

## Simple Regression II: Model Fit and Inference (Soc 504)

In the previous notes, we derived the least squares and maximum likelihood estimators for the parameters in the simple regression model. Estimates of the parameters themselves, however, are only useful insofar as we can determine a) how well the model fits the data and b) how good our estimates are, in terms of the information they provide about the possible relationship between variables in the population. In this set of notes, I discuss how to interpret the parameter estimates in the model, how to evaluate the fit of the model, and how to make inference to the population using the sample estimates.

### 1 Interpretation of Coefficients/Parameters

The interpretation of the coefficients from a linear regression model is fairly straightforward. The estimated intercept ( $b_0$ ) tells us the value of  $y$  that is expected when  $x = 0$ . This is often not very useful, because many of our variables don't have true 0 values (or at least not relevant ones-like education or income, which rarely if ever have 0 values). The slope ( $b_1$ ) is more important, because it tells us the relationship between  $x$  and  $y$ . It is interpreted as the expected change in  $y$  for a one-unit change in  $x$ . In the deathrates example, the intercept estimate was  $-9.45$ , and the slope estimate was  $.084$ . This means that the log of the death rates is expected to be  $-9.45$  at age 0 and is expected to increase  $.084$  units for each year of age.

The other parameter in the model, the standard error of the regression ( $s_e$ ), is important because it gives us (in  $y$  units) an estimate of the average error associated with a predicted score. In the model for deathrates, the standard error of the regression was  $.576$ .

### 2 Evaluating Model Fit

Before we make inferences about parameters from our sample estimates, we would like to decide just how well the model fits the data. A first approach to this is to determine the amount of the variance in the dependent variable is 'accounted for' by the regression line. If we think of the formula for variance:

$$\sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1},$$

we can consider the numerator to be called the "Total Sum of Squares" (TSS). We have an estimate of the variance that is unexplained by the model-the "Residual Sum of Squares" (SSE), which is just the numerator of the standard error of the regression function:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The difference between these two is the “Regression Sum of Squares,” (RSS), or the amount of variance accounted for by the model. RSS can be represented as:

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

Some basic algebra reveals that, in fact,  $RSS + SSE = TSS$ . A measure of model fit can be constructed by taking the ratio of the explained variance to the total variance:

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

This measure ranges from 0 to 1. For a poor-fitting model, the error (SSE) will be large (possibly equal to the TSS), making RSS small. For example, if we were to allow the mean of  $y$  to be our best guess for  $y$  ( $\hat{Y} = \bar{Y}$ ), then we would be supposing that the relationship between  $x$  and  $y$  did not exist. In this case, RSS would be 0, as would  $R^2$ . For a perfect model, on the other hand,  $RSS = TSS$ , so  $R^2 = 1$ . In the deathrates example, the  $R^2$  was .92, indicating a good linear fit.

Interestingly, in the simple linear regression model the signed square root of  $R^2$  is the correlation between  $x$  and  $y$  (sign based on sign of  $b_1$ ). In multiple regression, This computation yields the ‘multiple correlation,’ which we will discuss later.

These results are almost all that is needed to complete an ANOVA table for the regression model. The typical model ANOVA output looks like (this is from my death rate example):

ANOVA TABLE	Df	SS	MS	F	Sig
Regression	1	315.54	315.54	949.78	$p = 0$
Residual	79	26.25	.33223		
Total	80	341.79			

The general computation of the table is:

ANOVA TABLE	DF	SS	MS	F	Sig
Regression	k-1	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$\frac{RSS}{Df(R)}$	$\frac{MSR}{MSE}$	(from F table)
Residual	n-k	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\frac{SSE}{Df(E)}$ (called MSE)		
Total	n-1	$\sum_{i=1}^n (Y_i - \bar{Y})^2$			

where  $k$  is the number of parameters, and  $n$  is the number of data points (sample size).

For the  $F$  test, the numerator and denominator degrees of freedom are  $k - 1$  and  $n - k$ , respectively. The  $F$  test is a joint test that all the coefficients in the model are 0. In simple linear regression,  $F = t^2$  (for  $b_1$ ).

There are numerous ways to further assess model fit, specifically by examining the error terms. We will discuss these later, in the context of multiple regression.

### 3 Inference

Just as the mean in the ‘models’ that you discussed in previous classes (and that we discussed earlier) have a sampling distribution from which you can assume your estimate of  $\bar{x}$  is a sampled value, regression coefficients also have a sampling distribution. For these sampling distributions to be general, there are several assumptions that the OLS (or MLE) regression model must meet. They include:

1. Linearity. This assumption says that the relationship between  $x$  and  $y$  must be linear. If this assumption is not met, then the parameters will be biased.
2. Homoscedasticity (constant variance of  $y$  across  $x$ ). This assumption says that the variance of the error term cannot change across values of  $x$ . If this assumption is violated, then the standard errors for the parameters will be biased. Note that if the linearity assumption is violated, the homoscedasticity assumption need not be.
3. Independence of  $e_i$  and  $e_j$ , for  $i \neq j$ . This assumption says that there cannot be a relationship between the errors for any two individuals. If this assumption is violated, then the likelihood function does not hold, because the probability multiplication rule says that joint probabilities are multiple of the individual probabilities ONLY if the events are independent.
4.  $e_i \sim N(0, \sigma_e)$ . Because the structural part of  $y$  is fixed, given  $x$ , this is equivalent to saying that  $y_i \sim N(b_0 + b_1x_i, \sigma_e)$ . This assumption is simply that the errors are normally distributed. If this assumption is not met, then the likelihood function is not appropriate. In terms of least squares, this assumption guarantees that the sampling distributions for the parameters are also normal. However, as  $n$  gets large, this assumption becomes less important, by the CLT.

If all of these assumptions are met, then the OLS estimates/MLE estimates are BLUE (Best Linear Unbiased Estimates), and the sampling distributions for the parameters can be derived. They are:

$$b_0 \sim N\left(\beta_0, \frac{\sigma_e^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2}\right)$$

$$b_1 \sim N\left(\beta_1, \frac{\sigma_e^2}{\sum (X_i - \bar{X})^2}\right)$$

With the standard errors obtained (square root of the sampling distribution variance estimators above), we can construct  $t$ -tests on the parameters just as we conducted  $t$ -tests

on the mean in previous courses. Note that these are  $t$ -tests, because we must replace  $\sigma_{\epsilon}$  with our estimate,  $s_e$ , for reasons identical to those we discussed before. The formulas for the standard errors are often expressed as (where  $(MSE)^{\frac{1}{2}}$  and  $s_e$  are identical):

$$S.E.(\hat{b}_0) = \sqrt{\left(\frac{MSE \sum X_i^2}{n \sum (X_i - \bar{X})^2}\right)}$$

$$S.E.(\hat{b}_1) = \sqrt{\left(\frac{MSE}{n \sum (X_i - \bar{X})^2}\right)}$$

Generally, we are interested in whether  $x$  is, in fact, related to  $y$ . If  $x$  is not related to  $y$ , then  $b_1$  will equal 0, and our best guess for the value of  $y$  is  $\bar{y}$ . Thus, we typically hypothesize that  $b_1 = 0$  and construct the following  $t$ -test:

$$t = \frac{b_1 - (H_0 : \beta_0 = 0)}{s.e.(b_1)}$$

In the deathrate example, the standard error for  $b_0$  was .189, and the standard error for  $b_1$  was .0027. Computing the  $t$ -tests yields values of  $-49.92$  and  $30.82$ , respectively. Both of these are large enough that we can reject the null hypothesis that the population parameters,  $\beta_0$  and  $\beta_1$ , are 0. In other words, if the null hypothesis were true, these data would be almost impossible to obtain in a random sample. Thus, we reject the null hypothesis.

As we discussed before, constructing a confidence interval for this estimate is a simple algebraic contortion of the  $t$ -test in which we decide *a priori* the ‘confidence level’ we desire for our interval:

$$(1 - \alpha)\%C.I. = b_1 \pm t_{\frac{\alpha}{2}} s.e.(b_1)$$

These tests work also for the intercept, although, as we discussed before, the intercept is often not of interest. However, there are some cases in which the intercept may be of interest. For example, if we are attempting to determine whether one variable is an unbiased measure of another, then the intercept should be 0. If there is a bias, then the intercept should pick this up, and the intercept will not be 0. For example, if people tend to under-report their weight by 5 pounds, then the regression model for observed regressed on reported weight will have an intercept of 5 (implying you can take an individual’s reported weight and add 5 pounds to it to get their actual weight-see Fox).

Sometimes, beyond confidence intervals for the coefficients themselves, we would like to have prediction intervals for an unobserved  $y$ . Because the regression model posits that  $y$  is a function of both  $b_0$  and  $b_1$ , a prediction interval for  $y$  must take variability in the estimates of both parameter estimates into account. There are various ways to do such calculations, depending on exactly what quantity you are interested in (but these are beyond the scope of this material).

With the above results, you can pretty much complete an entire bivariate regression analysis.

## 4 Deriving the Standard Errors for Simple Regression

The process of deriving the standard errors for the parameters in the simple regression problem using maximum likelihood estimation involves the same steps as we used in finding the standard errors in the normal mean/standard deviation problem before:

1. Construct the second derivative matrix of the log likelihood function with respect to each of the parameters. That is, in this problem, we have three parameters, and hence the Hessian matrix will contain 6 unique elements (I've replaced  $s_e^2$  with  $\tau$ :

$$\begin{bmatrix} \frac{\partial^2 LL}{\partial b_0^2} & \frac{\partial^2 LL}{\partial b_0 \partial b_1} & \frac{\partial^2 LL}{\partial b_0 \partial \tau} \\ \frac{\partial^2 LL}{\partial b_1 \partial b_0} & \frac{\partial^2 LL}{\partial b_1^2} & \frac{\partial^2 LL}{\partial b_1 \partial \tau} \\ \frac{\partial^2 LL}{\partial \tau \partial b_0} & \frac{\partial^2 LL}{\partial \tau \partial b_1} & \frac{\partial^2 LL}{\partial \tau^2} \end{bmatrix}$$

2. Take the negative expectation of this matrix to obtain the Information matrix.
3. Invert the information matrix to obtain the variance-covariance matrix of the parameters. For the actual standard error estimates, we substitute the MLEs for the population parameters.

### 4.1 Deriving the Hessian Matrix

We can obtain the elements of the Hessian matrix using the first derivatives shown above. Using the first derivative of the log likelihood with respect to  $b_0$ , we can take the second derivative with respect to  $b_0$ . The first derivative with respect to  $b_0$  was:

$$\frac{\partial LL}{\partial b_0} = -\frac{1}{2\tau} \left( 2nb_0 - 2 \sum Y + 2b_1 \sum X \right).$$

The second derivative with respect to  $b_0$  is thus simply:

$$\frac{\partial^2 LL}{\partial b_0^2} = -\frac{1}{2\tau} (2n) = \frac{-n}{\tau}.$$

The second derivative with respect to  $b_1$  is:

$$\frac{\partial^2 LL}{\partial b_0 \partial b_1} = -\frac{1}{2\tau} \left( 2 \sum X \right) = \frac{\sum X}{\tau}$$

The second derivative with respect to  $\tau$  is:

$$\frac{\partial^2 LL}{\partial b_0 \partial \tau} = -\frac{1}{\tau^2} \left( nb_0 - \sum Y + b_1 \sum X \right)$$

The three remaining second partial derivatives require the other two first partial derivatives. Using the first partial derivative with respect to  $b_1$ :

$$\frac{\partial LL}{\partial b_1} = -\frac{1}{2\tau} \left( -2 \sum XY + 2b_0 \sum X + 2b_1 \sum X^2 \right),$$

we can easily take the second partial derivative with respect to  $b_1$ :

$$\frac{\partial LL}{\partial b_1^2} = -\frac{1}{2\tau} \left( 2 \sum X^2 \right) = \frac{-\sum X^2}{\tau}$$

We can also obtain the second partial derivative with respect to  $\tau$ :

$$\frac{\partial^2 LL}{\partial b_1 \partial \tau} = \tau^{-2} \left( -\sum XY + b_0 \sum X + b_1 \sum X^2 \right)$$

For finding the second partial derivative with respect to  $s_e^2$ , we need to re-derive the first derivative with respect to  $s_e^2$  rather than  $s_e$ . So, letting  $\tau = s_e^2$  as we did above, we get the following:

$$\frac{\partial LL}{\partial \tau} = -\frac{n}{2\tau} + \frac{1}{2\tau} \sum (Y - (b_0 + b_1 X))^2.$$

The second partial derivative, then, is:

$$\frac{\partial^2 LL}{\partial \tau^2} = \frac{1}{2} \left[ n\tau^{-2} - 2\tau^{-3} \sum (Y - (b_0 + b_1 X))^2 \right].$$

We now have all 6 second partial derivatives, giving us the following Hessian matrix (after a little rearranging of the terms):

$$\begin{bmatrix} \frac{-n}{\tau} & \frac{\sum X}{\tau} & -\frac{(nb_0 - \sum Y + b_1 \sum X)}{\tau^2} \\ \frac{\sum X}{\tau} & \frac{-\sum X^2}{\tau} & \frac{(\sum XY - b_0 \sum X - b_1 \sum X^2)}{\tau^2} \\ -\frac{(nb_0 - \sum Y + b_1 \sum X)}{\tau^2} & -\frac{(\sum XY - b_0 \sum X - b_1 \sum X^2)}{\tau^2} & \frac{n}{2\tau^2} - \frac{\sum (Y - (b_0 - b_1 X))^2}{\tau^3} \end{bmatrix}$$

## 4.2 Computing the Information Matrix

In order to obtain the information matrix, we need to negate the expectation of the Hessian matrix. There are a few ‘tricks’ involved in this process. The negative expectation of the first element is simply  $\frac{n}{\tau}$ . The negative expectation of the second element (and also the fourth, given the symmetry of the matrix) is  $\frac{n\mu}{\tau}$  (recall the trick we used earlier—that if  $\bar{X} = \frac{\sum X}{n}$ , then  $\sum X = n\bar{X}$ ). We will skip the third and sixth (and hence also the seventh and eighth) elements for the moment. The negative of the expectation of the fifth element is  $\frac{\sum X^2}{\tau}$ . (We leave it unchanged, given that there is no simple way to reduce this quantity). Finally, to take the negative of the expectation of the ninth and last element, we must first note that the expression can be rewritten as:  $\frac{n\tau - \sum (Y - (b_0 + b_1 X))^2}{2\tau^3}$ . The expectation of the sum, though, is nothing more than the error variance itself,  $\tau$ , taken  $n$  times. Thus, the expectation of the numerator is  $n\tau - 2n\tau$ . After some simplification, including some cancelling with terms in the denominator, we are left with  $\frac{n}{2\tau^2}$  for the negative expectation.

All the other elements in this matrix (3, 6, 7, and 8) go to 0 in expectation. Let's take the case of the third (and seventh) element. The expectation of  $b_0$  is  $\beta_0$ , which is equal to  $\mu_Y - \beta_1\mu_X$ . The expectation of  $\sum Y$  is  $n\mu_Y$ , and the expectation of  $\sum X$  is  $n\mu_X$ . Finally, the expectation of  $b_1$  is  $\beta_1$ . Substituting these expressions into the numerator yields:

$$n(\mu_Y - \beta_1\mu_X) - n\mu_Y + n\beta_1\mu_X.$$

This term clearly sums to 0, and hence the entire third and seventh expressions are 0. A similar finding results for elements six and eight.

### 4.3 Inverting the Information Matrix

Our information matrix obtained in the previous section turns out to be:

$$\begin{bmatrix} \frac{n}{\tau} & \frac{n\mu_X}{\tau} & 0 \\ \frac{n\mu_X}{\tau} & \frac{\sum X^2}{\tau} & 0 \\ 0 & 0 & \frac{n}{2\tau^2} \end{bmatrix}$$

Although inverting  $3 \times 3$  matrices is generally not easy, the 0 elements in this one make the problem relatively simple. We can break this problem into parts. First, recall that the multiple of a matrix with its inverse produces an identity matrix. So, in this case, the sum of the multiple of the elements of the last row of the information matrix by the elements of the last column of the inverse of the information matrix must be 1. But, all these elements are 0:

$$[0 \ 0 \ I(\theta)_{33}^{-1}] \begin{bmatrix} 0 \\ 0 \\ \frac{n}{2\tau^2} \end{bmatrix} = 1.$$

The only way for this to occur is for  $I(\theta)_{33}^{-1}$  to be  $\frac{2\tau^2}{n}$ .

If we do a little more thinking about this problem, we will see that, because of the 0 elements, to invert the rest of the matrix, we can treat the remaining non-zero elements of the matrix as a  $2 \times 2$  sub-matrix, with the elements that are 0 in the information matrix also being 0 in the variance-covariance matrix. The inverse of a  $2 \times 2$  matrix  $\mathbf{M}$  can be found using the following rule:

$$M^{-1} \equiv \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \frac{1}{|M|} \begin{bmatrix} D & -B \\ -C & A \end{bmatrix}.$$

You can derive this result algebraically for yourself by setting up a system of four equations in four unknowns.

The determinant of the submatrix here is  $\frac{n \sum X^2 - n^2 \mu_X^2}{\tau^2}$ , so, after inverting (the constant in front of the inverse formula above is  $\frac{1}{|M|}$ ), we obtain  $\frac{\tau^2}{n \sum X^2 - n^2 \mu_X^2}$ . This expression can be simplified by recognizing that an  $n$  can be factored from the numerator, leaving us with

$n$  multiplied by numerator for the so-called computation formula for the variance of  $X$  ( $= \sum (X - \mu_X)^2$ ). Thus, we have  $\frac{\tau^2}{n \sum (X - \mu_X)^2}$ . The inverse matrix thus becomes:

$$\begin{bmatrix} \frac{\tau \sum X^2}{n \sum (X - \mu_X)^2} & \frac{-\tau \mu_X}{\sum (X - \mu_X)^2} \\ \frac{-\tau \mu_X}{\sum (X - \mu_X)^2} & \frac{\tau}{\sum (X - \mu_X)^2} \end{bmatrix}$$

We now have all the elements of the variance-covariance matrix of the parameters. These terms should look familiar, after replacing  $\tau$  with  $\sigma_\epsilon^2$  and the population-level parameters with the sample ML estimates:

$$\begin{bmatrix} \frac{s_e^2 \sum X^2}{n \sum (X - \bar{X})^2} & \frac{-s_e^2 \bar{X}}{\sum (X - \bar{X})^2} \\ \frac{-s_e^2 \bar{X}}{\sum (X - \bar{X})^2} & \frac{s_e^2}{\sum (X - \bar{X})^2} \end{bmatrix}$$