Predicting Student Performance

Linear Models (PHP2601), Prof. Ani Eloyan

Daniel Posmik, Jizhou Tian, Aristofanis Rontogiannis

2024-12-06

Table of contents I

EDA and the Linear Model

Introduction

We will be analyzing educational data.

We are interested in understanding the predictors of student performance as measured by exam scores.

We will be using a publicly available dataset from Kaggle that contains information about students and their exam scores.

Hypothesis to be Tested

We want to further explore a specific hypothesis about a subset of predictor variables. Suppose we maintain that the following variables are significant predictors:

- Hours Studied
- Attendance
- Sleep Hours
- Previous Scores
- Tutoring Sessions

We can formalize this question as follows:

- $H_0 : \begin{bmatrix} 1_{[0,\cdots,p+1]}, & 0_{[p+2,\cdots,P]} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 & \cdots & \beta_P \end{bmatrix}^T = \beta_0 + \cdots + \beta_{p+1} = 0$
- $H_A : \{\beta_1 \neq 0\} \cap \cdots \cap \{\beta_5 \neq 0\}$

Exploratory Data Analysis (EDA)

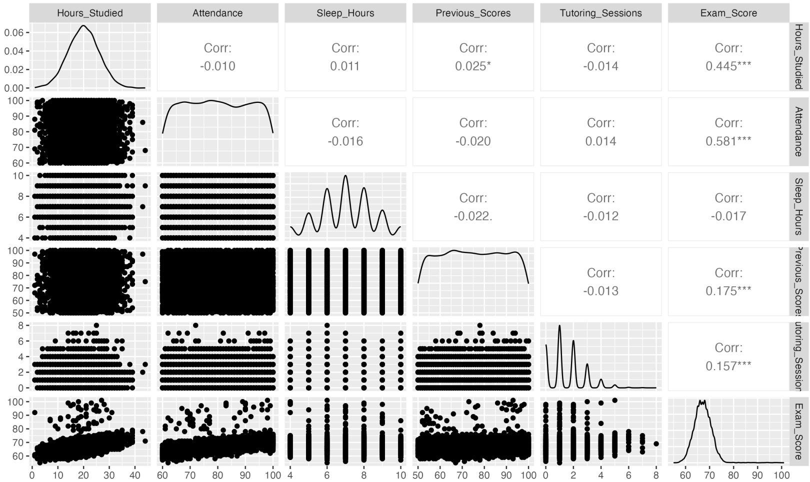

Figure 1: Correlation Matrix

The Linear Model

Let us begin by discussing the assumptions of linear regression model. In a Gauss-Markov setting, we assume that our linear model is of the form:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & X_{13} & \cdots & X_{1(p+1)} \\ 1 & X_{22} & X_{23} & \cdots & X_{2(p+1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & X_{n3} & \cdots & X_{n(p+1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $\mathbb{E}[\epsilon] = 0$ and Var$[\epsilon] = \sigma^2 I$ denote the zero-mean and constant variance assumptions. In our case, we begin with $p = 5$, i.e. our design matrix has $p + 1$ columns, accounting for the intercept term.

Variable Transformations

We will transform the variables to ensure that the assumptions of the linear model are met.

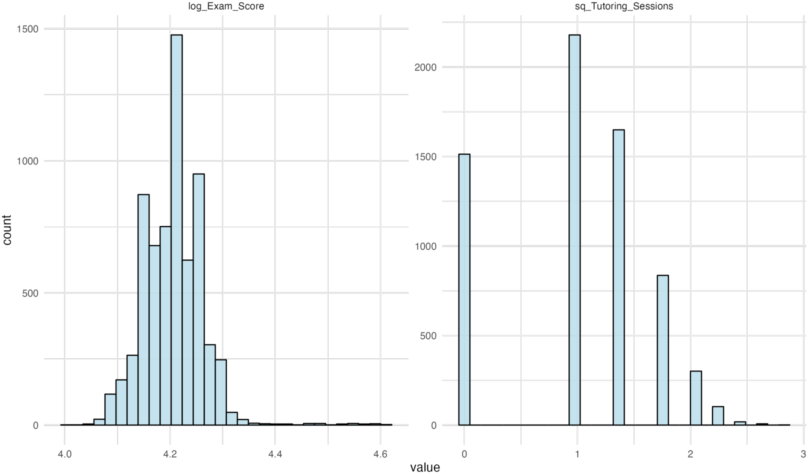

Figure 2: Variable Transformation

Solving for $\hat{\beta}$

we can solve for $\hat{\beta}$ via the normal equations:

$$\begin{aligned}
\hat{\beta} =& (X^T X)^g X^T Y \\
=& \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1(p+1)} & X_{2(p+1)} & \cdots & X_{n(p+1)} \end{bmatrix} \begin{bmatrix} 1 & X_{12} & \cdots & X_{1(p+1)} \\ 1 & X_{22} & \cdots & X_{2(p+1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \cdots & X_{n(p+1)} \end{bmatrix} \right)^g \cdot \\
& \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1(p+1)} & X_{2(p+1)} & \cdots & X_{n(p+1)} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}
\end{aligned}$$

In our case, all predictors but Sleep Hours are significant predictors of exam scores, even at a 1% level of significance.

Estimability and BLUE

A necessary condition for the hypothesis to be testable is that $\mathbf{K}^T\beta$ is estimable. We say $\exists\ A$ s.t. $X^TA = K^T$, i.e. the rows of K are linearly dependent on the rows of X. We are now ready to state an important intermediate distributional result. Since $\mathbf{K}^T\beta$ is estimable, its best linear unbiased estimator (BLUE) is given by:

$$\mathbf{K_i}^T\hat{\beta} \sim N(\mathbf{K_i}^T(X^TX)^gX^TX\beta, \sigma^2\mathbf{K_i}^T(X^TX)^g\mathbf{K_i}) \quad \text{and}$$
$$\mathbf{K}^T\hat{\beta} \sim N(\mathbf{K}^T(X^TX)^gX^TX\beta, \sigma^2\mathbf{K}^T(X^TX)^g\mathbf{K})$$

Indeed, we can can test our hypothesis by constructing a quadratic form. While this is certainly not the only way to test our hypothesis, it is a tractable method to incorporate the precision of each $\hat{\beta}_i$ into our hypothesis testing framework.

Quadratic Form in our Joint Testing Procedure

$$(K\beta)^T(\sigma^2 H)^{-1}(K\hat{\beta}) \sim \chi^2_{\mathsf{df=rank}(H)}(\lambda)$$

where the non-centrality parameter $\lambda = \frac{1}{2}(K\beta)^T(\sigma^2 H)^{-1}(K\beta)$ by the well-known distributional result of a normal quadratic form. We are now ready to construct the F-test statistic as follows:

$$F := \frac{\left((K\beta)^T(\sigma^2 H)^{-1}(K\beta)\right)/\mathsf{rank}(H)}{\mathsf{RSS}/(n-p)} \sim \frac{\chi^2(\lambda)}{\chi^2} \sim F_{\mathsf{rank}(H),n-p}(\lambda)$$

We have successfully constructed a statistical test that allows us to test our hypothesis with a simple F-test. In `R`, we can use the `anova()` function to perform this test.

Results

Table 1: F-Test Results for the Hypothesis Test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
6606	20.696858	NA	NA	NA	NA
6601	7.477088	5	13.21977	2334.162	0

The result shows that under the null hypothesis, the probability of getting a more extremeresult than our calculate F-test statistics Pr$(> F)$ is $2.2e - 16$. This evidence would lead us to reject the null hypothesis and conclude that our subset of predictors is indeed a significantpredictor of exam scores

LASSO Regression

Non-Linear Model