DUMMY/QUALITATIVE/BINARY VARIABLES

Example: race, sex, marital status, labor force participation, structural change => presence or absence of a quality.

- Qualitative explanatory variables
- Qualitative dependent variable

<u>Qualitative Explanatory Variables</u>:

Example: wage differential for male.

Quantify the qualitative variable 'gender', denoted by D_F, as follows:

D_F = 1 if the person is female
= 0 otherwise.

Why (1,0)? Interpretation of parameters in regression is very natural.

Consider the regression equation:

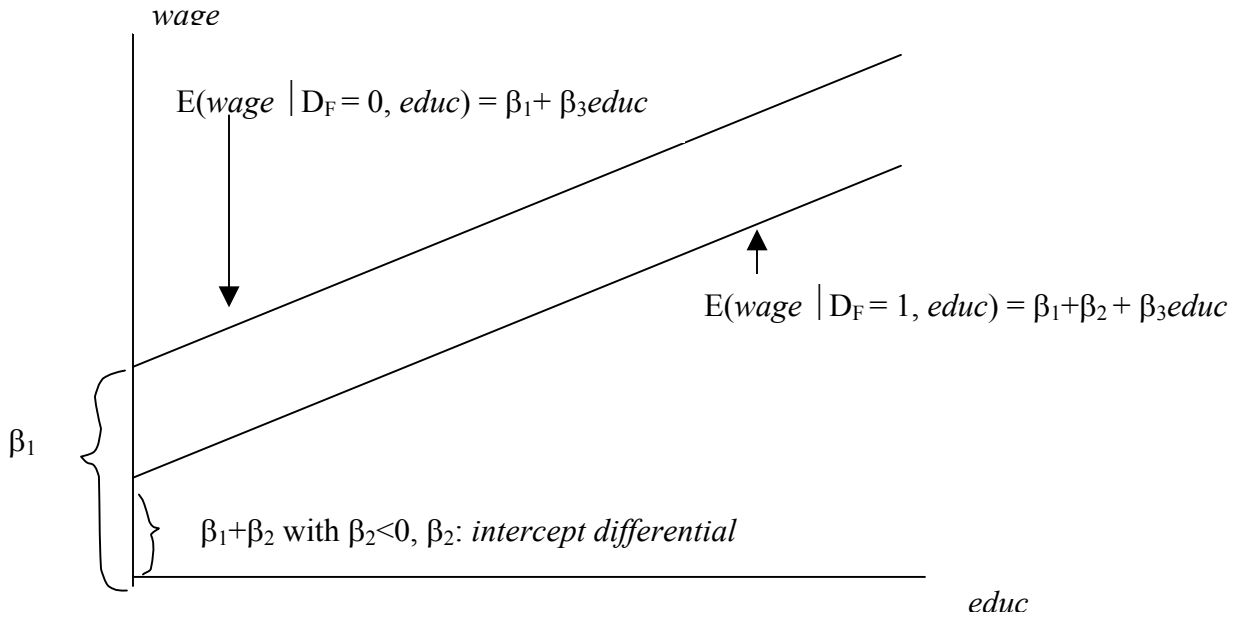
$wage = \beta_1 + \beta_2 D_F + \beta_3 educ + u$

=> β_2 is the mean difference in wage of females from that of males with the same education.

$E(wage \mid D_F = 1, educ) = \beta_1 + \beta_2 + \beta_3 educ$
$E(wage \mid D_F = 0, educ) = \beta_1 + \beta_3 educ$

⇨ $\beta_2 = E(\text{wage} \mid D_F = 1, \text{educ}) - E(\text{wage} \mid D_F = 0, \text{educ})$

Geometrically: => an **intercept shift**, *assuming slope unchanged*.



Note:

1. One dummy variable, D_F, for two categories: male and female. What if two variables for two categories? D*ummy variable trap* => perfect multicollinearity. Define

   D_M = 1 if the observation is male
   = 0 otherwise.

   and form the regression equation

   $wage = \beta_1 + \beta_2 D_F + \beta_4 D_M + \beta_3 educ + u$

   Note that $D_M + D_F = 1$ => a perfect linear relationship between regressors (D_M, D_F, the intercept term). Individual effects of D_M and D_F (viz., β_2 and β_4) cannot be identified/estimated

separately. In general, m-1 dummy variables for m categories (example: highest level of education dummy: not even high school diploma, high school, college, graduate degree, Ph.D).

2. It is up to you to decide for which category you define the dummy (to be =1). Depends on your focus. We could have worked with D_M. The interpretations change for the regression equation:

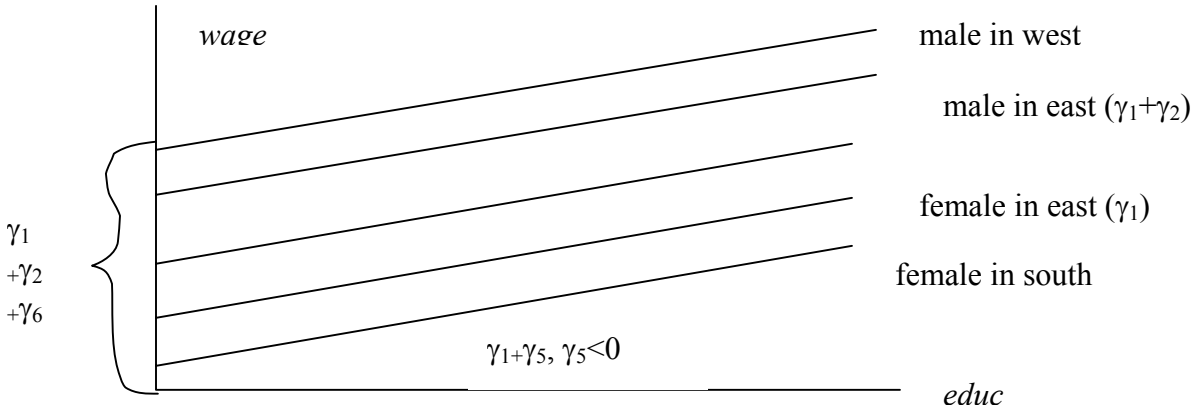
   $wage = \gamma_1 + \gamma_2 D_M + \gamma_3 educ + u$

   γ_2 is the difference in wage of males from that of females with the same education. But the ultimate results are identical. Note: $\gamma_1 = \beta_1 + \beta_2$; $\gamma_2 = -\beta_2$; $\gamma_3 = \beta_3$. (Try out the equation above in SAS and compare with the results of regression with D_F). In the current analysis the female group (for which dummy value = 0) is referred to as the *base, benchmark, reference, or control* group.

Dummy variables for more than two categories:

Example: 'region'

$$wage = \gamma_1 + \gamma_2 D_M + \gamma_3 educ + \gamma_4 northcen + \gamma_5 south + \gamma_6 west + u$$



Females in the eastern region constitute the reference group here. Male wage differential is γ_2 (across all regions). Wage differential (across gender) for the northcentral region relative to the eastern region is region is γ_4 ($\gamma_4<0$), the same for the south is γ_5 ($\gamma_5<0$), and for the west is γ_6 .

One can also obtain the wage differential of the south relative to:

- East (γ_5)
- West ($\gamma_5-\gamma_6$)
- Northcentral ($\gamma_5-\gamma_4$)

Interactions Effects:

Interaction between two dummy variables:

$wage = \beta_1 + \beta_2 D_F + \beta_3 educ + \beta_4 D_{NW} + \beta_5 D_F D_{NW} + u$

$D_{NW} = 1$ if the person is non-white
$= 0$ otherwise.

=> female wage differential ($\beta_2 + \beta_5 D_{NW}$) depends on race. And non-white wage differential ($\beta_4 + \beta_5 D_F$) depends on gender. White male is the reference group here. Can test the hypothesis average wages are identical for white and non-white with the same education.

Interaction between dummy variables and quantitative variables:

$wage = \beta_1 + \beta_2 D_F + \beta_3 educ + \beta_4 D_F \, educ + u$

- ⇨ effect of education on wage depends on gender. Now **slope w.r.t. education changes too** across gender.
- ⇨ Can test:
  - Return to education is the same for male and female. ($\beta_4=0$)

Comparing Regression Functions Across Groups/regimes using Dummy Variables: a (superior) Substitute of the Chow Test:

Objective: test whether the same regression equation (i.e., the same slope coefficients and intercept) explains the data from two groups of observations (e.g., pre and post war period, white and non-white).

$savings = \beta_1 + \beta_2 \, income + u$ (1)

Can do Chow-test by estimating two different regression equations. Better alternative using dummy variables. We can rewrite the equation above using dummy variables as follows to accommodate the possibility of different slope and intercept coefficients in the two periods:

$$savings = \beta_1 + \gamma_1 D2 + \beta_2\, income + \gamma_2\, income^{*}\, D2 + u \qquad (2)$$

where D2 = 1 if the observation is from pre-1955 years
= 0 otherwise.

Now we can answer the question above by testing whether H_0: $\gamma_1 = \gamma_2 = 0$ in (2) and we can carry out the test using our familiar F-test comparing restricted vs unrestricted regressions. Here our unrestricted equation is (2) and restricted equation is (1).

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/2}{(1 - R^2_{unrestricted})/(n-k)} = \frac{(.9526 - .9185)/2}{(1-.9526)/(18-4)} = \frac{.0171}{.0034} = 5.03$$

=> p-value (using $F_{2,\,14}$) =.02. So we can reject H_0.

We can also derive the regression equations for the two periods.

Note: implicit assumption: no heteroskedasticity or autocorrelation.