STA 303: Summary of Models

Alison Gibbs

Winter 2012

1 Components of a Generalized Linear Model

- Response variable: Y
- Explanatory variables: $X_1, \ldots, X_p$
- Link function: $g(\cdot)$
- Model: $g(\mathrm{E}(Y)) = f(\mathbf{X}; \boldsymbol{\beta})$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and $f(\mathbf{X}; \boldsymbol{\beta})$ is a linear function of the β's

2 One-way and Two-way Analysis of Variance

1. Underlying probability distribution: Normal
2. Response variable – continuous
3. Explanatory variables – categorical
4. Model: $Y = f(\mathbf{X}; \boldsymbol{\beta}) + e$ or $\mathrm{E}(Y) = f(\mathbf{X}; \boldsymbol{\beta})$ where the explanatory variables are indicator variables with coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and $f(X; \boldsymbol{\beta})$ is a linear function of the β's
5. Link function: identity
6. Conditions for valid inference (assuming correct form of model[1]):
   - independent observations
   - same variance
   - normally distributed error terms (so no outliers)
7. Estimation: least squares
8. Inference: t and F tests based on the Normal distribution

[1]Correct form of model includes: necessary explanatory variables are in the model, unnecessary explanatory variables are not in the model, continuous explanatory variables are transformed as appropriate

3 Logistic Regression

2.1 Binary

1. Underlying probability distribution: Bernoulli
2. Response variable – binary

2.2 Binomial

1. Underlying probability distribution: Binomial
2. Response variable – binomial counts out of m trials

3. Explanatory variables – anything
4. Model: $\log\left(\frac{\pi}{1-\pi}\right) = f(\mathbf{X};\boldsymbol{\beta})$ where $f(\mathbf{X};\boldsymbol{\beta})$ is a linear function of the β's
5. Link function: logit
6. Conditions for valid inference (assuming correct form of model):
   - independent observations
   - variance follows Bernoulli / binomial distribution form
   - no outliers
   - large sample size
7. Estimation: maximum likelihood estimation
8. Inference: Likelihood ratio tests, Wald tests and confidence intervals based on large-samples properties of maximum likelihood estimators

4 Poisson Regression and Log-linear Models

1. Underlying probability distribution: Poisson(μ)
2. Response variable – counts
3. Explanatory variables – anything for Poisson regression; categorical variables for log-linear models on contingency tables
4. Model: $\log(\mu) = f(\mathbf{X};\boldsymbol{\beta})$ where $f(\mathbf{X};\boldsymbol{\beta})$ is a linear function of the β's
5. Link function: log
6. Conditions for valid inference (assuming correct form of model):
   - independent observations
   - variance = mean
   - no outliers
   - large sample size
7. Estimation: maximum likelihood estimation
8. Inference: Likelihood ratio tests, Wald tests and confidence intervals based on large-samples properties of maximum likelihood estimators

5 Repeated Measures / Longitudinal Data (General Linear Mixed Model)

1. Underlying probability distribution: Normal (for both errors and random effects)
2. Response variable – continuous; multiple measures of the response on each subject
3. Explanatory variables (fixed effects) – anything
4. Model: $Y = f(\mathbf{X}; \boldsymbol{\beta}) + u + e$ where $f(\mathbf{X}; \boldsymbol{\beta})$ is a linear function of the β's, u is the random effect, e is random noise
5. Link function: identity
6. Conditions for valid inference (assuming correct form of model):
   - Correct form of model includes correct covariance structure for observations on same subject
   - observations on different subjects independent observations
   - error variance can be modelled to vary across X's
   - normally distributed error terms and random effects (so no outliers)
   - large sample size for likelihood ratio test to compare models with different variance-covariance structures
7. Estimation: restricted maximum likelihood (maximum likelihood estimation for variance and covariance parameters and generalized least squares for coefficients of fixed effects)
8. Inference: t and F tests based on the Normal distribution for fixed effects

6 Statistical Methods Covered that are Not Linear Models

- Two-sample t-test
- Two-sample test of proportions
- Pearson's chi-square test on two-way contingency tables
- Likelihood ratio test based on the multinomial distribution for two-way contingency tables