Lesson: Introduction to least squares estimation

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

realization from a (normal) random variable

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

measured response

Unknown, to be estimated

measured predictors

Estimating Model Parameters

The line of best fit to the data is the line that minimizes the sum of the squared vertical distances between the line and the observed points:

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$$

measured response

Unknown, to be estimated

measured response

Estimating Model Parameters

The ***surface*** of best fit to the data is the surface that minimizes the sum of the squared vertical distances between the surface and the observed points:

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \parallel \mathbf{Y} - X\boldsymbol{\beta} \parallel^2$$
$$= (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta})$$

The problem is the find a $\boldsymbol{\beta}$ so that $X\boldsymbol{\beta}$ is as close as possible to $\boldsymbol{y}$.

Estimating Model Parameters

Lesson: Linear algebra for least squares

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

Let X be an $m \times n$ matrix, $\mathbf{v}$ be $n \times 1$, and $\mathbf{y}$ be $m \times 1$. Then:

1. *Lemma 1*: Then $X^T X$ is symmetric, i.e., $(X^T X)^T = X^T X$.

2. *Lemma 2*: Let $\mathbf{y} = X\mathbf{v}$. Then $\frac{\partial y}{\partial \mathbf{v}} = X$ and $\frac{\partial y^T}{\partial \mathbf{v}} = X^T$

3. *Lemma 3*: Let $c = \mathbf{v}^T (X^T X)\mathbf{v}$. Then $\frac{\partial c}{\partial \mathbf{v}} = 2X^T X\mathbf{v}$

Lesson: Deriving the least squares solution

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

Definition: The *residuals* are defined as:

Definition: The *fitted values* are defined as:

Definition: The *hat matrix*, H, is defined as:

The hat matrix is useful in theoretical calculations related to least squares

Least Squares Estimation: We define the best estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ as the one that minimized the sum of the squared residuals:

Differentiating with respect to $\boldsymbol{\beta}$, we get:

Lesson:Justifying least squares: the Gauss-Markov theorem and maximum likelihood estimation

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

The maximum likelihood estimator.

Suppose that $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$**.** Then:

1. marginal pdf:

2. joint pdf:

3. log-likelihood:

Lesson: Sums of squares and estimating the error variance

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

Photo by Nils Rasmusson on Unsplash

3
2
1
0.00
0.25
0.50
0.75
1.00
x

Sums of squares:

1. RSS: Residual sum of squares:

2. ESS: *Explained* (or regression) sum of squares:

3. TSS: *Total* sum of squares:

Sums of Squares

The residual sum of squares RSS can be interpreted as a measure of how much variation in y is left ***unexplained by the model***—that is, how much cannot be attributed to a linear relationship.

Estimating σ^2

The parameter σ^2 determines the amount of spread about the true regression line. Two separate examples:

Photo by visuals on Unsplash

An estimate of σ^2 will be used in statistical inference (e.g., confidence interval formulas and hypothesis testing), presented in the next two sections.

Given our interpretation of the RSS, it seems plausible that contribute to an estimator of σ^2:

Photo by visuals on Unsplash

Note:

1. The divisor $n-(p+1)$ in is the number of degrees of freedom (df) associated with RSS and $\hat{\sigma}^2$.

2. The RSS has $n-(p+1)$ df because $p+1$ parameters must first be estimated to compute it, which results in a loss of $p+1$ df.

3. Replacing each y_i in the formula for $\hat{\sigma}^2$ by the r.v. Y_i gives a random variable.

Lesson: The coefficient of determination, “R-squared”

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

The *coefficient of determination,* R^2, is defined as:

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

Note:

- $0 \leq R^2 \leq 1$
- Assuming that the model is correct, R^2 is interpreted as the proportion of observed variation in y explained by the model.

Warnings about R^2:

1. R^2 can be close to 1 but the model is the wrong fit for the data.

2. R^2 can be close to 0 even when the model is the correct fit for the data.

3. R^2 should **not** be used to compare models with a different number of predictors.

4. R^2 says nothing about the causal relationship between the predictors and the response.

Lesson: The problem of non-identifiability

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

The least squares estimate is the solution to the *normal* equations:

$$X^T X \boldsymbol{\beta} = X^T \mathbf{Y}.$$

1. When $(X^T X)^{-1}$ exists, there is a unique solution, $\hat{\boldsymbol{\beta}}$.

2. When $(X^T X)^{-1}$ does not exist, there will be infinitely many solutions.

Definition: When $(X^T X)^{-1}$ does not exist, the regression model is said to be ***non-identifiable*** (or, ***unidentifiable***).

When does this happen?

Photo by Nils Rasmusson on Unsplash

Why might we have non-identifiability?

1. One variable is just a multiple of another.

2. One variable is a ***linear combination*** of several others.

3. There are more variables than members in the sample.

Note: ***Near non-identifiability*** is trickier than exact non-identifiability.