Dummy Independent Variable Regression

- **Dummy variables** are variables with two (or more) possible values (usually 0 and 1) that indicate the absence (0) or presence (1) of a characteristic.
- They are used to represent categorical data in regression models.
- Intercept dummy variables only affect the y-intercept of the regression line.

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \epsilon \implies E(Y|D=0) = \beta_0 + \beta_1 X \text{ and } E(Y|D=1) = (\beta_0 + \beta_2) + \beta_1 X$$

- Slope dummy variables affect the slope of the regression line.

$$Y = \beta_0 + \beta_1 X + \beta_2 DX + \epsilon \implies E(Y|D=0) = \beta_0 + \beta_1 X \text{ and } E(Y|D=1) = \beta_0 + (\beta_1 + \beta_2) X$$

- To include categorical variables with k levels, create $k-1$ dummy variables.

Dummy (Dependent) Variable Regression

- A linear probability model (LPM) is a dummy variable regression where the dependent variable is linear in the parameters.

$$D = \beta_0 + \beta_1 X + \epsilon$$

- Conditional expectation of the dependent variable equals the conditional success probability.

$$E(D|X) = 1 \cdot P(D = 1|X) + 0 \cdot P(D = 0|X) = P(D = 1|X) = \beta_0 + \beta_1 X$$

- Errors is a binary random variable and heteroskedastic.

$$\epsilon = D - \beta_0 - \beta_1 X = \begin{cases} -\beta_0 - \beta_1 X & \text{if } D = 0 \\ 1 - \beta_0 - \beta_1 X & \text{if } D = 1 \end{cases} \quad Var(\epsilon|X) = P(1-P) = (\beta_0 + \beta_1 X)(1 - \beta_0 - \beta_1 X)$$