20 Sept' 202-

EC401 ASSIGNMENT 1

Compiled by - DIVYA
2122770708 7

DATA

Koop and Tobias (2004) Labour Market Experience Data is a panel of 2178 individuals with a total of 17,919 Observations. I have fixed the Time Trend takes values from 0 to 14. I have fixed the time trend at 14, recent most year. This leaves us with 1499 observations.
I have applied filter Time Trend = 14, to get this subset of data and stored it in excel file titled "kt.xlsx".

IMPORTING LIBRARIES

I imported 'numpy' - for linear algebra; 'pandas' - for dealing with data frames; 'matplotlib.pyplot' - for visualization and; 'statsmodel.api' for OLS and 'spicy.stats' for statistical tests.

DATA DESCRIPTION

1. Imported data with time trend fixed at 14 as 'kt'. The choice of 14 was driven by the fact that it was recent most recent data of within the panel and levels of education had stabilised by then.
2. Displaying first 10 entries of the data.

```
kt.head(10)
```

	PERSONID	EDUC	LOGWAGE	POTEXPER	TIMETRND	ABILITY	MOTHERED	FATHERED	BRKNHOME	SIBLINGS
1	2	15	2.60	12	14	1.50	12	12	0	1
2	4	13	2.12	11	14	0.26	12	10	1	4
3	6	15	2.70	14	14	0.44	12	16	0	2
4	7	15	2.35	9	14	0.91	12	12	0	1
5	8	13	2.01	18	14	0.51	12	15	1	2
6	10	11	2.60	16	14	0.26	12	12	0	2
7	12	13	2.91	16	14	-1.30	13	12	0	5
8	13	12	3.88	17	14	-0.63	12	12	1	4
9	15	13	3.22	12	14	0.28	10	12	1	3
10	16	12	2.56	12	14	-0.72	14	12	0	1

* the default index as 0, command was given to change index from 0 to 1

3. Now to get details about columns - count and datatype.

```
✓ kt.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1499 entries, 1 to 1499
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   PERSONID   1499 non-null   int64
 1   EDUC       1499 non-null   int64
 2   LOGWAGE    1499 non-null   float64
 3   POTEXPER   1499 non-null   int64
 4   TIMETRND   1499 non-null   int64
 5   ABILITY    1499 non-null   float64
 6   MOTHERED   1499 non-null   int64
 7   FATHERED   1499 non-null   int64
 8   BRKNHOME   1499 non-null   int64
 9   SIBLINGS   1499 non-null   int64
dtypes: float64(2), int64(8)
memory usage: 117.2 KB
```

In the dataset 'kt';

→ There are a total of 10 columns

→ Each column has 1499 non null enterus

→ Person Id, Education, experience, timetrend, mother's educations, fathers education, Broken home and siblings are take integer values

→ Logwage and Ability are float

4. Adding a column titled 'Constant' with value = 1 everywhere; or adding ~~ia~~ 1499 X 1 vector of 1s, to the dataframe kt.

5. Descriptive Statistics about the ~~reverse~~ ~~kt~~ variables in kt I have ignored person Id as s.d. ~~count~~, mean etc don't make sense for Person Id.

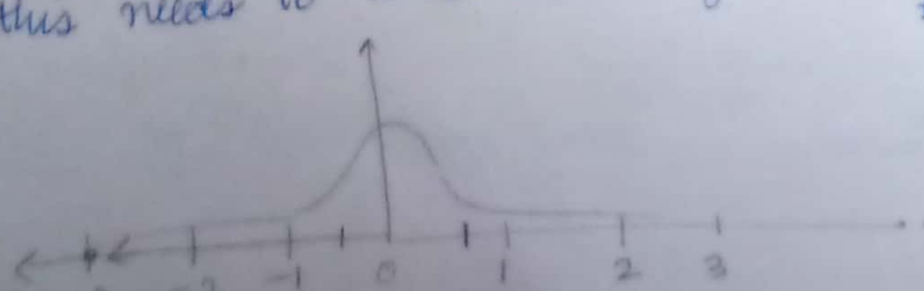
```
✓ kt.loc[:, kt.columns != 'PERSONID'].describe()
```

	EDUC	LOGWAGE	POTEXPER	TIMETRND	ABILITY	MOTHERED	FATHERED	BRKNHOME	SIBLINGS	CONSTANT
count	1499.000000	1499.000000	1499.000000	1499.0	1499.000000	1499.000000	1499.000000	1499.000000	1499.000000	1499.0
mean	13.110740	2.495484	13.392929	14.0	0.120734	11.609740	11.919947	0.152101	3.063376	1.0
std	2.192598	0.546943	3.028962	0.0	0.932445	3.026882	3.819037	0.359239	2.036873	0.0
min	9.000000	0.210000	4.000000	14.0	-3.960000	0.000000	0.000000	0.000000	0.000000	1.0
25%	12.000000	2.180000	11.000000	14.0	-0.460000	11.000000	10.000000	0.000000	2.000000	1.0
50%	12.000000	2.500000	13.000000	14.0	0.280000	12.000000	12.000000	0.000000	3.000000	1.0
75%	15.000000	2.840000	16.000000	14.0	0.840000	12.000000	14.000000	0.000000	4.000000	1.0
max	20.000000	4.320000	22.000000	14.0	2.010000	20.000000	20.000000	1.000000	15.000000	1.0

Inferences

→ On average, mother's have a slightly less level of education than fathers; median level of education is same = 12; but for mothers 75 % quantile is 12 but for father's it is 14, Thus a larger proportion of fathers tend to have education level greater than 14. relative to mothers.

→ Ability is concentrated around median, as the ~~med~~ min is -3.96 and about 25 % obs are left of -0.46, also the max is 2.01 but only 25 % obs are right of 0.84; this needs to checked though. as it could be the case that the values are concentrated beyond -0.46 and 0.84 and min max are just outlier



Since descriptive statistics not are prime concern, going ahead.

DEFINING X1

$X_1 = [\text{Constant}, \text{Educ}, \text{Potenper}, \text{Ability}]$ is 1499 X4 matrix

DEFINING X2

$X2 = [\text{MotherEd}, \text{FatherEd}, \text{Siblings}]$ 1499 X 3 matrix

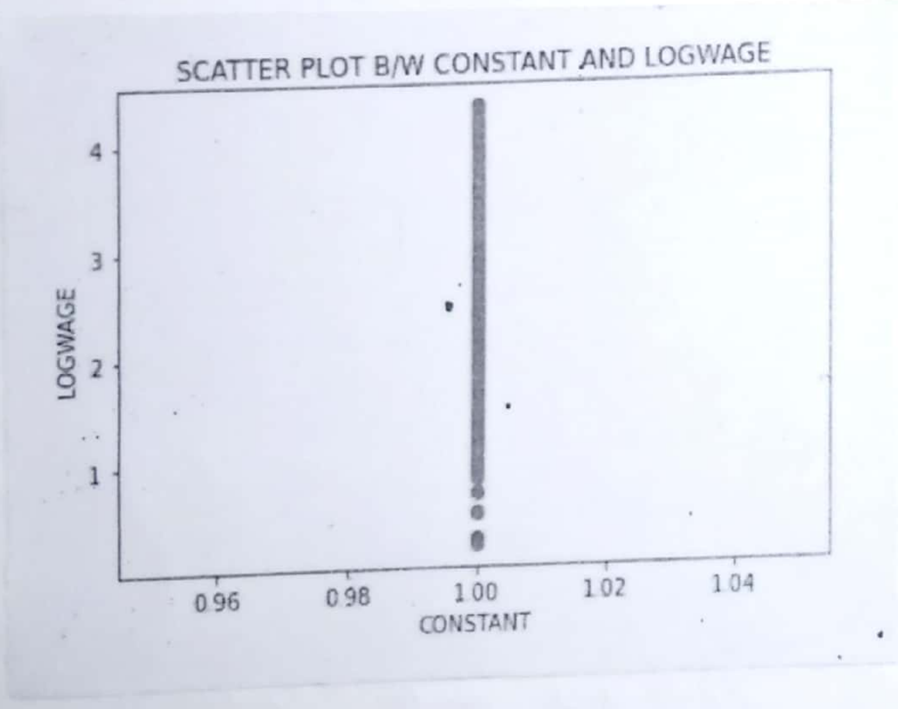
DEFINING Y

$Y = [\text{LogWage}]$ 1499 X 1 matrix

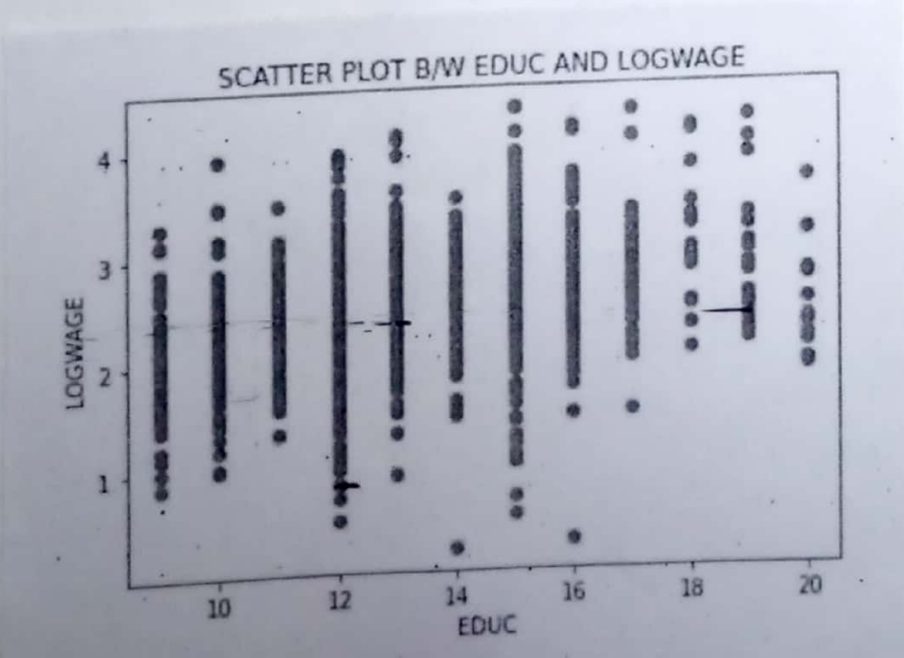
DEFINING X

$X = [X_1\ X_2]$ = 1499 X 7 matrix.

Question 1 Scatter Plot of Y on X1

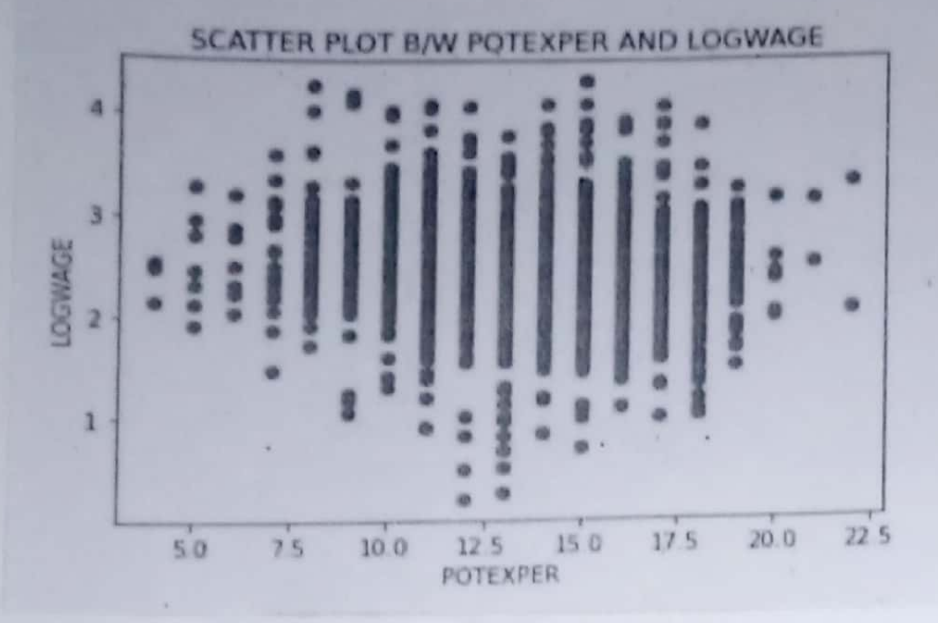


This graph doesn't offer much and is equivalent to plotting Y on a number line. We can infer logwage takes only positive values varying b/w 0 & 5.

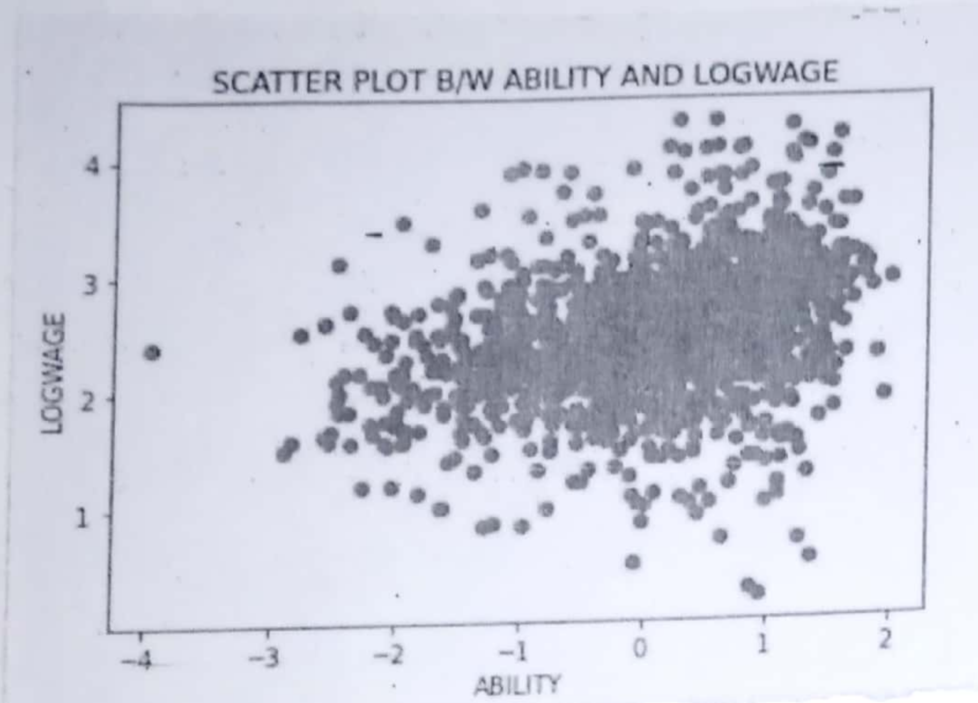


Education and log wage seem to be slightly positively correlated.

13



Experience and logwage seem to be slightly negatively correlated



Ability and logwage appear to be positively correlated – moderately.

Question 2 Correlation b/w Each Variable in X and Y.

```
kt_corr=X.apply(Y["LOGWAGE"].corr)

CONSTANT         NaN
EDUC        0.319125
POTEXPER   -0.133602
ABILITY     0.311470
MOTHERED    0.211839
FATHERED    0.228093
SIBLINGS   -0.085604
dtype: float64
```

→ NaN to be interpreted as no correlation, which is true as variable "constant" is a constant.

→ Level of education is positively correlated as expected.

→ Experience is negatively correlated with logwage which is a little counter-intuitive as experience higher should be associated with greater wages. But something else like Age could be confounding this relationship.

→ Ability and logwage are positively correlated as expected

→ Higher than avg levels of mothers education and fathers education are associated with higher than normal levels of logwage

→ Siblings is negatively associated correlated with logwage.

4

Question 3 Estimated Regression Formula

$$\text{Logwage} = \beta_0 \cdot 1 + \beta_1 \cdot \text{Educ} + \beta_2 \cdot \text{Potenper} + \beta_3 \cdot \text{Ability} + \beta_6 \text{ Siblings} + \beta_4 \cdot \text{MotherEd} + \beta_5 \cdot \text{FatherEd} + e$$

Linear Multiple Reg. Model

$$Y = \sum \beta_i X_i + e$$

Expected Signs

$\beta_0 > 0$ - Due to minimum wage laws, subsistence wages etc.

$\beta_1 > 0$ - ~~conditional~~ Increase in level of education, given everything else ceteris paribus should lead to increase in wage

$\beta_2 > 0$ - Increase in ~~level of edu~~ experience, ceteris paribus should lead to increase in logwage as experience is valued in the industry.

$\beta_3 > 0$ - Increase in ability, ceteris paribus, should lead to increase in logwage ~~as~~.

$\beta_4, \beta_5 > 0$ ~~Increase in~~ Individuals with greater father's or mothers education, conditional on everything else being same will have higher log wages due to synergy and guidance and opportunities exploited by such Individual.

$\beta_6 < 0$ This is because of resource constraints faced by family and resources that received per head are likely to fall with increase in no. of siblings

Question 4 β's ~~using Inbuilt~~ without using Inbuilt.

I have ~~β ..~~ used matrix multiplication to calculate $\hat{\beta}$

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

$\hookrightarrow 7 \times 1$

		0
β_0	0	0.963104
β_1	1	0.073498
β_2	2	0.029154
β_3	3	0.100931
β_4	4	0.005060
β_5	5	0.008557
β_6	6	0.001752

Signs are as expected except the coefficient associated with siblings
However we can't make any conclusion yet - need to look at pvalues for significance

Question 5 : Using Inbuilt OLS Estimation functions

Regressed Y on X

```
lm_y_on_x=sm.OLS(Y,X).fit()
```

OLS Regression Results

Dep. Variable:	LOGWAGE	R-squared:	0.145
Model:	OLS	Adj. R-squared:	0.142
Method:	Least Squares	F-statistic:	42.19
Date:	Tue, 20 Sep 2022	Prob (F-statistic):	1.01e-47
Time:	15:29:59	Log-Likelihood:	-1104.5
No. Observations:	1499	AIC:	2223.
Df Residuals:	1492	BIC:	2260.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CONSTANT	0.9631	0.182	5.305	0.000	0.607	1.319
EDUC	0.0735	0.009	8.092	0.000	0.056	0.091
POTEXPER	0.0292	0.006	4.952	0.000	0.018	0.041
ABILITY	0.1009	0.018	5.536	0.000	0.065	0.137
MOTHERED	0.0051	0.006	0.821	0.412	-0.007	0.017
FATHERED	0.0086	0.005	1.767	0.077	-0.001	0.018
SIBLINGS	0.0018	0.007	0.253	0.801	-0.012	0.015

Omnibus:	97.109	Durbin-Watson:	1.903
Prob(Omnibus):	0.000	Jarque-Bera (JB):	246.656
Skew:	-0.357	Prob(JB):	2.75e-54
Kurtosis:	4.855	Cond. No.	354.

Coefficients generated in Q4 are approximately same as ones generated by using inbuilt regression functions

* In Q5, they have been rounded off and hence have used 'approximately' though they should be exactly the same

$0 \notin CI$

$\left.\right\}$ $p < 0.05$ Significant coefficients

$\left.\right\}$ $p > 0.05$ Insignificant

$0 \in CI$

Question 6 : Regress each of three variables in X_2 on X_1; compute residuals; arrange them in $X2^*$; Sample Mean; Explain

Regressing MotherEd on X1

```
lm_mothered=sm.OLS(X2.MOTHERED,X1).fit()
```

```
lm_mothered.summary()
```

OLS Regression Results

Dep. Variable:	MOTHERED	R-squared:	0.227
Model:	OLS	Adj. R-squared:	0.225
Method:	Least Squares	F-statistic:	145.9
Date:	Tue, 20 Sep 2022	Prob (F-statistic):	6.02e-83
Time:	14:22:28	Log-Likelihood:	-3594.2
No. Observations:	1499	AIC:	7196.
Df Residuals:	1495	BIC:	7218.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CONSTANT	8.5475	0.919	9.302	0.000	6.745	10.350
EDUC	0.2397	0.047	5.093	0.000	0.147	0.332
POTEXPER	-0.0161	0.031	-0.524	0.600	-0.076	0.044
ABILITY	1.1190	0.090	12.376	0.000	0.942	1.296

Omnibus:	295.740	Durbin-Watson:	1.497
Prob(Omnibus):	0.000	Jarque-Bera (JB):	770.336
Skew:	-1.042	Prob(JB):	5.29e-168
Kurtosis:	5.826	Cond. No.	252.

```
lm_fathered=sm.OLS(X2.FATHERED,X1).fit()
```

OLS Regression Results

Dep. Variable:	FATHERED	R-squared:	0.241
Model:	OLS	Adj. R-squared:	0.239
Method:	Least Squares	F-statistic:	157.8
Date:	Tue, 20 Sep 2022	Prob (F-statistic):	7.13e-89
Time:	15:21:05	Log-Likelihood:	-3928.9
No. Observations:	1499	AIC:	7866.
Df Residuals:	1495	BIC:	7887.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CONSTANT	7.2298	1.149	6.293	0.000	4.976	9.483
EDUC	0.3742	0.059	6.359	0.000	0.259	0.490
POTEXPER	-0.0279	0.038	-0.726	0.468	-0.103	0.047
ABILITY	1.3109	0.113	11.597	0.000	1.089	1.533

Omnibus:	122.046	Durbin-Watson:	1.543
Prob(Omnibus):	0.000	Jarque-Bera (JB):	186.377
Skew:	-0.620	Prob(JB):	3.38e-41
Kurtosis:	4.202	Cond. No.	252.

Regressing Father Ed on X1

Regressing Siblings on X1

```
X2_asterisk.describe()
```

	mothered_resid	fathered_resid	siblings_resid
count	1.499000e+03	1.499000e+03	1.499000e+03
mean	-1.357212e-14	-1.521457e-14	-3.762464e-15
std	2.662055e+00	3.328148e+00	1.957301e+00
min	-1.295448e+01	-1.412118e+01	-4.071056e+00
25%	-1.072323e+00	-1.748878e+00	-1.314880e+00
50%	2.580314e-01	2.548543e-01	-3.512178e-01
75%	1.537228e+00	2.018712e+00	9.422830e-01
max	7.520909e+00	9.118324e+00	1.132803e+01

Storing residuals from the above three mentioned regressions in X2_asterisk

The sample mean of each of the residuals is approximately 0. Though to check this and be able to say with confidence, I run one sample t-test.

Regressing Father Ed on X1

```
lm_siblings=sm.OLS(X2.SIBLINGS,X1).fit() ···
```

```
lm_siblings.summary() ···
```

Regressing Siblings on X1

OLS Regression Results

Dep. Variable:	SIBLINGS	R-squared:	0.077
Model:	OLS	Adj. R-squared:	0.075
Method:	Least Squares	F-statistic:	41.34
Date:	Tue, 20 Sep 2022	Prob (F-statistic):	1.15e-25
Time:	14:23:42	Log-Likelihood:	-3133.2
No. Observations:	1499	AIC:	6274.
Df Residuals:	1495	BIC:	6296.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CONSTANT	1.9729	0.676	2.920	0.004	0.648	3.298
EDUC	-0.0161	0.035	-0.464	0.643	-0.084	0.052
POTEXPER	0.1004	0.023	4.438	0.000	0.056	0.145
ABILITY	-0.3570	0.066	-5.371	0.000	-0.487	-0.227

Omnibus:	296.901	Durbin-Watson:	1.628
Prob(Omnibus):	0.000	Jarque-Bera (JB):	645.544
Skew:	1.118	Prob(JB):	6.64e-141
Kurtosis:	5.311	Cond. No.	252.

Storing residuals from the above three mentioned regressions in X2_asterisk

The sample mean of each of the residuals is approximately 0.

Though to check this and be able to say with confidence, I run one sample t-test.

Hypothesis Testing

Under Null H_0^1 pop mean here mean of mothered_resid = 0

```
stats.ttest_1samp(a=X2_asterisk.mothered_resid, popmean=0)
Ttest_1sampResult(statistic=-2.0130528615766356e-13, pvalue=0.9999999999998395)
```

Since p values large, fail to reject Ho

Illy

H_0^2 = mean of fathered_resid = 0

```
stats.ttest_1samp(a=X2_asterisk.fathered_resid, popmean=0)
Ttest_1sampResult(statistic=-1.685156193819453e-13, pvalue=0.9999999999998656)
```

p values large fail to reject H_0^2

Illy

H_0^3 = mean of siblings_resid = 0

```
stats.ttest_1samp(a=X2_asterisk.siblings_resid, popmean=0)
Ttest_1sampResult(statistic=-7.351023470063113e-14, pvalue=0.9999999999999414)
```

pvalues large fail to reject H_0^3.

Thus none of the residual means are significantly different from 0.

QUESTION 7   Regressing Y on X1

```
lm_y_on_x1=sm.OLS(Y,X1).fit()
```

OLS Regression Results

Dep. Variable:	LOGWAGE	R-squared:	0.140
Model:	OLS	Adj. R-squared:	0.139
Method:	Least Squares	F-statistic:	81.39
Date:	Tue, 20 Sep 2022	Prob (F-statistic):	8.93e-49
Time:	15:25:34	Log-Likelihood:	-1108.6
No. Observations:	1499	AIC:	2225.
Df Residuals:	1495	BIC:	2246.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CONSTANT	1.0717	0.175	6.123	0.000	0.728	1.415
EDUC	0.0779	0.009	8.686	0.000	0.060	0.095
POTEXPER	0.0290	0.006	4.952	0.000	0.018	0.041
ABILITY	0.1172	0.017	6.804	0.000	0.083	0.151

Omnibus:	100.501	Durbin-Watson:	1.895
Prob(Omnibus):	0.000	Jarque-Bera (JB):	260.425
Skew:	-0.364	Prob(JB):	2.81e-57
Kurtosis:	4.908	Cond. No.	252.

Coefficents here are diff from ones reported in Q 5, happens because of cov b/w X_2 and X1 Variables.

All significant $0 \notin CI$

Question 8   Y on X_1 and X_2

This exercise is exactly same as performed under Q 5.

Q 5 would have yielded different coefficents with the regression formula was differently defined – eg Brokenhome dummy was included.

However since Q5. Y regressed on X_1 and X_2; coeffients are same in Q5 and Q 8.

Question 9   Y on X1 and X2_asterisk

```
✓ lm_y_on_x1_and_x2_asterisk=sm.OLS.from_formula('LOGWAGE ~ CONSTAI
✓ lm_y_on_x1_and_x2_asterisk.summary()
```

OLS Regression Results

Dep. Variable:	LOGWAGE	R-squared:	0.145
Model:	OLS	Adj. R-squared:	0.142
Method:	Least Squares	F-statistic:	42.19
Date:	Tue, 20 Sep 2022	Prob (F-statistic):	1.01e-47
Time:	14:29:40	Log-Likelihood:	-1104.5
No. Observations:	1499	AIC:	2223.
Df Residuals:	1492	BIC:	2260.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CONSTANT	1.0717	0.175	6.133	0.000	0.729	1.414
EDUC	0.0779	0.009	8.701	0.000	0.060	0.095
POTEXPER	0.0290	0.006	4.960	0.000	0.018	0.040
ABILITY	0.1172	0.017	6.816	0.000	0.083	0.151
X2_asterisk.mothered_resid	0.0051	0.006	0.821	0.412	-0.007	0.017
X2_asterisk.fathered_resid	0.0086	0.005	1.767	0.077	-0.001	0.018
X2_asterisk.siblings_resid	0.0018	0.007	0.253	0.801	-0.012	0.015

Omnibus:	97.109	Durbin-Watson:	1.903
Prob(Omnibus):	0.000	Jarque-Bera (JB):	246.656
Skew:	-0.357	Prob(JB):	2.75e-54
Kurtosis:	4.855	Cond. No.	252.

Refer Q 7

→ Coefficents of X1 are exactly same as ones obtained by regressing Y on X1

→ Coefficents on X2_asterisk variables are exactly same as ones obtained by regressing Y on X_1 and X2

Refer Q 8 / Q 5

~~This is Frisch Waugh Lovell Theorem~~

FWG Theorem holds

Scanned with CamScanner