Unit 1 Lecture 1

Eugene Katsevich

August 31, 2021

1 Announcements

- See the Syllabus for course information and logistics.
- Homework 0 (just to get computational tools set up) is due this Wednesday, September 1. It will be submitted but not graded. For those who are new to either Git/Github or R/RStudio, you are welcome to attend the STAT 471 computing tutorial (at 5:15-6:45pm in JMHH 360). Its recording will also be made available on Canvas.
- Office hours are starting this week.
- Homework 1 is now out and due September 13 at 11:59pm.

2 Introduction to linear models and GLMs

2.1 Introduction (Agresti 1.1)

The overarching statistical goal addressed in this class is to learn about relationships between a response y and predictors $x_1, x_2, \ldots, x_p$. This abstract formulation encompasses an extremely wide variety of applications. The most widely used set of statistical models to address such problems are *generalized linear models*, which are the focus of this class.

Let's start by recalling the *linear model*, the most fundamental of the generalized linear models. In this case, the response is continuous ($y \in \mathbb{R}$) and modeled as

$$y = \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \tag{1}$$

where

$$\epsilon \sim (0, \sigma^2), \quad \text{i.e. } \mathbb{E}[\epsilon] = 0 \text{ and } \mathrm{Var}[\epsilon] = \sigma^2. \tag{2}$$

We view the predictors $x_1, \ldots, x_p$ as fixed, so the only source of randomness in y is ϵ. Another way of writing the linear model is

$$\mu \equiv \mathbb{E}[y] = \beta_1 x_1 + \cdots + \beta_p x_p \equiv \eta.$$

Not all responses are continuous, however. In some cases, we have binary responses ($y \in \{0, 1\}$) or count responses ($y \in \mathbb{Z}$). In these cases, there is a mismatch between the

$$\textit{linear predictor } \eta \equiv \beta_1 x_1 + \cdots + \beta_p x_p$$

and the

$$\textit{mean response } \mu \equiv \mathbb{E}[y].$$

The linear predictor can take arbitrary real values ($\eta \in \mathbb{R}$), but the mean response can lie in a restricted range, depending on the response type. For example, $\mu \in [0, 1]$ for binary y and $\mu \in [0, \infty)$ for count y.

For these kinds of responses, it makes sense to model a *transformation* of the mean as linear, rather than the mean itself:

$$g(\mu) = g(\mathbb{E}[y]) = \beta_1 x_1 + \cdots + \beta_p x_p = \eta. \tag{3}$$

This transformation g is called the link function. For binary y, a common choice of link function is the *logit link*, which transforms a probability into a log-odds:

$$\text{logit}(\pi) \equiv \log \frac{\pi}{1-\pi}.$$

So the predictors contribute linearly on the log-odds scale rather than on the probability scale. For count y, a common choice of link function is the *log link*.

Models of the form (3) are called *generalized linear models* (GLMs). They specialize to linear models for identity link function, i.e. $g(\mu) = \mu$. The focus of this course are methodologies to learn about the coefficients $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)^T$ of a GLM based on a sample $(\boldsymbol{X}, \boldsymbol{y}) \equiv \{(x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1}^n$ drawn from this distribution. Learning about the coefficient vector helps us learn about the relationship between the response and the predictors. This course is (tentatively) broken up into six units.

- **Unit 1. Linear model: Estimation.** The *least squares* point estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ based on a dataset $(\boldsymbol{X}, \boldsymbol{y})$ under the linear model assumptions (1) and (2).
- **Unit 2. Linear model: Inference.** Under the additional assumption that $\epsilon \sim N(0, \sigma^2)$, how to carry out statistical inference (hypothesis testing and confidence intervals) for the coefficients.
- **Unit 3. Linear model: Misspecification.** What to do when the linear model assumptions are not correct: What issues can arise, how to diagnose them, and how to fix them.
- **Unit 4. GLMs: General theory.** Estimation and inference for GLMs (generalizing Units 1 and 2). GLMs fit neatly into a unified theory based on *exponential families*.
- **Unit 5. GLMs: Special cases.** Looking more closely at the most important special cases of GLMs, including logistic regression and Poisson regression.
- **Unit 6. Further topics.** Linear mixed models (extending linear models to situations where correlations exist among samples); penalized GLMs (extending GLMs to situations where there are more predictors than samples); multiple testing (how to correct for multiplicity when testing many hypotheses—in GLMs or otherwise).

We will use the following notations in this course. Vector and matrix quantities will be bolded, whereas scalar quantities will not be. Capital letters will be used for matrices, and lowercase for vectors and scalars. No notational distinction will be made between random quantities and their realizations. The letters $i = 1, \ldots, n$ and $j = 1, \ldots, p$ will index samples and predictors, respectively. The predictors $\{x_{ij}\}_{i,j}$ will be gathered into an $n \times p$ matrix $\boldsymbol{X}$. The rows of $\boldsymbol{X}$ correspond to samples, with the ith row denoted $\boldsymbol{x}_{i*}$. The columns of $\boldsymbol{X}$ correspond to predictors, with the jth column denoted $\boldsymbol{x}_{*j}$. The responses $\{y_i\}_i$ will be gathered into an $n \times 1$ response vector $\boldsymbol{y}$. The notation $\equiv$ will be used for definitions.

2.2 Types of predictors; interpreting linear model coefficients (Agresti 1.2)

The types of predictors x_j (e.g. binary or continuous) has less of an effect on the regression than the type of response, but it is still important to pay attention to the former.

Intercepts. It is common to include an *intercept* in a linear regression model, a predictor x_0 such that $x_{i0} = 1$ for all i. When an intercept is present, we index it as the 0th predictor. The simplest kind of linear model is the *intercept-only model* or the *one-sample model*:

$$y = \beta_0 + \epsilon. \tag{4}$$

The parameter β_0 is the mean of the response.

Binary predictors. In addition to an intercept, suppose we have a binary predictor $x_1 \in \{0, 1\}$ (e.g. $x_1 = 1$ for patients who took blood pressure medication and $x_1 = 0$ for those who didn't). This leads to the following linear model:

$$y = \beta_0 + \beta_1 x_1 + \epsilon. \tag{5}$$

Here, β_0 is the mean response (say blood pressure) for observations with $x_1 = 0$ and $\beta_0 + \beta_1$ is the mean response for observations with $x_1 = 1$. Therefore, the parameter β_1 is the difference in mean response between observations with $x_1 = 1$ and $x_1 = 0$. This parameter is sometimes called the *effect* or *effect size* of x_1, though a causal relationship might or might not be present. The model (5) is sometimes called the *two-sample model*, because the response data can be split into two "samples": those corresponding to $x_1 = 0$ and those corresponding to $x_1 = 1$.

Categorical predictors. A binary predictor is a special case of a categorical predictor: A predictor taking two or more discrete values. Suppose we have a predictor $w \in \{w_0, w_1, \ldots, w_{C-1}\}$, where $C \geq 2$ is the number of categories and $w_0, \ldots, w_{C-1}$ are the *levels* of w. E.g. suppose $\{w_0, \ldots, w_{C-1}\}$ is the collection of U.S. states, so that $C = 50$. If we want to regress a response on the categorical predictor w, we cannot simply set $x_1 = w$ in the context of the linear regression (5). Indeed, w does not necessarily take numerical values. Instead, we need to add a predictor x_j for each of the levels of w. In particular, define $x_j \equiv \mathbb{1}(w = w_j)$ for $j = 1, \ldots, C - 1$ and consider the regression

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{C-1} x_{C-1} + \epsilon. \tag{6}$$

Here, category 0 is the *base category*, and β_0 represents the mean response in the base category. The coefficient β_j represents the difference in mean response between the jth category and the base category.

Quantitative predictors. A quantitative predictor is one that can take on any real value. For example, suppose that $x_1 \in \mathbb{R}$, and consider the linear model

$$y = \beta_0 + \beta_1 x_1 + \epsilon. \tag{7}$$

Now, the interpretation of β_1 is that an increase in x_1 by 1 is associated with an increase in y by β_1. We must be careful to avoid saying "an increase in x_1 by 1 *causes* y to increase by β_1" unless we make additional causal assumptions. Note that the units of x_1 matter. If x_1 is the height of a person, then the value and the interpretation of β_1 changes depending on whether that height is measured in feet or in meters.

Ordinal predictors. There is an awkward category of predictor in between categorical and continuous called *ordinal*. An ordinal predictor is one that takes a discrete number of values, but these values have an intrinsic ordering, e.g. $x_1 \in \{\texttt{small}, \texttt{medium}, \texttt{large}\}$. It can be treated as categorical at the cost of losing the ordering information, or as continuous if one is willing to assign quantatitive values to each category.

Multiple predictors. A linear regression need not contain just one predictor (aside from an intercept). For example, let's say x_1 and x_2 are two predictors. Then, a linear model with both predictors is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon. \tag{8}$$

When there are multiple predictors, the interpretation of coefficients must be revised somewhat. For example, β_1 in the above regression is the effect of an increase in x_1 by 1 *while holding* x_2 *constant* or *while adjusting for* x_2 or *while controlling for* x_2. If y is blood pressure, x_1 is a binary predictor indicating blood pressure medication taken and x_2 is sex, then β_1 is the effect of the medication on blood pressure while controlling for sex. In general, the coefficient of a predictor depends on what other predictors are in the model. As an extreme case, suppose the medication has no actual effect, but that men generally have higher blood pressure and higher rates of taking the medication. Then, the coefficient β_1 in the single regression model (5) would be nonzero but the coefficient in the multiple regression model (8) would be equal to zero. In this case, sex acts as a *confounder*.

Interactions. Note that the multiple regression model (8) has the built-in assumption that the effect of x_1 on y is the same for any fixed value of x_2 (and vice versa). In some cases, the effect of one variable on the response may depend on the value of another variable. In this case, it's appropriate to add another predictor called an *interaction*. Suppose x_2 is quantitative (e.g. years of job experience) and x_2 is binary (e.g. sex, with $x_2 = 1$ meaning male). Then, we can define a third predictor x_3 as the product of the first two, i.e. $x_3 = x_1 x_2$. This gives the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon. \tag{9}$$

Now, the effect of adding another year of job experience is β_1 for females and $\beta_1 + \beta_3$ for males. The coefficient β_3 is the difference in the effect of job experience between males and females.

2.3 Model matrices, model vectors spaces, and identifiability (Agresti 1.3-1.4)

The matrix $\boldsymbol{X}$ is called the *model matrix* or the *design matrix*. Concatenating the linear model equations (1) and (2) across observations give us an equivalent formulation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \ \mathrm{Var}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}_n$$

or

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\eta}.$$

As $\boldsymbol{\beta}$ varies in $\mathbb{R}^p$, the set of possible vectors $\boldsymbol{\eta} \in \mathbb{R}^n$ is defined

$$C(\boldsymbol{X}) \equiv \{\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\}.$$

$C(\boldsymbol{X})$, called the *model vector space*, is a subspace of $\mathbb{R}^n$: $C(\boldsymbol{X}) \subseteq \mathbb{R}^n$. Since

$$\boldsymbol{X}\boldsymbol{\beta} = \beta_1 \boldsymbol{x}_{*1} + \cdots + \beta_p \boldsymbol{x}_{*p},$$

the model vector space is the column space of the matrix $\boldsymbol{X}$.

The *dimension* of $C(\boldsymbol{X})$ is the rank of $\boldsymbol{X}$, i.e. the number of linearly independent columns of $\boldsymbol{X}$. If $\text{rank}(\boldsymbol{X}) < p$, this means that there are two different vectors $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ such that $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta}'$. Therefore, we have two values of the parameter vector that give the same model for $\boldsymbol{y}$. This makes $\boldsymbol{\beta}$ *not identifiable*, and makes it impossible to reliably determine $\boldsymbol{\beta}$ based on the data. For this reason, we will generally assume that $\boldsymbol{\beta}$ is *identifiable*, i.e. $\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{X}\boldsymbol{\beta}'$ if $\boldsymbol{\beta} \neq \boldsymbol{\beta}'$. This is equivalent to the assumption that $\text{rank}(\boldsymbol{X}) = p$. Note that this cannot hold when $p > n$, so for the majority of the course we will assume that $p \leq n$. In this case, $\text{rank}(\boldsymbol{X}) = p$ if and only if $\boldsymbol{X}$ has *full-rank*.

As an example when $p \leq n$ but when $\boldsymbol{\beta}$ is still not identifiable, consider the case of a categorical predictor. Suppose the categories of w were $\{w_1, \ldots, w_{C-1}\}$, i.e. the baseline category w_0 did not exist. In this case, the model (6) would not be identifiable because $x_0 = 1 = x_1 + \cdots + x_{C-1}$ and thus $x_{*0} = 1 = x_{*1} + \cdots + x_{*,C-1}$. Indeed, this means that one of the predictors can be expressed as a linear combination of the others, so $\boldsymbol{X}$ cannot have full rank. A simpler way of phrasing the problem is that we are describing $C-1$ intrinsic parameters (the means in each of the $C-1$ groups) with C model parameters. There must therefore be some redundancy. For this reason, if we include an intercept term in the model then we must designate one of our categories as the baseline and exclude its indicator from the model.

2.4 Least squares estimation (Agresti 2.1.1)

Now, suppose that we are given a dataset $(\boldsymbol{X}, \boldsymbol{y})$. How do we go about estimating $\boldsymbol{\beta}$ based on this data? The canonical approach is the *method of least squares*:

$$\widehat{\boldsymbol{\beta}} \equiv \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2. \tag{10}$$

The quantity

$$\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 = \sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2 \tag{11}$$

is called the *residual sum of squares (RSS)*, and it measures the lack of fit of the linear regression model. We therefore want to choose $\widehat{\boldsymbol{\beta}}$ to minimize this lack of fit. Note that if $\boldsymbol{\epsilon}$ is assumed to be $N(0, \sigma^2\boldsymbol{I}_n)$, then the least squares solution would also be the maximum likelihood solution. Indeed, for $y_i \sim N(\mu_i, \sigma^2)$, the log-likelihood is

$$\log\left[\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)\right] = \text{constant} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2.$$

Letting $L(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, we can do some calculus to derive that

$$\frac{\partial}{\partial\boldsymbol{\beta}}L(\boldsymbol{\beta}) = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{12}$$

Setting this vector of partial derivatives equal to zero, we arrive at the *normal equations*:

$$-\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0 \quad \Longleftrightarrow \quad \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}.$$

If $\boldsymbol{X}$ is full rank, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is invertible and we can therefore conclude that

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

3 R demo (Agresti 2.6)

The R demo will be based on the `ScotsRaces` data from the textbook. Data description (quoted from the textbook):

> "Each year the Scottish Hill Runners Association publishes a list of hill races in Scotland for the year. The table below shows data on the record time for some of the races (in minutes). Explanatory variables listed are the distance of the race (in miles) and the cumulative climb (in thousands of feet)."

```
library(tidyverse)
```

```
# read the data into R
scots_races = read_tsv("../../data/ScotsRaces.dat", col_types = "cddd")
scots_races
```

```
## # A tibble: 35 x 4
##    race                   distance climb  time
##    <chr>                     <dbl> <dbl> <dbl>
##  1 GreenmantleNewYearDash      2.5  0.65  16.1
##  2 Carnethy5HillRace           6    2.5   48.4
##  3 CraigDunainHillRace         6    0.9   33.6
##  4 BenRhaHillRace              7.5  0.8   45.6
##  5 BenLomondHillRace           8    3.07  62.3
##  6 GoatfellHillRace            8    2.87  73.2
##  7 BensofJuraFellRace         16    7.5  205.
##  8 CairnpappleHillRace         6    0.8   36.4
##  9 ScoltyHillRace              5    0.8   29.8
## 10 TraprainLawRace             6    0.65  39.8
## # ... with 25 more rows
```

```
# Exploration

# pairs plot

# Q: What are the typical ranges of the variables?
# Q: What are the relationships among the variables?


# mile time versus distance

# Q: How does mile time vary with distance?
# Q: What races deviate from this trend?
# Q: How does climb play into it?
```

```
# Linear model

# Q: What is the effect of an extra mile of distance on time?
```

```
# Linear model with interaction

# Q: What is the effect of an extra mile of distance on time
#  for a run with low climb?

# Q: What is the effect of an extra mile of distance on time
#  for a run with high climb?
```