STATS305B – Lecture 1

Jonathon Taylor
Scribed by Michael Howes

01/03/22

Contents

1 Announcements **1**

2 Course overview **1**
2.1 Models for discrete data . . . 1
2.2 Regression . . . 2
2.3 Survival analysis . . . 2

3 Distributions **3**
3.1 Multinomial and Poisson . . . 3
3.2 Exponential families . . . 3

1 Announcements

- The course webpage can be viewed at http://web.stanford.edu/class/stats305b/intro.html.
- Jonathon's office hours are 2 – 4 pm on Wednesdays.

2 Course overview

Here are some brief descriptions of the main topics we'll cover.

2.1 Models for discrete data

Suppose $X_k \overset{\text{iid}}{\sim} F$ for $1 \leq k \leq N$. Let $R(X_k)$ and $C(X_k)$ be discrete random variables (R for row, C for column). For example, $R(X_k), C(X_k)$ may be $\{0, 1\}$ valued and record the presence/absence of a trait or $R(X_k), C(X_k)$ may take more than one value and record the label of a trait.

Suppose $R(X_k)$ can take values $R_1, R_2, \ldots, R_I$ and $C(X_k)$ can take values $C_1, \ldots, C_J$. Let $Y_{i,j}$ be the number of times $R(X_k)$ takes value R_i and $C(X_k)$ takes value C_j. We can record these counts in a table

	C_1	C_2	$\ldots$	C_J	Row total
R_1	Y_{11}	Y_{12}	$\ldots$	Y_{1J}	$Y_{1\cdot}$
R_2	Y_{21}	Y_{22}	$\ldots$	Y_{2J}	$Y_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
R_I	Y_{I1}	Y_{I2}	$\ldots$	Y_{IJ}	$Y_{I\cdot}$
Column total	$Y_{\cdot 1}$	$Y_{\cdot 2}$	$\ldots$	$Y_{\cdot J}$	$Y_{\cdot\cdot}$ = Grand total = N

We will often write i for C_i and j for R_j. The notation $Y_{i\cdot}$ and $Y_{\cdot j}$ is useful shorthand to mean the variable replaced with $\cdot$ has been marginalized. The distribution of the above table is described by

$$\pi_{ij} = P_F(R = i, C = j), \text{ for } 1 \leq i \leq I \text{ and } 1 \leq j \leq J.$$

Some common questions we might ask about the distribution are:

1. Independence: $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for all $1 \leq i \leq I$, $1 \leq j \leq J$.
2. Homogeneity (I): If we didn't sample the rows randomly but instead took multiple samples of C from different populations, then π_{ij} doesn't make sense since R is not random. We instead have I different distributions for $\pi_{\cdot j}$. We could ask if they are all the same.
3. Homogeneity (II): suppose $I = J$ and the values R_i, C_i are common. We could then ask if $\pi_{i\cdot} = \pi_{\cdot i}$ for $1 \leq i \leq I$.

Some common models for this sort of data are the multinomial model and the Poisson model. We will see that the multinomial and Poisson models are connected.

2.2 Regression

In standard linear regression from 305A we have $Y \in \mathbb{R}^{n\times q}$ and $X \in \mathbb{R}^{n\times p}$, and we modelled (X_i, Y_i) as independent draws with $Y_i|X_i \sim \mathsf{N}(X_i^T\beta, \sigma^2)$. We estimated β with $\widehat{\beta} = \text{argmin}_b \|Xb - y\|_2^2$. In this course we will consider extensions of this where the response Y is not real valued.

1. Binary response $Y \in \{0,1\}^n$. $Y_i|X_i \sim \mathsf{Bernoulli}(F(X_i^T\beta))$ where F is a CDF. For example,
   - Probit: $F \sim \mathsf{N}(0,1)$.
   - Logit: $F(x) = \frac{e^x}{1+e^x}$.

   Once F is fixed we can work with the likelihood function to choose β. We can either do maximum likelihood or we can include some sort of penalty term. The maximum likelihood estimator is

   $$\widehat{\beta} = \underset{\beta}{\text{argmin}} -2\log(L(\beta|X,Y)) = \underset{\beta}{\text{argmin}}\, \text{DEV}(\beta|X,Y),$$

   where L is the likelihood function and DEV is the deviance (to be defined later). Some natural questions to ask are
   - What is the (asymptotic) distribution of $\widehat{\beta}$?
   - Can we use this asymptotic distribution to do inference?
2. Multinomial $Y \in \{1,\ldots,k\}^n$. Some models we'll use are baseline logistic and order logistic.

2.3 Survival analysis

Let T be a survival time (simply a non-negative random variable). We will model T via it's survival function $P(T > t) = 1 - \text{CDF}$. One model is via hazard functions where we have

$$P_\beta(T > t|X) = \exp\left(-\int_0^t h_\beta(s;X)ds\right).$$

There are non-parametric methods and also the *Cox proportional hazards model* where

$$\frac{h_\beta(s;X_1)}{h_\beta(s;X_0)} = \exp\left((X_1 - X_0)^T\beta\right) = \frac{\exp(X_1^T\beta)}{\exp(X_0^T\beta)}.$$

We will consider some complications like censoring (aka truncation). Censoring can occur when conducting a study with an end date. The end date may pass before we have viewed the survival time of some subjects. We have partial information (the survival time is greater than the end date of the study) but we do not have full information (the exact survival time). How can we use the partial information?

3 Distributions

3.1 Multinomial and Poisson

Some distributions that we will use in this class are:

- Binomial (2 classes): $Y \sim \mathsf{Binomal}(n, \pi)$, $\pi \in [0,1]$, $n = 1, 2, \ldots$. Then $Y \in \{0, 1, \ldots, n\}$ and $\mathbb{P}(Y = j) = \binom{n}{j}\pi^j(1-\pi)^{n-j}$.
- Multinomial (k classes): $Y \sim \mathsf{Multinomial}(n, \pi)$, $\pi = (\pi_1, \ldots, \pi_k)$, $\pi_i \geq 0$ and $\sum_{i=1}^k \pi_i = 1$, then $Y \in \{(Y_1, \ldots, Y_k) \in \mathbb{Z}^{k,+} : \sum_{i=1}^k Y_i = n\}$ and
$$\mathbb{P}(Y = (y_1, \ldots, y_k)) = \binom{n}{(y_1, \ldots, y_k)} \prod_{i=1}^k \pi_i^{y_i} = \frac{n!}{\prod_{i=1}^k y_i!} \prod_{i=1}^k \pi_i^{y_i}.$$
- Poisson (unbounded count): $Y \sim \mathsf{Poisson}(\lambda)$ $\lambda > 0$, then $Y \in \mathbb{Z}^+ = \{0, 1, 2, \ldots\}$ and
$$\mathbb{P}(Y = j) = \frac{e^{-\lambda}\lambda^j}{j!}.$$

The multinomial and Poisson distributions have the following relationship. Suppose $Y_1, \ldots, Y_k$ are independent and $Y_i \sim \mathsf{Poisson}(\lambda_i)$. Then $\sum_{i=1}^k Y_i \sim \mathsf{Poisson}(\sum_{i=1}^k \lambda_i)$ and
$$(Y_1, \ldots, Y_k) | \sum_{i=1}^k Y_i \sim \mathsf{Multinomial}(N, \pi),$$
where $N = \sum_{i=1}^k Y_i$ and $\pi_i = \frac{\lambda_i}{\sum_{j=1}^k \lambda_j}$.

3.2 Exponential families

A family of distribution P_η are said to be an *exponential family* if
$$\frac{dP_\eta(y)}{dm} = e^{\eta^T S(y) - \Lambda(\eta)},$$
where,

- m is a measure called the *carrier* or *reference*.
- η are called the *natural parameters*.
- $S(y)$ are called the *sufficient statistics*.
- $\Lambda(\eta)$ is called the *cummulant generating function* of S.

Since the density for an exponential family integrates to 1, we must have,
$$\Lambda(\eta) = \log\left(\int e^{\eta^T S(y)} dm\right).$$

Remark 1. Often we will have $\{\eta : \Lambda(\eta) < \infty\} = \mathbb{R}^{\dim(\eta)}$ which makes optimizing the natural parameters η easier than optimizing some other parameters which may be constrained.

Examples 1. We have already seen examples of exponential families:

- Binomial:
$$P_\pi(j) = \binom{n}{j}\pi^j(1-\pi)^{n-j} = \binom{n}{j}\left(\frac{\pi}{1-\pi}\right)^j(1-\pi)^n.$$
To turn this into an exponential family, let $m = \binom{n}{j}$ with respect to counting measure on $\{0, 1, \ldots, n\}$ and let $\eta = \log(\pi/(1-\pi))$, so $\pi = \frac{e^\eta}{1+e^\eta}$. Also let $S(j) = j$ and $e^{-\Lambda(\eta)} = (1-\pi)^n$ so
$$\Lambda(\eta) = -n\log(1-\pi) = -n\log\left(\frac{1}{1+e^\eta}\right) = n\log(1+e^\eta).$$
Now rather than $\pi \in (0,1)$ we have $\eta \in \mathbb{R}$. This makes it easy to do optimization once we have a model. The exponential family automatically suggests a model. Suppose we have (Y_i, X_i) and we want $Y_i|X_i \sim \mathsf{Bernoulli}(\pi(X_i))$. We can use the natural parameters η and make a model $Y_i|X_i \sim \mathsf{Bernoulli}(\eta(X_i))$. Since η is unconstrained we can take $\eta(X_i) = X_i^T\beta$, giving us a linear model. Under this model we have
$$\mathbb{P}(Y_i = 1|X_i) = \frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}} = F(X_i^T\beta),$$
where F is the logistic distribution function. Thus, one way to derive *logistic regression* is by working with the Bernoulli distribution after writing it as an exponential family.

- Poisson:
$$\mathbb{P}(Y = j) = \frac{\lambda^j e^{-\lambda}}{j!}.$$
This is an exponential family with $m = \frac{1}{j!}$ with respect to counting measure on $\mathbb{Z}^+ = \{0, 1, \ldots\}$, $S(j) = j$, $\eta = \log(\lambda)$ and $\Lambda(\eta) = \lambda = e^\eta$. As before, we can use this to make a model. Suppose we want $Y_i|X_i \sim \mathsf{Poisson}(\lambda(X_i))$. We could again try $\eta(X_i) = X_i^T\beta$ in which can we would have $\lambda(X_i) = e^{X_i^T\beta}$. This is the *log-linear model*. Both logistic regression and the log linear model are canonical examples of generalized linear models. They are also canonical generalized linear models in a technical sense which we will discuss later in the course.

- Multinomial:
$$\mathbb{P}(Y = y) = \binom{n}{(y_1, \ldots, y_k)}\prod_{i=1}^k \pi_i^{y_i}.$$
To write this as an exponential family take $m = \binom{n}{(y_1,\ldots,y_k)}$ with respect to counting measure on $\{y \in \mathbb{Z}^{k,+} : \sum_{i=1}^k y_i = n\}$. Let $S(y) = y \in \mathbb{Z}^{k,+}$, $\eta = (\log(\pi_1), \ldots, \log(\pi_k)) \in \mathbb{R}^k$ and so $\pi_i = e^{\eta_i}$ and
$$e^{-\Lambda(\eta)} = \left(\frac{1}{\pi_1 + \ldots + \pi_k}\right)^n = \left(\frac{1}{e^{\eta_1} + \ldots + e^{\eta_k}}\right)^n.$$
Note that if $\eta' = \eta + \delta\mathbf{1}$, then η' and η give the same probability distribution. Thus, the model is unidentifiable. We really only have $k-1$ natural parameters. One work around is to look at the law of $(y_1, \ldots, y_{k-1})$ (this is what we do with the binomial distribution). In this case, m is $\binom{n}{(y_1,\ldots,y_{k-1},n-y_1-\ldots-y_{k-1})}$ with respect to counting measure on $\{y \in \mathbb{Z}^{k-1,+} : \sum_{i=1}^{k-1} y_i \le n\}$. The sufficient statistic is $S(y) = y$ and $\eta_i = \log(\pi_i/\pi_k)$ for $i = 1, \ldots, k-1$ and so
$$\pi_i = \begin{cases} \frac{e^{\eta_i}}{1+\sum_{j=1}^{k-1} e^{\eta_j}} & \text{if } i < k, \\ \frac{1}{1+\sum_{j=1}^{k-1} e^{\eta_j}} & \text{if } i = k. \end{cases}$$

The model is then $Y_s|X_s \sim \mathsf{Multinomial}(1, \pi(X_s))$ where $\pi(X_s)$ can be described by $\eta_s = \eta(X_s)$ where $\eta_s \in \mathbb{R}^{k-1}$ and $\eta_{s,i} = X_s^T \beta_i$ and so $(\eta)_{N\times(k-1)} = X_{N\times p}\beta_{p\times(k-1)}$ where N is the number of subjects, p is the number of features and k is the number of classes. This model is called *baseline logistic regression.*