In-Class Exercise 8

For this in-class exercise, work with your group of 2-3 people, to answer the following questions. These questions are not necessarily easy and sometimes they will not have a clear "correct" answer. The goal is to get you thinking about the material we've learned. Some of these questions may require you to discuss and debate with your group members to come up with an answer or can cover topics that we have not yet covered in class.

Be prepared to share your answers with the class and add to the discussion.

After class submit your a do-file with your answers in comments to Moodle for grading. You will be graded as a group on your submission. Only one group member needs to submit the assignment, but make sure add all group member names.

For this exercise, use the `wage1` dataset by running:

```
sysuse wage1, clear
```

These are data from the 1976 Current Population Survey.

Despite convergence in education and experience, the gender wage gap persists. A major economic theory suggests this isn't just about gender, but about the interaction between gender and parenthood. Women often face a "penalty" for having children (lower wages relative to childless women), while men often receive a "premium" (higher wages relative to childless men).

Problem 1

Estimate a regression of $ln(wage)$ on $educ$, $exper$ and $exper^2$. Generate variables as necessary.

1. What are the coefficients and their interpretations?
2. Write out the expression for the marginal effect of an additional year of experience on log wages. Calculate the marginal effect for a worker with 5 years of experience and one with 15 years of experience. What do you observe?
3. At what year do you get the maximum return to experience? Show your work.

! Solution

```
gen exper2 = exper^2
reg lwage educ exper exper2
```

1. **Coefficients:**
   - `educ`: 0.090. An additional year of education is associated with approximately a 9% increase in wages.
   - `exper`: 0.041. The first year of experience increases wages by approx 4.1%.
   - `exper2`: −0.0007. The negative sign indicates diminishing returns to experience (concave shape).
2. **Marginal Effect:**

$$\frac{\partial ln(wage)}{\partial exper} = \hat{\beta}_{exper} + 2\hat{\beta}_{exper^2} \cdot exper$$

   - At experience = 5: $0.0410 - 2(0.00071)(5) \approx 0.034$ (3.4% return).
   - At experience = 15: $0.0410 - 2(0.00071)(15) \approx 0.020$ (2.0% return).
   - Observation: The return to experience falls as experience increases.
3. **Turning Point (Max Return):** The maximum wage occurs where the marginal effect is zero:

$$0 = 0.0410 - 0.0014 \cdot exper \Longrightarrow exper^{*} = \frac{0.0410}{0.0014} \approx 29.3 \quad \text{years}$$

   (Calculation using full precision: $0.041005/(2 \times 0.000713) \approx 28.7$ years).

Problem 2

Add a dummy variable for `female` to the regression in Problem 1.

1. What is the coefficient on `female`? Interpret it.
2. Does this coefficient prove discrimination? As a group, brainstorm one important variable that is not in this model (an "observable") that might explain this gap if we could control for it.

! Solution

```
reg lwage educ exper exper2 female
```

1. **Coefficient:** `female`: −0.337.
   - Interpretation: Holding education and experience constant, women earn approximately 33.7% less than men. (Exact calculation: $e^{-0.337} - 1 \approx -28.6\%$).
2. **Discrimination?**
   - No, this does not prove discrimination because we might have omitted variable bias (`OVB`).
   - **Missing variables:** Occupation, industry, hours worked (if wage is weekly), ability, negotiation behavior, interruptions in career history, etc.

Problem 3

Now, we want to see if the labor market treats fathers and mothers differently. Create an interaction term between gender and parenthood. Run the regression:

$$lnwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \beta_5 has\_child + \beta_6 (female \times has\_child) + \varepsilon$$

To generate `has_child`, create a binary variable using the `numdep` variable (the number of dependents in the household).

! Hint

Rather than creating the interaction term manually, you can use Stata's factor variable notation. For example, `c.female#c.has_child` will create the interaction term for you. If you want to include the main effects as well, you can use `i.female##i.has_child`.

Calculate the estimated wage differential (in percent) for the following comparisons. Show your work using the coefficients (β):

1. Fatherhood Premium: What is the effect of having a child for a man (relative to a childless man)?
2. Motherhood Penalty: What is the effect of having a child for a woman (relative to a childless woman)?

! Solution

```
gen has_child = numdep > 0
reg lwage educ exper exper2 i.female##i.has_child
```

Results: - $\hat{\beta}_{child} \approx 0.031$ (Not statistically significant, $p = 0.56$). - $\hat{\beta}_{female} \approx -0.255$. - $\hat{\beta}_{interaction} \approx -0.154$.

1. **Fatherhood Premium:** $\hat{\beta}_{child} \approx 3.1\%$.
   - Men with children earn 3.1% more than men without children (holding other factors constant), though this is not statistically significant in this sample.
2. **Motherhood Penalty:** $\hat{\beta}_{child} + \hat{\beta}_{interaction} = 0.031 - 0.154 = -0.123$.
   - Women with children earn approximately 12.3% less than women without children.

Problem 4

Your group suspects that for men, parenthood has statistically zero effect on wages (meaning they receive neither a penalty nor a premium).

1. State the Null Hypothesis (H_0) required to test if being a parent has no effect specifically for men.
2. Now, assume you want to test if parenthood has no effect for anyone (men or women). State the joint Null Hypothesis involving multiple coefficients ($H_0 : \beta = 0 \quad \text{and} \quad \beta = 0$)
3. Explain how you would calculate the F-statistic for this second test using the R^2 of a "restricted" and "unrestricted" model

> ! Solution

1. **Test for Men:** $H_0 : \beta_5 = 0$ (coefficient on `has_child`).
   - Since `female`=0 for men, the interaction term drops out, and only β_5 matters.
2. **Test for Anyone:** $H_0 : \beta_5 = 0 \quad \text{AND} \quad \beta_6 = 0$.
   - If both the main effect of children and the interaction with gender are zero, children do not affect wages for either group.
3. **F-Statistic:**
   - **Unrestricted Model:** The full model from Problem 3 ($R^2_{UR} \approx 0.4060$).
   - **Restricted Model:** Run the regression *excluding* `has_child` and `female_x_child` ($R^2_R \approx 0.3996$).
   - Formula:

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(n - k - 1)}$$

     - $q = 2$ (number of restrictions).
     - $n = 526$, $k = 6$.
     - Result: $F \approx 2.79$ ($p \approx 0.06$). We fail to reject the null at the 5% level, but might reject at the 10% level.

Problem 5

One member of your group argues that the relationship between Education and Wages is biased because we are looking at too many variables at once. They want a simple 2D scatterplot of the "pure" relationship between log(Wage) and Education, purged of experience and gender effects.

Produce this plot using the Frisch-Waugh-Lovell theorem.

Why is the slope of this line numerically identical to the coefficient on educ in the full multiple regression? Why are the standard errors different?

! Solution

Steps: 1. Regress $ln(wage)$ on controls $(exper, exper^2, female) \to$ get residuals $\tilde{y}$. 2. Regress $educ$ on controls $(exper, exper^2, female) \to$ get residuals $\tilde{x}$. 3. Plot $\tilde{y}$ vs $\tilde{x}$. The slope of the regression line through these residuals is the partial effect of education.

```
reg lwage exper exper2 female
predict r_lwage, resid

reg educ exper exper2 female
predict r_educ, resid

twoway (scatter r_lwage r_educ) (lfit r_lwage r_educ)
```

Why identical? - The FWL theorem proves mathematically that the coefficient from the residual regression is exactly the same as the coefficient in the multivariate regression ($\hat{\beta} \approx 0.084$). The multivariate coefficient *is* the relationship between X and Y after "partialing out" the other variables.

Why different standard errors? - The manual residual regression (Step 3) generates standard errors based on $N - 2$ degrees of freedom (since it sees only 1 regressor and a constant). - The true multivariate model uses $N - K - 1$ degrees of freedom (accounting for estimation of all control variables). - The manual method underestimates the variance of the error term because it claims to satisfy "fewer" constraints than the data actually supported.