Predicting Student Performance

Linear Models (PHP2601), Prof. Ani Eloyan

Daniel Posmik, Jizhou Tian, Aristofanis Rontogiannis

2024-12-09

Table of contents I

EDA and the Linear Model

Introduction

We will be analyzing educational data to understand the predictors of student performance. Specifically, we seek to **understand whether five predictors – as a subset of an exhaustive list of potential predictors – are significant predictors of student performance**.

Testing the significant of a subset of predictors is becoming increasingly important in modern statistical questions, especially with more information becoming available.

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

Hypothesis to be Tested

We are interested in:

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

We can formalize this question as follows:

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

Observe the 0-indexed variables from $p+2$ to P.

Exploratory Data Analysis (EDA)

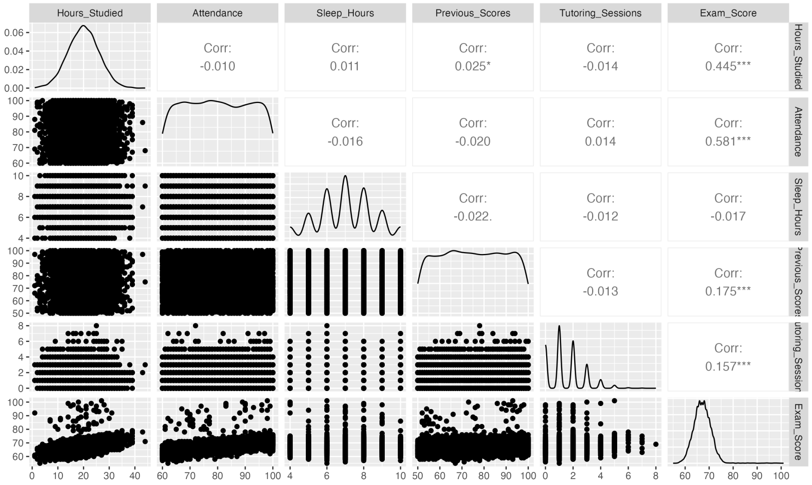

Figure 1: Correlation Matrix

Variable Transformations

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

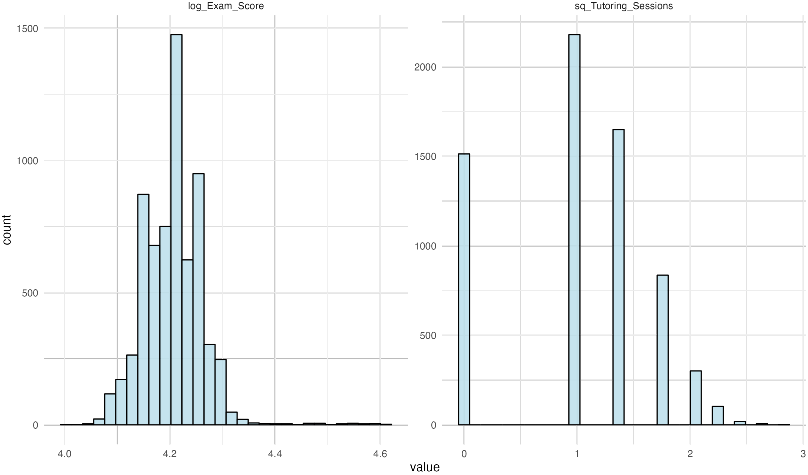

Figure 2: Variable Transformation

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 $\mathrm{Var}[\epsilon] = \sigma^2 I$ 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.

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 of the Hypothesis

Question: **Can we estimate an object $K^T\beta$ with our data X?**

Formally, we say that if $\exists\ A$ s.t. $X^TA = K^T$, i.e. K^T can be expressed as a linear combination of X and some matrix A, then $K^T\beta$ is estimable.

In our case, this is straightforward to verify. Can we think of an example when this is not true? (Hint: Dimension "mismatch")

Distribution of $K^T\beta$

Since $K^T\beta$ estimable, its best linear unbiased estimator (BLUE) is given by:

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

This object $K^T\beta$ may seem a bit arbitrary, even useless, at first. However, it is in fact the building block for the test statistic we will construct now!

Quadratic Form in our Joint Testing Procedure

Suppose $H := K(X^T X)^g K^T$, then

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

where the non-centrality parameter $\lambda = \frac{1}{2}(K\beta)^T(\sigma^2 H)^{-1}(K\beta)$ is the well-known distributional result of a normal quadratic form.

Finally, our F Statistic:

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

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 extreme result 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 significant predictor of exam scores

LASSO Regression

Non-Linear Model