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