Intro to Regression

Simple Linear Regression

Linear Regression relates inputs to outputs. For instance, predicting grades based on study time. Notations include:

- Y - Target variable
- y_j - j-th observation for Y
- X_i - i-th predictor
- x_{ij} - i-th predictor's j-th observation

Simple Linear Regression

In **Simple Linear Regression**, one predictor exists. The relationship is:

$$Y \sim N(\beta_0 + \beta_1 X_1, \sigma)$$

or

$$E(Y) = \beta_0 + \beta_1 X_1$$

Coefficients β_0 and β_1 represent the intercept and slope, respectively. Including σ, there are three unknown parameters.

Simple Linear Regression

Suppose you have data points $(y_1, x_{11}), (y_2, x_{12}), ..., (y_n, x_{1n})$. Maximum Likelihood Estimation can estimate β_0, β_1, and σ using the following likelihood:

$$L(\beta_0, \beta_1, \sigma|.) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_j - (\beta_0 + \beta_1 x_{1j}))^2}{2\sigma^2}\right)$$

Interpreting Regression Coefficients: An Example

Let's consider a simple example to understand the interpretation of regression coefficients. Assume we have conducted a study on the following data:

- Number of Hours Studied: $x_1 = [1, 2, 3, 4, 5]$
- Test Scores: $y = [53, 59, 61, 65, 70]$

After fitting a simple linear regression model, we find:

- Intercept $(\beta_0) = 50$
- Slope $(\beta_1) = 4$

Interpreting the Coefficients

In this example:

- The **Intercept** **($\beta_0 = 50$)** indicates that if a student does not study at all ($x_1 = 0$), their expected test score would be 50.
- The **Slope** **($\beta_1 = 4$)** signifies that for each additional hour of study, we can expect the test score to increase by 4 points, holding all else constant.

For example:

- If a student studies for 3 hours, the expected test score can be calculated as $50 + 4 \times 3 = 62$.
- If a student studies for 5 hours, the expected test score can be calculated as $50 + 4 \times 5 = 70$.

We can also estimate σ and we can use that to talk about uncertainty. BUT not yet. There's an extra layer we will unpack at a later date

Simple Linear Regression

For regression, we can minimize the target function:

$$T(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_j - (\beta_0 + \beta_1 x_{1j}))^2$$

This is called "Least Squares" or "Ordinary Least Squares (OLS)"

How does this compare to viewing regression as a probability model"

- The estimates for β_0 and β_1 are in fact the exact same.
- There is no σ so there is no uncertainty built into the model.

Multiple Linear Regression: Introduction

Multiple Linear Regression allows for more than one predictor variable to model the relationship with the response variable.

$$E(Y_j) = \beta_0 + \beta_1 x_{1j} + \ldots + \beta_p x_{pj}$$

Here p is the number of predictor variables.

Coefficients in Multiple Linear Regression

Each coefficient has its own interpretation:

- β_0 is the intercept, indicating the expected value of Y when all X_i are zero.
- β_i represents the expected change in Y when X_i increases by 1, holding all other predictor variables constant.

Optimizing in Multiple Linear Regression

We want maximize the. likelihood function:

$$L(\beta_0, \beta_1, \sigma|.) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_j - (\beta_0 + \beta_1 x_{1j} + \ldots + \beta_p x_{pj}))^2}{2\sigma^2}\right)$$

or minimize the target function to find the best-fit coefficients.

$$T(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} (y_j - (\beta_0 + \beta_1 x_{1j} + \ldots + \beta_p x_{pj}))^2$$

Example: Housing Prices

Let's consider an example where we predict the price of a house based on two predictors:

- Square Footage (X_1)
- Number of Bedrooms (X_2)

After analysis, we find:

$$\text{Price} = 50,000 + 120 \times (\text{Square Footage}) + 25,000 \times (\#\text{ of Bedrooms})$$

Interpreting Coefficients: Housing Prices

In this example:

- $\beta_0 = 50,000$ indicates that a house with zero square footage and zero bedrooms is theoretically priced at \$50,000.
- $\beta_1 = 120$ implies that for each additional square foot, the house price will increase by \$120 when the number of bedrooms is constant
- $\beta_2 = 25,000$ signifies that adding one bedroom will increase the house price by \$25,000, holding square footage constant.