

Econometrics Lecture 8

EC2METRIE

Dr. Anna Salomons

Utrecht School of Economics (U.S.E.)

16 January 2017

This class

- ▶ **Linear probability model**
- ▶ **Studenmund**
 - ▶ Chapter 12, section 1 (i.e. excluding sections 2 and 3)
 - ▶ Note that this material is not part of the project paper, i.e. is only relevant for the exam.

Dummy variable

- ▶ A dummy variable is a variable that can only **take on the value 0 or 1**.
- ▶ So far, we have seen how to interpret regression models with a **dummy variable as an independent variable**.
- ▶ However, sometimes we want to answer economic questions that require using a **dummy variable as a dependent variable**.

Dummy variable as independent variable: example

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 male_i + \varepsilon_i$$

where $wage_i$ are wages measured in dollars.

- ▶ $male_i$ is a dummy variable, =1 if the person is male, =0 if the person is female.
- ▶ Interpretation of β_3 : men earn β_3 dollars more than women, holding constant education and age.

Dummy variable as independent variable: example

```
. sum lnprice attractive school age rich alcohol bar street other
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lnprice	3016	5.839489	.7155389	2.302585	8.665613
attractive	3016	.137931	.3448848	0	1
school	3016	.3169761	.4653752	0	1
age	3016	27.40981	7.729452	12	54
rich	3016	.8428382	.3640136	0	1
alcohol	3016	.846817	.3602236	0	1
bar	3016	.8047082	.3964909	0	1
street	3016	.1747347	.379803	0	1
othersite	3016	.020557	.1419194	0	1

Dummy variable as independent variable: example

```
. reg lnprice attractive school age rich alcohol bar street
```

Source	SS	df	MS
Model	501.703241	7	71.6718916
Residual	1041.9644	3008	.34639774
Total	1543.66764	3015	.511995901

Number of obs = 3016
 F(7, 3008) = 206.91
 Prob > F = 0.0000
 R-squared = 0.3250
 Adj R-squared = 0.3234
 Root MSE = .58856

lnprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
attractive	.2394121	.0315921	7.58	0.000	.1774678	.3013563
school	.1637754	.0238151	6.88	0.000	.11708	.2104709
age	-.0210136	.0014531	-14.46	0.000	-.0238627	-.0181645
rich	.2924201	.0304404	9.61	0.000	.232734	.3521061
alcohol	.2403329	.0358481	6.70	0.000	.1700436	.3106222
bar	.2160627	.0785642	2.75	0.006	.0620178	.3701076
street	-.2621039	.0793876	-3.30	0.001	-.4177633	-.1064444
_cons	5.752484	.0912836	63.02	0.000	5.5735	5.931469

Dummy variable as independent variable: interpretation

Interpretation of a coefficient is based on the **marginal (or partial) effect**:

- ▶ Effect of *age* on *ln price*:

$$\frac{\partial \ln price}{\partial age} = \beta_{age} = -0.02$$

If the sex worker is 1 year older, she earns 2% less per transaction, cet. par.

- ▶ For a dummy variable, this is

$$\frac{\partial \ln price}{\partial alcohol} = \beta_{alcohol} = 0.24$$

If the client consumed alcohol, the transaction price is 24% higher, cet. par.

Dummy variable as independent variable: interpretation

Interpretation of a coefficient is based on the **marginal (or partial) effect**

- ▶ If a dummy variable measures more than one category, interpretation is relative to the omitted category:

$$\frac{\partial \ln price}{\partial bar} = \beta_{bar} = 0.21$$

If the transaction originated in a bar, the price is 21% higher than if it originated in another site (not the bar, nor the street), *cet. par.*

Dummy variable as dependent variable

- ▶ Sometimes, we want to estimate a model that has a **dummy variable as a dependent variable**.
- ▶ This is often when we want to model a choice, e.g. we may want to model the choice:
 - ▶ to participate in the labor force or not (0=no participation; 1=participation);
 - ▶ to have a child or not (0=no child, 1=child);
 - ▶ to go to college or not (0=no college, 1=college);
 - ▶ to buy a particular product or not (0=do not buy, 1=buy);
 - ▶ to report a crime or not (0=not reported, 1=reported);
 - ▶ to hire someone or not (0=not hired; 1=hired).


Example of a dummy dependent variable model



Are Emily and Brendan more employable than Lakisha and Jamal?

Example of a dummy dependent variable model

- ▶ **Racial discrimination in the labor market:** very famous study by Marianne Bertrand & Sendhil Mullainathan (American Economic Review, 2004).¹
- ▶ **Fictitious CVs** were sent in response to 1,300 help-wanted ads listed in the Boston Globe and the Chicago Tribune.
- ▶ CVs had **randomly assigned names**, which were either **distinctly white-sounding** (*Emily Walsh, Brendan Baker*) or **distinctly African-American-sounding** (*Lakisha Washington, Jamal Jones*).

¹A non-technical summary of the study is available [here](#). 

Marianne & Sendhil



Do Lakisha and Jamal get fewer call-backs? Summary statistics

To investigate this, we need to use a dummy-dependent variable model:

$$call_i = f(black_i, pcol_i, linc_emp_i)$$

variable name	variable label
call	=1 if applicant was called back
black	=1 if applicant has an African-American sounding name
pcol	percentage college educ or more in applicant's zipcode
linc_emp	log median household income in employer's zipcode

Variable	Obs	Mean	Std. Dev.	Min	Max
call	1870	.0641711	.2451231	0	1
black	1870	.4994652	.5001335	0	1
pcol	1870	21.39705	16.90033	3.084686	78.01243
linc_emp	1870	10.65373	.4408802	9.170247	11.81431

Dummy dependent variable model

- ▶ But **how can we estimate such a model**, which has a 0-1 variable as the dependent variable?
- ▶ It turns out, one possibility is to use the OLS estimator: this is called the **Linear Probability Model (LPM)**
- ▶ Non-linear estimators, such as logit and probit, are also an option but not part of this course.

Using OLS: linear probability model

- ▶ We can **use OLS** to estimate a model with a dummy dependent variable: this is called the **Linear Probability Model** (LPM)

- ▶ This means we can simply write the model as:

$$call_i = \beta_0 + \beta_1 black_i + \beta_2 pcol_i + \beta_3 linc_emp_i + \varepsilon_i$$

- ▶ The dependent variable $call_i$ is 0 or 1, the independent variables can be either dummy variables as well ($black$) or continuous variables ($fraccol$, $linc_emp$).

LPM: Bernoulli distribution

- ▶ The **dependent variable follows a Bernoulli distribution** (=a binomial distribution with $n = 1$):

$$f(Y; \theta) = \theta^Y (1 - \theta)^{1-Y}$$

with

$$\Pr(Y = 1) = \theta^1 (1 - \theta)^{1-1} = \theta$$

$$\Pr(Y = 0) = \theta^0 (1 - \theta)^{1-0} = 1 - \theta$$

- ▶ So in our example, the probability of getting called back is θ , and the probability of not getting called back is $1 - \theta$.

LPM: Bernoulli distribution

- ▶ To find the mean of this dependent variable:

$$\begin{aligned}\mu &= E(Y_i) = \sum \Pr(Y_i = y_i) \times y_i \\ &= (1 - \theta) \times 0 + \theta \times 1 \\ &= \theta\end{aligned}$$

- ▶ This is 0.064 (see summary statistics): so the average probability of getting getting called back is 6.4%.

LPM: Bernoulli distribution

- And the variance:

$$\begin{aligned} \text{Var}(Y_i) &= E(y_i - \mu)^2 \\ &= \sum \Pr(Y_i = y_i) \times (y_i - \mu)^2 \\ &= (1 - \theta) \times (0 - \theta)^2 + \theta \times (1 - \theta)^2 \\ &= (1 - \theta) \times \theta^2 + \theta \times (1 - \theta)^2 \\ &= \theta^2 - \theta^3 + \theta \times (1 - 2\theta + \theta^2) \\ &= \theta^2 - \theta^3 + \theta - 2\theta^2 + \theta^3 \\ &= \theta^2 + \theta - 2\theta^2 \\ &= \theta - \theta^2 = \theta(1 - \theta) \end{aligned}$$

- Which is $0.064(1 - 0.064) \approx 0.060$. Hence the standard deviation is $\sqrt{0.060} \approx 0.245$ (see summary statistics)

LPM: interpretation

$$call_i = \beta_0 + \beta_1 black_i + \beta_2 pcol_i + \beta_3 linc_emp_i + \varepsilon_i$$

- ▶ This can be shown to be (linearly) **modeling the probability of getting called back** (hence linear probability model):

$$\Pr(\mathbf{call}_i = \mathbf{1}) = \beta_0 + \beta_1 black_i + \beta_2 pcol_i + \beta_3 linc_emp_i + \varepsilon_i$$

- ▶ A one-unit change in the dependent variable corresponds to a 100% change.
- ▶ This means a change in the dependent variable is **in percentage points**, if we multiply the equation by 100

LPM interpretation: proof

We can prove that the LPM models the probability of the dependent variable taking on the value 1. First, we know that (under the OLS assumptions), in linear regression it holds that:

$$E(y_i|x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

Writing this for the LPM, where y_i can take on the value 0 or 1:

$$\begin{aligned} E(y_i|x_i) &= \sum y_i \Pr(Y = y_i|x_i) \\ &= 0 \times \Pr(Y = 0|x_i) + 1 \times \Pr(Y = 1|x_i) \\ &= \Pr(Y = 1|x_i) \end{aligned}$$

LPM interpretation: proof

We have shown that for the LPM, it holds that:

$$E(y_i|x_i) = \Pr(Y = 1|x_i)$$

Hence

$$\Pr(Y = 1|x) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

And

$$\Pr(Y = 1) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

LPM: interpretation

$$\Pr(\text{call}_i = 1) = \beta_0 + \beta_1 \text{black}_i + \beta_2 \text{fraccol}_i + \beta_3 \text{linc_emp}_i + \varepsilon_i$$

- ▶ The **marginal (or partial) effect of each variable is given by its coefficient** (just like in ordinary linear regression):

$$\begin{aligned} \frac{\partial \Pr(\text{call}_i = 1)}{\partial \text{black}_i} &= \frac{\partial (\beta_0 + \beta_1 \text{black}_i + \beta_2 \text{pcol}_i + \beta_3 \text{linc}_i + \varepsilon_i)}{\partial \text{black}_i} \\ &= \beta_1 \end{aligned}$$

- ▶ **Interpretation:** if an applicant has an African-American sounding name, the probability that they get called back for an interview is $\beta_1 \times 100$ percentage points higher than if they have a white-sounding name, *cet. par.*

LPM estimates

```
. reg call black pcol linc_emp
```

Source	SS	df	MS
Model	1.2526628	3	.417554268
Residual	111.046802	1866	.059510612
Total	112.299465	1869	.060085321

Number of obs = 1870
 F(3, 1866) = 7.02
 Prob > F = 0.0001
 R-squared = 0.0112
 Adj R-squared = 0.0096
 Root MSE = .24395

call	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
black	-.0313094	.0112851	-2.77	0.006	-.0534421	-.0091768
pcol	.0009546	.0003341	2.86	0.004	.0002994	.0016097
linc_emp	-.0272847	.0128032	-2.13	0.033	-.0523948	-.0021746
_cons	.3500684	.1370253	2.55	0.011	.0813294	.6188074

LPM estimates: interpretation

All estimated coefficients are significant, interpretations:

- ▶ If an applicant has an African-American sounding name, the probability that they get called back for an interview is 3.13 percentage points lower than if they have a white-sounding name, cet. par.
- ▶ If an applicant lives in an area with 1 percentage point more college educated workers, the probability that they get called back for an interview is 0.095 percentage points higher, cet. par.
- ▶ If income in the employer's area is 1% higher, the probability that an applicant gets called back for an interview is 0.027 percentage points lower, c.p.

LPM

- ▶ **Advantages:** easy to compute and easy to interpret.
- ▶ Most important **disadvantages:**
 - ▶ The errors are **heteroskedastic** by construction.
 - ▶ **Predicted probabilities may lie outside the 0-1 interval**
(and of course it makes no sense to interpret these as probabilities).

LPM: heteroskedastic errors

In the LPM, the **variance of the error term is not constant**, i.e. the errors are **heteroskedastic**. This is not a big problem:

- ▶ We know that heteroskedasticity **biases the standard errors**.
- ▶ But we can easily correct this using the option **,robust**.
- ▶ So whenever you estimate a linear probability model, use this option.

LPM: adjustment for heteroskedasticity

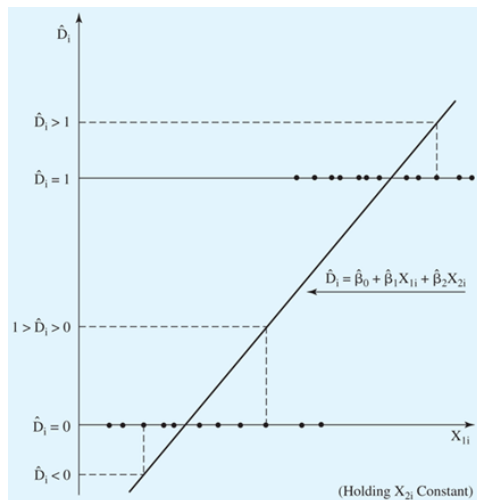
```
. reg call black pcol linc_emp, robust
```

Linear regression

```
Number of obs =    1870
F( 3, 1866) =    6.90
Prob > F      =    0.0001
R-squared     =    0.0112
Root MSE     =    .24395
```

call	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
black	-.0313094	.0113075	-2.77	0.006	-.0534862	-.0091327
pcol	.0009546	.0003787	2.52	0.012	.0002119	.0016972
linc_emp	-.0272847	.013683	-1.99	0.046	-.0541202	-.0004492
_cons	.3500684	.1462512	2.39	0.017	.0632352	.6369016

LPM: predicted probabilities outside the 0-1 interval



LPM: predicted probabilities outside the 0-1 interval

- ▶ Since OLS is a linear estimator, predicted probabilities can lie outside the 0-1 interval.
- ▶ Therefore, whenever estimating a LPM, you should always **check your predicted probabilities**.
- ▶ **If many are <0 or >1 , you should use different estimator** (logit / probit) instead of OLS – not part of this course.

LPM: predicted probabilities outside the 0-1 interval

```
. predict yhat
(option xb assumed; fitted values)
```

```
. sum yhat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat	1870	.0641711	.0258888	-.0006465	.1388698

```
. sum yhat if yhat<0 | yhat>1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat	2	-.0003539	.0004138	-.0006465	-.0000613

LPM: predicted probabilities outside the 0-1 interval

- ▶ Conclusion: **LPM appropriate here** since only 0.10% ($=2/1870 \cdot 100\%$) of the predicted values are below 0 or above 1.
- ▶ In cases where more than a few percent of the predicted values lie outside the 0-1 interval, LPM is not appropriate.

Conclusions on labor market discrimination

- ▶ The **marginal effect of an African-American sounding name on the probability of being called back** is significantly negative.
- ▶ **Conclusion:** firms consider Emily and Brendan to be more employable than Lakisha and Jamal, **African-Americans are discriminated against in the labor market.**
- ▶ This experiment has also been performed in the Netherlands, with similar findings: for identical CVs, applicants with immigrant-sounding names are discriminated against in favor of applicants with Dutch-native sounding names.²

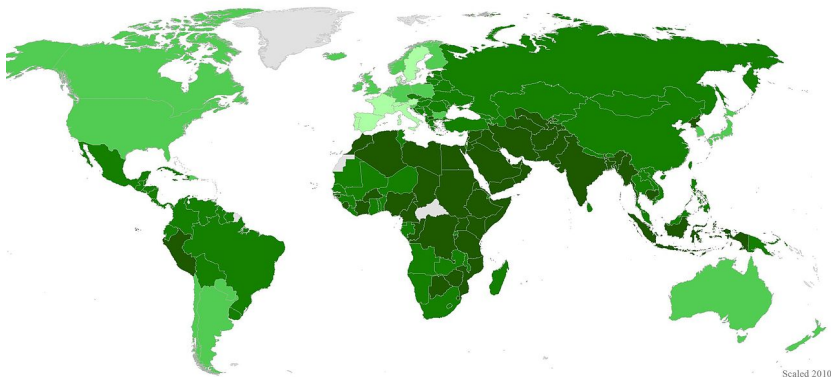
²Study ("Liever Mark dan Mohammed", in Dutch) available [here](#).

Discrimination: some further disturbing evidence

- ▶ This type of experimental approach has been replicated many times in different settings: one study finding that university professors are discriminating against female and minority students (available [here](#)).
- ▶ See the [New York Times](#) article, "*Professors Are Prejudiced, Too*" for a non-technical summary.

Example: crimes against women

Women's Physical Security



■ No Data

■ Women physically secure

■ Women have high levels of physical security

■ Women have medium levels of physical security

■ Women have low levels of physical security

■ Women lack physical security

Dummy dependent variable model: an example

- ▶ **Underreporting of crimes:** only 1% of Indian women who experience sexual violence report it to the police (Amnesty International 2014).
- ▶ We therefore want to investigate what determines the **reporting of crimes against women** (0=not reported, 1=reported).
- ▶ Note that this implies estimating a dummy dependent variable model.

Dummy dependent variable model: an example

- ▶ In particular, we want to know if **increased political representation of women** makes fighting violence against women a higher priority, thereby increasing the reporting of such crimes.
- ▶ In West Bengal, there has been a project which reserved the position of **chief of Village Council** of 1/3 of all villages for women- i.e., only a woman could be elected. Which villages were affected was randomized, which makes this a **natural experiment!**

└ Another example of a dummy dependent variable model

West Bengal



Dummy dependent variable model: an example

We want to estimate the following model:

$$report_i = \beta_0 + \beta_1 res_woman_i + \beta_2 age_i + \varepsilon_i$$

the subscript i indicates individuals; $report_i$ is a dummy variable equal to 1 if a crime was reported; res_woman_i is a dummy variable equal to 1 if the individual lives in a village where the chief position was reserved for a woman; and age_i is the age of the respondent.

- We can look at different types of crimes, those against women and property crimes.

Dummy dependent variable model: an example

$$evetease_i = \beta_0 + \beta_1 res_woman_i + \beta_2 age_i + \varepsilon_i$$

$$bicycle_i = \gamma_0 + \gamma_1 res_woman_i + \gamma_2 age_i + u_i$$

- ▶ $evetease_i$ is a dummy variable, =1 if harassment of women is reported. $bicycle_i$ is a dummy variable, =1 if a stolen bicycle is reported.
- ▶ If women's political representation has an impact, we would expect to see more reporting of crimes against women ($\beta_1 > 0$) but not of gender-neutral crimes ($\gamma_1 = 0$) in villages with female leaders.

Dummy dependent variable model: summary statistics

variable name variable label

res_woman	=1 if head of village council is reserved for a woman
evetease	=1 if will probably or definitely file a report for eveteasing
bicycle	=1 if will probably or definitely file a report for stolen bicycle

```
. sum res_woman evetease bicycle
```

Variable	Obs	Mean	Std. Dev.	Min	Max
res_woman	4607	.3542435	.4783354	0	1
evetease	4607	.7712177	.4200943	0	1
bicycle	4607	.7668765	.4228661	0	1

Reporting eveteasing: LPM estimates

```
. reg evetease res_woman age, robust
```

Linear regression

Number of obs = 4607
 F(2, 4604) = 3.15
 Prob > F = 0.0429
 R-squared = 0.0014
 Root MSE = .41989

evetease	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
res_woman	.0256412	.0127807	2.01	0.045	.000585	.0506975
age	-.0018945	.0012414	-1.53	0.127	-.0043283	.0005394
_cons	.8085021	.0311677	25.94	0.000	.7473985	.8696057

Reporting eveteasing: LPM

```
. predict yhat1
(option xb assumed; fitted values)
```

```
. sum yhat1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat1	4607	.7712177	.0156233	.7308292	.8151987

Is the LPM model appropriate? Yes, because no predicted values < 0 or > 1 .

Reporting eve-teasing: conclusion

- ▶ With a female leader of the village council, harassment of women is significantly more likely to be reported: the estimate is around 2.5 percentage points, *cet. par.* on age.
 - ▶ National Indian statistics indicate that the number of rapes and other sexual assaults have steadily increased since 2012 (as reported [here](#)). This has been interpreted as an increase in reporting frequencies as sexual violence against women has become less of a taboo and more openly discussed (e.g. see [here](#) (in Dutch)).
- ▶ The evidence would be even stronger if we additionally found no such effect for crimes that are not related to gender, such as property crimes (stolen bicycles).

Reporting stolen bicycles: LPM estimates

```
. reg bicycle res_woman age, robust
```

Linear regression

Number of obs = 4607
 F(2, 4604) = 2.79
 Prob > F = 0.0614
 R-squared = 0.0012
 Root MSE = .4227

bicycle	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
res_woman	.0151662	.0129328	1.17	0.241	-.0101883	.0405207
age	-.0025666	.0012569	-2.04	0.041	-.0050308	-.0001025
_cons	.824323	.0315535	26.12	0.000	.7624631	.8861829

Reporting stolen bicycles: LPM

```
. predict yhat2  
(option xb assumed; fitted values)
```

```
. sum yhat2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat2	4607	.7668765	.0148493	.7190915	.813823

Is the LPM model appropriate? Yes, because no predicted values < 0 or > 1 .

Reporting stolen bicycles: interpretation

- ▶ We find **no statistically significant effect** of the political representation of women on the probability of reporting a property crime.
- ▶ This further **strengthens our previous conclusion** that increased political representation of women may be a way to reduce gender violence.
- ▶ (Note that **the effect of age is significant**, indicating that people who are 1 year older are around 0.25 percentage points less likely to report a stolen bicycle, cet. par.)

Reducing gender violence in India: further evidence

- ▶ Cable TV access may also be improving women's position in rural Indian societies: Jensen, Robert and Emily Oster (2009). "The Power of TV: Cable Television and Women's Status in India," *Quarterly Journal of Economics*, 124(3), p. 1057-1094.
 - ▶ Non-technical summaries of this paper can be found [here](#) and [here](#).
- ▶ Finally, there is further recent evidence that women may have different policy priorities: female jurors provide [harsher](#) sentencing for violence against women than male jurors do.

Estimating Linear Probability Models

- ▶ Whenever you use OLS to estimate a model with a dummy dependent variable, you are estimating a Linear Probability Model (LPM).
- ▶ When estimating a LPM, don't forget to:
 - ▶ Correct for heteroskedasticity (, **robust**): else the standard errors are biased.
 - ▶ **Check the fitted values**: if more than a few percent fall outside of the 0-1 range, the OLS estimator is not appropriate.

Causality revisited

- ▶ In applied econometric analysis we are often after **causal effects**: e.g. does increasing women's political representation lead to an increased incidence of reporting of gender crimes?
- ▶ In this course, we have seen that **uncovering these effects from observational data can be challenging**.
- ▶ The most important reason is the assumption that $\text{Corr}(\varepsilon_i, X_i) = 0$: **omitted variable bias** threatens the interpretation of our estimates as causal.³

³More generally, **model specification is crucial**: this is why you have spent a lot of time in your project paper honing the specification of your model. ▶

Observational versus experimental data

Consider two cases where OLS is used to estimate the following model:

$$call_i = \beta_0 + \beta_1 black_i + \varepsilon_i$$

- Case 1:** Using an **observational** dataset of job applicants, a dummy for being African-American ($black_i$) and whether they received a call-back for their latest job application ($call_i$).
- Case 2:** Using data from the CV **experiment**, where fake CVs were sent out to help-wanted ads and call-backs were recorded ($call_i$): the last names of applicants, some of which are African-American sounding ($black_i$), are randomly assigned to CVs.

Observational versus experimental data

- ▶ If we find a $\beta_1 < 0$ in case 1, this could be because African-American applicants on average have lower education levels and/or less relevant work experience, whereas education and experience themselves increase call-back rates. We can then try to control for such omitted variables but we should always be worried that we are not controlling for everything that affects call-back rates and is correlated with race.
- ▶ In case 2, we are not worried about omitted variables since experimentation randomly assigns values of the independent variable of interest (here, the racial identity of the applicant's name) to individual observations (here, applicant CVs) such that $\text{Corr}(\varepsilon_i, \text{black}_i) = 0$. As such, we can interpret $\beta_1 < 0$ as causal.

Natural experiments

- ▶ Experiments are not always possible, however: another approach increasingly used by economists are so-called **natural experiments**.
- ▶ Natural experiment = **random variation contained within observational data** (as opposed to experiments designed by the researchers themselves).
- ▶ One example is the political representation of women in West-Bengal: since the reservation of seats for women was random across towns, this is a natural experiment.

Causality: caution

- ▶ Finally, note that we are not *always* interested in causal effects.
- ▶ For example, we may just want to know how large the labor market disadvantage is for African Americans (whether that be due to lower educational attainment, lower work experience, or other factors).
- ▶ In that case, we do not want to use a (natural) experiment and the OLS estimator of β_1 using observational data (i.e. case 1) uncovers the parameter of interest.

Course overview

- ▶ The following slides contain an overview of the main topics covered in this course.
- ▶ (Of course, this overview is non-exhaustive.)

Cross-sectional analysis – weeks 1-5

- ▶ Rewriting expressions for expected value, (co)variance, and correlation.
- ▶ OLS and its assumptions (first 4 assumptions required for unbiasedness of $\hat{\beta}$; all 6 required for unbiasedness of $se(\hat{\beta})$).
- ▶ Violations of OLS assumptions (should know consequences, diagnosis, solutions):
 - ▶ Perfect (multi)collinearity
 - ▶ Omitted variable bias & causality
 - ▶ Heteroskedasticity

Cross-sectional analysis – weeks 1-5

- ▶ Interpretation of coefficients in level-level, log-log, level-log, log-level models; and models with quadratic terms.
- ▶ Hypothesis testing: t-test, F-test, Chow test.
- ▶ Interpretation of coefficients on dummy variables, and interaction terms.
- ▶ Multicollinearity (should know consequences, diagnosis, solution).

Cross-sectional analysis – week 8

- ▶ Interpretation of coefficients in dummy dependent variable model (LPM).
- ▶ Disadvantages of LPM.

Time-series analysis – weeks 6-7

- ▶ Interpretation of static and dynamic models (calculation of short-run versus long-run effect).
- ▶ Violations of OLS assumptions:
 - ▶ Serial correlation (should know consequences, diagnosis, solution).
 - ▶ Non-stationarity (should know consequences, diagnosis, solution): timetrends and unit roots.
- ▶ Model specification (levels or first differences) to avoid spurious regression.

Exam structure

- ▶ 2 hours
- ▶ 4 questions
 - ▶ Econometric theory questions: the OLS estimator & OLS assumptions, calculation rules for expectations, variances, etc.
 - ▶ Applied econometric questions: interpreting Stata output, performing statistical tests, providing advice on model specification, etc.
 - ▶ Approximately 75% of points for cross-sectional analysis, 25% for time-series.
- ▶ Sample exams are available on Blackboard and two former exam questions are practiced in the tutorial of week 8.

General exam advice

- ▶ **Partial credit is awarded:** if you do not know the full answer, at least outline some of the first steps towards it – if these are correct, partial credit is typically awarded.
- ▶ It also follows that **full credit is only awarded for a complete answer.** Two common examples:
 - ▶ In the case of statistical tests, you are expected to write down the model, H_0 & H_A , the test statistic and its critical value based on the chosen α (alternatively, if provided, the test's p-value), the rejection rule, and a conclusion in words.
 - ▶ When asked to provide an economic interpretation, consider the model specification (i.e. statements along the lines of "a one unit increase in x" are incorrect as an economic interpretation requires assessing the units, whether the model is in logs or levels, whether the variable enters only linearly or also quadratically, etc.).

General exam advice

- ▶ **Read questions carefully** and **be precise** in your answers: here are some typical mistakes that can cost you a lot of points throughout the exam:
 - ▶ incorrectly reading the units of a variable;
 - ▶ forgetting to add "cet. par " in coefficient interpretations for multivariate models;
 - ▶ copying the wrong numbers from Stata output (e.g. RSS , k , n , when calculating an F-statistic);
 - ▶ mixing up population and sample notation (e.g. confusing the error term with the residual);
 - ▶ confusing OLS' statistical and algebraic properties.

LPM: heteroskedastic errors proof

Why are the errors in a LPM heteroskedastic? First, consider that there are only 2 possible cases for the error term:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ \varepsilon_i &= \begin{cases} -\beta_0 - \beta_1 X_i & \text{if } Y_i = 0 \\ 1 - \beta_0 - \beta_1 X_i & \text{if } Y_i = 1 \end{cases} \end{aligned}$$

Now, write out the expression for the error variance:

$$\begin{aligned} \text{Var}(\varepsilon_i) &= E(\varepsilon_i - E(\varepsilon_i))^2 \\ &= E(\varepsilon_i)^2 \quad (\text{since } E(\varepsilon_i) = 0) \end{aligned}$$

LPM: heteroskedastic errors proof

$$\text{Var}(\varepsilon_i) = E(\varepsilon_i)^2$$

This expectation can be written as the sum of the two possible error terms, squared, and weighted by their respective probabilities:

$$\begin{aligned}\text{Var}(\varepsilon_i) &= \Pr(Y_i = 0) \times (-\beta_0 - \beta_1 X_i)^2 \\ &\quad + \Pr(Y_i = 1) \times (1 - \beta_0 - \beta_1 X_i)^2\end{aligned}$$

Note that $\Pr(Y_i = 0) + \Pr(Y_i = 1) = 1$ such that $\Pr(Y_i = 0) = 1 - \Pr(Y_i = 1)$. This gives:

$$\begin{aligned}\text{Var}(\varepsilon_i) &= [1 - \Pr(Y_i = 1)] \times (-\beta_0 - \beta_1 X_i)^2 \\ &\quad + \Pr(Y_i = 1) \times (1 - \beta_0 - \beta_1 X_i)^2\end{aligned}$$

LPM: heteroskedastic errors proof

$$\begin{aligned} \text{Var}(\varepsilon_i) &= [1 - \Pr(Y_i = 1)] \times (-\beta_0 - \beta_1 X_i)^2 \\ &\quad + \Pr(Y_i = 1) \times (1 - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

Consider that in a LPM, $\Pr(Y_i = 1) = E(Y_i) = \beta_0 + \beta_1 X_i$, hence we can write this as

$$\begin{aligned} \text{Var}(\varepsilon_i) &= [1 - \beta_0 - \beta_1 X_i] \times (-\beta_0 - \beta_1 X_i)^2 \\ &\quad + (\beta_0 + \beta_1 X_i) \times (1 - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

factoring,

$$\begin{aligned} \text{Var}(\varepsilon_i) &= (1 - \beta_0 - \beta_1 X_i) \times (\beta_0 + \beta_1 X_i) \\ &\quad \times [(\beta_0 + \beta_1 X_i) + (1 - \beta_0 - \beta_1 X_i)] \end{aligned}$$

LPM: heteroskedastic errors proof

$$\begin{aligned} \text{Var}(\varepsilon_i) &= (1 - \beta_0 - \beta_1 X_i) \times (\beta_0 + \beta_1 X_i) \\ &\quad \times [\beta_0 + \beta_1 X_i + 1 - \beta_0 - \beta_1 X_i] \end{aligned}$$

$$\text{Var}(\varepsilon_i) = (1 - \beta_0 - \beta_1 X_i) \times (\beta_0 + \beta_1 X_i)$$

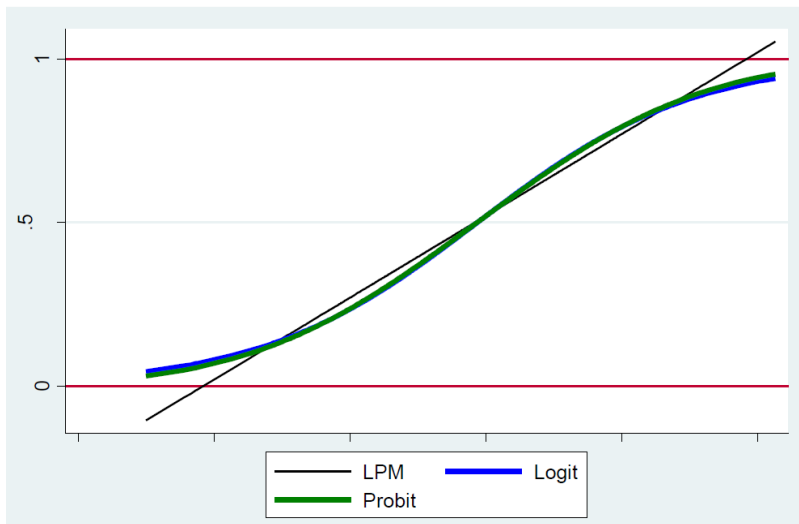
From this expression it can be seen that the variance of the error term depends on X_i , i.e. the error variance is not a constant σ^2 : there is **heteroskedasticity in the LPM**.

These extra slides are included here for those interested in logit and probit estimation (e.g. you may encounter it one day while reading a research paper for your Bachelor thesis or for another course). This not exam material hence **no questions will be asked about it on the exam.**

Logit and probit

- ▶ **Instead of modeling the probability linearly** (with OLS), we can **model it non-linearly**.
- ▶ These non-linear models use **functions which lie between 0 and 1**: as such, the predicted probabilities never lie outside the 0-1 interval.
- ▶ **Logit and probit** models use different functions:
 - ▶ Logit uses the cumulative standard logistic distribution function
 - ▶ Probit uses the cumulative standard normal distribution function
 - ▶ Both functions have similar shapes.

Linear (LPM) vs non-linear functions (logit and probit)



Logit and probit

- ▶ Since these functions are non-linear, we **cannot use a linear estimator** such as OLS!
- ▶ Instead, we use **maximum likelihood estimation** (an iterative procedure).
- ▶ The disadvantage is we **cannot directly interpret the estimated coefficients**, unlike in linear regression.

Logit and probit interpretation

- ▶ Slope coefficients can be interpreted as the effect of a unit of change in the X variable on the predicted **transformed dependent variable** (either logits or probits) with the other variables in the model held constant.
- ▶ To find the effect on the probabilities (as given by the coefficients in the LPM), we need to **calculate the marginal effects**.
- ▶ This is easily done in Stata using the **mfxf command** following logit or probit estimation.

Dummy dependent variable model: logit estimates

```
. logit call black pcol linc_emp
```

```
Iteration 0: log likelihood = -445.60886
Iteration 1: log likelihood = -435.9182
Iteration 2: log likelihood = -435.46422
Iteration 3: log likelihood = -435.46335
Iteration 4: log likelihood = -435.46335
```

Logistic regression

Log likelihood = -435.46335

```
Number of obs   =      1870
LR chi2(3)      =      20.29
Prob > chi2     =      0.0001
Pseudo R2      =      0.0228
```

call	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
black	-.5329805	.1952285	-2.73	0.006	-.9156213	-.1503396
pcol	.0141096	.0050038	2.82	0.005	.0043024	.0239169
linc_emp	-.4618463	.2147819	-2.15	0.032	-.8828112	-.0408815
_cons	2.127714	2.279064	0.93	0.351	-2.339169	6.594597

Dummy dependent variable model: logit marginal effects

However, to be able to interpret the coefficients as effects on the probability of being called back, we **calculate the marginal effects**:

```
. mfx
```

Marginal effects after logit

```
y = Pr(call) (predict)
    = .05969399
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]		x
black*	-.0301434	.01094	-2.76	0.006	-.051577	-.00871	.499465
pcol	.000792	.00028	2.87	0.004	.000251	.001333	21.397
linc_emp	-.0259237	.01189	-2.18	0.029	-.049224	-.002624	10.6537

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Dummy dependent variable model: logit predicted values

Predicted values from the logit model always lie in the 0-1 interval:

```
. predict yhat_logit  
(option pr assumed; Pr(call))  
  
. sum yhat_logit
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat_logit	1870	.0641711	.0264945	.0214964	.1744567

Dummy dependent variable model: probit estimates

```
. probit call black pcol linc_emp
```

```
Iteration 0:   log likelihood = -445.60886
```

```
Iteration 1:   log likelihood = -435.29963
```

```
Iteration 2:   log likelihood = -435.19382
```

```
Iteration 3:   log likelihood = -435.1938
```

Probit regression

Number of obs = 1870

LR chi2(3) = 20.83

Prob > chi2 = 0.0001

Pseudo R2 = 0.0234

Log likelihood = -435.1938

call	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
black	-.2669647	.0929276	-2.87	0.004	-.4490994	-.0848299
pcol	.007084	.0025036	2.83	0.005	.002177	.011991
linc_emp	-.2222178	.1034262	-2.15	0.032	-.4249294	-.0195062
_cons	.7956528	1.101896	0.72	0.470	-1.364024	2.95533

Dummy dependent variable model: probit marginal effects

To be able to interpret the coefficients as effects on the probability of being called back, we **calculate the marginal effects**:

```
. mfx
```

Marginal effects after probit

```
y = Pr(call) (predict)
    = .06014466
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	x
black*	-.0319879	.01107	-2.89	0.004	-.053686	-.01029		.499465
pcol	.0008455	.0003	2.85	0.004	.000264	.001427		21.397
linc_emp	-.0265212	.01226	-2.16	0.031	-.050558	-.002485		10.6537

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Dummy dependent variable model: probit predicted values

Predicted values from the probit model always lie in the 0-1 interval:

```
. predict yhat_probit  
(option pr assumed; Pr(call))
```

```
. sum yhat_probit
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat_probit	1870	.0641725	.0269917	.019002	.1675266

Summary

- ▶ The **marginal effect of an African-American sounding name on the probability of being called back** is: -0.031 in the LPM; -0.030 in the logit model; -0.032 in the probit model.
- ▶ In this case, the different estimators produce very similar results.
- ▶ Note that logit and probit estimators also have drawbacks which are not discussed here.