# Reading Questions

*Econ 224*

*Fall 2018*

## Instructions

This document lists the reading assignments for Econ 224 along with the associated reading questions. For each reading assignment there will be an in-class quiz. The dates of these quizzes are listed below. Quiz questions will be randomly selected from the reading questions listed on this document, so if you thoroughly prepare your answers in advance, you will be sure to get 100% on each quiz of the semester. The abbreviation *ISL* refers to "An Introduction to Statistical Learning" by James et al. while *MM* refers to "Mastering 'Metrics" by Angrist and Pischke. Note that a complete answer to each of these questions requires at most a short paragraph, and more typically a few sentences.

### Quiz #1: Thursday, August 30th – ISL 2

1. We use $Y$ to denote the variable we want to predict and $X$ to denote a variable used to predict $Y$. List the different names that ISL uses interchangeably for $Y$. Do the same for $X$.
2. Define reducible and irreducible error. Which of these errors do the authors of ISL say that their book is focused on minimizing?
3. What is a parametric method? In particular, what are the two steps involved in using a parametric method? Give an example.
4. Contrast parametric and nonparametric methods. What is the main advantage and disadvantage of each?
5. What is the difference between supervised and unsupervised learning?
6. What is the difference between regression and classification?
7. Explain the difference between *training* MSE and *test* MSE. Which of these do we want our statistical learning method to minimize?
8. Define bias and variance. How would we expect each of these quantities to change as we increase the flexibility of our statistical learning model?
9. In place of MSE, what measure of prediction accuracy is used in classification problems?
10. What is a Bayes classifier?
11. Is it possible for a statistical learning model to attain an error rate *lower* than the Bayes error rate? Why or why not?
12. Briefly explain the K-nearest Neighbors classifier. What trade-off involved in choosing a value of K?

### Quiz #2: Tuesday, September 4th – MM Intro & 1.1

1. Define the term *ceteris paribus.*
2. What is the "fundamental empirical conundrum" when trying to learn the causal effect of health insurance on health?
3. In the NHIS example, what is the *outcome*, what is the *treatment*, who makes up the *treatment group* and who makes up the *control group*?
4. List some of the major differences in demographic characteristics between the insured and uninsured in the NHIS.
5. Briefly explain the idea of *potential outcomes* using a simple example. What notation do we use to represent these?
6. The difference in average health by insurance can be written as the sum of two terms. What are they? Briefly explain the meaning of each and relate them to the potential outcomes notation.

7. What is the relevance of the LLN for random assignment?
8. Explain the meaning of $E[Y_i|D_i = d]$.
9. Briefly explain how random assignment eliminates selection bias. Explain both in words and using the notation of conditional expectation and potential outcomes.
10. What are the two key findings of the RAND HIE?

## Quiz #3: Thursday, September 6th – MM 1.2 & Appendix

1. What are some limitations of using the results of the RAND HIE to extrapolate to the causal effect of increasing insurance coverage in the US today?
2. What was the OHP lottery and why was it carried out? Why does it provide evidence for the costs and benefits of insurance coverage for the currently uninsured?
3. Summarize the key findings of the OHP lottery.
4. Define the term *unbiased estimator*. Is the sample mean an unbiased estimator of something? If so, what?
5. Write down the formulas for the sample and population variance of $Y_i$. What does each of these measure? What Greek letter do we use to represent the population variance?
6. If we multiply $Y_i$ by a constant $c$, what happens to the variance? What happens to the standard deviation?
7. Define the term *standard error*. In terms of the relevant population parameters and sample size, what is the standard error of the sample mean?
8. Explain the difference between *standard error* and *estimated standard error*.
9. Explain how to construct an approximate 95% confidence interval for a population mean based on the Central Limit Theorem.
10. Write down the formula for the standard error of a difference of sample means from independent populations if: (1) both populations have the same variance, (2) each population has a different variance.

## Quiz #4: Tuesday, September 11th – ISL 3.1-3.2

1. What are the formulas for calculating $\widehat{\beta}_0$ and $\widehat{\beta}_1$ in a *simple* linear regression?
2. Explain the difference between the *population* regression line and the *least squares* regression line.
3. In a simple linear regression, what is the formula for the standard error of $\widehat{\beta}_1$? Based on this formula, how is the standard error related to: $\text{Var}(\epsilon)$, sample size, and the sample standard deviation of $X$?
4. Write the general expression for a linear regression model with an intercept and $p$ predictor variables. What optimization problem does least squares solve to estimate the regression coefficients?
5. Write down the formulas for: residual sum of squares, total sum of squares, and $R^2$.
6. Explain how to carry out an F-test of the null hypothesis that none of the predictors $X_1, \dots, X_p$ is helpful in predicting $Y$.
7. Explain how to carry out an F-test of the null hypothesis that only a particular subset of $q$ out of the total set of $p$ predictors $X_1, \dots, X_p$ is helpful in predicting $Y$.
8. If we add more regressors to our model, what happens to the RSS? What happens to the $R^2$?
9. Write down the formula relating the residual standard error to the residual sum of squares. Does the residual standard error always decrease if we add more regressors to the model? Why or why not?
10. In the linear regression model, what is a source of reducible error? What is a source of irreducible error?

## Quiz #5: Thursday, September 13th – ISL 3.3-3.5

1. Consider the regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $x_i$ is a dummy variable that equals 1 if person $i$ is female. The book describes two ways to code the category *male*: (a) $x_i = 0$ if person $i$ is male, (b) $x_i = -1$ if person $i$ is male. What is the meaning of $\beta_1$ in each case? Does the choice between (a) and (b) affect our predictions?

2. Suppose we wanted to include the categorical predictor `class` in a regression. This predictor takes on one of the following values: *freshman*, *sophomore*, *junior*, or *senior*. Explain how we could use dummy variables to encode `class`. How many dummy variables would we need?
3. Define the term *interaction effect*. How does including an interaction term in a regression relax the additive assumption?
4. What is the *hierarchical principle*?
5. Define *polynomial regression* and give a simple example. How does it extend the linear model?
6. What goes wrong in a regression setting if the error terms are correlated and or have non-constant variance?
7. Define the term *outlier* in the context of linear regression. Briefly explain a way of identifying outlying observations.
8. Define the term *high leverage point*. How is this different from an *outlier*? Why should we worry about high leverage points?
9. What is collinearity and why is it a problem? How can we detect it, and how can we address it?
10. Explain the KNN regression method. What trade-off is involved in choosing a value for $K$?
11. If the true form of $f$ is approximately linear, which would we expect to perform better: least squares linear regression or KNN regression? Why?
12. Explain the *curse of dimensionality* as it relates to KNN regression. When would we expect least squares regression to outperform KNN regression?

## Quiz #6: Tuesday, September 18th – MM 2.1-2.2

## Quiz #7: Thursday, September 20th – MM 2.3 & Appendix