Second course in Statistics. Lecture 21.

Linear regression analysis

- Model specification
- Assumptions
- Parameter estimation: least squares method and maximum likelihood estimation (MLE)
- Statistical inference: testing and confidence interval
- Diagnostics

Introduction

Regression is used to study relationships between variables. Linear regression is used for a special class of relationships, namely, those that can be described by straight lines, or by generalizations of straight lines to many dimensions.

One variable takes on the special role of a response variable, while all the others are viewed as predictor variables as having values set by the data collector. The value of the response is a function of predictors. A hypothesized model specifies the behaviour of the response given values of the predictors.

The model generally also specifies some of the characteristics of the failure to provide exact fit through hypothesized error terms.

Linear relationship

The word 'linear' means that the selected model is linear in the parameters. The phrase 'linear in the parameters' means that no parameter in the model appears as an exponential or is multiplied by or divided by another parameter.

Example: Which of the followings show linear relationship?

1) $Y = \beta_0 + \beta_1 x + \varepsilon$,

2) $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$,

3) $Y = \beta_0 + \beta_1 \ln(x) + \varepsilon$,

4) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$,

5) $Y = \beta_0 \exp(\beta_1 x) + \varepsilon$

Simple linear regression

Definition: When there is a single predictor variable and the regression equation is assumed to be a linear function, this model is called simple linear model. Simple linear model has the form:

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1,2,\cdots,n$

β_0 is the intercept, the value of Y_i when $x_i = 0$

β_1 is the slope, the rate of change in Y_i for a unit change in x_i

ε_i is a random error term, the difference between observed Y_i and predicted $\hat{Y}_i = \beta_0 + \beta_1 x_i$. These statistical errors are devices that account for the failure of a model to provide an exact fit.

Assumptions

1. The values predicted are the expected values (average values) at given predictors.
2. The response variable is random variable whose values are observed by selecting values of the predictor variables in a desired range.
3. The predictor variables are a set of fixed values representing points of observation for the response variable.
4. The variability in the response variable that cannot be accounted for by the equation is due to random error $\varepsilon_i \quad i = 1,2,\cdots n.$ $E(\varepsilon_i) = 0,$ $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ and $\text{var}(\varepsilon_i) = \sigma^2$ for $i = 1,2,\cdots n.$
5. Further $\varepsilon_i \sim NID(0, \sigma^2)$ for $i = 1,2,\cdots n$ is imposed for the inference of distribution of the estimators.

Least squares estimation

Definition: Least squares method (LS) finds estimate for the parameters in the selected equation by minimizing the sum of the squared deviations of the observed values of the response variable from those predicted by the equation. These values are known as the least squares estimates (LSE) of the parameters.

Minimize $\sum e_i^2 = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$

leading to $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Where $S_{xx} = \sum x^2 - \dfrac{(\sum x)^2}{n}, S_{xy} = \sum xy - \dfrac{(\sum x)(\sum y)}{n}$

Least squares estimation

Estimated regression line is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{y}_i$ is the value on the fitted line at $x = x_i$

Alternative form for estimated regression line: $\hat{Y}_i = \overline{Y} + \hat{\beta}_1 \left(x_i - \overline{x} \right)$

The estimated errors (a.k.a. residuals) $e_i = y_i - \hat{y}_i$

The residuals give the vertical distances between the fitted line and the actual y-values.

Least squares estimation

Properties of the least squares estimators

$E(Y_i) = \hat{Y}_i$

e_i should satisfy the conditions in the assumption.

$\sum_{i=1}^{n} e_i = 0;$

$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i;$

$\sum_{i=1}^{n} x_i e_i = 0.$

Least squares estimation

Estimation of variance σ^2

$$s^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

s^2 is called the residual variance, or mean square error (MSE), and the positive square root s is called the residual standard deviation. S^2 is an unbiased estimator of σ^2.

The residual variance s^2 is an absolute measure of how well the estimated regression line fits the means of the observed response variables. In general the smaller this value, the better the fit.

Least squares estimation

Explanations of variability of response variable

It can be shown that $\Sigma\left(y-\bar{y}\right)^2 = \Sigma\left(\hat{y}-\bar{y}\right)^2 + \Sigma\left(y-\hat{y}\right)^2$

$\Sigma\left(y-\bar{y}\right)^2$ total variability in y-values sum of squares about the mean

$\Sigma\left(\hat{y}-\bar{y}\right)^2$ variability by model sum of squares due to the model

$\Sigma\left(y-\hat{y}\right)^2$ unexplained variability sum of squares for error

Coefficient of determination $R^2 = \dfrac{\Sigma\left(\hat{y}-\bar{y}\right)^2}{\Sigma\left(y-\bar{y}\right)^2}$ the proportion of total variability explained by the regression model.

Example 1

University students learn rather quickly that the better their grade point averages (GPA), the better their chances of finding good jobs upon graduation. Suppose the data are listed in table 1 represent the GPA of 15 recent graduates and their starting annual salaries. We choose starting salaries as response and GPA as predictor. Determine a regression equation for average starting salary as a function of GPA and estimate the parameters in the model specified.

Table 1 (salaries in thousands of dollars)

GPA	Starting salary
2,95	18,5
3,20	20,0
3,40	21,1
3,60	22,4
3,20	21,2
2,85	15,0
3,10	18,0
2,85	18,8
3,05	15,7
2,70	14,4
2,75	15,5
3,10	17,2
3,15	19,0
2,95	17,2
2,75	16,8