Homework 1 Questions

Note: Please respect intellectual property and do not make this document available in the public domain

1. Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$. Suppose you know that β_0 = 0. What is the formula for the ordinary least squares estimator of β_1?

2. A university administrator declares "A broad-based curriculum is very critical to help you adapt to the economy of the future. Many employers are attracted to graduates with a broad-based education because they will have knowledge and skills in multiple fields that can be applied to the workplace". To investigate this claim, you decide to do an experiment. You visit the job advertisements in the Straits Times Classified section and responds to job advertisements by employers there. You create 2 fictitious resumes which differ only in the applicant's course proportion of courses read outside the major but which are otherwise identical: for one resume, the applicant is stated to read 80% of all courses within her major while for the other resume, the applicant is stated to read only 30% of all courses within her major. Taking more courses within a person's major implies that a person has a more specialized curriculum while taking fewer courses within a person's major would imply that the person has a more broad-based curriculum. You include a different photo on each of these 2 fictitious resumes. You then send the resumes randomly to employers which have listed job advertisements in the Straits Times and record the number of times each resume receives an interview offer. The first resume is sent to 1032 companies and the second resume is sent to 1535 companies. You find that the resume that is stated to read only 30% of all courses within the major receives a statistically significantly higher call-back rate for an interview than the resume that is stated to read 80% of all courses within the major.

   Which of the following statements is true?

3. A principal of a school wants to know whether time spent in school (as measured by hours spent per day in school) leads to improved student performance (as measured by subject test scores). She wants to study this question because she wants to know if she should implement a longer school day. She is thinking of either having students spend 4 hours of their time each day in school, 6 hours, 8 hours, or 10 hours. She will pick the option which gives her the highest average student performance. Suppose she designs an experiment to answer this question. She randomly assigns students from her school at the start of the academic year to one of four groups. Group 1 spends 4 hours per day in school, Group 2 spends 6 hours per day in school, and so on. Each group contains more than 500 students. At the end of the academic year, she then administers a Science test to students in all 4 groups and computes the average test score for each of these groups, before picking the duration which produces the highest test score.

   Which of the following statements is incorrect?

4. A school principal wants to know whether time spent in school (as measured by hours spent per day in school) leads to improved student performance (as measured by test scores). She wants to study this question because she wants to know if she should implement a longer school day. She is thinking of either having students spend 4 hours of their time each day in school, 6 hours, 8 hours, or 10 hours. She will pick the option which gives her the highest average student performance. Suppose she designs an experiment to answer this question. She randomly assigns students from her school at the start of the academic year to one of four groups. Group 1 spends 4 hours per day in school, Group 2 spends 6 hours per day in school, and so on. Each group contains more than 400 students. At the end of the academic year, she then administers a Science test to students in all 4 groups and computes the average test score for each of these groups, before picking the duration which produces the highest test score.

   Suppose that the average test scores of students in group 1 are numerically different from the average test scores of students in group 4. Suppose, however, that the difference in the average test scores are not statistically significant at the 60% level, which of the following is correct?

5. Consider the regression model $Q_i = \beta_0 + \beta_1 P_i + u_i$, where Q is the dependent variable and P is the regressor. Is the R^2 from the regression of *Q* on *P* *exactly* the same as the R-squared value from the regression of *P* on *Q*?

6. Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$. Suppose the 3 least squares assumptions are satisfied, except that the first assumption is replaced with $E(u_i|X_i = x) = 27$. Which of the following is true?

7. Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$. Suppose the 3 least squares assumptions are satisfied, except that the first assumption is replaced with $E(u_i|X_i = x) = 50$. Which of the following is true?

8. You would like to know how various factors affect the academic achievement of students in university. To this end, you will examine data collected from a university in the United States. The stata data file Harfort.dta, which you will find in Canvas (under Files -> Homework -> Homework 1) contains data on the academic achievement of students in university and a variety of student characteristics for a sample of 8,000 students.

   The variables in the dataset are defined as follows:

   gpa is the student's grade point average (or GPA) at the point of graduation (i.e. in the 4th year), computed as $GPA = \frac{sum(course\ grade\ point \times\ credits\ assigned\ to\ course)}{sum(\ credits\ assigned\ to\ all\ courses\ used\ in\ calculating\ the\ numerator)}$. It is a continuous variable which measures academic achievement in university and potentially ranges from 0 to 5. *male* is a binary variable equal to 1 if the student is male and equal to 0 if the student is female. *honours* is a binary variable equal to 1 if the student graduated with an honours degree and equal to zero if the student did not graduate with an honours degree. *minor* is a binary variable equal to 1 if the student graduated with a minor and equal to zero if the student did not graduate with a minor. *white* is a binary variable equal to 1 if the student is white and is equal to 0 if the student is not white, *incoming_score* is a continuous variable measuring the student's achievement on a common high school leaving examination and represents the incoming academic score of the student. Finally, *household_income* is a continuous variable measuring the student's monthly household income in US dollars.

Use this dataset to answer questions 8 to 20.

Run a regression of grade point average (gpa) on the incoming academic score of students. Remember to include heteroskedasticity-robust standard errors. The output tells us that:

```
. regress gpa incoming_score, robust

Linear regression                               Number of obs     =      8,000
                                                F(1, 7998)        =       7.34
                                                Prob > F          =     0.0067
                                                R-squared         =     0.0010
                                                Root MSE          =     .59022
```

gpa	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
incoming_score	-.000604	.0002229	-2.71	0.007	-.0010409	-.0001671
_cons	3.78287	.0367134	103.04	0.000	3.710902	3.854837

9. Examine the data on grade point average (gpa) as well as the incoming academic score of students. Again, run a regression of grade point average (gpa) on the incoming academic score of students. Remember to include heteroskedasticity-robust standard errors. Nicole is a student with an incoming academic score of 88. Corona is a student with an incoming academic score of 24. How much more or less is Corona's grade point average predicted to be compared with Nicole's, assuming they are alike in all other aspects apart from incoming academic score? Round your answers to 3 significant places.

10. Run a regression of grade point average (gpa) on the incoming academic score of students. Remember to include heteroskedasticity-robust standard errors. Using the normal table, construct a 43.8% confidence interval for the coefficient on *incoming_score*. Round your answers to 3 significant figures. Pick the closest option.

11. Run a regression of grade point average (gpa) on the incoming academic score of students. Remember to include heteroskedasticity-robust standard errors. Suppose grade point average now increases by a factor of 30 (i.e. a gpa of 1 now becomes 30). At the same time, incoming academic score increases by a factor of 5 (i.e. an incoming academic score of 1 now becomes 5). What would the estimated coefficient on the modified *incoming_score* variable be if this were the case? Round your answer to 4 significant figures. Pick the closest option.

12. Run a regression of grade point average (gpa) on the monthly household income of students. Remember to include heteroskedasticity-robust standard errors. Walter is a student with a monthly household income of $15,700. Teck Hua is a student with a monthly household income of $460. How much more or less is Teck Hua's grade point average predicted to be compared to Walter's, assuming they are alike in all other aspects apart their household incomes? Round your answers to 2 decimal places.

13. Run a regression of grade point average (gpa) on the monthly household income of students. Remember to include heteroskedasticity-robust standard errors. Jianhao has attained a grade point average of 3.37 and he has a monthly household income of $2500. How large is the residual specific for Jianhao? Round your answer to 3 decimal places. Pick the closest option.

14. Refer to the regression in question 13. Suppose that we modify the variable *household_income* so that it is now a continuous variable measuring the student's monthly household_income in hundreds of US dollars (i.e. a household income of $1000 is now recorded as 10). What will be the 96.6% confidence interval for the coefficient on this modified *household_income* variable? Round your answers to 4 decimal places. Pick the closest option.

15. Refer again to question 13. Suppose that we modify the variable *household_income* so that it is now a continuous variable measuring the student's monthly household_income in hundreds of US dollars (i.e. a household income of $1000 is now recorded as 10). What would be the coefficient and the standard error on the modified *household_income* variable now? Round your answers to 3 significant figures. Pick the closest option.

16. You have ran a regression of grade point average (gpa) on the monthly household income of students. Suppose that the Root Mean Squared Error of the Regression is 0.56675. Suppose the values of the grade point average in the dataset are now divided by a factor of 400 (i.e. a grade point average of 1 now becomes 0.0025). At the same time, suppose that the variable *household_income* is now a continuous variable measuring the student's monthly household_income in thousands of US dollars (i.e. a household income of $1000 is now recorded as 1). What will the new Root Mean Squared Error of the regression be? Round your answer to 3 significant figures. Choose the nearest answer.

17. You have ran a regression of grade point average (gpa) on the monthly household income of students. Suppose that the R-squared value of the regression is 0.0789. Further, suppose the values of the grade point average in the dataset are now divided by a factor of 400 (i.e. a grade point average of 1 now becomes 0.0025). At the same time, suppose that the variable *household_income* is now a continuous variable measuring the student's monthly household_income in thousands of US dollars (i.e. a household income of $1000 is now recorded as 1). What will the new R-squared value of the regression be? Round your answer to 3 significant figures. Choose the nearest answer.

18. Suppose you are interested to know whether monthly household income differs between male and female students. Run a regression which will allow you to examine this. Remember to include heteroskedasticity robust standard errors. What do your results tell you?

19. Suppose you are interested to know whether monthly household income differs between students who graduated with a minor and students who graduated without a minor. Run a regression which will allow you to examine this. Remember to include heteroskedasticity robust standard errors. What do your results tell you?

20. Suppose you are interested to know whether monthly household income is greater for students with higher grade point average scores. Run a regression which will allow you to examine this. Remember to include heteroskedasticity robust standard errors. What do your results tell you?