The Illusion of Predictability: Simple Linear Regression Analysis of California Water Right Allocations*

An Assessment of Appropriative Status on Allocated Water Volume

Lasya Ramachandruni

September 24, 2025

This study employs Simple Linear Regression (SLR) to analyze the California Water Rights List and assess whether legal classification as an Appropriative right is a significant predictor of allocated water volume. Our findings reveal a highly statistically significant difference ($p < 2 \times 10 - 16$ p<2×10 −16): holders of Appropriative rights receive, on average, approximately 17,881 acre-feet per year more than holders of other types of water rights. However, thorough model diagnostics uncovered severe violations of core SLR assumptions, namely heteroscedasticity and non-normality, compounded by the presence of highly influential outliers. These issues resulted in a near-zero R-squared value, despite the statistical significance of the group mean comparison. Consequently, while the SLR model robustly detects differences in mean allocations between groups, the resulting prediction intervals—spanning over ± 400,000 ±400,000 acre-feet per year—are so wide that the model cannot be used for any reliable prediction of individual water allocations in practice.

Introduction

The distribution and management of water rights represent a critical challenge in California, impacting environmental sustainability and economic activity. This study analyzes the quantitative relationship between legal right type and the volume of water allocated. Drawing on the California Water Rights List, our research addresses a core question: **Does a water right designated as "Appropriative" predict a larger mean allocated water volume (Y) compared to other types of rights (X)?** This study employs Simple Linear Regression

*Project repository available at: https://github.com/LasyaRamachandruni/MATH261A-project-template/.

(SLR) to analyze the California Water Rights List ("California Water Rights List" 2025), using R (R Core Team 2025), RStudio (RStudio Team 2025), and the tidyverse (Wickham 2023) and car (Fox and Weisberg 2023) packages. Diagnostics and methodology closely follow Fox and Weisberg (Fox and Weisberg 2019) and Kutner et al. (Kutner, Nachtsheim, and Neter 2005). We employ Simple Linear Regression (SLR) with a single binary predictor derived from the categorical right type. This provides a direct test of **mean group differences**.

The analysis establishes a highly statistically significant difference in mean allocations, confirming that Appropriative rights are, on average, associated with larger volumes. However, a transparent and reproducible diagnostic review shows **severe violations** of key assumptions. Consequently, while the model reliably compares **group means**, it is **unsuitable for predicting** individual allocations. The sections that follow outline the data (Data), the model and checks (Methods), the quantitative findings and figures (Results), and sources (References).

Data

Source:
The raw data is provided by the California Water Rights List ("California Water Rights List" 2025) via the California Open Data Portal ("California Open Data Portal" 2025). Data wrangling and cleaning were performed in R (R Core Team 2025) with the tidyverse (Wickham 2023). Each observational unit corresponds to a single, official water right record registered within the state.

Variables:
- **Outcome (Y):** The dependent variable is `face_value_amount`, representing the annual volume of water allocated to each right, measured in acre-feet per year. This variable is quantitative and highly right-skewed, with a handful of extremely large values (extreme outliers).
- **Predictor (X):** The main independent variable is `is_approp`, a binary indicator defined as 1 if a record is classified as "Appropriative", and 0 if it represents any other kind of legal right. This was constructed based on the categorical variable `water_right_type` as reported in the source data.

Data Cleaning & Preprocessing:
Prior to any statistical analysis, the raw dataset was rigorously preprocessed. All records containing missing or undefined values for either `face_value_amount` or `water_right_type` were excluded, ensuring that only complete cases entered the modeling stage. The categorical variable `water_right_type` was then converted into a binary indicator (`is_approp`) that flags Appropriative rights with a value of 1, and all other types (e.g., Riparian, Statement) with 0. No further transformations were performed at this stage, as the initial purpose was to conduct a baseline group mean comparison before considering additional remedies such as log transformation or robust regression.

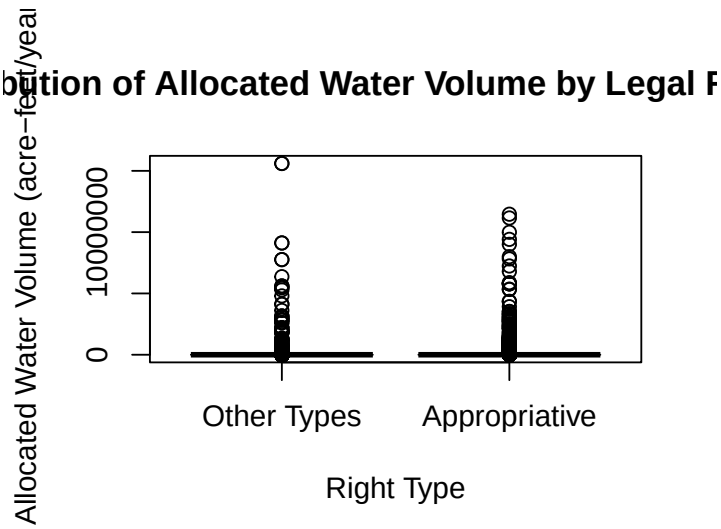

Figure 1: Boxplot of allocated water volume by legal right type. Appropriative rights exhibit a higher median allocation, but both groups show substantial variability and many outliers.

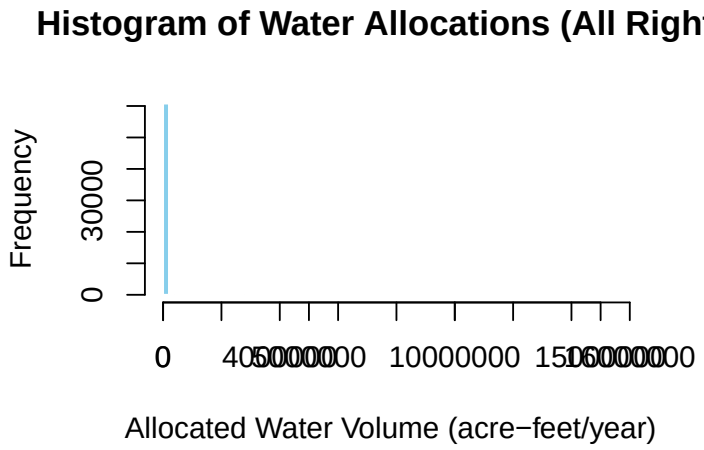

Figure 2: Histogram of water allocations (all legal types), showing heavy right skew and extreme upper-tail outliers.

Characteristics & Limitations:
Exploratory analyses using boxplots and histograms (Figures Figure 1 and Figure 2) provide insight into the cleaned dataset. Figure Figure 1 shows that Appropriative rights have a higher average and median allocation than other rights, but exhibit wide variability, with several notable outlier records above the box. Figure Figure 2 further demonstrates the highly right-skewed distribution of water allocations overall, as most rights are clustered at low volumes while a minority receive very large allocations. Together, these visualizations confirm that even with mean differences by group, the large variability and outlier prevalence pose significant challenges for regression modeling and interpretation.

Methods

Model diagnostic and regression procedures followed best practices as outlined by Fox and Weisberg (Fox and Weisberg 2019), Kutner et al. (Kutner, Nachtsheim, and Neter 2005), and Gelman et al. (Gelman, Hill, and Vehtari 2020).

We fit a Simple Linear Regression (SLR) model to evaluate the association between legal right type and allocated water volume. For each water right record i, the model is specified as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where:

- Y_i: Allocated water volume for record i
- X_i: Appropriative indicator (1 if Appropriative, 0 otherwise)
- β_0: Mean allocation for Other Types
- β_1: Mean difference (Appropriative vs Other)
- ϵ_i: Random error (unexplained variation)

Model parameters were estimated using Ordinary Least Squares (OLS), minimizing the sum of squared residuals.

Assumptions and Diagnostics:
The model relies on four primary statistical assumptions: - *Independence*: Likely valid for cross-sectional administrative records. - *Linearity/Additivity*: Assured by the binary nature of X. - *Constant Variance (Homoscedasticity)*: Checked using residual plots; evidence of violation detected. - *Normality of Errors*: Evaluated using Q-Q plots; major departures observed due to heavy tails and outliers.

Influence Assessment:
Influential observations were identified using Cook's distance, with spikes indicating records that disproportionately impact parameter estimates.

Diagnostic plots were generated to visually assess these assumptions and guide interpretation of model validity. We assess **Cook's distance** to identify observations exerting disproportionate influence on the slope and intercept (see Figure 5).

Results

The fitted Simple Linear Regression model reveals a statistically significant difference in mean allocated water volumes between Appropriative and other water right types. Coefficient estimates, fit metrics, and uncertainty measures are summarized below.

Table 1: Summary of fitted SLR model: coefficient estimates, uncertainty, and fit metrics.

Term	Value
Intercept (Other Types Mean)	4,542
Slope (Difference for Appropriative)	18,851
R^2	0.00159
MSE	41,186,033,068
t-statistic (Slope)	9.86
p-value (Slope)	6.35e-23

All visualizations and statistical checks were performed using the car (Fox and Weisberg 2023) and tidyverse (Wickham 2023) packages in RStudio (RStudio Team 2025)

Interpretation:
Appropriative right holders receive, on average, approximately 18,851 acre-feet/year more than other types. However, the R-squared value is extremely low (0.00159), showing that legal status explains only a small fraction of the variance.

Model Uncertainty:
- 95% CI for mean allocation (Other Types): [2,691, 6,393]
- 95% CI for mean difference (Appropriative vs Other): [15,104, 22,597]
- 95% prediction intervals for new observations:
- $X = 0$: [-393,239, 402,322]
- $X = 1$: [-374,397, 421,182]

Diagnostics and Assumption Checks:

To examine the validity of these findings, we conducted formal and visual diagnostic checks:

```
[1] 11678 11679
```

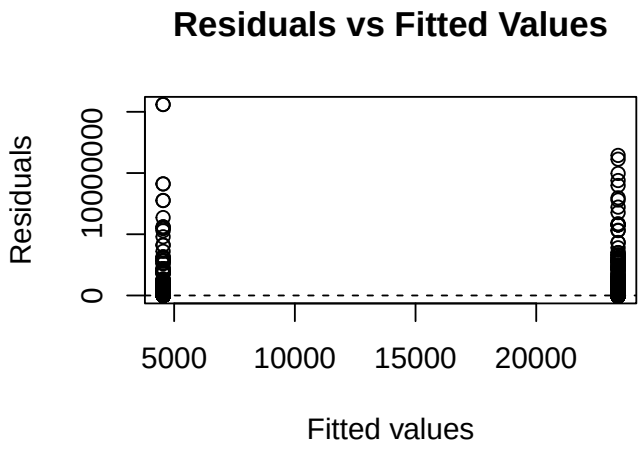

Figure 3: Residuals versus fitted values: dramatic increase in variance confirms severe heteroscedasticity.

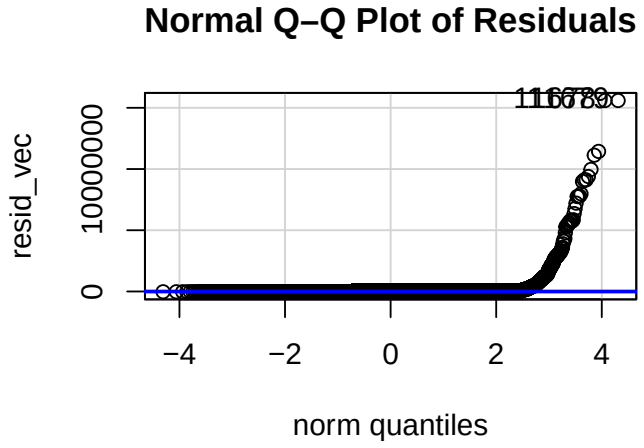

Figure 4: Normal Q–Q plot of residuals: pronounced tail deviations highlight non-normality and presence of outliers.

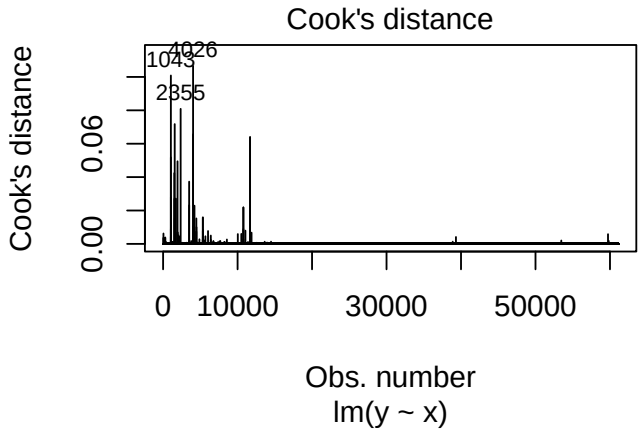

Figure 5: Cook's distance plot: vertical spikes mark observations with disproportionate influence on model fit.

Figures Figure 3 and Figure 4 highlight clear violations of homoscedasticity and normality, together with pronounced outlier effects. Figure Figure 5 identifies highly influential observations whose presence alters the model's parameters substantially.

Despite a strong group mean difference, the extreme variability within groups, the presence of influential records, and wide prediction intervals reveal the limits of simple regression for this dataset. For policymakers, this means legal rights alone cannot predict individual allocations reliably; for data scientists, it underscores the necessity of diagnostic rigor and transparent communication.

Potential improvements include log transformation, robust regression, or the inclusion of additional predictors (such as geography or year) to account for the data's underlying complexity.

References

"California Open Data Portal." 2025. 2025. https://data.ca.gov/.

"California Water Rights List." 2025. California Open Data. 2025. https://data.ca.gov/.

Fox, John, and Sanford Weisberg. 2019. *An r Companion to Applied Regression*. 3rd ed. Sage.

———. 2023. *Car: Companion to Applied Regression*. https://cran.r-project.org/package=car.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press.

Kutner, Michael H., Christopher J. Nachtsheim, and John Neter. 2005. *Applied Linear Statistical Models*. 5th ed. McGraw-Hill Irwin.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2025. *RStudio: Integrated Development Environment for r*. RStudio, PBC. https://rstudio.com/.

Wickham, Hadley. 2023. *Tidyverse: Easily Install and Load the Tidyverse*. https://cran.r-project.org/package=tidyverse.