The Association Between Felonies in NYC and Weather and Temporal Conditions

Brian Weinstein

Columbia University
STAT W4201: Advanced Data Analysis

May 2, 2016

Abstract

Background: The New York City Police Department recently released incident-level felony data to the NYC Open Data portal. The dataset includes timestamped information felonies committed in NYC.

We first examine the association between the daily number of felonies committed in NYC in 2015 and temperature, presence of precipitation, day of week, federal and New York holidays, and school days. Second, we examine the association between day-to-day changes in the number of felonies and large increases in temperature ($> 8°$F from the previous day).

Methods and Results: Using simple linear regression, the data provides overwhelming evidence that felonies is associated with temperature. For every 1°F increase in temperature, there are, on average, 1.38 additional felonies per day (95% CI 1.23 to 1.53 felonies, two sided p-value $< 2 \times 10^{-16}$).

Using multiple linear regression, after accounting for temperature, there is convincing evidence that the number of felonies on a given day is associated with precipitation. On average there are 21 fewer felonies on days with precipitation than on days without (95% CI 4 to 38 fewer felonies; two-sided p-value 0.014990). The data also provides overwhelming evidence that felonies is associated with day of week (p-value 3×10^{-6} from an extra sum of squares F-test).

Initially, it did not appear that the number of felonies was associated with holidays or with school days, but after removing 17 observations with high studentized residuals, these variables became significant. From the revised model, there is suggestive, but inconclusive evidence that felonies is associated with holidays, with holidays generally having 13 fewer felonies than non-holidays (95% CI 0 to 27 fewer felonies; two-sided p-value 0.05275). There is moderate evidence that felonies is associated with school days, with school days generally having 7 more felonies that non-school days (95% CI 1 to 13 more felonies; two-sided p-value 0.03306).

After accounting for day of week, the data provides no evidence that day-to-day changes in the number of felonies is associated with large increases in temperature (two-sided p-value 0.10006).

Conclusions: There is a clear association between warmer temperatures and an increased number of felonies. Presence of precipitation, day of week, holidays, and school days are also related to the number of felonies that occur. There is no evidence that the day-to-day changes in the number of felonies is associated with large increases in temperature.

1 Introduction

After many years of pressure, the New York City Police Department (NYPD) recently released incident-level felony data to the NYC Open Data portal as part of their initiative to improve their accessibility, transparency, and accountability. Prior to this release, felony data had only been provided in an aggregated format (by week and police precinct), and was done so only in PDF and Excel files on a weekly and quarterly basis.

In this paper, we use the newly-released data to examine the association between the daily number of felonies committed in New York City (NYC) and: day of week, outside air temperature, precipitation, federal and New York (NY) holidays, and public school days.

1.1 Questions of Interest

In this paper we study three main questions of interest:

1. Are felonies associated with temperature? After taking temperature into account, is felonies associated with precipitation, school days, holidays, and day of week?
2. Although there's no causal relationship, for a given set of these conditions, how many felonies can the NYPD reasonably expect?
3. After taking into account temperature, precipitation, school days, holidays, and day of week; are the day-to-day changes in the number of felonies associated with large ($> 8°$F) increases in temperature?

1.2 Dataset

The dataset contains 365 observations, one for each date in 2015. The class, description, and source for each variable in the dataset is outlined below. Only those variables used/referenced in the analyses are included here — redundant and untranformed variables that were removed during exploratory analysis are not described below.

- `felonies` (integer) is a count of the number of felonies committed on each day in NYC in 2015. The values are derived counts from the "NYPD 7 Major Felony Incidents" dataset in the NYC Open Data Portal [data.cityofnewyork.us/d/hyij-8hr7]. The felonies included in the dataset that contribute to the overall daily count are burglary, felony assault, grand larceny, grand larceny of motor vehicle, murder and non-negligent manslaughter, rape, and robbery.
- `temp_min_degF` (numeric) is the minimum daily temperature on the given date (in degrees Fahrenheit), as reported by the New York Central Park Belvedere Tower weather station. The data was requested via the National Centers for Environmental Information [ncdc.noaa.gov/cdo-web/search].
- `any_precip` (factor) is an indicator variable, taking value "1" if there was any precipitation on the given date, and "0" otherwise. See `temp_min_degF` for source information.
- `is_holiday` (factor) is an indicator variable, taking value "1" if the given date is a NY or federal holiday, or value "0" otherwise. The NY and federal holidays were defined using the lists provided by the NY State Department of Civil Service [cs.ny.gov/attendance_-leave/2015_legal_holidays.cfm] and U.S. Office of Personnel Management [opm.gov/policy-data-oversight/snow-dismissal-procedures/federal-holidays/#url=2015], respectively.
- `is_school_day` (factor) is an indicator variable, taking value "1" if NYC Public Schools were open and in session on the given date, or value "0" otherwise. Although the NYC Department of Education publishes this data to the NYC Open Data Portal, the historical data is only retained there for the current school year. Instead we scrape the attendance data from XML files in Aaron Schumacher's "NYCattends" Github repository [github.com/ajschumacher/NYCattends/tree/master/xml].

- `day_of_week` (factor) is a categorical variable indicating the day of week (Sunday="1", Monday="2" , . . . , Saturday="7").
- `felonies_diff` (numeric) indicates for a given date the difference in the number of felonies as compared to the previous day. On Jan. 3, 2015, for example, `felonies_diff` = 6, since there were 6 more felonies committed on Jan. 3 than on Jan. 2.
- `temp_min_degF_diff` (numeric) indicates for a given date the difference in `temp_min_degF` as compared to the previous day. On Jan. 3, 2015, for example, `temp_min_degF_diff` = -1.98, since the daily minimum temperature (`temp_min_degF`) was 1.98°F lower on Jan. 3 than on Jan. 2.
- `temp_jump` (factor) is an indicator variable, taking value "1" if `temp_min_degF_diff` > 8, or value "0" otherwise. An increase of > 8°F puts the day of interest in the top 10% of day-to-day temperature increases in 2015.

1.3 Report Overview

In Section 2 we briefly present some exploratory analysis and data transformations.

Then in Section 3 we model the daily number of felonies. Assumptions are discussed in Section 3.1; an exploratory model is presented in Section 3.2; then the analysis and model checking and improvement are presented in Sections 3.3 and 3.4.

In Section 4 we model the day-to-day changes in the number of felonies; again discussing our assumptions, analysis, and model checking and improvement in Sections 4.1, 4.2, and 4.3.

Lastly, in Section 5 we discuss our statistical conclusions and the scope of inference.

2 Exploratory Analysis and Data Cleaning

We first examine pairwise scatterplots of some of the numeric variables in the raw dataset (`felonies`, `temp_min_degF`, `temp_max_degF`, and `school_attendance_pct`), as shown in Figure 1. From this figure we first notice that there are approximately linear relationships between `felonies` and `temp_min_degF`, and between `felonies` and `temp_max_degF`. There is strong collinearity between `temp_min_degF` and `temp_max_degF` (correlation: 0.969), however, so we remove one of these variables (`temp_max_degF`) from the covariates that will be used in the regression model.

Figure 1 also shows that there is no linear relationship between `felonies` and `school_attendance_pct` (the percent of students present in school on a given day). Any non-school-day has 0% attendance, so instead of using this as a numeric variable, we convert it to the `is_school_day` indicator variable, taking value "1" if NYC Public Schools were open and in session on the given date (i.e., if `school_attendance_pct` is > 0), or value "0" otherwise.

Faceted boxplots of the categorical variables are shown in Figure 2.

3 Modeling the Number of Felonies Per Day

In this section we use simple and multiple linear regression to model the number of felonies per day, addressing the first and second questions of interest.

3.1 Assumptions

We first assume that each occurrence of a felony is an independent Bernoulli event with very low probability p. The sum of these Bernoulli events is the number of felonies that occur on a given day. This sum follows a binomial distribution where n is the number of opportunities for a felony to occur — as a rough approximation this might be on the order of the population of NYC (~ 8.4 million). Since n is large enough, we can approximate this binomial distribution with a normal distribution.

We also assume the four assumptions of linear regression:

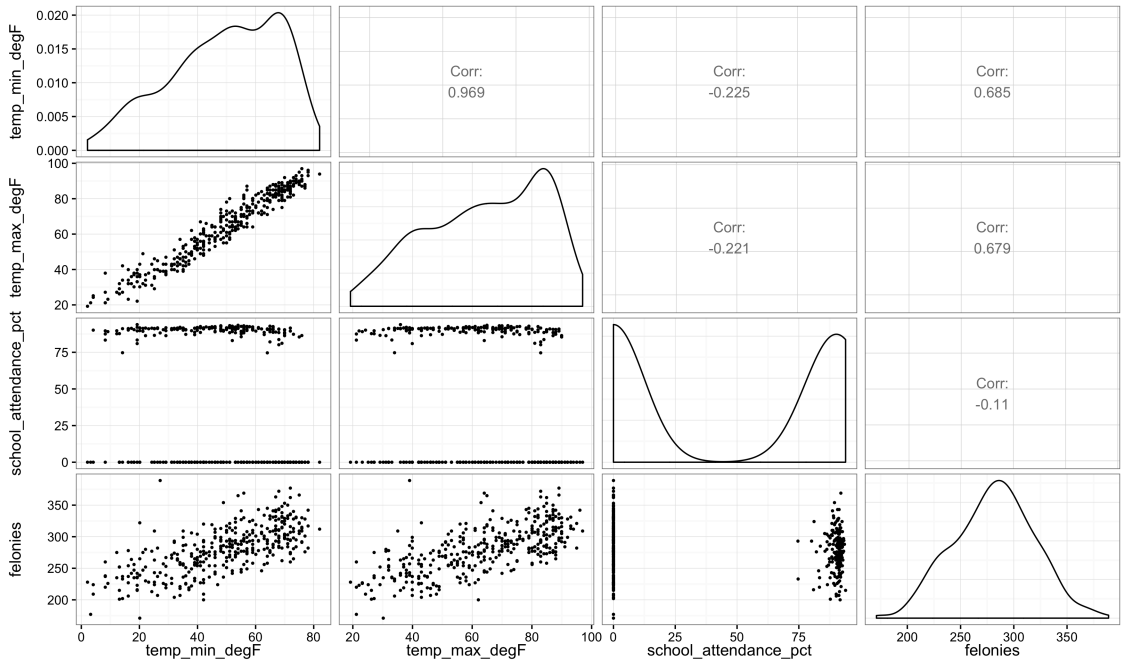

Figure 1: Pairwise scatterplots of some of the numeric variables in the raw dataset

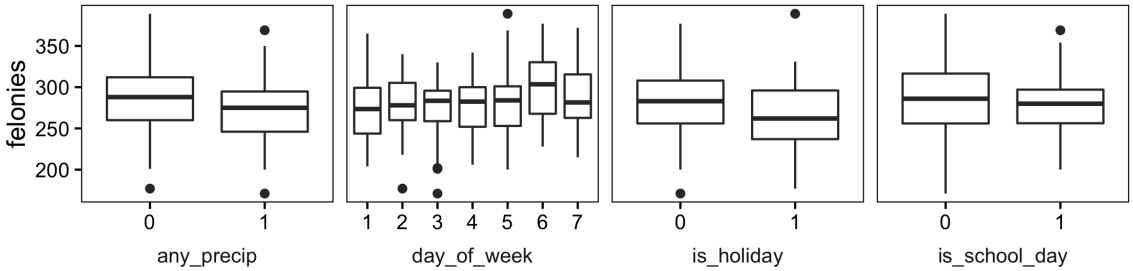

Figure 2: Faceted boxplots of the categorical variables in the dataset.

- Linearity: `felonies` can be expressed as linear combination of the independent variables
- Homoscedasticity (constant variance): $\text{Var}(Y|X_1, \ldots, X_p)$ is the same at all values of $X_1, \ldots, X_p$
- Normality: residuals in the fitted model are normally distributed
- Independence: residuals in the fitted model are independent

3.2 Exploratory Model

As an exploratory step, we initially perform a simple linear regression of felonies on temperature. The regression summary is shown in Table 1.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	213.12	4.07	52.35	$< 2 \times 10^{-16}$
temp_min_degF	1.38	0.08	17.91	$< 2 \times 10^{-16}$

Table 1: Regression summary from the simple linear regression of `felonies`.

There is overwhelming evidence of an association between felonies and temperature. For every 1°F increase in temperature, there are 1.38 additional felonies per day (95% CI 1.22938 to 1.532569, two sided p-value $< 2 \times 10^{-16}$ for a test that the coefficient is 0).

What's most interesting here, however, are the observations with high residuals, as shown in the residual plot in Figure 3. Some days with large residuals aren't modeled well by temperature alone — Jan. 1, 2015, for example, a federal holiday, had many more felonies than we'd expect

given the temperature on that day. Including other covariates in the model, like `is_holiday` (an indicator as to whether the day is a NY/federal holiday), might help to account for some of this behavior. We next add these additional variables in Section 3.3.

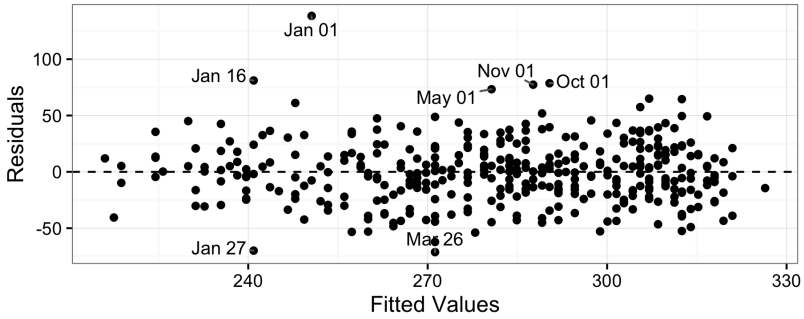

Figure 3: Residual plot for the regression of `felonies` on `temp_min_degF`.

There are some potentially problematic observations with high leverage or large studentized residuals, but there was no change in interpretation after removing these observations from the dataset. Also note that we tested higher order terms of `temp_min_degF` in a multiple regression, but only the first-order term was significant.

3.3 Statistical Analysis

We next incorporate additional covariates, performing a multiple linear regression of felonies on the variables shown in the regression summary in Table 2.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	210.71	5.67	37.19	$< 2 \times 10^{-16}$
temp_min_degF	1.38	0.09	15.75	$< 2 \times 10^{-16}$
any_precip1	-21.01	8.59	-2.44	0.01499
is_holiday1	-6.01	7.68	-0.78	0.4349
is_school_day1	5.95	3.75	1.59	0.1133
day_of_week2	0.07	5.74	0.01	0.9905
day_of_week3	-3.30	5.70	-0.58	0.5636
day_of_week4	-3.47	5.72	-0.61	0.5444
day_of_week5	-0.18	5.68	-0.03	0.9752
day_of_week6	19.01	5.70	3.34	0.0009384
day_of_week7	14.21	5.02	2.83	0.00494
temp_min_degF:any_precip1	0.16	0.17	0.95	0.3434

Table 2: Regression summary from the multiple linear regression of `felonies`.

After accounting for temperature, there is convincing evidence that felonies is associated with precipitation. On average there are 21 fewer felonies on days with precipitation than on days without (95% CI 4 to 38 fewer felonies; two-sided p-value 0.014990). The data also provides overwhelming evidence that felonies is associated with day of week (p-value 3×10^{-6} from an extra sum of squares F-test) — compared to Sundays, Fridays have 19 more felonies on average (95% CI 8 to 30 more felonies; two-sided p-value 0.000938), and Saturdays have 14 more felonies on average (95% CI 4 to 24 more felonies; two-sided p-value 0.004940).

After accounting for temperature, the holiday indicator, school day indicator, and the interaction between temperature and precipitation are not significant (two-sided p-values: 0.434865, 0.113258, and 0.343385, respectively).

3.4 Model Checking and Improvement

To check the validity of our model, we examine a residual plot and Q-Q plot in Figure 4. The residual plot doesn't reveal any significant violations of the linearity, constant variance, or independence

assumptions, and the Q-Q plot shows that the residuals aren't perfectly normal, but that normality isn't a bad approximation.

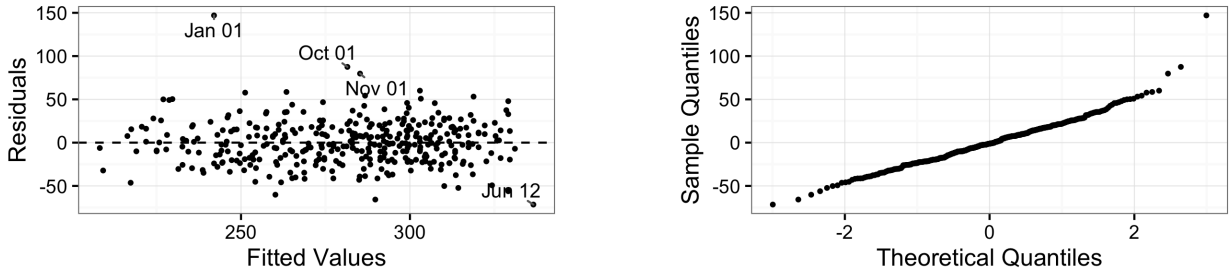

Figure 4: Residual plot and Q-Q plot for the regression of `felonies` on the variables shown in Table 2.

Next, we check for influential observations by examining leverages, studentized residuals, and Cook's distances. From these case influence statistics there are 17 potentially problematic observations. Partial residual plots for `is_holiday` and `is_school_day` (two of the insignificant variables from the regression) are shown in Figure 5. If we ignore the 17 potentially problematic observations (coded with blue triangles in Figure 5), it appears as though holidays have fewer felonies than non-holidays, and school days tend to have more felonies than non-school days.

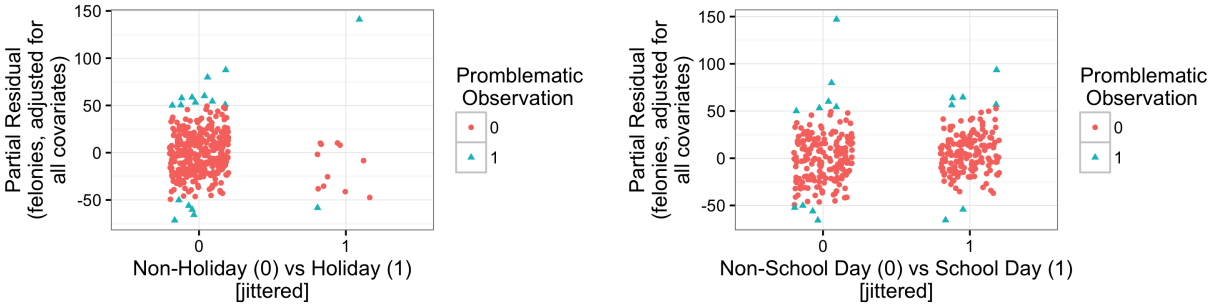

Figure 5: Partial residual plots for `is_holiday` and `is_school_day`. In each plot, the 17 potentially problematic observations are coded with blue triangles.

We re-fit the regression from Section 3.3, but now excluding the 17 potentially problematic observations. The regression summary is shown in Table 3.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	202.30	4.73	42.78	$< 2 \times 10^{-16}$
temp_min_degF	1.48	0.07	20.17	$< 2 \times 10^{-16}$
any_precip1	-20.41	7.20	-2.83	0.004863
is_holiday1	-13.27	6.83	-1.94	0.05275
is_school_day1	6.69	3.13	2.14	0.03306
day_of_week2	4.01	4.82	0.83	0.4055
day_of_week3	0.11	4.73	0.02	0.9819
day_of_week4	-0.19	4.74	-0.04	0.9683
day_of_week5	-2.77	4.80	-0.58	0.5639
day_of_week6	22.99	4.79	4.80	2.409×10^{-6}
day_of_week7	19.50	4.22	4.62	5.448×10^{-6}
temp_min_degF:any_precip1	0.14	0.14	1.00	0.3161

Table 3: Regression summary from the multiple linear regression of `felonies`, after excluding 17 problematic observations.

After removing the 17 problematic observations, the holiday indicator and school day indicator

are now significant. There is suggestive, but inconclusive evidence that `felonies` is associated with holidays — on average, holidays have 13 fewer felonies than non-holidays (95% CI 0 to 27 fewer felonies; two-sided p-value 0.05275). There is moderate evidence that felonies is associated with school days — on average, school days have 7 more felonies that non-school days (95% CI 1 to 13 more felonies; two-sided p-value 0.03306). The interaction between temperature and precipitation still isn't significant, even after removing the 17 observations (two-sided p-value 0.31615).

4 Modeling the Day-to-Day Change in the Number of Felonies

Addressing the third question of interest as a supplementary analysis, we now explore if day-to-day changes in the number of felonies, as compared to the previous day, are associated with large increases in temperature after taking our other covariates into account. The theory here is that spikes in temperature (e.g., at the beginning of a heat wave or on the first few days of spring) might influence the number of felonies that occur. In this section, `temp_jump` indicates if the temperature has increased by more than 8°F as compared to the previous day (see Section 1.2).

4.1 Assumptions

In this section we use multiple linear regression to model `felonies_diff` (the difference in the number of felonies as compared to the previous day; see Section 1.2). Again we assume the 4 assumptions of linear regression, as outlined in Section 3.1.

4.2 Statistical Analysis

We now perform a multiple linear regression of `felonies_diff` on the variables shown in the regression summary in Table 4.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-19.10	6.66	-2.87	0.004397
temp_jump1	21.74	16.38	1.33	0.1854
temp_min_degF	0.12	0.10	1.26	0.2068
any_precip1	-1.91	3.72	-0.51	0.6089
is_holiday1	0.95	9.40	0.10	0.9195
is_school_day1	6.95	4.60	1.51	0.1318
day_of_week2	12.22	7.00	1.75	0.08167
day_of_week3	2.91	6.98	0.42	0.6765
day_of_week4	11.05	7.00	1.58	0.1151
day_of_week5	9.95	6.94	1.43	0.1529
day_of_week6	25.60	6.97	3.68	0.0002747
day_of_week7	1.80	6.14	0.29	0.7702
temp_jump1:temp_min_degF	-0.28	0.35	-0.81	0.4187

Table 4: Regression summary from the multiple linear regression of `felonies_diff`.

After accounting for temperature, precipitation, holidays, school days, and day of week, the data provides no evidence that changes in the number of felonies is associated with large increases in temperature (two-sided p-value 0.185381). The interaction between temperature and the temperature jump indicator is also not significant (two-sided p-value 0.418679).

The data provides overwhelming evidence that `felonies_diff` is associated with `day_of_week` (p-value 0.003788 from an extra sum of squares F-test) — on Fridays, for example, there are on average 26 more felonies than on the preceding Thursday (95% CI 12 to 39 more felonies; two-sided p-value 0.000275).

4.3 Model Checking and Improvement

Neither a residual plot nor a Q-Q plot (not shown) reveal any significant violations of our assumptions. Case influence statistics reveal 12 potentially influential observations, but there is no change

in interpretation once they're removed.

To simplify the model, we now iteratively remove the insignificant variables from the regression, and end up with a model in which only `temp_jump` and `day_of_week` are remaining. The regression summary is shown in Table 5.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	-13.34	4.38	-3.04	0.002517
temp_jump1	8.94	5.42	1.65	0.1001
day_of_week2	16.48	6.11	2.70	0.007358
day_of_week3	7.00	6.11	1.14	0.253
day_of_week4	16.15	6.12	2.64	0.008634
day_of_week5	14.50	6.08	2.38	0.01773
day_of_week6	30.21	6.12	4.94	1.229×10^{-6}
day_of_week7	1.75	6.13	0.28	0.7759

Table 5: Regression summary from the multiple linear regression of `felonies_diff`, after removing the insignificant variables.

The temperature jump indicator still is not significant (two-sided p-value 0.10006). Only the `day_of_week` variable has a significant relationship with day-to-day changes in the number of felonies (p-value 9×10^{-6} from an extra sum of squares F-test) — Fridays, for example, generally have 30 more felonies than the preceding Thursday (95% CI 18 to 42 more felonies; two-sided p-value 1×10^{-6}).

5 Conclusions

5.1 Statistical Conclusions

The data provides overwhelming evidence that felonies is associated with temperature. For every 1°F increase in temperature, there are, on average, 1.38 additional felonies per day (95% CI 1.23 to 1.53 felonies, two sided p-value $< 2 \times 10^{-16}$).

After accounting for temperature, there is convincing evidence that felonies is associated with precipitation. On average there are 21 fewer felonies on days with precipitation than on days without (95% CI 4 to 38 fewer felonies; two-sided p-value 0.014990). The data also provides overwhelming evidence that felonies is associated with day of week (p-value 3×10^{-6} from an extra sum of squares F-test).

Initially, it did not appear that the number of felonies was associated with holidays or with school days, but after removing 17 observations with high studentized residuals, these variables became significant. From the revised model, there is suggestive, but inconclusive evidence that felonies is associated with holidays — on average, holidays have 13 fewer felonies than non-holidays (95% CI 0 to 27 fewer felonies; two-sided p-value 0.05275). Again from the revised model, there is moderate evidence that felonies is associated with school days — on average, school days have 7 more felonies that non-school days (95% CI 1 to 13 more felonies; two-sided p-value 0.03306).

In both the initial and revised models, the interaction between temperature and precipitation was not significant (two-sided p-value was > 0.3 in both models).

From a supplementary analysis, after accounting for day of week, the data provides no evidence that day-to-day changes in the number of felonies is associated with large (> 8°F) increases in temperature, as compared to the previous day (two-sided p-value 0.10006).

5.2 Scope of Inference

As this is purely observational data, these statistical associations cannot be used to draw any causal connections. Further, any generalization of these results to cities other than NYC for time periods other than 2015 is speculative.