Methods of Data Analysis II

Derek Li

Contents

1 Review

1.1 Linear Model

A model is linear when each term is either a constant or the product of a parameter and a predictor variable. A linear equation is constructed by adding the results for each term.

Example 1.1. $Y = \beta_0 + \beta_1 X_1 + (\beta_2 X_2)^2, Y = \beta_0 + \beta_1\beta_2 X_2 + \beta_3 X_3, Y = \beta_0 + \beta_1 X_1 + X_3^{\beta_2}$ are not linear models, while $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 e^{X_3}$ is a linear model.

Recall the general linear model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Thus,

$$Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(X_i^T\beta, \sigma^2) \text{ and } Y \sim \mathcal{N}_n(X\beta, \sigma^2 I).$$

Recall the assumptions of linear regression:

1. Errors are independent (observations are independent);
2. Errors are identically distribute with $\mathbb{E}[\varepsilon_i] = 0$;
3. Homoscedasticity: $\text{Var}[\varepsilon_i] = \sigma^2$;
4. A straight-line relationship exists between ε_i and y_i.

1.2 ANOVA

Analysis of variance (ANOVA) is used to test differences between two or more means and to test general rather than specific differences among means. In fact, ANOVA is just a specific case of the general linear model. We test

$$H_0 : \mu_1 = \cdots = \mu_n \text{ against } H_1 : \text{At least one mean is different from the others.}$$

Recall the assumptions of ANOVA:

1. Errors are independent;
2. Errors are normally distributed with $\mathbb{E}[\varepsilon_i] = 0$;
3. Homoscedasticity: $\text{Var}[\varepsilon_i] = \sigma^2$.

Note that the normality assumption can be relaxed if sample size is large (Central Limit Theorem). The normality assumption is most important when N (sample size) is small, highly non-normal or small effect size.

Define group mean as

$$\widehat{\mu}_j = \text{mean}\{\widehat{y}_j, j = 1, \cdots, J_i\},$$

and grand mean as

$$\widehat{\mu} = \text{mean}\{\widehat{y}_{ij}, i = 1, \cdots, k, j = 1, \cdots, J_i\}.$$

We have

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

or

$$\frac{s_b^2}{s_w^2} \sim F_{(k-1,N-k)},$$

where s^2 is the sample variance.

Here are some comments on ANOVA:

- ANOVA is not estimating but it uses a linear model structure to get the variation between and within groups.
- ANOVA tells us if at least one of the group means is different but does not give us insights into how the group means are different.
- ANOVA will reject H_0 for any large dataset so that it can be useful when a dataset is small. For large datasets, we fit a random effects model.

1.3 Likelihood Ratio Test

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

$$H_0 : Y_{ij} \sim \mathcal{N}(\mu_0, \sigma^2) \text{ against } H_1 : Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2).$$

Likelihood under H_1 is always larger than under H_0 because having different $\widehat{\mu}_i$ cannot possibly be worse than constraining $\widehat{\mu}_i = \widehat{\mu}_0$. But if H_0 is true, the H_1 likelihood should not be much larger.

Note that ANOVA uses F distribution and the LR test uses χ^2 test.

In R,

```
lmtest::lrtest(lm0, lm1)
```

2 Generalized Linear Model

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

Here are some comments of the generalized linear model:

1. The data $Y_1, \cdots, Y_n$ are independently distributed and thus errors are independent (but not necessarily normally distributed).
2. The dependent variable Y_i is typically from an exponential family (e.g., binomial, Poisson, multinomial, normal).
3. GLM does not assume a linear relationship between the dependent variable and the independent variables, but a linear relationship between the transformed response (in terms of the link function) and the explanatory variables.
4. Explanatory variables can be even the power terms or some other non-linear transformations of the original independent variables.
5. The homogeneity of variance does not need to be satisfied.
6. It uses MLE rather than OLS to estimate the parameters, and thus relies on large-sample approximations.

2.1 Components of GLM

GLM has three parts:

1. Random component: the response and an associated probability distribution.
2. Systematic component: explanatory variables and relationships among them (e.g., interaction terms).
3. Link function: the relationship between the systematic component (or linear predictor) and the mean of the response.

It is the link function that allows us to generalize the linear models for count, binomial and percent data. It ensures the linearity and constrains the predictions to be within a range of possible values.

2.2 Model

We write

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

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

Note that OLS is a version of GLM when G is a normal distribution, θ is the variance parameter, denoted σ^2, and h is the identity function.

Example 2.1 (Binomial or Logistic Regression). $Y_i \sim \text{Bin}(N_i, \mu_i), h(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) = X_i^T \beta$, where G is a binomial distribution or a Bernoulli if $N_i = 1$, h is the ***logit link***, and $\mu_i \in [0, 1]$.

In R, binomial models in GLM require `y` to be a matrix with two columns `Success` and `N-Success`:

```
Fail = N - Success
y = as.matrix(shuttle[, c('Success', 'Fail')])
Fit = glm(y ~ x, family = binomial(link = 'logit'))
```

2.3 Parameter Estimation

Since Y_i are independently distributed, the joint density

$$\pi(Y_1, \cdots, Y_N; \beta, \theta) = \prod_{i=1}^{N} f_G(Y_i; \mu_i, \theta),$$

where f_G is the marginal density. Thus, we have

$$\ln L(\beta, \theta; y_1, \cdots, y_N) = \sum_{i=1}^{N} \ln f_G(y_i; \mu_i, \theta)$$

and

$$\widehat{\beta}, \widehat{\theta} = \arg\max_{\beta,\theta} L(\beta, \theta; y_1, \cdots, y_N).$$

2.4 Comparing Nested Models

We want to test $H_0 : \beta_k = C_k, \forall k \in \Omega, \Omega \subset \{1, \cdots, P\}$ against $H_1 : \beta$ is unconstrained. We write $\widehat{\beta}^{(C)}$ as the constrained MLEs under H_0 and

$$2[\ln L(\widehat{\beta}; y) - \ln L(\widehat{\beta}^{(C)}; y)] \sim \chi^2_{|\Omega|}.$$

2.5 Logistic Regression

Suppose $Y_i \sim \text{Bin}(N_i, \mu_i)$ and thus

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{p=1}^{P} X_{ip}\beta_p, \frac{\mu_i}{1-\mu_i} = \prod_{p=1}^{P} e^{\beta_p^{X_{ip}}},$$

where μ_i is the probability, $\ln\left(\frac{\mu_i}{1-\mu_i}\right)$ is a log-odds, and $\frac{\mu_i}{1-\mu_i}$ is an odds. If $\mu_i \approx 0$, then $\mu_i \approx \frac{\mu_i}{1-\mu_i}$.

Suppose $X_{1p} = X_{2p}$ for all p except $X_{2q} = X_{1q} + 1$, then

$$\beta_q = \ln\left(\frac{\mu_2}{1-\mu_2}\right) - \ln\left(\frac{\mu_1}{1-\mu_1}\right), e^{\beta_q} = \left(\frac{\mu_2}{1-\mu_2}\right) \Big/ \left(\frac{\mu_1}{1-\mu_1}\right),$$

where β_q is the log-odds ratio, e^{β_q} is the odds ratio, and $e^{\text{intercept}}$ is baseline odds, when $X_1 = \cdots = X_n = 0$.

2.6 Poisson Regression

2.6.1 Infinitely Divisible

Normal, Poisson, Gamma distribution are infinitely divisible since if $Y_i = \sum_{k=1}^{K} X_{ik}$ and $Y_i \sim D(\theta_1)$, then $X_{ik} \sim D(\theta_2)$. However, uniform, binomial and log-normal distribution are not infinitely divisible since we cannot construct a same distribution from a sum.

2.6.2 Components

The components of a GLM for a count response are:

1. Random component: $Y_i \sim \text{Poisson}(\lambda)$ and $\mathbb{E}[Y] = \text{Var}[Y] = \lambda$.
2. Systematic component: the model matrix and parameters, $X^T\beta$.
3. Link function:

(1) ***Identity link***: $\mu = X^T\beta$, but for binomial data, with an identity link, the model can yield $\mu < 0$ where we only want $\mu \geqslant 0$.

(2) ***Log link*** (the most common and canonical link): $\ln(\mu) = X^T\beta$.

2.6.3 Offsets

Offsets allow us adjust for the time period under consideration, which is the exposure period. An offset term can be thought of a s the log of the time period under study when we are doing Poisson regression with a log link. When using an offset, we often say we are fitting a rate model. In R, we use `offset()` function.

Example 2.2.

$$\ln\left(\frac{\text{Children}}{\text{Month}}\right) = X^T\beta \Rightarrow \ln(\text{Children}) = X^T\beta + \ln(\text{Month}).$$

2.6.4 Model

The Poisson regression is

$$Y_i \sim \text{Poisson}(O_i\mu_i), \ln(\mu_i) = X_i^T\beta,$$

where Y_i is the response variable (number of some event occurred), μ_i is the intensity/rate per time, O_i is the offset term (number of time) and X_i is the intercept.

Or

$$Y_i \sim \text{Poisson}(\rho_i), \ln(\rho_i) = X_i^T\beta + \ln(O_i).$$

In R, for example,

```
model = glm(formula = y ~ offset(logYears) + x1 + x2, family = poisson(link = log),
            data = data)
```

2.7 Gamma Regression

2.7.1 Components

The components of a GLM for a waiting time response are:

1. Random component: $Y_i \sim \text{Gamma}(\phi, v)$ and $\mathbb{E}[Y] = \phi v, \text{Var}[Y] = \phi^2 v$.
2. Systematic component: the model matrix and parameters, $X^T\beta$.
3. Link function:

(1) ***Log link***: $\ln(\mu) = X^T\beta$.

(2) ***Inverse link*** (the canonical link): $\frac{1}{\mu} = \mu^{-1} = X^T\beta$.

Note that ϕ is the range parameter and v is the shape parameter and we have

$$\frac{1}{\sqrt{v}} = \frac{\text{SD}(X)}{\mathbb{E}[X]}.$$

2.7.2 Model

The Gamma regression is

$$Y_i \sim \text{Gamma}\left(\frac{\mu_i}{v}, v\right), \ln(\mu_i) = X_i^T\beta,$$

and `glm` reports a dispersion parameter $\frac{1}{v}$.

2.8 Weibull Regression

Weibull is the standard for event times. In R, `survreg`'s scale is the Weibull shape parameter.

2.9 Gompertz Regression

In R, we use `flexsurvreg(Surv(life) ~ x, data = data, dist = "gompertz")`.

2.10 Comments on GLM

- It is not easy to test which distribution is best, since they are not nested.
- GLMs are not often used with continuous data: they are almost always binomial or Poisson with the notable exception of the Weibull for event times.
- Standard practice is to transform continuous data to normality (logs, Box-Cox).
- Assessing model fit is difficult for binary and count data and residuals do not have nice properties. Histograms can be useful as can exploratory plots.

2.11 Summary

Distribution of Y_i	Range	Use Case	Link Function	Mean Function
Normal	$\mathbb{R}$	Linear response	Identity: $X^T\beta = \mu$	$\mu = X^T\beta$
Binomial	$\{0, 1\}$	Count of 0/1 responses in fixed number of trials	Logit: $X^T\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{e^{X^T\beta}}{1+e^{X^T\beta}}$
Poisson	$\mathbb{W}$	Count of occurrences in fixed time or space	Log: $X^T\beta = \ln(\mu)$	$\mu = e^{X^T\beta}$
Gamma	$\mathbb{R}^+$	Wait times	Inverse: $X^T\beta = \mu^{-1}$	$\mu = (X^T\beta)^{-1}$

3 Linear Mixed Model

If independence assumption is violated, we cannot use linear model, but then we can add a random effect to fit a model:

$$\begin{aligned} Y_{ij}|U_i &\sim \mathcal{N}(\mu_{ij}, \tau^2) \\ \mu_{ij} &= X_{ij}^T\beta + U_i \\ [U_1, \cdots, U_M]^T &\sim \text{MVN}(\mathbf{0}, \Sigma) \end{aligned}$$

where observations Y_{ij} for repeated measures j on individuals i, $X_{ij}^T\beta$ are fixed effects, and U_i are random effects for $i = 1, \cdots, M$.

Equivalently, we can write

$$Y_{ij} = X_{ij}^T\beta + \varepsilon_{ij}$$

where $\varepsilon_{ij} = U_i + Z_{ij}$ and $Z_{ij} \sim \mathcal{N}(0, \tau^2)$, i.e., errors are normally distributed but correlated.

Or

$$Y_i = X_i^T\beta + Z_i b_i + \varepsilon_i$$

where Y_i is vector of outcomes for subject i, X_i and Z_i are model matrices for the fixed and random effects, respectively, the vector β describes the effect of covariates on the expectation of the outcome, the vector b_i is the random effects for unit (assumed to be normally distributed with mean zero), and ε_i is the vector of residual errors, normally distributed with a given variance and the errors within units are mutually independent.

Linear mixed models assume that

1. There is a continuous response variable.
2. We have modeled the dependency structure correctly.
3. Units or subjects are independent, even through observations within each subject are taken not to be.
4. Both the random effects and within-unit residual errors follow normal distributions.
5. The random effects errors and within-unit residual errors have constant variance.

Note that linear mixed models are robust to violations of some of assumptions.

3.1 Random Intercept

Example 3.1. Y_{ij} is the vocal pitch for the ith subject on the jth vocal response, $X_{ij}^T\beta$ has an intercept and an effect for gender and politeness (formal/informal), U_i is subject j's baseline pitch, τ^2 is random chance and potentially other unmeasured counfounders.

We can consider each subject's mean vocal pitch and in model we will assume different random intercepts for each subject. The mixed model estimates these intercepts. The R code is

```
library(lme4)
sub_only = lmer(frequency ~ (1 | subject), data = polite_data)
```

`(1 | subject)` is the R syntax for a random intercept since we assume there is a different intercept for each subject. `1` stands for the intercept and the term to the right of `|` should be a nominal or factor variable to be used for the random effect. The formula means it should expect that there is multiple response per subject and these responses will depend on each subject's baseline level, which effectively resolves the non-independence that stems from having multiple responses by the same subject.

3.2 Random Slope

Example 3.2. We had assumed that the effect of formal/informal conditions were the same for all subjects and thus had one coefficient for this variable. However, the effect of the condition might be different for different subjects. For example, it might be expected that some people are more polite in formal scenarios, others less. Hence, we need a random slope model, where subjects and items are not only allowed to have different intercepts but where they are also allowed to have different slopes for the effect of politeness (i.e., different effects of condition on pitch).

3.3 Generalized Linear Mixed Model

3.3.1 Pros and Cons

Pros:

1. Powerful class of modes that combine the characteristic of generalized linear models and linear mixed models.
2. They can be used with a range of response distributions (Poisson, Binomial, Gamma).
3. They can be used in a range of situations where observations are grouped in some way (not all independent).
4. Fast and can be extended to handle somewhat more complex situations (e.g., zero-inflated Poisson).

Cons:

1. Some of the standard ways we have learned to test models do not apply.
2. Greater risk of making sensible models that are too complex for our data to support.

3.3.2 Assumptions

Assume that

1. Units/Subjects are independent, even though observations within each subject are taken not to be.
2. Random effects come from a normal distribution.
3. Random effects errors and within-unit residual errors have constant variance (variances of data transformed by the like function are homogeneous across categories).
4. The chosen link function is appropriate or the model is correctly specified (responses of transformed data linear w.r.t. continuous predictors).

3.3.3 Model

The generalized linear mixed model is

$$\begin{aligned} Y_{ij}|U &\sim G(\mu_{ij}, \theta) \\ h(\mu_{ij}) &= X_{ij}^T\beta + U_i \\ U &\sim \text{MVN}(\mathbf{0}, \Sigma) \end{aligned}$$

Example 3.3 (Bernoulli).

$$\begin{aligned} Y_{it} &\sim \text{Bernoulli}(\rho_{it}) \\ \text{logit}(\rho_{it}) &= \varepsilon + X_{it}^T\beta + U_i \\ U_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where Y_{it} is presence of bacteria in individual i at time t, μ is the intercept, X_{it} has indicator variables for week and treatment type, and U_i is an individual level random effect, $U_i > 0$ if i is more likely than the average to have the bacteria (allows for within-individual dependence).

Example 3.4 (Poisson).

$$\begin{aligned} Y_i|U &\sim \text{Poisson}(O_i\lambda_i) \\ \ln(\lambda_i) &= X_i^T\beta + U_i \\ U_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where U_i represents woman i's fertility or preference of family size, σ^2 is the extra-Poisson variation or overdispersion.

Example 3.5 (Zero-Inflated Poisson). Let $X \sim \text{Poisson}(\lambda), Y = X$ with probability $(1-\theta)$ and $Y = 0$ with probability θ. $P(Y = 0) = \theta + (1-\theta)e^{-\lambda}$ and $P(Y = y) = (1-\theta)\frac{\lambda^y e^{-\lambda}}{y!}, y \geqslant 1$. The linear mixed model is

$$\begin{aligned} Y_i|U &\sim \text{ZIP}(O_i\lambda_i, \rho) \\ \ln(\lambda_i) &= X_i^T\beta + U_i \\ U_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where U_i represents woman i's fertility or preference of family size, σ^2 is the extra-Poisson variation or overdispersion, and ρ is the proportion of couples which are infertile.

Example 3.6 (Gamma).

$$\begin{aligned} Y_{ij} &\sim \text{Gamma}\left(\frac{\mu_{ij}}{v}, v\right) \\ \ln(\mu_{ij}) &= \beta_0 + t_{ij} + \beta_1 + U_i \\ U_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where Y_{ij} is the weight of pig i at time t_{ij}, $\exp(\beta_0)$ is the population average weight at birth, $\exp(\beta_1)$ is the average proportion weight gain per year, and U_i is pig i's deviation from the population average.

3.4 Case-Control Study

Property 3.1. Let X_i, Y_i, Z_i be the smoking status, cancer status and "in the study" indicators respectively. We want $P(Y_i = y|X_i = x)$. Modeling the case-control data gives us $P(Y_i = y|X_i = x, Z_i = 1)$. We have

$$\frac{\text{odds}(Y_i|X_i = x_1, Z_i = 1)}{\text{odds}(Y_i|X_i = x_0, Z_i = 1)} = \frac{\text{odds}(Y_i|X_i = x_1)}{\text{odds}(Y_i|X_i = x_0)}.$$

Proof. We have

$$P(Y_i|X_i, Z_i = 1) = \frac{P(Y_i, Z_i = 1|X_i)}{P(Z_i = 1|X_i)} = \frac{P(Z_i = 1|Y_i, X_i)P(Y_i|X_i)}{P(Z_i = 1|X_i)}.$$

Assume Z_i is independent of X_i given Y_i, then

$$P(Y_i|X_i, Z_i = 1) = \frac{P(Y_i|X_i)P(Z_i = 1|Y_i)}{P(Z_i = 1|X_i)}$$

and

$$\frac{P(Y_i = 1|X_i, Z_i = 1)}{P(Y_i = 0|X_i, Z_i = 1)} = \frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)} \cdot \frac{P(Z_i = 1|Y_i = 1)}{P(Z_i = 1|Y_i = 0)}.$$

Hence

$$\text{logodds}(Y_i|X_i, Z_i = 1) = \text{logodds}(Y_i|X_i) + C.$$

□

3.4.1 Logistic Model for Case-Control Data

We have

$$P(Y_i = 1|X_i, Z_i = 1) = \lambda_i^*$$
$$\ln\left(\frac{\lambda_i^*}{1-\lambda_i^*}\right) = \beta_0^* + \sum_{p=1}^{P} X_{ip}\beta_p^*$$

By the property,

$$\beta_p^* = \begin{cases} \beta_0 + \ln\left(\frac{P(Z_i=1|Y_i=1)}{P(Z_i=1|Y_i=0)}\right), & p = 0 \\ \beta_p, & p \neq 0 \end{cases},$$

we have

$$P(Y_i = 1|X_i) = \lambda_i$$
$$\ln\left(\frac{\lambda_i}{1-\lambda_i}\right) = \beta_0 + \sum_{p=1}^{P} X_{ip}\beta_p$$

4 Generalized Additive Model

A generalized additive model is a generalized linear model where the linear predictor depends linearly on unknown smooth functions of some predictor variables.

$$g(\mathbb{E}[Y]) = \beta_0 + f_1(x_1) + \cdots + f_m(x_m).$$

We can use the package `mgcv` to fit the GAM, and it uses a rank reduced framework where we replace our unknown smoothing functions with basis expansions:

$$f_j(x_j) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x_j).$$