#11-F2017

Poisson Regression and Log-linear Model

Summary of the last lecture

- Multinomial regression
  - Nominal logistic regression
  - Ordinal logistic regression

Key terms of this lecture

- Poisson regression and log-linear models
  - Revisit: GLM
  - Interpretation
  - Example

Reading

- Dobson and Bartnett (2008) Chapter 9

Count Data

- The Poisson model can be used
  - to model **count** data
  - to draw inferences about rates (i.e. incidence rates person-years) in cohort studies.
    - British doctors' smoking and coronary death. Table 9.1 on P 169.
- Notation:
  - Y_i =number of events (disease, death, etc.) in cell i.
  - n_i =person-years in cell i.
  - $\mathbf{X}_i$ = covariates in cell i.
  - λ_i = event rate in cell i.

- Distributional assumption:

$$Y_i \sim \text{Poisson}(\mu_i), \text{ where } \mu_i = n_i\lambda_i.$$

where n_i is known, and our interest is λ_i, rather than μ_i.

- The mean μ_i requires careful definition often it needs to be described as a rate. e.g., for occupational injuries, each worker is exposed for the period he or she is a work, so the rate is specified in terms of units "exposure" it may be defined in terms of person-years "at risk"
- The effect of explanatory variables on the response Y is modelled through the mean parameter μ, not on Y directly.

Revisit: GLM

- Recall: Poisson model

$$\begin{aligned} f(y_i|\mu_i) &= \frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!} \\ &= \exp\{y_i\log(\mu_i) - \mu_i - \log(y_i!)\}. \end{aligned}$$

where

- $\theta_i =$ $\log(\mu_i) = \eta_i$
- $b(\theta_i) =$ $\mu_i = \exp(\theta_i)$
- $a_i(\phi =$ $\phi/\omega = 1$
- $v(\mu_i) =$ $b''(\theta_i) = \mu_i$

- Three components for GLM with canonical link.
  1. Poisson distribution in exponential family
  2. linear predictor: $\eta_i = x_i^T\beta$
  3. Link function: $\log(\mu_i) = \eta_i$

Interpretation

- Model:

$$\log(\mu_i) = \log(n_i) + \beta_0 + \beta_1 X_{i1} + \quad + \beta_p X_{ip}.$$

- Interpretation of β_j for $j = 1, \quad , p$.
  - log RR (log relative risk, or log rate ratio) for one unit increase in X_j given that all other Xs are held constant.

$$RR = \frac{\lambda_i | X_{ij} = x + 1}{\lambda_i | X_{ij} = x} = \exp(\beta_j$$

  with holding other covariates constant.

- Programs:
  - R:

```
glm(Y~offset(log(n))+X1+X2  , family=poisson)
```

  - SAS:

```
data.  ;
lpy=log(n);
run;
proc genmod; model Y=X1 X2    /dist=poi offset=lpy;
```

Example

- (Dobson and Barnett) Example 9.2.1. In 1951, all British doctors were sent a brief questionnaire about whether they smoked tobacco. The table shows the numbers of deaths from coronary heart disease among male doctors 10 years after the survey.

Quantify discrete or Categorical variable

Agecat (continuous)

	Age (Categorical)	Smokers		Non-smokers	
	group	Deaths	Person-years	Deaths	Person-yeas
1	35-44	32	52407	2	18790
2	45-54	104	43248	12	10673
3	55-64	206	28612	28	5710
4	65-74	186	12663	28	2585
5	75-84	102	5317	31	1462

Example

- Questions:
  1. Is the death rate higher for smokers than non-smokers?
  2. If so, how much?
  3. Is there a differential age effect?
- Exploratory step:
  - Draw a scatter plot: age vs.,

$$\log\left(\frac{\text{Deaths}_i}{n_i}\right)$$

  - Decide what effects should be included: age^2?
  - Is the differential effect related to age? i.e., interaction between age and smoking status?

Example (cont'd)

- What do you see from the scatter plot? [Table 9-123.sas, Table 9-1.R]
  - The rates increase with age but more steeply than in a straight line – suggesting non-linear effect of age.
  - Death rates appear to be generally higher among smokers than non-smokers, but they do not rise as rapidly with age – suggesting age × smoker interaction.

Let μ = Expected # of deaths = n rate

- An appropriate model is Model rate as $\log(\text{rate}) = \beta_0 + A + S + A*S + A^2$

Then $\log(\mu) = \log(n) + \beta_0 + A + S + A*S + A^2$

or $\log(\text{Expected \# of deaths}) = \log(\text{Person years}) + \beta_0 + \beta_1 S + \beta_2 A + \beta_3 A^2 + \beta_4 S*A$

Example (cont'd)

- The statistics for the goodness-of-fit tests:
  - Compare to the saturated model, use GOF test,

$$X^2 = 1.550, \quad (SAS)$$

$$D = 1.6354. \quad (R \text{ and } SAS)$$

with $df = N - p = 10 - 5 = 5$, the test is not significant. It suggests that the model is a good fit to the data.

Pearson Residual: $r_i = \frac{O_i - E_i}{\sqrt{E_i}}$

standardized Pearson Residual: $r_i = \frac{O_i - E_i}{\sqrt{E_i(1-h_i)}}$, $H = X(X^TX)^{-1}X^T$

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$D = 2\sum_i [O_i \log(O_i/E_i)] \quad (\text{Model has an intercept } \beta_0)$$

or $= 2\sum_i [O_i \log(O_i/E_i) - (O_i - E_i)]$ (Model has no intercept)

- Compare to the minimal model, use overall test with
  $H_0 \quad \beta_1 = \beta_2 = \beta_3 = \beta_4$,

$$\begin{aligned} C &= 2[l(b) - l(b_{min})] \\ &= 2[l(b_{max}) - l(b_{min}) - \{l(b_{max}) - l(b)\}] \\ &= \text{Null Dev} - \text{Res. Dev.} \\ &= 935.0673 - 1.6354 \\ &= 933.43 \end{aligned}$$

  with $df = 5 - 1 = 4$, the test is highly significant. It suggests that the covariates have important effects.

- Pseudo

$$R^2 = \frac{l(b_{min}) - l(b)}{l(b_{min})} = \frac{(-495.067) - (-28.352)}{-495.067} = 94\%.$$

  It also suggests a good fit.

- Interpret the estimates of the parameters.
  - The estimates are $\hat{\beta}_{smoke} = 1.441$, $\hat{\beta}_{smkage} = -0.308$. For example, the risk of coronary deaths was about $e^{(1.441-0.308)} = 3.10$ times higher for smokers than non-smokers when age group 1 was considered. However the effect is attenuated as age increases.

Since there is an Age * Smoke interaction, the interpretations is made at a specific age level, here, only Consider Age = 1 The comparison result varies with Age

Poisson Regression versus Log-linear Models

- The effect of explanatory variables on the Poisson response Y is modelled through the parameter mean μ. There are two situations.
  - Case 1. In this case, the events relate to varying amounts of "exposure" The other explanatory variables (in addition to "exposure") may be continuous or categorical. Here "exposure" is not constant and is relevant to the model. **Poisson regression** is used in this case.
  - Case 2: "exposure" is constant (and not relevant to the model) and the explanatory variables are usually categorical. The data are summarized in a cross-classified or so called contingency table, called "contingency table" The explanatory variables are used to define the table. The response variable is the frequency or count in each cell of the table. **Log-linear models** are used in this case.

Contingency Tables and Log-linear Models

- Two-way contingency table. It is important to consider how the design at the study may determine constraints on the data.
- Example: cross-section study of *malignant melanoma*, see Tables 9.4, 9.5, 9.9 and 9.10. For a sample of $n = 400$ patients, the site of the tumor and its historical type were recorded. [See `Table9-45 sas`, `Table9-9-10 sas`, `Table9_10 R`]
  - It is a 4×3 contingency table.
  - Question: whether there is any association between tumor type and site.
  - According to the row and column percents, it appears that Hutchinson's melanotic freckle is more common on the head and neck but there is little evidence of association between other tumor types and sites. See column percentage in Table 9.5.

Tumor type	Head & Neck	Trunk	Extremities	Total
Hutchinson's melanotic freckle	32.4	1.9	4.4	8.5
Superficial spreading melanoma	23.5	50.9	50.9	46.25
Nodular	27.9	31.1	32.3	31.25
Indeterminate	16.2	16.0	12.4	14.00
Total	100	100	100	100

Log-linear Models for Two-way Table

- Consider a $I \times J$ contingency table
  - (i, j) cell frequency: y_{ij} $i = 1, \quad , I$ $j = 1, \quad , J$
- Consider how the design of the study may determine constraints to the model.

In this example, there are $J = 4$ rows and $K = 3$ Columns and the constrait is $\sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} = n$, where $n = 400$ is fixed by design

Case I: Cross-Section Study

← It is a type of observation Study where data are collected at a specific time point.
Cross-sectional studies differ from case-control studies and longitudial studies.

- In a cross-section study, the constraint is

$$\sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij} = n,$$

where $Y_{ij} \sim$ Poisson($\mu_{ij} = E(Y_{ij}$ independently, then, the sum has the Poisson distribution and

$$E(n) = \tilde{\mu} = \sum_i \sum_j \mu_{ij}$$

Here $\tilde{\mu}$ is different from the intercept μ used late.

- The joint probability distribution of the Y_{ij}'s, conditional on n, is multinomial distribution

$$f(\mathbf{Y}|n) = n! \prod_{i=1}^{I} \prod_{j=1}^{J} \theta_{ij}^{y_{ij}} / y_{ij}!,$$

where $\theta_{ij} = \mu_{ij}$ μ, it is the probability of an observation in (i,j)th cell. Then $E(Y_{ij} = \mu_{ij} = n\theta_{ij}$ hence,

$$\log \mu_{ij} = \log n + \log \theta_{ij}$$

It is a log-linear model. It is like Poisson regression model except $\log n$ is the same for all Y_{ij}s.

- Consider hypothesis that the row and column variables are independent, so that

$$\theta_{ij} = \theta_{i\cdot}\theta_{\cdot j},$$

where $\theta_{i\cdot}$ and $\theta_{\cdot j}$ are marginal probabilities with $\sum_i \theta_{i\cdot} = 1$ and $\sum_j \theta_{\cdot j} = 1$, then

$$\log \mu_{ij} = \log n + \log \theta_{i\cdot} + \log \theta_{\cdot j}$$

- In general, the saturated log-linear model for an $I \times J$ table is

$$\log(\mu_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

for two categorical variables X and Y This notation is convenient for tables of higher dimensions. It is similar to the two-way ANOVA for a continuous response Y Notice the equivalence of

$$X \perp Y \iff H_0 \quad \lambda_{ij}^{XY} = 0.$$

The corresponding independence log-linear model is given by

$$\log(\mu_{ij} = \mu + \lambda_i^X + \lambda_j^Y$$

- Since the term $\log n$ has to be in all models, the minimal mode is

$$\log(\mu_{ij}) = \mu,$$

we don't see offset here.

- Revisit Example: cross-section study of *malignant melanoma*, see Tables 9.4, 9.9 and 9.10 [Table9-9-10 sas]. The saturated model with 12 parameters fits the 12 data points exactly. The deviance for the additive model is $X^2 = 65.8 \sim \chi^2(6)$ (SAS and R) and $D = 51.79 \sim \chi^2(6)$ (R). We reject H_0 and conclude that there is an association between tumor type and site, or the two variables are not independent. Equivalent to test Tumor Type * Site interaction significant or not

- Note: SAS has different values of log-likelihood, but differences of these values are the same. How to calculate different deviances D ?

Residual deviance: $D_{res} = 2\{\ell(b_{max}) - \ell(b)\} = 2\{-29.556 - (-55.453)\} = 51.794$,

GOF test → Since the saturated model (including interaction) is the maximal model, this result also implies that the main effect (additive) model is not a good fit and an interaction effect may exist.

Overall test; H_0: $\lambda_i^X = 0$ and $\lambda_j^Y = 0$

Overall test → SAS code: Table9-4a.sas does overall test to compare current model (the main effects model) with the minimal model (intercept only).

$D_{overall} = 2\{\ell(b) - \ell(b_{min})\} = 2[1124.3272 - 1002.6232]$ (SAS)

~~$= 2[(-55.453) - (-177.16)]$, (textbook)~~

$= 2 \times (121.704) = 243.408$

Another way to compute $D_{overall}$ = Null Dev. − Residual Dev.

$= 295.208 - 51.794 = 243.414 \approx 243.408$ (Some Rounding error).

where Null Dev. $= 2\{\ell(b_{max}) - \ell(b_{min})\} = 2\{-29.556 - (-177.16)\} = 295.208$.

Case II: Prospective Study

The overall test is significant, indicating tumor type and site are important predictors.

- In this case, the row totals are fixed, the constraint is

$$\sum_{j=1}^{J} Y_{ij} = y_{i\cdot}$$

and $y_{i\cdot}$ is fixed.

- The joint probability distribution of each row, conditional on $y_{i\cdot}$ is multinomial distribution

$$f(y_{i1}, \quad , y_{iJ}|y_{i\cdot}) = y_{i\cdot}! \prod_{j=1}^{J} \theta_{ij}^{y_{ij}} / y_{ij}!,$$

where $\sum_{j=1}^{J} \theta_{ij} = 1$. So the joint distribution for all the cells in rows is the

product multinomial distribution

$$f(\mathbf{Y}|y_{1\cdot}, \quad , y_{I\cdot} = \prod_{i=1}^{I} y_{i\cdot}! \prod_{j=1}^{J} \theta_{ij}^{y_{ij}}/y_{ij}!.$$

In this case, $E(Y_{ij}) = y_{i\cdot}\theta_{ij}$ hence,

$$\log E(Y_{ij}) = \log \mu_{ij} = \log y_{i\cdot} + \log \theta_{ij}$$

This is called the product multinomial model.

- Consider hypothesis that the response pattern is the same for all I groups, we have

$$\theta_{ij} = \theta_{\cdot j}, \quad j = 1, \quad , J.$$

- The hypothesis of homogeneity of the response distributions can be tested by comparing the model

$$\log(\mu_{ij} \quad = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

corresponding to $E(Y_{ij}) = y_{i\cdot}\theta_{ij}$ (different from the previous model where

overall total n is fixed) and the model

$$\log(\mu_{ij} \quad = \mu + \lambda_i^X + \lambda_j^Y$$

corresponding to $E(Y_{ij}) = y_{i\cdot}\theta_{\cdot j}$ It is equivalent to the test for H_0: all $\lambda_{ij}^{XY} = 0$.

- The minimal model is

$$\log(\mu_{ij}) = \mu + \lambda_i^X,$$

this is an equal probability model, because the row totals, corresponding to the subject i, are fixed by the design of the study.

- Parameter constraints, sum-to-zero:

$$\sum_i \lambda_i^X = 0, \ \sum_j \lambda_j^Y = 0, \ \sum_i \lambda_{ij}^{XY} = 0, \ \sum_j \lambda_{ij}^{XY} = 0.$$

Example for Case II: Prospective Study

- Example: Randomized controlled trial of influenza vaccine compare two populations. Table 9.6 on P 174 [Table9-6 sas].

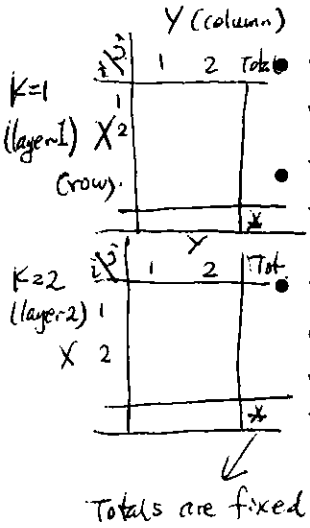
	Response			
	Small	Moderate	Large	Total
Placebo	25	8	5	38
Vaccine	6	18	11	35

- Patients were randomly allocated to two treatment groups. The responses were titre levels of *hemagglutinin inhibiting antibody* categorized as "small" "moderate" "large" The row totals are fixed.

- Question: If the distribution (pattern) is the same for each treatment group?

We can test H_0 Trmt * Response $=0$ or $H_0: \lambda_{ij}^{TR}=0$
in the full model $\log(\mu_{ij}) = \mu + \text{Trmt} + \text{Response} + \text{Trmt} * \text{Response}$
The reduced model is: $\log(\mu_{ij}) = \mu + \text{Trmt} + \text{Response}$
We compare the two models. $N = 2\times 3 = 6$
$df_{res.}$ in Deviance $= 2$

Three-Way Contingency Table



- Consider a three-dimensional table containing the crossclassification of variables X Y and Z
- The cross-sections at different levels (or layers) of Z are called partial tables. There are two cases.
- Case I: The i row totals at each layer k and a level of X are fixed. (e.g. in a case-control retrospective study). Suppose that in a three-dimensional table with I rows, J columns and K layers, $\sum_j y_{ijk} = y_{i\cdot k}$ is fixed, $\sum_j \theta_{ijk} = 1$. Then, the joint probability for the Y_{ijk}'s is

$$f(\mathbf{Y}|y_{i\cdot k}, i = 1, \quad , I, k = 1, \quad , K) = \prod_{i=1}^{I}\prod_{k=1}^{K} y_{i\cdot k}! \prod_{j=1}^{J} \theta_{ijk}^{y_{ijk}}/y_{ijk}!.$$

For each combination of (i, k),

$$\mu_{ijk} = E(Y_{ijk}) = y_{i\cdot k}\theta_{ijk},$$

and

$$\log(\mu_{ijk}) = \log y_{i\cdot k} + \log \theta_{ijk}.$$

- Case II: The partial table (or layer) totals are fixed.. Then, the joint probability for the Y_{ijk}'s is

$$f(\mathbf{Y}|y_{..k}, k = 1, \quad , K) = \prod_{k=1}^{K} y_{..k}! \prod_{i=1}^{I} \prod_{j=1}^{J} \theta_{ijk}^{y_{ijk}} / y_{ijk}!$$

with $\sum_i \sum_j \theta_{ijk} = 1$ for $k = 1, \quad , K$ Then,

$$\mu_{ijk} = E(Y_{ijk}) = y_{..k}\theta_{ijk},$$

and

$$\log(\mu_{ijk}) = \log y_{..k} + \log \theta_{ijk}.$$

Z(GD)

Example for Three-Way Contingency Table

gastric (k=1)

Aspirin use (X)

	Non-user	user	
Control	62	6	68
ulcer case	39	25	64

Y

- Consider a case control study of gastric and duodenal ulcers and aspirin use. Table 9.7 on P 174. Ulcer patients were compared to control patients. Ulcer patients were classified according to the site of the ulcer: gastric or duodenal. [See Table9-78 sas, Table9_11 R] Row fixed ⇒ CC × GD fixed
- This is a $2 \times 2 \times 2$ contingency table. Questions are Test

duodenal (k=2)

Aspirin use (X)

	Non-user	user	
control	53	8	61
ulcer case	49	8	57

Y

1. Is control or case associated with aspirin use? AP × CC
2. Is any association with aspirin use the same for both ulcer sites? AP × GD

- **Exploratory analysis**: examine row percents. It appears that aspirin use is more common among ulcer patients than among controls for gastric ulcer but not for duodenal ulcer It suggests that aspirin use may be a risk factor for gastric ulcer but not for duodenal ulcer

Coding: case-control status:

$$CC = \begin{cases} 1 & \text{case} \\ 0 & \text{control} \end{cases}$$

Aspirin use: $AP = \begin{cases} 1 & \text{user} \\ 0 & \text{non-user} \end{cases}$ (ulcer site)

$$GD = \begin{cases} 1 & \text{duodenal} \\ 0 & \text{gastric} \end{cases}$$

- **Statistical analysis**: There are two approached:
  1. Analyze 2×2 tables for gastric ulcer and duodenal ulcer separately.
  2. Conduct a full data analysis.
- We use the second approach. Since the marginal row totals are fixed, i.e., the rows in the table are classified by these two factors, the minimal model is

$$\log \mu_{ijk} = \mu + CC + GD + CC \times GD,$$

This model does not contain the effect of AP.
where $CC = case - control$, $GD = gastric - duodenal$, $AP = aspirin$.
The saturated model for the 2x2 table with 4 totals 68, 64, 61, 57, these values can be exactly fitted, corresponding to the fixed marginal totals.

- Models of interest are:

$$\begin{aligned}
\log \mu_{ijk} &= \mu + CC + GD + CC \times GD + AP, \\
\log \mu_{ijk} &= \mu + CC + GD + CC \times GD + AP + AP \times CC, \\
\log \mu_{ijk} &= \mu + CC + GD + CC \times GD + AP + AP \times CC + AP \times GD.
\end{aligned}$$

- See Appendix: R and SAS code for the examples in this lecture note.

Deviances.

Model	df $(N-p)$	D
GD + CC + GD x CC	4 (8-4)	126.708
GD + CC + GD x CC + AP	3 (8-4-1)	21.789
GD + CC + GD x CC + AP + AP x CC	2 (8-4-1-1)	10.538
GD + CC + GD x CC + AP + AP x CC + AP x GD	1 (8-4-3)	6.283

$\Leftarrow D = 2[\ell(b_{max}) - \ell(b)]$

The above model are equivalent to the logistic regression models by using AP as a binomial response variable. See problem 4.6 (c) in the textbook

- Results are in Table 9.11, see `Table9-11-12 sas`.
  1. The comparison of aspirin use between cases and controls. This is to test H_0: $\lambda^{AP \times CC} = 0$. Use the second and third row entries in the table,

$$\Delta D = 21.789 - 10.538 = 11.251.$$

  The $df = 1$, the test is significant, indicating that aspirin is a risk factor for ulcers.
  2. To compare two sites, we test H_0: $\lambda^{AP \times GD} = 0$, using

$$\Delta D = 10.538 - 6.283 = 4.255.$$

  The p-value=0.04, showing a weak evidence for the association.
  3. The goodness-of-fit statistics are $X^2 = 6.49$, $D = 6.28$ with $df = N - p = 8 - 7 = 1$, the tests are significant, indicating a non particularly good fit. Note: the saturated model is

$$\log \mu_{ijk} = GD + CC + GD \times CC + AP + AP \times CC + AP \times GD + AP \times CC \times GD.$$