1 Linearity in Parameters

The OLS estimation technique requires that our model be *linear in parameters*. What this means is, each β term must appear essentially as a constant: we cannot have β_1^2 or $\log(\beta_1)$ or $\beta_1\beta_2$, for instance. This is because the OLS technique is only able to solve explicitly for each β if they appear in a linear fashion.

However, this does not necessitate that the model be linear in *variables*. There is no reason why we can't specify a model of the form

$$y = \beta_1 + \beta_2 \log(x) + u$$

if we think it is useful to do so.[1] And it certainly might be useful to do so. Consider the relationship between healthcare expenditure and life expectancy. We would expect more healthcare expenditure to be correlated with higher life expectancy, but at a diminishing rate (since there is a natural limit to life expectancy that medical treatment cannot overcome). Therefore it wouldn't make sense to impose a linear relationship between healthcare expenditure and life expectancy; we would want to use a log to capture diminishing returns to healthcare expenditure. Figure 1 illustrates this example.

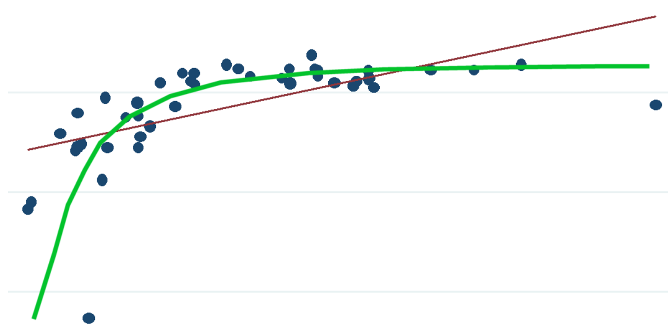

FIGURE 1: Specifying a logarithmic relationship (green) generates a better fit of the data compared to a linear relationship (red).

2 Functional Forms

There are an infinite number of different ways we could specify such a model. I will focus on three due to their salient economic interpretations. It will be assumed throughout that

[1] If you'd like, you can define $v \equiv \log(x)$ and rewrite the model as

$$y = \beta_1 + \beta_2 v + u,$$

from which it is obvious that the model exhibits the same form as that with which we are familiar.

the zero conditional mean assumption holds; this implies that x and u are uncorrelated. In practice, this means that when we change x, there is no change in u on average. This is useful because we will be taking derivatives soon, for which the zero conditional mean assumption implies that $du/dx = 0$.

2.1 Linear-Log Regression

A **linear-log** regression is of the form

$$y = \beta_1 + \beta_2 \log(x) + u. \tag{1}$$

It is named as such because the dependent variable is linear (it's simply y) and the regressor variable is within a logarithm.

Although $\log(x)$ may do a better job of capturing the data or a salient economic phenomenon, we are not really interested in finding out how a change in $\log(x)$ will affect y; we are interested in how a change in x will affect y. We'll have to do a little work to squeeze that information out of the regression, but it's not so bad. First, we can take the derivative of both sides with respect to x, which yields

$$\frac{dy}{dx} = \frac{\beta_2}{x}.$$

Now multiply both sides by dx, and then multiply the right-hand size by 100/100. Moving around the 100 factors, we end up with

$$dy = \frac{\beta_2}{100}\left(\frac{dx}{x} \times 100\right).$$

Notice that the term in parentheses is the percentage change in x.

Although calculus operations are in terms of infinitesimally small changes dy and dx, the equation still constitutes a valid approximation for small changes Δy and Δx. In practice, we will typically consider $\%\Delta x = 1$. The interpretation in words is: the change in the level of y equals $\beta_2/100$ times the percentage change in x,

$$\Delta y \approx \frac{\beta_2}{100} \times \%\Delta x. \tag{2}$$

2.2 Log-Linear Regression (Semi-Elasticity)

A **log-linear** regression is of the form

$$\log(y) = \beta_1 + \beta_2 x + u. \tag{3}$$

To interpret, start by taking the derivative of both sides with respect to x, which yields

$$\frac{d\log(y)}{dx} = \beta_2.$$

In order to squeeze y out of the left-hand side, we will appeal to the chain rule of calculus. In particular, we can write

$$\frac{d\log(y)}{dy}\frac{dy}{dx} = \beta_2.$$

We know that $d\log(y)/dy = 1/y$, so let's make that substitution. Let's also multiply both sides by $100 \times dx$, which yields

$$\frac{dy}{y} \times 100 = (100 \times \beta_2)dx.$$

In words: the percentage change in y equals $100 \times \beta_2$ times the change in the level of x,

$$\%\Delta y \approx 100\beta_2 \times \Delta x. \tag{4}$$

In this form, coefficient β_2 is referred to as the **semi-elasticity** of y with respect to x.

Log-Log Regression (Elasticity)

A **log-log** regression is of the form

$$\log(y) = \beta_1 + \beta_2 \log(x) + u. \tag{5}$$

To interpret, take the derivative of both sides with respect to x, which gives

$$\frac{d\log(y)}{dx} = \frac{\beta_2}{x}.$$

Use the chain rule again on the right-hand side so that

$$\frac{d\log(y)}{dy}\frac{dy}{dx} = \frac{\beta_2}{x}.$$

We know that $d\log(y)/dy = 1/y$, so let's make that substitution. Also multiply both sides by dx and both sides by 100. Doing so yields

$$\frac{dy}{y} \times 100 = \beta_2 \left(\frac{dx}{x} \times 100 \right).$$

In words: the percentage change in y is equal to β_2 times the percentage change in x,

$$\%\Delta y \approx \beta_2 \times \%\Delta x. \tag{6}$$

In this form, coefficient β_2 is referred to as the **elasticity** of y with respect to x, which you hopefully remember from a microeconomics course.

3 Summary

Again, there are a multitude of other functional forms we could consider, e.g. quadratic forms, that are useful in certain contexts. Those will be discussed later as they become pertinent. But for now, here is a table of the four functional forms we have seen thus far.

Model	Dependent Variable	Regressor	Interpretation of β_2
linear	y	x	$\Delta y = \beta_2 \times \Delta x$
linear-log	y	$\log(x)$	$\Delta y \approx \frac{\beta_2}{100} \times \%\Delta x$
log-linear (semi-elasticity)	$\log(y)$	x	$\%\Delta y \approx 100\beta_2 \times \Delta x$
log-log (elasticity)	$\log(y)$	$\log(x)$	$\%\Delta y \approx \beta_2 \times \%\Delta x$

4 Life Expectancy and Healthcare Expenditure

Download `hcle.dta` from my website and import the data into Stata. The dataset has three variables: country, life expectancy at birth, and healthcare spending per-capita in 2015. If we estimate a linear regression of the form

$$\text{lifeexpect} = \beta_1 + \beta_2 \text{hcspending} + u,$$

then we find goodness-of-fit measure $R^2 = 0.363$. On the other hand, if we do a linear-log regression like suggested earlier,

$$\text{lifeexpect} = \beta_1 + \beta_2 \log(\text{hcspending}) + u,$$

then we find goodness-of-fit measure $R^2 = 0.542$, implying a better fit.

Note that there is an **important caveat** in comparing the R^2 of different models: it is *not* meaningful to compare the R^2 of models that have different dependent variables! If the regressors are different but the dependent variables are the same, as is the case here, then it is fine to compare R^2.

```
. regress lifeexpect hcspending

      Source |       SS           df       MS      Number of obs   =        44
-------------+----------------------------------   F(1, 42)        =     23.93
       Model |  394.963176         1  394.963176   Prob > F        =    0.0000
    Residual |  693.098579        42  16.5023471   R-squared       =    0.3630
-------------+----------------------------------   Adj R-squared   =    0.3478
       Total |  1088.06175        43  25.3037617   Root MSE        =    4.0623

------------------------------------------------------------------------------
  lifeexpect |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  hcspending |   .0014559   .0002976     4.89   0.000     .0008553    .0020565
       _cons |   73.98101   1.148863    64.39   0.000     71.66251    76.29951
------------------------------------------------------------------------------

. gen loghc = log(hcspending)

. regress lifeexpect loghc

      Source |       SS           df       MS      Number of obs   =        44
-------------+----------------------------------   F(1, 42)        =     49.67
       Model |  589.550424         1  589.550424   Prob > F        =    0.0000
    Residual |  498.511331        42  11.8693174   R-squared       =    0.5418
-------------+----------------------------------   Adj R-squared   =    0.5309
       Total |  1088.06175        43  25.3037617   Root MSE        =    3.4452

------------------------------------------------------------------------------
  lifeexpect |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       loghc |   4.646454   .6592863     7.05   0.000      3.31596    5.976947
       _cons |   42.30245    5.19564     8.14   0.000     31.81723    52.78768
------------------------------------------------------------------------------
```