Linear Regression

Tianhang Zheng

https://tianzheng4.github.io

Bias and Variance (Corrections)

$$E_D\left[\left(Y - g(X)\right)^2\right] = bias^2 + variance + \sigma^2$$

$$\left.\begin{array}{ll} \text{Bias:} & E_D[g(X;D)] - f(X) \\ \text{Variance:} & \mathrm{E_D}[(\mathrm{E}_D[g(X;D)] - g(X;D))^2] \end{array}\right\}\ \text{Depends on model complexity}$$

Irreducible error: σ

Linear Regression

Linear Regression (LR) is one of the simplest methods for modeling

Linear Regression assumes that the dependence of Y on X_1, X_2, X_3 … is linear

In most cases, regression function is not linear (but interpretable)

Simple Linear Regression

Linear Regression with a single predictor (Assume the ideal model is a linear function)

$$Y = \beta_0 + \beta_1 X + \epsilon$$

β_0 is called intercept and β_1 is called slope, which are two parameters.

ϵ is the error term: $\epsilon \sim N(0, \sigma^2)$

Simple Linear Regression

The objective is to learn (estimate) β_0 and β_1

The estimates of β_0 and β_1 are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{y}$ is an estimate (prediction) of outcome given $X = x$

$e = y - \hat{y}$ is the residual

Least Squares Method

The least squares method is commonly used for estimating β_0 and β_1

Given a training dataset $D_{tr} = \{(x_i, y_i)\}_{i=1}^{N}$, the residual sum of squares (RSS) can be defined as

$$RSS = \sum_i e_i^2 = (y_i \; - \hat{y}_i)^2$$

Least Squares Method

$$\min_{\hat{\beta}_0, \hat{\beta}_1} RSS = \sum_i e_i^2 = \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Take the derivative and set the derivative as 0

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Analyzing Least Squares Method

Under the assumption of linear regression model (ideal model)

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$E_{\epsilon_i}(\hat{\beta}_1) = E_{\epsilon_i}\left[\frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}\right] = E_{\epsilon_i}\left[\frac{\sum_i (x_i - \overline{x})(\epsilon_i + \beta_1 (x_i - \overline{x}))}{\sum_i (x_i - \overline{x})^2}\right]$$

Analyzing Least Squares Method (Unbiased)

$$E_{\epsilon_i}[\sum_i (x_i - \overline{x})] = 0$$

$$E_{\epsilon_i}\left[\frac{\sum_i (x_i - \overline{x})(\epsilon_i + \beta_1 (x_i - \overline{x}))}{\sum_i (x_i - \overline{x})^2}\right]$$

$$= \frac{1}{\sum_i (x_i - \overline{x})^2} E_{\epsilon_i}\left[\sum_i (x_i - \overline{x})(\epsilon_i + \beta_1 (x_i - \overline{x}))\right]$$

$$= \frac{\beta_1}{\sum_i (x_i - \overline{x})^2} E_{\epsilon_i}[\sum_i (x_i - \overline{x})^2] = \beta_1$$

Why unbiased?

Analyzing Least Squares Method

The variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$SE^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2} \qquad SE^2(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2}\right]$$

$$\sigma^2 = Var(\epsilon)$$

$\hat{\beta} \sim N(\beta, SE^2(\hat{\beta}))$ Why Gaussian distribution?

Confidence Level

$\hat{\beta} \sim N(\beta, SE^2(\hat{\beta}))$ means that $\beta \sim N(\hat{\beta}, SE^2(\hat{\beta}))$

A 95% confidence interval is defined as a range of values with 95% probability, and the interval for the least square method is

$$[\hat{\beta} - 2SE(\hat{\beta}), \hat{\beta} + 2SE(\hat{\beta})]$$

There is 95% probability that this interval contains the true β

Hypothesis Testing

H_0: There is no relationship between X and Y?

H_1: There is some relationship between X and Y?

$H_0: \beta_1 = 0 \qquad H_1: \beta_1 \neq 0$

Hypothesis Testing

For the hypothesis testing, we need a t-statistic (not z-statistic)

Because we do not know the true σ!

We can only estimate σ by $\hat{\sigma}^2 = \frac{1}{|D|}\sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$

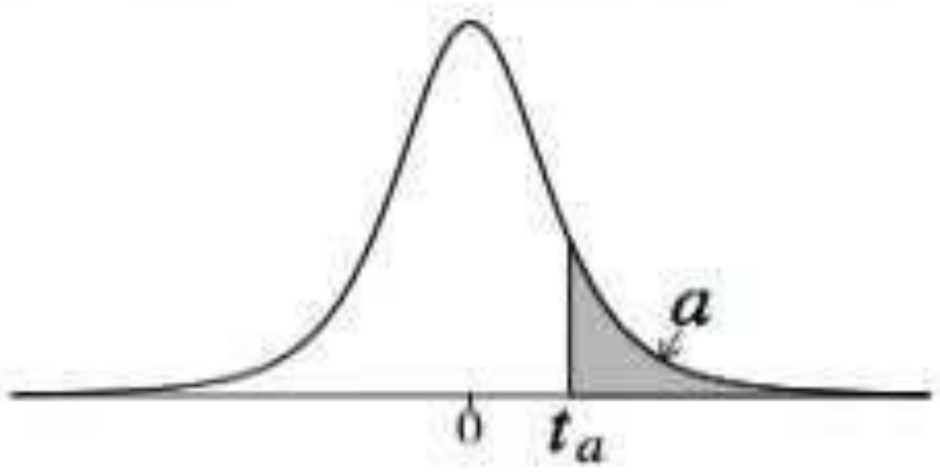
Hypothesis Testing

$$\widehat{SE}^2\ (\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_i (x_i - \overline{x})^2}$$

We could compute a t-statistics by $t = \dfrac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)}$

This variable should satisfy t-distribution with n-2 degrees

Hypothesis Testing



	Area to the right (a)								
df	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781

Hypothesis Testing

If $|t|$ is very large, which means α is very small

Then we could reject H_0

This means there is some relationship between X and Y

Prediction Error

The residual sum of squares (RSS)

$$RSS = \sum_i e_i^2 = (y_i - \hat{y}_i)^2$$

The residual standard error (RSE)

$$RSE = \frac{1}{n-2}\sqrt{RSS}$$

Prediction Error

The residual sum of squares (RSS)

$$RSS = \sum_i e_i^2 = (y_i \ - \hat{y}_i)^2$$

The residual standard error (RSE)

$$RSE = \frac{1}{n-2}\sqrt{RSS}$$

R-Squared

The proportion of the variance that can be explained by a model

$$R^2 = 1 - \frac{RSS}{TSS}$$

TSS is the total sum of squares (total variance of y)

$$TSS = \sum_i (y_i - \bar{y})^2$$

R-Squared

For linear regression, R-squared is the square of the correlation

$$R^2 = r^2$$

r is the correlation between X and Y

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - y)^2}}$$

Q & A