

Notes on Regression Analysis

Gary Lorden

The setup is the following (formally slightly different from Rice).

We observe

$$\begin{pmatrix} Y_1 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1r} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ x_{n1} & \cdots & x_{nr} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_r \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{pmatrix}$$

given numbers unknown regression coefficients iid "errors" $N(0, \sigma^2)$

which can be written in the simpler form

$$Y = X\beta + \epsilon.$$

We assume that $n > r$ and note that since $\epsilon_1, \dots, \epsilon_n$ have mean zero, the "true mean vector"

$$(1) \quad EY = X\beta = \beta_1 X_{[1]} + \beta_2 X_{[2]} + \cdots + \beta_r X_{[r]},$$

where $X_{[1]}, \dots, X_{[r]}$ denote the columns of X . Relation (1) shows that the true mean vector is an unknown linear combination of known vectors - that is

$$\begin{aligned} EY \in M &= \text{"column space" of } X \\ &= \text{subspace spanned by the columns of } X. \end{aligned}$$

We assume that this subspace has dimension r , the maximum possible - i.e. that the columns of X are linearly independent. Then the rank of X is r . (This is called the "full rank case.")

Estimation of the β_j 's

Maximum likelihood estimation of β_1, \dots, β_r boils down to the same sort of thing as in fitting a straight line: we maximize the likelihood by finding $\hat{\beta}_1, \dots, \hat{\beta}_r$ that minimize over all possible β_1, \dots, β_r the residual sum of squares

$$\sum_{i=1}^n [Y_i - (\beta_1 x_{i1} + \cdots + \beta_r x_{ir})]^2 = |Y - (\beta_1 X_{[1]} + \cdots + \beta_r X_{[r]})|^2$$

Letting $\hat{Y} = X\hat{\beta} = \hat{\beta}_1 X_{[1]} + \cdots + \hat{\beta}_r X_{[r]}$, we see that we need

$$|Y - \hat{Y}|^2 = \min_{\eta \in M} |Y - \eta|^2,$$

1

i.e. we need to find the vector in M that is closest to the observed vector Y . This vector, \hat{Y} , is the **projection** of Y on M , the unique $\hat{Y} \in M$ such that $Y - \hat{Y} \perp M$. Since M is spanned by the columns of X , we have $Y - \hat{Y} \perp X_{[1]}, \dots, X_{[r]}$.

It is now easy to see that

$$X^t(Y - \hat{Y}) = 0.$$

So $0 = X^t(Y - X\hat{\beta}) = X^tY - X^tX\hat{\beta}$, which implies that

$$(2) \quad X^tX\hat{\beta} = X^tY$$

Relation (2) is a matrix equation, which can be regarded also as r linear equations in the r “unknowns,” $\hat{\beta}_1, \dots, \hat{\beta}_r$. These are called “the Normal Equations,” and because we are assuming we’re in the “full rank case,” the rank of X^tX , which equals the rank of X , is r . So X^tX , a symmetric $r \times r$ matrix, has rank r and is invertible (non-singular), and therefore

$$(3) \quad \hat{\beta} = (X^tX)^{-1}X^tY$$

is the unique **least squares** solution.

By plugging the observed Y into (3), we obtain the maximum likelihood estimates of $\hat{\beta}_1, \dots, \hat{\beta}_r$.

Distribution of SSR and estimation of σ^2 .

The sum of squares of residuals is given by

$$(4) \quad SSR = |Y - X\hat{\beta}|^2 = |Y - \hat{Y}|^2 \stackrel{\text{Dist.}}{=} \sigma^2 \chi_{n-r}^2,$$

which is (probabilistically) independent of \hat{Y} and the $\hat{\beta}_j$'s

This is the key to confidence intervals and hypothesis tests about β_1, \dots, β_r .

t-statistics

Sometimes we want an interval or test for a single regression coefficient β_j , but often we are interested in making inferences about a linear combination $\sum c_j \beta_j$. (For example, the c_j 's might be values of input variables.) To use matrix algebra, write the c_j 's as a $1 \times r$ row vector: $c = (c_1, \dots, c_r)$.

Then we use $c\hat{\beta}$ as a least squares estimator of $c\beta$, and tests and confidence intervals are derived by calculating the mean and variance of $c\hat{\beta}$ - specifically

$$c\hat{\beta} = c(X^tX)^{-1}X^tY = qY \quad (\text{say})$$

and

$$\begin{aligned} E(c\hat{\beta}) &= EqY = qEY && \left(\begin{array}{l} \text{which means } (q_1, \dots, q_n) \left(\begin{array}{c} EY_1 \\ \vdots \\ EY_n \end{array} \right) \end{array} \right) \\ &= qX\beta = c(X^tX)^{-1}X^tX\beta = c\beta && (c\hat{\beta} \text{ is unbiased!}) \end{aligned}$$

and

$$\text{Var}(c\hat{\beta}) = \text{Var}(qY) = \text{Var}\Sigma q_i Y_i = \Sigma \text{Var}(q_i Y_i) = \Sigma q_i^2 \sigma^2 = qq^t \sigma^2,$$

so we calculate

$$\begin{aligned} qq^t &= c(X^t X)^{-1} X^t (c(X^t X)^{-1} X^t)^t \\ &= c(X^t X)^{-1} X^t (X(X^t X)^{-1} c^t) \quad (\text{since } (X^t X)^{-1} \text{ is symmetric}) \\ &= c(X^t X)^{-1} c^t \end{aligned}$$

Thus

$$c\hat{\beta} \stackrel{\text{Dist.}}{=} N(c\beta, \sigma^2 c(X^t X)^{-1} c^t),$$

and since

$$SSR = \left| Y - \hat{Y} \right|^2 \stackrel{\text{Dist.}}{=} \sigma^2 \chi_{n-r}^2$$

independent of $c\hat{\beta}$, we conclude that

$$(5) \quad \frac{c\hat{\beta} - c\beta}{\sqrt{c(X^t X)^{-1} c^t s}} \stackrel{\text{Dist.}}{=} \text{Student's } t \text{ with } n - r \text{ degrees of freedom,}$$

where $s = \sqrt{\frac{SSR}{n-r}}$ is an estimate of the unknown σ .

Special case: If we want to look at β_j , then we choose $c = (0, 0, \dots, \underset{j\text{th entry}}{1}, \dots, 0, 0)$ so that $c(X^t X)^{-1} c^t = d_j = j\text{th diagonal entry of } (X^t X)^{-1}$. So a $100(1 - \alpha)\%$ confidence interval for β_j is

$$(6) \quad \hat{\beta}_j \pm \sqrt{d_j} s t_{n-r, \frac{\alpha}{2}}$$

and to test $H : \beta_j = b_0$ vs. $K : \beta_j \neq b_0$ we reject H if

$$(7) \quad t = \frac{\hat{\beta}_j - b_0}{\sqrt{d_j} s} \text{ satisfies } |t| \geq t_{n-r, \frac{\alpha}{2}}.$$

The generalizations of (6) and (7) for $\Sigma c_j \beta_j$ are based on (5):

$$c\hat{\beta} \pm \sqrt{c(X^t X)^{-1} c^t} s t_{n-r, \frac{\alpha}{2}} \text{ is the confidence interval}$$

and

$$t = \frac{c\hat{\beta} - b_0}{\sqrt{c(X^t X)^{-1} c^t} s} \text{ is the } t\text{-statistic for testing } H : c\beta = b_0,$$

where b_0 is some given constant (often 0).

Prediction Intervals: Suppose we are going to make an additional observation Y_{n+1} with input values given by $c = (c_1, \dots, c_r)$, i.e. $Y_{n+1} = c_1 \beta_1 + \dots + c_r \beta_r + \epsilon_{n+1}$, where

ϵ_{n+1} is $N(0, \sigma^2)$. Then we have the “guess” that Y_{n+1} will be near $c\hat{\beta}$, and moreover we know

$$Y_{n+1} - c\hat{\beta} = (Y_{n+1} - c\beta) + (c\beta - c\hat{\beta}) = \epsilon_{n+1} + (c\beta - c\hat{\beta})$$

The two terms on the right-hand side are independent normal random variables, each having mean zero. Adding their variances, we get

$$Y_{n+1} - c\hat{\beta} = N(0, \sigma^2 + c(X^t X)^{-1}c^t \sigma^2),$$

and since $s = \sqrt{\frac{SSR}{n-r}}$ is an estimate of σ independent of $Y_{n+1} - c\hat{\beta}$, we have

$$\frac{Y_{n+1} - c\hat{\beta}}{\sqrt{1 + c(X^t X)^{-1}c^t} s} \stackrel{\text{Dist.}}{=} \text{Student's } t \text{ with } n - r \text{ degrees of freedom.}$$

Therefore, a $100(1 - \alpha)\%$ confidence **prediction interval** for Y_{n+1} is

$$c\hat{\beta} \pm \sqrt{1 + c(X^t X)^{-1}c^t} s t_{n-r, \frac{\alpha}{2}}.$$