Determinants of Hail Size

By:
Katherine Faulkner, Chana Garmier, Carly Kovacik
# Table of Contents

Executive Summary: 3

Description of Data: 4-5

Statistical Analysis: 5-8

Major Findings: 8

Discussion: 9

R Code Appendix: 10-20
Executive Summary:

The purpose of this data analysis is to compare the size of hail stones based on the location (latitude and longitude) of recorded hail stones and the time of day during a severe weather outbreak that took place across the United States on July 16, 2009. The two major outbreaks of severe weather were located across the southern Great Plains as a developing squall line and in New England as a cluster of single cell thunderstorms. There were also a few randomly distributed outbreaks in the northern Great Plains and across the eastern coastline. Across these areas, The National Weather Service collected data that included hail size, hour, and location. The hypothesis for this analysis predicts that hail size will increase as the squall line or the center of the storm approaches the recorded locations. Incorporating the factor of time into the analysis will help determine when the squall line or center of the storm reaches the recorded locations and should correspond to the largest hail size being recorded in late afternoon, due to solar radiation. In conducting this analysis, multiple variations of linear models were executed to provide a best fit for the data and to optimally show the strongest statistical significance between hail size, location of recorded hail, and time of day. The predictor variables included latitude, longitude, and time of day. The response variable was the individual hail stones. Size was measured in millimeters. Time was measured in hours. Latitude was measured in degrees north. Longitude was measured in degrees west. The data was first uploaded into the computer program, R, and then the relationship between hail size and each predictor variable was plotted to gain a general visual understanding of the data. The linear models were then executed and interpreted based upon the multiple R squared values and p values. The major finding from this study was that hail size is not significantly related to any of the predictor variables used. Thus, it can be concluded that the hail data gathered on July 16, 2009 was too random to determine a best fit linear model using R. The relationships provide no statistical significance and therefore, the models should be thrown out.
Description of Data:

Hail stones begin as frozen cloud droplets that are cycled through intense updrafts of a cumulonimbus cloud until they become too heavy to be suspended by the updraft. Hail events usually accompany severe thunderstorms during the spring and summer months in the United States. This particular set of hail data was chosen for analysis because it was recorded from a severe weather outbreak in the middle of summer. This makes it easier to predict a relationship between the size of hail, location of the recorded hail, and the hour of the day. This particular data set was chosen over an event of a single cell thunderstorm outbreak (ordinary thunderstorms) because these storms are generally weaker and shorter lived. Also, a severe weather outbreak may last for more than one day which can make it possible to record multiple data sets as the storms develop and strengthen through a given time period. This makes it more probable to find a relationship among the data. In this particular case, data was used from all of the storms that were picked up by radar and recorded throughout the entire day of July 16, 2009:

<table>
<thead>
<tr>
<th>size</th>
<th>hour</th>
<th>Lat</th>
<th>Lon</th>
</tr>
</thead>
<tbody>
<tr>
<td>175</td>
<td>16.58333</td>
<td>36.63333</td>
<td>86.80000</td>
</tr>
<tr>
<td>75</td>
<td>18.25000</td>
<td>33.43333</td>
<td>86.55000</td>
</tr>
<tr>
<td>100</td>
<td>19.38333</td>
<td>43.05000</td>
<td>75.65000</td>
</tr>
<tr>
<td>75</td>
<td>19.71667</td>
<td>37.58333</td>
<td>95.68333</td>
</tr>
<tr>
<td>175</td>
<td>20.05000</td>
<td>43.56667</td>
<td>74.96667</td>
</tr>
<tr>
<td>75</td>
<td>20.10000</td>
<td>35.18333</td>
<td>95.51667</td>
</tr>
</tbody>
</table>

Due to the fact that the National Weather Service recorded the data of all the storms present on this day, it is near impossible to separate the data from squall line storms or ordinary storms. This indicates that the models executed in this analysis were an attempt to establish a relationship between multiple types of storms. This may hinder any chances of establishing a true relationship among the variables in the data set. In this analysis, the dependent variable is hail size, and the independent variables are the latitude, longitude, and hour, since the hail size will vary by location and time of day. Therefore, the main purpose of executing the R codes will
be to determine how hail size relates to location and time, respectively, using both graphical and numerical features.

**Statistical Analysis:**

The beginning of the analysis consisted of summarizing the hail size data to evaluate general statistical values such as mean, maximum, and minimum size. This provided an understanding of the ranges of the dependent variable. Before choosing which linear models to test, the hail size was plotted against each independent variable to gain a visual understanding of each relationship. Hail data was first plotted against hour:

The lowess line was added to the graph to determine if there was a significant linear relationship between size and hour. It is evident from the plot that no ordinary function will explain a significant portion of this data. However, this does not yet lead to the conclusion that there is no linear model to explain the data. Next, hail size was plotted against latitude and the lowess line was added to it:
It appears that a linear function would match the distribution of hail size versus latitude. It can therefore be predicted that there will be a linear model that explains the relationship between the hail size and latitude data reasonably well. Hail size was also plotted against longitude and the lowess line was added to the graph:
The lowess line seems to faintly represent a type of quadratic function. This may indicate that applying a linear model with a longitude squared component will lead to a statistically significant relationship between hail size and longitude. Now that these relationships have been established, it will be easier to estimate a best fit linear model.

The next step in the analysis of the data was to use different linear regression models to determine which independent variables best explained hail size. The first attempt at creating a regression (see appendix for lm command) including size and hour was the most general type of regression (lm 1). The same basic regression model was then performed for latitude (lm 2) and longitude (lm 3). These results did not prove to be statistically significant so more regressions combining certain variables were then performed. The next attempted regression was a sine relationship between size, latitude, and longitude (lm 4). This regression was chosen based upon the fact that many meteorological processes have been explained using the sine function. However, this regression failed to explain a significant portion of the data. Size was then fit to a model that included a factor of hour squared (lm 5). Once again, the regression was inconclusive and no significant relationship was observed. The other independent variables,
latitude (lm 6) and longitude (lm 7), were also modeled in the same manner. Both models also indicated there was no significant relationship. Three linear regressions were then tested with size compared to each independent variable cubed (lm 8-10). These models proved to be no better than any of the previous regression models. Another three regressions involving each independent variable raised to the fourth power was generated (lm 11-13). There was still no major correlation. Then a regression including all three variables was conducted and there was still no evidence of a significant relationship (lm 14). Based upon the inconclusive results of every regression, hail size was determined to be too random for any of the independent variables to predict. The group of models consistently explained less than 10% of the data, and therefore were all disregarded.

**Major Findings:**

Had this analysis been successful, the resulting information would have enhanced the ability to forecast hail size. However, this analysis did not result in any statistical method to predict hail size based on location and time. The objective of fitting linear regressions was to obtain a multiple R squared value that was greater than .70 and a p value that was less than .05. Lm 1 had a multiple R squared value of .012 and a p value of .33. Lm 2 resulted in a multiple R squared value of .019 and a p value of .21. Lm 3 produced a multiple R squared value of .012 along with a p value of .32. Lm 4 had a multiple R squared value of .076 and a p value of .04. Lm 5 generated a multiple R squared value of .030 and a p value of .29. Lm 6 produced a multiple R squared value of .020 and a p value of .45. Lm 7 created a multiple R squared value of .026 and a p value of .35. Lm 8 resulted in a multiple R squared value of .027 and a p value of .34. Lm 9 generated a multiple R squared value of .020 and a p value of .45. Lm 10 created a multiple R squared value of .025 and a p value of .36. Lm 11 had a multiple R squared value of .023 and a p value of .39. Lm 12 produced a multiple R squared value of .020 and a p value of .45. Lm 13 resulted in a multiple R squared value of .025 and a p value of .37. Lm 14 created a multiple R squared value of .034 and a p value of .44. Unfortunately, none of the models were able to produce the desired multiple R squared values and only one produced a desirable p value (lm4). Therefore, none of the models are a good predictor of hail size given the available data for this particular severe weather outbreak.
**Discussion:**

In conclusion, the analyzed hail data was either completely random with no observable pattern, or the available predictor variables were not statistically significant. Based on the major findings of this analysis, improvements in the approach of this study can be conducted in the future. For instance, this particular study was analyzed using data recorded for only one day. Modeling more than one day’s worth of data could eliminate error or produce an observable trend in the relationship between hail size and its predictor variables. Also, the predictor variables that were used in this particular study could be altered or even replaced. A possible solution to this could be to compare hail size to the moisture content in the air near the observed hail. Results may have also improved if the data used in this study had been narrowed down to the path of a particular storm or to a specific region in the United States. Narrowing the data in this manner may help prevent any outliers, which negatively contribute to the results. This analysis did not result in a more applicable way to forecast hail or explain the relationships between hail size, location, and time of day. The question of how to determine hail size during a severe weather outbreak remains unsolved. A subsequent study that uses different predictor variables, data for more than one day, and a narrower area of study may be able to enhance the ability of forecasting hail in the future.
R Code Appendix:

```R
> haildata = read.table("clipboard", header=T)
> head(haildata)

size     hour      Lat      Lon
1  175 16.58333 36.63333 86.80000
2   75 18.25000 33.43333 86.55000
3  100 19.38333 43.05000 75.65000
4   75 19.71667 37.58333 95.68333
5  175 20.05000 43.56667 74.96667
6   75 20.10000 35.18333 95.51667

> size = haildata$size
> hour = haildata$hour
> Lat = haildata$Lat
> Lon = haildata$Lon

> summary(size)

     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.  
    75.00   88.00    100.00   107.60   100.00   300.00

> plot(size~hour)
> plot(size~Lat)
> plot(size~Lon)
```
**Lm 1:**

> summary(lm(size~hour))

Call:

lm(formula = size ~ hour)

Residuals:

```
  Min 1Q Median 3Q Max
-57.239 -20.700 -8.684 -5.473 190.892
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 135.727     29.142   4.657  1.26e-05 ***
hour        -1.210      1.234  -0.980     0.33
```

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.1 on 80 degrees of freedom

Multiple R-squared: 0.01187,    Adjusted R-squared: -0.0004818

F-statistic: 0.961 on 1 and 80 DF,  p-value: 0.3299

---

**Lm 2:**

> summary(lm(size~Lat))

Call:

lm(formula = size ~ Lat)

Residuals:

```
  Min 1Q Median 3Q Max
```
Coefficients:

   Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.194     48.913   0.944    0.348
Lat           1.582     1.255   1.261    0.211

Residual standard error: 43.93 on 80 degrees of freedom
Multiple R-squared: 0.01948, Adjusted R-squared: 0.007228
F-statistic:  1.59 on 1 and 80 DF,  p-value: 0.2110

Lm 3:

> summary(lm(size~Lon))

Call:
lm(formula = size ~ Lon)

Residuals:

       Min      1Q  Median      3Q     Max

Coefficients:

   Estimate Std. Error t value Pr(>|t|)
(Intercept) 145.2573    37.9853   3.824 0.000259 ***
Lon      -0.4212     0.4209 -1.001 0.320014

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.09 on 80 degrees of freedom
Multiple R-squared: 0.01236, Adjusted R-squared: 1.599e-05
F-statistic: 1.001 on 1 and 80 DF,  p-value: 0.32
**Lm 4:**

```r
> fit=lm(size ~ sin(Lon) + sin(Lat))
> summary(fit)
```

Call:
```
lm(formula = size ~ sin(Lon) + sin(Lat))
```

Residuals:
```
       Min      1Q Median       3Q      Max
-52.466 -22.572  -7.919   4.023  189.557
```

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 106.910      4.827  22.150   <2e-16 ***
sin(Lon)   -14.784      6.857  -2.156   0.0341 *
sin(Lat)   -7.867      6.968  -1.129   0.2623
```

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.91 on 79 degrees of freedom

Multiple R-squared: 0.07624,   Adjusted R-squared: 0.05285

F-statistic:  3.26 on 2 and 79 DF,  p-value: 0.04362

**Lm 5:**

```r
> summary(lm(size~hour+c(hour^2)))
```

Call:
```
lm(formula = size ~ hour + c(hour^2))
```

Residuals:
```
       Min      1Q Median       3Q      Max
-38.757 -21.777 -10.466   -5.105  188.947
```
Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 79.99948 | 53.36162   | 1.499   | 0.138    |
| hour           | 3.98236  | 4.34858    | 0.916   | 0.363    |
| c(hour^2)      | -0.11686 | 0.09387    | -1.245  | 0.217    |

Residual standard error: 43.95 on 79 degrees of freedom

Multiple R-squared: 0.03088,  Adjusted R-squared: 0.006346

F-statistic: 1.259 on 2 and 79 DF,  p-value: 0.2897

**Lm 6:**

```r
> summary(lm(size~Lat+c(Lat^2)))
```

Call:

`lm(formula = size ~ Lat + c(Lat^2))`

Residuals:

```
  Min     1Q    Median     3Q    Max
```

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -92.25383| 549.49825  | -0.168  | 0.867    |
| Lat            | 8.77987  | 28.48178   | 0.308   | 0.759    |
| c(Lat^2)       | -0.09263 | 0.36617    | -0.253  | 0.801    |

Residual standard error: 44.19 on 79 degrees of freedom

Multiple R-squared: 0.02028,  Adjusted R-squared: -0.004525

F-statistic: 0.8176 on 2 and 79 DF,  p-value: 0.4452
**Lm 7:**

```r
> summary(lm(size~Lon+c(Lon^2)))
```

**Call:**

```
lm(formula = size ~ Lon + c(Lon^2))
```

**Residuals:**

```
Min 1Q Median 3Q Max
-43.009 -25.653 -13.825 -2.678 197.788
```

**Coefficients:**

```
  Estimate Std. Error t value Pr(>|t|)
(Intercept) 523.46530 359.04104   1.458    0.149
Lon -9.19976  8.29767 -1.109    0.271
```

Residual standard error: 44.06 on 79 degrees of freedom

Multiple R-squared: 0.02619, Adjusted R-squared: 0.001541

F-statistic: 1.062 on 2 and 79 DF, p-value: 0.3505


**Lm 8:**

```r
> summary(lm(size~hour+c(hour^3)))
```

**Call:**

```
lm(formula = size ~ hour + c(hour^3))
```

**Residuals:**

```
Min 1Q Median 3Q Max
-37.261 -21.954 -10.205 -5.848 189.365
```

**Coefficients:**

```
  Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.333235  49.078233  1.881  0.0636 .
```

15
hour         1.665762   2.894415   0.576   0.5666
c(hour^3)  -0.001723   0.001569  -1.098   0.2755
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 44.05 on 79 degrees of freedom
Multiple R-squared: 0.02673,    Adjusted R-squared: 0.002086
F-statistic: 1.085 on 2 and 79 DF,  p-value: 0.343

**Lm 9:**

```r
> summary(lm(size~Lat+c(Lat^3)))

Call:
lm(formula = size ~ Lat + c(Lat^3))

Residuals:
     Min      1Q  Median      3Q     Max

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -27.846458  370.874466 -0.075   0.940
Lat                         4.474600   14.416848  0.310   0.757
 c(Lat^3)              -0.000635   0.003152 -0.201   0.841

Residual standard error: 44.2 on 79 degrees of freedom
Multiple R-squared: 0.01999,    Adjusted R-squared: -0.004823
F-statistic: 0.8056 on 2 and 79 DF,  p-value: 0.4504
```
**Lm 10:**

```r
> summary(lm(size~Lon+c(Lon^3)))
```

Call:
```
lm(formula = size ~ Lon + c(Lon^3))
```

Residuals:
```
     Min      1Q  Median      3Q     Max
-42.813 -25.995 -13.890 -2.597  197.844
```

Coefficients:
```
                    Estimate Std. Error   t value  Pr(>|t|)
(Intercept)  3.820e+02  2.328e+02  1.641   0.105
Lon      -4.559e+00  4.036e+00 -1.130   0.262
```

Residual standard error: 44.07 on 79 degrees of freedom
Multiple R-squared: 0.02547, Adjusted R-squared: 0.0007994
F-statistic: 1.032 on 2 and 79 DF, p-value: 0.3609

**Lm 11:**

```
summary(lm(size~hour+c(hour^4)))
```

Call:
```
lm(formula = size ~ hour + c(hour^4))
```

Residuals:
```
     Min      1Q  Median      3Q     Max
```

Coefficients:
```
                    Estimate Std. Error   t value  Pr(>|t|)
(Intercept)  1.006e+02  4.665e+01  2.156   0.0341 *
```
hour  8.022e-01  2.423e+00  0.331  0.7415
c(hour^4)  -3.398e-05  3.521e-05  -0.965  0.3375
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 44.12 on 79 degrees of freedom
Multiple R-squared: 0.02338,    Adjusted R-squared: -0.001342
F-statistic: 0.9457 on 2 and 79 DF,  p-value: 0.3928

Lm 12:
> summary(lm(size~Lat+c(Lat^4)))
Call:
  lm(formula = size ~ Lat + c(Lat^4))
Residuals:
     Min      1Q  Median      3Q     Max
-39.11 -24.52 -11.36  -2.16  197.01
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.397e+00  2.821e+02   0.016    0.988
Lat            3.036e+00  9.748e+00   0.311    0.756
c(Lat^4)       -6.090e-06  4.048e-05  -0.150    0.881
Residual standard error: 44.2 on 79 degrees of freedom
Multiple R-squared: 0.01977,    Adjusted R-squared: -0.005051
F-statistic: 0.7965 on 2 and 79 DF,  p-value: 0.4545
**Lm 13:**

>`summary(lm(size~Lon+c(Lon^4)))`

Call:

`lm(formula = size ~ Lon + c(Lon^4))`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-42.608</td>
<td>-26.274</td>
<td>-13.953</td>
<td>-2.530</td>
<td>197.880</td>
</tr>
</tbody>
</table>

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|---------|
| (Intercept)| 3.113e+02| 1.706e+02  | 1.825   | 0.0718  |
| Lon        | -3.010e+00| 2.627e+00 | -1.146  | 0.2554  |
| c(Lon^4)   | 9.320e-07| 9.336e-07 | 0.998   | 0.3212  |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.09 on 79 degrees of freedom

Multiple R-squared: 0.02467,  Adjusted R-squared: -2.587e-05

F-statistic: 0.999 on 2 and 79 DF,  p-value: 0.3729

---

**Lm 14:**

>`summary(lm(size~hour+Lat+Lon))`

Call:

`lm(formula = size ~ hour + Lat + Lon)`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
</table>

---

19
Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 61.70453 | 104.10527  | 0.593   | 0.555    |
| hour       | -1.38417 | 1.31866    | -1.050  | 0.297    |
| Lat        | 1.83242  | 1.73015    | 1.059   | 0.293    |
| Lon        | 0.07814  | 0.59059    | 0.132   | 0.895    |

Residual standard error: 44.16 on 78 degrees of freedom

Multiple R-squared: 0.03385, Adjusted R-squared: -0.003305

F-statistic: 0.9111 on 3 and 78 DF, p-value: 0.4396