When Waste Becomes Risk: Modeling the Drivers of Toxic Chemical Release Occurrence in U.S. Facilities*

Prasanna Sai Rohit Durbha

December 10, 2025

This study analyzes facility–chemical records from the 2024 EPA Toxics Release Inventory (TRI) to understand how production activity, waste management practices, chemical hazard classifications, and industry sector relate to on-site chemical releases. Initially, the study fits a multiple linear regression model using total release amounts as the continuous response. Diagnostic plots and residual patterns reveal that the model breaks the assumptions of linearity, constant variance, and normality, driven by skewed and zero-inflated distribution of release totals. The model provides little predictive accuracy and shows that it is not suitable for explaining variation in raw TRI release quantities. To address these limitations, The study then transforms the continuous response variable to model the probability of any on-site release using logistic regression. This approach aligns the method with the binary nature of most TRI outcomes and yields interpretable, meaningful results. Production-related waste, one-time releases, and hazard indicators such as PBT (persistent, bioaccumulative, and toxic) and Clean Air Act chemical status significantly increase the odds of reporting a release. Calibration and residual diagnostics show that the logistic model fits the data better than the linear regression model did on the raw release amounts and captures meaningful structure in release behavior. Together, these findings demonstrate that logistic regression offers a more appropriate and informative framework for understanding the factors associated with on-site chemical releases in the TRI data.

*Project repository available at: https://github.com/rohitDurbha/MATH261A_Paper2_RohitDurbha.

1 Introduction

The Toxics Release Inventory (TRI) is a long-running U.S. Environmental Protection Agency (EPA) program that requires certain industrial facilities to report, each year, how they manage and release specified toxic chemicals (TRI Dataset 2024). TRI data are widely used by communities, researchers, and policymakers to monitor potential environmental hazards, track progress in pollution prevention, and evaluate regulatory policies. Every TRI reporting form includes details such as facility location, industry sector, chemical identity, on-site releases to air, water, and land, transfers to other sites, and quantities managed through recycling, energy recovery, and treatment. This information could also form the basis for determining rising pollution levels and declines in flora and fauna near these facilities. Public discussions of TRI data often emphasize the amount of chemicals released such as "total pounds released to air in a state.". But from a regulatory and risk-screening perspective, a more basic question is also important: **under what conditions do facilities release toxic chemicals at all?** Not every facility–chemical combination in TRI results in an on-site release; many facilities may instead recycle, treat, or transfer chemicals off site. Moreover, when we look at on-site release amounts across all facility–chemical records, the distribution is zero-inflated, i.e., most records are exactly zero, and a relatively small number have very large positive values. This structure presents substantial challenges for traditional multiple linear regression, which assumes approximately normal, homoscedastic errors. This study focuses on the following research question: **Among facilities reporting to the 2024 TRI, which facility-level, operational-level, and chemical-level characteristics are associated with a higher probability that a facility reports any on-site release of a given toxic chemical?**

The ideal scenario to evaluate this research question would be to predict the actual on-site release total. To explore this, We initially fit a multiple linear regression (MLR) model with the continuous on-site release total as the response and several facility and chemical predictors. Diagnostic plots reveal violations of linear model assumptions due to zero inflation, skewness, and heteroskedasticity. We then reformulate the problem as a binary outcome, whether any on-site release occurred and fit a logistic regression model. This logistic framework is more appropriate because it directly models the probability of a release event and respects the discrete nature of the outcome.

The findings indicate that operational intensity, non-routine one-time releases, federal facility status, and chemical hazard classifications are all associated with better odds of reporting an on-site release. The remainder of the report is structured as follows. Section 2 describes the 2024 TRI Basic Data File, the observational units, and the variables used. Section 3 introduces both the multiple linear and logistic regression models and explains their assumptions. Section 4 presents the main model results and interprets the key coefficients. Section 5 reflects on the implications and limitations of the findings.

---

2 Data

2.1 Observational units and population

The data comes from the **2024 TRI Basic Data File**, which the EPA describes as containing "the data elements most frequently requested by TRI data users," including facility information, chemical identification, on-site release quantities, off-site transfers, and summary waste-management data (TRI Dataset 2024). Each **row** in this Basic Data File represents a **single facility–chemical–year record**. Concretely:

- A facility is a specific industrial site (for example, a chemical plant or metal fabrication facility) that meets TRI reporting thresholds (A facility meets TRI reporting thresholds if it is covered by the TRI program or is a federal facility, has 10 or more full-time employees, uses or processes TRI-listed chemicals, and exceeds the annual activity thresholds for a listed chemical or chemical category).
- A chemical is one TRI-listed substance (or chemical category) covered under the TRI program. (Mercury, Lead, Barium)
- For the 2024 reporting year, a facility that handles three TRI chemicals will contribute three rows or one row per chemical.

Therefore, the observational unit in this analysis is: One facility–chemical record for the 2024 TRI reporting year.

Some facilities appear multiple times in the dataset if they handle multiple chemicals. Some chemicals also appear in many facilities. The analysis generalizes to the population of all facilities subject to TRI reporting requirements in 2024 in the United States, recognizing that TRI is not a simple random sample but effectively a census of covered facilities and chemicals (TRI Dataset 2024).

2.2 Response variables

This study considers **two** different response variables, one for a multiple linear regression model and the other for a logistic regression model. They are as follows:

1. **Continuous response for MLR** : The TRI Basic Data File provides several columns summarizing different on-site release pathways (for example, fugitive air emissions, stack air emissions, discharges to surface water, and various land disposal categories). These are about 14 columns of values which are combined into one variable called **ON-SITE RELEASE TOTAL** (in pounds) which reports the total quantity of the chemical released on site via all reported media during the year. This continuous measure is the natural candidate for a multiple linear regression model, because it directly quantifies environmental burden at the facility–chemical level.

2. **Binary response for logistic regression** : Since the response variable's values range from 0 to a value in millions, it becomes a difficult task to fit a multiple regression model to this variable, and simplifying the variable helps us refine the research question to whether any release occurs. The transformed variable is as follows:

   - **RELEASE_FLAG** = 1 if ON-SITE RELEASE TOTAL > 0 (i.e., any positive on-site release is reported for that facility–chemical in 2024);

   - **RELEASE_FLAG** = 0 if ON-SITE RELEASE TOTAL = 0 (no on-site release of that chemical is reported for that facility).

   This binary outcome is much better suited to logistic regression, which models the probability that a release occurs rather than the magnitude of the release.

2.3 Predictor variables: definitions

The TRI Basic Data File includes many variables. For this analysis we select a set of predictors that are both substantively meaningful and consistently reported across facilities (TRI Dataset 2024). For each variable, we explain what it represents and why it is relevant to predicting releases.

- **Production Ratio** (numeric) : This is a self-reported ratio comparing the facility's production in the current year to production in a baseline or previous year. A value greater than 1 indicates increased production, while a value less than 1 indicates reduced production. Higher production typically implies more chemical use and more waste streams, which could increase both routine and non-routine releases.

- **Production Waste** (numeric) : This is the total production-related waste for the chemical and aggregates quantities released, recycled, used for energy recovery, and treated, but only those associated with normal production activities. Production-related waste summarizes the overall scale of the chemical's throughput in the facility. Facilities generating more waste are likely to handle larger quantities of the chemical and therefore may face greater challenges in preventing releases.

- **Treatment On Site** (numeric) : This is the amount of the chemical waste that is treated on site (for example, through physical, chemical, or biological treatment) to reduce its hazard or volume. Facilities with substantial on-site treatment may either indicate high waste burdens (increasing potential release opportunities) or strong pollution control infrastructure (possibly reducing releases). Including this variable allows the model to test which pattern dominates in practice.

- **Recycling On Site** (numeric) : This represents the quantity of the chemical that is recycled within the facility rather than being released or transferred off site. On-site recycling can indicate more integrated resource management. It may either correlate

with better pollution prevention (if recycling replaces disposal) or with large operational scale. The sign of the coefficient helps distinguish these possibilities.

- **One-time Release** (numeric) : This variable captures non-routine or one-time releases, such as accidental spills, equipment failures, or other unusual events, reported under TRI's one-time release categories. One-time releases are directly related to environmental incidents. While they are themselves a type of release, they can also serve as a signal of underlying process instability or risk. Including this variable helps separate routine operational behavior from exceptional events.
- **Industry Sector** (categorical) : This factor encodes the facility's industry sector based on North American Industry Classification System (NAICS) codes mapped into EPA TRI categories (for example, Chemicals, Petroleum, Fabricated Metals, Food, Paper). Different industries use different processes, chemicals, and technologies. Some sectors are intrinsically more emission-intensive than others. By including industry sector as a categorical predictor, the model can control for broad structural differences across industries.
- **Federal Facility** (categorical: YES/NO) : This indicator identifies whether the facility is owned or operated by a federal agency. Federal facilities may operate under different regulatory regimes or reporting practices compared to private facilities. Including this variable allows us to test whether federal operations are more or less likely to report on-site releases.
- **Clean Air Act Chemical** (categorical: YES/NO) : This flag indicates whether the reported chemical is also regulated as a hazardous air pollutant under the Clean Air Act. Chemicals regulated under multiple statutes may face stronger air emission controls, but they are also frequently used in large-scale industrial operations. This variable helps assess whether such regulations correlate with observed release behavior.
- **Carcinogen** (categorical: YES/NO) : This variable identifies chemicals that are classified as carcinogens under OSHA or other regulatory criteria. Carcinogens are subject to stricter occupational and environmental controls. Including this indicator tests whether carcinogenic chemicals are more or less likely to be released.
- **PBT** (categorical: YES/NO) : This flag marks chemicals that are persistent, bioaccumulative, and toxic (PBT). A particularly hazardous group of substances that do not break down easily in the environment and accumulate in organisms. PBTs are among the highest-priority chemicals from an environmental health perspective. If facilities handling PBTs are more likely to release them on site, that carries especially important implications for policy and community risk.

2.4 Data Collection and Preprocessing

The data used in this study come from the U.S. Environmental Protection Agency's 2024 Toxics Release Inventory (TRI) Basic Data File, which compiles facility-level reports on chemical management and environmental releases submitted under Section 313 of the Emergency Planning and Community Right-to-Know Act. The Basic Data File includes facility identifiers, industry sector classifications, chemical characteristics, quantities of waste managed in various ways, and all reported on-site release amounts (TRI Dataset 2024). Because TRI reporting covers all eligible facilities rather than a sampled subset, the dataset represents a near-census of industrial chemical activity in the United States for 2024.

Before analysis, we performed a series of preprocessing steps to prepare the data for regression modeling. Several waste-management variables such as production-related waste, on-site treatment quantities, and one-time releases were imported as character strings due to formatting inconsistencies in the raw file. These fields were converted to numeric form, and all categorical predictors were recoded as factors. We then inspected the data for missing values across the full set of predictors and the response variables. TRI records with incomplete reporting on any of these fields were removed, resulting in 10,560 complete facility–chemical observations. Because each row corresponds to one facility reporting on one chemical in 2024, the cleaned dataset reflects the subset of facilities and chemicals for which all required information was available to support the regression models.

There are other alternative EPA datasets that could have been used but were not selected for this analysis. The TRI National Analysis Dataset provides aggregated summaries at the facility, state, and industry levels, but does not preserve the facility–chemical granularity required for modeling release behavior. The Risk-Screening Environmental Indicators (RSEI Dataset 2024) dataset incorporates toxicity-weighted risk scores derived from TRI data, enabling better risk assessment, but its derived nature makes it less suitable for studying the operational predictors of raw release events. Similarly, the EPA Enforcement and Compliance History Online (Echo Dataset 2024) system provides detailed compliance histories and inspection records for facilities, but does not include chemical-level waste-management data needed for the regression models used here. The Basic Data File was therefore chosen because it contains the complete, unaggregated facility–chemical information necessary to evaluate how production activities, waste management practices, and chemical hazard classifications relate to the occurrence of on-site releases.

Table 1 reports how many facility–chemical records are available before and after cleaning. The original file includes all reported records for 2024. After removing rows with missing values on the response or any of the predictors used in this study, 10,560 complete records remain. Since each row represents one facility handling one chemical in 2024, this final sample reflects the set of facilities and chemicals for which all variables required for the model are observed.

Table 1: Dataset Summary Pre vs Post Cleaning

Metric	Value
Rows before cleaning	77295
Rows after cleaning	10560
Columns after cleaning	123

Table 2 shows the range of values for each numerical variable in the cleaned sample. All of these variables span wide intervals. Some facilities report very small amounts of waste or none at all, while others report very large quantities. These wide ranges tell us that the observations come from facilities of many different sizes and operations. They also help explain why the continuous release values do not fit well with a linear regression model, since most records report zero release while a small number report very large amounts.

Table 2: Ranges of Numeric Variables

Variable	Range
ON-SITE RELEASE TOTAL	[0.00, 131567794.90]
PRODUCTION WSTE	[0.00, 131613570.70]
TREATMENT ON SITE	[0.00, 46496938.00]
RECYCLING ON SITE	[0.00, 92524111.50]
ONE-TIME RELEASE	[0.00, 3590000.00]
PRODUCTION RATIO	[0.00, 416.20]

Table 3 lists the levels for each categorical variable. The industry sector variable includes many different types of facilities, showing that the sample covers a wide mix of industrial activities. The hazard flags, such as whether a chemical is a carcinogen or a PBT chemical, include both yes and no responses. The federal facility indicator also includes both types of facilities. These variables help describe important differences among facilities and chemicals, which the regression model accounts for.

Table 3: Levels of Categorical Variables

Variable	Levels
INDUSTRY SECTOR	Beverages, Chemical Wholesalers, Chemicals, Coal Mining, Computers and Electronic Products, Electric Utilities, Electrical Equipment, Fabricated Metals, Food, Furniture, Hazardous Waste, Leather, Machinery, Metal Mining, Miscellaneous Manufacturing, Natural Gas Processing, Nonmetallic Mineral Product, Other, Paper, Petroleum, Petroleum Bulk Terminals, Plastics and Rubber, Primary Metals, Printing, Publishing, Textile Product, Textiles, Tobacco, Transportation Equipment, Wood Products
FEDERAL FACILITY	NO, YES
CLEAN AIR ACT CHEMICAL	NO, YES
CARCINOGEN	NO, YES
PBT	NO, YES

2.5 Descriptive Statistics

Before fitting the regression models, it is necessary to summarize the variables related to on-site releases and waste-management activities. These summaries help show the main patterns in the data, including the distribution of release amounts, the range of values for related quantities, and the differences across industry sectors. The following figures provide this context and show the features of the data that are relevant for the modeling choices that follow.

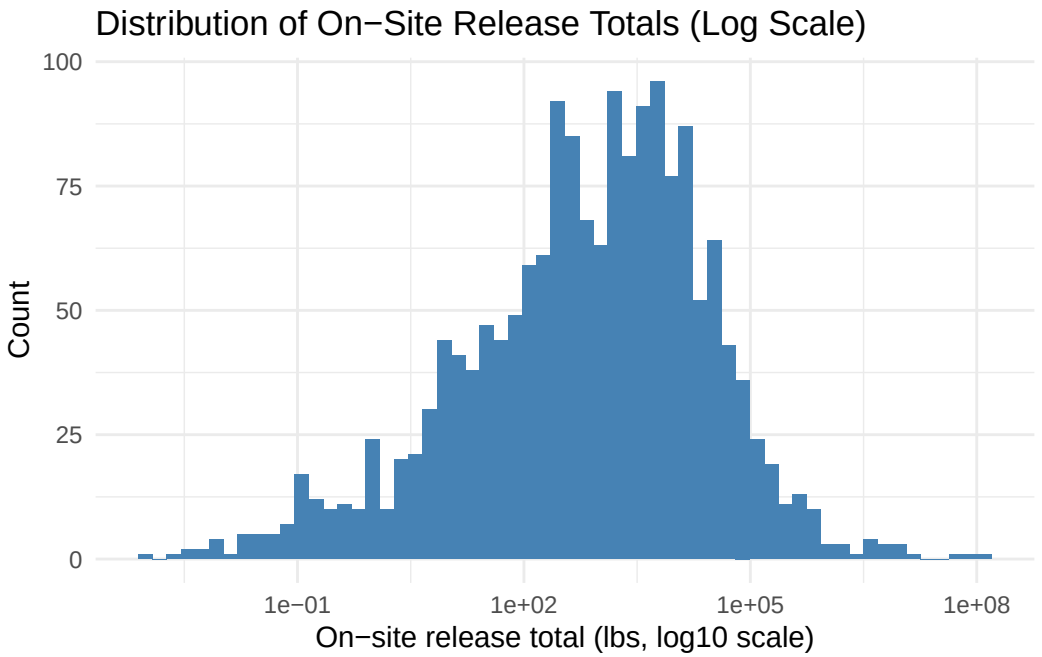

Figure 1: Histogram Showing the Highly Skewed Distribution of Facility On-Site Release Totals (log10 Scale)

Figure 1 shows the distribution of on-site release totals on a log scale. Most facility–chemical records report very small release amounts or none at all, while a smaller number report much larger values. Using a log scale makes it easier to see this spread. The shape of the distribution confirms that release totals vary widely across facilities, which is consistent with the idea that some operations handle far larger quantities of chemicals than others. This wide spread also helps explain why the continuous release totals do not fit well into a linear regression model.

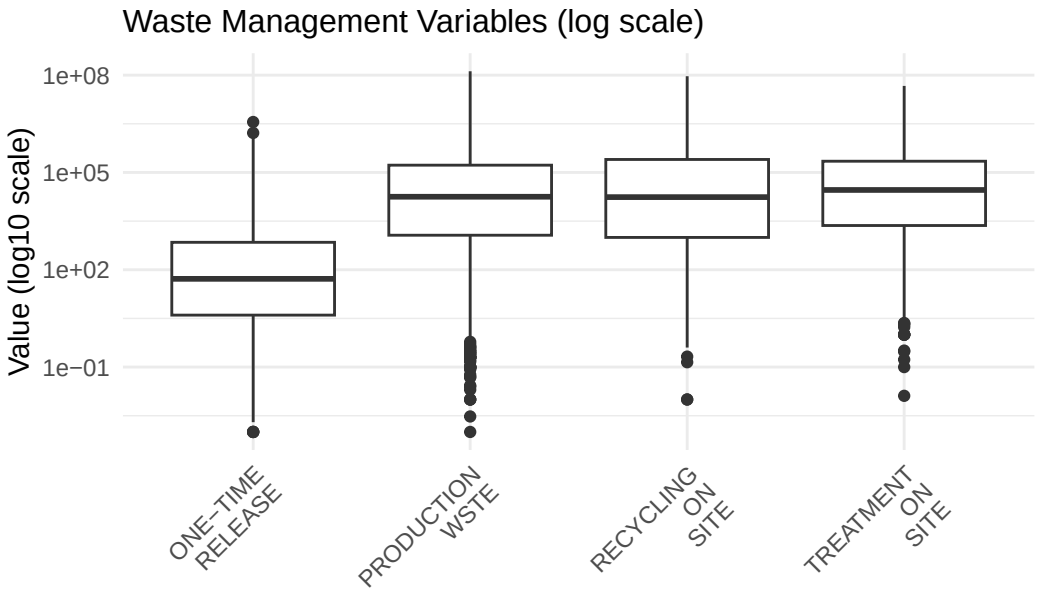

Figure 2: Boxplots showing the distribution of key waste-management quantities across facilities on a log10 scale.

Figure 2 shows boxplots for the main waste-management variables on a log scale. Each of these variables has many small values and a few very large ones. The boxes in the middle of each plot show that most facilities report relatively modest amounts, while the long whiskers and points show the presence of much larger values. These plots illustrate that the facilities handling TRI-listed chemicals differ greatly in size and activity, which is important context for later modeling.

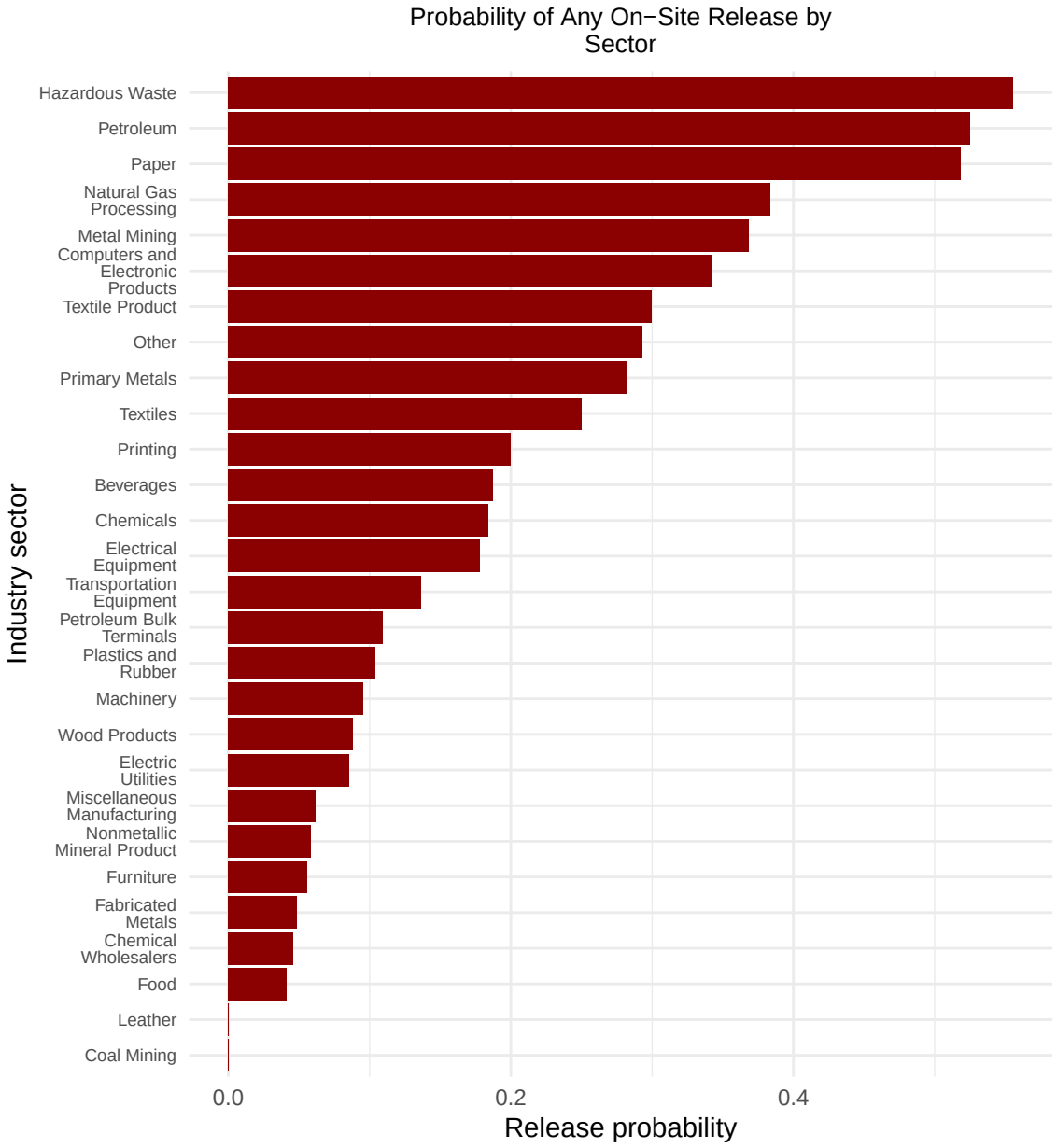

Figure 3: Estimated probability of a facility reporting any on-site release, grouped by industry sector.

Figure 3 shows the probability that a facility reports any on-site release, separated by industry sector. The sectors are ordered from highest to lowest probability. Some industries, such as hazardous waste management and petroleum-related activities, have higher release rates, while others report releases less often. This figure shows that release behavior differs across

industries.

---

3 Methods

The goal of this section is to describe the two statistical models used to study the research question. The first model treats the on-site release total as a continuous response and uses a multiple linear regression framework. This provides a starting point for assessing how the predictors relate to release amounts, even though the distribution of the response raises concerns for this approach. The second model focuses on whether any release occurs, which is a binary outcome, and uses logistic regression.

3.1 Multiple Linear Regression (MLR) model

As an initial approach, we treat the continuous on-site release total as the outcome of interest (or the response variable) and fit a **multiple linear regression** model (M. L. Regression 2013). Let

$$Y_i = \text{on-site release total for facility–chemical record } i,$$

$X_{i1}, X_{i2}, \ldots, X_{ip} =$ the production, waste, and chemical characteristics recorded for that observation.

We model the relationship as

$$\begin{aligned} Y_i = \beta_0 &+ \beta_1(\text{Production Ratio})_i + \beta_2(\text{Production Waste})_i + \beta_3(\text{Treatment on Site})_i \\ &+ \beta_4(\text{Recycling on Site})_i + \beta_5(\text{One-Time Release})_i + \text{Industry Sector Effects}_i \\ &+ \beta_F(\text{Federal Facility})_i + \beta_C(\text{Clean Air Act Chemical})_i \\ &+ \beta_K(\text{Carcinogen})_i + \beta_P(\text{PBT})_i + \varepsilon_i. \end{aligned}$$

In this model, β_0 is the intercept. It represents the expected on-site release total for a record in the baseline levels of all categorical variables and with all numerical predictors equal to zero.The coefficients β_1 through β_5 describe how the average release amount changes with the numerical predictors. For example, β_1 measures the change in the expected release total associated with a one-unit increase in the production ratio, holding the other predictors fixed. The same interpretation applies to production waste, treatment on site, recycling on site, and one-time releases. The term Industry Sector Effects$_i$ represents the set of indicator variables created from the industry sector category. Each coefficient in this group compares a given sector with the baseline sector. These terms allow the model to capture differences in release behavior across sectors. The coefficients β_F, β_C, β_K, and β_P measure the differences in

expected releases associated with being a federal facility, reporting a Clean Air Act chemical, reporting a carcinogen, or handling a PBT chemical, respectively. Each of these describes how the average release amount changes when that indicator equals one rather than zero. The error term ε_i accounts for variation in release totals that is not explained by the predictors in the model.

3.2 Assumptions of the Multiple Linear Regression Model

For this model to provide reliable estimates and valid inference, the below assumptions must hold:

1. **Linearity** : The average on-site release should change in a linear way with each predictor. This means that increases in production, waste, or hazard indicators should influence the response through a straight-line relationship within the range of the data.
2. **Independence** : The errors should be independent across observations. Each facility–chemical record is treated as a separate observation.
3. **Constant variance** : The spread of the errors should remain roughly the same across all fitted values. If the variance grows or shrinks systematically with the predicted release amount, this assumption is violated.
4. **Normality** : The errors should follow an approximately normal distribution. This supports the validity of the standard tests for the regression coefficients.

Although the true errors are not observed, the residuals from the fitted model are used to check these assumptions. As shown later in the results section, the distribution of on-site release totals makes these assumptions difficult to satisfy, which helps motivate the use of a logistic regression model in the next part of the analysis.

3.3 Logistic regression model

Logistic Regression is chosen as the second method of this regression analysis. Here, the response is whether a facility–chemical record reports any on-site release. This is a binary outcome mapped to **RELEASE_FLAG** variable, so we use a logistic regression model. For each record, let

$$Y_i = \begin{cases} 1, & \text{if the record reports any on-site release,} \\ 0, & \text{otherwise.} \end{cases}$$

Let p_i denote the probability that $Y_i = 1$ for record i. The logistic regression model relates this probability to the predictors through the logit link:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_1(\text{Production Ratio})_i + \alpha_2(\text{Production Waste})_i + \alpha_3(\text{Treatment on Site})_i$$
$$+\alpha_4(\text{Recycling on Site})_i + \alpha_5(\text{One-Time Release})_i + \text{Industry Sector Effects}_i$$
$$+\alpha_F(\text{Federal Facility})_i + \alpha_C(\text{Clean Air Act Chemical})_i + \alpha_K(\text{Carcinogen})_i + \alpha_P(\text{PBT})_i.$$

In this model, α_0 is the intercept. The coefficients α_1 through α_5 describe how the log-odds of reporting any release change with the numerical predictors. Each coefficient represents the change in the log-odds of a release associated with a one-unit increase in that predictor, holding the others fixed. The industry sector term represents a set of indicator variables created from the sector classification. Each coefficient in this group compares a sector's log-odds with the baseline sector. The coefficients α_F, α_C, α_K, and α_P describe how the log-odds of reporting any release differ for federal facilities, for chemicals regulated under the Clean Air Act, for carcinogens, and for PBT chemicals, compared with their reference categories. To interpret these coefficients in terms of odds, we exponentiate them. For example, $\exp(\alpha_P)$ is the factor by which the odds of reporting any on-site release change when the record involves a PBT chemical rather than a non-PBT chemical, holding the other predictors fixed.

3.4 Assumptions of the Logistic Regression Model

1. **Choosing the right link function** :The logit link should describe how the predictors relate to the probability of a release. This means the model should capture the main pattern in how the predictors influence the likelihood of reporting any release.
2. **Independence** : Each facility–chemical record is treated as a separate observation, and the outcomes are assumed to be independent across records.
3. **Linearity in the logit** : For numerical predictors, the log-odds should change in a roughly linear way with the predictor. This is checked using plots of the logit against each numerical variable.

As with the multiple linear regression model, these assumptions are evaluated using diagnostic checks.

3.5 Software Used

All analyses were conducted in R 4.5 (R Core Team 2024) using RStudio. The TRI Basic Data File was imported from CSV format, and data-cleaning steps were carried out using functions from the tidyverse package. Summary tables were produced with knitr and kableExtra. All figures, including histograms, boxplots, and bar charts, were generated using ggplot2. The multiple linear regression model was fit using the lm() function. This function returns least-squares estimates of the model coefficients along with standard errors and fitted values, which

were used to evaluate the suitability of the linear model for the continuous release totals. Logistic regression models were fit using the glm() function with a logit link. Both modeling approaches used the cleaned dataset as input, and fitted values and residuals were examined to assess whether the assumptions for each model were met.

---

4 Results

4.1 Fitted Model

4.1.1 Multiple Linear Regression

The fitted model results in Table 4 relates on-site release totals to measures of production activity, waste management, facility type, chemical hazard indicators, and industry sector. Many of the numerical predictors have large estimated coefficients because the response is measured in pounds and spans several orders of magnitude. In this context, the coefficient magnitudes should be understood relative to the scale of TRI releases.

The estimated coefficient for production waste is large and positive, indicating that facilities reporting greater production-related waste tend to report larger on-site releases, holding the other predictors constant. A similar pattern appears for one-time releases, where the positive coefficient suggests that records with larger one-time release quantities tend to have higher total on-site releases. In contrast, the coefficients for treatment on site and recycling on site are negative, implying that facilities reporting more waste treated or recycled on site tend to have lower release totals. This is consistent with the idea that treatment and recycling activities offset the need for disposal or release. Because industry sector is represented by a collection of indicator variables, each sector coefficient measures how the average release total for that sector differs from the baseline sector. For example, the coefficient for fabricated metals is positive and significant, meaning that facilities in that sector are estimated to release more, on average, than facilities in the baseline sector, after controlling for the other predictors. Binary indicators such as federal facility, carcinogen, Clean Air Act chemical, and PBT chemical have relatively small estimated effects compared with the waste-related predictors, and several are not significant at conventional levels. This suggests that once production and waste quantities are accounted for, these hazard flags contribute less to explaining variation in release totals.

Table 4: Coefficient Estimates from the Multiple Linear Regression Model

Term	Estimate	Std. Error	t value	p value
Intercept	-25410.197	215651.386	-0.118	0.906
Production Ratio	-1536.703	2028.328	-0.758	0.449
Production Waste	0.698	0.004	156.504	0.000
Treatment on Site	-0.844	0.011	-74.847	0.000
Recycling on Site	-0.709	0.007	-97.572	0.000
One-Time Release	1.342	0.191	7.042	0.000
Industry SectorChemical Wholesalers	9396.010	217722.275	0.043	0.966
Industry SectorChemicals	-70109.959	216047.255	-0.325	0.746
Industry SectorCoal Mining	-8127.128	888580.064	-0.009	0.993
Industry SectorComputers and Electronic Products	-36560.887	260508.854	-0.140	0.888
Industry SectorElectric Utilities	2969.025	228663.432	0.013	0.990
Industry SectorElectrical Equipment	60.270	238004.969	0.000	1.000
Industry SectorFabricated Metals	3820.618	218734.213	0.017	0.986
Industry SectorFood	13868.597	217075.256	0.064	0.949
Industry SectorFurniture	-51.695	296172.356	0.000	1.000
Industry SectorHazardous Waste	27358.554	359249.424	0.076	0.939
Industry SectorLeather	25410.197	646527.324	0.039	0.969
Industry SectorMachinery	2963.170	221538.145	0.013	0.989
Industry SectorMetal Mining	1120926.458	244502.846	4.585	0.000
Industry SectorMiscellaneous Manufacturing	12073.752	240586.665	0.050	0.960
Industry SectorNatural Gas Processing	36954.920	220118.627	0.168	0.867
Industry SectorNonmetallic Mineral Product	-9854.618	219076.797	-0.045	0.964
Industry SectorOther	-10416.796	227772.649	-0.046	0.964
Industry SectorPaper	-16457.712	245365.273	-0.067	0.947
Industry SectorPetroleum	-107010.267	218636.246	-0.489	0.625
Industry SectorPetroleum Bulk Terminals	5830.443	217403.769	0.027	0.979
Industry SectorPlastics and Rubber	12658.348	221058.634	0.057	0.954
Industry SectorPrimary Metals	10437.951	219929.692	0.047	0.962
Industry SectorPrinting	38701.358	347520.093	0.111	0.911
Industry SectorTextile Product	15574.150	347493.589	0.045	0.964
Industry SectorTextiles	-28728.436	329266.467	-0.087	0.930
Industry SectorTransportation Equipment	278.557	221627.052	0.001	0.999
Industry SectorWood Products	9318.469	219578.102	0.042	0.966
Federal Facility (Yes)	-119813.308	134122.472	-0.893	0.372
Clean Air Act Chemical (Yes)	33537.325	18787.429	1.785	0.074
CARCINOGENYES	-25582.665	22177.910	-1.154	0.249

PBTYES	201241.785	57428.430	3.504	0.000

Statistical significance for each coefficient is assessed using a two-sided t-test of

$$H_0 : \beta_j = 0 \qquad \text{versus} \qquad H_A : \beta_j \neq 0.$$

For example, the coefficient for production waste is significant, with a very small p-value. This test evaluates whether the relationship between production-related waste and on-site release totals is strong enough to conclude that waste quantities truly explain variation in releases, rather than the observed association being due to random variability in the error term ε_i. The small p-value provides strong evidence against H_0, indicating that production waste is an important predictor of on-site release totals in this dataset.

A contrasting example is the coefficient for federal facility. Its p-value is large, providing little evidence to reject the null hypothesis that federal and non-federal facilities have the same average release totals once the other predictors are included. In this case, the model does not provide evidence that federal status influences release magnitude beyond what is explained by production and waste variables.

4.1.2 Logistic Regression

The logistic regression model in Table 5 relates the probability of any on-site release to production activity, waste-management quantities, hazard classifications, and industry sector. Each coefficient represents the change in the log-odds of reporting a release associated with a one-unit change in the predictor, holding the other variables constant. The reduction in deviance from the null model indicates that the predictors collectively improve the explanation of release occurrence, which connects to the research question of identifying conditions under which releases are reported.

Production-related waste and one-time releases have positive coefficients, meaning that larger waste streams and non-routine release activity are associated with higher odds of reporting a release. The production ratio also increases the log-odds of a release, while recycling on site shows a negative association after adjusting for the other predictors. Treatment on site shows limited evidence of association. Several industry sectors differ from the baseline sector, reflecting processes specific to those sectors that relate to release likelihood.

Table 5: Coefficient Estimates from the Logistic Regression Model for Any On-Site Release

Term	Estimate	Std. Error	z value	p value
Intercept	-2.6491	0.8988	-2.9473	0.0032
Production Ratio	0.6780	0.0949	7.1427	0.0000

Production Waste	0.0002	0.0000	13.4944	0.0000
Treatment on Site	0.0000	0.0000	0.5273	0.5980
Recycling on Site	0.0028	0.0010	2.7592	0.0058
One-Time Release	0.0001	0.0000	3.6965	0.0002
Industry SectorChemical Wholesalers	-0.9659	0.9166	-1.0538	0.2920
Industry SectorChemicals	-0.6368	0.9008	-0.7069	0.4796
Industry SectorCoal Mining	-24.1838	356123.9998	-0.0001	0.9999
Industry SectorComputers and Electronic Products	-1.6467	1.1058	-1.4891	0.1365
Industry SectorElectric Utilities	-1.2662	0.9993	-1.2671	0.2051
Industry SectorElectrical Equipment	-1.5088	1.1137	-1.3548	0.1755
Industry SectorFabricated Metals	-2.1812	0.9520	-2.2913	0.0219
Industry SectorFood	-2.0475	0.9392	-2.1801	0.0292
Industry SectorFurniture	-1.0633	1.4677	-0.7244	0.4688
Industry SectorHazardous Waste	-0.5105	1.8486	-0.2761	0.7824
Industry SectorLeather	-23.9170	251817.6949	-0.0001	0.9999
Industry SectorMachinery	-1.1390	0.9525	-1.1958	0.2318
Industry SectorMetal Mining	0.5671	0.9808	0.5782	0.5631
Industry SectorMiscellaneous Manufacturing	-1.2008	1.1596	-1.0355	0.3004
Industry SectorNatural Gas Processing	0.3225	0.9134	0.3531	0.7240
Industry SectorNonmetallic Mineral Product	-1.8200	0.9712	-1.8740	0.0609
Industry SectorOther	-0.6226	0.9575	-0.6502	0.5155
Industry SectorPaper	0.3517	1.0265	0.3427	0.7319
Industry SectorPetroleum	0.9270	0.9059	1.0232	0.3062
Industry SectorPetroleum Bulk Terminals	-0.1990	0.9052	-0.2198	0.8260
Industry SectorPlastics and Rubber	-0.5439	0.9322	-0.5834	0.5596
Industry SectorPrimary Metals	-0.3632	0.9230	-0.3936	0.6939
Industry SectorPrinting	-174.3865	2481805.1613	-0.0001	0.9999
Industry SectorTextile Product	0.8548	1.2237	0.6986	0.4848
Industry SectorTextiles	-3.8179	5.8018	-0.6581	0.5105
Industry SectorTransportation Equipment	-1.4487	0.9644	-1.5021	0.1331
Industry SectorWood Products	-0.7291	0.9240	-0.7891	0.4301
Federal Facility (Yes)	2.8482	0.4861	5.8595	0.0000
Clean Air Act Chemical (Yes)	0.2668	0.1039	2.5687	0.0102
CARCINOGENYES	0.3681	0.1007	3.6569	0.0003
PBTYES	2.7272	0.1900	14.3525	0.0000

Logistic regression coefficients are estimated using maximum likelihood rather than ordinary least squares. Under standard regularity conditions, maximum likelihood estimators are asymptotically normal, so each coefficient $\hat{\alpha}_j$ has an approximately normal sampling dis-

tribution for large samples. Because the variance is obtained from the likelihood rather than from an estimate of the error variance, the appropriate test statistic is:

$$z = \frac{\hat{\alpha}_j}{\mathrm{SE}(\hat{\alpha}_j)},$$

This is compared to the standard normal distribution under $H_0 : \alpha_j = 0$. This is why logistic regression uses a z-test, while multiple linear regression uses a t-test. Therefore, the statistical significance for each coefficient in the logistic regression model:

$$H_0 : \alpha_j = 0 \qquad \text{versus} \qquad H_A : \alpha_j \neq 0.$$

This test evaluates whether predictor X_j is associated with a change in the log-odds of reporting any on-site release after controlling for the other variables. For example, the coefficient for production waste is large and significant, with a very small p-value. This provides strong evidence against H_0, indicating that production waste meaningfully increases the odds that a release occurs. In practical terms, facilities with large production-related waste streams are much more likely to report at least one release.

In contrast, the coefficient for federal facility has a large p-value, indicating little evidence to reject the null hypothesis that federal and non-federal facilities have the same release probability once production and waste-management variables are included. Here the data do not support the claim that federal status alone affects the likelihood of reporting a release.

Hazard indicators also contribute to explaining release occurrence. Records involving PBT chemicals and Clean Air Act chemicals have higher log-odds of reporting a release, while carcinogens do not differ meaningfully from non-carcinogens once the other variables are controlled. To illustrate the interpretation of a specific hypothesis test, the Wald test (L. Regression 2002) for the PBT coefficient evaluates:

$$H_0 : \ \alpha_P = 0 \qquad \text{versus} \qquad H_A : \ \alpha_P \neq 0.$$

A small p-value provides evidence against H_0, and the odds ratio $\exp(\alpha_P)$ shows how the odds of reporting any on-site release differ for records involving PBT chemicals after adjusting for production, waste, and industry sector. These results support the research objective by identifying facility-level and chemical-level characteristics that relate to whether releases occur.

4.2 Goodness of fit and interpretation

The fitted-value plot in Figure 4 (full scale) shows that the multiple linear regression model does not capture meaningful variation in on-site release totals. Nearly all predicted values lie close to zero while observed values span several orders of magnitude. The diagonal reference line represents perfect prediction, yet almost no points fall near it. This pattern indicates that the model systematically underpredicts release totals across the entire range of the data. In Figure 5 (zoomed view), which restricts the axes to focus on lower release amounts, the predictions remain tightly compressed near zero while observed values vary widely. Even within this range, the model fails to translate changes in production waste, treatment, recycling, or one-time releases into corresponding changes in predicted release totals. The fitted values collapse to a narrow band of near-zero predictions despite several predictors being significant, showing that the assumed linear form cannot match the heavy-tailed, zero-inflated structure of TRI release data.

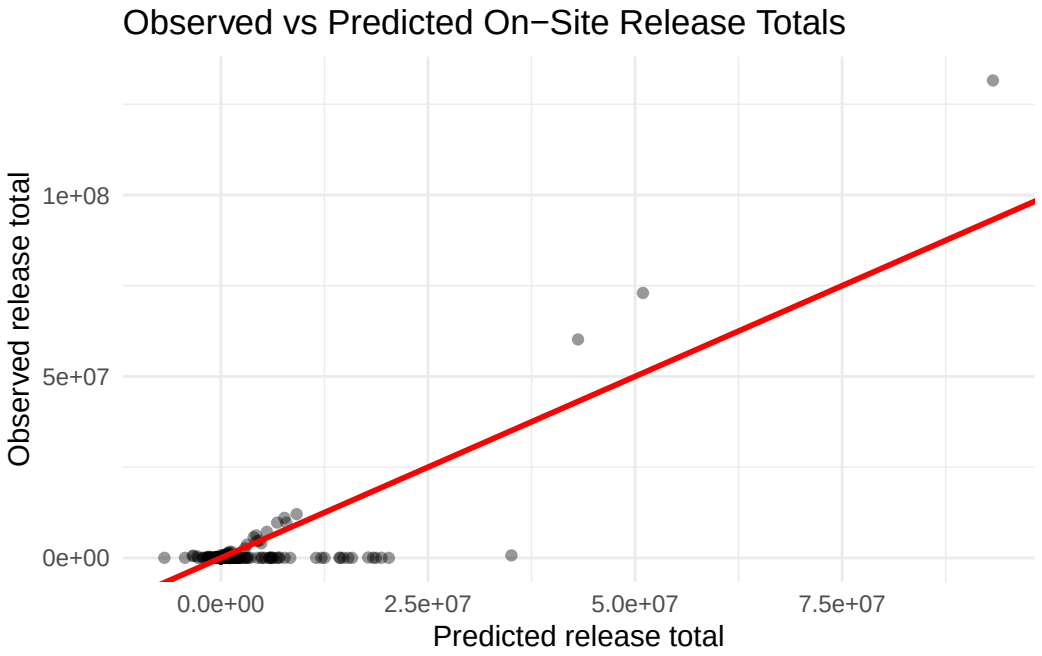

Figure 4: Observed on-site release totals plotted against predicted values from the multiple linear regression model.

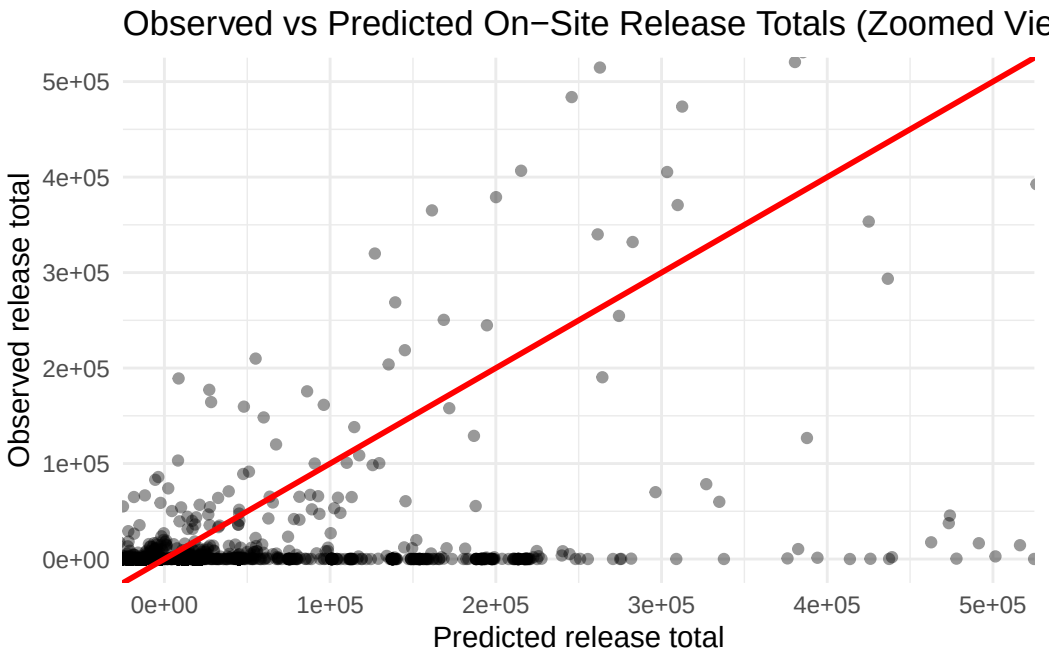

Figure 5: Observed on-site release totals plotted against predicted values from the multiple linear regression model (Zoomed in)

These fitted-value patterns help in identifying which facility and chemical characteristics predict release behavior. Although the linear regression output suggests associations between release totals and predictors such as production waste and one-time releases, the fitted-value plots in Figure 4 and Figure 5 show that the model cannot reproduce meaningful variation in the response. Because the model produces nearly identical predictions across all levels of the predictors, it cannot be used to draw substantive conclusions about how operational or chemical characteristics relate to release magnitude. This demonstrates that modeling continuous release totals with a Gaussian linear model is not appropriate for the structure of the TRI data. These results motivate the use of logistic regression, which models the probability of any release occurring and aligns more closely with the underlying distribution.

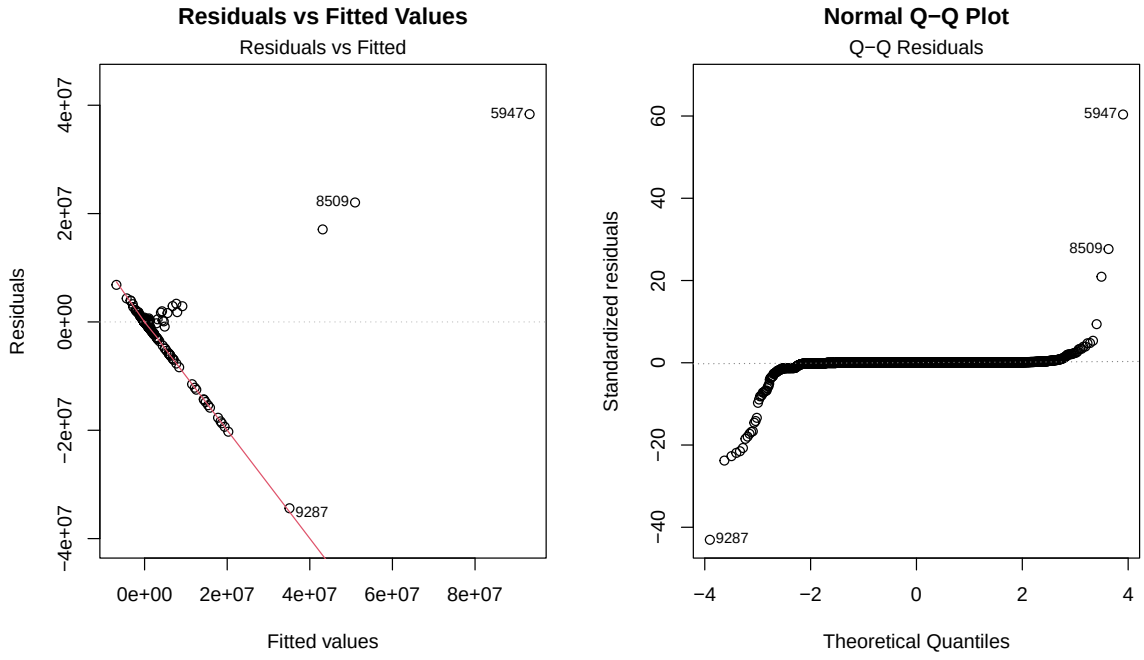

Figure 6: Diagnostic plots for the regression model. The left panel shows residuals versus fitted values, used to check linearity and constant variance of errors. The right panel is a Q–Q plot assessing whether the residuals follow an approximately normal distribution.

The residual diagnostics in Figure 6 (Residuals vs Fitted Values) show patterns that indicate the multiple linear regression assumptions do not hold for the TRI release data. The residuals form a curved, downward-sloping band rather than a cloud centered around zero, which signals that the linear functional form does not describe how the predictors relate to release totals. The spread of the residuals also increases with the fitted values, indicating non-constant variance. Several points with large positive or negative residuals appear far from the main cluster, showing the effect of release totals on the fitted model. These patterns indicate that the model does not account for the distributional structure of the response and that deviations from the fitted line are not random. Because the residuals do not behave as independent, constant-variance errors, the fitted values and coefficient interpretations are not reliable for studying the magnitude of releases.

The Q–Q plot in Figure 6 (Normal Q–Q Plot of Residuals) adds to this justification. The residuals depart from the reference line across the entire range of theoretical quantiles, with deviations in both tails. This pattern means that the error distribution differs substantially from normality, a requirement for valid inference in linear regression. The few large residuals correspond to similar observations seen in the residuals-versus-fitted plot, indicating that a

small number of large releases affect the model. Because most observations have zero or near-zero releases while a small group has large values, the linear model cannot satisfy its distributional assumptions, even after controlling for production activity, waste-management measures, and chemical characteristics. The model does not capture the structure of the data and therefore cannot be used to draw conclusions about the relationship between releases and the predictors. These limitations motivate shifting from modeling how much is released to modeling whether any release occurs. This shift aligns with the distribution of the data and leads directly to the logistic regression model, whose fitted results and diagnostics provide a more suitable basis for examining the determinants of on-site releases.

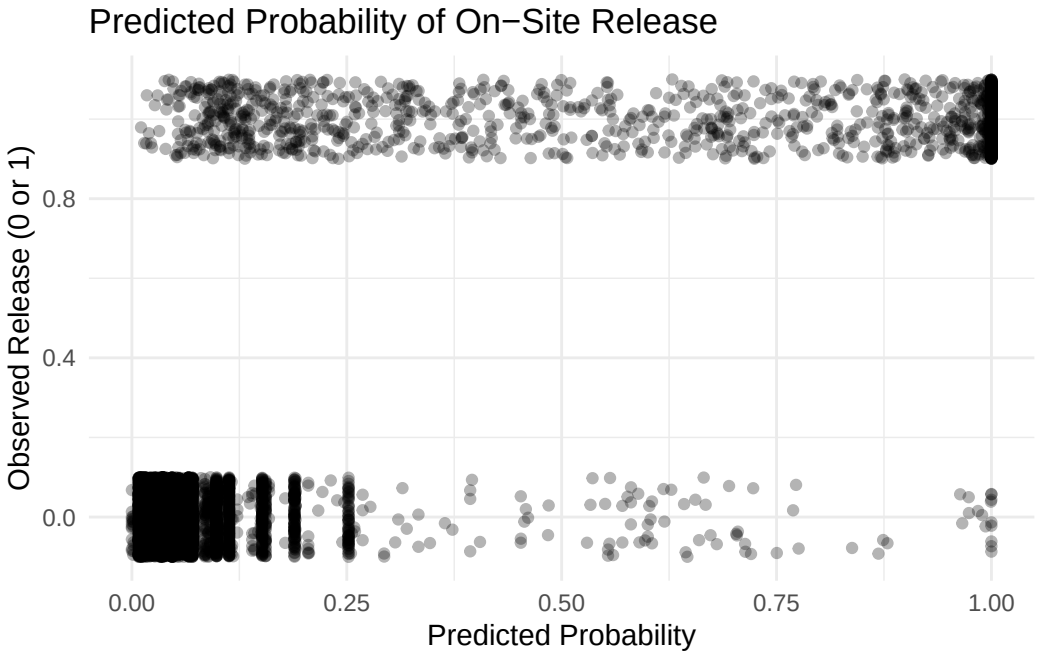

Figure 7: Predicted probabilities from the logistic regression model plotted against the observed binary outcomes.

The logistic regression diagnostics in Figure 7 show how well the model distinguishes facility–chemical records with releases from those without. The plot of predicted probabilities against the observed outcomes shows separation, where most observations with RELEASE_FLAG = 1 occur at higher predicted probabilities, while most observations with RELEASE_FLAG = 0 occur at lower predicted probabilities. This indicates that the predictors provide information about whether a release occurs, which directly addresses the research question of identifying factors associated with the presence of any on-site release. In contrast to the multiple linear regression model, estimated probabilities are not compressed near zero, and the model captures systematic differences between release-positive and release-negative records.

The calibration plot Figure 8 summarizes how closely predicted probabilities match observed release frequencies. The plot groups fitted probabilities into bins and compares the mean predicted probability with the actual proportion of releases within each bin. A well-calibrated model produces points near the 45-degree reference line. In this case, observed release rates rise steadily with predicted probabilities, and most points lie near the reference line. This indicates that the estimated probabilities align with frequencies, providing evidence that the model captures the pattern in how production and waste measures, hazard indicators, and industry sector relate to release occurrence. These diagnostics show why logistic regression is better than the multiple linear regression model. The outcome is binary, the probability scale is bounded between zero and one, and the logit link handles the uneven distribution of release outcomes more effectively than a continuous Gaussian model. The logistic regression model also avoids the assumption violations seen in the MLR residual plots and provides predictions that correspond to observed release behavior. Since the research question focuses on understanding which facility and chemical characteristics are associated with reporting any on-site release, these results show that logistic regression provides a more appropriate framework for interpretation and inference.

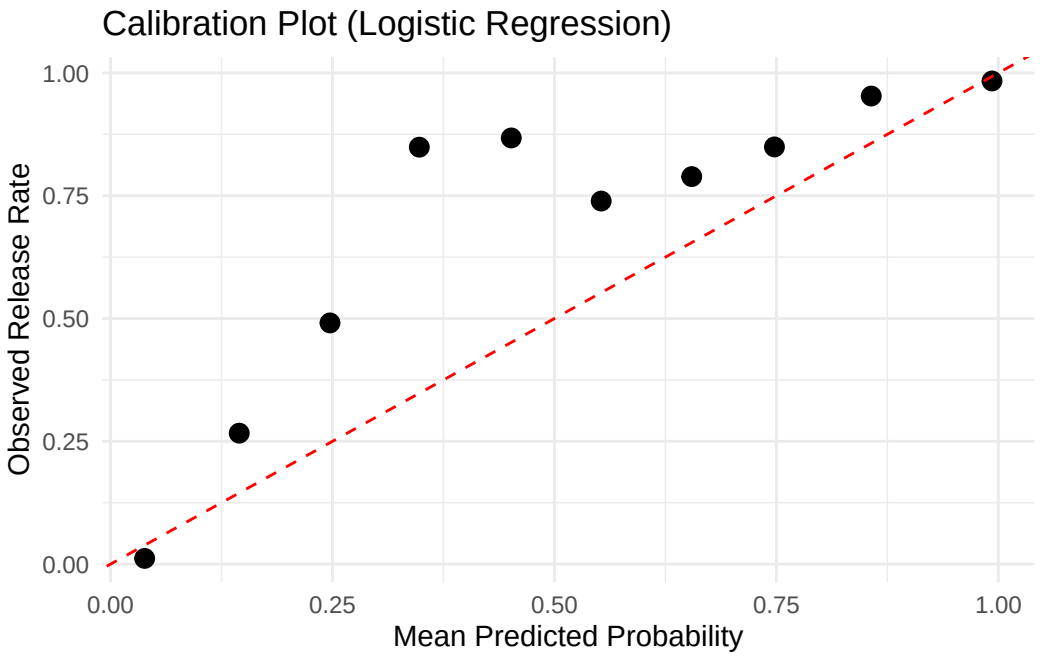

Figure 8: Calibration plot comparing mean predicted probabilities with observed release rates across probability bins.

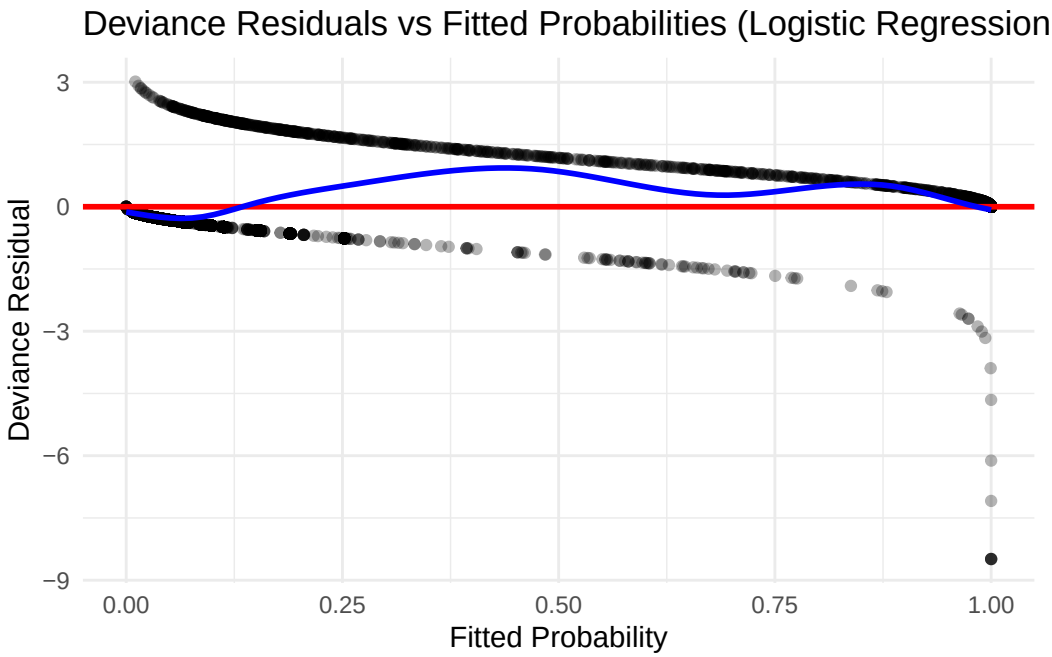

Figure 9: Deviance residuals plotted against fitted probabilities for the logistic regression model.

The logistic regression model provides a better fit than the linear regression model as it targets a binary outcome (whether any on-site release occurred) rather than attempting to model skewed release totals. Deviance residuals are used instead of ordinary residuals because logistic regression does not model additive normal errors. Deviance residuals measure each observation's contribution to the log-likelihood and provide the correct scale for diagnosing lack of fit in a binary-outcome generalized linear model. The deviance residual plot in Figure 9 shows residuals centered around zero across most fitted probabilities, with mild curvature in the smoothed trend line. Although a cluster of large negative residuals appears near predicted probabilities close to one, this behavior is typical when the model encounters observations that contradict strong predicted probabilities. The residual structure is more interpretable than in the linear model, whose residuals displayed magnitude and nonlinearity. Hypothesis tests based on z-statistics indicate that several predictors such as production waste and one-time releases significantly increase the likelihood of reporting any release. Others, such as federal facility status, show no evidence of association after adjusting for production and chemical characteristics.

5 Discussion

The purpose of this study was to examine which facility-, operational-, and chemical-level characteristics are associated with the presence of any on-site release in the 2024 Toxics Release Inventory. The analysis compared two modeling approaches: a multiple linear regression model for continuous release totals and a logistic regression model for the binary outcome of whether any release occurred. The results show that the continuous-release model does not provide a suitable framework for this dataset, while the logistic model offers a more interpretable description of release behavior.

The multiple linear regression model could not capture meaningful variation in the response. Diagnostic plots showed strong deviations from linear-model assumptions, including nonlinearity, non-constant variance, and non-normal residuals. The fitted values were compressed near zero across the entire range of release amounts, indicating that the model could not represent the heavy-tailed and zero-inflated structure of the TRI release totals. As a result, although several predictors had significant coefficients, the model did not yield reliable or useful predictions. These limitations demonstrate that a linear Gaussian framework is not appropriate for modeling release magnitude in this context.

Shifting the focus to the probability of any release through logistic regression addressed some of these issues. The logistic model aligned the response scale with the form of the data and offered a clearer interpretation of how facility characteristics relate to release occurrence. The results show that production-related waste, one-time releases, and a higher production ratio are associated with increased odds of reporting a release. Several industry sectors also differ from the baseline sector, indicating that release patterns vary across broad categories of industrial activity. Hazard indicators, including PBT status and Clean Air Act chemical designation, contribute additional information about release likelihood. Together, these findings provide evidence that both operational scale and chemical characteristics help explain when releases are reported.

The model diagnostics support the use of logistic regression for this problem. Predicted probabilities were calibrated to observed release frequencies, and deviance residuals did not show the systematic patterns observed in the multiple linear regression residual plots. This indicates that the logistic model fits the data adequately and that its coefficient estimates can be interpreted in relation to the research question. Since the goal of the study was to identify predictors of release occurrence rather than to model release quantity, the logistic model aligns well with the question being addressed.

Several limitations should be noted. The analysis relies on the subset of facility–chemical records with complete reporting across all variables. Differences in reporting quality across facilities may affect the results, and the model does not account for potential correlations among chemicals reported by the same facility. In addition, the predictors included in the model capture broad features of production and hazard classification but do not include more detailed operational or regulatory information that may also influence release behavior.

Future work could extend this analysis in several ways. Penalized logistic regression methods such as LASSO or ridge regression could be used to manage the large number of sector indicators and to evaluate variable selection in a more formal way. Interaction effects could also explain the hidden relations with the chosen predictor variables. Alternative link functions or generalized additive models could capture possible nonlinear relationships between production measures and release probability.

References

Echo Dataset. 2024. "Enforcement and Compliance History Online (ECHO)." https://echo.epa.gov/.

R Core Team. 2024. "R for Statistical Computing." https://www.R-project.org/.

Regression, Logistic. 2002. "An Introduction to Logistic Regression Analysis and Reporting." https://doi.org/10.1080/00220670209598786.

Regression, Multiple Linear. 2013. "A Study on Multiple Linear Regression Analysis." https://10.1016/j.sbspro.2013.12.027.

RSEI Dataset. 2024. "Risk-Screening Environmental Indicators (RSEI) Model and Dataset." https://www.epa.gov/rsei.

TRI Dataset. 2024. "TRI National Analysis Dataset." https://www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-present.