Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- 5. Dummy Variables

Dummy Variables

- So far, the dependent and independent variables in our multiple regression models have had a *quantitative* meaning.

Quantitative Information

Table 1.3

Minimum Wage, Unemployment, and related data for Puerto Rico

obsno	*year*	*avgmin*	*avgcov*	*unemp*	*gnp*
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Dummy Variables

- In empirical work, we must also incorporate *qualitative* factors into regression models.
- *Qualitative factors* often, but not always, come in the form of *binary information*, i.e. a person is female or male, is either married or not.

Qualitative Information

Table 1.1

A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Describing Qualitative Information

- When *qualitative factors* come in the form of *binary information* the relevant information can be captured by defining a **binary variable** or a zero-one variable.
- In econometrics, binary variables are most commonly called **dummy variables**.
- In defining a dummy variable, we must decide which event is assigned the value one and which is assigned the value zero.

Describing Qualitative Information

- **Example 1**: In a gender case we can define

  female = 1 if the person is female,

  female = 0 if the person is male.

  Of course we can also define:

  male = 0 if the person is female,

  male = 1 if the person is male.

  But note that both variables, *female* and *male*, convey the same information.

Describing Qualitative Information

- **Example 2**: In a marital status case we can define

  married = 1 if the person is married,

  married = 0 if the person is not married.

- Using this trick we can incorporate *qualitative information* in our regression models.

Describing Qualitative Information

Table 1.1

A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Describing Qualitative Information

- Why do we use the values zero and one to describe qualitative information?
- In a sense, these values are arbitrary: any two different values would do the job.
- The real benefit of capturing qualitative information using zero-one variables is that it leads to regression models where the parameters have very natural interpretations.

Single Dummy Independent Variable

- How do we incorporate binary information into regression models?
- In the simplest case, with only a single dummy explanatory variable, we just add it as an independent variable in the equation.
- **Example**: Consider de simple model of hourly wage determination

$$wage = \beta_0 + \beta_1 educ + u$$

Single Dummy Independent Variable

- To measure gender wage discrimination we can just simply introduce a dummy variable for gender, for example the *female* variable defined above,

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

- What is the **interpretation** of δ_0? δ_0 is the difference in hourly wage between females and males, *given* the same amount of education (and the same error term *u*).

Single Dummy Independent Variable

- Thus, the coefficient δ_0 determines whether there is discrimination against women: If $\delta_0 < 0$ then, for the same level of other factors, women earn less than men on average.
- In terms of expectations, if we assume the zero conditional mean assumption $E(u|\, female, educ) = 0$, then

$$\delta_0 = E(wage|\, female = 1, educ) - E(wage|\, female = 0, educ)$$

Single Dummy Independent Variable

- Since *female* = 1 corresponds to females and *female* = 0 corresponds to males, we can write this more simply as

$$\delta_0 = E(wage|\, female, educ) - E(wage|male, educ)$$

- The key here is that the level of education is the same in both expectations; the difference, δ_0, is due to gender only.

Single Dummy Independent Variable

- Since *female* = 1 corresponds to females and *female* = 0 corresponds to males, we can write two models, one for females

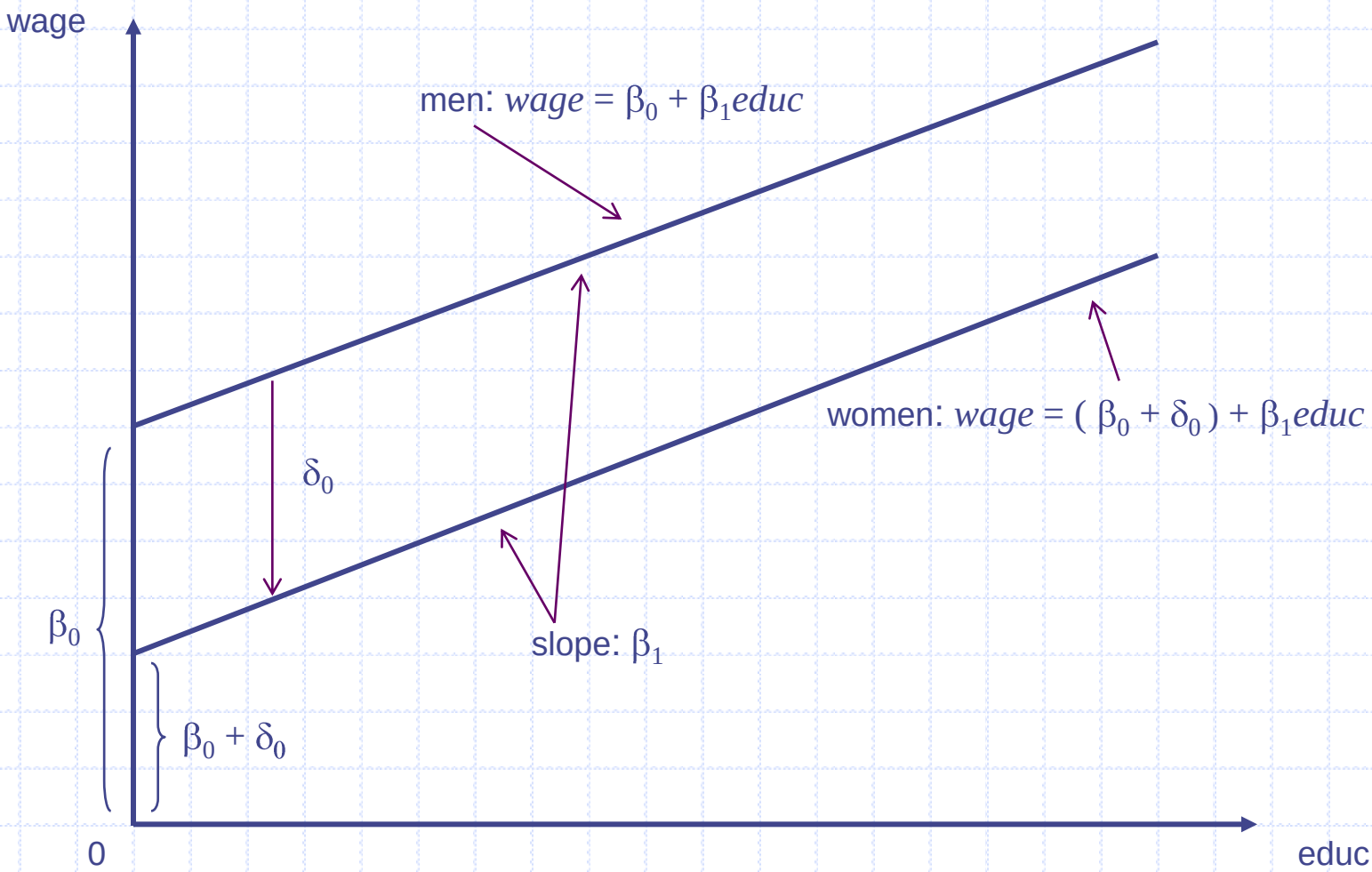
$$wage = \beta_0 + \delta_0 + \beta_1 educ + u$$

and other for males

$$wage = \beta_0 + \beta_1 educ + u$$

- Hence the situation can be depicted graphically as an **intercept shift** between males and females.

Example of $\delta_0 < 0$



Single Dummy Independent Variable

- Why we do not also include a dummy variable, say *male*, in the above equation? So we have,

$$wage = \beta_0 + \delta_0 female + \gamma_0 male + \beta_1 educ + u$$

- The answer is simple: this would be redundant.
- In the original equation the intercept for males is β_0, and the intercept for females is $\beta_0 + \delta_0$.

Single Dummy Independent Variable

- Since there are just two categories, we only need two different intercepts. This means that, in addition to β_0, we need to use only *one* dummy variable; we have chosen to include the dummy variables for females.
- Using two dummy variables would introduce perfect collinearity because *female* + *male* = 1, which means that *male* is a perfect linear function of *female*.

Single Dummy Independent Variable

- Including dummy variables for both genders is the simplest example of the so-called **dummy variable trap**, which arises when a dummy variable for each category is introduced in an equation, in addition to the intercept.
- In the previous example we have chosen to introduce *female*, which makes males to be the **base group** or **benchmark group**, that is, the group against which comparisons are made.

Single Dummy Independent Variable

- This is why β_0 is the intercept for males, and δ_0 is the *difference* in intercepts between females and males.
- What happen if we use *male* instead of *female* in the wage equation?

$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u$$

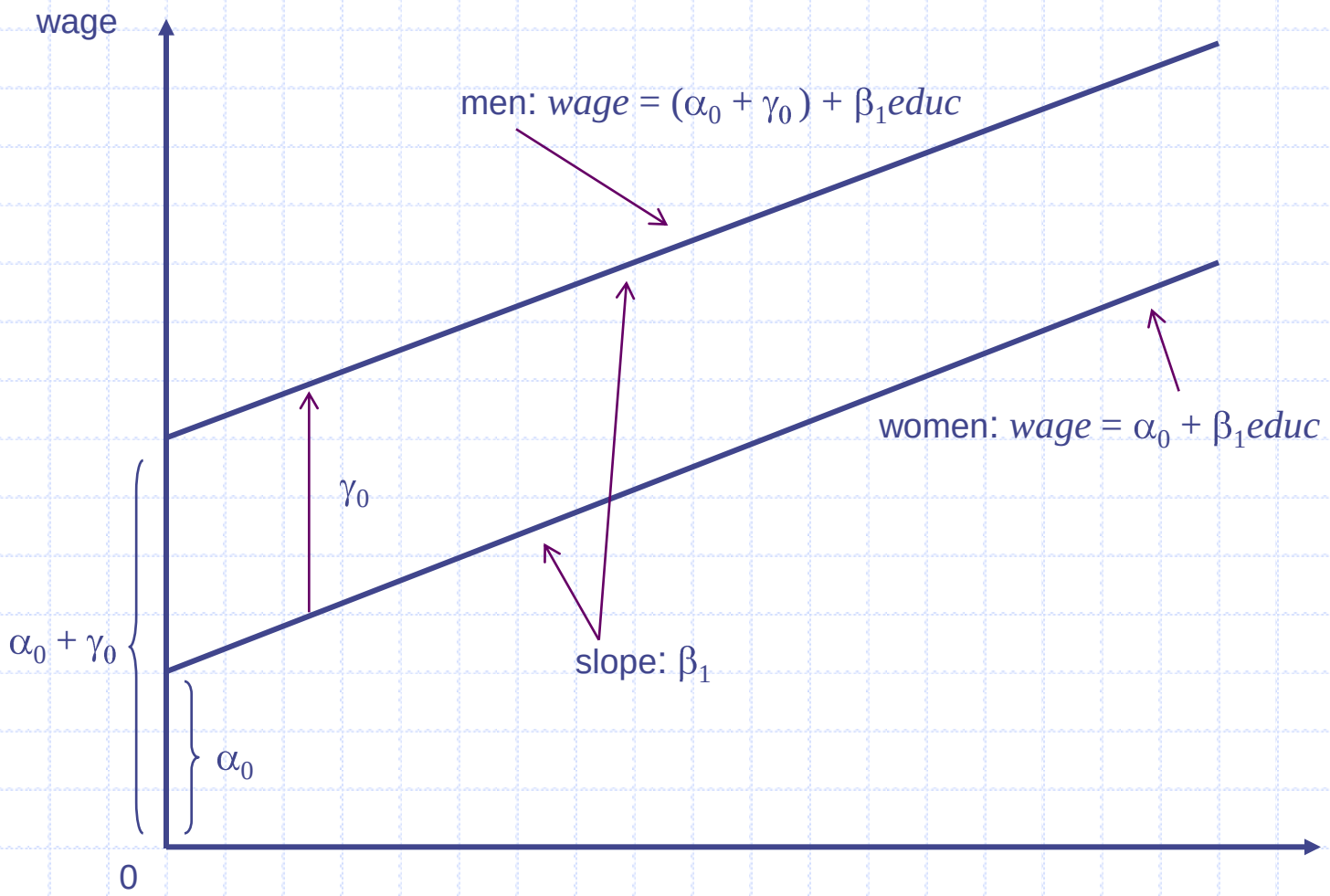
- Nothing, except the interpretation of α_0 and γ_0.

Single Dummy Independent Variable

- α_0 is the intercept for females, which is now the **base group**, and $\alpha_0 + \gamma_0$ is the intercept for males.
- This implies the following relation between coefficients:

$$\alpha_0 = \beta_0 + \delta_0 \quad \text{and} \quad \alpha_0 + \gamma_0 = \beta_0 \quad \Rightarrow \quad \gamma_0 = -\delta_0$$

Example of $\gamma_0 > 0$



Single Dummy Independent Variable

- In any application, it does not matter how we choose the base group, since this only affects the interpretation of the coefficients associated to the dummy variables, but it is important to keep track of which group is the base group.
- Choosing a base group is usually a matter of convenience.

Single Dummy Independent Variable

- It would be possible also to drop the intercept and to include a dummy variable for each group or category.
- The equation would then be

  $wage = \mu_0 male + \nu_0 female + \beta_1 educ + u$

  where the intercept for men is μ_0 and the intercept for women is ν_0.
- There is no dummy variable trap in this case because we do not have an overall intercept.

Single Dummy Independent Variable

- **Important**: Nothing changes about the mechanics of OLS or the statistical theory when some of the independent variables are dummy variables.
- The only difference is in the interpretation of the coefficient of the dummy variable.

Single Dummy Independent Variable

- **Hypothesis testing** proceeds as usual.
- In the above example the null hypothesis of *no* difference between men and women is $H_0: \delta_0 = 0$.
- The alternative that there is discrimination against women is $H_1: \delta_0 < 0$.
- We can test this by means of a one sided *t* test.

Dependent Variable is log(y)

- A common specification in applied work has the dependent variable as log(y), with one or more dummy variables appearing as independent variables.

  Example:

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + u$$

Dependent Variable is log(*y*)

- How do we interpret the dummy variable coefficients in this case?
- Because δ_0 is the difference in *log* hourly wage between females and males, *given* the same amount of education (and the same error term *u*), δ_0 has now a *percentage* interpretation.

Dependent Variable is log(y)

- When log(y) is the dependent variable in a model, the coefficient on a dummy variable, when multiplied by 100, is interpreted as the percentage difference in y, with respect to the base group, holding all other factors fixed.

Several Dummy Independent Variables

- We can use several dummy independent variables in the same equation.

- We can distinguish to cases:

1. Multiple categories or groups for a given attribute.
2. Several attributes.

Multiple Categories or Groups

- To measure geographical wage differentials we can define the following regional dummy variables,

  $region_j = 1$ if observation belongs to region j,

  $region_j = 0$ otherwise,

  for $j = 0,1,2,3,\ldots,16$ if there are 17 regions in the country, i.e. 17 *CCAA* in Spain.

Multiple Categories or Groups

- And estimate the following equation:

$$wage = \beta_0 + \sum_{j=1}^{16} \theta_j region_j + \beta_1 edu + u$$

θ_j is the difference in hourly wage between *region j* and *region* 0, *given* the same amount of education (and the same error term *u*).

region 0 (whatever it is) is the **base group**, that is comparisons are made against this region.

Multiple Categories or Groups

- The previous example illustrates a **general principle** for including dummy variables to indicate different categories or groups:

- If the regression model is to have different intercepts for, say g **categories or groups**, we need to include $g - 1$ **dummy variables** in the model along with an intercept.

Multiple Categories or Groups

- The intercept for the base group is the overall intercept in the model, and the dummy variable coefficient for a particular group represents the estimated difference in intercepts between that group and the base group.
- Including g dummy variables along with an intercept will result in the **dummy variable trap**.

Multiple Categories or Groups

- An alternative is to include g dummy variables and to exclude an overall intercept.
- This is not advisable, not only because testing differences relative to a base group becomes more difficult, but because this only works in the case when we only have one attribute.

Several Attributes

- Consider now the possibility to take into account two possible sources of wage discrimination, gender and marital status.
- We say in this case that we have two attributes to take into consideration.

Several Attributes

- Then introduce the dummy variable, *female* to take into account gender, and the dummy variable *married*, to take into account marital status,

$$wage = \beta_0 + \delta_0 female + \phi_0 married + \beta_1 educ + u$$

- ϕ_0 is the difference in hourly wage between those who are and are not married, *given* gender and the same amount of education (and also the same error term *u*).

Several Attributes

- Note that for each attribute we have a base group, male for gender and not married for marital status.
- The overall intercept in the equation picks up the effect of both base groups, male and not married, so single men is the base group.
- In the simplest you should introduce, for each attribute, a number of dummy variables equal to the number of categories less one.

Ordinal Variables

- An **ordinal variable** is a variable that represents a *ranking*, i.e. University rankings.
- Any ordinal variable can be turned into a set of dummy variables, and these can be introduced in a regression model, after a base group have been selected.
- If there are a lot of categories, it may make sense to group some together, i.e. top 10 *ranking*, 11 – 25,…etc.

Interactions Among Dummy Variables

- Just as variables with quantitative meaning can be interacted in regression models, so can dummy variables.
- To allow for a possibility of an interaction between gender and marital status on wage discrimination we can add an **interaction term** between *female* and *married* in the above example.

Interactions Among Dummy Variables

- The equation to estimate is

$$wage = \beta_0 + \delta_0 female + \phi_0 married + \varphi_0 female.married + \beta_1 educ + u$$

- This allows the marriage premium to depend on gender.
- This model allows us to obtain the estimated wage differential among all four groups, *married-females*, *single-females*, *married-men* and *single-men*.

Allowing for Different Slopes

- So far we have only allowed for different intercepts for any number of groups in a multiple regression model.
- There are also occasions for interacting dummy variables with explanatory variables that are not dummy variables to allow for a **difference in slopes**.

Allowing for Different Slopes

- How can we take into account different returns to education according to gender?

- Consider the following model,

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female.educ + u$$

Allowing for Different Slopes

- For males, *female* = 0, so

$$wage = \beta_0 + \beta_1 educ + u$$

Intercept β_0, and slope β_1

- For females, *female* = 1, so

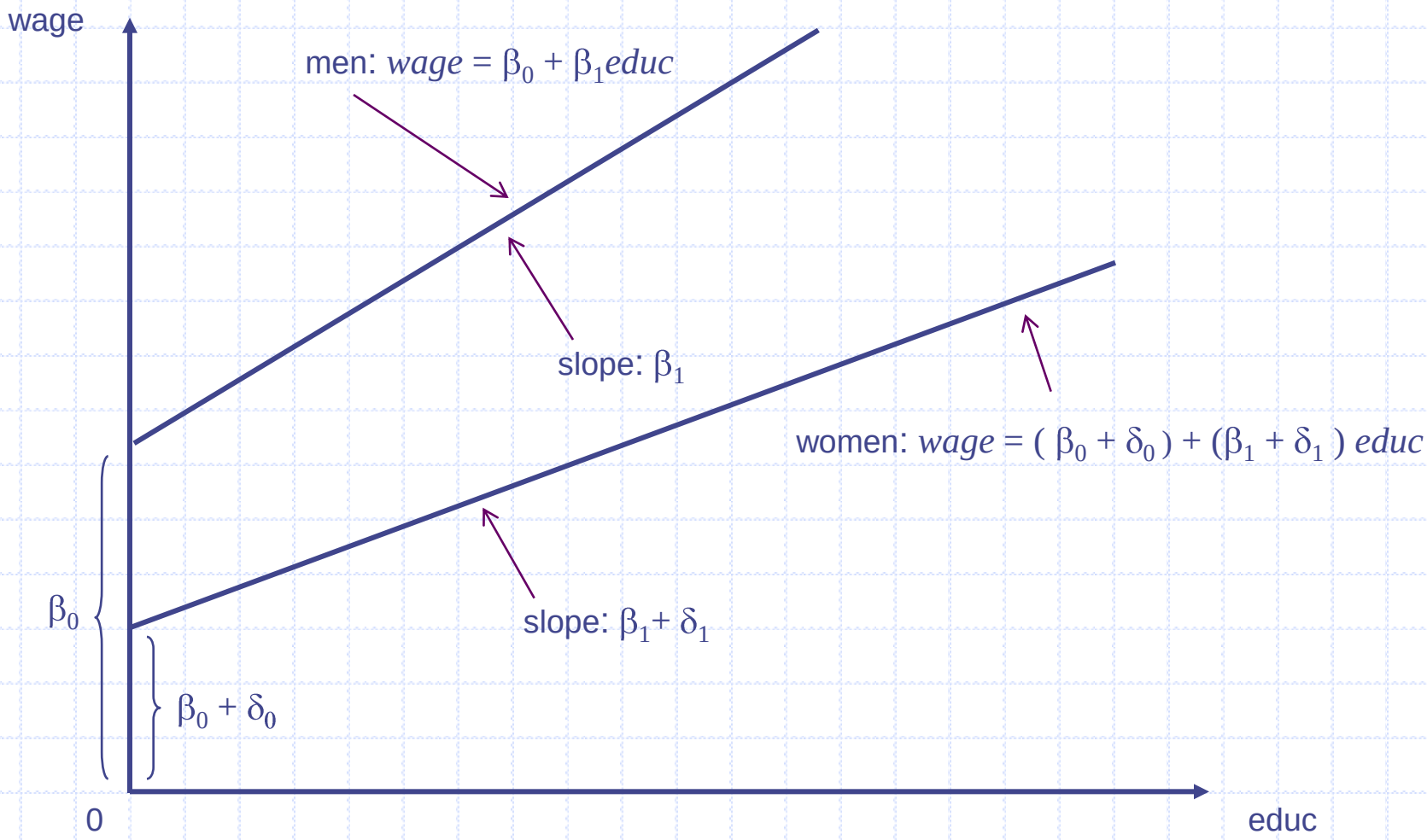
$$wage = \beta_0 + \delta_0 + (\beta_1 + \delta_1)\, educ + u$$

Intercept $\beta_0 + \delta_0$, and slope $\beta_1 + \delta_1$

Allowing for Different Slopes

- Therefore, δ_0 measures the difference in intercepts between women and men.
- And, δ_1 measures the difference in the return to education between women and men.

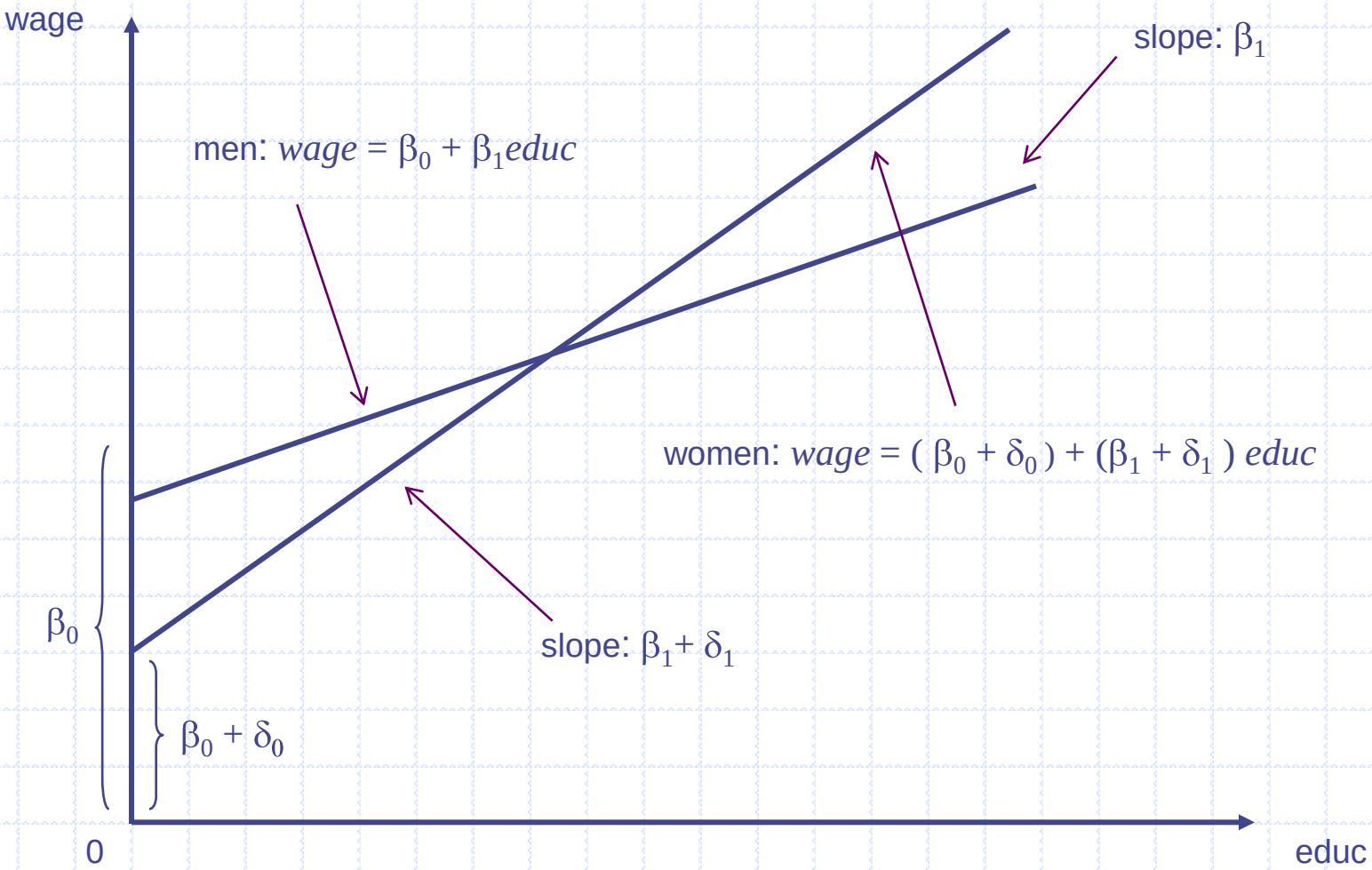
Example of $\delta_0 < 0$ and $\delta_1 < 0$



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

- This case shows a lower intercept and a lower slope for women than for men.
- This means that women earn less than men at all levels of education, and the gap increases as *educ* gets larger.
- An additional year of education shows a lower return for women than for men.

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



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

- This case shows a lower intercept for women than for men but a higher slope for women than for men.
- This means that women earn less than men at low levels of education, but the gap narrows as education increases. At some point, a woman earns more than a man, given the same levels of education.
- An additional year of education shows a higher return for women than for men.

Allowing for Different Slopes

- A test that the return to education is the same for women and men is stated as H_0: $\delta_1 = 0$, which means that the slope of *wage* with respect to *educ* is the same for men and women.
- Note that this hypothesis puts no restrictions on the difference in intercepts, δ_0. A wage differential between men and women is allowed under this null, but it must be the same at all levels of education.

Testing for Differences Across Groups

- Sometimes, we wish to test the **null hypothesis** that **two populations or groups follow the same regression function**, against the alternative that one or more of the slopes differ across groups.
- This is simply a test for the joint significance of the dummy and its interactions with all other independent variables.

Testing for Differences Across Groups

- So we estimate the model with all the interactions (unrestricted model) and without them (restricted model) and form an *F* statistic.
- **Example**: To test for differences in wage determination due to gender, we estimate the equation

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female.educ + u$$

Testing for Differences Across Groups

And test the hypothesis $H_0: \delta_0 = \delta_1 = 0$.

The restricted model is

$$wage = \beta_0 + \beta_1 educ + u$$

From the estimation of both equations we form the F statistic, either from the SSR or from the R^2.

- The same procedure works in cases with more explanatory variables.

The Chow Test

- It turns out that you can compute the proper *F* statistic without running the unrestricted model with the dummy and the interactions with all *k* continuous variables.
- The key insight is that the unrestricted *SSR* can be obtained from the *SSR* from the separate regressions for each group.

The Chow Test

- **Example**: In the wage determination example it can be shown that the *SSR* of the equation

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female.educ + u$$

say SSR_{ur}, is equal to the sum of the *SSR* obtained from estimating

$$wage = \beta_0 + \beta_1 educ + u$$

for women, say SSR_1, and for men, say SSR_2.

$$SSR_{ur} = SSR_1 + SSR_2$$

The Chow Test

- Hence a test of the hypothesis H_0: $\delta_0 = \delta_1 = 0$ can be constructed by estimating

$$wage = \beta_0 + \beta_1 educ + u$$

three times:

(*i*) for the whole sample, this is usually called the **pooled regression**, $SSR_r = SSR_p$, and

(*ii*) and (*iii*) for each subgroup, women, SSR_1, and men, SSR_2, so $SSR_{ur} = SSR_1 + SSR_2$.

The Chow Test

- And forming the F statistic

$$F = \frac{\left[SSR_p - \ SSR_1 + SSR_2 \right]}{SSR_1 + SSR_2} \cdot \frac{n - 2(k+1)}{k+1}$$

- So the test procedure proceeds as usual.

The Chow Test

- This particular F statistic is usually called the **Chow statistic** in econometrics.
- Since the Chow test is just an F test, it is only valid under homoskedasticity.
- In particular, under the null hypothesis, the error variances for the groups must be equal.
- As usual, normality is not needed for asymptotic analysis.

The Chow Test

- The Chow test is really just a simple F test for exclusion restrictions, the dummy and all the interactions with the explanatory variables, but we have realized that the SSR for the unrestricted model is just the sum of the SSR for each of the groups considered, so

$$SSR_{ur} = SSR_1 + SSR_2.$$

The Chow Test

- Note that we have $k + 1$ restrictions, the slope coefficients (interactions) plus the intercept (dummy).
- Note also that the unrestricted model would estimate 2 different intercepts and 2 different slope coefficients, so *df* of the model is $n - 2k - 2$.
- Hence the statistic, $F = \frac{\left[SSR_p - SSR_1 + SSR_2 \right]}{SSR_1 + SSR_2} . \frac{n - 2(k+1)}{k+1}$

The Chow Test

- One important limitation of the Chow test, regardless of the method used to implement it, is that under the null there are no differences at all between the groups.
- In many cases, it is more interesting to allow for an intercept difference between the groups under the null and then to test for slope differences only.

The Chow Test

- **Example**: In the wage determination example,

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female.educ + u$$

this means testing $H_0: \delta_1 = 0$ only.

The Chow Test

- We can easily perform this kind of test,

  (*i*) either by including the group dummy and all the interactions and test joint significance of the interaction terms only, or

  (*ii*) by forming the restricted SSR, SSR_p, from the regression that allows an intercept shift only. In other words, we run a pooled regression and just include the dummy variable that distinguishes the two groups.

The Chow Test

- Note that the Chow test, in either form, can be generalized to more than two groups in a natural way.
- In this case, running separate regressions for each group and the pooled regression is probably the easiest way to perform the test from a practical point of view.

Linear Probability Model

- So far we have seen how to incorporate qualitative information as explanatory variables in a multiple linear regression.
- But in all models up until now, the dependent variable *y* has had *quantitative* meaning.
- What happens if we want to use multiple regression to explain a *qualitative* event?
- In the simplest case the event we would like to explain is a binary outcome.

Linear Probability Model

- In this case, our dependent variable, y, takes on only two values, zero and one.
- **Example**: We want to explain labor force participation, so we can define

  $y = 1$ if the person is in the labor force,

  $y = 0$ if the person is out of the labor force.

Linear Probability Model

- What does it mean to write down a multiple regression model, such as,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

when y is a binary variable?

- Since y can take on only two values, β_j cannot be interpreted as the change in y given a one-unit increase in x_j, holding all other factors fixed: y either changes from zero to one or from one to zero.

Linear Probability Model

- Nevertheless, the β_j still have useful interpretations.
- If we assume that the zero conditional mean assumption MLR.3 holds, $E(u|x_1,x_2,\ldots,x_k) = 0$, then we have, as always,

$$E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

where $\boldsymbol{x}$ is shorthand for all of the explanatory variables.

Linear Probability Model

- The key point is that when y is a binary variable taking on the values zero and one,

$$E(y|\boldsymbol{x}) = 1.P(y = 1|\boldsymbol{x}) + 0.\ P(y = 0|\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$$

- So, it is always true that $P(y = 1|\boldsymbol{x}) = E(y|\boldsymbol{x})$: the probability of "success", i.e. the probability that $y = 1$, is the same as the expected value of y.

Linear Probability Model

- Thus, we have the important equation

$$P(y = 1 \mid \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

which says that the probability of success, say $p(\boldsymbol{x}) = P(y = 1 \mid \boldsymbol{x})$, is a linear function of the x_j.

- This is an example of a **binary response model**, and $P(y = 1 \mid \boldsymbol{x})$ is also called the **response probability**.

Linear Probability Model

- Because probabilities must sum to one, $P(y = 0 | \boldsymbol{x}) = 1 - P(y = 1 | \boldsymbol{x})$ is also a linear function of the x_j.
- The MLR model with a binary dependent variable is called the **linear probability model (LPM)** because the response probability is linear in the parameters β_j.

Linear Probability Model

- In the LPM, β_j measures the change in the probability of success when x_j changes by one unit, holding other factors fixed:

$$\Delta P(y = 1 \mid \boldsymbol{x}) = \beta_j \Delta x_j$$

- With this in mind, the MLR model can allow us to estimate the effects of various explanatory variables on qualitative events.
- The mechanics of OLS are, however, unchanged.

Linear Probability Model

- If we write the estimated equation as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

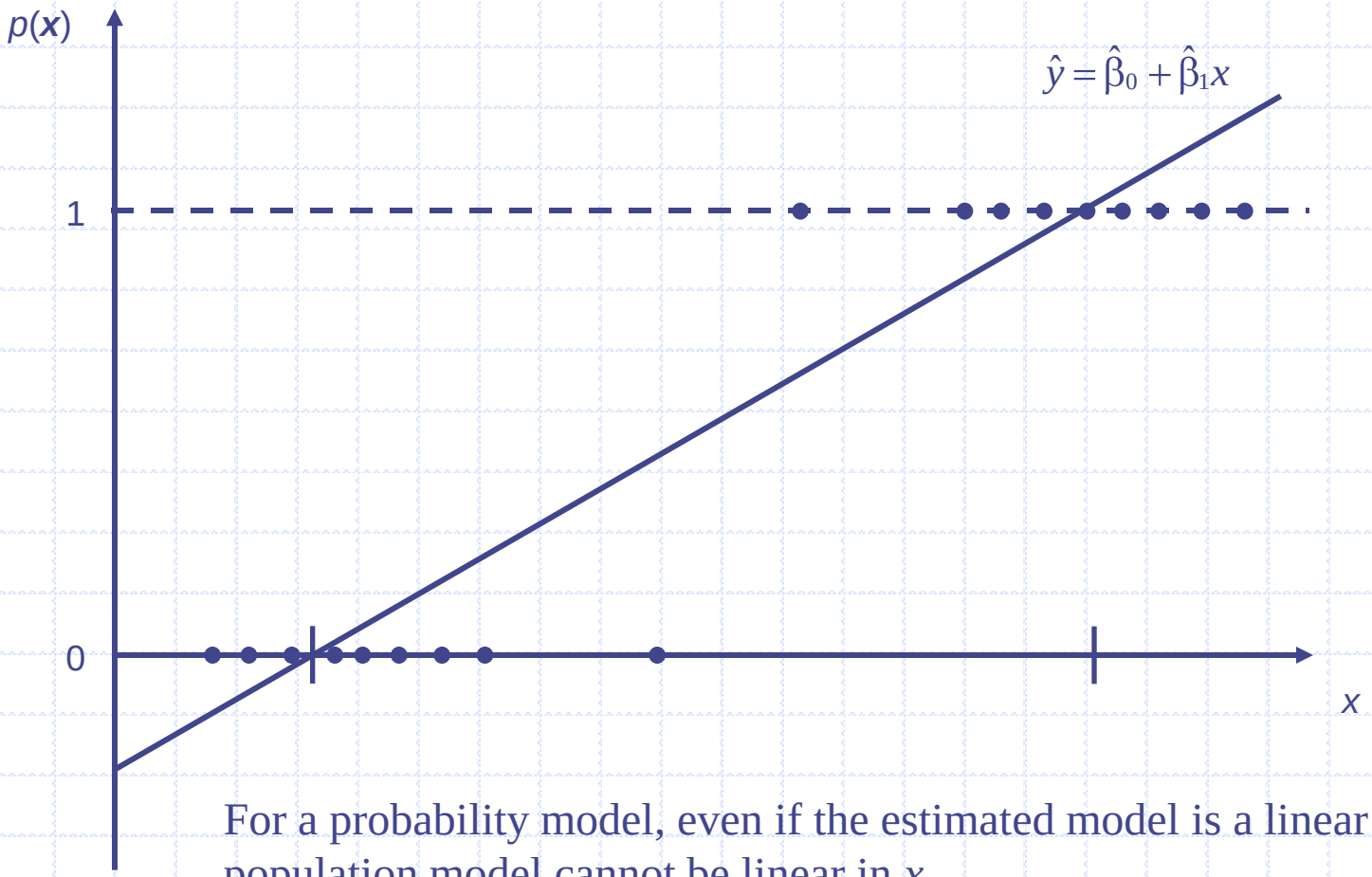
  we must now remember that $\hat{y}$ is the predicted probability of success.

- The LPM is simple to estimate and simple to interpret but has some shortcomings that should be known.

Linear Probability Model

1. For certain values of the x´s, $\hat{y}$ can be outside the interval [0, 1]. Since these are predicted probabilities, this is nonsense.
2. A related problem is that a probability cannot be linearly related to the independent variables for all their possible values, since eventually we will be outside the interval [0, 1]. Probability models are essentially nonlinear.

Linear Probability Model



For a probability model, even if the estimated model is a linear one, the population model cannot be linear in *x*.

Linear Probability Model

3. Due to the binary nature of y, the LPM exhibits heteroskedasticity, so the LPM does violates one of the Gauss-Markov assumptions.

   When y is a binary variable, its variance, conditional on $\boldsymbol{x}$, is

   $$\text{Var}(y \mid \boldsymbol{x}) = \text{E}(y^2 \mid \boldsymbol{x}) - \text{E}(y \mid \boldsymbol{x})^2 = p(\boldsymbol{x}) - p(\boldsymbol{x})^2$$

   so

   $$\text{Var}(y \mid \boldsymbol{x}) = p(\boldsymbol{x}).[1 - p(\boldsymbol{x})]$$

Linear Probability Model

- This means that, except in the case where the probability does not depend on any of the x´s, there *must* be heteroskedasticity in the LPM.
- This does not cause bias in the OLS estimators, but standard formulas for OLS standard errors are not valid.
- Hence, standard inference with the usual t and F statistics is not justified, even in large samples.

Linear Probability Model

4. **Normality is not an acceptable assumption**, since the normal has continuous support on the real line and now *y* takes only two values, 0 and 1.

   In fact *y* has, conditional on the *x´*s, a **bernoulli distribution**.

   This means that estimators will not be normal in finite samples, even they will be normal in large samples.

Linear Probability Model

- It is possible to correct the standard errors for heteroskedasticity, and modern *software* is able to do this.
- In this case correct inference can be performed.
- In particular, the *t* statistics calculated with heteroskedasticity corrected standard errors have a standard normal distribution in large samples and they can be used to perform hypothesis testing or to construct CI.

Linear Probability Model

- Despite all these drawbacks the LPM is a good place to start when *y* is binary, and usually provides sensible estimates.
- In terms of prediction of probabilities is best to use values of the independent values that are near the averages in the sample.
- We can also include *dummy* variables as independent variables even if *y* is binary. The coefficient measures the predicted difference in probability with respect to the base group.

Policy Analysis and Program Evaluation

- A typical use of *dummy* variables is when we are looking for a program effect.
- For example, we may have individuals that received job training, or welfare benefits and we are interested in measuring changes in behavior after the program is in effect.
- We need to remember that in the social sciences the control and treatment groups are not randomly assigned.

Self-selection Problems

- Usually individuals choose whether to participate in a program or not, which may lead to a **self-selection** problem.
- If we can control for everything that is correlated with both, participation and the outcome of interest, then there is no problem.
- Often, though, there are unobservables that are correlated with participation, and this leads to the usual omitted variable bias.