ECON 453
University of Arizona
Dr. Satheesh Aradhyula
Final Exam: Part 2 (50 points)

Instructions:

- Answer all questions in the boxes provided.
- Upload two files to d2L:
  (1) This file with your answers written. Print this file, write your answers, scan it, and then upload it to d2L as a pdf file. This file should have 6 pages.
  (2) In a separate file, upload your well labeled R code that includes R output (for all questions) as well. Without this, you will not get credit for your answers.
- This is an open book exam. You may use your notes, any book, and any publicly available online resource. However, you are expected to work independently. Do not discuss or communicate with others about the exam.
- Print the following honor code, sign, and submit it along with your exam:

I hereby affirm that the work done on this exam is entirely my own and I have not given nor received aid from any other individual in this regard. I understand that discussing the exam, participating either passively or actively in any chat or communication, obtaining help from others whether solicited or unsolicited are not permitted and may result in a severe reduction in my grade for this exam.

Signed  Date: 12/10/2022

Print: Darien Oliva

Note: Part 1 of the Final exam (which has 25 multiple choice questions) is separate from this. Part 1 should be taken separately on d2l. Go to quizzes section to take that exam. That exam must be completed in 70 minutes.

Question 1. Use data in sheet *Housing_Starts* in the file *final_exam_part2_data_*version_a.*xslx*. Attach your R code with output so we can locate your estimates if needed. Attach your R code with output.

1. Housing starts are the number of new residential construction projects that have begun during a month. It is a leading indicator of economic strength. Data on monthly housing starts (in 1,000s) in the US from January 2011 to November 2018 are available in the file *final_exam_part2_data.xslx*, sheet *Housing_Starts* on d2L. Use this data for answering the following questions.

1a) Fit a linear trend model (without monthly dummies) in R. Note: This model will have only two beta coefficients. Write the estimated model in the box below:

Write your answer here →	model1 <- lm(data1$Hstarts ~ data1$Date, data=1)

Comment on the significance of the trend in this model. Justify your answer using p-value(s).

Write your answer here →	p-value = $2.2e^{-16}$ which indicates the relationship between Date and housing Starts is statistically significant.

Using the estimated model, forecast housing starts for January 2019. You may do this in R.

Write your answer here →	Based on the forecasting Graph for Housing Starts, The amount of housing starts in Jan 2019 is approximately 88.

1b) Using the same data estimate a linear trend with seasonality (i.e., tread and monthly dummies). In your estimation, use January as the reference month. Note: This model will have 13 beta coefficients.

Based on the estimated regression results, housing starts for which month(s) are not significantly different from January?

Write your answer here →	December, November, February, March

* See p-values!

Using the estimated model, forecast housing starts for January 2019. You may do this in R.

Write your answer here →	Approximately 88.

According to the estimated model, housing starts are the highest in which month?

Write your answer here →	July

According to the estimated model, housing starts are the lowest in which month?

Write your answer here →	February

1c) Between the linear trend model in (1a) and linear trend model with seasonality in (1b), which is the preferred model? Justify your answer using results from R.

Write your answer here →	From examining both models, R shows that the linear trend model without seasonality is a better fit. This is because the Adjusted R-Squared is higher in the first model, and the P-value is lower in the first model, both showing us that the first model is a better fit.

Question 2. Use data in sheet *gym* in the file *final_exam_part2_data_*version_a.*xslx*. Attach your R code with output so we can locate your estimates if needed.

2. A local gym sells memberships on an annual basis. The manager is concerned about the attrition rate at her gym. She would like to identify the profile of members who renew their annual membership. Data in membership renewal (Renewed = 1 if the member renewed the membership, 0 otherwise), member's age, member's income (in $1,000s) and whether member joined on a single or family plan.

2a) Fit a linear probability model (LPM) using *Renewed* as the response variable and *Age*, *Income* and *Single* (equals 1 if on a single plan, 0 otherwise) as predictor variables. Write the estimated model in the box below:

Write your answer here →	model2 <- lm(Renewed ~ Age + Income + D_Plan, data = data2)

Using LPM, predict the probability of renewing for a 40-year-old with an income of $80,000 and on a single plan.

Write your answer here →	Probability = 0.3736

* See R-code.

2b) Fit a logistic regression model using *Renewed* as the response variable and *Age*, *Income* and *Single* (equals 1 if on a single plan, 0 otherwise) as predictor variables. Using estimated logistic model, predict the probability of renewing for a 40-year-old with an income of $80,000 and on a single plan. You may want to use R for these computations.

Write your answer here →	Probability = -0.8084

* See model in R code!

Would you classify "40-year-old with an income of $80,000 and on a single plan" individual as a renew or no-renew?

Write your answer here →	no - renew

Calculate odds of a "40-year-old with an income of $80,000 and on a single plan" individual renewing his/her/their membership? Show your code in R.

Write your answer here →	odds = -0.4470

* See R-code

2c) Based on prediction ability and other qualities which model (LPM or Logistic) is preferred? Make needed computations in R to help you decide. Summarize your findings (along with relevant numbers from R output) in the box below.

Write your answer here →	From examining both models, R shows that the first model (the LPM) is preferred because with that model a lower p-value was generated p-value → 2.2 e-16

Question 3. Use data in sheet *Houses* in the file *final_exam_part2_data_*version_a.*xslx*. Attach your R code with output so we can locate your estimates if needed.

3. A realtor is analyzing the relationship between the sale price of a home (price in $) its square footage (Sqft), the number of bedrooms (Beds), the number of bathrooms (Baths) and a colonial dummy variable (Colonial = 1 if a colonial-style home, 0 otherwise).

3a) Estimate a linear model where Price is the dependent variable and Sqft, Beds, Baths, and Colonial are explanatory variables. Write the estimated model in the box below:

Write your answer here →	model3 <- lm(price ~ Sqft + Beds + Baths + colonial, data = data3)

3b) Interpret the estimated coefficient attached to Sqft.

Write your answer here →	It's 91.68, meaning on average, for each additional square foot in a home the price is expected to increase by $91.68

* All Else Constant.

* This is statistically significant, as indicated by p-value.

3c) Interpret the estimated coefficient attached to Colonial.

Write your answer here →	It's 74,557.88, meaning colonial homes are expected to sell for $74,557.88 than non-colonial homes, on average, and while controlling other variables. * Statistically significant indicated by p-value.

3d) Comment of the significance (or lack there of) of variables. Use p-values where needed.

Write your answer here →	Variables Sqft, Bath, and colonial are all statistically significant as indicated by low p-values. This shows a strong relationship between these variables and sales price. Beds is not statistically significant as indicated by high p-value. Meaning there is not a strong relationship between number of bedrooms and sales price.

* See code for P-values.

3e) Construct a 95% confidence interval for expected price for a 2500 square-foot colonial-style home with three bedrooms and two bathrooms.

Write your answer here →	lower: $466,436.40	fit: $615,990.90	upper: $765,545.40

> predict.lm(model3, data.frame(Sqft = 2500, Beds = 3, Baths = 2, Colonial = 1), interval = "predict", level 0.95).