1

STAT2008/STAT6038

Introduction to Multiple Linear Regression

Multiple Regression

Data: one dependent variable (continuous) two or more independent variables (continuous)

Example: if we are examining the effect of a food additive on the amount and quality of marketable beef produced per head of cattle, we must take into account the breed, age and other factors associated with each animal included in the study.

Model – general form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

which is estimated by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

$\hat{\beta}_0$ = estimated intercept

$\hat{\beta}_i$ = estimated partial regression coefficient

As before, use least squares method to estimate parameters, minimise the error (residual) sum of squares.

Least Squares solution

Involves solving a system of simultaneous equations after differentiating and equating partial derivatives to zero.

However, we can only find a solution if:

Number of predictors is less than number of observations - overfitting

None of the independent variables are perfectly correlated with each other - multicollinearity

Matrix Notation

5

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

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$$

Assumptions

6

Consider the values of the predictor variables, x_{ij}, to be fixed.

We assume that the random error vector, ϵ, satisfies:

$$E(\epsilon) = 0 \quad \text{and} \quad Var(\epsilon) = \sigma^2 I.$$

We assume uncorrelated (independence) and homoscedastic (constant variance) errors.

We will generally assume that the ϵ_i's are normally distributed.

Assumptions

7

We assume that the underlying true relationship between the response and the predictors is a linear one.

By this we mean "linear in the parameters".

As we saw in simple linear regression, it was often useful to transform the response and/or the predictor and this did not change the underlying "linearity" of our model structure

Polynomials - Linear?

8

So, for example, the models:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

are both linear

Design Matrices

9

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{11} \cdot x_{21} \\ 1 & x_{12} & x_{22} & x_{12} \cdot x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{1n} \cdot x_{2n} \end{pmatrix}$$

Linear?

10

$$Y = \beta_0 x_1^{\beta_1} x_2^{\beta_2}$$

We can re-write this model as:

$$\ln(Y) = \ln(\beta_0) + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) = \beta_0' + \beta_1 \ln(x_1) + \beta_2 \ln(x_2).$$

Linear?

11

$$Y = \frac{\beta_0}{1 + \beta_1 x_1 + \beta_2 x_2}$$

can be linearized by taking reciprocals and writing the new model as

$$\frac{1}{Y} = \frac{1}{\beta_0} + \frac{\beta_1}{\beta_0} x_1 + \frac{\beta_2}{\beta_0} x_2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2.$$

Once the γ parameters have been estimated, the relationships

$$\beta_0 = \frac{1}{\gamma_0}; \qquad \beta_1 = \frac{\gamma_1}{\gamma_0}; \qquad \beta_2 = \frac{\gamma_2}{\gamma_0}$$

can be used to find corresponding estimates of the β parameters

Transformations

12

Linearising transformations have an effect on the nature of the error variable.

Can often have the desirable effect of attaining homoscedasticity or normality.

Be careful! It doesn't always have the desired effect.

We will shortly discuss diagnostic procedures to examine these aspects of our model in detail.

Interpreting a Partial Regression Coefficient

Imagine a case with two predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

β_1 represents the expected change in Y when X_1 is increased by one unit, but X_2 is held constant or otherwise controlled.

Partial Regression Coefficients

14

The predictor variables themselves may be interrelated.

The "slopes" parameters in a multiple linear regression may not really be measurable in the population

Multicollinearity (more on this later)

Additive and multiplicative effects

Combined effects of X_1 and X_2 are additive – if both X_1 and X_2 are increased by one unit, expected change in Y would be $(\beta_1+\beta_2)$.

If we add other terms to the model, such as an interaction term between $X_1$and X_2, or squared or higher order terms in one or more of the X's, then the effects of the variables involved are multiplicative and more difficult to interpret.

Intercept

16

Strictly speaking the interpretation of the intercept is the same as it was for simple linear regression; it is the expected value of Y when all the X variables are equal to 0.

The hazards of the interpretation of the intercept are the same as they were for the case of simple linear regression.

Generally consider β_0 as a structural and not a directly interpretable component of the model.

Least Squares Estimation

17

Define the distance function

$$d(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2.$$

The least-squares estimators are those values which minimise this distance function.

With more predictors included we will arrive at a different b_0 and b_1 than we did in the simple regression case.

Minimising the distance function...

18

Had we carried out the individual differentiations in the previous section we would have seen that the residuals must now satisfy p linear constraints;

$$\sum_{i=1}^{n} e_i = 0; \quad \sum_{i=1}^{n} e_i x_{1i} = 0; \quad \ldots \quad \sum_{i=1}^{n} e_i x_{ki} = 0.$$

LS Estimation – Matrix Notation

19

$$d(b) = (Y - Xb)^T (Y - Xb),$$

where b is now a vector of length p where $p = k + 1$.

$$\frac{\partial d}{\partial b} = -2X^T Y + 2(X^T X)b = 0 \quad \Longrightarrow \quad b = (X^T X)^{-1} X^T Y.$$

$$\hat{Y} = Xb = X(X^T X)^{-1} X^T Y = HY$$

$$e = Y - \hat{Y} = Y - HY = (I - H)Y.$$

Look familiar?

Variance of the estimators

20

variance covariance matrix

$$Var(b) = \sigma^2 (X^T X)^{-1}.$$

Under the normal error assumption, we know that

$$b \sim N\{\beta, \sigma^2 (X^T X)^{-1}\}.$$

Once we have an estimator for σ, we can easily construct tests and confidence intervals for the β_k's, since the joint distribution of b implies that the marginal distributions of each b_j is

$$b_j \sim N(\beta_j, \sigma^2 c_{jj}),$$

where c_{jj} is the j^{th} diagonal element of the matrix $(X^T X)^{-1}$.

Regression Scale

21

Estimate σ^2 by starting from the sum of squared errors,

$$SSE = e^T e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

The appropriate number of degrees of freedom by which we divide the SSE in order to arrive at our unbiased estimate of σ^2 is now $n - p$:

$$s^2 = \frac{SSE}{n - p}.$$