ECOM20001: Econometrics 1

Assignment 3

Student Information

To receive an assignment grade, you must fill out the information in this table and include this page as your assignment cover page.

Name	Student ID Number	Tutor	Tutorial Day & Time	Tutorial Location
Sally Probability	422552	Richard Hayes	Tue 10:15am	The Spot 4452
Markus Statistics	653223	Sahiba Narang	Wed 4:15pm	The Spot 3054

Due Date and Weight

- **Submit via the LMS by 8am on Monday, 27 May 2019.**
- No late assignments will be accepted.
- This assignment is worth 5% of your final mark in ECOM20001.
- There are 50 marks in total.

What You Must Submit via the LMS

- **Assignment answers**, no more than 10 A4 pages with 12 point font. 5 points out of 50 will be deducted if you answers exceed 10 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 that it can be viewed and tested by markers. The R code Appendix does not count toward your 10 page answer limit. You may alter and shrink the R code font to less than 12 point font so that it is easier to read. 2 points out of 50 will be deducted if you do not include your R code.

Additional Instructions

- You may submit this assignment in groups of one or two. Students in a group of two are allowed to be in different tutorials.
- You must complete the assignment in no more than 10 A4 pages with 12 point Arial, Times New Roman, Helvetica, Cambria or Calibri font. The assignment cover page does not count as one of the 10 A4 pages.
- To save time, you may cut and paste RStudio output directly into your answers in reporting empirical results. You are also free to create your own better-formatted tables based on your RStudio output, which is of course good practice in learning how to present empirical results.
- Figures may also may be copied and pasted directly into your assignment answers. They may be scaled down in size to meet the 10 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 labeled, not well-scaled, or missing legends.
- This R code in the Appendix at the end of your assignment (as discussed on the previous page) must be clearly commented and easy for the subject tutors to follow. If the code is not well commented and easy to follow, 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 Assignment1 folder on your computer, and then go to the LMS site for ECOM 20001 and download the following data file into the Assignment1 folder:

- as3_smoke.csv

This micro dataset[1] has the following 13 variables:

- id: baby identifier
- birthweight: baby's birthweight in grams
- smoker: equals one if mother is a smoker, 0 otherwise
- alcohol: equals one if mother drank alcohol during pregnancy, 0 otherwise
- drinks: number of drinks per week during pregnancy
- nprevisit: total number of prenatal visits
- tripre1: equals one if 1st prenatal care in 1st trimester, 0 otherwise
- tripre2: equals one if 1st prenatal care in 2nd trimester, 0 otherwise
- tripre3: equals one if 1st prenatal care in 3rd trimester, 0 otherwise
- tripre0: equals one if no prenatal visits, 0 otherwise
- unmarried: equals one if mother is unmarried
- educ: years of educational attainment of mother
- age: age of mother
- gambles: equals one if mother is a problem gambler, 0 otherwise

In total, the dataset contains this information for n=3000 babies and their mothers.

About the Assignment

The data used in this assignment should be familiar as the dataset used in tutorials 7-9. In this assignment, investigate the use of logarithmic regressions and polynomial regressions to further study the relationship between smoking, prenatal visits and a baby's birthweight.

[1] Recall from Tutorial 7 that this dataset is from Almond, D and K. Chay (2005): "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, 120(3): 1031-1083.

Questions

1. **(8 marks)** Create a new variable called log_birthweight, which is computed as log_birthweight= log(birthweight). Run the following 4 regressions where log_birthweight is the dependent variable, and where each regression includes a constant and one of the following four sets of regressors:
   - Reg (1): smoker
   - Reg (2): smoker, alcohol, drinks, gambles
   - Reg (3): smoker, alcohol, drinks, gambles, nprevisit, tripre1, tripre2, tripre3
   - Reg (4): smoker, alcohol, drinks, gambles, nprevisit, tripre1, tripre2, tripre3, unmarried, educ, age

Report the results for Reg (1) - (4) in a table using stargazer() in R. In each regression, and for the remainder of the assignment, work under the assumption of heteroskedastic standard errors. In Reg (4), interpret the coefficient on smoker, and comment on whether it is statistically significantly different from 0 at the 5% level.

2. **(3 marks)** Using the ggplot() command in R, produce a scatter plot where log_birthweight is on the vertical axis, and nprevisit is on the horizontal axis. In your scatter plot, present a quadratic regression line that highlights any potential nonlinearities in the relationship between the two variables. Does the relationship appear to be nonlinear?

3. **(5 marks)** Construct a new variable named nprevisit_sq, which is the square of nprevisit_sq. That is, nprevisit_sq = nprevisit X nprevisit. Run another regression where log_birthweight is the dependent variable, and where the regression includes a constant and the following regressors:
   - Reg (5): smoker, alcohol, drinks, gambles, nprevisit, nprevisit_sq, tripre1, tripre2, tripre3, unmarried, educ, age

Do the coefficients on nprevisit and nprevisit_sq correspond to the relationship you saw in the scatter plot in question 2?[2]

[2] To save space in your solutions, you may report all the results for Reg (1) to Reg (5) from questions 1 and 3 in a single table using stargazer(), reporting heteroskedasticity-robust standard errors.

4. **(8 marks)** Report estimates and 95% confidence intervals of the following two partial effects from Reg (5), holding all other variables fixed:
   - Changing nprevisit from 2 to 4.
   - Changing nprevisit from 12 to 14

Briefly explain why you obtain differences in these partial effects, despite the change in nprevisit being 2 in each case.

5. **(3 marks)** Construct a variable called log_nprevisit, which is computed as log_nprevisit= log(1+nprevisit).[3] Construct a scatter plot using the ggplot() command in R with log_birthweight on the vertical axis and log_nprevisit on the horizontal axis. In your scatter plot, present a quadratic regression line that highlights any potential nonlinearities in the relationship between the two variables. Does the relationship appear to be nonlinear?

6. **(8 marks)** Generate another variable called log_nprevisit_sq, which is the square of log_nprevisit: log_nprevisit_sq = log_nprevisit X log_nprevisit. Run the following regressions where log_birthweight is the dependent variable, and where each regression includes a constant and one of the following 5 sets of regressors:
   - Reg (1): smoker
   - Reg (2): smoker, alcohol, drinks, gambles
   - Reg (3): smoker, alcohol, drinks, gambles, log_nprevisit, tripre1, tripre2, tripre3
   - Reg (4): smoker, alcohol, drinks, gambles, log_nprevisit, tripre1, tripre2, tripre3, unmarried, educ, age
   - Reg (5): smoker, alcohol, drinks, gambles, log_nprevisit, log_nprevisit_sq, tripre1, tripre2, tripre3, unmarried, educ, age

Report the results for Reg (1) - (5) in a table using stargazer() in R.[4] Interpret the coefficient estimate on log_nprevisit in Reg (4), and comment on whether it is statistically significant at the 5% level. Also briefly explain why you think there is such a large change in coefficient and standard error on log_nprevisit between Reg (4) and Reg (5) in light of your findings from question 5.

---

[3] We work with log(1+nprevisit) and not simply just log(nprevisit) because log(nprevisit) is undefined if nprevisit=0. We avoid this with log(1+nprevisit), and it is fine because log(1+x) is approximately equal to log(x) for x sufficiently large.

[4] This should be a separate table from the one produced in questions 1-3 above.

7. **(5 marks)** Construct one final variable, log_nprevisit_age, which is the following interaction variable between the log of prenatal visits and a mother's age: log_nprevisit_age = log_nprevisit X age. Run another regression where log_birthweight is the dependent variable, and where the regression includes a constant and the following regressors:

   - Reg (6): smoker, alcohol, drinks, gambles, log_nprevisit, log_nprevisit_age, tripre1, tripre2, tripre3, unmarried, educ, age

Based on your regression results, is the elasticity of birthweight with respect to nprevisit larger or smaller in magnitude for older mothers?[5]

8. **(8 marks)** Based on your estimates for Reg (6) compute the elasticity of birthweight with respect to nprevisit for a mother with age=20 and age=40. Also report the 95% CI for each elasticity.

9. **(2 marks)** R-code: we will review and mark your R code according to the following scheme:

   - 2/2 if R code is correct and organised and commented like the solution code for the assignment.
   - 1/2 if R code is correct, but hard to follow or not well commented.
   - 0/2 if R code is incorrect and/or a complete mess, or not submitted.

[5] If you find it useful, you may report all the results for Reg (1) to Reg (6) from questions 6 and 7 in a single table using stargazer() in R, reporting heteroskedasticity-robust standard errors. So, in total, your assignment should have two regression tables: one for questions 1-3, and a separate one for questions 6-7.