Linear regression

Single-variable linear regression

Given two random variables X and Y, we can use regression to predict Y from X and estimate the error bars around the prediction.

Assume that $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d. from some unknown joint distribution $\mathbb{P}.$

$\mathbb{P}$ can be described by one of the following:

- Joint PDF $h(x,y)$
- Marginal density of X: $h(x) = \int h(x,y) dy$ and conditional density $h(y|x) = \frac{h(x,y)}{h(x)}$

The conditional density $h(y|x)$ contains all information about Y given X.

The conditional expectation of Y given $X=x$ is $\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty}yh(y|x)dy$.

Conditional median of Y given $X=x$: $\int_{-\infty}^{m(x)}yh(y|x)dy = \frac{1}{2}$

Other ways of describing the conditional distribution are conditional quantiles and variance.

Regression

The regression function of Y given X can be defined s:

$$\nu(x) = \mathbb{E}[Y|X=x]$$

In the continuous case: $\mathbb{E}[Y|X=x] = \sum_{\Omega_Y}y \mathbb{P}(Y=y|X=x)$

In the discrete case: $\mathbb{E}[Y|X=x] = \int_{\Omega_Y}yh(y|x)$

The simplest regression function is the linear (or affine) function because it is the simplest function where y has a dependence on x:

$$\nu(x) = \mathbb{E}[Y|X=x] = a + bx$$

Theoretical linear regression

The theoretical linear regression of Y on x is the line $y=a^*+b^*x$ such that $\mathbb{E}[(Y-a-bX)^2]$ is minimized.

Setting the partial derivatives of $\mathbb{E}[(Y-a-bX)^2]$ to zero gives:

$$b^* = \frac{{\rm Cov}(X,Y)}{{\rm Var}(X)}$$ $$a^* = \mathbb{E}[Y] - b^*\mathbb{E}[X] = \mathbb{E}[Y] - \frac{{\rm Cov}(X,Y)}{{\rm Var}(X)}\mathbb{E}[X]$$

Noise

Data points will be exactly on the line (if ${\rm Var}(Y|X=x)>0$). The deviation from the regression line is called noise and defined as:

$$\varepsilon = Y - (a^* + b^*X)$$

Properties of noise:

- satisfies $Y = a + bX + \varepsilon$
- $\mathbb{E}[\epsilon] = 0$
- ${\rm Cov}(X, \epsilon) = 0$

According to this definition, each data point satisfies the following relation:

$$ y_i = a^* + b^* x_i + \epsilon_i $$

Assume that the noise are independent and have variance $\mathrm{Var}[\varepsilon_i] = \sigma^2$. Then, the following equations hold:

$$ \hat{b} = b^* + \frac{\bar{(x - \bar{x})\varepsilon}}{\bar{x^2} - \bar{x}^2} $$

$$ \hat{a} = a^* + \bar{\varepsilon} - \bar{x} \frac{\bar{(x - \bar{x})\varepsilon}}{\bar{x^2} - \bar{x}^2} $$

The variances of the estimator are as follows:

$$ \mathrm{Var}[\hat{b} - b^*] = \frac{n - 1}{n^2} \frac{\sigma^2}{\sigma_x^2} \mathrm{,\ where\ } \sigma_x^2 = \bar{x^2} - \bar{x}^2 $$

$$ \mathrm{Var}[\hat{a} - a^*] = \frac{\sigma^2}{n} \left(1 + \frac{\bar{x}^2}{\sigma_x^2}\frac{n - 1}{n} \right)$$

Quantifying error

- Residual squared error (RSS)

$$ \sum_{i = 1}^{n} (y_i - \hat{a} - \hat{b} x_i)^2 $$

- R-square (normalized error metric, does not change based on scaling of data):

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

where

$$ \mathrm{TSS} = \sum_{i = 1}^{n} (y_i - \hat{y})^2 $$

In other words, R^2 describes how much of the variation in y is captured by the regression.

Predictive distribution

Predictive error is the difference between the prediction and true value.

$$ \hat{Y}(x) - a^* - b^*x = \frac{1}{n} \sum_{i = 1}^{n} \varepsilon_i \left\{ 1 + \frac{(x_i - \bar{x})(x - \bar{x})}{\sigma_x^2} \right\} $$

The expectation is 0, of course:

$$ \mathbb{E}[\hat{Y}(x) - a^* - b^*x] = 0 $$

The variance is:

$$ \mathrm{Var}[\hat{Y}(x) - a^* - b^*x] = \mathbb{E}[(\hat{Y}(x) - a^* - b^*x)^2] = \frac{\sigma^2}{n} \left(\frac{(x - \bar{x})^2}{\sigma_x^2} \frac{n-1}{n} + 1 \right) $$

The distribution is Gaussian if ε_i are Gaussian. If it is Gaussian, then we can easily compute confidence intervals.

Multivariate linear regression

Setup

$$\textbf{Y}_i = \textbf{X}_i^T \boldsymbol \beta^* + \varepsilon_i, i = 1, \ldots, n$$

where

- $\textbf{X}_i$ is the vector of explanatory variables or covariates
- $\textbf{Y}_i$ is the response/dependent variable
- $\boldsymbol \beta^* = (a^*, \textbf{b}^T)^T$ (β_1^* is the intercept)
- $\varepsilon_{i=1, \ldots, n}$: noise terms

Then, the least squares estimator (LSE) of $\hat{\boldsymbol \beta}$ is the minimizer of the sum of errors squared:

$$\hat{\boldsymbol \beta} = {\rm argmin}_{\boldsymbol \beta \in \mathbb{R}^p} \sum_{i = 1}^n ({Y}_i-\textbf{X}_i^T\beta)^2$$

Matrix form

- $\textbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}_n$ (vector of observations)
- $\mathbb{X}$ be the design matrix whose rows are $\textbf{X}_1^T, \ldots, \textbf{X}_n^T$ (aka the design matrix, $n \times p$)
- $\boldsymbol \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ (vector of noise)
- Then, $\textbf{Y} = \mathbb{X}\boldsymbol \beta^* + \boldsymbol \varepsilon$. $\boldsymbol \beta^*$ is the unknown model parameter.

The least squares estimator of the unknown model parameter $\hat{\boldsymbol \beta}$ is:

$$\hat{\boldsymbol \beta} = {\rm argmin}_{\boldsymbol \beta \in \mathbb{R}^p} ||\textbf{Y} - \mathbb{X}\boldsymbol \beta||_2^2$$

Each row of $\mathbb{X}$ represents one set of explanatory variables, and the corresponding row/element of $\textbf{Y}$ represents the response variable for that set of explanatory variables. The corresponding row/element of $\boldsymbol \varepsilon$ represents the error between the true response variable and the value predicted by the regression model.

$\mathbb{X}$ is an $n \times p$ matrix, where n is the number of observations, and p is the number of covariates, including one constant covariate.

Evaluating the least-squares estimator

By setting the gradient of the sum of errors squared to zero, we find that the LSE $\hat{\boldsymbol \beta}$ must satisfy:

$$\mathbb{X}^T\mathbb{X} \hat{\boldsymbol \beta} = \mathbb{X}^T \textbf{Y}$$

To isolate $\hat{\boldsymbol \beta}$, we can multiply both sides by $(\mathbb{X}^T\mathbb{X})^{-1}$ from the left. To do this, $\mathbb{X}^T\mathbb{X}$ must be invertible. $\mathbb{X}$ having rank equal to the number of covariates will guarantee that $\mathbb{X}^T\mathbb{X}$ is invertible.

If ${\rm rank}(\mathbb{X}) < p$, where p is the number of covariates, there will be an infinite collection of estimators that satisfy the least-squares condition.

If ${\rm rank}(\mathbb{X}) = p$, there will be a unique LSE $\hat{\boldsymbol \beta}$.

$$\hat{\boldsymbol \beta} = (\mathbb{X}^T\mathbb{X})^{-1} \mathbb{X}^T \textbf{Y}$$

Deterministic design

When we use deterministic design, we make the following assumptions:

- $\mathbb{X}$ is deterministic, and ${\rm rank} \mathbb{X} = p$
- $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. (The model is homoscedastic)
- The noise vector $\boldsymbol \varepsilon$ is Gaussian. $\boldsymbol \varepsilon \sim \mathcal{N}_n(0, \sigma^2I_n)$

$$\textbf{Y} = \mathbb{X}\boldsymbol \beta + \boldsymbol \varepsilon$$

Implications

- This way, the only random element in the equation for the response variable Y is the noise $\boldsymbol \varepsilon$.
- The response variable Y is therefore a Gaussian random variable.
- The LSE $\hat{\boldsymbol \beta}$ is also a Gaussian random variable.

LSE properties

The LSE is equal to the maximum likelihood estimator (MLE).

Distribution

$$\hat{\boldsymbol \beta} \sim \mathcal{N}_p(\boldsymbol \beta^*, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$$ The distribution of the LSE $\hat{\boldsymbol \beta}$ is a p-dimensional Gaussian with mean $\boldsymbol \beta^*$ and variance $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$.

Quadratic risk

$$\mathbb{E}[||\hat{\boldsymbol \beta} - \boldsymbol \beta||_2^2] = \sigma^2{\rm tr}(\sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$$

The quadratic risk is defined as the typical error in the LSE $\hat{\boldsymbol \beta}$ compared to the true parameter $\boldsymbol \beta$.

${\rm tr}(\mathbb{X})$ is the trace, defined as the sum of elements on the main diagonal of X.

Prediction error

$$\mathbb{E}[||\textbf{Y} - \mathbb{X} \hat{\boldsymbol \beta}||_2^2] = \sigma^2(n-p)$$

The prediction error is defined as the typical error between model predictions $\mathbb{X} \hat{\boldsymbol \beta}$ and observations $\textbf{Y}$.

Variance estimator

Unbiased estimator of σ^2: $\hat{\sigma}^2 = \frac{||\textbf{Y} - \mathbb{X} \hat{\boldsymbol \beta}||_2^2}{n - p} = \frac{1}{n - p} \sum_{i = 1}^{n} \hat{\varepsilon}_i^2$

Theorem

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n - p}^2$$ $$\hat{\boldsymbol \beta} \perp \hat{\sigma}^2$$

Significance testing

Hypothesis testing setup example:

$$H_0: \beta_j = 0$$ $$H_1: \beta_j \neq 0$$

If γ_j is the jth diagonal coefficient of $(\mathbb{X}^T\mathbb{X})^{-1}$

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2\gamma_j}} \sim t_{n - p}$$

The test statistic is $T_n^{(j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2\gamma_j}}$.