main

April 5, 2025

1 ECON3003 Econometrics II Project

1.1 Huang Minxing SC122481

2 Question 1

Consider the wage equation:

$\text{logsal} = \beta_1 + \beta_2 \text{logsalbegin} + \beta_3 \text{educ} + \beta_4 \text{gender} + \beta_5 \text{minority} + \epsilon$

Estimate the wage equation (1) by OLS for the sample of job categories 1 and 3 employees and interpret the estimated coefficients. This should include both the economic meaning of each of the slope coefficients and their individual significance.

[11]:
```
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
from scipy import stats
%matplotlib inline
```

[7]:
```
df = pd.read_csv('ECON3003_Project_Data.csv')
df_filtered = df[df["jobcat"].isin([1, 3])]

model = smf.ols("logsal ~ logsalbegin + educ + gender + minority",␣
 ↪data=df_filtered).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 logsal   R-squared:                       0.813
Model:                            OLS   Adj. R-squared:                  0.812
Method:                 Least Squares   F-statistic:                     481.4
Date:                Sat, 05 Apr 2025   Prob (F-statistic):          1.50e-159
Time:                        13:48:28   Log-Likelihood:                 141.18
No. Observations:                 447   AIC:                            -272.4
Df Residuals:                     442   BIC:                            -251.8
Df Model:                           4
Covariance Type:            nonrobust
```

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.1336      0.323      6.600      0.000       1.498       2.769
logsalbegin    0.8087      0.037     21.673      0.000       0.735       0.882
educ           0.0291      0.004      6.688      0.000       0.021       0.038
gender         0.0285      0.021      1.365      0.173      -0.013       0.070
minority      -0.0540      0.022     -2.509      0.012      -0.096      -0.012
==============================================================================
Omnibus:                       40.918   Durbin-Watson:                   1.755
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               63.064
Skew:                           0.625   Prob(JB):                     2.02e-14
Kurtosis:                       4.351   Cond. No.                         659.
==============================================================================

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

- The results show that the initial salary has a significant positive effect on current salary, with a coefficient of 0.8087. Economically speaking, this implies that a 1% increase in beginning salary is associated with an average 0.81% increase in current salary.
- The **Education** also has a positive and significant effect on wages. An additional 1 year of education would increase salary by 2.9%.
- The coefficient on **Gender** is positive, but not significant, since $p = 0.173$. This shows that we cannot conclude a gender wage gap exists in this subsample after controlling for other factors.
- The coefficient on **Minority** is negative and significant at 5% level ($p = 0.012$). Minority employees earn about 5.4% less, on average, than non-minority employees, which may indicate discriminatory wage penalties against minority groups.

3 Question 2

Test the null hypothesis $H_0 : \beta_3 = \beta_4$ for the two job categories against the alternative $H_1 : \beta_3 \neq \beta_4$, using the F test, the likelihood ratio (LR) test, and the Lagrange Multiplier (LM) test.

[8]:
```python
# Unrestricted model
df_filtered = df[df["jobcat"].isin([1, 3])].copy()
model_unrestricted = smf.ols("logsal ~ logsalbegin + educ + gender + minority",␣
↪data=df_filtered).fit()

df_filtered["educ_plus_gender"] = df_filtered["educ"]+df_filtered["gender"]

# Restricted model
model_restricted = smf.ols("logsal ~ logsalbegin + educ_plus_gender +␣
↪minority", data=df_filtered).fit()
```

[9]:
```python
# F Test
f_test = model_unrestricted.compare_f_test(model_restricted)

print("== F-Test ==")
print(f"F-statistic: {f_test[0]:.4f}, p-value: {f_test[1]:.4f}")
```

```
== F-Test ==
F-statistic: 0.0008, p-value: 0.9778
```

[12]:
```python
# LR Test
lr_stat = 2 * (model_unrestricted.llf - model_restricted.llf)

# We use chi2 to compute the p-value, and df = 1 is because we are testing one␣
 ↪restriction
lr_pval = stats.chi2.sf(lr_stat, df=1)

print("\n== Likelihood Ratio (LR) Test ==")
print(f"LR statistic: {lr_stat:.4f}, p-value: {lr_pval:.4f}")
```

```

== Likelihood Ratio (LR) Test ==
LR statistic: 0.0008, p-value: 0.9776
```

[14]:
```python
# LM Test
# We use tools from the statsmodels library to perform the LM test
from statsmodels.stats.diagnostic import linear_lm

# Build the restricted model
resid = model_restricted.resid
X_restricted = model_restricted.model.exog
lm_test_stat = df_filtered.shape[0] * model_unrestricted.rsquared - df_filtered.
 ↪shape[0] * model_restricted.rsquared
lm_pval = stats.chi2.sf(lm_test_stat, df=1)

print("\n== Lagrange Multiplier (LM) Test ==")
print(f"LM statistic: {lm_test_stat:.4f}, p-value: {lm_pval:.4f}")
```

```

== Lagrange Multiplier (LM) Test ==
LM statistic: 0.0001, p-value: 0.9903
```

[15]:
```python
# In total, we have three tests: F-test, LR test, and LM test.
print("== F-Test ==")
print(f"F-statistic: {f_test[0]:.4f}, p-value: {f_test[1]:.4f}")

print("\n== Likelihood Ratio (LR) Test ==")
print(f"LR statistic: {lr_stat:.4f}, p-value: {lr_pval:.4f}")
```

```
print("\n== Lagrange Multiplier (LM) Test ==")
print(f"LM statistic: {lm_test_stat:.4f}, p-value: {lm_pval:.4f}")
```

```
== F-Test ==
F-statistic: 0.0008, p-value: 0.9778

== Likelihood Ratio (LR) Test ==
LR statistic: 0.0008, p-value: 0.9776

== Lagrange Multiplier (LM) Test ==
LM statistic: 0.0001, p-value: 0.9903
```

To test the null hypothesis $H_0 : \beta_3 = \beta_4$, which states that the effects of education and gender on log wages are equal, we employed the F test, Likelihood Ratio (LR) test, and Lagrange Multiplier (LM) test. Across all three tests, the p-values are very large (above 0.97), far exceeding the conventional significance level of 0.05. This means that we could not reject the null hypothesis. This suggests that there is **no statistically significant difference between the coefficients on education and gender** in explaining log wages in this sample. That is to say, the data does not provide evidence that these two variables have different marginal effects on wages.

4 Question 3

Perform a diagnostic test of heteroskedasticity for equation (1) across the two job categories using the Breusch-Pagan test. Report and comment on the test results.

[17]:
```
from statsmodels.stats.diagnostic import het_breuschpagan

model = smf.ols("logsal ~ logsalbegin + educ + gender + minority",␣
 ↪data=df_filtered).fit()

residuals = model.resid # We get the residuals from the fitted model
exog = model.model.exog  # We get the exogenous variables from the fitted model

# Breusch-Pagan test for heteroskedasticity
bp_test = het_breuschpagan(residuals, exog)

# The test returns four values:
bp_stat = bp_test[0]        # LM statistic
bp_pval = bp_test[1]        # p-value
f_stat = bp_test[2]         # F statistic
f_pval = bp_test[3]         # F p-value

# Print the results
print("== Breusch-Pagan Test for Heteroskedasticity ==")
print(f"LM Statistic: {bp_stat:.4f}, p-value: {bp_pval:.4f}")
print(f"F Statistic: {f_stat:.4f}, p-value: {f_pval:.4f}")
```

```
== Breusch-Pagan Test for Heteroskedasticity ==
LM Statistic: 13.5191, p-value: 0.0090
F Statistic: 3.4462, p-value: 0.0087
```

To assess whether the residuals from the baseline regression model exhibit heteroskedasticity, we conducted the **Breusch-Pagan test**.

The test returned the following results:

- **LM Statistic** = 13.5191, **p-value** = 0.0090
- **F Statistic** = 3.4462, **p-value** = 0.0087

Since both p-values are less than 0.05, we **reject the null hypothesis of homoskedasticity** at the 5% significance level. This suggests that the variance of the error term is **not constant** across observations — in other words, **heteroskedasticity is present** in the model.

As a result, standard OLS standard errors may be unreliable. It is suggested that we should report the **robust standard errors** (i.e., Whit's standard errors) to the presence of heteroskedasticity (Just what we need to do in **Question 4**)

5 Question 4

If the Breusch-Pagan test in 3 gave evidence of heteroskedasticity, then re-esitmate equation (1) using standard errors that are robust to the presence of heteroskedasticity (i.e., White's standard errors), and comment on the results. If the Breusch-Pagan test in 3 gave no or little evidence of heteroskedasticity, then skip this step.

[18]:
```python
# We set up the model again to show the heteroskedasticity-robust standard␣
↪errors
model = smf.ols("logsal ~ logsalbegin + educ + gender + minority",␣
↪data=df_filtered).fit()

# Use White's correction for heteroskedasticity
model_robust = model.get_robustcov_results(cov_type='HC1')

# Show the summary with robust standard errors
print(model_robust.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 logsal   R-squared:                       0.813
Model:                            OLS   Adj. R-squared:                  0.812
Method:                 Least Squares   F-statistic:                     440.8
Date:                Sat, 05 Apr 2025   Prob (F-statistic):          9.55e-153
Time:                        14:40:00   Log-Likelihood:                 141.18
No. Observations:                 447   AIC:                            -272.4
Df Residuals:                     442   BIC:                            -251.8
Df Model:                           4
Covariance Type:                  HC1
```

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.1336      0.330      6.461      0.000       1.485       2.783
logsalbegin    0.8087      0.038     21.515      0.000       0.735       0.883
educ           0.0291      0.004      7.212      0.000       0.021       0.037
gender         0.0285      0.021      1.360      0.175      -0.013       0.070
minority      -0.0540      0.019     -2.882      0.004      -0.091      -0.017
==============================================================================
Omnibus:                       40.918   Durbin-Watson:                   1.755
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               63.064
Skew:                           0.625   Prob(JB):                     2.02e-14
Kurtosis:                       4.351   Cond. No.                         659.
==============================================================================

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

Since the Breusch-Pagan test indicated the presence of heteroskedasticity in the original model, we re-estimated equation (1) using **White's heteroskedasticity-robust standard errors (HC1)**.

The coefficient estimates remain unchanged, as expected, because OLS is still unbiased. However, the **standard errors and p-values have changed**, which affects the statistical inference.

- The variables **logsalbegin**, **educ**, and **minority** remain statistically significant at the 1% level.
- The variable **gender**, however, is **not statistically significant** ($p = 0.175$), suggesting that gender does not have a robust effect on log wages once we account for heteroskedasticity.
- The R-squared of the model remains high at **0.813**, indicating a good overall fit.

Using robust standard errors ensures that our inference is valid despite the presence of heteroskedasticity, and highlights the importance of checking model assumptions when making conclusions.

6 Question 5

Use the dummy variable approach to test whether the wage equation is the same for job category 1 and job category 3.

[ ]:
```python
df_job13 = df[df['jobcat'].isin([1, 3])].copy()
df_job13['jobcat3'] = (df_job13['jobcat'] == 3).astype(int)
model_base = smf.ols("logsal ~ logsalbegin + educ + gender + minority",␣
 ↪data=df_job13).fit()

model_interact = smf.ols(
    "logsal ~ logsalbegin * jobcat3 + educ * jobcat3 + gender * jobcat3 +␣
 ↪minority * jobcat3",
    data=df_job13
).fit()
```

```python
f_test_result = model_interact.compare_f_test(model_base)

# Print the results of the Chow test
print("== Chow Test via Dummy Variable Interaction ==")
print(f"F-statistic: {f_test_result[0]:.4f}")
print(f"p-value: {f_test_result[1]:.4f}")
print(f"Degrees of freedom: df_diff = {f_test_result[2]}")
```

```
== Chow Test via Dummy Variable Interaction ==
F-statistic: 10.0637
p-value: 0.0000
Degrees of freedom: df_diff = 5.0
```

The method we use is basically the **Chow Test**.

We constructed a pooled model that includes all individuals from job categories 1 and 3, and interacted a dummy variable for job category 3 with all explanatory variables in the wage equation.

The F-test comparing the restricted model (no interactions) with the unrestricted model (with interactions) yielded the following result:

- **F-statistic** = 10.0637
- **p-value** = 0.0000
- **Degrees of freedom (df_diff)** = 5

Since the p-value is effectively zero, we **reject the null hypothesis** that the wage equations are the same across the two job categories. This provides strong statistical evidence that the wage determination process differs **significantly** between job category 1 and job category 3.

Therefore, it is appropriate to model these two groups separately or allow for job-specific coefficients in the wage equation.