The Basics of Multiple Regression

5.1. The Basics

Education is not the only factor that affects pay. As shown in Figure 4.2, even for workers with the *same* education, there is remarkable variation in wages. Surely, some of this variation is due to work experience, unionization, industry, occupation, region, and demographics, such as gender, race, marital status, etc. These easily can be accounted for using multiple regression.

For example, one could think of wages as a function of education and work experience:

$$Wage = f(Education, Experience).$$

The longer one spends on a job, the better one gets. If people are paid for their productivity, then workers with more work experience should be more productive, and therefore, paid more. That is,

$$\frac{\Delta Wage}{\Delta Experience} > 0,$$

other things equal.

The complete relationship between wages, education, and experience can be written as

$$\ln(Wage_i) = \beta_1 + \beta_2 Education_i + \beta_3 Experience_i + u_i, \qquad (1)$$

where wages are measured in natural logs. This is a multiple regression model of wages. Because there is more than one explanatory variable, each parameter is interpreted as a partial derivative, or the change in the dependent variable for a change in the explanatory variable, holding all other variables constant. For example,

$$\beta_3 = \frac{\partial \ln(Wage)}{\partial Experience} \approx \left.\frac{\Delta \ln(Wage)}{\Delta Experience}\right|_{Education} \qquad (2)$$

is the effect of experience on the log wage, *holding education constant*. Other ways of saying "holding experience constant" are "controlling for experience" or "accounting for the effect of experience." Because pay is measured in natural logs, β_3 also can be interpreted as

$$\beta_3 = \left.\frac{\%\Delta Wage}{\Delta Experience}\right|_{Education}, \qquad (3)$$

or the "return to experience" in the labor market.

If we group all workers according to their education level (less than high school, high school, some college, college graduates, and more than college), we can compare wages and work experience *within* education categories. This is really what multiple regression does. By looking *within* categories, you are *holding education constant*. From the univariate analysis in Chapter 4, we know that wages increase with education level. Table 5.1 shows that within any given education category (i.e., reading across rows), hourly wages rise with greater work

experience. This suggests β_3 is positive, so that wages increase with work experience controlling for education, but also that work experience explains some of the residual variation in wages within education levels.

Table 5.1

Means, Standard Deviations and Frequencies of Hourly Wages

Years of Education	Years of Work Experience					
	exp<=5	5<x<=10	10<x<=20	20<x<=30	exp>30	Total
Educ<12	6.610577	8.3096154	8.506556	8.6632116	11.499039	9.310918
	2.2335515	4.5823646	3.2871646	3.8071621	9.4367496	6.0386959
	4	10	22	25	25	86
Educ=12	8.5617234	10.222842	11.647422	15.137898	13.069812	12.522641
	3.8917755	6.1940197	7.0772693	7.2318973	6.6605462	6.9705726
	26	45	102	93	96	362
Educ=13	6.0346955	11.595442	12.601342	17.252274	16.064233	13.730448
	2.2878627	5.0236677	6.807096	8.68351	9.4877654	8.0749169
	18	27	67	52	38	202
13<Educ<=16	12.018377	11.547343	19.680886	18.701486	18.666967	16.868053
	5.1686776	4.4477838	10.56787	8.6071216	12.906913	9.5829901
	40	38	78	66	31	253
Educ>16	17.574786	22.328942	28.116649	22.697912	26.153953	24.389087
	6.5705128	11.500545	13.368682	10.918406	10.881327	11.893388
	9	15	35	32	9	100
Total	10.274019	12.073586	15.587707	16.724456	14.907941	14.769702
	5.5195622	7.2312446	10.452645	8.8239055	9.4999288	9.257249
	97	135	304	268	199	1003

Likewise,

$$\beta_2 = \frac{\partial \ln(Wage)}{\partial Education} \approx \frac{\Delta \ln(Wage)}{\Delta Education}\bigg|_{Experience} \tag{4}$$

is the effect of education on the log wage, holding experience constant. β_2 also can be expressed as

$$\beta_2 = \frac{\%\Delta Wage}{\Delta Education}\bigg|_{Experience}, \tag{5}$$

or the return to education in the labor market.

If we group workers according to years of work experience (0-5, 5-10, 11-20, 21-30, >30), we can compare wages and education *within* work experience categories. Again, this is what multiple regression does. By looking *within* experience categories, we are *holding experience constant*. In Table 5.1, within any given experience category (reading down columns), the hourly wage rises with education. This suggests β_2 is positive, so that wages increase with education even when controlling for work experience.

Importantly, multiple regression recognizes possible *interdependence* among explanatory variables. For example, for any individual, education and work experience are determined *in part* by the underlying decision to allocate time. Individuals can go to school or work. Those with more education will have less work experience, and vice versa, holding other factors such as age constant. Thus, education and experience are interdependent. In fact, they are inversely correlated since the sample correlation coefficient, r= –0.186.

This interdependence implies that some of the population variation in education and experience is common. The Venn diagram in Figure 5.1 illustrates this. The two circles represent the variation in education and experience, respectively. Area B is the intersection and represents the variation shared by the variables. This is the co-variation between education and experience. Area A is the remaining variation in education and is due to influences other than experience, and hence, is *independent* of experience. Similarly, area C is the remaining variation in experience, *independent* of education.

When estimating parameters, least squares uses only the *independent* variation in each explanatory variable to estimate that variable's parameter. To estimate β_2, only the independent part of education is used. The formula for the least squares estimator of β_2 is

$$\hat{\beta}_2 = \frac{\widehat{Cov}(\ln(Wage), \text{Independent Part of } Education)}{\widehat{Var}(\text{Independent Part of } Education)}. \qquad (6)$$

and for β_3,

$$\hat{\beta}_3 = \frac{\widehat{Cov}(\ln(Wage), \text{Independent Part of } Experience)}{\widehat{Var}(\text{Independent Part of } Experience)}. \qquad (7)$$

Table 5.2 shows parameter estimates, standard errors and 95% confidence intervals for simple and multiple regression models of the log wage.

Table 5.2 Regression of Log Wages against Education and Experience

Explanatory Variable	(1)	(2)	(3)
Education	0.0933		0.1035
	(0.0067)	----	(0.0066)
	(0.0801, 0.1064)		

Experience		0.0084	0.0129
	----	(0.0017)	(0.0015)
		(0.0051, 0.0117)	(0.0098, 0.0159)
Constant	1.2597	2.3456	0.8629
	(0.0919)	(0.0389)	(0.1008)
	(1.0793, 1.4399)	(2.2692, 2.4220)	(0.6651, 1.0607)
R^2	0.162	0.024	0.217

For the simple regression model 1,

$$\ln(Wage_i) = \beta_1 + \beta_2 Education_i + u_i,$$

an additional year of education is estimated to raise log wages by 0.0933 or in terms of relative change in wages, by a factor of exp(0.0933)=1.098 with a 95% confidence interval of (exp(0.0801), exp(0.1064)) = (1.08, 1.11). Economists often say that the increase in percent wages is 9.3%, an approximation. This is a moderately good return to a one-year investment!

Alternatively, for model 2,

$$\ln(Wage_i) = \beta_1 + \beta_2 Experience_i + u_i,$$

an additional year of experience is estimated to raise the log wage by 0.0084 or to raise the wage by a factor of exp(0.0084)=1.0084 or 0.84%. To put this finding in a more meaningful context, an additional 10 years of experience raises the log wages by 0.084, or raises wages by a factor of exp(0.084)=1.088 or 8.8%. R^2 is 0.024, which means that variation in experience alone explains just 2.4% of the sample variation in log wages.

The estimates for the multiple regression model 3,

$$\ln(Wage_i) = \beta_1 + \beta_2 Education_i + \beta_3 Experience_i + u_i,$$

show that together, education and experience explain 21.7% of the variation in log wages. This is much more than both explain individually (16.2% and 2.4%). So, the whole is greater than the sum of its parts!

Accounting for the effect of experience on wages, an additional year of education is estimated to raise the log wage by 0.1035 or the actual wage by a factor of exp(0.1035)=1.109 or 10.9% (Economists 10.4%). In addition, accounting for the effect of education on wages, an additional year of experience is estimated to raise the log wage by 0.0129 or wages by a factor of 1.013 or 1.3%. Surprisingly, the return to an additional year of experience is significantly less than the return to an additional year of education. In fact, based on these estimates, it would require an additional 8.4 years of work experience to raise wages by the same percent as an additional year of education (10.9/1.3=8.4). Education seems like a good deal!

Interestingly, the estimated effects of education and experience on wages change substantially from simple to multiple regression. An additional year of education is estimated to raise the wages by 9.8% in model (1) but by 10.9% in model (3). That is, the estimated return rises by more than a percentage point once differences in work experience are taken into account.

Because this is a big difference in the return on an investment---you would much prefer a 10.4% to a 9.3% return---it is natural to ask: "Why did this happen?"

The answer is at the heart of multiple regression. There are many less-educated (but more-experienced) workers that earn as much as more-educated (but less-experienced) workers. Without accounting for differences in experience, the better educated appear to get a lower return to education. Is this "low return" *really* because of education? No, it is because of experience.

Simple linear regression does not account for experience; however, multiple regression does. Once differences in experience across workers are taken into account, an additional year of education has a much bigger payoff and the estimated return to education rises. Because education and experience are correlated (or *interdependent*), simple regression confuses or "confounds" the effect of education on wages with the effect of experience on wages. By acknowledging potential correlation between the explanatory variables, multiple regression neatly sorts out each variable's independent effect. Section 6 will discuss "confounding effects" in more detail.

5.2. Gender and Wages

A question of great public interest is whether there is gender inequality in earnings, and, if so, what accounts for it. The basic comparison of *average* wages for men and women in section 3 showed that women earn $4.90 per hour less than men. Because men's average earnings were $17.05, this implies that, on average, women earn about 28.7% less than men (4.90/17.05=0.287).

If pay is based solely on productivity, then this differential could be economically rational only if there were some innate underlying difference in productivity between the sexes. In this case, men would have to be more productive than women to justify their higher wages. If one believes the sexes are equal, then gender difference in wages must be caused by something else. One view is that there is labor market discrimination against women. Another is that there are other, *confounding* factors that affect wages but happen to be correlated with gender. Multiple regression can account for these additional factors. If gender-based wage differentials exist even after controlling for many possible confounding influences, then more credence might be given to the discrimination explanation.

The effect of gender on wages can be modeled simply as

$$\ln(Wage_i) = \beta_1 + \beta_2 Female_i + u_i, \tag{8}$$

where *Female* is an indicator variable that is 1 if the worker is female and 0 otherwise (which, of course, means male). A 0-1 indicator variable such as this is frequently referred to as a categorical or dummy variable. β_2 is the relative change in the wage from going from the 0 category to the 1 category. The way to think about β_2 in this context is to observe the wage of worker first as a male, and then "transform" the worker into a female and observe the wage. If the wage rises, β_2 is positive and women earn more than men. That is, there is a labor marker "premium" to being a woman. If the wage does not change, β_2 is zero and gender has no effect on wages. Finally, if the wage falls, β_2 is negative and men earn more than women. That is, there is a labor market discount to being a woman.

Based on the comparison of mean wages above, one expects β_2 to be negative. The least squares estimate is shown in column (1) of Table 5.3.

Table 5.3 Regression of Log Wages against Female, Education and Experience

Explanatory Variable	(1)	(2)	(3)
Female	0.0933		0.1035
	(0.0067)	----	(0.0066)
	(0.0801, 0.1064)		
Education	0.0933		0.1035
	(0.0067)	----	(0.0066)
	(0.0801, 0.1064)		
Experience		0.0084	0.0129
	----	(0.0017)	(0.0015)
		(0.0051, 0.0117)	(0.0098, 0.0159)
Constant	1.2597	2.3456	0.8629
	(0.0919)	(0.0389)	(0.1008)
	(1.0793, 1.4399)	(2.2692, 2.4220)	(0.6651, 1.0607)
R^2	0.162	0.024	0.217

$\hat{\beta}_2 = -0.3193$, which says that females earn 31.9% less than males. The 95% confidence interval is (-0.3918, -0.2468) and does not include zero. Therefore, this discount to being female is statistically significantly different from zero.

In theory, one explanation for this wage differential could be differences in education between men and women. If men had more education than women on average, then that could explain the gender differential. Unfortunately, this is not the case. From section 3 above, the sample mean education is 13.48 years for men, but 13.42 years for women. That is, men and women have *almost identical* education levels! In fact, the sample correlation between *Education* and *Female* is -0.01. Effectively, the variables are uncorrelated. Therefore, differences in education will not explain the gender difference in wages.

This can be seen in two ways. First, Table 5.4 compares mean hourly wages for males and females *within* education categories. Remember, this is akin to what least squares does to estimate the effect of *Female* on wages holding *Education* constant in a multiple regression.

Table 5.4

Means, Standard Deviations and Frequencies of Hourly Wages

Years of Education	Gender		
	Male	Female	Total

Educ<12	10.609443	7.225408	9.310918
	6.9104135	3.46186	6.0386959
	53	33	86
Educ=12	14.280749	10.601934	12.522641
	7.5401576	5.7210515	6.9705726
	189	173	362
Educ=13	16.397839	10.955284	13.730448
	8.9813645	5.8753599	8.0749169
	103	99	202
13<Educ<=16	19.477352	14.110256	16.868053
	10.390313	7.7854763	9.5829901
	130	123	253
Educ>16	26.977737	20.165499	24.389087
	13.00519	8.371838	11.893388
	62	38	100
Total	17.048445	12.14377	14.769702
	10.240742	7.132306	9.257249
	537	466	1003

It is easy to see that *within* all categories, males are paid more females. Therefore, education cannot explain the gender wage differential.

Second, column (2) of Table 5.3 gives parameter estimates for the multiple regression model

$$\ln(Wage_i) = \beta_1 + \beta_2 Female_i + \beta_3 Education + u_i . \quad (9)$$

β_2 measures the effect of being female on wages (in relative terms), *controlling for education*. $\hat{\beta}_2 = -0.3137$, which says that females earn 31.4% less than males, even after accounting for any differences in education between the sexes! This estimate is virtually unchanged from its value in column (1). The 95% confidence interval is (-0.3797, -0.2477) and does not include zero. Therefore, this discount to being female is statistically significantly different from zero.

The estimated relationship between wages, education, and gender also can be illustrated graphically. Figure 5.1 demonstrates that regressing log wages on education and gender has the effect of fitting separate parallel lines to the relationship between log hourly wages and education for males and females. Parallel lines mean that the increase in log wages for an additional year of education (or the return to education) is the same for males and females, and averages about 0.093 (or 9.3%). The distance between the lines for males and females represents the effect of gender: the line for males is 0.3137 log dollars (or 31.37%) higher than the line for females. Figure 5.2 shows the same relationship, but expressed in terms of the wage rather than the log wage.

Another potential explanation for the gender wage differential could be differences in work experience between men and women. If men had more experience

than women on average, then that could explain the gender differential. Again, this is not the case. From section 3, the sample mean experience is 19.72 years for men, but even more, 20.63 years, for women. That is, women have slightly more experience than men. The sample correlation between *Experience* and *Female* is 0.041. Effectively, the variables are uncorrelated. Therefore, differences in experience will not explain the gender difference in wages.

This is illustrated nicely in column (2) of Table 5.2, which gives parameter estimates for the multiple regression model

$$\ln(Wage_{ie}) = \beta_1 + \beta_2 Female_i + \beta_3 Education + \beta_4 Experience + u_i . \quad (9)$$

β_2 measures the effect of being female on wages (in relative terms), *controlling for education and experience*. $\hat{\beta}_2 = -0.3252$, which says that females earn 32.5% less than males, even after accounting for any differences in education and experience between the sexes. The 95% confidence interval is (-0.3887, -0.2617) and does not include zero. Therefore, this discount to being female is statistically significantly different from zero. The R^2 is 0.289, which means that variation in education, experience, and gender explains 28.9% of the sample variation in log wages.