Hurricane Sandy: 311 open data analysis of pre and post

**Final Paper: Foundation Module**

**TEAM MEMBERS**

1. Kim, Tae ([thk301@nyu.edu](mailto:thk301@nyu.edu))
2. Putro, Dimas Rinarso ([drp354@nyu.edu](mailto:drp354@nyu.edu))
3. Zhuang, Yuzheng ([yz2611@nyu.edu](mailto:yz2611@nyu.edu))

# Table of contents

Table of contents 2

Table of figure 3

1. Introduction 4

2. Time series plot 4

3. ACF and PACF test 5

4. The use of maps and visual representations 6

5. Bivariate regression analysis 6

5.1. Number of tree VS number of complaints 6

5.2. Number of accidents VS number of complaints 8

6. Conclusions 8

Table of figures 9

7. References 21

# Table of figure

Figure 1 All time series 2011-2014 9

Figure 2 2011 time series plot of top 3 complaints to agency in New York City 9

Figure 3 2012 time series plot of top 3 complaints to agency in New York City 10

Figure 4 2013 time series plot of top 3 complaints to agency in New York City 10

Figure 5 ACF (left) and PACF (Right) of the DPR complaints for Oct and Nov 2011 (top), 2012 (middle), 2013 (bottom) 11

Figure 6 Map of DPR complaints 2011 12

Figure 7 Map of DPR complaints 2012 12

Figure 8 Map of DPR complaints 2013 13

Figure 9 number of trees VS number of DPR complaints 2011 13

Figure 10 number of trees VS number of DPR complaints 2012 14

Figure 11 number of trees VS number of DPR complaints 2013 14

Figure 12 Statistics summary for tree VS DPR complaints 2011 15

Figure 13 Statistics summary for tree VS DPR complaints 2012 16

Figure 14 Statistics summary for tree VS DPR complaints 2012 17

Figure 15 Accident VS DPR complaints 2013 18

Figure 16 Accident VS DPR complaints 2012 18

Figure 17 Statistics summary accident VS DPR complaints 2013 19

Figure 18 Statistics summary accident VS DPR complaints 2013 20

# Introduction

New York City has its own story on how it bounced back from one of the biggest disasters ever hit the East Coast since the 20th Century, Hurricane Sandy. With 147 direct deaths reported, in which 72 of m occurred in the mid-Atlantic and Northeastern US, Hurricane Sandy caused the biggest direct fatalities related to tropical cyclones outside the southern areas since 1972 (Eric et al., 2013). This assignment paper examines 311 dataset changes and patterns in the year 2011, 2012, and 2013 respective to the time range close to Hurricane Sandy (between October 1st – December 31st), to see unseen patterns that can be extracted from the event.

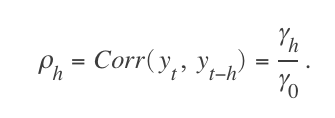
In this team I worked together with Tae Kim ([thk301@nyu.edu](mailto:thk301@nyu.edu)) and Yuzheng Zhuang ([yz2611@nyu.edu](mailto:yz2611@nyu.edu)). My part is to filter and clean the data for other team members. The other tasks were to make visual presentation using Bokeh library in Python and produce correlation analysis using Pandas, Numpy and Statsmodel package in Python.

# Time series plot

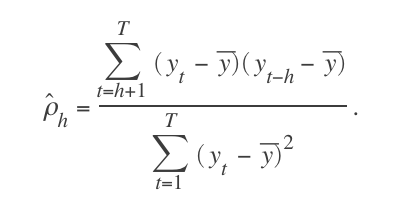
Our first step is to categorize the complaints by agency and plot the time series to find visually see the patterns. From figure 1 to figure 4 we could see that regardless of the scale of number of complaints, NYPD and TLC were showing minimum response to the event of disaster, constantly maintained the trend without noticeable peaks. DPR, on the other hand showed peaks in 2011 and 2012 in figure 2 and figure 3 respectively, with in 2012 it reached maximum value of 4400. This provided us hints that DPR trends in 2011 and 2012 differ compared to 2013. Another clue provided by this time series plot is that the negatives trends NYPD complaints in 2012 when DPR reached it peak value. From this we understood that we wanted to limit our scope of observation to only DPR and the next step to statistically confirm the state of abnormalities for DPR.

# ACF and PACF test

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) was preferred to auto regression method since we wanted to particularly set a test to see, through the correlogram tools, whether in specific year DPR presents abnormal behavior, specified by the level of randomness between data point in certain time interval. This could be achieved by observing the relation of the data at *t* and *t*-*h*, where *h* is a time lag and in this case was set in the range of 1-20 days, and then see whether it falls within certain threshold of level of confidence (95%):



The denominator γ0 is the lag 0 covariance. The autocorrelation function (ACF) for a time series *yt*, t = 1,...,N, is the sequence *ρh, h* = 1, 2,...,N – 1. The partial autocorrelation function (PACF) is the sequence ϕh,h*,h* = 1, 2, ..., N – 1.The lag-*h* autocorrelation would be then obtained by:



To visually observe this, we could see the correlogram presented in figure 5. We could see from ACF (left) figure that in 2011 and 2012, the lag equal to 0 and 1 fell outside the confidence level threshold, while in 2013 lag = 1 was within the boundaries. In addition, we could see that in 2013, the autocorrelation value for DPR fluctuates in smaller periodic in respect to time intervals. This was caused by increased randomness driven by number of complaints scattered. We also figured out that PACF, while showing similar pattern difference 2011-2012 vs 2013, was more difficult to observe see, thus, for setting up threshold of abnormality for natural disaster events, ACF was considered to be simple, relevant model complementing the use of manual time series observation. From the figure none of the correlation value pass the confidence limit test (all within 95% confidence level), although ACF for 2012 when when lag *h* is equal to 1 passed the confidence limit test. Therefore, in principle we could not fit any model. Since the data does not show any auto correlation, we decided that ACF and PACF is not be the most accurate test since it failed to predict the patterns in the time series data. However, it might be useful to provide visual representation on how the characteristic of the data in 2011, 2012 and 2013 differs.

# The use of maps and visual representations

Since we know that statistically there were different in patterns of DPR’s number of complaints in 2011-2012 versus 2013, we could not tell what is the particular reason why the pattern was different and what variables were correlated. Thus, we did visual representations by plotting the number of DPR complaints per zip code by longitude and latitude given in 311 dataset, and plot it on the map. The result can be seen in figure 6, 7 and 8. From here we could observe that while Sandy’s effect is apparent in 2012, there were significant influx of complaints in Staten Island on 2011, while in 2013 is mostly flat. This might be related to the snow storms that in vast majority occurred in Staten Island in October 2011 (Silive, 2011). The next step is to find good predictor for regression analysis and build the model.

# Bivariate regression analysis

## Number of tree VS number of complaints

First, we obtained the number of trees data in New York City and try to find the correlation by running bivariate regression model with both linear and second order polynomial regression. This is to confirm that the 2 disasters, snowstorm in 2011 and Sandy in 2012 were assumed largely affected by trees. In this paper we use linear regression equation as follows:

*Y'* = ß*x + const*

Where ß is linear coefficient and *const* is the constants. Second order polynomial, on the other hand, was given in standard polynomial equation:



or, equally:



where *a2* and *a1* in the statistics represent ß0 and ß1is and *a0* represents the constants. Our finding shows that in 2011 and 2012, in figure 9 and 10, characterized by increased R–squared value at 0.59 and 0.81 for both linear and second order polynomial fit, respectively. This indicates that the model explains most of the variability of the sample data around its mean. R-squared with the value 0.59 means 59% of the number of DPR complaints were affected by the change of by the number of value. Linear fit, represented by black bold line, pointed an increased value from ß = 1.30 to ß = 1.83 from 2011 to 2012, while it decreased to to from ß = 0.77 in 2013, when there was no disaster events. 2011-2012 presented indication of super linearity correlation, meaning that the number of DPR complaints has more increased rate relative to the number of trees. However, 95% confidence level, as shown in figure 12-14 fell between 0.932 - 1.670 and 1.543 - 2.126 in 2011 and 2012 respectively, which tells us in 2011 there under the 95% level of confidence there was a possibility of sub linear characteristics (ß <1).

The other interesting finding from the tree and DPR number of complaints is that the both of the model, linear and second order polynomial could maintain the similar result, confirming that tree and the number of complaints do scale linearly and therefore no need to use more sophisticated model. This resonates with some emphasis given in class that “in most cases linear approach is the best model to start with”, as linear regression approach was also an interesting subject explored by other urban planners such as in identify the famous “economic of scale” concept proposed in (Luis and Geoffrey, 2007).

## Number of accidents VS number of complaints

Again we ran bivariate regression test with both linear and second order polynomial regression against number of road accidents for each year in 2012 and 2013 due to the limitation that 2011 data is not available. Moreover we were not interested in comparing 2011 because it also refers to other disaster which can state bias. We would like to compare Sandy and non-disaster state, hence 2012-2013 was chosen. Our finding in figure 15 and 16, show correlations characterized by weak R–squared value at 0.01 and 0.04 for 2012 and 2013, respectively. The low R-squared hinted that there relation of accident and DPR complaints were not statistically significant. Moreover, since we know from observation 5.1 that trees are strongly correlated to DPR, we could imply that number of trees does not necessarily correlate with number of accident.

# Conclusions

In this foundation module final paper we could see how basic applied statistics learned in the class could help us detect abnormality in 311 complaints data in the case of disaster. Moreover, it helped us analyzing some unique phenomenon that have happened during Sandy and be able to provide us hints on which visual representations might be applicable and which new set of predictors could be added before we can evaluate statistically by observing R-squared and so on. For instance, it guided us to the conclusions that tree is the root cause of top 311 complaints during disaster, although, however, it did not cause traffic accident. This capability of manually observe and pick more meaningful model and predictors is essential feature in basic applied statistics when compared to the likes of neural network and other pattern recognition methods.

# Table of figures

# Macintosh HD:Users:dimasrinarso:Dropbox:5004:tae_ads_foundation:combined_all.png

Figure 1 All time series 2011-2014

# Macintosh HD:Users:dimasrinarso:Dropbox:5004:tae_ads_foundation:2011_3agencies.png

Figure 2 2011 time series plot of top 3 complaints to agency in New York City

# Macintosh HD:Users:dimasrinarso:Dropbox:5004:tae_ads_foundation:2012_3agencies.png

Figure 3 2012 time series plot of top 3 complaints to agency in New York City

# Macintosh HD:Users:dimasrinarso:Dropbox:5004:tae_ads_foundation:2013_3agencies.png

Figure 4 2013 time series plot of top 3 complaints to agency in New York City

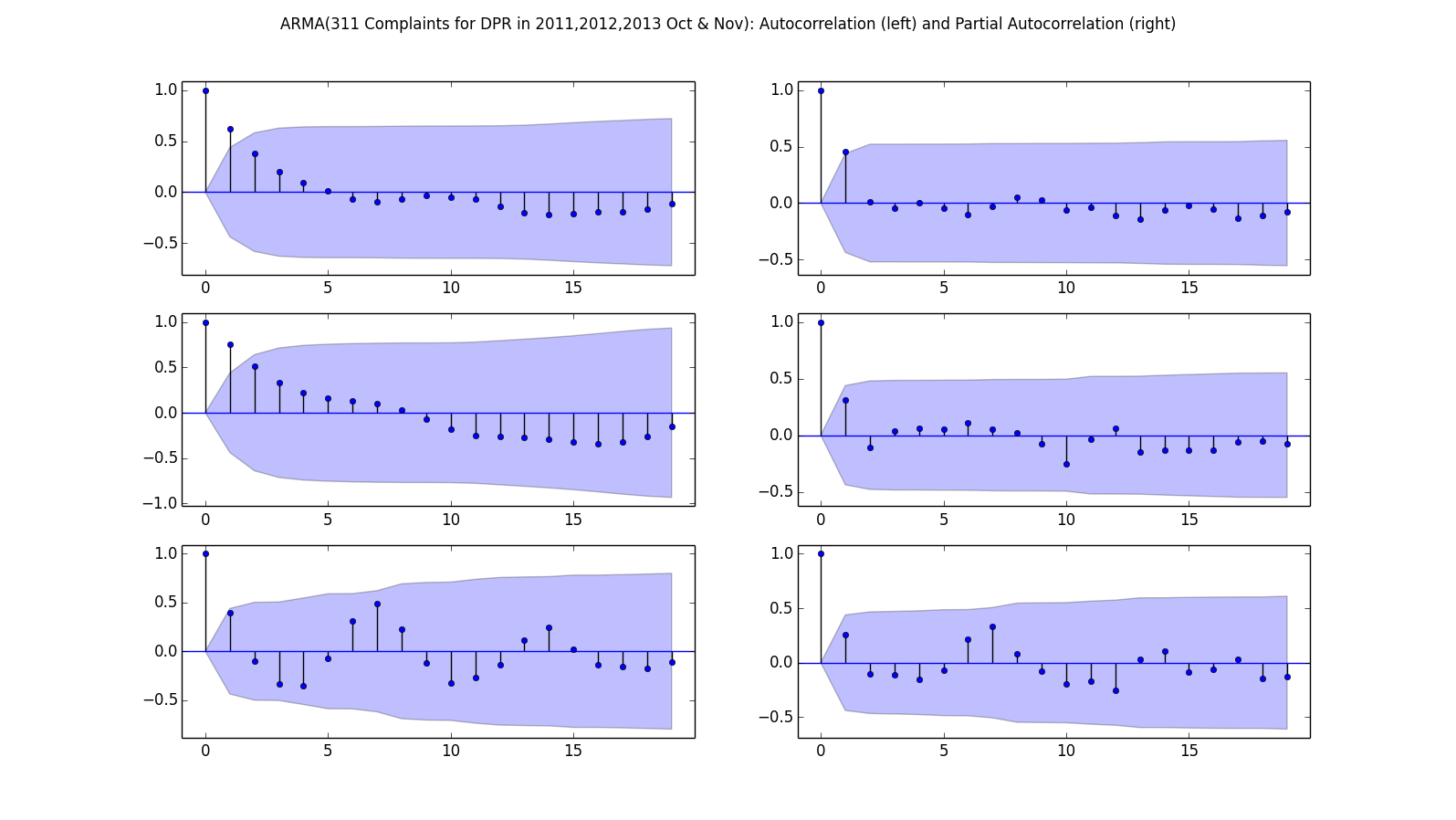


Figure 5 ACF (left) and PACF (Right) of the DPR complaints for Oct and Nov 2011 (top), 2012 (middle), 2013 (bottom)

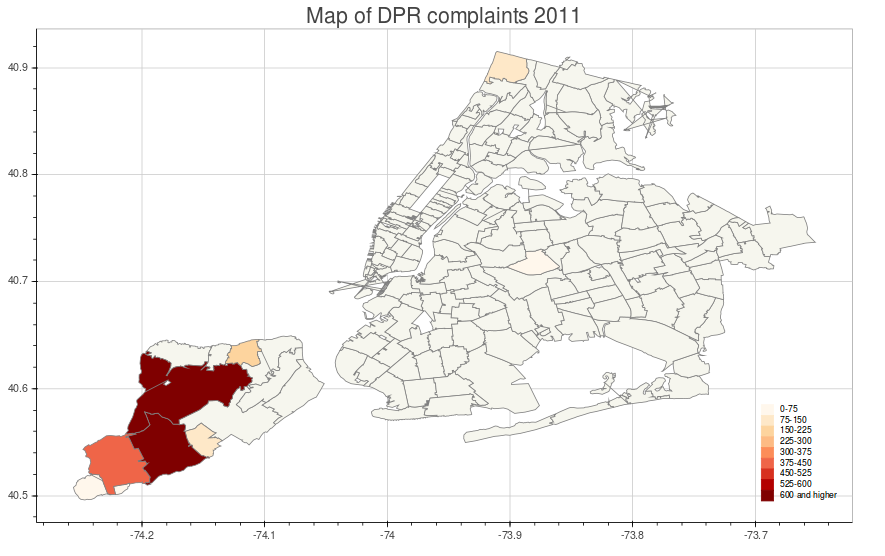


Figure 6 Map of DPR complaints 2011

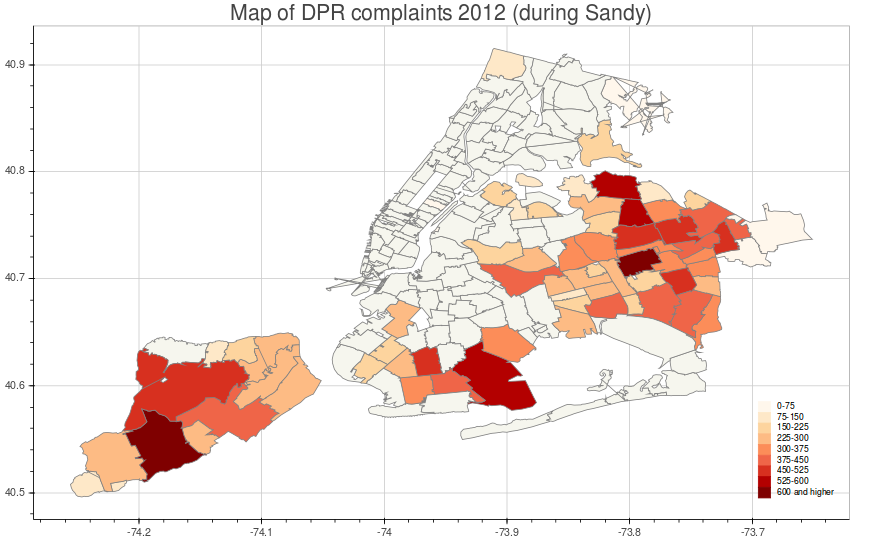


Figure 7 Map of DPR complaints 2012

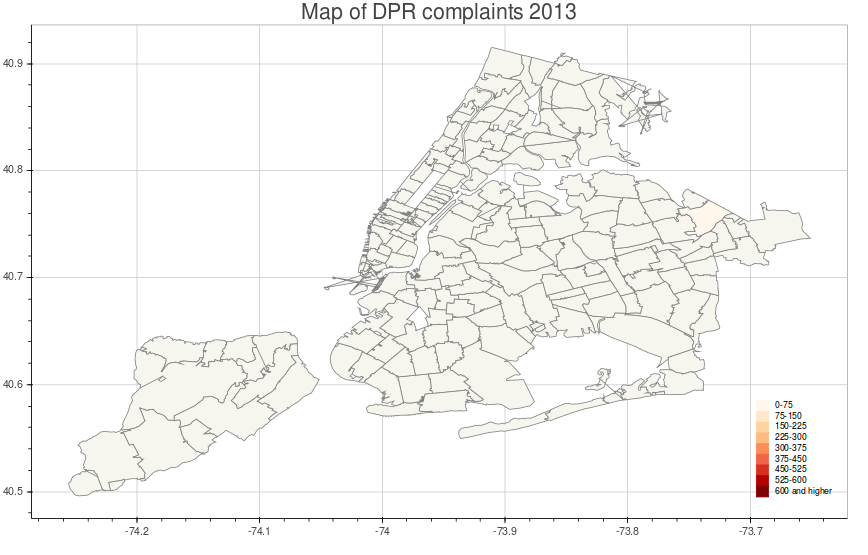


Figure 8 Map of DPR complaints 2013

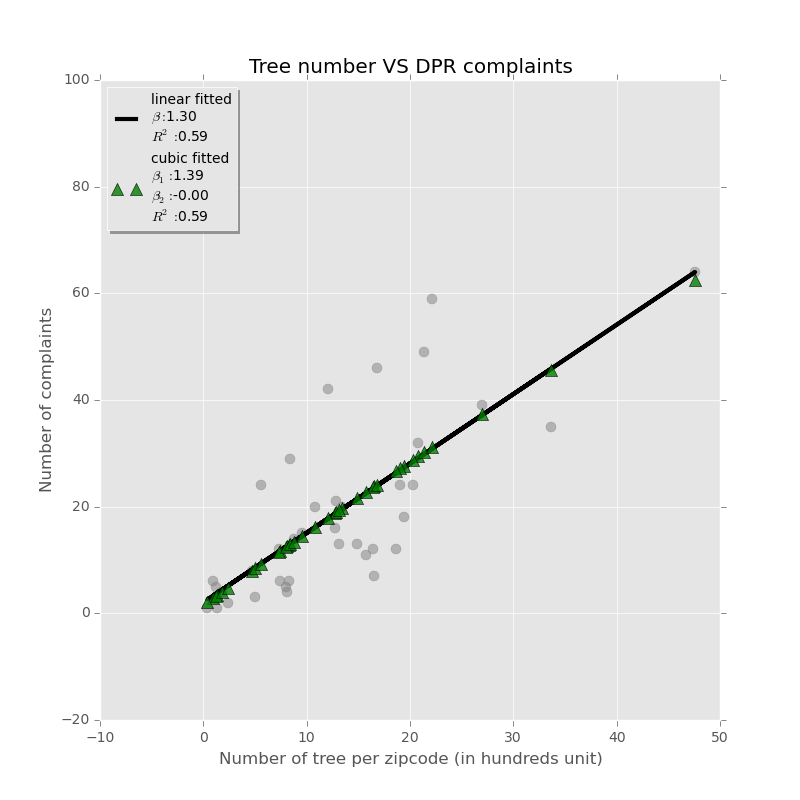


Figure 9 number of trees VS number of DPR complaints 2011

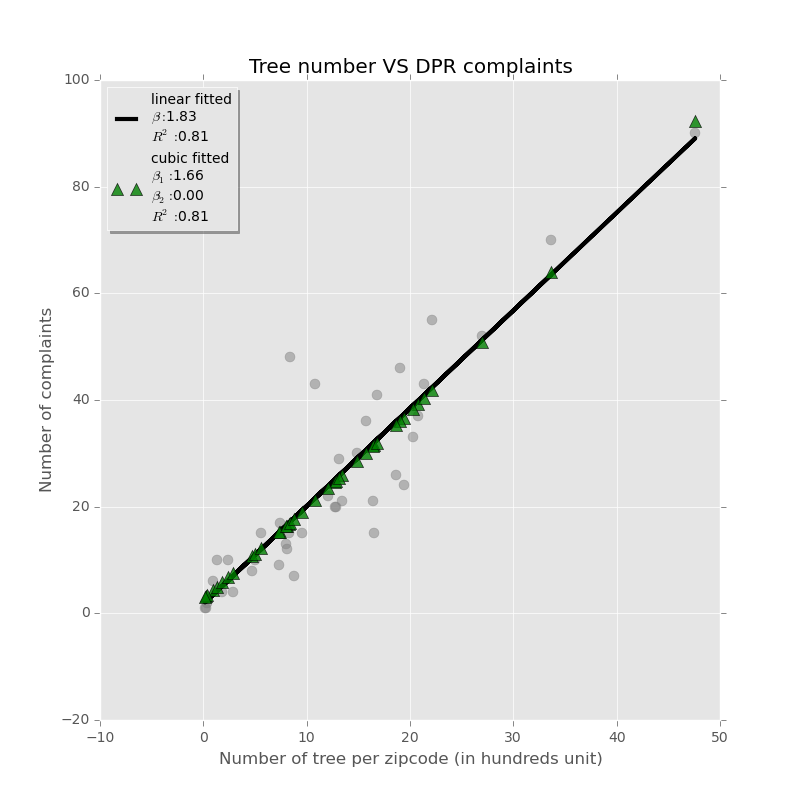


Figure 10 number of trees VS number of DPR complaints 2012

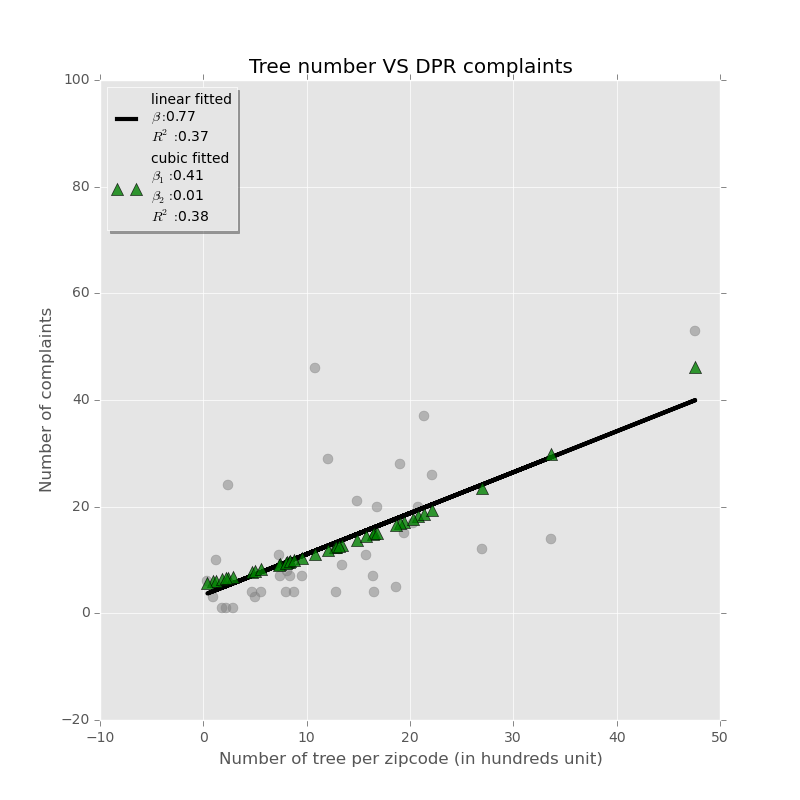


Figure 11 number of trees VS number of DPR complaints 2013

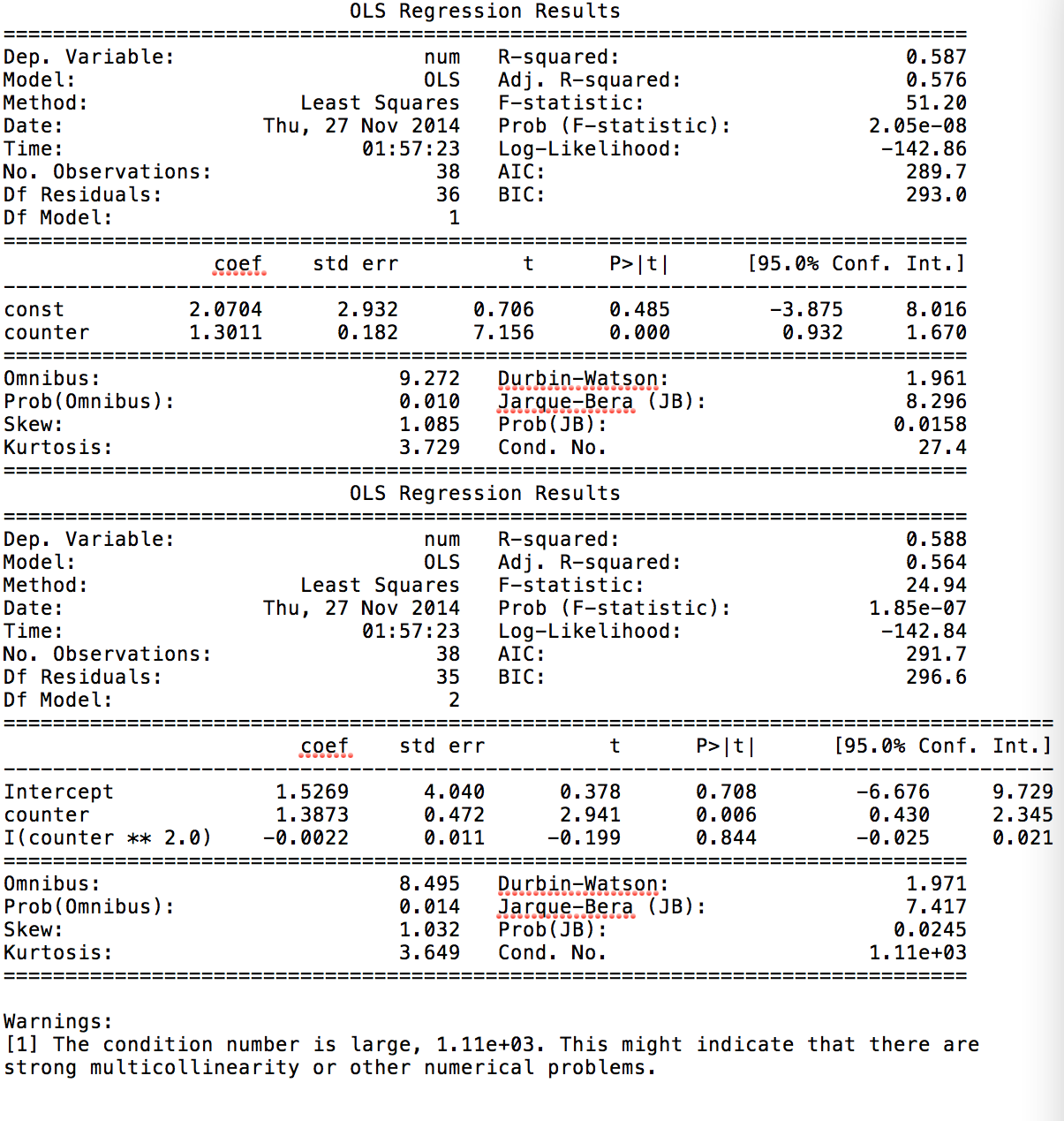


Figure 12 Statistics summary for tree VS DPR complaints 2011

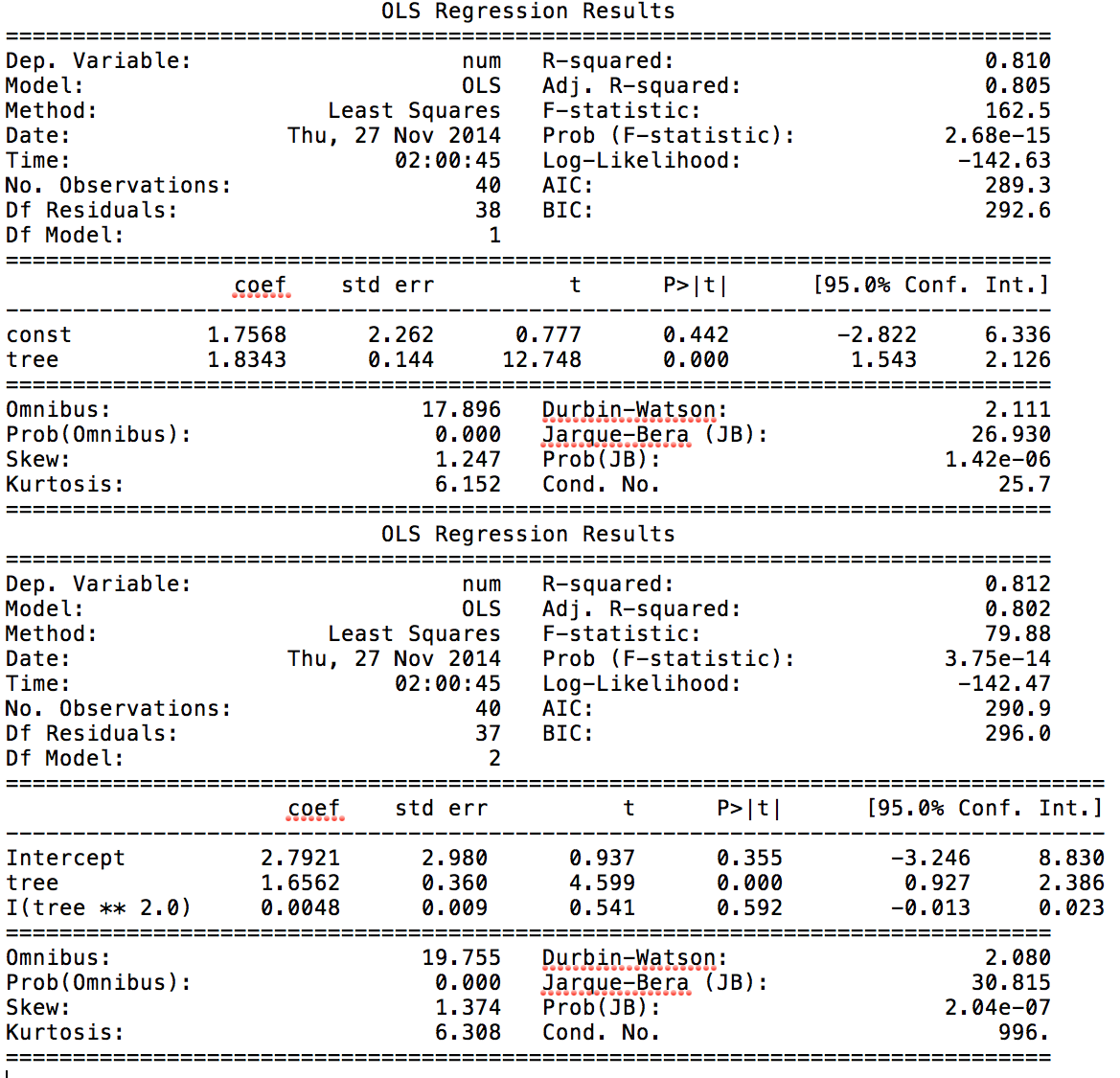


Figure 13 Statistics summary for tree VS DPR complaints 2012

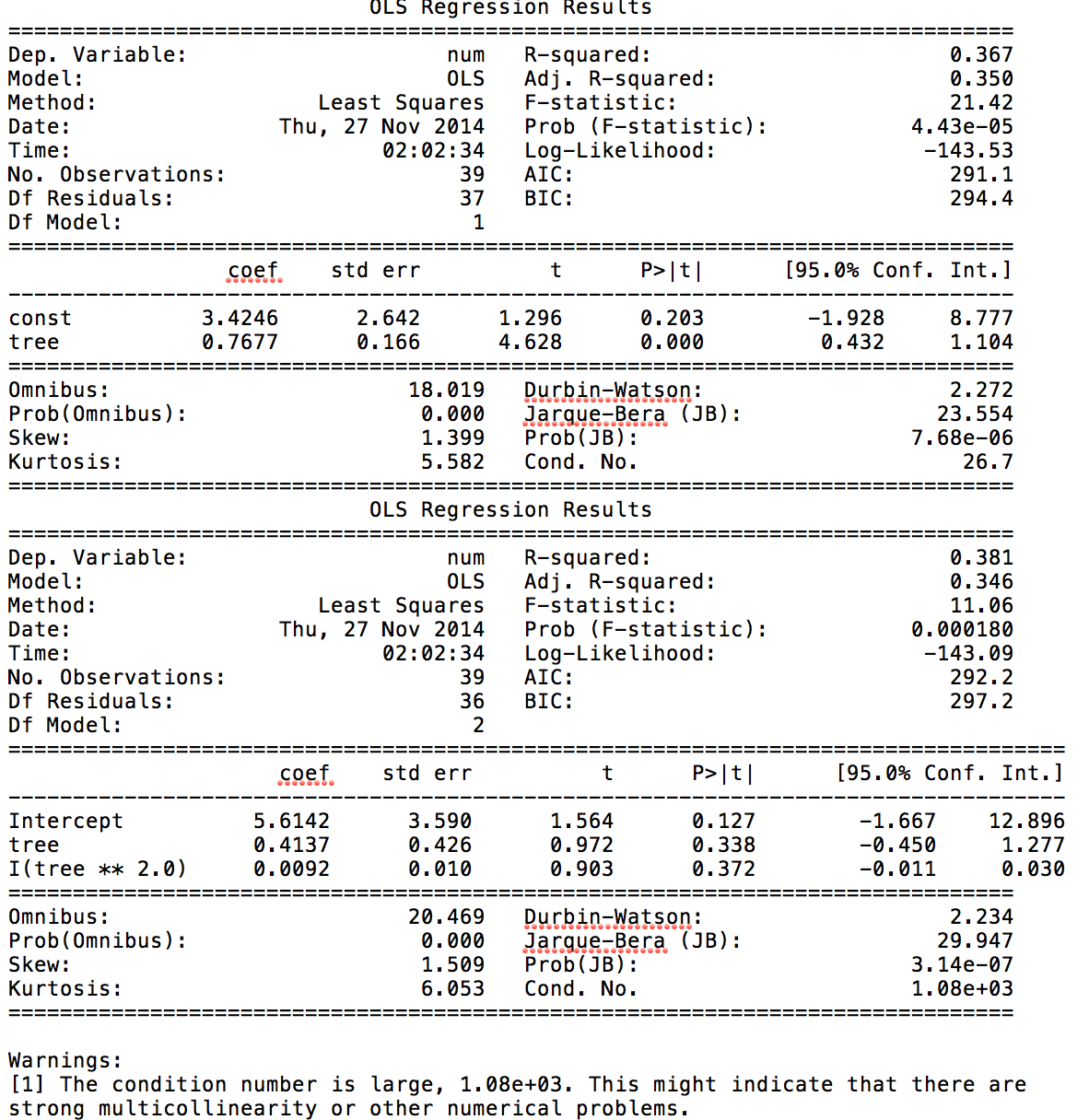


Figure 14 Statistics summary for tree VS DPR complaints 2012

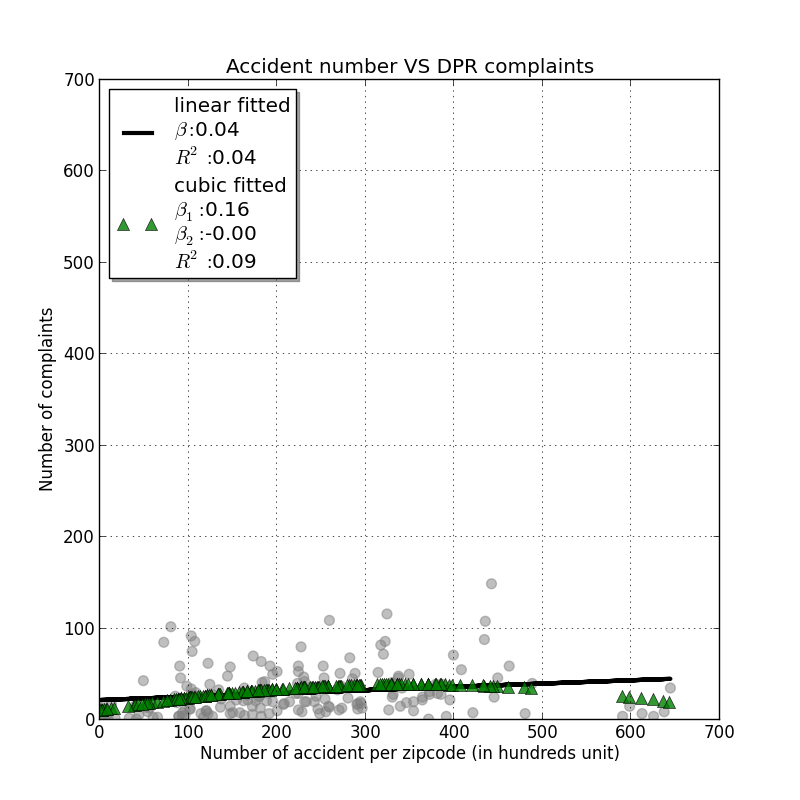


Figure 15 Accident VS DPR complaints 2013

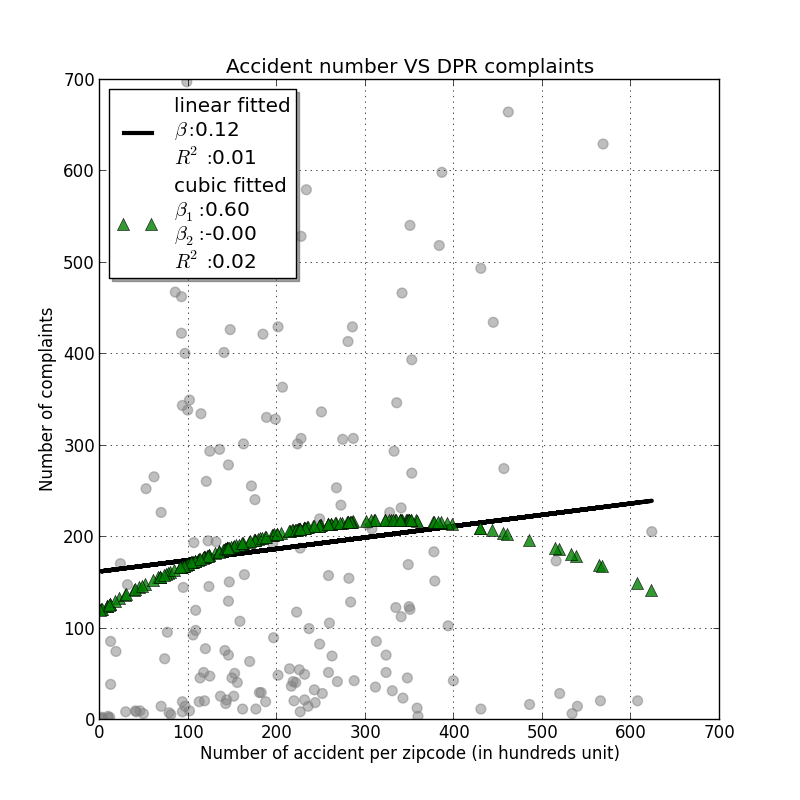


Figure 16 Accident VS DPR complaints 2012

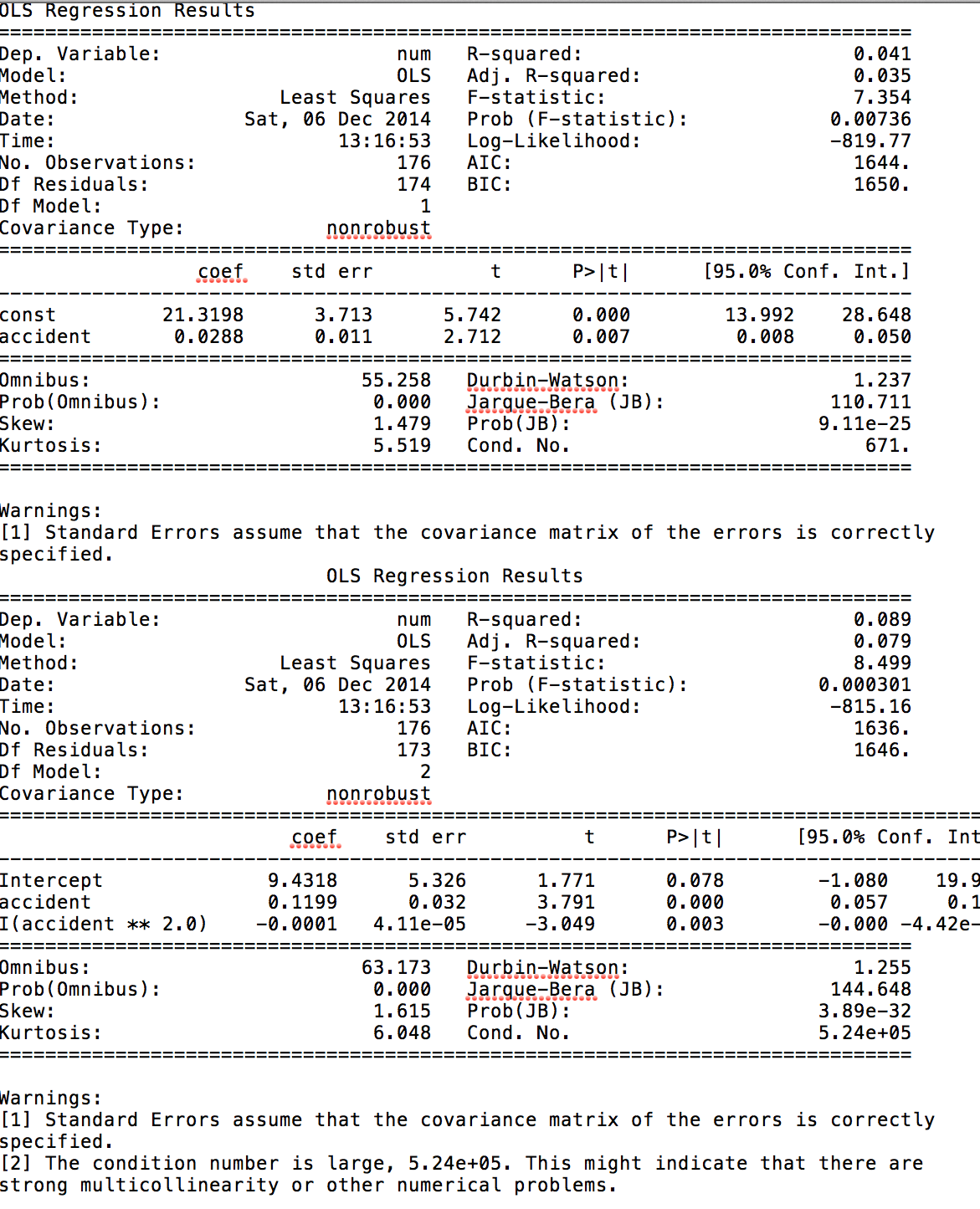


Figure 17 Statistics summary accident VS DPR complaints 2013

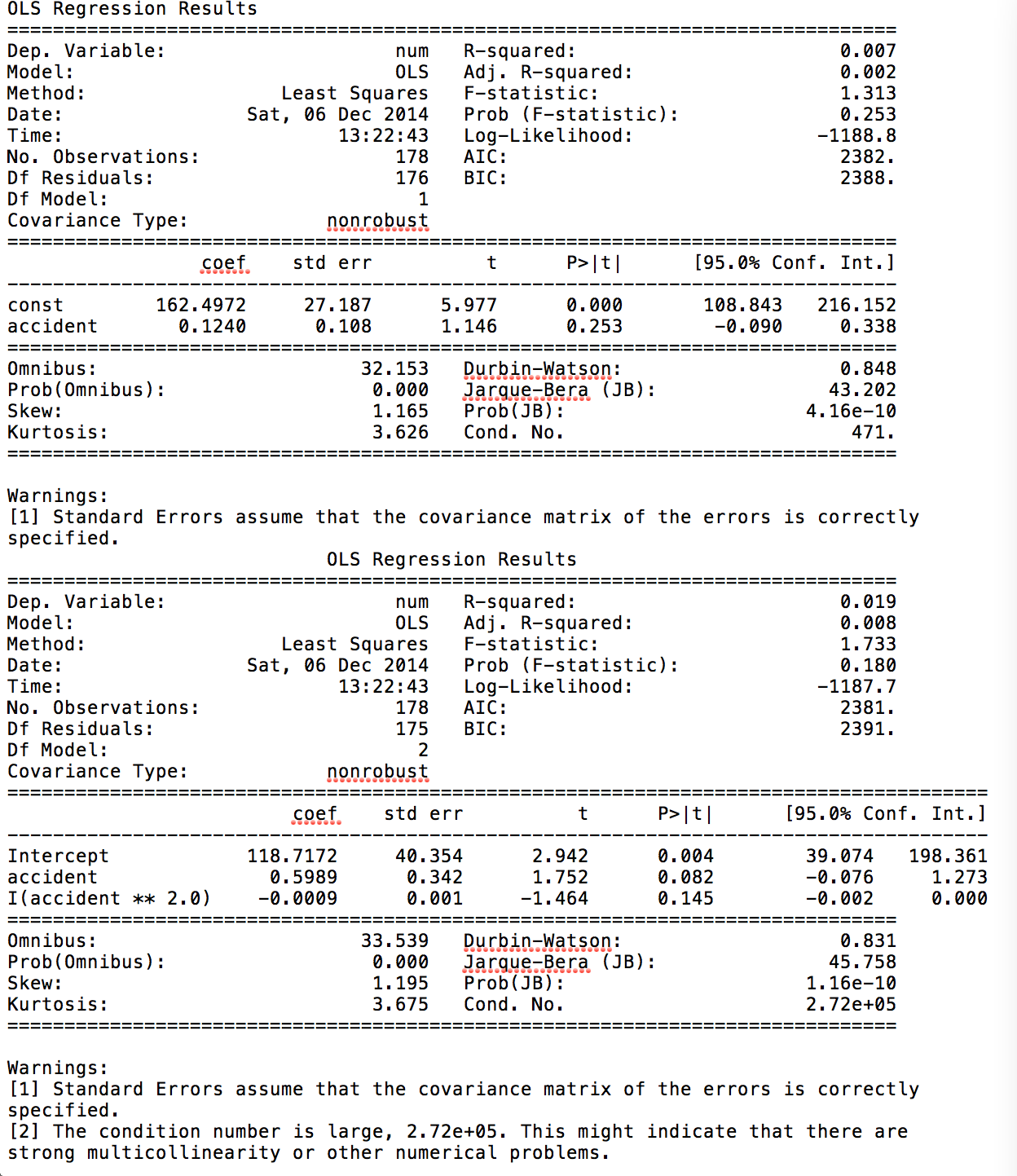


Figure 18 Statistics summary accident VS DPR complaints 2013

# References

1. Blake, Eric S., Kimberlain, Todd B., J. Berg, Robery, P. Cagnialosi, John, and L. Beven II, John, 2013. *“Tropical Cyclone Report: Hurricane Sandy,”* National Hurricane Center*, National Hurricane Center* February 12, 2013.
2. Bloomberg, Michael, 2013. *A Stronger, More Resislient New York.* NYC Special initiative for Rebuilding and Resiliency, 2013, pp 1-438.
3. Silive 2011, *October snow storm pounds Staten Island with 2 inches*. Last updated on November 03 2011.

<http://www.silive.com/news/index.ssf/2011/10/october_surprise_2_inches_of_s.html>.

1. Luís M. A. Bettencourt and Geoffrey­ B. West, “Bigger Cities Do More with Less”, U.S. Patent And Trademark Office (data on patents filed between 2000–2005 for U.S. metropolitan areas).