Scientific Research Publishing

Journal of Transportation Technologies, 2026, 16(1), 54-76
https://www.scirp.org/journal/jtts
ISSN Online: 2160-0481
ISSN Print: 2160-0473

Modeling Incident Duration Using Lasso and Ridge Regressions

Zhubin Najafi, Hualiang Teng

Department of Civil and Environmental Engineering and Construction, University of Nevada, Las Vegas, Las Vegas, USA
Email: zhubin@unlv.nevada.edu, hualiang.teng@unlv.edu

How to cite this paper: Najafi, Z. and Teng, H.L. (2026) Modeling Incident Duration Using Lasso and Ridge Regressions. *Journal of Transportation Technologies*, **16**, 54-76.
https://doi.org/10.4236/jtts.2026.161004

Received: October 10, 2025
Accepted: December 5, 2025
Published: December 8, 2025

Abstract

Traffic incidents significantly disrupt freeway operations, causing delays, congestion, fuel waste, and economic losses. Effective incident management requires not only rapid detection and clearance but also accurate real-time prediction of incident duration, a capability currently lacking in most Traffic Management Centers (TMCs). This study develops machine learning models to predict incident duration based on real-time responses and evolving incident conditions. The analysis uses incident data from the I-15 corridor in Las Vegas, Nevada, encompassing 643 recorded incidents, of which 272 were further documented in a novel Video Snapshots Dataset (VSDS) that captures 15-second visual records of incident characteristics. Key incident attributes, including total and average blockage duration, were extracted to enrich model training. Three predictive approaches were evaluated: Lasso Regression and Ridge Regression with 10-fold cross-validation, with comparison with Multiple Linear Regression. Among these, Ridge Regression achieved the highest predictive accuracy, demonstrating its effectiveness for real-time estimation of incident duration. These findings provide a foundation for enhancing TMC operations by enabling more reliable travel time updates and proactive traffic management strategies.

Keywords

Incident Duration Prediction, Incident Response Modeling, Incident Management, Machine Learning, Lasso Regression, Ridge Regression

1. Introduction

Highway incidents pose substantial challenges to roadway operations, safety, and the economy. Effective analysis of these incidents is essential for planning and implementing incident management strategies that mitigate their impacts. Ac-

cording to the National Highway Traffic Safety Administration (NHTSA), approximately 6.3 million police-reported incidents occurred in 2015, resulting in 2.5 million injuries and more than 35,000 fatalities, an average of 96 per day [1]. Between 2012 and 2021, fatal traffic incidents in the United States increased by 27 percent [2]. Beyond the human toll, incidents are a primary contributor to non-recurring congestion, generating millions of dollars annually in costs associated with delays, property damage, energy consumption, and lost productivity

Incident clearance is highly variable, depending on time of day, location, and incident severity. Blockages may last minutes or several hours, creating substantial uncertainty for travelers. To address this, Departments of Transportation (DOTs) have implemented incident management strategies aimed at reducing incident frequency and severity, as well as ensuring safe and timely clearance. Such strategies improve safety, reduce secondary incidents, alleviate congestion, and improve overall travel time.

Over the past two decades, incident duration prediction has been a major research focus, driven by its importance for mitigating congestion, delays, and safety risks. Numerous factors influence incident duration, including lane blockages, weather conditions, time of day, incident severity, and emergency response actions, making accurate prediction challenging. Most studies define incident duration as the first three phases of incident management, ranging from detection to clearance, while excluding the recovery phase, as shown in **Figure 1**. This approach streamlines analysis and aligns with traffic management practices [3] [4].

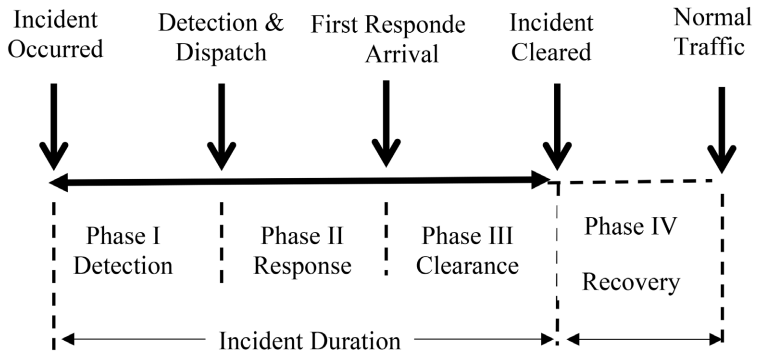

Figure 1. Four components of incident management.

Early studies used statistical regression techniques such as Multiple Linear Regression (MLR) to quantify the effects of factors such as number of vehicles involved, lane closures, weather, and time of day on incident duration [5]-[7]. Later extensions included stepwise regression [8], time sequential methods [9], structural equation models [10], quantile regression [11], probit models, and switching regression [12]. Hazard-based duration models (HBDM), an adaptation of survival analysis, were also widely applied, providing insights into how covariates affect the probability of incident clearance at a given time [13] [14]. While these approaches offered interpretability and statistical rigor, they struggled with nonlinear relationships and missing or imprecise TMC records.

With the rise of data-driven methods, machine learning has been increasingly adopted for incident duration prediction. Support Vector Machines (SVM/SVR) is effective in handling nonlinear relationships and high-dimensional data, often outperforming traditional regression when feature interactions are complex [4] [6] [7] [15]. Artificial Neural Networks (ANN) is applied to capture nonlinear dependencies and adapt to evolving traffic patterns, with extensions using Genetic Algorithms (GA) to optimize architecture and parameters [4] [6] [7] [15]. K-Nearest Neighbors (KNN) is a simple yet effective method that leverages historical similarity between incidents [4] [16]. Bayesian Models & Gaussian Process Regression (GPR) provide probabilistic frameworks that provide uncertainty quantification and continuous model updating as new data arrive [15] [17]. Fuzzy Models is useful for handling uncertainty and imprecision inherent in incident records [18].

Decision Trees have been widely used due to their interpretability, though they are prone to overfitting [4] [19]. Ensemble approaches, including Random Forests (RF) and other ensemble tree methods, improve prediction accuracy by aggregating multiple weak learners, capturing nonlinearities and feature interactions while mitigating overfitting risks [20] [21].

Clustering techniques, such as cluster-based lognormal distribution models, group similar incidents and model durations within each cluster, improving performance when incident data positively skewed and non-negative, characteristics typical of incident duration data [22] [23].

Despite extensive research, existing approaches in estimating incident duration face two key limitations. First, on incident data, many models use incomplete or error-prone operator logs, which omit evolving incident conditions. Few studies account for changes in lane closures, response activities, or secondary blockages over time. Visual data such as video snapshots remain underexplored, despite their potential to enrich model accuracy. In this study, we developed a Video Snapshots Dataset (VSDS) that captures evolving incident characteristics at 15-second intervals, enabling more accurate and dynamic prediction of incident duration. Second, traditional statistical regression approaches face challenges such as multicollinearity and overfitting problems. To address these issues, Lasso and Ridge regression were applied in this study where regularization techniques are incorporated. These methods reduce variance and improve generalization.

In the remaining part of the paper, the first section is dedicated to the description of regression modeling that is used in estimating incident duration. Data collection is presented in Section 3. Section 4 describes the development of the Ridge and Lasso regression models and the linear regression model. Future work is presented in the last section.

2. Methodology

2.1. Lasso Regression

Lasso stands for "Least Absolute Shrinkage and Selection Operator". In Lasso Regression, a penalty term proportional to the absolute values of the coefficients is

added to the linear regression objective function. This penalty encourages some coefficients to become exactly zero, effectively performing variable selection and making the model sparse. Lasso is particularly useful when dealing with datasets containing many variables, as it tends to set irrelevant or redundant variables' coefficients to zero, resulting in a simpler model.

In Lasso Regression, the objective function is:

$$\text{Objective}: \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}\right)\right)^2+\lambda\sum_{j=1}^{n}\left|\beta_j\right|\right\}$$

where β_0 is the intercept, β_j are the predictor coefficients, λ is the regularization parameter, and $\lambda\sum_{j=1}^{n}\left|\beta_j\right|$ is the Lasso penalty.

A higher value of λ leads to more coefficients being pushed towards zero. When $\lambda = 0$, then the Lasso Regression penalty is also 0. That means the Lasso Regression will only minimize the sum of squared residuals and the Lasso Regression line will be the same as the Least Squares Line as both minimize the sum of squared residuals only.

2.2. Ridge Regression

Ridge Regression adds a penalty term proportional to the square of the coefficients to the linear regression objective function. This penalty discourages large coefficient values. This means that larger coefficients are penalized more, while smaller coefficients are encouraged. This leads to a more balanced and stable model.

The objective function is:

$$\text{Objective}: \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}\right)\right)^2+\lambda\sum_{j=1}^{n}\beta_j^2\right\}$$

where β_0 is the intercept, β_j are the predictor coefficients, λ is the regularization parameter, and $\lambda\sum_{j=1}^{n}\beta_j^2$ is the Ridge penalty. Unlike Lasso Regression, Ridge does not drive coefficients exactly to zero but shrinks them toward zero as λ increases. Unlike Lasso, Ridge Regression does not force coefficients to become exactly zero, only shrinks the coefficients towards zero, so it maintains all variables in the model. It can be especially helpful when dealing with multicollinearity, as it tends to distribute the impact of correlated variables more evenly.

A small λ will not have much impact on the coefficients, while a large λ will strongly shrink them. Again, when $\lambda = 0$, then the Ridge Regression penalty is also 0. That means the Ridge Regression will only minimize the sum of squared residuals and the Ridge Regression line will be the same as the Least Squares Line.

The developed Lasso and Ridge regression model developed in this study will be compared with the Multiple Linear Regression (MLR), which is expressed as:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{n}\beta_j X_j$$

A purely linear model may not be ideal in this case, given the large number of

independent variables and the potential for multicollinearity. When present, multicollinearity can cause many predictors to become statistically insignificant, often reflected in p-values greater than 0.05. Despite this limitation, it is important to compare it with Lasso and Ridge regression model.

3. Data Collection

The data needed for this study were collected through three main steps: processing the incident database, creating the Video Snapshots Dataset (VSDS), and digitizing incident characteristics.

3.1. Incident Database (IDB) Processing

The primary data source is the Incident Database (IDB) maintained by the Regional Transportation Commission of Southern Nevada's Freeway and Arterial System of Transportation (RTC FAST). The IDB contains all reported freeway incidents across the Las Vegas Valley. Since this study focuses on I-15, the database was restructured according to the following criteria:

1) Study segments: I-15 from St Rose to the Speedway.
2) Directions: Both Northbound (NB) and Southbound (SB).
3) Period: September 1, 2014 to August 31, 2015.
4) Time of occurrence: 5:00 A.M. - 8:00 P.M.
5) Time of impact: 5:00 A.M. - 10:00 P.M.

Nighttime data were excluded due to frequent maintenance activities, unreliable work-zone scheduling, and reduced visibility, which hinder digitization accuracy. In addition, incidents during nighttime hours generally have a smaller impact because of lower traffic volumes. Rainy days were also excluded because altered driving behaviors and physical conditions (e.g., reduced friction, poor visibility) affect incident characteristics differently. Given the predominantly sunny climate in Las Vegas, and consistent with the law of rare events, rainy-day incidents were considered too infrequent to be relevant for model development.

The original one-year database recorded 764 incidents on I-15. After applying the above filters, 643 incidents were retained for analysis. Proportional sampling ensured balanced representation across months.

It is important to note that the IDB is susceptible to operator errors. FAST staff visually monitor incidents via large display screens and manually enter details into the system, resulting in potential inaccuracies in incident location, lane blockages, and event timing. Therefore, all entries used in this study were carefully verified and cross-checked to improve reliability.

3.2. Video Snapshots Dataset (VSDS)

To complement operator logs, this study constructed the Video Snapshots Dataset (VSDS). The Nevada Department of Transportation (NDOT) operates freeway cameras approximately every half-mile along major corridors, providing live feeds to NDOT, police and fire departments, and other agencies for incident detection

and management. Although continuous video is not archived due to storage constraints, RTC FAST stores 15-second video snapshots of most incidents in its Performance Monitoring and Measurement System (PMMS).

These snapshots were retrieved and systematically reviewed. A minimum of 20 incidents were randomly selected per month, yielding 272 incidents (42% of the 643 study cases). The VSDS thus provides a detailed visual record of incident characteristics. Figure 2 shows one of the snapshots in the series of video records for Accident No. 13732, used as an example in constructing the dataset.

Incident recorded on 02/09/2015.

The snapshot series spans from 2:36 PM to 4:37 PM, with the current frame captured at 2:39:15 PM.

At this time of the incident one Freeway Safety Patrol (FSP) unit was on scene, lane 5 (the right regular lane) was blocked, and three vehicles were involved.

Figure 2. First video snapshot for Accident No. 13732.

3.3. Digitization of Incident Characteristics

Each selected incident was digitized by extracting detailed characteristics from the video snapshots. The recorded information included incident timing, lane blockage sequence and duration, responder arrival/departure times, and vehicle classifications. In cases where vehicles could not be identified as Freeway Service Patrol (FSP), Nevada Highway Patrol (NHP), Emergency Vehicles (EV), or Tow Trucks (TOW), they were recorded as miscellaneous (MISC).

Table 1 summarizes digitized data for Accident No. 13732. Additional analyses were also performed on the VSDS, such as computing total blockage duration, average blockage duration, and related measures across all 272 incidents. A sample analysis for Accident No. 13732 is shown in Table 2.

Table 1. Incident characteristics summarized in the video snapshots dataset.

Incident Condition	Accident No	13732
	Date Stamp	02/09/2015
	Roadway ID	99
	Segment ID	1
	Severity	Severe
	Truck Involved	No
	Quick Clearance	No

Continued

	Veh Moved By Itself			No
	Injury			No
	Validity of Snapshots			High
Road Condition	Direction			NB
	Existing # of Lanes			5
	Existing Shoulder			R
Location Condition	Incident Lane Location			5
	Blockage Lane's Type			R Reg Lane(s)
	Number of Thru Lanes Blocked			2
Timing	Meridiem	PM		
	Incident Start Time			2:36:00
	Arrival Time	FSP	1	2:39:15
		NHP	1	2:46:00
		MISC	1	2:54:45
		MISC	2	3:04:30
		NHP	2	3:05:15
		NHP	3	3:05:45
		TOW	1	3:15:45
	Moving to Shoulder Time			2:53:00
	Departure Time	MISC	1	2:56:00
		MISC	2	3:06:45
		FSP	1	3:07:00
		NHP	1	3:12:45
		NHP	2	3:39:30
		TOW	1	4:03:00
		NHP	3	4:11:45
	Partial Blockage Lane	A	5	2:36:00
	Partial Opening Lane	A	5	2:55:45
	Partial Blockage Lane	B	4	2:41:00
	Partial Opening Lane	B	4	2:54:15
	All Lanes' Opening Time			2:55:45
	Incident End Time			4:03:00
	Snapshots End Time			4:37:00

Table 2. Data analysis summary of incidents in the video snapshots dataset.

Data Analysis			Comment
	Incident End Check	0	1 if Incident End Time = N/A, 0 otherwise.
	No. of Vehicles Remained	0	Number of vehicles remained in the incident scene by the end of snapshots
	Incident Duration	87.00	minutes
	Number of Thru Lanes Blocked	2	
	Total Blockage Duration	33.00	minutes
	Avg. Blockage Duration	16.50	minutes
	No. of FSP	1	# of Freeway Safety Patrol vehicles arrived
	First Arrival	3.25	minutes from the start of the incident
	Avg. Arrival	3.25	minutes from the start of the incident
	Avg. Departure	31.00	minutes from the start of the incident
	Avg. Duration	27.75	minutes
	No. of NHP	3	# of Nevada Highway Patrol vehicles arrived
	First Arrival	10.00	minutes from the start of the incident
	Avg. Arrival	23.00	minutes from the start of the incident
	Avg. Departure	65.33	minutes from the start of the incident
	Avg. Duration	42.33	minutes
	No. of EV	0	# of Emergency Vehicles arrived
	First Arrival	-	minutes from the start of the incident
	Avg. Arrival	-	minutes from the start of the incident
	Avg. Departure	-	minutes from the start of the incident
	Avg. Duration	-	minutes
	No. of TOW	1	# of Tow trucks arrived
	First Arrival	39.75	minutes from the start of the incident
	Avg. Arrival	39.75	minutes from the start of the incident
	Avg. Departure	87.00	minutes from the start of the incident
	Avg. Duration	47.25	minutes
	No. of MISC	2	# of Miscellaneous vehicles arrived
	First Arrival	18.75	minutes from the start of the incident
	Avg. Arrival	23.63	minutes from the start of the incident
	Avg. Departure	25.38	minutes from the start of the incident
	Avg. Duration	1.75	minutes

4. Lasso Regression, Ridge Regression and Linear Regression Model Development

In modeling incident duration, a broad set of candidate predictors was initially considered to capture the complex dynamics of traffic incidents and their impacts on roadway operations. However, not all variables contributed meaningful information or were suitable for modeling. Variables were therefore screened for relevance and reliability. Categorical attributes were transformed into numerical representations where necessary to ensure compatibility with machine learning algorithms. Administrative fields; no predictive value.

4.1. Excluded Variables

Several variables were excluded from the model. Table 3 summarizes these variables and their rationale:

Table 3. Excluded variables and rationale.

Variable(s)	Type/Category	Rationale for Exclusion
Accident No, Date Stamp, Roadway, Segment ID	Identifiers	Administrative fields with no predictive value
Severity, Quick Clearance, Veh Moved By Itself, Num of Vehicles Remained	Operator-entered attributes	Subjective, inconsistent, or unreliable inputs
Validity of Snapshots, Snapshots End Time, Incident End Check	Identifiers	Quality-control variable with no predictive value
Partial Opening Lane 1 - 6	Lane indicators	Replaced by blockage variables
Incident End Time	Time stamp	Replaced by computed Incident Duration

4.2. Included Variables

The model incorporates a range of variables that capture temporal, spatial, roadway, operational, and responder-related factors. Together, these variables provide a comprehensive representation of incident characteristics and their potential influence on incident duration. Categorical variables were transformed using one-hot encoding, allowing their inclusion in regression models without imposing artificial ordering. This preprocessing step ensures that all variable types are properly structured for analysis and effectively integrated into the modeling process.

The temporal and spatial variables include:

- Time of Day (TOD): Recoded as a binary variable Meridiem (0 = AM, 1 = PM).
- Direction: Binary variable (0 = I-15 Southbound, 1 = I-15 Northbound).
- Incident Location: Original roadway and segment identifiers were replaced by sequence IDs (SeqID), which were further grouped into seven zones, labeled *IncidentLocationZone* in the model (Table 4).

Incident location is coded into zones along I-15 as shown in Table 4.

Table 4. Defined incident location zones.

Incident Location Zone	Location	Frequency of Incidents
1	Primm to South of Silverado Ranch	1
2	Silverado Ranch to south of I-215 Interchange	1
3	I-215 Interchange to south of Tropicana Ave	29
4	Tropicana Ave to south of Sahara	88
5	Sahara to north of US 95 Interchange	135
6	D Street to south of Cheyenne	13
7	Cheyenne to Speedway	5

Roadway and lane-related variables include:

- Existing Number of Lanes: Integer variable (1 - 8).
- Existing Shoulder: Categorical variable indicates the presence of a shoulder, coded as Right, Left, Both, or None.
- Incident Lane Location: Recoded into Left, Middle, Right, or Unknown (when video snapshots begin after the vehicles have already been moved to the shoulder). Lane assignments follow roadway cross-section (Table 5).

The incident lane location is coded as shown in Table 5:

Table 5. Assigned incident lane locations.

Number of existing lanes	Incident Lane		
	Left	Middle	Right
3 lanes	1	2	3
4 lanes	1	2-3	4
5 lanes	1	2-3-4	5
6 lanes	1-2	3-4	5-6
7 lanes	1-2	3-4-5	6-7
8 lanes	1-2	3-4-5-6	7-8

Incident characteristics are coded as follows:

- Truck Involved: Binary (0 = No, 1 = Yes).
- Injury: Binary (0 = No injury, 1 = Injury reported).
- Number of Thru Lanes Blocked: Integer (0 - 8).
- Blockage Type: Categorical, with dummy variables assigned based on Table 6.

Lane blockage is represented as shown in Table 6:

The operational time-based variables include:

- Moving to Shoulder Time (Moving to Shoulder Time).

Table 6. Assigned lane blockage types.

Type of blockage	Variable in the model
L Reg Lane(s)	Blockage_LeftReg
C Reg Lane(s)	Blockage_CenterReg
R Reg Lane(s)	Blockage_RightReg
L Shoulder	Blockage_LeftShoulder
R Shoulder	Blockage_RightShoulder
Ramp	Blockage_Ramp
All	Blockage_AllLanes
L Exp Lane	Blockage_LeftExpress
R Exp Lane	Blockage_RightExpress
L & R Exp Lanes	Blockage_Left_RightExp
R Exp + L Reg Lanes	Blockage_RightExp_LeftReg
L & R Exp + L Reg Lanes	Blockage_Left_RightExp_LeftReg

- Total Blockage Duration (Total Blockage Duration).
- Average Blockage Duration (Avg Blockage Duration).
- All Lanes Opening Time (All Lanes Opening Time).
- First responder-related variables (Average Arrival and Average Departure): The presence of first responders can significantly affect traffic flow through rubbernecking and additional lane blockages caused by their vehicles. Arrival and departure information was included for FSP, NHP, EV, TOW, and MISC vehicles. To better capture operational impact, average departure times were used instead of presence durations. This choice reflects that traffic flow typically begins to recover only once responders leave the scene, not while they remain on site.

While these variables provide a comprehensive overview of the incident's impact, some exhibit high correlation (Figure 3).

4.3. Fitting Lasso Regression, Ridge Regression and MLR Models

Three regression models are fitted for comparison and evaluation: 1) Ridge Regression; 2) Lasso Regression, and Multiple Linear Regression.

4.3.1. Lasso Regression with 10-Fold Cross-Validation

In Lasso regression, the regularization parameter λ is important. The impact of λ on coefficient shrinkage is shown in Figure 4. As λ increases, many coefficients are reduced to zero, yielding a sparse model that retains only the most influential predictors.

The trajectories of the colored lines illustrate how predictor coefficients shrink as λ increases. The numbers at the top of the plot (41, 37, 14, 3) indicate the count of non-zero coefficients at different values of λ. For example, at $\log(\lambda) \approx -4$, 41

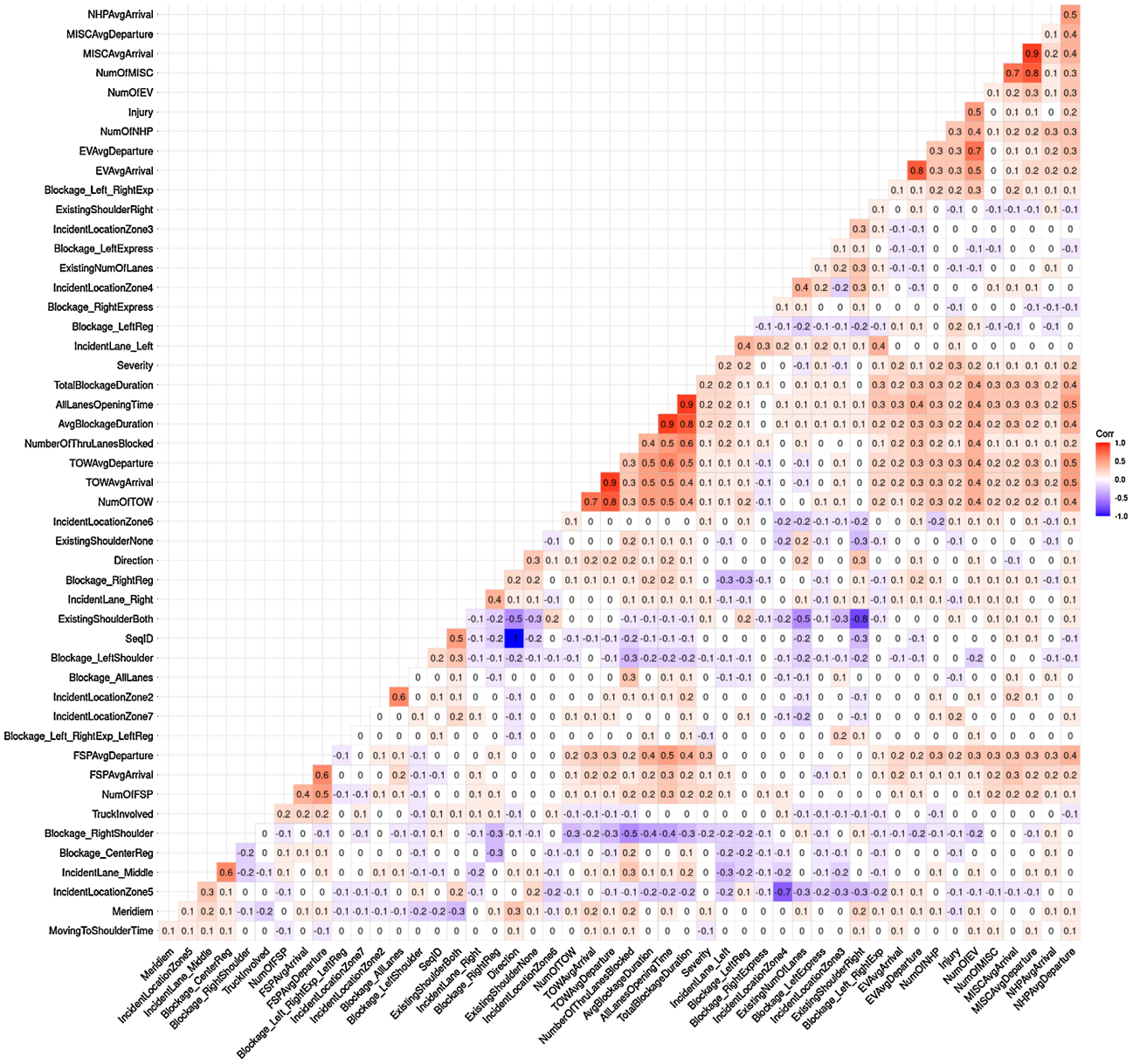

Figure 3. Correlation matrix of independent variables.

predictors remain non-zero, while at log (λ) ≈ 0, only 14 remain, demonstrating Lasso's variable selection property.

At low values of λ (left side), many predictors are retained, increasing model complexity and the risk of overfitting. At high values of λ (right side), most coefficients are driven to zero, resulting in a much simpler model that risks underfitting. Thus, the optimal λ is selected through cross-validation, which balances predictive accuracy and model parsimony.

Model performance was further assessed using 10-fold cross-validation, with Mean Squared Error (MSE) plotted against log(λ) (**Figure 5**). Each red point represents the average MSE across 10 folds, with gray error bars indicating standard error. The left vertical dotted line identifies the λ that minimizes cross-validated

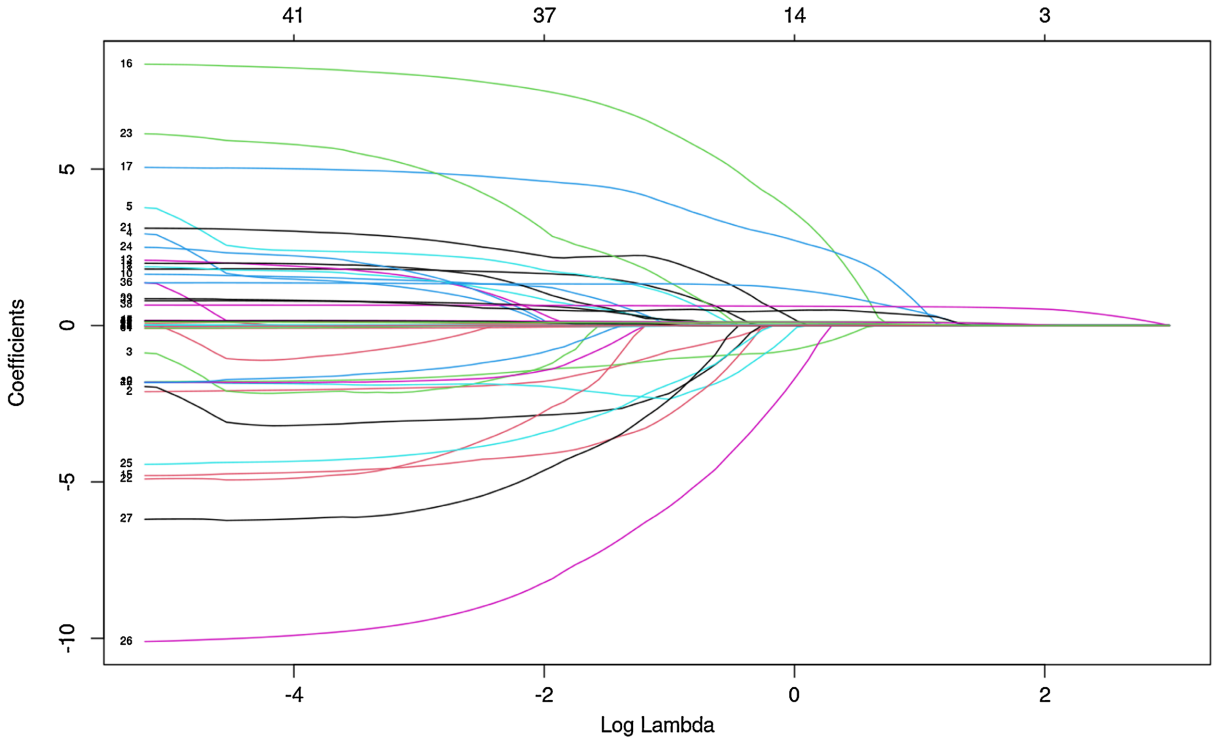

Figure 4. Impact of lambda on coefficients in lasso regression.

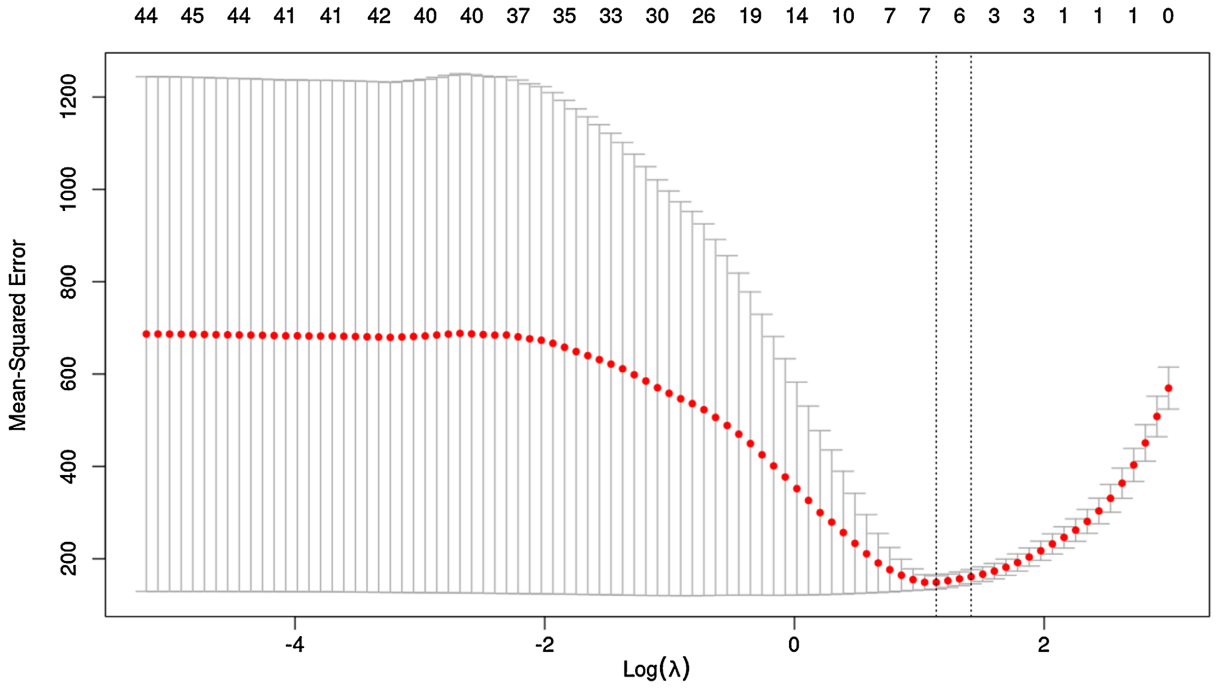

Figure 5. Mean squared error of lasso regression.

MSE (optimal λ), while the right dotted line represents the largest λ within one standard error of the minimum (1-SE λ), yielding a simpler but slightly less accurate model.

MSE is high for small values of λ due to model complexity and overfitting. MSE

decreases as regularization increases, reaching a minimum near the optimal λ. Beyond this point, MSE rises again, reflecting underfitting when too few predictors remain. Using the optimal $\lambda = 2.3549$, the minimum MSE achieved was 139.533.

4.3.2. Ridge Regression with 10-Fold Cross-Validation

In the Ridge regression, the regularization parameter λ is important as well. Figure 6 illustrates the coefficient trajectories as λ increases. Each colored line represents one of the 46 predictors plus the intercept (total 47). At small λ, coefficients are relatively large and fluctuate, which may lead to overfitting. As λ increases, coefficients shrink toward zero, reducing model complexity and mitigating overfitting. The largest coefficients at low λ likely correspond to the most influential predictors, whereas coefficients that shrink quickly are less important.

The trade-off between model complexity and regularization is evident: small λ values allow the model to fit the data closely but risk overfitting, while large λ values produce simpler models that may underfit. The optimal λ typically lies between these extremes, balancing predictive accuracy and parsimony.

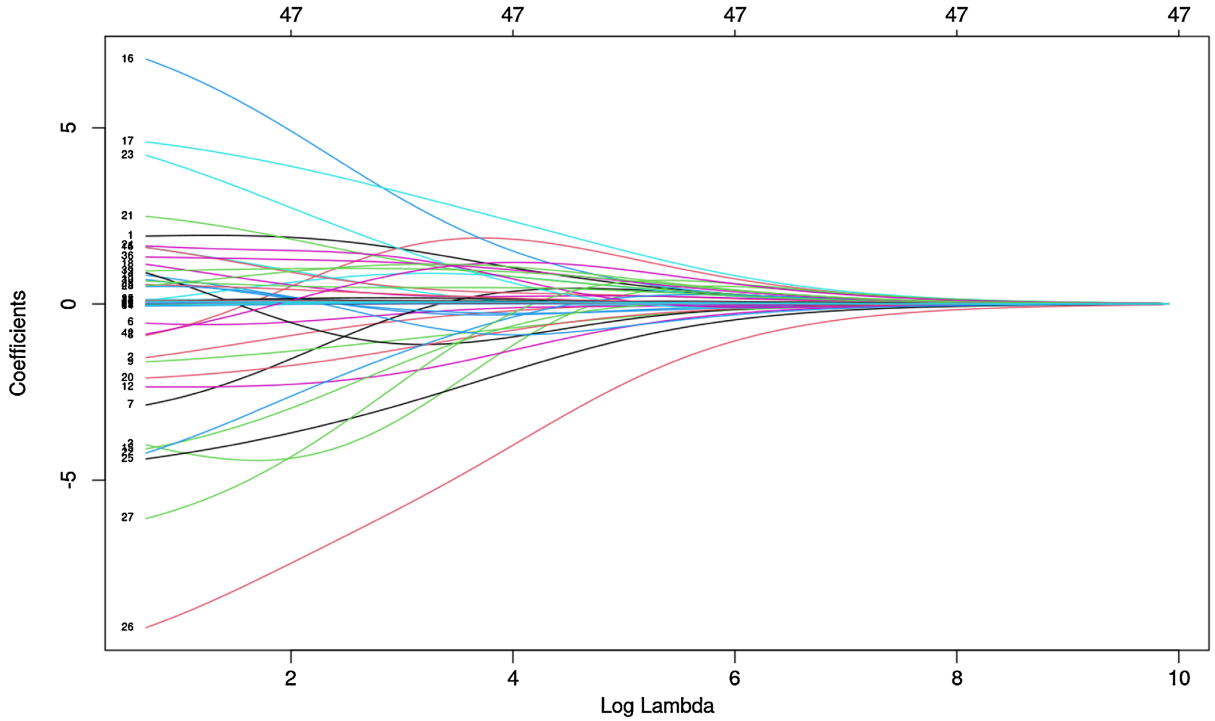

Figure 6. Impact of lambda on coefficients in ridge regression.

Figure 7 shows the Mean Squared Error (MSE) plotted against $\log(\lambda)$ using 10-fold cross-validation. Each red point represents the average MSE across folds, with gray error bars indicating the standard error. The left vertical dotted line marks the λ that minimizes cross-validated MSE (optimal λ), while the right dotted line indicates the largest λ within one standard error of the minimum (1-SE λ), which favors a simpler model with near-optimal performance.

As expected, MSE is high at low λ due to overfitting, decreases near the optimal

λ, and rises again at high λ as the model underfits. Unlike Lasso, all 47 features remain non-zero across λ values. Using the optimal λ = 9.730597, the minimum MSE achieved was 180.4374.

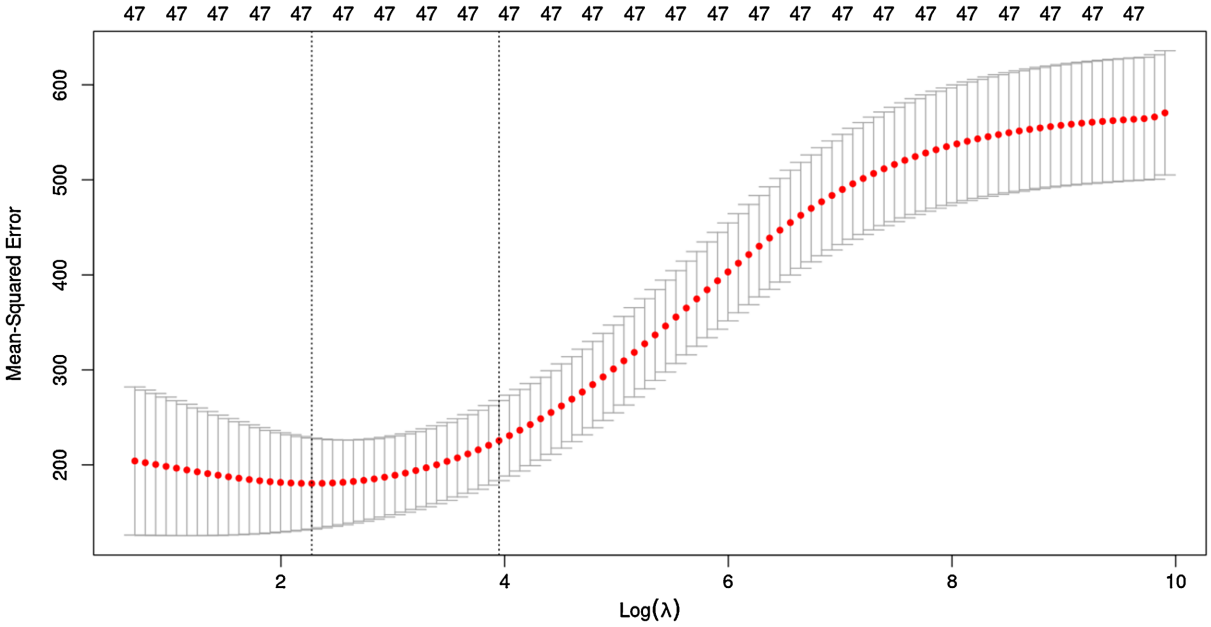

Figure 7. Mean squared error of ridge regression.

4.3.3. Multiple Linear Regression

A purely linear model may not be ideal in the case of estimating incident duration, given the large number of independent variables and the potential for multicollinearity. When present, multicollinearity can cause many predictors to become statistically insignificant, often reflected in p-values greater than 0.05. Despite this limitation, it remains important to evaluate the linearity assumption.

This is first examined through the Histogram of Residuals, which displays the distribution of errors (observed minus predicted values). The histogram appears roughly symmetric around zero but is slightly right-skewed, with a few positive outliers. Ideally, residuals should follow a normal distribution centered around zero. In this model, most residuals cluster near zero, but the skewness indicates that some predictions underestimate the observed values. The residuals range from approximately −20 to +40, reflecting variability in prediction errors (Figure 8).

Notably, the right tail of the histogram extends farther than the left, with residuals exceeding +30 or +40, indicating substantial underestimation in certain cases. This suggests the model may not be capturing some underlying patterns or nonlinear relationships.

To further assess the normality assumption, a Quantile–Quantile (Q–Q) Plot is examined (Figure 9). Between quantiles of about −1.5 and 1.5, the points align closely with the reference line, indicating approximate normality in this range. However, deviations occur in the tails: the lower quantiles (below −2) show mild

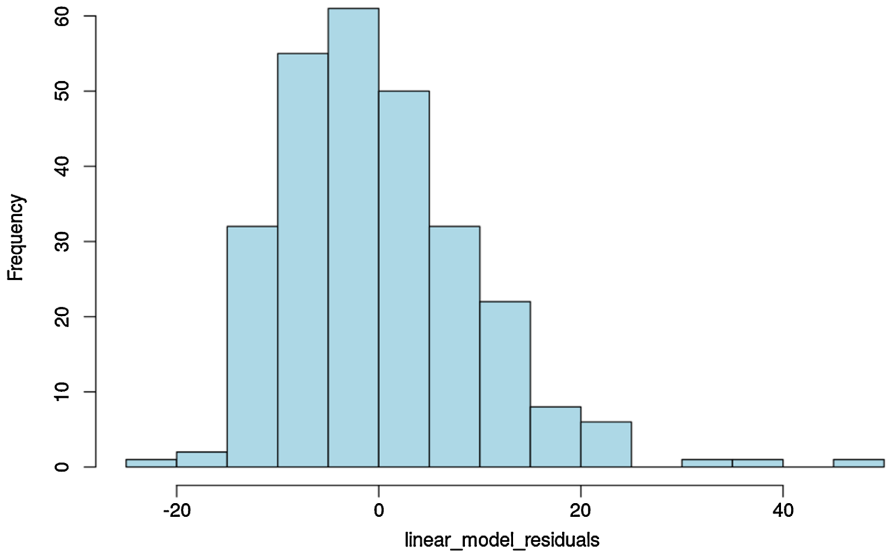

Figure 8. Histogram of residuals.

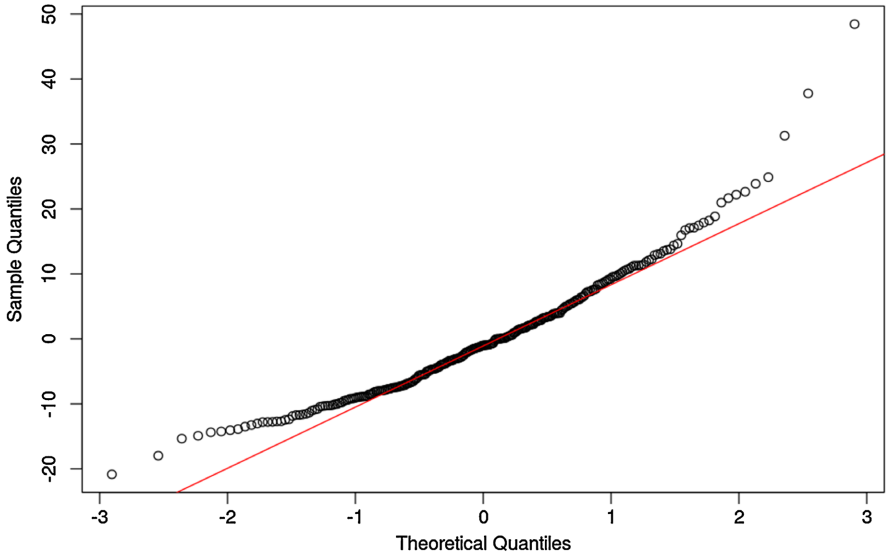

Figure 9. Normal Q-Q plot.

departures, suggesting some underestimation, while the upper quantiles (above 2) deviate strongly, reflecting large positive residuals. These departures confirm that the residuals are not perfectly normally distributed, consistent with the skew observed in the histogram. Despite these limitations, MLR provides a valuable reference point for comparing advanced techniques.

The results of the Lasso, Ridge regressions and MLR models are presented in Table 7.

Several contrasts emerge when comparing the three modeling approaches. MLR includes all predictors, but many coefficients are not statistically significant, reflecting the challenges of multicollinearity. Ridge Regression retains all variables but shrinks coefficients toward zero, stabilizing estimates in the presence of correlated predictors. Lasso Regression, in contrast, eliminates some predictors by

driving their coefficients exactly to zero, producing a sparse and more interpretable model. The relative strengths of each method are summarized in Table 8.

Table 7. Coefficients of regression models.

Independent Variables	*Coefficients*			*p-value*
	Ridge	*Lasso*	*MLR*	
(Intercept)	19.509	17.806	11.465	0.565
Meridiem	1.93	0	1.792	0.239
Direction	−1.526	0	−2.162	0.225
IncidentLocationZone2	−4.025	0	2.107	0.908
IncidentLocationZone3	0.827	0	6.014	0.597
IncidentLocationZone4	1.614	0	6.784	0.546
IncidentLocationZone5	−0.532	0	4.366	0.694
IncidentLocationZone6	−2.855	0	0.977	0.933
IncidentLocationZone7	−0.879	0	2.907	0.808
ExistingNumOfLanes	−1.64	0	−1.840	0.095.
ExistingShoulderRight	0.695	0	1.701	0.881
ExistingShoulderBoth	0.1	0	0.027	0.998
ExistingShoulderNone	−2.35	0	−1.741	0.880
IncidentLane_Left	0.879	0	2.231	0.398
IncidentLane_Middle	1.591	0	2.009	0.402
IncidentLane_Right	−4.127	0	−4.846	0.124
TruckInvolved	6.955	0	8.394	0.002**
Injury	4.598	1.161	5.056	0.018*
NumberOfThruLanesBlocked	1.12	0	1.997	0.111
Blockage_LeftReg	−0.036	0	0.065	0.995
Blockage_CenterReg	−2.094	0	−1.756	0.879
Blockage_RightReg	2.488	0	3.202	0.778
Blockage_AllLanes	−4.184	0	−4.878	0.737
Blockage_LeftShoulder	4.217	0	6.430	0.591
Blockage_RightShoulder	1.646	0	2.775	0.811
Blockage_LeftExpress	−4.392	0	−4.477	0.702
Blockage_RightExpress	−9.186	0	−10.179	0.414
Blockage_Left_RightExp	−6.077	0	−6.165	0.611
Blockage_Left_RightExp_LeftReg	0.489	0	0.000	
MovingToShoulderTime	0.0002	0	−0.002	0.747

Continued

TotalBlockageDuration	−0.011	0	−0.033	0.381
AvgBlockageDuration	−0.014	0	0.027	0.808
AllLanesOpeningTime	0.009	0	−0.022	0.812
NumOfFSP	0.65	0	0.778	0.435
FSPAvgArrival	−0.06	0	−0.096	0.224
FSPAvgDeparture	0.128	0.115	0.141	0.001***
NumOfNHP	1.33	0.556	1.363	0.053.
NHPAvgArrival	0.083	0	0.022	0.758
NHPAvgDeparture	0.553	0.601	0.652	<2e−16***
NumOfEV	0.935	0.363	0.875	0.407
EVAvgArrival	0.088	0	0.157	0.372
EVAvgDeparture	−0.031	0	−0.094	0.466
NumOfTOW	−0.862	0	−1.923	0.266
TOWAvgArrival	0.066	0	−0.007	0.939
TOWAvgDeparture	0.108	0.094	0.165	0.043*
NumOfMISC	0.521	0	0.161	0.895
MISCAvgArrival	0.025	0	−0.041	0.711
MISCAvgDeparture	0.074	0.045	0.117	0.179

Note: for MLR, significance codes are provided, reflecting p-values for hypothesis tests of whether each coefficient differs significantly from zero: ***Highly significant (0.001 level); **Very significant (0.01 level); *Significant (0.05 level); Marginally significant (0.1 level); (no symbol): Not significant.

Table 8. Key differences among regression models.

Criterion	*MLR*	*Lasso Regression*	*Ridge Regression*
Multicollinearity	Struggles	Handles it well	Handles it well
Feature Selection	No	Yes	No
Number of Predictors	Works best with fewer predictors	Works well with many predictors	Works well with many predictors
Interpretability	High	High	Moderate
Overfitting Risk	High with many predictors	Low, reduces overfitting	Low, reduces overfitting
Handling of Sparsity	Poor	Excellent	Poor
Cross-validation Performance	May overfit	Best if many features are irrelevant	Better, especially with collinearity
Computation	Fast	More computationally intensive	Moderate

Model performance was evaluated using two standard metrics: Mean Squared Error (MSE) and R-squared. MSE assesses the accuracy of predictions, with lower values indicating better predictive fit. R-squared reflects the proportion of variance explained by the model, with values closer to 1 denoting stronger explanatory power.

The choice of the "best" model ultimately depends on the analytical objective. When predictive accuracy is the priority, minimizing MSE provides the most reliable criterion. Alternatively, if the focus is on explaining variability in the data and understanding underlying relationships, R-squared becomes the more informative measure.

MSE measures the average of the squared differences between the actual values (y_j) and the predicted values ($\hat{y}_j$) produced by the model. MSE is expressed in squared units of the dependent variable (e.g., minutes² for incident duration). While this makes it less intuitive than its square root, RMSE, it is still valuable for quantifying prediction accuracy and comparing model performance. R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. In practice, both MSE and R^2 should be considered together: MSE assesses predictive accuracy, while R^2 reflects explanatory power. Table 9 summarizes the key evaluation metrics across the three models.

Table 9. Comparison of model performance metrics for incident duration estimation.

Metric	*MLR*	*Lasso Regression*	*Ridge Regression*
Best Lambda	—	2.354929	9.730597
MSE	92.40456	139.533	180.4374
RMSE	9.6127	11.8124	13.4327
SSR	25,134.04	34,667.02	26,104.61
R-squared	0.8038	0.7752895	0.8307908

Based on these results and the study objective, the Ridge Regression demonstrates superior performance and robustness. Although its MSE (180.44) is higher than that of MLR (92.40) and Lasso (139.53), Ridge achieves the highest R^2 value (0.8308), indicating it explains the largest share of variance in incident duration. Its relatively low SSR (26,104.61), particularly compared to Lasso, further supports its effectiveness.

4.4. Interpretations of Ridge Regression

The coefficients from the Ridge Regression model indicate the expected change in incident duration (in minutes) for a one-unit increase in each independent variable, holding other factors constant. In general:

- Positive coefficients suggest that an increase in the predictor is associated with longer incident durations.
- Negative coefficients suggest that an increase in the predictor is associated with

shorter incident durations.

To illustrate, two coefficients are highlighted below:

- Meridiem (+1.93): Incidents occurring in the PM period are expected to last about 1.93 minutes longer than those in the AM. This likely reflects the higher congestion during evening peak hours, which slows response times and clearance operations.
- Direction (−1.526): Incidents in the Northbound direction tend to be shorter in duration compared to Southbound incidents. This can be attributed to the heavier congestion, complex lane configurations, and event-driven traffic in the Southbound lanes near major attractions, which complicates incident management and prolongs clearance times.

These examples demonstrate how Ridge Regression captures operational and contextual factors that influence incident duration.

4.5. Model Limitations

While Ridge Regression emerged as the preferred model in this analysis, several limitations should be acknowledged. First, Ridge does not perform variable selection; all predictors remain in the model even if their contributions are minimal. This reduces interpretability. Therefore, if the goal is to identify the most critical factors influencing incident duration, approaches such as Lasso Regression may be more appropriate. However, the primary objective of this study was to estimate incident duration from real-time responses and evolving incident conditions, and Ridge Regression performs well in capturing situational changes inherent to the dynamic nature of incident scenes.

Second, although Ridge effectively addresses multicollinearity, it does not fully resolve other statistical challenges such as skewed residuals, outliers, or potential non-normality. As observed in the diagnostic analyses, predictions for extreme cases may still be inaccurate. Furthermore, while Ridge achieved the highest R^2 value, it did not produce the lowest Mean Squared Error (MSE), highlighting a tradeoff between explanatory power and predictive accuracy.

Finally, Ridge Regression relies on the assumption of linearity and remains sensitive to the scale of predictors. Complex, non-linear interactions—such as compounding effects of secondary and tertiary incidents—may not be fully captured. Moreover, the model was calibrated to a specific corridor (I-15 in Las Vegas) and time period, which may limit its generalizability to other road networks or operational contexts without re-estimation.

5. Future Work

Based upon the findings of this study, several promising directions for future research are proposed:

First, the current analysis can be extended to account for the effects of secondary and tertiary incidents, as well as other complex, non-linear dynamics that are not adequately captured by Ridge Regression. This may enhance predictive per-

formance, especially for extreme or atypical incidents.

Second, future studies should explore the integration of real-time traffic flow and sensor data, including variables such as vehicle speed, density, and queue length. The weather data such as day time, night time, raining day and snow day should also be included. Combining these with incident-specific characteristics may improve the predictive power of models and support the development of adaptive, real-time decision-support tools for traffic management centers.

Third, expanding the model beyond the I-15 corridor in Las Vegas, Nevada to include other freeways, corridors, and geographic regions would improve generalizability. Comparative studies across multiple urban areas could reveal whether the identified patterns are location-specific or reflect broader trends in urban traffic behavior.

Fourth, developing methods for variable importance and interpretability in regularized models would be valuable. While Ridge Regression performs well in prediction, it retains all predictors, limiting clarity on which factors matter most. Combining Ridge with techniques such as SHAP (SHapley Additive exPlanations) values or hybrid Ridge-Lasso frameworks could yield more interpretable models without compromising accuracy.

Finally, future research should explore the potential for real-time deployment of predictive models in traffic management operations. The incident duration, combined with other variables can be used to calculate total traffic dela. Integration with incident detection and response systems could allow TMCs to forecast incident duration dynamically and allocate resources more efficiently, ultimately reducing secondary incidents, delays, and congestion.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] NHTSA (2017) Summary of Motor Vehicle Crashes (Final Edition): 2015 Data. Report No. DOT HS 812 376, National Highway Traffic Safety Administration, U.S. Department of Transportation.

[2] NHTSA (2023) Summary of Motor Vehicle Crashes: 2021 Data. Report No. DOT HS 813 515, National Highway Traffic Safety Administration, U.S. Department of Transportation.

[3] Garib, A., Radwan, A.E. and Al-Deek, H. (1997) Estimating Magnitude and Duration of Incident Delays. *Journal of Transportation Engineering*, **123**, 459-466. https://doi.org/10.1061/(asce)0733-947x(1997)123:6(459)

[4] Valenti, G., Lelli, M. and Cucina, D. (2010) A Comparative Study of Models for the Incident Duration Prediction. *European Transport Research Review*, **2**, 103-111. https://doi.org/10.1007/s12544-010-0031-4

[5] Khattak, A., Wang, X. and Zhang, H. (2012) Incident Management Integration Tool: Dynamically Predicting Incident Durations, Secondary Incident Occurrence and Incident Delays. *IET Intelligent Transport Systems*, **6**, 204-214. https://doi.org/10.1049/iet-its.2011.0013

[6] Pereira, F.C., Rodrigues, F. and Ben-Akiva, M. (2013) Text Analysis in Incident Duration Prediction. *Transportation Research Part C*: *Emerging Technologies*, **37**, 177-192. https://doi.org/10.1016/j.trc.2013.10.002

[7] Corbally, R., Yang, L. and Malekjafarian, A. (2024) Predicting the Duration of Motorway Incidents Using Machine Learning. *European Transport Research Review*, **16**, Article No. 14. https://doi.org/10.1186/s12544-024-00632-6

[8] Yu, B. and Xia, Z. (2012) A Methodology for Freeway Incident Duration Prediction Using Computerized Historical Database. In: Fang, F.C., Wei, H. and Wang, Y.P., Eds., *CICTP* 2012: *Multimodal Transportation Systems—Convenient*, *Safe*, *Cost-Effective*, *Efficient*, American Society of Civil Engineers, 3463-3474. https://doi.org/10.1061/9780784412442.351

[9] Khattak, A.J., Schofer, J.L. and Wang, M. (1995) A Simple Time Sequential Procedure for Predicting Freeway Incident Duration. *I V H S Journal*, **2**, 113-138. https://doi.org/10.1080/10248079508903820

[10] Lee, J.Y., Chung, J.H. AND Son, B. (2009) Incident Clearance Time Analysis for Korean Freeways Using Structural Equation Model. *Proceedings of the Eastern Asia Society for Transportation Studies*, **7**. https://doi.org/10.11175/eastpro.2009.0.360.0

[11] Khattak, A.J., Liu, J., Wali, B., Li, X. and Ng, M. (2016) Modeling Traffic Incident Duration Using Quantile Regression. *Transportation Research Record*: *Journal of the Transportation Research Board*, **2554**, 139-148. https://doi.org/10.3141/2554-15

[12] Ding, C., Ma, X., Wang, Y. and Wang, Y. (2015) Exploring the Influential Factors in Incident Clearance Time: Disentangling Causation from Self-Selection Bias. *Accident Analysis & Prevention*, **85**, 58-65. https://doi.org/10.1016/j.aap.2015.08.024

[13] Qi, Y. and Teng, H. (2008) An Information-Based Time Sequential Approach to Online Incident Duration Prediction. *Journal of Intelligent Transportation Systems*, **12**, 1-12. https://doi.org/10.1080/15472450701849626

[14] Kalair, K. and Connaughton, C. (2021) Dynamic and Interpretable Hazard-Based Models of Traffic Incident Durations. *Frontiers in Future Transportation*, **2**, Article 669015. https://doi.org/10.3389/ffutr.2021.669015

[15] Mohammed, Z.A., Abdullah, M.N. and Al Hussaini, I.H. (2021) Predicting Incident Duration Based on Machine Learning Methods. *Iraqi Journal of Computers*, *Communications*, *Control and Systems Engineering*, **21**, 1-15.

[16] Kuang, L., Yan, H., Zhu, Y., Tu, S. and Fan, X. (2019) Predicting Duration of Traffic Accidents Based on Cost-Sensitive Bayesian Network and Weighted K-Nearest Neighbor. *Journal of Intelligent Transportation Systems*, **23**, 161-174. https://doi.org/10.1080/15472450.2018.1536978

[17] Zou, Y., Lin, B., Yang, X., Wu, L., Muneeb Abid, M. and Tang, J. (2021) Application of the Bayesian Model Averaging in Analyzing Freeway Traffic Incident Clearance Time for Emergency Management. *Journal of Advanced Transportation*, **2021**, Article ID: 6671983. https://doi.org/10.1155/2021/6671983

[18] Dimitriou, L. and Vlahogianni, E.I. (2015) Fuzzy Modeling of Freeway Accident Duration with Rainfall and Traffic Flow Interactions. *Analytic Methods in Accident Research*, **5**, 59-71. https://doi.org/10.1016/j.amar.2015.04.001

[19] Hamad, K., Khalil, M.A. and Alozi, A.R. (2019) Predicting Freeway Incident Duration Using Machine Learning. *International Journal of Intelligent Transportation Systems Research*, **18**, 367-380. https://doi.org/10.1007/s13177-019-00205-1

[20] Shan, L., Yang, Z., Zhang, H., Shi, R. and Kuang, L. (2019) Predicting Duration of

Traffic Accidents Based on Ensemble Learning. In: Gao, H., Wang, X., Yin, Y. and Iqbal, M., Eds., *Collaborative Computing: Networking, Applications and Worksharing*, Springer, 252-266. https://doi.org/10.1007/978-3-030-12981-1_18

[21] Hamad, K., Al-Ruzouq, R., Zeiada, W., Abu Dabous, S. and Khalil, M.A. (2020) Predicting Incident Duration Using Random Forests. *Transportmetrica A: Transport Science*, **16**, 1269-1293. https://doi.org/10.1080/23249935.2020.1733132

[22] Weng, J., Qiao, W., Qu, X. and Yan, X. (2015) Cluster-Based Lognormal Distribution Model for Accident Duration. *Transportmetrica A: Transport Science*, **11**, 345-363. https://doi.org/10.1080/23249935.2014.994687

[23] Zhao, H., Gunardi, W., Liu, Y., Kiew, C., Teng, T. and Yang, X.B. (2022) Prediction of Traffic Incident Duration Using Clustering-Based Ensemble Learning Method. *Journal of Transportation Engineering, Part A: Systems*, **148**, Article ID: 04022044. https://doi.org/10.1061/jtepbs.0000688