Regression Lab

Evgueni Smirnov

September 22, 2023

Learning Goals

- **Statistical Modeling:** Gain a foundational understanding of statistical models, including linear and polynomial regression, and how they can be applied to analyze data.
- **Data Interpretation:** Learn to interpret model parameters in meaningful ways to make informed decisions or recommendations.
- **Data Visualization:** Acquire skills in visualizing data and model results effectively, using tools like scatterplots to support statistical reasoning.
- **Programming:** Develop hands-on programming skills in Python, particularly in using libraries like scikit-learn for statistical modeling.
- **Critical Thinking:** Foster critical thinking by evaluating model performance, understanding the limitations of models, and suggesting ways to improve them.
- **Ethical Considerations:** Cultivate an awareness of the ethical implications of statistical modeling and how to address them.

1 Linear Regression Model Analysis

In this assignment, your task is to investigate the existence of gender bias in the starting salaries of recent graduates. A dataset is given by the following input variables, each with its given range:

- GPA (range: 0 to 4)
- IQ (range: 70 to 130)
- Gender (coded as 1 for Female and 0 for Male)

Additionally, the dataset contains two interaction terms:

- Interaction between GPA and IQ
- Interaction between GPA and Gender

The output variable Y represents starting salary after graduation (in thousands of euros). A linear regression model has been fitted to the data with the following coefficients:

$$\begin{aligned}
&\hat{\beta}_0 = 50 \quad \text{(Intercept)}\\
&\hat{\beta}_1 = 20 \quad \text{(Coefficient for GPA)}\\
&\hat{\beta}_2 = 0.07 \quad \text{(Coefficient for IQ)}\\
&\hat{\beta}_3 = 10 \quad \text{(Coefficient for Gender)}\\
&\hat{\beta}_4 = 0.01 \quad \text{(Coefficient for Interaction between GPA and IQ)}\\
&\hat{\beta}_5 = -3 \quad \text{(Coefficient for Interaction between GPA and Gender)}
\end{aligned}$$

1.1 Gender Bias Analyis

Using the linear regression model given above, investigate the role of gender in determining the starting salary after graduation. Specifically, analyze whether the model indicates a gender bias in starting salaries. Explain your reasoning based on the coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, and $\hat{\beta}_5$.

1.2 Model Conversion

Convert the linear regression model given above into a model tree.

2 Data Generation and Model Fitting

2.1 First Data Generation

In this exercise, you are tasked with creating simulated data and fitting simple linear regression modesl to that data.To create the data use the `random` module from `numpy`. You can study the module on:

`https://numpy.org/doc/stable/reference/random/index.html`

Make sure to run function `random.seed(seed)` prior to data generating and model fitting to ensure consistent experimental results for a fixed value of the variable seed. Use value of 42 for `seed`.

Follow the steps below:

1. Using the `random.normal` function, create a vector x containing 100 observations drawn from a normal distribution with a mean of 0.0 and a variance of 1.

2. Using the `random.normal` function, create a vector *eps* containing 100 observations drawn from a normal distribution with a mean of 0.0 and a variance of 0.25.

3. Using x and *eps*, generate a vector y according to the model:

$$y = -0.5 + 0.75x + eps$$

a What is the length of the vector y?

b What are the values of β_0 and β_1 in this linear model?

2.2 First Data Visualization

Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

2.3 Fitting First Linear Regression

Fit a least squares linear model `LinearRegression()` from (module `linear_model` of `sklearn`) to predict y using x. Comment on the model obtained:

(a) How do the estimations of $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1 ?

(b) Display the least squares line on the scatterplot obtained in Subsection 2.2.

(c) Compute R^2 statistics (using function `r2_score` from the `sklearn.metrics` module).

2.4 Fitting Second Linear Regression

Now fit a polynomial regression model that predicts y using x and x^2. Comment on the model obtained:

(a) What is the estimated value for $\hat{\beta}_2$?

(b) How do the estimations of $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1 ?

(b) Display the least squares line on the scatterplot obtained in Subsection 2.2.

(d) Compute R^2 statistics.

(e) Is there evidence that the quadratic term improves the model fit? Explain your answer.

2.5 Second Data Generation

Using x and *eps*, generate a vector y according to the model:

$$y = -0.5 + 0.75x + x^2 + eps$$

2.6 Second Data Visualization

Create a new scatterplot displaying the relationship between x and y. Comment on what you observe.

2.7 Fitting Third Linear Regression

Fit a least squares linear model `LinearRegression()` from the module `linear_model` to predict y using x. Comment on the model obtained:

(a) How do the estimations of $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1 ?

(b) Display the least squares line on the scatterplot obtained in Subsection 2.6.

(c) Compute R^2 statistics.

2.8 Fitting Fourth Linear Regression

Now fit a polynomial regression model that predicts y using x and x^2. Comment on the model obtained:

(a) How do the estimations of $\hat{\beta}_0$, $\hat{\beta}_1$, amd $\hat{\beta}_2$ compare to β_0, β_1, and β_2 ?

(b) Display the least squares line on the scatterplot obtained in Subsection 2.6.

(c) Compute R^2 statistics.

(d) Is there evidence that the quadratic term improves the model fit? Explain your answer.

3 LASSO Regression Model Analysis

Repeat the asssignment from Section 2 by fitting Lasso regression models (use the function `Lasso()` from the module `linear_model`). What is the reason for the poor performance of the Lasso models? How can it be improved?

Submission Requirements

Please follow the guidelines below for submitting your assignments:

- **Analytical Report:** Submit a PDF file that serves as your Analytical Report. This should contain all your written answers, interpretations, and graphical visualizations for each exercise. They have to be labeled using the numbering system of the exercies. Ensure that the graphics are clearly labeled and appropriately integrated into your explanations.
- **Jupyter Notebook PDF:** Additionally, submit a PDF version of your Jupyter Notebook that contains all the code used for data generation, analysis, and visualization. Make sure that the code is well-commented for readability.
- **Self-Evaluation PDF:** Submit another PDF containing your self-evaluation of the Analytical Report and the Jupyter Notebook code. End this file with a summary of what you have learned from completing the assignments. The rubrics for self-evaluation are given in Appendices 1 and 2 (see below).

Note: Submitting all three files is essential for a complete submission. Failure to submit any of them will result in a deduction of points.

Academic Integrity Declaration

By submitting the Analytical Report, you are declaring that you have not used large language models or any other automated tools to generate written answers and interpretations on a semantic level and/or a language level. The work you submit must be your own. Failure to adhere to this academic integrity guideline will be considered a violation and may result in a grade penalty or other disciplinary actions.

Please include this declaration at the end of your Analytical Report PDF.

> I, [Your Name], hereby declare that I have not used large language models or any automated tools for generating the written answers and interpretations in this Analytical Report.

Appendix 1: Grading Rubrics for Analytical Report

3.1 Assignment 1: Linear Regression Model Analysis

- **Gender Bias Analyis**
  - Excellent understanding and interpretation: 20 points
  - Good understanding but minor errors: 12 points
  - Partial understanding with major errors: 5 points
  - Poor understanding or incomplete: 0 points
- **Model Conversion**
  - Correct conversion: 10 points
  - Partially correct conversion: 5 points
  - Incorrect conversion or no answer: 0 points

Total: 30 points

3.2 Assignment 2: Python Experience

- **Data Generation**
  - Correctly generated data: 15 points
  - Partially correct: 10 points
  - Incorrect or incomplete: 0 points
- **Scatterplot**
  - Accurate plot and insightful commentary: 10 points
  - Accurate plot but lacking commentary: 5 points
  - Incorrect or incomplete: 0 points
- **Linear Regression Model**
  - Accurate model fitting, insightful commentary, and R^2 computation: 20 points

  - Minor errors in model or commentary: 15 points
  - Major errors or incomplete: 5 points

- **Polynomial Regression Model**
  - Accurate model fitting, insightful commentary, and R^2 computation: 25 points
  - Minor errors in model or commentary: 20 points
  - Major errors or incomplete: 5 points

Total: 70 points

Overall Total Points: 100 points

Appendix 2: Code Evaluation Rubrics (0-100 points)

- Code organization: 10 points
- Proper commenting and documentation: 30 points
- Proper use of Python libraries and functions: 20 points
- Correctness of the implemented logic: 20 points
- Efficiency of code: 20 points