Count Data

Rahul Telang

Count data

- Counts are non-negative integers. They represent the number of occurrences of an event within a fixed period.
- Examples:
  - Number of “jumps” in stock returns per day.
  - Number of trades in a time interval.
  - Number of a given disaster
  - Number of crimes on campus per semester.
- The events also can be rare (number of earthquakes)

Using Count data

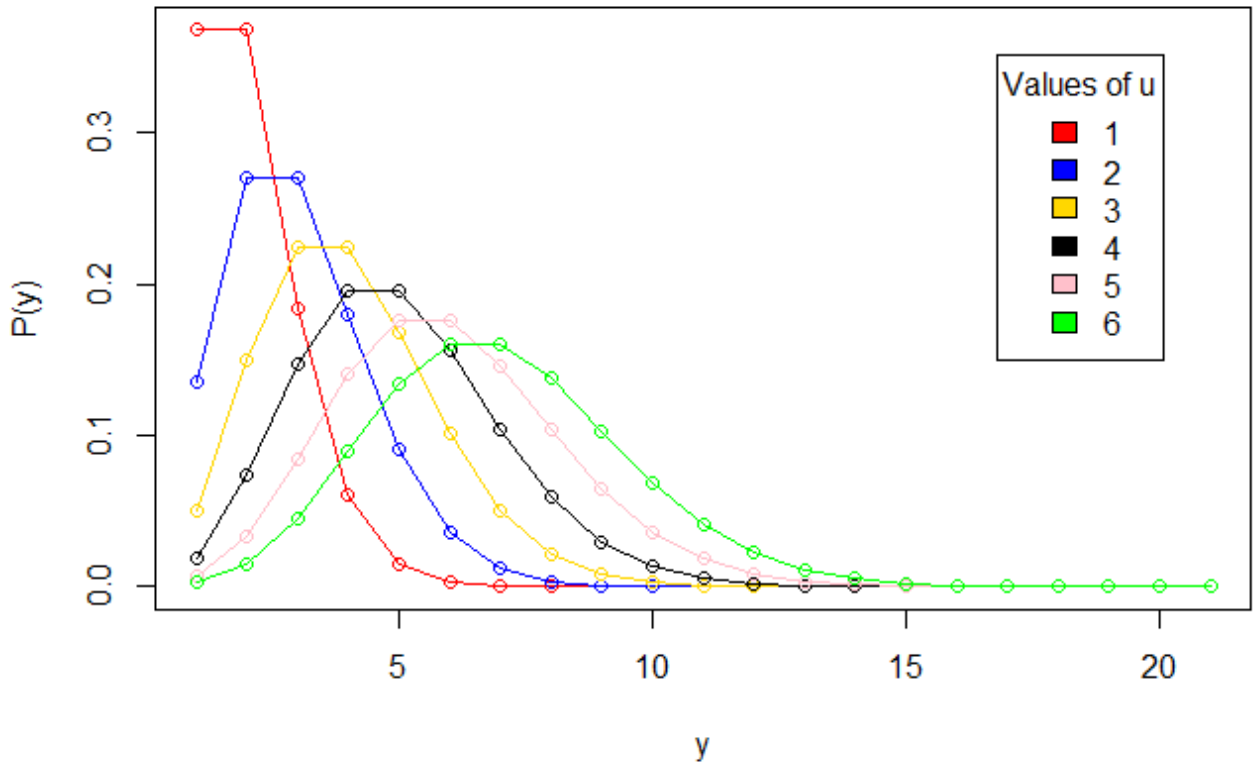
- When the data is “count”, applying a linear regression framework
  - $Y = X\beta + \varepsilon$

is inappropriate because Y is

  - discrete,
  - Y does not have negative outcomes (Y>=0)
- In short, assuming a normal distribution for ε is incorrect.
- What is an appropriate distribution for count data?

Poisson Distribution

- Poisson distribution is a starting point for count data. Poisson distribution has a pdf $P(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$



Poisson distribution

- This is a discrete distribution with
  - $P(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$ with mean E(y) = λ
- In short, the data generation process for count data is Poisson distribution.
- When one fits Poisson distribution, one estimates λ (the parameter of Poisson distribution).

$$\max_{\lambda} L\,ogL = \max_{\lambda} \sum_{n} Log(f(y_n/\lambda))$$

- A likelihood functions to estimate λ would involve maximizing
  - where f(y/λ) = P(y/λ) is pdf for Poisson distribution

Poisson regression

- Can Poisson be represented in GLM framework?
- Since Poisson falls under exponential family, GLM can be used. In this framework, if covariates affect the outcome, we want
  - $E[y_i] = \alpha + \beta x_i$
- where E(y) is represented a linear combination of covariates.
- In a GLM setting, since the natural parameter is linear combination - $\eta = x'\beta$. We need to connect $E(y) = \mu$ with η.

Poisson

- What could be the possible link function such that η = g(μ)?
- A log(.) function is logical because log(.) ensures that it converts range $(-\infty, \infty)$ to $(0, \infty)$ which is outcome we observe. Thus, any prediction will be bounded in this range.
- It can be shows that link function g(.) = log(μ) satisfies the requirement for GLM for Poisson.

- Since μ = E(y) = λ (the mean of Poisson). This also means that
  - log(λ)= $\alpha + \beta_1 x_1$ (because our link function is log()
  - $\lambda = e^{(\alpha+\beta_1 x_1)}$
  - $\lambda = e^{\propto} e^{\beta_1 x_1}$
  - $\lambda = \lambda_0 e^{\beta_1 x_1}$
- λ_0 is the baseline Poisson mean and covariates shift the mean. A unit increase in covariate changes the mean λ by e^{β} unit.
- When there are no covariates, we simply estimate α and hence λ_0.

When there are no covariates?

- In GLM, the goal is to estimate the impact of covariates on mean λ. But what if there are no covariates?
- One of the advantage of assuming a distribution for the data is that one can infer something useful and modify the model to fit the context.

Predicting Billboard effectiveness

As a marketing manager, you are interested in measuring the effectiveness of your ads. Suppose you are in charge of billboard advertising (you can pretty much extend it to any other medium) .

You pay money to get “monthly showing,” which comprises a specific set of billboards carrying the advertiser’s message in a given market.

How do you measure effectiveness of your campaign?

1. Reach (proportion of users who have seen the ad atleast once)
2. frequency (average number of exposure for those who have seen it),
3. gross rating points (GRPs) – (average number of exposers per 100 users in the market).

How do you collect this data?

- Respondents record their daily travel on maps. An “exposure” is deemed to occur each time the respondent travels by a billboard in the showing, on the street or road closest to that billboard, going towards the billboard’s face.

Measurement

Your problem is to know whether the benefits are worth the cost?

You recruit a sample of 250 users who are willing to do this for a **week** and your job is to “project” this into **monthly** numbers.

i.e. after a week you can calculate the metrics.

And if you were to run this campaign for a month, how would your metrics change?

Another way to think is, you are doing a pilot for a week and want to decide whether you want the campaign to run for the next three weeks or not?

Distribution of Billboard Exposures (1 week)

# Exposures	# People
0	48
1	37
2	30
3	24
4	20
5	16
6	13
7	11
8	9
9	7
10	6
11	5

# Exposures	# People
12	5
13	3
14	3
15	2
16	2
17	2
18	1
19	1
20	2
21	1
22	1
23	1

Data

- To be able to make predictions we have understand the distribution of data
- This is the count data. Hence a Poisson distribution can explain the data
  - $P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$
- To estimate parameters, we write a GLM equation where E(y) = λ and λ is parameter of Poisson distribution. With a logarithm link functions,

  $\log(\lambda) = \alpha + \beta_1 x_1$, or

  $\lambda = \lambda_0 e^{\beta_1 x_1}$ where $\lambda_0 = e^{\alpha}$
- In GLM, we estimate outcome with covariates. What are covariates here?
- Since there are no covariates, we are simply estimating λ_0 which is the baseline mean when covariates are 0.

Code

- We define regression equation
  - Expr = ‘exposures ~ 1’

  We are running a regression with a dummy.
- We have to create exogenous and endogenous variable.
  - y, X = dmatrices(expr, df, return_type='dataframe')

- We then use GLM
- See https://www.statsmodels.org/stable/generated/statsmodels.genmod.generalized_linear_model.GLM.html

  - poisson_training_results = sm.GLM(y, X, family = sm.families.Poisson()).fit()

Results

- We estimate $\lambda_0 = e^{1.49}$
- How to make predictions?
- Since we have a dummy as covariate, the model predict values as $e^{1.49}$=4.45.
- It will make the predictions that 4.45 is the mean exposure.

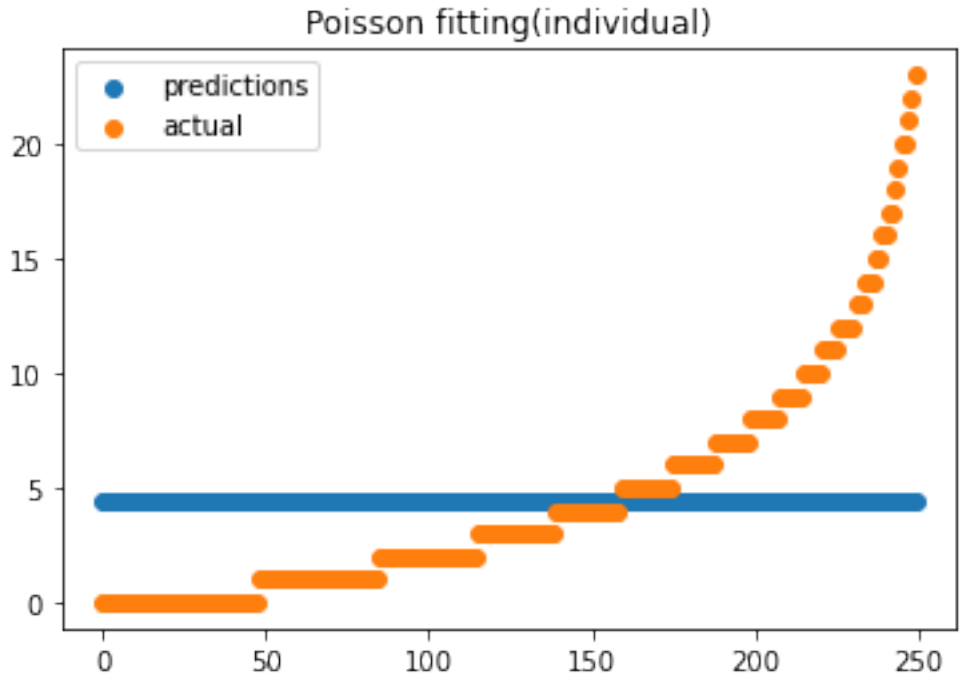
Generalized Linear Model Regression Results

Dep. Variable:	exposures	No. Observations:	250
Model:	GLM	Df Residuals:	249
Model Family:	Poisson	Df Model:	0
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-929.04
Date:	Sun, 16 Apr 2023	Deviance:	1203.3
Time:	14:18:00	Pearson chi2:	1.30e+03
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.4943	0.030	49.873	0.000	1.436	1.553

Data and predictions

- GLM makes predictions for outcomes on for a user given covariates. Since there are no covariates, prediction for all users is given by mean λ .
- With this aggregate data, it is hard to know if Poisson distribution is a good fit.

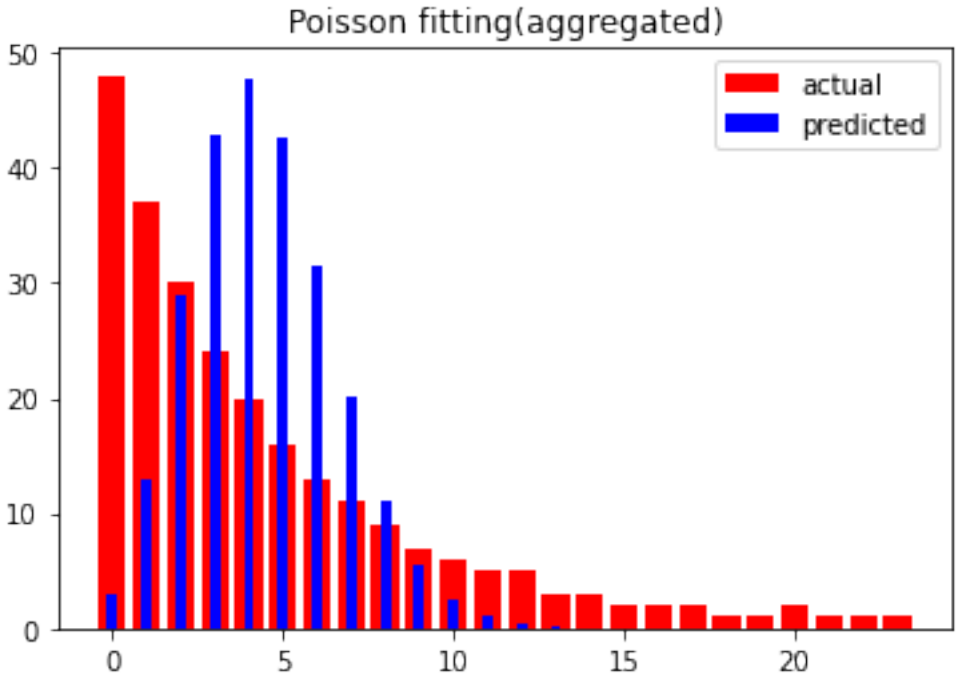


Poisson Prediction

- Recall Poisson predicts probability of exposures
  - P(y=1,..,n) = $\frac{e^{-\lambda}\lambda^{y}}{y!}$
- The mean rate can outline the probability of each discrete exposure (y = 1,...,n) .
- We can compare that predicted exposure with the actual data.

Predictions

- We calculate P(x=0,1…).
- Our data shows that 48 users had 0 exposure. We can readily calculate P(x=0). Since we have 250 users in the sample, number of people who have 0 exposure is simply 250*p(x=0). This will be our predicted value.
- We use Poisson pdf to calculate the probabilities.
  - stats.poisson.pmf(x_range, lmbda)

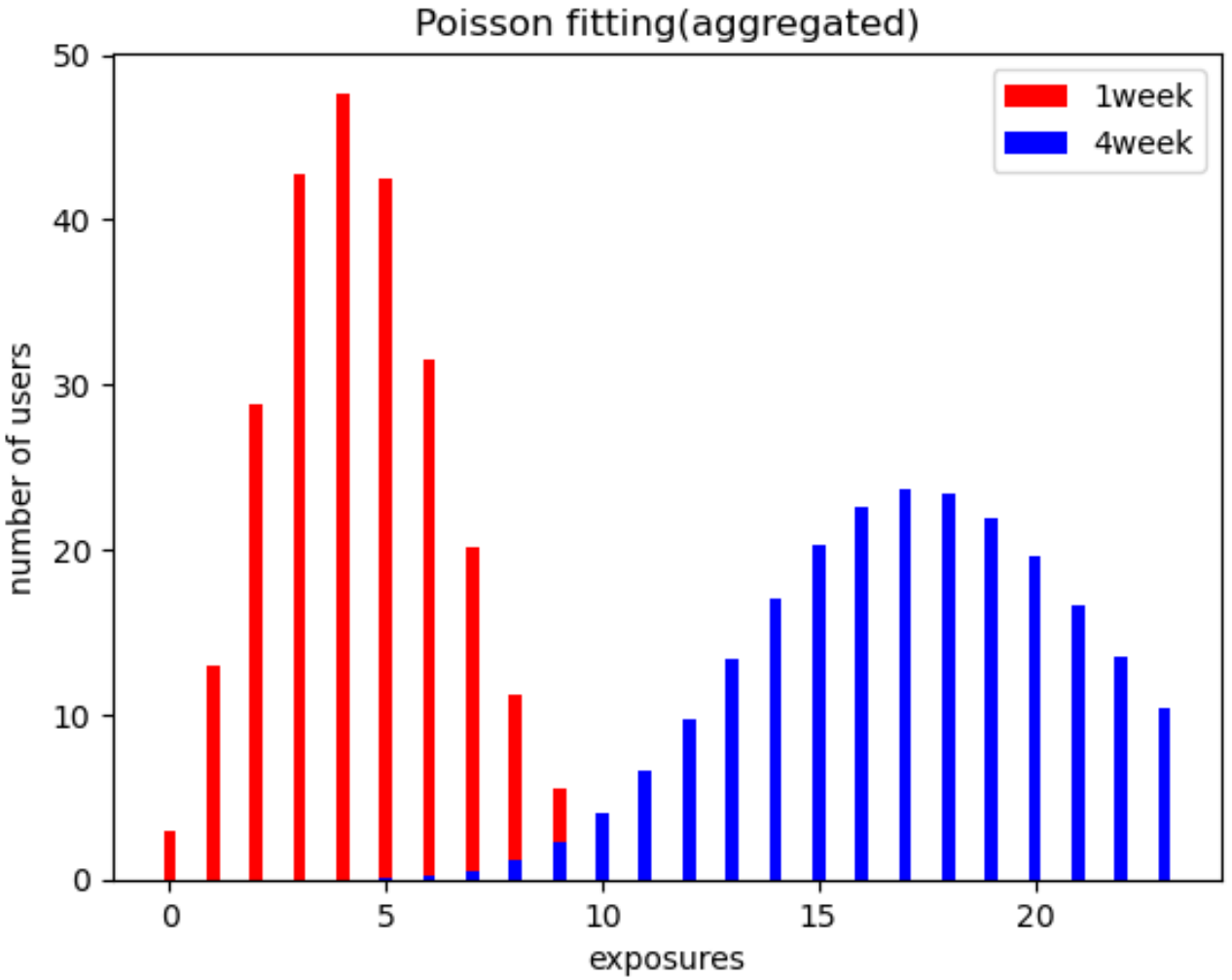


Predictions

- Given estimate λ = 4.45, we can predict how many users will have 0,1..,n exposures.
- Just by estimating the mean, we see that the predictions are not great. We will explore ways to make this a better fit.
- We still have not answered the original question.
  - How do we project weekly data into monthly projections?
- For Poisson distribution, if λ is the mean exposure rate in one unit, then λt is the mean rate in “t” units of time.
- In short, in 4 weeks of time, we can calibrate the model with 4λ mean rate.

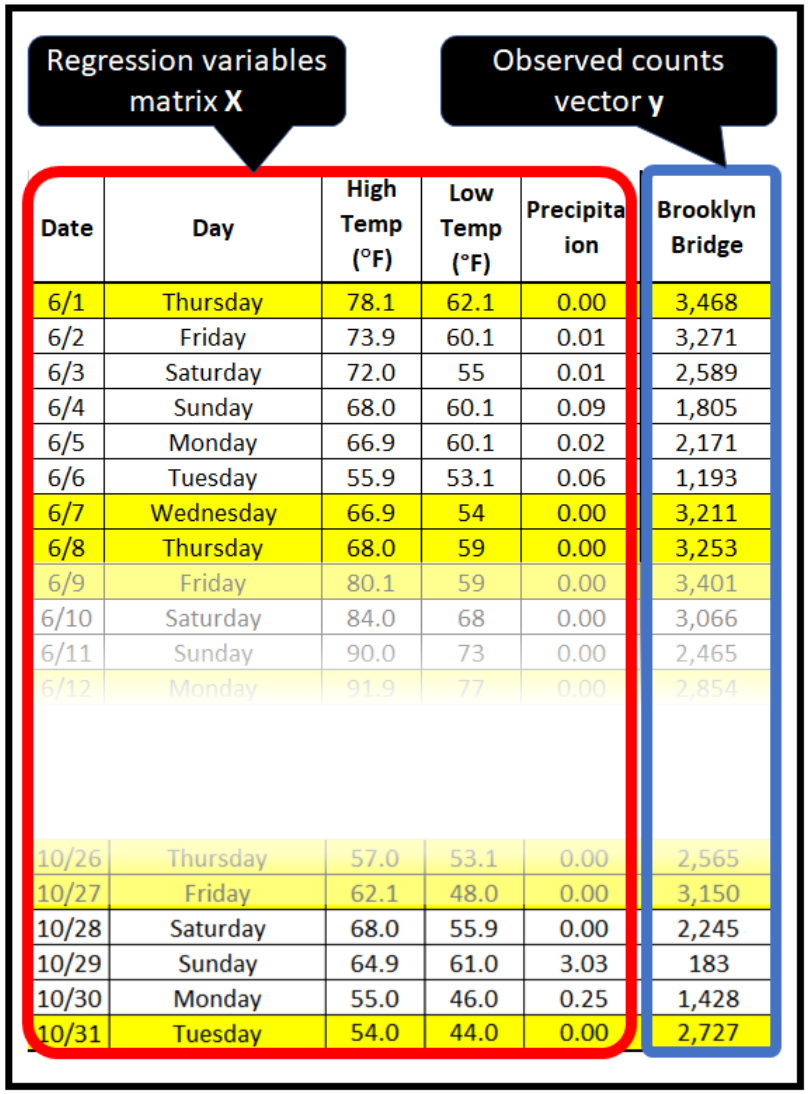
Predictions in 4 weeks

- Number of users who have higher number of exposure increase significantly when the campaign is run for 4 weeks.
- One can potentially do cost benefit analysis to about the campaign after 1 week of data.



Example 2 – Predicting number of Bicyclists

- We have daily data on number of bicyclists on Brooklyn bridge
- Our goal is to predict number of bicyclist counts for a given day.
- Since it is the count data, we again use Poisson regression to model the number of bicyclists.
- We have data on the date, day, temperature and precipitation and we want to predict how many bicyclists pass through Brooklyn bridge on a given day.



Date	Day	High Temp (°F)	Low Temp (°F)	Precipitation	Brooklyn Bridge
6/1	Thursday	78.1	62.1	0.00	3,468
6/2	Friday	73.9	60.1	0.01	3,271
6/3	Saturday	72.0	55	0.01	2,589
6/4	Sunday	68.0	60.1	0.09	1,805
6/5	Monday	66.9	60.1	0.02	2,171
6/6	Tuesday	55.9	53.1	0.06	1,193
6/7	Wednesday	66.9	54	0.00	3,211
6/8	Thursday	68.0	59	0.00	3,253
6/9	Friday	80.1	59	0.00	3,401
6/10	Saturday	84.0	68	0.00	3,066
6/11	Sunday	90.0	73	0.00	2,465
6/12	Monday	91.9	77	0.00	2,854
10/26	Thursday	57.0	53.1	0.00	2,565
10/27	Friday	62.1	48.0	0.00	3,150
10/28	Saturday	68.0	55.9	0.00	2,245
10/29	Sunday	64.9	61.0	3.03	183
10/30	Monday	55.0	46.0	0.25	1,428
10/31	Tuesday	54.0	44.0	0.00	2,727

Code

- We convert the date and day data into numeric numbers to use in regression.
- Series.dt can return many useful attributes

```
ds = df.index.to_series()
df['MONTH'] = ds.dt.month
df['DAY_OF_WEEK'] = ds.dt.dayofweek
df['DAY'] = ds.dt.day
```

- We split the data in test and training samples so we calibrate and then test our model.

Poisson regression

- We create the regression we want to estimate
- expr = "BB_COUNT ~ DAY + DAY_OF_WEEK + MONTH + HIGH_T + LOW_T + PRECIP“
- (Covariates impact the count of bicyclists)
- We use dmatrices to generate X and Y variables

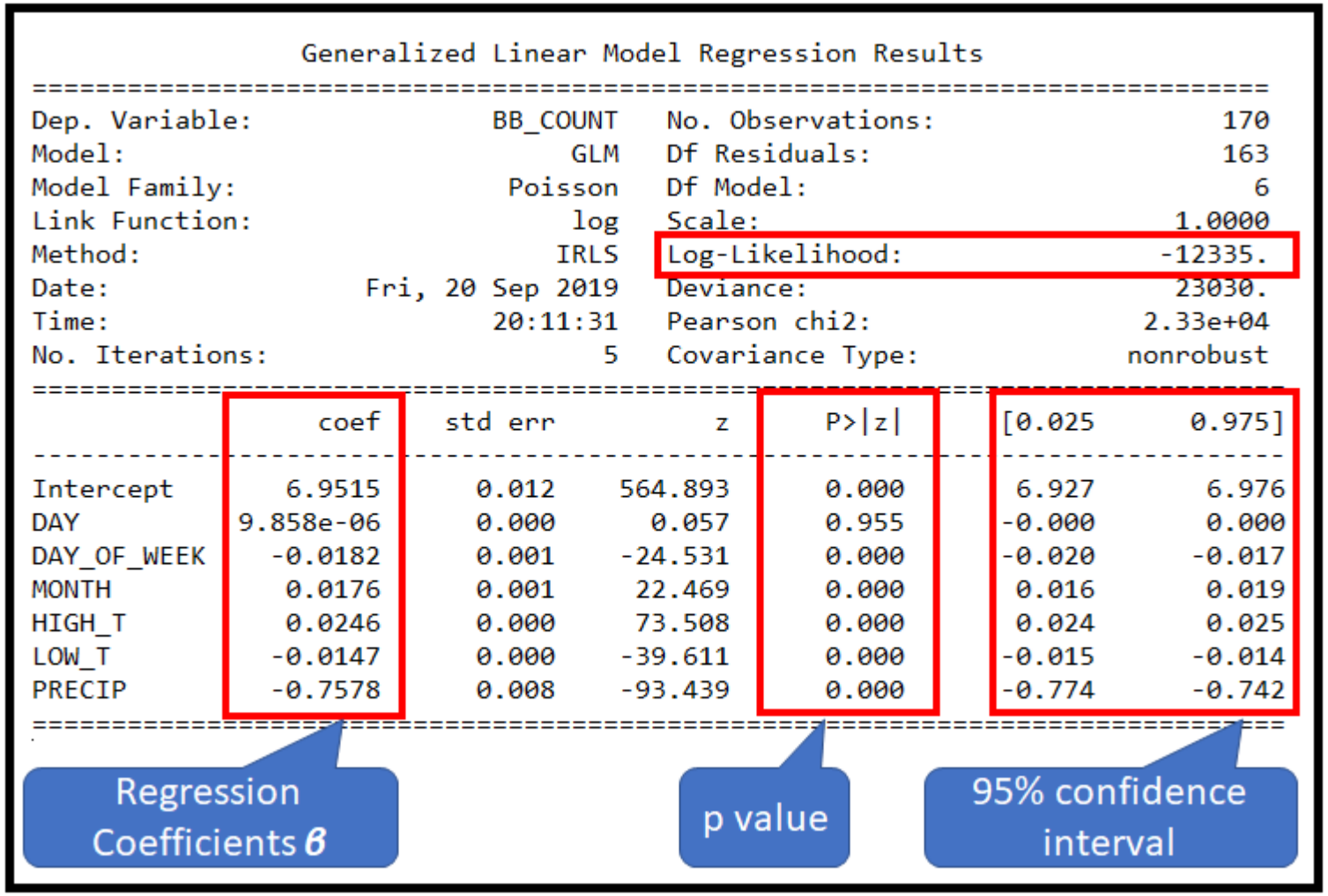
```
y_train, X_train = dmatrices(expr, df_train, return_type='dataframe')
y_test, X_test = dmatrices(expr, df_test, return_type='dataframe’)
```

- We use GLM to estimate Poisson model

```
poisson_training_results = sm.GLM(y_train, X_train, family=sm.families.Poisson()).fit()
```

Results

- As month increases by one unit, the mean of count increases by exp(0.0176) = 1.19 = (1.19-1)= 19%.
- Precipitation decreases count by exp(-0.75) = 0.47 = (1-0.47)=53%.
- The key to note is that the mean gets shifted heavily by covariates.



Generalized Linear Model Regression Results

Dep. Variable:	BB_COUNT	No. Observations:	170
Model:	GLM	Df Residuals:	163
Model Family:	Poisson	Df Model:	6
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-12335.
Date:	Fri, 20 Sep 2019	Deviance:	23030.
Time:	20:11:31	Pearson chi2:	2.33e+04
No. Iterations:	5	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	6.9515	0.012	564.893	0.000	6.927	6.976
DAY	9.858e-06	0.000	0.057	0.955	-0.000	0.000
DAY_OF_WEEK	-0.0182	0.001	-24.531	0.000	-0.020	-0.017
MONTH	0.0176	0.001	22.469	0.000	0.016	0.019
HIGH_T	0.0246	0.000	73.508	0.000	0.024	0.025
LOW_T	-0.0147	0.000	-39.611	0.000	-0.015	-0.014
PRECIP	-0.7578	0.008	-93.439	0.000	-0.774	-0.742

predictions

- Now we use these estimates to make predictions for test data.

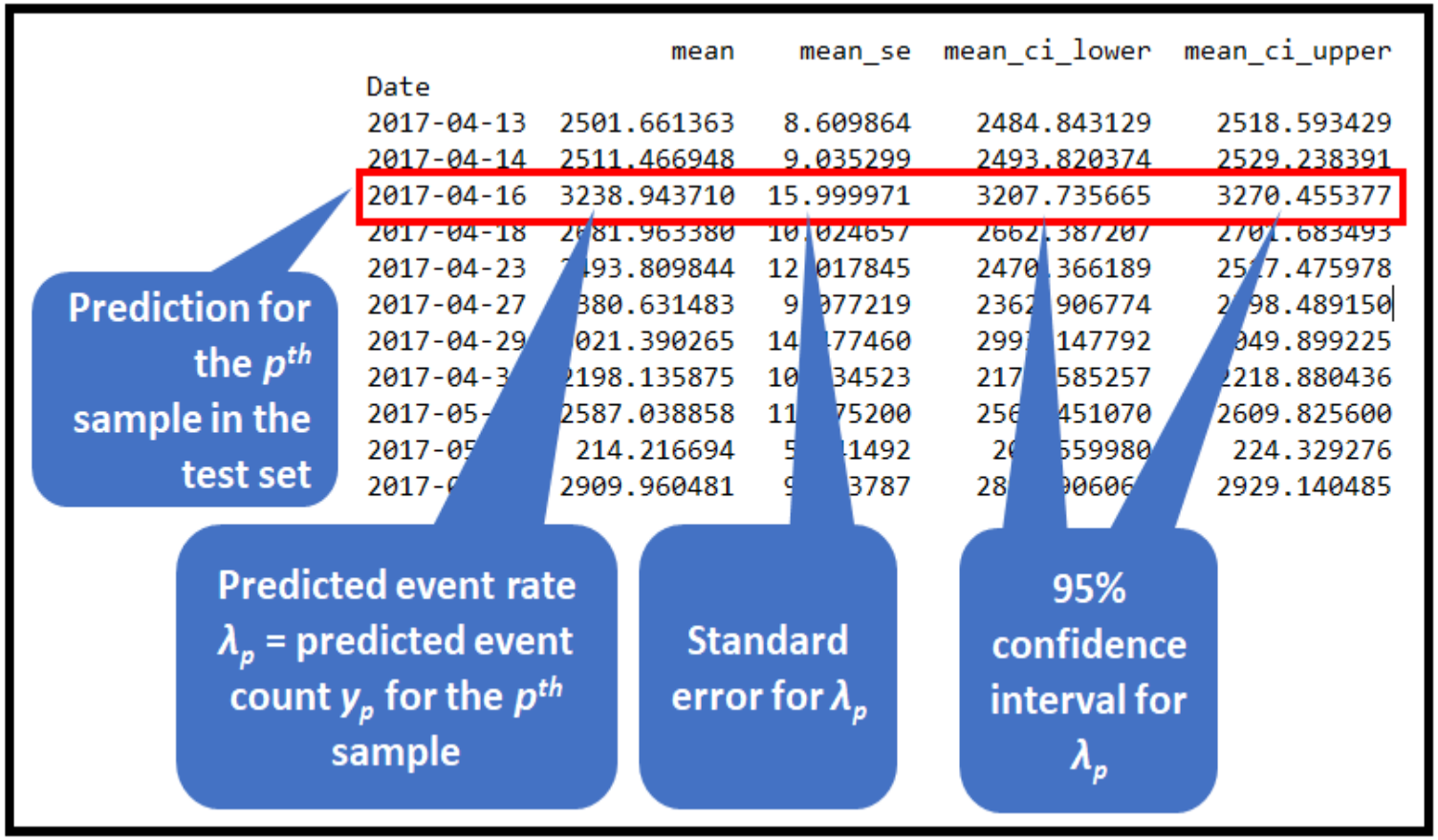
```
poisson_predictions =
poisson_training_results.get_predic
tion(X_test)

predictions_summary_frame =
poisson_predictions.summary_fram
e()

print(predictions_summary_frame)
```

You can refer to

https://www.statsmodels.org/dev/generated/statsmodels.discrete.discrete_model.Poisson.html



Observed vs predicted

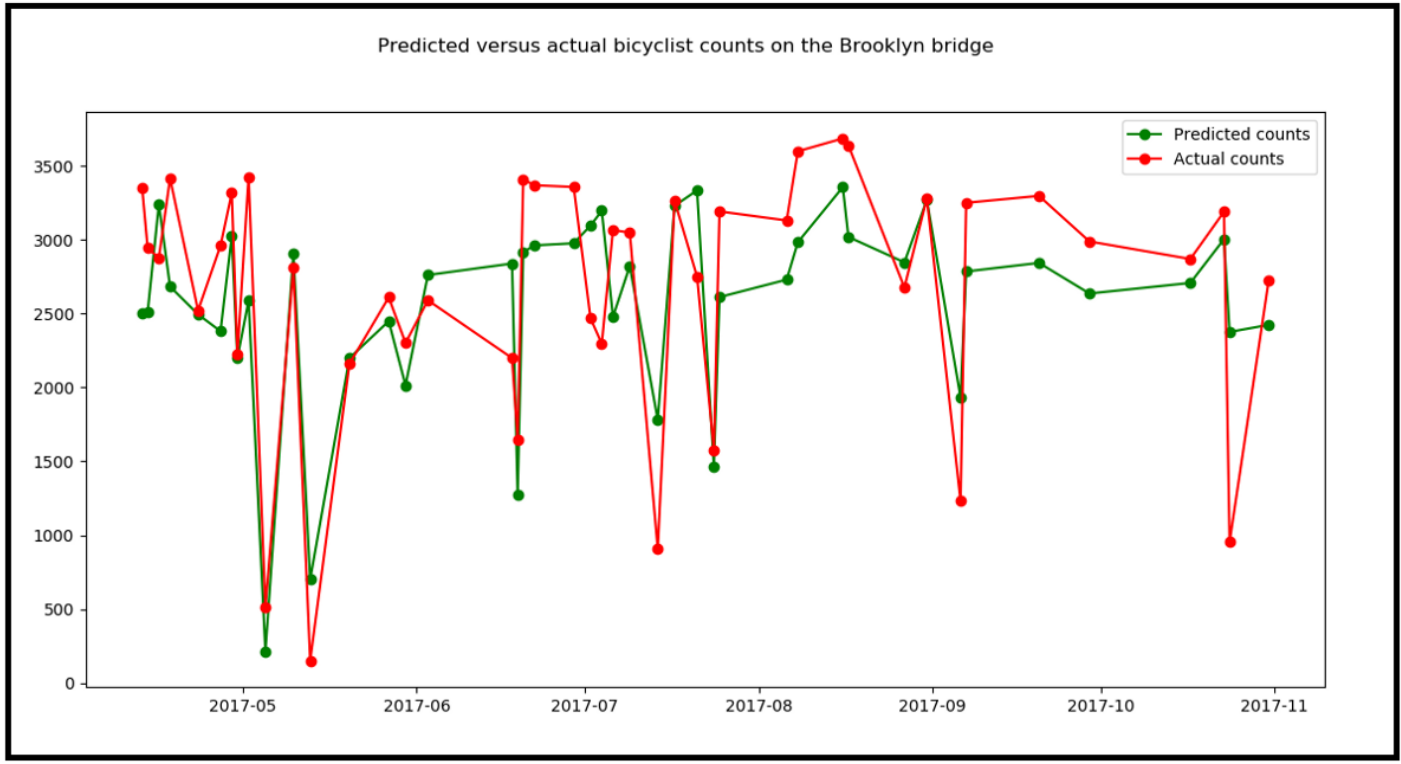
- We can plot the predicted counts versus the actual counts for the test data.

```
predicted_counts=predictions_summary_frame['mean’]

actual_counts = y_test['BB_COUNT’]


predicted, = plt.plot(X_test.index, predicted_counts, 'go-',
label='Predicted counts’)

actual, = plt.plot(X_test.index, actual_counts, 'ro-',
label='Actual counts’)
```



Limitation of Poisson

- One of the limitation of Poisson is that the mean (λ) is same as variance. In short mean = variance.
- In reality, this may not always hold.
- Deviance number suggests the fit. Recall, lower deviance is better.

Generalized Linear Model Regression Results

Dep. Variable:	BB_COUNT	No. Observations:	170
Model:	GLM	Df Residuals:	163
Model Family:	Poisson	Df Model:	6
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-12335.
Date:	Fri, 20 Sep 2019	Deviance:	23030.
Time:	20:11:31	Pearson chi2:	2.33e+04
No. Iterations:	5	Covariance Type:	nonrobust

Fit of Goodness

- Deviance can be thought of as how well the predictions do compared to actual data.
- The Chi square value (from the table) for 163 degree of freedom at 5% level is 193.79. The current results show that deviance is 23300 which is on the higher side.
- The assumption of mean=variance remains a challenge is using Poisson.
- We will come back to relaxing this assumption.

Summary

- For any count data, Poisson is the starting distribution. Since Poisson fits in the exponential family, GLM can be used.
- GLM estimates the impact of covariates and parameters of distribution.
- Understanding the distribution of data is important for making predictions.