Short Guides to Microeconometrics
Fall 2022

Kurt Schmidheiny
Unversity of Basel

Functional Form in the Linear Model

matrix-free

Version: 14-10-2022, 08:09

1 Introduction

Despite its name, the classical *linear* regression model, is not limited to a linear relationship between the dependent and the explanatory variables.

Consider K variables x_{i1} x_{i2} ... x_{iK} for each observation i. The L functions $f_1(x_{i1}... \, x_{iK}), f_2(x_{i1}... \, x_{iK}), ..., f_L(x_{i1}... \, x_{iK})$ map the K explanatory variables into L new variables z_{i1}, z_{i2}, ..., z_{iL}. The function $g(y_i)$ is a function of the dependent variable. The non-linear econometric model

$$g(y_i) = \beta_0 + \beta_1 f_1(x_{i1}... \, x_{iK}) + \cdots + \beta_L f_L(x_{i1}... \, x_{iK}) + u_i$$

can therefore be written as

$$g(y_i) \quad = \quad \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_L z_{iL} + u_i \; .$$

The latter is the usual multiple linear regression model with $L+1$ regressors as long as all necessary assumptions about the error term and the L *transformed* explanatory variables z_{i1}, z_{i2}, ..., z_{iL} are satisfied. All properties of OLS are therefore preserved.

Note: While the original model is potentially *non-linear in the variables* x_{ik}, it is *linear in the parameters* β. Also note that the error term u_i is additive.

2 Examples

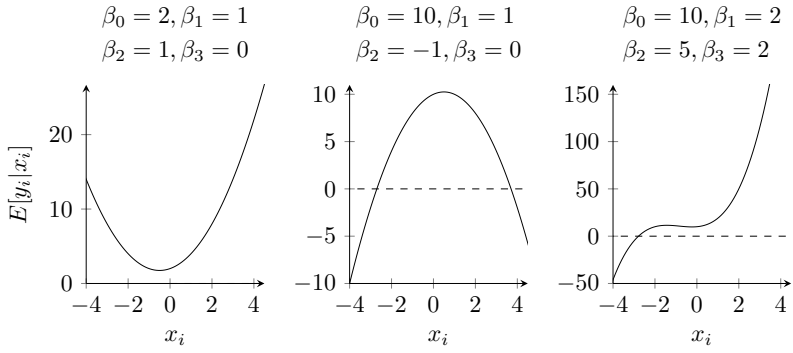
2.1 Polynomial

Functional form (3rd order):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i$$

Expected value under *OLS3c*:

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$



Marginal effect on the expected dependent variable:

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = \beta_1 + 2\beta_2 x_i + 3\beta_3 x_i^2.$$

Marginal effect on an individual given its error term u_i:

$$\frac{\partial y_i}{\partial x_i} = \beta_1 + 2\beta_2 x_i + 3\beta_3 x_i^2.$$

Note that the marginal effect depends on the value of the explanatory variable x_i. The individual parameters β_1, β_2 and β_3 often have no direct interpretation.

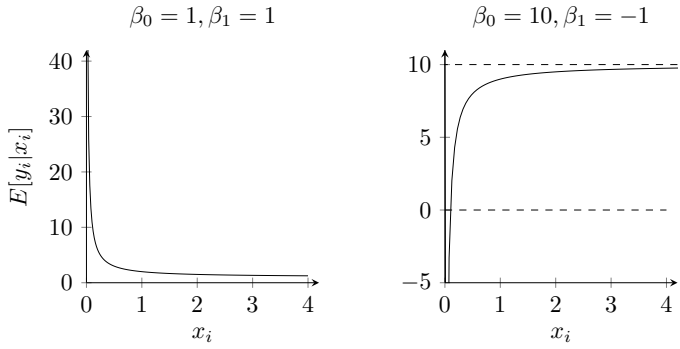
2.2 Inverse

Functional form:

$$y_i = \beta_0 + \beta_1 \frac{1}{x_i} + u_i.$$

Expected value under *OLS3c*:

$$E[y_i|x_i] = \beta_0 + \beta_1 \frac{1}{x_i}$$



Marginal effect:

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = -\frac{\beta_1}{x_i^2}$$

and

$$\frac{\partial y_i}{\partial x_i} = -\frac{\beta_1}{x_i^2}.$$

Note that a positive sign of β_1 means a negative relationship and vice-versa.

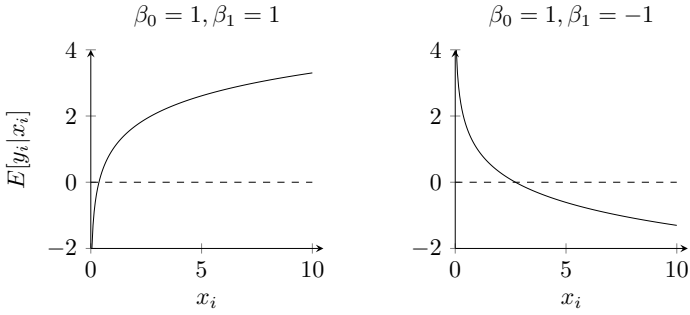
2.3 Explanatory Variable in Logs (linear-log)

Functional form:

$$y_i = \beta_0 + \beta_1 \ln(x_i) + u_i.$$

Expected value under *OLS3c*:

$$E[y_i|x_i] = \beta_0 + \beta_1 \ln(x_i)$$



Marginal effect:

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = \frac{\beta_1}{x_i}, \qquad \frac{\partial y_i}{\partial x_i} = \frac{\beta_1}{x_i}.$$

The coefficient β_1 is approximately the effect of a 100 % change in the explanatory variable x_i on the dependent variable y_i

$$\beta_1 \approx \frac{\Delta E[y_i|x_i]}{\Delta x_i / x_i}, \quad \beta_1 \approx \frac{\Delta y_i}{\Delta x_i / x_i}$$

where Δ is a small discrete change. Hence β_1 divided by 100 is approximately the effect of a 1 % change in x_i. β_1 is also called a semi-elasticity. For example, $\beta_1 = 56$ means that an increase in x_i by 1 % leads to an increase in the dependent variable y_i by (approximately) 0.56 units.[1]

[1] The exact effect of a discrete change of the explanatory variable by Δx_i is

$$\Delta E[y_i|x_i] = \beta_1 \cdot \ln\left(1 + \frac{\Delta x_i}{x_i}\right) \quad \text{and} \quad \Delta y_i = \beta_1 \cdot \ln\left(1 + \frac{\Delta x_i}{x_i}\right).$$

For example, $\beta_1 = 56$ means that an increase in x_i by $0.01 = 1\,\%$ leads to an increase in the dependent variable y_i by exactly $56 \cdot \ln(1 + 0.01) = 56 \cdot 0.00995 = 0.557$ units.

2.4 Dependent Variable in Logs (log-linear)

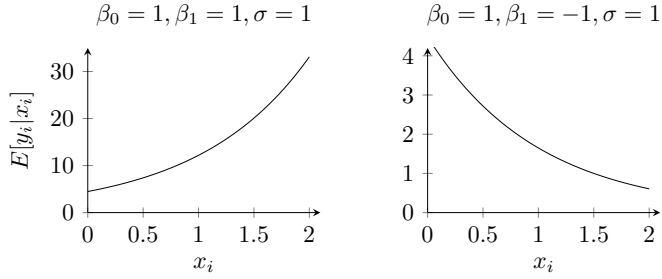
Functional form:

$$\ln(y_i) = \beta_0 + \beta_1 x_i + u_i.$$

Exptected value:

$$E[y_i|x_i] = e^{\beta_0+\beta_1 x_i} \cdot \alpha_i.$$

where $\alpha_i = E[e^{u_i}|x_i] > 1$. If the error is independent of the explanatory variables (*OLS3b*), $\alpha_i = \alpha$ is constant. If the error is normally distributed with homoscedastic error $V[u_i|x_i] = \sigma^2$ (*OLS3a*, *OLS4a*), $\alpha_i = \alpha = e^{\sigma^2/2}$.



Marginal effect:

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = E[y_i|x_i] \cdot \beta_1, \qquad \frac{\partial y_i}{\partial x_i} = y_i \cdot \beta_1.$$

The coefficient $\beta_1 \cdot 100\,\%$ is approximately the percentage effect on the dependent variable y_i of a change in the variable x_i by one unit

$$\beta_1 \approx \frac{\frac{\Delta E[y_i|x_i]}{E[y_i|x_i]}}{\Delta x_i}, \qquad \beta_1 \approx \frac{\frac{\Delta y_i}{y_i}}{\Delta x_i}$$

where Δ is a small discrete change. β_1 is also called a semi-elasticity. For example, $\beta_1 = 0.06$ means that an increase in x_i by one unit leads to an increase in the dependent variable y_i by (approximately) 6 %.[2]

[2]The exact effect of a discrete change of the explanatory variable by Δx_i units is

$$\frac{\Delta E[y_i|x_i]}{E[y_i|x_i]} = e^{\beta_1 \Delta x_i} - 1 \quad \text{and} \quad \frac{\Delta y_i}{y_i} = e^{\beta_1 \Delta x_i} - 1.$$

For example, $\beta_1 = 0.06$ means that an increase in x_i by one unit leads to an increase in the dependent variable y_i by exactly $exp(0.06) - 1 = 0.0618 = 6.18\,\%$.

2.5 Dependent and Explanatory Variables in Logs (log-log)

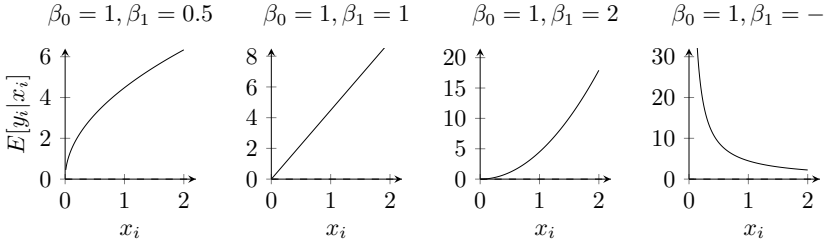
Functional form:

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + u_i.$$

Expected value:

$$E[y_i|x_i] = e^{\beta_0 + \beta_1 \ln(x_i)} \cdot \alpha_i.$$

where $\alpha_i = E[e^{u_i}|x_i] > 1$. Assuming that the error is independent of the explanatory variables (*OLS3b*), $\alpha_i = \alpha$ is a constant. Assuming that the error is normally distributed with homoscedastic error $V[u_i|x_i] = \sigma^2$ (*OLS3a* and *OLS4a*), $\alpha_i = \alpha = e^{\sigma^2/2}$. For $\sigma = 1$, hence $\alpha = 1.65$:



Marginal effect:

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = E[y_i|x_i] \cdot \frac{\beta_1}{x_i}, \qquad \frac{\partial y_i}{\partial x_i} = y_i \cdot \frac{\beta_1}{x_i}.$$

The coefficient β_1 is approximately the percentage effect on the dependent variable y_i of a 1 % change in the explanatory variable x_i

$$\beta_1 \approx \frac{\frac{\Delta E[y_i|x_i]}{E[y_i|x_i]}}{\frac{\Delta x_i}{x_i}}, \qquad \beta_1 \approx \frac{\frac{\Delta y_i}{y_i}}{\frac{\Delta x_i}{x_i}}$$

where Δ is a small discrete change. β_1 is also called an elasticity. For example, $\beta_1 = 0.8$ means that an increase in x_i by 1 % leads to an increase in the dependent variable y_i by (approximately) 0.8 %.[3]

[3] The exact effect of a discrete change of the explanatory variable by Δx_i is

$$\frac{\Delta E[y_i|x_i]}{E[y_i|x_i]} = e^{\beta_1 \ln\left(1+\frac{\Delta x_i}{x_i}\right)} - 1 \quad \text{and} \quad \frac{\Delta y_i}{y_i} = e^{\beta_1 \ln\left(1+\frac{\Delta x_i}{x_i}\right)} - 1.$$

For example, $\beta_1 = 0.8$ means that an increase in x_i by $0.01 = 1$ % leads to an increase in the dependent variable y_i by exactly $exp(0.8 \cdot \ln(1 + 0.01)) - 1 = 0.00799 = 0.799$ %.

2.6 Dummy Variables

Functional form:

$$y_i = \beta_0 + \beta_1 d_i + u_i$$

where $d_i \in \{0, 1\}$ is a dummy variable that either takes value 0 or 1.

Expected value under *OLS3c*:

$$E[y_i|d_i] = \begin{cases} \beta_0 & \text{if } d_i = 0 \\ \beta_0 + \beta_1 & \text{if } d_i = 1 \end{cases}$$

Marginal effect:

$$\frac{\partial E[y_i|d_i]}{\partial d_i} = \beta_1 \quad \text{and} \quad \frac{\partial y_i}{\partial d_i} = \beta_1.$$

Note that the notion of a marginal, i.e. infinitesimally small change, is not useful for a dummy variable d_i which can only increase by exactly one unit. The coefficient β_1 is better interpreted as the difference in the means of the (treatment) group with $d_i = 1$ and the (control) group with $d_i = 0$

$$\beta_1 = E[y_i|d_i = 1] - E[y_i|d_i = 0]$$

which constitutes the average treatment effect (ATE).

Note that the model

$$y_i = \beta_0 + \beta_1 treat_i + \beta_2 control_i + u_i$$

where the dummy variable $treat_i$ takes value 1 for the treatment group and 0 for the control group while the dummy variable $control_i$ takes the opposite values violates *OLS5* and the parameters β_0, β_1 and β_2 cannot be separately identified. This is called the dummy variable trap.

2.7 Interaction Terms

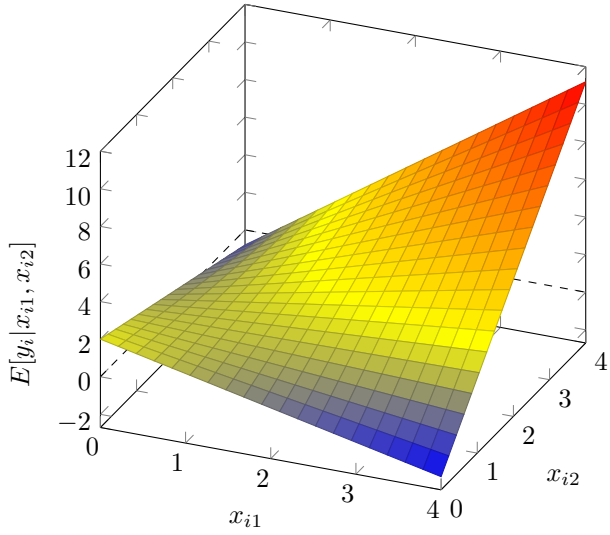
Functional form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i1} \cdot x_{i2}) + u_i$$

Expected value under *OLS3c*:

$$E[y_i|x_{i1}, x_{i2}] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i1} \cdot x_{i2})$$

$$\beta_0 = 2, \beta_1 = -1, \beta_2 = -0.7, \beta_3 = 1$$



Marginal effects:

$$\frac{\partial E[y_i|x_{i1}, x_{i2}]}{\partial x_{i1}} = \frac{\partial y_i}{\partial x_{i1}} = \beta_1 + \beta_3 x_{i2}$$

$$\frac{\partial E[y_i|x_{i1}, x_{i2}]}{\partial x_{i2}} = \frac{\partial y_i}{\partial x_{i2}} = \beta_2 + \beta_3 x_{i1}$$

2.8 Interactions with Dummy Variables

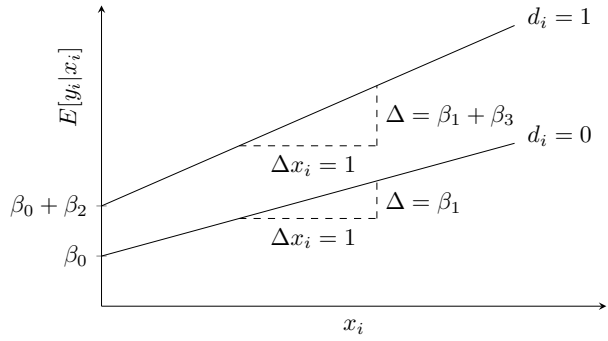
Functional form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 (x_i \cdot d_i) + u_i$$

where $d_i \in \{0, 1\}$ is a dummy variable that either takes value 0 or 1 and x_i is a continuous explanatory variable.

Expected value under *OLS3c*:

$$E[y_i|x_i, d_i] = \begin{cases} \beta_0 + \beta_1 x_i & \text{if } d_i = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i & \text{if } d_i = 1 \end{cases}$$



Marginal effect:

$$\frac{\partial E[y_i|x_i, d_i]}{\partial x_{i1}} = \frac{\partial y_i}{\partial x_{i1}} = \begin{cases} \beta_1 & \text{if } d_i = 0 \\ \beta_1 + \beta_3 & \text{if } d_i = 1 \end{cases}$$

Note that the interaction of all variables (here constant and one explanatory variable x_i) with the dummy variable d_i estimates separate linear relationships for the two groups defined by d_i. This yields identical estimates as two separate regressions for both groups. The standard errors may differ between the interacted joint estimation and the separate regressions unless robust standard errors are computed.

2.9 Spline Functions

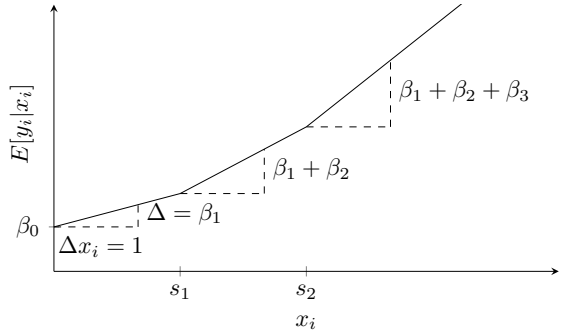
Functional form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_{i1}(x_i - s_1) + \beta_3 d_{i2}(x_i - s_2) + u_i$$

where $d_{i1} = 1$ if $x_i \geq s_1$ and $d_{i2} = 1$ if $x_i \geq s_2$. s_1 and s_2 are known thresholds.

Expected value under *OLS3c*:

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i + \beta_2 d_{i1}(x_i - s_1) + \beta_3 d_{i2}(x_i - s_2)$$



Marginal effect:

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = \frac{\partial y_i}{\partial x_i} = \begin{cases} \beta_1 & \text{if } x_i < s_1 \\ \beta_1 + \beta_2 & \text{if } s1 \leq x_i < s_2 \\ \beta_1 + \beta_2 + \beta_3 & \text{if } x_i \geq s_2 \end{cases}$$

3 Implementation in Stata

Non-linear Functional forms can be estimated with OLS by generating the transformed variables. For example,

```
webuse auto7.dta
generate mpg2 = mpg^2
reg price mpg mpg2
```

estimates a second order polynomial.

Dummy variables are easily created from categorical variables with the `xi` command. For example,

```
xi i.manufacturer
reg price _Imanufactu_*
```

creates 23 dummy variables for the 24 categories in the variable `manufacturer` (excluding the first one for use as reference category) and regresses `price` on all 23 dummy variables plus a constant. This can also be done in one step,

```
xi: reg price i.manufacturer
```

Interactions with dummy variables are also directly created with the `xi` command. For example,

```
xi: reg price i.foreign*mpg
```

estimates separate slopes and intercepts for foreign and domestic cars. As of version 11, dummy variables and interactions can also be formed as "factor variables". The above example is then

```
reg price i.foreign##c.mpg
```

The variables used for spline functions are conveniently created with the `mkspline` command. For example,

```
mkspline mpg_1 20 mpg_2 25 mpg_3 = mpg
reg price mpg_*
```

regresses `price` on `mpg` using a piecewise linear function. Also consider the option `marginal`.

4 Implementation in R

Non-linear Functional forms can be estimated with OLS by specifying the Functional form in the estimated model. For example,

```
> data(mtcars)
> lm(mpg~wt+I(wt^2), data=mtcars)
```

estimates a second order polynomial for the variable `wt`. Note that most mathematical functions need to be wrapped within the `I()` function.

Categorical variables are automatically included as a set of dummy variables if they are defined as factor variables. For example,

```
> mtcars$carbf <- factor(carb)
> lm(mpg~wt+carbf, data=mtcars)
```

regresses `mpg` on `wt` and on 5 dummy variables for 5 categories in `carbf` (excluding the first category for use as reference group) plus a constant.

Interactions with dummy variables from categorical variables can be directly estimated when the categorical variable is defined as factor variable. For example,

```
> mtcars$amf <- factor(mtcars$am, labels=c("automatic", "manual"))
> lm(mpg~amf+wt:amf, data=mtcars)
```

estimates separate slopes of `wt` and intercepts for cars with automatic and manual transmission. Altenatively,

```
> summary(lm(mpg~amf+wt/amf, data=mtcars))
```

reports the difference between the two slopes. This is equivalent to

```
> lm(mpg~am+wt+wt:am, data=mtcars)
```

which does not use factor variables.

Linear (and polynomial) spline functions are implemented in the `splines` package. See the help for details,

```
> library("splines")
> ?splines
```

References

Stock, James H. and Mark W. Watson (2020), Introduction to Econometrics, 4th Global ed., Pearson. Chapter 8.

Wooldridge, Jeffrey M. (2009), Introductory Econometrics: A Modern Approach, 4th ed. South-Western. Section 6.2 and 6.4.

Jaccard James and Robert Turrisi (2003), Interaction Effects in Multiple Regression, 2nd ed., Quantitative Applications in the Social Sciences 07-72, Sage.

Kennedy, Peter (2003), A Guide to Econometrics, 5th ed., Blackwell Publishing, chapter 7.