

Relationship between Daily Mean Gage Height and Discharge on Boneyard Creek

Jeff Gao, Sean Li, Ava Rizzo, Isaac Gross, Danny Monahan

Date Submitted: December 10, 2025

Abstract: This brief study examines the relationship between daily mean gage height (ft) and river discharge (cubic feet/second) measured at Boneyard Creek. Using daily observations from the provided dataset (U.S. Geological Survey, 2022), a simple linear regression model is used with discharge as the independent variable and gage height as the dependent variable. Descriptive statistics, the correlation coefficient, and confidence intervals for the linear regression model are analyzed and reported. Results indicate the lack of a linear relationship between the discharge data and gage height, leading to a failure of rejecting the null hypothesis that no linear relationship exists between the two measured variables.

Introduction:

Monitoring stream gage height and discharge is essential for flood forecasting, watershed management, and infrastructure planning (U.S. Geological Survey, 2021). Since gage height often correlates with discharge through channel geometry and roughness, accurately modeling this relationship can support decision-making during storm events and aid long-term stream management. In this report, the question asked is: How well can daily mean gage height be predicted from daily discharge measurements for Boneyard Creek? Understanding this relationship provides a site-specific tool for quick estimation of water level from discharge measurements or vice versa. This report develops a simple linear regression model that relates daily mean gage height to daily discharge. The project guidelines are followed to present data description, model fitting, hypothesis testing, diagnostics, interpretation, and limitations.

Data and Methods: The data was gathered from the file
6-daily_Boneyard_discharge_and_gage_height.c

sv (U.S. Geological Survey, 2022). The two variables used in this study are:

- discharge: daily average discharge (cubic feet/second)
- gage_height: daily mean gage height (feet)

The dataset contains 3 distinct groups of gage height values, therefore, values that were not in between 5-12.5 were discarded to enable a more accurate linear regression model. The null hypothesis is that there is no linear relationship between gage height and daily discharge, with the alternate hypothesis being that a linear relationship does exist. The simple linear regression is built with the equation

$$gageHeight = \beta_0 + \beta_1 \times discharge + \varepsilon$$

where β_0 is the intercept, β_1 is the slope, and ε are random errors with mean 0 and constant variance. Here, gage height is the dependent variable, and discharge is the independent variable. A scatterplot with a 95% confidence interval (Figure 10) was also produced, showing the fitted regression line, the upper and lower bounds of the 95%

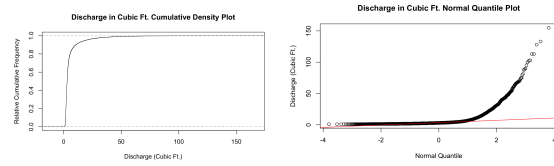
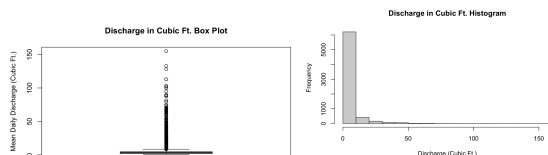
confidence interval, as well as 4 diagnostic plots (Figures 12-15) to verify the accuracy of the linear model.

Results and Discussion: Part 1 of this project analyzed the mean, median, variance, standard deviation, range, IQR, skewness, and kurtosis of each variable (Figure 1). Those descriptive statistics are included in the figure below for reference.

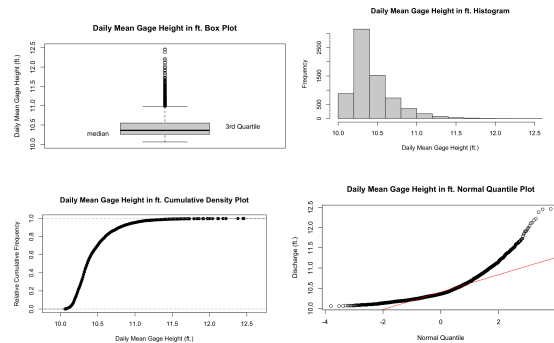
Statistic	Discharge_cfs	Gage Height_ft
Mean	5.483	10.440
Median	3.000	10.360
Variance	76.915	0.071
Standard Deviation	8.770	0.267
Range	154.200	2.400
Interquartile Range(IQR)	2.530	0.290
Skewness	6.074	1.875
Kurtosis	57.867	8.421

Fig 1. Summary table of data measurements

Part 2 examined graphical representations of the data in the form of box plots, histograms, CDFs, and normal quantile plots for discharge and daily mean gage height (Figures 2-9) (Harvard Chan Bioinformatics Core, R Documentation). Those plots are also included in this report for reference to demonstrate the continuity of our project. All plots have been updated as we modified the range of our dataset for Part 3.



Figs 2-5. Box plot, Histogram, CDF, Normal Quartile plot for daily discharge



Figs 6-9. Box plot, Histogram, CDF, Normal Quartile plot for gage height

Part 3 modeled the relationship between discharge and gage height using simple linear regression (Figures 10-14). The data is limited to gage height between 5-12.5ft, leading to a more consistent linear model. Simple linear regression was used to determine whether the two variables are related, which resulted in the equation

$$gageHeight = 10.31 + 0.02442 \times discharge + \epsilon$$

Where the slope represents a predicted 0.024 ft increase with every increase in discharge by 1 cubic ft/s, and the intercept represents the predicted gage height value is 10.31ft when discharge is at 0 cubic ft/s (University of Illinois Urbana-Champaign). The output from the simple linear regression model (Figure 10) displays an R^2 value of 0.645, which suggests that around 64.5% of the variations in Gage height is correlated to variations in discharge volume and that some sort of linear relationship exists between the two variables.

However, the scatterplot with fitted trend line and a 95% confidence interval (Figure 11) suggests that the linear relationship between the two variables is unclear, with numerous outliers throughout the plot.

Residuals:				
Min	1Q	Median	3Q	Max
-1.63091	-0.10819	-0.02686	0.08522	0.79273
Coefficients:				
Estimate	Std. Error	t value	Pr(> t)	
Intercept	1.031E+01	2.248E-03	4584.1	<2e-16
discharge	2.442E-02	2.174E-04	112.3	<2e-16
Residual standard error: 0.1589 on 6942 degrees of freedom				
Multiple R-squared: 0.6451		Adjusted R-squared: 0.645		
F-statistic: 1.262e+04 on 1 and 6942 DF		p-value: <2.2e-16		

Fig 10. SLR output in R

The diagnostic plots (Figures 12-15) generated by the linear model further confirm the uncertainty of a linear relationship. The Residuals vs fitted graph (Figure 12) displays clear deviations from the zero line- with a sharp increase from negative to positive and then a gradual decline into the negative region. The lack of consistency about the horizontal line at 0 displays a lack of linear relationship between the observed and fitted data. The Q-Q plot (Figure 13) shows that the dataset deviates from a normal distribution on both tails. The scale-location plot (Figure 14) shows a heavy concentration of data points along the fitted line, suggesting that the spread/variance of a large portion of the dataset is dependent on the fitted line while labeling points 1712, 1593, and 1815 as influential outliers. Lastly, the Residuals vs.

Conclusions: In conclusion, the data shows a tightly clustered IQR, indicating that the central 50% of observations are concentrated within a relatively narrow range and that variability among the majority of values is low. The diagnostic plots suggest that the data

Leverage graph (Figure 15) labels the same points as far beyond the dashed line representing Cook's Distance, suggesting that they are influential points. Overall, the scatterplot and the four diagnostic plots of the linear regression model suggest that there is a lack of linear relationship between the two variables, meaning the data fails to reject the null hypothesis.

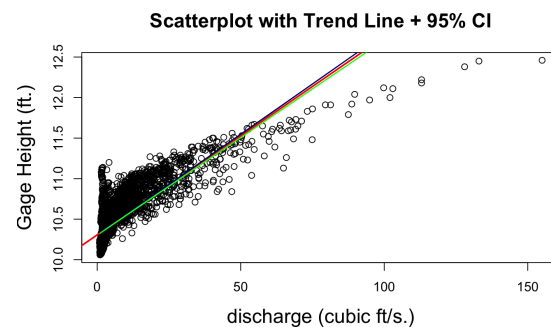
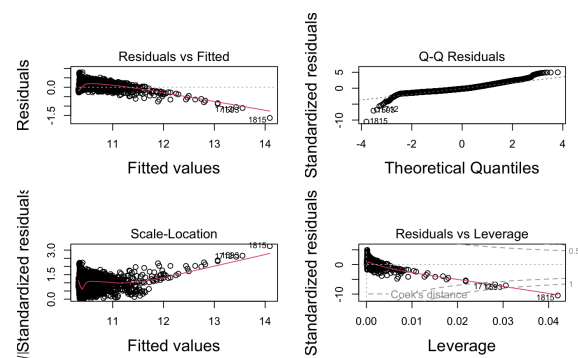


Fig 11. Scatterplot with 95% confidence interval



Figs 12-15. Diagnostic plots for linear regression

is approximately normally distributed, with heavy outliers towards the left and right tails. The residuals vs. fitted plot is the main evidence for the uncertainty of the linear relationship. Ultimately, the data's inconclusive results fail to reject the null hypothesis of a lack of linear relationship.

References:

Harvard Chan Bioinformatics Core. (n.d.). Plotting and data visualization in R (basics). Introduction to R. https://hbctraining.github.io/Intro-to-R/lessons/basic_plots_in_r.html

R Documentation (n.d.). : R: Histograms, Stat.ethz.ch. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/hist.html>

U.S. Geological Survey. (2022). *National water information system: Discharge and gage height data for Boneyard Creek* (Station No. 03337000) [Data set]. Retrieved October 2022. https://waterdata.usgs.gov/nwis/inventory?agency_code=USGS&site_no=03337000

U.S. Geological Survey. (2021). *Why we use gage height*. Water Data For The Nation Blog. https://waterdata.usgs.gov/blog/gage_height/

University of Illinois Urbana-Champaign. (n.d.). *Simple linear regression*. Data Science Discovery. <https://discovery.cs.illinois.edu/learn/Towards-Machine-Learning/Linear-Regression/>

Yair Daon. (2013). Complete.obs of cor() function. Stack Overflow. <https://stackoverflow.com/questions/18892051/complete-obs-of-cor-function>

Contribution Statement:

- Gao, Jeff, Data Engineer, Prepared filtered CSV file and wrote R code to create graphs and summaries in accordance with rubric.
- Li, Sean, Data Visualizer, Modified R code to plot the confidence intervals for the SLR model, conducted analysis based on the 4 analysis plots, proofread and edited the final version of the paper.
- Rizzo, Ava, Report Writer, Structured and wrote the analysis paper, confirmed results with R code, analyzed variables and plots
- Monahan, Danny, Editor, helped with report, helped with R code to create graphs, helped with analysis paper
- Gross, Isaac, Filled in gaps throughout the document, confirmed results with R code, compared paper with rubric in canvas, translated findings from R into the paper