STA303 - Methods of Data Analysis II

Heather Tan

Jan - Apr 2020

Contents

1 Linear model recap

Why model? The goal of a model is to provide a (relative) simple summary of a data set. We can use it to describe data and make predictions.

1.1 The general Linear Model

$$Y = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$
$$DATA = MODEL + ERROR$$

The model is **general** since it can have multiple response variables. It is **linear** because the model is linear in parameters.
E.g.

$$y_i = \beta_0 + \gamma_1\delta_1 x_{1i} + exp(\beta_2)x_{2i} + \epsilon_i = \beta_0 + \gamma_1\delta_1 x_{1i} + e^{\beta_2}x_{2i} + \epsilon_i$$

is **linear** since it can be replace to $y_i = \beta_0 + c_1 x_{1i} + c_2 x_{2i} + \epsilon_i$

$$y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \epsilon_i$$
$$y_i = \beta_0 exp(\beta_1 x_{1i}) + \epsilon_1$$

is **not linear** since it can be replaced as $y_i = c_0 + c_1 x_i^{c_2} + \epsilon$ or $y_i = c_0 \cdot e^{c_1 x_1} + \epsilon$

1.2 Linear regression assumptions

1. Errors are independent
2. Errors are identically distributed with $E[\epsilon_i] = 0$
3. Homoscedasticity $Var[\epsilon_i] = \sigma^2$
4. **A straight-line relationship exists between ϵ_i and y_i**

$1 - 3 \implies \epsilon_i \sim N(0, \sigma^2)$

2 ANOVA

Analysis of Variance is a statistical method used to test differences between two or more means. It is useful since in real life we often want to compare more than two groups
Two-sample t-tests can be problematic:

- increasing the risk of Type I error$(H_1|H_0 True)$
- At $\alpha = 0.05$, with 100 comparison, 5 will show a difference when none exists

2.1 ANOVA Hypothesis

$H_0 : \mu_1 = \cdots = \mu_n$

H_1: at least one mean is different from others

2.2 ANOVA assumptions

$\epsilon_i \sim N(0, \sigma^2)$

2.2.1 A not on Normality

If N is large, Central limit Theorem can be used to relax the normality assumption. The normality assumption is most important when

- n is small
- highly non-normal
- small effective size

2.3 ANOVA test

- **Group Mean:** $\hat{\mu_j} = mean\{\hat{y_j}, j = 1, \cdots, J_i\}$
- **Grand Mean:** $\hat{\mu_0} = mean\{\hat{y_{ij}}, i = 1, \cdots, kj = 1, \cdots, J_i\}$
- want to test:**Are the $\hat{\mu_i}$ close to $\hat{\mu_0}$?**

$$\frac{\sum_j n_j(\hat{\mu}_j - \hat{\mu}_0)^2/(k-1)}{\sum_{ij}(\hat{y}_{ij} - \hat{\mu}_i)^2/(N-k)} \sim F_{N-1,M-1}$$

$$\frac{\text{variability between groups}}{\text{variability within groups}} \sim \text{F-distribution}_{obs-1,obs-groups}$$

$$\frac{s_b^2}{s_w^2} \sum F_{N-1,N-k}$$

2.3.1 Last commands on Anova

- ANOVA is testing, not estimating but it uses a linear model structure to get the variation between and within groups
- ANOVA tells us if at least on of the group means is different but doesn't give us insight into how the group means are different
- ANOVA will reject H_0 for large dataset
- **ANOVA can be useful when a dataset is small, for large data set, fit a random effects model.**

3 Designing matrics

Whenever you have categorical variables in your dataset, R is making you dummy variables for the levels of those categorical variables. Basically a true or false for if each observation take the given level of the categorical variable

4 Likelihood ratio(LR) test

The likelihood-ratio test lets us compare the goodness of fit of two competing models based on the ratio of their likelihoods.

Additional info link: `https://www.zhihu.com/question/51287367/answer/149207450`

4.1 Maximum Likelihood Estimation

- Model Parameters: β, σ
- Likelihood function: $L(\beta, \sigma|Y) = \pi(Y; \beta, \sigma)$
- MLE's: $(\hat{\beta}, \hat{sigma}) = \overset{argmax}{\beta, \sigma} \; L(\beta, \sigma|Y)$

4.2 LR Hypothesis

- $H_0 : Y_{ij} \sim N(\mu_0, \sigma^2)$ i.e errors are iid noise
- $H_1 : Y_{ij} \sim N(\mu_j, \sigma^2)$ i.e erros vary by explanatory variable
- Very low chance of observing a LR(likelihood ratio) too big if H_0 were true

4.3 LR test interpreting

- likelihood under H_1 is always higher than under H_0
  - having different $\hat{\mu}_i$ is possibly to have higher likelihood than all same $\hat{\mu}$
  - if H_0 true, the alternative hypothesis likelihood shouldn't be much larger
- 2 times the difference in log likelihood will follow a chi-square distribution if H_0 true.

5 Generalized linear Models

Generalized linear models are a flexible class of models that let us generalise from the linear model to include more types of response variables, such like count, binary and proportion data.

5.1 Assumptions of the GLM

- GLM does not assume a linear relationship between the dependent variable and the independent variables, but it does assume a linear relationship between the transformed response(in terms of ink function) and the explanatory variables.
- **Dependent variable:**
  - Independently distributed
  - Can have a distribution from any of the exponential family: binomial, poisson, multinomial or normal
  - Dependent variable is not assumed to have a linear relationship between it and independent variable. i.e LHS can have non linear transformation
- **Error:** errors are independent but not necessarily normal distributed
- **Independent variable:**
  - Have a linear relationship between the transformed response and independent variables i.e RHS must be a linear formula
  - Can be the power terms or some other none linear transformations of the original independent variables.
- The homogeneity of variance does not need to be satisfied
- It use MLE in stead of OLS to estimate the parameters, thus relies on large-sample approximations.

5.2 Components of GLM

1. **random component:** the response and an associated probability distribution. i.e y and its distribution
2. **systematic component:** explanatory variables and relationships among them. i.e x and β
3. **link function:** which tell us about the relationship between the systematic component(or linear predictor) The link function allows us to generalize the linear models for count, binomial and percent data. i.e the link between x and y

5.3 Model of GLM

$$Y_i \sim G(\mu_i, \theta)$$
$$h(\mu_i) = X_i^T \beta$$

- G is the distribution of the response variable
- μ_i is a location parameter for the obseration i
- θ are the additional parameters for the density of G
- h is a link function
- X_i are covariates for observation i
- β is a vector of regression coefficients

5.3.1 OLS comparison

OLS:

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = X_i^T \beta$$

OLS is just a flavor of GLM when

- G is a normal distribution
- θ is the variance parameter, denoted as σ^2
- h is the density function

5.4 Binomial (or logistic) regression

$$Y_i \sim Bin(N_i, \mu_i)$$
$$log(\frac{\mu_i}{1-\mu_i}) = X_i\beta$$

- g is a Binomial Distribution or Bernoulli if $N_i = 1$
- h is the logit link. What's logit?: `https://zhuanlan.zhihu.com/p/27188729`
- $X_i^T\beta$ can be negative
- μ_i is between 0 and 1

5.4.1 Inference: Paramter estimation

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$h(\mu_i) = X_i^T \beta$$
$$\pi(Y_1, \cdots, Y_n; \beta, \theta) = \Pi_{i=1}^N f_G(Y_i; \mu_i, \theta)$$
$$logL(\beta, \theta; y_1, \cdots, y_N) = \sum_{i=1}^N log f_G(y_i; \mu_i, \theta)$$

- The Y_i are independetnly distributed
- **Joint density** π of random variables $(Y_1, \cdots, Y_n)$ is the product of the marginal densities f_G
- **Likelihood function** L given observed data $y_1, \cdots, y_N$ is a function of the parameters.
- **MLE:** $\hat{\beta}, \hat{\theta} = argmax_{\beta,\theta} L(\beta, \theta; y_1, \cdots, y_N)$

$$\frac{\partial}{\partial \beta_p} logL(\beta, \theta; y_1, \cdots, y_N) = \sum_{i=1}^{N} [\frac{\partial}{\partial \mu} log f_G(Y_i; \mu, \theta)]_{\mu = h^{-1}(X_i^T \beta)} [\frac{\partial}{\partial \eta} h^{-1}(\eta)]_{\eta = X_i^T \beta} \cdot X_{ip}$$

- Information Matrix $I(\hat{\beta}|Y) = \frac{\partial}{\partial \beta \partial \beta^T} - logL(\beta|Y)|\hat{\beta}$
- MLEs are approximately Normal $\hat{\beta} \sim MVN(\beta, I(\hat{\beta}|Y)^{-1})$
- Standard errors are the root of diagonals of inverted information matrix.

5.4.2 Likelihood ratio tests

$$2[logL(\hat{\beta}; y) - logL(\beta; y)] \sim \chi^2_P$$

where P is the number of parameters in β

5.4.3 Comparing Nested Models

Model A is nested in Model B if the parameters in Model A are a subset of the parameters in Model B.

- $H_0 : \beta_k = C_k \forall k \in \Omega, \quad \Omega \subset \{1, \cdots, P\}$
- $H_1 : \beta$ unconstrained
- **Nested:** H0 is a special case of H1
- Write $\hat{\beta}^{(C)}$ as the constrained MLES under H0

$$2[logL(\hat{\beta}; y) - logL(\hat{\beta}^{(C)}; y)] \sim \chi^2_{|\Omega|}$$

5.4.4 Interpreting Logistic Models

$$Y_i \sim Bin(N_i, \mu_i)$$
$$log(\frac{\mu_i}{1-\mu_i}) = \sum_{p=1}^{P} X_{ip}\beta_p$$
$$(\frac{\mu_i}{1-\mu_i}) = \Pi_{p=1}^{P} exp(\beta_p)^{X_{ip}}$$

- μ_i is a probability
- $log(\frac{\mu_i}{1-\mu_i})$ is a log-odds
- $(\frac{\mu_i}{1-\mu_i})$ is an odds
- $\mu \approx 0 \implies \mu_i \approx \mu_i/(1-\mu_i)$

Suppose $X_{1p} = X_{2p} \forall p$ except $X_{2q} = X_{1q} + 1$

$$\beta_p = log(\frac{\mu_2}{1-\mu_2}) - log(\frac{\mu_1}{1-\mu_1})$$
$$exp(\beta_p) = (\frac{\mu_2}{1-\mu_2})/(\frac{\mu_1}{1-\mu_1})$$

- β_p is the log-odds ratio
- $exp(\beta_p)$ is the odds ratio
- $exp(intercept)$ is baseline odds, when $X_{i2} \cdots X_{ip} = 0$

6 w4 - Linear Mixed Models(LMMs)

why we cannot assume independence?: multiple responses from the same subject cannot be regarded as independent from each other.

6.1 Model set up

$$Y_{ij}|U_i \sim N(\mu_{ij}, \tau^2)$$
$$\mu_{ij} = x_{ij}\beta + U_i$$
$$[U_1, \cdots, U_M]^T \sim MVN(0, \Sigma)$$

- Observations Y_{ij} for repeated measures j on individuals i
- fixed effects $X_{ij}\beta$
- random effects U_i for i in 1 to M(new parts that makes this a linear mixed model)

6.2 Assumption of LMMs

1. There is a continuous response variable
2. we have modeled the dependency structure correctly
3. Our units/subjects are independent, even though observations within each subjects are taken not to be

4. Both the random effects and within-unit residual errors follow normal distribution
5. The random effects errors and within-unit residual errors have constant variance

6.3 Alternative formulations

$$Y_{ij} = X_{ij}\beta + \epsilon_{ij}$$
$$\epsilon_{ij} = U_i + Z_{ij} \quad Z_{ij} \sim N(0, \tau^2)$$
OR
$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

- Y_i is a vector of outcomes for subject i
- X_i and Z_i are model matrices for the fixed and random effects
- The vector β describes the effect of covariates on the mean/expectation of the outcome and the vector b_i is the random effects for units
- ϵ_i is the vector of residual errors, normally distributed with a given variance and the errors within units are mutually independent.

7 w5-Generalised Linear Mixed Models(GLMMs)

Pros:

- powerful class of methods that conbined the characteristics of generalized linear models and linear mixed models
- Can be used with a range of response distributions(Poi, Bi, Gam)
- Can be used in a range of situation where observations are grouped in some way
- fas and can be extended to handle somewhere more complex situations

Cons:

- Some of the standard ways we've learned to test models don't apply
- Greater risk of making sensible models that are too complex for our data to support

7.1 Assumption

1. X are independent, even though Y within each subject are taken not to be
2. Random effects come from a normal distribution
3. the random effects errors and within-unit residual errors have constant variance
4. The chosen link function is appropriate/ the model is correctly specified

8 In-class Questions

8.1 Week 5

Q: The shipd dataset found in the MASS package gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation. Which model do you think you should fit?
A: Generalised linear model(GLM) - Poisson

Q: How might we model the risk of mortality for people of different income and education level?
A: GLM-Binomial

Q: how might we model the number of species on an different island in the Galapagos based on features of each island?
A: (Ch3 faraway,week 3 page) GLM- Poisson

Q: How might we model the heights of children based on their age, birth weight and mother's age at birth
A: Linear model

Q: How might we model the achievement scores of students in a standardise test across 15 different hs in TDSB?
A: Linear mixed model

Q: How might we model the waiting time for ride share cars, based on suburb/neibourhood of the request and time of the day
A: GLM-Gamma

9 Notes of Coding

- **::** - helps to access the exact function from that specific package
- R code "myLik" calculates $-logL(\beta, \sigma|Y)$, "optim" minimizes