

Appendix E Regression Analysis

Regression analysis can be used to describe a relationship between a dependent or outcome variable and one or more independent variables. For example, our goal can be to describe how workers' productivity depends on the type of compensation and workers' ability. As a first step, we may specify the model as follows:

$$Y_i = a + bD_i + cA_i + u_i$$

where i indexes workers, Y_i represents worker i 's productivity, D_i is an indicator variable equal to 1 if worker i is a piece-rate worker and 0 if he is a salary worker, and A_i is worker i 's ability. u_i represents all other factors that explain variation in productivity across workers besides how workers are paid and their ability.

The coefficients a , b , and c can be interpreted as follows:

- (1) $a = E[Y_i | D_i=0, A_i=0]$, so a represents the average productivity of a salary worker ($D_i=0$) of ability 0 ($A_i=0$).
- (2) $b = \partial E[Y_i] / \partial D_i$, where the symbol ∂ indicates a partial derivative and indicates that all other factors are held constant. Since D takes the value of 1 or 0, b represents the average difference in productivity between piece-rate and salary workers, holding all other variables constant. In our model, b tells us the average difference between productivity of piece-rate and salary workers of same ability.
- (3) $c = \partial E[Y_i] / \partial A_i$, so c tells us the average difference in productivity between two workers who differ in their ability by a small amount, but who are otherwise paid in the same way.

Once we specified the model, we need to collect data on Y , D , and A . Suppose that we obtained the following data:

Worker (i)	Productivity (Y_i)	Piece-Rate (D_i)	Ability (A_i)
1	10.82	1	3
2	13.18	1	4
3	10.42	1	2
4	11.90	1	3
5	10.20	1	3
6	10.17	0	3
7	7.68	0	2
8	7.38	0	2
9	4.43	0	1
10	8.06	0	2

Our next task is to estimate coefficients a , b , and c , given the data. This can be accomplished in a number of ways, but the most commonly method used is the ordinary least squares (OLS). The OLS method is to choose parameters a , b , and c to minimize the sum of squared difference between the observed and predicted values of Y :

$$\text{Min}_{a,b,c} \sum_i (Y_i - a - bD_i - cA_i)^2$$

The first-order conditions for a , b and c will give us values that best ‘fit’ the data.

In addition to obtaining estimates that best fit the data in our sample, we also want to know how likely it is to obtain similar estimates if we were to sample another group of workers. Some variation in the estimates is expected because of idiosyncratic differences between workers, and this sampling noise can be conveniently captured by estimating variances of estimates.

Most statistical software packages, such as Stata, will routinely produce estimated coefficients of the model and their variances. For example, the Stata output for our model is as follows:

productivity	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
piecerate	1.635	.6780478	2.41	0.047	.0316715 3.238328
ability	2.125	.4205074	5.05	0.001	1.130658 3.119342
_cons	3.294	.9212855	3.58	0.009	1.115506 5.472494

The first column (*Coef.*) reports the estimated coefficients b , c , and a , respectively. Therefore, the average productivity of a salary worker with no ability is about 3.3 (i.e. $a=3.294$). A piece rate worker of same ability as a salary worker has on average 1.6 higher productivity ($b=1.635$). Lastly, workers’ productivity increases by about 2.1 for every additional unit of ability, independently of how workers are paid ($c=2.125$).

The second column (*Std. Err.*) indicates the standard errors of estimates b , c , and a , respectively. These standard errors, in conjunction with the estimated coefficients, can be used to test hypotheses about the value of coefficients. One commonly used hypothesis is to test whether there is any relationship between the outcome and a given independent variable (i.e. whether the coefficient is equal to zero).

For example, we may be interested to know whether how workers are paid has an impact on workers’ productivity (i.e. whether $b=0$). Given the assumption that u has a normal distribution, we can use the t-test to test this hypothesis. The t-statistic is conveniently displayed in the third column of the table (t). To determine whether this t-statistic falls inside the region of t-values that are consistent with hypothesis $b=0$ or inside the region that is not consistent with the hypothesis, we need to consult the Student t table. As a rule of thumb, if the t-statistic is greater than 2 or smaller than -2,

the hypothesis that the coefficient is equal to zero (i.e. no relationship exists) can be rejected at the 5 percent confidence level. When this is the case, we say that the coefficient is statistically significant. In our example, all of our three estimates coefficients are statistically significant (i.e. different from zero) because their t-values exceed 2 in the absolute value. Therefore, both how workers are paid and workers' ability significantly influence the average productivity of the workers.

The fourth column ($P > |t|$) gives us the exact confidence level at which we can reject the hypothesis that the coefficient is equal to zero. This value is also known as the p-value. Again, as a rule of thumb, we can conclude that the coefficient is significant if its p-value is equal to or less than 5 percent. The last two columns give us the expected interval for values of coefficients if we were to draw new samples of workers. The interval is specified at the 5 percent confidence level. For example, the piece rate workers are on average 1.6 units more productive than salary workers, controlling for the worker's ability, but this estimate can range between 0.03 and 3.3 in other samples.

The estimates presented in the table give us opportunity to say at least three things about the impact of each independent variable on the outcome variable:

1. Whether the relationship exists or not (statistical significance);
2. Whether the relationship is positive or negative; and
3. How strong is the relationship (economic significance).

In our example, the results indicate that how workers are paid influences their productivity ($t\text{-value} > |2|$). Moreover, the results show that piece-rate workers are more productive than salary workers ($b=1.6 > 0$). This difference seems to be economically significant: the percentage gain between piece rate and salary workers of similar ability is $1.6/3.3=0.48$, or 48 percent.

While the regression analysis is useful in describing relationships between variables, it is not necessarily informative about whether the independent variables causally affect the outcome variable. In other words, the regression results tell us that the outcome and a given independent variable are correlated, but we cannot say without further assumptions that this correlation indicates a causal relationship. The main reason for this is known as the omitted variable problem. For our example, we know that how workers are paid and their productivity are positively correlated, controlling for workers' ability, but how can we be sure that there are no other variables that can explain this correlation? Put differently, are salary workers a good control group for piece rate workers, even if they had same ability?

Identification strategies, such as randomized experiments, deal specifically with the issue of how to uncover causal relationships. You are referred to the appendices that discuss these strategies in detail.