FAHRİ ULKAT 090220756

Question 1

Define the following terms in your own words.

1. The null hypothesis: Hypothesis to be tested called null hypothesis.
2. The alternative hypothesis: The scenario when the null is not satisfied, opposing the null.
3. A test statistic: A calculated value from sample data to decide to reject null or fail to reject null.
4. Type-I error: The situation when we reject the true null hypothesis.
5. Type-II error: The situation when we fail to reject false null hypothesis.
6. Significance level: The threshold probability for rejectin the null, representing the risk of type-I error.
7. The 95% confidence interval: It means 95 times out of 100 times of chosen statistic will be in this interval.
8. Heteroskedasticity: A condition where the error variance change across values of independent variable.
9. The Gauss-Markov theorem: A theorem stating that under certain assumptions OLS estimator are Best Linear Unbiased Estimator (BLUE).

Question 2

In [29]:
```python
from IPython.display import Image, display

display(Image(filename="C:\\Users\\fahri\\Desktop\\Screenshot_1.png"))
```

Question 2

Suppose a researcher, using wage data on 250 randomly selected male workers and 280 female workers, estimates the OLS regression

$$\widehat{Wage} = \underset{(0.23)}{12.52} + \underset{(0.36)}{2.12} \times Male, \quad R^2 = 0.06, \quad SER = 4.2$$

where $Wage$ is measured in dollars per hour and $Male$ is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. Define the gender gap as the difference in mean earnings between men and women.

a. Interprete the estimated coefficient on $Male$. What is the estimated gender gap?
b. Is the estimated gender gap significantly different than zero? (Compute the p-value)
c. Construct a 95% confidence interval for the gender gap.
d. In the sample, what is the mean wage of women? Of men?
e. Another researcher uses the same data but regresses $Wage$ on $Female$, a binary variable that is equal to 1 if the person is a female and 0 if the person is a male. What are the regression estimates calculated from this regression?

$$\widehat{Wage} = \widehat{\gamma}_0 + \widehat{\gamma}_1 \times Female, \quad R^2 = ?, \quad SER = ?$$

a. The Coefficient 2.12 means that being male makes your wage higher of 2.12. Estimated gender gap equals the coefficient, so it is 2.12.

In [30]:
```python
import numpy as np
from scipy.stats import norm

#coefs
b0 = 12.52
b1 = 2.12

#standard errors
seb0 = 0.23
seb1 = 0.36


r2 = 0.06
ser = 4.2

#Option B
t_stat = (b1-0)/seb1

print(f"t-stat = {t_stat}")

p_value = 2 * (1 - norm.cdf(abs(t_stat), loc=0, scale=1))

print(f"p-value = {p_value}")
```

```
t-stat = 5.888888888888889
p-value = 3.888007249486236e-09
```

b. Since p-value is much smalle than 0.05 we can reject the null hypothesis

In [31]:
```python
upper_limit = 2.12 + 1.96*0.34
lower_limit = 2.12 - 1.96*0.34

print(f"c. CI: {lower_limit} < gender gap < {upper_limit}")
```

```
c. CI: 1.4536000000000002 < gender gap < 2.7864
```

d.

1. Mean avarage for women: 12.52 + 2.12 * 0 = 12.52
2. Mean avarage for women: 12.52 + 2.12 * 1 = 14.64

We know that mean wage for men equasl 14.64 so new beta0 will be 14.64 and beinf femala will decerase the mean so new beta1 will be -2.12. Thus,

- Wage = 14.64 - 2.12 x Female

R2 and SER will be same because changind dummy variable does not change relationship between Y and D.

- R2 = 0.06 SER = 4.2

Question 3

In [32]:
```python
display(Image(filename="C:\\Users\\fahri\\Desktop\\Screenshot_3.png"))
```

Question 3

Each month the Bureau of Labor Statistics in the U.S. Department of Labor conducts the Current Population Survey (CPS), which provides data on labor force characteristics of the population, including the level of employment, unemployment, and earnings. Approximately 65,000 randomly selected U.S. households are surveyed each month. The sample is chosen by randomly selecting addresses from a database comprised of addresses from the most recent decennial census augmented with data on new housing units constructed after the last census. The exact random sampling scheme is rather complicated (first small geographical areas are randomly selected, then housing units within these areas randomly selected); details can be found in the Handbook of Labor Statistics and is described on the Bureau of Labor Statistics website (www.bls.gov). The R package `AER` provides several data sets constructed from the CPS. For this exercise, you will utilize the data set called `CPSSW8'. Use the following code chunk to load data:

```r
library(AER)
data("CPSSW8")
names(CPSSW8)
```

```
[1] "earnings"  "gender"    "age"       "region"     "education"
```

If you are using Python for this exercise, use the `pandas` module to import the data contained in the `CPSSW8.xlsx` file.

a. Run a regression of (average hourly) earnings on education and compute heteroskedasticity robust standard errors.
b. Is the estimated education effect significantly different than zero? Compute the t-statistic and the p-value.
c. Construct a 90% confidence interval for the coefficient of education.

In [33]:
```python
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt
import warnings
import statsmodels.formula.api as smf
import statsmodels.api as sm
import scipy.stats as stats


warnings.simplefilter("ignore", UserWarning)
```

a.

In [34]:
```python
df = pd.read_excel(r"C:\Users\fahri\Desktop\itü\econ\ecn301e\ProblemSet06\ProblemSet6\ProblemSet6\CPSSW8.xls

df.head()
```

Out[34]:

	earnings	gender	age	region	education
0	20.673077	male	31	South	14
1	24.278847	male	50	South	12
2	10.149572	male	36	South	12
3	8.894231	female	33	South	10
4	6.410256	female	56	South	10

In [35]:
```python
model=smf.ols(formula='earnings ~ education',data=df)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               earnings   R-squared:                       0.180
Model:                            OLS   Adj. R-squared:                  0.180
Method:                 Least Squares   F-statistic:                 1.346e+04
Date:                Mon, 25 Nov 2024   Prob (F-statistic):               0.00
Time:                        21:15:40   Log-Likelihood:            -2.2317e+05
No. Observations:               61395   AIC:                         4.464e+05
Df Residuals:                   61393   BIC:                         4.464e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -5.3763      0.209    -25.778      0.000      -5.785      -4.967
education      1.7451      0.015    116.011      0.000       1.716       1.775
==============================================================================
Omnibus:                     9721.871   Durbin-Watson:                   1.828
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            17015.805
Skew:                           1.033   Prob(JB):                         0.00
Kurtosis:                       4.543   Cond. No.                         78.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [36]:
```python
robust_se = model.fit(cov_type = 'HC1')
print(robust_se.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               earnings   R-squared:                       0.180
Model:                            OLS   Adj. R-squared:                  0.180
Method:                 Least Squares   F-statistic:                 1.159e+04
Date:                Mon, 25 Nov 2024   Prob (F-statistic):               0.00
Time:                        21:15:40   Log-Likelihood:            -2.2317e+05
No. Observations:               61395   AIC:                         4.464e+05
Df Residuals:                   61393   BIC:                         4.464e+05
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -5.3763      0.212    -25.307      0.000      -5.793      -4.960
education      1.7451      0.016    107.669      0.000       1.713       1.777
==============================================================================
Omnibus:                     9721.871   Durbin-Watson:                   1.828
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            17015.805
Skew:                           1.033   Prob(JB):                         0.00
Kurtosis:                       4.543   Cond. No.                         78.5
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

In [37]:
```python
from statsmodels.iolib.summary2 import summary_col

models=['Homoskedastic Model', 'Heteroskedastic Model']
results_table=summary_col(results=[results, robust_results],
                          float_format='%0.3f',
                          stars=True,
                          model_names=models)
results_table
```

Out[37]:

	Homoskedastic Model	Heteroskedastic Model
Intercept	-5.376***	-5.376***
	(0.209)	(0.212)
education	1.745***	1.745***
	(0.015)	(0.016)
R-squared	0.180	0.180
R-squared Adj.	0.180	0.180

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

b.

In [38]:
```python
education_coef = robust_se.params['education']
education_se = robust_se.bse['education']
```

In [47]:
```python
t_stat = education_coef / education_se
t_stat
```

Out[47]: `107.66885682517291`

In [46]:
```python
p_value = 2 * (1 - stats.norm.cdf(abs(t_stat)))
p_value
```

Out[46]: `0.0`

Estimated education effect significantly different than zero because p-value < 0.05

c.

In [39]:
```python
z_critical = stats.norm.ppf(1 - 0.05)
z_critical
```

Out[39]: `1.6448536269514722`

In [40]:
```python
error = z_critical * education_se
lower_bound = education_coef - error
upper_bound = education_coef + error
print(f'90% Confidence Interval for the coefficient of education: ({lower_bound}, {upper_bound})')
```

```
90% Confidence Interval for the coefficient of education: (1.7184885147645164, 1.7718096869730429)
```