ISL. Ch. 5.1-5.4 MASS Ch. 8.8, 6.6
B4. Linear, shrinkage and regularization methods	a. Selecting the tuning parameters for both lasso and ridge regression. b. Applying lasso and ridge regression to make predictions.	ISL Ch. 6.2, 6.5. 6.6 CAS Ch. 4.3
B5. Semi-parametric models	a. Applying GAM, Additive models, nonlinear models to model real continuous response variable with both categorical and real continuous variables.	MASS Ch. 8.8 ISL Ch. 7.1-7.9
B6. Tree based methods	a. Understanding tree-based regression models. b. Understanding bagging, boosting and random forests. c. Candidates need to be able to use the Tree libraries and functions such as tree, rpart, and party, but **do not need to be able to do trees from first principals.**	MASS Ch. 9.1-9.3 CAS Ch. 4.4-4.5 ISL Ch. 8.1-8.4
B7. Unsupervised learning and its application	a. Applying principal components Analysis to model clusters. b. Applying k-means to model clusters.	ISL Ch.10.1-10.5 MASS Ch. 11.2

©2021, The CAS Institute, All Rights Reserved
The CAS Institute, 4350 North Fairfax Drive, Suite 250, Arlington, VA 22203
TheCASInstitute.org

All packages and functions used in the assigned readings will be examined except where exceptions have been noted **by the use of bold text above**. Candidates will be allowed to use other R libraries (such as dplyr and caret) and can request additional libraries be made available for use during the exam provided such requests are made via email to iCAS Director Amy Brener, abrener@thecasinstitute.org, a minimum of 30 days prior to the exam date.

References

ISL-
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer,

CAS-
Arthur R, Charpentier (2015). *Computational Actuarial Science with R*. Boca Raton CRC Press.

MASS-
Venables, W. N., Ripley, B. D., & Venables, W. N. (2002). *Modern Applied Statistics with S*.

An internet search with provide options for both purchase of physical copies as well as free downloads.

©2021, The CAS Institute, All Rights Reserved
The CAS Institute, 4350 North Fairfax Drive, Suite 250, Arlington, VA 22203
TheCASInstitute.org