Multiple Regression Analysis of Robbery Trends in Toronto

Name: Zarraf Afnan

Student ID: 100880521

Introduction

Crime analysis is an essential aspect of urban planning and law enforcement strategies. This study investigates robbery trends in Toronto by examining factors influencing the frequency of reported robberies across different locations. The dataset for this analysis was obtained from Kaggle, particularly from a collection of reported robbery incidents in Toronto spanning from 2014 to around mid-2022. The primary objective of this analysis is to determine how different environmental factors (e.g., crime-prone areas, premises type) influence robbery frequency. The dependent variable in this study is Robbery_Count, representing the number of robbery incidents. Independent variables include:

- High_Crime_Zone: Whether the incident took place in a historically high-crime region (Binary Form: 1= Yes, 0= No).
- Premises Type: The type of location where the robbery occurred (Apartment, Commercial, or Outside).

A multiple regression model was chosen as the statistical method to analyze the relationship between these predictors and robbery counts.

Methods

A multiple linear regression model was selected for this study as it allows for the simultaneous examination of multiple independent variables and their impact on robbery incidents. The following steps were performed using SAS software:

1. Data Cleaning & Preparation

- The dataset was imported into SAS, and categorical variables were converted into dummy variables.
- The dataset was aggregated by year and month to analyze temporal trends.

2. Model Selection

- Initially, all predictors were included in the model (YEAR, MONTH, High_Crime_Zone, Premises_Type).
- Insignificant predictors (YEAR, MONTH) were removed to improve model efficiency.
- The final model included High_Crime_Zone, Premises_Type (Apartment, Commercial,

Outside).

3. Model Evaluation & Assumption Checks

- Checking for Similar Variables: The model was tested to see if any independent variables were too similar to each other. The results showed no major issue indicating that each variable provides new information to the model.

-Checking for Normal Distribution: The model was tested to see if the errors (i.e. the differences between predicted and actual values) followed a normal pattern. The tests showed that the errors were not normally distributed, but this is not a major issue for multiple regression, so the results were still considered valid.

Results

Model Fit & Significance:

- R-Square: 0.6404
- Adjusted R-Square: 0.6387
- F-statistic: 371.32 ($p < 0.0001$)
- Root Mean Square Error (RMSE): 17.85

These results indicate that 64.04% of the variance in robbery counts is explained by the model, suggesting a strong fit.

Key Predictor Variables

Variable	Coefficient (β)	p-value	Interpretation
Intercept	18.835	< 0.0001	Expected robbery count when all variables are 0.
High_Crime_Zone	-3.803	0.0021	Negative effect: Fewer robberies in high-crime zones.
Prem_Apartment	-4.637	0.0083	Negative effect: Fewer robberies in apartments.
Prem_Commercial	19.819	< 0.0001	Positive effect: More robberies in commercial areas.
Prem_Outside	55.224	< 0.0001	Largest effect: Most robberies occur outside.

Discussion & Conclusion

This study applied multiple regression analysis to explore factors influencing robbery trends in Toronto. The results indicate that the type of location significantly impacts the number of reported robberies. Robberies were most frequent in outdoor and commercial areas, whereas apartment buildings had significantly lower incidents. This suggests that open spaces and businesses may be more vulnerable due to increased accessibility and the presence of valuable goods. Another interesting and surprising result was that certain high-crime areas reported fewer robberies. This could be due to a stronger police presence as a hindrance for criminal activities or differences in how crimes are reported in such areas. Understanding such patterns can help city officials and law enforcement develop more effective crime prevention strategies and eventually improve public safety.

Strengths of the Study:

- Robust Model Fit: The model explains 64% of the variation in robbery counts, making it a reliable tool for understanding robbery patterns.
- Meaningful Insights: Findings align with expectations regarding urban crime patterns.
- Statistically Significant Results: All key predictors (location types and high-crime zones) were statistically significant, meaning they have a real impact on robbery rates.

Limitations & Future Improvements:

- Residuals are non-normal
- Potential Reporting Bias: Official reports may not capture unreported crimes.
- Additional Variables: Future studies could include socio-economic factors, time of day, or police response times to get a more complete picture of the robbery trends.

Final Thoughts

This study indicates that the number of robbery incidents is strongly affected by location, with outdoor and commercial areas being at the highest risk. These results suggest that law enforcement officers may need to prioritize increasing patrols, monitoring high-risk areas, and improving security measures to reduce crime. Business owners and city planners can also use this information to make informed decisions about where to enhance safety measures like better lighting, surveillance cameras or public awareness campaigns. While this analysis certainly offers useful insights, future research can also explore additional factors to develop more effective crime prevention strategies and improve urban safety.

Appendices

The following pages provide the SAS code and its output.

```
/* Generated Code (IMPORT) */
/* Source File: Robbery.xlsx */
/* Source Path: /home/u63700025/assignment */
/* Code generated on: 3/21/25, 1:31 PM */

%web_drop_table(WORK.IMPORT);


/* Importing the Excel File */
PROC IMPORT DATAFILE="/home/u63700025/assignment/Robbery.xlsx"
   OUT=WORK.ROB_DATA
   DBMS=XLSX
   REPLACE;
   GETNAMES=YES;
RUN;

/* Data Preparation */
DATA WORK.ROB_CLEAN;
   SET WORK.ROB_DATA;

   /* Converting Date Variables */
   OCC_DATE = INPUT(SCAN(occurrencedate, 1, 'T'), YYMMDD10.);
   FORMAT OCC_DATE DATE9.;
   YEAR = YEAR(OCC_DATE);
   MONTH = MONTH(OCC_DATE);

   /* Creating Binary High Crime Zone Variable */
   IF Division IN ('D14', 'D51', 'D52', 'D11', 'D41', 'D31', 'D23') THEN High_Crime_Zone = 1;
   ELSE High_Crime_Zone = 0;

   /* Creating Dummy Variables for Premises Type */
   Prem_Apartment = (premises_type = 'Apartment');
   Prem_Commercial = (premises_type = 'Commercial');
   Prem_Outside = (premises_type = 'Outside');
   /*Educational or other premises types(Reference Category) */
RUN;

/* Aggregating Data for Regression */
PROC SQL;
   CREATE TABLE WORK.ROB_MODEL AS
   SELECT
      YEAR, MONTH, High_Crime_Zone,
      Prem_Apartment, Prem_Commercial, Prem_Outside,
      COUNT(*) AS Robbery_Count
   FROM WORK.ROB_CLEAN
   GROUP BY YEAR, MONTH, High_Crime_Zone,
         Prem_Apartment, Prem_Commercial, Prem_Outside;
QUIT;

/* Checking for VIF Values */
PROC REG DATA=WORK.ROB_MODEL;
   MODEL Robbery_Count =
      YEAR MONTH
```

```
      High_Crime_Zone
      Prem_Apartment Prem_Commercial Prem_Outside / VIF;
   TITLE "Multicollinearity Check (VIF Values)";
RUN;
QUIT;

/* Running Multiple Regression Analysis */
PROC REG DATA=WORK.ROB_MODEL;
   MODEL Robbery_Count =
      YEAR MONTH
      High_Crime_Zone
      Prem_Apartment Prem_Commercial Prem_Outside;
   TITLE "Multiple Regression Model for Robbery Trends in Toronto";
RUN;
QUIT;



/* Scatterplot to check linearity */
PROC SGSCATTER DATA=WORK.ROB_MODEL;
   PLOT Robbery_Count*(YEAR MONTH);
   TITLE "Scatterplots of Robbery Count vs Year and Month";
RUN;

/* Normality of Residuals */
PROC UNIVARIATE DATA=WORK.ROB_MODEL NORMAL;
   VAR Robbery_Count;
   HISTOGRAM Robbery_Count / NORMAL;
   TITLE "Distribution of Robbery Count";
RUN;

PROC REG DATA=WORK.ROB_MODEL;
   MODEL Robbery_Count = High_Crime_Zone Prem_Apartment Prem_Commercial Prem_Outside;
   TITLE "Refined Multiple Regression Model (Only Significant Predictors)";
RUN;
QUIT;



%web_open_table(WORK.IMPORT);
```

Multicollinearity Check (VIF Values)

The REG Procedure
Model: MODEL1
Dependent Variable: Robbery_Count

Number of Observations Read	839
Number of Observations Used	839

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	473998	79000	248.14	<.0001
Error	832	264881	318.36671		
Corrected Total	838	738879			

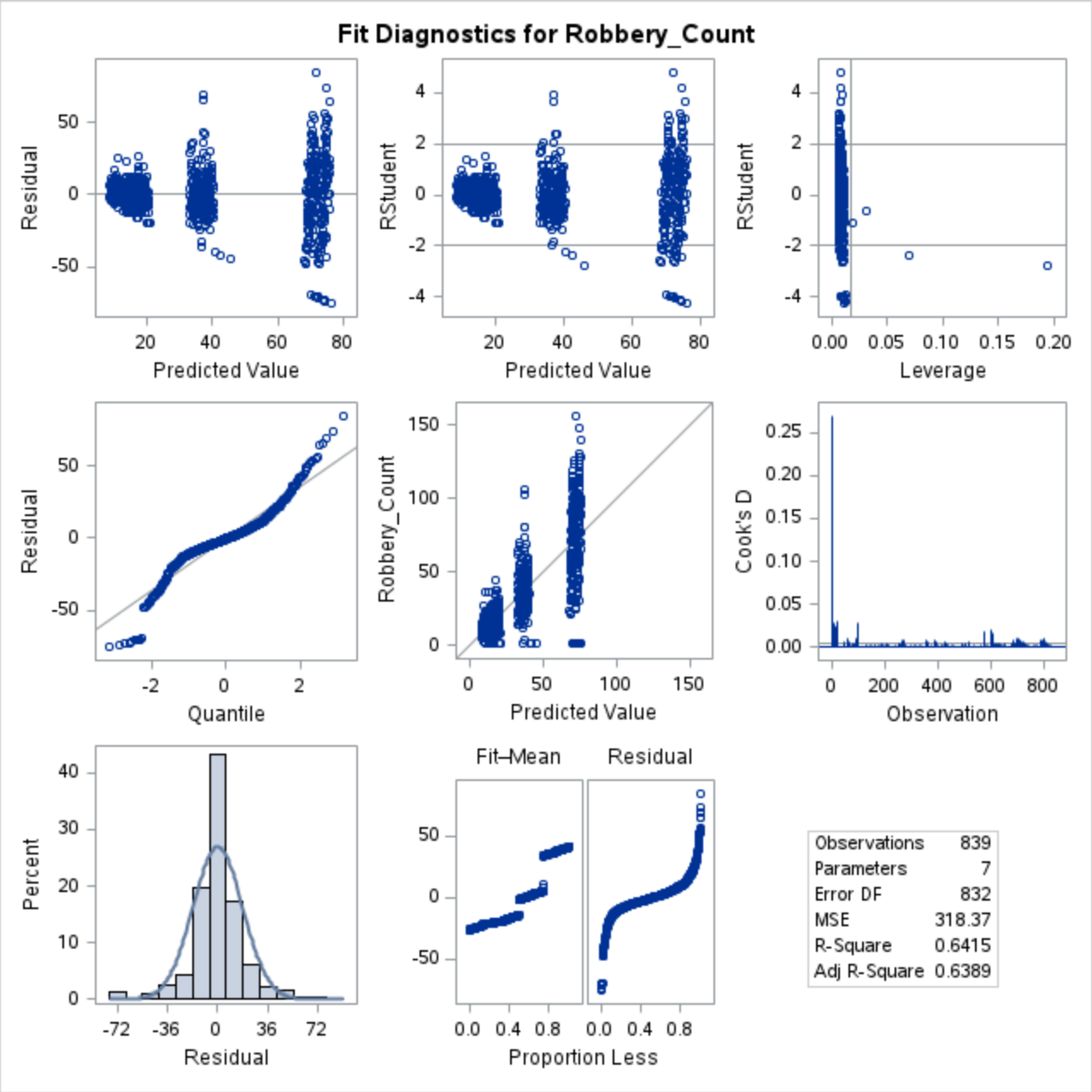
Root MSE	17.84283	R-Square	0.6415
Dependent Mean	34.83909	Adj R-Sq	0.6389
Coeff Var	51.21498		

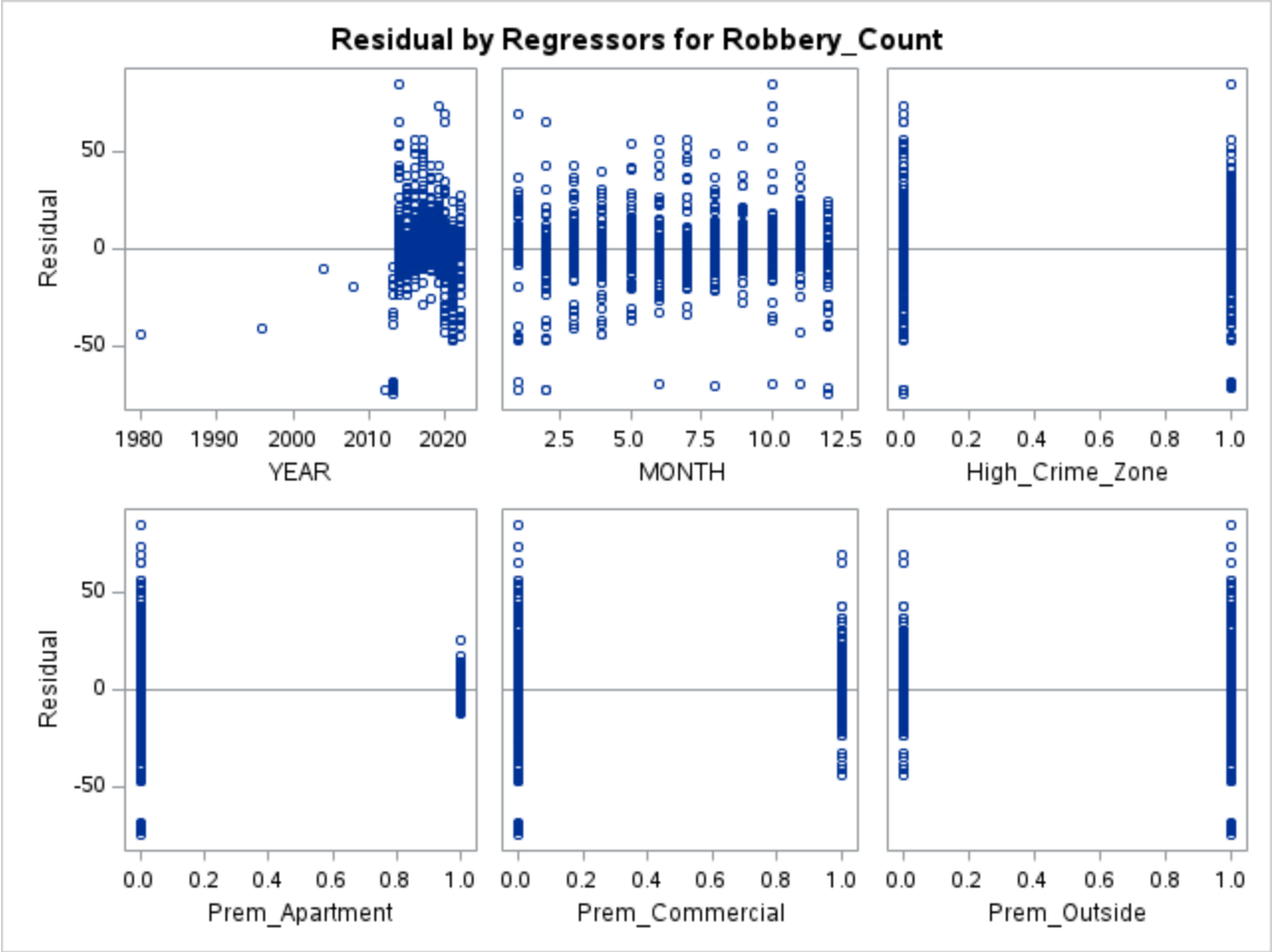
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	429.91041	416.26832	1.03	0.3020	0
YEAR	1	-0.20441	0.20628	-0.99	0.3220	1.00734
MONTH	1	0.21014	0.17815	1.18	0.2385	1.00593
High_Crime_Zone	1	-3.79872	1.23230	-3.08	0.0021	1.00047
Prem_Apartment	1	-4.61656	1.75184	-2.64	0.0086	1.49820
Prem_Commercial	1	19.77498	1.74406	11.34	<.0001	1.50419
Prem_Outside	1	55.22090	1.73540	31.82	<.0001	1.50800

Multicollinearity Check (VIF Values)

The REG Procedure
Model: MODEL1
Dependent Variable: Robbery_Count


Fit Diagnostics for Robbery_Count
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Robbery_Count
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 839
Parameters 7
Error DF 832
MSE 318.37
R-Square 0.6415
Adj R-Square 0.6389



Multiple Regression Model for Robbery Trends in Toronto

The REG Procedure
Model: MODEL1
Dependent Variable: Robbery_Count

Number of Observations Read	839
Number of Observations Used	839

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	473998	79000	248.14	<.0001
Error	832	264881	318.36671		
Corrected Total	838	738879			

Root MSE	17.84283	**R-Square**	0.6415
Dependent Mean	34.83909	**Adj R-Sq**	0.6389
Coeff Var	51.21498		

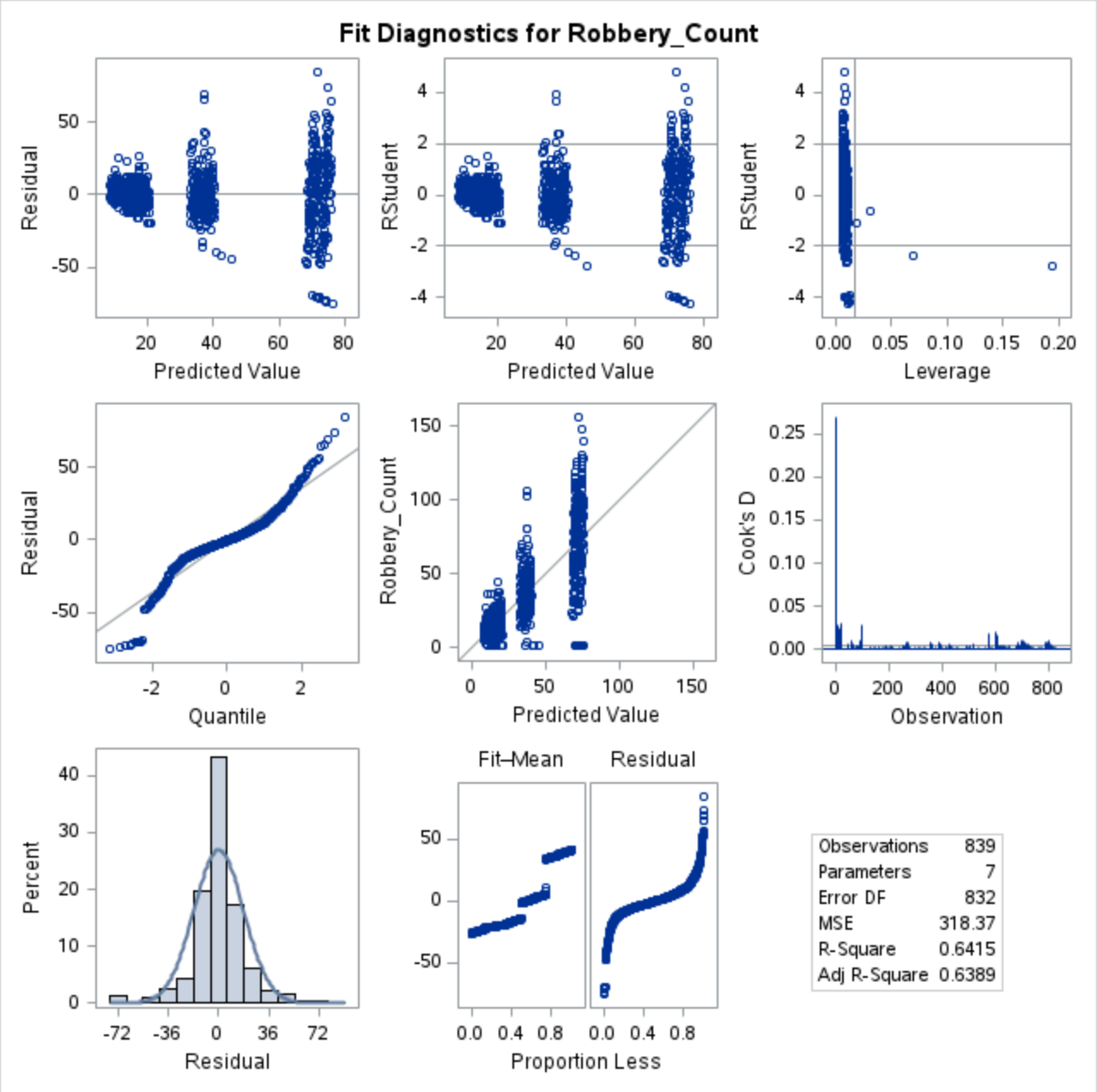
Parameter Estimates

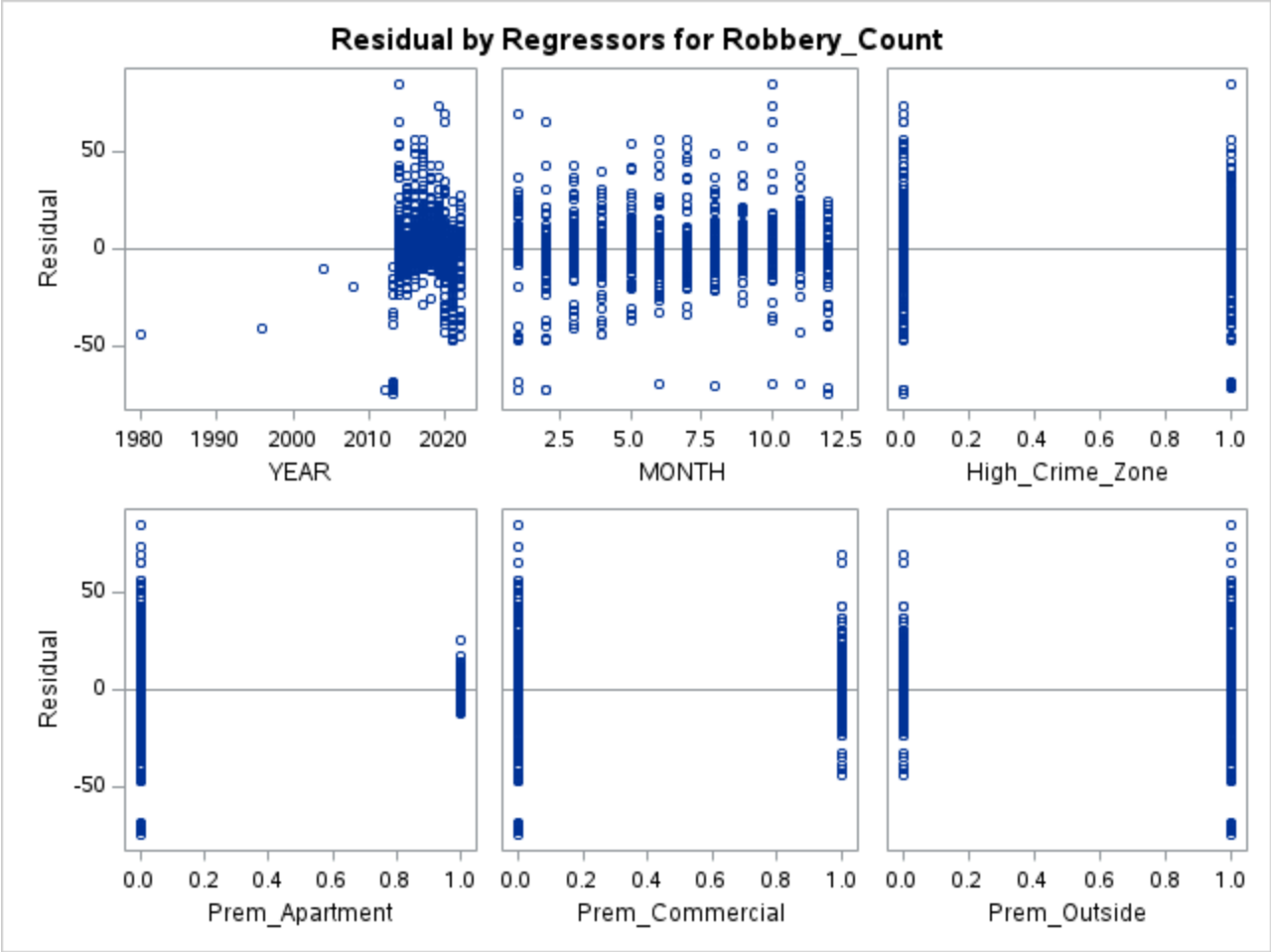
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	429.91041	416.26832	1.03	0.3020

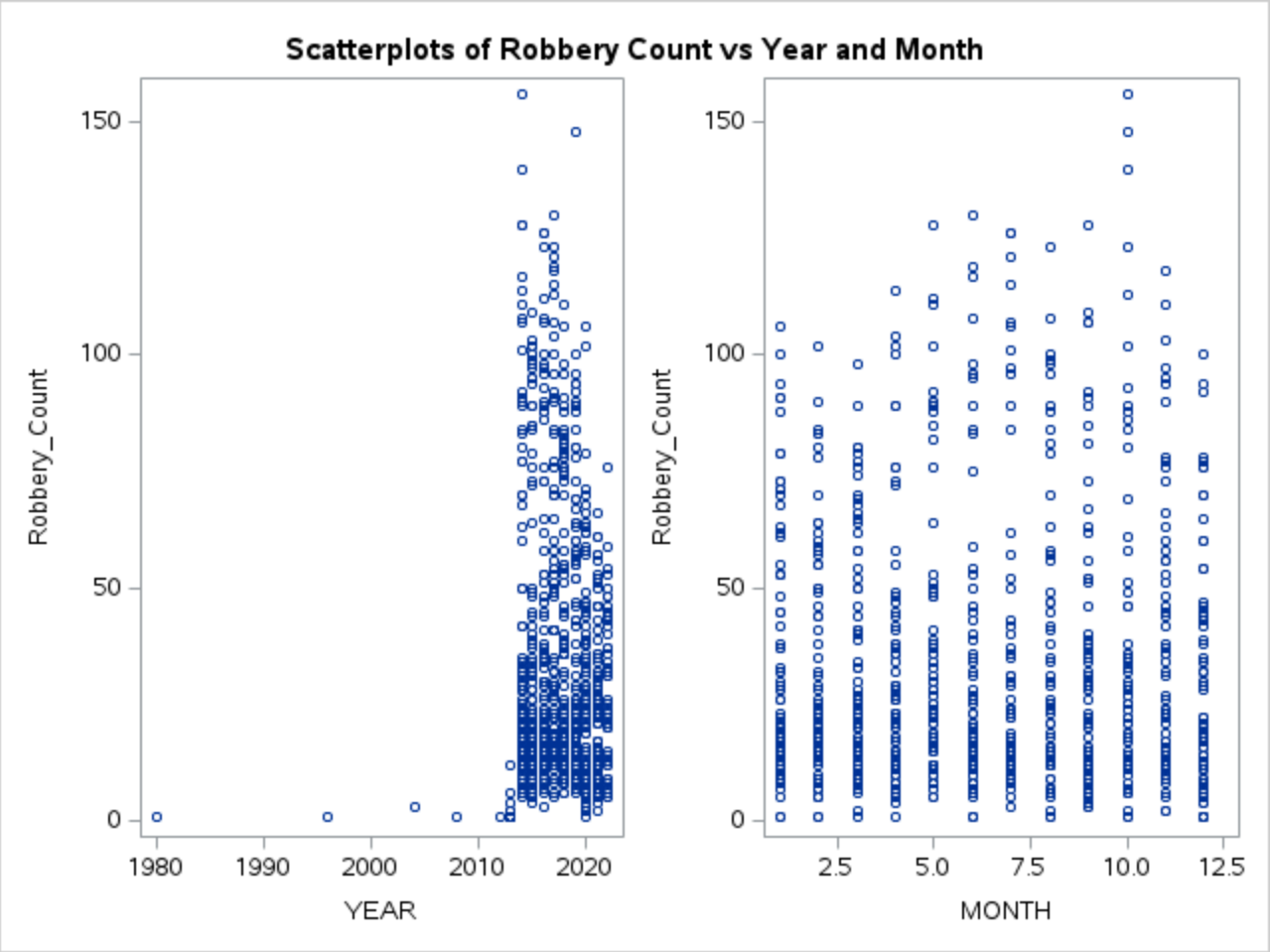
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
YEAR	1	-0.20441	0.20628	-0.99	0.3220
MONTH	1	0.21014	0.17815	1.18	0.2385
High_Crime_Zone	1	-3.79872	1.23230	-3.08	0.0021
Prem_Apartment	1	-4.61656	1.75184	-2.64	0.0086
Prem_Commercial	1	19.77498	1.74406	11.34	<.0001
Prem_Outside	1	55.22090	1.73540	31.82	<.0001

Multiple Regression Model for Robbery Trends in Toronto

The REG Procedure
Model: MODEL1
Dependent Variable: Robbery_Count




Residual by Regressors for Robbery_Count
Residual
50
0
-50
1980
1990
2000
2010
2020
YEAR
2.5
5.0
7.5
10.0
12.5
MONTH
0.0
0.2
0.4
0.6
0.8
1.0
High_Crime_Zone
Prem_Apartment
Prem_Commercial
Prem_Outside



Distribution of Robbery Count

The UNIVARIATE Procedure
Variable: Robbery_Count

Moments			
N	839	**Sum Weights**	839
Mean	34.8390942	**Sum Observations**	29230
Std Deviation	29.6937285	**Variance**	881.717515
Skewness	1.30930868	**Kurtosis**	1.05187573
Uncorrected SS	1757226	**Corrected SS**	738879.278
Coeff Variation	85.231058	**Std Error Mean**	1.02514135

Basic Statistical Measures			
Location		**Variability**	
Mean	34.83909	**Std Deviation**	29.69373
Median	24.00000	**Variance**	881.71752
Mode	12.00000	**Range**	155.00000
		Interquartile Range	35.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	33.98467	**Pr > |t|**	<.0001
Sign	**M**	419.5	**Pr >= |M|**	<.0001

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Signed Rank	S	176190	Pr >= |S|	<.0001

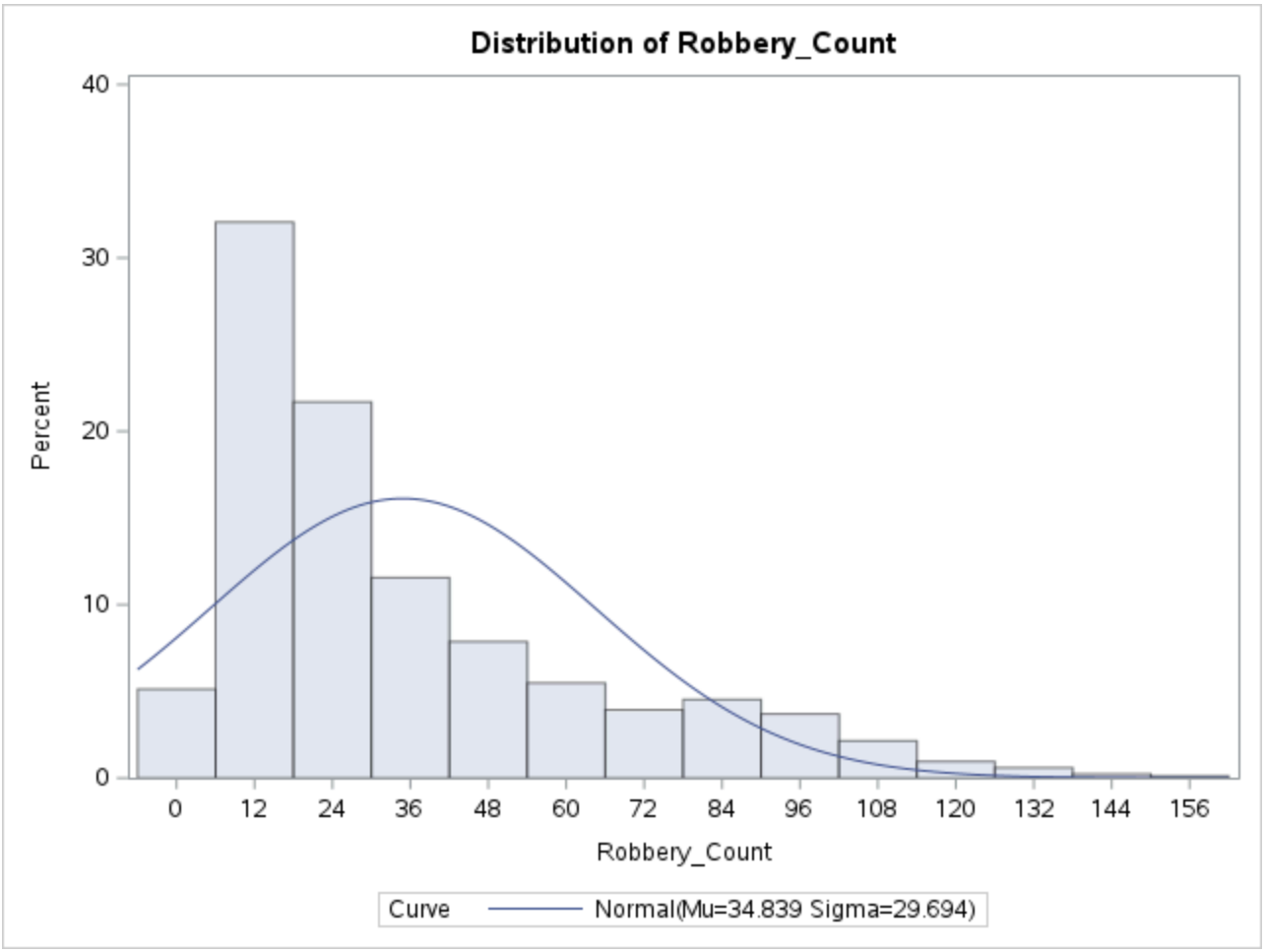
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.852744	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.166743	Pr > D	<0.0100
Cramer-von Mises	W-Sq	7.64482	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	43.14222	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	156
99%	123
95%	98
90%	84
75% Q3	48
50% Median	24
25% Q1	13
10%	8
5%	5
1%	1
0% Min	1

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	643	128	89
1	22	130	353
1	21	140	97
1	20	148	577
1	19	156	101

Distribution of Robbery Count

The UNIVARIATE Procedure



Distribution of Robbery Count

The UNIVARIATE Procedure
Fitted Normal Distribution for Robbery_Count

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	34.83909
Std Dev	Sigma	29.69373

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.1667431	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	7.6448203	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	43.1422168	**Pr > A-Sq**	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	1.00000	-34.23885
5.0	5.00000	-14.00274
10.0	8.00000	-3.21495
25.0	13.00000	14.81098
50.0	24.00000	34.83909
75.0	48.00000	54.86721
90.0	84.00000	72.89314

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
95.0	98.00000	83.68093
99.0	123.00000	103.91704

Refined Multiple Regression Model (Only Significant Predictors)

The REG Procedure
Model: MODEL1
Dependent Variable: Robbery_Count

Number of Observations Read	839
Number of Observations Used	839

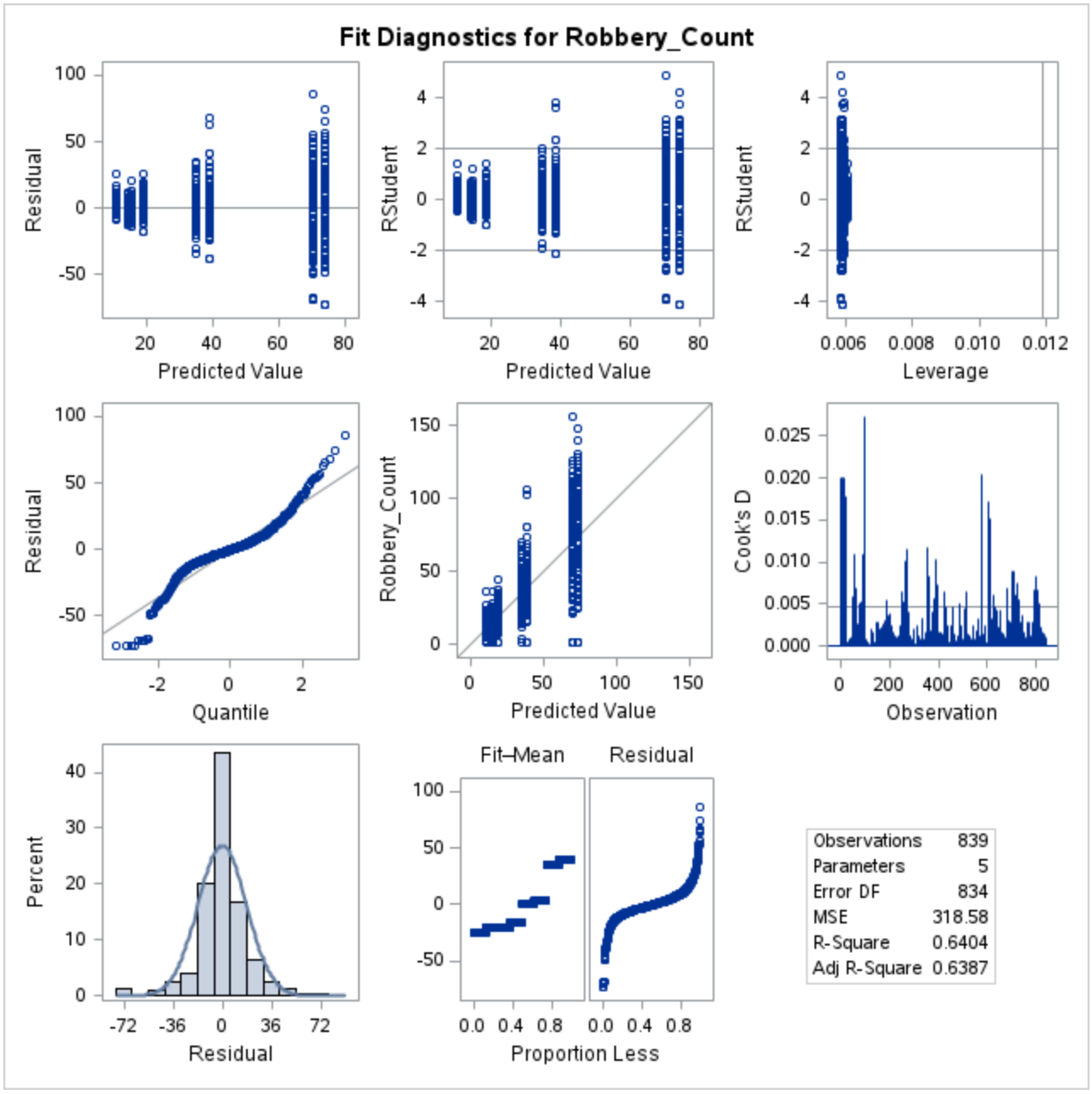
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	473182	118295	371.32	<.0001
Error	834	265698	318.58224		
Corrected Total	838	738879			

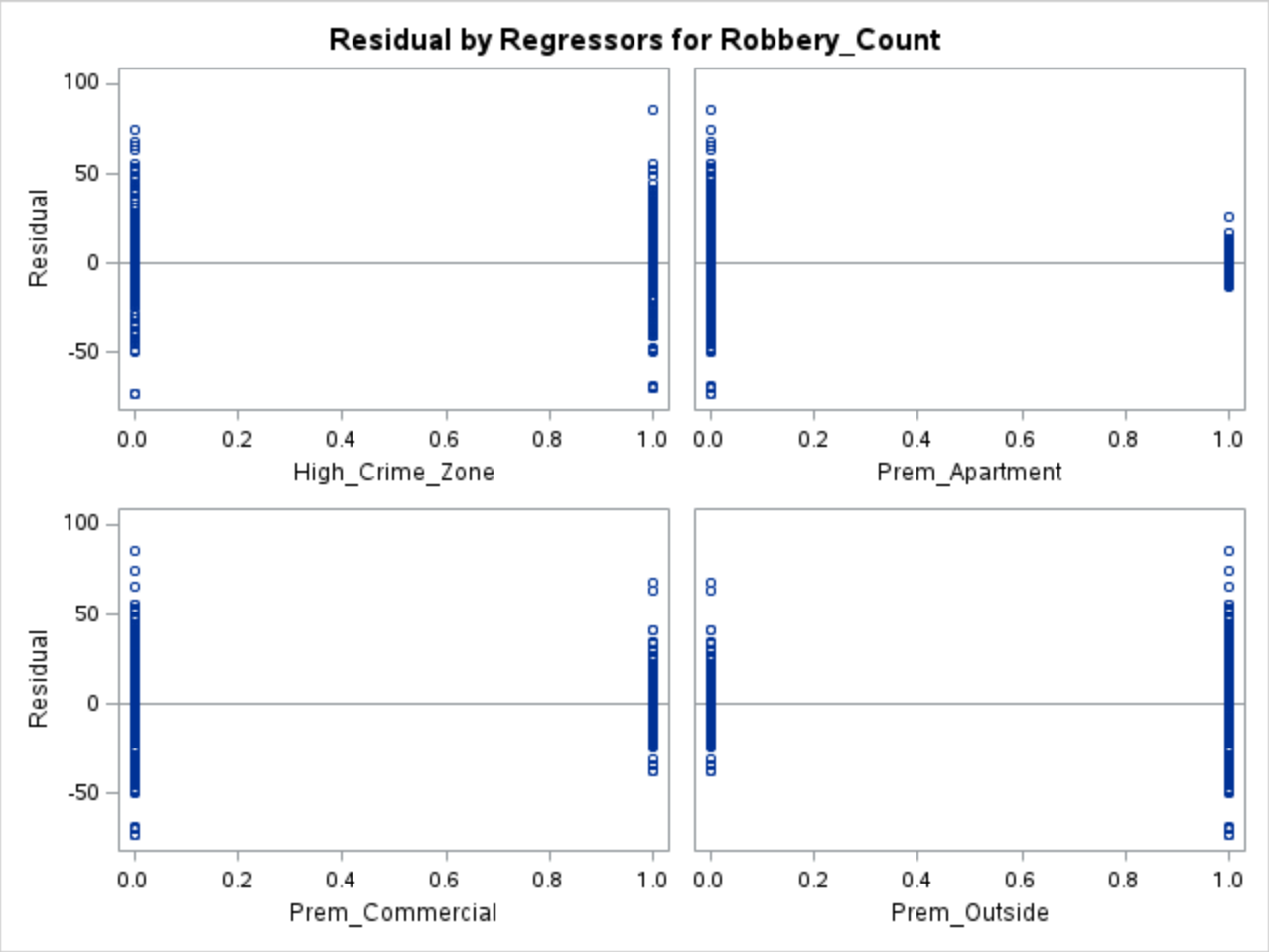
Root MSE	17.84887	R-Square	0.6404
Dependent Mean	34.83909	Adj R-Sq	0.6387
Coeff Var	51.23231		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.83501	1.37856	13.66	<.0001
High_Crime_Zone	1	-3.80306	1.23244	-3.09	0.0021
Prem_Apartment	1	-4.63737	1.75238	-2.65	0.0083
Prem_Commercial	1	19.81890	1.74396	11.36	<.0001
Prem_Outside	1	55.22447	1.73583	31.81	<.0001

Refined Multiple Regression Model (Only Significant Predictors)

The REG Procedure
Model: MODEL1
Dependent Variable: Robbery_Count


Fit Diagnostics for Robbery_Count
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Robbery_Count
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 839
Parameters 5
Error DF 834
MSE 318.58
R-Square 0.6404
Adj R-Square 0.6387


Residual by Regressors for Robbery_Count
Residual
100
50
0
-50
0.0
0.2
0.4
0.6
0.8
1.0
High_Crime_Zone
Prem_Apartment
Prem_Commercial
Prem_Outside