Poisson Regression

12.01.2021

Hannah Do

0. Introduction

Linear Regression		Continuous Values
Logistic Regression	***predicts***	Multiple Classes
Poisson Regression		Count / Rate at which an event occurs

1. Dataset - Count Data

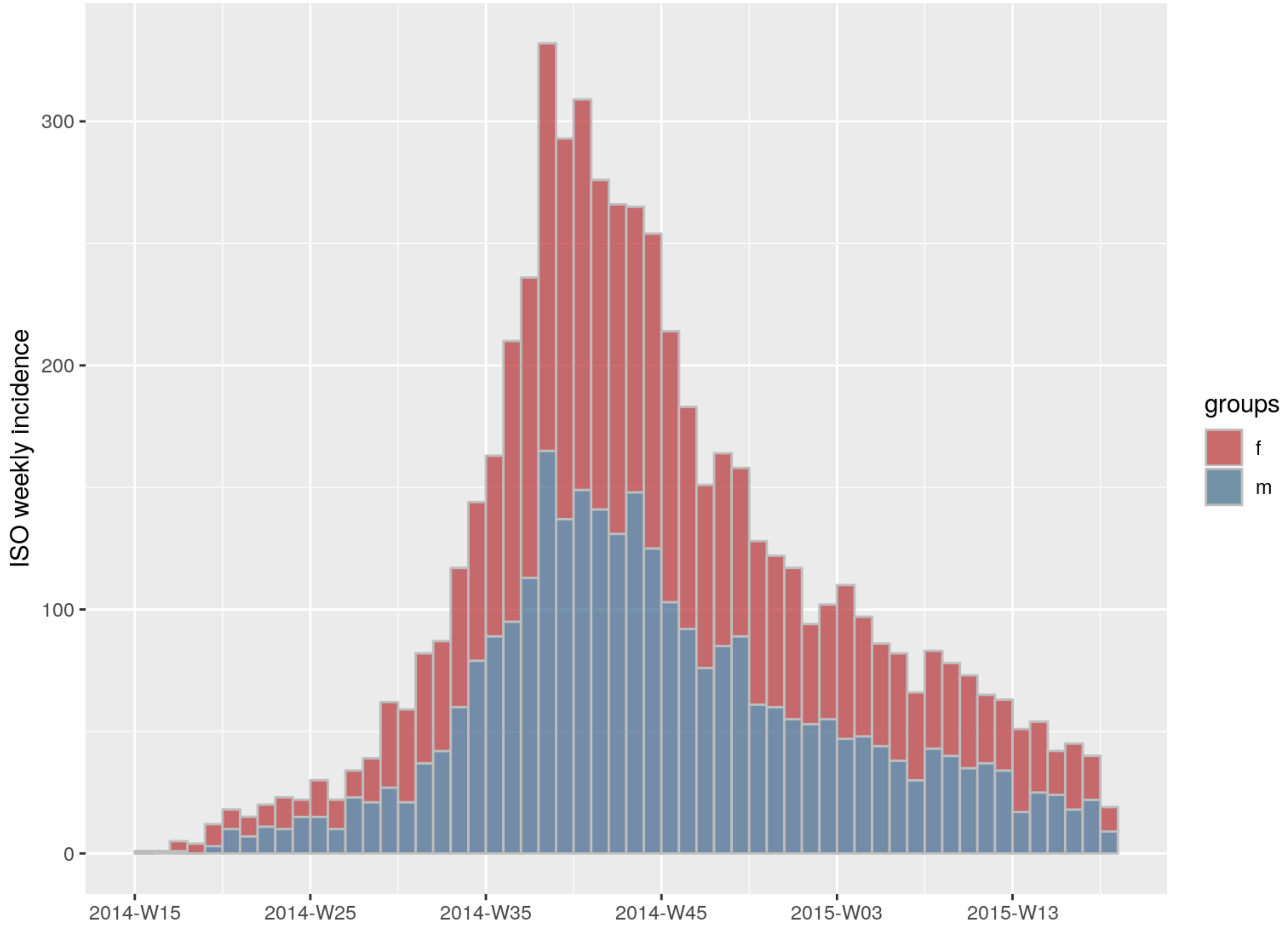
COUNT BASED DATA

Count based data contains events that occur at a certain rate.

Distribution of counts is discrete, not continuous, and limited to non-negative values.

Modeling counts / rates requires expected *number of events over a given time period* (Poisson distribution)

2. Dataset Example



Weekly epidemic incidence by gender (Ebola)
(https://www.repidemicsconsortium.org/incidence/)

Incidence Rate

= Number of events / person-time

Person-time can be

- Person-days
- person-week
- person-years

3. Distribution Function

Distribution of Poisson processes

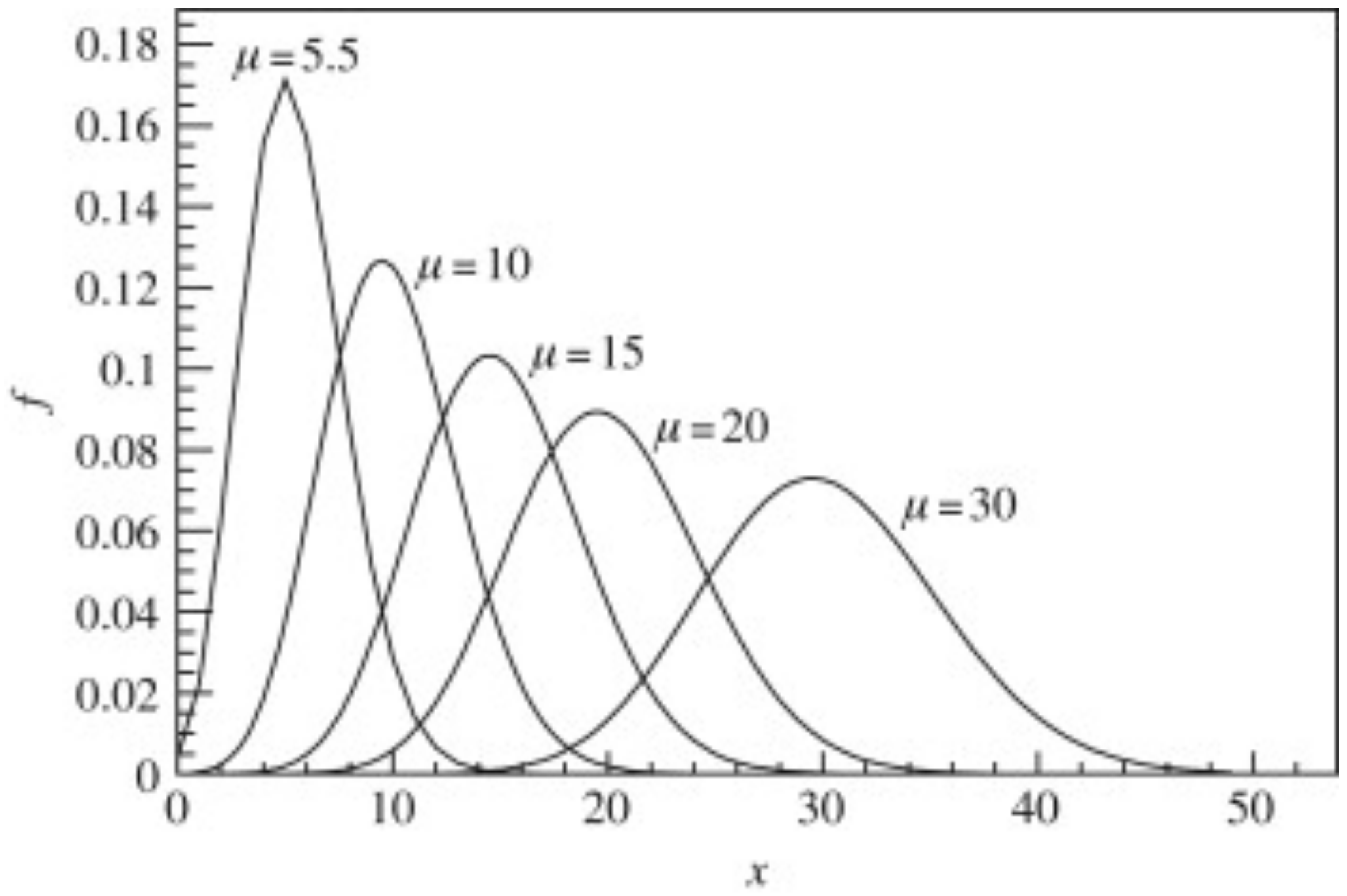
Model for a series of discrete event where the average time between events is known

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Probability Density Function of a Poisson distribution

- e : Euler's number (e = 2.71828...)
- x : Number of occurrences
- λ : Expected value of x when equal to its variance

PDF depicts probability functions in terms of continuous random variable values presenting in between a clear range of values.



Poisson probability density distribution for different values of λ. Marked as μ in the image above.

Essential statistics for data analysis
Syed Naeem Ahmed, in Physics and Engineering of Radiation Detection (Second Edition), 2015

4. Estimating the best mean

Maximum Likelihood Method

Poisson distribution is a discrete probability distribution, its likelihood function for a set of *n* measurements can be written as the following :

$$
\begin{aligned}
L(\mu) &= \prod_{i=1}^{n} f(x_i, \mu) \\
&= \prod_{i=1}^{n} \left[\frac{\mu^{x_i}\ \mathrm{e}^{-\mu}}{x_i!}\right] \\
&= \frac{\mu \sum x_i\ \mathrm{e}^{-n\mu}}{x_1!x_2!\ldots x_n!}.
\end{aligned}
$$

The log-likelihood function of $L(\mu)$ is

$$l \equiv \ln(L) = \left(\sum_{i=1}^{n} x_i\right)\ln(\mu) - n\mu - \ln(x_1!x_2!\ldots x_n!)$$

Following the maximum likelihood method ($\partial l/\partial\mu=0$), we get

(9.3.34)

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\left[\left(\sum_{i=1}^{n} x_i\right)\ln(\mu) - n\mu - \ln(x_1!x_2!\ldots x_n!)\right] &= 0 \\
\frac{1}{\mu^*}\sum_{i=1}^{n} x_i - n &= 0 \\
\mu^* &= \frac{1}{n}\sum_{i=1}^{n} x_i.
\end{aligned}
$$

which shows that the *simple mean* is the most probable value of a Poisson distributed variable.

Essential statistics for data analysis
Syed Naeem Ahmed, in Physics and Engineering of Radiation Detection (Second Edition), 2015

5. Estimating the error around the mean

Following equation also determines the error in the previous equation (μ) :

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \mu^2} &= -\frac{1}{\mu} \textstyle\sum_{i=1}^{n} x_i \\
\Delta\mu &= \left[-\frac{\partial^2 l}{\partial \mu^2}\right]^{-1/2} \\
&= \left[-\frac{\mu^{*2}}{\sum_{i=1}^{n} x_i}\right]^{-1/2} \\
&= \frac{1}{n}\left[\textstyle\sum_{i=1}^{n} x_i\right]^{-1/2}.
\end{aligned}
$$

The result implies that the statistical error we can expect is simply the *square root* of the measured quantity.

https://www.sciencedirect.com/topics/mathematics/poisson-distribution

6. Implementation of Poisson Regression

Predict Volume (count) of Dow Jones (S&P index)

```
df = pd.read_csv('dow_jones_index.data', header=0, \
     infer_datetime_format=True, parse_dates=[0], index_col=['date'])
```

	high	low	volume	percent_change_price
date				
1/7/2011	16.72	15.78	239655616	3.79267
1/14/2011	16.71	15.64	242963398	-4.42849
1/21/2011	16.38	15.60	138428495	-2.47066
1/28/2011	16.63	15.82	151379173	1.63831
2/4/2011	17.39	16.18	154387761	5.93325

https://www.sciencedirect.com/topics/mathematics/poisson-distribution

6. Implementation of Poisson Regression

```python
expr = """volume ~ DAY  + DAY_OF_WEEK + MONTH + high + low + percent_change_price"""

#Set up the X and y matrices
y_train, X_train = dmatrices(expr, df_train, return_type='dataframe')
y_test, X_test = dmatrices(expr, df_test, return_type='dataframe')

#Using the statsmodels GLM class, train the Poisson regression model on the training data set.
poisson_training_results = sm.GLM(y_train, X_train, family=sm.families.Poisson()).fit()

#Print the training summary.
print(poisson_training_results.summary())
```

Data conversion to Poisson distribution

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 volume   No. Observations:                  375
Model:                            GLM   Df Residuals:                      368
Model Family:                 Poisson   Df Model:                            6
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:            -1.0248e+10
Date:                Wed, 01 Dec 2021   Deviance:                   2.0496e+10
Time:                        06:20:40   Pearson chi2:                 2.97e+10
No. Iterations:                     8
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept               16.9084      0.000   3.86e+04      0.000      16.908      16.909
DAY                     -0.0070   5.93e-07  -1.17e+04      0.000      -0.007      -0.007
DAY_OF_WEEK              0.9134      0.000   8361.876      0.000       0.913       0.914
MONTH                   -0.1665    4.5e-06   -3.7e+04      0.000      -0.166      -0.166
high                     0.2534   6.47e-06   3.92e+04      0.000       0.253       0.253
low                     -0.3024   6.73e-06   -4.5e+04      0.000      -0.302      -0.302
percent_change_price    -0.0195   1.48e-06  -1.32e+04      0.000      -0.020      -0.020
========================================================================================
```

IRLS is **used to find the maximum likelihood estimates of a generalized linear model**, and in robust regression to find an M-estimator, as a way of mitigating the influence of outliers in an otherwise normally-distributed data set. For example, by minimizing the least absolute errors rather than the least square errors.

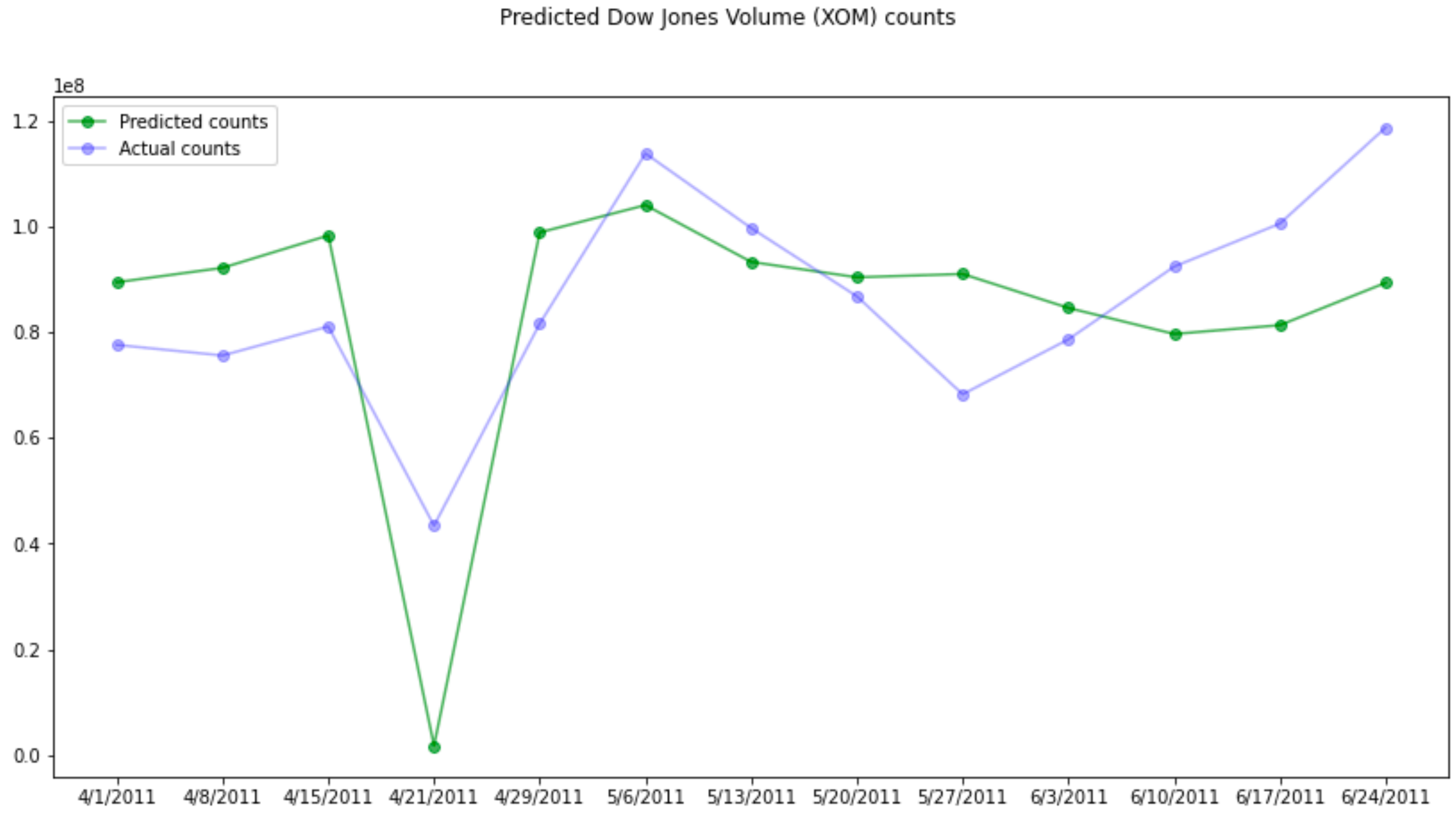
Code referenced from : https://timeseriesreasoning.com/contents/poisson-regression-model/

6. Implementation of Poisson Regression

Predictions with mean & confidence intervals

```
predictions_summary_frame.head(100)
```

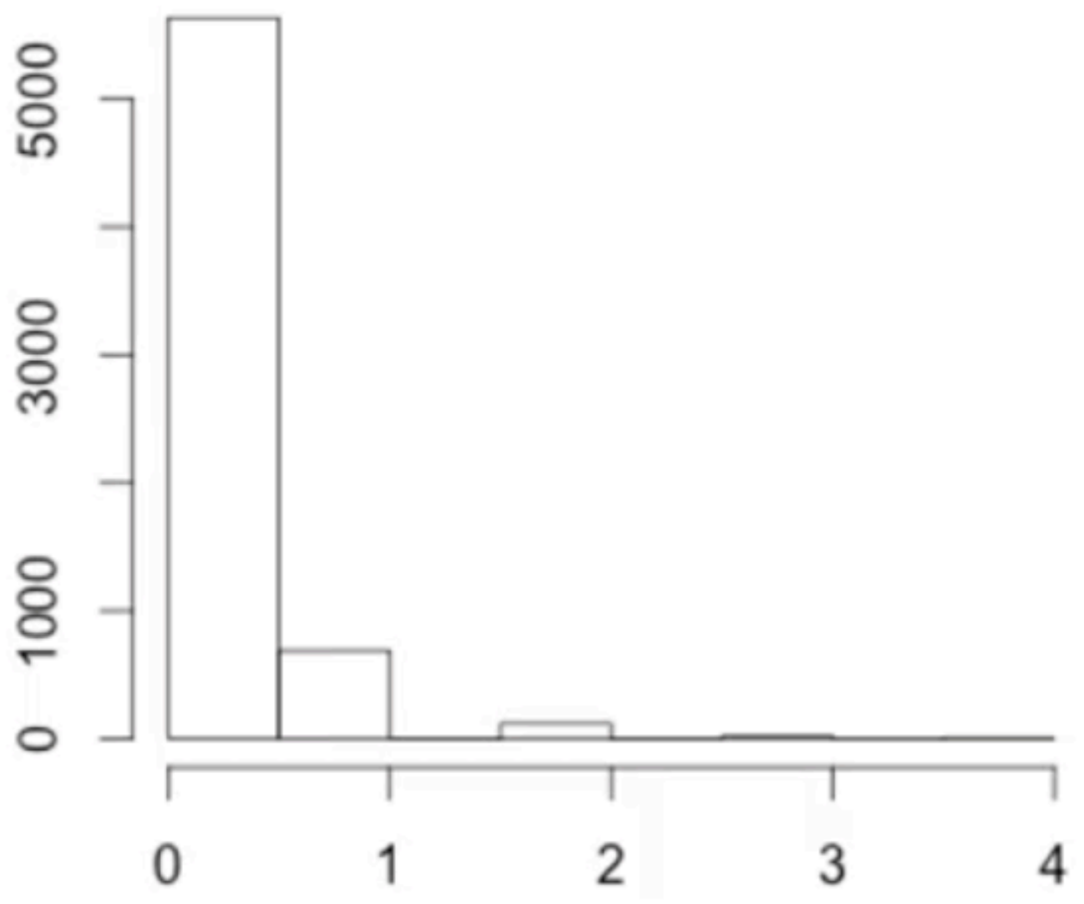
date	mean	mean_se	mean_ci_lower	mean_ci_upper
4/1/2011	8.943011e+07	11170.484046	8.940821e+07	8.945200e+07
4/8/2011	9.215631e+07	11024.941214	9.213470e+07	9.217792e+07
4/15/2011	9.827723e+07	8583.777363	9.826040e+07	9.829405e+07
4/21/2011	1.745294e+06	394.550190	1.744521e+06	1.746068e+06



$$Y = b_0 + b_1 x_1 + b_1 x_1 + \ldots + b_n x_n$$ Regular linear regression equation

$$\ln(Y) = b_0 + b_1 x_1 + b_1 x_1 + \ldots + b_n x_n$$ Poisson regression equation

7. Summary



Poisson Distribution

1. Closely portrays count or rate data
2. Skewed depending on lambda value
3. Mean is equal to Variance

Poisson Regression

1. Count/rate data must be converted to Poisson Distribution
2. Model returns Prediction of the Means and its Confidence Intervals

I. References

1. Poisson Regression Part I | Statistics for Applied Epidemiology | Tutorial 9 https://www.youtube.com/watch?v=0XfXHYDY0BA

2. Poisson distribution (ScienceDirect), Mathematical Modeling (Fourth edition), 2013. https://www.sciencedirect.com/topics/mathematics/poisson-distribution

3. Time Series Analysis, Regression and Forecasting https://timeseriesreasoning.com/contents/poisson-regression-model/