STAT475/675 Final Review

http://www.sfu.ca/~zza115/teaching.html
http://people.stat.sfu.ca/~joanh/stat475-675web.html

Zhiyang Zhou (zhiyang_zhou@sfu.ca)

2018-04-09

One binary variable

- Ber(π) and Binom(n, π)
- MLE for π: $\hat{\pi} = w/n$
- CI for π:
  - Wald: may exceed $[0, 1]$
  - Wilson (score-type): always between 0 and 1
  - Agresti-Coull: recommended for $n \geq 40$; may exceed $[0, 1]$
  - Clopper-Pearson (exact): conservative
  - interpretation

One Poisson variable

- $Y_1, \ldots, Y_n \sim$ Poisson(μ)
- MLE for μ: $\hat{\mu} = \bar{Y}$
- CI for μ
  - Wald
  - score-type
  - Clopper-Pearson

Two binary variables and 2×2 contingency table

- with binomial or multinomial sampling
  - Concepts
    - joint probability, marginal probablity, conditional probability
    - independence: $\pi_{ij} = \pi_{i+}\pi_{+j}$
    - difference ($\frac{\pi_{11}}{\pi_{1+}} - \frac{\pi_{21}}{\pi_{2+}}$), relative risk (RR), odds ratio (OR)
    - why OR is prefered?
  - Inference
    - MLE for π_{ij}, π_{+j}, π_{i+}
    - CI for difference: Wald, Agresti-Caffo
    - CI for RR or (OR): construct Wald CI for ln(RR) (or ln(OR)) and take exp
      - log form is more approriate to be normally approximated
      - the lower bound is above 0
      - interpretation
    - test independence
      - independence $\Leftrightarrow$ RR = OR = 1
      - find out whether CI covers 1
      - χ^2-test and LRT: $n_{ij} \geq 5$ and $n \gg 1$
- with hypergeometric sampling
  - Lady tasting tea
  - Fisher's exact test

- Permutation χ^2 test

$I \times J$ contingency table (with purposive or multinomial sampling)

- $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j vs $H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$
- χ^2-test and LRT statistics χ^2_{obs} and G^2_{obs} with $(I-1)(J-1)$
- Conclusion
  - $p-$ value $< \alpha$: there is a strong evidence against H_0, i.e., a signficant association between X and Y

$2 \times 2 \times K$ contingency table

- X-Y partial table: fixed at a level of Z
  - X-Y conditional OR $\theta_{XY(k)}$
  - homogeneous conditional X-Y association: $\theta_{XY(k)} \equiv$ constant
    * Breslow-Day test
    * Mantel-Haenszel estimator
  - $X \perp\!\!\!\perp Y|Z \Leftrightarrow \theta_{XY(k)} = 1$ for all k
    * Cochran-Mantel-Haenszel test
- X-Y marginal table: ignore Z
  - X-Y marginal OR: θ_{XY}
  - $X \perp\!\!\!\perp Y \Leftrightarrow \theta_{XY} = 1$
- Simpson's paradox: $X \perp\!\!\!\perp Y \not\Leftrightarrow X \perp\!\!\!\perp Y|Z$ in general
- Test on mutual independence of (X, Y, Z): χ^2-test & LRT

Unified framework

- Generalized linear model (GLM)
  - random component: $Y \sim f_{\mu_1,\ldots,\mu_I}$, a parametric distribution (belonging to the exponential family) characterized by mean functions $\mu_i = \mu_i(x_1, \ldots, x_p), \quad i = 1, \ldots, I$
  - systematic component: linear function with respect to β's
  - link function: monotone and differentiable over the range of $(\mu_1, \ldots, \mu_I)$
- Generalized linear mixed model (GLMM)
  - compared with GLM, β's are randomized

Logistic regression

- GLM components
  - random component: $Y \sim \text{Ber}(\pi) = \text{Binom}(1, \pi)$ with $\pi = \pi(x_1, \ldots, x_p)$
  - systemetic Component: linear function with respect to β's
  - link function: logit
- Inference
  - MLE for β's and then π and $\ln \text{OR}$
  - CI: detour
- Coding schemes for a predictor with m (≥ 2) levels
  - qualitative: replace it with $m - 1$ dummy binary predictors
    * R
    * SAS

    * ANOVA-type
  - quantitative: take it as a single ordinal predictor
  - remark: parameters under different coding schemes have different values and meanings

Multicategory logit model

- GLM components
  - random component: $Y \sim \text{Multinom}(1, \pi_1, \ldots, \pi_J)$, $J \geq 3$
  - systemetic Component: linear function with respect to β's
  - link function
    * $\ln \frac{\pi_j}{\pi_1}$, $j = 2, \ldots, J$: baseline-category logit model
    * $\text{logit}(\sum_{i=1}^{j} \pi_i(x_1, \ldots, x_p))$, $j = 1, \ldots, J-1$: cumulative logit model
    * $\ln \frac{\pi_{j+1}}{\pi_j}$, $j = 1, \ldots, J-1$: adjacent-categories logit model
- Nominal response: baseline-category logit model
  - odds of Category j vs Category i for $x_1, \ldots, x_p$ regardless of baseline category:

$$\frac{\pi_j}{\pi_i} = \exp((\alpha_j - \alpha_i) + (\beta_{j1} - \beta_{i1})x_1 + \cdots + (\beta_{jp} - \beta_{ip})x_p)$$

  - OR of Category j vs Category 1 for $x_1 + c_1, \ldots, x_p + c_p$ and $x_1, \ldots, x_p$:

$$\frac{\pi_j}{\pi_i} = \exp(\beta_{j1}c_1 + \cdots + \beta_{jp}c_p)$$

- Ordinal response: cumulative logit & adjacent-categories logit model
  - proportional case: $\beta_{j1} = \cdots = \beta_{jp}$

Loglinear regression

- GLM components
  - random component: $Y \sim \text{Poisson}(\mu)$ with $\mu = \mu(x_1, \ldots, x_p)$
  - systemetic Component: linear function with respect to β's
  - link function: ln
- Inference
  - MLE for β's and then π and $\ln \text{OR}$
  - CI: detour
- Coding schemes for a predictor with m (≥ 2) levels
  - qualitative: replace it with $m - 1$ dummy binary predictors
    * R
    * SAS
    * ANOVA-type
  - quantitative: take it as a single ordinal predictor
  - remark: parameters under different coding schemes have different values and meanings

Loglinear regression for contingency Table

- Saturated loglinear model (XYZ)
- Loglinear model of homogeneous association (XY, YZ, XZ)
  - When $I = J = 2$,

$$\ln \theta_{XY(k)} = \ln \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} = \beta_{11}^{XY} + \beta_{22}^{XY} - \beta_{12}^{XY} - \beta_{21}^{XY}$$

        * $\theta_{XY(k)}$ stays still for all k, i.e., homogeneous conditional OR holds
        * if $\beta_{ij}^{XY} = 0$ for all i, j, then
            · $\theta_{XY(k)} = 1$ for all k
            · $X \perp\!\!\!\perp Y|Z$

- Loglinear model of independence (X, Y, Z)

Loglinear-logit connection

- loglinear$(XYZ) \Leftrightarrow Y \sim$ logit(XZ) or multi-logit(XZ)
    - $\ln \mu_{ijk} = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$
    - $\ln \frac{\pi_{ijk}}{\pi_{ij'k}} = \beta_{j0} + \beta_{ji}^X + \beta_{jk}^Z + \beta_{jik}^{XZ}$
    - $\beta_{j0} = \lambda_j^Y - \lambda_{j'}^Y$
    - $\beta_{ji}^X = \lambda_{ij}^{XY} - \lambda_{ij'}^{XY}$
    - $\beta_{jk}^Z = \lambda_{jk}^{YZ} - \lambda_{j'k}^{YZ}$
    - $\beta_{jik}^{XZ} = \lambda_{ijk}^{XYZ} - \lambda_{ij'k}^{XYZ}$
- loglinear$(XY, YZ, XZ) \Leftrightarrow Y \sim$ logit(X, Z) or multi-logit(X, Z)
    - $\ln \mu_{ijk} = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
    - $\ln \frac{\pi_{ijk}}{\pi_{ij'k}} = \beta_{j0} + \beta_{ji}^X + \beta_{jk}^Z$
    - $\beta_{j0} = \lambda_j^Y - \lambda_{j'}^Y$
    - $\beta_{ji}^X = \lambda_{ij}^{XY} - \lambda_{ij'}^{XY}$
    - $\beta_{jk}^Z = \lambda_{jk}^{YZ} - \lambda_{j'k}^{YZ}$
- Given $X \perp\!\!\!\perp Z|Y$, loglinear$(XY, YZ) \Leftrightarrow Y \sim$ logit(X, Z) or multi-logit(X, Z)
    - $\ln \mu_{ijk} = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$
    - $\ln \frac{\pi_{ijk}}{\pi_{ij'k}} = \beta_{j0} + \beta_{ji}^X + \beta_{jk}^Z$
    - $\beta_{j0} = \lambda_j^Y - \lambda_{j'}^Y$
    - $\beta_{ji}^X = \lambda_{ij}^{XY} - \lambda_{ij'}^{XY}$
    - $\beta_{jk}^Z = \lambda_{jk}^{YZ} - \lambda_{j'k}^{YZ}$

Modified Poisson model

- Poisson rate regression
    - GLM components
        * random component: $Y \sim$ Poisson(μ) with $\mu = \mu(t, x_1, \ldots, x_p)$
        * systemetic Component: linear function with respect to $\ln t$ and β's
        * link function: ln
- Zero-inflated Poisson (ZIP) model
    - GLM components
        * random component: with $\mu = \mu(x_1, \ldots, x_p)$,

$$Y \begin{cases} = 0 \text{ with probabity } \pi \\ \sim \text{Poisson}(\mu) \text{ with probabity } 1 - \pi \end{cases}$$

        * systemetic Component
            · $\ln \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
            · $\text{logit}^{-1} \pi = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_J z_J$
        * link function: ln for μ & logit for π

Probit Regression

- GLM components
  - random component: $Y \sim \text{Ber}(\pi) = \text{Binom}(1, \pi)$ with $\pi = \pi(x_1, \ldots, x_p)$
  - systemetic Component: linear function with respect to β's
  - link function: Φ^{-1}, where Φ is the cdf of $N(0, 1)$

Marginal modelling

- Response without specific distribution
- Quasi-Poisson
  - Motivation: overdispersion
  - GLM components
    - random component: $Y \sim (\mu, \rho\mu)$
    - systemetic Component: linear function with respect to β's
    - link function: ln
  - Remark
    - The loglinear model and quasi-Poisson model offer identical estimate for the mean funtion.
    - CIs from the quasi-Poisson model completely cover corresponding ones from the loglinear model.
- Generalized estimating equation (GEE)
  - Motivation: existence of within-cluster correlation
  - GLM components
    - random component: Y_{ij} with $\text{E}(Y_{ij}) = \mu_{ij}$ and $\text{cov}(Y_{ij}, Y_{i'j'}) = 0$ if $i \neq i'$
    - systemetic Component: linear function with respect to β's
    - link function: based on the response
  - Remark
    - Even the working correlation is misspecified, the estimate of mean function is still consistent.

Model evaluation and selection

- Model checking
  - inferential methods
    - $H_0 : M$ is correct vs H_1 : otherwise
      - special case: checking independence for contingency tables
    - χ^2-test & LRT
  - graphical method: residual plots
    - Pearson's residual
    - standardized (adjusted) Pearson's residual
    - extreme residuals
      - $|e_i| \geq 2$ (or $|e_i^*| \geq 2$): 5% if the model is correct
      - $|e_i| \geq 3$ (or $|e_i^*| \geq 3$): extremely rare (0.1%) if the model is correct
      - $|e_i| \geq 4$ (or $|e_i^*| \geq 4$): unexpected at all if the model is correct
- Model comparison and variable selection
  - LRT: $M_0 \subset M_1$,
$$H_0 : M_0 \text{ vs } H_1 : M_1$$
    - under H_0, $\mathcal{G}^2(M_0|M_1) = \mathcal{G}^2(M_0|M_s) - \mathcal{G}^2(M_1|M_s) = -2\ln\frac{\max L_{M_0}}{\max L_{M_1}} \approx \chi^2(\text{df}_{\text{res}}(M_0) - \text{df}_{\text{res}}(M_1))$
      - M_s is the saturated model
      - $\text{df}_{\text{res}}(M_0) - \text{df}_{\text{res}}(M_1)$ = the difference on numbers of non-redundant parameters

        - M_0 ought to be nested into M_1
    - information criteria: $\text{IC}(k) = -2\ln(L(\hat{\beta}|\text{data})) + kr$
    - quasi-information criteria

Summary

- Have a big picture on contingency tables and GLMs
- Review lecture notes carefully: understand the concepts and examples
- Review assignments and midterm papers: understand the R outputs

TA Office Hours

- 11am - 12pm, April 11, AQ4145
- 11am - 12pm, April 18, AQ4145
- 10am - 12pm, April 20, AQ4145