Linear, Multiple, Interaction, and Polynomial Regression Analysis of Internet Firewall Traffic Data

Gabil Gurbanov Vusal Shirinbayli
Applied Machine Learning & Data Analytics
Instructor: Dr. Samir Rustamov

***Abstract*—This project applies regression-based modeling to a real-world Internet Firewall dataset to explain and predict transmitted traffic volume (bytes). We implement (i) simple linear regression with a single predictor, (ii) multiple linear regression with statistical inference via t-tests and a global F-test, (iii) interaction analysis between a qualitative variable (firewall action) and a quantitative predictor (packets), and (iv) polynomial regression models. Models are evaluated using RSS, RSE, R^2, adjusted R^2, correlation r, prediction intervals, and RMSE using a train-test split. In addition, we study model complexity through a bias–variance tradeoff analysis across polynomial degrees. The empirical results show that packet count is the dominant determinant of transmitted bytes; additional predictors offer only minor improvements, while higher-degree polynomial models reduce bias but increase variance and generalization error. These findings provide a statistical explanation of why simpler models can be preferable for traffic-volume prediction in firewall logs with heavy-tailed distributions and outliers.**

***Index Terms*—Linear Regression, OLS, Multiple Regression, Interaction Terms, Polynomial Regression, Bias–Variance Tradeoff, Firewall Logs, Model Diagnostics**

I. Introduction (Motivation)

Network firewalls generate high-volume traffic logs capturing packet counts, bytes transferred, durations, port information, and action decisions (e.g., allow/deny). Quantifying how these variables relate is important for (i) traffic engineering and capacity planning, (ii) identifying unusual sessions or flows that deviate from typical packet-to-byte relationships, and (iii) interpreting how firewall policies correlate with traffic characteristics.

From a machine learning perspective, firewall logs provide a structured real-world setting where regression models can be used not only for prediction, but also for *interpretability*: coefficients and hypothesis tests provide evidence about which variables influence the response and whether the influence differs across conditions (e.g., firewall action). A key practical question is whether a simple model (fast and interpretable) is sufficient, or whether additional predictors and non-linear complexity yield meaningful improvement.

This study aims to:

- quantify how well packet-level activity explains transmitted bytes;
- test statistical significance of predictors (t-tests and global F-test);
- evaluate whether multiple predictors provide substantive improvement;
- analyze interactions between qualitative and quantitative variables;
- compare polynomial models of increasing degree and interpret results using the bias–variance framework.

II. Dataset and Exploratory Analysis

We use the Internet Firewall dataset (Kaggle and UCI) containing $n = 65{,}532$ observations. The response variable is *bytes*. The key quantitative predictors used in this report are *packets* and *elapsed_time_sec*. The qualitative variable *action* indicates the firewall decision (allow/deny). Additional firewall fields (e.g., ports, sent/received packet counters) exist in the dataset and appear strongly correlated with the main volume variables; to avoid redundancy and multicollinearity, we focus on the most interpretable predictors aligned with the assignment requirements.

A. Preprocessing and Data Quality Notes

The dataset is numerical and categorical, requiring minimal transformation for linear regression. There were no missing values after checking and preprocessing. Typical preprocessing steps applied include: selecting target and predictors, handling categorical variables via indicator coding (e.g., $C(action)$), and ensuring consistent train-test split for the polynomial experiments. Because network traffic often exhibits heavy-tailed distributions, we inspect plots at both full scale and zoomed scale (to reduce dominance of extreme outliers in visualization).

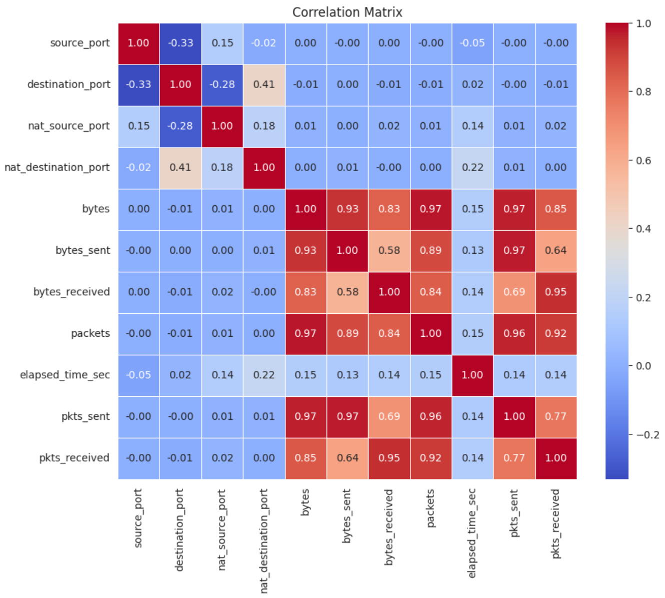

Fig. 1. Correlation matrix. Traffic volume variables are highly correlated (e.g., $corr(bytes, packets) = 0.97$), indicating multicollinearity risk when multiple volume-related predictors are used together.

Fig. 1 highlights a strong correlation between packets and bytes, consistent with the approximation:

$$bytes \approx packets \times (\text{avg. packet size}).$$

This relationship motivates simple linear regression as a strong baseline. At the same time, the presence of multiple highly correlated volume features (e.g., bytes_sent, pkts_sent, pkts_received) suggests that including all of them simultaneously would inflate multicollinearity, making coefficient interpretations less stable even if prediction is accurate.

III. Methods

All regression models are fit using Ordinary Least Squares (OLS). OLS estimates parameters by minimizing:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

We report:

- **Inference (single and multiple)**: coefficient estimates, t-statistics, p-values; global F-test for multiple regression.
- **Accuracy (single)**: RSS, RSE, R^2, correlation r.
- **Fit (multiple)**: R^2, adjusted R^2, residual diagnostics.
- **Prediction**: example prediction and 95% prediction interval.
- **Generalization (polynomial)**: train-test RMSE comparison, plus bias–variance trend across degrees.

A. Simple Linear Regression

We fit:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1}$$

where Y is bytes and X is packets. Inference uses:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}.$$

We report p-values to judge statistical significance at $\alpha = 0.05$.

B. Multiple Linear Regression

We fit:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \tag{2}$$

where X_1 is packets and X_2 is elapsed_time_sec. Overall usefulness is tested via the global F-test:

$$H_0 : \beta_1 = \beta_2 = \cdots = 0 \quad \text{vs} \quad H_1 : \text{at least one } \beta_j \neq 0.$$

We also discuss adjusted R^2, which penalizes adding predictors that do not improve explanatory power.

C. Interaction Model

To test whether the packets–bytes relationship changes across firewall action, we use:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D) + \varepsilon.$$

Here, β_3 measures slope change across categories; a non-significant interaction implies the slope remains effectively consistent across allow/deny.

D. Polynomial Regression and Complexity

We fit polynomial models:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \varepsilon,$$

and evaluate with RMSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}.$$

We compare train RMSE vs test RMSE across degrees to interpret bias–variance tradeoff and overfitting.

IV. Results (Rubric-Mapped)

This section explicitly answers all required questions for (i) single predictor regression, (ii) multiple regression, (iii) interaction analysis, and (iv) polynomial regression. Numeric values are taken from the implemented notebook outputs.

A. Single Predictor Regression (30%)

(a) Coefficients and RSS. The fitted model yields:

$$\hat{\beta}_0 = -12{,}585.57, \quad \hat{\beta}_1 = 1066.53,$$

with $RSS = 1.0464 \times 10^{20}$. The positive slope indicates that, on average, each additional packet increases transmitted bytes by approximately 1066.53 bytes, reflecting a stable average packet size effect in aggregated logs.

(b) t-statistic and p-value. For packets: $t = 1109.02$, $p < 0.001$. Therefore, packet count is statistically significant and strongly associated with bytes. The extremely large t-statistic is consistent with the large sample size and strong correlation observed in Fig. 1.

(c) Accuracy (RSE, R^2, r).

$$R^2 = 0.9494, \quad RSE = 1{,}263{,}655.85, \quad r = 0.9744.$$

Thus, packets alone explain about 94.94% of the variance in bytes, and the Pearson correlation confirms a strong positive linear relationship.

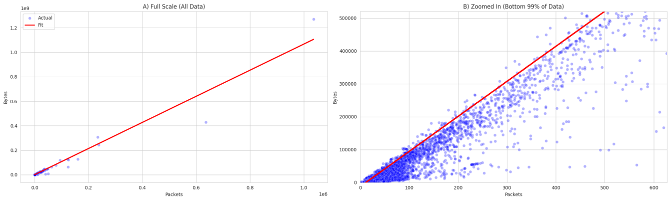

Fig. 2. Simple regression fit: bytes vs packets. The relationship is strongly linear, but heavy-tailed traffic introduces outliers at high packet counts.

B. Interpretation and Practical Meaning

The high R^2 and large t-statistic indicate the model is both statistically significant and practically meaningful. However, the large RSE reflects that raw bytes vary widely across sessions, even for similar packet counts; this is expected when traffic contains heterogeneous applications and variable payload sizes. Therefore, while the linear trend is strong, prediction uncertainty remains high for individual observations.

C. Multiple Predictors Regression (30%)

(a) Is at least one predictor useful? The global F-test p-value is < 0.001, so we reject the null hypothesis that all coefficients (excluding intercept) are zero. Hence, at least one predictor is useful.

(b) Do all predictors help or only a subset? Individual tests show:

$$p_{\text{packets}} < 0.001, \quad p_{\text{elapsed\_time\_sec}} < 0.001.$$

Both predictors are statistically significant. Nonetheless, statistical significance does not automatically imply a large practical improvement; thus we also compare R^2 and adjusted R^2.

(c) How well does the model fit?

$$R^2 = 0.9612, \quad Adj.\ R^2 = 0.9612.$$

The improvement over the baseline ($\Delta R^2 \approx 0.0118$) is modest: packets explains nearly everything, while elapsed_time_sec provides additional, smaller explanatory value.

(d) Prediction and accuracy. For an example observation, the model predicts:

$$\hat{Y} = 97{,}123.95 \text{ bytes}$$

with 95% prediction interval:

$$[-2{,}070{,}740.09,\ 2{,}264{,}987.99].$$

Because bytes cannot be negative, a practical lower bound is 0. The wide interval reflects high variability in session characteristics, heavy-tailed distributions, and outliers.

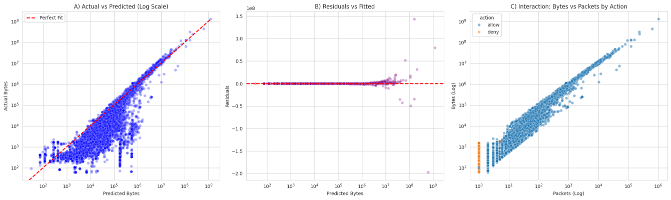

Fig. 3. Multiple regression diagnostics (residuals vs fitted). Increasing variance suggests mild heteroskedasticity; outliers dominate at high fitted values.

(e) Interaction analysis. The interaction term $X \times D$ has $p > 0.05$. Hence, we do not find strong evidence that firewall action changes the slope of packets. In other words, packets remain the dominant predictor regardless of allow/deny, and action does not meaningfully modify the packets–bytes linear relationship in the observed sample.

D. Subset Selection Perspective

Given multicollinearity in traffic-volume fields, an important practical modeling step is selecting a subset of predictors that improves fit without redundancy. In our analysis, packets provides maximal explanatory power; elapsed_time_sec adds a small improvement; adding additional volume-related predictors may yield diminishing returns and reduce interpretability due to multicollinearity.

E. Forward and Backward Feature Selection

To further evaluate whether all predictors are necessary, we applied both forward selection and backward elimination using p-value criteria at $\alpha = 0.05$.

Candidate predictors:

$$\{\text{packets},\ \text{elapsed\_time\_sec},\ \text{pkts\_sent},\ \text{pkts\_received}\}$$

Forward Selection Procedure:

- Step 1: Added *packets* ($p < 10^{-10}$)
- Step 2: Added *pkts_sent* ($p < 10^{-10}$)
- Step 3: Added *elapsed_time_sec* ($p = 7.33 \times 10^{-26}$)
- Step 4: Added *pkts_received* ($p = 1.30 \times 10^{-6}$)

Final Forward Model:

$$\{\text{packets},\ \text{pkts\_sent},\ \text{elapsed\_time\_sec},\ \text{pkts\_received}\}$$

Backward Elimination Procedure:

Starting with all predictors, no variable had $p > 0.05$. Therefore, no predictors were removed.

Final Backward Model:

$$\{\text{packets},\ \text{elapsed\_time\_sec},\ \text{pkts\_sent},\ \text{pkts\_received}\}$$

Interpretation:

Both forward and backward procedures selected the same full set of predictors. This indicates that each variable contributes statistically significant explanatory power. However, due to strong multicollinearity among traffic-volume features (see Fig. 1), the practical incremental improvement beyond *packets* alone is modest.

F. Polynomial Regression (20%)

We compare the linear baseline ($d = 1$) to a quadratic model ($d = 2$) using test RMSE:

$$RMSE_{linear} = 418{,}728.70, \quad RMSE_{poly2} = 511{,}282.77.$$

The polynomial model performs worse on test data, indicating overfitting: higher flexibility fits noise/outliers rather than stable signal.

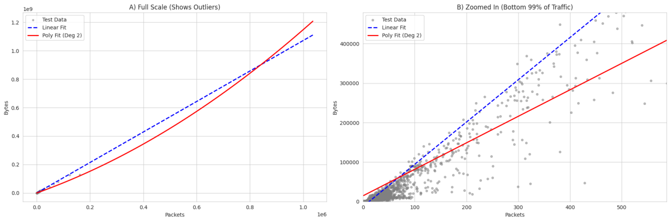

Fig. 4. Linear vs quadratic polynomial regression. The quadratic curve may fit certain ranges but does not improve generalization overall.

V. Bias–Variance Tradeoff and Model Complexity

To deepen the polynomial analysis, polynomial degrees from 1 to 6 were evaluated. Fig. 5 shows train and test RMSE on a log scale. Training RMSE tends to decrease with degree (reduced bias), while test RMSE shows a U-shaped trend: it decreases initially then increases (higher variance).

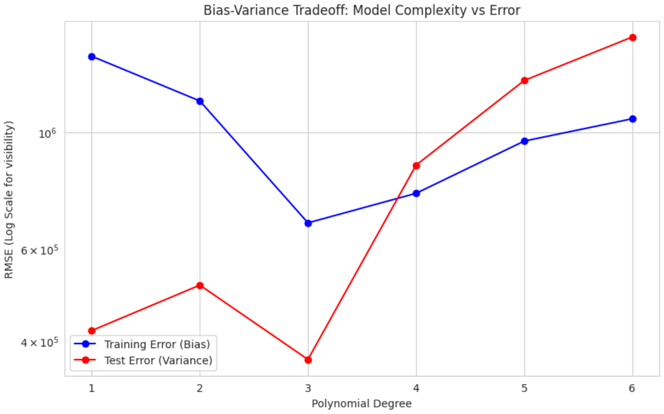

Fig. 5. Bias–variance tradeoff: training and test RMSE across polynomial degrees. Higher degrees reduce training error but increase test error (overfitting).

A. Interpretation

This pattern matches the classical bias–variance tradeoff: models that are too simple underfit (high bias), while models that are too complex overfit (high variance). In network traffic with outliers, increasing polynomial degree often amplifies sensitivity to extreme points. This provides an empirical explanation for why $d = 2$ can perform worse than $d = 1$ even if the training fit appears improved.

VI. Summary of Key Metrics

Table I consolidates rubric-critical numerical results for transparent grading and quick cross-checking.

TABLE I
Key Results Summary

Item	**Value**
Single LR $\hat{\beta}_0$	-12,585.57
Single LR $\hat{\beta}_1$	1066.53
Single LR RSS	1.0464×10^{20}
Single LR t (packets)	1109.02
Single LR p (packets)	< 0.001
Single LR R^2	0.9494
Single LR RSE	1,263,655.85
Single LR r	0.9744
Multi LR R^2	0.9612
Multi LR Adj. R^2	0.9612
Multi LR F-test p	< 0.001
Prediction $\hat{Y}$	97,123.95 bytes
95% Pred. Interval	[-2,070,740.09, 2,264,987.99]
RMSE (Linear)	418,728.70
RMSE (Poly deg-2)	511,282.77

A. Additional Interpretability Notes

Although R^2 is high, prediction intervals remain wide, highlighting the difference between *explaining variance* and *predicting individual outcomes*. This is typical in noisy real-world systems where unobserved factors (protocol type, payload, application behavior) influence bytes beyond packets and duration.

VII. Discussion and Error Analysis

Multicollinearity. Fig. 1 indicates strong relationships among traffic volume variables. In multicollinear settings, coefficient estimates may become unstable and standard errors may inflate, which can complicate interpretation. This is one reason to prefer parsimonious models. A typical diagnostic (not required here) is Variance Inflation Factor (VIF), which quantifies how much variance is inflated due to predictor correlation.

Residual structure and heteroskedasticity. Fig. 3 suggests increasing residual variance at higher fitted values (mild heteroskedasticity). This is common in network traffic due to heavy tails and extreme sessions. Potential remedies (beyond assignment scope) include log-transforming the response (e.g., $\log(1 + \text{bytes})$) or using robust standard errors to correct inference under heteroskedasticity.

Model assumptions. Classical OLS assumes linearity, independence, constant variance (homoscedasticity), and approximately normally distributed errors. In large samples, inference can be robust to mild departures from normality, but heteroskedasticity can affect standard errors. Our diagnostic plots suggest linearity is strong, while constant variance is only approximate.

Negative results and baseline comparison. Polynomial regression is a negative result: it performs worse than the baseline on test RMSE (Section IV-C). This is not a failure but an important outcome: it supports the conclusion that the system is already well explained by a linear mechanism and that increased flexibility mainly fits noise/outliers.

Practical implications. In firewall monitoring, a simple linear model can provide a fast approximation for expected bytes given packet count. When used for anomaly detection, one could flag sessions with unusually large residuals (actual bytes far from predicted bytes), although that extension is beyond this assignment.

VIII. Conclusion

We implemented and evaluated regression models on Internet Firewall traffic data. The single-predictor model (bytes vs packets) explains most variance ($R^2 = 0.9494$) with highly significant coefficients. Multiple regression yields a modest improvement ($R^2 = 0.9612$), while interaction analysis provides no strong evidence that firewall action changes the packets-to-bytes slope. Polynomial regression and higher-degree models increase variance and worsen generalization, confirmed by both test RMSE and the bias–variance tradeoff analysis. Overall, the relationship between packets and bytes is predominantly linear in this dataset.

Team Contributions

Gabil Gurbanov: regression modeling and statistical inference (t-tests, F-tests), interaction analysis, evaluation metrics and interpretation, result validation, report formatting/writing and compilation.

Vusal Shirinbayli: data preparation and cleaning, visualization generation, polynomial experiments and bias–variance analysis,

References

[1] UCI Machine Learning Repository, “Internet Firewall Data Set.”

[2] Kaggle, “Internet Firewall Data Set.”

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.