Open Access Library Journal
2022, Volume 9, e8873
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Modeling Vehicle Crash Frequency When Multicollinearity Exists in Vehicle Crash Data: Ridge Regression versus Ordinary Least Squares Linear Regression

Azad Abdulhafedh

University of Missouri, Columbia, Missouri, USA
Email: dr.azad.s.a@gmail.com

How to cite this paper: Abdulhafedh, A. (2022) Modeling Vehicle Crash Frequency When Multicollinearity Exists in Vehicle Crash Data: Ridge Regression versus Ordinary Least Squares Linear Regression. *Open Access Library Journal*, **9**: e8873.
https://doi.org/10.4236/oalib.1108873

Received: May 12, 2022
Accepted: June 20, 2022
Published: June 23, 2022

Abstract

Ridge Regression is an important statistical method in modeling vehicle crash frequency when crash data contains collinear predictors. The term multicollinearity refers to the condition in which two or more predictors are highly correlated with one another. This would make the explanatory variables become very sensitive to small changes in the model. Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of the regression model. Common methods to address multicollinearity include: variable selection and ridge regression. Variable selection simply entails dropping predictors that are highly correlated in the model. But sometimes this is not possible, especially when a variable that contributes to the collinearity might be a main predictor in the model. However, using ridge regression will allow retention of all explanatory variables of interest, even if they are highly collinear, and provide information regarding which coefficients are the most sensitive to multicollinearity. Ridge regression works by adding a degree of bias to the regression estimates that reduce the standard errors and produce estimates that are much more reliable. This paper uses a five-year vehicle crash data extending from 2011 to 2015 on the interstate highway (I-90) in the state of Minnesota, USA. The data has shown multicollinearity between some independent variables. Results show that the Ridge regression is an effective tool to address the existing multicollinearity and produce accurate regression estimates compared with multiple linear regression.

Subject Areas

Applied Statistical Mathematics, Mathematical Analysis

Keywords

Vehicle Crash Frequency, Ridge Regression, Ordinary Least Squares Linear Regression, Multiple Linear Regression, Multicollinearity, Shrinkage Estimator

1. Introduction

Vehicle crash data may suffer from multicollinearity when some or all independent variables in a regression model are correlated. The primary assumption in the regression analysis is that the explanatory variables should be independent from each other [1]. When multicollinearity exists, this assumption is compromised, because the variance of the regression estimates will become very large and the standard error goes up, the corresponding *t*-value goes down and hence comes up with a high p-value, which could make a significant variable insignificant by increasing its standard error. Ridge regression can be utilized to add a degree of bias to the regression estimates that reduce the standard errors and produce regression estimates that are much more reliable. If the degree of correlation between two or more variables is high enough, it can cause problems when fitting the model and insufficient results might be obtained. Obviously, the removal of any correlation between independent variables in a regression model is highly desirable because the interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each one unit change in an independent variable when holding all other independent variables constant. However, when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. Hence, it becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable because the independent variables tend to change in unison [1]-[6]. There are many ways to address multicollinearity, and each method has its pros and cons. Common methods include: variable selection and ridge regression. Variable selection will drop predictors that are highly correlated in the model [7]-[12]. However, sometimes this is not possible, especially when a variable contributing to the collinearity might be a main predictor in the model. On the other hand, using ridge regression will allow keeping all explanatory variables of interest in the model, even if they are highly collinear, and will provide information regarding which coefficients are the most sensitive to multicollinearity [13] [14].

2. Background Literature

The ridge regression technique proposed by Hoerl and Kennard in 1970 has become a common tool for analysis of data characterized with high multicollinearity. The Ridge regression method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias [15] [16]. The ridge regression was investigated by (Pasha and Shah, 2004) in multicollinear data, together with the ridge estimator's properties. By regressing the number of persons on five variables, the eigen values, variance inflation factors and standardization problem were studied through empirical comparison of OLS with ridge regression model, and some methods have been proposed for identifying the bias parameter, *k* [17]. In a study conducted by (Al-Hassan, 2008), seven approaches to estimation of the ridge parameter were examined. This research suggested a simulation approach on the basis of the minimal MSE measure. Based on the simulation approach, two estimators were proposed and found to be effective under specific conditions [18]. Other estimators of the Ridge parameter, *k*, have been introduced in the study of (Mansson *et al.*, 2010). This study considered three approaches: 1) The prediction sum of square (PRESS), MSE and maximum MSE were considered as the performance criteria; 2) Various error variances were employed (with sigma between 0.5 and 5) and; 3) The number of regressors considered ranged from 4 - 12. Based on results of the simulation, it was confirmed that augmenting correlations between independent variables leads to negative effects on the PRESS and MSE. However, raising the number of regressors has positive effects on both the PRESS and MSE. The MSE decreases as sample size is increased, even if associations between independent variables are high [19]. In spatial context where usually data have many irregularities, a study by (Lauridsen and Mur, 2006) mainly aimed at investigating this situation, and investigated the effect of multicollinearity. These researchers planned and solved a Monte Carlo simulation. It was illustrated that the extra impacts on tests of adding extra variable in general disappear for growing multicollinearity [20]. Chopra *et al.* (2013) employed ridge regression to predict the compressive strength of concrete. Values of the regression coefficients have been varied and data were reduced. They found that the traditional least squares method did not prove to be useful for forecasting the compressive strength of concrete. They concluded that the ridge regression work better in their research [21]. (Zaka and Akhter, 2013) used Relative Least Squares Method (RLSM), a ridge regression method and least squares (LSM) method to determine the parameters of power function distribution. This study employed Total Deviation and the MSE to determine the finest of the three estimators investigated. They determined the optimum estimation method on the basis of different sample sizes and values of parameters and recommended the use of the LSM method for estimating parameters of the power function distribution [22].

3. Ridge Regression VS. Ordinary Least Squares Linear Regression

Ridge regression is an effective approach to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

Ordinary least squares linear regression cannot produce accurate estimates when the number of predictors exceeds the number of observations. This leads to overfitting a model and failure to produce unique solutions. More importantly, ordinary least squares also have undesirable issues dealing with multicollinearity in data. Ridge regression works much better because it does not require unbiased estimators; while least squares produce unbiased estimates, and variances can be so large that they may be inaccurate. Ridge regression adds some bias to make the estimates reasonably reliable to real data. Ridge regression uses a type of shrinkage estimator called a ridge estimator or shrinkage estimator, which theoretically produce new estimators that are shrunk closer to the real parameters. The ridge estimator works very good at improving the least-squares estimate when multicollinearity is present. A ridge parameter (k) controls the strength of the penalty term. When $k = 0$, ridge regression equals least squares regression. If $k = \infty$, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between zero and infinity (∞). Ordinary least squares linear regression (OLS) requires that the inverse of the matrix $X'X$ exists. $X'X$ is arranged so that it represents a correlation matrix of all predictors. However, in certain situations $(X'X)^{-1}$ may not exist. Specifically, if the determinant of $X'X$ is equal to 0, then the inverse of $X'X$ does not show up. Thus, if the inverse of $X'X$ cannot be calculated, the OLS coefficients are indeterminate. In other words, the parameter estimates will have remarkably high variances and, consequently, will not be interpretable. The causes that make the $(X'X)^{-1}$ to be indeterminate, could be due to the number of parameters in the model exceeds the number of observations or the multicollinearity between the predictors. Ridge regression estimates tend to be more stable than the OLS estimates because they are little affected by small changes in the data on which the fitted model is based [23]-[31].

4. Multicollinearity

Multicollinearity is the existence of linear relationships among the independent variables that would create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, give false, non-significant p-values, and degrade the predictability of the model. The source of the multi-collinearity might come from the following [32] [33]:

- Data collection. When the data are collected from a narrow population of the independent variables, then the multicollinearity might be created by the sampling methodology. Obtaining more data on an expanded range would cure this multicollinearity problem. An example of this situation is when you try to fit a line to a single point.
- Physical constraints of the model or population. This source of multicollinearity will exist no matter what sampling technique is used. For example, some manufacturing or service processes have constraints on independent variables (as to their range), either physically, politically, or legally, which will create multicollinearity in the dataset.

- Over-defined model. In this case, there will be more variables than observations and, hence causing multicollinearity. So, this situation should be avoided.
- Model choice or specification. This may cause multicollinearity that comes from using independent variables that are powers or interactions of an original set of variables.
- Outliers. Extreme values or outliers can cause multicollinearity as well as hide it. This should be corrected by removing the outliers before ridge regression is applied.

5. Detection of Multicollinearity in the Data

The Detection of the multicollinearity in the data can be achieved by several ways as follows [31] [32] [33]:

- Visual inspection of pairwise scatter plots of independent variables and looking for near-perfect linear relationships between them.
- Considering the Variance Inflation Factors (VIF), which provide an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity. (VIFs) start at a value of (1) and have no upper limit. A value of (1) indicates that there is no correlation between this independent variable and any others. (VIFs) between (1) and (5) suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. (VIFs) greater than (10) represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.
- Considering the Eigenvalues of the correlation matrix of the independent variables, if they are near zero, then this indicates multicollinearity.
- Checking for large condition numbers (CNs) of the independent variables. The CN is calculated by taking the maximum eigenvalue and dividing it by the minimum eigenvalue. As a rule of thumb, CN > 5 indicates moderate multicollinearity. However, CN > 30 indicates severe multicollinearity.
- Investigating the signs of the regression coefficients that are produced from the ordinary least square regression, if they are opposite in sign from what one would expect, then this may indicate multicollinearity.

Depending on what the source of multicollinearity is, the solutions will vary. For example, if the multicollinearity has been created by the data collection, then try to collect additional data over a wider population. If the choice of the linear model has increased the multicollinearity, then simplify the model by using variable selection techniques. If an outlier or two has induced the multicollinearity, remove those observations. When these steps are not possible, one might try the ridge regression.

6. The Derivation of the Ridge Regression Model

Ridge regression can analyze data even when severe multicollinearity is present and helps prevent overfitting. This type of regression reduces the large, problematic variances that multicollinearity causes by introducing a small bias in the regression estimates, which produces much more accurate coefficient estimates when multicollinearity is present. Ridge regression solves the multicollinearity problem through a shrinkage parameter *k*. The assumptions of the ridge regression are the same as those used in regular multiple regression model (*i.e.*, linearity, constant variance (no outliers), and independence of variables). Since ridge regression does not provide confidence limits, normality need not be assumed.

Let us say, Y is regressed against X_1 and X_2 where X_1 and X_2 are highly correlated. Then the effect of X_1 on Y is hard to distinguish from the effect of X_2 on Y because any increase in X_1 tends to be associated with an increase in X_2. In addition, individual *t*-tests and p-values can be misleading. This means a p-value can be high which indicates that the variable is not significant, even though the variable is important and significant. The linear multiple regression equation in matrix form is [34] [35]:

$$Y = XB + \varepsilon \tag{1}$$

where,

Y: the dependent variable,

X: the vector of the independent variables,

B: the vector of the regression coefficients to be estimated,

ε : represents the residual errors.

The regression coefficients (B hat) are estimated by using the matrix formula as follows:

$$B^\wedge = \left(X'X\right)^{-1} X'Y \tag{2}$$

The ridge regression penalizes the size of the regression coefficients, and since the variables in ridge regression are standardized then:

$$X'X = R \tag{3}$$

where,

R: the correlation matrix of the independent variables.

The variance-covariance matrix of the estimates in ridge regression is:

$$V\left(B^\wedge\right) = \sigma^2 R - 1 \tag{4}$$

For the standardized variables, $\sigma^2 = 1$, and therefore:

$$V\left(B^\wedge\right) = \frac{1}{R_j^2} \tag{5}$$

where,

R_j^2: the R-squared value obtained from regressing X_j on the other independent variables.

Ridge regression proceeds by adding a small value, *k*, to the diagonal elements of the correlation matrix (Marquardt and Snee 1975) as follows:

$$B\sim = \left(R + kI\right)^{-1} X'Y \tag{6}$$

where,

k: the shrinkage parameter of the ridge regression, $0 < k < 1.0$.

I: the identity matrix.

The estimated ridge coefficient and the amount of bias in this estimator are given by:

$$E\left(B\sim-B\right)=\left[\left(X'X+kI\right)^{-1}X'X-I\right]B \tag{7}$$

The ridge regression has the effect of shrinking the estimates toward zero introducing bias but reducing the variance of the estimate. The ridge covariance matrix can now be written as:

$$V\left(B\sim\right)=\left(X'X+kI\right)^{-1}X'X\left(X'X+kI\right)^{-1} \tag{8}$$

In order to choose an appropriate value of *k*, Hoerl and Kennard (1970), the inventors of ridge regression, suggested using a graphic which they called the ridge trace. This plot shows the ridge regression coefficients as a function of *k*. When viewing the ridge trace, the analyst picks a value for *k* for which the regression coefficients have stabilized. Hoerl and Kennard (1970) proved that there is always a value of $k > 0$ such that the mean square error (MSE) is smaller than the MSE obtained using OLS. Often, the regression coefficients will vary widely for small values of *k* and then stabilize. Choose the smallest value of *k* possible (which introduces the smallest bias) after which the regression coefficients have seem to remain constant. Note that increasing *k* will eventually drive the regression coefficients to zero. To obtain the first value of *k*, we can use the least squares coefficients. This produces a new value of *k*. Using this new *k*, a new set of coefficients is found, and so on. This method involves that the estimated coefficients and (VIFs) are plotted against a range of specified values of *k*. From this plot, Hoerl and Kennard suggest selecting the value of *k* that [15] [16]:

1) Stabilizes the system such that it reflects an orthogonal system;

2) Leads to coefficients with reliable values;

3) Ensures that coefficients with improper signs at $k = 0$ have switched to the proper sign;

4) Ensures that the residual sum of squares is not inflated to an unreasonable value.

However, these criteria are very subjective. Therefore, it is best to use another method in addition to the ridge trace plot. A more reliable method is generalized cross validation (GCV). Cross validation simply entails looking at subsets of data and calculating the coefficient estimates for each subset of data, using the same value of *k* across subsets. This is then repeated multiple times with different values of *k*. The value of *k* that minimizes the differences in coefficient estimates across these data subsets is then selected [36]-[40]. The value of *k* that minimizes this equation can be computed using R, SAS, or other software's such as NCSS.

7. Data Description

Data were obtained from the Highway Safety Information System (HSIS) database maintained by the Federal Highway Administration (FHWA) of the United

States Department of Transportation. This paper used a 5-year crash period extending from 2011 to 2015 on the interstate highway (I-90) in the state of Minnesota. The interstate I-90 is a multi-lane divided highway that connects the eastern and western coasts of the US, and it passes through the southern part of Minnesota with a length of 444 km (276 mile). All crashes that occurred on the I-90 during the study period were considered in the analysis including fatal, different levels of severity injury, and property damage crashes. Different risk factors related to the road geometry, the driver behavior, the environment, and the vehicles involved in the crashes were carefully examined, classified, and selected. Table 1 shows the summary statistics of the selected risk factors, their name interpretation, their sub-classifications, their means, and their standard deviations.

The total observed crash frequency of I-90 from 2011 to 2015 is 994. The I-90 in the State of Minnesota was disaggregated equally into 276 sections, each section of one mile length. The vehicle crashes were counted at each section, and it

Table 1. Risk factors included in the study with summery statistics.

HSIS Variable Name	Name Interpretation	Variable sub classification	Mean	Standard Deviation
Rd_char	The characteristics of the road section where the crash occurred	1) Straight 2) Upgrade 3) Downgrade 4) Horizontal curve	1.673	1.102
Rdsurf	The condition of the road surface where the crash occurred	1) Dry 2) Wet 3) Snowy, or muddy	2.409	0.862
Weather	Weather conditions when the crash occurred	1) Clear 2) Rain 3) Snow, sleet 4) Fog	1.671	0.857
Light	The type of light existed at the time of the crash	1) Daylight 2) Dark, Lights On 3) Dark, No Lights	1.686	0.712
Drv_age	The age of the driver of the vehicle involved	1) <21 years 2) between 21 to 65 3) >65 years	1.772	0.597
Drv_sex	Sex of the driver of the vehicle involved	1) Male 2) Female	1.439	0.521
Vehtype	Type or body of vehicle involved in the crash	1) Passenger Car 2) Van or Minivan 3) Bus 4) Truck	1.218	0.629

AADT	Annual Average Daily Traffic on the crash road section	Numeric values in 1000s of vehicles.
		Min. = 5.77
		Max. = 28.845
			14.117	5.587

ranges from 0 to 7 crashes as shown in Table 2. For example, sections with zero crash frequency are 545, sections with only one crash frequency are 332, and sections with only two crash frequencies are 49 and so on.

Table 3 shows the descriptive statistics of the dependent variable (crash frequency) at the I-90 in Minnesota (2011-2015). It can be seen from Table 3 that the total number of observed crashes (crash count) is 994, the average or mean crashes per section is 0.773, the standard error of the mean is 0.041, the minimum number of crashes per section is 0.00, and the maximum number of crashes per section is 7.0, the skewness (a measure of how asymmetric a distribution can be when the curve appears distorted to the left or right in a statistical distribution) is 2.32, the Kurtosis (a measure of the tailedness of the probability distribution of the crash data) is 9.81, and the Shapiro-Wilk (a value of a test for normal distribution exhibiting high power, leading to good results) is 0.673.

8. Methodology

The first step is to examine the correlation between all the explanatory (independent) variables in the model. First, the Pearson correlation test is used in order

Table 2. Sections of crash frequency at I-90 in MN from 2011-2015.

Crash Frequency Section	Total Observed Crashes per section
0	545
1	332
2	49
3	43
4	9
5	7
6	4
7	5
Total	994

Table 3. I-90 Descriptive statistics of crash frequency on the I-90 in MN, USA (2011-2015).

Descriptive Statistics of Crash Frequency on I-90 in Minnesota (2010-2014)	
Count	994
Mean	0.773
Standard Error	0.041
Minimum	0.0
Maximum	7.0
Skewness	2.32
Kurtosis	9.81
Shapiro-Wilk	0.673

to identify the highly correlated variables (*i.e.*, correlation of 50% or more) as shown in Table 4. The highly correlated variables are highlighted in yellow in Table 4, which are the road characteristics, road surface, Annual Average Daily Traffic (AADT), weather, and light.

In addition to the Pearson correlation test, other methods are also used to find the correlated variables as shown in Table 5 including the variance inflation factor (VIF), the eigen values, and the Condition Numbers for the independent variables (risk factors) included in the model.

It can be seen from Table 5 that the VIFs of the predictors (rd_char, rdsurf, aadt, weather, and light) are critical as these values are bigger than 10. The Eigenvalues of (rd_char, rdsurf, aadt, weather, and light) are also critical as they are near zero. The Condition Factors of the same independent variables are also critical as they are greater than 30. All these methods indicate that the risk factors (rd_char, rdsurf, aadt, weather, and light) are the most important variables in the data. The other variables (drv_age, drv_sex, and vehtype) are less important in the data.

Table 4. Pearson correlation matrix of the explanatory variables used in the analysis.

Variables or Risk Factor	Rd_char	Rdsurf	AADT	Weather	Light	Drv_age	Drv_sex	Vehtype
Rd_char	1.000	0.716	0.856	0.639	0.778	0.097	0.069	0.055
Rdsurf	0.716	1.000	0.792	0.611	0.039	0.073	0.081	0.063
AADT	0.856	0.792	1.000	0.843	0.092	0.163	0.033	0.067
Weather	0.639	0.611	0.843	1.000	0.767	0.013	0.045	0.138
Light	0.778	0.039	0.092	0.767	1.000	0.082	0.119	0.038
Drv_age	0.097	0.073	0.163	0.013	0.082	1.000	0.043	0.016
Drv_sex	0.069	0.081	0.033	0.045	0.119	0.043	1.000	0.029
Vehtype	0.055	0.063	0.067	0.138	0.038	0.016	0.029	1.000

Table 5. The variance inflation factors (VIFs), the eigen values, and the condition numbers for the independent variables.

Independent Variable	Variance Inflation Factor	Eigenvalue	Condition Number
rd_char	44.053	0.00172	44.817
rdsurf	37.075	0.00149	62.923
aadt	76.055	0.00202	41.628
weather	82.104	0.00251	38.782
light	71.071	0.00196	64.337
drv_age	1.040	2.955	6.945
drv_sex	1.008	5.831	4.119
vehtype	1.015	7.713	3.752

The next step is to find the coefficient's estimates of the explanatory variables in the data using both the Ordinary Least Squared (OLS) Multiple linear regression, and the Ridge regression models. This step was done using the R software. The *t*-statistics and the p-value obtained in this step very good ways of testing the significance of the explanatory variables used in the models. If the *t*-statistics is significant for any variable (as indicated by the associated p-value), then this variable is significant, and should be kept in the model, and if not, then this variable can be omitted from the model. Table 6 shows the coefficient estimates,

Table 6. Results of the analysis for both multiple linear regression and ridge regression models.

Independent Variables	OLS Multiple Linear Regression			Ridge Regression		
	Coeff. Estimate	t-Statistics	P-value	Coeff. Estimate	*t*-statistics	P-value
Intercept	0.4493		0.000	0.1417		0.000
Rd_char						
1) Straight	0.202	0.393	0.149	0.193	0.317	0.044
2) U. grade	0.471	6.794	0.222	0.513	4.312	0.001
3) D. grade	3.319	1.614	0.125	2.561	4.729	0.001
4) H. Curve	2.772	2.637	0.216	3.389	9.476	0.002
Rd_surf						
1) Dry	−0.482	0.263	0.243	−1.172	1.489	0.041
2) Wet	2.322	4.866	0.323	1.778	2.533	0.002
3) Muddy	1.782	11.412	0.193	1.914	8.612	0.001
Weather						
1) Clear	−0.541	0.482	0.175	1.398	−1.627	0.031
2) Rain	6.439	6.748	0.133	2.773	9.163	0.003
3) Snow	3.743	6.188	0.202	4.493	8.961	0.001
4) Fog	5.016	14.871	0.382	8.144	7.643	0.002
Light						
1) Day Light	2.016	1.179	0.092	0.199	2.191	0.007
2) Light ON	3.153	3.244	0.082	2.371	5.333	0.044
3) No Light	4.095	11.877	0.076	5.876	14.742	0.001
Drv_age						
1) <21 yr.	3.289	11.853	0.004	4.479	17.248	0.001
2) (21 to 65)	−0.541	1.096	0.139	−0.354	1.435	0.003
3) >65 yr.	3.177	11.847	0.102	7.618	11.987	0.002
Drv_sex						
1) Male	−3.337	13.449	0.131	−1.899	11.169	0.002
2) Female	−5.228	12.767	0.067	−2.993	14.119	0.002
Vehtype						
1) P. Car	−5.301	14.285	0.139	−5.612	16.339	0.044
2) Van	−2.699	13.312	0.129	−3.411	10.417	0.022
3) Bus	5.890	12.449	0.102	4.712	8.746	0.003
4) Truck	2.969	8.771	0.202	5.844	9.352	0.002
AADT	6.974	3.683	0.122	4.811	6.213	0.023

Table 7. Observed crashes vs. predicted crashes for both ridge regression and OLS multiple regression.

Crash Frequency sections	Observed Crashes	Predicted Crashes by Ridge Regression	Predicted Crashes OLS by Multiple Regression
0	545	443	294
1	332	283	209
2	49	29	11
3	43	18	9
4	9	4	2
5	7	3	1
6	4	1	1
7	5	2	1
Total	994	783 (79% Prediction)	528 (53% Prediction)

Table 8. The R squared value and the standard errors of the ridge regression and OLS multiple linear regression models.

Goodness of Fit	Ridge Regression Model	OLS Multiple Linear Regression Model
R	0.801	0.362
R-Squared	0.762	0.276
Adjusted R-Squared	0.761	0.274
Standard Error of Estimates	0.572	0.987

t-statistics, and p-values of all independent variables for both OLS Multiple Regression and the Ridge Regression models.

The next step is to determine the predicted crashes at each type of the road sections for both the OLS Multiple regression and the Ridge regression models. The prediction of crashes at all sections is shown in Table 7. The R-squared value, the Adjusted R-squared, and Standard Errors of both the Ridge Regression and OLS Multiple Linear Regression models are shown in Table 8.

9. Findings and Discussion

Since the *t*-statistics shown in Table 6 are significant at the 95% confidence level for all the explanatory variables used in the Ridge Regression model (*i.e.*, their p-values are less than 0.05), then these factors are significant, and should be kept in the model. However, the t-statistics for all the explanatory variables are insignificant in the OLS Multiple Linear Regression model (*i.e.*, their p-values are greater than 0.05). This clearly indicates that the Ridge Regression can effectively be used to identify the significant independent variables in crash data, whereas the OLS Multiple Regression can make the significant variables to be insignificant as shown in Table 6. Therefore, using Ridge Regression is paramount in crash data modeling that suffers from multicollinearity. Also, the coefficient's estimates and their signs for the data shown in Table 6 can be used to explore the contribution of each explanatory variable to the resulting dependent variable (*i.e.*, crash frequency). The positive sign of the estimate indicates that the associated explanatory variable would increase the likelihood of the crash occurrence, and the negative sign indicates negative contribution of the variable to the crash occurrence. For example, when inspecting the road characteristics factors in both the Multiple Regression and Ridge Regression models, the positive sign of the upgrade, downgrade, and horizontal curves means that the occurrence of crashes at road segments with these features are more likely to happen than at the straight portions of the road. The grades and curves affect the operation of vehicles and their speed, and this obviously could increase the probability of the vehicle accidents. The wet and muddy conditions of the road surface would decrease the coefficient of friction between the tires and the road surface, and hence would increase the crash probabilities, as indicated by the positive sign of the wet and muddy coefficient estimates compared to the negative sign of the dry condition estimate. For the weather factors estimates, the positive sign of the snow, and fog conditions indicates increased crash frequency at these conditions, as the driver vision within the fog could decrease, and the friction coefficient within the snow could substantially decrease, and hence, causing the increased probability of more accidents. The accidents could also increase in the dark with no light, as indicated by the positive sign of the (No light) factor estimate in the table. The driver age group of (21 to 65 years) has negative estimate, indicating that this group is less likely to increase the crash occurrence, whereas the young drivers (less than 21 years), and the elderly (more than 65 years) can positively contribute to the increased crash frequency, as indicated by their positive sign estimates. The driver sex has negative estimates for both males and females, indicating no preferences on crash occurrence in term of driver sex. The vehicle type factors show that both the passenger cars and vans or mini vans have negative sign estimates, meaning that their contribution to the accidents is less likely to increase, compared to the buses and trucks with positive estimates that can increase the crash occurrence likelihood. The annual average daily traffic (AADT) has positive estimate sign, indicating that the increased daily traffic volume at any section can increase the crash frequency as vehicles are more likely to interact with each other in higher volume conditions.

The prediction performance of the two models can be presented by comparing the observed crashes versus the predicted crashes for each model at each crash section, as shown in Table 7 for both methods. The Ridge Regression crash prediction results are much better than the OLS Multiple Linear Regression in all crash sections (*i.e.*, within sections of 0, 1, 2, 3, 4, 5, 6, 7). The prediction accuracy of the Ridge Regression is 79% compared to 53% for the OLS Multiple Linear Regression. These prediction results of the Ridge Regression demonstrate that it is an effective approach in predicting highway crash frequency and can improve the accuracy of the prediction results upon the results obtained from the traditional statistical models, such as the OLS Multiple Regression.

In order to further prove the accuracy of the results, Table 8 shows the R-squared value, the Adjusted R-squared, and Standard Errors of both the Ridge Regression and OLS Multiple Linear Regression models. The R-squared is 0.762 for Ridge Regression and 0.276 for OLS Multiple Linear Regression. R-squared can range from 0% to 100%. The higher the R-squared, the better the fit. Clearly, the Ridge Regression model fits much better the data than the OLS Multiple Linear Regression. The residual standard error is used to measure how well a regression model fits a dataset. The smaller the residual standard error, the better a regression model fits a dataset. The standard error of the Ridge Regression is 0.572, which is smaller than the value of 0.987 that belongs to the OLS Multiple Linear Regression model. These values indicate an excellent fit of the Ridge Regression model into the crash data.

10. Conclusion

Ridge Regression is presented in this paper as an effective statistical technique for analyzing vehicle crash data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors and produces accurate estimates compared with the ordinary least squared multiple regressions. In this paper, two crash prediction methods were chosen for the analysis of the crash data on the interstate highway I-90 in Minnesota, namely: the Ridge Regression Model, and the OLS Multiple Linear Regression Model. The analysis showed that the OLS Multiple linear regression model might not be well suited to fit the crash data because of the multicollinearity between the independent variables in the crash data. The Ridge regression model can take the multicollinearity into account, and hence, can produce much better prediction results. Hence, this paper recommends employing the Ridge regression in crash frequency modeling so that the correlation problems between the explanatory variables would not be a concern, as it can effectively handle the correlation problem without affecting the output.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Washington, S.P., Karlaftis, M.G. and Mannering, F. (2010) Statistical and Econometric Methods for Transportation Data Analysis. 2nd Edition, Chapman Hall/CRC, Boca Raton.

[2] Cule, E. and De Iorio, M. (2012) A Semi-Automatic Method to Guide the Choice of Ridge Parameter in Ridge Regression. arXiv: 1205.0686
http://arxiv.org/pdf/1205.0686.pdf

[3] Ahn, J.J., Kim, Y.M., Yoo, K., Park, J. and Oh, K.J. (2012) Using GA-Ridge Regres-

sion to Select Hydro-Geological Parameters Influencing Groundwater Pollution Vulnerability. *Environmental Monitoring and Assessment*, **184**, 6637-6645. https://doi.org/10.1007/s10661-011-2448-1

[4] Alkhamisi, M., Khalaf, G. and Shukur, G. (2006) Some Modifications for Choosing Ridge Parameters. *Communications in Statistics*: *Theory and Methods*, **35**, 2005-2020. https://doi.org/10.1080/03610920600762905

[5] Abdulhafedh, A. (2022) Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. *Open Access Library Journal*, **9**, e8414. https://doi.org/10.4236/oalib.1108414

[6] Cameron, A.C. and Trivedi, P.K. (1998) Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511814365

[7] Abdulhafedh, A. (2017) Road Crash Prediction Models: Different Statistical Modeling Approaches. *Journal of Transportation Technologies*, **7**, 190-205. https://doi.org/10.4236/jtts.2017.72014

[8] Hilbe, J. (2014) Modeling Count Data. Cambridge University Press, London.

[9] Alkhamisi, M.A. and Shukur, G. (2008) Developing Ridge Parameters for SUR Model. *Communications in Statistics*: *Theory and Methods*, **37**, 544-564. https://doi.org/10.1080/03610920701469152

[10] Al-Hassan, Y.M. (2008) A Monte Carlo Evaluation of Some Ridge-Type Estimators. *Jordan Journal of Applied Sciences*, **10**, 101-110.

[11] Anders, B.J. (2001) Ridge Regression and Inverse Problems. Stocks University, Sweden.

[12] Dorugade, A.V. and Kashid, D.N. (2010) Alternative Method for Choosing Ridge Parameter for Regression. *Applied Mathematical Sciences*, **4**, 447-456.

[13] El-Dereny, M. and Rashwan, N.I. (2011) Solving Multicollinearity Problem Using Ridge Regression Models. *International Journal of Contemporary Mathematical Sciences*, **6**, 585-600.

[14] Abdulhafedh, A. (2016) Crash Frequency Analysis. *Journal of Transportation Technologies*, **6**, 169-180. https://doi.org/10.4236/jtts.2016.64017

[15] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[16] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12**, 69-82. https://doi.org/10.1080/00401706.1970.10488635

[17] Pasha, G.R. and Shah, M.A. (2004) Application of Ridge Regression to Multicollinear Data. *Journal of Research* (*Science*), **15**, 97-106.

[18] Mansson, K., Shukur, G. and Golam Kibria, B.M. (2010) A Simulation Study of Some Ridge Regression Estimators under Different Distributional Assumptions. *Communications in Statistics*: *Simulation and Computation*, **39**, 1639-1670. https://doi.org/10.1080/03610918.2010.508862

[19] Lauridsen, J. and Mur, J. (2006) Multicollinearity in Cross-Sectional Regressions. *Journal of Geographical Systems*, **8**, 317-333. https://doi.org/10.1007/s10109-006-0031-z

[20] Chopra, P., Sharma, R.K. and Kumar, M. (2013) Ridge Regression for the Prediction of Compressive Strength of Concrete. *International Journal of Innovations in Engi-

neering and Technology (*IJIET*), **2**, 106-111.

[21] Zaka, A. and Akhter, A.S. (2013) Methods for Estimating the Parameters of the Power Function Distribution. *Pakistan Journal of Statistics and Operation Research*, **9**, 213-224. https://doi.org/10.18187/pjsor.v9i2.488

[22] Farrar, D.E. and Glauber, R.R. (1967) Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, **49**, 92-107. https://doi.org/10.2307/1937887

[23] Abdulhafedh, A. (2016) Crash Severity Modeling in Transportation Systems. PhD Dissertation. University of Missouri-Columbia, MO, USA. https://doi.org/10.32469/10355/59817

[24] Frank, I.E. and Friedman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135. https://doi.org/10.1080/00401706.1993.10485033

[25] Abdulhafedh, A. (2017) Road Traffic Crash Data: An Overview on Sources, Problems, and Collection Methods. *Journal of Transportation Technologies*, **7**, 206-219. https://doi.org/10.4236/jtts.2017.72015

[26] Judge, G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. and Lee, T.C. (1985) The Theory and Practice of Econometrics. 2nd Edition, Wiley, New York.

[27] Abdulhafedh, A. (2021) Incorporating K-Means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, **3**, 12-30.

[28] Fu, W.J. (1998) Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, **7**, 397-416. https://doi.org/10.1080/10618600.1998.10474784

[29] Duzan, H. and Shariff, N.S. M. (2015) Ridge Regression for Solving the Multicollinearity Problem: Review of Methods and Models. *Journal of Applied Sciences*, **15**, 392-404. https://doi.org/10.3923/jas.2015.392.404

[30] Abdulhafedh, A. (2022) Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables. *Open Access Library Journal*, **9**, e8346. https://doi.org/10.4236/oalib.1108346

[31] Khalaf, G. (2012) A Proposed Ridge Parameter to Improve the Least Square Estimator. *Journal of Modern Applied Statistical Methods*, **11**, Article 15. https://doi.org/10.22237/jmasm/1351743240

[32] Singh, R. (2012) Solution of Multicollinearity by Ridge Regression. *International Journal of Research in Computer Application & Management*, **2**, 130-136.

[33] Abdulhafedh, A. (2017) Incorporating the Multinomial Logistic Regression in Vehicle Crash Severity Modeling: A Detailed Overview. *Journal of Transportation Technologies*, **7**, 279-303. https://doi.org/10.4236/jtts.2017.73019

[34] Gorman, J.W. and Toman, R.J. (1966) Selection of Variables for Fitting Equations to Data. *Technometrics*, **8**, 27-51. https://doi.org/10.1080/00401706.1966.10490322

[35] Heinze, G. and Schemper, M. (2002) A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*, **21**, 2409-2419. https://doi.org/10.1002/sim.1047

[36] Goldstein, M. and Smith, A.F.M. (1974) Ridge-Type Estimators for Regression Analysis. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **36**, 284-291. https://doi.org/10.1111/j.2517-6161.1974.tb01006.x

[37] Golub, G.H., Heath, M. and Wahba, G. (1979) Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, **21**, 215-223. https://doi.org/10.1080/00401706.1979.10489751

[38] Abdulhafedh, A. (2017) A Novel Hybrid Method for Measuring the Spatial Autocorrelation of Vehicular Crashes: Combining Moran's Index and Getis-Ord G_i^* Statistic. *Open Journal of Civil Engineering*, **7**, 208-221.
https://doi.org/10.4236/ojce.2017.72013

[39] Abdulhafedh, A. (2017) Identifying Vehicular Crash High Risk Locations along Highways via Spatial Autocorrelation Indices and Kernel Density Estimation. *World Journal of Engineering and Technology*, **5**, 198-215.
https://doi.org/10.4236/wjet.2017.52016

[40] Khalaf, G. and Shukur, G. (2005) Choosing Ridge Parameter for Regression Problems. *Communications in Statistics*: *Theory and Methods*, **34**, 1177-1182.
https://doi.org/10.1081/STA-200056836