STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2019

Dr. Shivon Sue-Chee

UNIVERSITY OF TORONTO

March 25-29, 2019

Week 11- Summary of Case Study VI

Framingham Heart Study
A Project of the National Heart, Lung, and Blood Institute and Boston University

CVD Status: P, A
Chol. level: H, L

	Diff. in prop	LRT	Log-linear
Assume	Row totals fixed	Overall total fixed	Totals are random
Dist. of Y	Binomial	Multinomial	Poisson
H_0	$\pi_1 = \pi_2$	$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$	Additive model
Test Stat.	Z	$\chi^2_{(I-1)(J-1)}$	$\chi^2_{(I-1)(J-1)}$

$Z^2 \sim \chi^2_1$ I, J > 2

A Three-way Contingency Table

Case Study VII Data:

- 1992 survey of high-school seniors in Ohio
- Table of counts of seniors who used alcohol, cigarettes and marijuana.

111

1

		Marijuana use	
Alcohol use	Cigarette use	Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

2

222

Q: Are alcohol (A), cigarettes (C) and marijuana (M) use associated?

Forms of independence in $I \times J \times K$ Tables

Independence	π_{ijk}	Short form	
Mutually indep.	$\pi_{ijk} = \pi_{i\cdot\cdot}\pi_{\cdot j\cdot}\pi_{\cdot\cdot k}$	(X,Y,Z)	Complete
Jointly indep.	$\pi_{ijk} = \pi_{ij\cdot}\pi_{\cdot\cdot k}$	(XY,Z)	Block
Conditionally indep.	$\pi_{ijk} = \pi_{i\cdot k}\pi_{\cdot jk}/\pi_{\cdot\cdot k}$	(XZ,YZ)	Partial
Uniform assoc.	$\pi_{ijk} = \pi_{ij\cdot}\pi_{i\cdot k}\pi_{\cdot jk}$	(XZ,YZ, XY)	Homo
Saturated	π_{ijk}	XYZ	

XZ, Y
YZ, X

Three-way Tables

- Learning Objectives
  - Write out the models used and the assumptions for inference
  - Carry out the inference procedures completely
  - Interpret the respective R outputs

Model 1: Complete Independence

- $P(ACM) = P(A)P(C)P(M)$; Alcohol, cigarette and marijuana use are mutually independent
- Hypotheses:

$$H_0 : \pi_{ijk} = \pi_{i\cdot\cdot}\pi_{\cdot j\cdot}\pi_{\cdot\cdot k} \text{ for all } i, j, k$$
$$H_a : \pi_{ijk} \neq \pi_{i\cdot\cdot}\pi_{\cdot j\cdot}\pi_{\cdot\cdot k}$$

- Short form: (A,C,M) -all 3 main effects only
- $I = J = K = 2$

$$\boxed{\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M}$$

Additive

where $\mathbf{I} : \{1 = Yes, 0 = No\}$

Model 1: Complete Independence

- In general, we have the constraint $n = \sum_i \sum_j \sum_k y_{ijk}$ or $\sum_i \sum_j \sum_k \hat{\pi}_{ijk} = 1$. Then by ML estimation,

$$\sum_i \sum_j \sum_k \hat{\mu}_{ijk} = n = \sum_i \sum_j \sum_k y_{ijk}$$

$$\Longrightarrow \hat{\pi}_{ijk} = \frac{y_{ijk}}{n} \text{ or } \boxed{\hat{\mu}_{ijk} = y_{ijk}}$$

MLE

- For complete independence model, using an additional $(I-1)+(J-1)+(K-1)$ constraints

$$\begin{aligned} \hat{\mu}_{ijk} &= n\hat{\pi}_{ijk} = n\hat{\pi}_{i\cdot\cdot}\hat{\pi}_{\cdot j\cdot}\hat{\pi}_{\cdot\cdot k} \\ &= n\frac{y_{i\cdot\cdot}}{n}\frac{y_{\cdot j\cdot}}{n}\frac{y_{\cdot\cdot k}}{n} \end{aligned}$$

H_0

Model Class 2: Block Independence

- $P(AC|M) = P(AC)$; Joint probability of alcohol and cigarette use is independent of marijuana use; Alcohol and cigarette use are associated
- Hypotheses:

$$H_0 : \pi_{ijk} = \pi_{ij\cdot}\pi_{\cdot\cdot k}$$
$$H_a : \pi_{ijk} \neq \pi_{ij\cdot}\pi_{\cdot\cdot k}$$

- Short form: (AC,M) - all 3 main effects and 1 interaction

$$\boxed{\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M + \beta_4 \mathbf{I}_{AC}}$$

where $\mathbf{I}_{AC} = \mathbf{I}_A * \mathbf{I}_C$

- Others in this class: (AM, C), (CM, A)

Model 2: Block Independence

- By ML estimation, for block independence model

$$\begin{aligned}\hat{\mu}_{ijk} = n\hat{\pi}_{ijk} &= n\hat{\pi}_{ij\cdot}\hat{\pi}_{\cdot\cdot k} \\ &= n\frac{y_{ij\cdot}}{n}\frac{y_{\cdot\cdot k}}{n}\end{aligned}$$

Model Class 3: Partial Independence

- $P(AC|M) = P(A|M)P(C|M)$; Alcohol and cigarette use are conditionally independent given marijuana use; Alcohol and marijuana use are associated, and cigarette and marijuana use are associated
- Hypotheses:

$$H_0 : \pi_{ijk} = \pi_{i \cdot k}\pi_{\cdot jk}/\pi_{\cdot\cdot k}$$
$$H_a : \pi_{ijk} \neq \pi_{i \cdot k}\pi_{\cdot jk}/\pi_{\cdot\cdot k}$$

- Short form: (AM,CM) - all 3 main effects and 2 interactions

$$\boxed{\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M + \beta_4 \mathbf{I}_{AM} + \beta_5 \mathbf{I}_{CM}}$$

- Others in this class: (AC, CM), (AC, AM)

C A

Model 3: Partial Independence

- We have $P(AC|M) = P(A|M)P(C|M)$.

$$\Longrightarrow \quad \frac{\pi_{ijk}}{\pi_{\cdot\cdot k}} = \frac{\pi_{\cdot jk}}{\pi_{\cdot\cdot k}} \frac{\pi_{i\cdot k}}{\pi_{\cdot\cdot k}}$$

$$\text{or} \quad \pi_{ijk} = \frac{\pi_{\cdot jk}\pi_{i\cdot k}}{\pi_{\cdot\cdot k}}$$

- Then by ML estimation

$$\begin{aligned} \hat{\mu}_{ijk} = n\hat{\pi}_{ijk} &= n\frac{\hat{\pi}_{\cdot jk}\hat{\pi}_{i\cdot k}}{\pi_{\cdot\cdot k}} \\ &= n\frac{(y_{\cdot jk}/n)(y_{i\cdot k}/n)}{(y_{\cdot\cdot k}/n)} \\ &= \frac{y_{\cdot jk}y_{i\cdot k}}{y_{\cdot\cdot k}} \end{aligned}$$

Model 4: Uniform association

- There is an association among all pairs
- For all levels of the 3rd variable, the association between the pair is the same
- Short form: (AM,AC,CM) - all 3 main effects and 3 two-way interactions but no three-way interaction

$$\boxed{\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M + \beta_4 \mathbf{I}_{AM} + \beta_5 \mathbf{I}_{AC} + \beta_6 \mathbf{I}_{CM}}$$

- Solutions for π_{ijk} (μ_{ijk}) are found numerically with no simple expression in terms of y_{ijk}'s
- No simple interpretation ito. independence structure

Saturated Model

- Total number of parameters:

$$1 + \underbrace{3}_{\text{1-way}} + \underbrace{3}_{\text{2-way}} + \underbrace{1}_{\text{3-way}} = 8$$

- Total number of observed counts:

$$\begin{aligned} 1 &+ (I-1) + (J-1) + (K-1) \\ &+ (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1) \\ &+ (I-1)(J-1)(K-1) = IJK = 2 * 2 * 2 = 8 \end{aligned}$$

$$\begin{aligned} \log(\mu_{ijk}) = \beta_0 &+ \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M \\ &+ \beta_4 \mathbf{I}_{AM} + \beta_5 \mathbf{I}_{AC} + \beta_6 \mathbf{I}_{CM} + \boxed{\beta_7 \mathbf{I}_{ACM}} \end{aligned}$$

- Saturated model always fits the data perfectly

On the Saturated Model

$$\log(\mu_{ijk}) = \beta_0 + \beta_1 \mathbf{I}_A + \beta_2 \mathbf{I}_C + \beta_3 \mathbf{I}_M + \beta_4 \mathbf{I}_{AM} + \beta_5 \mathbf{I}_{AC} + \beta_6 \mathbf{I}_{CM} + \beta_7 \mathbf{I}_{ACM}$$

- Total # of parameters=Total # of observed counts
- Has a separate parameter for each observation
- Always gives a perfect fit
- Explains all the variation by its systematic component
- Sounds good but not a helpful model
- Does not smooth the data or is not parsimonious
- Serves as a baseline for checking model fit

Results from R output

G-o-F Test

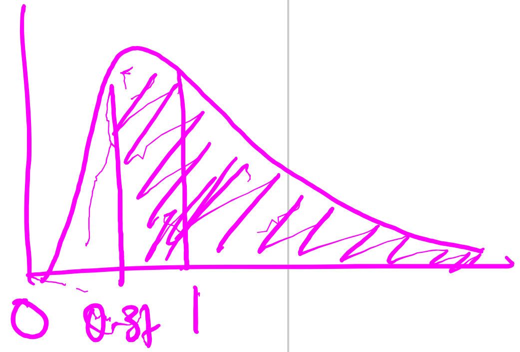
H_0: Fitted
H_a: Saturated

Model	df	G^2=Deviance	p-value
(A,C,M)	4	1286.02	$<$ 0.0001
(AC,M)	3	843.83	$<$ 0.0001
(AM, C)	3	939.56	$<$0.0001
(A,CM)	3	534.21	$<$0.0001
(AC,AM)	2	497.37	$<$0.0001
(AC,CM)	2	92.02	$<$0.0001
(AM,CM)	2	187.75	$<$0.0001
(AC,AM,CM)	1	0.37	0.5408
(ACM)	0	0.00	-

Complete
Block
Partial A C M
Uniform
Saturated

AIC BIC
1343.06
558.41
63.43
Adequate
$P(\chi^2_1 > 0.37) = 0.5408$

The simplest model that fits the data adequately is the "Uniform Association" model (AC,AM,CM).



Exercise: Fitted values and Interpretations

Refer to RM.

Q: Complete the fitted equation and prove the fitted values ~~above~~ below

Fitted equation:

$$\log(\hat{\mu}_{ijk}) = 6.81 - 5.53\mathbf{I}_{A_2} - 3.02\,1_{C_2} - 0.52\,1_{M_2} + 2.98\,1_{A_2} * 1_{M_2} + 2.05\,1_{A_2} * 1_{C_2} + 2.85\,1_{C_2} * 1_{M_2}$$

Some fitted values, $\hat{\mu}_{ijk}$:

$$\hat{\mu}_{ijk} = e^{\log(\hat{\mu}_{ijk})} = \exp\{\ \}$$

1 - Yes

2 - No

			Log-linear models	
A use	C use	M use	(AC, AM, CM,)	(ACM)
Yes	Yes	Yes	910.4	911
Yes	Yes	No	538.6	538
Yes	No	Yes	44.6	44
Yes	No	No		

0 — y_{111}

$$\hat{\mu}_{111} = e^{6.81} \quad (1_{A_2} = 0,\ 1_{C_2} = 0,\ 1_{M_2} = 0)$$

$$\hat{\mu}_{112} = e^{6.81 - 0.52}$$

$$1_{A_2} = 0 = 1_{C_2},\ 1_{M_2} = 1$$

$$1_{A_2} = \begin{cases} 1 & \text{if Alcohol was NOT used} \\ 0 & \text{if '' was used} \end{cases}$$

Fitted values and Interpretations

- **Use estimates of β's to calculate odds.**
  - Eg, the odds of marijuana use for alcohol and cigarette use at (i, j) are:

$$\boxed{\frac{\hat{\pi}_{ij1}}{\hat{\pi}_{ij2}} = \frac{\hat{\mu}_{ij1}}{\hat{\mu}_{ij2}}}$$

    - *Example 1*: For students who use alcohol and cigarettes, the estimated odds of using marijuana are:

$(A = 1 = C = M) \longrightarrow$ $\hat{\mu}_{111} = 910.38$

$(A = 1 = C, M = 2) \longrightarrow$ $\hat{\mu}_{112} = 538.61$

$\frac{\hat{\mu}_{111}}{\hat{\mu}_{112}} = \frac{910.38}{538.61} = 1.69 > 1$

    - *Example 2*: For students who use neither alcohol nor cigarettes, the estimated odds of using marijuana are:

$(A = 2 = C, M = 1) \longrightarrow$ $\hat{\mu}_{221} = 1.38$

$(A = 2 = C = M) \longrightarrow$ $\hat{\mu}_{222} = 279.62$

$\frac{\hat{\mu}_{221}}{\hat{\mu}_{222}} = \frac{1.38}{279.62} = 0.0054$

Inference for Log-linear models

- Q: What procedures do we use to:
  - Estimate parameters in log-linear models?
    - A: Maximum likelihood estimation
  - Carry out inference (significance tests and C.I.s)?
    - A: Wald tests and C.I.s, and LRT

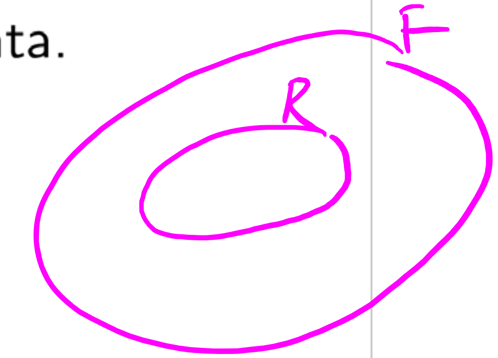
What are the conditions for inference to be valid?

1. Independent quantities being counted
2. Large enough sample sizes for MLE asymptotic tests to hold.
   - RULE-OF-THUMB: (Most) $\widehat{\mu}_{ijk} \geq 5$ for all i, j, k.
3. Cross-classified counts follow a Poisson distribution, i.e., $\boxed{\text{Var}(y_{ijk}) = \mu_{ijk}}$.
   - If not, then the deviance is very large ("extra-Poisson" variation).
   - Deviance/df should be about 1.
4. Correct form of the model / Model fits the data.
   - $\log(E(Y))$ is linear in the β's
   - All relevant variables included.
   - No outliers
   - Agreement of predicted and observed counts
   - Check deviance goodness-of-fit test

What is the frame of a Likelihood Ratio Test?

- **Idea**: Compare likelihood of data under FULL (F) model, $\mathcal{L}_F$ to likelihood under REDUCED (R) model, $\mathcal{L}_R$ of same data.

$$\text{Likelihood ratio}: \frac{\mathcal{L}_R}{\mathcal{L}_F}, \quad \text{where } \mathcal{L}_R \leq \mathcal{L}_F$$



- **Hypotheses**: $H_0 : \beta_1 = \cdots = \beta_k = 0$

  (Reduced model is appropriate; fits data as well as Full model)

  H_a: at least one $\underline{\beta_1, \cdots, \beta_k} \neq 0$

  (Full model is better)

- **Test Statistic**: $G^2 = -2 \log \mathcal{L}_R - (-2 \log \mathcal{L}_F) = -2 \log \left(\frac{\mathcal{L}_R}{\mathcal{L}_F} \right)$

- For large n, under H_0, G^2 is an observation from a Chi-square distribution with $\underline{k}$ df.

LRT
- Global (Fitted vs Null)
- Deviance G-OF (Fitted vs Saturated)

Comparing models

- LRTs for models with and without set of indicator variables for effect of interest
- Particularly useful if > 2 levels in categorical explanatory variables

$I = J = 2,\ k = 3$

- *Example*: Suppose we have a $\overset{1}{2} \times \overset{1}{2} \times \overset{2}{3}$ table and we fit the Uniform association model (XY, XZ, YZ)

Z

$$
\begin{aligned}
\log \mu_{ijk} = & \beta_0 + \beta_1 \mathbf{I}_{X=1} + \beta_2 \mathbf{I}_{Y=1} + \boxed{\beta_3 \mathbf{I}_{Z=1} + \beta_4 \mathbf{I}_{Z=2}} \\
& + \beta_5 \mathbf{I}_{X=1} * \mathbf{I}_{Y=1} + \beta_6 \mathbf{I}_{X=1} * \mathbf{I}_{Z=1} + \beta_7 \mathbf{I}_{X=1} * \mathbf{I}_{Z=2} \\
& + \beta_8 \mathbf{I}_{Y=1} * \mathbf{I}_{Z=1} + \beta_9 \mathbf{I}_{Y=1} * \mathbf{I}_{Z=2}
\end{aligned}
$$

Is the YZ interaction needed?
$H_0 : \beta_8 = \beta_9 = 0$ vs H_a: at least 1 of β_8, β_9 is not 0

Comparing models

2-way Int: Effect of Factor 1 on outcome varies with Factor 2.

- *Exercise*: For the Uniform association model (AC,AM,CM), is the CM interaction needed in the Uniform association model?/ Does the (AC, AM) model fit just as well?

$$\log \mu_{ijk} = \beta_0 + \beta_1 \mathbf{I}_{A=1} + \beta_2 \mathbf{I}_{C=1} + \beta_3 \mathbf{I}_{M=1} + \beta_5 \mathbf{I}_{A=1} * \mathbf{I}_{C=1} + \beta_6 \mathbf{I}_{A=1} * \mathbf{I}_{M=1} + \beta_7 \mathbf{I}_{C=1} * \mathbf{I}_{M=1}$$

3-way Int: Effect of the combination of Factor 1 & 2 on the outcome varies with Factor 3.

- A: $H_0: \beta_7 = 0$ (Reduced, (AC,AM))
  $H_a: \beta_7 \neq 0$ (Full, (AC,AM,CM)) $(k=1)$

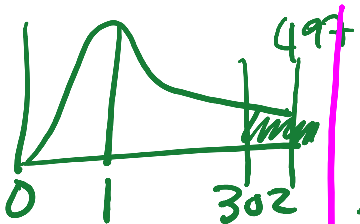
Wald

$Z^2 = (17.382)^2 = 302.13$

$Z^2 \sim \chi^2_1$

p-value $= P(\chi^2_1 > 302.13) \simeq 0$

(Same concl.)



LRT

Deviance $= 497.3 - 0.374 \simeq 497 \sim \chi^2_1$

p-value $= P(\chi^2_1 > 497) \simeq 0$

Evidence that there is an association between C use & M use, along with AC and AM.

What is the Deviance G-O-F test?

- Uses LRT: Compares
  - Fitted Model of Interest (REDUCED, R) model to
  - Saturated Model (FULL, F) model.
- Sometimes called "Drop-in-Deviance" test.

- Hypotheses:

  H_0: (Fitted model fits data as well as Saturated model)

  H_a: (Saturated model is better)
- Test Statistic:

$$Deviance = -2\log\left(\frac{\mathcal{L}_R}{\mathcal{L}_F}\right) = -2\log\left(\frac{\mathcal{L}_M}{\mathcal{L}_S}\right)$$

- Under H_0, *Deviance* is an observation from a chi-square distribution with $df = \#parameters(S) - \#parameters(M)$.

Deviance G-O-F statistic for 3-way tables

- The joint distribution of cell counts is

$y_{ijk} \sim P(\mu_{ijk})$

$$P(\mathbf{Y}=\mathbf{y}) = \prod_k \prod_j \prod_i \frac{\mu_{ijk}^{y_{ijk}} e^{-\mu_{ijk}}}{y_{ijk}!}$$

- Log-likelihood function:

$$\log \mathcal{L} = \sum_k \sum_j \sum_i (y_{ijk} \log \mu_{ijk} - \mu_{ijk} - \log y_{ijk}!)$$

- Likelihood ratio statistic: (Practice Question)

$-2\log L_S + 2\log L_F$

y_{ijk} $\hat{\mu}_{ijk}$

$$Deviance = 2\sum_k \sum_j \sum_i y_{ijk} \log\left(\frac{y_{ijk}}{\hat{\mu}_{ijk}}\right)$$

- Hints: Under the saturated model, $\boxed{\hat{\mu}_{ijk} = y_{ijk};}$ $\sum_k \sum_j \sum_i y_{ijk} = n$

How to interpret Deviance?

- Is the form of the fitted model adequate or do I need something more complicated?
- Compares fitted model to saturated model
- Small deviance / Large p-values implies:
  - Fitted model is adequate, OR
  - Test is not powerful enough to detect inadequacies
- Large deviance / Small p-values implies:
  - Fitted model is not adequate; consider a more complex model OR
  - Underlying distribution is not adequately modelled by the Poisson distribution / Poisson model not correct / $Var(y_{ijk}) > \mu_{ijk}$ OR
  - There are severe outliers in the data

Are there outliers?

- Check residuals

1. Raw residual: $y_{ijk} - \hat{\mu}_{ijk}$
2. Pearson residual: sum of squares gives Pearson chi-square test statistic

$$\frac{y_{ijk} - \hat{\mu}_{ijk}}{\sqrt{\hat{\mu}_{ijk}}}$$

$\sim N(0,1)$ with large n

$\sum_{i=1}^{n} Z^2$

3. Deviance residual: sum of the squares is the Deviance

$\sum_{i=1}^{n} D^2$

$$\text{sign}(y_{ijk} - \hat{\mu}_{ijk})\sqrt{2\left\{y_{ijk}\log\left(\frac{y_{ijk}}{\hat{\mu}_{ijk}}\right) - y_{ijk} + \hat{\mu}_{ijk}\right\}}$$

Pearson and Deviance residuals

- Easier to interpret: Pearson
- More reliable: Deviance
- Usually similar? Yes
- Differences are more prominent when used to compare models
- If Poisson means are large, the sampling distributions are ... Approx Normal
- Rule-of-thumb: Outlier if Pearson or Deviance residual > 3 (if sample size is small, consider those > 2)

Presence of "Extra-Poisson Variation"

- Check if $\frac{Deviance}{df} > 1$
- Q: How much > 1 is important?
- A: If Deviance GOF test is statistically significant.
- If other problems are ruled out, then include a dispersion parameter in the model, i.e.,

$$\text{Var}(Y_{ijk}) = \psi \mu_{ijk}$$

$$\hat{\psi} = \frac{\text{Pearson}}{df}$$

quasi-Poisson

- OR use Negative Binomial regression

$$Var(Y_{ijk}) = \mu_{ijk}(1 + \psi \mu_{ijk})$$

(Agresti, Chp. 14)

Summary of models

$y = X\beta$ $\qquad \pi = \frac{e^{\mu}}{1+e^{\mu}}$ $\qquad \mu = e^{\mu}$

	OLS	Logistic	Log-linear
Link	Identity	Logit	Log
Regression	Linear	(Non-linear	Non-linear)
$\mu\{Y\vert\mathbf{X}\}$ is	linear in β's	not linear in β's	not linear in β's
Models	Mean of Y	Log odds	Log of means
Natural response	Yes	Yes	No
Response is	Normal	Binomial	Poisson
Indep. Obs.	Yes	Yes	Yes
$Var(Y_i\vert\mathbf{X}) =$	σ^2	$\pi_i(1-\pi_i)$	μ_i

(variance is constant) (variance changes with i)

Week's Summary

- Three-way contingency tables:
  - Log-linear model approach
  - Types of independence or association/ interactions
    - (i) Complete
    - (ii) Block
    - (iii) Partial
    - (iv) Uniform association
    - (v) 3-way interaction
  - Deviance goodness-of-fit test
  - Using fitted equation to find odds
  - Model diagnostics
- Things to do:
  - Assignment #3 → Mar. 28
  - Participation 8 → Apr. 3
  - Practice Problems on Poisson Regression (Log-linear models)