Stats 401 Lab 7

Naomi Giertych

10/19/2018

Announcements

- Midterm Monday in class

Outline

- Probability Model
- Covariance Review
- Exam Practice and Questions

Probability Model

(Looking ahead to HW 6)

- Recall: Probability Model is an assignment of probabilities to possible outcomes. We don't observe these probabilities, but we observe a random sample of them, e.g. a response variable and p predictor variables.
- This means there exists a probablity model of the form $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Note: $\boldsymbol{\beta}$ is unknown and fixed, $\boldsymbol{\epsilon}$ is an unknown random variable with $E[\boldsymbol{\epsilon}] = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$, and $\mathbb{X}$ is the **observed** explanatory matrix.
- However, when we observe data, we only have the sample version of the linear model $\boldsymbol{y} = \mathbb{X}\boldsymbol{b} + \boldsymbol{e}$

Examples of Writing these models in different forms

Subscript form:

$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i$, where

- Y_i is the ith observation of the response variable Y
- $\beta_1, \beta_2, \ldots, \beta_p$ are the true coefficients of the explanatory variables
- $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$ are the observed p predictor variables for observation i; note: $x_{i,p}$ is set to 1 so our model includes an intercept
- ϵ_i is the true error of the ith observation

Examples of Writing these models in different forms

Full matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where

- Y_i is the ith observation of the response variable Y
- $\beta_1, \beta_2, \dots, \beta_p$ are the true coefficients of the explanatory variables
- $x_{i,1}, x_{i,2}, \dots, x_{i,p}$ are the observed p predictor variables for observation i; note: $x_{i,p}$ is set to 1 so our model includes an intercept
- ϵ_i is the true error of the ith observation

Examples of Writing these models in different forms

Matrix Notation:

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{Y}$ is the $(n \times 1)$ vector of the response variable
- β is the $(p \times 1)$ vector of the true coefficients of the explanatory variables
- $\mathbb{X}$ is the $(n \times p)$ design matrix of the p explanatory variables
- $\boldsymbol{\epsilon}$ is the $(n \times 1)$ vector of the true errors

$\hat{\beta}$

Recall: $\boldsymbol{b} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{y}$

Similar to the probability model for $\boldsymbol{y} = \mathbb{X}\boldsymbol{b} + \boldsymbol{e}$, $\boldsymbol{b}$ has a probability model.

$\hat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}$

Key idea: We have a population model for the response variables ($\boldsymbol{Y}$). We can estimate the true coefficients using $\hat{\boldsymbol{\beta}}$, e.g. $\hat{\boldsymbol{\beta}}$ is an estimator for $\boldsymbol{\beta}$. However, since we only get a random draw of $\boldsymbol{Y}$, known as $\boldsymbol{y}$, we can only use a realization of $\hat{\boldsymbol{\beta}}$ known as $\boldsymbol{b}$, e.g. $\boldsymbol{b}$ is an estimate of $\boldsymbol{\beta}$.

Multivariate Random Variables

Recall in last lab, we discussed bivariate random variables and the bivariate normal distributions, and we extended these concepts to multivariate random variables.

- For example, we might have the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$
- (Another natural example of this random vector variable would be the vector of the p predictor variables of my probability model.)

Multivariate Random Variables

- Summary statistics for a multivariate random variable include the expected value vector and the variance-covariance matrix
- The expected value vector $\mathsf{E}(\mathbf{X}) = (\mathsf{E}(X_1), \ldots, \mathsf{E}(X_p))$ tells us the means for each component of $\mathbf{X}$
- The variance-covariance matrix gives the variances for each component along the diagonal and the pairwise covariances in the other entries:

$$\mathbb{V} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \ldots & \mathrm{Var}(X_p) \end{bmatrix}$$

Lab Activity (Part 1)

Fitting a probability model.

The director of the CDC wants to assess how well rates of hospital-acquired infections (`Infection.risk`) can be predicted using properities of a hospital. She expects to use the average length of stay (`Length.of.stay`), the average number of cultures for each patient without signs or symptoms of hospital-acquired infection, times 100 (`Culture`), and the number of X-ray procedures divided by number of patients without signs or symptoms of pneumonia, times 100 (`X.ray`).

- Write the probability model in subscript form, in full matrix form, and using matrix notation.

Lab activity (Part 2)

Fitting a sample linear model.

She collects a dataset for 113 hopitals with the variables mentioned above and fits the linear model below. Explain how this linear model is different from the one in Part 1. Write this model in subscript form, using the numbers below.

```
##                Estimate Std. Error
## (Intercept)        0.31       0.54
## Length.of.stay     0.24       0.05
## Culture            0.05       0.01
## X.ray              0.01       0.01
```

Lab Activity (Part 3)

Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ be a random vector with mean vector (2, 4, 6) and variance/covariance metrix

$$\mathbb{V} = \begin{bmatrix} 6 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 1 \end{bmatrix}$$

W_1 is the sum of Y_1, Y_2,, and Y_3.

W_2 is the sum of Y_1, (Y_2 multiplied by 2), and (Y_3 multiplied by -1).

W_3 is the sum of Y_1, and (Y_2 and Y_3 both multiplied by -1).

1. State the above in matrix notation.
2. Find the expectation of the random vector $\boldsymbol{W}$.
3. Find the variance/covariance matrix of $\boldsymbol{W}$.

Lab Activity Solutions (Part 1)

Subscript form:

$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 + \epsilon_i$, where

- Y_i is the infection risk for hospital i
- $\beta_1, \beta_2, \beta_3$, are the true coefficients for the length of stay, culture, and x-ray. β_4 is the true intercept or the infection risk for a hospital with a length of stay of 0 days, the number of cultures of 0, and 0 x-rays.
- $x_{i,1}, x_{i,2}$, and $x_{i,3}$ are the observed values of the length of stay, culture and x-ray, respectively, for hospital 1 i; note: $x_{i,4}$ is set to 1 so our model includes an intercept.
- ϵ_i is the true error of hospital i.

Lab Activity Solutions (Part 1)

Full matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{113} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & 1 \\ x_{2,1} & x_{2,2} & x_{2,3} & 1 \\ \vdots & \vdots & \ddots & \vdots \\ x_{113,1} & x_{113,2} & x_{113,p} & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{113} \end{bmatrix}$$

where

- Y_i is the infection risk for hospital i
- $\beta_1, \beta_2, \beta_3$, are the true coefficients for the length of stay, culture, and x-ray. β_4 is the true intercept or the infection risk for a hospital with a length of stay of 0 days, the number of cultures of 0, and 0 x-rays.
- $x_{i,1}, x_{i,2}$, and $x_{i,3}$ are the observed values of the length of stay, culture and x-ray, respectively, for hospital 1 i; note: $x_{i,4}$ is set to 1 so our model includes an intercept.
- ϵ_i is the true error of hospital i.

Lab Activity Solutions (Part 1)

Matrix notation:

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{Y}$ is the (113×1) vector of the infection rates for the 113 hospitals
- $\boldsymbol{\beta}$ is the (4×1) vector of the true coefficients of the length of stay, cultures, x-rays, and intercept respectively
- $\mathbb{X}$ is the (113×4) design matrix of the the length of stay, cultures, x-rays, and intercept respectively
- $\boldsymbol{\epsilon}$ is the (113×1) vector of the true errors

Lab Activity Solutions (Part 2)

$y_i = 0.24x_{i,1} + 0.05x_{i,2} + 0.01x_{i,3} + 0.31 + e_i$, where

- y_i is the infection risk for hospital i
- 0.05, 0.01, and 0.31 are the estimates for the coefficients for the length of stay, culture, and x-ray. 0.24 is the estimate of the intercept or the infection risk for a hospital with a length of stay of 0 days, the number of cultures of 0, and 0 x-rays.
- $x_{i,1}, x_{i,2}$, and $x_{i,3}$ are the observed values of the length of stay, culture and x-ray, respectively, for hospital 1 i; note: $x_{i,4}$ is set to 1 so our model includes an intercept.
- ϵ_i is the residual of hospital i.

Lab Activity Solutions (Part 3)

$W_1 = [1, 1, 1] \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$. Let $\mathbb{A}_1 = [1, 1, 1]$ $W_2 = [1, 2, -1] \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$ Let $\mathbb{A}_2 = [1, 2, -1]$ $W_3 = [1, -1, -1] \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$ Let $\mathbb{A}_3 = [1, -1, -1]$

Let $\mathbb{A} = \begin{bmatrix} \mathbb{A}_1 \\ \mathbb{A}_2 \\ \mathbb{A}_3 \end{bmatrix}$. In other words,

$$\mathbb{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & -1 \\ 1 & -1 & -1 \end{bmatrix}$$

Lab Activity Solutions (Part 3)

Then $E[\boldsymbol{W}] = \begin{bmatrix} 12 \\ 4 \\ -8 \end{bmatrix}$

Then $Var(\boldsymbol{W}) = \begin{bmatrix} 31 & 24 & -9 \\ 24 & 5 & -10 \\ -9 & -10 & 39 \end{bmatrix}$

Exam Questions

- What questions do you have about concepts or from the practice midterm or hw?

Lab Ticket

Using the same random variables as Lab Activity Part 3.

1. Find the probability that Y_1 is bigger than Y_2
2. Find the probability that W_1 is bigger than W_3.

Lab Ticket Solutions

1. $\mathbb{A} = [1, -1, 0]$. $E[\mathbb{A}\boldsymbol{Y}] = -2$ and $Var(\mathbb{A}\boldsymbol{Y}) = 6$. Then $P(Y_1 > Y_2) =$ `pnorm(0, -2, sqrt(6), lower.tail = F)` $= 0.2071081$.
2. $\mathbb{A} = [0, 2, 2]$. $E[\mathbb{A}\boldsymbol{Y}] = 20$ and $Var(\mathbb{A}\boldsymbol{Y}) = 60$. Then $P(W_1 > W_3) =$ `pnorm(0, 20, sqrt(60), lower.tail = F)` $= 0.9950884$.