Exam Study Guide **MGSC 310** **Prof. Doosti Fall 2019**
Topics on this list may change.

Concepts

1. Definitions and Statistical Learning (Chapters 1 and 2)
   a. What is f? Why estimate f? How do we estimate f?
   b. Supervised versus un-supervised learning
   c. Model accuracy
   d. Bias-variance tradeoff
   e. Continuous response variable (y) versus classification problems

2. R scripting
   a. Writing useful statistical scripts in R
   b. Using comment lines to format and organize work

3. Descriptive Statistics and Graphics
   a. Interpretation and computation of descriptive statistics: means, medians, standard deviations, covariance, correlations etc.
   b. Histograms and density plots
   c. Boxplots
   d. Scatter Diagrams
   e. Correlation structure

4. Linear Regression (Chapter 3)
   a. Dependent or response variable (y) and independent, explanatory or predictor variables: $x_1, x_2, \ldots x_p$
   b. Standard linear model: $y = \beta_0 + \beta_1 * x_1 + \ldots + \beta_p * x_p + \varepsilon$
   c. Assumptions on the errors, ε
      i. Normally distributed with mean 0 and standard deviation, σ
      ii. Uncorrelated (with each other)
   d. Least squares estimates, $b_0, b_1, \ldots, b_p$ for $\beta_0, \beta_1, \ldots, \beta_p$ respectively, are there that minimizing the error sum of squares, RSS (aka residual sum squares)
   e. Writing the estimated model from the R output
   f. Interpreting the meaning of coefficients from the estimated regression model in the context of the problem.
   g. Evaluating the overall fit: R^2 (R squared) = 1– RSS/TSS where TSS = $\Sigma(y_i - y\_bar)^2$, and RSS = $\Sigma(y_i - y\_hat)^2$.
   h. Evaluating significance of individual variables (t-tests) and p-value.
   i. The standard error of regression aka the residual standard error (RSE)
   j. Use of Indicator/Dummy/Binary variables (remember to drop one level in the factor)
   k. Splitting data into training and validation (test) sets.
   l. Computing/obtaining predicted values using multiple regression
   m. Computing/obtaining residual values
   n. Linear model extensions: adding non-linear terms (squared terms or log transformed)
   o. Potential problems of linear regression:
      i. Collinearity
      ii. Heteroskedasticity vs. Homoskedasticity

5. Classification (Chapter 4)
    a. Why classification, that is, why not regression?
    b. Logistic regression (simple and multiple)
    c. Writing the estimated model from the output of the regression equation
    d. Interpreting the coefficients (in terms of odds)
    e. Evaluating significance of individual variables (t-tests) and p-value.
    f. Making predictions with logistic regression
    g. Model comparison: confusion matrices
    h. Accuracy, true positive rate, and true negative rate (Sensitivity and Specificity)
    i. ROC curve and AUC: Threshold selection and model comparison

6. Resampling Methods (Cross Validation) (Chapter 5)
    a. Why use resampling?
    b. Training data and Validation (holdout) data
    c. Training error and Test (validation) error estimate (based on validation data)
    d. Validation set approach
    e. Leave one out cross validation (LOOCV)
    f. K-fold cross validation

7. Linear Model Selection and Regulation (Chapter 6)
    a. Best subset selection (of variables)
    b. Idea of computational intensity
    c. Stepwise (forward and backward) selection
    d. Model selection in multiple regression using model selection diagnostics:
        i. Adjusted $R^2 = 1 - (1 - R^2) * (n-1)/(n-p-1)$ (maximize)
    e. Shrinkage methods: Ridge regression and Lasso
    f. Choosing best lambda (penalty term) for Ridge and Lasso
    g. Getting the coefficients for a specific lambda.
    h. Selecting the best model