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

Avigyan Chakraborty (BS2516)

November, 2025

Abstract

This study investigates the relationship between website page size and load time using data collected from the official websites of Indian colleges. For each website, multiple measurements of page size and load time were taken under identical network conditions to minimize random variation due to bandwidth fluctuations. Exploratory analysis revealed a positive but non-linear association between the two variables. After applying logarithmic transformations to both, the relationship became approximately linear, motivating the use of a power-law (log–log) regression model.

The fitted model explained about 75.8% of the variability in load time and showed a strong positive correlation ($r = 0.87$). Residual analysis indicated approximate normality, supporting the validity of inference based on t-distribution. When tested on new websites, all observed values fell within their respective 95% prediction intervals, confirming the model's reliability within its domain. The findings highlight that while page size is a key determinant of load time for minimalistic websites, other factors such as caching, server configuration, and use of CDNs play a major role in more complex, resource-intensive sites.

1 Introduction

In the modern web ecosystem, the loading speed of a website plays a crucial role in determining user experience and overall accessibility. Websites that take longer to load often experience higher bounce rates and lower engagement, making load time an important performance metric. Among the various factors influencing how fast a webpage loads, the total size of the page and its associated resources is one of the most significant.

The objective of this study is to examine the relationship between the size of a webpage and its corresponding load time. To achieve this, data were collected from the official websites of various Indian colleges. The data include multiple load-time measurements for each site, allowing us to eliminate the noise due to network fluctuations and derive a stable estimate of load duration and total page size by considering the average.

An initial visual inspection of the raw data suggested a positive but non-linear association between the two variables. After applying logarithmic transformations to both metrics, the relationship became approximately linear, motivating the use of a log–log regression model. The subsequent analysis explores this relationship in detail, estimates the model parameters, and assesses the fit using confidence and prediction intervals.

2 Data Description

The dataset used in this project was generated using an automated Python script that reads a list of 82 websites containing 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 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 to `averaged_data.csv` in a structured format. Considering the average helps us to eliminate the noises created due to network fluctuations. 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

To gain a better understanding of how the metrics might be related, we make a scatter plot by plotting **load_time_avg** along the Y-axis and **page_size_avg** along the X-axis.

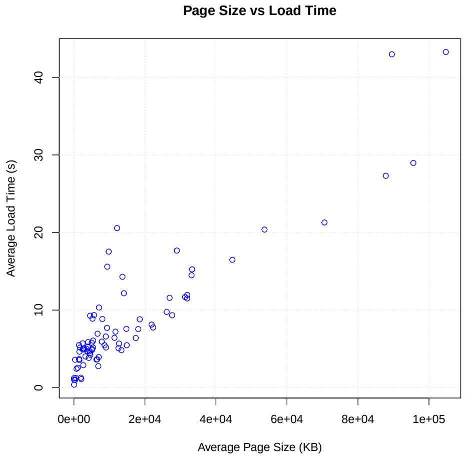

Figure 1: Load Time vs. Page Size.

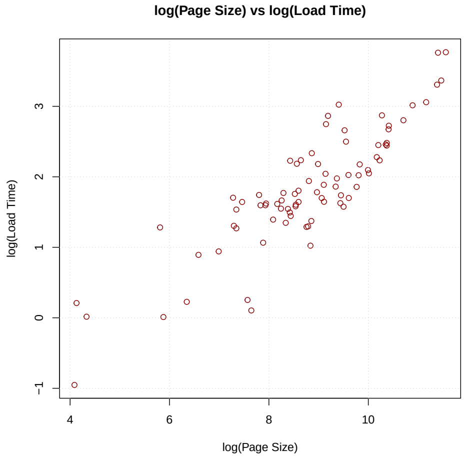

Figure 2: log(Load Time) vs. log(Page Size).

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 now estimate the parameters $\hat{\alpha}$ and $\hat{\beta}$ 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.

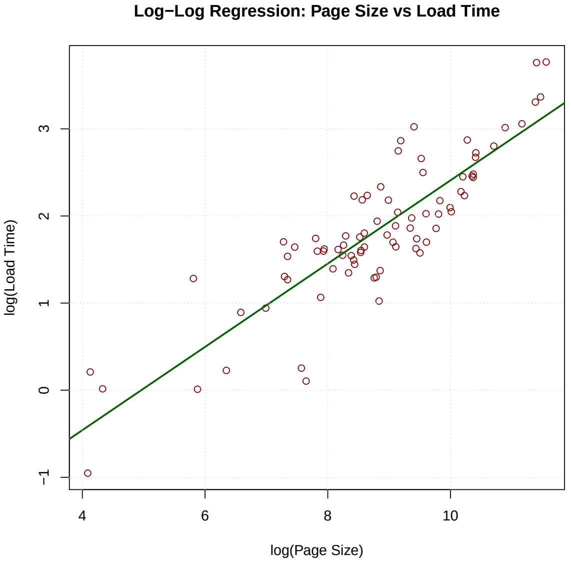

Figure 3: Scatter plot of log(Page Size) vs. log(Load Time) with fitted regression line.

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 $\widehat{\log(y_i)}$ is the fitted value corresponding to that observation.

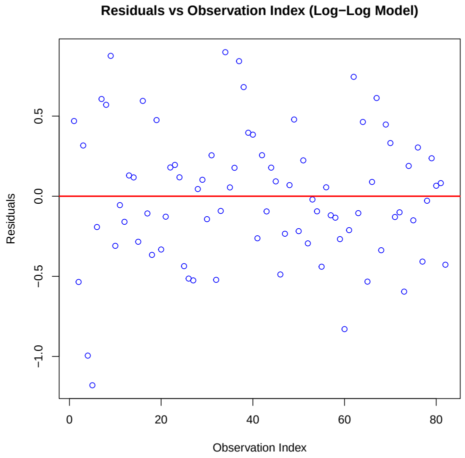

Figure 4: Plot of residuals vs. Observation index.

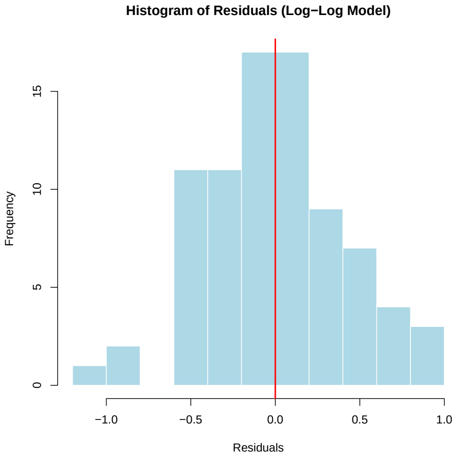

Figure 5: Histogram of residuals for the log–log regression model.

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 $X = x$.

We now put the prediction interval to test using a few new websites as test cases.

Website	**Prediction Interval for** $\log(Y)$	**Actual** $\log(y)$	**Inside interval?**
Amazon	$(0.1462, 1.8253)$	0.432	Yes
Project Euler	$(-1.0737, 0.6683)$	0.377	Yes
Suckless	$(-0.9869, 0.7490)$	0.586	Yes

Table 10: Comparing bounds of prediction interval with actual value of log(Load Time).

6 Conclusion

The analysis demonstrates a strong positive relationship between average web page size and average load time. After applying logarithmic transformations to both variables, the fitted model

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

explained about 75.8% of the variation in the log-transformed load time. The estimated correlation coefficient of $r = 0.87$ further confirms a strong linear association between these variables.

Residual analysis showed that the error terms were approximately normally distributed, validating the use of t-based confidence and prediction intervals. When tested on new websites, all observed log(Load Time) values fell within their respective 95% prediction intervals, indicating that the model performs reliably within its observed domain.

It should be noted that all data were collected from websites of Indian colleges, which are generally minimalistic and have relatively small page sizes compared to large commercial or media websites. Furthermore, several external factors influencing page load time, such as, the use of Content Delivery Networks (CDNs), server response time, caching mechanisms, client-side rendering, and network routing were not modeled explicitly. However, since all measurements were made using the same network and under similar conditions, much of the variation due to network speed was effectively smoothed out. As a result, the regression primarily captures the structural relationship between page size and load time for lightweight, static websites. Nevertheless, the model may not generalize well to resource-intensive, dynamically rendered websites that depend heavily on JavaScript execution, third-party assets, or geographically distributed servers.

A Appendix

A.1 Dataset

Table 11: Raw Average Data for Websites.

Website	Load Time (s)	Page Size (KB)
Indian Statistical Institute, Kolkata	10.3276	7066.44
Indian Statistical Institute, Delhi	0.3858	59.6
Indian Statistical Institute, Bangalore	1.0156	76.0
Indian Statistical Institute, Chennai	1.2878	1940.2
Indian Statistical Institute, Pune	1.1098	2093.1
Indian Statistical Institute, Tezpur	7.584	14778.2
Chennai Mathematical Institute	1.2328	62.14
Tata Institute of Fundamental Research	9.2808	4573.1
TIFR Centre for Applicable Mathematics	3.6046	333.84
TIFR Centre for Interdisciplinary Sciences	8.1344	21874.06
International Centre for Theoretical Sciences	11.5872	26943.2
Harish-Chandra Research Institute	4.4566	4541.56
Institute of Mathematical Sciences	3.5582	1548.02
Raman Research Institute	2.4424	722.84
Physical Research Laboratory	9.3362	27668.32
Inter-University Centre for Astronomy and Astrophysics	5.4912	1449.8
S.N. Bose National Centre for Basic Sciences	4.975	5124.2

Table 11 – continued from previous page

Website	**Load Time (s)**	**Page Size (KB)**
Institute of Physics, Bhubaneswar	7.7596	22306.3
Indian Institute of Science	9.3522	5671.52
IISER Pune	2.9018	2656.14
IISER Kolkata	4.8636	5098.6
IISER Bhopal	20.3772	53687.0
IISER Tirupati	5.0602	2817.76
IISER Berhampur	15.249	33293.1
IIT Bombay	1.2544	571.04
IIT Delhi	5.0844	12550.1
IIT Madras	3.6296	6350.46
IIT Kanpur	7.7106	9316.92
IIT Kharagpur	6.9596	6664.0
IIT Guwahati	11.506	31890.84
IIT Roorkee	28.965	95601.9
IIT Hyderabad	5.4708	14873.24
IIT Indore	4.6852	4374.22
IIT BHU	20.5672	12158.16
IIT Gandhinagar	5.7974	5024.42
IIT Patna	4.9366	2779.06
IIT Mandi	17.538	9794.48
IIT Bhubaneswar	42.9686	89554.08
IIT Tirupati	4.645	1548.1
IIT Palakkad	5.7126	2446.7
IIT Dhanbad (ISM)	5.466	8630.1
IIT Dharwad	8.859	8012.36
NIT Trichy	5.1724	5413.0
NIT Surathkal	5.8766	3992.5
NIT Calicut	21.2912	70576.94
NIT Durgapur	3.9468	7002.16
NIT Silchar	6.43	11423.16
NIT Meghalaya	14.49	33129.52
NIT Agartala	14.2918	13669.86
NIT Raipur	4.2412	4623.38
NIT Kurukshetra	4.9326	2517.9

Table 11 – continued from previous page

Website	Load Time (s)	Page Size (KB)
NIT Srinagar	7.5526	18134.58
NIT Arunachal Pradesh	4.705	3800.4
NIT Nagaland	6.5974	8995.5
NIT Sikkim	6.4048	17421.0
NIT Goa	16.4818	44634.9
NIT Puducherry	11.6826	31309.92
Delhi University	5.9378	7839.2
BITS Pilani	3.8468	4183.3
Jawaharlal Nehru University	2.7822	6876.62
Banaras Hindu University	9.7682	26132.32
University of Calcutta	15.5968	9403.62
Indian Institute of Engineering Science and Technology, Shibpur	11.9522	31924.2
Anna University	8.8914	5229.74
Savitribai Phule Pune University	3.6592	6562.66
Aligarh Muslim University	5.2848	3853.84
University of Madras	43.2654	104757.7
Panjab University	5.1862	9035.3
Central University of Rajasthan	5.1708	1740.1
Maulana Azad National Institute of Technology (MANIT)	17.672	28977.98
Visvesvaraya National Institute of Technology (VNIT)	7.2212	11702.72
Goa University	4.027	3244.68
Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR)	4.8342	13388.84
Saha Institute of Nuclear Physics (SINP)	3.6876	1473.14
National Centre for Biological Sciences (NCBS)	8.8122	18537.66
Institute of Plasma Research (IPR)	12.1688	14067.02
Bhabha Atomic Research Centre (BARC) Training School	5.6862	12703.36

Table 11 – continued from previous page

Website	Load Time (s)	Page Size (KB)
Institute of Genomics and Integrative Biology (IGIB)	2.5644	1085.5
Indira Gandhi Centre for Atomic Research (IGCAR) Training School	27.3082	87850.6
Vellore Institute of Technology (VIT)	6.0696	5411.14
Jamia Millia Islamia	5.0288	3526.96
Visva-Bharati University	1.0108	356.8

A.2 R Codes

A.2.1 Code for printing summary table

```r
data <- read.csv("../../data/averaged_data.csv")

model <- lm(log(load_time_avg) ~ log(page_size_avg), data = data)

cat("Summary statistics for Page Size (KB):\n")
cat("Mean:", mean(data$page_size_avg), "\n")
cat("Median:", median(data$page_size_avg), "\n")
cat("Minimum:", min(data$page_size_avg), "\n")
cat("Maximum:", max(data$page_size_avg), "\n")
cat("Standard deviation:", sd(data$page_size_avg), "\n\n")

cat("Summary statistics for Load Time (s):\n")
cat("Mean:", mean(data$load_time_avg), "\n")
cat("Median:", median(data$load_time_avg), "\n")
cat("Minimum:", min(data$load_time_avg), "\n")
cat("Maximum:", max(data$load_time_avg), "\n")
cat("Standard deviation:", sd(data$load_time_avg), "\n\n")

cat("Summary statistics for log(Page Size):\n")
cat("Mean:", mean(log(data$page_size_avg)), "\n")
cat("Median:", median(log(data$page_size_avg)), "\n")
cat("Minimum:", min(log(data$page_size_avg)), "\n")
cat("Maximum:", max(log(data$page_size_avg)), "\n")
cat("Standard deviation:", sd(log(data$page_size_avg)), "\n\n")

cat("Summary statistics for log(Load Time):\n")
cat("Mean:", mean(log(data$load_time_avg)), "\n")
cat("Median:", median(log(data$load_time_avg)), "\n")
cat("Minimum:", min(log(data$load_time_avg)), "\n")
cat("Maximum:", max(log(data$load_time_avg)), "\n")
cat("Standard deviation:", sd(log(data$load_time_avg)), "\n\n")
```

A.2.2 Code for making the Load Time vs. Page Size bivariate plot

```r
data <- read.csv("../../data/averaged_data.csv")

plot(data$page_size_avg, data$load_time_avg,
  main = "Page Size vs Load Time",
  xlab = "Average Page Size (KB)",
  ylab = "Average Load Time (s)",
  col = "blue")

grid()
```

A.2.3 Code for making the log(Load Time) vs. log(Page Size) bivariate plot

```r
data <- read.csv("../../data/averaged_data.csv")

plot(log(data$page_size_avg), log(data$load_time_avg),
  main = "log(Page Size) vs log(Load Time)",
  xlab = "log(Page Size)",
  ylab = "log(Load Time)",
  col = "darkred")

grid()
```

A.2.4 Code for fitting the linear regression

```r
data <- read.csv("../../data/averaged_data.csv")

model <- lm(log(load_time_avg) ~ log(page_size_avg), data = data)

summary(model)

alpha_hat <- coef(model)[1]
beta_hat  <- coef(model)[2]
r_squared <- summary(model)$r.squared

cat("Value of alpha_hat:", alpha_hat, "\n")
cat("Value of beta_hat:", beta_hat, "\n")
cat("Value of R^2:", r_squared, "\n")

plot(log(data$page_size_avg), log(data$load_time_avg),
     main = "Log-Log Regression: Page Size vs Load Time",
     xlab = "log(Page Size)",
     ylab = "log(Load Time)",
     col = "darkred")

abline(model, col = "darkgreen", lwd = 2)

grid()
```

A.2.5 Code for plotting residuals

```r
data <- read.csv("../../data/averaged_data.csv")

model <- lm(log(load_time_avg) ~ log(page_size_avg), data = data)

y_hat <- fitted(model)
residuals <- log(data$load_time_avg) - y_hat
serial_no <- 1:length(residuals)

plot(serial_no, residuals,
     main = "Residuals vs Observation Index (Log-Log Model)",
     xlab = "Observation Index",
     ylab = "Residuals",
     col = "blue")

abline(h = 0, col = "red", lwd = 2)
```

A.2.6 Code for making a histogram for the residuals

```r
data <- read.csv("../../data/averaged_data.csv")

model <- lm(log(load_time_avg) ~ log(page_size_avg), data = data)

y_hat <- fitted(model)
residuals <- log(data$load_time_avg) - y_hat

hist(residuals,
     main = "Histogram of Residuals (Log-Log Model)",
     xlab = "Residuals",
     col = "lightblue",
    border = "white")

abline(v = 0, col = "red", lwd = 2)
```

A.3 Python Codes

The Python scripts can be found at the following GitHub repository.