ECON 452* -- NOTE 6

Using Dummy Variable Regressors for Multi-Category Categorical Variables

❑ Dummy Variable Regressors for Multi-Category Variables

- **Consider** a ***four-way*** **partitioning** of a population or sample into **four *mutually exclusive* and *exhaustive* industry groups --** ***industry 1***, ***industry 2***, ***industry 3***, and ***industry 4***.

  - Let $\mathbf{IN1_i}$ be the **indicator (dummy) variable** for ***industry 1***:

    $IN1_i$ = 1 if observation i is in industry 1
    = 0 if observation i is not in industry 1.

  - Let $\mathbf{IN2_i}$ be the **indicator (dummy) variable** for ***industry 2***:

    $IN2_i$ = 1 if observation i is in industry 2
    = 0 if observation i is not in industry 2.

  - Let $\mathbf{IN3_i}$ be the **indicator (dummy) variable** for ***industry 3***:

    $IN3_i$ = 1 if observation i is in industry 3
    = 0 if observation i is not in industry 3.

  - Let $\mathbf{IN4_i}$ be the **indicator (dummy) variable** for ***industry 4***:

    $IN4_i$ = 1 if observation i is in industry 4
    = 0 if observation i is not in industry 4.

- **Adding-Up Property of the Industry Indicator Variables:**

$$IN1_i + IN2_i + IN3_i + IN4_i = 1 \qquad \forall\ i$$

- **Implications of the Adding-Up Property**

Any *three* of the four industry dummy variables $IN1_i$, $IN2_i$, $IN3_i$ and $IN4_i$ *completely represents* the *four-way partitioning* of a population and sample into four industry groups.

❑ Model 1 -- The Benchmark Model

Contains three regressors in the two explanatory variables X_1 and X_2, both of which are assumed to be *continuous* variables.

- The **population regression equation for Model 1** takes the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad \textbf{(1)}$$

- The **population regression function**, or conditional mean function, **for Model 1** takes the form

$$E(Y_i \mid X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \qquad (1')$$

- Model 1 does not allow for any coefficient differences among subgroups of the relevant population, such as coefficient differences among industries.

  Model 1 assumes that all three regression coefficients β_j (j = 0, 1, 2) are the same for all population members.

  Model 1 assumes that the population regression function is the same for all population members.

❑ Model 4: Different Industry Intercept Coefficients

Model 4.1 -- Version 1 of Model 4: No Industry Base Group

Allows for **different *industry intercepts*** by introducing all four industry dummy variables $IN1_i$, $IN2_i$, $IN3_i$, and $IN4_i$ as additional additive regressors in Model 1.

- The **population regression equation for Model 4.1** is:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 IN1_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i \qquad \textbf{(4.1)}$$

The distinguishing characteristic of Model 4.1 is that it contains **no *intercept coefficient***. That is because there is no industry base group in Model 4.1.

- The **population regression function**, or conditional mean function, **for Model 4.1** is obtained by taking the conditional expectation of regression equation (4.1) for any given values of the regressors X_{i1}, X_{i2}, $IN1_i$, $IN2_i$, $IN3_i$, and $IN4_i$:

$$E(Y_i \mid X_{i1}, X_{i2}, IN1_i, IN2_i, IN3_i, IN4_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 IN1_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i \qquad (4.1')$$

- The **population regression function**, or CMF, **for *industry 1*** implied by Model 4.1 is obtained by setting the industry 1 indicator variable $IN1_i = 1$ in (4.1'), which implies that $IN2_i = 0$ and $IN3_i = 0$ and $IN4_i = 0$:

$$E(Y_i \mid X_{i1}, X_{i2}, IN1_i = 1) = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

The ***industry 1 intercept* coefficient = ϕ_1**.

$$E(Y_i \mid X_{i1}, X_{i2}, IN1_i, IN2_i, IN3_i, IN4_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 IN1_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i \qquad (4.1')$$

- The **population regression function for *industry 2*** implied by Model 4.1 is obtained by setting the industry 2 indicator variable $IN2_i = 1$ in (4.1'), which implies that $IN1_i = 0$ and $IN3_i = 0$ and $IN4_i = 0$:

$$E(Y_i \mid X_{i1}, X_{i2}, IN2_i = 1) = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_2 = \phi_2 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

The ***industry 2 intercept* coefficient = $\boldsymbol{\phi_2}$.**

- The **population regression function for *industry 3*** implied by Model 4.1 is obtained by setting the industry 3 indicator variable $IN3_i = 1$ in (4.1'), which implies that $IN1_i = 0$ and $IN2_i = 0$ and $IN4_i = 0$:

$$E(Y_i \mid X_{i1}, X_{i2}, IN3_i = 1) = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_3 = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

The ***industry 3 intercept* coefficient = $\boldsymbol{\phi_3}$.**

- The **population regression function for *industry 4*** implied by Model 4.1 is obtained by setting the industry 4 indicator variable $IN4_i = 1$ in (4.1'), which implies that $IN1_i = 0$ and $IN2_i = 0$ and $IN3_i = 0$:

$$E(Y_i \mid X_{i1}, X_{i2}, IN4_i = 1) = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_4 = \phi_4 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

The ***industry 4 intercept* coefficient = $\boldsymbol{\phi_4}$.**

- ***Hypothesis Test:*** Test the proposition that there are **no differences in mean Y *across industries*** for population members with given values of X_1 and X_2. There are **no *inter-industry* differences** in the conditional mean values of Y for given values of X_1 and X_2.

  In terms of the regression coefficients in Model 4.1, this hypothesis states that **the *four* industry coefficients ϕ_1, ϕ_2, ϕ_3, and ϕ_4 are *all equal***.

- The ***null* and *alternative* hypotheses** are as follows:

  H_0: $\phi_2 = \phi_1$ *and* $\phi_3 = \phi_1$ *and* $\phi_4 = \phi_1$
  $\phi_2 - \phi_1 = 0$ *and* $\phi_3 - \phi_1 = 0$ *and* $\phi_4 - \phi_1 = 0$

  H_1: $\phi_2 \neq \phi_1$ *and/or* $\phi_3 \neq \phi_1$ *and/or* $\phi_4 \neq \phi_1$
  $\phi_2 - \phi_1 \neq 0$ *and/or* $\phi_3 - \phi_1 \neq 0$ *and/or* $\phi_4 - \phi_1 \neq 0$

- The ***restricted* model implied by the null hypothesis H_0** is obtained by imposing on Model 4.1 (the unrestricted model) the coefficient restrictions specified by H_0.

  Model 4.1, the unrestricted model, is:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 IN1_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i \qquad \textbf{(4.1)}$$

  The ***restricted* model** is obtained by setting $\phi_2 = \phi_1$ and $\phi_3 = \phi_1$ and $\phi_4 = \phi_1$ in Model 4.1:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 IN1_i + \phi_1 IN2_i + \phi_1 IN3_i + \phi_1 IN4_i + u_i$$

i.e.,

$$
\begin{aligned}
Y_i &= \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 (IN1_i + IN2_i + IN3_i + IN4_i) + u_i \\
&= \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 + u_i \\
&= \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \\
&= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad \textbf{(1)}
\end{aligned}
$$

- The ***test statistic*** appropriate for this hypothesis test is a **Wald F-statistic**.

Model 4.2 -- Version 2 of Model 4: Base Group is Industry 1

Model 4.2 allows for **different *industry intercepts*** by introducing the three industry dummy variables $IN2_i$, $IN3_i$, and $IN4_i$ as additional additive regressors in Model 1. The ***industry* base group in Model 4.2 is *industry 1***. The **industry 1 dummy variable $IN1_i$ is *excluded*** from the regressor set.

- The **population regression equation for Model 4.2** is:

$$Y_i = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_2 IN2_i + \psi_3 IN3_i + \psi_4 IN4_i + u_i \qquad \textbf{(4.2)}$$

- The **population regression function**, or conditional mean function, **for Model 4.2** is obtained by taking the conditional expectation of regression equation (4.2) for any given values of the regressors X_{i1}, X_{i2}, $IN2_i$, $IN3_i$, and $IN4_i$:

$$E(Y_i \mid X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_2 IN2_i + \psi_3 IN3_i + \psi_4 IN4_i \qquad (4.2')$$

- The **population regression function**, or CMF, **for *industry 1*** -- the industry base group -- in Model 4.2 is obtained by setting all three included industry dummy variables in (4.2') equal to zero, i.e., by setting $IN2_i = 0$ and $IN3_i = 0$ and $IN4_i = 0$ in (4.2'):

$$E(Y_i \mid X_{i1}, X_{i2}, IN1_i = 1)$$

$$= E(Y_i \mid X_{i1}, X_{i2}, IN2_i = 0, IN3_i = 0, IN4_i = 0) = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

The ***industry 1 intercept* coefficient = $\boldsymbol{\phi_1}$** = the equation intercept coefficient

$$E(Y_i \mid X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_2 IN2_i + \psi_3 IN3_i + \psi_4 IN4_i \qquad (4.2')$$

- The **population regression function for *industry 2*** implied by Model 4.2 is obtained by setting the industry 2 dummy variable $IN2_i = 1$ in (4.2'), which by definition requires that $IN3_i = 0$ and $IN4_i = 0$ in (4.2'):

$$\begin{aligned} E(Y_i \mid X_{i1}, X_{i2}, IN2_i = 1) &= E(Y_i \mid X_{i1}, X_{i2}, IN2_i = 1, IN3_i = 0, IN4_i = 0) \\ &= \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_2 \\ &= (\phi_1 + \psi_2) + \beta_1 X_{i1} + \beta_2 X_{i2} \end{aligned}$$

The ***industry 2 intercept* coefficient = $\phi_1 + \psi_2$**
The ***industry 1 intercept* coefficient = ϕ_1**

Therefore, the $IN2_i$ coefficient ψ_2 in Model 4.2 is:

ψ_2 = *industry 2 intercept* coefficient – *industry 1 intercept* coefficient

$$E(Y_i \mid X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_2 IN2_i + \psi_3 IN3_i + \psi_4 IN4_i \qquad (4.2')$$

- The **population regression function for *industry 3*** implied by Model 4.2 is obtained by setting the industry 3 dummy variable $IN3_i = 1$ in (4.2'), which by definition requires that $IN2_i = 0$ and $IN4_i = 0$ in (4.2'):

$$\begin{aligned} E(Y_i \mid X_{i1}, X_{i2}, IN3_i = 1) &= E(Y_i \mid X_{i1}, X_{i2}, IN2_i = 0, IN3_i = 1, IN4_i = 0) \\ &= \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_3 \\ &= (\phi_1 + \psi_3) + \beta_1 X_{i1} + \beta_2 X_{i2} \end{aligned}$$

The ***industry 3 intercept* coefficient = $\phi_1 + \psi_3$**
The ***industry 1 intercept* coefficient = ϕ_1**

Therefore, the $IN3_i$ coefficient ψ_3 in Model 4.2 is:

ψ_3 = *industry 3 intercept* coefficient – *industry 1 intercept* coefficient

$$E(Y_i \mid X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_2 IN2_i + \psi_3 IN3_i + \psi_4 IN4_i \qquad (4.2')$$

- The **population regression function for *industry 4*** implied by Model 4.2 is obtained by setting the industry 4 dummy variable $IN4_i = 1$ in (4.2'), which by definition requires that $IN2_i = 0$ and $IN3_i = 0$ in (4.2'):

$$\begin{aligned} E(Y_i \mid X_{i1}, X_{i2}, IN4_i = 1) &= E(Y_i \mid X_{i1}, X_{i2}, IN2_i = 0, IN3_i = 0, IN4_i = 1) \\ &= \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_4 \\ &= (\phi_1 + \psi_4) + \beta_1 X_{i1} + \beta_2 X_{i2} \end{aligned}$$

The ***industry 4 intercept* coefficient = $\phi_1 + \psi_4$**
The ***industry 1 intercept* coefficient = ϕ_1**

Therefore, the $IN4_i$ coefficient ψ_4 in Model 4.2 is:

ψ_4 = *industry 4 intercept* coefficient – *industry 1 intercept* coefficient

- ***Hypothesis Test:*** Test the proposition that there are **no differences in mean Y *across industries*** for population members with given values of X_1 and X_2. There are **no *inter-industry* differences** in the conditional mean values of Y for given values of X_1 and X_2.

  In Model 4.2, this hypothesis requires that **the *three* industry coefficients ψ_2, ψ_3, and ψ_4 are *all zero***.

  The ***null* and *alternative* hypotheses** are as follows:

  H_0: $\psi_2 = 0$ *and* $\psi_3 = 0$ *and* $\psi_4 = 0$
  $\phi_2 - \phi_1 = 0$ *and* $\phi_3 - \phi_1 = 0$ *and* $\phi_4 - \phi_1 = 0$

  H_1: $\psi_2 \neq 0$ *and/or* $\psi_3 \neq 0$ *and/or* $\psi_4 \neq 0$
  $\phi_2 - \phi_1 \neq 0$ *and/or* $\phi_3 - \phi_1 \neq 0$ *and/or* $\phi_4 - \phi_1 \neq 0$

- The ***restricted* model implied by the null hypothesis H_0** is obtained by imposing on Model 4.2 (the unrestricted model) the coefficient restrictions specified by H_0.

  Model 4.2, the unrestricted model, is:

  $$Y_i = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_2 IN2_i + \psi_3 IN3_i + \psi_4 IN4_i + u_i \qquad \textbf{(4.2)}$$

  The ***restricted* model** is obtained by setting $\psi_2 = 0$ and $\psi_3 = 0$ and $\psi_4 = 0$ in Model 4.2:

  $$Y_i = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad \textbf{(1)}$$

- The ***test statistic*** appropriate for this hypothesis test is a **Wald F-statistic**.

Model 4.3 -- Version 3 of Model 4: Base Group is Industry 3

Model 4.3 allows for **different *industry intercepts*** by introducing the three industry dummy variables $IN1_i$, $IN2_i$, and $IN4_i$ as additional additive regressors in Model 1. The ***industry* base group in Model 4.3 is *industry 3***. The industry 3 dummy variable $IN3_i$ is excluded from the regressor set.

- The **population regression equation for Model 4.3** is:

$$Y_i = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_1 IN1_i + \omega_2 IN2_i + \omega_4 IN4_i + u_i \qquad \textbf{(4.3)}$$

- The **population regression function**, or conditional mean function, **for Model 4.3** is obtained by taking the conditional expectation of regression equation (4.3) for any given values of the regressors X_{i1}, X_{i2}, $IN1_i$, $IN2_i$, and $IN4_i$:

$$E(Y_i \mid X_{i1}, X_{i2}, IN1_i, IN2_i, IN4_i) = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_1 IN1_i + \omega_2 IN2_i + \omega_4 IN4_i \qquad (4.3')$$

- The **population regression function for *industry 3* --** the industry base group -- in Model 4.3 is obtained by setting all three included industry dummy variables in (4.3') equal to zero, i.e., by setting $IN1_i = 0$ and $IN2_i = 0$ and $IN4_i = 0$ in (4.3'):

$$\begin{aligned} E(Y_i \mid X_{i1}, X_{i2}, IN3_i = 1) &= E(Y_i \mid X_{i1}, X_{i2}, IN1_i = 0, IN2_i = 0, IN4_i = 0) \\ &= \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} \end{aligned}$$

The ***industry 3 intercept* coefficient = $\boldsymbol{\phi_3}$** = the equation intercept coefficient

$$E(Y_i \mid X_{i1}, X_{i2}, IN1_i, IN2_i, IN4_i) = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_1 IN1_i + \omega_2 IN2_i + \omega_4 IN4_i \qquad (4.3')$$

- The **population regression function for *industry 1* in Model 4.3** is obtained by setting $IN1_i = 1$ and $IN2_i = 0$ and $IN4_i = 0$ in equation (4.3'):

$$\begin{aligned} E(Y_i \mid X_{i1}, X_{i2}, IN1_i = 1) &= E(Y_i \mid X_{i1}, X_{i2}, IN1_i = 1, IN2_i = 0, IN4_i = 0) \\ &= \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_1 \\ &= (\phi_3 + \omega_1) + \beta_1 X_{i1} + \beta_2 X_{i2} \end{aligned}$$

The ***industry 1 intercept* coefficient = $\phi_3 + \omega_1$**
The ***industry 3 intercept* coefficient = ϕ_3**

Therefore, the $IN1_i$ coefficient ω_1 in Model 4.3 is:

ω_1 = *industry 1 intercept* coefficient – *industry 3 intercept* coefficient

$$E(Y_i \mid X_{i1}, X_{i2}, IN1_i, IN2_i, IN4_i) = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_1 IN1_i + \omega_2 IN2_i + \omega_4 IN4_i \qquad (4.3')$$

- The **population regression function for *industry 2*** implied by Model 4.3 is obtained by setting the industry 2 dummy variable $IN2_i = 1$ in (4.3'), which by definition requires that $IN1_i = 0$ and $IN4_i = 0$ in (4.3'):

$$\begin{aligned} E(Y_i \mid X_{i1}, X_{i2}, IN2_i = 1) &= E(Y_i \mid X_{i1}, X_{i2}, IN1_i = 0, IN2_i = 1, IN4_i = 0) \\ &= \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_2 \\ &= (\phi_3 + \omega_2) + \beta_1 X_{i1} + \beta_2 X_{i2} \end{aligned}$$

The ***industry 2 intercept* coefficient = $\phi_3 + \omega_2$**
The ***industry 3 intercept* coefficient = ϕ_3**

Therefore, the $IN2_i$ coefficient ω_2 in Model 4.3 is:

ω_2 = *industry 2 intercept* coefficient – *industry 3 intercept* coefficient

$$E(Y_i \mid X_{i1}, X_{i2}, IN1_i, IN2_i, IN4_i) = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_1 IN1_i + \omega_2 IN2_i + \omega_4 IN4_i \qquad (4.3')$$

- The **population regression function for *industry 4*** implied by Model 4.3 is obtained by setting the industry 4 dummy variable $IN4_i = 1$ in (4.3'), which by definition requires that $IN1_i = 0$ and $IN2_i = 0$ in (4.3'):

$$\begin{aligned} E(Y_i \mid X_{i1}, X_{i2}, IN4_i = 1) &= E(Y_i \mid X_{i1}, X_{i2}, IN1_i = 0, IN2_i = 0, IN4_i = 1) \\ &= \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_4 \\ &= (\phi_3 + \omega_4) + \beta_1 X_{i1} + \beta_2 X_{i2} \end{aligned}$$

The ***industry 4 intercept* coefficient = $\phi_3 + \omega_4$**
The ***industry 3 intercept* coefficient = ϕ_3**

Therefore, the $IN4_i$ coefficient ω_4 in Model 4.3 is:

ω_4 = *industry 4 intercept* coefficient – *industry 3 intercept* coefficient

- ***Hypothesis Test:*** Test the proposition that there are no differences in mean Y across industries for population members with given values of X_1 and X_2. There are **no *inter-industry* differences** in the conditional mean values of Y for given values of X_1 and X_2.

  In Model 4.3, this hypothesis requires that **the *three* industry coefficients ω_1, ω_2, and ω_4 are *all zero***.

  The ***null* and *alternative* hypotheses** are as follows:

  H_0: $\omega_1 = 0$ *and* $\omega_2 = 0$ *and* $\omega_4 = 0$
  $\phi_1 - \phi_3 = 0$ *and* $\phi_2 - \phi_3 = 0$ *and* $\phi_4 - \phi_3 = 0$

  H_1: $\omega_1 \neq 0$ *and/or* $\omega_2 \neq 0$ *and/or* $\omega_4 \neq 0$
  $\phi_1 - \phi_3 \neq 0$ *and/or* $\phi_2 - \phi_3 \neq 0$ *and/or* $\phi_4 - \phi_3 \neq 0$

- The ***restricted* model implied by the null hypothesis H_0** is obtained by imposing on Model 4.3 (the unrestricted model) the coefficient restrictions specified by H_0.

  Model 4.3, the unrestricted model, is:

$$Y_i = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_1 IN1_i + \omega_2 IN2_i + \omega_4 IN4_i + u_i \qquad \textbf{(4.3)}$$

  The ***restricted* model** is obtained by setting $\omega_1 = 0$ and $\omega_2 = 0$ and $\omega_4 = 0$ in Model 4.3:

$$Y_i = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \; = \; \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad \textbf{(1)}$$

- The ***test statistic*** appropriate for this hypothesis test is a **Wald F-statistic**.

Compare Models 4.1, 4.2 and 4.3 – They are Observationally Equivalent

- The **population regression equation for Model 4.1** is:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \phi_1 IN1_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i \qquad \textbf{(4.1)}$$

Test for *industry effects* in Model 4.1: a joint F-test of

H_0: $\phi_2 = \phi_1$ *and* $\phi_3 = \phi_1$ *and* $\phi_4 = \phi_1$
$\phi_2 - \phi_1 = 0$ *and* $\phi_3 - \phi_1 = 0$ *and* $\phi_4 - \phi_1 = 0$

H_1: $\phi_2 \neq \phi_1$ *and/or* $\phi_3 \neq \phi_1$ *and/or* $\phi_4 \neq \phi_1$
$\phi_2 - \phi_1 \neq 0$ *and/or* $\phi_3 - \phi_1 \neq 0$ *and/or* $\phi_4 - \phi_1 \neq 0$

- The **population regression equation for Model 4.2** is:

$$Y_i = \phi_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \psi_2 IN2_i + \psi_3 IN3_i + \psi_4 IN4_i + u_i \qquad \textbf{(4.2)}$$

Test for *industry effects* in Model 4.2: a joint F-test of

H_0: $\psi_2 = 0$ *and* $\psi_3 = 0$ *and* $\psi_4 = 0$

H_1: $\psi_2 \neq 0$ *and/or* $\psi_3 \neq 0$ *and/or* $\psi_4 \neq 0$

- The **population regression equation for Model 4.3** is:

$$Y_i = \phi_3 + \beta_1 X_{i1} + \beta_2 X_{i2} + \omega_1 IN1_i + \omega_2 IN2_i + \omega_4 IN4_i + u_i \quad \textbf{(4.3)}$$

Test for *industry effects* in Model 4.3: a joint F-test of

H_0: $\omega_1 = 0$ *and* $\omega_2 = 0$ *and* $\omega_4 = 0$

H_1: $\omega_1 \neq 0$ *and/or* $\omega_2 \neq 0$ *and/or* $\omega_4 \neq 0$

- ***Result:*** These three F-tests for industry effects are ***identical***; they yield exactly the **same sample value F_0** of the general F-statistic, and hence **yield identical inferences** about the presence or absence of industry effects on the conditional mean value of Y for given values of X_1 and X_2.

❑ Model 5: Models with Several Discrete/Categorical Explanatory Variables

Consider a linear regression model in which **two or more explanatory variables** are ***discrete* or *categorical* variables**.

To illustrate, suppose the two discrete explanatory variables are ***gender*** and ***industry***.

- ***Gender*** can be represented by means of the following **two dummy variables**:

  $\mathbf{F_i}$ is a ***female* indicator (dummy) variable**, defined as follows:

  $F_i = 1$ if observation i is female, $= 0$ if observation i is not female.

  $\mathbf{M_i}$ is a ***male* indicator (dummy) variable**, defined as follows:

  $M_i = 1$ if observation i is male, $= 0$ if observation i is not male.

  Adding-Up Property of the Gender Indicator Variables $\mathbf{F_i}$ and $\mathbf{M_i}$

  $$F_i + M_i = 1 \qquad \forall\ i$$

- ***Industry*** can be represented by means of the following ***industry*** **dummy variables** (assuming a four-level categorization of the variable industry):

$IN1_i$ = 1 if observation i is in industry 1, = 0 otherwise.

$IN2_i$ = 1 if observation i is in industry 2, = 0 otherwise.

$IN3_i$ = 1 if observation i is in industry 3, = 0 otherwise.

$IN4_i$ = 1 if observation i is in industry 4, = 0 otherwise.

Adding-Up Property of the Industry Indicator Variables:

$$IN1_i + IN2_i + IN3_i + IN4_i = 1 \qquad \forall\ i$$

Model 1 -- The Benchmark Model

Contains two regressors in the two explanatory variables X_1 and X_2, both of which are assumed to be continuous variables.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad \textbf{(1)}$$

- The **population regression function**, or conditional mean function, **for Model 1** takes the form

$$E(Y_i \mid X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \qquad (1')$$

- Model 1 assumes that the **population regression function** is the same for all population members. For example, it allows no gender or industry differences in any of the regression coefficients β_j ($j = 0, 1, 2$).

Model 5.1 -- Version 1 of Model 5: No Gender or Industry Base Group

Allows for **different *male* and *female intercepts*** by introducing both the gender dummy variables F_i and M_i as additional additive regressors in Model 1.

Allows for **different *industry intercepts*** by introducing all four industry dummy variables $IN1_i$, $IN2_i$, $IN3_i$, and $IN4_i$ as additional additive regressors in Model 1.

- The **population regression equation for Model 5.1** is:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \theta_f F_i + \theta_m M_i + \phi_1 IN1_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i \qquad \textbf{(5.1)}$$

  The distinguishing characteristic of Model 5.1 is that it contains **no equation *intercept coefficient***. That is because there is **no base group** in Model 5.1 for either gender or industry.

- **Problem with Model 5.1**: It violates the **full rank assumption A5**. It exhibits ***perfect multicollinearity***.

  Reason:

  The two ***gender* dummy variables** by definition satisfy the adding-up property

  $$F_i + M_i = 1 \qquad \forall\ i$$

  The four ***industry* dummy variables** by definition satisfy the same adding-up property:

  $$IN1_i + IN2_i + IN3_i + IN4_i = 1 \qquad \forall\ i$$

- **Estimation Strategies for Model 5:** There are at least two alternative strategies that can be adopted to make Model 5 susceptible to estimation.

  Strategy 1: **Select a base group** for each of the categorical variables gender and industry, and reformulate Model 5 accordingly.

  Strategy 2: Introduce an equation intercept coefficient in regression equation 5.1, and use **restricted OLS estimation** to estimate the resulting equation subject to two linear coefficient restrictions: one on the coefficients of the gender dummy variables; and another on the coefficients of the industry dummy variables.

  Estimate by **restricted (constrained) OLS** the regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta_f F_i + \theta_m M_i + \phi_1 IN1_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i$$

  subject to the two linear coefficient restrictions

$$\theta_f + \theta_m = 0 \qquad \text{(c1)}$$

$$\phi_1 + \phi_2 + \phi_3 + \phi_4 = 0 \qquad \text{(c2)}$$

Model 5.2 -- Version 2 of Model 5: Base Groups for Gender and Industry

Derivation of Model 5.2

- **Select *males*** as the **base group for *gender***.

  Substitute for the male dummy variable M_i in equation (5.1) the equivalent expression

  $$M_i = 1 - F_i \qquad \forall\ i$$

- **Select *industry 1*** as the **base group for *industry***.

  Substitute for the industry 1 dummy variable $IN1_i$ in equation (5.1) the equivalent expression

  $$IN1_i = 1 - IN2_i - IN3_i - IN4_i \qquad \forall\ i$$

- Make these substitutions in regression equation (5.1):

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \theta_f F_i + \theta_m M_i + \phi_1 IN1_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i \qquad \textbf{(5.1)}$$

$$\begin{aligned} &= \beta_1 X_{i1} + \beta_2 X_{i2} + \theta_f F_i + \theta_m (1 - F_i) \\ &\qquad + \phi_1 (1 - IN2_i - IN3_i - IN4_i) + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i \end{aligned}$$

$$\begin{aligned} &= \beta_1 X_{i1} + \beta_2 X_{i2} + \theta_f F_i + \theta_m - \theta_m F_i \\ &\qquad + \phi_1 - \phi_1 IN2_i - \phi_1 IN3_i - \phi_1 IN4_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i \end{aligned}$$

Now **collect terms**: there are two constant terms, two terms in F_i, two terms in $IN2_i$, two terms in $IN3_i$, and two terms in $IN4_i$.

$$\begin{aligned} Y_i &= \beta_1 X_{i1} + \beta_2 X_{i2} + \theta_f F_i + \theta_m - \theta_m F_i \\ &\quad + \phi_1 - \phi_1 IN2_i - \phi_1 IN3_i - \phi_1 IN4_i + \phi_2 IN2_i + \phi_3 IN3_i + \phi_4 IN4_i + u_i \end{aligned}$$

$$\begin{aligned} &= (\theta_m + \phi_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + (\theta_f - \theta_m) F_i \\ &\quad + (\phi_2 - \phi_1) IN2_i + (\phi_3 - \phi_1) IN3_i + (\phi_4 - \phi_1) IN4_i + u_i \end{aligned} \qquad \textbf{(5.2)}$$

- Re-name some of the coefficients in regression equation (5.2). Define

$$\begin{aligned} \beta_0 &= \theta_m + \phi_1 \\ \lambda_f &= \theta_f - \theta_m \\ \pi_2 &= \phi_2 - \phi_1 \\ \pi_3 &= \phi_3 - \phi_1 \\ \pi_4 &= \phi_4 - \phi_1 \end{aligned}$$

- ***Result:*** The **population regression *equation* for Model 5.2**, equation (5.2), can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \lambda_f F_i + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i + u_i \qquad \textbf{(5.2)}$$

- ***Interpretation of the coefficients in Model 5.2***

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \lambda_f F_i + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i + u_i \qquad \textbf{(5.2)}$$

$\beta_0 = \theta_m + \phi_1$ = intercept for males in industry 1

$\lambda_f = \theta_f - \theta_m$ = female intercept – male intercept

$\pi_2 = \phi_2 - \phi_1$ = industry 2 intercept – industry 1 intercept

$\pi_3 = \phi_3 - \phi_1$ = industry 3 intercept – industry 1 intercept

$\pi_4 = \phi_4 - \phi_1$ = industry 4 intercept – industry 1 intercept

- ***Key Features of Model 5.2***

The omitted base group for gender is males, and for industry is industry 1.

The **male indicator variable $\mathbf{M_i}$** and the **industry 1 indicator variable $\mathbf{IN1_i}$** ***are excluded*** from the regressor set of Model 5.2.

Model 5.2 allows for both **different *male and female intercepts*** and **different *industry intercepts***.

Model 5.2 constrains the slope coefficients β_1 and β_2 on the continuous regressors X_{i1} and X_{i2} to be the same both **for *males and females*** and **for all four *industry groups***.

- The **population regression *function* for Model 5.2** is obtained by taking the conditional expectation of regression equation (5.2) for any given values of the regressors X_{i1}, X_{i2}, F_i, $IN2_i$, $IN3_i$, and $IN4_i$, and using the zero conditional mean error assumption $E(u_i \mid X_{i1}, X_{i2}, F_i, IN2_i, IN3_i, IN4_i) = 0$ for all i:

$$E(Y_i \mid X_{i1}, X_{i2}, F_i, IN2_i, IN3_i, IN4_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \lambda_f F_i + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i \qquad (5.2')$$

- The ***female* population regression function for Model 5.2** is obtained by setting the female indicator $F_i = 1$ in (5.2'):

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

$$= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \lambda_f + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i$$

$$= \beta_0 + \lambda_f + \beta_1 X_{i1} + \beta_2 X_{i2} + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i \qquad \mathbf{(5.2f)}$$

The female population regression function gives the ***female* conditional mean Y** value for *given* values of the regressors X_1, X_2, IN2, IN3, and IN4.

- The ***male* population regression function for Model 5.2** is obtained by setting the female indicator $F_i = 0$ in (5.2'):

$$E(Y_i \mid F_i = 0, X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i \qquad \mathbf{(5.2m)}$$

The male population regression function gives the ***male* conditional mean Y** value for *given* values of the regressors X_1, X_2, IN2, IN3, and IN4.

$$
\begin{aligned}
E(Y_i \mid F_i = 1, X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) \\
&= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \lambda_f + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i \\
&= \beta_0 + \lambda_f + \beta_1 X_{i1} + \beta_2 X_{i2} + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i \qquad \textbf{(5.2f)}
\end{aligned}
$$

$$
E(Y_i \mid F_i = 0, X_{i1}, X_{i2}, IN2_i, IN3_i, IN4_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i \qquad \textbf{(5.2m)}
$$

- Compare the ***female* and *male* population regression functions for Model 5.2**:

  Only the ***intercept coefficient* differs** between the male and female regression functions implied by Model 5.2.

  The ***slope coefficients* are all identical** in the male and female regression functions for Model 5.2.

- The ***female-male difference*** **in conditional mean Y for given values of the regressors** is obtained by subtracting the male population regression function (5.2m) from the female population regression function (5.2f):

  Define the 1×6 row vector $x_i^T = [X_{i1} \quad X_{i2} \quad IN2_i \quad IN3_i \quad IN4_i]$ containing the values of the regressors X_1, X_2, IN2, IN3, and IN4 for observation i.

  Then the ***difference*** between the ***female*** **conditional mean Y** for *given* values of the regressors X_1, X_2, IN2, IN3, and IN4 and the ***male*** **conditional mean Y** for the *same* values of the regressors X_1, X_2, IN2, IN3, and IN4 is:

$$
\begin{aligned}
& E(Y_i \mid F_i = 1, x_i^T) - E(Y_i \mid F_i = 0, x_i^T) \\
&\quad = \beta_0 + \lambda_f + \beta_1 X_{i1} + \beta_2 X_{i2} + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i \\
&\qquad - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i) \\
&\quad = \beta_0 + \lambda_f + \beta_1 X_{i1} + \beta_2 X_{i2} + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i \\
&\qquad - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \pi_2 IN2_i - \pi_3 IN3_i - \pi_4 IN4_i \\
&\quad = \lambda_f \qquad \textbf{(5.2*)}
\end{aligned}
$$

  Note: The ***female-male difference*** **in the conditional mean value of Y** for given values of the regressors X_{i1}, X_{i2}, $IN2_i$, $IN3_i$, and $IN4_i$ is ***a constant***; it does not depend on the values of the regressors X_1 and X_2 or on industry.

- ***Interpretation of the coefficients in Model 5.2***

Rewrite the **population regression *equation* for Model 5.2**:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \lambda_f F_i + \pi_2 IN2_i + \pi_3 IN3_i + \pi_4 IN4_i + u_i \qquad \textbf{(5.2)}$$

β_0 = intercept for ***males* in industry 1**

$\beta_0 + \lambda_f$ = intercept for ***females* in industry 1**

λ_f = ***female* industry 1** intercept – ***male* industry 1** intercept

$\beta_0 + \pi_2$ = intercept for ***males* in industry 2**

$\beta_0 + \lambda_f + \pi_2$ = intercept for ***females* in industry 2**

λ_f = ***female* industry 2** intercept – ***male* industry 2** intercept

$\beta_0 + \pi_3$ = intercept for ***males* in industry 3**

$\beta_0 + \lambda_f + \pi_3$ = intercept for ***females* in industry 3**

λ_f = ***female* industry 3** intercept – ***male* industry 3** intercept

$\beta_0 + \pi_4$ = intercept for ***males* in industry 4**

$\beta_0 + \lambda_f + \pi_4$ = intercept for ***females* in industry 4**

λ_f = ***female* industry 4** intercept – ***male* industry 4** intercept

π_2 = ***male*** **industry 2** intercept – ***male*** **industry 1** intercept
= ***female*** **industry 2** intercept – ***female*** **industry 1** intercept

π_3 = ***male*** **industry 3** intercept – ***male*** **industry 1** intercept
= ***female*** **industry 3** intercept – ***female*** **industry 1** intercept

π_4 = ***male*** **industry 4** intercept – ***male*** **industry 1** intercept
= ***female*** **industry 4** intercept – ***female*** **industry 1** intercept

Inter-industry differences in the conditional mean value of Y are ***equal*** **for** ***males*** **and** ***females***. The effects of industry on Y are identical for males and females in Model 5.2.