QMSS – Data Analysis (5015) Midterm, Fall 2018 Name : Yue Ma

Part 1

1. (a) In Model 1, interpret the coefficient on Number of Children (8 points). (b) Is it statistically significant and how do you know? Please talk about t-statistics in your answer. (9 points) (c) In Model 1, does the correlation (or R-squared) between Ideal Number of Children and Number of Children seem low or high to you? Based on your prior expectation, give one reason for why you think that relationship is either low or high (4 points).

(a) Generally, in regression, the coefficient tells us how much the dependent variable is expected to increase or decrease when that independent variable increases or decreases by one. In Model 1, the ideal number of children for a family to have would, on average, be 0.12 more with 1 actual children more the respondent has. That is to say, respondent with 5 actual children would, on average, like to have 0.12 more children in their family to their expectation than those with 4 actual children.

(b) It is statistically significant. The t-statistic is the ratio of the departure of the estimated value of a parameter from its hypothesized value to its standard error. P-value is the probability of observing a test statistic at least as large as the one calculated assuming the null hypothesis is true. In Model 1, the Std. Error is 0.02, so the t-statistics of the coefficient is 6 (=0.12/0.02) or as negative as -6, and the p-value obtained in the t-test results is less than 0.001 (the alpha value, and with ***), t =6 is only likely to happen by chance less than 0.1% of the time assuming the null is correct, so it means that we have enough evidence against the null, and conclude that there's a statistically significant difference. The null hypothesis is that there is no corelated relationship between the ideal number of children and the number of children people have.

(c) The correlation (or R-squared) between *Ideal Number of Children and Number of Children* seems low based on my prior expectation. Generally, R-squared is a statistical measure of how close the data are to the fitted regression line. 0% indicates that the model explains none of the variability of the response data around its mean, and 100% indicates that the model explains all the variability of the response data around its mean. In Model1, the R-squared is 0.04, which means that the model explains only 4% of the variability of the response data around its mean. From my perspective, many parents who have children now get a lot of happiness spending time with their family, so they are more likely to have more children in the near future.

2. In Model 1, interpret the constant/intercept in this model-- i.e., to whom does it refer? Also note statistical significance. (10 points)

The constant/intercept is the expected mean value of Y when all X=0. If X sometimes equals 0, the intercept is simply the expected mean value of Y at that value. In Model 1, when the respondent has no actual children now, then the number of the ideal children to his or her family to have is 2.39. The Std. Error of the intercept is 0.05, so the t-statistics of the coefficient is 47.8 (=2.39/0.05), or as negative as -47.8, and the p-value of the intercept is less than 0.001 (the alpha value, and with ***), t=47.8 is only likely to happen by chance less than 0.1% of the time assuming the null is correct, so it means that we have enough evidence against the null and to say that the intercept in Model 1 is significantly different from zero, and it is meaningful.

3. (a) In Model 2, interpret the coefficient on Age . Note statistical significance. (8 points). (b) In Model 2, Age has a negative relationship with Ideal Number of Children . Does this relationship make sense to you? Give me one explanation for why you think Age would have a negative relationship with Ideal Number of Children (4 points).

(a) The researcher further includes two variables - the number of siblings the respondent has and the respondent's age. The coefficient on *Age* is -0.01. And the Std. Error on *Age* is 0.00, p value on *Age* is less than 0.001 (the alpha value, and with ***). So we can reject the null hypothesis and conclude that there's a statistically significant difference. Assuming the number of siblings the respondent has and the number of children the respondent has are constant, the coefficient represents difference in the predicted value of the ideal children number for each one-unit difference in the respondent's age. -0.01 means: with the number of siblings and the number of children constant, the ideal number of children for a family to have would, on average, be 0.01 less with 1 age more the respondent is. That is to say, respondent with and age on 33, on average, like to have 0.01 less children in their family to their expectation than those with an age on 32.

(b) This negative relationship makes sense to me. Without considering other factors, with increasing age, people are more likely to develop themselves from various aspects, including interests, career path, mental maturity, financial status, social status, etc. With the quality of life increasing, people will have more choices to enjoy life, and raising children is just one of the large amount of ways to enjoy better life. As a result, the number of ideal children to the family of these group of people will decrease slightly. Instead, they will pay more attention to other aspects in their life. For instance, they may develop new hobbies because they do not need to immerse themselves a lot in large workload of jobs and have more free time; they may prefer to spend more time with their children or with grand-children, etc. As a result, it indeed makes sense that Age has a negative relationship with Ideal Number of Children.

Part 2

6. Interpret the coefficient on Income (1000s) in Model 1 . Note statistical significance. (6 points). Interpret the coefficient on Ln(Income) in Model 2 . Note statistical significance. (9 points). Why do economists often prefer the specification in Model 2 over the specification in Model 1? Give two reasons. (8 points).

In Model 1, the coefficient on *Income (1000s)* is 0.09. And the Std. Error on *Income (1000s)* is 0.02, p value on *Income (1000s)* is less than 0.001 (the alpha value, and with ***). So we can reject the null hypothesis and conclude that there's a statistically significant difference. The coefficient represents difference in the predicted size of the incentive (Fee in $) for each one-unit difference in the respondent's total family income. 0.09 means: the size of the incentive would, on average, be 0.09 dollars more with 1000s of dollars income more the respondent's family has. That is to say, respondent with total family income on 11000 dollars, on average, like to have 0.09 dollars more monetary incentive than those with total family income on 10000 dollars.

In Model 2, the coefficient on *Ln(Income)* is 0.06. And the Std. Error on *Ln(Income)* is 0.01, p value on *Ln(Income)* is less than 0.001 (the alpha value, and with ***). So we can reject the null hypothesis and conclude that there's a statistically significant difference. The coefficient represents difference in the predicted size of the natural log of the incentive (*Ln(Fee)*) for each one-unit difference in the natural log

of the total family income (*Ln(Income)*). 0.06 means: the log of amount of monetary incentive would, on average, be 0.06 more with 1 unit more on the log of the total family income.

The two reasons are as below:
(1) **Coefficients in log-log regressions ≈ proportional percentage changes**
Based on the characteristics of logarithm, the equation is correct: LN(X (1+r)) = LN(X) + LN(1+r) ≈ LN(X) + r, as a result, the change in natural log is likely to equal to the percentage change. In real economic situations, the marginal effect of one variable on the expected value of another is linear in terms of percentage changes rather than absolute changes. So, applying a natural log transformation to both dependent and independent variables may be appropriate. Moreover, there is also a great advantage in such cases that small changes in the natural log of a variable are directly interpretable as percentage changes, to a very close approximation. So it is more interpretable to use logarithm just like the one in Model 2 rather than Model 1.

(2) **Make the data more smooth and easy to analyze.**
Due to the characteristic of the function ln(), It is really helpful to reduce the sharpness of the data by applying a natural log transformation to both dependent and independent variables. Then, narrow the range of the variables. In some cases, the variable may be highly skewed to the right, which may produce problems in a regression analysis. By taking the natural log, the variables tend to behave more in line with the normality assumption.

7. Interpret the coefficient on "Miscellaneous Work Status" in Model 3 . Note statistical significance. (8 points)

In Model 3, the coefficient on *Miscellaneous Work Status* is-0.26. And the Std. Error is 0.07, p value is less than 0.001 (the alpha value, and with ***). So we can reject the null hypothesis and conclude that there's a statistically significant difference. The coefficient represents difference in the predicted size of the natural log of the incentive (*Ln(Fee)*) for each one-unit difference in the potential work statuses of the respondent. -0.26 means: with other factors fixed, people with miscellaneous work status would get 0.26 unit less of log of the fee than people with full time.

8. (a) Explain what a partial F-test tests for (9 points), and (b) explain whether, from a statistical standpoint, this F-test indicates that the researcher made a good decision to include the Work Status dummy variables. (8 points)

(a) A partial F-test is the appropriate test to use when the simultaneous test of the statistical significance of a group of variables is needed. It assesses whether the improvement in model fit (as assessed by a reduction in prediction error) using the full model is too large to be ascribed to chance alone.

(b) The F-test result is 3.48 ($p<0.001$). It means that the variables are jointly significant, and the researcher made a good decision to include the *Work Status* dummy variables.

9. Interpret the coefficient on How Attractive Respondent Is (Interviewer Assessment) in Model 4 . Note statistical significance. (5 points). What conclusion do you draw from the sign and statistical significance of this variable and its relationship to the size of incentive given to the respondent? Provide one possible

conclusion. (4 points)

(a) In Model 4, the coefficient on *How Attractive Respondent Is* is 0.03. And the Std. Error is 0.01, p value is less than 0.05 (the alpha value, and with *). So we can reject the null hypothesis and conclude that there's a statistically significant difference. The coefficient represents difference in the predicted size of the natural log of the incentive (*Ln(Fee)*) for each one-unit difference in the attractive level of the respondent. 0.03 means: with other factors fixed, people with 1 higher attractive level would get 0.03 more unit of natural log of the fee than those with lower attractive level.

(b) Respondents would be likely to get more fee it they are more attractive. Maybe more attractive look will help the respondents leave the interviewer a more good first impression, as a result, they tend to receive more fee.