# IDS 702

## Categorical predictors

# Categorical variable terms

- Levels: values of a categorical variable
- Binary variable: categorical variable with only two levels
- Factor (in R): categorical variable that stores levels and labels
- Dummy variable: numeric variable that represents a categorical variable
- Reference/baseline level: value to which other values of categorical variable are compared (important for coefficient interpretation)

# Binary variable

- Example: binary variable to represent home ownership
- 2 levels: owns a home or does not own a home
- Dummy variable:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

# Categorical variables (>2 levels)

- A single dummy variable cannot represent all values
- We also cannot have a dummy variable for every level (coefficients cannot be estimated uniquely in this case)
- We need # levels -1 dummy variables for each categorical variable
- Example: region (East, West, South) (What's the reference category?)

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West.} \end{cases}$$

# Interpreting categorical variables

- Estimates are interpreted in relation to the reference category

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

# Nested F test / type III test

- We may want to assess the association between a categorical variable and the outcome
- Since we have dummy variables, this requires testing if a subset of the coefficients are equal to 0
- This is called a nested F test or Type III test
- The test compares a reduced model to the full model