

Multiple Regression

I. Introduction

In the last chapter we looked at the simple linear regression model where we have only one explanatory (or independent) variable. This can be easily expanded to a multivariate setting. Our model can be written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

So we would have k explanatory variables. The interpretation of the β 's is the same as in the simple regression framework. For example β_1 is the marginal influence of X_1 on the dependent variable Y, **holding all the other explanatory variables constant**.

This is easy to do in Excel. It is similar to multiple regression except that one needs to have all the X variables side by side in the spreadsheet.

Inference about individual coefficients is exactly the same as in simple regression.

Suppose we have the following data for 10 hospitals:

Y	X1	X2
Cost	Size	Visibility
2750	225	6
2400	200	37
2920	300	14
1800	350	33
3520	200	11
2270	250	21
3100	175	21
1980	400	22
2680	350	20
2720	275	16

Cost is the cost per case for each hospital, Size is the size of the hospital in the number of beds, and visibility is a scale that measures how much the administrator knows about competitor hospitals. In this case we might expect larger hospitals to have lower costs per case, and when the administrator has more knowledge about his competition costs will be lower as well.

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.834034
R Square	0.695612
Adjusted R Square	0.608644
Standard Error	323.9537
Observations	10

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	1678818	839409	7.998485	0.01556
Residual	7	734622	104946		
Total	9	2413440			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4240.131	435.6084	9.733813	2.56E-05	3210.081	5270.181
Size	-3.76232	1.442784	-2.60768	0.035032	-7.17395	-0.35068
Visibility	-29.8955	11.66298	-2.56328	0.037372	-57.4741	-2.31699

So in this case we'd say that each bed lowers the per case cost of the hospital by \$3.76, and every one unit increase in the visibility scale lowers costs by 29.90. Note that these are not the same results we would get if we did two simple regressions:

If we only included Size:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.640237
R Square	0.409904
Adjusted R Square	0.336142
Standard Error	421.9245
Observations	10

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	989277.9	989277.9	5.557108	0.046149
Residual	8	1424162	178020.3		
Total	9	2413440			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	3804.717	522.4326	7.282694	8.53E-05	2599.984	5009.449	2599.984	5009.449
Size	-4.3696	1.853606	-2.35735	0.046149	-8.64403	-0.09518	-8.64403	-0.09518

While if we only included Visibility:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.632394
R Square	0.399922
Adjusted R Square	0.324912
Standard Error	425.4781
Observations	10

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	965187.2	965187.2	5.331595	0.049765
Residual	8	1448253	181031.6		
Total	9	2413440			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	3315.282	332.1822	9.980311	8.61E-06
Visibility	-34.8896	15.11012	-2.30902	0.049765

Why the changes in the coefficients?

Note the R-squared do not add up. The multiple Regression R^2 is .695, and the two simple R^2 's are .40 and .41.

II. The Coefficient of Determination

Recall from last time that the coefficient of determination, or R^2 , indicates how much of the variation in the dependent variable is explained by the variation in the explanatory variable. This is the same in a multivariate regression setting:

$$R^2 = SSR/SST$$

Where SSR =Regression sum of squares = $\sum (\hat{Y}_i - \bar{Y})^2$

And SST = Total sum of squares: = $\sum (Y_i - \bar{Y})^2$

Note that the only difference here is that \hat{Y} is constructed from the multivariate equation.

One problem with the R^2 , however, is that it turns out to be the case that as you add explanatory variables, the R^2 increases even if these explanatory variables do not add anything of value to the equation.

Suppose we add a variable to the above analysis: Size of Pillow and get the following:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.836031
R Square	0.698948
Adjusted R Square	0.548422
Standard Error	347.9873
Observations	10

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	3	1686869	562289.6	4.643369	0.052481
Residual	6	726571.1	121095.2		
Total	9	2413440			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4214.048	478.7347	8.80247	0.000119	3042.626	5385.471
Size	-3.87199	1.607127	-2.40926	0.052627	-7.80449	0.060515
Visibility	-30.4994	12.74526	-2.393	0.053806	-61.686	0.687151
size of pillow	1.247354	4.837621	0.257844	0.805143	-10.5899	13.08459

Note that size of pillow is not a relevant variable, the t-stat is .25 so it is not significantly different from zero. But note the R^2 increases to .699 from .696, not a large change, but it increases. The problem is that we could continue to add irrelevant things to the model and the R^2 will continue to rise. We need to account for this somehow.

As we add more and more “things” to our model, the R^2 will increase, not necessarily because we are explaining more of the variation, but simply because of the way the R^2 is constructed.

So we’d like to come up with a way to account for this. The Adjusted R^2 is computed to reflect both the number of explanatory variables in the model and the sample size. The Adjusted R^2 basically includes a penalty for the number of right-hand-side (RHS) variables.

Adjusted R^2 (also denoted as \bar{R}^2) = $1 - \left[(1 - R^2) * \frac{n - 1}{n - k - 1} \right]$

Note that the final term $(n-1)/(n-k-1)$ will get larger as k (the number of RHS variables) increases, and thus, the adjusted R^2 will become smaller. So in order for the adjusted R^2 to increase as variables are added to the model, the R^2 needs to increase by enough to overcome the penalty associated with adding more variables.

Note in my example above $R^2 = .699$, $n=10$, $k= 3$ so:

$$\text{Adjusted } R^2 = 1 - [(1 - .699) * 9 / 6] = .548$$

Notice in the model with only Size and Visibility, the Adjusted R^2 is .609. So we get the result that adding an irrelevant variable does nothing to adjusted R^2

Excel gives you the adjusted R^2 in its output.

III. Testing for the Significance of the Multiple Regression Model

F-test

Another general summary measure for the regression model is the F-test for overall significance. This is testing whether or not any of our explanatory variables are important determinants of the dependent variable. This is a type of ANOVA test.

Our null and alternative hypotheses are:

Ho: $\beta_1 = \beta_2 = \dots = \beta_k = 0$ (none of the variables are significant)

H1: At least one $\beta_j \neq 0$

Here the F statistic is:

$$F = \frac{SSR/k}{SSE/(n-k-1)}$$

Notice that this statistic is the ratio of the regression sum of squares to the error sum of squares. If our regression is doing a lot towards explaining the variation in Y then SSR will be large relative to SSE and this will be a “big” number. Whereas if the variables are not doing much to explain Y, then SSR will be small relative to SSE and this will be a “small” number.

This ratio follows the F distribution with k and n-k-2 degrees of freedom.

The middle portion of the Excel output contains this information (this is the model with school and experience, not shoe size):

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	1678818	839409	7.998485	0.01556
Residual	7	734622	104946		
Total	9	2413440			

$$F = (1678818/2)/(734622/(10-2-1)) = 839409/104946 = 7.998$$

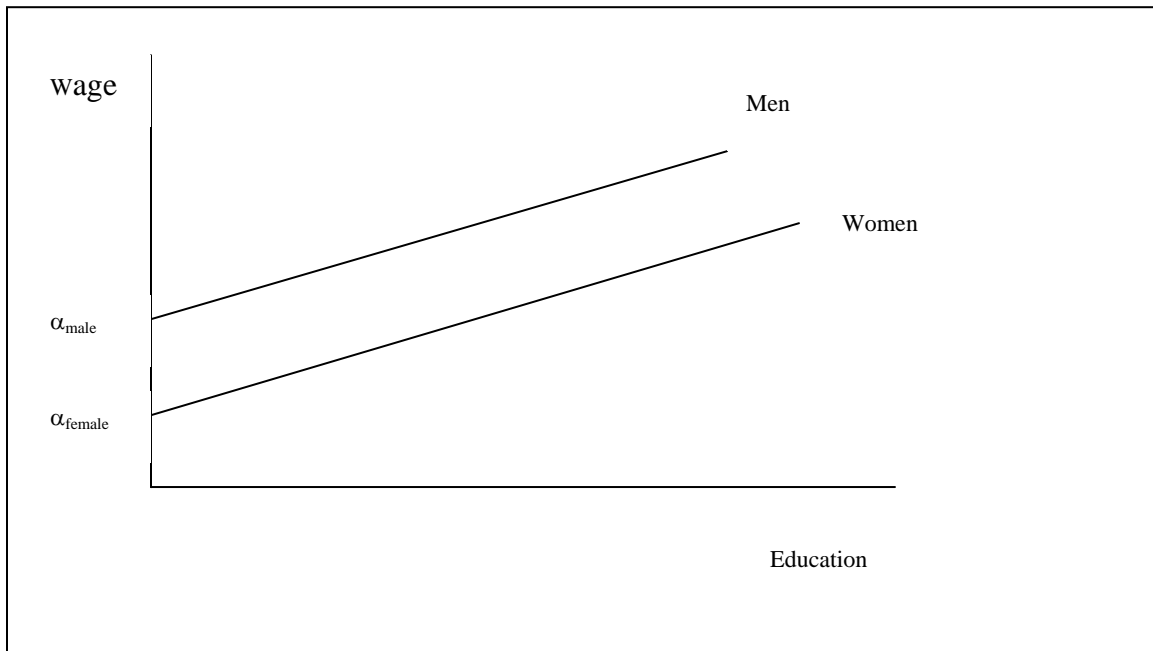
The “Significance F” is the p-value. So we’d reject Ho and conclude there is evidence that at least one of the explanatory variables is contributing to the model. Note that this is

a pretty weak test: it could be only one of the variables or it could be all of them that matter, or something in between. It just tells us that something in our model matters.

IV. Dummy Variables in Regression

Up to this point we've assumed that all the explanatory variables are numerical. But suppose we think that, say, earnings might differ between males and females. How would we incorporate this into our regression?

The simplest way to do this is to assume that the only difference between men and women is in the intercept (that is the coefficients on all the other variables are equal for men and women).



Assume for now the only other variable that matters is education. The idea is that we think men make more than women independent of education. That is the male intercept (α_{male}) is greater than the female intercept (α_{female}). We can incorporate this into our regression by creating a dummy variable for gender. Suppose we let the variable Male = 1 if the individual is a male, and 0 otherwise. Then our equation becomes:

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Experience}_i + \beta_3 \text{Male}_i + \varepsilon_i$$

So if the individual is male the variable Male is “on” and if she is female Male is “off”. The coefficient β_3 indicates how much extra the wage of males is than female (this can be positive or negative in theory). In terms of our graph, $\alpha_{\text{female}} = \beta_0$, and $\alpha_{\text{male}} = \beta_0 + \beta_3$.

So the dummy variable indicates how much the intercept shifts up or down for that group.

This can be done for more than two categories. Suppose we think that earnings also differ by race then we can write:

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Experience}_i + \beta_3 \text{Male}_i + \beta_4 \text{Black}_i + \beta_5 \text{Asian}_i + \beta_6 \text{Other}_i + \varepsilon_i$$

Where Black is a dummy variable equal to 1 if the individual is black, Asian =1 if the individual is Asian, other =1 for other nonwhite race. Note that white is omitted from this group. Just like female is omitted. Thus the coefficients β_4 , β_5 , and β_6 indicate how wages differ for blacks, Asian, and other, relative to whites:

Note that if there are x different categories, we include x-1 dummy variables in our model. The omitted group is always the comparison.

Suppose we estimate this and get:

$$\widehat{\text{Wage}} = -20 + 3.12 * \text{School} + .39 * \text{Experience} + 4 * \text{Male} - 5 * \text{black} - 3 * \text{Asian} - 3 * \text{other}.$$

- So males make \$4/hour more than females
- Blacks make \$5 less than whites
- Asians make \$3 less than whites
- Other races make \$3 less than whites.

Note that this is controlling for differences in other characteristics. That is, a male with the same level of schooling, experience and race will earn \$4 more than the identical female. Similarly for race.

V. Interaction Effects

Suppose we're interested in explaining total charges and we think LOS and gender are the explanatory variables. We could estimate our model and get something like:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.891447
R Square	0.794678
Adjusted R Square	0.770522
Standard Error	973.619
Observations	20

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
--	-----------	-----------	-----------	----------	-----------------------

Regression	2	62370927	31185463	32.89835	1.43E-06
Residual	17	16114878	947934		
Total	19	78485805			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	1488.768	533.0511	2.792919	0.01249	364.1272	2613.41
LOS	755.0667	108.7859	6.940847	2.38E-06	525.5481	984.5853
Female	-544.165	473.0625	-1.1503	0.265944	-1542.24	453.9114

Interpret this.

Now what if we think the effect of LOS on charges is different for males than females. How might we deal with this? Note that the idea is that not only is there an intercept difference, but there is a slope difference as well. To get at this we can interact LOS and Female – that is create a new variable that multiplies the two together, then we would get something like the following:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.927739
R Square	0.860699
Adjusted R Square	0.834581
Standard Error	826.6315
Observations	20

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	3	67552690	22517563	32.95319	4.43E-07
Residual	16	10933115	683319.7		
Total	19	78485805			

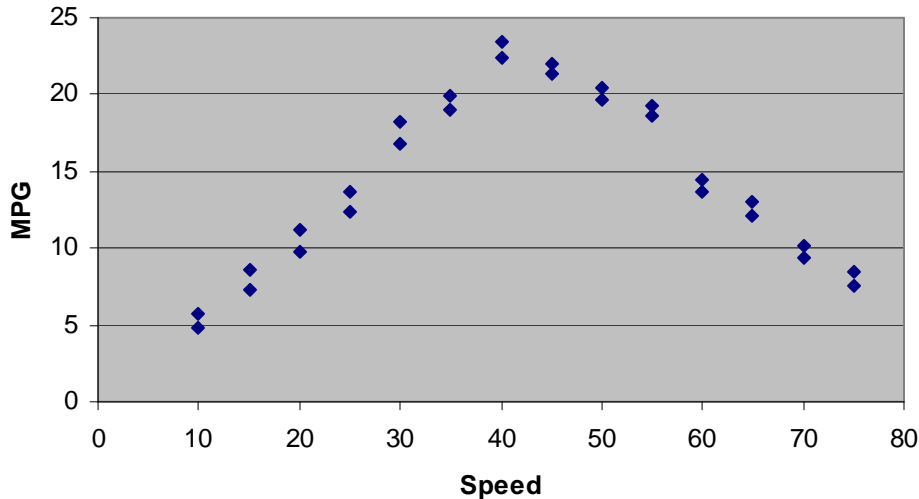
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	1959.009	483.7202	4.04988	0.000929	933.5682	2984.45
LOS	637.5065	101.7513	6.26534	1.13E-05	421.8034	853.2096
Female	-2280.05	747.4512	-3.05044	0.007632	-3864.58	-695.527
LOS*Female	667.8416	242.5195	2.753764	0.014124	153.7233	1181.96

Note the adjusted R^2 increases which suggests that adding this new variable is “worth it”. How do we interpret?

Females have charges that are \$2,280 lower than males holding constant LOS. A one unit increase in LOS increases charges for males by \$637 for males, while the effect for females is \$667 LARGER. That is each day in the hospital increases charges for females by $637.5+667.5 = \$1,305$. So females start at a lower point, but increase faster with LOS than do males.

VI. Non-linear terms in the regression model

Suppose we think there is a relationship between miles per gallon and speed. If miles per gallon is the dependent variable, we might think that speed has a nonlinear effect on MPG. Consider the following scatter diagram:



Note that MPG increases with speed initially but then decreases once speed reaches about 40 MPH. Note that if we had to draw a straight line through this, it would miss a lot. The best line would be a curved line. We can incorporate this into the regression model as follows:

$$\text{MPG} = \beta_0 + \beta_1 \text{Speed} + \beta_2 \text{Speed}^2 + \varepsilon$$

We include a speed squared term. This will capture the quadratic effect of the relationship.

Estimating this model yields:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.958546
R Square	0.91881
Adjusted R Square	0.912315
Standard Error	1.663417
Observations	28

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	782.8247	391.4123	141.4596	2.34E-14
Residual	25	69.17387	2.766955		
Total	27	851.9986			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-7.55555	1.424109	-5.30546	1.69E-05
Speed	1.271694	0.075732	16.792	3.99E-15
SpeedSQ	-0.0145	0.000872	-16.6325	4.97E-15

So $MPG = -7.55 + 1.27 * Speed - .0145 * Speed^2$

Speed increases MPG at a decreasing rate. The relationship between MPG and Speed is dependent on the current Speed.

Someone going 30 mph would be predicted to get $-73.55 + 1.27(30) - .0145(30^2)$
 $= -7.55 + 38.1 - 13.5 = 17.05$ MPG.

The easiest way to interpret this is to take the partial derivative of MPG with respect to Speed.

That is, how does MPG change as speed change?:

$\partial MPG / \partial Speed = 1.27 + 2(-.0145) * Speed$, or $1.27 - .029 * Speed$

At 1 MPH increasing speed by 1 MPH would result in a $1.27 - .029 = 1.241$ increase in MPG. At 20 MPH increasing speed by 1 MPH would result in a $1.27 - .029(20) = .69$ increase in MPH. As speed increases note that the speed squared term will begin to kick in until it begins to overtake the linear term. At 50 MPH, increasing speed by 1 MPH would result in a $1.27 - .029(50) = .18$ decrease in MPG.

Note that if we take this partial derivative and set it equal to zero and solve for speed we can find the speed for which MPG is maximized:

$1.27 - .029 * Speed = 0$

$.029 * Speed = 1.27$

$Speed = 1.27 / .029 = 43.79$

So a Speed of 43.79 MPH will maximize MPG