Properties of Simple Regression Model

February 6, 2011

Outline

Terminology

Units and Functional Form

Mean of the OLS Estimate

Omitted Variable Bias

Next we will address some properties of the regression model

Forget about the three different motivations for the model, none are relevant for these properties

Outline

Consider the following terminology from Wooldridge

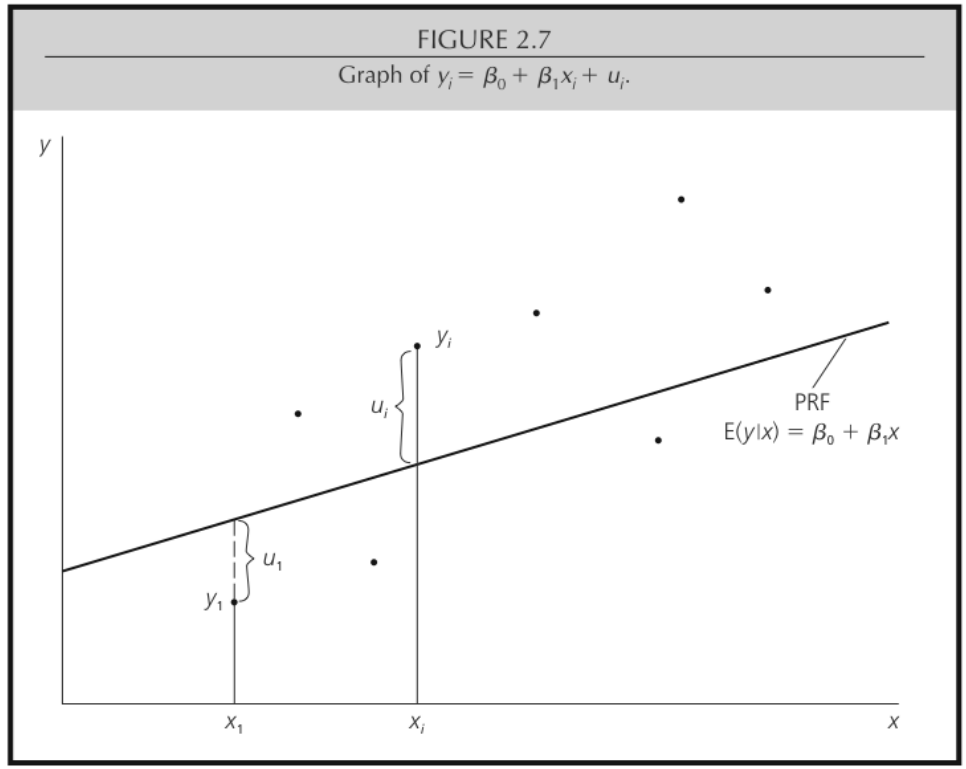
TABLE 2.1

Terminology for Simple Regression

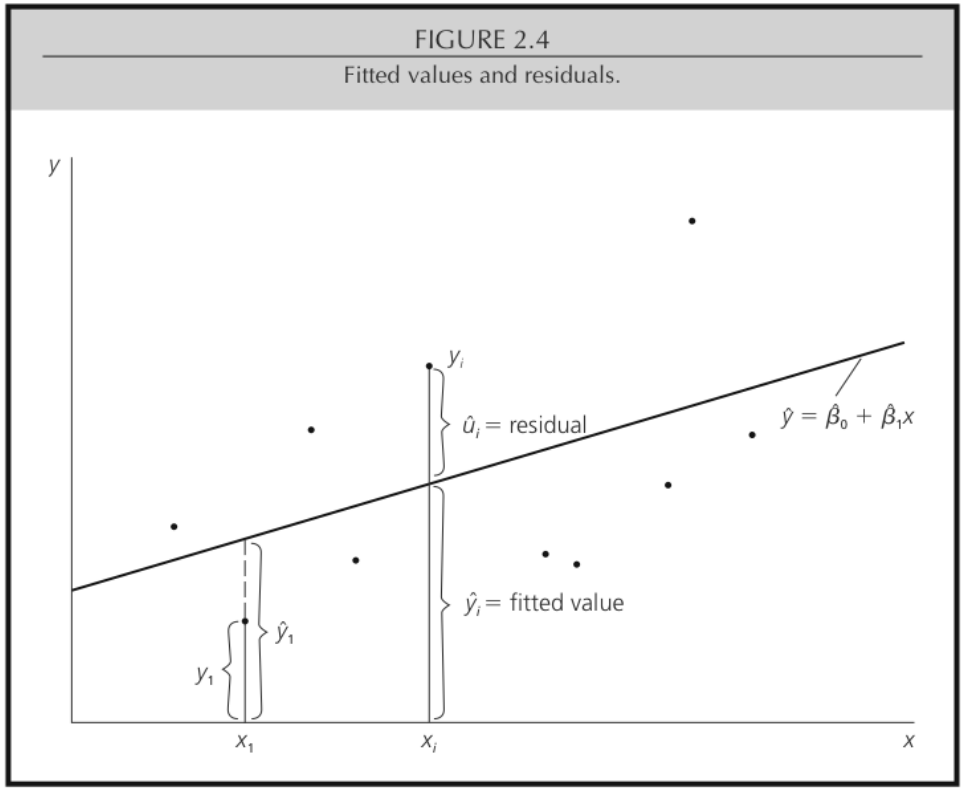
y	x
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor

Graphically the model is defined in the following way

Population Model



Sample Model



Also when we have the regression model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

we call this a regression of "*y* **on** *x*"

Normal Equations

It must be the case that

$$
\begin{aligned}
0 &= \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \\
0 &= \frac{1}{N}\sum_{i=1}^{N} x_i \widehat{u}_i \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)
\end{aligned}
$$

This is not approximately true, this is exactly true

This really is not surprising

These are the equations we started with when we derived $\widehat{\beta}_0$ and $\widehat{\beta}_1$

Goodness of Fit

We want to say something about how well our model fits the data

We will make use of the following three things

- Total Sum of Squares (SST)

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

- Explained Sum of Squares (SSE)

$$SSE = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

- Residual Sum of Squares (SSR)

$$SSR = \sum_{i=1}^{N} \hat{u}_i^2$$

It turns out that there is a nice relationship between these concepts

$$
\begin{aligned}
SST &= \sum_{i=1}^{N} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{N} ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \\
&= \sum_{i=1}^{N} (\hat{u}_i + (\hat{y}_i - \bar{y}))^2 \\
&= \sum_{i=1}^{N} \left[\hat{u}_i^2 + (\hat{y}_i - \bar{y})^2 + 2\hat{u}_i (\hat{y}_i - \bar{y}) \right]
\end{aligned}
$$

$$= \sum_{i=1}^{N} \hat{u}_i^2 + \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N} 2\hat{u}_i (\hat{y}_i - \bar{y})$$

$$= SSR + SSE + \sum_{i=1}^{N} 2\hat{u}_i (\hat{y}_i - \bar{y})$$

So this is really nice except for that last part. Lets deal with it

$$\begin{aligned}
\sum_{i=1}^{N} 2\hat{u}_i \left(\hat{y}_i - \bar{y}\right) =& 2\sum_{i=1}^{N} \hat{u}_i \left(\left\{\hat{\beta}_0 + \hat{\beta}_1 x_i\right\} - \left\{\hat{\beta}_0 + \hat{\beta}_1 \overline{x}\right\}\right) \\
=& 2\sum_{i=1}^{N} \hat{u}_i \left(\hat{\beta}_1 x_i - \hat{\beta}_1 \overline{x}\right) \\
=& 2\hat{\beta}_1 \sum_{i=1}^{N} \hat{u}_i x_i - 2\hat{\beta}_1 \overline{x} \sum_{i=1}^{N} \hat{u}_i \\
=& 0
\end{aligned}$$

Thus

$$SST = SSE + SSR$$

This gives us a really nice way of describing the goodness of fit of the model

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Lets look at examples of models that fit well versus those that don’t

Outline

Units

What happens when we change the units of measurement of a variable?

Lets think about the smoking example we looked at in the Statistics Review

Before we looked at whether someone smoked on Birthweight

Instead lets look at the number of cigarettes per day on birthweight from the data set bwght

The units here are sort of arbitrary, we can look at at cigarettes per day and weight in ounces

We could have just as easily looked at packs per day and weight in pounds.

Hopefully this shouldn’t change the basic results

Lets define the following four variables:

O	Weight in Ounces
L	Weight in Pounds
C	Cigarettes per day
P	Packs per day

Notice that

$$O = 16L$$

$$C = 20P$$

Next think about the following three Conditional Expectations:

$$
\begin{aligned}
E(O \mid C) &= \beta_0^1 + \beta_1^1 C \\
E(L \mid C) &= \beta_0^2 + \beta_1^2 C \\
E(O \mid P) &= \beta_0^3 + \beta_1^3 P
\end{aligned}
$$

What is the relationship between these things?

It should be the case that

$$\begin{aligned}\beta_0^1 + \beta_1^1 C &= E(O \mid C) \\ &= E(16L \mid C) \\ &= 16E(L \mid C) \\ &= 16\beta_0^2 + 16\beta_1^2 C\end{aligned}$$

Thus it should be the case that

$$\begin{aligned}\beta_0^1 &= 16\beta_0^2 \\ \beta_1^1 &= 16\beta_1^2\end{aligned}$$

It turns out that this is exactly true when you run the regression

Along Similar lines

$$\begin{aligned}\beta_0^3 + \beta_3^3 p =& E(O \mid P = p) \\ =& E(O \mid C = 20p) \\ =& \beta_0^1 + \beta_1^1 (20p)\end{aligned}$$

Then it should be the case that

$$\begin{aligned}\beta_0^3 &= \beta_0^1 \\ \beta_1^3 &= 20\beta_1^1\end{aligned}$$

Once again it turns out that this is precisely true in the data.

Functional Form

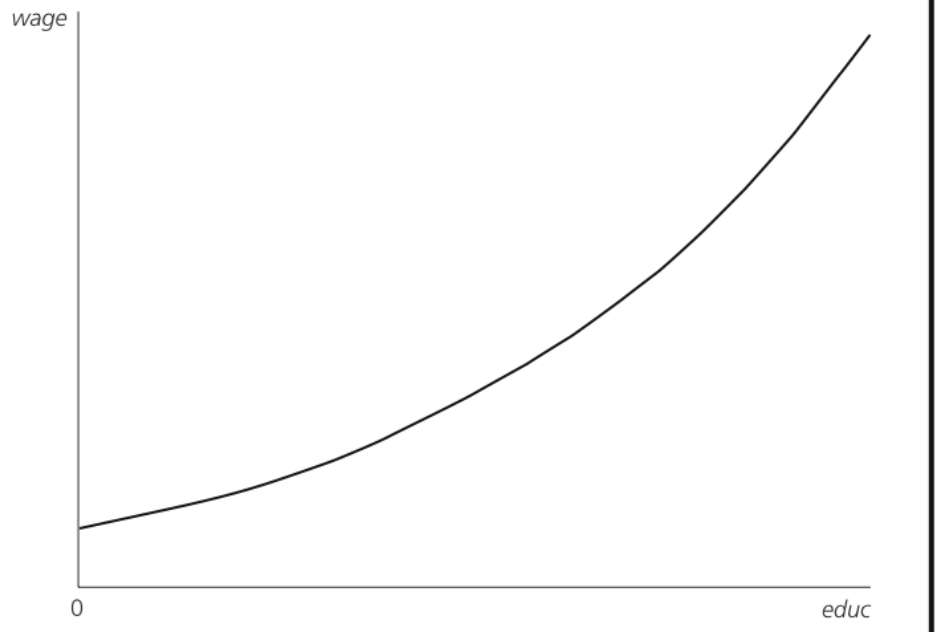
A really important assumption that we have made is that the model is linear

However that is really not as strong as an assumption as you might think because we can pick y and x to be whatever we want

For example consider the following figure

FIGURE 2.6

$wage = \exp(\beta_0 + \beta_1 educ)$, with $\beta_1 > 0$.



This clearly isn't linear, wages are growing exponentially with education

However, this is easy to fix if wages are growing exponentially with education, then the logarithm of wages are growing linearly with wages

Rather than regressing wages on schooling we regress the log of wages on schooling

$$log(wage_i) = \beta_0 + \beta_1 Ed_i + u_i$$

Figuring out exactly how to do this involves playing with the data

Here are some specifications and what they are called

TABLE 2.3

Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\Delta y = \beta_1 \Delta x$
level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Lets look at a number of different examples

Outline

Classical Linear Regression

In this section I will follow section 2.5 of Wooldridge very closely

Our goal is to derive the mean and variance of the OLS estimator

In doing so we need to make some assumptions about the population and the sample.

This set of assumptions is often referred to as the Classical Linear Regression Model

First we need to define the basic model.

Assumption (SLR.1-Linear in Parameters)

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

$$y = \beta_0 + \beta_1 x + u,$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

Notice that in making this assumption we have really moved to the “structural world.” That is we are really saying that this is the actual data generating process and our goal is to uncover the true parameters.

Now we need to assume something about the sample

Assumption (SLR.2-Random Sampling)

We have a random sample of size n, $(x_i, y_i), i = 1, ..., n$ *following the population model defined in SLR.1.*

This now defines the basic environment.

Next we need an assumption that allows us to estimate the model

Assumption (SLR.3-Sample Variation in the Explanatory Variable)

The sample outcomes on x, *namely,* $\{x_i, i = 1, ..., n\}$, *are not all the same value.*

Without this assumption we would have real trouble.

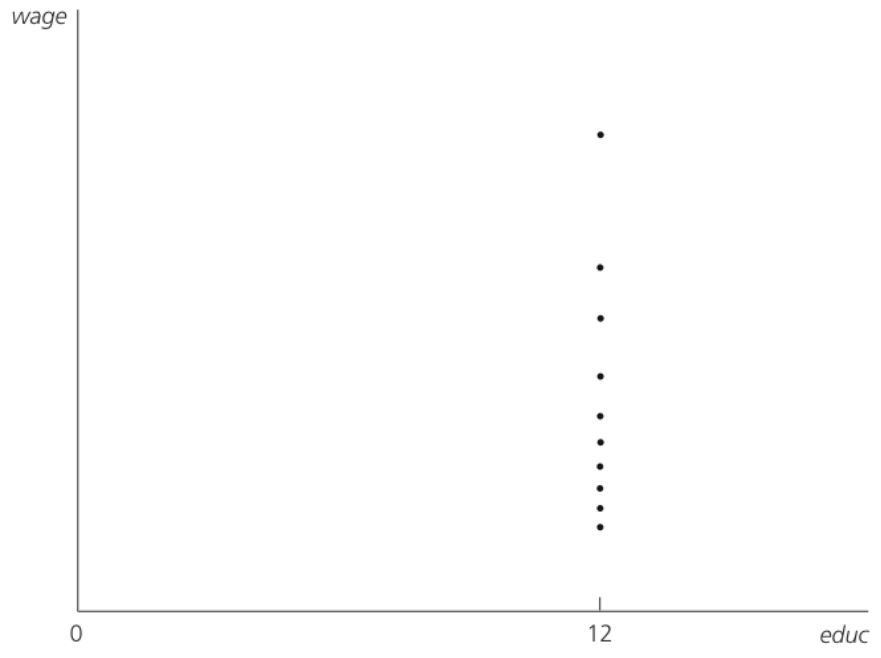
Practically the denominator for $\hat{\beta}_1$ is $\sum_{i=1}^{N} (x_i - \overline{x})^2$.

This would be zero if there is no variation in x_i.

Further suppose that the data looked like this:

FIGURE 2.3

A scatterplot of wage against education when $educ_i = 12$ for all i.



How would you estimate the slope?

Assumption (SLR.4-Zero Conditional Mean)

The error u has an expected value of zero given any value of the explanatory variable. In other words

$$E(u \mid x) = 0.$$

This will turn out to be the most important assumption for causal work.

Our goal now is to figure out

$$E\left(\hat{\beta}_0\right)$$

and

$$E\left(\hat{\beta}_1\right)$$

I am now going to be explicit about something that Wooldridge is a bit loose about.

We need to use the law of iterated expectations which states that for random variables W and Z,

$$E\left[E\left(Z \mid W\right)\right] = E[Z]$$

This can kind of give you a headache to think about. Rather than trying to derive it and get into details, lets just see why it is useful.

This means that

$$E\left[E\left(\hat{\beta}_1 \mid x_1, ..., x_n\right)\right] = E\left(\hat{\beta}_1\right)$$

Which further means that if we can show that

$$E\left(\hat{\beta}_1 \mid x_1, ..., x_n\right) = \beta_1$$

then

$$E\left(\hat{\beta}_1\right) = \beta_1.$$

This is what we will do (and what Wooldridge does as well).

First we will make use of the fact that

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \\
&= \frac{\sum_{i=1}^{N}(x_i-\bar{x})(\beta_0+\beta_1 x_i+u_i-\beta_0-\beta_1\bar{x}-\bar{u})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \\
&= \frac{\sum_{i=1}^{N}(x_i-\bar{x})(\beta_1 x_i-\beta_1\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} + \frac{\sum_{i=1}^{N}(x_i-\bar{x})(u_i-\bar{u})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \\
&= \beta_1 \frac{\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} + \frac{\sum_{i=1}^{N}(x_i-\bar{x})(u_i-\bar{u})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{N}(x_i-\bar{x})(u_i-\bar{u})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}
\end{aligned}
$$

Then notice that

$$
\begin{aligned}
E\left(\hat{\beta}_1 \mid x_1, ..., x_N\right) =& E\left(\beta_1 + \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(u_i - \bar{u}\right)}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \mid x_1, ..., x_N\right) \\
=& \beta_1 + \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right) E\left(u_i - \bar{u} \mid x_1, ..., x_N\right)}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \\
=& \beta_1
\end{aligned}
$$

This means that $\hat{\beta}_1$ is **UNBIASED**

This is a really nice property

Thinking about it, the only assumption that was really important for this was SLR.4

Lets go through a number of examples and think about causality and think about whether we believe this assumption

Outline

Omitted Variable Bias

In general the problem is that there is some other variable out there that affects y other than x.

(The material I am discussing here is covered in Wooldridge in Chapter 3 rather than Chapter 2)

To see why this is a problem suppose that in reality the unobserved variable depends on two things x and z so that it is still true that

$$y = \beta_0 + \beta_1 x + u,$$

but now suppose that

$$u = \beta_2 z + \varepsilon$$

and further we are worried that x and z are related.

Lets think of this as determined by the model

$$z = \delta_1 x + \xi$$

Now make the following assumptions about these new error terms

$$E(\varepsilon \mid x) = E(\xi \mid x) = 0.$$

Does this satisfy the assumptoins of the classical linear regression model?

We are assuming all of them except SLR.4

For that one notice that

$$\begin{aligned} E(u \mid x) &= E(\beta_2 z + \varepsilon \mid x) \\ &= \beta_2 E(z \mid x) + E(\varepsilon \mid x) \\ &= \beta_2 E(\delta_1 x + \xi \mid x) \\ &= \beta_2 \delta_1 x \end{aligned}$$

Lets think about the bias of OLS again

$$
\begin{aligned}
E\left(\hat{\beta}_1 \mid x_1, ..., x_n\right) =& E\left(\beta_1 + \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right) u_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \mid x_1, ..., x_n\right) \\
=& \beta_1 + \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right) E\left(u_i \mid x_1, ..., x_n\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\
=& \beta_1 + \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right) \beta_2 \delta_1 x_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\
=& \beta_1 + \beta_2 \delta_1 \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right) x_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\
=& \beta_1 + \beta_2 \delta_1
\end{aligned}
$$

Lets think about the sign of the bias

β_2	δ_1	Bias
+	+	+
+	-	-
-	+	-
-	-	+

What is the best solution for omitted variable bias?

Don’t omit the variable.

That is the motivation for the Multiple Regression model that we will consider next.