STAT2008/6038

Random and fixed predictors

Random Predictors

In all of the above discussion, we have assumed that the values of the predictor variable, x_i, have been fixed.

Suppose that we now assume that the pairs (X_i, Y_i) are randomly sampled from a population in such a way that they have a *bivariate normal distribution* with means and variances:

$$E(X) = \mu_x; \quad E(Y) = \mu_y; \quad Var(X) = \sigma_x^2; \quad Var(Y) = \sigma_y^2; \quad Cov(X, Y) = \rho\sigma_x\sigma_y.$$

Random Predictors

It can be shown that the conditional distribution of Y given X is also normal with:

$$E(Y|X) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x) = \beta_0 + \beta_1 x; \qquad Var(Y|X) = \sigma_y^2(1 - \rho^2) = \sigma^2.$$

So, we can see that the values of Y and X are linearly related with

$$\beta_0 = \mu_y - \beta_1\mu_x$$

$$\beta_1 = \rho\frac{\sigma_y}{\sigma_x},$$

and the expression for the variance shows that the conditional variability of Y does not depend on X, so that the data will be homoscedastic.

Random Predictors

In this situation, interest generally centers on the parameter ρ, though the above identities show that this is in some sense equivalent to interest in β_1. Further, we can see that

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_y^2} = \beta_1^2\frac{\sigma_x^2}{\sigma_y^2}.$$

Random Predictors

A comparison of these formulas with those for the fixed predictor case shows the strong similarity between the two situations.

Since ρ is a correlation, it must lie between -1 and 1, and can only equal these values if $\sigma^2 = 0$, i.e., if their is perfect linear association between the random variables X and Y.

So ρ measures the degree of linear association between the two random variables.

Estimation

Estimation of the parameters proceeds are calculated using the same formulas for the fixed predictor case.

Similarly ρ, which is r, the sample correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = b_1\sqrt{\frac{S_{xx}}{S_{yy}}},$$

where $S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = SST$.

Equivalence of the test statistics

typically interest centers on testing the hypotheses

$$H_0 : \rho = 0 \quad \text{versus} \quad H_A : \rho \neq 0$$

which amounts to testing whether there is any linear association between the two variables. It turns out that, if H_0 is true, then the test statistic

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

has a Student's t-distribution with $n-2$ degrees of freedom, and thus we can test the null hypothesis by comparing the observed value of T to the appropriate t-quantiles, $t_{n-2}(1-\alpha/2)$.

7

Equivalence of the test statistics

- We have noted previously that all tests of the significance of a regression are actually identical!

8

Equivalence of the test statistics

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{\left(b_1\sqrt{\frac{S_{xx}}{S_{yy}}}\right)\sqrt{n-2}}{\sqrt{1-\frac{SSR}{SST}}}$$

$$= \frac{b_1\sqrt{S_{xx}}}{\sqrt{\frac{SST-SSR}{n-2}}} = \frac{b_1\sqrt{S_{xx}}}{\sqrt{\frac{SSE}{n-2}}} = \frac{b_1\sqrt{S_{xx}}}{\sqrt{MSE}} = \frac{b_1\sqrt{S_{xx}}}{s} = \frac{b_1}{s(b_1)}.$$

9

Equivalence of the test statistics

Thus, this test is the same as the t-test and the F-test for the null hypothesis that $\beta_1 = 0$.

($\beta_1 = 0$ if and only if $\rho = 0$.)

10

Fixed Predictors

Even in the case of random predictors, we focus on the conditional distribution of the response given the predictor values.

It is the variation in the response that we are trying to explain.

The way in which the values of the predictor variables are arrived at is then secondary information!

Therefore we will always make the assumption that the x_i's are fixed.

11