Lab 2: Notation, Properties of Random Variables, and a Proof

Robin Mejia

Email: *rjmejia@berkeley.edu*

Office Hours: **Thurs, 9am-10:45am, 109 Haviland**

September 9, 2015

Logistic Regression Reminders

- We use logistic regression for binary outcomes $Y \in \{0, 1\}$, so here $\mathbb{E}(Y|X_1 = x_1, X_2 = x_2) = \mathbb{P}(Y = 1|X_1 = x_1, X_2 = x_2)$.

- Model of regression:
$logit[\mathbb{E}(Y|X_1 = x_1, X_2 = x_2)] = logit[\mathbb{P}(Y = 1|X_1 = x_1, X_2 = x_2)] = \log\left(\frac{\mathbb{P}(Y=1|X_1=x_1,X_2=x_2)}{1-\mathbb{P}(Y=1|X_1=x_1,X_2=x_2)}\right) = \log\left(\frac{\mathbb{P}(Y=1|X_1=x_1,X_2=x_2)}{\mathbb{P}(Y=0|X_1=x_1,X_2=x_2)}\right)$

Logistic Regression Reminders

- Expit function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

- Logit function:

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

- Interpret coefficients as odds ratios: how do you get back to an odds ratio from a logistic regression? How do you get a predicted value?

Question 1.2, Jewell and Hubbard

Suppose a study followed a simple random sample of ten individuals for approximately six months. During clinic visits over the follow-up period, blood pressure measurements (diastolic, a continuous variable) were taken, and a simple questionnaire related to issues of stress was administered. The questions regarding stress were summarized by a simple binary score (High stress $= 1$; Low stress $= 0$). The number of visits to the clinic during follow-up varied among individuals: 3 came only once, 6 came twice, and 1 came four times.

Questions

Using the notation introduced in this chapter:

1. What is the value of m?
2. What are the values of n_i for each i?
3. Symbolically, write down the entire data set (order the individuals from those with fewest to most visits) treating the blood pressure measurement as the outcome variable of interest–be specific about the dimension of the vectors and matrices you use.

Solutions

1. $m = 10$, the number of individuals in the study

2. For $i = \{1, 2, 3\}, n_i = 1$
   For $i = \{4, \ldots, 9\}, n_i = 2$
   For $i = 10, n_i = 4$

Solutions

3. The observed data looks like

$$\begin{bmatrix} i & j & X & Y \\ 1 & 1 & X_{11} & Y_{11} \\ 2 & 1 & X_{21} & Y_{21} \\ 3 & 1 & X_{31} & Y_{31} \\ 4 & 1 & X_{41} & Y_{41} \\ 4 & 2 & X_{42} & Y_{42} \\ 5 & 1 & X_{51} & Y_{51} \\ 5 & 2 & X_{52} & Y_{52} \\ 6 & 1 & X_{61} & Y_{61} \\ 6 & 2 & X_{62} & Y_{62} \\ 7 & 1 & X_{71} & Y_{71} \\ 7 & 2 & X_{72} & Y_{72} \\ 8 & 1 & X_{81} & Y_{81} \\ 8 & 2 & X_{82} & Y_{82} \\ 9 & 1 & X_{91} & Y_{91} \\ 9 & 2 & X_{92} & Y_{92} \\ 10 & 1 & X_{101} & Y_{101} \\ 10 & 2 & X_{102} & Y_{102} \\ 10 & 3 & X_{103} & Y_{103} \\ 10 & 4 & X_{104} & Y_{104} \end{bmatrix}$$

Continued #3

So we can say the data structure is

- Number of people: $i \in \{1, \ldots, 10\}$
- Number of visits to the clinic per person: $n_i \in \{1, 2, 4\}$
- Outcome—Blood Pressure Measurements: $Y_i^T = (Y_{i1}, \ldots, Y_{in_i})$
- Exposure—High or Low Stress: $X_i^T = (X_{i1}, \ldots, X_{in_i})$

Questions Continued

4. Symbolically, write down a variance-covariance matrix of the blood pressure measurements for one of the subjects who had more than one visit, stating any assumptions you may use.
5. Write down a linear model that relates properties of the blood pressure measurement at a particular time to the binary measure of stress at the same time on the same individual. Repeat using vector notation.
6. Write down the equivalent linear model for the entire data set using matrix notation.
7. Write down a linear model that allows the effect of stress on (the mean) of diastolic blood pressure to vary with time since the beginning of the study.

Solutions for #4, #5

4. For $n_i = 2$,

$$\mathbf{V_i} = \begin{bmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) \\ Cov(Y_{i2}, Y_{i1}) & Var(Y_{i2}) \end{bmatrix} = \begin{bmatrix} v_{i11} & v_{i12} \\ v_{i21} & v_{i22} \end{bmatrix}$$

If we assume independence and constant variance (probably not true here, but is other times)

$$\mathbf{V_i} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

where $\sigma^2 = Var(Y_{ij})$

5. $Y_{i1} = b_0 + b_1 X_{i1} + e_{i1}$ for a person with $n_i = 1$
If $n_i \in \{2, 4\}$,

$$\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & X_{i1} \\ \vdots & \vdots \\ 1 & X_{in_i} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_{i1} \\ \vdots \\ e_{in_i} \end{pmatrix}$$

Solutions for #6, #7

6. $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ where
   - $\mathbf{Y}$ is a $N \times 1$ vector (here 19×1
   - $\mathbf{X}$ is a $N \times (p+1)$ matrix (here 19×2)
   - β is a $(p+1) \times 1$ vector (here 2×1)
   - $\mathbf{e}$ is a $N \times 1$ vector (here 19×1)

7. $\mathbb{E}[Y_{ij}|X_{ij} = x_{ij}, X_{i1} = x_{i1}] = b_0 + b_{CS}X_{i1} + b_L(X_{ij} - X_{i1})$

Expectation

- Every random variable X has an expected value, denoted $\mathbb{E}[X]$. Sometimes, we use the notation μ_X. It has the following properties:
  1. Scaling: $\mathbb{E}[aX] = a\mathbb{E}[X] \ \forall a \in \mathbb{R}$
  2. Shifting: $\mathbb{E}[X + b] = \mathbb{E}[X] + b \ \forall b \in \mathbb{R}$

- We can calculate the expectation of a sum of two random variables, X and Y:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

  Note: this is true for all random variables X and Y, even if they are not independent from each other!

Expectation Generalized

We can generalize the expectation of a sum of two random variables to a sum of many (n) random variables, $X_1, \ldots, X_n$.

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

This property is often stated as:

> "The expectation of a sum is the sum of the expectations."

Variance

- Every random variable X has a variance, denoted $Var(X)$. Sometimes we see this as σ_X^2.
- To get from variances to standard deviations, we take the square root:

$$SD(X) = \sqrt{Var(X)}$$

- Although standard deviations are nice to work with because their units are interpretable, in theory we use variance because it has nice theoretical properties.
  1. Scaling: $Var(aX) = a^2 Var(X)\ \forall a \in \mathbb{R}$
  2. Shifting does not affect variance: $Var(X + b) = Var(X)\ \forall b \in \mathbb{R}$

Variance of *Independent* Random Variables

- We can calculate the variance of a sum of two independent random variables, X and Y:

$$Var(X + Y) = Var(X) + Var(Y)$$

- Therefore, we can generalize this to a sum of n independent random variables, $X_1, \ldots, X_n$:

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$$

What if my variables are *not* independent?

We introduce the idea of **covariance**! That is, how do two random variables change together?

$$\begin{aligned} cov(X_1, X_2) &= \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] \\ &= \mathbb{E}[X_1X_2 - \mathbb{E}[X_1]X_2 - \mathbb{E}[X_2]X_1 + \mathbb{E}[X_1]\mathbb{E}[X_2]] \end{aligned}$$

Remember, things that are expectations are constants now, and we can use the rule for sums:

$$\begin{aligned} &= \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] - \mathbb{E}[X_2]\mathbb{E}[X_1] + \mathbb{E}[X_1]\mathbb{E}[X_2] \\ &= \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] \end{aligned}$$

Note: the variance of X is just the covariance of X with itself, or $cov(X, X)$.

Variance of two dependent variables

Now we can use covariance to discuss the variance of a sum, $X + Y$, when X is not independent of Y

$$Var(X + Y) = Var(X) + Var(Y) + 2cov(X, Y)$$

One last property!

Correlation is the last property of two random variables that will be useful for your assignment.

$$corr(X_1, X_2) = \frac{cov(X_1, X_2)}{SD(X_1)SD(X_2)}$$

Question 1.3

Consider a simple experiment of measuring cholesterol twice on each of m (independent) individuals. In addition, we will assume a simple, random effects model of the form:

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

for the jth measurement ($j = 1, 2$) on individual i ($i = 1, \ldots, m$), where,

- $\mathbb{E}(\alpha_i) = 0$
- $\mathbb{E}(e_{ij}) = 0$
- $cov(e_{i1}, e_{i2}) = 0$
- $cov(\alpha_i, e_{ij}) = 0$
- μ is a constant and thus the mean of Y_{ij}
- the variance between individuals is $var(\alpha_i) = \sigma^2_\alpha$
- variance within individuals (between measurements) is $var(e_{ij}) = \sigma^2_e$

Question 1.3 (cont)

Using the rules and definitions shown in the first few slides (one does not necessarily need them all), demonstrate that the correlation of measurements made on the same subject, $corr(Y_{i1}, Y_{i2}) = \rho$, is:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}.$$

For a simple random effects model, this ρ is often called the *intraclass correlation coefficient*. Please show all work.

Solution

Our goal is to show that

$$corr(Y_{i1}, Y_{i2}) = \rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}$$

We know from a previous slide that

$$corr(Y_{i1}, Y_{i2}) = \frac{cov(Y_{i1}, Y_{i2})}{SD(Y_{i1})SD(Y_{i2})}$$

Therefore, we can break this process into two steps:

1. $cov(Y_{i1}, Y_{i2}) = \sigma_\alpha^2$
2. $SD(Y_{i1})SD(Y_{i2}) = \sigma_\alpha^2 + \sigma_e^2$

Step 1

We need to calculate

$$cov(Y_{i1}, Y_{i2}) = \mathbb{E}[(Y_{i1} - \mathbb{E}[Y_{i1}])(Y_{i2} - \mathbb{E}[Y_{i2}])]$$

We should first understand what $\mathbb{E}[Y_{i1}]$ is. Remember, we defined

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

so we can calculate the expectation of Y_{i1} using the rule for the expectations of sums of random variables:

$$\mathbb{E}[Y_{i1}] = \mathbb{E}[\mu + \alpha_i + \epsilon_{i1}] = \mathbb{E}[\mu] + \mathbb{E}[\alpha_i] + \mathbb{E}[\epsilon_{i1}] = \mu$$

This follows from properties of α_i and ϵ_{ij} from slide 10. Notice that μ doesn't depend on i or j, so $\mathbb{E}[Y_{i1}] = \mathbb{E}[Y_{i2}]$.

Step 1 (cont)

Now we can consider the covariance calculation

$$
\begin{aligned}
cov(Y_{i1}, Y_{i2}) &= \mathbb{E}[(Y_{i1} - \mathbb{E}[Y_{i1}])(Y_{i2} - \mathbb{E}[Y_{i2}])] \\
&= \mathbb{E}[(\mu + \alpha_i + \epsilon_{i1} - \mathbb{E}[\mu + \alpha_i + \epsilon_{i1}]) \\
&\times (\mu + \alpha_i + \epsilon_{i2} - \mathbb{E}[\mu + \alpha_i + \epsilon_{i2}])] \\
&= \mathbb{E}[(\mu + \alpha_i + \epsilon_{i1} - \mu)(\mu + \alpha_i + \epsilon_{i2} - \mu)] \\
&= \mathbb{E}[\alpha_i^2 + \alpha_i\epsilon_{i1} + \alpha_i\epsilon_{i2} + \epsilon_{i1}\epsilon_{i2}] \\
&= \mathbb{E}[\alpha_i^2] + \mathbb{E}[\alpha_i\epsilon_{i1}] + \mathbb{E}[\alpha_i\epsilon_{i2}] + \mathbb{E}[\epsilon_{i1}\epsilon_{i2}]
\end{aligned}
$$

How do we evaluate these four quantities?

Step 1(cont)

- $\mathbb{E}[\alpha_i \epsilon_{i1}] = \mathbb{E}[\alpha_i \epsilon_{i2}]$
  On slide 10, we are given $cov(\alpha_i, \epsilon_{ij}) = 0$, $\mathbb{E}[\epsilon_{ij}] = 0$, and $\mathbb{E}[\alpha_i] = 0$. Working with properties of covariance (slide 7):

  $$cov(\alpha_i, \epsilon_{ij}) = \mathbb{E}[\alpha_i \epsilon_{ij}] - \mathbb{E}[\alpha_i]\mathbb{E}[\epsilon_{ij}] = \mathbb{E}[\alpha_i \epsilon_{ij}]$$

  But, we know this quantity is just 0, so we can say

  $$\mathbb{E}[\alpha_i \epsilon_{i1}] = \mathbb{E}[\alpha_i \epsilon_{i2}] = 0$$

- $\mathbb{E}[\epsilon_{i1} \epsilon_{i2}]$
  This term is also 0 by the properties from slide 10: $cov(\epsilon_{i1}, \epsilon_{i2}) = 0$, and $\mathbb{E}[\epsilon_{ij}] = 0$. The logic is the same as the previous bullet point.

Step 1 (cont)

- $\mathbb{E}[\alpha_i^2]$
  We use yet another property of covariance for this calculation (slide 7) and $\mathbb{E}[\alpha_i] = 0$ from slide 10:

$$\begin{aligned} Var(\alpha_i) &= Cov(\alpha_i, \alpha_i) \\ &= \mathbb{E}[\alpha_i \alpha_i] - \mathbb{E}[\alpha_i]\mathbb{E}[\alpha_i] \\ &= \mathbb{E}[\alpha_i^2] - 0 * 0 = \mathbb{E}[\alpha_i^2] \end{aligned}$$

  Therefore,

$$\mathbb{E}[\alpha_i^2] = Var(\alpha_i) = \sigma_\alpha^2$$

We can then go back to our calculation from slide 14 and say

$$cov(Y_{i1}, Y_{i2}) = \sigma_\alpha^2 + 0 + 0 + 0 = \sigma_\alpha^2$$

Step 2

Now, we must show

$$SD(Y_{i1})SD(Y_{i2}) = \sigma_\alpha^2 + \sigma_e^2.$$

We begin by calculating $SD(Y_{ij}) = \sqrt{Var(Y_{ij})}$. Remembering that $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ and the scaling property for variance, we can see

$$Var(Y_{ij}) = Var(\mu + \alpha_i + \epsilon_{ij}) = Var(\alpha_i + \epsilon_{ij})$$

Now we have a sum of two random variables that are not necessarily independent, so we use the formula from slide 8 and the property $cov(\alpha_i, \epsilon_{ij}) = 0$ from slide 10:

$$Var(\alpha_i + \epsilon_{ij}) = Var(\alpha_i) + Var(\epsilon_{ij}) + 2cov(\alpha_i, \epsilon_{ij}) = Var(\alpha_i) + Var(\epsilon_{ij})$$

Step 2 (cont)

Now we can plug this into our formula for standard deviations:

$$\begin{aligned} SD(Y_{i1})SD(Y_{i2}) &= \sqrt{Var(\alpha_i) + Var(\epsilon_{i1})}\sqrt{Var(\alpha_i) + Var(\epsilon_{i2})} \\ &= \sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2}\sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2} \\ &= \sigma_\alpha^2 + \sigma_\epsilon^2 \end{aligned}$$

Putting the steps together

Going back to slide 12, we have just shown that:

$$corr(Y_{i1}, Y_{i2}) = \frac{cov(Y_{i1}, Y_{i2})}{SD(Y_{i1})SD(Y_{i2})} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_e}$$