MS in Applied Statistics: Study Guide for the Data Science concentration Comprehensive Examination

The Part II comprehensive examination is a three-hour closed-book exam that is offered on the second Saturday of the autumn and spring quarters.

This exam covers materials in MAT 456, and MAT 449 or MAT 450.

.

1. MAT 456 Applied Regression Analysis

Student will be examined in the following broad topics:

- Simple linear regression
- Inferences about regression parameters
- derive simple linear regression parameter estimates
- Diagnostics for regression models
- Estimation for parameters
- Model selection and validation
- Analyze and interpret printouts for proposed models and test them for fit.
- Read the outputs from SAS
- When to use transformations

The afore-mentioned topics represent the first 11 chapters of the required textbook "Applied Linear Regression Models" by M. Kutner, C. Nachtsheim, and J. Neter, 4th edition.

The level of difficulty of the exam questions will be similar to that of your exams questions, homework assignments in the textbook and examples in lecture notes.

Specific references to the lecture notes are given below:

1. **Chapter 1: Linear regression with one predictor variable**
   - Introduction to linear regression
2. **Chapter 2: Inferences in regression analysis**
   - Inferences concerning β_1
   - Inferences concerning β_0
   - Some considerations on making inferences concerning β_0 and β_1
   - Interval estimation of $E(Y_h)$
   - Prediction of a new observation
   - Analysis of variance approach to regression analysis
   - General linear test approach
   - Descriptive measures of association between X and Y in the regression model
   - Considerations in applying regression analysis

3. **Chapter 3: Diagnostics and remedial measures**
    - Diagnostics for predictor variable
    - Residuals
    - Diagnostics for residuals and transformations
    - Tests for constancy of error variance

4. **Chapter 4: Joint estimation of β_0 and β_1**
    - Simultaneous estimation of mean responses
    - Simultaneous prediction intervals for a new observation
    - Inverse prediction

5. **Chapter 6: Multiple regression I**
    - General linear regression model in matrix terms
    - Estimation of regression coefficients
    - Fitted values and residuals
    - Inferences about regression parameters
    - Analysis of variance results
    - Estimation of mean response and prediction of new observation
    - Diagnostics and remedial measures

6. **Chapter 7: Multiple regression II**
    - Extra sums of squares
    - Uses of extra sums of squares in tests for regression coefficients
    - Coefficients of partial determination
    - Multicollinearity and its effects

7. **Chapter 8: Regression models for quantitative and qualitative predictors**
    - Polynomial regression models
    - Interaction regression models
    - Qualitative predictors

8. **Chapter 9: Building the regression model I: model selection and validation**
    - Criteria for model selection
    - Automatic search procedures for model selection
    - Model validation

9. **Chapter 10: Building the regression model II: diagnostics**
    - Model adequacy for a predictor variable
    - Identifying outliers
    - Identifying influential cases

10. **Chapter 11: Building the regression model III: remedial measures and validation**
    - Unequal error variance remedial measures
    - Multicollinearity remedial measures
    - Remedial measures for influential cases

2. Mat 449 Statistical Data Management

You should be able to know

- How to write a simple R function
- How to import data from external sources using read.table.. etc
- How use the Infile statement in SAS (common options like dlm, missover…)
- SQL commands to create and modify data set
- How to combining different data set using SQL
- The difference between Macro variable and Macro function

The level of difficulty of the exam questions will be similar to your exam questions, homework assignments and example in lecture notes.

Specific references to the lecture notes are given below:

1. **Lecture 1-2 A very quick overview of R**

- Import data, read.table, read,csv
- Write a code for simple R function
- Loops for and while
- vectors and data frames

2. **Lecture 3. SAS**

- Naming variables
- INFILE statement and important options used with the INFILE statement
- input statement

3. **Lecture 4 SQL part** I

    - Terminology used in Data Step and SQL
    - Basic PROC SQL Steps and the Order of Clauses Within the PROC SQL Select Statement. Review the examples
    - Conditional Operators (Between-and, contain, in, like, wildcard %, …..)

4. **Lecture 5 SQL part II**
    - Subsetting Rows by Calculated Values
    - Column Labels, Column Formats

5. **Lecture 6 SQL part III Combining Tables Horizontally using PROC SQL**
    - combining the small data sets using inner join; outer join: left, right and full
    - COALESCE function in a basic SELECT clause

6. **Lecture 7 SQL part IV Combining Tables Vertically using PROC SQL**
    - combining the results of multiple PROC SQL queries in different ways by using the set operators EXCEPT, INTERSECT, UNION, and OUTER UNION

7. **Lecture 8 SAS Macro**
    - Macro variable definitions and referencing
    - Substituting Text with Macro Variables
    - Simple Macro functions

3. MAT 450 Advanced Statistical Methods

You should be able to know

- How to generate random numbers from given distributions and use acceptance/rejection method (how to find C)
- How to use resampling techniques (bootstrap)
- How to use the optimization methods (find the constraint matrix A, D matrix, d vector ….)
- How to find roots for MLE or any function using Newton Raphson method
- How to compute the transition matrix from a given diagram or vis versa
- How to use the kernel function for density or regression model

The afore-mentioned topics represent the first 8 lecture notes covered in MAT 450.

1. **Lecture 1. Simulation and random number generation**
    - Random number generation using Inversion Method
    - Acceptance-Rejection Methods

2. **Lecture 2-3. Resampling Techniques *Bootstrap estimation and Permutation***
   - Bootstrap small data may be given and you can be asked to illustrate the procedure
   - Difference between bootstrap and permutation test

3. **Lecture 4-5 Optimization Methods**
   - Root finding techniques
   - Nonlinear programming using proc NLP
   - Quadratic problem (find D, d, b0 and A matrix)

4. **Lecture 6-7 Smoothing Methods**
   - Kernel Smoothing
   - Spline Smoothing
   - Density Estimation
   - Estimate regression model using Kernel smoothing

5. **Lecture 8 MCMC**
   - Probability transition matrix
   - Computing Probabilities for Markov Chains
   - Steady-State Probabilities
   - Absorbing Markov Chains