Linear regression

Using python.

Iván Andrés Trujilllo Abella

trujilloiv@javeriana.edu.co

Linear regression

When it is useful the linear regression?
for instance, the relation among the income and compsution, body mas index and levels of microalbum, etc, the wage and years of experience, unlike logistic regression the indepent variable not is dicotomy.

A linear model

Where are going consider that our prediction $\hat{y}_i$ differ from the real value y_i in u_i:

$$y_i - \hat{y}_i = u_i \tag{1}$$

and we are going to model $\hat{y}_i = \beta_0 + \beta_1 x_i$, therefore:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{2}$$

Where y_i is our independent term, β_0 is a constant.

MLE

We have a collection of $\vec{y} = y_1, y_2, ..., y_n$, that our objective is build a model that allow us to find the value of β_0, β_1 and maximize the probability of seen the data($\vec{y}$).

Assume that

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{3}$$

and

$$u_i \sim N(0, \sigma^2) \tag{4}$$

Therefore the $E(y_i) = \beta_0 + \beta_1 x_i$ and $var(y_i) = \sigma^2$ (why?).

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \tag{5}$$

If we make distributional assumptions we could make inference and hypotesis testing.

PMF normal distributon

The PMF of a normal variable is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{6}$$

where μ is mean. Therefore for our problem.

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\right)^2\right) \tag{7}$$

Likelihood function

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i) \tag{8}$$

$$\prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2\right) \tag{9}$$

Now rememeber that $e^x.e^y = e^{x+y}$ therefore the *likelihood function* is rewritten as:

$$\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right) \tag{10}$$

Log likelihood function

Apply natural logarithm to the likelihood function

$$\ln(L(\beta_0, \beta_1, \sigma^2)) = -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \qquad (11)$$

now we need establish the partial derivatives for β_0 and β_1. Notice that to maximize the expression we need minimize $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$.

$$\frac{\partial ln(L)}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{12}$$

$$\frac{\partial ln(L)}{\partial \beta_1} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \tag{13}$$

We are going to reach the same equations minimizing the error among the prediction and the real value of y_i.

Solving the problem

We have an analytical solution!.

Ordinary least squares

$$u = \sum (y_i - \hat{y}_i)^2 \tag{14}$$

note that $\hat{y}_i = \beta_0 + \beta_1 x_i$.

$$u^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2 \tag{15}$$

the first order condition require $\frac{\partial u^2}{\partial \beta_0} = 0, \frac{\partial u^2}{\partial \beta_1} = 0$.

Chain rule

to get

$$\frac{\partial u^2}{\partial \beta_0} = \frac{\partial u^2}{\partial u}\frac{\partial u}{\partial \beta_0} \tag{16}$$

$$\frac{\partial u^2}{\partial \beta_1} = \frac{\partial u^2}{\partial u}\frac{\partial u}{\partial \beta_1} \tag{17}$$

Remember that $\frac{d\sum g_i(x)}{dx} = \frac{\sum dg_i(x)}{dx}$. Therefore $\frac{\partial u^2}{\partial u} = 2u$ and $\frac{\partial u}{\partial \beta_0} = -1$, $\frac{\partial u}{\partial \beta_1} = -x_i$.

Partial derivatives

$$\frac{\partial u^2}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_i) \tag{18}$$

$$\frac{\partial u^2}{\partial \beta_1} = -2 x_i \sum (y_i - \beta_0 - \beta_1 x_i) \tag{19}$$

note that this will be zero if we know the exactly parameters.

Solving

Analytically

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)}. \tag{20}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{21}$$

To see the full derivation of 20 and 21 **click here**

Gradient descent

...

$$\nabla J(\vec{\Theta}) = \begin{pmatrix} \frac{\partial J(\Theta)}{\partial \Theta_1} \\ \frac{\partial J(\Theta)}{\partial \Theta_2} \\ \vdots \\ \frac{\partial J(\Theta)}{\partial \Theta_k} \end{pmatrix}$$

...

$$\vec{\Theta}_{t+1} = \vec{\Theta}_t - \alpha \nabla J(\vec{\Theta}_t) \tag{22}$$

Linear regression

In this case we could say that

...

$$\hat{y}_i = \Theta_0 + \Theta_1 x_{i1} + \Theta_2 x_{i2} + ... + \Theta_k x_{ik} \tag{23}$$

and our Cost function is defined as

....

$$J(\vec{\Theta}) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - y_i)^2 \tag{24}$$

The factor $\frac{1}{N}$ is for avaraged and 2 for convenience.

Computing gradient

...

By rule chain

$$\frac{\partial J(\vec{\Theta})}{\partial \Theta_j} = \sum_{i=1}^{N} \frac{\partial J(\vec{\Theta})}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \Theta_j} \tag{25}$$

...

$$\frac{\partial J(\Theta)}{\partial \hat{y}_i} = -\frac{1}{N}(y_i - \hat{y}_i) \tag{26}$$

...

$$\frac{\partial \hat{y}_i}{\partial \Theta_j} = x_{ij} \tag{27}$$

Gradient

...

$$\begin{aligned}\frac{\partial J(\vec{\Theta})}{\partial \Theta_j} &= -\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)x_{ij} \\ &= \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_{ij}\end{aligned} \tag{28}$$

It is important note that are necessary an initial value of $\vec{\Theta}_0$.

Gradient

...

$$\nabla J(\vec{\Theta}) = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_{i1} \\ \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_{i2} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_{ik} \end{pmatrix} \tag{29}$$

This expression could be write in a matricial form.

Matricial way

...

$$\begin{aligned}\nabla J(\vec{\Theta}) &= \frac{1}{N}\boldsymbol{X}^T(\vec{\hat{y}} - \vec{y}) \\ &= \frac{1}{N}\boldsymbol{X}^T(\boldsymbol{X}\vec{\Theta} - \vec{y})\end{aligned} \tag{30}$$

As excersice put the dimension to the arrays.

...

Implementation from scratch(Click here)

Standardized coefficients

Suppose a X vector (exogenous) and y (endogenous) that are transformed in Z punctuation and for instance in the regression: x_1 have associated β_1. The interpretation is:

Interpreation

The increase of one standard deviation in x_1 is associated with the increase (reduction) of y in β_1 standard deviations.

Get standardized from OLS

β_1 is a no-standardized coefficient, and β_{1std} is obtained from:

$$\beta_{1std} = \frac{\sigma_x}{\sigma_y}\beta_1 \tag{31}$$

where σ_x and σ_y are estimated standard deviations.

Quantile regression

Interaction in regression

interaction term, we can not interpret the variables used in interaction.

Interacted terms

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{1i} * x_{2i} \tag{32}$$

Asumme that x_1 is binary and x_2 is numerical, regardless to the model there are two considerations when $x_{1i} = 1$ then:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{2i} \tag{33}$$

$$\hat{y}_i = (\hat{\beta}_0 + \hat{\beta}_1) + (\hat{\beta}_2 + \hat{\beta}_3) x_{2i} \tag{34}$$

when $x_{1i} = 0$ then we have:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_2 x_{2i} \tag{35}$$

According to this we can interpret to $\hat{\beta}_1$ as the difference in intercepts between categories and $\hat{\beta}_3$ is the differece in slopes between of them.

Interacted terms - Dummy variables

Assume that x_{1i}, x_{2i} are binaries.

	$x_1 = 1$	$x_1 = 0$
$x_2 = 1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$
$x_2 = 0$	$\beta_0 + \beta_1$	β_0

The above table we can get the difference with: $\hat{\beta}_2 + \hat{\beta}_3$, $\hat{\beta}_2$, $\hat{\beta}_1 + \hat{\beta}_3$, $\hat{\beta}_1$, and finally $\hat{\beta}_3$.

$\hat{\beta}_3$ is also difference in difference. The earning associated with $x2 = 1$ is $\hat{\beta}_3$ more or less (according to the sing) for $x_1 = 1$ in comparataion with $x_1 = 0$.