Gentle introduction to STAT 528

Daniel J. Eck

This is an advanced introduction to generalized linear models and categorical data analysis with a focus on analysing data from disciplines such as biostatistics, education, evolutionary biology, and medicine. In this course you will learn how to conduct scientifically relevant data analyses using methodologically rigorous statistical techniques. For example, we have just experienced nearly three semesters of online learning due to COVID-19. Have you wondered about the impact of online learning? In this class we will explore and estimate the causal effects of online learning using causal effect estimators. In addition we are living in a region where soybean is a major crop. As such we will estimate significant gene markers that improve the photosynthetic process in soybeans using mixed-effects models. These analyses will develop your critical thinking skills as a statistician. We will also place a strong emphasis on statistical properties of presented methods. Furthermore, practical advantages, limitations, and comparisons of methods will be discussed. Here is a brief list of methods and topics that we will study in this course:

- exponential family theory and applications
- categorical variables: possible values are categories
- categorical response variables and generalized linear models (GLM)
- contingency tables
- linear mixed effects models (LMM)
- generalized linear mixed-effects models (GLMM) and generalized estimating equations (GEE)
- aster models for life history analysis
- multivariate linear regression and envelope mehtodology

Categorical variable distributions are necessarily discrete and are not continuous. There are different kinds of categorical variables:

- **binary**: two possible categories (usually 1 denotes a "success" and 0 denotes a "failure")

- **nominal**: no natural ordering of the levels comprising the categorical variable
- **ordinal**: the levels of the categorical variable exhibit a natural ordering (bachelor's, master's, Ph.D.)

Ordinal variables may lack numerical distances between the levels (unlike interval variables) or meaningful ratios (unlike ratio variables).

Important Distributions

We will cover many of these in much more detail throughout the course.

Bernoulli

A random variable $Y \sim \text{Bernoulli}(p)$ has mass function

$$f(y) = \begin{cases} p & y = 1 \\ 1-p & y = 0 \end{cases}$$

where $\text{E}(Y) = p$ and $\text{Var}(Y) = 1 - p$. Here, $0 < p < 1$ is a success probability.

Binomial

A random variable $Y \sim \text{Binomial}(n, p)$ has mass function

$$f(y) = \binom{n}{y} p^y (1-p)^{1-y}, \qquad y = 0, 1, \ldots, n,$$

where $\text{E}(Y) = np$ and $\text{Var}(Y) = np(1-p)$. The Binomial distribution arises as a sum of Bernoulli trials. Let $Z_i, \ldots, Z_n \overset{iid}{\sim} \text{Bernoulli}(p)$. Then $\sum_{i=1}^{n} Z_i \sim \text{Binomial}(n, p)$. **Prove that the Binomial distribution arises as a sum of Bernoulli random variables**.

Also note that:

$$\frac{Y - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1), \qquad \text{as } n \to \infty.$$

Multinomial

Suppose that there are n independent trials, each one resulting in one of c categories, with probability vector $\mathbf{p} = (p_1, \ldots, p_c)$ such that $\sum_{j=1}^{c} p_j = 1$. Let N_j be the total number of observations in category level $j = 1, \ldots, c$, where $\sum_{j=1}^{c} N_j = n$. We will say

$$(N_1, \ldots, N_c) \sim \text{multinomial}(n, \mathbf{p})$$

with mass function

$$f(n_1, \ldots, n_c) = \binom{n}{n_1 \ldots n_c} p_1^{n_1} \cdots p_c^{n_c}, \qquad \sum_{j=1}^{c} n_j = n,$$

where $\mathrm{E}(N_j) = np_j$, $\mathrm{Var}(N_j) = np_j(1-p_j)$ and $(N_j, N_k) = -np_jp_k$ where $j \neq k$.

Poisson

A random variable $Y \sim$ Poisson(μ), $\mu > 0$, has mass function

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}, \qquad y = 0, 1, \ldots,$$

where $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \mu$. Also note that:

$$\frac{Y - \mu}{\sqrt{\mu}} \xrightarrow{d} N(0, 1), \qquad \text{as } n \to \infty.$$

Let $Y_1, \ldots, Y_c$ be independent random variables such that $Y_j \sim$ Poisson(μ_j) then the following conditional relationship holds

$$(Y_1, \ldots, Y_c) \,\Big|\, \sum_{j=1}^{c} Y_j = n \quad \sim \quad \text{multinomial}(n, \mathbf{p}),$$

where

$$p_k = \frac{\mu_k}{\sum_{j=1}^{c} \mu_j}.$$

Likelihoods

For a model with parameter θ, the **likelihood** $L(\theta)$ is the joint density of data at its observed values, as a function of θ, and the **log likelihood** $l(\theta) = \log(L(\theta))$ where $\log$ denotes the natural logarithm. The kernel of $l(\theta)$ includes only the factors that depend on θ. Statistical inference will involve only the kernel, so that $l(\theta)$ need only be specified up to an additive constant.

For our purposes, $l(\theta)$ will be well-defined and at least twice continuously differentiable. The joint density of the data will most often correspond to independent or independent and identically distributed data. Thus the likelihood and log likelihood will be

$$L(\theta) = \prod_{i=1}^{n} f_\theta(y_i), \qquad l(\theta) = \sum_{i=1}^{n} \log(f_\theta(y_1)).$$

A maximum likelihood estimate (MLE) $\hat{\theta}$ maximizes $l(\theta)$ and $L(\theta)$ by monotonicity. The estimate $\hat{\theta}$ is usually the unique solution of $l'(\theta) = 0$. An MLE also maximizes the kernel. The **score function** is

$$u(\theta) = \frac{\partial l(\theta)}{\partial \theta}$$

and the **Fisher information** matrix is

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\right),$$

where the expectation is over the assumed distribution for the data when the parameter value is θ. Note that these quantities can be found even when $l(\theta)$ is known only up to an additive constant. If the data are from a sample of size n, we consider asymptotic behavior as $n \to \infty$. Typically, the inverse Fisher information $I(\theta)^{-1}$ is the asymptotic variance of the MLE $\hat{\theta}$. We can also show that

$$\mathrm{E}(u(\theta)) = 0, \qquad \mathrm{Var}(u(\theta)) = I(\theta),$$

where the expectations are taken over the assumed distribution for the data when the parameter value is θ. **Prove the above.** When the parameter value is θ, $u(\theta)$ is often asymptotically normal (after appropriate standardization).

Exponential families

An *exponential family of distributions* is a parametric statistical model having log likelihood that takes the form

$$l(\theta) = \langle y, \theta \rangle - c(\theta), \tag{1}$$

where y is a vector statistic and θ is a vector parameter, $\langle y, \theta \rangle$ is the usual inner product, and $c(\theta)$ is the cumulant function.

This uses the convention that terms that do not contain the parameter vector can be dropped from a log likelihood; otherwise additional terms also appear in (1). The distributions mentioned in this Introduction can be rewritten as exponential families. Therefore exponential families constitute extremely useful classes of statistical distributions.

The mathematically simple representation of the log likelihood of an exponential family is very powerful. For example, we can derive important quantities directly from the cumulant function:

$$\mathrm{E}_\theta(Y) = \nabla c(\theta),$$
$$\mathrm{Var}_\theta(Y) = \nabla^2 c(\theta),$$

provided that θ is an interior point of the parameter space.

Generalized linear models

Generalized linear models (GLM) are an extension to linear regression to other classes of distributions. In this class we will place emphasis on GLMs corresponding to exponential families. Recall that in linear regression one is interested in estimated β, a regression coefficient vector in standard terminology, where

$$y = x^T\beta + \varepsilon, \qquad \text{and} \qquad \mathrm{E}(y|x) = x^T\beta$$

where ε is statistical error (usually assumed to follow a normal distribution) and β is a regression coefficient vector. In more general models we cannot always express the mean function as a linear function of covariates in a sensible fashion. However, all is not lost. In an exponential family GLM, the parameter vector θ is "linked" to the mean value parameter $E(y|x)$ through a change-of-parameter mappings $g(\theta)$ (called the inverse link function). We can reparameterize $\theta_i = x_i^T\beta$ where β is a lower dimension vector of regression coefficients and write

$$\mathrm{E}_\theta(y_i|x_i) = \mu_x = g(x_i^T\beta)$$

which implies that we can write

$$g^{-1}\left(\mathrm{E}_\theta(y_i|x_i)\right) = x_i^T\beta.$$

Hence the name generalized linear model. These models have very nice statistical properties when the underlying model is an exponential family. Note that most treatments of GLM will define g as the "link function" while we define g as a change-of-parameters map from canonical parameters to mean-value parameters. This is because we will build the exponential family theoretical foundations that underpins GLMs before we formally introduce GLMs.

Linear mixed effects models

Linear mixed effects models (LMM) are an extension to the classical linear model in which units are thought to be realizations from a super populations that has a lower-dimensional parametric form. Therefore, units have random effects as well as the standard fixed effects that we are accustomed to seeing in regression modeling. The basic LMM takes the form

$$Y = X\beta + Zb + \varepsilon \qquad \text{or} \qquad Y \mid b \sim N(X\beta + Zb, \sigma^2 I),$$

where Y is the response vector, X is a fixed-effects model matrix, β is a fixed-effects coefficient vector, Z is a model matrix of random-effects, b is a vector of random effects, and σ^2 is the variance of the error distribution. If we further assume that $b \sim N(0, \sigma^2 D)$ then the unconditional response Y is distributed as

$$Y \sim N(X\beta, \sigma^2(I + ZDZ^T)).$$

Multivariate regression model

Multivariate regression models are an extension to the classical linear model in which the response is now a vector and elements of the response are allowed to be correlated. The basic multivariate regression model takes the form

$$Y = \alpha + \beta X + \varepsilon,$$

where Y is the response vector, α is an intercept vector, X is a predictor vector, β is a coefficient matrix, and $\varepsilon \sim N(0, \Sigma)$ where $\Sigma > 0$. The multivariate regression model is very similar to the standard linear regression model with a scalar response. But estimation of important quantities requires matrix calculus which is not a part of standard statistics curricula.

Acknowledgments

This course is developed largely from Agresti [2013], Faraway [2016], Trevor Park's STAT 426 notes, Charles Geyer's notes on exponential families, and other topics.

References

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2013.

J. J. Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2016.