About Simple Linear Regressions

Ray Zhang

October 2, 2025

1 Introduction

My name is Ray Zhang, and I am a junior at Northwestern University, originally from San Antonio, Texas. I am expected to graduate in June 2027 with a B.A. in Data Science and Mathematics and a B.M. in Viola Performance.

In this document, I will outline simple linear regressions and explain how we calculate p-values from correlations in observed data.

2 Regression Model Basics

A simple linear regression model assumes that the dependent variable Y is linearly related to the independent variable X with an additive error term ϵ:

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

Here, ϵ represents random noise or variation not explained by X. For inference purposes, we typically assume that

$$\epsilon \sim N(0, \sigma^2)$$

where σ is the standard deviation of the error term. This assumption allows us to make probabilistic statements about the estimated regression coefficients.

3 Estimating the Slope and Standard Error

The slope of the regression line, $\hat{\beta}_1$, is estimated from the data as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The standard error of the slope, which quantifies uncertainty in $\hat{\beta}_1$, is given by

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

where $\hat{\sigma}$ is the residual standard deviation, estimated as

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

and $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

4 Calculating P-values

In Python, the `scipy.stats.pearsonr` function can test the linear relationship between two variables, returning a correlation coefficient r and its associated p-value.

The null and alternative hypotheses for this test are:

$$H_0 : \rho = 0 \quad \text{(no linear correlation)}$$

$$H_a : \rho \neq 0 \quad \text{(linear correlation exists)}$$

The t-statistic used to test these hypotheses is

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

which follows a Student's t-distribution with $n-2$ degrees of freedom. The p-value is derived from this distribution.