Econometrics I Homework 1

Eric Schulman

February 5, 2018

1 Part I

2.2

$$E(E(xy|x))$$
$$= E(xE(y|x))$$
$$= E(a(x + bE(x))|x)$$
$$= a(x + bE(x))$$

2.4

$$E(y|x = 0) = E(y^2|x = 0) = .8$$
$$E(y|x = 1) = E(y^2|x = 1) = .6$$
$$E(y^2|x = 0) - (E(y|x = 0))^2 = .16$$
$$E(y^2|x = 1) - (E(y|x = 1))^2 = .24$$

2.5 a)

$$E((e^2 - g(x))^2)$$

b)

$$\text{Minimize } E((e^2 - h(x))^2)$$

c)

$$E((e^2 - g(x))^2)$$
$$= E((e^2 - \sigma^2(x) + \sigma^2(x) - h(x))^2)$$
$$= E((e^2 - \sigma^2(x))^2) + E((\sigma^2(x) - h(x))^2) + 2E((\sigma^2(x) - h(x))(e^2 - \sigma^2(x)))$$

Since

$$E((\sigma^2(x) - h(x))(e^2 - \sigma^2(x)))$$
$$= E(E((\sigma^2(x) - h(x))(e^2 - \sigma^2(x))|x))$$
$$= E(E(e^2\sigma^2(x) - e^2g(x) + g(x)\sigma^2(x) + \sigma^2(x)\sigma^2(x)|x))$$
$$= E(\sigma^2(x)\sigma^2(x) - \sigma^2(x)g(x) + g(x)\sigma^2(x) + \sigma^2(x)\sigma^2(x)) = 0$$

We have

$$E((e^2 - g(x))^2) = E((e^2 - \sigma^2(x))^2) + E((\sigma^2(x) - h(x))^2) \geq E((e^2 - \sigma^2(x))^2)$$

2.7
$$E(y^2|x) - (E(y|x))^2 = \text{Var}(y|x) = \text{Var}(\beta x + e|x) = \text{Var}(e|x) = \sigma^2(x)$$

2.10 True

$$E(x^2 e) = E(E(ex^2|x)) = E(x^2 E(e|x)) = 0$$

2.11 False

Suppose x is distributed $N(0,1)$, $e|x$ is distributed $N(x^2,1)$. Then,

$$E(xe) = E(E(xe|x) = E(xE(e|x)) = E(x^3) = 0$$

However,

$$E(x^2 e) = E(E(x^2 e|x)) = E(x^2 E(e|x)) = E(x^4) = 3\sigma^4 = 3$$

i.e. the fourth moment of the normal distribution

2.12 False

Suppose $e|x$ has the distribution $N(0,x^2)$. Then $E(e|x) = 0$, but x and e are not independent.

2.13 False

Suppose x is distributed $N(0,1)$, $e|x$ is distributed $N(x^2,1)$. Then,

$$E(ex) = 0$$

(as shown in 2.11)

$$E(e|x) = x^2$$

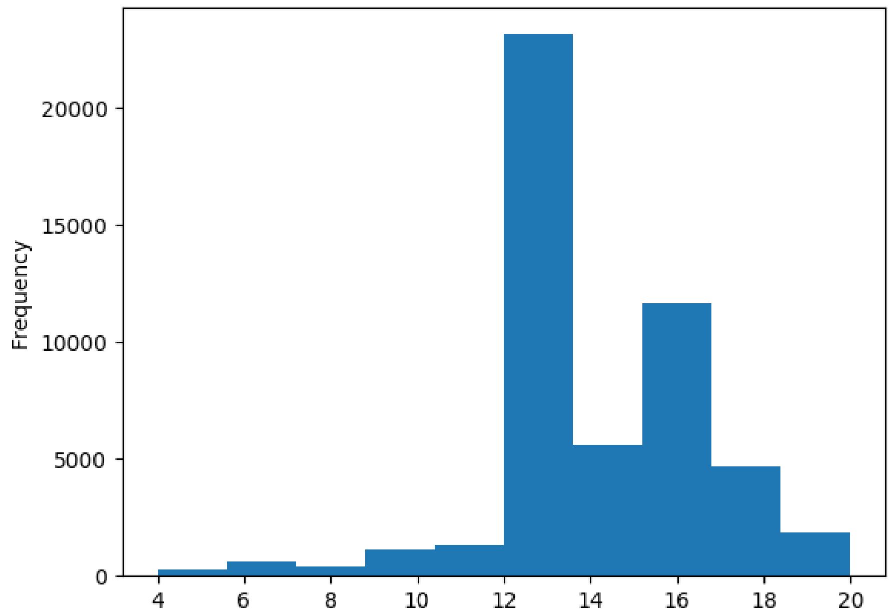
2.14 False

Consider a distribution where the third conditional moment is a function of x. i.e.

$$E(x|e) = 0$$
$$E(x^2|e) = 0$$
$$E(x^3|e) = h(x)$$

2 Part II

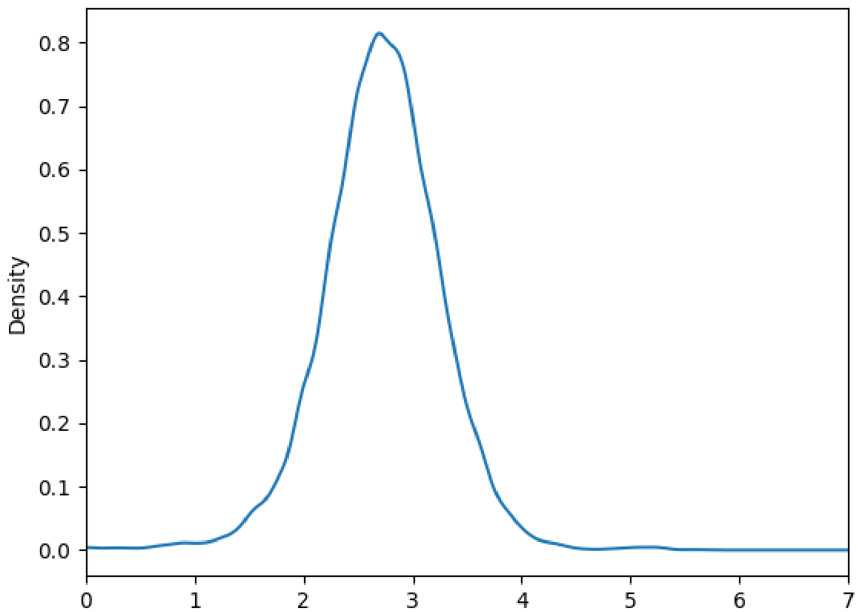
a) Below is the histogram. The histogram appears to have 2 peaks.

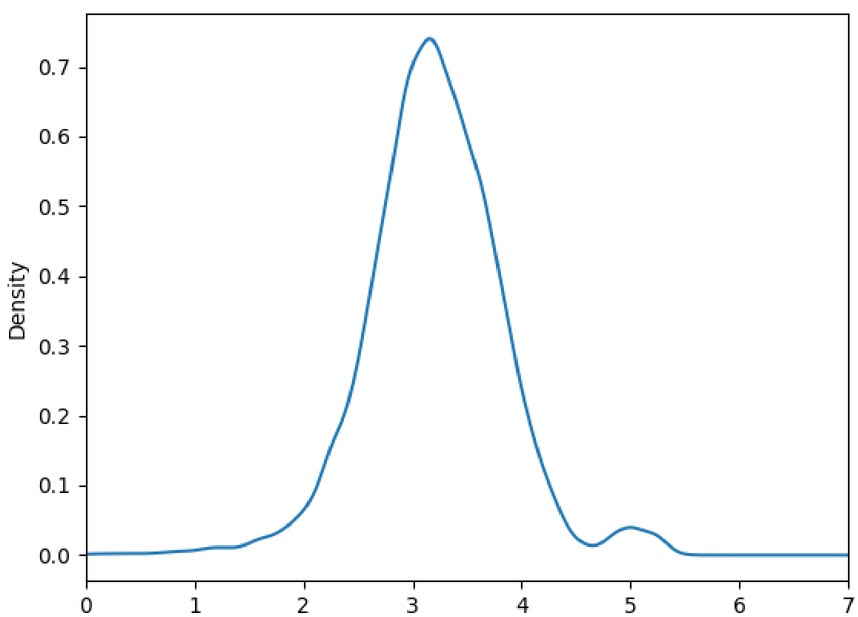


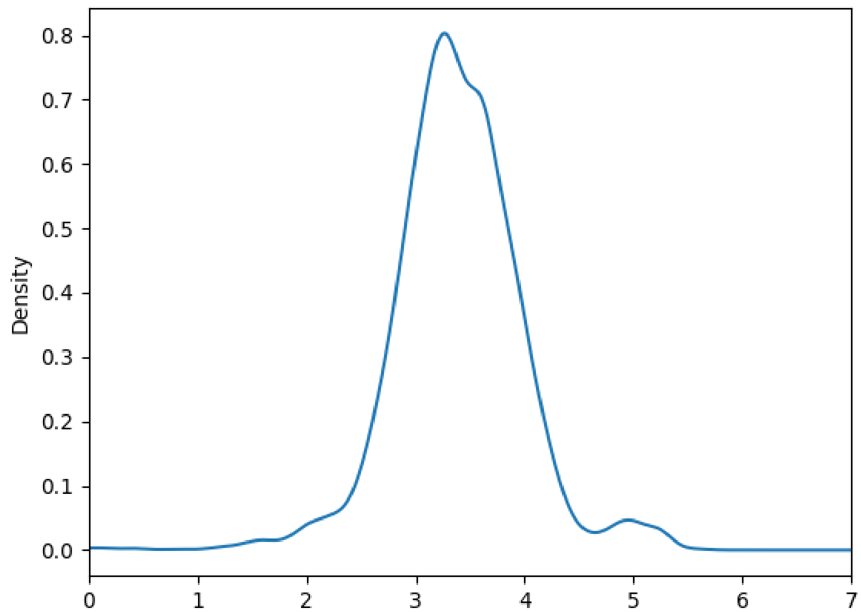
b) The table below show the results.

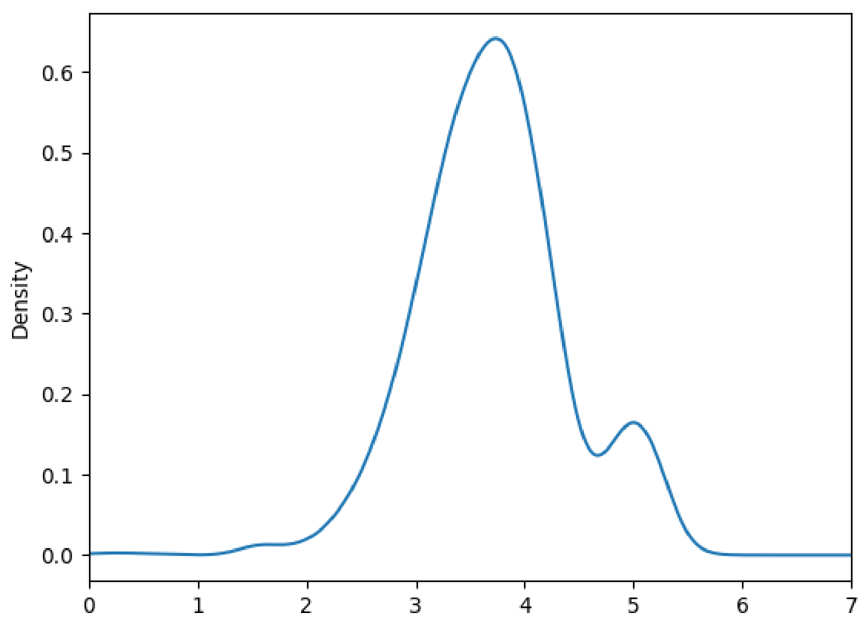
Years	Observations
12	13896
16	11640
18	4670
20	1875

c) Below are the kernel density plots in order of years of education (i.e. 12, 16, 18, 20).


Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
2
3
4
5
6
7


Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
4
5
6
7





d) The table below show the results.

Years	Mean	Variance
12	2.712	0.329
16	3.201	0.416
18	3.381	0.363
20	3.689	0.619

By the table, we can see 20 years of education has the highest variance, followed by 16 years of education. 12 years of education has the least variance.

e) The table below show the results.

Year	25th	50th	75th
12	2.403	2.733	3.055
16	2.839	3.180	3.585
18	3.062	3.362	3.710
20	3.275	3.690	4.089

Again the distribution for 20 years of education is has the most dispersion. The highest quartile of 12 years of education make less than the lowest quartile of 20 years of education.

f) The table below show the results.

Years	Difference
12	0
16	0.489
18	0.669
20	0.977

The more education you have the more you are expected to make according to the differences in the table.

g) The table below show the results.

Year 1	Year 2	Diff	SE	T-Value	Reject
12	12	0.0	0.00486691158299	0.0	False
12	16	-0.488918	0.00486691158299	-100.457462777	True
12	18	-0.669176	0.00486691158299	-137.494970241	True
12	20	-0.976912	0.00486691158299	-200.725245359	True
16	12	0.488918	0.00597543294539	81.821282852	True
16	16	0.0	0.00597543294539	0.0	False
16	18	-0.180258	0.00597543294539	-30.1665629463	True
16	20	-0.487994	0.00597543294539	-81.6667908266	True
18	12	0.669176	0.00882675326792	75.8122316274	True
18	16	0.180258	0.00882675326792	20.4218095382	True
18	18	0.0	0.00882675326792	0.0	False
18	20	-0.307736	0.00882675326792	-34.8640263334	True
20	12	0.976912	0.0181678405989	53.7714989472	True
20	16	0.487994	0.0181678405989	26.8603431317	True
20	18	0.307736	0.0181678405989	16.9385104793	True
20	20	0.0	0.0181678405989	0.0	False

h) The table below show the results.

Dep. Variable:	lwage	**R-squared:**	0.203
Model:	OLS	**Adj. R-squared:**	0.203
Method:	Least Squares	**F-statistic:**	2727.
Date:	Sat, 03 Feb 2018	**Prob (F-statistic):**	0.00
Time:	13:45:04	**Log-Likelihood:**	-30104.
No. Observations:	32081	**AIC:**	6.022e+04
Df Residuals:	32077	**BIC:**	6.025e+04
Df Model:	3		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2.7121	0.005	516.939	0.000	2.702	2.722
educ_16	0.4889	0.008	62.916	0.000	0.474	0.504
educ_18	0.6692	0.010	63.969	0.000	0.649	0.690
educ_20	0.9769	0.015	64.203	0.000	0.947	1.007

Omnibus:	10460.946	**Durbin-Watson:**	1.767
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	213631.163
Skew:	-1.068	**Prob(JB):**	0.00
Kurtosis:	15.460	**Cond. No.**	4.98

3 Python Code

```
import pandas
import matplotlib.pyplot as plt
import statsmodels.api as sm

FNAME = 'cps09mar.dta'
CATEGORIES = [12,16,18,20]

def main():
        #part a
        df = pandas.read_stata(FNAME)
        plt.figure()
        hist = df[(df.educ > 1)]
        hist['educ'].plot.hist()
        plt.savefig('part_a')

        #part b-e
        stats = dict()
        for y in CATEGORIES:
                stats[y] = helper(df,y)

        #write result to file
```

```
        result = open("hw1_results.csv","w+")
        result.write('num_obs,mean,var,q25,q50,q75\n')
        for y in CATEGORIES:
                result.write('%s,%s,%s,%s,%s,%s\n'% tuple(stats[y]) )

        result.close()

        #part f-g
        result = open("hw1_results.tex","w+")
        result.write('\\begin{tabular}{ c c c c c c c } \\\\')
        result.write(' Year 1 & Year 2 & Diff & SE & T-Value & Reject \\\\ \
        for y in CATEGORIES:
                for x in CATEGORIES:
                        diff = stats[y][1] - stats[x][1]
                        se =  (stats[y][2]/stats[y][0])**(.5)
                        t_value = diff / se
                        reject = (t_value > 2.02) or (t_value < -2.02)
                        result.write('%s & %s & %s & %s & %s & %s \\\\ \n'%
        result.write('\\end{tabular}')
        result.close()


def helper(df,y):
        df = df[df.educ == y]
        num_obs = df['educ'].count()

        plt.figure()
        df['lwage'].plot.density()
        plt.xlim(0, 7)
        plt.savefig('part_b_%s'%y)

        mean = df['lwage'].mean()
        var = df['lwage'].var()

        q25 = df['lwage'].quantile(q=.25)
        q50 = df['lwage'].quantile(q=.5)
        q75 = df['lwage'].quantile(q=.75)

        #last 2 fields to be filled in later
        return [num_obs,mean,var,q25,q50,q75]


def do_regression():
        df = pandas.read_stata(FNAME)
        df = df[(df.educ == 12) | (df.educ == 16)| (df.educ == 18)| (df.educ
```

```
        #set up X matrix
        X = pandas.get_dummies(df['educ'], prefix='educ')
        X = X[['educ_16', 'educ_18', 'educ_20']]
        X = sm.add_constant(X)

        y = df['lwage']

        #write result
        model_result = sm.OLS(y,X).fit()
        result_doc = open('hw1_reg.tex','w+')
        result_doc.write( model_result.summary().as_latex() )
        result_doc.close()


if __name__ == "__main__":
        main()
        do_regression()
```