PROJECT

ECON 382
DUE ON DECEMBER 19

Instructions

The project consists of two parts. Part 1 covers Chapters 4 and 5. Part 2 covers Chapters 6, 7 and 11. Make sure to present your estimation results using the `R` package `stargazer`. Submit your answers in a `word` (or `html`) document via email. If you are familiar with `R Markdown`, you can create an `R Markdown` file for your answers. Knit your `Markdown` file as a `word` (or `html`) document. Do not(!) print the knitted document.

Data

The data set `401ksubs.csv` contains information on net financial wealth (`nettfa`), age of survey respondent (`age`), annual family income (`inv`), family size (`fsize`), a binary variable for eligibility in a 401(k) plan (`e401k`), participation in certain pension plans for people in the United States and other variables. The wealth and income variables are both recorded in thousands of dollars.

Part 1

**Use only the data for single-person households, i.e., `fsize == 1` to answer the questions in Part 1.

1. How many single-person households are there in the data set?
2. Use OLS to estimate the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u \tag{1}$$

   and interpret the slope coefficients. Is there any sign contrary to your expectation?
3. Does the intercept from the regression have an interesting meaning? Explain.
4. Find the p-value for the test $H_0 : \beta_2 = 1$ against $H_1 : \beta_2 < 1$. Do you reject H_0 at the 1% significance level?
5. What is the youngest age of people in this sample? How many people are at that age?

6. Use OLS to estimate the the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u. \tag{2}$$

   Are you concerned that the coefficient on `age` is negative? Explain.

7. Because the youngest people in the sample are 25, it makes sense to think that, for a given level of income, the lowest average amount of net total financial assets is at age 25. Recall that the partial effect of `age` on `nettfa` is $\beta_2 + 2\beta_3 age$, so the partial effect at age 25 is $\beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3$; call this θ_2. Find $\hat{\theta}_2$ and obtain the two-sided p-value for testing $H_0 : \theta_2 = 0$. You should conclude that $\hat{\theta}_2$ is small and very statistically insignificant. [*Hint*: One way to do this is to estimate the model $nettfa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$, where the intercept, α_0, is different from β_0. There are other ways, too.]

8. Because the evidence against $H_0 : \theta_2 = 0$ is very weak, set it to zero an estimate the model

$$nettfa = \alpha_0 + \beta_1 inc + \beta_3(age - 25)^2 + u. \tag{3}$$

   In terms of goodness-of-fit, does this model fit better than that the model in (2)?

9. For the estimated equation in (3), set $inc = 30$ and graph the relationship between `nettfa` and `age`, but only for $age \geq 25$. Describe your finding.

Part 2

**Use the entire dataset to answer the questions in Part 2.

1. What fraction of the families in the sample are eligible for participation in a 401(k) plan?

2. Estimate a linear probability model explaining 401(k) eligibility in terms of income, age and gender. Include income and age in quadratic form. Would you say that 401(k) eligibility is independent of income and age? What about gender? Explain.

3. Obtain the fitted values from the linear probability model. Are any fitted values negative or greater than one?

4. Estimate a model for `nettfa` that includes income, age and `e401k` as explanatory variables. The income and age variables should appear in quadratics. Explain the estimated dollar effect of 401(k) eligibility?

5. To the model estimated in 4., add the interactions `e401k·(age-41)` and `e401k·(age-41)`^{2}. Note that the average age in the sample is 41, so that in the new model, the coefficient on `e401k` is the estimated effect of 401(k) eligibility at the average age. Which interaction term is significant?

6. Based on your findings from 4. and 5., do the estimated effects of 401(k) eligibility at age 41 differ much? Explain.

7. Drop the interaction terms. Define five family size dummy variables: `fsize1`, `fsize2`, `fsize3`, `fsize4` and `fsize5`. The variable `fsize5` is unity for families with five or more members. Include these dummies in the model from 4., be sure to choose a base group, i.e., dummy trap. Are the family dummies significant at the 1% level?

8. Using OLS, estimate a lienar probability model for `e401k`, using as explanatory variables `inc`, `inc`^{2}, `age`, `age`^{2} and `male`. Obtain both the usual OLS standard errors and heteroskedasticity-robust versions. Are there any important differences?