

Multiple Regression II (Soc 504)

1 Evaluation of Model Fit and Inference

Previously, we discussed the theory underlying the multiple regression model, and we derived the OLS estimator for the model. In that process, we discussed the standard error of the regression, as well as the standard errors for the regression coefficients. We will now develop the t-tests for the parameters, as well as the ANOVA table for the regression and the regression F test.

However, first, we must be sure that our data meet the assumptions of the model. The assumptions of multiple regression are no different from those of simple regression:

1. Linearity
2. Homoscedasticity
3. Independence of errors across observations
4. Normality of the errors.

Typically, assumptions 1-4 are lumped together as: $Y \sim N(X\beta, \sigma_e^2 I)$, or equivalently as $e \sim N(0, \sigma_e^2 I)$. If linearity holds, then the expectation of the error is 0 (see preceding formula). If homoscedasticity holds, then the variance is constant across observations (see the identity matrix in the preceding formulas). If independence holds, then the off-diagonal elements of the error variance matrix will be 0 (also refer to the identity matrix in the preceding formulas). Finally, if the errors are normal, then both Y and e will be normally distributed, as the formulas indicate.

If the assumptions are met, then inference can proceed (if not, then we have some problems-we will discuss this soon). This information is not really new-much of what we developed for the simple regression model is still true. The basic ANOVA table is (k is the number of x variables INCLUDING the intercept):

ANOVA TABLE	DF	SS	MS	F	Sig
Regression	k-1	$(X\beta - \bar{Y})^T(X\beta - \bar{Y})$	$\frac{SSR}{Df(Reg)}$	$\frac{MSR}{MSE}$	(from F table)
Residual	n-k	$e^T e$	$\frac{SSE}{Df(E)}$ (called MSE)		
Total	n-1	$(Y - \bar{Y})^T(Y - \bar{Y})$			

$$R^2 = \frac{SSR}{SST} \text{ or } 1 - \frac{SSE}{SST} \text{ and Multiple Correlation} = \sqrt{R^2}.$$

Note that in the sums of squares calculations above, I have simply replaced the scalar notation with matrix notation. Also note that the mean of Y is a column vector ($n \times 1$) in which the elements are all the mean of Y . There is therefore no difference between these calculations and those presented in the simple regression notes-in fact, if you're more comfortable with scalar notation, you can still obtain the sums of squares as before.

The multiple correlation before was simply the bivariate correlation between X and Y . Now, however, there are multiple X variables, so the multiple R has a different interpretation. Specifically, it is the correlation between the observed values of Y and the model-predicted values of Y ($\hat{Y} = X\hat{\beta}$).

The F -Test is now formally an overall (global) test of the model's fit. If at least one variable has a significant coefficient, then the model F should be significant (except in cases in which multicollinearity is a problem-we will discuss this later). There is more use for F in the multiple regression context, as we will discuss shortly.

The t -tests for the parameters themselves are conducted just as before. Recall from the last set of notes that the standard errors (more specifically, the variances) for the coefficients are obtained using: $\hat{\sigma}_e^2(X^T X)^{-1}$. This is a $k \times k$ variance-covariance matrix for the coefficients. However, we are generally only interested in the diagonal elements of this matrix. The diagonal elements are the covariances of the coefficients with themselves. Hence, their square roots are the standard errors. The off-diagonal elements of this matrix are the covariances of the coefficients with other coefficients and are not important for our t -tests. Since we don't know σ_e^2 , we replace it with the MSE from the ANOVA table.

For example, suppose we have data that look like:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{bmatrix} \quad Y = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 3 \\ 3 \end{bmatrix}, \quad \text{so } (X^T Y) = \begin{bmatrix} 12 \\ 50 \end{bmatrix}$$

So,

$$(X^T X) = \begin{bmatrix} 6 & 21 \\ 21 & 91 \end{bmatrix}$$

and

$$(X^T X)^{-1} = \begin{bmatrix} \frac{91}{105} & \frac{-21}{105} \\ \frac{-21}{105} & \frac{6}{105} \end{bmatrix}$$

If we compute $(X^T X)^{-1}(X^T Y)$, we get the OLS solution:

$$\begin{bmatrix} \frac{42}{105} \\ \frac{48}{105} \end{bmatrix},$$

which will ultimately yield an MSE of .085714 (if you compute the residuals, square them, sum them, and divide by 4).

If we compute $MSE(X^T X)^{-1}$, Then we will get:

$$\begin{bmatrix} .0743 & -.0171 \\ -.0171 & .0049 \end{bmatrix},$$

the variance-covariance matrix of the coefficients. If we take the diagonal elements and square root them, we get .273 and .07 as the standard errors of b_0 and b_1 , respectively. The off-diagonal elements tell us how the coefficients are related to each other. We can get the correlation between b_0 and b_1 by using the formula: $Corr(X, Y) = \frac{Cov(X, Y)}{S(X)S(Y)}$. In this case, the ‘S’ values are the standard errors we just computed. So, $Corr(b_0, b_1) = \frac{-.0171}{(.273)(.07)} = -.895$. This indicates a very strong negative relationship between the coefficients, which is typical in a bivariate regression setting. This tells us that if the intercept were large, the slope would be shallow (and vice versa), which makes sense.

If we wish to make inference about a single parameter, then the t test on the coefficient of interest is all we need to determine whether the parameter is ‘significant.’ However, occasionally we may be interested in a test of a set of parameters, rather than a single parameter. For example, suppose our theory suggests that 3 variables are important in predicting an outcome, net of a host of control variables. Ideally we should be able to construct a joint test of the importance of all 3 variables. In this case, we can conduct ‘nested F tests.’ There are numerous approaches to doing this, but here we will discuss two equivalent approaches. One approach uses R^2 from two nested models; the other uses the ANOVA table information directly.

The F -test approach is computed as:

$$F = \frac{SSE(R) - SSE(F)}{df(R) - df(F)} \div \frac{SSE(F)}{df(F)}$$

Here, R refers to the reduced model (with fewer variables), and F refers to the full model (with all the variables). The degrees of freedom for the reduced model will be greater than for the full model, because the degrees of freedom for the error are $n - k$, where k is the number of regressors (including the intercept). The numerator df for the test is $df(R) - df(F)$; the denominator df is $df(F)$. Recognize that this test is simply determining the proportional increase in error the reduced model generates relative to the full model.

An equivalent formulation of this test can be constructed using R^2 :

$$F = \frac{R_F^2 - R_R^2}{df(R) - df(F)} \div \frac{1 - R_F^2}{df(F)}$$

Simple algebra shows that these are equivalent tests; however, when there is no intercept in the models, we must use the first calculation.

2 Interpretation of the Model Parameters

The interpretation of the parameters in this model is ultimately no different than the interpretation of the parameters in the simple linear regression with one key difference. The

parameters are now interpreted as the relationship between X and Y , *net of* the effects of the other variables. What exactly does this mean? It means simply that the coefficient for X in the model represents the relationship between X and Y , holding all other variables constant. Let's look at this a little more in depth. Suppose we have a data set consisting of 3 variables: X_1 , X_2 , and Y , and the data set looks like:

X_1	X_2	Y
0	1	2
0	2	4
0	3	6
0	4	8
0	5	10
1	6	12
1	7	14
1	8	16
1	9	18
1	10	20

Suppose we are interested in the relationship between X_1 and Y . The mean of Y when $X_1 = 0$ is 6, and the mean of Y when $X_1 = 1$ is 16. In fact, when a regression model is conducted, here are the results:

ANOVA TABLE	Df	SS	MS	F	Sig
Regression	1	250	250	25	$p = .001$
Residual	8	80	10		
Total	9	330			

Effect	Coefficient	Standard Error	t	p-value
Intercept	6	1.41	4.24	.003
Slope	10	2	5.00	.001

The results indicate there is a strong relationship between X_1 and Y . But, now let's conduct a multiple regression that includes X_2 . Those results are:

ANOVA TABLE	Df	SS	MS	F	Sig
Regression	2	330	165	∞	$p = 0$
Residual	7	0	0		
Total	9	330			

Effect	Coefficient	Standard Error	t	p-value
Intercept	0	0	∞	0
b_{X_2}	2	0	∞	0
b_{X_1}	0	0	∞	0

Notice how the coefficient for X_1 has now been reduced to 0. Don't be confused by the fact that the t-test for the coefficient is large: this is simply an artifact of the contrived nature of the data such that the standard error is 0 (and the computer considers $\frac{0}{0}$ to be ∞). The coefficient for X_2 is 2, and the R^2 for the regression (if you compute it) is 1, indicating perfect linear fit without error.

This result shows that, once we have controlled for the effect of X_2 , there is no relationship between X_1 and Y : that X_2 accounts for all of the relationship between X_1 and Y . Why? Because the effect of X_1 we first observed is simply capturing the fact that there are differences in X_2 by levels of X_1 . Let's look at this in the context of some real data.

Race differences in birthweight (specifically black-white differences) have been a hot topic of investigation over the last decade or so. African Americans tend to have lighter babies than whites. Below is a regression that demonstrates this. The data are from the National Center for Health Statistics, which has recorded every live birth in the country for over a decades. The data I use here are a subset of the total births for 1999.

ANOVA TABLE	Df	SS	MS	F	Sig
Regression	1	$2.94e + 8$	$2.94e + 8$	815.15	$p = 0$
Residual	36,349	$1.31e + 10$	361,605.4		
Total	36,350	$1.34e + 10$			

Effect	Coefficient	Standard Error	t	p-value
Intercept	3364.1	3.44	977.07	0
Black	-245.1	8.58	-28.55	0

These results indicate that the average white baby weighs 3364.1 grams, while the average African American baby weighs $3364.1 - 245.1 = 3119$ grams. The racial difference in these birth weights is significant, as indicated by the t -test. The question is: why is there a racial difference? Since birth weight is an indicator of health status of the child, we may begin by considering that social class may influence access to prenatal care, and prenatal care may increase the probability of a healthy and heavier baby. Thus, if there is a racial difference in social class, this may account for the racial difference in birthweight. So, here are the results for a model that 'controls on' education (a measure of social class):

ANOVA TABLE	Df	SS	MS	F	Sig
Regression	2	$3.71e + 8$	$1.85e + 8$	515.57	$p = 0$
Residual	36,348	$1.31e + 10$	359,525.5		
Total	36,350	$1.34e + 10$			

Effect	Coefficient	Standard Error	t	p-value
Intercept	3149.8	15.14	208.06	0
Black	-236.96	8.58	-27.62	0
Education	16.67	1.15	14.54	0

Education, in fact, is positively related to birthweight: more educated mothers produce heavier babies. The coefficient for ‘Black’ has become smaller, indicating that, indeed, part of the racial difference in birthweight is attributable racial differences in educational attainment. In other words, at similar levels of education, the birthweight gap between blacks and whites is about 237 grams. The overall mean difference, however, is about 245 grams, if educational differences in these populations are not considered.

Let’s include one more variable: gestation length. Obviously, the longer a fetus is *in utero*, the heavier it will tend to be at birth. In addition to social class differences, there may be gestation-length differences between blacks and whites (perhaps another proxy for prenatal care). When we include gestation length, here are the results:

ANOVA TABLE	Df	SS	MS	F	Sig
Regression	3	$4.17e + 9$	$1.4e + 9$	5443.88	$p = 0$
Residual	36,347	$9.3e + 9$	255,108.2		
Total	36,350	$1.34e + 10$			

Effect	Coefficient	Standard Error	<i>t</i>	p-value
Intercept	-1695.4	41.7	-40.64	0
Black	-160.44	7.25	-22.12	0
Education	15.7	.97	16.26	0
Gestation	124.9	1.02	121.98	0

These results indicate that a large portion of the racial difference in birthweight is attributable to racial differences in gestation length: notice how the race coefficient has been reduced to -160.44 . Interestingly, the effect of education has also been reduced slightly, indicating that there are some gestational differences by education level. Finally, notice now how the intercept has become large and negative. This is because there are no babies with gestation length 0. In fact, in these data, the minimum gestation time is 17 weeks (mean is just under 40 weeks). Thus, a white baby with a mother with no education who gestated for 17 weeks would be expected to weigh a little more than 1,000 grams (just over two pounds). As we’ve discussed before, taken out of the context of the variables included in the model, the intercept is generally meaningless.

3 Comparing Coefficients in the Model

In multiple regression, it may be of interest to compare effects of variables in the model. Comparisons are fairly clear cut if one variable reaches significance while another one doesn’t, but generally results are not that clear. Two variables may reach significance, but you may be interested in which one is *more* important. To some extent, perhaps even a large extent, this is generally more of a substantive or situational question than a statistical one. For example, from a policy perspective, if one variable is changeable, while another one isn’t, then the more important finding may be that the changeable variable has a significant effect even after controlling for unchangeable ones. In that case, an actual comparison of coefficients

may not be warranted. But let's suppose we are testing two competing hypotheses, each of which is measured by its own independent variable.

Let's return to the birth weight example. Suppose one theory indicates that the racial difference in birthweight is a function of prenatal care differences, while another theory indicates that birthweight differences are a function of genetic differences between races. (As a side note, there in fact *is* a somewhat contentious debate in this literature about racial differences in genetics). Suppose that our measure of prenatal care differences is education (reflecting the effect of social class on ability to obtain prenatal care), and that our measure of genetic differences is gestation length (perhaps the genetic theory explicitly posits that racial differences in genetics lead to fewer weeks of gestation). We would like to compare the relative merits of the two hypotheses. We have already conducted a model that included gestation length and education. First, we should note that, because the racial differences remained even after including these two variables, neither is a sufficient explanation. However, both variables have a significant effect, so how do we compare their effects?

It would be unreasonable to simply examine the relative size of the coefficients (about 15.7 for education and 124.9 for gestation), because these variables are measured in different units. Education is measured in years, while gestation is measured in weeks. Our first thought may be, therefore, to recode one or the other variable so that the units match. So, what are the results if we recode gestation length into years? If we do so, we will find that the effect of education does not change, but the effect of gestation becomes 6,494.9. Nothing else in the model changes (including the *t*-test, the model *F*, *R*², etc.). But now, the gestation length effect appears huge relative to the education effect. But, to demonstrate why this approach doesn't work, suppose we rerun the model after recoding education into centuries. In that case, the effect of education becomes 1,570.5. Now, you could argue that measuring education in centuries and comparing to gestation length in years does not place the two variables in the same units. However, I would argue that neither of these units is more inappropriate than the other. Practically no (human) fetus gestates for a year, and no one attends school for a century. The problem is thus a little more complicated than simply a difference in units of measurement. Plus, I would add that it is often impossible to make units comparable: for example how would you compare years of education with salary in dollars?

What we are missing is some measure of the *variability* in the measures. As a first step, I often compute the effect of variables at the smallest and largest values possible in the measure. For example, gestation length has a minimum and maximum of 17 and 52, respectively. Thus, net effect of gestation for a fetus that gestates for 17 weeks is 2,123 grams, while the net effect of gestation for a fetus that gestates for 52 weeks is 6,495 grams. Education, on the other hand, ranges from 0 to 17. Thus, the net effect of education at 0 years is 0 grams, while the net effect at 17 years is 267 grams. On its face, then, the effect of gestation length appears to be larger.

However, even though there is a wider range in the net effect of gestation versus education, we have not considered that the extremes may be just that—extreme—in the sample. A 52-week gestation period is a very rare (if real) event, as is having no education whatsoever. For this reason, we generally *standardize* our coefficients for comparisons using some function of the standard deviation of the *X* variables of interest and sometimes the standard deviation of *Y* as well. For fully standardized coefficients (often denoted using β as opposed to *b*), we

compute:

$$\hat{\beta}_X = b_X \frac{s.d.(X)}{s.d.(Y)}.$$

If we do that for this example, we get standardized effects for education and gestation of .07 and .53, respectively. The interpretation is that, for a standard unit increase in education, we expect a .07 standard unit increase in birth weight; whereas for gestation, we expect a .53 standard unit increase in birth weight for a one standard unit increase in gestation. From this perspective, the effect of gestation length indeed seems more important.

As a final note on comparing coefficients, we must realize that it is not always useful to compare even the standardized coefficients. The standardized effect of an X variable that is highly nonnormal may not be informative. Furthermore, we cannot compare standardized coefficients of dummy variables, because they only take two values (more on this in the next set of notes). Finally, our comparisons cannot necessarily help us determine which hypothesis is more correct: in part, this decision rests on how well the variables operationalize the hypothesis. For example, in this case, I would not conclude that genetics is more important than social class (or prenatal care). It could easily be argued (and perhaps more reasonably) that gestation length better represents prenatal care than genetics!

4 The Maximum Likelihood Solution

So far, we have derived the OLS estimator, and we have discussed the standard error estimator, for multiple regression. Here, I derive the Maximum Likelihood estimator and the standard errors.

Once again, we have normal likelihood function, but we will now express it in matrix form:

$$p(Y | \beta, \sigma_e, X) \equiv L(\beta, \sigma_e | X, Y) = (2\pi\sigma_e^2)^{\frac{-n}{2}} \exp \left\{ \frac{-1}{2\sigma_e^2} (Y - X\beta)^T (Y - X\beta) \right\}$$

Taking the log of this likelihood yields:

$$LL(\beta, \sigma_e) \propto -n \log(\sigma_e) - \frac{1}{2\sigma_e^2} \{ (Y - X\beta)^T (Y - X\beta) \}$$

We now need to take the partial derivative of the log of the likelihood with respect to each parameter. Here, we will consider that we have two parameters: the vector of coefficients and the variance of the error. The derivative of the log likelihood with respect to β should look somewhat familiar (much of it was shown in the derivation of the OLS estimator):

$$\frac{\partial LL}{\partial \beta} = \frac{-1}{2\sigma_e^2} (-2X^T)(Y - X\beta) = \frac{1}{\sigma_e^2} (X^T)(Y - X\beta).$$

If we set this expression equal to 0 and multiply both sides by σ_e^2 , we end up with the same result as we did for the OLS estimator. We can also take the partial derivative with respect to σ_e :

$$\frac{\partial LL}{\partial \sigma_e} = \frac{-n}{\sigma_e} + \sigma_e^{-3} \{(Y - X\beta)^T(Y - X\beta)\}$$

The solution, after setting the derivative to 0 and performing a little algebra, is:

$$\sigma_e^2 = \frac{\{(Y - X\beta)^T(Y - X\beta)\}}{n}.$$

These solutions are the same as we found in the simple regression problem, only expressed in matrix form.

The next step is to take the second partial derivatives in order to obtain the standard errors. Let's first simplify the first partial derivatives (and exchange σ_e^2 with τ):

$$\frac{\partial LL}{\partial \beta} = -\frac{1}{\tau}(X^T Y - X^T X\beta).$$

and

$$\frac{\partial LL}{\partial \tau} = -\frac{n}{2}\tau^{-1} + \frac{1}{2}\tau^{-2}e^T e.$$

The second partial derivative of LL with respect to β is:

$$\frac{\partial^2 LL}{\partial \beta^2} = -\frac{1}{\tau}(X^T X).$$

The second partial derivative of LL with respect to τ is:

$$\frac{\partial^2 LL}{\partial \tau^2} = \frac{n}{2}\tau^{-2} - \tau^{-3}e^T e.$$

The off-diagonal elements of the Hessian Matrix are:

$$\frac{\partial^2 LL}{\partial \beta \partial \tau} = \frac{\partial^2 LL}{\partial \tau \partial \beta} = \frac{1}{\tau^2} (-X^T Y + X^T X\beta).$$

Thus, the Hessian Matrix is:

$$\begin{bmatrix} -\frac{1}{\tau}(X^T X) & \frac{1}{\tau^2} (-X^T Y + X^T X\beta) \\ \frac{1}{\tau^2} (-X^T Y + X^T X\beta) & \frac{n}{2}\tau^{-2} - \tau^{-3}e^T e \end{bmatrix}.$$

We now need to take the negative expectation of this matrix to obtain the information matrix. The expectation of β is β , and so, if we substitute the computation of $\beta (= (X^T X)^{-1}(X^T Y))$, the numerator of the off-diagonal elements is 0.

The expectation of the second partial derivative with respect to β remains unchanged. However, the second partial derivative with respect to τ changes a little. First, the expectation of $e^T e$ is $n\tau$, just as we discussed while deriving the ML estimators for the simple regression model. This gives us:

$$\frac{n}{2\tau^2} - \frac{n\tau}{\tau^3}.$$

Simple algebraic manipulation gives us:

$$-\frac{n}{2\tau^2}.$$

Thus, after taking the negative of these elements, our information matrix is:

$$\begin{bmatrix} \frac{1}{\tau}(X^T X) & 0 \\ 0 & \frac{n}{2\tau^2} \end{bmatrix}.$$

As we've discussed before, we need to invert this matrix and square root the diagonal elements to obtain the standard errors of the parameters. Also as we've discussed before, the inverse of a diagonal matrix is simply a matrix with the diagonal elements inverted. Thus, our variance-covariance matrix is:

$$\begin{bmatrix} \tau(X^T X)^{-1} & 0 \\ 0 & \frac{2\tau^2}{n} \end{bmatrix},$$

which should look familiar after substituting σ_e^2 in for τ .