Functional Forms and Dummies

Panel Data Econometrics
Prof. Alexandre Gori Maia
State University of Campinas

Topics

- Functional Forms
- Binary Variables
- Binaries in logarithmic models

Reference

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;

$Y_i = \beta_0 + \beta_1 X_i^2 + e_i$

It is the same then...
Functional Forms - Examples

1) Linear model:
\[ Y_i = \beta_0 + \beta_1 X_i + e_i \]

2) Log-log model:
\[ \ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + e_i \]

3) Log-lin model:
\[ \ln(Y_i) = \beta_0 + \beta_1 X_i + e_i \]

4) Lin-log model:
\[ Y_i = \beta_0 + \beta_1 \ln(X_i) + e_i \]
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 \( X \);

\[
E[Y / 0] = \alpha + \beta(0) = \alpha
\]
\( \alpha \) is the expected value of \( Y \) when \( X=0 \)

\[
\frac{\Delta Y}{\Delta X} = \frac{dY}{dX} = \frac{d(\alpha + \beta X)}{dX} = \beta
\]
\( \beta \) is the marginal variation in \( Y (\Delta Y) \) for each unit variation in \( X (\Delta X=1) \).
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 $Y$;

\[ Y_{i+1} = e^{\alpha + \beta X_{i+1}} \]
\[ Y_i = e^{\alpha + \beta X_i} \]
\[ \Delta Y = \beta Y_i \]
\[ \Delta X = 1 \]

Small (infinitesimal) changes in $\ln Y$ mean relative changes in $Y$. This means:

\[ \Delta \ln(Y) = \frac{\Delta Y}{Y_i} \]
\[ \beta = \frac{\ln(Y_i)}{\Delta X} = \frac{\Delta Y / Y_i}{\Delta X} \]
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:

$$\ln(Y_i) = 6.006 + 0.121X_i + \hat{e}_i$$

We can expect that, for each additional year of education ($\Delta X=1$), the wage grows by 12.1% ($\Delta Y=0.121 Y_i$).
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 $X$.

\[
\begin{align*}
\Delta \ln(X) &= \frac{\Delta X}{X_i} \\
\Delta \ln(Y) &= \frac{\Delta Y}{Y_i}
\end{align*}
\]

If

\[
\beta = \frac{\Delta \ln(Y)}{\Delta \ln(X)} = \frac{\Delta Y / Y_i}{\Delta X / X_i}
\]
Example – Stata

- Adjusting logarithmic models in Stata:

```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:

```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:

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

# linear-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 $\beta_2$ shows the difference between the expected values of $Y$ for the category A ($D=1$) and the reference category B ($D=0$).

<table>
<thead>
<tr>
<th>Category</th>
<th>$D_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
</tr>
</tbody>
</table>

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.

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

The coefficient $\beta_2$ 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 $\beta_3$ shows the difference between the expected values of $Y$ for the category B ($D_2=1$) and the reference category C.

<table>
<thead>
<tr>
<th>Category</th>
<th>$D_{1i}$</th>
<th>$D_{2i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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:

Suppose the initial model

\[ \text{renda}_i = \alpha + \beta_1 \text{anosest}_i + \beta_2 \text{idade}_i + e_i \]

Including a binary for sex:

\[ \text{feminino} = 1, \text{se mulher}; 0 \text{ se homem} \]

Including binaries for race (black is the reference):

\[ \text{branca} = 1, \text{se cor branca}; 0 \text{ c.c.} \]
\[ \text{parda} = 1, \text{se cor parda}; 0 \text{ c.c.} \]
\[ \text{amarela} = 1, \text{se cor amarela}; 0 \text{ c.c.} \]

Including binaries for region (South is the reference):

\[ \text{no} = 1, \text{se região norte}; 0 \text{ c.c.} \]
\[ \text{ne} = 1, \text{se região nordeste}; 0 \text{ c.c.} \]
\[ \text{se} = 1, \text{se região sudeste}; 0 \text{ c.c.} \]
\[ \text{co} = 1, \text{se região centro-oeste}; 0 \text{ c.c.} \]
Binaries in Logarithmic Models

Suppose the log-lin model:

\[ \ln(Y_i) = \alpha + \beta D_i + u_i \]

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:

\[ \frac{Y_1 - Y_0}{Y_0} = \frac{e^\alpha e^\beta - e^\alpha}{e^\alpha} = e^\beta - 1 \]

Where \( Y_1 \) is the expected value of \( Y \) when \( D = 1 \) and \( Y_0 \) when \( D = 0 \);

The interpretation of \( \beta \) is now given by:

\[ \Delta Y = \frac{Y_1 - Y_0}{Y_0} = e^\beta - 1 \]
Binary in Log-Lin – Example

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

Supposing the model:

\[ \ln(\text{renda}_i) = \alpha + \beta_1 \text{anosest}_i + \beta_2 \text{idade}_i + e_i \]

In relation to 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:

```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 lnc02 lngdp ind brics
```

• Model with binary variable in R:

```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(lnc02 ~ lngdp + ind + brics, data=countries)
summary(loglog2)
```
Example - Python

• Model with binary variable in Python:

```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 significant change?