ECON613: Applied Econometrics in Microeconomics

Problem Set # 1

Yueru Li

Due Date: 11 PM, Fri Feb 4

Problem 1. OLS Estimates

1. Calculate the correlation between Y and X

   I calculated the corrlation with the following formula

   $$corr(Y,X) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2 \times \sum_{i=1}^{n}(y-\bar{y})^2}}$$

   I dropped the entries with 0 wage. This applied on the sample yields a correlation of 0.143492.

2. Calculate the coefficient on this regression. Remember $\hat{\beta} = (X^TX)^{-1}X^TY$

   I included a constant term and a liner term x. The result is

   $$\begin{aligned} \beta_0 \equiv \quad & \beta_{intercept} && = 14141.1794 \\ \beta_1 \equiv \quad & \beta_{age} && = 230.9923 \end{aligned}$$

3. Calculate the standard errors of β

   (a) Using the standard formulas of the OLS

   I first check if the data satisfy Homoscedasticity. From the plotted fitted vs residual plot as shown in figure 1, it is clear that Homoscedasticity is violated. So I calculate the robust standard errors.

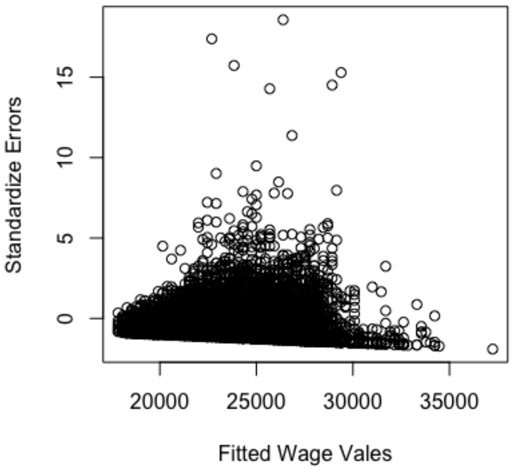

Figure 1: Checking Homoscedasticity

$$\sqrt{var(\beta_0)} = 526.69029$$
$$\sqrt{var(\beta_1)} = 14.20466$$

(b) Using bootstrap with 49 and 499 replications respectively.

i. Using 49 bootstrap

$$\sqrt{var(\beta_0)} = 542.45271$$
$$\sqrt{var(\beta_1)} = 14.73729$$

ii. Using 499 bootstrap

$$\sqrt{var(\beta_0)} = 571.53772$$
$$\sqrt{var(\beta_1)} = 15.15403$$

(c) Comment on the differences between the two strategies

The OLS formula requires that we have a closed form solution to the estimator. While we can see it it gives us a tighter standard error, some times when it is difficult or impossible to have a closed form solution to the variance, bootstrap's non-parametric way of estimating variance becomes convenient and provides reasonably good approximations.

Problem 2. Detrend Data

1. Create a categorical variable ag, which bins the age variable into the following groups: "18-25", "26- 30", "31-35", "36-40", "41-45", "46-50", "51-55", "56-60", and "60+".

   Please refer to the code. I dropped those who are below 18.

2. Plot the wage of each age group across years. Is there a trend? For the quantile analysis of the wage we can see that 80000 dollars is already at 99^{th} percentile. I limit my plot to 99^{th} to prevent the distribution to be squeezed by the outliers. Please refer to figure 2. The plot for those who are above 60 is clustered together because the up until around 90 percentile the wage is still zero.

   It appears that the 25% to 75% workers are gaining income, however it is unclear if this is because of inflation or real wage growth.

3. Consider $Y_{it} = \beta X_{it} + \gamma_t + e_{it}$. After including a time fixed effect, how do the estimated coefficients change?

   We focus on the estimated terms for intercept, age and age squared.

$$\beta_0 = 23701.21149$$
$$\beta_1 = -239.43116$$

   We can see that the intercept changed a lot, increase by almost 10000. What is more striking is the change of the coefficient on *age*. The coefficient went from 230 to -239. If we believe that the year fixed effect should belong to the model, then the previous estimate yields great bias. There are other reasons for bias including very likely a non-linear relationship between age and wage, but that is a separate issue not discussed in this problem.

Problem 3. Numerical Optimization

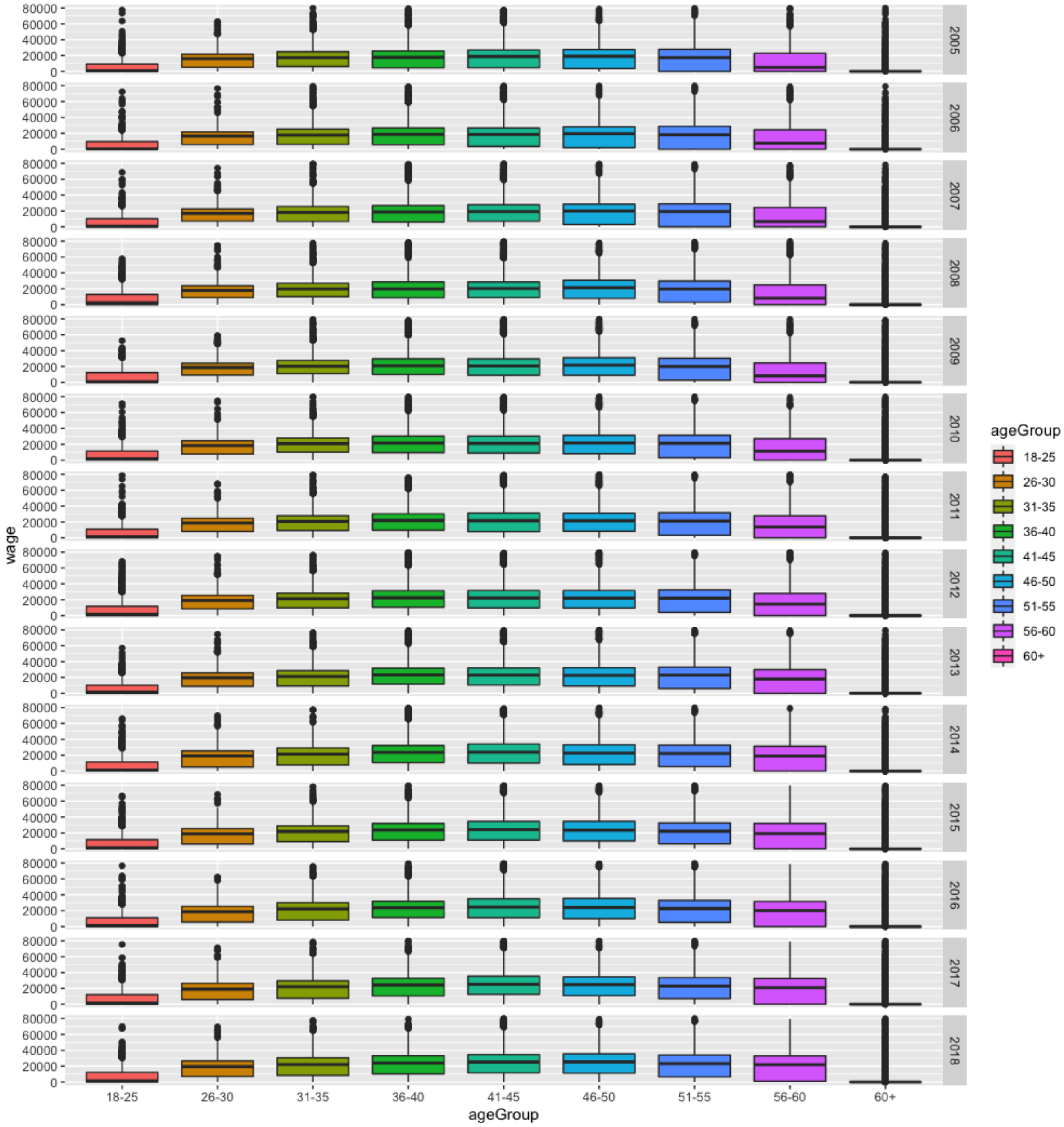

Figure 2: Income By Age By Year

1. Exclude all individuals who are inactive.

   I dropped those who are below 18, or with employment status "Inactive" or "Retired"

2. Write a function that returns the likelihood of the probit of being employed.

   Please refer to the code.

3. Optimize the model and interpret the coefficients. You can use pre-programmed optimization pack- ages.

   Including a quadratic term of age caused some problem to the optimization, yielding negative entries on the diagonal of the Hessian matrix. So in this problem and in next problems estimating age's influence on labor participation, I only include a linear term of age.

   Please refer to the code for the implementation. Here is the list of the result of this optimization.

$$\beta_0 = 1.058080250$$
$$\beta_1 = 0.006618533$$

   $\beta_0 > 0$ does not immediately have a meaning for interpretation because it is related to the probability of employment if we extrapolate our data to individuals of age 0. But it's positiveness and the positiveness of β_1 jointly suggests that at the age of 18, there is a positive probability that one is employed. The fact that $\beta_1 > 0$ suggests that as one's age increases, the probability that he get employed also increases. We do not have an immediate quantitative interpretation of the coefficients.

4. Can you estimate the same model including wages as a determinant of labor market participation? Explain.

   No. We only observe wage if a person chooses to participate in employed. This is censored data. In other words, we have no data on the wage of those who choose not to be employed had they choose to be employed. So we can't predict employment based on wage.

Problem 4. Discrete Choice

1. Exclude all individuals who are inactive.

   Please refer to the code. I dropped those who are below 18, or with employment status ”Inactive” or ”Retired”

2. Write and optimize the probit, logit, and the linear probability models.

   Please refer to the code for implementation. For the same reason I mentioned in exercise 3, I only included a linear term of age.

3. Interpret and compare the estimated coefficients. How significant are they?

   Please refer to the following table for the coefficients and corresponding standard errors for each model. I only keep 4 significant digits and the values in the parenthesis are the standard errors for the corresponding estimated coefficient.

Variable	Probit	Logit	Linear
β_0	0.7613(0.0230)	1.140(0.0446)	0.8003(1.785e-05)
β_1	0.0121(0.0004)	0.0250(0.0008)	0.0023(5.597e-09)
β_{2006}	0.0135(0.0229)	0.0264(0.0444)	0.0024(1.680e-05)
β_{2007}	0.0772(0.0231)	0.1520(0.0451)	0.0134(1.649e-05)
β_{2008}	0.1071(0.0233)	0.2079(0.0457)	0.0179(1.656e-05)
β_{2009}	0.0243(0.0228)	0.0429(0.0442)	0.0038(1.656e-05)
β_{2010}	0.0191(0.0226)	0.0338(0.0438)	0.0029(1.629e-05)
β_{2011}	0.0509(0.0227)	0.0935(0.0441)	0.0081(1.611e-05)
β_{2012}	0.0083(0.0222)	0.0085(0.0429)	0.0005(1.569e-05)
β_{2013}	-0.0411(0.0224)	-0.0866(0.0431)	-0.0084(1.638e-05)
β_{2014}	-0.0358(0.0224)	-0.0757(0.0431)	-0.0074(1.626e-05)
β_{2015}	-0.0575(0.0224)	-0.1190(0.0430)	-0.0116(1.637e-05)

   Let’s first interpret the results in base line year, year 2005.

   First we interpret the coefficient on age, β_1. Notice that it is positive across three models. For Probit and Logit model, we can only conclude that increasing age increases the probability of being employed. However, for the linear probability model, we can quantitatively interpret the coefficient as a 1 year increase in age increases the proability of being employed by 0.23 percentage points.

For interpreting the coefficient on the intercept β_0, notice that they are all positive. We can't interpret this as the employment at age 0, because that is extrapolating beyond the support of the data we used to fit the models. But we can use this plus the positive coefficient on age to conclude that at the age of 18, there is a positive probability that one is employed. For the linear probability model, we can conclude that at the age of 18, there is a $0.8003 + 18 * 0.0023 = 84.17\%$ probability that one is employed.

Now that we incorporate the year fixed effects. Notice that this does not affect our interpretation of β_1, but it does affect our β_0 interpretation indirectly. Because $\beta_0 + \beta_{year}$ will form our new intercept for our year of interest. However, notice that none of the year fixed effects is big enough to change the qualitative interpretations we made for Probit or Logit models. And it is straight forward to incorporate the new intercept into our interpretation for the linear model.

Notice that while some of the year fixed effects might be not significant. The main parameter of interest, β_0 and β_1 are always statistically significant with a large number of standard deviations away from zero. We can see that

(a) For β_0

Probit: $0.7631/0.0230 = 33.1$, Logit: $1.140/0.0446 = 25.56$ and Linear probability model: $0.8003/1.785e - 5 = 44834.7$. Given the asymptotic normality of MLE estimators and OLS estimators, these are all statistically significant coefficients.

(b) For β_1 Probit: $0.0121/0.0004 = 30.25$, Logit: $0.0250/0.0008 = 31.25$ and Linear probability model: $0.0023/5.597e - 5 = 41.09344$. Given the asymptotic normality of MLE estimators and OLS estimators, these are all statistically significant coefficients.

The interpretations of the coefficients are robust across the three models.

Problem 5. Marginal Effects

1. Compute the marginal effect of the previous probit and logit models.

Since I am asked to perform the following calculation for models estimated in Exercise 4. I continue perform the same restrictions to my data. I dropped those who are below 18, or with employment status ”Inactive” or ”Retired”

For this problem, I calculate the marginal effect at the mean. Because of the existence of year fixed effects, I choose to make the evaluation at the baseline year, year 2005.

For detailed implementation, please refer to the code. Here is a brief exposition of what I did. Because I only included a linear term of age, the calculation is ratehr simple.

(a) ME for Probit

$$ME_{Probit}(a\bar{g}e) = \phi(\beta_0 + a\bar{g}e\beta_1) * \beta_1$$

(b) ME for Logit

$$ME_{Logit}(a\bar{g}e) = \beta_1 * \frac{e^{-\beta_0 - \beta_1 * a\bar{g}e}}{(1 + e^{-\beta_0 - \beta_1 * a\bar{g}e})^2}$$

Since I evaluated at base year, all the year fixed effects go away.

The result I get is the following. First of all, the average age in the sample $a\bar{g}e \approx 40.81$

$$ME_{Probit}(40\bar{.}81) \approx 0.0022$$
$$ME_{Logit}(40.\bar{8}13) \approx 0.0023$$

2. Construct the standard errors of the marginal effects. Hint: Boostrap may be the easiest way.

   With two bootstraps of 49 iterations, I estimated that the standard error for $ME_{Probit}(40\bar{.}81)$ to be 0.0001 and the standard error for $ME_{Logit}(40\bar{.}81)$ to be $9.588e - 05$. I kept only 4 significant digits.