Statistical Analysis of the Relationship Between Website Page Size and Load Time

Avigyan Chakraborty (BS2516)

November, 2025

Abstract

This is where you'll write your abstract content later. You can replace this text with your actual abstract when you're ready. The abstract should provide a brief overview of your statistics assignment, including the main objectives, methods, and key findings.

You can write multiple paragraphs here. The formatting will handle the spacing and make it look professional and well-organized for your statistics assignment.

Contents

1 Introduction

Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance. Your content starts here. This section will introduce your statistical analysis project and its significance.

2 Data Description

The dataset used in this project was generated using an automated script which reads the list of websites, containing 82 entries, from `website_list.json` and makes five HTTP requests to each site and records two key metrics for each attempt:

- **Page load time (in seconds):** The total time required for the page to fully load under a web-browser instance.

- **Page size (in kilobytes):** The total size of loaded webpage including all its resources.

These raw measurements were then dumped into `website_load_data.json`. This data was then piped into the script `make_avg_csv.py` which calculates the average of five entries of each metric for each website and then writes the result into `averaged_data.csv` in a structured manner. The scripts and datasets used in this project are available in the following GitHub repository.

2.1 Summary Statistics

Variable	Mean	Median	Minimum	Maximum	Std. Dev.
Page Size (kb)	15427.26	6939.39	59.6	104757.7	22067.91
Load Time (s)	8.593405	5.755	0.3858	43.2654	7.929519

Table 1: Summary statistics of average page size and load time

3 Model Formulation

In order to have a better understanding at how the metrics might be related to each other, we make a scatter plot by plotting **load_time_avg** along Y-axis and **page_size_avg** along X-axis.
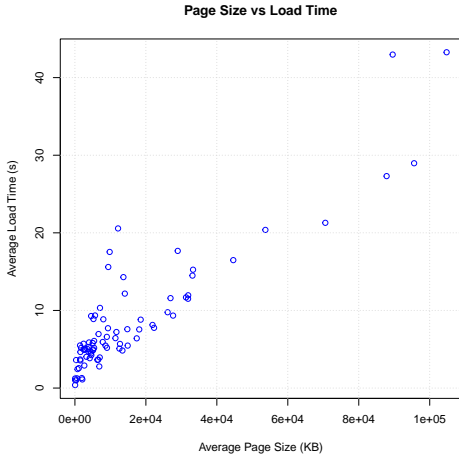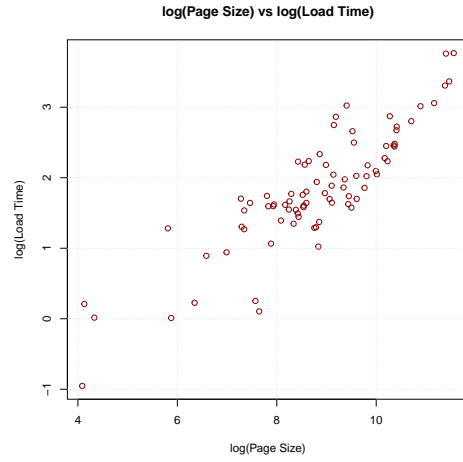

Figure 1: Load Time vs. Page Size

Figure 2: log(Load Time) vs. log(Page Size)

The relationship between the metrics is not very clear in Fig. 1. However, after applying a logarithmic transformation to both, the condition improves substantially as there's a vivid linear relationship between them as seen in Fig. 2.

Thus considering a power-law model of the form

$$\text{Load Time} = e^{\alpha} \cdot \text{Page Size}^{\beta} \cdot e^{\varepsilon}$$

would be suitable as this suggests

$$\log(\text{Load Time}) = \alpha + \beta \log(\text{Page Size}) + \varepsilon,$$

that is, a linear relationship between log(Load Time) and log(Page Size).

Let X denote the average page size (in kilobytes) and Y denote the average page load time (in seconds) for each website.

We will now estimate the parameters α and β using the method of least squares on the log-transformed data.

4 Regression Analysis

Fitting the linear regression model

$$\log(Y) = \alpha + \beta \log(x) + \varepsilon$$

to the data yields the following estimates:

$\hat{\alpha}$	$\hat{\beta}$	R^2
-2.371373	0.4780497	0.7581765

Table 2: Estimated Parameters of the Model

The fitted equation is therefore approximately

$$\log(\hat{Y}) = -2.371 + 0.478 \log(x),$$

with coefficient of determination $R^2 = 0.758$.

Variable	Mean	Median	Minimum	Maximum	Std. Dev.
log(Page Size)	8.776999	8.844928	4.087656	11.55941	1.531387
log(Load Time)	1.824468	1.750042	-0.9524362	3.767353	0.8407613

Table 3: Summary statistics of log(Page Size) and log(Load Time)

So, using values from Table 3 and Table 2, we get,

$$r = \frac{S_{x,Y}}{S_x S_y} = \hat{\beta} \cdot \frac{S_x}{S_Y} = 0.478 \times \frac{1.531}{0.84} = 0.87.$$

Therefore the sample correlation $r = 0.87$ indicates there is a strong positive linear association between the two variables.
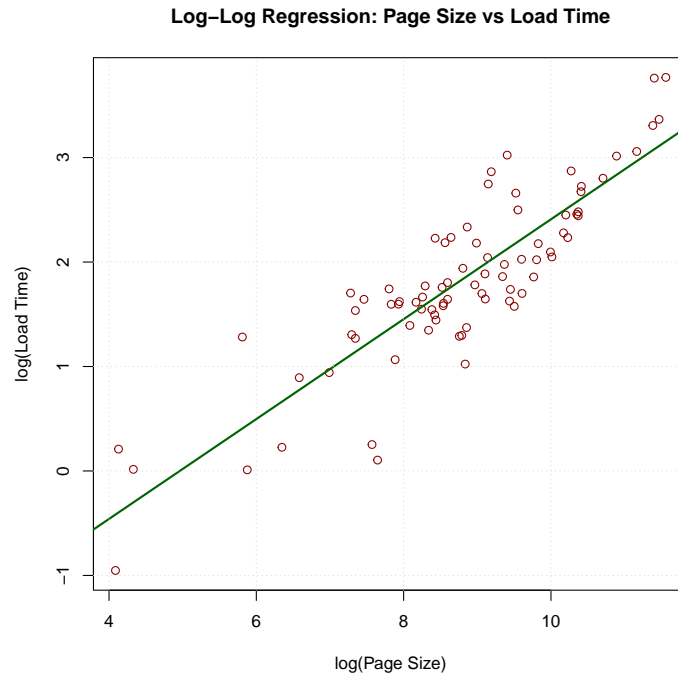
Figure 3: Scatter plot of log(Page Size) vs. log(Load Time) with fitted regression line

5   Interval Estimates