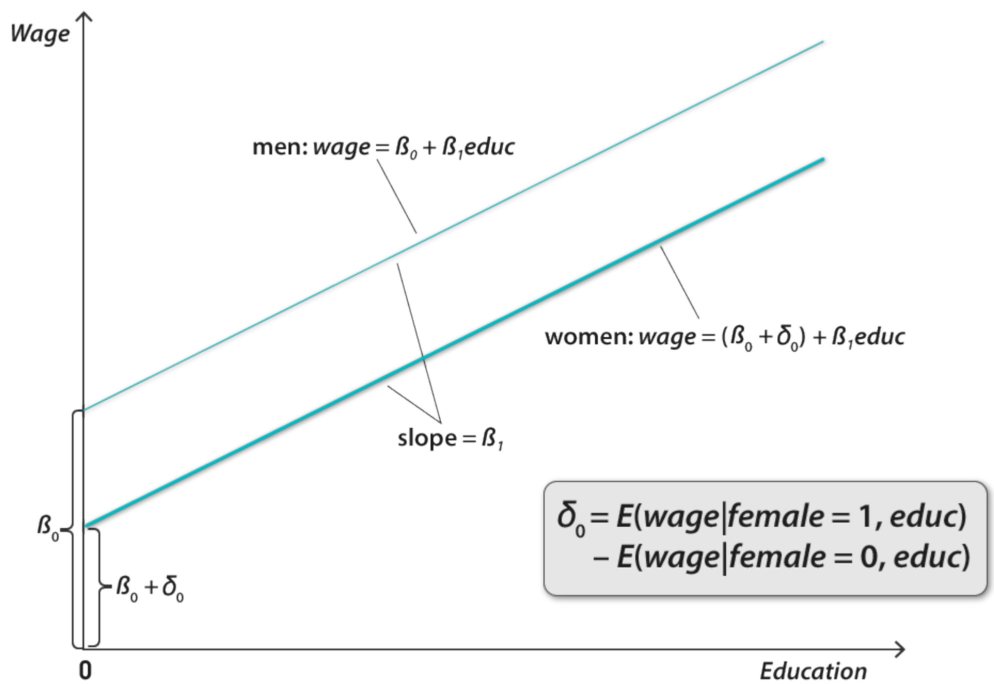
Qualitative Data

- Many variables are categorical in nature:
  - E.g., gender, race, industry, region, letter grade
  - Variables take on limited number of values, assigning each observation to a "category."
  - In experimental traditions and in R, these variables are called factors.
- Categorical variables may be **nominal** or **ordinal**.

Incorporating Qualitative Data

- To put categorical variables into regression model, we typically use indicator variables (**dummy variables**).
  - Value 1 ("true") is used for certain states and 0 ("false") otherwise.
- Example: population model that predicts wage as a function of education, with an indicator variable for female:
  - $\text{wage} = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + u$
  - For male subject with given value of education, expected wage will be $\beta_0 + \beta_1 \text{educ}$.
  - For female subject, expected wage will be $\beta_0 + \delta_0 + \beta_1 \text{educ}$.



Omitting Base Category

- Did not include indicator variables for both male and female
  - $\text{wage} = \beta_0 + \gamma_0 \text{male} + \delta_0 \text{female} + \beta_1 \text{educ} + u$
  - Wouldn't be able to estimate model because they'd be perfectly collinear
- Must omit one category: the **base category**
  - Could have chosen male or female, model would be equivalent
  - $\text{wage} = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + u$
  - $\text{wage} = \beta_0 + \gamma_0 \text{male} + \beta_1 \text{educ} + u$

Omitting Base Category (cont.)

- Could leave both categories in but omit intercept
  - $\text{wage} = \gamma_0 \text{ male} + \delta_0 \text{ female} + \beta_1 \text{ educ} + u$
  - Harder to test if categories are different
  - Usual formula for *R*-squared no longer valid

Interpreting Coefficients

Fitted wage equation including female indicator variable:

$$\begin{aligned} \hat{wage} = {} & \underset{(.025)}{-1.57} \underset{(.26)}{- 1.81 female} + \underset{(.049)}{.572 educ} \\ & + \underset{(.012)}{.025 exper} + \underset{(.021)}{.141 tenure} \\ n = {} & 526,\ R^2 = .364 \end{aligned}$$

- Holding education, experience, and tenure fixed, women earn $1.81 less per hour than men.

Comparing Group Means

- We may want to compare mean of a variable for two different groups.
  - Put indicator variable for one category in population model by itself.

$$\hat{wage} = \underset{(.21)}{7.10} - \underset{(.26)}{2.51} female$$
$$n = 526,\ R^2 = .116$$

- Not holding other factors constant, women earn \$2.51 less than men (i.e., difference between mean wage of men and women is \$2.51).
- *t*-statistic in this case is test of whether two group means are equal.

Treatment as a Dummy Variable

- Randomly assign subjects to control group and one or more treatment groups.
  - Have control group as base category, dummies for each treatment group.
- Example: clinical trial where subjects are randomly assigned to take new blood pressure medication or placebo
  - $\text{Blood pressure} = \beta_0 + \beta_1 \text{ medication} + u$
  - β_1 represents difference in blood pressure between treatment and control.
  - *t*-test would test hypothesis that treatment has no effect.

Ordinal Variables

How do we put ordinal variables into regression model?

- Generally wrong to place ordinal variable directly into population model
  - Would impose linear structure on variable
- Use indicator variables for each category, allowing effect of each one to vary independently
- Example: $\mathrm{MBR} = \beta_0 + \beta_1 \mathrm{CR} + \text{other factors}$
  - *MBR* = Municipal bond rate
  - *CR* = Credit rating from 0 to 4 (0 = worst, 4 = best)

Ordinal Variables (cont.)

- This specification not appropriate—credit rating only contains ordinal information
- Better way to incorporate information is to define dummies:
  - $\mathrm{MBR} = \beta_0 + \delta_1 \mathrm{CR}_1 + \delta_2 \mathrm{CR}_2 + \delta_3 \mathrm{CR}_3 + \delta_4 \mathrm{CR}_4 + \text{other factors}$