Econ 613 - Applied Econometrics - 2022 Spring
Homework 2

Yican Liu

February 4, 2022

1

1.1

I use R code, *cor* to estimate the correlation. The correlation between age and wage is -0.1789.

1.2

The coefficient is

$$\text{Wage} = -180.18 \times \text{Age} + 22075 + \epsilon$$

1.3

First, I use the standard formulas of OLS. First, I estimate the standard deviation of the regression residuals. Second, I use the formula

$$\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

to estimate the standard error of estimated regression coefficients. The results is 6.9687 for the coefficient of age and 357.8275 for the coefficient of constant.

Second I use bootstrap with 49 and 499 replications. For each replication, I draw 100 observations to estimate the regression coefficient. The result is

- With 49 replication: The mean of regression coefficients are -172.4 for age and 21687 for constant. The standard deviations are 22.6133 for the coefficient of age and 1145.711 for the coefficient of constant.
- With 499 replication: The mean of regression coefficients are -181.9 for age and 22147 for constant. The standard deviations are 22.5118 for the coefficient of age and 1273.761 for the coefficient of constant.

2

2.1

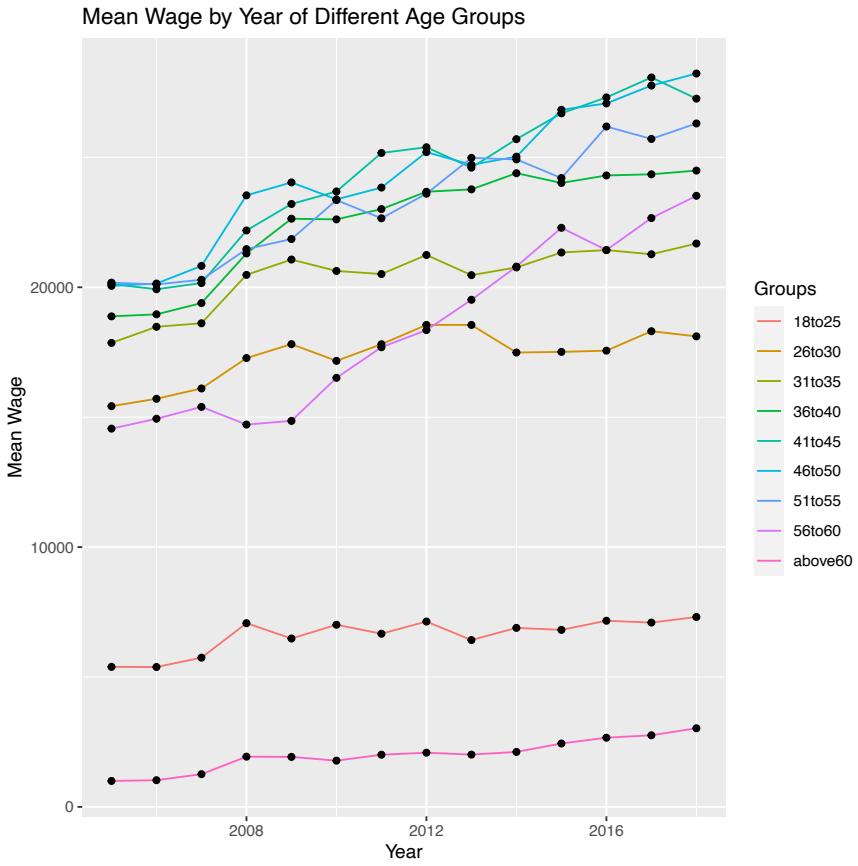
I generate the categorical variables as follows

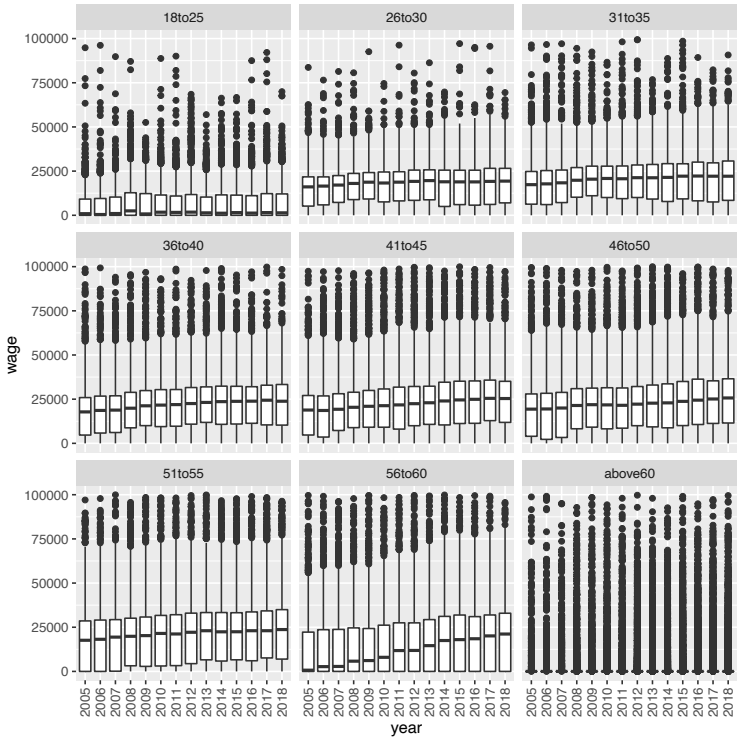
	age	wage	year	cat.age	cat.age.var
1	31	12334	2005	31to35	3
3	32	50659	2005	31to35	3
4	28	19231	2005	26to30	2
5	90	0	2005	above60	9
6	37	31511	2005	36to40	4
7	35	24873	2005	31to35	3
8	41	30080	2005	41to45	5
9	16	0	2005		0
10	55	43296	2005	51to55	7
11	55	20426	2005	51to55	7
12	57	0	2005	56to60	8
13	52	0	2005	51to55	7
16	51	0	2005	51to55	7
17	47	0	2005	46to50	6
19	55	49240	2005	51to55	7
20	17	0	2005		0
21	41	15005	2005	41to45	5
22	39	35192	2005	36to40	4
25	80	0	2005	above60	9
26	30	22852	2005	26to30	2
27	32	1832	2005	31to35	3
29	42	28247	2005	41to45	5
30	36	21134	2005	36to40	4
33	69	0	2005	above60	9
34	75	0	2005	above60	9
35	76	0	2005	above60	9
36	74	0	2005	above60	9
37	51	0	2005	51to55	7
38	41	0	2005	41to45	5
39	56	21051	2005	56to60	8
40	54	16168	2005	51to55	7
41	27	19688	2005	26to30	2
42	31	11666	2005	31to35	3
43	70	0	2005	above60	9
44	70	0	2005	above60	9

The variable is *cat.age* and *cat.age.var*

2.2

I make two figures. In the first figure, I plot the time-series of mean value of wage by different age group. In the second figure, I plot the boxplot to show the distribution. Please see the following two figures.


Mean Wage by Year of Different Age Groups
Mean Wage
20000
10000
0
2008
2012
2016
Year
Groups
18to25
26to30
31to35
36to40
41to45
46to50
51to55
56to60
above60


18to25
26to30
31to35
36to40
41to45
46to50
51to55
56to60
above60
wage
100000
75000
50000
25000
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
year

As is shown in this figure, there is a slightly increasing trend in mean wage of each age group.

2.3

I include the time fixed effect by adding the the dummy variable of each year in the regression. I set the year of 2005 as the benchmark and generate dummy variables from year 2006 to year 2018 as time control. The estimated regression coefficient is -186.8793 for age and 20675.0583 for constant. See the following figure for the output of this code.

```
> print(beta.est)
                [,1]
age        -186.87927
dummy2006    21.93723
dummy2007   294.80257
dummy2008  1425.19060
dummy2009  1720.36049
dummy2010  1869.52505
dummy2011  2116.01760
dummy2012  2601.22748
dummy2013  2478.84340
dummy2014  2749.67501
dummy2015  3120.96921
dummy2016  3410.11335
dummy2017  3479.03189
dummy2018  3636.15153
cons      20675.05832
```

when controlling for the time fixed effect, the estimated regression coefficients do not change a lot compared to the previous seciton.

3

3.1

Please see my code for this part. I delete the individuals who are retired or inactive in the data as

- data.datind2007[data.datind2007$empstat!="Inactive" & data.datind2007$empstat!="Retired",]

3.2

Please see my code for the code of estimating the likelihood.

```r
# Question 3.2
# We estimate the likelihood
likelihood <- function(beta, y, x) {
  x.beta <- beta[1] + beta[2]*x
  prob.y.est <- pnorm(x.beta)
  likelihood.negative <- -sum( (y*log(prob.y.est)) + (1-y)*log(1-prob.y.est) )
  return(likelihood.negative)
}
```

3.3

I use the R function, *optim*, in this question. The estimated coefficients are 1.0447 for constant and 0.0069 for age. Please see the following code:

```r
# Question 3.3
# For age and employment, I write the function as below
mle.opt.age.and.emp <- optim(fn=likelihood,
                             par= c(0, 0),
                             lower = c(-Inf, -Inf),
                             upper = c(Inf, Inf),
                             x = data.datind2007$age,
                             y = data.datind2007$dummy.emp
                             )
print(mle.opt.age.and.emp$par)
```

3.4

I try to use the same way in the previous question to estimate the model with wages as determinant. However, the R code could not generate estimation results. This is because of the feature of the data. If someone works, her/his wage should not be zero. If someone does not works, her/his wage should be zero. Thus, the coefficients by optimizing the likelihood will not converge to a point.

4

4.1

Please see my code and I drop the observations with "Inactive" and "Retired" from the dataset.

4.2

I generate year dummies to control the time fixed effect. I use the similar maximum likelihood as the previous section. The estimated regression coefficients are

- Probit: 0.0170 for age and 0.6126 for constant.
- Logistic: 0.0488 for age and 0.2034 for constant.

4.3

Holding everything else fixed (in our case, the control is the year). As age increase by 1, the probability will increase by 0.0170 when we use probit model to estimate, and the probability will increase by 0.0488 when we use logistic model.

5

5.1

The marginal effects of probit and logit model is the same as the regression coefficient of the previous probit and logit regression, which are 0.0170 for probit model and 0.0488 for logit model.

5.2

I use bootstrap to estimate the standard errors of the marginal effects. For each replication, I draw 5000 observations and I do 1000 replications for each model.

- Probit: mean coefficient for age is 0.0213, and the standard deviation is 0.0032.
- Logit: mean coefficient for age is 0.0470 and the standard deviation is 0.0034.