10-Regression

ENSY SILVER[1]

Thursday 10th September, 2020

[1]Thanks to my family, my friend and freedom.

1 Introduction

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.

2 The method of least squares

Theorem 2.1. *Given n points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, the straight line $y = a + bx$ minimizing*

$$L = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

has slope

$$b = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and y-intercept

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \overline{y} - b\overline{x}$$

3 Residuals

The difference between an observed y_i and the value of the least squares line when $x = x_i$ is called the ith residual. Its magnitude reflects the failure of the least squares.

Theorem 3.1. *Let a and b be the least squares coefficients associated with the sample $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. For any value of x, the quantity $\hat{y} = a + bx$ is known as the predicted value of y. For any given i, $i = 1, 2, \cdots, n$, the difference $y_i - \hat{y}i = y_i - (a + bx_i)$ is called the i^{th} residual. A graph of $y_i - \hat{y}i$ versus x_i, for all i, is called a residual plot.*

4 Nonliner models

Not all dependent variables have linear relations.

1. If $y = ae^{bx}$, then $\ln y$ is linear with x.
2. If $y = ax^b$, then $\log y$ is linear with $\log x$.
3. If $y = L/(1 + e^{a+bx})$, then $\ln\left(\frac{L-y}{y}\right)$ is linear with x.
4. If $y = 1/(a + bx)$, then $1/y$ is linear with x.
5. If $y = x/(a + bx)$, then $1/y$ is linear with $1/x$.

5 The linear model

Theorem 5.1. *Let $f_{Y|x}(y)$ denote the pdf of the random variable Y for a given value x, and let $E(Y|x)$ denote the expected value associated with $f_{Y|x}(y)$. The function*

$$y = E(Y|z)$$

is called the regression curve of Y on x.

5.1 Simple linear model

A pdf $f_{Y|x}(y)$ and a regression curve $E(Y|x)$ form a simple linear model if it satisfies the following assumptions.

1. $f_{Y|x}(y)$ is a normal pdf for all x.
2. The standard deviation, σ, associated with $f_{Y|x}(y)$ is the same for all x.
3. The means of all the conditional Y distributions are collinear—that is,

$$y = E(Y|x) = \beta_0 + \beta_1 x$$

4. All of the conditional distributions represent independent random variables.

Theorem 5.2. *Let $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$ be a set of points satisfying the simple linear model, $E(Y|x) = \beta_0 + \beta_1 x$. The maximum likelihood estimators for β_0, β_1, and σ^2 are given by*

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^n x_i Y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$
$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Theorem 5.3. *Let $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$ be a set of points satisfying the simple linear model, $E(Y|x) = \beta_0 + \beta_1 x$. Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$ be the maximum likelihood estimators for $\beta_0, \beta_1, \sigma^2$, respectively. Then*

1. *$\hat{\beta}_0$ and $\hat{\beta}_1$ are both normally distributed.*
2. *$\hat{\beta}_0$ and $\hat{\beta}_1$ are both unbiased: $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$.*
3. $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$.
4. $\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}$.

Corollary 5.3.1. *Let $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$ satisfy the assumptions of the simple linear model. Then*

1. *$\hat{\beta}_1$, $\overline{Y}$ and $\hat{\sigma}^2$ are mutually independent.*
2. *$\frac{n\hat{\sigma}^2}{\sigma^2}$ has a χ^2 distribution with $n-2$ degrees of freedom.*

Corollary 5.3.2. *Let $\hat{\sigma}^2$ be the maximum likelihood estimator for σ^2 in a simple linear model. Then $\frac{n}{n-2}\hat{\sigma}^2$ is an unbiased estimator for σ^2.*

Corollary 5.3.3. *The random variables $\hat{Y}$ and σ^2 are independent.*

Corollary 5.3.4. *The unbiased estimator for σ^2 based on σ^2 is denoted S^2, where*

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2$$

Applying corollary 5.3.4, we have the distribution about $\hat{\beta}_1$.

Theorem 5.4. *Let $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$ be a set of points satisfying the simple linear model, $E(Y|x) = \beta_0 + \beta_1 x$. Let $S^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. Then*

$$T_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

has a Student t distribution with $n-2$ degrees of freedom.

Applying theorem 5.4, we could draw inference about β_1. Similarly, inferences about β_0 and σ^2 can be deduced from χ^2 distribution.

5.2 Drawing inferences about $E(Y|x)$

Intuition tells us that a reasonable point estimator for $E(Y|x)$ is the height of the regression line at x—that is, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. By theorem 5.3, the $\hat{Y}$ is unbiased, and the variance is

$$\mathrm{Var}(\hat{Y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]$$

Now, we construct a Student t random variable based on $\hat{Y}$. Specifically,

$$T_{n-2} = \frac{\hat{Y} - (\beta_0 + \beta_1 x)}{S\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \tag{1}$$

Then we can build confidence interval for $E(Y|x) = \beta_0 + \beta_1 x$ from 1.

5.3 Drawing inferences about future observations

Consider the difference $\hat{Y} - Y$. Clearly,

$$E(\hat{Y} - Y) = 0$$

and

$$\mathrm{Var}(\hat{Y} - Y) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \tag{2}$$

Then we can build confidence interval for Y from 2.

5.4 Testing the equality of two samples

Theorem 5.5. *Let $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$ and $(x_1^*, Y_1^*), (x_2^*, Y_2^*), \cdots, (x_n^*, Y_m^*)$ be two independent sets of points, each satisfying the assumptions of the simple linear model—that is, $E(Y|x) = \beta_0 + \beta_1 x$ and $E(Y^*|x^*) = \beta_0^* + \beta_1^* x^*$.*

$$T = \frac{\hat{\beta}_1 - \hat{\beta}_1^* - (\beta_1 - \beta_1^*)}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \frac{1}{\sum_{i=1}^{m}(x_i^* - \overline{x^*})^2}}}$$

and

$$S = \sqrt{\frac{\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 + \sum_{i=1}^{m}[Y_i - (\hat{\beta}_0^* + \hat{\beta}_1^* x_i^*)]^2}{n + m - 4}}$$