

Spring 2025 ENVECON/IAS 118 - Introductory Applied Econometrics Problem Set 3

Due on Gradescope at Midnight, March 4

Submission Instructions

For the purposes of this class, we will be using Berkeley's *Datahub* to conduct our analysis remotely using these notebooks. The data files can be accessed directly through *Datahub* and do not require you to install anything on your computer.

If instead you already have an installation of R/RStudio on your personal computer and prefer to work offline, you can download the data for this assignment from bCourses (Make sure to install/update all packages mentioned in the problem sets in order to prevent issues regarding deprecated or outdated packages).

To submit your completed assignment, go to the file dropdown menu and select the "Save and export notebook as" dropdown menu. In this menu make sure to select "PDF", "Webpdf" or "PDF via Chrome" (if that option appears instead).

The figures in the problem statement may not show up in the pdf you generate. Once you have downloaded this pdf, make sure it shows all your answers (including code and output). Please **do not print the entire dataset in your submission**. If you viewed the data earlier, remove that line of code and re-run the code cell (as datasets get bigger this adds many pages to pdf submissions and increases the likelihood we miss your answer).

Upload your submission to Gradescope:
<https://www.gradescope.com/courses/927499>

*Note: Coding Bootcamp [Part 3](#) and [Part 4](#) covers all necessary R methods.

Preamble

When writing R code, it's a good habit to start your notebooks or R scripts with a preamble, a section where you load all necessary packages, set paths or change the working directory, or declare other options.

Use the below code cell to load in packages you will use throughout the problem set (at least `haven`, `tidyverse`, and `ggplot2` this week).

*Note: **never** try to install packages on Datahub. All packages that you need are already installed and can be loaded immediately using the `library()` function. Attempting to install packages will create conflicts with the package versions on the server and potentially corrupt your notebook.*

```
In [46]: # insert your code
```

Question 1: Relationship between Housing Prices (in USD) and Characteristics of US Cities.

This exercise is to be completed using R. We will establish a simple linear relationship between **housing prices and characteristics of cities** in a sample of cities. This is called a hedonic regression, relating price to characteristics. The idea is that if a characteristic is valued in a city, demand for housing increases as people move there, and then housing price increases, all else constant. Vice versa, if people do not value a characteristic, like crime, for example.

Note: in economics, log always refers to the natural log, $\ln()$.

Data description

We will use September 2021 data from Zumper on one-bedroom apartment prices and 2019 data from the FBI on crime for US cities and other characteristics of US cities, such as number of bars, air quality index, wealth of the city measured by GDP, population, whether the city has a winning record majors sports team, as well as the number of sports teams in the major basketball, baseball and American football leagues. The data has 96 cities.

Readme for data variables, several sources - collected by Villas-Boas, Fall 2021

Variable name	Definition	Source
city	City name	
state	State name	
pricesept2021	One bedroom housing price, in USD	www.Zumper.com
successteams	Dummy variable =1 if at least one NBA, NFL, or MLB team in a city had a winning record last season (2020 season), =0 otherwise	Google search

Variable name	Definition	Source
violentcrime2019	Violent crimes (in thousands)	FBI
numberbars	Number of bars, count	www.yellowpages.com
aqi2020	Annual 2020 air quality index (AQI)	EPA
gdp	Gross domestic product (billion \$)	BEA
popul2019	2019 population (in thousands of people)	FBI
nteam	Number of major professional sports teams	Google search

0. Data Setup (Ungraded) The dataset is in Stata format (.dta) and was created for the purpose of this problem set only. It is available on bcourses and Datahub and is called **Villas-Boas_2025pset3.dta**.

Read the data into R using `my_data <- read_dta("Villas-Boas_2025pset3.dta")` and create a new variable, *gdpPc*, as the GDP per capita, defined as `(gdp)/(popul2019)`.

Explore each of the variables, including their summary statistics, by using the function `summary()`

In [28]: `# insert your code`

1. First, generate a variable called **goodAQI** that is equal to one for cities with `aqi2020 <= 50`, and equal to zero otherwise. Compute an estimate for the mean of the housing prices for the goodAQI group (**goodAQI=1**) in the data frame. Construct a 95% confidence interval for this mean. Give an interpretation of these results in a sentence.

Use `mean()`, `sd()`, `qt()`, and/or `qnorm()`, to get the necessary information to construct the CI (do not use any canned functions to calculate the confidence interval). Make sure to show all of your intermediate steps and calculations in your answer.

In [29]: `# insert your code`

➡ Type your interpretation here.

2. Next we will compare housing across the two groups. Let D be the difference in prices between the cities with good AQI (**goodAQI=1**) and not good AQI (**goodAQI=0**) groups. State an estimator \hat{D} for D and use the estimator to compute an estimate of D . Compute a standard error for \hat{D} . Derive a 90% confidence interval for D and interpret it in one sentence.

Use `mean()` , `sd()` , `qt()` , and/or `qnorm()` , to get the necessary information to construct the CI (do not use any canned functions to calculate the confidence interval). Make sure to show all of your intermediate steps and calculations in your answer.

In [30]: `# insert your code`

→ Type your answer here.

3. Next, we will test whether the average housing price **pricesept2021** for the good AQI city group is statistically different at the 10% significance level ($\alpha = 0.1$) from average housing values in the not good AQI city group. That is, in terms of the hypotheses, test the null hypothesis that average housing prices are equal across these two groups against the alternative hypothesis that the average prices are not equal. Make sure to follow the 5 step-procedure for hypothesis testing, including interpreting your result in one sentence.

Use `mean()` , `sd()` , `qt()` , and/or `qnorm()` , to get the necessary information to conduct the hypothesis test (do not use any canned functions to conduct the hypothesis test). Make sure to show all of your intermediate steps and calculations in your answer.

3.i: Step 1 State the null (H_0) and alternative (H_A) hypotheses.

→ Type your answer here.

3.ii. Step 2: Calculate the test statistic

In [31]: `# insert your code`

→ Type your answer here.

3.iii. Step 3: Find the critical value

In [32]: `# insert your code`

→ Type your answer here.

3.iv. Step 4: Define the rejection rule

→ Type your answer here.

3.v. Step 5: Decide and interpret

→ Type your answer here.

4. Let's now look at air quality in the data more closely. **The U.S. AQI is EPA's index for reporting air quality.** Draw a histogram for **aqi2020** and add a vertical red line at the EPA standard for Spare the Air Day AQI = 100 and a green line at AQI = 50. (<https://www.airnow.gov/aqi/aqi-basics/>)

For example, in the Bay Area, a Spare the Air Alert is called when air quality is forecast to be unhealthy, or above 100 in the AQI, in any one of the reporting zones. An alert may span over two days if air quality is expected to remain unhealthy for prolonged periods. If air quality is unhealthy in the Bay Area, it is almost always because of two kinds of air **pollutants**: **Ozone** and **fine particulate matter, or PM2.5**.

Hint: see the "Lines" section of Coding Bootcamp Part 4

In [33]: `# insert your code`

5. (a) Regress **pricesept2021** on a constant, **successsteams**, **violentcrime2019**, **aqi2020**, **numberbars**, **gdpPc**. (b) Generate a series of the predicted values of price and plot those against the price data series: What do you see in terms of fit?

In [34]: `# insert your code`

→ Type your answer here.

6. What is the percent variation of housing prices that the model is explaining, and what percent is the model **NOT** explaining?

→ Type your answer here.

7. Compute the residuals series and plot the residuals on the vertical axis against **gdpPc** in the x axis, using `ggplot()`. When plotting, exclude the outlier city with `gdpPc > 6`, by setting the ggplot scale limits as follows: `lims(x = c(0, 6), y = c(-1000, 1500))`.

Is the constant variance assumption for the residuals valid or not for different levels of GDP per capita (*gdpPc*) when you look at the scatter plot of the estimated residuals?

In [35]: `# insert your code`

→ Type your answer here.

8. Using the triple Sign Size Significance (SSS), let's interpret two of the coefficients from the model in Question 5.

(a) What can you say of the effect of **aqi2020** on housing prices holding other factors constant?

→ Type your answer here.

(b) What about the coefficient on **numberbars**? Use the (SSS) interpretation again.

→ Type your answer here.

9. Estimate the correlation between **gdpPc** and the air quality index **aqi2020** across cities. Consider this information along with the estimated coefficients in Question 5's regression. Without running any additional regressions, what will happen to the estimated coefficient of **aqi2020** if you do not include GDP per capita (**gdpPc**) in the estimated regression in question 5? Go through the Omitted Variable formula and explain briefly.

In [36]: `# insert your code`

→ Type your answer here.

10. Now estimate the model in Question 5 but do not include **gdppc**. What is the new estimate of the coefficient on **aqi2020**, and do you confirm your answer in Question 9?

In [37]: `# insert your code`

→ Type your answer here.

11. What happens to the R squared (R^2) when you do not include the **gdppc** variable in the equation compared to the R squared in Question 5?

In [38]: `# insert your code`

→ Type your answer here.

Question 2: Insurance Takeup

This question is adapted from the following paper (but you do not need to read the paper to complete the assignment):

Cai, Jing, Alain De Janvry, and Elisabeth Sadoulet. "Subsidy policies and insurance demand." American Economic Review 110.8 (2020): 2422-2453.

The data for this exercise comes from households in 134 villages in the Jiangxi province, which is considered a representative sample of rice producers in Jiangxi. Households were asked whether the household head is literate and also whether they took up a weather insurance product in 2011.

The product in this study is an area-yield index weather insurance that covers natural disasters, including heavy rains, floods, windstorms, extremely high or low temperatures, and droughts. If any of these disasters occur and leads to a 30 percent or more average loss in yield in a given area, farmers that take up insurance in that area are eligible to receive payouts from the insurance company.

	Percent choosing Insurance	Total number of respondents
Overall	52.85 %	3474
Literate	53.03%	2492
Not Literate	49.95 %	979

where *literate* is a dummy variable equal to 1 if the household head is literate, and equal to 0 otherwise

Let p be the fraction of respondents that choose to take up the insurance product.

1. Use the survey results to estimate p . Also estimate the standard error of your estimate.

Use `mean()`, `sd()`, `qt()`, and/or `qnorm()` in your answer. (Do not use any canned functions to calculate the test statistic or standard error). Make sure to show all of your intermediate steps and calculations in your answer.

```
In [47]: # insert your code
```

→ Type your answer here.

2. Construct a 95% confidence interval for p . Interpret your results in a complete sentence.

Use `mean()`, `sd()`, `qt()`, and/or `qnorm()`, to get the necessary information to construct the CI (do not use any canned functions to calculate the confidence interval). Make sure to show all of your intermediate steps and calculations in your answer.

```
In [48]: # insert your code
```

→ Type your answer here.

3. Construct a 99% confidence interval for p . Is it larger or narrower than the 95% confidence interval? Why? Explain your reasoning in 1-2 sentences.

Use `mean()` , `sd()` , `qt()` , and/or `qnorm()` , to get the necessary information to construct the CI (do not use any canned functions to calculate the confidence interval). Make sure to show all of your intermediate steps and calculations in your answer.

In [49]: `# insert your code`

→ Type your answer here.

4. Is there statistical evidence that more than 50% of respondents chose the insurance product? Use the 5 steps for hypothesis testing with a 5% significance level.

Use `mean()` , `sd()` , `qt()` , and/or `qnorm()` , to conduct the hypothesis test (do not use any canned functions). Make sure to show all of your intermediate steps and calculations in your answer.

4.i: Step 1 State the null () and alternative () hypotheses.

→ Type your answer here.

4.ii: Step 2: Calculate the test statistic

In [50]: `# insert your code`

→ Type your answer here.

4.iii: Step 3: Find the critical value

In [51]: `# insert your code`

→ Type your answer here.

4.iv: Step 4: Define the rejection rule

→ Type your answer here.


4.v: Step 5: Decide and Interpret

→ Type your answer here.

5. Is there statistical evidence that choosing to take up the insurance product is more likely for respondents that are literate compared to respondents that are not, at the 1% significance level? Explain. (To answer this question use the 5 steps for hypothesis testing).


Use `mean()` , `sd()` , `qt()` , and/or `qnorm()` , to conduct the hypothesis test (do not use any canned functions). Make sure to show all of your intermediate steps and calculations in your answer.

5.i: Step 1: Define the hypotheses

 Type your answer here.


5.ii: Step 2: Calculate the test statistic

In [52]: `# insert your code`


 Type your answer here.

5.iii: Step 3: Find the critical value


In [53]: `# insert your code`

 Type your answer here.

5.iv: Step 4 Define the decision rule

 Type your answer here.

5.v: Step 5: Decide and interpret

 Type your answer here.

Please remember to submit your Jupyter Notebook displaying all codes and output.

Link to the paper if interested: <https://www.aeaweb.org/articles?id=10.1257/aer.20190661>.