Qualitative Information

William Brasic

The University of Arizona

Quantitative Information

Definition 1: Quantitative Information

Quantitative information comes in the form of numerical values that can be measured and quantified.

- Discrete examples include:
  - Age
- Continuous examples include:
  - Income

Qualitative Information

Definition 2: Qualitative Information

Qualitative information comes in the form of non-numerical categories or labels.

- Types of qualitative data:
  - Binary data: Two categories, e.g., yes/no, true/false.
  - Ordinal data: Ordered categories, e.g., rating scales (poor, fair, good, excellent).

Indicator Variables

Definition 3: Indicator Variables

Indicator Variables equal one when a condition is met and zero otherwise.

- Also called dummy or binary variables.
- We often denote an indicator variable as $\mathbb{1}(Condition)$.
- Examples include:
  - $\mathbb{1}(Male)$.
  - $\mathbb{1}(Graduated\ College)$.
- We create indicator variables by one-hot-encoding.

Indicator Variables

Question 1

What is the average of an indicator variable?

Indicator Variables

Question 1

What is the average of an indicator variable?

Answer to Question 1

The average of an indicator variable represents the fraction of the sample satisfying the condition.

SLR with Indicators

Example 1: SLR with Indicators

Suppose we are interested in the model

$$\begin{aligned} income_i &= \beta_0 + \beta_1 \cdot \mathbb{1}(female_i) + u_i \\ \widehat{income}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \mathbb{1}(female_i) \\ &= 40000 - 10000 \cdot \mathbb{1}(female_i). \end{aligned}$$

When $\mathbb{1}(female_i)$ evaluates to:

- **False** (i.e., the individual is male), we get $\widehat{income}_i = \$40,000$ (so our regression line is horizontal at this level).
- **True** (i.e., the individual is female), we get $\widehat{income}_i = 40000 - 10000 = \$30,000$ (so our regression line is horizontal at this level).

MLR with Indicators

Example 2: MLR with Indicators

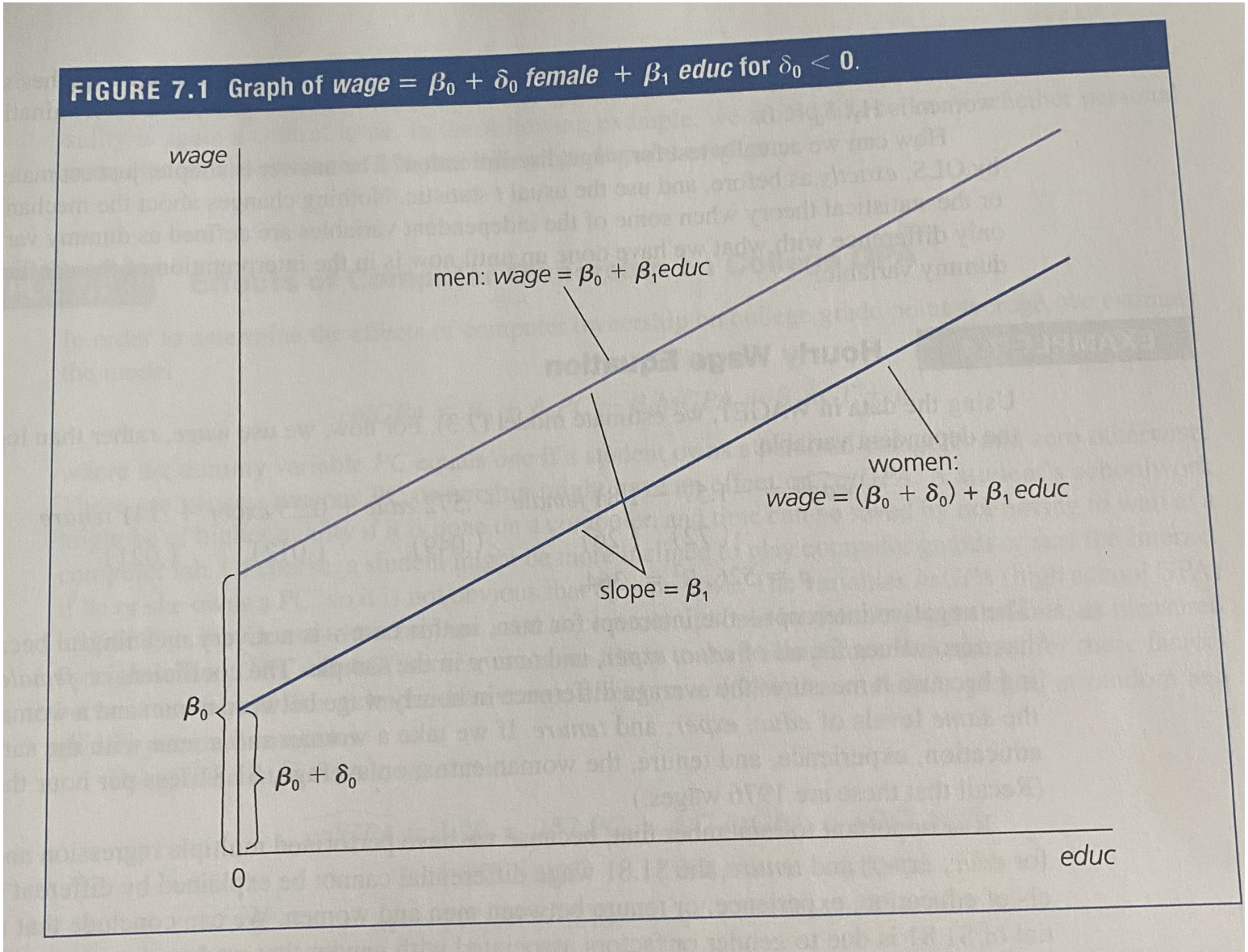
Suppose we are interested in the model

$$\begin{aligned} income_i &= \beta_0 + \beta_1 \cdot \mathbb{1}(female_i) + \beta_2 educ_i + u_i \\ \widehat{income}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \mathbb{1}(female_i) + \widehat{\beta}_2 educ_i \\ &= 40000 - 10000 \cdot \mathbb{1}(female_i) + 5000 educ_i. \end{aligned}$$

When $\mathbb{1}(female_i)$ evaluates to:

- **False** (i.e., the individual is male), we get $\widehat{income}_i = 40000 + 5000 educ_i$
- **True** (i.e., the individual is female), we get $\widehat{income}_i = 40000 - 10000 + 5000 educ_i = 30000 + 5000 educ_i$

MLR with Indicators



MLR with Indicators

Question 2

Why did we exclude males from the prior equation? Can't we include it and still obtain parameter estimates?

MLR with Indicators

Question 2

Why did we exclude males from the prior equation? Can't we include it and still obtain parameter estimates?

Answer to Question 2

No! Then, MLR Assumption 3 of No Perfect Multicollinearity fails. This means $X'X$ is no longer invertible so we can't obtain $\widehat{\beta}$.

- In this case, we must exclude males and they are called the *base case*.

MLR with Indicators

Example 3: MLR with Indicators

Suppose we collect data on $y_i = income_i$, $x_{i1} = \mathbb{1}(male_{i1})$, and $x_{i2} = \mathbb{1}(female_{i2})$, and $x_{i3} = age_{i3}$ for $n = 3$ observations. Our feature matrix could look like

$$X = \begin{bmatrix} 1 & 1 & 0 & 25 \\ 1 & 1 & 0 & 35 \\ 1 & 0 & 1 & 45 \end{bmatrix}.$$

Then, $\mathbb{1}(male_i) = 1 - \mathbb{1}(female_i)$ so these covariates are a perfect linear combination of each other. Consequently, including both in the regression will cause the columns of X to be linearly dependent, making $X'X$ non-invertible. That is why we a either exclude x_{i1} or x_{i2} from the regression, but never both.

MLR with Multiple Indicators

Answer to Question 1

Suppose we collect data on income and education level. In our survey we ask the participants to choose their education level from the following options:

1. Less than high school
2. High school or GED
3. Bachelors
4. Masters
5. Doctorate

Now we have multiple categories. How would we model the effect of education level on income?

MLR with Multiple Categories

Answer to Question 3

Let $\mathbb{1}(less\ than\ high\ school)$ be the base case (exclude it from our design matrix) and estimate the model

$$\begin{aligned} income_i = \beta_0 &+ \beta_1 \cdot \mathbb{1}(high\ school_i) \\ &+ \beta_2 \cdot \mathbb{1}(bachelors_i) + \beta_3 \cdot \mathbb{1}(masters_i) \\ &+ \beta_4 \cdot \mathbb{1}(doctorate_i) + u_i. \end{aligned}$$

Then, since there are no other covariates:

- β_0 represents the average income for someone with less than a high school education.
- β_1 represents the difference in average income between someone with a high school degree and someone with less than a high school education.
- ...

MLR with Multiple Categories

Example 4: MLR with Multiple Categories

Suppose upon estimation we get

$$\widehat{income}_i = 30000 + 20000 \cdot \mathbb{1}(high\ school_i) + 40000 \cdot \mathbb{1}(bachelors_i) + 50000 \cdot \mathbb{1}(masters_i) + 70000 \cdot \mathbb{1}(doctorate_i).$$

This means:

- The average income for someone with less than a high school degree is predicted to be $\$30,000$.
- The average income for someone with a doctorate degree is predicted to be $30000 + 70000 = \$100,000$.
- The difference in average income between someone having less than a high school degree and a masters degree is $\$50,000$.

Indicators and Interactions

Question 4

Suppose we wish to estimate the effect of experience on income by males and females. Stated differently, we want to determine if experience impacts an individual's income differently for males versus females. How would we model this in log-linear form?

Indicators and Interactions

Answer to Question 4

Use interactions! Our model would be

$$\ln(income_i) = \beta_0 + \beta_1 \cdot \mathbb{1}(female_i) + \beta_2 exper_i + \beta_3 \cdot \mathbb{1}(female_i) \times exper_i + u_i.$$

- β_0 represents intercept for males
- $\beta_0 + \beta_1$ represents the intercept for females
- β_2 represents the slope for males
- $\beta_2 + \beta_3$ represents the slope for females
- When β_1 is negative and β_3 is positive, regression line for females will intersect and surpass males at some point (draw picture).

Indicators and Interactions

Example 5: Indicators and Interactions

After estimating the prior model, we differentiate with respect to experience and get

$$\frac{\partial \ln\left(\widehat{income}\right)}{\partial exper} = \widehat{\beta}_1 + \widehat{\beta}_3 \cdot \mathbb{1}(female_i).$$

Thus, for a year increase in experience, income increases for:

- **Females** by $100\left(\widehat{\beta}_1 + \widehat{\beta}_3\right)\%$.
- **Males** is $100\left(\widehat{\beta}_1\right)\%$.

Hypothesis Testing with Indicators

Question 5

Continuing from the prior example:

1. What does the test of $H_0 : \beta_3 = 0$ versus $H_A : \beta_3 \neq 0$ represent and how do we carry it out?
2. What does the test of $H_0 : \beta_1 = \beta_3 = 0$ versus $H_A : \beta_1 \neq 0$ or $\beta_3 \neq 0$ represent and how do we carry it out?

Hypothesis Testing with Indicators

Answer to Question 5

1. The test of $H_0 : \beta_3 = 0$ versus $H_A : \beta_3 \neq 0$ represents testing if the effect of experience on income for males is the same as that for females. We carry out this test via a T-test.
2. The test of $H_0 : \beta_1 = \beta_3 = 0$ versus $H_A : \beta_1 \neq 0$ or $\beta_3 \neq 0$ represents testing if gender does not impact income. We carry out this test via a F-test.

Thank You!