Week 5: Poisson Regression

MATH-516 Applied Statistics

Tomas Masak

March 20th 2023

Section 1

Log-linear Model

Poisson is Exponential Family

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for} \quad x \in \{0, 1, 2 \ldots\} \text{ and } \lambda \in (0, \infty)$$
$$= \exp\left(x \log \lambda - \lambda + \log(1/x!)\right)$$

hence

- $\theta = \log \lambda$, $\varphi = 1$, $b(\theta) = e^\theta$, and $c(x, \varphi) = \log(1/x!)$
- $\mathrm{var}(Y) = \lambda$ and $\mu = \mathbb{E}X = \lambda \Rightarrow V(\mu) = \mu$
- canonical link must satisfy $g(\mu) = \theta = \log \mu$
  - i.e. $\mu_n = e^{X_n^\top \beta}$ this is why Poisson regression is sometimes referred to as the log-linear model

Interpreting Parameters

- let $X_n^\top \beta = \beta_1 + X_{n,2}\beta_2 + \ldots + X_{n,p}\beta_p$
- the intercept captures the expected frequency (count) with zero regressors:
  - $\mathbb{E}[Y_n \mid X_{n,2} = \ldots = X_{n,p} = 0] =: \lambda_0$
  - $\Rightarrow \lambda_0 = e^{\beta_1}$ ... expected frequency with all-zero regressors
- other coefficients capture proportional change in expected frequency (between two observations that differ by 1 in the corresponding regressor):
  - let $x = (x_1, \ldots, x_p)^\top$ and $\tilde{x}^{(j)} = (x_1, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p)^\top$
  - $\mathbb{E}(Y_n \mid X_n = x) =: \lambda_n \Rightarrow \lambda_n = e^{\beta^\top x}$
  - $\mathbb{E}(Y_n \mid X_n = \tilde{x}^{(j)}) =: \tilde{\lambda}_n^{(j)} \Rightarrow \tilde{\lambda}_n^{(j)} = e^{\beta^\top \tilde{x}^{(j)}}$
  - dividing: $e^{\beta_j} = \tilde{\lambda}_n^{(j)} / \lambda_n$ ... proportional change when $X_{n,j} \mapsto X_{n,j} + 1$

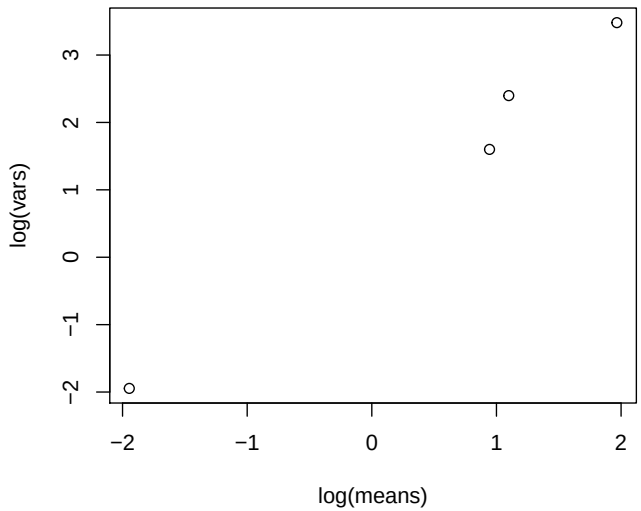
Small Example

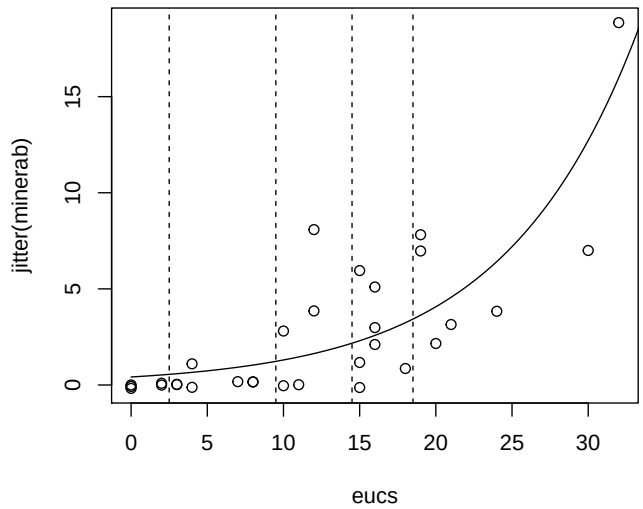
Noisy miners are small but aggressive native Australian birds. A study counted the birds in two hectare transects.

```
op <- par(mar=c(4,4,1,1))
plot(log(means),log(vars)) # looks linear with slope 1 and intercept 0, which is exa
gm1 <- glm(minerab ~ eucs, data=nminer, family=poisson(lin="log"))
sum(residuals(gm1, type="pearson")^2)/(31-2) # indicates that the model is not great
```

```
## [1] 2.327851
```

```
plot(jitter(minerab) ~ eucs, data=nminer)
abline(v=mybreaks, lty=2)
x <- seq(0,35,length=100)
points(x,exp(coef(gm1)[1] + x*coef(gm1)[2]),type="l")
```





Section 2

Contingency Tables

All Variables Binned

- often (e.g. sample surveys), all variables are categorical
- say we have two variables $Y \in \{1, \dots, I\}$ and $Z \in \{1, \dots, J\}$
- $(Y_1, Z_1)^\top, \dots, (Y_N, Z_N) \overset{\perp\!\!\!\perp}{\sim} (Y, Z)^\top$ is our random sample
- then the frequencies $N_{ij} = \sum_n \mathbb{I}_{[Y=i, Z=j]}$ define a (two-way) contingency table

	$Z = 1$	$\cdots$	$Z = J$	$\sum$
$Y = 1$	N_{11}	$\cdots$	N_{1J}	N_{1+}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Y = I$	N_{I1}	$\cdots$	N_{IJ}	N_{I+}
$\sum$	N_{+1}	$\cdots$	N_{+J}	$N_{++} \equiv N$

- denote expected frequencies $m_{ij} = \mathbb{E}N_{ij}$
  - and accordingly m_{i+}, m_{+j} and m_{++}
- denote probabilities of observing (i, j) by $\pi_{ij} = P(X = i, Z = j)$
  - and accordingly π_{i+}, π_{+j} and $\pi_{++} = 1$
- if $(Y, Z) \perp\!\!\!\perp N$, we have $m_{ij} = m_{++}\pi_{ij}$
  - what we observe does not depend on how many times we draw

Two Probabilistic Models

1. $N_{ij} \overset{\perp\!\!\!\perp}{\sim} Po(m_{ij})$
   - N itself is random here
   - no. of observations IJ is fixed (asymptotics?)
   - if $m_{ij} = e^{\alpha + x_{ij}^\top \beta}$, we have a log-linear model
     - elegant, exponential family, etc.
     - X_{ij} is the row of the design matrix (depends on the parametrization chosen for the factors)
2. $(N_{11}, \ldots, N_{IJ})^\top \sim Mult(N, \pi)$ with $\pi = (\pi_{11}, \ldots, \pi_{IJ})^\top$
   - N is fixed here and it is the no. of observations drawn from $Mult(1, \pi)$ $\Rightarrow$ classical MLE asymptotics
   - not as elegant to work with, luckily we don't need to...
   - $\pi_{ij} = m_{ij}/m_{++} = e^{\alpha + x_{ij}^\top \beta} / \sum_{ij} e^{\alpha + x_{ij}^\top \beta}$ does not depend on α
     - α only affects the frequencies (either N or equivalently any single field such as N_{11}), not the probabilities
     - we often seek interpretation in terms of probabilities

Claim. Since, a vector of independent Poissons given their sum is Multinomial, likelihoods based on 1. and 2. are equivalent w.r.t. β.

Independence Model for Two-way Table

- pseudo-contrast parametrization (no interaction)

$$\log m_{ij} = \alpha + \beta_i^Y + \beta_j^Z \quad \text{with} \quad \beta_1^Y = \beta_1^Z = 0$$

- how does the model matrix look like?
- $\log m_{11} = \alpha \Rightarrow e^{\alpha}$ is the expected frequency of the first entry

$$\pi_{ij} = m_{ij}/m_{++} \Rightarrow \log \pi_{11} = \alpha - \log m_{++}$$
$$\Rightarrow \log \pi_{ij} = \log \pi_{11} + \beta_i^Y + \beta_j^Z$$

- $e^{\beta_i^Y} = \pi_{ij}/\pi_{1j} = P(Y = i, Z = j)/P(X = 1, Z = j)$ for all j
  $\Rightarrow e^{\beta_i^Y}$ is odds of $Y = i$ against $Y = 1$
  - similarly $e^{\beta_j^Z}$ is odds of $Z = j$ against $Z = 1$ (irrespective of Y)
- clearly, Y and Z are independent
  - it can be calculated that $\pi_{ij} = \pi_{i+}\pi_{+j}$ (equivalent to independence)

Dependence Model for Two-way Table

- pseudo-contrast again (with interaction this time)

$$\log(m_{ij}) = \alpha + \beta_i^Y + \beta_j^Z + \beta_{ij}^{YZ} \quad \text{with} \quad \beta_1^Y = \beta_1^Z = \beta_{i1}^{YZ} = \beta_{1j}^{YZ} = 0$$

- as before, e^α is the expected frequency of the first entry and

$$\log \pi_{ij} = \log \pi_{11} + \beta_i^Y + \beta_j^Z + \beta_{ij}^{YZ} \quad =: \Delta_{ij}$$

- $e^{\beta_i^Y} = \pi_{i1}/\pi_{11} = P(Y = i, Z = 1)/P(X = 1, Z = 1)$
  $\Rightarrow e^{\beta_i^Y}$ is odds of $Y = i$ against $X = 1$ given $Z = 1$
  - similarly $e^{\beta_j^Z}$ is odds of $Z = 1 \mapsto Z = j \mid X = 1$
- to isolate β_{ij}^{YZ}, one takes $\Delta_{ij} - \Delta_{1j} - \Delta_{i1}$ and obtains
  $\Rightarrow \beta_{ij}^{YZ} = \frac{\pi_{ij}/\pi_{1j}}{\pi_{i1}/\pi_{11}}$ is the odds ratio
  - how many times the odds of $Y = i$ against $Y = 1$ change when Z changes from 1 to j
  - or equivalently it is the change in odds of $Z = j$ against $Z = 1$ for $Y = 1 \mapsto Y = j$
- clearly, Y and Z are dependent now

Test of Independence in a Two-way Table

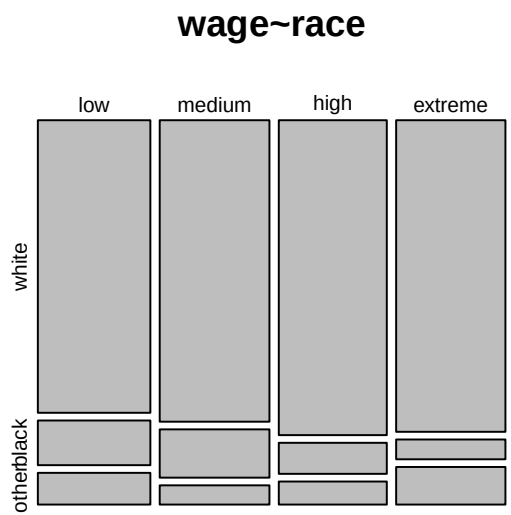
- in introductory statistics classes, two tests of independence for 2×2 tables are usually taught:
  - Pearson χ^2 test (asymptotic, ROT: $m_{ij} \geq 5$)
  - Fisher factorial test (exact, applicable when expected frequencies are small)
- many other tests (McNemmar, Cochran-Mantel-Haenszel) exist, some of them able to handle multi-way tables
- **deviance test**: model-submodel test between the dependence and independence models above
  - asymptotic (ROT: $m_{ij} \geq 3 (\vee 5)$)
  - can be easily generalized to multi-way tables
  - goodness of fit test
    - the larger model is saturated, i.e. it always holds
    - in case of all regressors discrete, the saturated model can be consistently estimated
    - the saturated model is rarely useful (what if the dependence is simply due to a confounder we missed? X dependent on Z and $X \perp\!\!\!\perp Z \mid W$ for some W do not exclude each other)

Example: Wage & Race

```
tab1 <- table(Wage$wage_cat, Wage$race)
tab1
```

```
##
##             white black other
##    low        609    93    66
##    medium     612    98    39
##    high       629    62    46
##    extreme    630    40    76
```

```
mosaicplot(tab1,main="wage~race")
```



Example: Wage & Race

```
Data <- as.data.frame(tab1)
names(Data) <- c("wage","race", "freq")
gm1 <- glm(freq ~ wage+race, data=Data, family=poisson(link="log"))
gm2 <- glm(freq ~ wage*race, data=Data, family=poisson(link="log"))
anova(gm1,gm2,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: freq ~ wage + race
## Model 2: freq ~ wage * race
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         6     48.041
## 2         0      0.000  6   48.041 1.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- wage and race are dependent, but could this be due to some other variables, e.g. education?

Three-way Tables

- we now have 3 variables: Y,Z and W
  - assume we are interested mostly in Y and Z and their relationship, W is just something we need to control for to get the relationship right
- possible models (that depend on all the variables):
  1. `freq~Y+W+Z`
     - $Y \perp\!\!\!\perp W \perp\!\!\!\perp Z$
  2. `freq~Y*W+Z` (or similarly `freq~Y+W*Z`)
     - $(Y, W)^\top \perp\!\!\!\perp Z$
  3. `freq~Y*W+W*Z*`
     - Y and Z dependent through W, but $Y \perp\!\!\!\perp Z \mid W$
  4. `freq~Y*W+W*Z+X*Z` (i.e. `freq~(.)^2` in short)
     - Y dependent on Z, even when conditioning on W, but the conditional relationship is the same regardless of the value of W
  5. `freq~Y*W*Z` (i.e. `freq~(.)^3`)
     - the saturated model

Interpretation in Three-way Tables

- pseudocontrast parametrization for the three-way interaction model:

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^Y + \beta_j^Z + \beta_k^V + \beta_{ij}^{YZ} + \beta_{ik}^{YW} + \beta_{kj}^{WZ} + \beta_{ijk}^{YZW}$$

  with constraints: β's with at least one index being 1 are zero
- $e^{\beta_i^Y}$ is odds of $Y = i$ against $Y = 1$
  - irrespective of Z and W if there are no interactions
  - for $Z = 1$ if there is `Y:Z` interaction
  - for $Z = 1$ and $W = 1$ if there are `Y:Z` and `Y:W` interactions
- $e^{\beta_{ij}^{YZ}}$ is the odds ratio - how many times the odds of $Y = i$ against $Y = 1$ change when $Z = 1 \mapsto j$
  - irrespective of W if the three-way interaction is not present
  - given $W = 1$ if the three-way interaction is present
- $e^{\beta_{ijk}^{YZW}}$ is the ratio of conditional odds ratios
  - a bit awkward interpretation
  - if this interaction is present, it implies that conditional relationships between Y and Z depend on the value of the conditioning variable W
- in the above, permute (Y, Z, W) to obtain the remaining interpretations

Example: Wage, Education & Race

```
tab2 <- table(Wage$education,Wage$wage_cat, Wage$race)
Data <- as.data.frame(tab2)
names(Data) <- c("education","income","race", "freq")
gm1 <- glm(freq ~ (.)^2, data=Data, family=poisson(link="log"))
gm2 <- glm(freq ~ (.)^3, data=Data, family=poisson(link="log"))
anova(gm1,gm2,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: freq ~ (education + income + race)^2
## Model 2: freq ~ (education + income + race)^3
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        24     30.298
## 2         0      0.000 24   30.298   0.1751
```

- go for the simpler model
- can we simplify further, in particular to a model without `race*wage`?

Example: Wage, Education & Race

```
library(car)
Anova(gm1,type=2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: freq
##                  LR Chisq Df Pr(>Chisq)
## education          496.65  4  < 2.2e-16 ***
## income               0.68  3  0.8783807
## race              3112.41  2  < 2.2e-16 ***
## education:income   802.34 12  < 2.2e-16 ***
## education:race      69.90  8   5.15e-12 ***
## income:race         27.48  6  0.0001178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- we cannot simplify any further, `race` and `wage` are still dependent
- the relationship between `race` and `wage` does not depend on `education`

Example: Wage, Education & Race

`summary(gm1)` provides

```
Coefficients:
                           Estimate Std. Err. z value Pr(>|z|)
incomemedium:raceblack      0.02067  0.15921   0.130  0.896708
incomehigh:raceblack       -0.41332  0.18091  -2.285  0.022331 *
incomeextreme:raceblack    -0.74089  0.21882  -3.386  0.000710 ***
incomemedium:raceother     -0.54964  0.21585  -2.546  0.010883 *
incomehigh:raceother       -0.61067  0.21491  -2.841  0.004491 **
incomeextreme:raceother    -0.42511  0.21219  -2.003  0.045133 *
```

showing that, for example:

- the odds of `extreme` salary (against the `low` bottomline) are more than double for whites as opposed to blacks ($e^{-0.74} \approx 0.48$) with the same `education` (regardless of the `education` level)
- the odds of `medium` salary (against the `low` bottomline) are about 2 % higher for blacks as opposed to whites ($e^{0.02} \approx 10.2$) again with the same `education` and regardless of the `education` level

Example: Wage, Education & Race

Don't forget about stress-testing:

- the model has some mild issues, it doesn't fit well low-education low-income whites and high-education high-income non-whites
- otherwise all looks good
- `sum(fitted(gm1)<5)` shows there are 8 (out of 60) table entries with fitted frequencies lower than 5, which is not negligible
  - we could simulate from the fitted model (a.k.a. parametric bootstrap) to ascertain whether the residual deviance (30.298 on 24 df) is suspiciously high (it is not, so the model seems adequate)

Section 3

Project 3

The Goal

- home advantage is a real thing in football and other sports
- during Covid, English Premier League (EPL) games played behind closed doors for 18 months
    - for simplicity, let's say that fans weren't allowed into stadiums from March 12, 2020 until the beginning of the 2021-22 season

Question: Did home advantage persist through Covid?

- specifically, test whether the home advantage reduced during Covid
- secondarily, quantify the home advantage before Covid and probe the development after Covid

Data

Sub-folder `Premier_League` of folder `Data` contains match results for 4 EPL seasons:

- 2018-2019 - pre-covid
- 2019-2020 - pre-covid until March 12, 2020, then cancelled
- 2020-2021 - in-covid
- 2021-2022 - post-covid

freely available online and googlable.

Tasks for You

1. Load and tidy up the data
   - combine the four `.csv` files into a single data frame
   - variable names, etc.
2. Wrangle your data frame into one that we can analyze by a log-linear model.
   - every match should be coded twice, once from the home team perspective, once from the away team perspective, i.e. both scores from a single match would be the response
3. Build a model and use it to answer the primary and the secondary question above.
   - quantify the home advantage by interpreting the coefficient and providing an intepreted confidence interval
4. Perform basic residual and stability analyses.
   - w.r.t 2019-20 season
   - w.r.t. a larger model
   - promoted/relegated teams discarded?
   - shouldn't we include a draw effect?