[STATS 413] HW4

Huy Le

February 2026

1 Problem 1: Education and Income

Throughout, `Income` is measured in *thousands of dollars*.

(a)

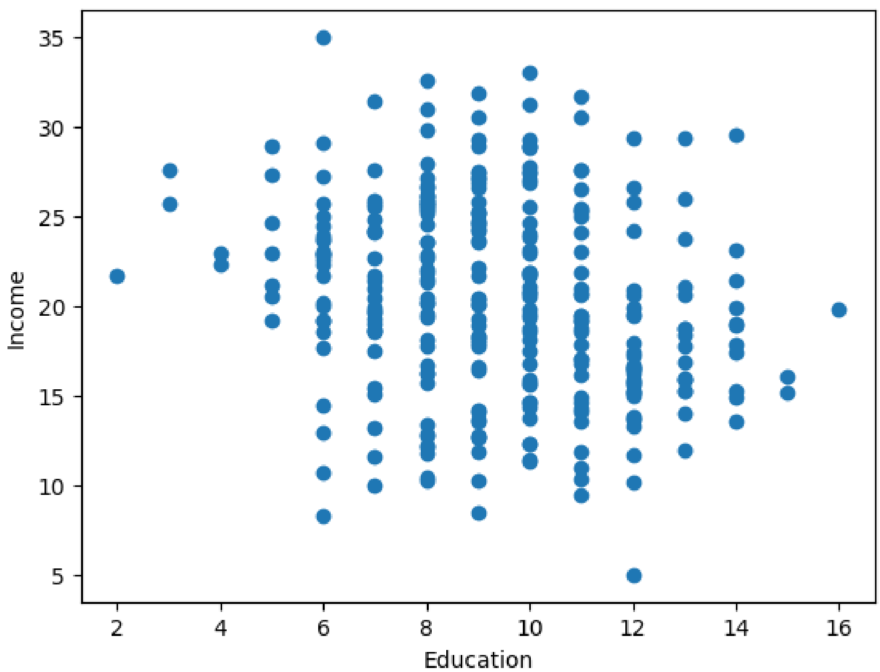

Figure 1: Scatterplot of Income (in thousands of dollars) vs. Education (years).

The overall association in the pooled data appears to be *negative* (higher Education tends to correspond to slightly lower Income).

(b)

From the OLS output, the fitted regression line is

$$\widehat{\text{Income}} = 25.2100 - 0.5172 * \text{Education}.$$

(c)

Test $H_0 : \beta_1 \geq -0.1$ vs. $H_1 : \beta_1 < -0.1$. Using $\hat{\beta}_1 = -0.5172$ and $SE(\hat{\beta}_1) = 0.128$, the test statistic is

$$t = \frac{\hat{\beta}_1 - (-0.1)}{SE(\hat{\beta}_1)} \approx \frac{-0.5172 + 0.1}{0.128} \approx -3.26.$$

The corresponding one-sided p-value is essentially 0, so we reject H_0. Therefore, we say that, at 95% confidence, the slope is less than -0.1.

(d)

```python
# (d) Analysis II: Multiple regression with Education, Millenial, and interaction
df['Millenial'] = df['Millenial'].astype(int)
df['Education_x_Millenial'] = df['Education'] * df['Millenial']
df[["Millenial", "Education_x_Millenial"]]

model2 = sm.OLS(df['Income'], sm.add_constant(df[['Education', 'Millenial', 'Education_x_Millenial']]))
results2 = model2.fit()
print(results2.summary())
```

[39] ✓ 0.0s

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Income   R-squared:                       0.565
Model:                            OLS   Adj. R-squared:                  0.561
Method:                 Least Squares   F-statistic:                     128.3
Date:                Fri, 06 Feb 2026   Prob (F-statistic):           2.86e-53
Time:                        22:18:04   Log-Likelihood:                -812.53
No. Observations:                 300   AIC:                             1633.
Df Residuals:                     296   BIC:                             1648.
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    21.1297      1.238     17.072      0.000      18.694      23.565
Education                 0.3990      0.146      2.740      0.007       0.112       0.686
Millenial                -8.1530      1.899     -4.294      0.000     -11.890      -4.416
Education_x_Millenial    -0.0859      0.198     -0.435      0.664      -0.475       0.303
==============================================================================
Omnibus:                        0.639   Durbin-Watson:                   1.870
Prob(Omnibus):                  0.726   Jarque-Bera (JB):                0.398
Skew:                           0.029   Prob(JB):                        0.820
Kurtosis:                       3.169   Cond. No.                         118.
==============================================================================
```

Figure 2: Output/plot for part (d).

(e)

```python
# (e) Prediction equations
print("Regression equation on millenials: ")
print(f"Income = {results2.params['const']:.3f} + {results2.params['Education']:.3f}*Education + {results2.params['Millenial']:.3f}*1 + {results2.params['Edu
print(f"       = {results2.params['const'] + results2.params['Millenial']:.3f} + ({results2.params['Education'] + results2.params['Education_x_Millenial']:.3
print("\nRegression equation on non-millenials: ")
print(f"Income = {results2.params['const']:.4f} + {results2.params['Education']:.4f}*Education")
```

✓ 0.0s   Python

```
Regression equation on millenials:
Income = 21.130 + 0.399*Education + -8.153*1 + -0.086*Education*1
       = 12.977 + (0.313)*Education

Regression equation on non-millenials:
Income = 21.1297 + 0.3990*Education
```

Figure 3: Output/plot for part (e).

(Millenial $= 0$):

$$\widehat{\text{Income}} = 21.1297 + 0.3990\,\text{Education}.$$

(Millenial $= 1$):

$$\widehat{\text{Income}} = (21.1297-8.1530)+(0.3990-0.0859)\,\text{Education} = 12.9767+0.3131\,\text{Education}.$$

(f)

```python
# (f) Test difference in expected income between non-Millenials with 8 vs 7 years > $100
# difference in income = (const + Education*8) - (const + Education*7) = Education
diff = results2.params['Education']
se_diff = results2.bse['Education']
t_stat = (diff - 0.1) / se_diff
p_value = 1 - t.cdf(t_stat, results2.df_resid)
print(f"p-value: {p_value:.4f}")
print(f"Since p-value = {p_value:.4f} < 0.05, we reject the null hypothesis that the difference in expected income")
print("between non-Millenials with 8 years vs 7 years of education is less than $100.")
print("Therfore, there is evidence (at α = 0.05) that the difference in expected income")
print("between non-Millenials with 8 years vs 7 years of education is greater than $100.")
```

✓ 0.0s

```
p-value: 0.0205
Since p-value = 0.0205 < 0.05, we reject the null hypothesis that the difference in expected income
between non-Millenials with 8 years vs 7 years of education is less than $100.
Therfore, there is evidence (at α = 0.05) that the difference in expected income
between non-Millenials with 8 years vs 7 years of education is greater than $100.
```

Figure 4: Output/plot for part (f).

(g)

The slope difference is the interaction coefficient β_3. The 90% CI is $[-0.4120,\ 0.2401]$.

(h)

The overall F-test for Analysis II has F-statistic about 128.35 with p-value essentially 0. Therefore, at $\alpha = 0.05$, Analysis II provides a significant improvement over a model with no predictors.

(i)

The apparent contradiction is because Analysis I ignores the generation variable. If we run separate regressions for non-Millenials and Millenials (as in Analysis II), both fitted lines have positive slopes, meaning that within each generation more education is associated with higher income.
However, when the two groups are mixed together and we fit a single line (Analysis I), the fact that Millenials have lower income overall can pull the pooled regression line downward, producing an overall negative slope even though the within-group slopes are positive.

(j)

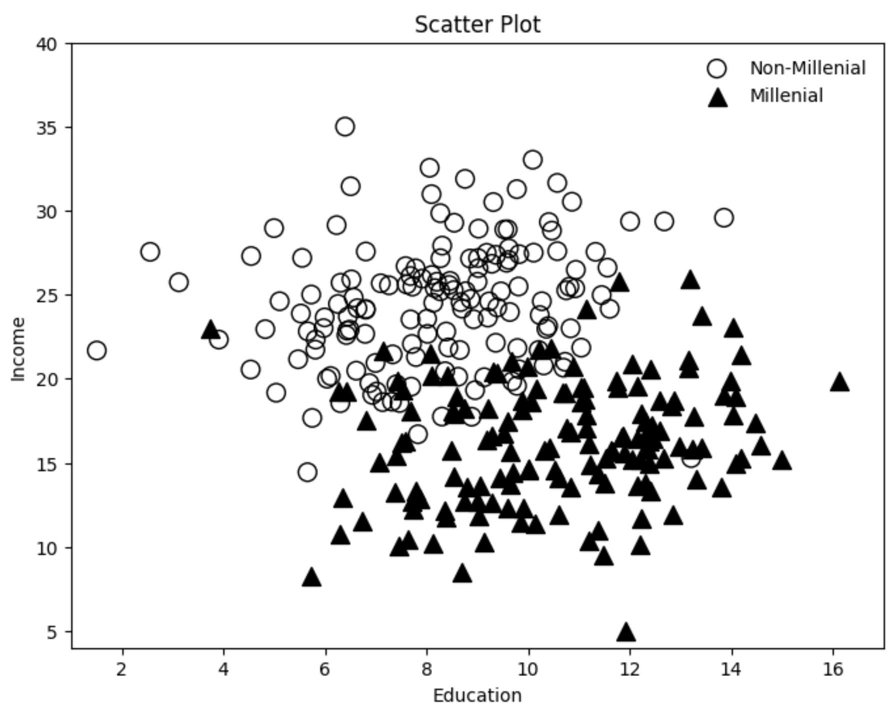

Figure 5: Jittered scatterplot of Income vs. Education, with Millenials and non-Millenials shown using different plotting symbols.

The plot shows that if we separate the data into two groups (non-Millenials and Millenials) and consider the within-group trends, the relationship between Education and Income is positive for each group. However, because Analysis I does not account for the generation factor, mixing the two groups together and fitting one regression line can yield an overall negative slope that is counterintuitive.

2 Problem 2: Deviation-from-the-mean regression

(a)

The normal equations are

$$X^\top(y - X\hat{\beta}) = 0 \quad \Longleftrightarrow \quad \begin{cases} 1_n^\top(y - 1_n\hat{\beta}_1 - X_2\hat{\beta}_2) = 0, \\ X_2^\top(y - 1_n\hat{\beta}_1 - X_2\hat{\beta}_2) = 0_{p-1}. \end{cases}$$

For the first equation, divide by n and use $\bar{y} = \frac{1}{n}1_n^\top y$ and $\bar{x}_2 = \frac{1}{n}X_2^\top 1_n$ to get

$$\bar{y} - \hat{\beta}_1 - \bar{x}_2^\top\hat{\beta}_2 = 0.$$

For the second equation, divide by n to obtain

$$\frac{1}{n}X_2^\top y - \bar{x}_2\hat{\beta}_1 - \frac{1}{n}X_2^\top X_2\hat{\beta}_2 = 0_{p-1},$$

as desired.

(b)

Let $H_1 = \frac{1}{n}1_n1_n^\top$ and we define the centered variables

$$\tilde{y} = (I_n - H_1)y, \qquad \tilde{X}_2 = (I_n - H_1)X_2.$$

From part (a), the first normal equation gives $\hat{\beta}_1 = \bar{y} - \bar{x}_2^\top\hat{\beta}_2$ Substitute this into the second normal equation, we have that

$$X_2^\top\left(y - 1_n\bar{y}\right) = X_2^\top\left(X_2 - 1_n\bar{x}_2^\top\right)\hat{\beta}_2$$

Noting that $y - 1_n\bar{y} = (I_n - H_1)y = \tilde{y}$ and $X_2 - 1_n\bar{x}_2^\top = (I_n - H_1)X_2 = \tilde{X}_2$, we can rewrite the equation as

$$\tilde{X}_2^\top\tilde{y} = \tilde{X}_2^\top\tilde{X}_2\,\hat{\beta}_2$$

Assuming $\tilde{X}_2^\top\tilde{X}_2$ is invertible, then

$$\hat{\beta}_2 = (\tilde{X}_2^\top\tilde{X}_2)^{-1}\tilde{X}_2^\top\tilde{y}$$

Thus, including an intercept is equivalent to centering the features and outcomes.