ECON 452* -- Introduction to Notes 5 to 8

A Linear Regression Model with Both Continuous and Categorical Explanatory Variables

Consider a multiple linear regression equation that has a ***continuous*** **regressand Y**; two ***continuous*** **explanatory variables**, $\mathbf{X_1}$ and $\mathbf{X_2}$; and two ***categorical*** explanatory variables, **gender** and **industry**.

- Y_i = a *continuous* regressand.
- X_{i1} = a *continuous* explanatory variable.
- X_{i2} = a second *continuous* explanatory variable.
- The *categorical* explanatory variable **gender** is represented by the binary female indicator (dummy) variable F_i, where by definition $F_i = 1$ if observation i is female, and $F_i = 0$ if observation i is male. The **base group** for gender is **males**.
- The *categorical* explanatory variable **industry** is a four-category categorical explanatory variable identifying which of industries 1, 2, 3 and 4 an individual observation is in. The categorical variable industry is completely represented by the following three industry indicator (dummy) variables, where the **base group industry** is arbitrarily chosen to be **industry 1**:

  $IN2_i = 1$ if observation i is in industry 2, = 0 otherwise (meaning observation i is in industry 1, 3 or 4)

  $IN3_i = 1$ if observation i is in industry 3, = 0 otherwise (meaning observation i is in industry 1, 2 or 4)

  $IN4_i = 1$ if observation i is in industry 4, = 0 otherwise (meaning observation i is in industry 1, 2 or 3)

Research Objective: To investigate ***whether* and *how*** a population regression function differs between females and males.

- The **population regression function for *females***, for whom the female indicator $F_i = 1$, is:

$$\begin{aligned} & E(Y_i \mid F_i = 1, X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) \\ & \quad = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i1}^2 + \alpha_4 X_{i2}^2 + \alpha_5 X_{i1} X_{i2} + \alpha_6 IN2_i + \alpha_7 IN3_i + \alpha_8 IN4_i \end{aligned} \qquad \textbf{(1f)}$$

The corresponding **population regression equation for *females*** can be written as:

$$Y_1 = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i1}^2 + \alpha_4 X_{i2}^2 + \alpha_5 X_{i1} X_{i2} + \alpha_6 IN2_i + \alpha_7 IN3_i + \alpha_8 IN4_i + u_i \qquad \text{for } F_i = 1$$

Estimate the female PRE on the subsample of female observations with the following *Stata* **regress** command:

```
regress y x1 x2 x1sq x2sq x1x2 in2 in3 in4 if f == 1
```

- The **population regression function for *males***, for whom the female indicator $F_i = 0$, is:

$$\begin{aligned} & E(Y_i \mid F_i = 0, X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) \\ & \quad = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 IN2_i + \beta_7 IN3_i + \beta_8 IN4_i \end{aligned} \qquad \textbf{(1m)}$$

The corresponding **population regression equation for *males*** can be written as:

$$Y_1 = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 IN2_i + \beta_7 IN3_i + \beta_8 IN4_i + u_i \qquad \text{for } F_i = 0$$

Estimate the male PRE on the subsample of male observations with the following *Stata* **regress** command:

```
regress y x1 x2 x1sq x2sq x1x2 in2 in3 in4 if f == 0
```

- The ***female-male difference*** **in the conditional mean value of Y** is obtained by subtracting the **male population regression function (1m)** from the **female population regression function (1f)**:

$$E(Y_i \mid F_i = 1, X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) - E(Y_i \mid F_i = 0, X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i)$$

$$\begin{aligned}
&= \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i1}^2 + \alpha_4 X_{i2}^2 + \alpha_5 X_{i1} X_{i2} + \alpha_6 IN2_i + \alpha_7 IN3_i + \alpha_8 IN4_i \\
&\quad - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 IN2_i + \beta_7 IN3_i + \beta_8 IN4_i)
\end{aligned}$$

$$\begin{aligned}
&= \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i1}^2 + \alpha_4 X_{i2}^2 + \alpha_5 X_{i1} X_{i2} + \alpha_6 IN2_i + \alpha_7 IN3_i + \alpha_8 IN4_i \\
&\quad - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i1}^2 - \beta_4 X_{i2}^2 - \beta_5 X_{i1} X_{i2} - \beta_6 IN2_i - \beta_7 IN3_i - \beta_8 IN4_i
\end{aligned}$$

$$\begin{aligned}
&= (\alpha_0 - \beta_0) + (\alpha_1 - \beta_1) X_{i1} + (\alpha_2 - \beta_2) X_{i2} + (\alpha_3 - \beta_3) X_{i1}^2 + (\alpha_4 - \beta_4) X_{i2}^2 + (\alpha_5 - \beta_5) X_{i1} X_{i2} \\
&\qquad + (\alpha_6 - \beta_6) IN2_i + (\alpha_7 - \beta_7) IN3_i + (\alpha_8 - \beta_8) IN4_i
\end{aligned}$$

where each $(\alpha_j - \beta_j)$, j = 0, 1, 2, …, 8, is a **female-male coefficient difference**.

Limitations of the Separate Regressions Approach

Question: Why not just separately estimate the male and female regression equations on their respective subsamples of male and female observations, as we have done on the previous slide?

Reasons:

1. Separate estimation of the female and male population regression equations on the subsamples of female and male observations **does not allow us to *test* for female-male coefficient differences**.

Example 1: To test that **industry effects are *equal* for females and males**, we want to perform the following joint hypothesis test:

H_0: $\alpha_6 = \beta_6$ *and* $\alpha_7 = \beta_7$ *and* $\alpha_8 = \beta_8$
$\alpha_6 - \beta_6 = 0$ *and* $\alpha_7 - \beta_7 = 0$ *and* $\alpha_8 - \beta_8 = 0$

H_1: $\alpha_6 \neq \beta_6$ *and/or* $\alpha_7 \neq \beta_7$ *and/or* $\alpha_8 \neq \beta_8$
$\alpha_6 - \beta_6 \neq 0$ *and/or* $\alpha_7 - \beta_7 \neq 0$ *and/or* $\alpha_8 - \beta_8 \neq 0$

Example 2: To test that the **marginal effect of the continuous explanatory variable X_1 is the same for females and males for any given values of X_{i1} and X_{i2}**, we must perform the following joint hypothesis test:

H_0: $\alpha_1 = \beta_1$ *and* $\alpha_3 = \beta_3$ *and* $\alpha_5 = \beta_5$
$\alpha_1 - \beta_1 = 0$ *and* $\alpha_3 - \beta_3 = 0$ *and* $\alpha_5 - \beta_5 = 0$

H_1: $\alpha_1 \neq \beta_1$ *and/or* $\alpha_3 \neq \beta_3$ *and/or* $\alpha_5 \neq \beta_5$
$\alpha_1 - \beta_1 \neq 0$ *and/or* $\alpha_3 - \beta_3 \neq 0$ *and/or* $\alpha_5 - \beta_5 \neq 0$

2. Separate estimation of the female and male population regression equations **does not allow us to constrain some or all of the regression coefficients to be equal for females and males**.

Example: Suppose we retain the null hypothesis that industry effects are equal for females and males. We then might want to impose the coefficient restrictions $\alpha_6 = \beta_6$ *and* $\alpha_7 = \beta_7$ *and* $\alpha_8 = \beta_8$ in estimating the female and male regression functions. But the separate regressions approach does not provide a way of doing this.

What's Ahead in Notes 5 to 8?

We will learn that the foregoing limitations of the separate regressions approach to investigating female-male differences in regression functions can be overcome by formulating a **pooled full-interaction regression equation in the female indicator variable $\mathbf{F_i}$**, and then estimating this pooled regression equation on the combined (or pooled) sample of male and female observations.

- The **pooled full-interaction regression equation in the female indicator $\mathbf{F_i}$** takes the form:

$$\begin{aligned} Y_1 = {} & \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 IN2_i + \beta_7 IN3_i + \beta_8 IN4_i \\ & + \delta_0 F_i + \delta_1 F_i X_{i1} + \delta_2 F_i X_{i2} + \delta_3 F_i X_{i1}^2 + \delta_4 F_i X_{i2}^2 + \delta_5 F_i X_{i1} X_{i2} + \delta_6 F_i IN2_i + \delta_7 F_i IN3_i + \delta_8 F_i IN4_i + u_i \end{aligned} \qquad \textbf{(2)}$$

where i = 1, …, $N = N_m + N_f$.

The $\boldsymbol{\beta_j}$ **coefficients** are the ***male*** **regression coefficients** for all j = 0, 1, …, 8.

The $\boldsymbol{\delta_j}$ **coefficients** are the ***female-male*** **coefficient *differences***, i.e., $\delta_j = \alpha_j - \beta_j$ for all j = 0, 1, …, 8.

The ***female*** **regression coefficients** are estimated indirectly as $\alpha_j = \beta_j + \delta_j$ for all j = 0, 1, …, 8.

$$Y_1 = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 IN2_i + \beta_7 IN3_i + \beta_8 IN4_i$$
$$+ \delta_0 F_i + \delta_1 F_i X_{i1} + \delta_2 F_i X_{i2} + \delta_3 F_i X_{i1}^2 + \delta_4 F_i X_{i2}^2 + \delta_5 F_i X_{i1} X_{i2} + \delta_6 F_i IN2_i + \delta_7 F_i IN3_i + \delta_8 F_i IN4_i + u_i \quad \textbf{(2)}$$

- The pooled full-interaction regression equation (2) **facilitates testing for female-male differences in regression functions**.

  Example 1: To test that **industry effects are *equal* for females and males**, we perform the following joint hypothesis test on the pooled regression equation (2):

  H_0: $\delta_6 = 0$ *and* $\delta_7 = 0$ *and* $\delta_8 = 0$ OR $\delta_j = 0$ for all j = 6, 7, 8

  H_1: $\delta_6 \neq 0$ *and/or* $\delta_7 \neq 0$ *and/or* $\delta_8 \neq 0$ OR $\delta_j \neq 0$ for j = 6, 7, 8

  Example 2: To test that the **marginal effect of the continuous explanatory variable X_1 is the same for females and males for any given values of X_{i1} and X_{i2}**, we perform the following joint hypothesis test on pooled regression equation (2):

  H_0: $\delta_1 = 0$ *and* $\delta_3 = 0$ *and* $\delta_5 = 0$ OR $\delta_j = 0$ for all j = 1, 3, 5

  H_1: $\delta_1 \neq 0$ *and/or* $\delta_3 \neq 0$ *and/or* $\delta_5 \neq 0$ OR $\delta_j \neq 0$ for j = 1, 3, 5

$$\begin{aligned} Y_1 = {} & \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 IN2_i + \beta_7 IN3_i + \beta_8 IN4_i \\ & + \delta_0 F_i + \delta_1 F_i X_{i1} + \delta_2 F_i X_{i2} + \delta_3 F_i X_{i1}^2 + \delta_4 F_i X_{i2}^2 + \delta_5 F_i X_{i1} X_{i2} + \delta_6 F_i IN2_i + \delta_7 F_i IN3_i + \delta_8 F_i IN4_i + u_i \end{aligned} \qquad \textbf{(2)}$$

- The pooled full-interaction regression equation (2) **enables us to constrain some of the regression coefficients to be equal for females and males.**

  Example: Suppose we ***retain*** **the null hypothesis** that **industry effects are equal for females and males**. We can then impose these restrictions in estimating the female and male regression functions by simply imposing on the pooled regression equation (2) the coefficient restrictions $\boldsymbol{\delta_6 = 0}$ ***and*** $\boldsymbol{\delta_7 = 0}$ ***and*** $\boldsymbol{\delta_8 = 0}$.

  The resulting ***restricted*** **pooled regression equation** is:

$$\begin{aligned} Y_1 = {} & \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \beta_6 IN2_i + \beta_7 IN3_i + \beta_8 IN4_i \\ & + \delta_0 F_i + \delta_1 F_i X_{i1} + \delta_2 F_i X_{i2} + \delta_3 F_i X_{i1}^2 + \delta_4 F_i X_{i2}^2 + \delta_5 F_i X_{i1} X_{i2} + u_i \end{aligned} \qquad \textbf{(3)}$$

- **Table formats** for reporting coefficient estimates of **Model 5.5**:

Table 2: OLS Estimates of Model 5.5 on Pooled Sample of Females and Males

	Females		Males		Female-Male Differences	
Regressor Name	Coef. Estimate	t-ratio	Coef. Estimate	t-ratio	Coef. Estimate	t-ratio
Intercept						
X_1						
X_2						
X_3						
X_1-sq						
X_2-sq						
X_1X_2						
IN2						
IN3						
IN4						
No. of obs =						
RSS =						
R-squared =						
ANOVA F =						
p-value of F =						

Table 2: OLS Estimates of Model 5.5 on Pooled Sample of Females and Males

	Females		Males		Female-Male Differences	
Regressor	$\hat{\beta}_j + \hat{\delta}_j$	t-ratio	$\hat{\beta}_j$	t-ratio	$\hat{\delta}_j$	t-ratio
Intercept						
X_1						
X_2						
X_3						
X_1-sq						
X_2-sq						
X_1X_2						
IN2						
IN3						
IN4						
No. of obs =						
RSS =						
R-squared =						
ANOVA F =						
p-value of F =						

Table 5: Hypothesis Test Results for Model 5.5

#	Null Hypothesis H_0	Interpretation of H_0	q[1/]	p-value[2/]
1	$\beta_1 + \delta_1 = 0$ & $\beta_4 + \delta_4 = 0$ & $\beta_6 + \delta_6 = 0$	ME of X_1 is zero for females	3	0.0000
2	$\beta_4 + \delta_4 = 0$ & $\beta_6 + \delta_6 = 0$	ME of X_1 is constant for females	2	0.0274
3	$\beta_1 = 0$ & $\beta_4 = 0$ & $\beta_6 = 0$	ME of X_1 is zero for males	3	0.0014
4	$\beta_4 = 0$ & $\beta_6 = 0$	ME of X_1 is constant for males	2	0.0083
5	$\delta_1 = 0$ & $\delta_4 = 0$ & $\delta_6 = 0$	ME of X_1 equal for females & males	3	0.0038
6	$\delta_4 = 0$ & $\delta_6 = 0$	F-M difference in ME of X_1 a constant	2	0.1494
7	$\beta_2 + \delta_2 = 0$ & $\beta_5 + \delta_5 = 0$ & $\beta_6 + \delta_6 = 0$	ME of X_2 is zero for females	3	0.0000
8	$\beta_5 + \delta_5 = 0$ & $\beta_6 + \delta_6 = 0$	ME of X_2 is constant for females	2	0.0000
9	$\beta_2 = 0$ & $\beta_5 = 0$ & $\beta_6 = 0$	ME of X_2 is zero for males	3	0.0741
10	$\beta_5 = 0$ & $\beta_6 = 0$	ME of X_2 is constant for males	2	0.3185
11	$\delta_2 = 0$ & $\delta_5 = 0$ & $\delta_6 = 0$	ME of X_2 equal for females & males	3	0.0063
12	$\delta_5 = 0$ & $\delta_6 = 0$	F-M difference in ME of X_2 a constant	2	0.03119
13	$\beta_3 + \delta_3 = 0$	ME of X_3 is zero for females	1	0.0372
14	$\beta_3 = 0$	ME of X_3 is zero for males	1	0.2461
15	$\delta_3 = 0$	ME of X_3 equal for females & males	1	0.0000
16	$\beta_7 + \delta_7 = 0$ & $\beta_8 + \delta_8 = 0$ & $\beta_9 + \delta_9 = 0$	No industry effects for females	3	0.0000
17	$\beta_7 = 0$ & $\beta_8 = 0$ & $\beta_9 = 0$	No industry effects for males	3	0.0000
18	$\delta_7 = 0$ & $\delta_8 = 0$ & $\delta_9 = 0$	Industry effects equal, females & males	3	0.0083
19	$\delta_j = 0$ for all $j = 0, 1, \ldots, 9$	F-M mean Y difference = 0	10	0.0000
20	$\delta_j = 0$ for all $j = 1, \ldots, 9$	F-M mean Y difference is constant	9	0.0000

Notes: 1/. q denotes the number of coefficient restrictions specified by the null hypothesis H_0. 2/. The p-values are two-tail p-values for the calculated sample value of the test statistic.