Poisson & GLM

Liangjiayi Wang

November 2022

1 Introduction-Poisson

The Poisson distribution is typically used to model counts; this is a natural choice for a number of reasons, including the fact that counts, like the Poisson distribution, take on non-negative integer values.

2 Functions

Suppose that a random variable Y takes on non-negative integer values, Poisson i.e. $Y \in 0, 1, 2, ...$ If Y follows the Poisson distribution, then

$$P(Y = k) = \frac{e^{-\lambda} * \lambda^k}{k!} for\ k = 0, 1, 2....$$

Here, $\lambda > 0$ is the expected value of Y , i.e. E(Y). It turns out that λ also equals the variance of Y , i.e. $\lambda = E(Y) = Var(Y)$. This means that if Y follows the Poisson distribution, then the larger the mean of Y , the larger its variance.

(In actual, the λ will not be a constant value)

Notice that we take the $log(\lambda(X_1, ..., X_p))$ to be linear in $X_1, ..., X_p$, rather than having $\lambda(X_1, ..., X_p)$ itself be linear in $X_1, ..., X_p$, in order to ensure that $\lambda(X_1, ..., X_p)$ takes on non-negative values for all values of the covariates.

$$E(Y|X_1...X_p) = \sum_{k=0}^{\infty} k * P(Y = k) = \lambda(X_1, ..., X_p)$$

$$Var(Y) = E(Y - E(Y))^2 = E(Y^2) - (E(Y))^2 = \sum_{k=0}^{\infty} k^2 * P(Y = k) - \lambda^2 = \lambda$$

$$log(\lambda(X_1, ..., X_p)) = \beta_0 + \beta_1 X_1 + ... + \beta_p * X_p$$

$$\lambda(X_1, ..., X_p) = e^{\beta_0 + \beta_1 X_1 + ... + \beta_p * X_p}$$

3 Likelihood

$$l(\beta_0, \beta_1, ..., \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} \lambda(x_i)_i^y}{y_i!}$$

where

$$\lambda(x_i) = e^{\beta_0 + \beta_1 X_{i1} + ... + \beta_p * X_{ip}}$$

4 Interpretation

We might model Y as a Poisson distribution with mean $E(Y) = \lambda = 5$ (i.e, average user=5). This means that:

The probability of no users during this particular hour is

$$Pr(Y = 0) = \frac{e^{-5} * 5^0}{0!} = e^{-5} = 0.0067$$

The probability of 1 users during this particular hour is

$$Pr(Y = 1) = \frac{e^{-5} * 5^1}{1!} = 5e^{-5} = 0.034$$

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	4.12	0.01	683.96	0.00
`workingday`	0.01	0.00	7.5	0.00
`temp`	0.79	0.01	68.43	0.00
`weathersit[cloudy/misty]`	-0.08	0.00	-34.53	0.00
`weathersit[light rain/snow]`	-0.58	0.00	-141.91	0.00
`weathersit[heavy rain/snow]`	-0.93	0.17	-5.55	0.00

TABLE 4.11. *Results for a Poisson regression model fit to predict* `bikers` *in the* `Bikeshare` *data. The predictors* `mnth` *and* `hr` *are omitted from this table due to space constraints, and can be seen in Figure 4.15. For the qualitative variable* `weathersit`*, the baseline corresponds to clear skies.*

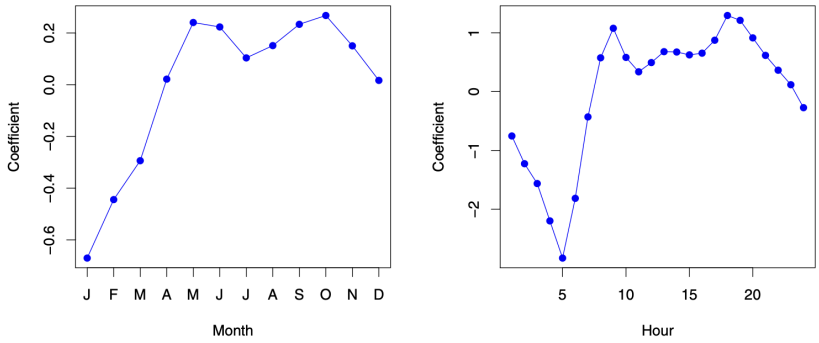

FIGURE 4.15. *A Poisson regression model was fit to predict* `bikers` *in the* `Bikeshare` *data set.* Left: *The coefficients associated with the month of the year. Bike usage is highest in the spring and fall, and lowest in the winter.* Right: *The coefficients associated with the hour of the day. Bike usage is highest during peak commute times, and lowest overnight.*

An increase in X_j by one unit is associated with a change in $E(Y) = \lambda$ by a factor of $exp(\beta_j)$.

For example, a change in weather from clear to cloudy skies is associated with a change in mean bike usage by a factor of $exp(-0.08) = 0.923$, i.e. on average, only 92.3% as many people will use bikes when it is cloudy relative to when it is clear.

If the weather worsens further and it begins to rain, then the mean bike usage will further change by a factor of $exp(-0.5) = 0.607$, i.e. on average only 60.7% as many people will use bikes when it is rainy relative to when it is cloudy.

4.1 Mean-variance relationship

Under the Poisson model, $\lambda = E(Y) = Var(Y)$. We implicitly assume that mean bike usage in a given hour equals the variance of bike usage during that hour.

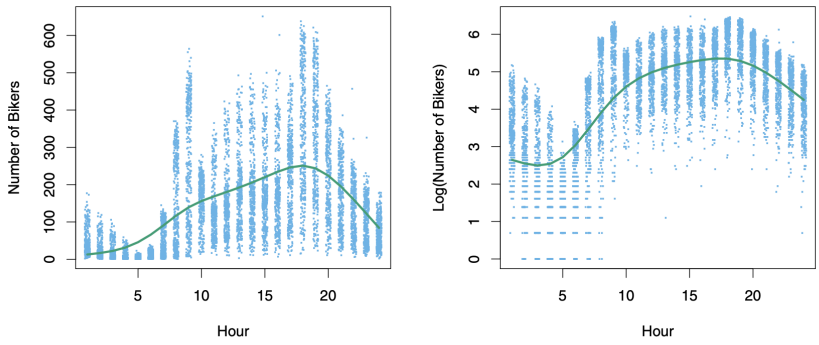

FIGURE 4.14. Left: *On the* `Bikeshare` *dataset, the number of bikers is displayed on the y-axis, and the hour of the day is displayed on the x-axis. Jitter was applied for ease of visualization. For the most part, as the mean number of bikers increases, so does the variance in the number of bikers. A smoothing spline fit is shown in green.* Right: *The log of the number of bikers is now displayed on the y-axis.*

By contrast, under a linear regression model, the variance of Y should always takes on a constant value.

Note: It is reasonable to suspect that when the expected value of bikers is small, the variance of bikers should be small as well. And in this case, when biking conditions are favorable, both the mean and the variance in bike usage are much higher than when conditions are unfavorable. However, this is a major violation of the assumptions of a linear model, which state that $Y = \sum_{j=1}^{p} X_j\beta_j + \epsilon$, where ϵ is a mean-zero error term with variance σ^2 that is constant, and not a function of the covariates. Therefore, the heteroscedasticity of the data calls into question the suitability of a linear regression model.

4.2 Non-negative fitted values

There are no negative predictions using the Poisson regression model. This is because the Poisson model itself only allows for non-negative values.

5 Additional Examples: Fiji baby

The dataset has 70 rows representing grouped individual data. Each row has entries for:

- The cell number (1 to 71, cell 68 has no observations),
- marriage duration (1=0-4, 2=5-9, 3=10-14, 4=15-19, 5=20-24, 6=25-29),
- residence [Suva is the capital city of Fiji, Urban means other urban areas except Suva.] (1=Suva, 2=Urban, 3=Rural),
- education (1=none, 2=lower primary, 3=upper primary, 4=secondary+),
- mean number of children ever born (e.g. 0.50),
- variance of children ever born (e.g. 1.14), and
- number of women in the cell (e.g. 8).

The dataset only contains grouped data instead of babies each individual woman has. But we only summarized mean of number of babies, variance of number of babies and the sample size of women with identical predictors in one row.

Are we able to still run Poisson regression? Yes, but using the nice "offset" feature of "glm" function. Suppose the l-th woman in a group has Y_l babies. The group total is $Y = \sum_{l=1}^{n} Y_l$ and we have n women in this group. In Poisson regression, if we know each Y_l, we assume

$$\log E[Y_l] = X'\beta.$$

Note that all Y_l's have shared predictors X. Then we have

$$\log E[Y] = \log E[\sum_{l=1}^{n} Y_l] = \log \sum_{l=1}^{n} E[Y_l] = \log nE[Y_l] = \log n + X'\beta.$$

This means if we only observe Y, but fail to observe each Y_l, we can still treat Y as the observed count data, but with an additional offset term $\log n$. Let us now add the offset term and fit the Poisson regression.

Table 1: Nominal logistic regression

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.117	0.055	-2.132	0.033	-0.226	-0.010
dur5-9	0.997	0.053	18.902	0.000	0.894	1.101
dur10-14	1.369	0.051	26.815	0.000	1.270	1.470
dur15-19	1.614	0.051	31.522	0.000	1.514	1.715
dur20-24	1.785	0.051	34.852	0.000	1.685	1.886
dur25-29	1.976	0.050	39.501	0.000	1.879	2.075
resrural	0.152	0.028	5.353	0.000	0.096	0.207
resurban	0.112	0.032	3.459	0.001	0.049	0.176
educlower	0.023	0.023	1.014	0.311	-0.021	0.067
educupper	-0.101	0.031	-3.268	0.001	-0.162	-0.041
educsec+	-0.310	0.055	-5.618	0.000	-0.420	-0.203

- What is the expected number of children for a Suvanese woman with no education who have been married 0-4 years?

$$\exp\{-0.1173\} = 0.89.$$

- As we move from duration 0-4 to 5-9 the log of the mean increases by almost one, which means that the number of CEB gets multiplied by $\exp\{0.9977\} = 2.71$. By duration 25-29, women in each category of residence and education have $\exp\{1.977\} = 7.22$ times as many children as they did at duration 0-4.

- The effects of residence show that Suvanese women have the lowest fertility. At any given duration since first marriage, women living in other urban areas have 12% larger families ($exp\{0.1123\} = 1.12$) than Suvanese women with the same level of education. Similarly, at any fixed duration, women who live in rural areas have 16% more children ($\exp\{0.1512\} = 1.16$), than Suvanese women with the same level of education.

- Finally, we see that higher education is associated with smaller family sizes net of duration and residence. At any given duration of marriage, women with upper primary education have 10% fewer kids, and women with secondary or higher education have 27% fewer kids ($1 - \exp\{-0.3096\} = 0.27$), than women with no education who live in the same type of place of residence.

Do we have interaction effect between marital duration and education here?

From the model itself, we did not include any interaction term. However, the model is additive in the log scale. In the original scale the model is multiplicative, and postulates relative effects which translate into different absolute effects depending on the values of the other predictors.

Consider the effect of education. Women with secondary or higher education have 27% fewer kids than women with no education.

```
##      dur  res educ n prediction
## 1    0-4 Suva none 1  0.8894987
## 2    0-4 Suva sec+ 1  0.6523026
## 3    5-9 Suva none 1  2.4105084
## 4    5-9 Suva sec+ 1  1.7677157
## 5  10-14 Suva none 1  3.4983737
## 6  10-14 Suva sec+ 1  2.5654879
## 7  15-19 Suva none 1  4.4667458
## 8  15-19 Suva sec+ 1  3.2756312
## 9  20-24 Suva none 1  5.3005682
## 10 20-24 Suva sec+ 1  3.8871043
## 11 25-29 Suva none 1  6.4192929
## 12 25-29 Suva sec+ 1  4.7075068
```

Note that the effect of 27% fewer kids means different number of kids when the marital duration changes. If we had used OLS regression for these data we would have ended up with a large number of interaction effects to accommodate the fact that residence and educational differentials increase with marital duration.

6 Introduction-GLM

- Each approach uses predictors $X_1, ..., X_p$ to predict a response Y. We assume that, conditional on $X_1, ..., X_p$, Y belongs to a certain family of distributions.

  For linear regression, we typically assume that Y follows a Gaussian or normal distribution.

  For logistic regression, we assume that Y follows a Bernoulli distribution.

  For Poisson regression, we assume that Y follows a Poisson distribution.

- Each approach models the mean of Y as a function of the predictors. In linear regression, the mean of Y takes the form:

  For linear:

  $$E(Y|X_1, ..., X_p) = \beta_0 + \beta_1 * X_1 + ... + \beta_p * X_p$$

  For logistic:

  $$E(Y|X_1, ..., X_p) = P(Y = 1|X_1...X_p) = \frac{e^{\beta_0+\beta_1*X_1+...+\beta_p*X_p}}{1 + e^{\beta_0+\beta_1*X_1+...+\beta_p*X_p}}$$

  For Poisson:

  $$E(Y|X_1, ..., X_p) = \lambda(X_1...X_p) = e^{\beta_0+\beta_1*X_1+...+\beta_p*X_p}$$

  Equations can be expressed using a link function, η, which applies a transformation to $E(Y|X_1, ..., X_p)$ so that the transformed mean is a linear function of the predictors. That is, $\eta(E(Y|X_1, ..., X_p)) = \beta_0 + \beta_1 * X_1 + ... + \beta_p * X_p$ The link functions for linear, logistic and Poisson regression are $\eta(\mu) = \mu, \eta(\mu) = log(\mu/(1 - \mu)), and \eta(\mu) = log(\mu)$, respectively.