Krish_Patel_Project

November 21, 2024

[40]:
```
import pandas as pd
import statsmodels.formula.api as smf
```

[41]:
```
df = pd.read_csv("CollegeDistance.csv")
df.head()
```

[41]:

	female	black	hispanic	bytest	dadcoll	momcoll	ownhome	urban	cue80	\
0	0	0	0	39.15	1	0	1	1	6.2	
1	1	0	0	48.87	0	0	1	1	6.2	
2	0	0	0	48.74	0	0	1	1	6.2	
3	0	1	0	40.40	0	0	1	1	6.2	
4	1	0	0	40.48	0	0	0	1	5.6	

	stwmfg80	dist	tuition	ed	incomehi
0	8.09	0.2	0.88915	12	1
1	8.09	0.2	0.88915	12	0
2	8.09	0.2	0.88915	12	0
3	8.09	0.2	0.88915	12	0
4	8.09	0.4	0.88915	13	0

Q1: Run a regression of years of completed education (ED) on distance to the nearest college (Dist), where Dist is measured in tens of miles.

[42]:
```
reg = smf.ols('ed ~ dist', data=df)
model_q1 = reg.fit()
print(model_q1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     ed   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     28.48
Date:                Thu, 21 Nov 2024   Prob (F-statistic):           1.00e-07
Time:                        15:49:02   Log-Likelihood:                -7632.2
No. Observations:                3796   AIC:                         1.527e+04
Df Residuals:                    3794   BIC:                         1.528e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
```

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     13.9559      0.038    369.945      0.000      13.882      14.030
dist          -0.0734      0.014     -5.336      0.000      -0.100      -0.046
==============================================================================
Omnibus:                     7187.794   Durbin-Watson:                   1.769
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              361.676
Skew:                           0.410   Prob(JB):                     2.90e-79
Kurtosis:                       1.729   Cond. No.                         3.73
==============================================================================

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

Q1-A: What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How does the average value of years of completed schooling change when colleges are built close to where students go to high school? The regression equation estimated is ed = 13.9559 - 0.0734 * dist. As the distance from 4year College in 10's of miles increases, the year of education completed decrease. So, When colleges are built closer to high schools, reducing the distance by 10 miles would increase the average years of completed schooling by approximately 0.0734 years.

[43]:
```
intercept = model_q1.params['Intercept']
slope = model_q1.params['dist']
print(f"Estimated intercept: {intercept}")
print(f"Estimated slope: {slope}")
```

```
Estimated intercept: 13.955856114789412
Estimated slope: -0.07337270712920115
```

Q1-B: Bob's high school was 20 miles from the nearest college. Predict Bob's years of completed education using the estimated regression. How would the prediction change if Bob lived 10 miles from the nearest college?

[44]:
```
edu_20 = model_q1.predict(exog={'dist':2})
edu_10 = model_q1.predict(exog={'dist':1})

print(f"Predicted years of education for 20 miles: {edu_20.iloc[0]}")
print(f"Predicted years of education for 10 miles: {edu_10.iloc[0]}")
```

```
Predicted years of education for 20 miles: 13.80911070053101
Predicted years of education for 10 miles: 13.882483407660212
```

Q1-C: Does distance to college explain a large fraction of the variance in educational attainment across individuals? Explain No, the distance to college doesn't explain a large fraction of the variance in educational attainment across individuals. Since the R^2 is 0.007, this value indicates that the varaible 'dist'(distance to college in 10's of miles) explains only 0.7% of the variance in years of completed education across individuals. This is a very small proportion,

suggesting that distance alone is not a significant factor in explaining variations in educational attainment.

Q1-D: Is the estimated regression slope coefficient statistically significant? That is, can you reject the null hypothesis H0 : 1 = 0 versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?

[59]:
```
p_value = model_q1.pvalues['dist']
print(f"p_value: {p_value:.10f}")
```

```
p_value: 0.0000001004
```

The estimated slope coefficient for the regression of years of completed education (ed) on distance to the nearest college (dist) is −0.0734. This means that for every 10-mile increase in distance to the nearest college, the average number of years of completed education decreases by approximately 0.073 years. The p-value is extremely small, much lower than all significance levels (0.10, 0.05, and 0.01). Therefore, we reject the null hypothesis at the 10%, 5%, and 1% significance levels. This indicates that the slope coefficient is statistically significant and that the distance to the nearest college has a meaningful impact on educational attainment. In conclusion, the data provides strong evidence that being farther from a college decreases the number of years of completed education.

Q1-E:Construct a 95% confidence interval for the slope coefficient. 95% Confidence interval for the slope coefficient is: [-0.100, -0.046]

Q1-F: An education advocacy group argues that, on average, a person's educational attainment would increase by approximately 0.15 year if distance to the nearest college is decreased by 20 miles. Is the advocacy groups' claim consistent with the estimated regression? Explain. $\Delta \text{ed} = -0.0734 * -2 = 0.1468$ years

The advocacy group's claim is consistent with the estimated regression model, as the predicted value (0.1468 years) is almost identical to their value (0.15 years).

Q2: Run a regression of ED on Dist, but include some additional regressors to control for characteristics of the student, the student's family, and the local labor market. In particular, include as additional regressors Bytest, Female, Black, Hispanic, Incomehi, Ownhome, DadColl, Cue80, and Stwmfg80. What is the estimated effect of Dist on ED?

[45]:
```
reg = smf.ols('ed ~ dist + bytest + female + black + hispanic + incomehi +␣
  ↪ownhome + dadcoll + cue80 + stwmfg80', data=df)
model_q2 = reg.fit()
print(model_q2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     ed   R-squared:                       0.279
Model:                            OLS   Adj. R-squared:                  0.277
Method:                 Least Squares   F-statistic:                     146.3
Date:                Thu, 21 Nov 2024   Prob (F-statistic):          6.94e-260
Time:                        15:49:02   Log-Likelihood:                -7025.9
```

```
No. Observations:                3796   AIC:                          1.407e+04
Df Residuals:                    3785   BIC:                          1.414e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.8275      0.250     35.271      0.000       8.337       9.318
dist          -0.0315      0.012     -2.550      0.011      -0.056      -0.007
bytest         0.0938      0.003     29.669      0.000       0.088       0.100
female         0.1454      0.051      2.874      0.004       0.046       0.245
black          0.3680      0.071      5.156      0.000       0.228       0.508
hispanic       0.3985      0.074      5.352      0.000       0.253       0.545
incomehi       0.3952      0.061      6.529      0.000       0.277       0.514
ownhome        0.1521      0.067      2.277      0.023       0.021       0.283
dadcoll        0.6961      0.069     10.129      0.000       0.561       0.831
cue80          0.0232      0.010      2.409      0.016       0.004       0.042
stwmfg80      -0.0518      0.020     -2.608      0.009      -0.091      -0.013
==============================================================================
Omnibus:                      118.266   Durbin-Watson:                   1.924
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               97.867
Skew:                           0.320   Prob(JB):                     5.60e-22
Kurtosis:                       2.543   Cond. No.                         539.
==============================================================================

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

Q2-A: Is the estimated effect of Dist on ED in the regression in Q2 substantively different from the regression in Q1? Based on this, does the regression in Q1 seem to suffer from important omitted variable bias? The coefficient for Dist is -0.0315 in Q2, meaning that for every additional 10 miles of distance to the nearest college, the years of education completed decreases by 0.0315 years. Since the ceof in Q1 was comparatively larger than the ceof in Q2, the regression in Q1 suffers from omitted variable bias, as it ignores important factors that influence education attainment.

Q2-B: The value of the coefficient on DadColl is positive. What does this coefficient measure? The coefficient on DadColl is 0.6961. This implies that on average, students whose fathers attended college complete 0.6961 more years of education than those whose fathers did not, holding all other factors constant.

Q2-C: Explain why Cue80 and Swmfg80 appear in the regression. Are the signs of their estimated coefficients (+ or -) what you would have believed? Interpret the magnitudes of these coefficients. Cue80 (county unemployment rate in 1980) represents local labor market conditions that could influence educational attainment. Higher unemployment rates may encourage students to pursue further education due to reduced job opportunities. A 1%

increase in the county unemployment rate is associated with an average increase of 0.0232 years in completed education, holding other factors constant. This aligns with the theory that higher unemployment encourages education.

Stwmfg80 (state manufacturing hourly wage in 1980) captures economic conditions that affect the opportunity cost of education. A higher manufacturing wage may incentivize individuals to enter the workforce instead of pursuing additional education. A $1 increase in the state manufacturing hourly wage is associated with an average decrease of 0.0518 years in completed education. This suggests that higher wages may reduce the motives to stay in school.

Q2-D: Bob is a black male. His high school was 20 miles from the nearest college. His baseyear composite test score (Bytest) was 58. His family income in 1980 was 26,000, and his family owned a home. His mother attended college, but his father did not. The unemployment rate in his county was 7.5%, and the state average manufacturing hourly wage was 9.75. Predict Bob's years of completed schooling using the regression in Q2.

[46]:
```
bobs_values = model_q2.predict(exog={'black':0,
                                     'hispanic': 0,
                                     'dist': 2,
                                     'female': 0,
                                     'bytest': 58,
                                     'incomehi': 1,
                                     'ownhome': 1,
                                     'momcoll': 1,
                                     'dadcoll': 0,
                                     'cue80': 7.5,
                                     'stwmfg80': 9.75})

print(f"Predicted years of education for bob is {bobs_values.iloc[0]}")
```

```
Predicted years of education for bob is 14.422543543303458
```

Q2-E: Jim has the same characteristics as Bob except that his high school was 40 miles from the nearest college. Predict Jim's years of completed schooling using the regression in Q2.

[48]:
```
jims_values = model_q2.predict(exog={'black':0,
                                     'hispanic': 0,
                                     'dist': 4,
                                     'female': 0,
                                     'bytest': 58,
                                     'incomehi': 1,
                                     'ownhome': 1,
                                     'momcoll': 1,
                                     'dadcoll': 0,
                                     'cue80': 7.5,
                                     'stwmfg80': 9.75})

print(f"Predicted years of education for Jim is {jims_values.iloc[0]}")
```

```
Predicted years of education for Jim is 14.359466075645905
```

Q2-F: It has been argued that, controlling for other factors, blacks and Hispanics complete more college than whites. Is this result consistent with the regressions in Q2? Yes, the results indicate that, controlling for other factors, blacks and Hispanics complete more years of education on average than whites. These findings support the argument presented. Since the coefficient for "black" is 0.368, this means that holing other factors constant, being black is associated with an additional 0.368 year of completed education compared to being white. Similarly, the coefficient for "hispanic" is 0.399, this means that holding other factors constant, being hispanic is associated with an additional 0.399 years of completed education compared to the white.