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 $X_1, \ldots, X_p$ 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 $X_1, \ldots, X_p$ 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 β_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

1. Let $\bar{Y}_1$ be the average earnings for a large random sample of individuals who graduated from a private college, and let $\bar{Y}_0$ be the average earnings for a large random sample of individuals who graduated from a public college. Does the difference $\bar{Y}_1 - \bar{Y}_0$ estimate the causal effect of attending a private school? Why or why not?
2. Briefly explain the logic of the Dale and Krueger matching strategy for estimating the payoff from attending a more selective college.
3. In the Dale and Krueger example, what is the *outcome variable*, what is the *treatment variable*, and what are the key *control variables*?
4. The *College and Beyond* survey includes data for over 14,000 students, but Dale and Krueger only use 5,583 students in their selectivity-group analysis. Explain why.
5. Based on the results of Table 2.2, how does the estimated effect of attending a private school vary depending on whether or not selectivity-group dummies are included in the regression?
6. Explain the idea behind the "self-revelation" model. How to the results of this model compare to those from the Barron's matches?
7. The regression results in Table 2.4 replace P_i with the average SAT score of students at the school attended. Summarize the results of these regressions.
8. Based on the Dale and Krueger paper, what should we conclude about the causal effect of attending a more selective college?

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

1. Write down the omitted variables bias formula and explain its meaning in words using a simple example.
2. Define the term *conditional expectation function*.
3. Consider two cases: (a) $\mathbb{E}[Y|X_1, ..., X_k]$ is linear, and (b) $\mathbb{E}[Y|X_1, ..., X_k]$ is non-linear. In each case, what does a linear regression of Y on $(X_1, ..., X_k)$ yield?
4. Covariance has three important properties. What are they?
5. Suppose we want to minimize the function $RSS(a, b) = \mathbb{E}[(Y_i - a - bX_i)^2]$. What values should we choose for a and b?
6. What two properties do the residuals in a linear regression satisfy?
7. Consider the regression $Y_i = \alpha + \beta X_i + \epsilon$ where X_i is a *dummy variable*. Express β in terms of conditional expectations and explain the meaning of your expression in words.

8. Write down the *regression anatomy* formula, and explain how it allows us to interpret the coefficient β_k in a multivariate regression in terms of a simple bivariate regression.
9. Let log denote the natural logarithm. What is the interpretation of the slope coefficient β in the regression $\log Y_i = \alpha + \beta X_i + \epsilon_i$?
10. Define the term *heteroskedasticity*.

Quiz #8: Tuesday, September 25th – ISL 4.1-4.3

1. Why do we use methods other than linear regression in classification problems?
2. Suppose we use logistic regression with a single predictor X to predict the probability $p(X)$ that $Y = 1$. Write down the *logistic function* that expresses $p(X)$ in terms of X and the coefficients β_0 and β_1.
3. Define the term *odds*. What range of values can the odds take?
4. Explain the meaning of β_0 and β_1 in a logistic regression that uses a single regressor X to predict the probability $p(X)$ that $Y = 1$. If X increases by one unit, how do the log-odds change?
5. In a logistic regression that uses a single predictor X to predict the probability $p(X)$ that $Y = 1$, how do we estimate the coefficients β_0 and β_1?
6. In a logistic regression that uses a single predictor X to predict the probability $p(X)$ that $Y = 1$, what is the meaning of the null hypothesis $H_0 : \beta_1 = 0$?
7. Write down the expression for the *log odds* in a logistic regression that uses $X_1, ..., X_p$ to predict the probability that $Y = 1$.

Quiz #9: Thursday, September 27th – ISL 4.4-4.5

1. Explain how to use Bayes' Theorem to classify an observation into one of $K \geq 2$ classes.
2. Write down the formula for the probability density of a normal random variable with mean μ and variance σ^2.
3. Write down the *discriminant function* for LDA in the simple case where $p = 1$. Explain the meaning of each symbol in this expression, and explain intuitively how LDA works.
4. Define the term *confusion matrix*.
5. Define the terms *sensitivity* and *specificity*.
6. Why might a credit company prefer *not* to use the Bayes classifier to predict credit card default?
7. Define the term *ROC curve*.
8. Define the term *AUC*. Which indicates a better classifier: high AUC or low AUC? Briefly explain.
9. Briefly explain the difference between LDA and QDA. What trade-off is involved in choosing between them?
10. Briefly compare and contrast LDA, logistic regression, and KNN.

Quiz #10: Tuesday, October 9th – MM 3.1

1. Suppose that we find that non-white students at KIPP schools have higher test scores than non-white students at other nearby schools. Why might this fact *fail* to convince us that attending a KIPP school improves test scores?
2. Explain in words the flowchart from Figure 3.1 detailing the application and enrollment data from the KIPP lotteries. (You do not need to memorize the exact numbers in the flowchart for the quiz: I just want you be able to explain the basic qualitative pattern.)
3. If we were to compare the characteristics of students who were offered a place at KIPP to those who were not offered a place, would we expect to find substantial differences? Why or why not?
4. In the KIPP example, the instrumental variable is a dummy variable that equals one if a given applicant was offered a place at KIPP. What three properties do we need to assume that this instrument satisfies?
5. Explain in words how the IV estimator of the causal effect of attending a KIPP is constructed (Figure 3.2).

6. Use mathematical notation to define each of the following: *first stage*, *reduced form*, and *Local Average Treatment Effect (LATE)*.
7. List and define each of the four "types" of individuals that are relevant for interpreting LATE. (Table 3.2)
8. Define the term *monotonicity* in the context of LATE.
9. What does the LATE theorem tell us?
10. Define TOT using potential outcomes notation. How and why can it differ from LATE?

No Quiz on Thursday, October 12th!

No Quiz on Tuesday, October 16th!

Quiz #11: Thursday, October 18th – MM 3.2, 3.3, and Appendix

1. What was the purpose of the Minneapolis Domestic Violence Experiment (MDVE) and how was the experiment carried out?
2. Given that the MDVE was a randomized experiment, why do we use IV estimation here?
3. Define the term Intent to Treat (ITT).
4. When is TOT equal to LATE?
5. Describe the results from the regression of adult firstborns' highest grade completed on family size from the ALS study. Why might these results *not* imply a causal relationship?
6. Explain the IV results from the ALS study using twin births as an instrumental variable. Why and how do they differ from the regression results in question 6?
7. Explain the idea behind the sibling sex composition instrument in the ALS study. What advantages does this instrument have over the twin births instrument?
8. How does 2SLS generalize IV?
9. Explain in words and equations how 2SLS works: in particular describe both the first and second stage regressions.
10. Why doesn't "manual 2SLS" give the right standard errors?

Quiz #12: Tuesday, October 23rd – ISL 5.1, 5.2

1. Define *model assessment*.
2. Define *model selection*.
3. Define the terms *training set* and *validation set* (aka *hold-out set*).
4. Briefly explain the *validation set approach*. How does it work, and what problem is it intended to solve?
5. Briefly explain the leave-one-out cross-validation (LOO-CV) procedure.What is an advantage of LOO-CV relative to the validation set approach?
6. Briefly explain *k-fold cross-validation* procedure. What is the computational advantage of *k-fold* CV relative to LOO-CV? Is there any setting in which there is no advantage?
7. What trade-off is involved in choosing k for k-fold CV?
8. Briefly explain the bootstrap procedure for estimating the standard error of $\widehat{\alpha}$ in the portfolio example from Section 5.2

Quiz #13: Thursday, October 25th – ISL 6.1

1. Explain the *best subset selection* algorithm.
2. Suppose we have p predictor variables. How many models do we need to fit to carry out *best subset selection*.
3. Explain the *forward stepwise selection* algorithm.
4. Explain the *backward stepwise selection* algorithm.

5. What are the advantages and dis-advantages of forward and backward stepwise selection vis-a-vis best subset selection?
6. What are the two general appropaches for estimating the test error used in model selection?
7. Give the formulas for AIC and BIC in a least squares problem. Compare and contrast them.
8. Explain how cross-validation can be used for model selection.

Quiz #14: Tuesday, October 30th – ISL 6.2, 6.4

1. The *ridge regression* coefficient estimates are chosen to minimize a function with two terms. Write down the function and explain the meaning of each of these two terms.
2. In ridge regression we *do not* apply the shrinkage penalty to the intercept β_0. Why not?
3. If $\mathbf{v}$ is a vector with elements $v_1, ..., v_m$ what is the formula formula for the ℓ_2 norm $||\mathbf{v}||_2$ of $\mathbf{v}$? What is the meaning of $||\mathbf{v}||_2$?
4. Briefly explain how the ridge tuning parameter λ affects the ridge regression coefficient estimates. In particular, as we increase λ, what happens to the ℓ_2 norm of the ridge regression coefficient vector $\widehat{\beta}_\lambda^R$?
5. Suppose that I run an ordinary least squares linear regression to predict height in centimeters y using weight in kilograms x. The estimated slope coefficient is 0.5. If I changed the units of x to grams, how would the estimated slope coefficient change? What is the name of the property of least squares regression that you used to answer this question?
6. It is recommended to apply ridge regression only *after* standardizing the predictors X_j. Why? What does it mean to standardize a predictor?
7. When and why would we expect ridge regression to improve over least squares?
8. What advantages does ridge regression have over best subset selection?
9. Explain the difference between the optimization problem that the lasso coefficient estimates $\widehat{\beta}_\lambda^L$ solve compared to the optimization problem that the ridge coefficient estimates $\widehat{\beta}_\lambda^R$ solve.
10. There is an important qualitative difference between the way that lasso and ridge shrink coefficients towards zero. What is it?
11. There is an alternative formulation of ridge regression and lasso that makes it easy to compare them to best-subsets regression. Write down this formulation, and explain the relationship.
12. When would we describe a dataset as *high-dimensional*? What are some of the challenges of working with high-dimensional datasets?

Quiz #15: Thursday, November 1st – ISL 8.1, 8.2.1, 8.2.2

1. Explain how to construct a regression tree via *recursive binary splitting*. How do we use the resulting tree to make predictions?
2. Explain *cost complexity pruning* aka *weakest link pruning*.
3. Sketch the algorithm for building a regression tree that combines recursive binary splitting, cost complexity pruning, and cross-validation.
4. How does a classification tree differ from a regression tree?
5. What are the two preferred criteria for making binary splits in a classification tree? Give the formula for each.
6. When would we expect regression trees to out-perform linear regression and vice-versa?
7. Explain how *bootstrap aggregation* aka *bagging* is applied to regression trees.
8. What is an *out-of-bag* (OOB) observation? How can OOB observations be used to estimate the test error of a bagged model without the need for cross-validation?
9. What is one advantage and one disadvantage of bagging when applied to regression trees?
10. Briefly explain the *random forest* algorithm. How does it differ from bagging? What is the rationale behind this difference?

Quiz #16: Tuesday, November 6th – MM 4.1

1. What is the "seemingly paradoxical idea" upon which regression discontinuity (RD) is based? How does this idea apply in the case of the minimum legal drinking age?
2. Define the term "running variable" in RD analysis.
3. What are the "two signal features" of RD designs?
4. What is the difference between "sharp" and "fuzzy" RD?
5. The validity of causal effects estimates from RD relies upon our willingness to do what?
6. Why might a simple linear RD specification fail to produce reliable causal estimates? How can we address this problem?
7. In the MLDA example, why might we expect a change in slope at the cutoff?
8. Explain the statement "estimates away from the cutoff constitute a bold extrapolation" in the context of the MLDA example. Where would we expect RD estimates in this example to be most reliable? Why?
9. Summarize the results of the MLDA example.

Quiz #17: Thursday, November 8th – MM 4.2

Quiz #18: Tuesday, November 13th – MM 5.1

Quiz #19: Thursday, November 15th – MM 5.2