Ordinary Least Squares Regression.

Simple Regression. Algebra and Assumptions.

In this part of the course we are going to study a technique for analysing the linear relationship between two variables Y and X. We have n pairs of observations (Y_i X_i), i = 1, 2, ..,n on the relationship which, because it is not exact, we shall write as:

$$y_i = \alpha + \beta x_i + \in_i \quad i = 1,....,n$$

In this relationship α, β and ε_i i = 1,.., n are fundamentally unobservable and we would like to estimate α and β.

Approach:

The idea is to select estimates of α and β lets call them α* and β* which yield a straight line (called the regression line) Y = α* + β*X which minimises a measure of the aggregate distance of the points (Y_i X_i), i = 1, 2, ..,n to that line in X Y space, where Y is measured on the vertical axis. The measure we use is the sum of squared vertical distances which we shall call the Error Sum of Squares (ERSS) so that α* and β* are solutions to the problem:

$$\min_{\alpha^*,\beta^*} ERSS\left(= \sum_{i=1}^{n}\left(Y_t - \left(\alpha^* + \beta^* X_i\right)\right)^2 \right)$$

The multivariate calculus is employed to solve this problem by setting the partial derivatives of ERSS with respect to α* and β* to zero (called the first order conditions) and solving thus:

$$\frac{\partial ERSS}{\partial \alpha^*} = -2\sum_{i=1}^{n}\left(Y_i - \left(\alpha^* + \beta^* X_i\right)\right) = 0$$

$$\frac{\partial ERSS}{\partial \beta^*} = -2\sum_{i=1}^{n}\left(Y_i - \left(\alpha^* + \beta^* X_i\right)\right)X_i = 0$$

The solutions to which are:

$$\alpha^* = \overline{Y} - \beta^* \overline{X}$$

$$\beta^* = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)X_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)X_i}$$

Note the formula for β^* has many alternative equivalent versions which may be seen by observing that for the numerator:

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)X_i = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right) = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)Y_i = \sum_{i=1}^{n} X_i Y_i - n\overline{Y}\,\overline{X}$$

and for the denominator:

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)X_i = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right) = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

so that various combinations of numerator and denominator formulae yield 12 alternative representations of β^*.

Assumptions in the Ordinary Least Squares model.

Note that while α, β and ε_i , i = 1,.., n are fundamentally unobservable we only concern ourselves with estimating α and β which define the relationship between Y and X. The ε_i i = 1,.., n are considered "errors" which accommodate all the other influences on Y not accounted for in α +βX as such we assume them to be random and to obey the following four assumptions:

1) $E(\varepsilon_i) = 0$ for all i.

This assumption really says that the average or net effect of all the other influences on Y not accounted for in α +βX is constant and zero for each observation (i=1,..,n).

2) $V(\varepsilon_i) = \sigma^2 > 0$ for all i.

This assumption really says that the variability of the net effect of all the other influences on Y not accounted for in α +βX is constant and non zero for each observation (i=1,..,n).

This is sometimes referred to as the homoskedasticity assumption and is not as innocuous as it seems. For example if Y were the consumption behaviour of an individual and X was their disposable income it says that the scale of variability of the unobserved effects would be the same for both rich and poor individuals.

3) $E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$.
4) $E(\varepsilon_i X_j) = 0$. For i and j

These last two assumptions derive from the general assumption that the net effects of all the other influences on Y not accounted for in $\alpha + \beta X$ are independent of the X's and of each other. Again these are not innocuous assumptions, for example if the equation relates to the behaviour of individuals and individuals in the sample are related to one another in some fashion they are unlikely to be true. Similarly if the equation relates to behaviour through time today's random effects are unlikely to be independent of yesterdays effects.

Given these assumptions certain properties of the estimators follow.

Unbiasedness.

Under the above assumptions the ordinary least squares estimators α^* and β^* are unbiased so that $E(\alpha^*) = \alpha$ and $E(\beta^*) = \beta$ which may be demonstrated as follows. From the various formulae for β^* we may write:

$$\beta^* = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)Y_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)X_i} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\alpha + \beta X_i + \in_i\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)X_i}$$

which, after noting that the sum of deviations from mean is equal to 0, may be written as:

$$\beta^* = \beta + \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\in_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)X_i}$$

which can readily be seen to have an expectation β. For α* note that:

$$\alpha^* = \overline{Y} + \beta^* \overline{X} = \frac{1}{n}\sum_{i=1}^{n}\left(\alpha + \beta X_i + \in_i\right) - \beta^* \overline{X} = \alpha + \left(\beta - \beta^*\right)\overline{X} + \overline{\in}$$

since E(β - β*) = 0 and given assumption 1 E(α*) = α.

The variances of the estimators.

The variance of the estimators α* and β* facilitate inference and confidence interval estimation. The can be derived, given the above assumptions as follows. The variance of any random variable W is given by $E[(W-E(W))^2]$ so that for β* from the above we may write:

$$E\left[\left(\beta^* - E\left(\beta^*\right)\right)^2\right] = E\left[\left(\beta + \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\in_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} - \beta\right)^2\right] = E\left[\left(\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\in_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)^2\right]$$

Note from our assumptions that:

$$E\left(X_i - \overline{X}\right)\in_i \left(X_j - \overline{X}\right)\in_j = 0 \quad \textit{for all } i \neq j$$

$$= \sigma^2\left(X_i - \overline{X}\right)^2 \quad \textit{for all } i = j$$

which means that the variance may be written as:

$$E\left[\left(\beta^* - E\left(\beta^*\right)\right)^2\right] = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\sigma^2}{\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

By similar arguments the variance of α* is given by:

$$E\left[\left(\alpha^* - E\left(\alpha^*\right)\right)^2\right] = \sigma^2\left(\frac{1}{n} + \frac{\overline{X^2}}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)$$

The Residuals.

The things that are not "explained" by the model, the ε_i's are in fact implicitly estimated by the model by $Y_i - \alpha^* - \beta^* X_i$. These estimates of the "errors" are frequently referred to as the "Residuals". They too have properties which reflect the assumptions made regarding the true errors.

1. The residuals sum to 0.

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(Y_i - \alpha^* - \beta^* X_i \right) = \sum_{i=1}^{n} \left(Y_i - \left(\overline{Y} - \beta^* \overline{X} \right) - \beta^* X_i \right)$$

$$= \sum_{i=1}^{n} \left(Y_i - \overline{Y} \right) - \beta^* \sum_{i=1}^{n} \left(X_i - \overline{X} \right) = 0$$

This mimics the idea that the errors have an expectation 0.

2. The residuals are orthogonal to the X's.

$$\sum_{i=1}^{n} e_i X_i = \sum_{i=1}^{n} \left(Y_i - \alpha^* - \beta^* X_i \right) X_i = \sum_{i=1}^{n} \left(Y_i - \left(\overline{Y} - \beta^* \overline{X} \right) - \beta^* X_i \right) X_i$$

$$= \sum_{i=1}^{n} \left(Y_i - \overline{Y} \right) X_i - \beta^* \sum_{i=1}^{n} \left(X_i - \overline{X} \right) X_i = \sum_{i=1}^{n} \left(Y_i - \overline{Y} \right) X_i - \frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y} \right) X_i}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right) X_i} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) X_i = 0$$

Which mimics the idea that the errors are independent of the X_i 's.

Inference and Confidence Intervals.

Noting that β* may be written as:

$$\beta^* = \beta + \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right) \in_i}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right) X_i}$$

it can readily be seen that the estimator is a linear function of the errors ε_i, i = 1,..,n so that if they were normally distributed then so would β* be, in fact based upon the foregoing it is distributed as:

$$\beta^* \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)$$

in a similar fashion α* will be distributed as:

$$\alpha^* \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\overline{X^2}}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)\right)$$

so that inference and confidence interval computations can proceed accordingly.

An Estimator for σ^2.

As usual it is extremely rare for the value of the variance to be known so that generally it has to be estimated. The estimator σ^{2*} is given by problem:

$$\sigma^{2*} = \frac{ERSS}{n-2} = \frac{\sum_{i=1}^{n}\left(Y_i - \left(\alpha^* + \beta^* X_i\right)\ \right)^2\right)}{n-2}$$

which is most conveniently calculated by noting that:

$$ERSS = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \beta^{*2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

employing this estimate means that:

$$\frac{\left(\beta^* - \beta\right)}{\sqrt{\frac{\sigma^{*2}}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}} \sim t\left(n-2\right)$$

and similarly:

$$\frac{\left(\alpha^* - \alpha\right)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\overline{X^2}}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)}} \sim t\left(n-2\right)$$

so inference and confidence interval calculations can be conducted accordingly.

R^2 and all that.

A very common instrument for measuring the degree of explanatory power in an equation is the R^2 statistic which measures the proportion of the variability in the Y_i's that is attributable to the X_i's. Given the amount of variability in the Y_i's attributable to the errors is given by the error sum of squares (ERSS), the proportion of the variability of the Y_i's attributable to the X_i's is given by:

$$R^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - ERSS}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{\beta^{*2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

where the last equality is obtained by substituting in the formula for ERSS. Noting that 1-R^2 is:

$$1 - R^2 = \frac{ERSS}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

it follows that:

$$\frac{1-R^2}{R^2}\cdot(n-2) = \frac{\beta^{*2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\frac{ERSS}{n-2}} = \left(\frac{\beta^*}{st\,dev\left(\beta^*\right)}\right)^2$$

which is the square of the classic "t" statistic for β which will have an F(1,n-2) distribution.

Multiple Regression.

Here we consider the extension of the simple regression technique to analysing the linear relationship between K+1 variables Y and X_1, X_2, ..,X_K. We have n K+1-tuples of observations (Y_i X_{i1} , X_{i2}, ..,X_{iK}) i = 1, 2, ..,n on the relationship which, because it is not exact, we shall write as:

$$Y_i = \alpha + \sum_{k=1}^{K}\beta_k X_{ik} + \in_i \quad i = 1,...,n$$

In this relationship α, β_k, k = 1,..,K and ε_i i = 1,.., n are fundamentally unobservable and we would like to estimate the α and β_k. Essentially this deals with the case where Y is described by more than one variable X. We require one addition to our 4 assumptions, namely n > K+1.

In the same way that estimators for α, β were developed in the simple regression case (by minimising ERSS) estimators for α, β_k, k = 1,..,K (denoted α^*, β^*_k, k = 1,..,K) can be developed by minimizing with respect to α^*, β^*_k, k = 1,..,K the multivariate version of the error sum of squares given by:

$$ERSS = \sum_{i=1}^{n}\left(Y_i - \alpha^* + \sum_{k=1}^{K}\beta^*_k X_{ik}\right)^2$$

The formulae for these estimates and similarly formulae for Var(α^*) and Var(β^*_k), k = 1,..,K are complicated and will not be given here however these estimates are readily calculated using available software packages and just as in the simple regression case inference and confidence intervals may be pursued by noting that:

$$\frac{\left(\beta_k^* - \beta_k\right)}{\sqrt{\text{var}\left(\beta_k^*\right)}} \sim t\left(n-K-1\right)$$

and similarly:

$$\frac{\left(\alpha_k^* - \alpha_k\right)}{\sqrt{\text{var}\left(\alpha_k^*\right)}} \sim t\left(n-k-1\right)$$

so that inference and confidence interval calculations can be conducted accordingly.

R^2 in the Multivariate Case.

In the multivariate case R^2 can be calculated in exactly the same fashion as:

$$R^2 = 1 - \frac{ERSS}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

however in this case if we wish to use it in the context of a joint test of the significance of all of the X_{ik}'s we have to use:

$$\frac{R^2}{1-R^2}\cdot\frac{(n-K-1)}{K}\sim F(K,n-K-1)$$

which provides us with a one sided upper tailed test statistic for the significance of the K explanatory variables X_{ik}.