Functional Forms and Dummies

Panel Data Econometrics

Prof. Alexandre Gori Maia

State University of Campinas

UNICAMP

Topics

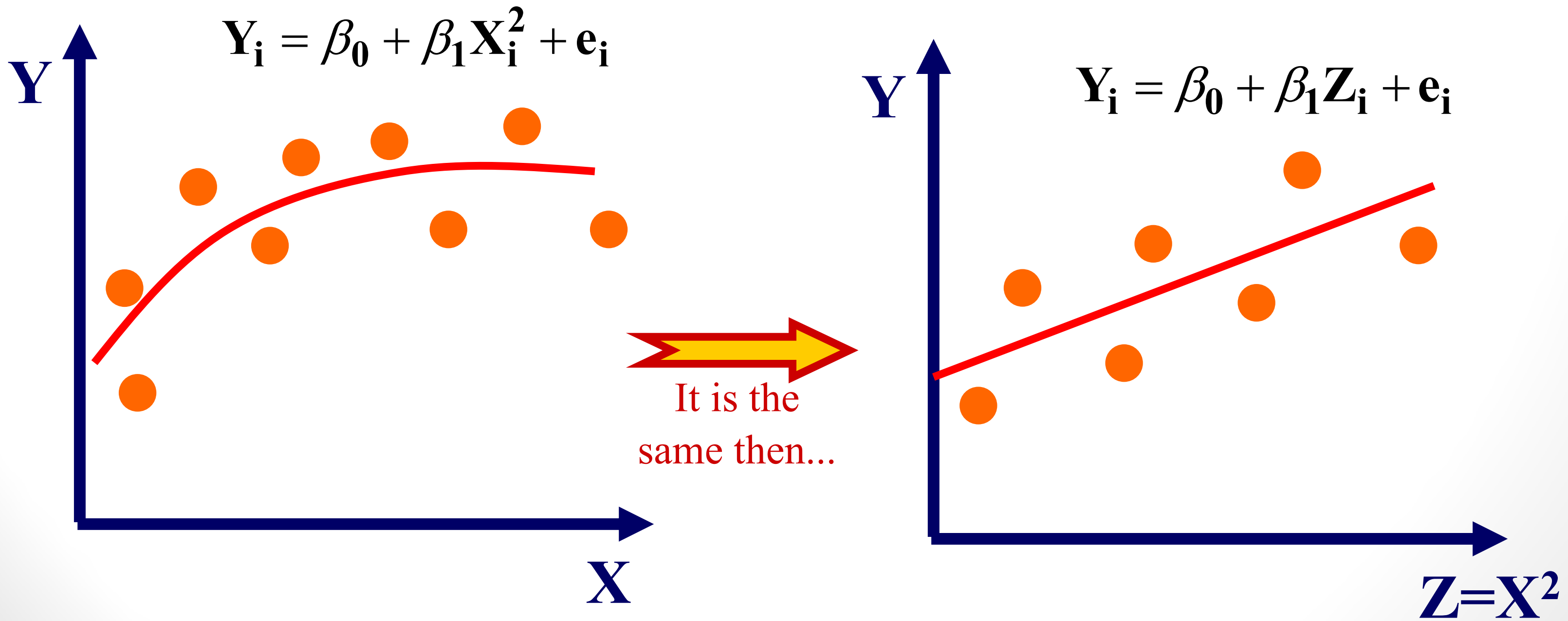
- Functional Forms
- Binary Variables
- Binaries in logarithmic models

Reference

Maia, A. G. 2014. Econometria: conceitos e aplicações. Insittuto de Economica. Caps. 1 a 8.

Linear Relation

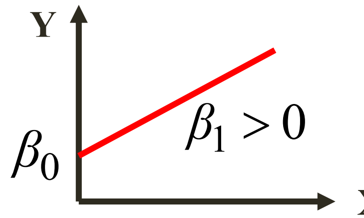
- When the relation between *Y* and *X* is non-linear, we can transform the variables to obtain a linear relation;
- The type of transformation (functional form) depends on the (i) theoretical assumptions; (ii) distribution of the variables;



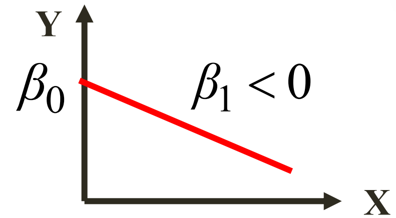
Functional Forms - Examples

1) Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

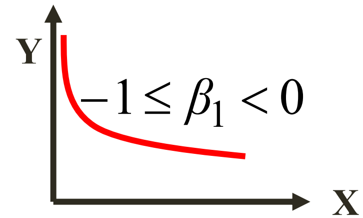


ou

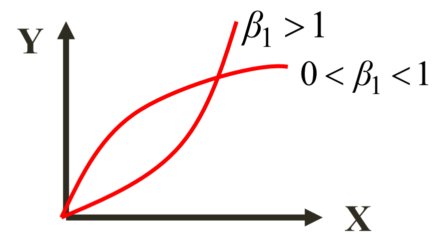


2) Log-log model:

$$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + e_i$$

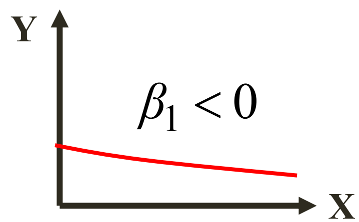


ou

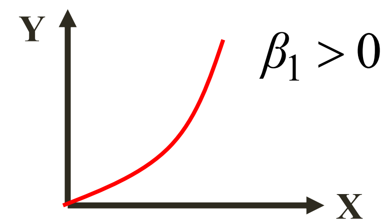


3) Log-lin model:

$$ln(Y_i) = \beta_0 + \beta_1 X_i + e_i$$

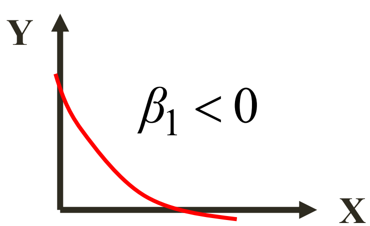


ou

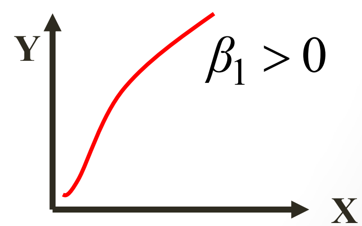


4) Lin-log model:

$$Y_i = \beta_0 + \beta_1 ln(X_i) + e_i$$

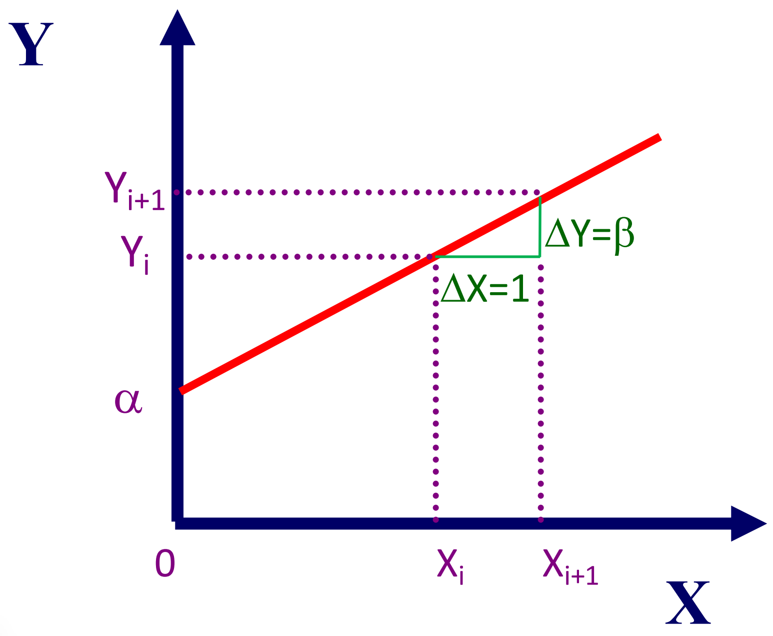


ou



Linear Model - Interpretation

- The linear model assumes that absolute changes in *X* imply in absolute changes in *Y*;
- The marginal variation in *Y* is the same for all values of *X*;



$$E[Y/0] = \alpha + \beta(0) = \alpha$$

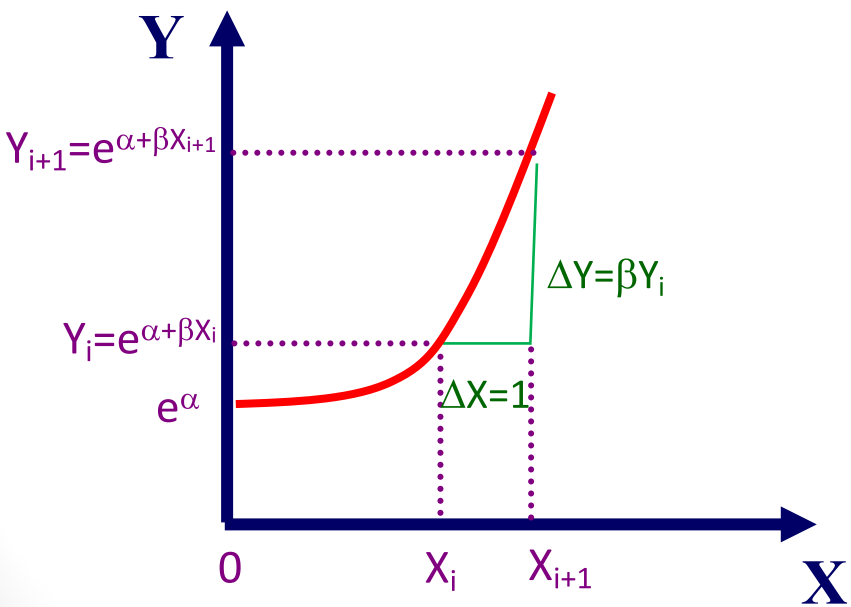
α is the expected value of *Y* when *X=0*

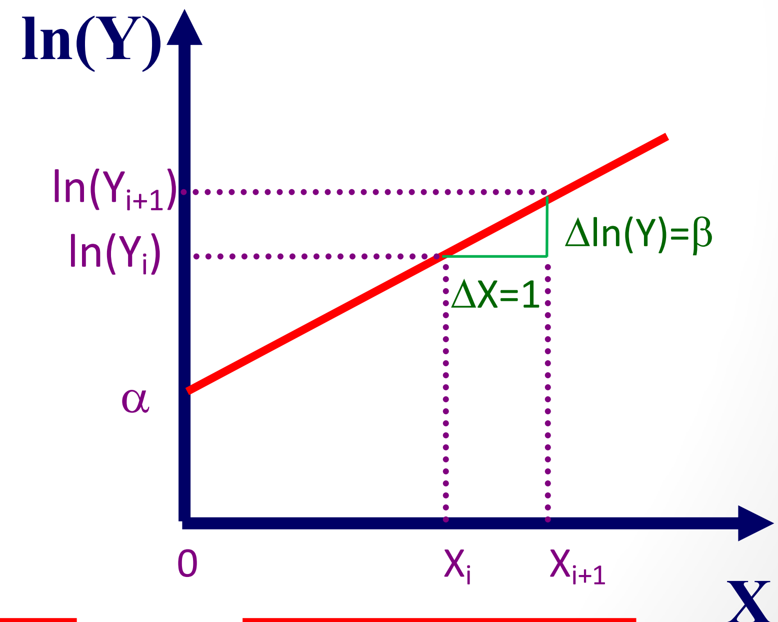
$$\frac{\Delta Y}{\Delta X} = \frac{dY}{dX} = \frac{d(\alpha + \beta X)}{dX} = \beta$$

β is the marginal variation in *Y* (*ΔY*) for each unit variation in *X* (*ΔX*=1).

Log-lin Model - Interpretation

- The log-lin model assumes that absolute changes in *X* imply in relative changes (%) in *Y*;
- The absolute change in *Y* is different for each value of *X*;
- For example, a change of 0,01 in ln(*Y*) means a change of 1% in *Y*;

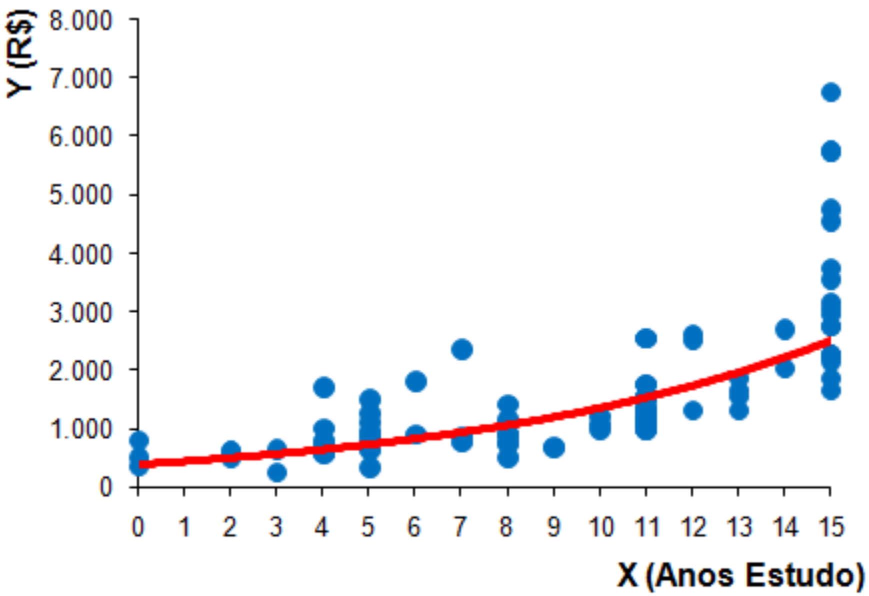




Small (infinitesimal) changes in ln*Y* mean relative changes in *Y*. This means: $\boxed{\Delta\ln(Y)=\dfrac{\Delta Y}{Y_i}}$ then... $\boxed{\beta=\dfrac{\ln(Y_i)}{\Delta X}=\dfrac{\Delta Y/Y_i}{\Delta X}}$

Log-Lin Model - Example

What is the marginal return of education?



Assuming that the wage (Y, in monthly R$) grows exponentially with the years of education (X):

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + e_i$$

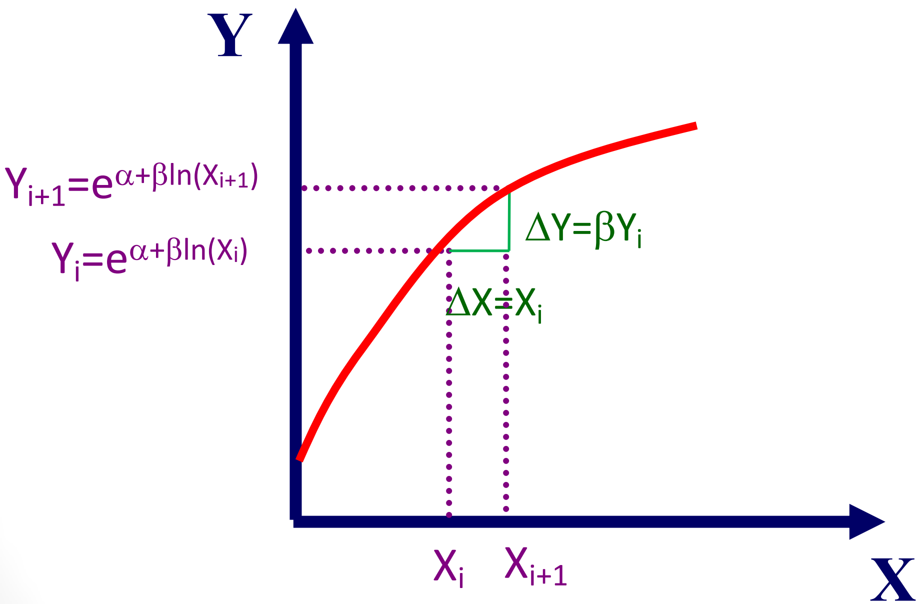
And a sample estimate:

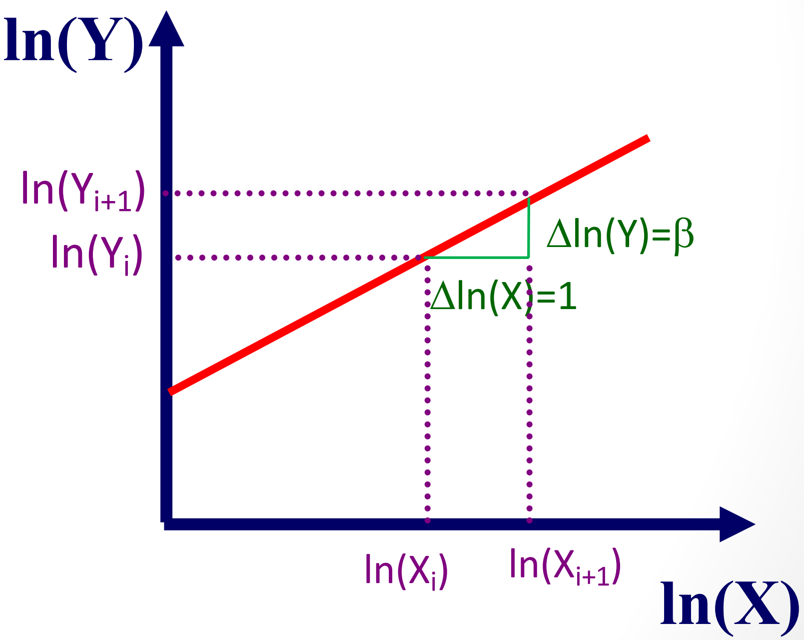
$$\ln(Y_i) = 6{,}006 + 0{,}121 X_i + \hat{e}_i$$

We can expect that, for each additional year of education (ΔX=1), the wage grows by 12,1% (ΔY=0,121 Y_i).

Log-Log Model - Interpretation

- The log-log model assumes that relative changes (%) in *X* imply in relative changes (%) in *Y*;
- The beta coefficient can be interpreted as a constant elasticity between *Y* and *X*, i.e., the percentage change in *Y* for 1% change in *X*.





If $\Delta\ln(X)=\dfrac{\Delta X}{X_i}$ and $\Delta\ln(Y)=\dfrac{\Delta Y}{Y_i}$ then $\boxed{\beta=\dfrac{\Delta\ln(Y)}{\Delta\ln(X)}=\dfrac{\Delta Y/Y_i}{\Delta X/X_i}}$

Example – Stata

- Adjusting logarithmic models in Stata:

```
* create log variables
generate lnco2 = log(co2)
generate lngdp = log(gdp)

* linear regression
regress co2 gdp ind

* log-lin regression
regress lnco2 gdp ind

* lin-log regression
regress co2 lngdp ind

* log-log model
regress lnco2 lngdp ind
```

Example – R

- Adjusting logarithmic models in R:

```
# create log variables
countries$lnco2 <- log(countries$co2)
countries$lngdp <- log(countries$gdp)

# linear model
linear <- lm(co2 ~ gdp + ind, data=countries)
summary(linear)

# log-lin model
loglin <- lm(lnco2 ~ gdp + ind, data=countries)
summary(loglin)

# lin-log model
linlog <- lm(co2 ~ lngdp + ind, data=countries)
summary(linlog)

# log-log model
loglog <- lm(lnco2 ~ lngdp + ind, data=countries)
summary(loglog)
```

Example - Python

- Adjusting logarithmic models in Python:

```
# modeulo for mathematical (log) operations
import numpy as np

# creating log variables
countries['lnco2'] = countries['co2'].apply(np.log)
countries['lngdp'] = countries['gdp'].apply(np.log)

# linear model
x = countries[['gdp','ind']]
y = countries['co2']
x = sm.add_constant(x)
linear = sm.OLS(y, x).fit()
print(linear.summary())

# log-lin model
x = countries[['gdp','ind']]
y = countries['lnco2']
x = sm.add_constant(x)
loglin = sm.OLS(y, x).fit()
print(loglin.summary())

 # lin-log model
x = countries[['lngdp','ind']]
y = countries['co2']
x = sm.add_constant(x)
linlog = sm.OLS(y, x).fit()
print(linlog.summary())

 # log-log model
x = countries[['lngdp','ind']]
y = countries['lnco2']
x = sm.add_constant(x)
loglog = sm.OLS(y, x).fit()
print(loglog.summary())
```

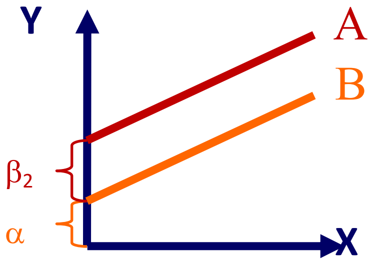
Binary Variables – 2 Categories

- In order to represent two nominal categories (A and B) as independent variables in a regression, we only need one binary variable (*D*).
- The reference of analysis is given by *D*=0;

$$Y_i = \alpha + \beta_1 X_i + \beta_2 D_i + e_i$$

The coefficient β_2 shows the difference between the expected values of Y for the category A (D=1) and the reference category B (D=0).

Category	D_i
A	1
B	0



For A: $Y_i = (\alpha + \beta_2) + \beta_1 X_i + e_i$

For B: $Y_i = \alpha + \beta_1 X_i + e_i$

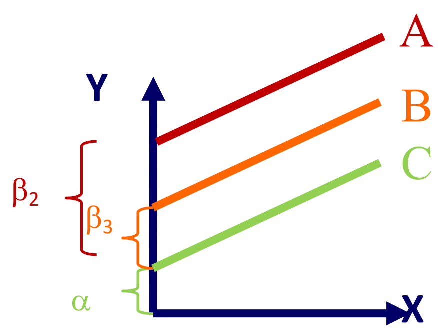
Binary Variables – *k* Categories

- In order to represent *k* nominal categories, we need *k*–1 binary variables.
- The reference of analysis is the nominal category without a binary variable;.

$$Y_i = \alpha + \beta_1 X_i + \beta_2 D_{1_i} + \beta_3 D_{2_i} + e_i$$

Category	D_{1i}	D_{2i}
A	1	0
B	0	1
C	0	0

The coefficient β_2 shows the difference between the expected values of Y for the category A (D_1=1) and the reference category C (D_1=0 and D_2=0).
The coefficient β_3 shows the difference between the expected values of Y for the category B (D_2=1) and the reference category C.



For A: $Y_i = (\alpha + \beta_2) + \beta_1 X_i + e_i$

For B: $Y_i = (\alpha + \beta_3) + \beta_1 X_i + e_i$

For C: $Y_i = \alpha + \beta_1 X_i + e_i$

Binary Variables – Example

The sample refers to the 164 thousand labors in Brazil in 2011:

Root MSE	1960.28180	R-Square	0.1549
Dependent Mean	1294.58777	Adj R-Sq	0.1548
Coeff Var	151.42131		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1340.83445	28.34673	-47.30	<.0001
anosest	1	164.28765	1.23027	133.54	<.0001
idade	1	33.83255	0.40005	84.57	<.0001
feminino	1	-616.38468	10.41720	-59.17	<.0001
branca	1	367.53558	18.00818	20.41	<.0001
parda	1	38.28339	17.67321	2.17	0.0303
amarela	1	618.12772	73.00807	8.47	<.0001
no	1	-18.11296	18.86931	-0.96	0.3371
ne	1	-149.30862	16.49369	-9.05	<.0001
se	1	64.90085	15.35264	4.23	<.0001
co	1	325.01732	19.36538	16.78	<.0001

Suppose the initial model

$$renda_i = \alpha + \beta_1 anosest_i + \beta_2 idade_i + e_i$$

Including a binary for sex:

$$feminino = 1, se\ mulher;\ 0\ se\ homen$$

Including binaries for race (black is the reference):

$$branca = 1, se\ cor\ branca;\ 0\ c.c.$$

$$parda = 1, se\ cor\ parda;\ 0\ c.c.$$

$$amarela = 1, se\ cor\ amarela;\ 0\ c.c.$$

Including binaries for region (South is the reference):

$$no = 1, se\ região\ norte;\ 0\ c.c.$$

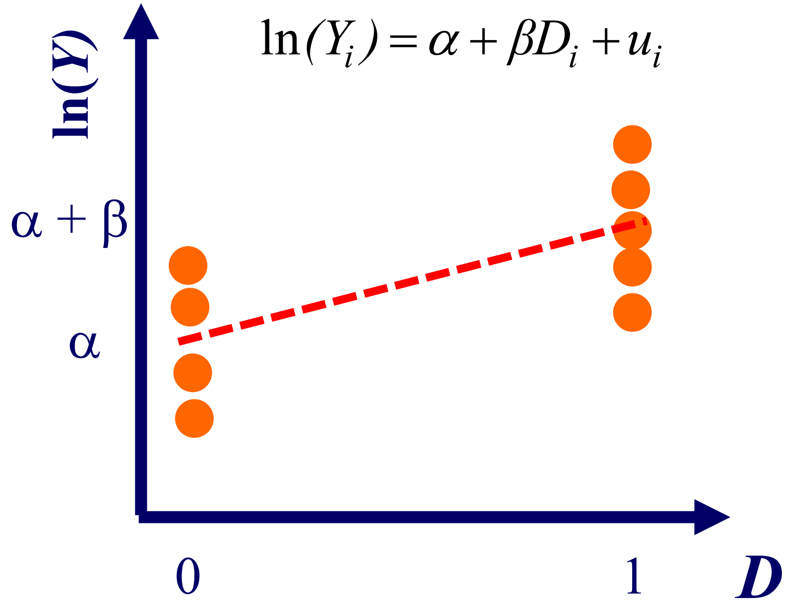
$$ne = 1, se\ região\ nordeste;\ 0\ c.c.$$

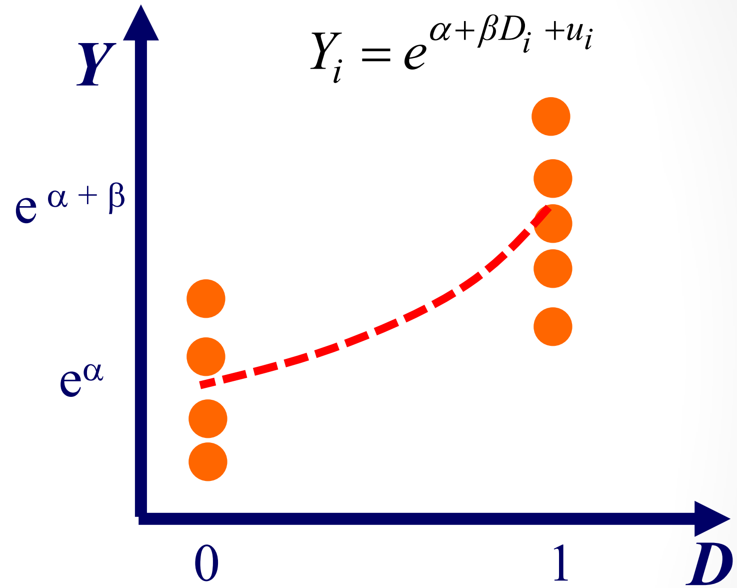
$$se = 1, se\ região\ sudeste;\ 0\ c.c.$$

$$co = 1, se\ região\ centro-oeste;\ 0\ c.c.$$

Binaries in Logarithimic Models

Suppose the log-lin model:





- Where Y_1 is the expected value of Y when D=1 and Y_0 when D=0;
- The interpretation of β is now given by:

For D=0:

$$\ln(Y_0) = \alpha$$
$$Y_0 = e^{\alpha}$$

For D=1:

$$\ln(Y_1) = \alpha + \beta$$
$$Y_1 = e^{\alpha+\beta}$$

Then:

$$\frac{Y_1 - Y_0}{Y_0} = \frac{e^{\alpha} e^{\beta} - e^{\alpha}}{e^{\alpha}} \Rightarrow \boxed{\frac{\Delta Y}{Y} = \frac{Y_1 - Y_0}{Y_0} = e^{\beta} - 1}$$

Binary in Log-Lin – Example

The sample refers to the 164 thousand labors in Brazil in 2011:

Root MSE	0.72061	R-Square	0.3600
Dependent Mean	6.71768	Adj R-Sq	0.3600
Coeff Var	10.72700		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.45369	0.01042	523.37	<.0001
anosest	1	0.10421	0.00045225	230.41	<.0001
idade	1	0.01596	0.00014706	108.55	<.0001
feminino	1	-0.44332	0.00383	-115.77	<.0001
branca	1	0.14130	0.00662	21.34	<.0001
parda	1	-0.00863	0.00650	-1.33	0.1841
amarela	1	0.22277	0.02684	8.30	<.0001
no	1	-0.16213	0.00694	-23.37	<.0001
ne	1	-0.37063	0.00606	-61.13	<.0001
se	1	-0.00458	0.00564	-0.81	0.4174
co	1	0.07583	0.00712	10.65	<.0001

Supposing the model:

$$\ln(renda_i) = \alpha + \beta_1 anosest_i + \beta_2 idade_i + e_i$$

In relation the male workers, the expected difference in % for sex is:

$$e^{-0,44332} - 1 = -0,3581$$

In relation to black workers, the expected differences in % for race are:

branca: $e^{0,14130} - 1 = 0,1518$

parda: $e^{-0,0086} - 1 = -0,0086$

amarela: $e^{0,2228} - 1 = 0,2495$

In relation to where workers in the South, the expected differences in (%) are:

NO: $e^{-0,1621} - 1 = -0,1497$

NE: $e^{-0,3706} - 1 = -0,3097$

SE: $e^{-0,0046} - 1 = -0,0046$

CO: $e^{0,0758} - 1 = 0,0788$

Example – Stata and R

- Model with binary variable in Stata:

```
* create binary variable: 1 for country in BRICS
generate brics = 0
replace brics = 1 if country=="BRA" | country=="RUS" | ///
                     country=="IND" | country=="CHN" | ///
                     country=="ZAF"

* log-log model with binary
regress lnco2 lngdp ind brics
```

- Model with binary variable in R:

```
# create binary variable: 1 for country in BRICS
countries$brics <- 0
countries$brics[countries$country=='BRA' | countries$country=='RUS' |
                countries$country=='IND' | countries$country=='CHN' |
                countries$country=='ZAF'] <- 1

# log-log model with binary
loglog2 <- lm(lnco2 ~ lngdp + ind + brics, data=countries)
summary(loglog2)
```

Example - Python

- Model with binary variable in Python:

```
# create binary: 1 for countr in BRICS
countries['brics'] = 0
countries.loc[(countries['country']=='BRA') | \
              (countries['country']=='RUS') | \
              (countries['country']=='IND') | \
              (countries['country']=='CHN') | \
              (countries['country']=='ZAF'), ['brics']] = 1

 # log-log model with binary variable
x = countries[['lngdp','ind','brics']]
y = countries['lnco2']
x = sm.add_constant(x)
loglog2 = sm.OLS(y, x).fit()
print(loglog2.summary())
```

Exercise

1) The dataset *Data_TravelCosts.csv* contains information on travel costs from several municipalities to a national park in Brazil (see MAIA, A. G. , ROMEIRO, A. . Validade e confiabilidade do método de custo de viagem: um estudo aplicado ao Parque Nacional da Serra Geral. Revista de Economia Aplicada, v. 12, p. 103-123, 2008):
   a) Which functional form presents the best goodness of fit measures?
   b) Create a binary variable *RS* that assumes 1 when the municipality is in the state of Rio Grande do Sul (the two first digits of the variable *code* must be equal to 43). Add this variable in the model. Is there any signficant change?