

ECOM20001: Econometrics 1

Assignment 2

Student Information

To receive an assignment grade, you must fill out the information in this table and include this table on the front cover page for your assignment. **Only students whose names and student ID numbers are included on the cover page will receive marks for the assignment.** Groups of up to 3 students are allowed.

Name	Student ID Number
Sally Probability	422552
Xiaosong Statistics	653223
Ipsa Regression	294480

Due Date and Weight

- **Submit via LMS by 8 am on 23 September 2022**
- No late assignments will be accepted.
- This assignment is worth 5% of your final mark in ECOM20001.
- There are 40 marks in total.

What You Must Submit via the LMS

- **Assignment answers** no more than 8 A4 pages with 12-point font.
5 marks will be deducted if your answers exceed 8 A4 pages.
- The **R code** that generates your results. Specifically, copy and paste your R code in an Appendix at the end of your assignment document (e.g., in the .docx file) so it can be viewed and tested by markers. The R code Appendix does not count toward your 8-page answer limit. You may alter and shrink the R code font to less than a 12-point font so that it is easier to read.
2 marks will be deducted if you do not include your R code.

Additional Instructions

- You may submit this assignment in groups of up to 3. Students in a group are allowed to be from different tutorials. **You must submit your group before 16 September at 8 am; otherwise, you submit as an individual.**
- You must complete the assignment in no more than 8 A4 pages with 12-point Arial, Times New Roman, Helvetica, Cambria or Calibri font. The assignment cover page does not count toward the 8 A4 page limit.
- To save time, you may copy RStudio output directly into your answers in reporting empirical results. You are also free to create your better-formatted tables based on your RStudio output, which is, of course, good practice in learning how to present empirical results.
- Figures may also be copied and pasted directly into your assignment answers. They may be scaled down in size to meet the 8-page limit, but please ensure that your figures are readable. If they are not, marks will be deducted.
- Marks will be deducted if interpretations of results are incorrect, imprecise, unclear, or not well-scaled. Similarly, marks will be deducted if figures or tables are incorrect, unclear, not properly labelled, not well-scaled, or missing legends.
- When in doubt, work with 3 digits past the decimal throughout.
- This R code in the Appendix at the end of your assignment (as discussed on the previous page) must be commented on and easy for the subject tutors to follow. If the code is not well commented and easy to follow, marks will be deducted. Commenting and code clarity must be at the level of tutorial code, or marks will be deducted.
- Students with a genuine reason for not being able to submit the assignment on time can apply for special consideration to have the assignment mark transferred to the exam at the following link:
 - <https://students.unimelb.edu.au/admin/special/>

Getting Started

Please create an Assignment 2 folder on your computer, go to the LMS site for ECOM 20001, and download the following data file into the Assignment 2 folder:

- [as2_sleep.csv](#)

This dataset contains the following 14 variables:

- **id**: individual identifier
- **sleep**: number of minutes sleep (overnight) per week
- **naps**: number of minutes napping per week
- **totwrk**: number of minutes working per week
- **educ**: number of years of educational attainment
- **age**: age
- **gdhlth**: equals 1 if self-reported health is “Excellent” or “Good”, 0 otherwise
- **smsa**: equals 1 if lives in a US urban area (SMSA), 0 otherwise
- **union**: equals 1 if part of a union, 0 otherwise
- **selfe**: equals 1 if self-employed, 0 otherwise
- **marr**: equals 1 if married, 0 otherwise
- **yrs marr**: number of years married
- **yngkid**: equals 1 if they have a young child less than 3 at home, 0 otherwise
- **male**: equals 1 if male, 0 otherwise

Data summary

This dataset contains annual information on 700 individuals and their self-reported sleep and work times, individual characteristics like education or age, job characteristics like if they are self-employed or in a union, and household characteristics related to marriage and children. The data are drawn from the article: Biddle, Jeff R., and Daniel S. Hamermesh (1990): “Sleep and the Allocation of Time,” *Journal of Political Economy*, 98(5), 922-943.

About the Assignment

In this assignment, we will investigate the determinants of sleep and how sleep time relates to work time. We work with a historical dataset, namely the 1975-76 Time Use Study from the United States.

Questions

1. **(2 marks)** Compute summary statistics (mean, standard deviation, min, max) for all variables in the dataset and report them in a table using `stargazer()`. Describe in words a typical observation based on the sample means.
2. **(3 marks)** Construct the following three scatter plots, where the first variable (e.g., before the vs.) goes on the vertical axis and the second variable (e.g., after the vs.) goes on the horizontal axis. Ensure correct and clear axes and graph titles and for each scatter plot, ensure you overlay a single-linear regression line that helps visualise the relationship between the two variables.
 - `sleep` vs `educ`
 - `sleep` vs `age`
 - `educ` vs `age`
3. **(4 marks)** Suppose you ran a single linear regression with the dependent variable `sleep`, and the independent variable is `educ`. Here, `age` is an omitted variable. Based on the signs of the estimated regression line slopes in the three scatter plots from question 2, carefully explain what the direction of the bias would be for the slope coefficient on `educ` in a single linear regression of `sleep` on `educ`. Also, determine whether the coefficient would be too large or too small in magnitude based on the sign you would expect for the coefficient on `educ` in a regression of `sleep` on `educ`.
4. **(5 marks)** Run the following 5 regressions where in each case, the dependent variable is `sleep`, the regression includes an intercept, and each bullet point below lists the other independent variables to be included:
 - Reg (1): `educ`
 - Reg (2): `educ`, `age`
 - Reg (3): `educ`, `age`, `gdhlth`
 - Reg (4): `educ`, `age`, `gdhlth`, `smsa`, `union`, `selfe`
 - Reg (5): `educ`, `age`, `gdhlth`, `smsa`, `union`, `selfe`, `marr`, `yrs marr`, `yngkid`

Construct your table using `stargazer()` in R. Report heteroskedasticity robust standard errors for each coefficient estimate, the adjusted R-squared for fit, and the number of observations running the regression.

5. **(10 marks)** Based on the regression results table from question 4, answer the following questions:
- Compare the results in Reg (1) and Reg (2). Does the change in the coefficient on **age** correspond to the patterns you documented in questions 2 and 3 above?
 - Compare the results across Reg (2) to Reg (3) in the table. What happens to the coefficient estimate for **educ** in terms of its value and whether it is statistically significantly different from 0 at the 5% level? Intuitively, explain why there is another change in the **educ** coefficient in this regression.
 - Compare the results across Reg (3) to Reg (5) in the table. Is there much change in the coefficient estimate on **educ** regarding its value and statistical significance (from 0 at the 5% level) after controlling for work context and household variables?
 - Focusing on the Reg (5) estimates, how much does a 3-unit change in **sleep** change when **educ** increases by 3 years (e.g., the length of an undergraduate degree)? What is the 99% confidence interval for this predicted change?
 - Compute the overall regression F-statistic for Reg (5). Report the F-statistic, degrees of freedom and p-value for the test, ensuring that you account for heteroskedasticity. Interpret the statistical significance of the results (assume a 5% level). What conclusion you can draw from the test for Reg (5)?
6. **(4 marks)** Construct another regression table with 2 regressions:
- Reg (1): the dependent variable is **sleep**
 - Reg (2): the dependent variable is **nap**

Reg (1) and (2) have the same set of independent variables to be used is **educ**, **age**, **gdhlth**, **smsa**, **union**, **selfe**, **marr**, **yrs marr**, **yngkid**.¹ Construct your table using the **stargazer()** command in R. For each regression report, heteroskedasticity-robust standard errors for each coefficient estimate, the adjusted R-squared for model fit, and the number of observations used in running the regression.

Compare the magnitude and statistical significance (from 0 at the 10% level) of the coefficient estimate on **educ** in the two regressions. Briefly interpret what your results mean in words.

¹ Reg (1) in this table simply repeats Reg (5) from the regression table from question 5.

7. **(3 marks)** Run the following 3 regressions where in each case, the dependent variable is **sleep**, the regression includes an intercept, and each bullet point below lists the other independent variables to be included. The sample to estimate the regression is noted in purple.

- Reg (1): **totwrk, educ, age, gdhlth, smsa, union, selfe, marr, yrs marr, yngkid**
sample: **entire sample**
- Reg (2): **totwrk, educ, age, gdhlth, smsa, union, selfe, marr, yrs marr, yngkid**
sample: **males only (male==1)**
- Reg (3): **totwrk, educ, age, gdhlth, smsa, union, selfe, marr, yrs marr, yngkid**
sample: **females only (male==0)**

Construct your table using **stargazer()** in R. Report heteroskedasticity robust standard errors for each coefficient estimate, the adjusted R-squared for fit, and the number of observations running the regression.

8. **(7 marks)** Based on the regression results table from question 7, answer the following questions:
- From Reg (1), comment on the statistical significance of the coefficient estimate on **totwrk** (from 0 at the 5% level) and interpret the associated predicted change in **sleep** from working one more hour per day for five days within a given week.
 - From Reg (1), test whether the coefficient on **marr** is equal to the coefficient on **yngkid**. Report the F-statistic, degrees of freedom and p-value for the test, ensuring that you account for heteroskedasticity. Provide an interpretation of the statistical significance of the results (assume a 5% level) and state in plain language what conclusion you can draw from the test.
 - From Reg (2) and (3), comment on the statistical significance of the coefficient estimate on **totwrk** (from 0 at the 5% level) in both regressions, interpret in words what the associated predicted changes in **sleep** is from working one more hour per day for five days within a given week. Also, interpret what any difference in the predicted changes between Reg (2) and (3) means in plain language.
9. **(2 marks)** R-code: we will review and mark your R code as follows:
- 2/2 if the R code is correct and organised and commented like the solution code for the assignment.
 - 1/2 if the R code is correct but hard to follow or not well commented.

Assignment 2

- 0/2 if the R code is incorrect and/or a complete mess or not submitted.