# What to do today (Mar 8)?

## 4. Analysis with Count Response (Chp 4)

*4.1 Poisson Model for Counts (Chp 4.1)*

*4.2 Poisson Regression Analysis (Chp 4.2)*

**4.3 Additional Topics on Count Responses (Chp 4.3-4)**

**4.3.1 Poisson rate regression**
**4.3.2 Overdispersion and zero inflation**
**4.3.3 Generalized linear models II**

## A Review for Midterm 2

# 4.3.1A Poisson rate regression: Introduction

**A "rate" variable is often of interest $Y/t$:** e.g.

- number of computer crashes in some area
- number of arrivals at an airport over some time periods

**When the baseline measure of the "exposure" varies over observations?**

- The measure needs to be incorporated into the analysis.
- One way to do this is to model $Y/t$ instead of just $Y$: $Y =$count of events; $t =$measure of opportunity for events.

**Poisson Rate Regression Model.** Consider the response $Y|t, x_1, \ldots, x_K \sim Poisson\big(\mu(x_1, \ldots, x_K; t)\big)$ and assume

$$\log[\mu(x_1, \ldots, x_K; t)] = \log(t) + \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K,$$

equivalently to $E(Y/t|\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K)$.

# 4.3.1B Poisson rate regression: Example

**Example.** Number of Credit Cards vs Income
(*https://onlinecourses.science.psu.edu/stat504/node/170*)

| Income[a] | Number Cases | Credit Cards |
|---|---|---|
| 24 | 1 | 0 |
| 27 | 1 | 0 |
| 28 | 5 | 2 |
| 29 | 3 | 0 |
| ... | ... | ... |
| 120 | 6 | 6 |
| 130 | 1 | 1 |

[a] in millions of lira
(the currency in Italy before euro)

Consider $\log(\mu/t) = \beta_0 + \beta_1 \text{income}$:
$\mu = E(\text{number of credit cards})$, $t = $ number of cases.

# 4.3.1B Poisson rate regression: Example

- The fitted model:

$$\log(\hat{\mu}/t) = -2.3866 + 0.0208 \times \text{income}$$

  where $\log(t) = \log(\text{num cases})$.

- Questions can be answered by the analysis:
  - What is the estimated average rate of incidence, i.e. the usage of credit cards given the income?
  - Is income a significant predictor? Does the overall model fit?

e.g. with income$= 65$,

$$\log(\hat{\mu}/t) = -2.3866 + 0.0208 \times 65 \Longrightarrow \log(\hat{\mu}) = -2.3866 + 0.0208 \times 65 + \log(t)$$

for a group of six people with income 65:

$$\log(\hat{\mu}) = -2.3866 + 0.0208 \times 65 + \log(6) \Longrightarrow \hat{\mu} = 2.126$$

## 4.3.2 Overdispersion and zero inflation

Recall that $Y$ has a **Poisson distribution**: $Y \sim Poisson(\mu)$, if its pmf is

$$P(Y = y) = p(y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots$$

A characteristic of the distribution is that its mean is equal to its variance: $E(Y) = \mu$; $Var(Y) = \mu$

In many situations, the variance of the observed counts is greater than the mean $\Longrightarrow$ **overdispersion** and the Poisson model is not appropriate:

- $Y$ has an "overdispersed" distribution if $Var(Y) > E(Y)$: e.g. "overdispered Poisson" distribution.
- how to deal with it? e.g. quasi-likelihood estimation (Chp 6)

## 4.3.2 Overdispersion and zero inflation

Another common problem with Poisson regression is excess zeros

... ...

- infection diseases: a population includes "susceptible" and "immune" individuals
- products with high quality

$\Longrightarrow$ the zero-inflated Poisson (ZIP) model (Lambert, 1992):

$$\begin{array}{rcccc} Y & = & 0 & \text{with prob} & \pi(z) \\ Y & \sim & Poisson(\mu(x)) & \text{with prob} & 1-\pi(z) \end{array}$$

with the logistic model $logit(\pi(z)) = \gamma_0 + \gamma_1 z$ and the loglinear model $\log(\mu(x)) = \beta_0 + \beta_1 x$.

**Remarks:**

- In package of "pscl" in R, use the function of *zeroinfl*().
- ZIP is a *mixture* distribution.

# 4.3.3 Generalized linear models II

**Generalized Linear Models (GLM)**: a unified framework for many regression analyses.

- including OLM, Logit, Loglinear models as special cases
- including other regression models.
  - To study $Y \leftarrow X, Z$? with binary response $Y = 1$, or 0: $P(Y = 1|X = x, Z = z) = \pi(x, z)$, $Y \sim Bernoulli\big(\pi(x, z)\big)$
    - *R: glmout< −glm(Y∼X*Z, family=binomial(link="probit"))*
  - To study $Y \leftarrow X, Z$? with count response $Y$ and two explanatory variables: $E(Y|X, Z) = \mu(X, Z)$ with $\log(\mu(X, Z)) = \alpha + \beta X + \gamma Z + \eta XZ$ but Poisson assumption is not appropriate
    - *R: glmout< −glm(Y∼X*Z, family=quasipoisson(link="log"))*

# 4.3.3 Generalized linear models II: GLIM Components

Recall how to conduct the analysis with *R:*
*glm(formula, family=xxx (link="xxx"))* $\Longrightarrow$

- **Random Component.** response r.v. $Y$ with $\mu(x_1, \ldots, x_k) = E(Y|x_1, \ldots, x_k)$ to be examined

- **Systematic Component.** $\alpha + \beta_1 x_1 + \ldots + \beta_K x_K$
Some $x_k$ can be based on others: e.g. $x_3 = x_1 x_2$.

- **Link Function.** $g(\mu) = \alpha + \beta_1 x_1 + \ldots + \beta_K x_K$
The link function $g(\cdot)$ links the *random componet* through its mean and the *systematic component*. **More on GLM later**

Quiz 2. [10 points] A clinical trial observed 41 successes from Treatment A group with size 60, and 42 failures from Treatment B group with size 80.

**Q1.**[4 points] Present the data using Table 1.

**Q2.**[6 points] Suggest two regression models for an analysis to establish how an individual's outcome is associated with his/her treatment.

**Table 1.**

| treatment | outcome ($Y$) | |
|---|---|---|
| ($X$) | Failure | Success |
| A | | |
| B | | |

## What if there's a third variable, say, age ($Z$) with three categories?

**Table 2.**

| age ($Z$) | treatment ($X$) | outcome ($Y$) Failure | Success |
|---|---|---|---|
| $< 25$ | A | 8 | 14 |
| | B | 7 | 13 |
| 25-55 | A | 5 | 19 |
| | B | 5 | 17 |
| $> 55$ | A | 6 | 8 |
| | B | 30 | 8 |

- Q3.1: 3-way contingency table: $2 \times 2 \times 3$
  - margianl OR of success for treatments A and B? answer in Q1 and Q2.2
  - conditional OR of success for treatments A and B with an age group?
  - MH test, Breslow-Day test? the common OR?
- Q3.2: logistic regression: $Y \sim X, Z$, $Y \sim XZ$?
- Q3.3: loglinear regression: (X,Y,Z), (XY,YZ,XZ), and (XYZ)?
  - parameter interpretation? fitted models?

# What will we do next?

1. *Introduction and Preparation*
2. *Analysis with Binary Variables (Chp 1-2)*
3. *Analysis with Multicategory Variables (Chp 3)*
4. *Analsyis with Count Response (Chp 4)*

5. **Model Selection and Evaluation (Chp 5)**
   - **5.1 Variable selection (Chp 5.1)**
   - **5.2 Tools to asses model fit (Chp 5.2-3)**
   - **5.3 Examples**

## Midterm 2. 10:30-11:20 Thu March 15

- To cover Chp1-4, including the supplementary material on multi-way contingency tables.

6. *Additional Topics (Chp 6)*

*Final Exam: 15:30-18:30 Monday April 23*