Stat 135, Fall 2006

Inference in the simple linear regression model

A. Adhikari

The statistical model for simple linear regression is

$$Y_i \;=\; \beta_0 \;+\; \beta_1 x_i \;+\; \epsilon_i, \qquad i = 1, 2, \ldots, n$$

Here n is the number of observations, and

- $x_1, x_2, \ldots, x_n$ are known constants.
- β_0 and β_1 are unknown constants (parameters of the model).
- $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are i.i.d. random errors. Their common distribution is normal with mean 0 and variance σ^2. The common variance σ^2 is an unknown constant and is a parameter of the model.

Implications of the Model.

1. For each i, Y_i is normal with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. The Y_i's are independent of each other.

2. $\bar{Y}$ is normal with mean $\beta_0 + \beta_1 \bar{x}$ and variance σ^2/n.

3. Estimates of the coefficients. The least-squares estimate of β_1 is the slope of the regression line, derived last time:

$$\hat{\beta}_1 \;=\; \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \;=\; \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The least-squares estimate of β_0 is the intercept of the regression line, also derived last time:

$$\hat{\beta}_0 \;=\; \bar{Y} \;-\; \hat{\beta}_1 \bar{x}$$

You can check that these are also the MLEs.

4. Distributions of the estimates. Since both $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear combinations of the Y_i's, they are both normally distributed.

5. Means of the estimates. Both the estimates are unbiased. Reason:

$$E(\hat{\beta}_1) \;=\; \frac{\sum_{i=1}^n (x_i - \bar{x})E(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \;=\; \frac{\sum_{i=1}^n (x_i - \bar{x})\beta_1(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \;=\; \beta_1$$

$$E(\hat{\beta}_0) \;=\; E(\bar{Y}) \;-\; E(\hat{\beta}_1)\bar{x} \;=\; \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} \;=\; \beta_0$$

6. Variances of the estimates. Use the second form of the expression for $\hat{\beta}_1$ to see that

$$Var(\hat{\beta}_1) \;=\; \frac{\sum_{i=1}^n (x_i - \bar{x})^2\sigma^2}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} \;=\; \sigma^2\Big[\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big]$$

Check that $Cov(Y_i, \bar{Y}) = \sigma^2/n$ for each i. Then use the second form of the expression for $\hat{\beta}_1$ again to observe that $\bar{Y}$ and $\hat{\beta}_1$ are uncorrelated:

$$Cov(\hat{\beta}_1, \bar{Y}) \quad = \quad \frac{\sum_{i=1}^n (x_i - \bar{x}) Cov(Y_i, \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad = \quad \frac{(\sigma^2/n) \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad = \quad 0$$

(Because they are jointly normal, they are in fact independent.) Now

$$Var(\hat{\beta}_0) \quad = \quad Var(\bar{Y}) \ + \ Var(\hat{\beta}_1)\bar{x}^2 \quad = \quad \sigma^2 \Big[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big]$$

You should check that

$$Cov(\hat{\beta}_0, \hat{\beta}_1) \quad = \quad -\,\sigma^2 \Big[\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big]$$

6. Estimate of the "mean value" at x_0, that is, the height of the true line at $x = x_0$. You are given some value x_0 of the variable x, and your task is to estimate the height of the true line at this value. The parameter you must estimate is $\beta_0 + \beta_1 x_0$. Your estimate will be the height of the regression line

$$M \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad = \quad \bar{Y} + \hat{\beta}_1 (x_0 - \bar{x})$$

The distribution of M is normal, with mean

$$E(M) = \beta_0 + \beta_1 x_0$$

and variance (obtained using the second form of the expression for M and the fact that $\bar{Y}$ and $\hat{\beta}_1$ are uncorrelated)

$$Var(M) \quad = \quad \frac{\sigma^2}{n} + \sigma^2 \Big[\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big] (x_0 - \bar{x})^2 \quad = \quad \sigma^2 \Big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big]$$

Summary

parameter	**estimate**	**mean**	**variance**
true intercept β_0	reg. line intercept $\hat{\beta}_0$	β_0	$\sigma^2 \Big[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big]$
true slope β_1	reg. line slope $\hat{\beta}_1$	β_1	$\sigma^2 \Big[\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big]$
true height at x_0 $\beta_0 + \beta_1 x_0$	reg. line height at x_0 $\hat{\beta}_0 + \hat{\beta}_1 x_0$	$\beta_0 + \beta_1 x_0$	$\sigma^2 \Big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big]$

7. Predicting a new observation at $x = x_0$. Suppose you are going to make a new observation (not one of the original n) at $x = x_0$. According to the model, the new observation will be the observed value of the random variable

$$Y_{new} \quad = \quad \beta_0 \; + \; \beta_1 x_0 \; + \; \epsilon$$

where ϵ is a normal $(0, \sigma^2)$ error independent of all the original n errors.

The value Y_{new} has two parts: the height of the true line at x_0, and the random error ϵ. You can estimate the height of the true line using M above. Because $E(\epsilon) = 0$, you can use the value of M as your prediction of Y_{new}:

$$\text{predicted value of } Y_{new} \text{ is } M = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

But now there are are two independent sources of error: the error in estimating the height of the true line, and the error in using 0 as a prediction for ϵ. Thus the variance of the prediction is

$$Var(M) + Var(\epsilon) \quad = \quad \sigma^2 \Big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \Big] \; + \; \sigma^2 \quad = \quad \sigma^2 \Big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + 1 \Big]$$

8. Estimate of σ^2. Define the ith residual to be $\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$. For every i the residual is normally distributed with mean 0. Of course the residuals are not independent of each other (their sum is 0). But here is a fact I will not prove:

Let the residual sum of squares (RSS) be $\sum_{i=1}^n \hat{\epsilon}_i^2$. Then RSS/σ^2 has the chi-squared distribution with $n - 2$ degrees of freedom. So

$$s^2 \quad = \quad \frac{RSS}{n-2} \quad = \quad \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n-2}$$

is an unbiased estimate of σ^2. Use s as an estimate of σ.

9. The t statistics. You can use 8 above and check the appropriate independence to show for example that

$$\frac{\hat{\beta}_1 - \beta_1}{s\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

has the t distribution with $n-2$ degrees of freedom. This can be used to test the hypothesis $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$, for example. Equivalently, it can be used to construct confidence intervals for β_1. For example, a 95% confidence interval for β_1 is

$$\hat{\beta}_1 \; \pm \; t^* s \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

where t^* is the 97.5th percentile of the t distribution with $n - 2$ degrees of freedom.

Inference for the other parameters can be performed in a similar manner. Just replace σ by s, and the normal distribution by the t distribution with $n-2$ degrees of freedom.

As always, if the sample sizes are large you can replace the t distribution by the normal.

10. Comparing two regressions. Sometimes it is useful to be able to compare two regression lines and ask if they are in fact estimates for the same true line. A more common question is to ask whether the slopes of two regression lines are in fact estimates of the same slope. For example, if your population consists of men and women, you may have fitted a line for the men and, independently, a line for the women. You may want to know whether the slope for the men is the same as the slope for the women. Or you may want to estimate the difference between the slopes.

Say the two true slopes are β_1 and γ_1, with estimates $\hat{\beta}_1$ and $\hat{\gamma}_1$ respectively. You get the estimates by performing separate regressions on the two datasets. The distribution of $\hat{\beta}_1 - \hat{\gamma}_1$ is normal with mean $\beta_1 - \gamma_1$ and variance equal to the sum of the variances of the two estimates.

The variance of the difference involves the values of σ_1^2 and σ_2^2, the error variances of the two models. **Assuming that the error variance σ^2 is the same for both models,** an unbiased estimate of the common σ^2 is provided by the pooled estimate

$$s_p^2 = \frac{RSS_1 + RSS_2}{(n-2)+(m-2)}$$

where RSS_1 and RSS_2 are the residual sums of squares for the two regressions, and n and m are the two sample sizes. Inference for the difference $\beta_1 - \gamma_1$ can be performed by replacing σ by s_p, and the normal distribution by the t distribution with $n+m-4$ degrees of freedom.

If the two sample sizes are large the assumption of equal variances (and hence the pooling) is not so important. You can think of $\hat{\beta}_1$ and $\hat{\gamma}_1$ as independent normal variables, and estimate the variance of the difference by simply adding to two variances.