ML and MAP estimation of Poisson distribution

Question:

Poisson distribution has been introduced to model deaths of soldiers in Prussian army from 1875

Table 1. Bortkewitsch's data table, giving numbers of deaths from horse-kicks

	Year																				
Corps	1875	1876	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886	1887	1888	1889	1890	1891	1892	1893	1894	Totals
G	0	2	2	1	0	0	1	1	0	3	0	2	1	0	0	1	0	1	0	1	16
I	0	0	0	2	0	3	0	2	0	0	0	1	1	1	0	2	0	3	1	0	16
II	0	0	0	2	0	2	0	0	1	1	0	0	2	1	1	0	0	2	0	0	12
III	0	0	0	1	1	1	2	0	2	0	0	0	1	0	1	2	1	0	0	0	12
IV	0	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	8
V	0	0	0	0	2	1	0	0	1	0	0	1	0	1	1	1	1	1	1	0	11
VI	0	0	1	0	2	0	0	1	2	0	1	1	3	1	1	1	0	3	0	0	17
VII	1	0	1	0	0	0	1	0	1	1	0	0	2	0	0	2	1	0	2	0	12
VIII	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	1	0	1	7
IX	0	0	0	0	0	2	1	1	1	0	2	1	1	0	1	2	0	1	0	0	13
X	0	0	1	1	0	1	0	2	0	2	0	0	0	0	2	1	3	0	1	1	15
XI	0	0	0	0	2	4	0	1	3	0	1	1	1	1	2	1	3	1	3	1	25
XIV	1	1	2	1	1	3	0	4	0	1	0	3	2	1	0	2	1	1	0	0	24
XV	0	1	0	0	0	0	0	1	0	1	1	0	0	0	2	2	0	0	0	0	8
Total	3	5	7	9	10	18	6	14	11	9	5	11	15	6	11	17	12	15	8	4	196

to 1894 across corps. Please find the data below.

Model the horse kick deaths using the Poisson distribution with different parameters for each of the corps. Learn Poisson distribution parameters for each of the corps using first 13 years of data and make predictions on remaining 7 years and compute the RMSE of predictions for each of the corps.

1. Use maximum likelihood estimation to learn the parameters.

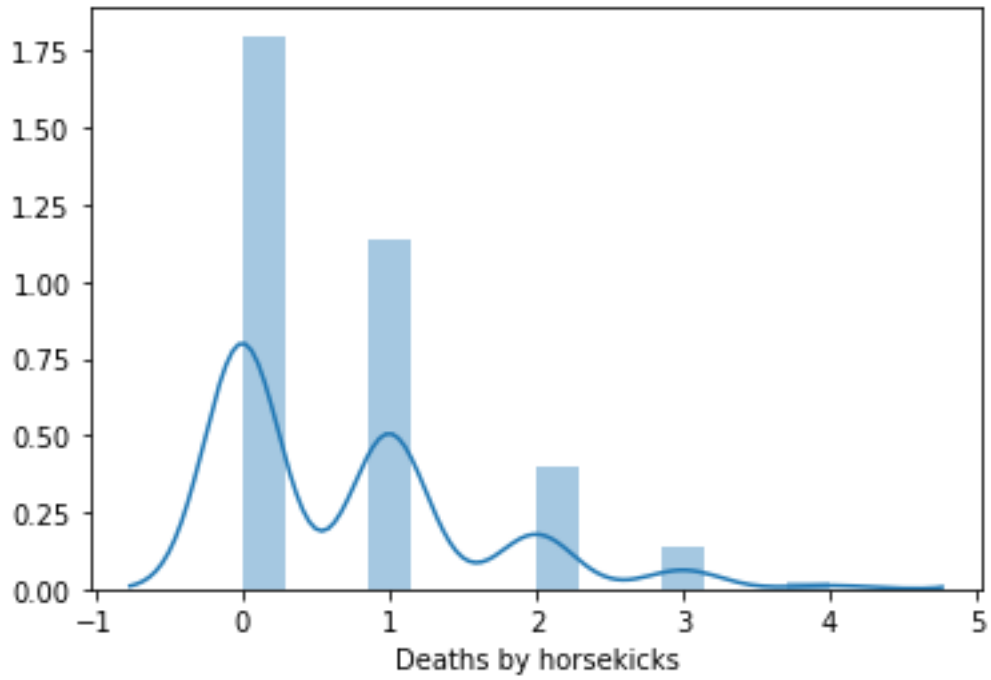
2. Use maximum aposteriori estimation to learn the parameters

a. Assume appropriate prior distribution over parameters and justify your assumption

b. Plot prior, likelihood and posterior and provide your observations in terms of mode of the distributions for corps 2, 4 and 6.

Solution :

By observation, the deaths count follow Poisson distribution. The below plot shows the density of total deaths due to horse kicks.



ML Estimate:

ML is a point estimate. It returns only one specific value considering the whole data. Since the deaths from horse kicks is random. The ML Estimate of Poisson is given by the mean of the observations.

$$\hat{\lambda}_{ML} = \frac{\sum_i^N y_i}{N}$$

The above equation to calculate the point estimate for all the corps and the results are summarized in the form of a table. The following formula is used to calculate RMSE.

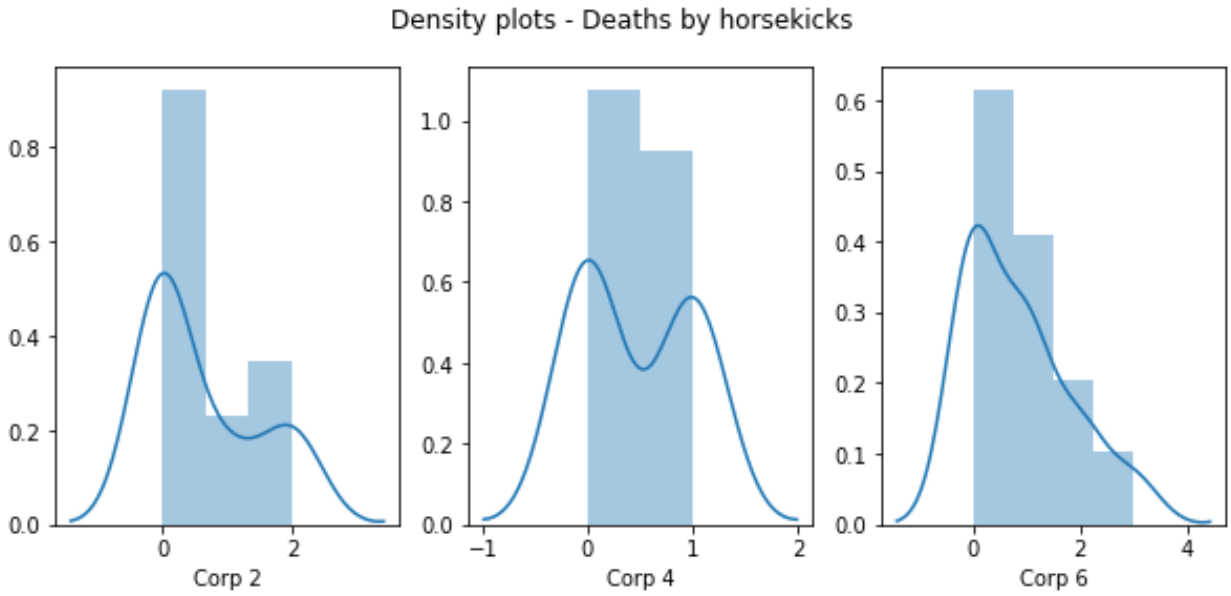
$$RMSE = \left[\frac{\sum_i^N (y_{pred} - y_i)^2}{N}\right]^{1/2}$$

Corps	ML Estimate (MLE)	RMSE	MLE (Rounded)
G	1.0	0.5547	1.0
I	0.6923	0.7845	1.0
II	0.6154	0.6202	1.0
III	0.6154	0.6202	1.0
IV	0.4615	0.3922	0
V	0.3846	0.6794	0
VI	0.8462	0.7338	1.0
VII	0.5385	0.6794	1.0
VIII	0.3077	0.4804	0
IX	0.6923	0.6202	1.0
X	0.5384	0.7338	1.0
XI	1.0	0.8321	1.0
XIV	1.4615	0.5547	1.0
XV	0.3077	0.7844	0

Table: MLE and RMSE calculations

Prior distribution:

Since the number of deaths are always positive, occurring randomly and following an exponential decrease, the gamma distribution is an optimum choice for prior.

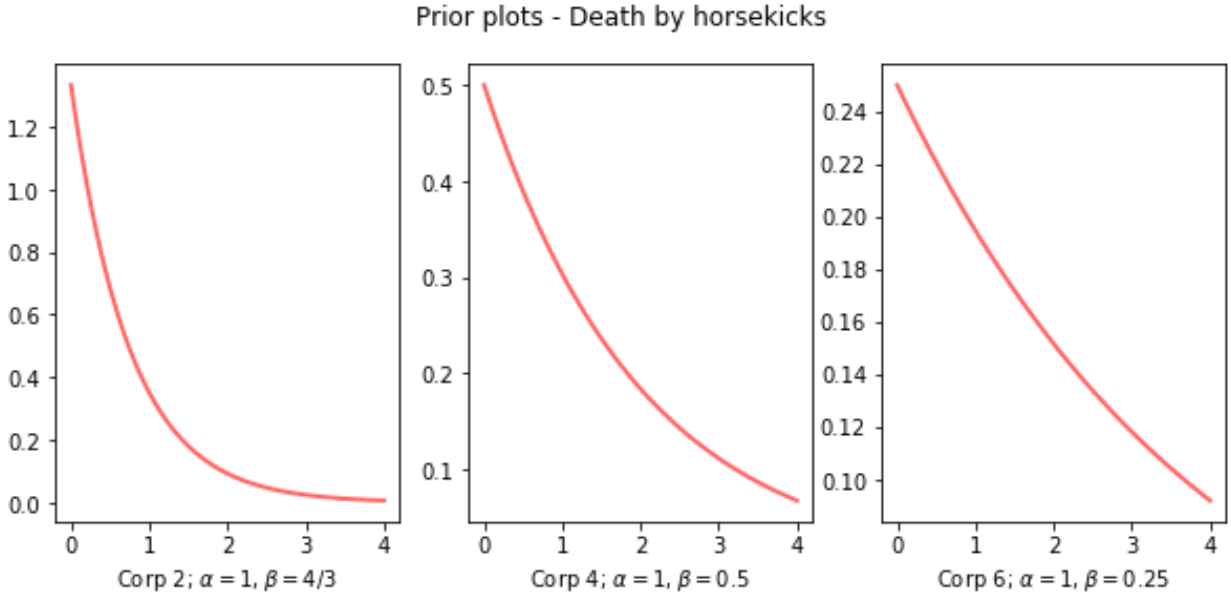


The figure represents density of horse kick deaths for corps **II, IV** and **VI.** Alpha and Beta values are assumed as necessary for each corp.

Corp	Alpha	Beta
II	1	4/3
IV	1	0.5
VI	1	0.25

Table Gamma prior parameters

The below plots are the Gamma prior plots for corps **II, IV** and **VI** with the above parameters.



The mode of the Gamma distribution is,

$$\boldsymbol{Mode}_{\gamma} = \frac{\alpha - \mathbf{1}}{\beta}$$

Since the highest concentration of deaths for all the corps is 0. The mode of the gamma prior will also be zero. Therefore, it is a suitable assumption.

MAP Estimate:

The Posteriori likelihood is calculated on the basis of bayes theorem.

The likelihood and posterior plots are plotted as below.

SOLUTION 3

we have assumed Gamma distribution as prior distribution

Now MAP,

$$Prob(\theta|x) = \frac{prob(x|\theta)\, prob(\theta)}{Prob(x)}$$

Now,

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \prod_{x_i \in X} prob(x_i|\theta)\, prob\,\theta$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left(\sum_{x_i \in X} \log prob(x_i|\theta) + \log prob(\theta) \right)$$

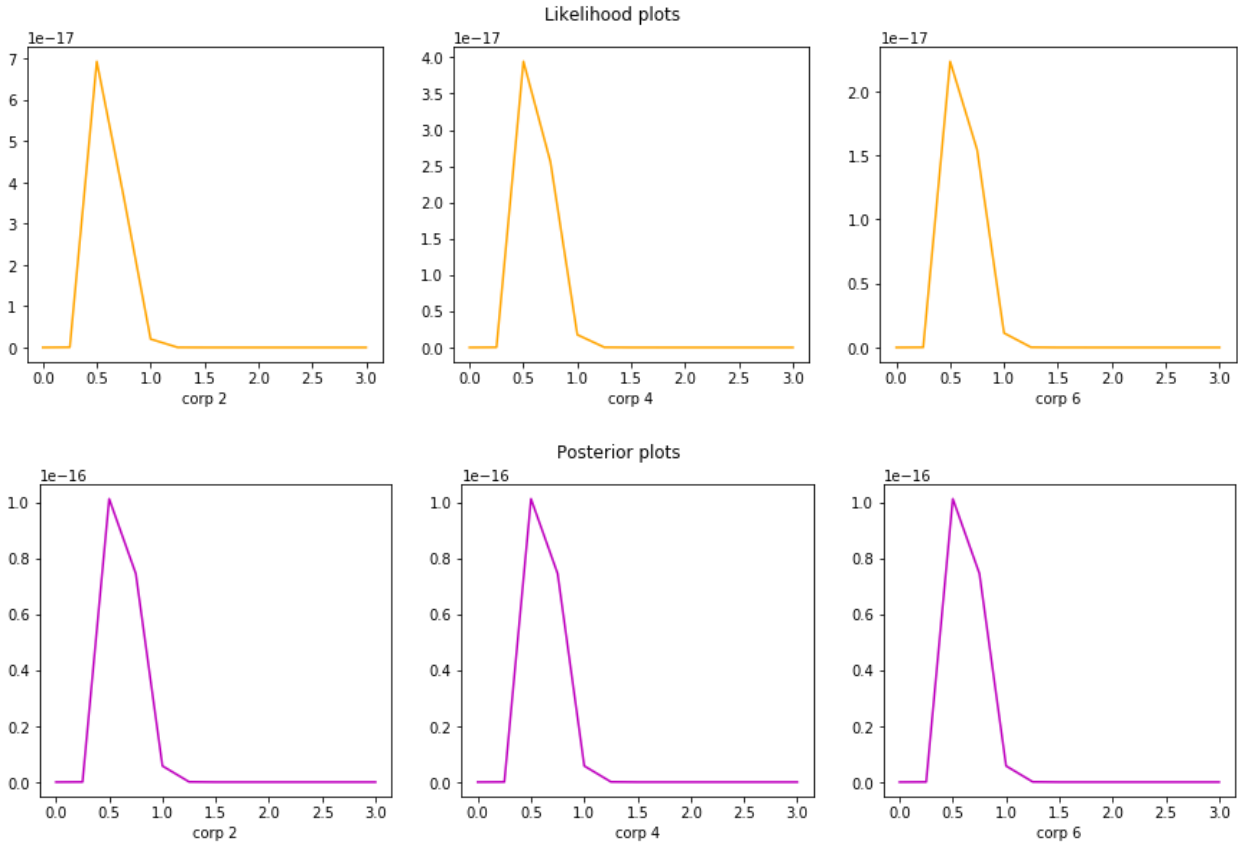
$$\hat{\theta}_{MAP}_{i} = \log\left(\frac{\beta^{\alpha}}{\lceil \alpha} \lambda_i^{\alpha-1} e^{-\beta\lambda_i} + \log \lambda_i^{x_i} e^{-\lambda_i} \right)$$

$$= \frac{d}{d\lambda}\left[(\alpha-1)\log\lambda_i - \beta\lambda_i + x\log\lambda_i - \lambda \right] = 0$$

$$\hat{\theta}_{i\,MAP} = \sum_{i=1}^{N} \hat{\theta}_{i\,MAP} = \left[\frac{\alpha - 1 + \sum_{i=1}^{N} x_i}{\beta + \sum_{i=1}^{N} 1} \right]$$

Hence,

$$\hat{\theta}_{i\,MAP} = \frac{\alpha - 1 + \sum_{i=1}^{N} x_i}{\beta + N}$$



The MAP and ML point estimates for corps **II, IV** and **VI** are provided in the following table.

Corp	MLE	MAP
II	0.6154	0.5581
IV	0.4615	0.4444
VI	0.8462	0.8302

Table MLE and MAP point estimates

Bike sharing demand

Question:

Forecast use of a city bikeshare system

You are provided hourly rental data spanning two years (Data (training and test) available here). The training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period. Fit a Poisson regression model to the count data (output). Treat year, month, weekday, hour, holiday, weather, atemp, humidity, windspeed etc. as input features that are combined linearly to determine the rate parameter of the Poisson distribution. Create a 80-20 split of the train data into training, and validation.

1) Explain maximum likelihood estimation in poisson regression and derive the loss function which is used to estimate the parameters.

2) Find statistics of the dataset like mean count per year, month etc.

3) Plot count against any 5 features.

4) Apply L1 and L2 norm regularization over weight vectors, and find the best hyper-parameter settings for the mentioned problem using validation data and report the accuracy on test data for no regularization, L1 norm regularization and L2 norm regularization.

5) Determine most important features determining count of bikes rented.

Solution :

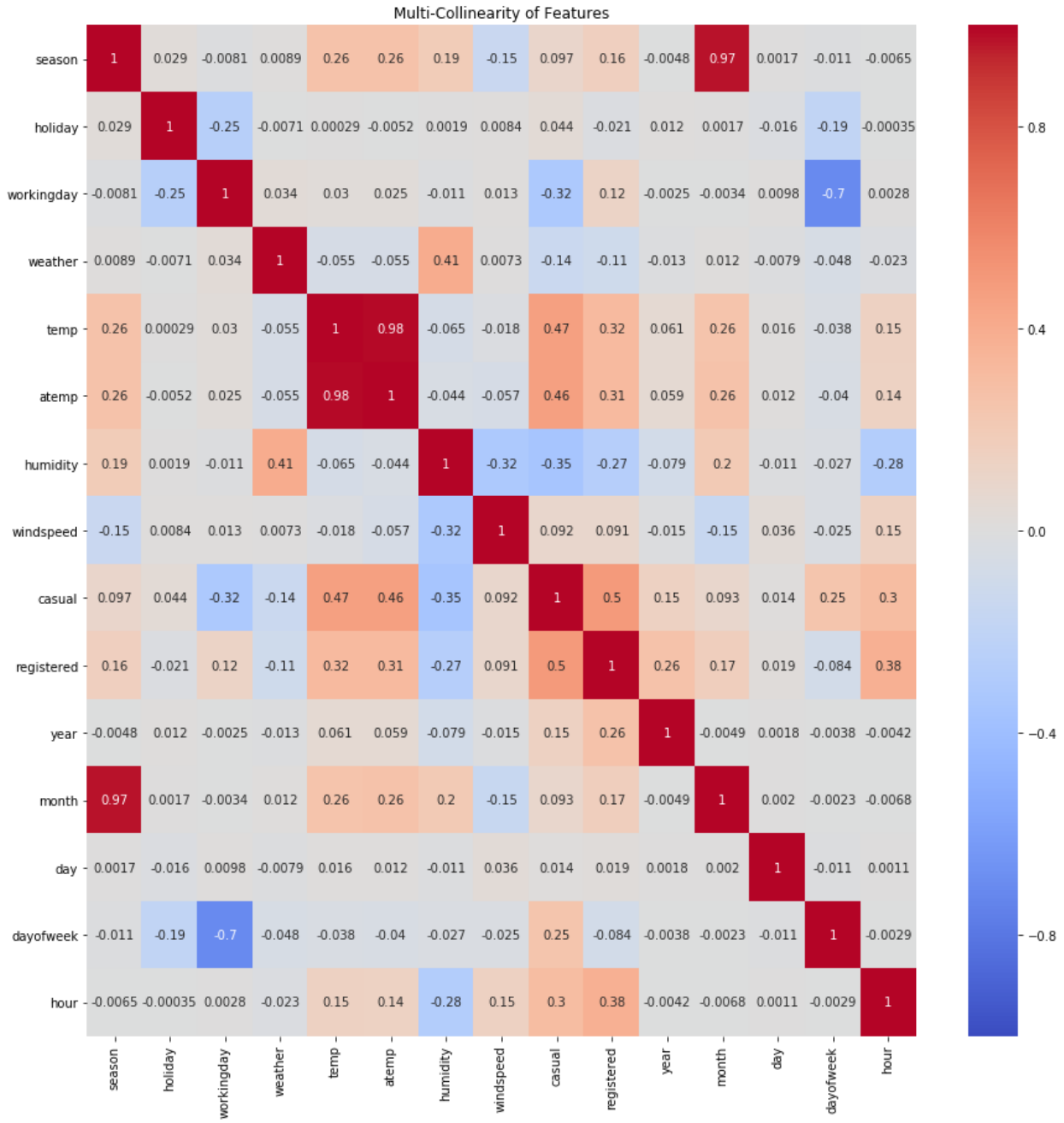
The following are the fields obtained from the training dataset.

- Datetime (Further parsed to hour, date, month and year)
- Season
- Holiday
- Working day
- Weather
- Temp
- Atemp (Absolute Temperature)
- Humidity
- Windspeed
- Casual
- Registered
- Count

After extracting the date and time information from the datetime field, the column is dropped. Hence, a total of 17 columns are obtained of which count (Number of bikes rented) is considered as dependent variable. A total of 10886 observations are available in the training set, which will be further divided into test and validation datasets before performing the prediction for test data.

EDA:

Some initial investigations are performed on data to discover the behavior of the independent variables among themselves to check for multi-collinearity. The correlation between the variables is calculated with the help of correlation function and the output is plotted on a heatmap.



The model predicts bad results if the multi-collinearity among independent variables is high. From the above heat map, it can be observed that the following features are correlated.

Features	Correlation value(Absolute value)
Temp, atemp	0.98
Season, month	0.97
Dayofweek, workingday	0.7
Casual, registered	0.5
Casual, temp	0.47
Registered, hour	0.38

Table Correlation among variables

In general, the casual and registered values have very high correlation values with most of the other features. From the values observed in the above table, the following variables are not considered in the model.

- Atemp
- Season
- Dayofweek
- Casual
- Registered
- Humidity

Train-Test split:

The data consists of hourly count of number of bikes rented across a period of 2 years. The training and test datasets are divided as first 19 days of the month and the remaining days on monthly basis respectively. The training set is further split into training and validation set respectively. The split ratio of training to validation is 80:20.

Metrics calculated:

RMSE and negative log likelihood are calculated for the model.

$$RMSE = \left[\frac{\sum_i^N (y_{pred} - y_i)^2}{N}\right]^{1/2}$$

$$NLL = -\log P\,(y/X, W)$$

Where NLL is the negative log-likelihood. The derivation for NLL is also included in the report.

Gradient descent: (Solution to 4-1)

Gradient descent is used to minimize the loss, which is the negative log likelihood.

SOLUTION.

Explaination of Maximum likelihood estimation in poisson regression and deriving the loss function which is used to estimate the parameters

Here we are given N observations with output y and M features X, for training.

$$(y^1, x_1^1, x_2^1 \text{ ----- } x_m^1)$$
$$(y^2, x_1^2, x_2^2 \text{ ----- } x_m^2)$$
$$\vdots$$
$$(y^N, x_1^N, \text{ ---------- } x_M^N)$$

Here, we define M dimensional vector w to represent weights which will map Inputs to outputs.

N x M-matrix has all inputs

y - N dimensional output vector.

$$\hat{y}_1 = w_1 x_1^1 + w_2 x_2^1 \text{ ----- } w_m x_m^1$$
$$\vdots$$
$$\hat{y}_N = w_1 x_1^N + w_2 x_2^N \text{ ----- } w_m x_m^N$$

Hence $\boxed{y \simeq \hat{y} = Xw}$

Now,

Y has a poisson distribution and assumes the log of its expected value can be modelled by a linear combination of unknown parameters.

Considering $u = Xw$ we have for poisson

$$y^t \sim Po(y^t; u^t)$$

link function of poisson is log. Hence,

$$\log u = w^T x^t$$

$$u^t = e^{(w^T x^t)}$$

$$\boxed{P(y|x,w) = \prod_{t=1}^{N} (u^t)^{y^t} \exp(-u^t)} \quad -(1)$$

Hence Above equation gives likelihood for poisson distribution,

Now, loss for poisson is defined as negative log likelihood.

Hence,

$$\boxed{\text{loss} = -\log P(y|x,w)} \quad -(2)$$

$$= \sum_t u^t - y^t \log u^t$$

Now taking derivative of loss fn. wrt w & equating it to 0.
Now for above no closed solution is possible. Hence, we go with gradient descent approach.

$$\rightarrow \frac{dL}{dw} = 0$$

$$\rightarrow \sum_t x^t u^t - x^t y^t$$

$$\rightarrow \sum_t x^t (u^t - y^t)$$

Hence $w_{new} = w + \Delta w$

where Δw can be calculated as:-

$$\Delta w = -\eta \sum_t x^t (u^t - y^t)$$

Here $\eta \rightarrow$ denotes small step size.

The NLL values of last 10 iterations is presented in the form of a table.

Iterations	**NLL value**
1	-4495.6783
2	-4495.6388
3	-4495.6012
4	-4495.5310
5	-4495.4983
6	-4495.4672
7	-4495.4374
8	-4495.4090
9	-4495.3819
10	-4495.3560

Table 10 iterations of gradient descent

The respective parameters are taken and the RMSE is calculated for the same for training, validation and test datasets. The metrics of test dataset are included in the last part of the solution.

Dataset	**RMSE**
Train	0.2173
Validation	0.2637

Table RMSE values

Normalisation:

All the variables are normalized using the following formula.

$$Norm = \frac{Actual - Min(Feature)}{Max(Feature) - Min(Feature)}$$

Where,

Norm – Normalised value, Actual – Actual values, Min (), Max () – Minimum and Maximum values of the particular feature.

Rate parameter:

Rate parameter is calculated using the following formula

$$Rate = \exp(W^T X)$$

The mean of the rate parameter for all the observations is calculated and obtained as 0.2026. Note that this is for normalized values.

L-1, L-2 Regularisation: (Solution to 4-4)

The Lasso and Ridge regression models are constructed and the numerical procedure for the same is attached.

The RMSE values for training and validation dataset are calculated using the same formula as mentioned in the above section and values are tabulated.

Regularisation	Dataset	RMSE
Lasso	Train	0.216
	Validation	0.2655
Ridge	Train	0.2173
	Validation	0.2636

Table RMSE values for Lasso and Ridge

SOLUTION.

L1 Regularisation.

$$L1 = \text{loss} + \lambda \sum_{i=1}^{p} |w_i|$$

$$= -\log P(y|x,w) + \lambda \sum_{i=1}^{p} |w_i|$$

$$\text{Now, } \frac{dL1}{dw} = \frac{d}{dw}\left(-\log P(y|x,w) + \lambda \sum_{i=1}^{p} |w_i|\right)$$

$|w|$ is differentiable everywhere except when $w=0$,

$$\frac{d|w|}{dw} = \begin{cases} 1 & w > 0 \\ -1 & w < 0 \end{cases}$$

Now lets solve this using gradient descent optimisation based on loss functions.

$$w_{new} = w - \eta \frac{dL}{dw}$$

$$= w - \eta \frac{\partial L}{\partial w}\left[-\log P(y|x,w)\right] + \sum_{i=1}^{p} \lambda \frac{d|w|}{dw}$$

$$= w - \eta \frac{\partial L}{\partial w}\left[-\log P(y|x,w)\right] + \lambda \quad w > 0$$

$$= w - \eta \frac{\partial L}{\partial w}\left[-\log P(y|x,w)\right] - \lambda \quad w < 0$$

$$= w = w_{curr} + w_{new}$$

L2 Regularisation.

$$L2 = \text{Loss} + \lambda \sum_{i=1}^{p} |w_i|^2$$

$$= -\log P(y|x,w) + \lambda \sum_{i=1}^{p} 2|w_i|$$

$$\text{Now, } \frac{dL2}{dw} = \frac{d}{dw}\left(-\log P(y|x,w)\right) + \lambda \sum_{i=1}^{p} 2|w_i|$$

Now lets solve this using gradient descent optimisation method based on loss functions.

$$w_{new} = w - \eta \frac{\partial L_2}{\partial w}$$

$$= w - \eta \frac{\partial L_2}{\partial w}\left[-\log P(y|x,w)\right] + \sum_{i=1}^{p} 2\lambda \frac{d|w|}{dw}$$

$$w = w_{curr} + w_{new}$$

Statistics: (Solutions to 4-2, 4-3)

Some of the statistics are calculated based on the training data. Also, plots are provided for the same.

- Average number of bikes rented per year = 1042738.0
- Hourly mean of bikes rented – Year wise

Year	**Hourly mean**
2011	144.22
2012	238.56

- Hourly mean of bikes rented – Month wise

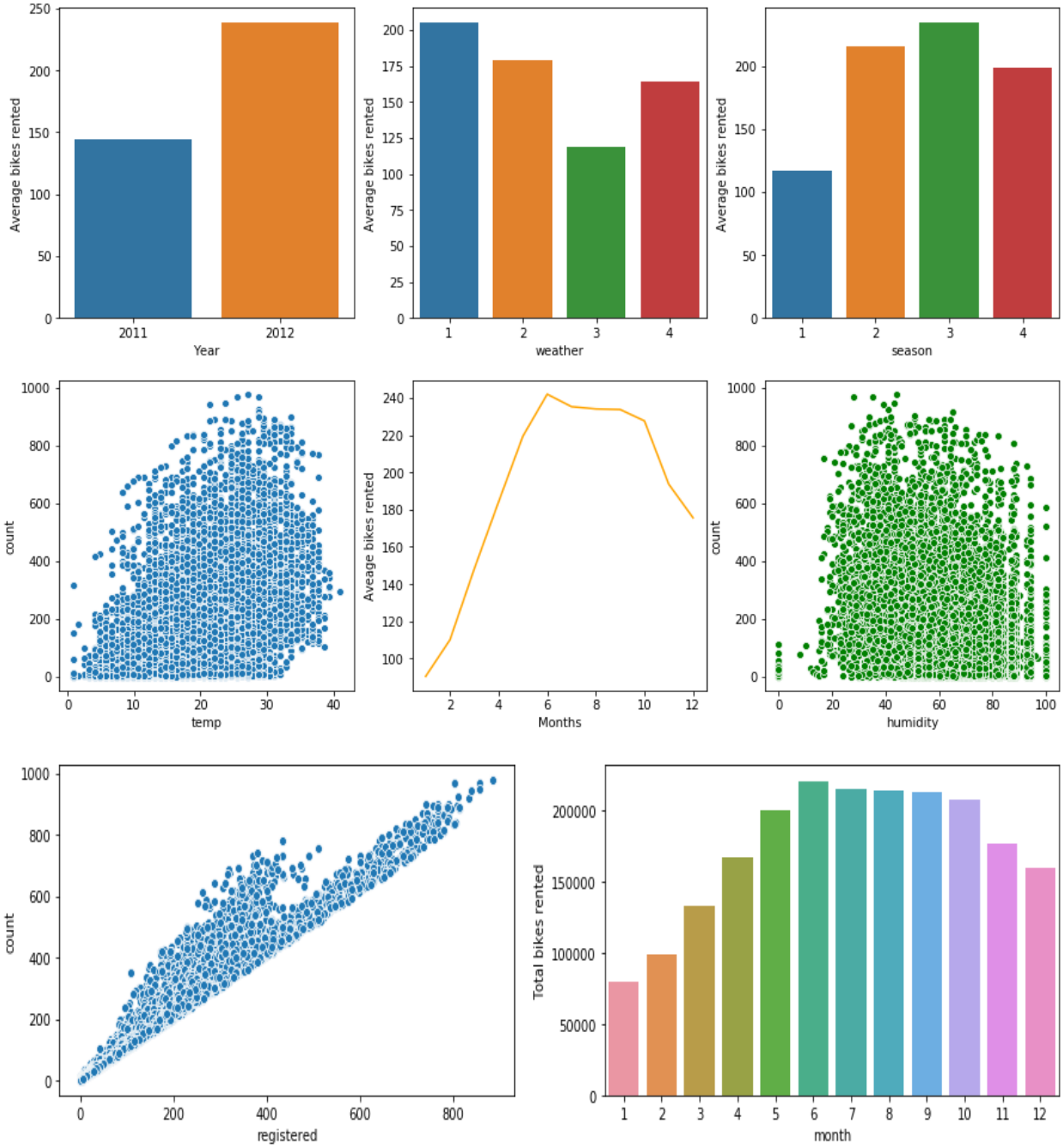
Month	Hourly mean
1	99.366
2	110.0
3	148.17
4	184.16
5	219.46
6	242.03
7	235.33
8	234.12
9	233.81
10	227.7
11	193.68
12	175.61

- Hourly mean of bikes rented – Season wise

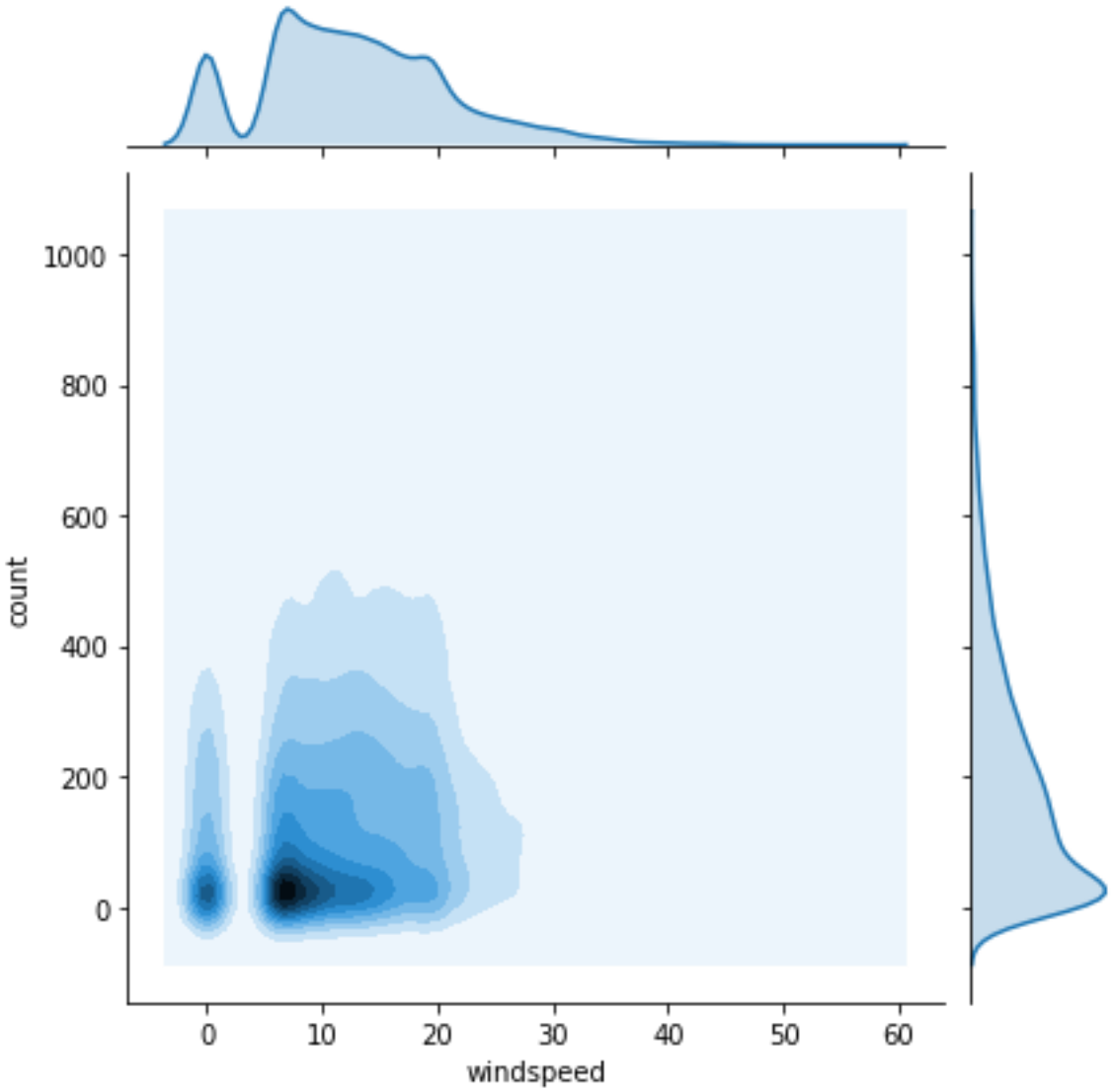
Season	Hourly mean
1	116.34
2	215.25
3	234.42
4	198.99

- Hourly mean of bikes rented – Weather wise

Weather	Hourly mean
1	205.24
2	178.96
3	118.85
4	164



- Scatter plots – Count vs Humidity, Temp, Registered
- Histogram plot - Count vs Month, Average bikes rented on hourly basis vs year, weather and season
- Joint plot – Count vs windspeed



Important features: (Solution to 4-5)

From the value of weights obtained in Lasso and ridge regression, the following features are written in descending order of importance (Top feature is more important).

1. Windspeed
2. Weather
3. Day
4. Month
5. Workingday
6. Holiday

7. Hour
8. Year
9. Temp

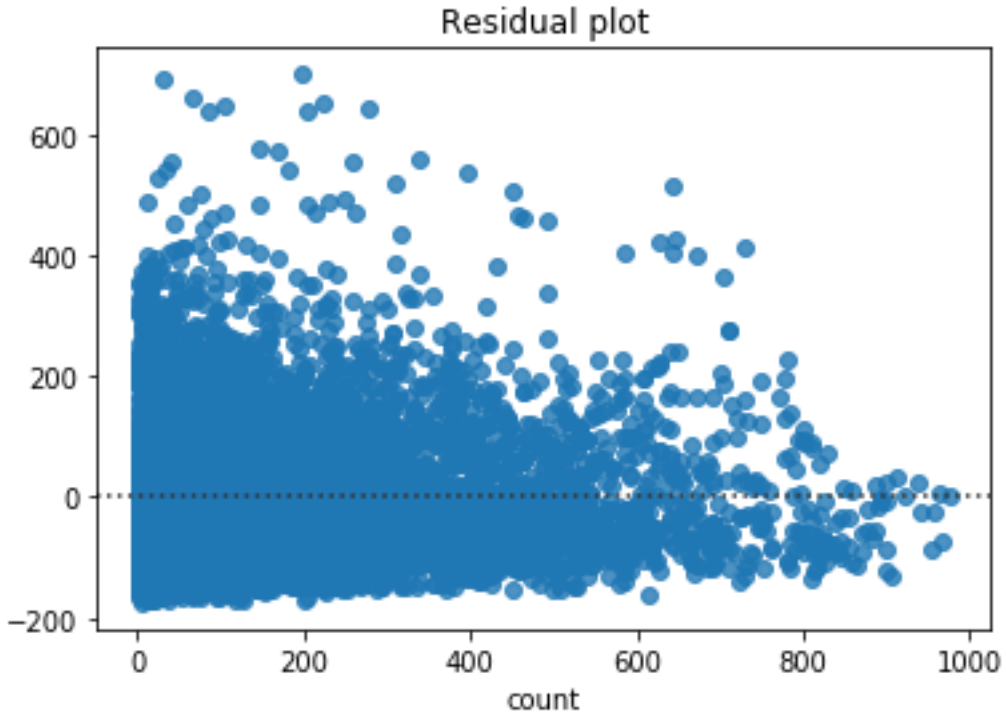
Results:

The final values of the weights for each of the above-mentioned features at convergence point is as shown.

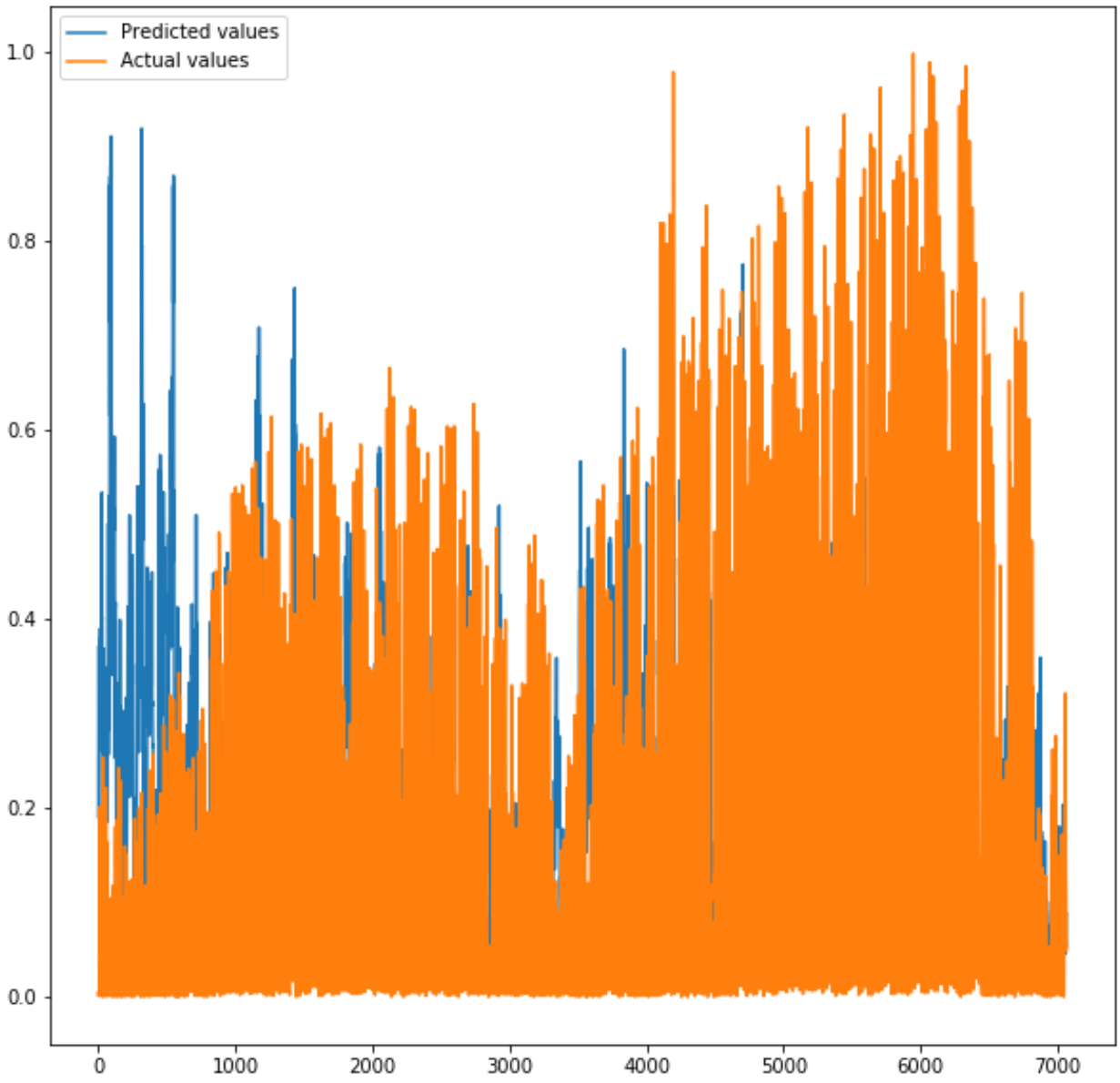
```
In [37]:  1  w

Out[37]: array([-0.54002184, -0.54848207, -1.29795405, -0.05774403, -1.51185166,
                -0.07530485, -0.77360886, -0.92177805,  0.14650406], dtype=float128)
```

The same weights are used against the features for test data to make the predictions.



The residual plot is plotted for the residuals (difference between actual and predicted values) versus predicted values.



The standard plot is plotted for predicted vs actual values.

RMSLE is used as metric for actual values of test dataset.

$$RMSLE = \left[\frac{\sum_i^N \left(\log\left(y_{pred}+1\right) - \log\left(y_i+1\right)\right)^2}{N}\right]^{1/2}$$

Dataset	**RMSLE**
Test data (Normalised values)	0.173
Test data (Actual values)	1.6145