STA 711 Homework 6

Due: Friday, March 21, 10:00pm on Canvas.

Instructions: Submit your work as a single PDF. You may choose to either hand-write your work and submit a PDF scan, or type your work using LaTeX and submit the resulting PDF. See the course website for a homework template file and instructions on getting started with LaTeX and Overleaf.

Multivariate normal distributions

The multivariate normal distribution will appear frequently in 711, for example as the asymptotic distribution of our coefficient estimates $\widehat{\beta}$. The purpose of this section is to derive a few important properties of this distribution.

To begin, let us formally define the multivariate normal. There are many equivalent definitions (one person's definition is another's "if and only if" theorem!). We will define the multivariate normal by first defining a multivariate standard normal, and then defining other multivariate normals as a transformation.

Definition (multivariate standard normal): Let $Z = (Z_1, ..., Z_k)^T$ be a k-dimensional random vector. We say that $Z \sim N(\mathbf{0}, \mathbf{I})$ (the multivariate normal with mean vector $\mathbf{0} \in \mathbb{R}^k$ and identity covariance matrix $\mathbf{I} \in \mathbb{R}^{k \times k}$) if

$$Z_1, ..., Z_k \overset{iid}{\sim} N(0, 1)$$

That is, the components are iid univariate standard normals.

Definition (multivariate normal): Let $X = (X_1, ..., X_k)^T$. We say that $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if there exists $Z = (Z_1, ..., Z_k)^T$ such that $Z \sim N(\mathbf{0}, \mathbf{I})$ and

$$X = \boldsymbol{\mu} + \mathbf{A}Z$$

where $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$.

1. Let $X \in \mathbb{R}^k$ such that $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{B} \in \mathbb{R}^{m \times k}$. Using the definitions above, show that
$$\mathbf{a} + \mathbf{B}X \sim N(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T)$$

2. Now let us prove some further properties of multivariate normal distributions:

   (a) Let $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $X \in \mathbb{R}^q$. Suppose that for some $1 \leq p < q$, $\boldsymbol{\Sigma}$ can be partitioned as
   $$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0}_{p \times (q-p)} \\ \mathbf{0}_{(q-p) \times p} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$
   where $\boldsymbol{\Sigma}_{11}$ is $p \times p$, $\boldsymbol{\Sigma}_{22}$ is $(q-p) \times (q-p)$, and $\mathbf{0}_{m \times n}$ denotes the matrix of zeros of the specified dimensions. Similarly partition
   $$X = \begin{pmatrix} X_{(1)} \\ X_{(2)} \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix},$$

into vectors of length p and $q-p$. Prove that

$$X_{(1)} \sim N(\boldsymbol{\mu}_{(1)}, \boldsymbol{\Sigma}_{11}), \quad X_{(2)} \sim N(\boldsymbol{\mu}_{(2)}, \boldsymbol{\Sigma}_{22}),$$

and $X_{(1)}$ and $X_{(2)}$ are independent.

(b) Using (a), conclude that if $X = (X_1, ..., X_q)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the entries X_i and X_j are independent *if and only* if $\boldsymbol{\Sigma}_{ij} = Cov(X_i, X_j) = 0$. (That is, while 0 covariance does not usually imply independence, it *does* for entries of a multivariate normal).

Logistic regression with earthquake data

In the second part of this assignment, you will work with a dataset from DrivenData, an online data competition site that hosts competitions aimed at improving education, health, safety, and general well being for individuals around the world.

Our data come from the 2015 Gorkha earthquake in Nepal. After the earthquake, a large scale survey was conducted to determine the amount of damage the earthquake caused for homes, businesses and other structures. This is one of the largest post-disaster surveys in the world, and researchers are interested in which building characteristics are associated with earthquake damage.

You will work with a subset of the earthquake data, consisting of 211774 buildings, containing the following variables:

- `Damage`: whether the building sustained any damage (1) or not (0)
- `Age`: the age of the building (in years)
- `Surface`: a categorical variable recording the surface condition of the land around the building. There are three different levels: `n`, `o`, and `t`. (The researchers who collected the data anonymized the level names to protect inhabitants' privacy).

You can load the data into R by

```
earthquake <- read.csv("https://sta711-s25.github.io/homework/earthquake_small.csv")
```

You will work with the following logistic regression model (you may assume all assumptions are met; no transformations or diagnostics are needed):

$$Damage_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Age_i + \beta_2 SurfaceO_i + \beta_3 SurfaceT_i + \beta_4 Age_i \cdot SurfaceO_i + \beta_5 Age_i \cdot SurfaceT_i$$

where $SurfaceO$ and $SurfaceT$ are indicator variables for whether surface is o or t, respectively.

3. (a) Fit the logistic regression model in R, and interpret the estimated slope $\widehat{\beta}_1$ in terms of the *odds* of damage.
   (b) Calculate the estimated probability of damage for a 50 year old building with surface condition = t.
4. The researchers want to know whether the relationship between Age and the probability of damage is the same for buildings in all three surface conditions. Use a Wald test to address the researchers' question; you should state the hypotheses in terms of one or more model parameters, calculate a test statistic and p-value, and make a conclusion.

Global F-test

Suppose that $V_1 \sim \chi^2_{d_1}$ and $V_2 \sim \chi^2_{d_2}$ are independent χ^2 random variables. Then

$$F = \frac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2}$$

where F_{d_1,d_2} denotes the F-distribution with numerator degrees of freedom d_1 and denominator degrees of freedom d_2.

The F-distribution is important for hypothesis testing in linear regression models. Suppose we observe independent data $(X_1, Y_1), ..., (X_n, Y_n)$, where $Y_i = \beta^T X_i + \varepsilon_i$, with $\beta = (\beta_0, ..., \beta_k)^T$ and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We wish to test the hypotheses

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \qquad H_A : \text{at least one of } \beta_1, ..., \beta_k \neq 0.$$

The F-test for these hypotheses is based on the F-statistic

$$F = \frac{(SSTO - SSE)/k}{SSE/(n-k-1)},$$

where $F \sim F_{k,n-k-1}$ under H_0, and

$$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

where

$$\widehat{Y} = \begin{pmatrix} \widehat{Y}_1 \\ \vdots \\ \widehat{Y}_n \end{pmatrix} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$$

The goal of this part of the assignment is to demonstrate that, indeed, $F \sim F_{k,n-k-1}$ under H_0.

5. The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is often called the *hat* matrix (because it puts a "hat" on Y). Show that $\mathbf{H}$ is idempotent.

6. The rank of an idempotent matrix is equal to its trace. Using this fact, show that

$$\text{rank}(\mathbf{H}) = k + 1$$

7. Let $\mathbf{J}_n$ be the $n \times n$ matrix of all 1s. Using the fact that $\mathbf{HX} = \mathbf{X}$, argue that $\mathbf{HJ}_n = \mathbf{J}_n$.

8. Using the previous question, show that $\mathbf{H} - \frac{1}{n}\mathbf{J}_n$ is idempotent.

9. Show that

$$SSTO = SSE + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

10. Show that

$$\sum_{i=1}^{n}(Y_i - \beta_0)^2 = SSE + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + n(\overline{Y} - \beta_0)^2$$

11. Show that under H_0, $\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_0)^2 \sim \chi^2_n$.

12. Using the above questions, find symmetric matrices A_1, A_2, A_3 such that under H_0,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_0)^2 = Z^T A_1 Z + Z^T A_2 Z + Z^T A_3 Z$$

where $Z \sim N(\mathbf{0}, \mathbf{I})$, $\frac{1}{\sigma^2}SSE = Z^T A_1 Z$, and $\frac{1}{\sigma^2}(SSTO - SSE) = Z^T A_2 Z$.

13. Using the matrices A_1, A_2, A_3 from the previous question, show that $rank(A_1) = n-k-1$, $rank(A_2) = k$, and $rank(A_3) = 1$.

14. By applying Cochran's theorem, show that $F = \dfrac{(SSTO - SSE)/k}{SSE/(n-k-1)} \sim F_{k,n-k-1}$ under H_0.