

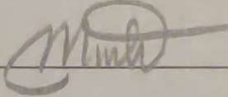
ECON 453  
University of Arizona  
Dr. Satheesh Aradhyula  
Final Exam: Part 2 (50 points)

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: 05/08/2023

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 90 minutes.

**Question 1.** Use data in sheet *Housing\_Starts* in the file *final\_exam\_part2\_data.xlsx*. Attach your R code with output so we can locate your estimates if needed. Attach your R code with output. I downloaded these data from: [https://www.census.gov/construction/nrc/historical\\_data/index.html](https://www.census.gov/construction/nrc/historical_data/index.html)

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 December 2022 are available in the above file in the 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 →

$$\text{hstarts} = 56.12329 + 0.55532 \text{trend-monthly}$$

(trend is the month starting from January 2011)

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

Write your answer here →

The trend is significant with p-value  $< 2e-16 < 0.05$   
(at 5% significance level)  
(and F-stat has p-value  $< 2.2e-16 < 0.05$ )

Using the estimated model, forecast housing starts for the current month (May, 2023). You may do this in R.

Write your answer here →

138.8660

- 1b) Using the same data estimate a linear trend with seasonality (i.e., trend 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?

$P < 0.05$

Write your answer here →

February (0.82), November (0.14), December (0.53)

Using the estimated model, forecast housing starts for the current month (May, 2023). You may do this in R.

Write your answer here →

146.1722

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

Write your answer here →

December (coefficient = -2.19)

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

Write your answer here →

June (coefficient = 24.25)

- 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 →

The linear model with seasonality in 1b should be preferred over 1a, as it has larger R-square ( $0.87 > 0.75$ ), larger adjusted R-square ( $0.86 > 0.75$ ), lower RSE ( $10.02 < 13.41$ ). The p-values of the new variables in 1b are also  $< 0.05$  for the majority, which means they are statistically significant.

( $\Rightarrow$ ) 1b model has better explanatory power)



Question 2. Use data in sheet gym in the file final\_exam\_part2\_data.xlsx. 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 →	$\text{Renewed} = -0.338 + 0.005 \text{ Age} + 0.008 \text{ Income} - 0.129 \text{ Single}$
--------------------------	---

Using LPM, predict the probability of renewing for a 50-year-old with an income of \$70,000 and on a family plan.

Write your answer here →	0.4783 (47.83%)
--------------------------	-----------------

- 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 50-year-old with an income of \$70,000 and on a family plan. You may want to use R for these computations.

Write your answer here →	0.4959 (49.59%)
--------------------------	-----------------

Would you classify "50-year-old with an income of \$70,000 and on a family plan" individual as a renew or no-renew?

Write your answer here →	No-renew (as 0.4959 < 0.5) ↳ threshold
--------------------------	---

Calculate odds of a "50-year-old with an income of \$70,000 and on a family plan" individual renewing his/her/their membership? Show your code in R.

Write your answer here →	0.9837
--------------------------	--------

- 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 →	Base solely on prediction ability, the LPM has a better accuracy than the Logistic model (79.5% > 78.5%) and thus should be preferred. However, on other qualities, Logistic Model may be preferred as it only produces results from 0 to 1, while LPM ranges from $-\infty$ to $\infty$ .
--------------------------	--

(any values outside  $[0,1]$  is not a good probability value).

Question 3. Use data in sheet *Houses* in the file *final\_exam\_part2\_data.xlsx*. 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 Ranch dummy variable (Ranch = 1 if a Ranch-style home, 0 otherwise).

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

Write your answer here →	$\text{Price} = 227635.79 + 98.24 \text{ Sqft} + 359.28 \text{ Beds} + 67373.21 \text{ Baths} - 66711.79 \text{ Ranch}$
--------------------------	---

- 3b) Interpret the estimated coefficient attached to Sqft.

Write your answer here →	<p>An 1-unit (square ft) increase in the house's square footage results in a \$98.24 increase in house price, holding other variables constant.</p>
--------------------------	---

- 3c) Interpret the estimated coefficient attached to Ranch.

Write your answer here →	<p>A Ranch-style home <sup>in price</sup> is \$66711.79 less than a non-Ranch-style one, keeping other variables constant.</p>
--------------------------	--

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

Write your answer here →	<p>Sqft, Baths, Ranch have p-values of 0.0009, 0.0017, 0.0069 respectively. As those are &lt; 0.05 at 5% significant level these 3 are statistically significant. Beds has a p-value of 0.9798 &gt; 5%. Thus at 5% significant level, beds is not statistically significant variable.</p>
--------------------------	---

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

Write your answer here →	<p>[495527.8, 589169.1]</p>
--------------------------	-----------------------------