A PRIMER ON LINEAR REGRESSION*

Marc F. Bellemare†

Introduction

This set of lecture notes was written to allow understanding the classical linear regression model, which is one of the most common tools of statistical analysis in the social sciences.

Among other things, a regression allows the researcher to estimate the association between a variable of interest D and an outcome of interest y while holding other included factors $x = (x_1, \dots, x_K)$ constant. For the purposes of this discussion, let D measure a given policy, y measure welfare, and the vector x measure the various control variables the researcher saw fit to include.

Example

For example, one might be interested in the impact of individuals' years of education D on their wage y while controlling for age, gender, race, state, sector of employment, etc. in x. Generally speaking, social scientists are interested in the causal impact of a specific variable of interest on an outcome of interest, i.e., in the causal impact of D on y.

Mechanics

The regression of y on $(D, x_1, \dots, x_K)$ is typically written as

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \gamma D_i + \epsilon_i, \quad (1)$$

where *i* denotes a unit of observation. In the example of wages and education, the unit of observation would be an individual, but units of observations can be individuals, households, plots of land, firms, farms, villages, communities, countries, etc. Just as the research question should drive the choice of what to measure for y, D, and x, the research question also drives the choice of the relevant unit of observation.

When estimating equation 1, the researcher will have data on N units of observations, so $i = 1, \dots, N$. Alternatively, we call N the sample size. For each of those N units, the researcher will have data on y, D, and x. In other words, we will ignore the problem of missing data, as observations with missing data are usually dropped by most statistical packages. In practice, however, it can be unwise to ignore missing data.

The role of regression analysis is to estimate the coefficients $(\alpha, \beta_1, \dots, \beta_K, \gamma)$. To differentiate the "true" coefficients from coefficient estimates, we will use a circumflex accent (more commonly known as a "hat") to denote estimated coefficients. Therefore, the estimated $(\alpha, \beta_1, \dots, \beta_K, \gamma)$ will be denoted $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K, \hat{\gamma})$.

---

* Version 4.0, August 2013.

† Assistant Professor, Department of Applied Economics, University of Minnesota, 1994 Buford Avenue, St. Paul, MN, 55108, mbellema@umn.edu.

Going back to our interest in estimating the impact of D on y at the margin, this impact is represented in the context of equation 1 by the parameter γ. Indeed, if you remember your partial derivatives, the marginal impact of D on y at is equal to $\frac{\partial y}{\partial D} = \gamma$. Moreover, the partial derivative measures what happens to y when only D varies.

In other words, γ measures the impact of a change in D on y holding everything else constant, or *ceteris paribus*. Here, it is important to note that "everything else" is somewhat misleading, as it is limited only to the factors that are included in the vector x of control variables. Whatever is not included among the variables x is *not* held constant by regression analysis.

Indeed, the relationship in equation 1 is not deterministic in the sense that even if we have data for y, D, and x and credible parameter estimates $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K, \hat{\gamma})$, we will still not be able to perfectly forecast for y. That is because there are several things about any given problem that we, as social science researchers, do not observe. Individuals have intrinsic motivations that even they may have difficulty expressing. They make mistakes. Individuals experience unforeseen events. So there are factors which are very important in determining y which we simply do not observe.

For all these reasons, we add an error term ϵ at the end of equation 1. The error term simply represents our ignorance about the problem. As such, it includes all of the things that we *did not think of including* in the right-hand side of equation 1, as well as all of the things that we *could not include* in the right-hand side of equation 1. The error term ϵ thus embodies our ignorance about the relationship between two variables.

So how does a linear regression actually work? To take an example I know well, suppose we are looking at only two variables: rice yield (i.e., kg/are), which will be our outcome of interest y since it represents agricultural productivity, and cultivated area (number of ares, or hundredths of a hectare, or 100 square meters), which will be our variable of interest D. The inverse relationship between farm or plot size and productivity an old empirical puzzle in development economics (Barrett et al., 2010). So, plotting some data on this question, we get figure 1 below.

The scatter plot in figure 1 directly shows that the relationship between yield and cultivated area is not deterministic. That is, the relationship between the two variables is not a straight line, and the fact that the relationship is scattered indicates that there are other factors beside cultivated area that contributed to determining rice productivity.

The role of the regression – and, as we will soon understand, of the error term – is to linearly approximate as best as possible the relationship between two variables. In other words, to do something that looks like the red line in figure 2.

Figure 1. Rice Productivity Scatter

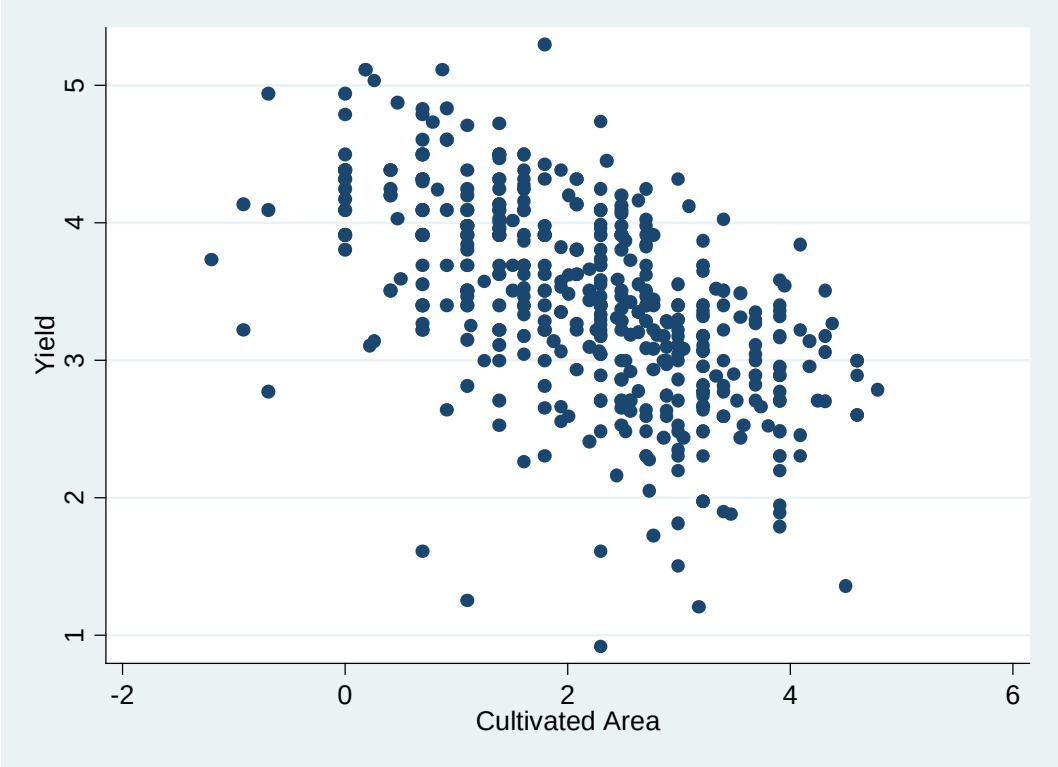

Figure 2. Rice Productivity Scatter and Regression Line

Note that we indeed find an inverse relationship between plot size and productivity, since the regression line slopes downward, which means that in this context, $\hat{\gamma} < 0$. Indeed, running a simple regression of rice yield on cultivated area yields $\hat{\alpha} = 4.187$ (and we can see from the graph that 4.187 would indeed be the value of rice productivity at a cultivated area of zero ares, or the intercept) and $\hat{\gamma} = -0.356$, with both coefficients statistically different from zero at less than the 1 percent level (i.e., there is a less than one percent chance α and γ are no different from zero). In other words, the finding here is that *on average*,

$$y = 4.187 - 0.356D, \tag{2}$$

or for every 1 percent increase in cultivated area, rice productivity decreases by 0.356 percent. This may seem counterintuitive, but remember that productivity is not total output – it is only a measure of average productivity on the plot.

So how does the apparatus of the linear regression determine the value of the intercept and the value of the slope of the regression line in figure 2? This is where our error term comes into play. Indeed, a linear regression will choose, among all possible lines, the one that minimizes the sum of the distances between each point in the scatter and the line itself, under the assumption that the error is on average equal to zero (i.e., that our predictions are right *on average*). Assuming that the error term is equal to zero on average and minimizing the sum of all point-line distances (technically, the sum of squared errors) allows us to obtain estimates $\hat{\alpha}$ and $\hat{\gamma}$ of the true parameters α and γ.

A few remarks. First off, the constant term (or the intercept) α does not have an economic interpretation in this case, since a cultivated area of zero would really entail a yield of zero – you cannot grow rice without land.

Second, since the only factor we included on the right-hand side of equation 1 was cultivated area, the error term includes a *lot* of things which may be potentially crucial in determining yield. For example, the plot's position on the toposequence, the quality of the soil, the source of irrigation of the plot, various characteristics of the household operating the plot, etc. So because we are typically interested in the impact of D on y controlling for a number of factors x, regression results will not be presented in the form of figure 2. Indeed, regression results are typically presented in the form of table 1 at the end of this document.

How do we interpret table 1? First off, note that N = 466. That is, we have data on 466 plots. The first column tells us what variables are included on the right-hand side of equation 1, viz. cultivated area; land value; total land owned by the household; household size (number of individuals); household dependency ratio (proportion of dependents within the household); whether the household head is a single female or a single male; whether the plot is irrigated by a dam, a spring, or rainfed; soil quality measurements (carbon, nitrogen, potassium percentages; soil pH; clay, silt, and sand percentages); and an intercept. The second column shows the estimated coefficients for the first specification of equation 1 (in this case, a pooled cross-section of all the plots and all the households, i.e., a specification which ignores the fact that some households own more than just one plot in the sample); and the third column shows the standard errors around each estimated coefficient.

These standard errors are used to determine whether each coefficient is statistically significantly different from zero or not. To make life simpler, table 1 shows whether coefficients are significant at the 10, 5, or 1

percent levels by using the symbols *, **, and *** respectively. Note that in all cases, there is a (significant) inverse relationship between cultivated area and rice productivity.

Taking column 1 as an example of how to interpret regression results, what can we say? First and foremost, note that for a 1 percent increase in cultivated area, there is an associated productivity decrease of 0.27 percent (alternatively, a doubling of the size of the plot is associated with a 27 percent decrease in productivity). Moreover, we can note three things. First off, the more valuable a plot, the more productive it will be; second, plots irrigated by a dam are more productive than plots without any irrigation; and third, plots irrigated by a spring are more productive than plots without any irrigation. In fact, comparing the magnitude of the coefficient estimates for irrigation by a dam and irrigation by a spring, we see that the impacts of these two types of irrigation are essentially the same.

Another thing of note in table 1 is how the coefficient on the variable of interest changes depending on what is included on the right-hand side of equation 1. Comparing the first two specifications (i.e., pooled cross-section vs. household fixed effects), note how the magnitude of the inverse relationship between productivity and cultivated area is reduced from -0.271 to -0.176 when household fixed effects (i.e., controls for household-specific unobservables characteristics, which is made possible here because there are 286 households for 466 plots; in other words, there are some households who own more than one plot in the sample) are included. This indicates that a great deal of the inverse relationship can be attributed to household-specific, otherwise unobservable factors. Likewise, comparing specifications 1 and 3 (i.e., pooled cross-section vs. soil quality), note again how the magnitude of the inverse relationship between productivity and cultivated area is reduced from -0.271 to -0.265 when soil quality measurements are included. Overall, this indicates that household-specific, unobserved factors are more important in driving the inverse relationship than the omission of soil quality measurements. In any event, a comparison of specifications 1 and 2 and of specifications 1 and 3 point to an important endogeneity problem (in this case, an omitted variables problem) caused by the omission, respectively, of household fixed effects and of soil quality measurements.

References

Barrett, Christopher B., Marc F. Bellemare, and Janet Y. Hou (2010), "Reconsidering Conventional Explanations of the Inverse Productivity—Size Relationship," *World Development* 38(1): 88-97.

Table 1 – Yield Approach Estimation Results (n=466)

	(1) Pooled Cross-Section		(2) Household Fixed Effects		(3) Soil Quality		(4) Household Fixed Effects and Soil Quality	
Variable	**Coefficient**	**(Std. Err.)**	**Coefficient**	**(Std. Err.)**	**Coefficient**	**(Std. Err.)**	**Coefficient**	**(Std. Err.)**
	Dependent Variable: Rice Yield (Kilograms/Are)							
Cultivated Area	-0.271***	(0.038)	-0.176***	(0.046)	-0.265***	(0.048)	-0.187***	(0.052)
Total Land Area	-0.055	(0.038)			-0.054	(0.047)		
Land Value	0.183***	(0.031)	0.303***	(0.069)	0.176***	(0.032)	0.287***	(0.063)
Household Characteristics								
Household Size	-0.007	(0.009)			-0.008	(0.008)		
Dependency Ratio	-0.073	(0.130)			-0.083	(0.144)		
Single Female	-0.056	(0.111)			-0.070	(0.119)		
Single Male	0.133	(0.134)			0.122	(0.155)		
Plot Characteristics								
Irrigated by Dam	0.389**	(0.171)	0.228	(0.202)	0.450**	(0.211)	0.545	(0.402)
Irrigated by Spring	0.365**	(0.175)	0.250	(0.214)	0.425**	(0.214)	0.541	(0.389)
Irrigated by Rain	0.184	(0.180)	0.024	(0.217)	0.249	(0.220)	0.313	(0.431)
Soil Quality Measurements								
Carbon					-1.361	(1.510)	-0.001	(1.844)
Nitrogen					1.668	(1.781)	-0.007	(2.750)
pH					-1.064	(7.163)	-17.969	(14.459)
Potassium					1.183	(1.412)	-5.528*	(3.035)
Clay					0.293	(3.115)	-5.183	(4.174)
Silt					0.521	(5.751)	4.485	(11.681)
Sand					-0.135	(5.607)	5.261	(7.106)
Intercept	-1.847***	(0.372)	-3.162***	(0.694)	-2.276	(1.552)	-3.328***	(0.819)
Number of Households	–		286		–		286	
Bootstrap Replications	–		–		500		500	
Village Fixed Effects	Yes		Dropped		Yes		Dropped	
R^2	0.45		0.97		0.46		0.97	
p-value (All Coefficients)	0.00		0.00		0.00		0.00	
p-value (Fixed Effects)	–		0.00		–		0.00	
p-value (Soil Quality)	–		–		0.79		0.52	

Source: Barrett et al. (2010). The symbols ***, ** and * indicate statistical significance at the one, five and ten percent levels, respectively.