§27. Linear Regression

The model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where

Y - dependent variable (or response)

x_i - independent variables (or predictors)

β_i - regression coefficients , β_0 - intercept

ε - noise term

Remark

This model covers nonlinear cases as well.

E.g.

$x_1 = x, \quad x_2 = x^2,$
$x_3 = xz, \quad x_4 = z,$
$x_5 = z^2$

So:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 xz + \beta_4 z + \beta_5 z^2 + \varepsilon$$

Least Squares Estimation of The Parameters

Say

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) \quad ; \quad i = 1, \ldots, n, \quad n > k$$

are observed data points.

Let

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{ik})$$

be the residuals.

Then

$$L = \sum_{i=1}^{k} \varepsilon_i^2$$

is the sum of squared 'errors'. This is the quantity we would like to minimize:

$$\frac{\partial L}{\partial \beta_j} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{1j} \right) x_{ij} = 0 \qquad (1)$$

and we can solve for β_i. It is best to write everything in matrix notation

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then the solutions of the least squares equations (1) are

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

The hat on $\hat{\beta}$ signifies values estimated from the data. For the fitted model predictions, we have $\hat{y} = X\hat{\beta}$ and the residuals are

$$e = y - \hat{y}$$

Assume that the errors, ε_i, are independent and identically distributed distributed random variables with mean 0, and variance σ^2.

Then

1. $\hat{\beta} = (X^\top X)^{-1} X^\top y$ is an unbiased estimator for β :

   $$E(\hat{\beta}) = \beta$$

2. An unbiased estimator for σ^2 is

   $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p}$$

   where $p = k + 1$ is the number of parameters in the model.

Significance of The Regression

Many different test for the 'quality' of the regression could be done. Here are some.

1. Tests on the individual repression coefficients:

   $$H_0 = \beta_i = \beta_{j0} \text{ versus } H_1 : \beta_j \neq \beta_{j0}$$

   The test statistic is:

   $$t_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 c_{jj}}} \qquad ; \qquad (n-p) \text{ df}$$

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

   The most important special case is $H_0 : \beta_i = 0$ versus $H_1 = \beta_i \neq 0$ and H_0 is not rejected. This indicates that the regressor, x_i, can be deleted from the model.

2. R^2 and R^2_{adj}

   $$SS_E = \sum_{i=1}^{m} {\varepsilon_i}^2; \qquad SS_T = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2; \qquad SS_R = SS_T - SS_E$$

   $$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Example

$R^2 = 0.98$ indicates that the model accounts for 98% of the variability in the data.

Note: R^2 can never decrease when a regressor is added. Adding a new regressor which only marginally increases R^2 could be counterproductive - the model is less interpretable; Occam's razor. Thus, the adjusted R^2 statistic is (heavily) used:

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

R^2_{adj} penalizes the analyst for adding terms to the model and is an easy way to prevent over fitting, including regressors which are not really useful. R^2_{adj} is used for variable selection.

Confidence and Prediction Intervals in Regression

A $100(1-\alpha)\%$ confidence interval on the **regression coefficient** β; is

$$\hat{\beta}_j - t_{\alpha/2}\sqrt{\hat{\sigma}^2 c_{jj}} \quad \leq \quad \beta_i \quad \leq \quad \hat{\beta}_j + t_{\alpha/2}\sqrt{\hat{\sigma}^2 c_{jj}} \qquad ; \quad (n-p) \text{ df}$$

A $100(1-\alpha)\%$ confidence interval on the **mean response** at the point $x_{01}, \ldots, x_{0k}$ is

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2}\sqrt{\hat{\sigma}^2 x_0^\intercal \left(x^\intercal x\right)^{-1} x_0} \quad \leq \quad \mu_{Y|x_0} \quad \leq \quad \hat{\mu}_{Y|x_0} + t_{\alpha/2}\sqrt{\hat{\sigma}^2 x_0^\intercal \left(x^\intercal x\right)^{-1} x_0}$$

A $100(1-\alpha)\%$ **prediction interval** for a future observation of the response, Y, at $x_{01}, \ldots, x_{0k}$ is

$$\hat{y}_0 - t_{\alpha/2}\sqrt{\sigma^2\left(1 + x_0^\intercal (x^\intercal x)^{-1} x_0\right)} \quad \leq \quad Y_0 \quad \leq \quad \hat{y}_0 + t_{\alpha/2}\sqrt{\sigma^2\left(1 + x_0^\intercal (x^\intercal x)^{-1} x_0\right)}$$

This prediction interval expresses both the error in estimating the mean of $Y \mid x_0$ as well as the inherent variability of Y at fixed $x = x_0$.

Remark
Never extrapolate far from the data!

Selection of Variables and Model Building

- Step-wise regression. Adding or removing variables at each step based on F-test
- Forward selection: variables are added me at a time
- Backwards elimination: start will all possible regressors and eliminate the insignificant ones one at a time.

Multicollinearity

It is expected that there might be dependencies between the regressors themselves.

Let R_j^2 be the coefficient of determination resulting from regressing x_j on the remaining $k-1$ regressors. The variance of $\hat{\beta}_j$ is effectively 'inflated' with variance inflation factor for β_j:

$$\text{VIF}\left(\beta_j\right) = \frac{1}{1 - R_j^2}$$

Estimation of the regression coefficients is very imprecise when multicollinearity is present. To combat multicollinearity it might be possible to collect more data or simply use a different (nonlinear regression) model.