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.

2 Data Description

The dataset used in this project was generated using an automated Python script that reads a list of websites containing 82 entries from `website_list.json`. The script 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 stored in `website_load_data.json`. This current data was then processed by the script `make_avg_csv.py`, which computes the average of the five measurements of each metric for every website and writes the summarized results into `averaged_data.csv` in a structured format. The scripts and datasets used in this project are available in the following GitHub repository.

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 of 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.

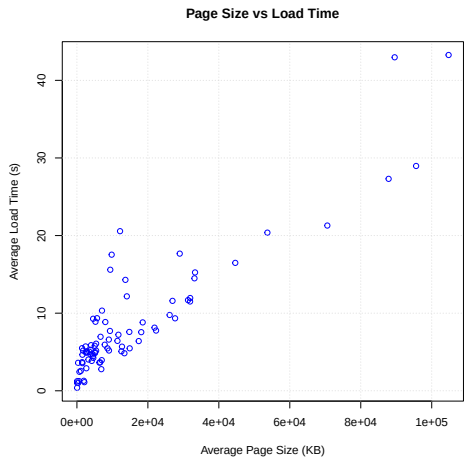

Figure 1: Load Time vs. Page Size

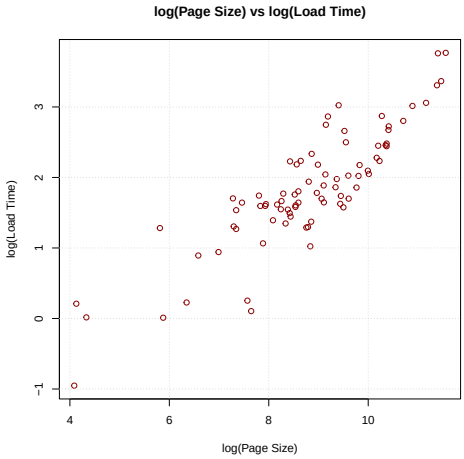

Figure 2: log(Load Time) vs. log(Page Size)

The relationship between the metrics is not very clear in Fig. 1. However, after applying a logarithmic transformation to both variables, the relationship becomes substantially more linear, as shown 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

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

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,

$$r = \frac{S_{\log(x),\log(Y)}}{S_{\log(x)} S_{\log(Y)}} = \hat{\beta} \cdot \frac{S_{\log(x)}}{S_{\log(Y)}} = 0.478 \times \frac{1.531}{0.84} = 0.87.$$

Therefore, the sample correlation $r = 0.87$ indicates a strong positive linear association between the two variables.

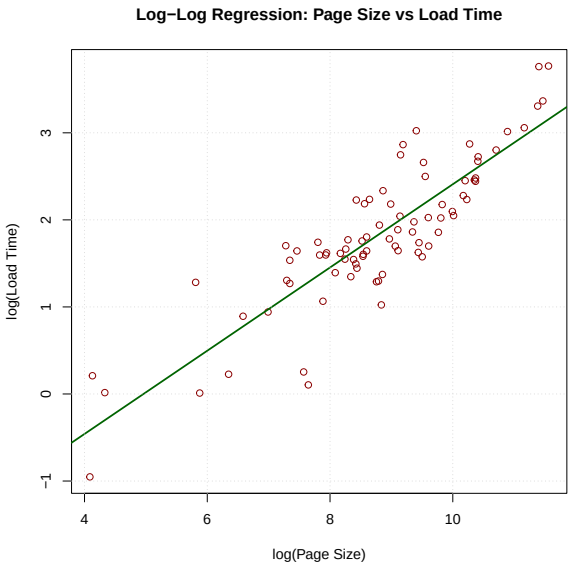

Figure 3: Scatter plot of log(Page Size) vs. log(Load Time) with fitted regression line

5 Interval Estimates

We now check how the residuals, $\hat{\varepsilon}_i = \log(y_i) - \widehat{\log(y_i)}$, behave where $\log(y_i)$ represents a realized value of log(Load Time) and $\log(\hat{y}_i)$ is the fitted value corresponding to that observation.

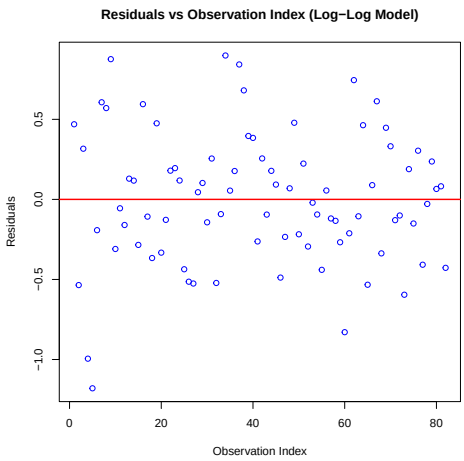

Figure 4: Plot of residuals versus fitted values

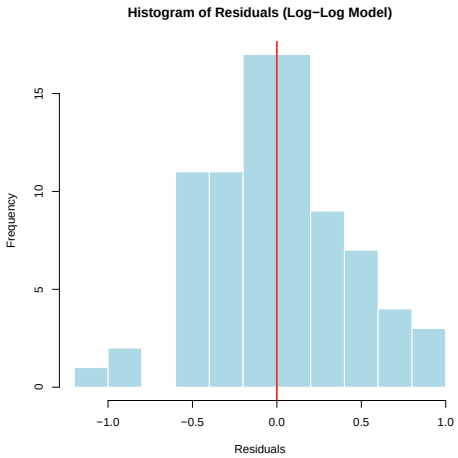

Figure 5: Histogram of residuals for the log–log regression model

From Fig. 5, it can be deduced that the distribution of residuals roughly follows a normal distribution with mean 0. It is therefore reasonable to assume that $\varepsilon \sim N(0, \sigma^2)$. This helps us to estimate confidence and prediction intervals using quantiles of t-distribution.

The formulas for the confidence intervals are as follows:

Parameter	Confidence Interval
α	$\hat{\alpha} \pm t_{\gamma/2}(n-2)s\sqrt{\frac{1}{n} + \frac{\overline{\log(x)}^2}{S_{\log(x)\log(x)}}}$
β	$\hat{\beta} \pm t_{\gamma/2}(n-2)\frac{s}{\sqrt{S_{\log(x)\log(x)}}}$

Table 4: $100(1-\gamma)\%$ confidence intervals for model parameters

We now list the formulas for the $100(1-\gamma)\%$ confidence and prediction intervals for $\mathbb{E}[\log(Y)]$ and $\log(Y)$, respectively, given a new value of $X = x$.

Quantity	Confidence Interval
$\mathbb{E}[\log(Y)]$	$(\hat{\alpha} + \hat{\beta}\log(x)) \pm t_{\gamma/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(\log(x)-\overline{\log(x)})^2}{S_{\log(x)\log(x)}}}$

Table 5: $100(1-\gamma)\%$ confidence interval for $E[\log(Y)]$ given a new value of $X = x$

Quantity	Prediction Interval
$\log(Y)$	$(\hat{\alpha} + \hat{\beta}\log(x)) \pm t_{\gamma/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(\log(x)-\overline{\log(x)})^2}{S_{\log(x)\log(x)}}}$

Table 6: $100(1-\gamma)\%$ prediction interval for $\log(Y)$ given a new value of $X = x$

Substituting dataset values into these formulas gives the realized upper and lower bounds of the 95% confidence and prediction intervals.

Parameter	Lower Limit	Upper Limit
α	-2.906478	-1.836269
β	0.4179795	0.5381199

Table 7: 95% confidence intervals for the regression parameters

Quantity	Confidence Interval
$\mathbb{E}[\log(Y)]$	$(-2.3714 + 0.4780 \cdot \log(x)) \pm 0.8279 \cdot \sqrt{0.0122 + \frac{(\log(x)-8.7770)^2}{189.9567}}$

Table 8: 95% confidence interval for $\mathbb{E}[\log(Y)]$ given a new value of $X = x$

Quantity	Prediction Interval
$\log(Y)$	$(-2.3714 + 0.4780 \cdot \log(x)) \pm 0.8279 \cdot \sqrt{1.0122 + \frac{(\log(x)-8.7770)^2}{189.9567}}$

Table 9: 95% prediction interval for $\log(Y)$ given a new value of $X = x$

With this, we conclude our interval estimate analysis.