Multiple Regression Analysis with Qualitative Information: Dummy variables. Wooldridge (2013), Chapter 7.

- Introduction
- Program Evaluation
- Perfect Multicollinearity and the Dummy variable trap
- Dummies for Multiple Categories
- Interactions between dummy variables
- Other Interactions with Dummies
- Testing for Differences Across Groups
- Linear Probability Model

Introduction

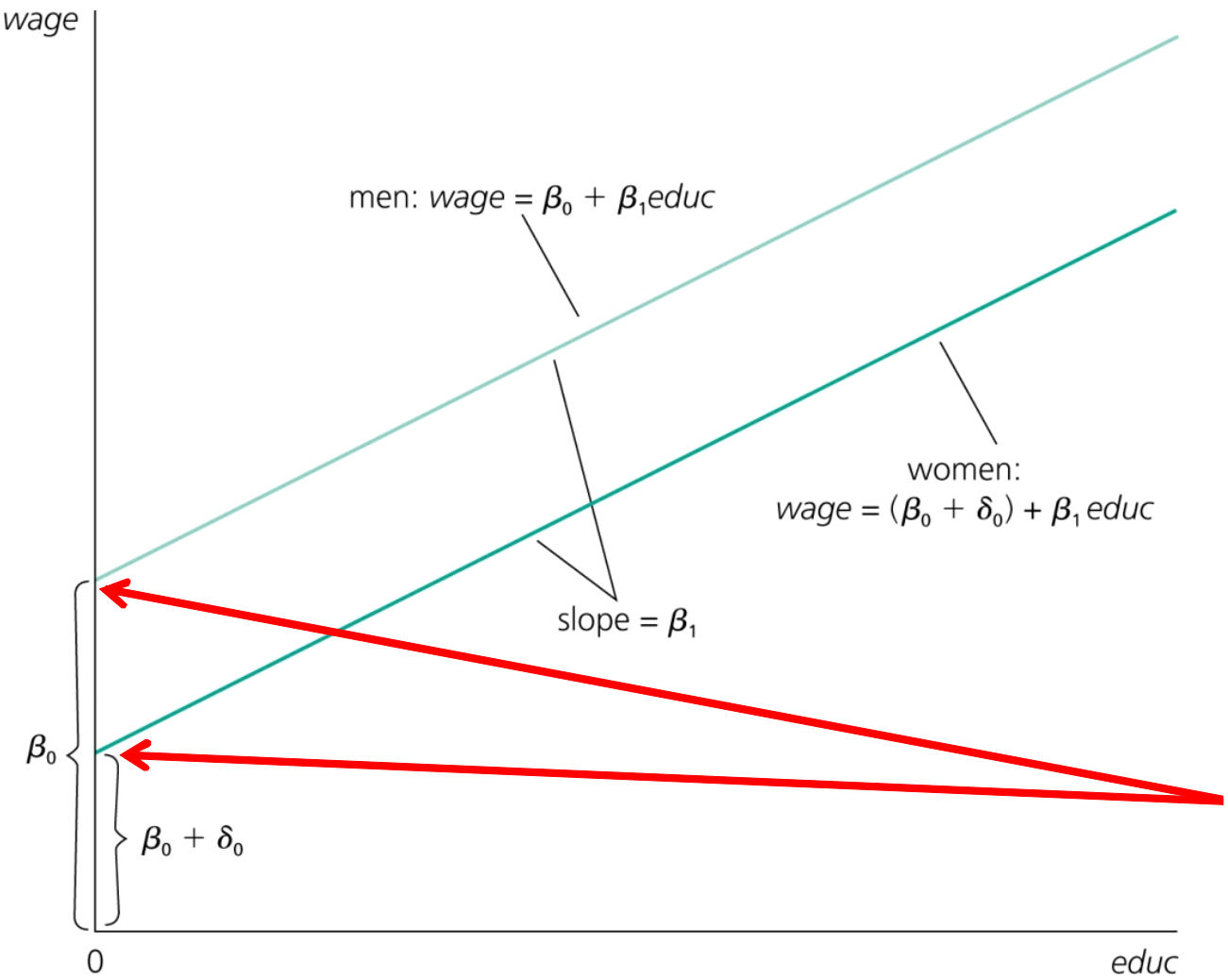
- A dummy variable (also known as an indicator variable or binary variable) is a variable that takes the values 0 or 1 and allows to take into account qualitative information in a regression model.
- Example of a dummy variable:

$$fem = \begin{cases} 1 & \text{if the individual is female} \\ 0 & \text{if the individual is male} \end{cases}$$

- We can use this dummy variable as a regressor to answer this question: do men earn more than women, after controlling for education?
- We can use the following model to answer the above question

$$\begin{aligned} wage &= \beta_0 + \beta_1 educ + \delta_0 fem + u, \\ E[u|fem] &= 0. \end{aligned}$$

Introduction



Intercept shift

$$\delta_0 < 0$$

Program Evaluation

- Dummy variables are widely used the *Program Evaluation Studies*.
- Suppose you wish to evaluate the effects of a "treatment" (for example, receiving training).
- One can define a dummy variable $T = 1$ if the individual has been treated and $T = 0$ if the individual has not been treated

$$y = \beta_0 + \beta_1 x + \delta_0 T + u,$$
$$E[u|x, T] = 0.$$

- Treated Group ($T = 1$) *vs* Control Group ($T = 0$).

Example: We would like to study if the average of hours of training per employee at the firm level are higher for companies that receive a job training grant. (**Data set:** Michigan manufacturing firms in 1988). Consider the model

$$hrsemp = \beta_0 + \delta_1 grant + \beta_1 \log(sales) + \beta_2 \log(employ) + u,$$

where

- $hrsemp$ = average hours of training per employee at the firm level.
- $grant = 1$ if the firm received a job training grant for 1988 and 0 otherwise.
- $sales$ =annual sales.
- $employ$ =number of employees.
- $n = 105$

The estimated regression line is given by

$$\widehat{hrsemp} = \underset{(43.41)}{46.67} + \underset{(5.59)}{26.25} grant + \underset{(3.54)}{0.98} \log(sales) + \underset{(3.88)}{6.07} \log(employ).$$

Perfect Multicollinearity and the Dummy variable trap

So far we have excluded the case that two or more explanatory variables are perfectly related (perfect collinearity is ruled out)
Perfect multicollinearity means an exact linear relationship between variables in the linear regression.
Example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
If x_1 is perfectly correlated with x_2 (say $x_1 = 2x_2$) then the OLS estimator cannot be computed.
If some of the regressors are dummy variables and we are not careful defining our regression model we might fall into the dummy variable trap (i.e. we might have perfect multicollinearity in our model).

Example of the Dummy variable trap: Define

$$fem = \begin{cases} 1 & \text{if the individual is female} \\ 0 & \text{if the individual is male} \end{cases},$$

$$male = \begin{cases} 1 & \text{if the individual is male} \\ 0 & \text{if the individual is female} \end{cases}$$

and consider the regression

$$wage = \beta_0 + \beta_1 fem + \beta_2 male + u.$$

Here we have perfect multicollinearity as $fem + male = 1$
Solution: Drop *fem* or *male*

Dummies for Multiple Categories

- Any categorical variable can be turned into a set of dummy variables.
- Because the base group is represented by the intercept, if there are m categories there should be $m - 1$ dummy variables (otherwise we fall in the dummy variable trap).
- If there are a lot of categories, it may make sense to group some together

Dummies for Multiple Categories

Example: Two Labour Economists, Hamermesh and Biddle, used measures of physical attractiveness in a wage equation. Each person in the sample was ranked by an interviewer for physical attractiveness.

Dummies for Multiple Categories

They estimated the following equations for men:

$$\begin{aligned} \widehat{\log(wage)} &= \hat{\beta}_0 - \underset{(0.046)}{0.164}\,belavg + \underset{(0.033)}{0.016}\,abvavg + \text{other factors} \\ n &= 700, \bar{R} = 0.403. \end{aligned}$$

The equation for women is:

$$\begin{aligned} \widehat{\log(wage)} &= \hat{\beta}_0 - \underset{(0.066)}{0.124}\,belavg + \underset{(0.049)}{0.035}\,abvavg + \text{other factors} \\ n &= 409, \bar{R} = 0.330. \end{aligned}$$

where
$belavg = 1$ if the person is below average, 0 otherwise
$abvavg = 1$ if the person is above average, 0 otherwise
base group: average

Interactions between dummy variables

Consider the model

$$y = \beta_0 + \delta_1 d_1 + \delta_2 d_2 + u,$$
$$E(u|d_1, d_2) = 0$$

- d_1 and d_2 are dummy variables.
- δ_1 is the effect of changing $d_1 = 0$ to $d_1 = 1$. In this specification this effect does not depend on the value of d_2,

$$\delta_1 = E\left[y|d_1 = 1, d_2 = D_2\right] - E\left[y|d_1 = 0, d_2 = D_2\right]$$

- To allow the effect of changing d_1 to depend on d_2, include the "interaction term" $d_1 \times d_2$

$$y = \beta_0 + \delta_1 d_1 + \delta_2 d_2 + \delta_3 d_1 \times d_2 + u$$
$$E(u|d_1, d_2) = 0$$

Interactions between dummy variables

$$y = \beta_0 + \delta_1 d_1 + \delta_2 d_2 + \delta_3 d_1 \times d_2 + u \tag{1}$$

In this case

$$E\left[y|d_1 = 1, d_2 = D_2\right] - E\left[y|d_1 = 0, d_2 = D_2\right] = \delta_1 + \delta_3 D_2$$

Other model that allows for interactions is

$$\begin{gathered} y = \beta_0 + \delta_1^* d_1 \times d_2 + \delta_2^*(1 - d_1) \times d_2 \\ + \delta_3^* d_1 \times (1 - d_2) + u, \\ E(u|d_1, d_2) = 0 \end{gathered} \tag{2}$$

Model (2) corresponds to a reparametrization of (1) as it can be shown that

$$\delta_1 = \delta_3^*,\ \delta_2 = \delta_2^* \text{ and } \delta_3 = \delta_1^* - \delta_2^* - \delta_3^*.$$

Model (2) might be easier to interpret as the following example shows.

Interactions between dummy variables

- **Example:** Suppose we are studying the Log-Hourly wage equation and you wish to allow the effect of being married to be different across men and women.
- Define the following dummy variables:

$$male = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases},$$

$$marr = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise} \end{cases}.$$

Interactions between dummy variables

Example: Consider the model

$$\begin{aligned}\log(wage) = \; & \beta_0 + \delta_0 marr \times male + \delta_1 marr \times (1 - male) \\ & + \delta_2 (1 - marr) \times (1 - male) \\ & + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \\ & \beta_5 tenure^2 + u,\end{aligned}$$

where
wage =average hourly earnings,
educ =years of education,
exper =years of potential experience,
tenure =years with current employer.
Base group: men not married

Interactions between dummy variables

Example:

$$\begin{aligned}\log(wage) = \; & \beta_0 + \delta_0 marr \times male + \delta_1 marr \times (1 - male) \\ & + \delta_2 (1 - marr) \times (1 - male) \\ & + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \\ & \beta_5 tenure^2 + u,\end{aligned}$$

Groups		Dummy	Intercept of the model
male	married	$marr \times male$	$\beta_0 + \delta_0$
female	married	$marr \times (1 - male)$	$\beta_0 + \delta_1$
female	not married	$(1 - marr) \times (1 - male)$	$\beta_0 + \delta_2$
male	not married	base group	β_0

Interactions between dummy variables

Example : Data Set: 1976 US Current Population Survey.
The estimated regression function is

$$\begin{aligned}
\widehat{\log(wage)} &= \underset{(0.100)}{0.321} + \underset{(0.055)}{0.213}marr \times male - \underset{(0.058)}{0.198}marr \times (1 - male) \\
&\quad - \underset{(0.056)}{0.110}(1 - marr) \times (1 - male) \\
&\quad + \underset{(0.007)}{0.079}educ + \underset{(0.005)}{0.027}exper - \underset{(0.00011)}{0.00054}exper^2 + \underset{(0.007)}{0.029}tenure + \\
&\quad - \underset{(0.00023)}{0.00053}tenure^2 \\
n &= 526, R^2 = 0.461.
\end{aligned}$$

Holding other things fixed, married women earn 19.8% less than not married men (= the base category)

Interactions between dummy variables

Remark: If we change the base group the estimates for the coefficients of the dummy variables will be different as the following example shows:

Example (cont):

$$\widehat{\log(wage)} = \underset{(0.016)}{0.123} + \underset{(0.046)}{0.411} marr \times male + \underset{(0.058)}{0.198}(1 - marr) \times male + \underset{(0.052)}{0.088}(1 - marr) \times (1 - male) + ...$$

Holding other things fixed, single men earn 19.8% more than married women (= the base category)

Other Interactions with Dummies

Can also consider interacting a dummy variable, d, with a continuous variable, x (slope shift):

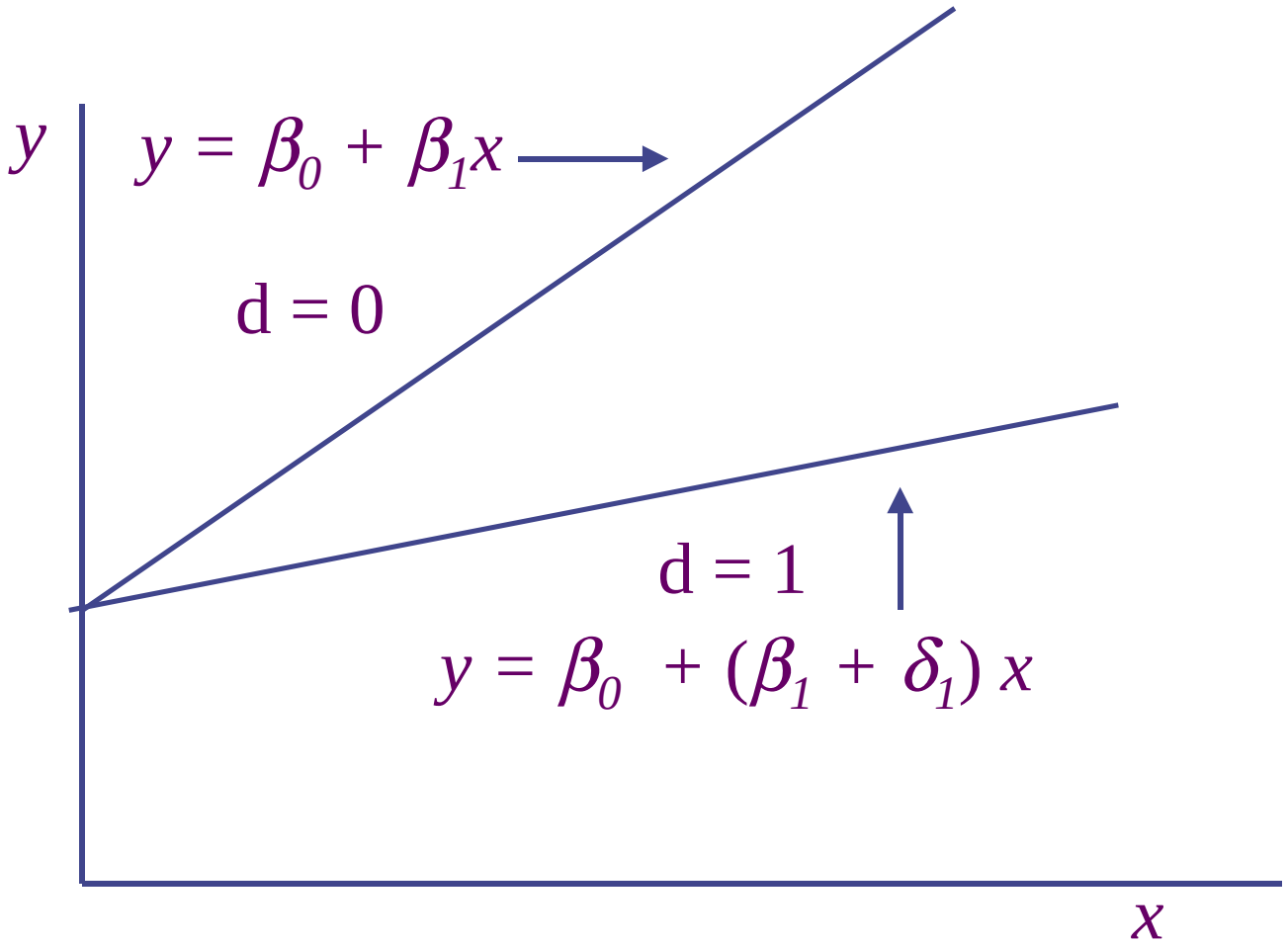
$$y = \beta_0 + \beta_1 x + \delta_1 d \times x + u$$

In this model :

d	Intercept	Slope
0	β_0	β_1
1	β_0	$\beta_1 + \delta_1$

Other Interactions with Dummies

Example of $\delta_1 < 0$



Other Interactions with Dummies

Can also consider a slope shift and an intercept shift:

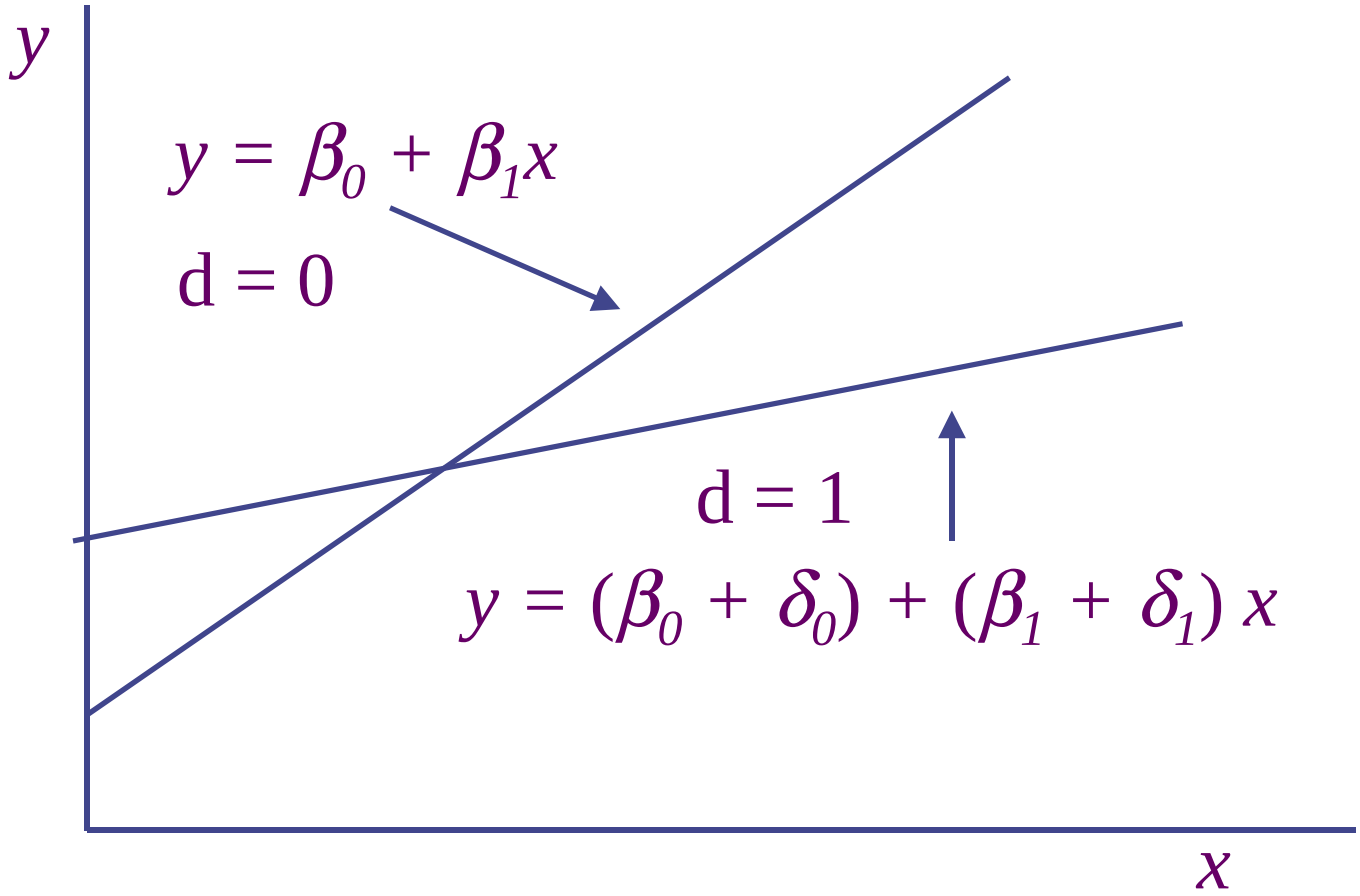
$$y = \beta_0 + \delta_0 d + \beta_1 x + \delta_1 d \times x + u$$

In this model :

d	Intercept	Slope
0	β_0	β_1
1	$\beta_0 + \delta_0$	$\beta_1 + \delta_1$

Other Interactions with Dummies

Example of $\delta_0 > 0$ and $\delta_1 < 0$



Other Interactions with Dummies

Example: Consider the model

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 female + \delta_1 female \times educ + \beta_1 educ + \beta_2 exper \\ &\quad + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u\end{aligned}$$

where
wage =average hourly earnings
female = 1 if female, 0 otherwise
educ =years of education
exper =years of potential experience
tenure =years with current employer.

Other Interactions with Dummies

Consider the following 2 estimated models

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(0.11869)}{0.38881} - \underset{(0.16754)}{0.22679}female - \underset{(0.01306)}{0.00556}female \times educ \\ &\quad + \underset{(0.00847)}{0.08237}educ + \underset{(0.00498)}{0.02934}exper - \underset{(0.00011)}{0.00058}exper^2 \\ &\quad + \underset{(0.00686)}{0.0319}\, tenure - \underset{(0.00024)}{0.00059}tenure^2, \\ n &= 526,\ R^2 = 0.441,\end{aligned}$$

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(0.10147)}{0.20157} + \underset{(0.00716)}{0.08453}educ + \underset{(0.00529)}{0.0293}\, exper \\ &\quad - \underset{(0.00011)}{0.00059}exper^2 + \underset{(0.00724)}{0.03712}tenure - \underset{(0.00025)}{0.00062}tenure^2, \\ n &= 526,\ R^2 = 0.3669.\end{aligned}$$

Test whether the variable female affects the conditional mean of $\log(wage)$ at 5% level.

Testing for Differences Across Groups

Consider the multiple regression model $y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + u$.

- **Objective:** Testing whether a regression function is different for a group versus another. Denote one group as A and the other as B. These groups are mutually exclusive and exhaustive.
- Denote $d = 1$ if the individual is in group A and zero if she is group B.

Our hypothesis can be tested as follows:

- Consider the more general model

$$y = \beta_0 + \delta_0 d + \sum_{i=1}^{k} (\beta_i x_i + \delta_i x_i d) + v$$

- Our null hypothesis can be thought of as simply testing for the joint significance of the dummy and its interactions with all other x variables: $H_0 : \delta_0 = \delta_1 = ... = \delta_k = 0$.
- **Example:** Suppose that we would like to test if the parameters of the model

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

are equal for males and females.

Testing for Differences Across Groups

- To test this we can consider the model

$$\log(wage) = \beta_0 + \beta_1 educ + \delta_0 female + \delta_1 female \times educ + v,$$

  where v is the error term and test $H_0 : \delta_0 = \delta_1 = 0$.
- So, you can estimate the model with all the interactions and without and form an F statistic:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/df},$$

  where SSR_r is the sum of squared residuals of the restricted model and SSR_{ur} is the sum of squared residuals of the unrestricted model, and df are the degrees of freedom of the model (sample size-number of parameters of the model) and q is the number of restrictions.
- However, this is equivalent to using a test known as *Chow Test*.

- **Key-idea:** You can compute the F statistic without running the unrestricted model with interactions with all k continuous variables.
- If we run the restricted model for group A and get SSR_A, then we run for group B and get SSR_B.
- Run the restricted model for all to get SSR, then

$$F_{Chow} = \frac{[SSR - (SSR_A + SSR_B)]}{SSR_A + SSR_B} \times \frac{[n - 2(k+1)]}{k+1}$$

Consider the model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

- The Chow test is the F test for exclusion restrictions described above.
- Recall that the usual F test in this setup is given by

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/df}.$$

- It can be shown that $SSR_{ur} = SSR_A + SSR_B$, also $SSR_r = SSR$.
- We have $q = k + 1$ restrictions (each of the slope coefficients and the intercept).
- The unrestricted model would estimate 2 different intercepts and 2 different slope coefficients, so $df = n - 2k - 2$.
- Note that $F_{Chow} \sim F(k+1, n-2k-2)$

Example: Consider the following regression lines.

- Regression with dummy variable

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(0.11685)}{0.82595} + \underset{(0.00899)}{0.07723}educ - \underset{(0.20325)}{0.36006}female \\ &\quad \underset{(0.01626)}{-0.00006}female \times educ, \\ SSR_{ur} &= 103.7983, n = 526\end{aligned}$$

- Regression without dummy variable

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(0.09823)}{0.58377} + \underset{(0.00774)}{0.08274}educ, \\ SSR_r &= 120.769, n = 526.\end{aligned}$$

- Regression for females

$$\begin{aligned} \widehat{\log(wage)} &= \underset{(0.16633)}{0.46589} + \underset{(0.01355)}{0.07716}educ, \\ SSR_A &= 40.3962, n_1 = 252. \end{aligned}$$

- Regression for males

$$\begin{aligned} \widehat{\log(wage)} &= \underset{(0.11684)}{0.82595} + \underset{(0.00899)}{0.07723}educ, \\ SSR_B &= 63.4021, n_2 = 274. \end{aligned}$$

- Test whether a regression function is different for a group versus another at 5% significance level using the F test for exclusion restrictions and the Chow test. Do the results differ?

Linear Probability Model

- Now the dependent variable y is a binary variable, that is it takes values 0 and 1.

Examples:

- Labour force participations.

$$y = \begin{cases} 1 & \text{if employed} \\ 0 & \text{otherwise} \end{cases}.$$

  We would like to study how labour force participation depends on the characteristics of the individuals.

- Financial crises

$$y = \begin{cases} 1 & \text{if the country is in a financial crisis} \\ 0 & \text{otherwise} \end{cases}.$$

  We would like to study the occurrence of a financial crisis depends on the characteristics of countries.

- Denote $\mathbf{x} = (x_1, ..., x_k)$.
- The objective of a regression model is to estimate $E(y|\mathbf{x})$.

Linear Probability Model

- $E(y|\mathbf{x}) = \mathcal{P}(y = 1|\mathbf{x})$, when y is a binary variable.
- In the *linear probability model* we assume that

$$\mathcal{P}(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

- So, the interpretation of β_j is the change in the probability of success when x_j changes:

$$\frac{\partial \mathcal{P}(y = 1|x)}{\partial x_j} = \beta_j, j = 1, ..., k$$

- The predicted y is the predicted probability of success.
- The linear probability model is estimated using OLS, that is regressing y on $x_1, ..., x_k$.

Example: Consider the model

$$\begin{aligned} inlf &= \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 \\ &\quad + \beta_5 age + \beta_6 kidslt6 + \beta_7 kidsge6 + u \end{aligned}$$

where

- *inlf* ("in the labour force") = binary variable indicating labour force participation by a married woman during 1975:
- *nwifeinc*=husband's earnings (, measured in thousands of dollars),
- *educ*= years of education,
- *exper*=past years of labor market experience,
- *age*=age of the woman,
- *kidslt6*=number of children less than six years old,
- *kidsge6*= number of kids between 6 and 18 years of age .

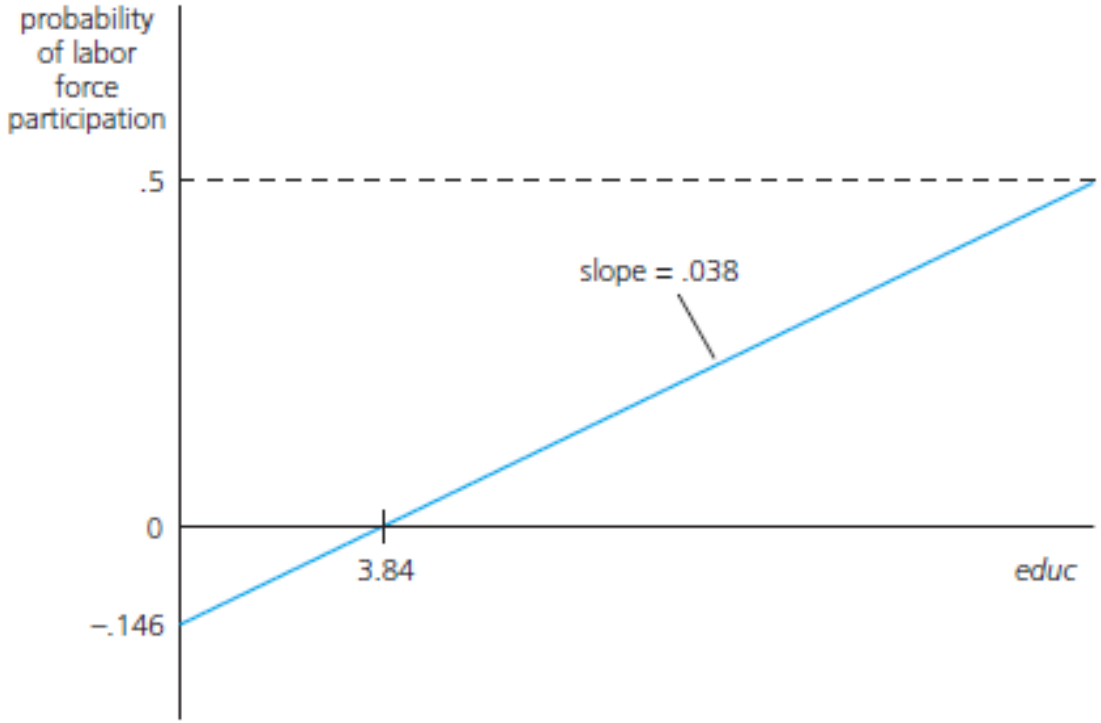
Linear Probability Model

Estimating the model we obtain:

$$\begin{aligned} \widehat{inlf} &= \underset{(0.154)}{0.586} - \underset{(0.0014)}{0.0034}nwifeinc + \underset{(0.007)}{0.038}educ + \underset{(0.006)}{0.039}exper - \underset{(0.00018)}{0.0006}exper^2 \\ &\quad - \underset{(0.002)}{0.016}age - \underset{(0.34)}{0.262}kidslt6 + \underset{(0.013)}{0.013}kidsge6, \\ n &= 743, R^2 = 0.264 \end{aligned}$$

Linear Probability Model

Graph for $nwifeinc = 50, exper = 5, age = 30, kindslt6 = 1, kidsge6 = 0$



- The maximum level of education in the sample is $educ = 17$. For the given case, this leads to a predicted probability to be in the labor force of about 0.5.
- For $educ < 3.84$ there is a negative predicted probability but no problem because no woman in the sample has $educ < 5$.

Linear Probability Model

Disadvantages of the linear probability model

- Potential problem that the fitted values can be outside $[0, 1]$.
- Even without predictions outside of $[0, 1]$, we may estimate effects that imply a change in x changes the probability by more than $+1$ or -1.
- This model will violate assumption of homoskedasticity, so will affect inference. Notice that

$$\begin{aligned} Var(y|\mathbf{x}) &= \mathcal{P}(y = 1|\mathbf{x})(1 - \mathcal{P}(y = 1|\mathbf{x})) \\ &= (\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \times \\ &\quad (1 - \beta_0 - \beta_1 x_1 - \ldots - \beta_k x_k). \end{aligned}$$

  Heteroskedasticity consistent standard errors need to be computed

Advantanges of the linear probability model

- Easy estimation and interpretation
- Estimated predictions often reasonably good in practice