Applied Statistics 2017, solved problems

Tomonari Sei

January 15, 2018

Contents

1 Descriptive statistics

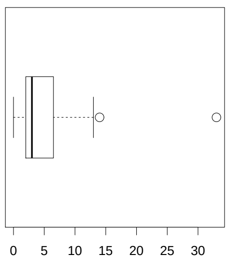
Q1-1

The following data is the number of coins that 23 students possess. Find their mean, median and frequency table. Draw the box plot.

3, 3, 2, 5, 0, 6, 3, 0, 4, 7, 2, 0, 1, 33, 0, 2, 3, 13, 2, 14, 4, 13, 7

Solution. In the R language, each statistic is obtained as follows.

```
> x = c(3,3,2,5,0, 6,3,0,4,7, 2,0,1,33,0, 2,3,13,2,14, 4,13,7)
> mean(x)
 [1] 5.521739
> median(x)
 [1] 3
> table(x)
x
 0  1  2  3  4  5  6  7 13 14 33
 4  1  4  4  2  1  1  2  2  1  1
> boxplot(x, horizontal=TRUE)
```



Note that the box and whisker are determined by five values: the minimum, 1st quartile (Q_1), median, 3rd quartile (Q_3), and maximum values. These values are confirmed by

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   3.000   5.522   6.500  33.000
```

The two outliers (14 and 33) are observations outside the interval

$$\begin{aligned}
&[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)] \\
&= [2.0 - 1.5 \times (6.5 - 2.0), 6.5 + 1.5 \times (6.5 - 2.0)] \\
&= [-4.75, 13.25].
\end{aligned}$$

□

Q1-2 (Anscombe's example)

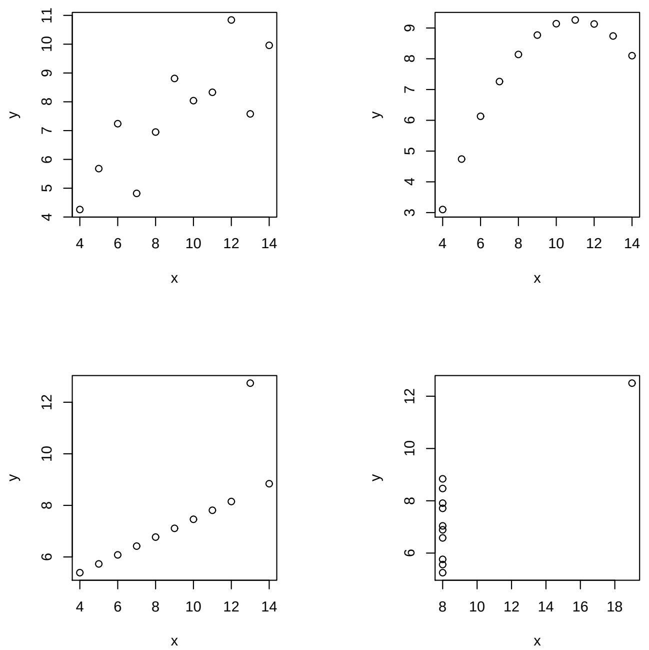
For each of the following four bivariate data, find the correlation coefficient. Draw the scatter plot. A CSV file is available from the course web page. On the statistical software R, an object `anscombe` is prepared.

x	y	x	y	x	y	x	y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Solution. In the R language, the correlation coefficients and scatter plots are obtained as follows.

```
X = read.csv("lec01-1.csv")
par(mfrow=c(2,2))
for(i in 1:4){
  show(cor(X[,i], X[,4+i]))
  plot(X[,i], X[,4+i], xlab="x", ylab="y")
}
```

The four datasets have almost the same value 0.816 of the correlation coefficient. However, the scatter plots are completely different.



□

Q1-3

For the following three-way contingency table, draw vectors (a_1, b_1), (c_1, d_1), (a_2, b_2), (c_2, d_2) in a two-dimensional plane, and give a geometric interpretation of Simpson's paradox.

n_{ij1}	1	2
1	a_1	b_1
2	c_1	d_1

n_{ij2}	1	2
1	a_2	b_2
2	c_2	d_2

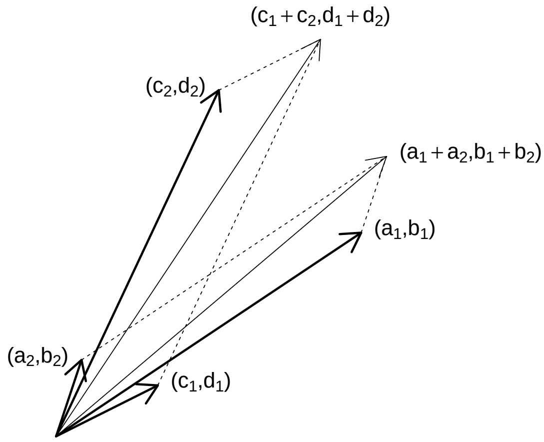
Solution. Simpson's paradox means that the odds ratio of each layer is

$$\frac{a_k d_k}{b_k c_k} < 1 \quad \text{for } k = 1, 2$$

while that of the marginal table is

$$\frac{(a_1 + a_2)(d_1 + d_2)}{(b_1 + b_2)(c_1 + c_2)} > 1$$

(or all inequality signs are reversed). A geometric interpretation of these conditions are shown in the following figure.



□

Q1-4

Let n be an odd integer. Denote the median of $x_1, \ldots, x_n \in \mathbb{R}$ by $f(x_1, \ldots, x_n)$. Then the following three conditions hold. Furthermore, for even n, show that there is not a function $f : \mathbb{R}^n \to \mathbb{R}$ satisfying these conditions.

(i) For any permutation $(i_1, \ldots, i_n)$, $f(x_{i_1}, \ldots, x_{i_n}) = f(x_1, \ldots, x_n)$.
(ii) For any increasing function $g : \mathbb{R} \to \mathbb{R}$, $g(f(x_1, \ldots, x_n)) = f(g(x_1), \ldots, g(x_n))$.
(iii) $f(-x_1, \ldots, -x_n) = -f(x_1, \ldots, x_n)$.

Solution. The median of $x_1, \ldots, x_n$ is the middle value after they are permuted in ascending order. The property (i) follows from the definition. For (ii) and (iii), we assume $x_1 \leq \cdots \leq x_n$ without loss of generality. Let $k = (n+1)/2$. Then $g(f(x_1, \ldots, x_n)) = g(x_k) = f(g(x_1), \ldots, g(x_n))$ and $f(-x_1, \ldots, -x_n) = -x_k = -f(x_1, \ldots, x_n)$.

For even n, consider the simplest case $n = 2$. The other cases are shown in a similar way. Assume $f(x_1, x_2)$ satisfies all the properties. Let a be any real number. Then we have $f(a, -a) = -f(-a, a) = -f(a, -a)$ by the properties (iii) and (i). Therefore $f(a, -a) = 0$. Now consider an increasing function g such that $g(1) = 1$, $g(-1) = -1$, and $g(0) \neq 0$. For example, $g(x) = 1 + (x-1)^3/4$. By (ii), we obtain $g(f(1, -1)) = f(g(1), g(-1))$, which implies $g(0) = f(1, -1) = 0$. This is a contradiction. □

2 Least squares and multiple regression

Q2-1

Let $\boldsymbol{y}$ and $\boldsymbol{x}$ be response and explanatory variables, respectively. Show that the regression equation is given by

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = r_{xy}(s_y/s_x),$$

where $\bar{x}, \bar{y}$ are the mean, s_x, s_y are the standard deviation, and r_{xy} is the correlation coefficient.

Solution. The coefficients $\hat{a}$ and $\hat{b}$ minimize

$$f(a,b) = \sum_{t=1}^{n} (y_t - a - bx_t)^2.$$

Using an identity

$$y_t - a - bx_t = (\bar{y} - a - b\bar{x}) + (y_t - \bar{y}) - b(x_t - \bar{x})$$

together with $\sum_t (y_t - \bar{y}) = 0$ and $\sum_t (x_t - \bar{x}) = 0$, we obtain

$$f(a,b) = n(\bar{y} - a - b\bar{x})^2 + \sum_t (y_t - \bar{y})^2 - 2b\sum_t (x_t - \bar{x})(y_t - \bar{y}) + b^2 \sum_t (x_t - \bar{x})^2.$$

This function is minimized by $\hat{a} = \bar{y} - \hat{b}\bar{x}$ and

$$\hat{b} = \frac{\sum_t (x_t - \bar{x})(y_t - \bar{y})}{\sum_t (x_t - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{r_{xy} s_x s_y}{s_x^2} = \frac{r_{xy} s_y}{s_x}.$$

□

Q2-2

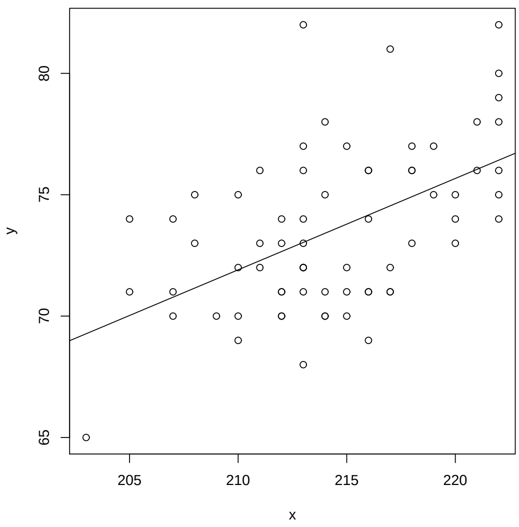
The "LPGA data" given in the course web site is the result of Japan Women's Open Golf Championship 2017, that consists of the score $\boldsymbol{x}$ up to the third day and the score $\boldsymbol{y}$ of the fourth day. Download it and determine the regression equation in that $\boldsymbol{y}$ is explained by $\boldsymbol{x}$. Make the scatter plot and add the regression line to it.

Solution. The regression line is

$$\hat{y}(x) = -7.1186 + 0.3763x.$$

Here is an R code:

```
X.ori = read.csv("lec02-1-LPGA.csv")
x = X.ori[,1]
y = X.ori[,2]
lm1 = lm(y ~ x)
plot(x, y)
abline(lm1$coef[1], lm1$coef[2])
```



□

Q2-3

The monthly average values y_t ($t = 1, \ldots, 12$) of the daily maximum temperature in Tokyo are

$$10.6, 12.2, 14.9, 20.3, 25.2, 26.3, 29.7, 31.6, 27.7, 22.6, 15.5, 13.8$$

in Celsius. Let the response variable be y_t and the explanatory variables be $x_{t1} = \cos(2\pi t/12)$ and $x_{t2} = \sin(2\pi t/12)$. Determine the regression equation.

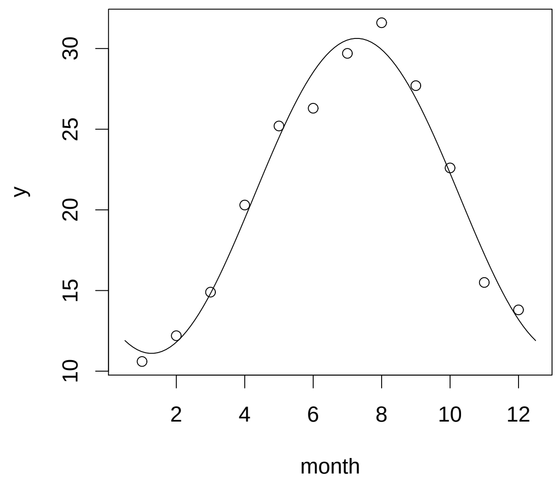
Solution. The predicted value is

$$\hat{y}_t = 20.9 - 7.67\cos(2\pi t/12) - 6.05\sin(2\pi t/12).$$

Here is an R code:

```
y = c(10.6,12.2,14.9,20.3,25.2,26.3,29.7,31.6,27.7,22.6,15.5,13.8)
x1 = cos(2*pi*(1:12)/12)
x2 = sin(2*pi*(1:12)/12)
lm(y ~ x1 + x2)
```

See the following figure.



$\square$

Q2-4

Let $\boldsymbol{y}$ be a response variable and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ be p explanatory variables. Prove that the regression equation

$$\hat{y}(x_1, \ldots, x_p) = \hat{a} + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p, \quad (x_1, \ldots, x_p) \in \mathbb{R}^p,$$

passes through the point $(\bar{x}_1, \ldots, \bar{x}_p, \bar{y})$. Furthermore, how are the coefficients represented if $\mathbf{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are mutually orthogonal.

Solution. The coefficients minimize

$$f(a, b_1, \ldots, b_p) = \sum_{t=1}^{n} (y_t - a - \sum_i b_i x_{ti})^2.$$

In a similar way to the case $p = 1$, we obtain $\bar{y} = \hat{a} + \sum_i \hat{b}_i \bar{x}_i$. If the orthogonality condition is satisfied, we have $f(a, b) = n(s_y^2 + (\bar{y} - a)^2 + \sum_i (-2b_i s_{x_i y} + b_i^2 s_{x_i}^2))$. Therefore $\hat{a} = \bar{y}$ and $\hat{b}_i = r_{x_i y} s_y / s_{x_i}$. $\square$

Q2-5 (Simpson's paradox revisited)

Let $\boldsymbol{y}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ be

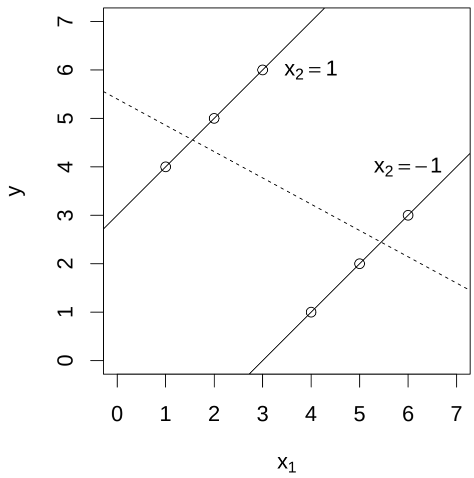
$$\boldsymbol{y} = (4, 5, 6, 1, 2, 3)', \quad \boldsymbol{x}_1 = (1, 2, 3, 4, 5, 6)', \quad \boldsymbol{x}_2 = (1, 1, 1, -1, -1, -1)',$$

respectively. Compare the regression coefficients when $\boldsymbol{y}$ is explained by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, with those when $\boldsymbol{y}$ is explained only by $\boldsymbol{x}_1$. Here the regression equation is assumed to contain the constant term.

Solution. By direct computation, the regression equations are

$$\hat{y}(x_1, x_2) = x_1 + 3x_2 \quad \text{and} \quad \hat{y}(x_1) = 3.5 - \frac{19}{35}(x_1 - 3.5),$$

respectively. The sign of the coefficient of x_1 is changed. Refer to the following figure.



□

Q2-6

For a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with rank p, prove that the matrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is an orthogonal projection matrix, that is, it satisfies $\boldsymbol{P}^2 = \boldsymbol{P}$ and $\boldsymbol{P}' = \boldsymbol{P}$. What is the set onto which $\boldsymbol{P}$ projects?

Solution. By definition,

$$\boldsymbol{P}^2 = (\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{P}.$$

It is also easy to show $\boldsymbol{P}' = \boldsymbol{P}$. Thus $\boldsymbol{P}$ is an orthogonal projection.

The range of $\boldsymbol{P}$ is the subspace spanned by the column vectors of $\boldsymbol{X}$. □

Q2-7

Let $\boldsymbol{X}$ be a design matrix and suppose that the QR decomposition $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ is already obtained. Explain an advantage of the QR decomposition when finding numerical values of regression coefficients. Here $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ is called QR decomposition if $\boldsymbol{Q}$ is column orthogonal (i.e., $\boldsymbol{Q}'\boldsymbol{Q}$ is the identity) and $\boldsymbol{R}$ is an upper triangular matrix with positive diagonal elements.

Solution. Let $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ be the QR decomposition of $\boldsymbol{X}$. Then the regression coefficient vector is

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{R})^{-1}\boldsymbol{R}'\boldsymbol{Q}'\boldsymbol{y} \\ &= \boldsymbol{R}^{-1}\boldsymbol{Q}'\boldsymbol{y}.\end{aligned}$$

Let $\boldsymbol{z} = \boldsymbol{Q}'\boldsymbol{y}$. Since $\boldsymbol{R}$ is an upper triangular matrix, the equation $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{z}$ is quickly solved by the backward substitution:

$$\hat{\beta}_p = \frac{z_p}{R_{pp}}, \quad \hat{\beta}_{p-1} = \frac{z_{p-1} - R_{p-1,p}\hat{\beta}_p}{R_{p-1,p-1}}, \quad \cdots, \quad \hat{\beta}_1 = \frac{z_1 - \sum_{j=2}^{p} R_{1j}\hat{\beta}_j}{R_{11}}.$$

This algorithm is numerically more stable than solving the normal equation directly. In terms of numerical linear algebra, the condition number of $\boldsymbol{R}$ is much smaller than that of $\boldsymbol{X}'\boldsymbol{X}$. Here we only give an example. Let

$$\boldsymbol{X} = \begin{pmatrix} 1 & 100 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} 1 \\ 0.03 \end{pmatrix}.$$

Then the two equations $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{Q}'\boldsymbol{y}$ and $(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ are

$$\begin{pmatrix} 1 & 100 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.03 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 100 \\ 100 & 10001 \end{pmatrix}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 100.03 \end{pmatrix},$$

respectively. Examine the Gaussian elimination method. □

Q2-8

Download the result of questionnaire from the course web site. Find the regression equation in that the sleeping time is explained by the commuting time.

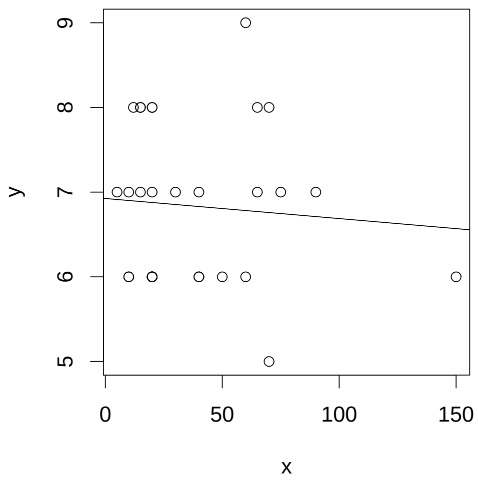
Solution. Here is an R program to obtain the regression line.

```
X = read.csv("questionnaire2017.csv")
x = X$Q2_commute_time
y = X$Q4_sleep
lm.8 = lm(y ~ x)
plot(x, y)
abline(lm.8)
```

The result is

$$\hat{y}(x) = 6.925 - 0.00237x,$$

where x is in minute whereas y is in hour.



A summary of the regression analysis is obtained as follows.

```
> summary(lm.8)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7588 -0.8655  0.1049  0.8999  2.2175

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.924764   0.278600  24.856   <2e-16 ***
x           -0.002371   0.005601  -0.423    0.675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9636 on 28 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.006357,Adjusted R-squared:  -0.02913
F-statistic: 0.1791 on 1 and 28 DF,  p-value: 0.6753
```

The p-value, which we will discuss later in this course, is not small. This means that we **cannot** conclude that x and y are related to each other.

Note that there is an missing value in this data but it is automatically deleted in the above analysis. □

3 Principal component analysis

Q3-1

Find the singular value decomposition of a matrix

$$\boldsymbol{X} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ -1 & -1 & h \\ -1 & -1 & -h \end{pmatrix},$$

where $h \geq 0$ is a constant. How the results change as h increases?

Solution. Perform the spectral decomposition $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$ and then calculate $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{D}^{-1}$ with $\boldsymbol{D} = \boldsymbol{\Lambda}^{1/2}$. We obtain the singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$, where

$$\boldsymbol{U} = \begin{pmatrix} -\frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 2\sqrt{2} & & \\ & 2 & \\ & & \sqrt{2}h \end{pmatrix}, \quad \boldsymbol{V} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

The sign of each column of $\boldsymbol{U}$ and $\boldsymbol{V}$ can be opposite. The order of singular values is changed at $h = \sqrt{2}$ and $h = 2$. If we focus on the first and second principal components, the 3rd and 4th individuals have the same scores if $h < \sqrt{2}$, but not if $h > \sqrt{2}$. □

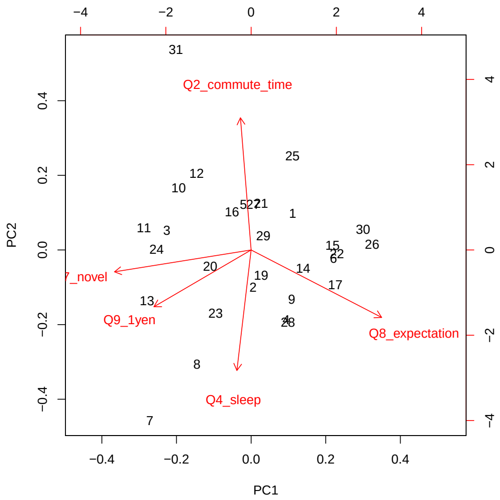
Q3-2

Perform the principal component analysis for a 5-variate data consisting of commuting time, sleeping time, novel reading, expectation, and 1-yen amount in the questionnaire result. To be more precise, after standardizing the data, determine the principal component score, rotation matrix and proportion of variance. Furthermore, draw the biplot and interpret the result if possible.

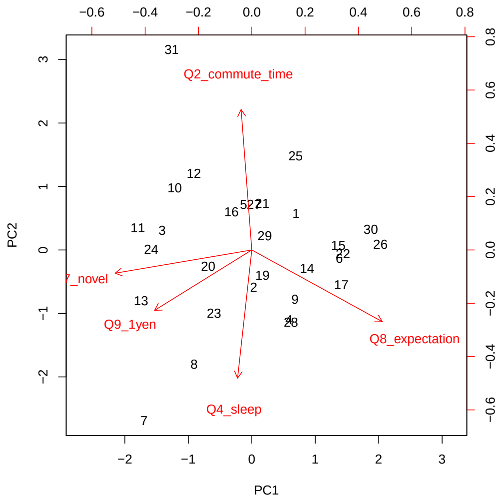
Solution. In R language,

```
> Xori = read.csv("questionnaire2017.csv")
> X = Xori[,c(3,5,8,9,10)]
> X = X[complete.cases(X),]   # remove missing values
> X = scale(X)   # must be scaled since the variates have different units
> pr1 = prcomp(X)  # PCA

> pr1$rotation  # rotation matrix (the same as V in X=UDV')  # output omitted
> pr1$x         # principal component scores (the same as UD)  # output omitted

> summary(pr1)   # proportion of variance
Importance of components:
                          PC1    PC2    PC3    PC4    PC5
Standard deviation     1.1414 1.0736 0.9695 0.9330 0.8569
```

```
Proportion of Variance 0.2606 0.2305 0.1880 0.1741 0.1469
Cumulative Proportion  0.2606 0.4911 0.6791 0.8531 1.0000

> biplot(pr1)   # biplot of U and VD (left figure)
> biplot(pr1, scale=0)  # biplot of UD and V (right figure)
```



biplot ($\boldsymbol{U}$ and $\boldsymbol{VD}$)



biplot ($\boldsymbol{UD}$ and $\boldsymbol{V}$)

□

Q3-3

In statistics, the following decompositions for real matrices are often used:

> spectral decomposition, singular value decomposition, Cholesky decomposition, and QR decomposition.

For each decomposition, answer the class of applicable matrices (arbitrary, square, symmetric, or positive definite). Furthermore, answer a similar question to Jordan canonical form, Schur canonical form, LU decomposition, and Sylvester canonical form.

Solution. See the following table.

	any	square	symmetric	positive definite	R function
spectral decomposition		$*$	✓	✓	`eigen`
singular value decomposition (SVD)	✓	✓	✓	✓	`svd`
Cholesky decomposition				✓	`chol`
QR decomposition	✓	✓	✓	✓	`qr`
Jordan canonical form		✓	✓	✓	
Schur canonical form		✓	✓	✓	
LU decomposition		✓	✓	✓	
Sylvester canonical form			✓	✓	

$*$ The spectral decomposition is available only if all the eigenvectors span the whole space. □

Q3-4

Let $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}' \in \mathbb{R}^{n\times n}$. Show that the principal component score of $\boldsymbol{X}$ depends only on $\boldsymbol{K}$. This is fundamental to the so-called kernel PCA.

Solution. Denote the spectral decomposition of $\boldsymbol{K}$ by $\boldsymbol{K} = \sum_{i=1}^{n} \lambda_i \boldsymbol{q}_i \boldsymbol{q}_i'$. Let $r = \min(n, p)$ and assume that $\lambda_1 > \cdots > \lambda_r > 0$. Then, for $1 \le i \le r$, the scores of the i-th principal component are given by $\sqrt{\lambda_i}\boldsymbol{q}_i$. Indeed, let $\boldsymbol{X} = \sum_{i=1}^{r} d_i \boldsymbol{u}_i \boldsymbol{v}_i'$ be the singular value decomposition. Then we have $\boldsymbol{K} = \sum_{i=1}^{r} d_i^2 \boldsymbol{u}_i \boldsymbol{u}_i'$ and therefore $d_i = \sqrt{\lambda_i}$ and $\boldsymbol{u}_i = \boldsymbol{q}_i$ for $1 \le i \le r$. □

Q3-5 (Eckart-Young theorem)

Denote the singular value decomposition of an $n \times p$ matrix $\boldsymbol{X}$ by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$, where singular values are in descending order. Fix $1 \le k \le \min(n, p)$ and consider a minimization problem:

$$\begin{aligned} &\text{Minimize} \quad \|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathrm{F}}^2 \\ &\text{subject to} \quad \operatorname{rank}(\boldsymbol{Y}) \le k. \end{aligned}$$

Show that the optimal solution is given by $\boldsymbol{Y} = \boldsymbol{U}_k \boldsymbol{D}_k \boldsymbol{V}_k'$, where $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ are the first k columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, and $\boldsymbol{D}_k$ is the upper-left $k \times k$ submatrix of $\boldsymbol{D}$. The quantity $\|\boldsymbol{A}\|_{\mathrm{F}} = (\operatorname{tr}(\boldsymbol{A}'\boldsymbol{A}))^{1/2}$ is called the Frobenius norm of $\boldsymbol{A}$.

Solution. Let $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{B}'$, where $\boldsymbol{A} \in \mathbb{R}^{n\times k}$, $\boldsymbol{B} \in \mathbb{R}^{p\times k}$, and $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_k$. This decomposition is always available due to the singular value decomposition of $\boldsymbol{Y}$.

We first show that $\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{B}'\|_{\mathrm{F}}^2$ is minimized by $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{B}$ for each $\boldsymbol{B}$. Indeed,

$$\begin{aligned} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{B}'\|_{\mathrm{F}}^2 &= \operatorname{tr}(\boldsymbol{X}'\boldsymbol{X} - 2\boldsymbol{X}'\boldsymbol{A}\boldsymbol{B}' + \boldsymbol{B}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}') \\ &= \operatorname{tr}(\boldsymbol{X}'\boldsymbol{X}) - 2\operatorname{tr}((\boldsymbol{X}\boldsymbol{B})'\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{A}'\boldsymbol{A}) \quad \because \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_k \\ &= \|\boldsymbol{A} - \boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{X}\|_{\mathrm{F}}^2 - \|\boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2, \end{aligned}$$

which is minimized by $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{B}$ for each $\boldsymbol{B}$. The minimum value is $\|\boldsymbol{X}\|_{\mathrm{F}}^2 - \|\boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2$.

Next we maximize $\|\boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2$ with respect to $\boldsymbol{B}$ using the Lagrange multiplier method. Define

$$F(\boldsymbol{B}, \boldsymbol{M}) = \operatorname{tr}(\boldsymbol{B}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}) - \operatorname{tr}(\boldsymbol{M}(\boldsymbol{B}'\boldsymbol{B} - \boldsymbol{I}_k)),$$

where the Lagrange multiplier $\boldsymbol{M}$ is a symmetric matrix. The stationary condition is

$$\frac{\partial F}{\partial \boldsymbol{B}} = 2(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{M}) = \boldsymbol{0}.$$

This implies that the column vectors of $\boldsymbol{B}$ spans an invariant subspace of the matrix $\boldsymbol{X}'\boldsymbol{X}$. Let $\boldsymbol{B} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)\boldsymbol{Q}'$, where $\boldsymbol{v}_i$'s are any k distinct eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{Q}$ is an orthogonal matrix. Now it is easy to show that $\|\boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 = \sum_{i=1}^{k} \lambda_i$, where λ_i is the eigenvalue corresponding to $\boldsymbol{v}_i$. This is maximized when $\lambda_1, \dots, \lambda_k$ are the largest k eigenvalues. As a consequence, we obtain $\boldsymbol{B} = \boldsymbol{V}_k\boldsymbol{Q}'$ and the optimal $\boldsymbol{Y}$ is

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{B}' = \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}' = \boldsymbol{X}\boldsymbol{V}_k\boldsymbol{V}_k' = \boldsymbol{U}_k\boldsymbol{D}_k\boldsymbol{V}_k'.$$

□

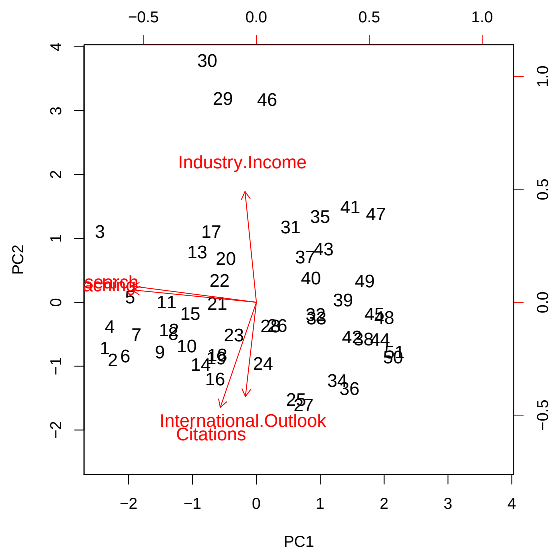
Q3-6

Access "World University Rankings" in the web site of Times Higher Education (THE). Apply the principal component analysis to the "performance breakdown" (Teaching, International Outlook, Research, Citations, Industry Income) of the top 50 universities.

Solution. Access "World University Rankings 2018" of Times Higher Education:
`https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking`

Click the tab "scores" to find the table. Use any favorite programming language to process. You may copy the necessary parts of the page and paste it to excel.

The following figure shows the biplot.



biplot (based on $\boldsymbol{UD}$ and $\boldsymbol{V}$)

In the figure, the top universities are located on the left side. In other words, the first PC score approximately determines the ranking. Note that the sign itself of any PC score has no meaning.

We observe that the first PC is strongly correlated with Teaching and Research. Indeed, the rotation matrix $\boldsymbol{V}$ is

```
                         PC1    PC2    PC3    PC4    PC5
Teaching              -0.687  0.069 -0.145 -0.155  0.691
Research              -0.693  0.090  0.091 -0.105 -0.702
Citations             -0.200 -0.582 -0.224  0.755 -0.019
Industry.Income       -0.063  0.613  0.491  0.605  0.115
International.Outlook -0.061 -0.522  0.824 -0.170  0.125
```

We also find three universities 29, 30, 46 away from the center of data cloud. They are Peking University, Tsinghua University, and University of Tokyo. These universities have more "industry income" or less "International Outlook" or less "Citations" than the others. In fact, the standardized scores of the three universities are

	Teaching	Research	Citations	Industry.Income	International.Outlook
Peking University	1.001	0.374	-2.380	1.853	-1.078
Tsinghua University	0.737	1.151	-2.729	1.844	-1.740
University of Tokyo	0.671	0.383	-3.687	-0.354	-2.226

□

4 Discriminant analysis

Q4-1

Define a 3-variate data by

$$\boldsymbol{X} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ -1 & -1 & h \\ -1 & -1 & -h \end{pmatrix},$$

where $h > 0$ is a constant. In the Mahalanobis distance constructed from $\boldsymbol{X}$, which is the farthest data from the mean vector?

Solution. The mean vector of the data is $(0, 0, 0)$. The covariance matrix and its inverse are

$$\boldsymbol{S} = \frac{1}{4}\begin{pmatrix} 6 & 2 & 0 \\ 2 & 6 & 0 \\ 0 & 0 & 2h^2 \end{pmatrix}, \quad \boldsymbol{S}^{-1} = \begin{pmatrix} 3/4 & -1/4 & 0 \\ -1/4 & 3/4 & 0 \\ 0 & 0 & 2/h^2 \end{pmatrix}.$$

The Mahalanobis distance between a point (x_1, x_2, x_3) and the mean vector is

$$d_{\boldsymbol{S}}((x_1, x_2, x_3), (0, 0, 0)) = \frac{3}{4}x_1^2 - \frac{1}{2}x_1x_2 + \frac{3}{4}x_2^2 + \frac{2}{h^2}x_3^2.$$

For each data point, the distance is equally 3.

Remark: In general, if the data points are $k+1$ points in k-dimensional space and they are affinely independent, the same result holds. See also Problem 4-2. □

Q4-2

Explain that the Mahalanobis distance is invariant under affine transformations.

Solution. The Mahalanobis distance between two points $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by

$$d_{\boldsymbol{S}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})'\boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{y}),$$

where $\boldsymbol{S}$ is the covariance matrix of given data. If we apply an affine transform $\boldsymbol{x} \mapsto \boldsymbol{b} + \boldsymbol{A}\boldsymbol{x}$ to the data, then the covariance matrix becomes $\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}'$. The two points are also transformed into $\boldsymbol{b} + \boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{b} + \boldsymbol{A}\boldsymbol{y}$. Then the distance between the transformed points is

$$\begin{aligned} d_{\boldsymbol{ASA}'}(\boldsymbol{b} + \boldsymbol{A}\boldsymbol{x}, \boldsymbol{b} + \boldsymbol{A}\boldsymbol{y}) &= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y})'(\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}')^{-1}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y}) \\ &= (\boldsymbol{x} - \boldsymbol{y})'\boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{y}) \\ &= d_{\boldsymbol{S}}(\boldsymbol{x}, \boldsymbol{y}). \end{aligned}$$

□

Q4-3

Determine the linear discriminant function based on the following two sets of bivariate data (artificial):

Group 1	variate 1	variate 2
individual 1	1	0
2	0	1
3	−1	0
4	0	−1

Group 2	variate 1	variate 2
individual 1	0	0
2	2	0
3	3	0
4	2	1
5	0	2

Solution. The mean vectors of the two groups are

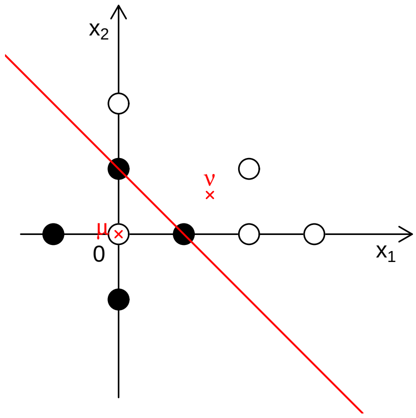
$$\boldsymbol{\mu} = (0, 0)', \quad \boldsymbol{\nu} = (7/5, 3/5)' = (1.4, 0.6)'.$$

respectively. The covariance matrix of the pooled data is

$$\boldsymbol{S} = \approx \begin{pmatrix} 1.506 & -0.037 \\ -0.037 & 0.666 \end{pmatrix}.$$

Therefore the linear discriminant function is

$$\begin{aligned} f(\boldsymbol{x}) &= (\boldsymbol{\mu} - \boldsymbol{\nu})'\boldsymbol{S}^{-1}\left(\boldsymbol{x} - \frac{\boldsymbol{\mu} + \boldsymbol{\nu}}{2}\right) \\ &= -\begin{pmatrix} 0.953 & 0.953 \end{pmatrix} \begin{pmatrix} x_1 - 7/10 \\ x_2 - 3/10 \end{pmatrix} \\ &\approx -0.953(x_1 + x_2 - 1). \end{aligned}$$



Remark: Since the data consists of integer values, the result can be written in fractions:

$$f(\boldsymbol{x}) = -\frac{81}{85}(x_1 + x_2 - 1).$$

□

Q4-4

Survey Fisher's iris data. Obtain the data, select two of the three varieties, and find the linear discriminant function discriminating them.

Solution. Omit. □

Q4-5

Using the Monte Carlo method, estimate the probability with its standard error that the following event occurs:

> Event: "when 10 dices are thrown simultaneously, the maximum frequency of the face is 4"

Also write a program code to compute the probability without Monte Carlo, and compare the results.

Solution. In R language, run the following code:

```
N = 1e4   # the number of experiments
result = numeric(N)   # generate a vector of size N
for(Ni in 1:N){
  x = sample(1:6, 10, replace=TRUE)   # throw 10 dices
  result[Ni] = ifelse(max(table(x))==4, 1, 0)  # 1 if maximum frequency is 4
}
p = sum(result) / N
print(p)  # estimate of probability
print(sqrt(p*(1-p)/N))  # error estimate
```

Then we obtain an estimate of the probability and its standard error as follows:

$$\hat{p} = 0.3118, \quad \sqrt{\hat{p}(1-\hat{p})/N} = 0.0046,$$

which depend on the random seed. Here $N = 10^4$ denotes the number of experiments.

The value we want to compute is

$$p = \sum_{i+j+k+l+m+r=10,\ \max(i,j,k,l,m,r)=4} \frac{10!}{i!j!k!l!m!r!} \left(\frac{1}{6}\right)^{10}.$$

One can obtain

$$p = \frac{18774000}{6^{10}} = 0.3104876$$

by a “brute-force” method.

If you are interested in a faster algorithm, refer to

> C. J. Corrado (2011). The exact distribution of the maximum, minimum and the range of multinomial/Dirichlet and multivariate hypergeometric frequencies, *Stat. Comput.*, **21**, 349–359.

□

Q4-6

For the world top 50 universities data seen in the last section, obtain the linear discriminant function that determines whether the university is in USA or not. Find its accuracy rate for the universities of rank 51 to 100.

Solution. Let $x_1, \ldots, x_5$ denote Teaching, Research, Citations, Industry Income, International Outlook, respectively. The linear discriminant function is

$$f(\boldsymbol{x}) = \sum_{i=1}^{5} a_i(x_i - b_i),$$

where the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are given by

$$\boldsymbol{a} = \boldsymbol{S}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu}), \quad \boldsymbol{b} = (\boldsymbol{\mu} + \boldsymbol{\nu})/2$$

in terms of the covariance matrix $\boldsymbol{S}$ of the pooled data and mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ of the two groups. The vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are given as follows.

i	1. Teaching	2. Research	3. Citations	4. Ind. Income	5. Int. Outlook
a_i	0.00552	0.03856	0.14169	−0.01364	−0.06906
b_i	72.60	81.64	93.20	60.75	72.95

For example, the University of Tokyo (rank 46) has $\boldsymbol{x} = (79.5, 85.2, 63.7, 52.7, 32.2)$ and therefore $f(\boldsymbol{x}) = -1.08 < 0$, that means the university is correctly classified into countries other than USA.

The following table shows the result when the discriminant function is applied to the universities of rank 51 to 100:

	classified into USA (Positive)	classified into others (Negative)
USA	9	7
others	0	34

The overall accuracy is $(9+34)/50 = 0.86$. The false positive rate (also called type-I error) is $0/(0+34) = 0$, and the false negative rate (also called type-II error) is $7/(9+7) = 0.44$. $\square$

5 Introduction of statistical inference

Q5-1

Let $X_1, \ldots, X_n$ be independent and normally distributed random variables with mean μ and σ^2. Find the distribution of the sample mean $\bar{X} = (1/n)\sum_{t=1}^n X_t$.

Solution. Since the random vector $(X_1, \ldots, X_n)$ is (jointly) normally distributed, $\bar{X} = n^{-1}\sum_{t=1}^n X_t$ is also normally distributed. Then it is enough to calculate its mean and variance. The result is $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$. $\square$

Q5-2 (simple regression model)

Let $x_1, \ldots, x_n$ be real numbers, and $Y_1, \ldots, Y_n$ be random variables determined by

$$Y_t = a + bx_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2),$$

where a, b, σ^2 are parameters. Find the distribution of Y_t. Furthermore, find the distribution of the regression coefficients $(\hat{a}, \hat{b})$ obtained by the least squares method when $\bar{x} = n^{-1}\sum_{t=1}^n x_t = 0$.

Solution. The distribution of Y_t is $\mathrm{N}(a + bx_t, \sigma^2)$. Its density function is

$$f(y_t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_t - a - bx_t)^2/(2\sigma^2)}.$$

The regression coefficients are determined by

$$\hat{a} = \bar{Y} - \hat{b}\bar{x} = \bar{Y}, \quad \hat{b} = \frac{\sum_t (x_t - \bar{x})(Y_t - \bar{Y})}{\sum_t (x_t - \bar{x})^2} = \frac{\sum_t x_t Y_t}{\sum_t x_t^2}$$

since $\bar{x}$ is assumed to be zero. The vector $(\hat{a}, \hat{b})'$ is normally distributed. The mean vector is calculated as

$$\mathrm{E}[\hat{a}] = \frac{1}{n}\sum_t \mathrm{E}[Y_t] = \frac{\sum_t (a + bx_t)}{n} = a,$$
$$\mathrm{E}[\hat{b}] = \frac{\sum_t x_t \mathrm{E}[Y_t]}{\sum_t x_t^2} = \frac{\sum_t (ax_t + bx_t^2)}{\sum_t x_t^2} = b.$$

The covariance matrix is

$$
\begin{aligned}
\mathrm{V}[\hat{a}] &= \frac{1}{n^2}\sum_t \mathrm{V}[Y_t] = \frac{\sigma^2}{n}, \\
\mathrm{Cov}[\hat{a},\hat{b}] &= \frac{1}{n(\sum_t x_t^2)}\sum_t x_t V[Y_t] = 0, \\
\mathrm{V}[\hat{b}] &= \frac{1}{(\sum_t x_t^2)^2}\sum_t x_t^2 V[Y_t] = \frac{\sigma^2}{\sum_t x_t^2}.
\end{aligned}
$$

Therefore $\hat{a}$ and $\hat{b}$ are independent and distributed according to $\mathrm{N}(a, \sigma^2/n)$ and $\mathrm{N}(b, \sigma^2/\sum_t x_t^2)$, respectively. □

Q5-3 (discriminant model)

Define the distribution of a random vector $(\boldsymbol{X}, y)$ by

$$
P(Y=1) = P(Y=-1) = \frac{1}{2}, \quad \boldsymbol{X}|\{Y=1\} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{X}|\{Y=-1\} \sim N(\boldsymbol{\nu}, \boldsymbol{\Sigma}),
$$

where $\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\Sigma}$ are parameters. Find the conditional distribution $P(Y|\boldsymbol{X}=\boldsymbol{x})$ given a real vector $\boldsymbol{x}$.

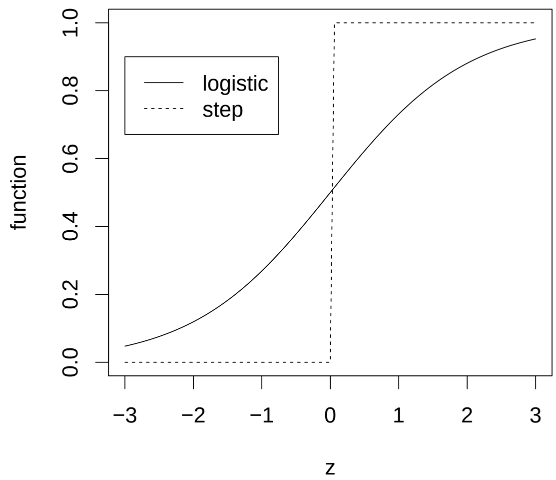
Solution. Let $f(\boldsymbol{x}|y)$ denotes the conditional density function of $\boldsymbol{X}$ given $Y=y$. The conditional distribution of Y given $\boldsymbol{X}$ is

$$
\begin{aligned}
P(Y=1|\boldsymbol{X}=\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|1)P(Y=1)}{f(\boldsymbol{x}|1)P(Y=1) + f(\boldsymbol{x}|-1)P(Y=-1)} \\
&= \frac{f(\boldsymbol{x}|1)}{f(\boldsymbol{x}|1) + f(\boldsymbol{x}|-1)} \\
&= \frac{e^z}{e^z+1},
\end{aligned}
$$

where

$$
\begin{aligned}
z &= \log\frac{f(\boldsymbol{x}|1)}{f(\boldsymbol{x}|-1)} \\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\nu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\nu}) \\
&= (\boldsymbol{\mu}-\boldsymbol{\nu})'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x} - \frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2}\right).
\end{aligned}
$$

Note that $P(Y=1|\boldsymbol{X}=\boldsymbol{x}) > 1/2$ if and only if $z > 0$. The following figure compares the logistic function $e^z/(e^z+1)$ and the step function $I_{(0,\infty)}(z)$.



□

Q5-4 (frequentist and Bayesian inference)

Assume that a baseball player makes X hits out of n boxes, and X has the binomial distribution $\mathrm{Bin}(n, p)$.

(i) Find the distribution of a statistic $\hat{p} = X/n$. Let $n = 10$ and draw a graph of the distribution of $\hat{p}$ when $p = 0.3$ and $p = 0$, respectively.

(ii) Assume that p is distributed according to the uniform distribution over $[0, 1]$. Then find the conditional distribution (posterior distribution) of p given $X = x$. Let $n = 10$ and draw a graph of the posterior distribution of p when $x = 3$ and $x = 0$, respectively.

Solution. The probability mass function of X given p is

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, \dots, n\}.$$

(i) The distribution of $\hat{p} = X/n$ is given by

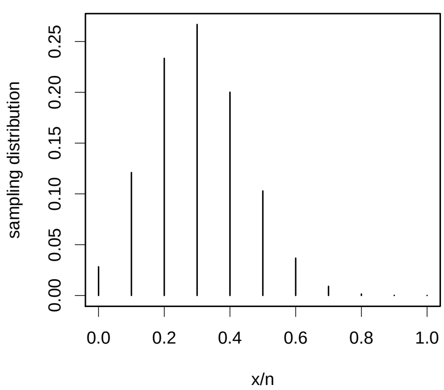
$$\mathrm{P}\left(\hat{p} = \frac{x}{n}\right) = f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, \dots, n\}.$$

The left panel of the following figures shows the distribution of $\hat{p}$ when $n = 10$.
(ii) The prior density function is $\pi(p) = 1$, $p \in [0, 1]$. The posterior density function given

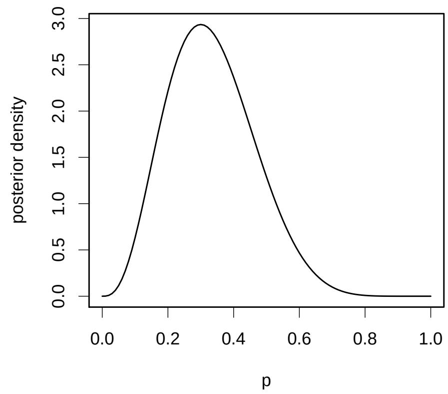
$X = x$ is

$$\begin{aligned}\pi(p|x) &= \frac{f(x|p)\pi(p)}{\int_0^1 f(x|p)\pi(p)dp} \\ &= \frac{p^x(1-p)^{n-x}}{\int_0^1 p^x(1-p)^{n-x}dp} \\ &= \frac{p^x(1-p)^{n-x}}{B(x+1, n-x+1)},\end{aligned}$$

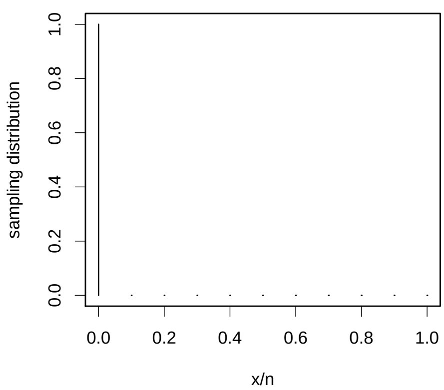
where $B(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}dp = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function. The right panel of the following figures shows the posterior density of p when $n = 10$.



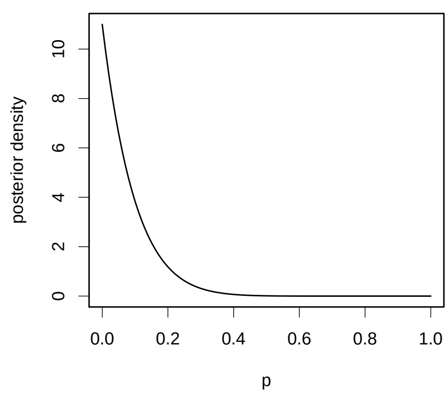
(i) Distribution of X/n given $p = 0.3$.



(ii) Posterior density of p given $X = 3$.



(i) Distribution of X/n given $p = 0$.



(ii) Posterior density of p given $X = 0$.

$\square$

Q5-5

Let $X_1, \ldots, X_n$ be independent random variables and distributed according to the exponential distribution with mean μ. Find the distribution of the sample mean $\bar{X}$.

Solution. The density function of the exponential distribution with mean μ is $f(x;\mu) = (1/\mu)e^{-x/\mu}$, and its moment generating function is

$$M(\theta) = \mathrm{E}[e^{\theta X_1}] = \int_0^\infty \frac{1}{\mu} e^{-x/\mu+\theta x} dx = \frac{1}{\mu(1/\mu - \theta)} = \frac{1}{1-\mu\theta}, \quad \theta < 1/\mu.$$

The moment generating function of $\bar{X} = n^{-1}(X_1 + \cdots + X_n)$ is

$$\mathrm{E}[e^{\theta n^{-1}\sum_{t=1}^n X_t}] = M(\theta/n)^n = \left(\frac{1}{1-\mu\theta/n}\right)^n.$$

On the other hand, the moment generating function of the gamma distribution is

$$M_{\alpha,\beta}(\theta) = \mathrm{E}[e^{\theta X}] = \int_0^\infty \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x+\theta x}}{\Gamma(\alpha)} dx = \left(\frac{\beta}{\beta-\theta}\right)^\alpha, \quad \theta < \beta.$$

Therefore the distribution of $\bar{X}$ is the gamma distribution with parameters $\alpha = n$ and $\beta = n/\mu$.

Different solution. Use the formula of convolution $f_{X+Y}(z) = \int f_X(x) f_Y(z-x)dx$ for independent random variables X and Y. Let $S_n = X_1 + \cdots + X_n$. First we have

$$f_{S_2}(z) = f_{X_1+X_2}(z) = \int_0^z \frac{e^{-x/\mu}}{\mu} \frac{e^{-(z-x)/\mu}}{\mu} dx = \frac{z}{\mu^2} e^{-z/\mu}.$$

Similarly, we can prove by induction

$$f_{S_n}(z) = \frac{z^{n-1}}{(n-1)!\mu^n} e^{-z/\mu}.$$

Since $\bar{X} = nS_n$, we obtain

$$f_{\bar{X}}(x) = nf_{S_n}(nx) = \frac{n^n x^{n-1}}{(n-1)!\mu^n} e^{-nx/\mu}.$$

□

5-A. Fundamental probability

Q5-A1

Denote the probability of an event A by $P(A)$. Assume that $P(A) \geq 0.9$ and $P(B) \geq 0.7$ for events A and B. Then show that $P(A \cap B) \geq 0.6$.

Solution. $P(A \cap B) = 1 - P(A^c \cup B^c) \geq 1 - P(A^c) - P(B^c) \geq 1 - 0.1 - 0.3 = 0.6$, where A^c denotes the complement of an event A. □

Q5-A2

There are n boxes and n balls, where the boxes are numbered from 1 to n. Each ball is randomly put into a box. Answer the following questions.

(a) Find the probability that there are k balls in the first box, where $0 \leq k \leq n$.

(b) To which value does the answer to (a) converge as $n \to \infty$? Here k is fixed.

(c) Let I_j be the box number into which the j-th ball is put ($j = 1, \dots, n$). Let $x_1, \dots, x_n$ be real numbers. Calculate expectation and variance of the random variable $X = n^{-1} \sum_{j=1}^{n} x_{I_j}$.

Solution. (a) By definition, the distribution is binomial with parameters $p = 1/n$ and n. Then the answer is $\binom{n}{k}(1/n)^k(1-1/n)^{n-k}$.
(b) Letting $n \to \infty$, we obtain

$$\binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!(n-1)^k}\left(1-\frac{1}{n}\right)^n \to \frac{e^{-1}}{k!},$$

which is the Poisson distribution with mean 1.
(c) Note that I_j's are independent and $P(I_j = i) = 1/n$ for each j and i. The expectation of X is

$$\mathrm{E}[X] = \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}[x_{I_j}] = \mathrm{E}[x_{I_1}] = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$, and the variance is

$$\mathrm{V}[X] = \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{V}[x_{I_j}] = \frac{1}{n}\mathrm{V}[x_{I_1}] = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Remark: Note that the value $\sqrt{\mathrm{V}[X]}$ is equal to the standard error of $\bar{x}$ when it is considered as an estimate of the population mean. This result is generalized to "the bootstrap method". We may encounter the method in a later lecture. □

Q5-A3 (Bayes' theorem)

The following table shows a survey on probability of symptoms S and their causes G when a product breaks down.

Cause G_i	G_1		G_2		G_3	
$P(G_i)$	0.60		0.10		0.30	
Symptom S_j	S_1	S_2	S_1	S_2	S_1	S_2
$P(S_j\|G_i)$	0.40	0.60	0.10	0.90	0.80	0.20

(Example: for a printer, S_1 = a paper jam, and G_1 = dirt of the roller.)

(a) Find the probability that the cause is G_1 when the symptom is S_1.

(b) Which of G_1, G_2, G_3 has the highest probability when the symptom is S_2?

Solution. (a) By Bayes' theorem,

$$P(G_1|S_1) = \frac{P(G_1)P(S_1|G_1)}{\sum_{j=1}^{3} P(G_j)P(S_1|G_j)} = \frac{0.24}{0.24 + 0.01 + 0.24} \approx 0.49.$$

(b) Since $P(G_i)P(S_2|G_i) = 0.36, 0.09, 0.06$ for $i = 1, 2, 3$, respectively, G_1 is the most probable. □

Q5-A4

For a nonnegative random variable X, a quantity

$$\frac{\sqrt{V[X]}}{E[X]}$$

is called the coefficient of variation, where $E[X]$ and $V[X]$ denote the expectation and variance, respectively.

(a) Calculate the coefficient of variation for the exponential random variable with the parameter $\lambda > 0$, whose density function is defined by $f(x) = \lambda e^{-\lambda x}$ $(x > 0)$.

(b) Calculate the coefficient of variation for the gamma random variable with the parameters $\alpha, \beta > 0$, whose density function is defined by $f(x) = \beta^\alpha x^{\alpha-1} e^{-\beta x}/\Gamma(\alpha)$. Here $\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} e^{-z} dz$ denotes the gamma function.

Solution. (a) This is a special case of (b) with $\alpha = 1$ and $\beta = \lambda$. The answer is 1.
(b) The expectation is

$$\mathrm{E}[X] = \int_0^\infty \frac{\beta^\alpha x^\alpha e^{-\beta x}}{\Gamma(\alpha)} dx = \frac{\Gamma(\alpha+1)}{\beta\Gamma(\alpha)} = \frac{\alpha}{\beta}.$$

The second moment is

$$\mathrm{E}[X^2] = \int_0^\infty \frac{\beta^\alpha x^{\alpha+1} e^{-\beta x}}{\Gamma(\alpha)} dx = \frac{\Gamma(\alpha+2)}{\beta^2\Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\beta^2}.$$

The variance is

$$\mathrm{V}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \frac{\alpha}{\beta^2}.$$

Thus the coefficient of variance is

$$\frac{\sqrt{\mathrm{V}[X]}}{\mathrm{E}[X]} = \frac{\sqrt{\alpha/\beta^2}}{\alpha/\beta} = \frac{1}{\sqrt{\alpha}}.$$

□

Q5-A5

Find the expectation and variance of X with a cumulative distribution function

$$F(x) = \frac{1}{1+e^{-x}}, \quad -\infty < x < \infty.$$

(Hint: use a transformation $u = 1/(1 + e^{-x})$ for calculation of variance.)

Solution. The probability density function is

$$f(x) = F'(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$

This function is symmetric (and exponentially decreases as $x \to \pm\infty$). Therefore the mean of X should be zero: $E[X] = 0$. The variance is

$$V[X] = \int_{-\infty}^{\infty} \frac{x^2 e^{-x}}{(1+e^{-x})^2} dx.$$

We use a transformation $u = 1/(1+e^{-x})$, equivalent to $x = \log u - \log(1-u)$. Then

$$\begin{aligned}
V[X] &= \int_0^1 \{\log u - \log(1-u)\}^2 u(1-u) \left(\frac{1}{u} + \frac{1}{1-u}\right) du \\
&= \int_0^1 \{\log u - \log(1-u)\}^2 du \\
&= 2\int_0^1 (\log u)^2 du - 2\int_0^1 (\log u)(\log(1-u)) du.
\end{aligned}$$

The first term is easily calculated as

$$\begin{aligned}
\int_0^1 (\log u)^2 du &= [u(\log u)^2]_0^1 - \int_0^1 u\{2(\log u)/u\} du \\
&= -2\int_0^1 (\log u) du \\
&= [-2u\log u]_0^1 + 2\int_0^1 du \\
&= 2.
\end{aligned}$$

However, it is difficult to evaluate the second term. We use Taylor's expansion as follows:

$$\begin{aligned}
\int_0^1 (\log u)(\log(1-u)) du &= \int_0^1 (\log u) \left\{-u - \frac{u^2}{2} - \frac{u^3}{3} - \cdots\right\} du \\
&= \sum_{n=1}^{\infty} \frac{1}{n} \int_0^1 (-\log u) u^n du \quad (\to \text{ see below}) \\
&= \sum_{n=1}^{\infty} \frac{1}{n} \left\{ \left[(-\log u) \frac{u^{n+1}}{n+1} \right]_0^1 + \int_0^1 \frac{u^n}{n+1} du \right\} \\
&= \sum_{n=1}^{\infty} \frac{1}{n(n+1)^2} \\
&= \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1} - \frac{1}{(n+1)^2} \right) \\
&= 1 - \sum_{n=1}^{\infty} \frac{1}{(n+1)^2} \\
&= 1 - \left(\frac{\pi^2}{6} - 1 \right) \quad (\to \text{ see below}) \\
&= 2 - \frac{\pi^2}{6}.
\end{aligned}$$

Finally, we have

$$V[X] = 2 \cdot 2 - 2\left(2 - \frac{\pi^2}{6}\right) = \frac{\pi^2}{3}.$$

Notes:

(i) We can exchange the integration and infinite series in this case because the integrand $(-\log u)u^n$ is non-negative over $u \in (0, 1)$. Refer to Lebesgue's monotone convergence theorem.

(ii) We used a formula $\sum_{n=1}^{\infty} 1/n^2 = \pi^2/6$. This is known as "the Basel problem".

□

Q5-A6

The probability density function of the multivariate normal distribution of mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is defined by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2},$$

where $\boldsymbol{\Sigma}$ is assumed to be positive definite. Prove that this function is actually a probability density function, i.e., show that $f(\boldsymbol{x}) \geq 0$ and $\int f(\boldsymbol{x})d\boldsymbol{x} = 1$. You can use the formula $\int_{-\infty}^{\infty} (2\pi)^{-1/2} e^{-x^2/2} dx = 1$.

(Hint: use matrix diagonalization or Cholesky decomposition.)

Solution. Since each factor of $f(\boldsymbol{x})$ is positive, we have $f(\boldsymbol{x}) > 0$. Let $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$, where $\boldsymbol{L}$ is the lower triangular matrix with positive diagonal elements. Then by changing the variable as $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$, we have

$$\begin{aligned}\int f(\boldsymbol{x})d\boldsymbol{x} &= \int f(\boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu})|\boldsymbol{L}|d\boldsymbol{z} = \int \frac{1}{(2\pi)^{d/2}|\boldsymbol{L}|} e^{-\boldsymbol{z}^\top \boldsymbol{z}/2} |\boldsymbol{L}| d\boldsymbol{z} \\ &= \prod_{i=1}^{d} \int \frac{1}{(2\pi)^{1/2}} e^{-z_i^2/2} dz_i = 1.\end{aligned}$$

□

Q5-A7

Suppose that random variables $X_1, \dots, X_n$ are independent and distributed according to the probability mass function

$$f(x) = (1-p)p^x, \quad x = 0, 1, \cdots.$$

where $0 < p < 1$. Answer the following questions.

(a) Obtain the moment generating function $M(\theta) = \mathrm{E}[e^{\theta X_1}]$ of X_1. Here determine the range of θ appropriately.

(b) Let $\bar{X}_n = (X_1 + \cdots + X_n)/n$. Use the law of large numbers to obtain the limit μ of $\bar{X}_n$.

(c) Use the central limit theorem to find the limit of the distributions of $\sqrt{n}(\bar{X}_n - \mu)$.

Solution. (a) The moment generating function is

$$M(\theta) = \mathrm{E}[e^{\theta X_1}] = \sum_{x=0}^{\infty}(1-p)p^x e^{\theta x} = \frac{1-p}{1-pe^{\theta}},$$

where the series converges if and only if $pe^{\theta} < 1$, that is $\theta < -\log p$.
(b) From the law of large numbers, the limit μ is the expectation of X_1. Using the result of (a), we have

$$\mu = E[X_1] = M'(0) = \frac{p}{1-p}.$$

(c) From the central limit theorem, the limit distribution is the normal distribution with mean zero and variance σ^2, where

$$\sigma^2 = V[X_1] = E[X_1^2] - E[X_1]^2 = M''(0) - \{M'(0)\}^2 = \frac{p}{(1-p)^2}.$$

□

6 Unbiased estimation and Cramér-Rao inequality

Q6-1

Let $X_1, \dots, X_n$ be independent random variables with an identical distribution of mean μ and variance σ^2 (not necessarily normal), where μ and σ^2 are unknown parameters.

(i) Show that the sample mean $\bar{X} = n^{-1} \sum_{t=1}^n X_t$ is an unbiased estimator of μ.

(ii) Show that the unbiased sample variance

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{t=1}^n (X_t - \bar{X})^2$$

is actually an unbiased estimator of σ^2.

Solution. (i) By linearity of expectation, we obtain

$$\mathrm{E}[\bar{X}] = \frac{1}{n} \sum_{t=1}^n \mathrm{E}[X_t] = \frac{1}{n} \sum_{t=1}^n \mu = \mu.$$

(ii) Note that $\mathrm{E}[(X_t - \mu)^2] = \sigma^2$ and $\mathrm{E}[(X_t - \mu)(X_s - \mu)] = 0$ if $t \neq s$. For each t, we have

$$\begin{aligned}
\mathrm{E}[(X_t - \bar{X})^2] &= \mathrm{E}\left[\left((1 - 1/n)(X_t - \mu) - (1/n) \sum_{s \neq t} (X_s - \mu)\right)^2\right] \\
&= (1 - 1/n)^2 \sigma^2 + (1/n)^2 (n-1) \sigma^2 \\
&= \frac{n-1}{n} \sigma^2.
\end{aligned}$$

Thus $\mathrm{E}[\hat{\sigma}^2] = (n-1)^{-1} \sum_{t=1}^n \mathrm{E}[(X_t - \bar{X})^2] = \sigma^2$. □

Q6-2

Let $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ be p-dimensional column vectors. Assume that the $n \times p$ matrix $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)})'$ is of rank p. Let $Y_1, \dots, Y_n$ be random variables determined by

$$Y_t = \boldsymbol{\beta}' \boldsymbol{x}^{(t)} + \varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2),$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$ are unknown parameters. Show that the regression coefficient $\hat{\boldsymbol{\beta}}$ obtained by the least squares method is an unbiased estimator of $\boldsymbol{\beta}$. Furthermore, show that

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{t=1}^n (Y_t - \hat{\boldsymbol{\beta}}' \boldsymbol{x}^{(t)})^2$$

is an unbiased estimator of σ^2.

Solution. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. The least-squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Its expectation is

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Therefore $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Let $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ be the orthogonal projection onto the subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{X}$. Then the predicted vector is $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}$ and the residual vector is

$$\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\varepsilon} \quad (\because \boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}).$$

The expectation of $\hat{\sigma}^2$ is

$$\begin{aligned}
\mathrm{E}[\hat{\sigma}^2] &= \frac{1}{n-p}\mathrm{E}[(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})] \\
&= \frac{1}{n-p}\mathrm{E}[\boldsymbol{\varepsilon}'(\boldsymbol{I}_n - \boldsymbol{P})'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\varepsilon}] \\
&= \frac{1}{n-p}\mathrm{E}[\boldsymbol{\varepsilon}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\varepsilon}] \\
&= \frac{1}{n-p}\mathrm{E}[\mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\}] \\
&= \frac{1}{n-p}\,\mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{P})(\sigma^2\boldsymbol{I}_n)\} \\
&= \frac{\sigma^2}{n-p}\,\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{P}) \\
&= \sigma^2 \qquad (\because\ \mathrm{tr}(\boldsymbol{I}_n) = n,\ \mathrm{tr}(\boldsymbol{P}) = p).
\end{aligned}$$

Remark: Problem 6-1 is a special case (except normality): $p = 1$ and $\boldsymbol{x}^{(t)} = 1$. □

Q6-3

Find the Fisher information of the following statistical models:

(i) Normal distribution of variance 1: $f(x;\theta) = (2\pi)^{-1/2}e^{-(x-\theta)^2/2}$.

(ii) Poisson distribution: $f(x;\theta) = (\theta^x/x!)e^{-\theta}$.

(iii) Cauchy distribution of median θ: $f(x;\theta) = \pi^{-1}(1 + (x-\theta)^2)^{-1}$.

Solution. (i) $I(\theta) = \mathrm{E}[-\partial_\theta^2 \log f(X;\theta)] = \mathrm{E}[1] = 1$.
(ii) $I(\theta) = \mathrm{E}[-\partial_\theta^2 \log f(X;\theta)] = \mathrm{E}[X/\theta^2] = 1/\theta$.

(iii) Since $\partial_\theta \log f(x;\theta) = 2(x-\theta)/(1+(x-\theta)^2)$, the Fisher information is

$$\begin{aligned}
I(\theta) &= \mathrm{E}[\{\partial_\theta \log f(X;\theta)\}^2] \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi(1+(x-\theta)^2)} \frac{4(x-\theta)^2}{(1+(x-\theta)^2)^2} dx \\
&= \int_{-\pi/2}^{\pi/2} \frac{1}{\pi(1+\tan^2\phi)} \frac{4\tan^2\phi}{(1+\tan^2\phi)^2} \frac{d\phi}{\cos^2\phi} \quad (\because x-\theta = \tan\phi) \\
&= \int_{-\pi/2}^{\pi/2} \frac{1}{\pi}(4\cos^2\phi \sin^2\phi) d\phi \\
&= \frac{1}{2},
\end{aligned}$$

where the following formula for non-negative integers m and n was used:

$$\begin{aligned}
\int_0^{\pi/2} (\cos\phi)^{2m} (\sin\phi)^{2n} d\phi &= \frac{1}{2}\int_0^1 u^{m-1/2}(1-u)^{n-1/2} du \quad (\because u = \cos^2\phi) \\
&= \frac{1}{2} B(m+\tfrac{1}{2}, n+\tfrac{1}{2}) \\
&= \frac{1}{2} \frac{\Gamma(m+\frac{1}{2})\Gamma(n+\frac{1}{2})}{\Gamma(m+n+1)} \\
&= \frac{\pi(2m-1)!!(2n-1)!!}{(m+n)! 2^{m+n+1}}.
\end{aligned}$$

Recall that $(2m-1)!! = (2m-1)(2m-3)\cdots 1$ and $(-1)!! = 1$. □

Q6-4

Let $f(x;\theta)$ be the probability density function of a random variable X. We abbreviate the derivative $\partial/\partial\theta$ by ∂_θ. Prove that

$$\mathrm{E}[\partial_\theta \log f(X;\theta)] = 0 \quad \text{and} \quad \mathrm{E}[\{\partial_\theta \log f(X;\theta)\}^2] = \mathrm{E}[-\partial_\theta^2 \log f(X;\theta)],$$

where the order of integral and differentiation is assumed to be reversed.

Solution.

$$\mathrm{E}[\partial_\theta \log f(X;\theta)] = \int \frac{\partial_\theta f(x;\theta)}{f(x;\theta)} f(x;\theta) dx = \int \partial_\theta f(x;\theta) dx = \partial_\theta \int f(x;\theta) dx = \partial_\theta 1 = 0.$$

$$\begin{aligned}
\mathrm{E}[\{\partial_\theta \log f(X;\theta)\}^2] &= \int \{\partial_\theta f(x;\theta)\}\{\partial_\theta \log f(x;\theta)\} dx \\
&= \partial_\theta \left\{ \int f(x;\theta) \partial_\theta \log f(x;\theta) dx \right\} - \int f(x;\theta) \partial_\theta^2 \log f(x;\theta) dx \\
&= -\mathrm{E}[\partial_\theta^2 \log f(X;\theta)].
\end{aligned}$$

□

Q6-5

The Kullback-Leibler divergence between probability density functions $f(x;\theta)$ and $f(x;\phi)$ is defined by

$$\mathrm{KL}(\theta,\phi) = \int f(x;\theta)\log\frac{f(x;\theta)}{f(x;\phi)}dx.$$

Show that $\mathrm{KL}(\theta,\phi)$ is non-negative and

$$\mathrm{KL}(\theta,\theta+\delta) = \frac{1}{2}I(\theta)\delta^2 + \mathrm{o}(\delta^2), \quad \delta \to 0.$$

Solution. By an identity $\int f(x;\theta)dx = \int f(x;\phi)dx = 1$ and an inequality $-\log z + z - 1 \geq 0$ for any $z > 0$, we have

$$\mathrm{KL}(\theta,\phi) = \int f(x;\theta)\left\{-\log\frac{f(x;\phi)}{f(x;\theta)} + \frac{f(x;\phi)}{f(x;\theta)} - 1\right\}dx \geq 0.$$

By Taylor's expansion, we have

$$\begin{aligned}\mathrm{KL}(\theta,\theta+\delta) &= \int f(x;\theta)\left\{-\delta\partial_\theta \log f(x;\theta) - \frac{\delta^2}{2}\partial_\theta^2 \log f(x;\theta) + \mathrm{o}(\delta^2)\right\}dx\\ &= \frac{\delta^2}{2}I(\theta) + \mathrm{o}(\delta^2)\end{aligned}$$

under regularity conditions. □

Q6-6

Let θ be an integer-valued parameter and X be a random variable taking values $\theta-1, \theta, \theta+1$ with probability $1/3$ each.

(i) Show that $\hat{\theta}(X)$ is an unbiased estimator of θ.

(ii) Show that there is no UMVUE (uniformly minimum variance unbiased estimator).

Solution. (i) It is easy to see $\mathrm{E}_\theta[X] = \{(\theta-1)+\theta+(\theta+1)\}/3 = \theta$.
(ii) Define an estimator $\phi(X)$ of θ by a difference equation

$$\frac{\phi(\theta-1)+\phi(\theta)+\phi(\theta+1)}{3} = \theta, \quad \theta \in \mathbb{Z} = \{0, \pm 1, \cdots\},$$

with initial conditions $\phi(-1) = \phi(0) = \phi(1) = 0$. Then all the values of $\phi(x)$ are determined. Indeed, we have

$$\phi(2) = \phi(3) = \phi(4) = 3, \quad \phi(5) = \phi(6) = \phi(7) = 6, \quad \cdots$$

and $\phi(-x) = -\phi(x)$. By definition, $\phi(X)$ is unbiased:

$$\mathrm{E}_\theta[\phi(X)] = \frac{\phi(\theta-1)+\phi(\theta)+\phi(\theta+1)}{3} = \theta.$$

The variance of ϕ is

$$\mathrm{V}_\theta[\phi(X)] = \begin{cases} 0 & \text{for } \theta = 0, \pm 3, \pm 6, \cdots, \\ 2 & \text{otherwise.} \end{cases}$$

Similarly, we can construct an unbiased estimator ψ such that $\mathrm{V}_\theta[\psi(X)] = 0$ for $\theta = 1, 4, \cdots$.

Now suppose that there is an UMVUE $\hat{\theta}^*$. Then it should satisfy $\mathrm{V}_\theta[\hat{\theta}^*] = 0$ for $\theta = 0$ and $\theta = 1$, but this is impossible. Indeed, $\mathrm{V}_{\theta=0}[\hat{\theta}^*] = 0$ implies $\hat{\theta}^* = \phi$ and therefore $\mathrm{V}_{\theta=1}[\hat{\theta}^*] = 2$. □

Q6-7

Show that unbiasedness is not invariant with respect to transformations of the parameter, by giving a concrete example.

Solution. For example, consider a normal model $\mathrm{N}(\theta, 1)$. The sample mean $\hat{\theta} = \bar{X}$ is an unbiased estimator of θ. However, putting $h(\theta) = \theta^2$, we have

$$\mathrm{E}[h(\hat{\theta})] = \mathrm{E}[\hat{\theta}^2] = \mathrm{E}[\bar{X}^2] = \mathrm{V}[\bar{X}] + \theta^2 = \frac{1}{n} + \theta^2 = \frac{1}{n} + h(\theta).$$

Therefore $h(\hat{\theta})$ is not an unbiased estimator of $h(\theta)$.

More generally, if h is a strictly convex function, we have

$$\mathrm{E}[h(\hat{\theta})] > h(\mathrm{E}[\hat{\theta}]) = h(\theta)$$

by Jensen's inequality. □

7 Maximum likelihood estimation

Q7-1

Find the maximum likelihood estimator of the following statistical models, that is, find θ maximizing $L(\theta) = \prod_{t=1}^n f(x_t; \theta)$ for given data $x_1, \ldots, x_n$. Determine its unbiasedness.

(i) Poisson distribution $f(x; \theta) = (\theta^x / x!) e^{-\theta}$, $x \in \mathbb{Z}_{\geq 0} = \{0, 1, \cdots\}$.

(ii) Normal distribution $f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}$, $x \in \mathbb{R}$, $\theta = (\mu, \sigma^2)$.

(iii) Exponential distribution $f(x; \theta) = \theta e^{-\theta x}$, $x > 0$.

Solution. (i) The log likelihood function is

$$\log L = \sum_{t=1}^n \log f(x_t; \theta) = \sum_{t=1}^n (x_t \log \theta - \theta) + (\text{const.}).$$

By solving the likelihood equation $(\partial/\partial\theta)\log L = 0$, we obtain $\hat{\theta} = n^{-1}\sum_{t=1}^n x_t$. This estimator is unbiased since $\mathrm{E}[\hat{\theta}] = n^{-1}\sum_{t=1}^n \mathrm{E}[X_t] = \theta$.
(ii) The log likelihood function is

$$\log L = \sum_{t=1}^n \log f(x_t;\mu,\sigma^2) = \sum_{t=1}^n \left(-\frac{(x_t-\mu)^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2\right).$$

By solving the system of equations $(\partial/\partial\mu)\log L = 0$ and $(\partial/\partial(\sigma^2))\log L = 0$, we obtain

$$\hat{\mu} = \bar{x} = n^{-1}\sum_{t=1}^n x_t, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^n (x_t-\bar{x})^2.$$

The estimator $\hat{\mu}$ is an unbiased estimator of μ whereas $\hat{\sigma}^2$ is not unbiased since $\mathrm{E}[\hat{\sigma}^2] = ((n-1)/n)\sigma^2$.
(iii) In the same way as (i) and (ii), we obtain $\hat{\theta} = 1/\bar{x} = n/\sum_{t=1}^n x_t$. This is not unbiased. To see this, use Jensen's inequality $\mathrm{E}[f(\bar{X})] \geq f(\mathrm{E}[\bar{X}])$ for the convex function $f(x) = 1/x$.

One can show that $\mathrm{E}[\hat{\theta}] = (n/(n-1))\theta > \theta$. Indeed, by noting that $Y := \sum_{t=1}^n X_t$ has the gamma distribution with the parameters $\alpha = n$ and $\beta = \theta$, we have

$$\mathrm{E}[1/\bar{X}] = \mathrm{E}[n/Y] = \int_0^\infty \frac{n}{y}\frac{\theta^n y^{n-1}}{(n-1)!}e^{-\theta y}dy = \frac{n\theta}{(n-1)!}\int_0^\infty z^{n-2}e^{-z}dz = \frac{n}{n-1}\theta.$$

□

Q7-2

Show that the following models are exponential families.

(i) Bernoulli distribution (binomial distribution with size 1) $f(x;p) = p^x(1-p)^{1-x}$, $x \in \{0,1\}$.

(ii) Negative binomial distribution $f(x;p) = \binom{r+x-1}{x}p^r(1-p)^x$, $x \in \mathbb{Z}_{\geq 0}$, where r is assumed to be known.

(iii) Multinomial distribution with k categories and size 1 $f(\boldsymbol{x};\boldsymbol{p}) = \prod_{i=1}^k p_i^{x_i}$, $x_i \in \{0,1\}$, $\sum_{i=1}^k x_k = 1$.

Solution. (i) $f(x) = p^x(1-p)^{1-x} = a(x)e^{\theta x - \psi(\theta)}$, where $a(x) = 1$, $\theta = \log(p/(1-p))$, $s(x) = x$, and $\psi(\theta) = -\log(1-p) = \log(1+e^\theta)$. The domain of ψ is $\mathbb{R}$.

Note: There are other parameterizations. But the dimension of θ should be one if we impose strict convexity of ψ (see Problem 7-5). For example, if we set $a(x) = 1$, $\theta_1 = \log p$, $\theta_2 = \log(1-p)$, $s_1(x) = x$, and $s_2(x) = 1-x$, then we have $\psi(\theta) = 1$, which is not strictly convex.
(ii) $f(x) = \binom{r+x-1}{x}p^r(1-p)^x = a(x)e^{\theta x-\psi(\theta)}$, where $a(x) = \binom{r+x-1}{x}$, $\theta = \log(1-p)$, $s(x) = x$, and $\psi(\theta) = -r\log p = -r\log(1-e^\theta)$. The domain of ψ is $\{\theta < 0\}$.
(iii) $f(\boldsymbol{x}) = \prod_{i=1}^k p_i^{x_i} = a(\boldsymbol{x})\exp(\sum_{i=1}^{k-1}\theta_i x_i - \psi(\boldsymbol{\theta}))$, where $\theta_i = \log(p_i/p_k)$ for $1 \leq i \leq k-1$ and $\psi(\boldsymbol{\theta}) = -\log p_k = \log(1+\sum_{i=1}^{k-1}e^{\theta_i})$. The domain of ψ is $\mathbb{R}^{k-1}$. □

Q7-3

Let $X_1, \dots, X_n$ be i.i.d. random variables with the probability density function $f(x; \theta) = \theta e^{-\theta x}$ $(x \geq 0)$. Sort the variables $X_1, \dots, X_n$ in ascending order and denote them by $X_{(1)}, \dots, X_{(n)}$. More precisely, on a probability space $(\Omega, \mathcal{F}, P)$, we define $X_{(k)}(\omega)$ for each $\omega \in \Omega$ by the k-th smallest value of $X_1(\omega), \dots, X_n(\omega)$. These variables $X_{(k)}$ $(1 \leq k \leq n)$ are called, in general, the order statistics. Answer the following questions.

(i) Obtain the joint density function of $X_{(1)}, \dots, X_{(n)}$.

(ii) Let $1 < k < n$. Find the maximum likelihood estimator of θ when only $X_{(1)}, \dots, X_{(k)}$ are observed.

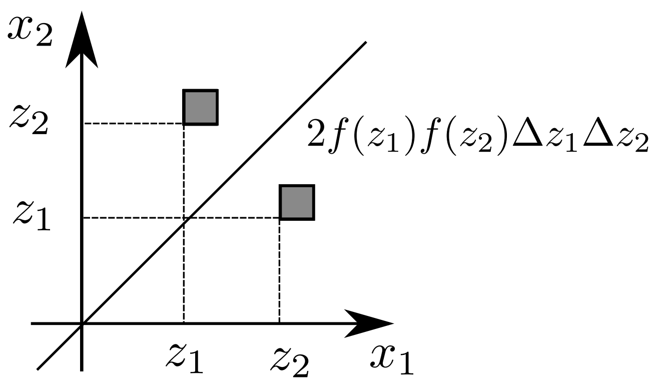
Solution. In genreal, if $X_1, \dots, X_n$ are i.i.d. continuous random variables with a density function $f(x)$, then the joint density function of the order statistics $X_{(1)}, \dots, X_{(n)}$ is given by

$$g(z_1, \cdots, z_n) = \begin{cases} n! \prod_{t=1}^n f(z_t) & \text{if } z_1 < \cdots < z_n, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed, for any $z_1 < \cdots < z_n$ and sufficiently small numbers $\Delta z_1, \dots, \Delta z_n > 0$, we have

$$\begin{aligned}
&P(X_{(i)} \in [z_i, z_i + \Delta z_i],\ 1 \leq i \leq n) \\
&= P\left(\bigcup_{\sigma \in S_n} \left\{ X_{\sigma(i)} \in [z_i, z_i + \Delta z_i],\ 1 \leq i \leq n \right\} \right) \\
&= \sum_{\sigma \in S_n} P\left(X_{(\sigma(i))} \in [z_i, z_i + \Delta z_i],\ 1 \leq i \leq n \right) \\
&= \sum_{\sigma \in S_n} \prod_{i=1}^n P\left(X_{(\sigma(i))} \in [z_i, z_i + \Delta z_i] \right) \\
&= n! \prod_{i=1}^n \left(f(z_i) \Delta z_i + o(\Delta z_i) \right),
\end{aligned}$$

where S_n denotes the set of all permutations. See the following figure for $n = 2$.



(i) If $f(x;\theta) = \theta e^{-\theta x}$, the joint density function is

$$g(z_1, \ldots, z_n) = \begin{cases} n! \prod_{t=1}^{n} \theta e^{-\theta z_t} & \text{if } z_1 < \cdots < z_n, \\ 0 & \text{otherwise.} \end{cases}$$

(ii) Denote the observed values by $x_{(1)}, \ldots, x_{(k)}$. The likelihood function is the joint density function of $X_{(1)}, \ldots, X_{(k)}$, that is,

$$\begin{aligned}
L(\theta) &= \int_{x_{(k)} < z_{k+1} < \cdots < z_n} g(x_{(1)}, \ldots, x_{(k)}, z_{k+1}, \ldots, z_n) dz_{k+1} \cdots dz_n \\
&= n! \left(\prod_{i=1}^{k} f(x_i;\theta) \right) \int_{x_{(k)} < z_{k+1} < \cdots < z_n} \left(\prod_{i=k+1}^{n} f(z_i;\theta) \right) dz_{k+1} \cdots dz_n \\
&= n! \left(\prod_{i=1}^{k} f(x_i;\theta) \right) \frac{1}{(n-k)!} \left(\int_{x_{(k)}}^{\infty} f(z;\theta) dz \right)^{n-k} \\
&= \frac{n!}{(n-k)!} \left(\prod_{i=1}^{k} \theta e^{-\theta x_i} \right) \left(\int_{x_{(k)}}^{\infty} \theta e^{-\theta z} dz \right)^{n-k} \\
&= \frac{n!}{(n-k)!} \theta^k e^{-\theta \sum_{i=1}^{k} x_{(i)}} e^{-(n-k)\theta x_{(k)}} \\
&= \frac{n!}{(n-k)!} \theta^k e^{-\theta \sum_{i=1}^{k-1} x_{(i)} - (n-k+1)\theta x_{(k)}}
\end{aligned}$$

The log-likelihood function is

$$\log L(\theta) = \log \frac{n!}{(n-k)!} + k \log \theta - \theta \left(\sum_{i=1}^{k-1} x_{(i)} + (n-k+1) x_{(k)} \right).$$

The likelihood equation is

$$\frac{\partial}{\partial \theta} \log L(\theta) = \frac{k}{\theta} - \left(\sum_{i=1}^{k-1} x_{(i)} + (n-k+1) x_{(k)} \right) = 0.$$

The MLE is

$$\hat{\theta} = \frac{k}{\sum_{i=1}^{k-1} x_{(i)} + (n-k+1) x_{(k)}}.$$

□

Q7-4

Consider the timestamps at which the students answered the questionnaire after the first lecture. Let us make a statistical model in the framework of survival analysis. Suppose that the number J of students attending the lecture is known ($J = 122$) and the timestamps are independent and identically distributed. For each student, when he/she did not answer the questionnaire by the time t, the conditional probability of answering between t and $t + \mathrm{d}t$ is assumed to be $\lambda(t)\mathrm{d}t$. The function $\lambda(t)$ is called the hazard function. Denote the cumulative distribution function and density function of the timestamp by $F(t)$ and $f(t) = F'(t)$, respectively. They have a relationship

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

(why?). The function $1 - F(t)$ is called the survival function.
Now suppose that the starting time of the questionnaire is 0 and the censoring time is $c > 0$. Let n be the number of students who answered and $t_1 < \cdots < t_n$ be the timestamps. Then, derive the likelihood function when the hazard function is

$$\lambda(t) = \frac{\phi}{\eta}\left(\frac{t}{\eta}\right)^{\phi-1}, \quad \eta > 0, \quad \phi > 0. \tag{$*$}$$

Furthermore, calculate the maximum likelihood estimate of (η, ϕ) for the real data by using a computer.
Note: The probability distribution having the hazard function in Eq. $(*)$ is called the Weibull distribution.

Solution. The likelihood function is given by

$$L = \left(\prod_{i=1}^{n} f(t_i)\right)(1 - F(c))^{J-n} \tag{1}$$

There are several ways to derive it. One of them is described as follows. See also Problem 7-3.

The probability that no students answer between 0 and t_1 is $(1 - F(t_1))^J$. Then the probability that one of them answers between t_1 and $t_1 + \mathrm{d}t_1$ is $J\lambda(t_1)\mathrm{d}t_1$. Next, the probability that the rest of students do not answer between t_1 and t_2 is $\{(1 - F(t_2))/(1 - F(t_1))\}^{J-1}$. By repeating the process, we finally obtain the overall probability

$$(1 - F(t_1))^J \left[\prod_{i=1}^{n}(J - i + 1)\lambda(t_i)\mathrm{d}t_i \left(\frac{1 - F(t_{i+1})}{1 - F(t_i)}\right)^{J-i}\right],$$

where $t_{n+1} = c$. This is equal to (1) except for the factor $\prod_{i=1}^{n}(J - i + 1)\mathrm{d}t_i$, which can be omitted since it does not depend on the parameters.

Now, if the hazard function is $\lambda(t) = \phi t^{\phi-1}/\eta^{\phi}$, then the log likelihood function is

$$\log L(\eta, \phi) = \sum_{i=1}^{n}\{\log\phi + (\phi - 1)\log t_i - \phi\log\eta - (t_i/\eta)^{\phi}\} - (J - n)(c/\eta)^{\phi}.$$

The maximum likelihood estimate of η given ϕ is shown to be

$$\eta = \eta(\phi) = \left(\frac{1}{n} \sum_{i=1}^{n} (t_i)^{\phi} + \frac{J-n}{n} c^{\phi} \right)^{1/\phi}.$$

To obtain the MLE of ϕ, we have to numerically maximize

$$\log L(\eta(\phi), \phi) = n \log \phi + (\phi - 1) \sum_{i=1}^{n} \log(t_i) - n \log \left(\frac{1}{n} \sum_{i=1}^{n} (t_i)^{\phi} + \frac{J-n}{n} c^{\phi} \right) - n.$$

For the given questionnaire data, the maximum likelihood estimate is

$$\hat{\eta} = 5.756 \times 10^7 \quad \text{[sec]}, \qquad \hat{\phi} = 0.266,$$

where the number J of the students is 122, and the start and censoring time are assumed to be 13:00:00 on Sep. 27 and 23:59:59 on Oct. 3, respectively. Here is an R code:

```
### data
X = read.csv("questionnaire2017.csv")
timeStamp = as.POSIXct(X[,1], format="%Y/%m/%d %H:%M:%S")  # observation
t0 = as.POSIXct("2017-09-27 13:00:00")  # start time
t1 = as.POSIXct("2017-10-03 23:59:59")  # censoring time
Ts = as.numeric(difftime(timeStamp, t0, units="secs"))
Ts = Ts[Ts > 0]   # remove negative values
c = as.numeric(difftime(t1, t0, units="secs"))
n = length(Ts)  # number of observations
J = 122         # number of students

### MLE
f = function(phi){
  n*log(phi) + phi*sum(log(Ts)) - n*log(mean((Ts)^phi) + (J-n)/n*c^phi)
}
phi.mle = optimize(f, c(0, 4), maximum=TRUE)$maximum
eta.mle = (mean((Ts)^phi.mle) + (J-n)/n*c^phi.mle)^(1/phi.mle)
```

□

Q7-5

For an exponential family

$$f(x; \boldsymbol{\theta}) = a(x) e^{\boldsymbol{\theta}' \boldsymbol{s}(x) - \psi(\boldsymbol{\theta})}, \quad x \in \mathbb{R}, \quad \boldsymbol{\theta} \in \mathbb{R}^d,$$

prove that the function $\psi(\boldsymbol{\theta})$ is convex. When is it strictly convex?

Solution. In order that f is a probability density function, $a(x)$ should be non-negative for all $x \in \mathbb{R}$ and strictly positive on a subset $\mathcal{X} \subset \mathbb{R}$ with positive measure.

The function $\psi(\boldsymbol{\theta})$ is determined by the condition $\int f(x; \boldsymbol{\theta}) \mathrm{d}x = 1$. Indeed, we have

$$\psi(\boldsymbol{\theta}) = \log \left(\int a(x) e^{\boldsymbol{\theta}' \boldsymbol{s}(x)} \mathrm{d}x \right).$$

Denote the domain of ψ by $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \psi(\boldsymbol{\theta}) < \infty\}$. As we observed in Problem 7-2, the set Θ may be a proper subset of $\mathbb{R}^d$. In the following, we assume that Θ has an interior point (i.e., there exist $\boldsymbol{\theta}_0 \in \Theta$ and $\delta > 0$ such that $\{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta\} \subset \Theta$).

The first derivative of ψ is

$$\mu_i := \frac{\partial \psi}{\partial \theta_i} = \frac{\int s_i(x) a(x) e^{\boldsymbol{\theta}' \boldsymbol{s}(x)} \mathrm{d}x}{\int a(x) e^{\boldsymbol{\theta}' \boldsymbol{s}(x)} \mathrm{d}x} = \int s_i(x) a(x) e^{\boldsymbol{\theta}' \boldsymbol{s}(x) - \psi(\boldsymbol{\theta})} \mathrm{d}x.$$

The second derivative is

$$\begin{aligned}\frac{\partial^2 \psi}{\partial \theta_i \partial \theta_j} = \frac{\partial \mu_i}{\partial \theta_j} &= \int s_i(x) \left(s_j(x) - \frac{\partial \psi}{\partial \theta_j} \right) a(x) e^{\boldsymbol{\theta}' \boldsymbol{s}(x) - \psi(\boldsymbol{\theta})} \mathrm{d}x \\ &= \int (s_i(x) - \mu_i)(s_j(x) - \mu_j) a(x) e^{\boldsymbol{\theta}' \boldsymbol{s}(x) - \psi(\boldsymbol{\theta})} \mathrm{d}x.\end{aligned}$$

This is the covariance matrix of $\boldsymbol{s}(x)$ with respect to $f(x; \boldsymbol{\theta})$ and therefore positive semi-definite, which implies ψ is convex.

From the proof above, we see that ψ is strictly convex if and only if the covariance matrix of $\boldsymbol{s}(x)$ is positive definite (for any $\boldsymbol{\theta} \in \Theta$). Furthermore, this condition holds if and only if the d functions $s_1(x), \ldots, s_d(x)$ are affinely independent, that is, for any non-zero real vector $(c_0, c_1, \cdots, c_d)$, the function $c_0 + c_1 s_1(x) + \cdots + c_d s_d(x)$ is not zero as a function of x. Indeed, if they are affinely independent, then $s_1(x) - \mu_1, \ldots, s_d(x) - \mu_d$ are linearly independent and therefore the covariance matrix of $s_i(x)$ is positive definite. Conversely, if the covariance matrix is positive definite, then $s_1(x) - \mu_1, \ldots, s_d(x) - \mu_d$ are linearly independent. Suppose now $c_0 + \sum_i c_i s_i(x) = 0$ for some c_i's. Then $\tilde{c_0} + \sum_i c_i (s_i(x) - \mu_i) = 0$, where $\tilde{c_0} = c_0 + \sum_i c_i \mu_i$. However, since $E[s_i(X)] = \mu_i$ for all i, we have $\tilde{c_0} = 0$. We also deduce that $c_i = 0$ by linear independence of $s_i(x) - \mu_i$.

Remark: the above proof is essentially the same as a proof of Hölder's inequality

$$\int F(x) G(x) dx \leq \left\{ \int F(x)^p dx \right\}^{1/p} \left\{ \int G(x)^q dx \right\}^{1/q}$$

for any non-negative functions $F(x)$ and $G(x)$ as long as $p, q > 0$ and $1/p + 1/q = 1$. Indeed, putting $f(x) = F(x)^p$, $g(x) = G(x)^q$ and $\lambda = 1 - 1/p$, the inequality is equivalent to

$$\int f(x)^{1-\lambda} g(x)^\lambda dx \leq \left\{ \int f(x) dx \right\}^{1-\lambda} \left\{ \int g(x) dx \right\}^\lambda,$$

which is implied from convexity of

$$\phi(\lambda) = \log \left(\int f(x)^{1-\lambda} g(x)^\lambda dx \right), \quad 0 \leq \lambda \leq 1.$$

□

8 Asymptotic normality and confidence intervals

Q8-1 (Basics of interval estimation)

Let $X_1, \dots, X_n$ be independent and distributed according to the normal distribution with mean μ and variance σ^2, where σ^2 is assumed to be known. Show that

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right]$$

is a 95% confidence interval of μ, where $\bar{X}$ denotes the sample mean of $X_1, \dots, X_n$.

Solution. By definition, a 95 % confidence interval $C(\boldsymbol{X})$ satisfies

$$\mathrm{P}_\mu(\mu \in C(\boldsymbol{X})) = 0.95.$$

Consider a random variable

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma},$$

which has the standard normal distribution $N(0, 1)$. We have

$$\begin{aligned}
0.95 &= \mathrm{P}_\mu(-1.96 \le Z \le 1.96) \\
&= \mathrm{P}_\mu\left(-1.96 \le \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \le 1.96\right) \\
&= \mathrm{P}_\mu\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= \mathrm{P}_\mu\left(-1.96\frac{\sigma}{\sqrt{n}} \le \mu - \bar{X} \le 1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= \mathrm{P}\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

Hence $[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}]$ is the 95% confidence interval.

Note: Strictly speaking, there are other 95% confidence intervals such as $[X_1 - 1.96\sigma, X_1 + 1.96\sigma]$. The interval we obtained above is the smallest one in a class of intervals $[\sum_i w_i X_i + a, \sum_i w_i X_i + b]$ with constants $w_1, \dots, w_n, a, b$. □

Q8-2 (Review of the central limit theorem)

Let $X_1, \dots, X_n$ be independent and distributed according to the uniform distribution on $[0, 1]$. Define

$$Z_n = \left(\sum_{i=1}^n \cos(2\pi X_i)\right)^2 + \left(\sum_{i=1}^n \sin(2\pi X_i)\right)^2.$$

Show that Z_n/n converges in distribution to the exponential distribution with mean 1 as $n \to \infty$.

Solution. The mean vector and covariance matrix of $(\cos(2\pi X_1), \sin(2\pi X_1))$ is obtained as follows:

$$\begin{aligned}
\mathrm{E}[\cos(2\pi X_1)] &= \int_0^1 \cos(2\pi x)dx = 0, \\
\mathrm{E}[\sin(2\pi X_1)] &= \int_0^1 \sin(2\pi x)dx = 0, \\
\mathrm{E}[\cos^2(2\pi X_1)] &= \int_0^1 \cos^2(2\pi x)dx = \frac{1}{2}, \\
\mathrm{E}[\cos(2\pi X_1)\sin(2\pi X_1)] &= \int_0^1 \cos(2\pi x)\sin(2\pi x)dx = 0, \\
\mathrm{E}[\sin^2(2\pi X_1)] &= \int_0^1 \sin^2(2\pi x)dx = \frac{1}{2}.
\end{aligned}$$

Let $U_n = \sum_{i=1}^n \cos(2\pi X_i)$ and $V_n = \sum_{i=1}^n \sin(2\pi X_i)$. Then $Z_n = U_n^2 + V_n^2$. By the central limit theorem, we have

$$\frac{1}{\sqrt{n}}\begin{pmatrix} U_n \\ V_n \end{pmatrix} \xrightarrow{\mathrm{d}} \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}\right).$$

Let U, V be independent random variables having $N(0, 1/2)$. Then $U^2 + V^2$ has the exponential distribution of mean 1. Indeed, for any $z \geq 0$,

$$P(U^2 + V^2 \geq z) = \iint_{u^2+v^2 \geq z} \frac{1}{\pi} e^{-(u^2+v^2)} dudv = \int_{\sqrt{z}}^\infty 2re^{-r^2} dr = e^{-z}.$$

Finally we have

$$P(Z_n/n \geq z) = P((U_n^2 + V_n^2)/n \geq z) \to P(U^2 + V^2 \geq z) = e^{-z},$$

which means that Z_n/n converges in distribution to the exponential distribution. □

Q8-3

Let $X_1, \ldots, X_n$ be independent and identically distributed with mean μ and variance σ^2.

(i) Let $\bar{X}$ be the sample mean. Find the asymptotic distribution of $\sqrt{n}(\bar{X} - \mu)$.

(ii) Obtain an approximate 95% confidence interval of μ, where σ^2 is unknown.

Solution. (i) By the central limit theorem, $\sqrt{n}(\bar{X} - \mu)$ converges in distribution to $N(0, \sigma^2)$.
(ii) It is deduced from (i) that the distribution of

$$Z_n := \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}}$$

is approximated by the standard normal distribution[1]. We have an approximate 95% confidence interval

$$\left[\bar{X} - \frac{1.96\hat{\sigma}}{\sqrt{n}}, \bar{X} + \frac{1.96\hat{\sigma}}{\sqrt{n}}\right]$$

in the same way as Q8-1. □

Q8-4

Let $X_1, \ldots, X_n$ be a Bernoulli trial with probability p.

(i) Let $\hat{p}$ be the maximum likelihood estimator. Find the asymptotic distribution of $\sqrt{n}(\hat{p} - p)$.

(ii) Obtain an approximate 95% confidence interval of p.

Solution. (i) The MLE is $\hat{p} = \bar{X} = (X_1 + \cdots + X_n)/n$. The asymptotic distribution of $\sqrt{n}(\hat{p} - p)$ is $\mathrm{N}(0, p(1-p))$ by the central limit theorem. Alternatively, we can use the asymptotic normality of $\hat{\theta}$, that is, the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ for any MLE is $\mathrm{N}(0, I(\theta)^{-1})$. Here the Fisher information is $I(p) = 1/(p(1-p))$.
(ii) It is deduced from (i) that the distribution of

$$Z_n := \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}}$$

is approximated by the standard normal distribution. In the same way as Q8-1, we obtain an approximate 95 % confidence interval

$$\left[\hat{p} - \frac{1.96\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + \frac{1.96\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right].$$

□

[1] Strictly speaking, the convergence of $\sqrt{n}(\bar{X} - \mu)/\hat{\sigma}$ to $\mathrm{N}(0,1)$ is a consequence of Slutsky's lemma, but this is beyond the scope of this lecture.

Q8-5 (difficult)

We prove consistency and asymptotic normality of MLE for exponential families. Let $X_1, \ldots, X_n$ be independent and have the probability density function $f(x;\theta) = a(x)e^{\theta s(x)-\psi(\theta)}$ ($\theta \in \mathbb{R}$). Let $\mu(\theta) = \psi'(\theta)$ and denote the Fisher information of θ by $I(\theta)$.

(i) Show that $I(\theta) = \psi''(\theta) = \mu'(\theta)$.

(ii) Show that the MLE $\hat{\theta}$ solves $\mu(\hat{\theta}) = \bar{s}$, where $\bar{s} = n^{-1}\sum_{t=1}^{n} s(X_t)$.

(iii) Show that the mean and variance of $s(X_1)$ is $\mu(\theta)$ and $I(\theta)$, respectively.

(iv) Use the law of large numbers to show that $\bar{s}$ converges in probability to $\mu(\theta)$.

(v) Show that $\hat{\theta}$ has consistency, that is, $\hat{\theta}$ converges in probability to θ. You can use the following lemma.

> Lemma Let Y_n be a sequence of random variables that converges in probability to a constant $c \in \mathbb{R}$. Then $h(Y_n)$ converges in probability to $h(c)$ for any continuous function $h(y)$.

(vi) Use the central limit theorem to show that $\sqrt{n}(\bar{s}-\mu(\theta))$ converges in distribution to $\mathrm{N}(0, I(\theta))$.

(vii) Show that $\sqrt{n}(\hat{\theta}-\theta)$ converges in distribution to $\mathrm{N}(0, I(\theta)^{-1})$. You can use the following theorem.

> Theorem (Delta method) Let Y_n be a sequence of random variables such that $\sqrt{n}(Y_n - c)$ converges in distribution to $\mathrm{N}(0, \sigma^2)$ for some constants $c \in \mathbb{R}$ and $\sigma^2 > 0$. Then $\sqrt{n}(h(Y_n) - h(c))$ converges in distribution to $\mathrm{N}(0, h'(c)^2\sigma^2)$ for any C^1-class function $h(y)$.

Solution. (i) Use a formula for the Fisher information (see Problem 6-4) to obtain

$$I(\theta) = \mathrm{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(X_1;\theta)\right] = \psi''(\theta).$$

(ii) The likelihood equation is

$$0 = \frac{\partial}{\partial\theta}\sum_{t=1}^{n}\log f(X_t;\theta) = \sum_{t=1}^{n}\{s(X_t) - \mu(\theta)\} = n(\bar{s} - \mu(\theta)).$$

Then the MLE $\hat{\theta}$ is the solution to $\mu(\hat{\theta}) = \bar{s}$.
(iii) Use the identity

$$\mathrm{E}_\theta[(\partial/\partial\theta)\log f(X_1;\theta)] = 0$$

(see Problem 6-4). We obtain $\mathrm{E}_\theta[s(X_1)] = \mu(\theta)$. The variance of $s(X_1)$ is

$$\begin{aligned}
\mathrm{V}_\theta[s(X_1)] &= \mathrm{E}_\theta[(s(X_1) - \mu(\theta))^2] \\
&= \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right)^2\right] \\
&= I(\theta),
\end{aligned}$$

where the last equality follows from the definition of the Fisher information.
(iv) By the weak law of large numbers, $\bar{s}$ converges in probability to $\mathrm{E}_\theta[s(X_1)] = \mu(\theta)$, since the condition $\mathrm{V}_\theta[s(X_1)] < \infty$ is verified.
(v) Note that $\hat{\theta}$ is written as $\hat{\theta} = \mu^{-1}(\bar{s})$, where μ^{-1} denotes the inverse map of μ. Since μ is continuous and $\mu'(\theta) = I(\theta) > 0$, the inverse map exists and is continuous. Therefore $\mu^{-1}(\bar{s})$ converges to $\mu^{-1}(\mu(\theta)) = \theta$ in probability.
(vi) It follows from the central limit theorem that $\sqrt{n}(\bar{s} - \mathrm{E}_\theta[s(X_1)])$ converges to the normal distribution $\mathrm{N}(0, \mathrm{V}_\theta[s(X_1)])$ in distribution. Then, by (iii), we have

$$\sqrt{n}(\bar{s} - \mu(\theta)) \xrightarrow{\mathrm{d}} \mathrm{N}(0, I(\theta)).$$

(vii) Apply the delta method to $Y_n = \bar{s}$, $c = \mu(\theta)$ and $h(y) = \mu^{-1}(y)$. Then we obtain

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}(h(\bar{s}) - h(c)) \xrightarrow{\mathrm{d}} \mathrm{N}(0, h'(c)^2 I(\theta)).$$

By using the inverse function theorem, we have

$$h'(c) = (\mu^{-1})'(c) = \frac{1}{\mu'(\mu^{-1}(c))} = \frac{1}{I(\mu^{-1}(c))} = \frac{1}{I(\theta)}.$$

Therefore,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \frac{1}{I(\theta)}\right).$$

□

Q8-6

For the questionnaire data, obtain an approximate 95% confidence interval of the mean sleeping time of students.

Solution. In the questionnaire data, the number of answering students was $n = 30$ (after a missing value is removed) and the frequency table of sleeping time is as follows:

sleeping time	5	6	7	8	9
frequency	1	12	9	7	1

The sample mean and standard deviation of sleeping time are $\bar{X} = 6.83$ [hours] and $\hat{\sigma} = 0.95$ [hours], respectively. Hence the 95% confidence interval based on the asymptotic normality of $\bar{X}$ is

$$\begin{aligned}\bar{X} \pm \frac{1.96\hat{\sigma}}{\sqrt{n}} &= 6.83 \pm \frac{1.96 \times 0.95}{\sqrt{30}} \\ &= 6.83 \pm 0.34 \\ &= [6.49, 7.17].\end{aligned}$$

□

9 Hypothesis testing

Q9-1 (Basics of hypothesis testing)

Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance 1. Consider a testing procedure of rejection region $R = \{\boldsymbol{x} \in \mathbb{R}^n \mid |\bar{x}| \geq c\}$ under H_0. Determine the value c that makes the significance level 0.05. What changes if the rejection region is $R = \{\boldsymbol{x} \in \mathbb{R}^n \mid \bar{x} \geq c\}$? Recall that the upper 2.5% and 5% points of the standard normal distribution are 1.96 and 1.64, respectively.

Solution. The constant c is determined by

$$\mathrm{P}_{\mu=0}(|\bar{X}| \geq c) = 0.05.$$

Since $\bar{X}$ has the distribution $\mathrm{N}(0, 1/n)$ under the null hypothesis, $Z := \sqrt{n}\bar{X}$ has $\mathrm{N}(0, 1)$. Then we obtain

$$\mathrm{P}(|Z| \geq \sqrt{n}c) = 0.05.$$

Therefore $c = 1.96/\sqrt{n}$.

If the rejection region is replaced with $R = \{\boldsymbol{x} \mid \bar{X} \geq c\}$, then c is determined by

$$\mathrm{P}_{\mu=0}(\bar{X} \geq c) = 0.05.$$

In the same way as above, we have $c = 1.64/\sqrt{n}$, where 1.64 is the upper 5% point of $\mathrm{N}(0, 1)$. □

Q9-2

Under the assumptions of Q9-1, find the log-likelihood ratio test statistic.

Solution. The likelihood function $L(\mu)$ is

$$L(\mu) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_t-\mu)^2/2}.$$

The MLE (maximum likelihood estimator) is $\hat{\mu} = \bar{x} = (1/n)\sum_{t=1}^n x_t$. Hence the LLR (the log-likelihood ratio test statistic) is

$$\begin{aligned}
2\log\frac{L(\hat{\mu})}{L(0)} &= 2\log\frac{\prod_{t=1}^n (2\pi)^{-1/2}e^{-(x_t-\bar{x})^2/2}}{\prod_{t=1}^n (2\pi)^{-1/2}e^{-x_t^2/2}} \\
&= -\sum_{t=1}^n (x_t-\bar{x})^2 + \sum_{t=1}^n x_t^2 \\
&= n\bar{x}^2.
\end{aligned}$$

Note: The test statistic $n\bar{X}^2$ has the chi-square distribution with degree 1 under the null hypothesis $\mu = 0$. □

Q9-3

Let $X_1, \ldots, X_n$ be independent random variables having a density function (or probability function) $f(x;\theta)$. For the following cases, write down the LLR (log-likelihood ratio test statistic) in terms of $n, \theta_0, \hat{\theta}$ when the null hypothesis is $\theta = \theta_0$ and the alternative hypothesis is $\theta \in \Theta \setminus \{\theta_0\}$. Here $\hat{\theta} \in \Theta$ is the MLE.

(i) Bernoulli distribution $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x \in \{0,1\}$, $\Theta = (0,1)$.

(ii) Poisson distribution $f(x;\theta) = (\theta^x/x!)e^{-\theta}$, $x \in \{0,1,\cdots\}$, $\theta = (0,\infty)$.

(iii) Exponential distribution $f(x;\theta) = \theta e^{-\theta x}$, $x \geq 0$, $\Theta = (0,\infty)$.

(iv) Normal distribution with unknown variance $f(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/(2\sigma^2)}$, $x \in \mathbb{R}$, $\Theta = \mathbb{R}\times(0,\infty)$.

Note: If the model is not an exponential family, it is not necessary that the LLR depends only on $n, \theta_0, \hat{\theta}$.

Solution. (i) The likelihood function is

$$L(\theta) = \prod_{t=1}^n \theta^{x_t}(1-\theta)^{1-x_t}$$

and the MLE is $\hat{\theta} = \bar{x} = n^{-1}\sum_{t=1}^n x_t$. The LLR is

$$\begin{aligned}
2\log\frac{L(\hat{\theta})}{L(\theta_0)} &= 2\log\frac{\prod_{t=1}^n \hat{\theta}^{x_t}(1-\hat{\theta})^{1-x_t}}{\prod_{t=1}^n \theta_0^{x_t}(1-\theta_0)^{1-x_t}} \\
&= 2n\left(\hat{\theta}\log\frac{\hat{\theta}}{\theta_0} + (1-\hat{\theta})\log\frac{1-\hat{\theta}}{1-\theta_0}\right).
\end{aligned}$$

(ii) The likelihood function is

$$L(\theta) = \prod_{t=1}^n \frac{\theta^{x_t}e^{-\theta}}{x_t!}$$

and the MLE is $\hat{\theta} = \bar{x} = n^{-1}\sum_{t=1}^n x_t$. The LLR is

$$\begin{aligned} 2\log\frac{L(\hat{\theta})}{L(\theta_0)} &= 2\log\frac{\prod_{t=1}^n \hat{\theta}^{x_t}e^{-\hat{\theta}}/(x_t!)}{\prod_{t=1}^n \theta_0^{x_t}e^{-\theta_0}/(x_t!)} \\ &= 2n\left(\hat{\theta}\log\frac{\hat{\theta}}{\theta_0} - \hat{\theta} + \theta_0\right). \end{aligned}$$

(iii) The likelihood function is

$$L(\theta) = \prod_{t=1}^n \theta e^{-\theta x_t}$$

and the MLE is $\hat{\theta} = 1/\bar{x} = n/\sum_{t=1}^n x_t$. The LLR is

$$\begin{aligned} 2\log\frac{L(\hat{\theta})}{L(\theta_0)} &= 2\log\frac{\prod_{t=1}^n \hat{\theta}e^{-\hat{\theta}x_t}}{\prod_{t=1}^n \theta_0 e^{-\theta_0 x_t}} \\ &= 2n\left(\log\frac{\hat{\theta}}{\theta_0} - 1 + \frac{\theta_0}{\hat{\theta}}\right). \end{aligned}$$

(iv) The likelihood function is

$$L(\mu,\sigma^2) = \prod_{t=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x_t-\mu)^2/(2\sigma^2)}$$

and the MLE is $\hat{\mu} = \bar{x} = n^{-1}\sum_{t=1}^n x_t$ and $\hat{\sigma}^2 = n^{-1}\sum_{t=1}^n (x_t-\bar{x})^2$. The LLR is

$$\begin{aligned} 2\log\frac{L(\hat{\mu},\hat{\sigma}^2)}{L(\mu_0,\sigma_0^2)} &= 2\log\frac{\prod_{t=1}^n (2\pi\hat{\sigma}^2)^{-1/2}e^{-(x_t-\hat{\mu})^2/(2\hat{\sigma}^2)}}{\prod_{t=1}^n (2\pi\sigma_0^2)^{-1/2}e^{-(x_t-\mu_0)^2/(2\sigma_0^2)}} \\ &= n\left(-\log\frac{\hat{\sigma}^2}{\sigma_0^2} - 1 + \frac{\hat{\sigma}^2 + (\hat{\mu}-\mu_0)^2}{\sigma_0^2}\right). \end{aligned}$$

Note: In general, for any exponential family $f(x;\theta) = a(x)e^{\theta s(x)-\psi(\theta)}$, the LLR is

$$2\log\frac{\prod_{t=1}^n f(x_t;\hat{\theta})}{\prod_{t=1}^n f(x_t;\theta_0)} = 2n\left((\hat{\theta}-\theta_0)\psi'(\hat{\theta}) - \psi(\hat{\theta}) + \psi(\theta_0)\right)$$

since the MLE satisfies the likelihood equation $\psi'(\hat{\theta}) = \bar{s} = n^{-1}\sum_{t=1}^n s(x_t)$. The LLR is equal to the Kullback-Leibler divergence (see Q6-5)

$$\int f(y;\hat{\theta})\log\frac{f(y;\hat{\theta})}{f(y;\theta_0)} = (\hat{\theta}-\theta_0)\psi'(\hat{\theta}) - \psi(\hat{\theta}) + \psi(\theta_0)$$

from $f(y;\hat{\theta})$ to $f(y;\theta_0)$, up to a multiplicative constant $2n$. $\square$

Q9-4 (ROC curve)

Let X have $\mathrm{N}(\mu, \sigma^2)$, the null hypothesis be $\mu = 0$ and the alternative hypothesis be $\mu = \delta > 0$. Consider a rejection region $R = \{x \in \mathbb{R} \mid |x| \geq c\}$. Letting $c \geq 0$ be a parameter, draw a graph between significance level and the power. What changes if $R = \{x \in \mathbb{R} \mid x \geq c\}$.

Solution. The number $c = c_\alpha$ satisfying $P_0(X \in R) = P_0(|X| \geq c) = \alpha$ is given by $c_\alpha = \sigma z_{\alpha/2}$, where $z_{\alpha/2}$ denotes the upper $\alpha/2$ point of the standard normal distribution, that is, $P(Z \geq z_{\alpha/2}) = \alpha/2$ for $Z \sim N(0,1)$. Then the power is

$$\begin{aligned} P_\delta(X \in R) &= P_\delta(|X| \geq \sigma z_{\alpha/2}) = P(|\delta + \sigma Z| \geq \sigma z_{\alpha/2}) \\ &= P\left(Z \geq -\frac{\delta}{\sigma} + z_{\alpha/2}\right) + P\left(Z \leq -\frac{\delta}{\sigma} - z_{\alpha/2}\right) \\ &= \Phi(\delta/\sigma - z_{\alpha/2}) + \Phi(-\delta/\sigma - z_{\alpha/2}), \end{aligned}$$

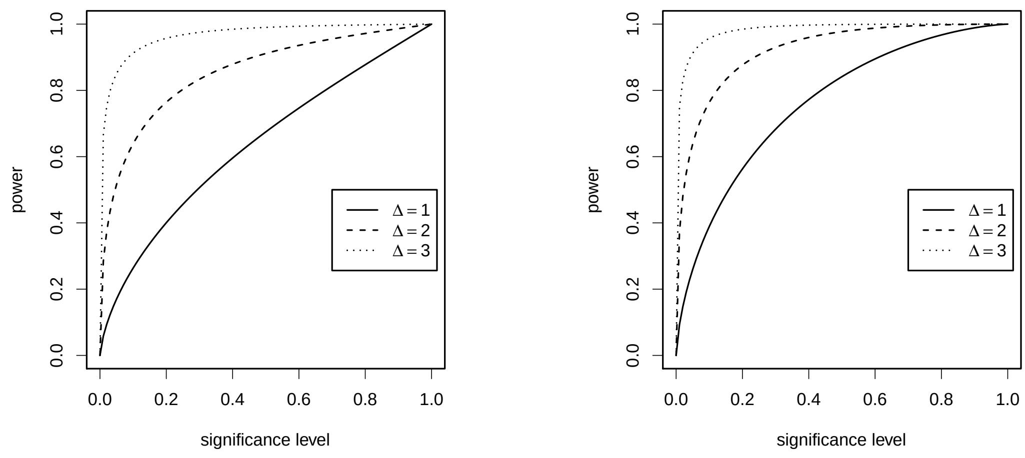
where Φ denotes the cumulative distribution function of the standard normal distribution. In particular, the power depends only on α and δ/σ.

If the rejection region is replaced with $R = \{x \geq c\}$, then the power is given by

$$P_\delta(X \geq \sigma z_\alpha) = \Phi(\delta/\sigma - z_\alpha)$$

against the significance level α.

The following figure shows the curve of the power against the significance level for several values of $\Delta = \delta/\sigma$.



(a) $R = \{|x| \geq c\}$.

(b) $R = \{x \geq c\}$.

□

Q9-5 (Neyman-Pearson lemma)

Consider a statistical model $f_n(\boldsymbol{x};\theta)$. Let the null and alternative hypotheses be $\theta=\theta_0$ and $\theta=\theta_1$, respectively. Suppose a condition

$$\mathrm{P}_{\theta_0}(f_n(\boldsymbol{X};\theta_1)/f_n(\boldsymbol{X};\theta_0)=c)=0, \quad \forall c\geq 0. \tag{$*$}$$

Then, show that a testing procedure

$$R=\{\boldsymbol{x}\mid f_n(\boldsymbol{x};\theta_1)/f_n(\boldsymbol{x};\theta_0)\geq c\}, \quad \mathrm{P}_{\theta_0}(\boldsymbol{X}\in R)=\alpha,$$

is the most powerful in all testing procedures with the significance level α.

Note: the condition $(*)$ is not satisfied for discrete distributions. For discrete distributions, the problem determining the most powerful testing procedure is equivalent to the knapsack problem, and therefore difficult to solve in general.

Solution. Let S be any rejection region of the significance level α, that is, $\mathrm{P}_{\theta_0}(\boldsymbol{X}\in S)\leq\alpha$. Then the difference of powers is

$$\begin{aligned}
\mathrm{P}_{\theta_1}(\boldsymbol{X}\in R)-\mathrm{P}_{\theta_1}(\boldsymbol{X}\in S) &= \int_R f_n(\boldsymbol{x};\theta_1)d\boldsymbol{x}-\int_S f_n(\boldsymbol{x};\theta_1)d\boldsymbol{x}\\
&= \int_{R\setminus S} f_n(\boldsymbol{x};\theta_1)d\boldsymbol{x}-\int_{S\setminus R} f_n(\boldsymbol{x};\theta_1)d\boldsymbol{x}\\
&\geq c\int_{R\setminus S} f_n(\boldsymbol{x};\theta_0)d\boldsymbol{x}-c\int_{S\setminus R} f_n(\boldsymbol{x};\theta_0)d\boldsymbol{x}\\
&= c\int_R f_n(\boldsymbol{x};\theta_0)d\boldsymbol{x}-c\int_S f_n(\boldsymbol{x};\theta_0)d\boldsymbol{x}\\
&= c\mathrm{P}_{\theta_0}(\boldsymbol{X}\in R)-c\mathrm{P}_{\theta_0}(\boldsymbol{X}\in S)\\
&\geq c\alpha-c\alpha\\
&= 0,
\end{aligned}$$

where the first inequality follows from the definition of R and the second inequality follows from $\mathrm{P}_{\theta_0}(\boldsymbol{X}\in R)=\alpha$ and $\mathrm{P}_{\theta_0}(\boldsymbol{X}\in S)\leq\alpha$. $\square$

Q9-6 (difficult)

Show that, for exponential families, the log-likelihood ratio test statistic converges in distribution to the chi-square distribution.

Solution. We only consider a one-parameter exponential family $f(x;\theta)=a(x)e^{\theta s(x)-\psi(\theta)}$ and a simple null hypothesis $\theta=\theta_0$. Refer to advanced textbooks for general cases. We know the maximum likelihood estimator $\hat{\theta}$ has asymptotic normality (see Problem 8-5). Denote the likelihood function by $L(\theta)=\prod_{t=1}^n f(x_t;\theta)$. The log-likelihood ratio test statistic is

$$
\begin{aligned}
T(\boldsymbol{x}) &= 2\log\frac{L(\hat\theta)}{L(\theta_0)} \\
&= 2\sum_{t=1}^{n}\log\frac{f(x_t;\hat\theta)}{f(x_t;\theta_0)} \\
&= 2\sum_{t=1}^{n}\{(\hat\theta-\theta_0)s(x_t)-\psi(\hat\theta)+\psi(\theta_0)\} \\
&= 2n\{(\hat\theta-\theta_0)\bar s-\psi(\hat\theta)+\psi(\theta_0)\} = n(\hat\theta-\theta_0)^2 I(\hat\theta_*),
\end{aligned}
$$

where the last equality follows from Taylor's formula applied to $\psi(\theta)$ around $\theta=\hat\theta$, and $\hat\theta_*$ is a value between θ_0 and $\hat\theta$. Note that $\bar s=\psi'(\hat\theta)$ and $I(\theta)=\psi''(\theta)$ (see Problem 9-3). By asymptotic normality of $\hat\theta$ (and Slutsky's lemma), we have

$$
\sqrt{nI(\hat\theta_*)}(\hat\theta(\boldsymbol{X})-\theta_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).
$$

Therefore $T(\boldsymbol{X})$ converges to $\chi^2(1)$ (by the continuous mapping theorem). $\square$

Remark. There are other two well-known test statistics that have the same asymptotic property as the likelihood ratio test. One is *the score test statistic* defined by

$$
T_{\text{score}}(\boldsymbol{x}) = \sum_{i,j=1}^{p} I_{(n)}^{ij}(\theta_0)\left(\frac{\partial}{\partial\theta_i}\log L(\theta_0)\right)\left(\frac{\partial}{\partial\theta_j}\log L(\theta_0)\right),
$$

where $(I_{(n)}^{ij})$ denotes the inverse matrix of the Fisher information matrix $\boldsymbol{I}_{(n)}=n\boldsymbol{I}$ of n observations. For the one-parameter exponential family, the statistic is

$$
T_{\text{score}}(\boldsymbol{x}) = (nI(\theta_0))^{-1}\left\{n(\bar s-\psi'(\theta_0))\right\}^2 = \frac{n(\bar s-\psi'(\theta_0))^2}{I(\theta_0)},
$$

which converges to $\chi^2(1)$ by the central limit theorem for $\bar s$. The other is *the Wald test statistic* defined by

$$
T_{\text{Wald}}(\boldsymbol{x}) = \sum_{i,j=1}^{p} I_{(n),ij}(\hat\theta)(\hat\theta_i-\theta_{0i})(\hat\theta_j-\theta_{0j}),
$$

where $(I_{(n),ij})=\boldsymbol{I}_{(n)}=n\boldsymbol{I}$. For the one-parameter exponential family, the statistic is

$$
T_{\text{Wald}}(\boldsymbol{x}) = nI(\hat\theta)(\hat\theta-\theta_0)^2,
$$

which converges to $\chi^2(1)$ since $\sqrt{nI(\hat\theta)}(\hat\theta-\theta_0)$ converges to $N(0,1)$. Compare it with the confidence interval of the MLE.

The likelihood-ratio and score tests are invariant under transformations of the parameter, but the Wald test is not. $\square$

Q9-7

(i) Consider a trinomial model with sample size n

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{3} \theta_i^{x_i}, \quad x_1 + x_2 + x_3 = n, \quad \theta_1 + \theta_2 + \theta_3 = 1.$$

Let the null hypothesis be $\theta_1 = \theta_3$. Find the log-likelihood ratio test statistic.

(ii) In order to study whether a newly developed beer product is better than the existing one, 40 people as experimental subjects tasted the two kinds of beer and answered which one was better. Here the subjects do not know which one is new. Suppose that the following table shows a result of the experiment. Examine the hypothesis testing if the significance level is $\alpha = 0.05$.

	new one is good	not different	existing one is good
frequency	17	10	13

Solution. (i) Let $\hat{\boldsymbol{\theta}} = \boldsymbol{x}/n$ be the MLE under the full model and $\tilde{\boldsymbol{\theta}}$ be the MLE under the restricted model $\theta_1 = \theta_3$. The LLR test statistic is given by

$$T(\boldsymbol{x}) = 2 \log \frac{f(\boldsymbol{x};\hat{\boldsymbol{\theta}})}{f(\boldsymbol{x};\tilde{\boldsymbol{\theta}})} = 2 \sum_{i=1}^{3} x_i \log \frac{\hat{\theta}_i}{\tilde{\theta}_i} = 2n \sum_{i=1}^{3} \hat{\theta}_i \log \frac{\hat{\theta}_i}{\tilde{\theta}_i}$$

The quantity $\tilde{\boldsymbol{\theta}}$ is obtained as follows. The likelihood under the null hypothesis is

$$L(\boldsymbol{\theta}) = \theta_1^{x_1+x_3}(1 - 2\theta_1)^{x_2}.$$

Then the likelihood equation is

$$\frac{d}{d\theta_1} \log L(\boldsymbol{\theta}) = \frac{x_1 + x_3}{\theta_1} - \frac{2x_2}{1 - 2\theta_1} = 0$$

and the solution is $\tilde{\theta}_1 = (x_1 + x_3)/(2n)$.
(ii) We have

$$\hat{\theta} = \frac{\boldsymbol{x}}{n} = \left(\frac{17}{40}, \frac{10}{40}, \frac{13}{40}\right)$$

and

$$\tilde{\theta} = \left(\frac{x_1 + x_3}{2n}, \frac{x_2}{n}, \frac{x_1 + x_3}{2n}\right) = \left(\frac{15}{40}, \frac{10}{40}, \frac{15}{40}\right).$$

See the following figure. Then

$$T(\boldsymbol{x}) = 2\left(17 \log \frac{17}{15} + 10 \log \frac{10}{10} + 13 \log \frac{13}{15}\right) = 0.535.$$

This is smaller than the critical value 3.84, the upper 5% point of the chi-square distribution with degree 1. Therefore we deduce that there is no significance.

In this case, the p-value is 0.465. This is actually larger than 0.05.

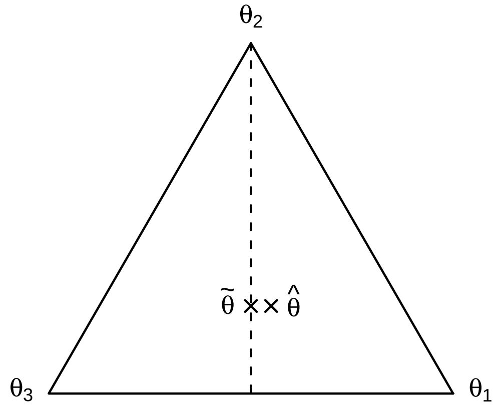


Figure: The estimates $\hat{\theta}$ and $\tilde{\theta}$. The dashed line denotes the null hypothesis.

$\square$

Remark. For this kind of problems, *Pearson's goodness-of-fit test* is often used. In general, consider a multinomial model

$$f_n(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^k \theta_i^{x_i}, \quad \sum_{i=1}^k x_i = n, \quad \boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta} \in \mathbb{R}^k_{\geq 0} \mid \sum_{i=1}^k \theta_i = 1\}$$

and a null hypothesis $\boldsymbol{\theta} \in \Theta_0$. Then, Pearson's goodness-of-fit test statistic is defined by

$$T_{\text{gof}}(\boldsymbol{x}) = \sum_{i=1}^k \frac{(x_i - n\tilde{\theta}_i)^2}{n\tilde{\theta}_i},$$

where $\tilde{\boldsymbol{\theta}}$ denotes the MLE under the null hypothesis, that is,

$$\tilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_0} f_n(\boldsymbol{x};\boldsymbol{\theta}).$$

The statistic also converges in distribution to the chi-square distribution with degree $d = \dim \Theta - \dim \Theta_0$.

For Q9-7, the statistic becomes

$$T_{\text{gof}}(\boldsymbol{x}) = \frac{(17-15)^2}{15} + \frac{(10-10)^2}{10} + \frac{(13-15)^2}{15} = 0.533,$$

close to the LLR statistic 0.535. $\square$

Remark of Remark. One can show that $T_{\text{gof}}(\boldsymbol{x})$ is equal to the score test statistic

$$T_{\text{score}}(\boldsymbol{x}) = \sum_{i,j=1}^{k-1} I_{(n)}^{ij}(\tilde{\boldsymbol{\theta}}) \left(\frac{\partial}{\partial \theta_i} \log f_n(\boldsymbol{x};\tilde{\boldsymbol{\theta}})\right) \left(\frac{\partial}{\partial \theta_j} \log f_n(\boldsymbol{x};\tilde{\boldsymbol{\theta}})\right)$$

(See the remark after the solution of Q9-6), where we consider $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})'$ as free variables and $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$. The Fisher information matrix is $I_{(n),ij} = n(\theta_i^{-1}\delta_{ij} - \theta_k^{-1})$ and its inverse matrix is $I_{(n)}^{ij} = (\theta_i \delta_{ij} - \theta_i\theta_j)/n$.

Let us prove $T_{\mathrm{score}}(\boldsymbol{x}) = T_{\mathrm{gof}}(\boldsymbol{x})$. Denoting $\theta_i = \tilde{\theta}_i$ for simplicity, we have

$$\begin{aligned} T_{\mathrm{score}}(\boldsymbol{x}) &= \sum_{i,j=1}^{k-1} \frac{\theta_i\delta_{ij} - \theta_i\theta_j}{n} \left(\frac{x_i}{\theta_i} - \frac{x_k}{\theta_k}\right)\left(\frac{x_j}{\theta_j} - \frac{x_k}{\theta_k}\right) \\ &= \frac{1}{n}\boldsymbol{u}' \underbrace{\begin{pmatrix} \boldsymbol{E}_{k-1} \\ -\boldsymbol{1}_{k-1}' \end{pmatrix} (\mathrm{Diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}') \begin{pmatrix} \boldsymbol{E}_{k-1} & -\boldsymbol{1}_{k-1} \end{pmatrix}}_{\boldsymbol{A}} \boldsymbol{u}, \end{aligned}$$

where $\boldsymbol{u} = (x_i/\theta_i)_{i=1}^k$, $\boldsymbol{E}_{k-1} = \mathrm{Diag}(\boldsymbol{1}_{k-1})$, $\boldsymbol{1}_{k-1} = (1, \ldots, 1)'$, and $\mathrm{Diag}(\boldsymbol{\theta})$ is the diagonal matrix with the diagonal elements $\boldsymbol{\theta}$. The matrix $\boldsymbol{A}$ is rewritten as

$$\begin{aligned} \boldsymbol{A} &= \begin{pmatrix} \mathrm{Diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}' & -\boldsymbol{\theta} + (\boldsymbol{1}'\boldsymbol{\theta})\boldsymbol{\theta} \\ -\boldsymbol{\theta}' + (\boldsymbol{1}'\boldsymbol{\theta})\boldsymbol{\theta}' & (\boldsymbol{1}'\boldsymbol{\theta}) - (\boldsymbol{1}'\boldsymbol{\theta})^2 \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{Diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}' & -\theta_k\boldsymbol{\theta} \\ -\theta_k\boldsymbol{\theta}' & \theta_k - \theta_k^2 \end{pmatrix} \qquad \left(\because \sum_{i=1}^k \theta_i = 1\right) \\ &= (\theta_i\delta_{ij} - \theta_i\theta_j)_{i,j=1}^k. \end{aligned}$$

Therefore

$$\begin{aligned} T_{\mathrm{score}}(\boldsymbol{x}) &= \frac{1}{n}\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u} \\ &= \sum_{i,j=1}^k (\theta_i\delta_{ij} - \theta_i\theta_j)\frac{x_ix_j}{n\theta_i\theta_j} \\ &= \left(\sum_{i=1}^k \frac{x_i^2}{n\theta_i}\right) - n \\ &= \sum_{i=1}^k \frac{(x_i - n\theta_i)^2}{n\theta_i}, \end{aligned}$$

which is equal to $T_{\mathrm{gof}}(\boldsymbol{x})$. $\square$

10 Normal linear model

Q10-1

Suppose that the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ is generated from a normal linear model

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n), \quad \boldsymbol{\mu} \in M, \quad \sigma^2 > 0, \tag{2}$$

where M is a linear subspace of $\mathbb{R}^n$. Let M_0 be a linear subspace of M and consider a null hypothesis $\mathrm{H}_0 : \boldsymbol{\mu} \in M_0$. Find the log-likelihood ratio test statistic $T(\boldsymbol{y})$ and show that $T(\boldsymbol{y})$ has an increasing relation with

$$F(\boldsymbol{y}) = \frac{\|\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}_0\boldsymbol{y}\|^2/(p - p_0)}{\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\|^2/(n-p)}.$$

Here $\|\cdot\|$ is the Euclidean norm, $\boldsymbol{P}$ and $\boldsymbol{P}_0$ are the orthogonal projection matrix to M and M_0, and p and p_0 are the dimension of M and M_0, respectively.

Remark: Under the null hypothesis, the statistic $F(\boldsymbol{Y})$ is shown to have the F-distribution. The hypothesis testing procedure based on the distribution is called the F-test.

Solution. The likelihood function is

$$L(\boldsymbol{\mu}, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\|\boldsymbol{y}-\boldsymbol{\mu}\|^2/2\sigma^2}, \quad \boldsymbol{\mu} \in M, \quad \sigma^2 > 0.$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\mu} \in M$ and $\sigma^2 > 0$ is given by $\hat{\boldsymbol{\mu}} = \boldsymbol{P}\boldsymbol{y}$ and $\hat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\|^2/n$. Note that $\hat{\sigma}^2$ is not unbiased. Similarly, the MLE under the null hypothesis $\boldsymbol{\mu} \in M_0$ is $\hat{\boldsymbol{\mu}}_0 = \boldsymbol{P}_0\boldsymbol{y}$ and $\hat{\sigma}_0^2 = \|\boldsymbol{y} - \boldsymbol{P}_0\boldsymbol{y}\|^2/n$. Then the log-likelihood ratio test statistic is

$$\begin{aligned}
T(\boldsymbol{y}) &= 2\log\frac{L(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2)}{L(\hat{\boldsymbol{\mu}}_0, \hat{\sigma}_0^2)} \\
&= -n\log\hat{\sigma}^2 - \frac{\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2} + n\log\hat{\sigma}_0^2 + \frac{\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0\|^2}{\hat{\sigma}_0^2} \\
&= -n\log\hat{\sigma}^2 + n\log\hat{\sigma}_0^2 \\
&= n\log\frac{\|\boldsymbol{y} - \boldsymbol{P}_0\boldsymbol{y}\|^2}{\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\|^2} \\
&= n\log\left(1 + \frac{\|\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}_0\boldsymbol{y}\|^2}{\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\|^2}\right),
\end{aligned}$$

where the last equality follows from a Pythagorean relation

$$\|\boldsymbol{y} - \boldsymbol{P}_0\boldsymbol{y}\|^2 = \|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\|^2 + \|\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}_0\boldsymbol{y}\|^2.$$

As a result, we have an increasing relation $T(\boldsymbol{y}) = n\log(1 + ((p - p_0)/(n-p))F(\boldsymbol{y}))$. $\square$

Q10-2

Rewrite the simple regression model

$$Y_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2), \quad i = 1, \dots, n,$$

as a normal linear model (2). Find the subspace M_0 corresponding to a null hypothesis $b = 0$.

Solution. The model is rewritten as

$$\boldsymbol{Y} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

Therefore M is the linear space spanned by $(1, \dots, 1)'$ and $(x_1, \dots, x_n)'$. The subspace M_0 corresponding to the null hypothesis $b = 0$ is spanned by $(1, \dots, 1)'$. □

Q10-3

Explain differences between the following two tests. Set up statistical models and null hypotheses by yourself.

(i) Test of the difference between the means of paired two samples.

(ii) Test of the difference between the means of unpaired two samples.

Solution. We derive the t-test statistic for each situation.
(i) Let $\{(x_i, y_i)\}_{i=1}^n$ be a paired sample. We assume a statistical model

$$Y_i - X_i \sim \mathrm{N}(a, \sigma^2),$$

where a and σ^2 are unknown parameters. This assumption is valid if X_i and Y_i are independent and

$$X_i \sim \mathrm{N}(\mu_i, \sigma^2/2), \quad Y_i \sim \mathrm{N}(\mu_i + a, \sigma^2/2)$$

for any unknown μ_i's. The null hypothesis is $a = 0$. Since $Y_i - X_i$ is considered as a random sample from $\mathrm{N}(a, \sigma^2)$, the t-test statistic is given by

$$T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sqrt{n}(\bar{y} - \bar{x})}{\hat{\sigma}}, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n ((y_i - x_i) - (\bar{y} - \bar{x}))^2,$$

with the degree of freedom $n - 1$. The statistic is also derived in the framework of normal linear models. Details are omitted here.

(ii) Let $\boldsymbol{x} = (x_i)_{i=1}^{n_1}$ and $\boldsymbol{y} = (y_j)_{j=1}^{n_2}$ be unpaired two samples. We assume a statistical model

$$X_i \sim \mathrm{N}(\mu, \sigma^2), \quad Y_j \sim \mathrm{N}(\mu + a, \sigma^2).$$

Note that μ cannot depend on the index i in contrast to the paired samples. The null hypothesis is $a = 0$. We show that the t-test statistic is given by

$$T_*(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{y} - \bar{x}}{\hat{\sigma}_*}, \quad \hat{\sigma}_*^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right), \qquad (*)$$

with the degree of freedom $n_1 + n_2 - 2$. The estimate $\hat{\sigma}_*^2$ is sometimes called the *pooled variance*.

Let us prove $(*)$. Let $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$ be the concatenated vector of $\boldsymbol{x}$ and $\boldsymbol{y}$. The subspaces M and M_0 of $\mathbb{R}^{n_1+n_2}$ corresponding to the full model and null hypothesis are

$$M = \mathrm{span} \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} \end{pmatrix} = \mathrm{span} \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} \end{pmatrix} \quad \text{and} \quad M_0 = \mathrm{span} \begin{pmatrix} \mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} \end{pmatrix},$$

where $\mathrm{span}(\boldsymbol{X})$ denotes the column space of a matrix $\boldsymbol{X}$. Let $\boldsymbol{P}$ be the orghotonal projection matrices to M. Let $\boldsymbol{e}$ be one of the two unit vectors orthogonal to M_0 in M. Specifically,

$$\boldsymbol{e} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \begin{pmatrix} -\frac{1}{n_1} \mathbf{1}_{n_1} \\ \frac{1}{n_2} \mathbf{1}_{n_2} \end{pmatrix}.$$

Then the t-statistic is defined by

$$T_*(\boldsymbol{x}, \boldsymbol{y}) := \frac{\boldsymbol{e}'\boldsymbol{z}}{\hat{\sigma}_*}, \quad \hat{\sigma}_*^2 := \frac{1}{n_1 + n_2 - 2} \|\boldsymbol{z} - \boldsymbol{P}\boldsymbol{z}\|^2.$$

By direct calculation, one can show that

$$\boldsymbol{P}\boldsymbol{z} = \begin{pmatrix} \bar{x} \mathbf{1}_{n_1} \\ \bar{y} \mathbf{1}_{n_2} \end{pmatrix}, \quad \|\boldsymbol{z} - \boldsymbol{P}\boldsymbol{z}\|^2 = \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2,$$

and

$$\boldsymbol{e}'\boldsymbol{z} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} (\bar{y} - \bar{x}).$$

Then we obtain $(*)$.

Even if $n_1 = n_2$, the statistic $T_*(\boldsymbol{x}, \boldsymbol{y})$ is different from $T(\boldsymbol{x}, \boldsymbol{y})$. Indeed, if $n_1 = n_2 = n$,

$$T_*(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sqrt{n}(\bar{y} - \bar{x})}{\hat{\tau}}, \quad \hat{\tau}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \{(x_i - \bar{x})^2 + (y_i - \bar{y})^2\}.$$

It is easy to see that $T(\boldsymbol{x}, \boldsymbol{y}) > T_*(\boldsymbol{x}, \boldsymbol{y})$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ have positive correlation. For example, let $n_1 = n_2 = 2$, $(x_1, y_1) = (0, 0)$ and $(x_2, y_2) = (50, 51)$. Then $T(\boldsymbol{x}, \boldsymbol{y}) = 1$ and $T_*(\boldsymbol{x}, \boldsymbol{y}) = 0.014$. The p-value for each statistic is 0.25 and 0.495, respectively. $\square$

Q10-4

Consider a multiple regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Answer the following questions about the test of the regression coefficient β_1 (other coefficients are similarly treated). Here we define $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ and $\boldsymbol{X}_0 = (\boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$. Denote the orthogonal matrices onto the column spaces of $\boldsymbol{X}$ and $\boldsymbol{X}_0$ by $\boldsymbol{P}$ and $\boldsymbol{P}_0$, respectively.

(i) Show that the covariance matrix of the least square estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ is given by $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. From this result, the standard error of the coefficient $\hat{\beta}_1$ is estimated by $\mathrm{se}(\hat{\beta}_1) = \hat{\sigma}\sqrt{((\boldsymbol{X}'\boldsymbol{X})^{-1})_{11}}$, where $\hat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\|^2/(n-p)$ is an unbiased estimator of σ^2 (see Q6-2).

(ii) Let $\boldsymbol{r} = \boldsymbol{x}_1 - \boldsymbol{P}_0\boldsymbol{x}_1$ and $\boldsymbol{e} = \boldsymbol{r}/\|\boldsymbol{r}\|$. Prove the following identity:

$$t(\boldsymbol{y}) := \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} = \frac{\boldsymbol{e}'(\boldsymbol{y} - \beta_1\boldsymbol{x}_1)}{\hat{\sigma}}.$$

Remark: The distribution of $t(\boldsymbol{Y})$ is shown to be the t-distribution. A testing procedure based on the result is called the t-test. The null hypothesis is usually $\beta_1 = 0$ and the value of $t(\boldsymbol{y})$ is studied. We also obtain a precise confidence interval of β_1.

Solution. Note that the projection matrices are given by $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{P}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'$.

(i) The least squares estimator is defined by the minimizer of $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ and explicitly given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

The expectation of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}[\boldsymbol{y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{\beta}.$$

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}\mathrm{Cov}[\hat{\boldsymbol{\beta}}] &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Cov}[\boldsymbol{y}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I}_n)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

(ii) By definition, the t-statistic $t(\boldsymbol{y})$ is

$$t(\boldsymbol{y}) = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\sqrt{((\boldsymbol{X}'\boldsymbol{X})^{-1})_{11}}}$$

We first show that $((\boldsymbol{X}'\boldsymbol{X})^{-1})_{11} = \|\boldsymbol{r}\|^{-2}$, where $\boldsymbol{r} = \boldsymbol{x}_1 - \boldsymbol{P}_0\boldsymbol{x}_1$, as defined in the problem. The matrix $\boldsymbol{X}$ is decomposed as

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{X}_0) = (\boldsymbol{r}, \boldsymbol{X}_0)\begin{pmatrix} 1 & \boldsymbol{0}' \\ * & \boldsymbol{I} \end{pmatrix},$$

where $*$ denotes elements that are not necessarily evaluated in the following calculation. Noting that $\boldsymbol{r}$ is orthogonal to each column of $\boldsymbol{X}_0$ (see the following figure), we have

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{X} &= \begin{pmatrix} 1 & * \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{r}' \\ \boldsymbol{X}_0' \end{pmatrix} \begin{pmatrix} \boldsymbol{r} & \boldsymbol{X}_0 \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}' \\ * & \boldsymbol{I} \end{pmatrix}. \\ &= \begin{pmatrix} 1 & * \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \|\boldsymbol{r}\|^2 & \mathbf{0}' \\ \mathbf{0} & * \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}' \\ * & \boldsymbol{I} \end{pmatrix}.\end{aligned}$$

The inverse matrix is

$$\begin{aligned}(\boldsymbol{X}'\boldsymbol{X})^{-1} &= \begin{pmatrix} 1 & \mathbf{0}' \\ * & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \|\boldsymbol{r}\|^{-2} & \mathbf{0}' \\ \mathbf{0} & * \end{pmatrix} \begin{pmatrix} 1 & * \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} \\ &= \begin{pmatrix} \|\boldsymbol{r}\|^{-2} & * \\ * & * \end{pmatrix}\end{aligned}$$

Hence we have proved that $((\boldsymbol{X}'\boldsymbol{X})^{-1})_{11} = \|\boldsymbol{r}\|^{-2}$. Now what we have to show is

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\|\boldsymbol{r}\|^{-1}} = \frac{\boldsymbol{e}'(\boldsymbol{y} - \beta_1 \boldsymbol{x}_1)}{\hat{\sigma}},$$

or equivalently

$$\|\boldsymbol{r}\|(\hat{\beta}_1 - \beta_1) = \boldsymbol{e}'(\boldsymbol{y} - \beta_1 \boldsymbol{x}_1).$$

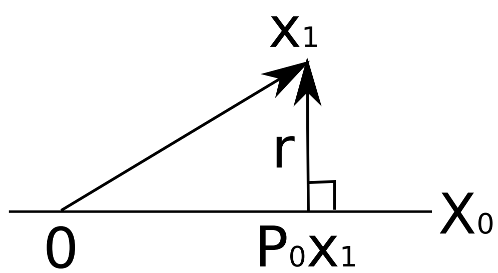
Indeed, we have

$$\boldsymbol{e}'\boldsymbol{x}_1 = \frac{1}{\|\boldsymbol{r}\|}\boldsymbol{r}'\boldsymbol{x}_1 = \frac{1}{\|\boldsymbol{r}\|}\boldsymbol{r}'(\boldsymbol{r} + \boldsymbol{P}_0\boldsymbol{x}_1) = \|\boldsymbol{r}\|$$

and

$$\boldsymbol{e}'\boldsymbol{y} = \boldsymbol{e}'\boldsymbol{P}\boldsymbol{y} = \boldsymbol{e}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{e}'(\boldsymbol{x}_1, \boldsymbol{X}_0)\hat{\boldsymbol{\beta}} = (\|\boldsymbol{r}\|, \mathbf{0}')\hat{\boldsymbol{\beta}} = \|\boldsymbol{r}\|\hat{\beta}_1.$$

This completes the proof.



□

Q10-5 (1-way ANOVA)

The following table shows a data of running times of "mini 4WD" (a toy car). Assume that the time is distributed according to the normal distribution with mean a_i and variance σ^2 if the motor A_i is used. Consider the F test of the null hypothesis $a_1 = a_2 = a_3$.

Motor	Running time [sec]			
"Torque tune" (A_1)	16.66	15.28	14.19	15.95
"Rev tune" (A_2)	15.60	15.99	15.73	15.55
"Plasma dash" (A_3)	13.53	13.47	14.96	13.47

Explain the following result showing an output of R execution.

```
> y1 = c(16.66, 15.28, 14.19, 15.95)
> y2 = c(15.60, 15.99, 15.73, 15.55)
> y3 = c(13.53, 13.47, 14.96, 13.47)
> y = c(y1, y2, y3)
> A = factor(c(1,1,1,1,2,2,2,2,3,3,3,3))
> lm.1 = lm(y ~ A)
> anova(lm.1)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
A          2 8.3500  4.1750  7.4401 0.01239 *
Residuals  9 5.0504  0.5612
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution. Let y_{it} ($1 \le i \le 3$, $1 \le t \le 4$) be the observed data. The statistical model is

$$Y_{it} = a_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathrm{N}(0, \sigma^2).$$

The F-test statistic for the null hypothesis $a_1 = a_2 = a_3$ is

$$F = \frac{\sum_{i=1}^{3}\sum_{t=1}^{4}(\bar{y}_i - \bar{y})^2/(3-1)}{\sum_{i=1}^{3}\sum_{t=1}^{4}(y_{it} - \bar{y}_i)^2/(12-3)} = \frac{8.35/2}{5.05/9} = \frac{4.17}{0.56} = 7.44.$$

In summary, we obtain the following analysis-of-variance (ANOVA) table:

	sum of squares	degree of freedom	variance	F-value	p-value
motor	8.35	2	4.17	7.44	0.012
residuals	5.05	9	0.56	—	—
total	13.40	11	1.22	—	—

The p-value is smaller than 0.05 (i.e., significant at the level 0.05), and therefore we will reject the null hypothesis $a_1 = a_2 = a_3$. In fact, the motor A_3 seems to have better performance than the others since $\bar{y}_1 = 15.52$, $\bar{y}_2 = 15.72$ and $\bar{y}_3 = 13.86$. □

11 Generalized linear model

Q11-1

The logistic regression model is defined by

$$Y_t \sim \text{Bernoulli}(\mu_t), \quad \log\left(\frac{\mu_t}{1-\mu_t}\right) = \boldsymbol{\beta}'\boldsymbol{x}_{(t)}, \quad t = 1, \ldots, n,$$

where Y_t is a $\{0,1\}$-valued response variable, $\boldsymbol{x}_{(t)}$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of parameters. Find the likelihood function.
Note: The predictor is often expressed as $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_{(t)}$. In this document, we set $\beta_0 = 0$ for simplicity and assume that, for example, the first component of $\boldsymbol{x}_{(t)}$ is one.

Solution. The probability function of Y_t is

$$f(y_t) = \mu_t^{y_t}(1-\mu_t)^{1-y_t} = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)} y_t}}{1+e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)}}}.$$

The likelihood function of $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}) = \prod_{t=1}^{n} f(y_t) = \prod_{t=1}^{n} \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)} y_t}}{1+e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)}}}.$$

□

Q11-2

The Poisson regression model is defined by

$$Y_t \sim \text{Poisson}(\mu_t), \quad \log \mu_t = \boldsymbol{\beta}'\boldsymbol{x}_{(t)}, \quad t = 1, \ldots, n,$$

where Y_t is a non-negative integer-valued response variable, $\boldsymbol{x}_{(t)}$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of parameters. Find the likelihood function.

Solution. The probability function of Y_t is

$$f(y_t) = \frac{\mu_t^{y_t}}{y_t!} e^{-\mu_t} = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)} y_t}}{y_t!} e^{-e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)}}}.$$

The likelihood function of $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}) = \prod_{t=1}^{n} f(y_t) = \prod_{t=1}^{n} \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)} y_t}}{y_t!} e^{-e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)}}}.$$

□

Q11-3

A (canonical) generalized linear model is defined by

$$f(y_t) = a(y_t, \phi) \exp\left(\frac{\theta_t y_t - \psi(\theta_t)}{\phi}\right), \tag{3}$$

$$\theta_t = \boldsymbol{\beta}' \boldsymbol{x}_{(t)}, \tag{4}$$

for a response variable Y_t and a vector of explanatory variables $\boldsymbol{x}_{(t)}$, where $a(y, \phi)$ and $\psi(\theta)$ are functions, $\boldsymbol{\beta}$ and $\phi > 0$ are parameters. Eq. (3) is an exponential family if ϕ is fixed. Eq. (4) is called a predictor.
Remark: ϕ is a constant or a parameter. It is known that the functions $a(y, \phi)$ and $\psi(\theta)$ are determined by $a(y, 1)$.

(i) Show that the mean and variance of Y_t is $\mu_t := \mathrm{E}[Y_t] = \psi'(\theta_t)$ and $\mathrm{Var}[Y_t] = \phi\psi''(\theta_t)$. The function $\theta_t = (\psi')^{-1}(\mu_t)$ is called a link function. Eq. (4) is written as $(\psi')^{-1}(\mu_t) = \boldsymbol{\beta}' \boldsymbol{x}_{(t)}$.

Remark: Instead of Eq. (4), a model defined by $g(\mu_t) = \boldsymbol{\beta}' \boldsymbol{x}_{(t)}$ for a function g is called a (not canonical) generalized linear model. The function g is also called a link function.

(ii) Show that the normal linear model, logistic regression model and Poisson regression model are generalized linear models, where $\phi = \sigma^2$, $\phi = 1$ and $\phi = 1$, respectively.

Solution. (i) This problem is similar to Problem 8-5 (iii). We abbreviate the index t, e.g., $\theta = \theta_t$. Differentiate the identity $\psi(\theta) = \phi \log\left(\int a(y, \phi) e^{\theta y/\phi} \mathrm{d}y\right)$ by θ to obtain

$$\psi'(\theta) = \phi \frac{\partial}{\partial \theta} \log\left(\int a(y, \phi) e^{\theta y/\phi} \mathrm{d}y\right) = \frac{\int y a(y, \phi) e^{\theta y/\phi} \mathrm{d}y}{\int a(y, \phi) e^{\theta y/\phi} \mathrm{d}y} = \mathrm{E}[Y] = \mu.$$

In a similar way, we have

$$\begin{aligned}\psi''(\theta) &= \frac{\partial}{\partial \theta}\left(\frac{\int y a(y, \phi) e^{\theta y/\phi} \mathrm{d}y}{\int a(y, \phi) e^{\theta y/\phi} \mathrm{d}y}\right) = \frac{\int (y^2/\phi) a(y, \phi) e^{\theta y/\phi} \mathrm{d}y}{\int a(y, \phi) e^{\theta y/\phi} \mathrm{d}y} - \frac{1}{\phi}\left(\frac{\int y a(y, \phi) e^{\theta y/\phi} \mathrm{d}y}{\int a(y, \phi) e^{\theta y/\phi} \mathrm{d}y}\right)^2 \\ &= \phi^{-1}(\mathrm{E}[Y^2] - \mathrm{E}[Y]^2) = \phi^{-1} \mathrm{Var}[Y].\end{aligned}$$

Thus we obtain $\mathrm{E}[Y] = \psi'(\theta)$ and $\mathrm{Var}[Y] = \phi\psi''(\theta)$.
(ii) The answer is summarized in the following table.

Model	$f(y)$	ϕ	$a(y, \phi)$	$\psi(\theta)$	$(\psi')^{-1}(\mu)$
Normal linear	$(2\pi\phi)^{-1/2} e^{-(y-\theta)^2/2\phi}$	σ^2	$(2\pi\phi)^{-1/2} e^{-y^2/2\phi}$	$\theta^2/2$	μ
Logistic	$e^{\theta y}/(e^\theta + 1)$	1	1	$\log(e^\theta + 1)$	$\log(\mu/(1-\mu))$
Poisson	$(e^{\theta y}/y!) e^{-e^\theta}$	1	$1/y!$	e^θ	$\log \mu$

□

Q11-4

Assume that Y_t follows the Poisson regression model. Find the statistical model of conditional distributions of Y_t given the total number $\sum_{t=1}^n Y_t = \nu$.

Solution. Let Y_t be a Poisson random variable with the mean parameter $\mu_t = e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)}}$. The conditional distribution given $\sum_{t=1}^n Y_t = \nu$ is the multinomial distribution with parameters $\pi_t = \mu_t/(\mu_1 + \cdots + \mu_n)$. Indeed, we have

$$\begin{aligned}
f(y_1,\ldots,y_n;\boldsymbol{\beta}) &:= \mathrm{P}(Y_1 = y_1,\cdots,Y_n = y_n | Y_1 + \cdots + Y_n = \nu) \\
&= \frac{\mathrm{P}(Y_1 = y_1,\ldots,Y_n = y_n)}{\mathrm{P}(Y_1 + \cdots + Y_n = \nu)} \\
&\propto \frac{1}{y_1!\cdots y_n!}\mu_1^{y_1}\cdots\mu_n^{y_n},
\end{aligned}$$

where the symbol $\propto$ means equality up to a multiplicative constant, and therefore

$$f(y_1,\ldots,y_n;\boldsymbol{\beta}) = \frac{\nu!}{y_1!\cdots y_n!}\pi_1^{y_1}\cdots\pi_n^{y_n}.$$

This model is called *the multinomial logistic regression model* with the link function $\log(\pi_t/\pi_n) = \log(\mu_t/\mu_n) = \boldsymbol{\beta}'(\boldsymbol{x}_{(t)} - \boldsymbol{x}_{(n)})$ for $t = 1,\ldots,n-1$. □

Remark. It is shown that the likelihood function of the conditional model is the same as that of the original Poisson regression model if the intercept term β_0 is included and estimated. Indeed, the Poisson regression model with the intercept term is

$$\tilde{f}(y_1,\ldots,y_n;\beta_0,\boldsymbol{\beta}) := \prod_{t=1}^n \frac{(e^{\beta_0}\mu_t)^{y_t}}{y_t!}e^{-e^{\beta_0}\mu_t}, \quad \mu_t = e^{\boldsymbol{\beta}'\boldsymbol{x}_{(t)}}.$$

The likelihood equation with respect to β_0 is

$$\sum_t \left(y_t - e^{\beta_0}\mu_t\right) = \nu - e^{\beta_0}\sum_t \mu_t = 0$$

and therefore

$$\hat{\beta}_0 = \hat{\beta}_0(\boldsymbol{\beta}) = \log\frac{\nu}{\sum_t \mu_t}$$

By substituting this to the expression of $\tilde{f}$, we have

$$\begin{aligned}
\tilde{f}(y_1,\ldots,y_n;\hat{\beta}_0,\boldsymbol{\beta}) &= \prod_{t=1}^n \frac{(e^{\hat{\beta}_0}\mu_t)^{y_t}}{y_t!}e^{-e^{\hat{\beta}_0}\mu_t} \\
&= \prod_{t=1}^n \frac{(\nu\pi_t)^{y_t}}{y_t!}e^{-\nu\pi_t} \\
&= \frac{\nu^\nu}{\nu!}e^{-\nu}\frac{\nu!}{y_1!\cdots y_n!}\pi_1^{y_1}\cdots\pi_n^{y_n},
\end{aligned}$$

which is equal to the likelihood of the multinomial distribution up to the factor $(\nu^\nu/\nu!)e^{-\nu}$. □

Q11-5

For the "university ranking" example discussed in Q3-6 and Q4-6, obtain a discriminant function determining whether a university is in USA or not, based on the logistic regression model.

Solution. The logistic regression is executed in R as follows:

```
> source("http://www.stat.t.u-tokyo.ac.jp/~sei/lec/THE2017.R")
> X = Xori[1:50, 5:9]
> y = ifelse(Xori[1:50, 3]=="United States", 1, 0)
> glm.THE = glm(y ~ X[,1]+X[,2]+X[,3]+X[,4]+X[,5], family=binomial())
> summary(glm.THE)
```

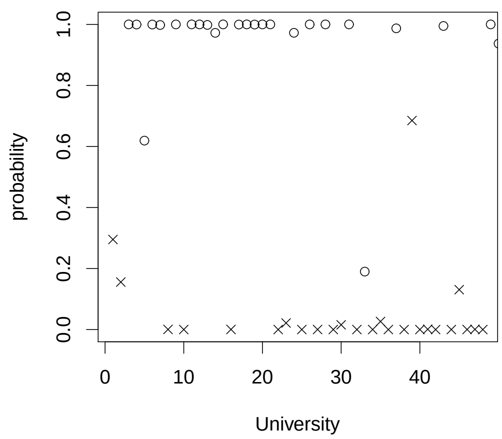
Here the file `THE2017.R` is a data file in the R format (made by the function `dump`). The estimated regression coefficients are

$$\hat{\beta}_0 = -80.208, \quad (\hat{\beta}_1, \ldots, \hat{\beta}_5) = (-0.031, 0.333, 0.998, -0.030, -0.506).$$

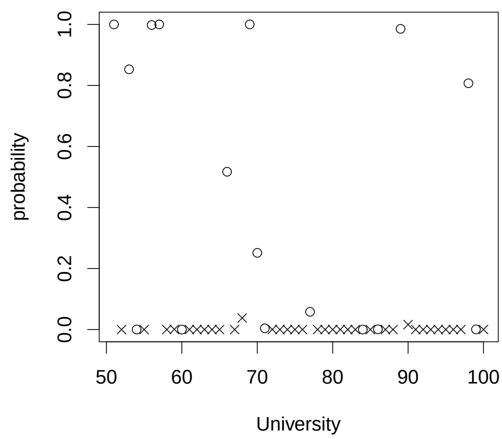
A warning message appears since the fitted value

$$\mathrm{P}(Y_t = 1|\boldsymbol{x}_{(t)}) = \frac{e^{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\boldsymbol{x}_{(t)}}}{1 + e^{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\boldsymbol{x}_{(t)}}}$$

is very close to 0 or 1 for some t. The fitted values are shown in Figure (a) below.



(a) Predicted values for training data.



(b) Predicted values for test data.

Figure (b) shows the predicted value $\mathrm{P}(Y_t = 1|\boldsymbol{x})$ for 51st to 100th universities. The discriminant function is given by

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } P(Y_t = 1|\boldsymbol{x}) \geq 0.5, \\ 0 & \text{otherwise,} \end{cases}$$

$$= \begin{cases} 1 & \text{if } \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\boldsymbol{x} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

□

Q11-6

The data file `FIFA.csv` uploaded in the course web site is a table of records of Japan national football team from 2014 to 2017, based on the web site

FIFA Ranking.net (http://fifaranking.net/nations/jpn/)

Each variate has the following meaning:

date	opponent	goal1	goal2	stadium
Date	Opponent	Goals of Japan	Goals of opponent	Home/Away

rank1	rank2
Rank of Japan	Rank of opponent

Suppose that we are constructing a model predicting the number of goals. Run the following R program and explain the meaning of the output.

```
> X = read.csv("FIFA.csv")
> glm.1 = glm(goal1 ~ stadium + rank1 + rank2, family=poisson, data=X)
> summary(glm.1)
```

Proof. Here is the output:

```
Call:
glm(formula = goal1 ~ stadium + rank1 + rank2, family = poisson,
    data = X)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6378  -0.8989  -0.2365   0.5573   2.2182

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.421513   1.206251  -2.007   0.0447 *
stadiumHome  0.420067   0.218898   1.919   0.0550 .
rank1        0.051114   0.024695   2.070   0.0385 *
rank2        0.003833   0.002246   1.707   0.0879 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 71.696  on 49  degrees of freedom
Residual deviance: 55.448  on 46  degrees of freedom
AIC: 180.69

Number of Fisher Scoring iterations: 5
```

In the following, we denote the explanatory and response variables by $\boldsymbol{x}_{(t)}$ and y_t. The Poisson regression model is defined by

$$f(y_t; \mu_t) = \frac{\mu_t^{y_t}}{y_t!} e^{-\mu_t}, \quad \mu_t = e^{\beta_0 + \boldsymbol{\beta}' \boldsymbol{x}_{(t)}},$$

and the null model (hypothesis) is $\mu_t =$ (constant).

In the output, "Coefficients" shows the maximum likelihood estimate $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ and its standard error. The "z value" is the ratio of the estimate to the standard error. For example, the z value of the intercept is

$$\frac{-2.421513}{1.206251} = -2.007.$$

Its p-value is $P(|Z| \geq 2.007) = 0.0447$, where $Z \sim N(0, 1)$.

In this data, the variable `stadium` is a factor object and automatically encoded as 1 if `stadium == Home` and 0 if `stadium == Away`.

The "Deviance Residuals" are defined by

$$\begin{aligned} r_t &:= \operatorname{sign}(y_t - \hat{\mu}_t)\sqrt{2\log\frac{f(y_t; y_t)}{f(y_t; \hat{\mu}_t)}} \\ &= \operatorname{sign}(y_t - \hat{\mu}_t)\sqrt{2\left(y_t \log\frac{y_t}{\hat{\mu}_t} - y_t + \hat{\mu}_t\right)}, \end{aligned}$$

for $t = 1, \ldots, n$. The "Residual Deviance" is equal to the log-likelihood ratio test statistic for the full model:

$$\begin{aligned} D &:= 2\log\frac{\prod_{t=1}^n f(y_t; y_t)}{\prod_{t=1}^n f(y_t; \hat{\mu}_t)} \\ &= 2\sum_{t=1}^n \left(y_t \log\frac{y_t}{\hat{\mu}_t} - y_t + \hat{\mu}_t\right) \\ &= \sum_{t=1}^n r_t^2. \end{aligned}$$

The "Null Deviance" is equal to the log-likelihood ratio test statistic for the null model:

$$\begin{aligned} D_0 &:= 2\log\frac{\prod_{t=1}^n f(y_t; y_t)}{\prod_{t=1}^n f(y_t; \bar{y})} \\ &= 2\sum_{t=1}^n y_t \log\frac{y_t}{\bar{y}}, \end{aligned}$$

where $\bar{y} = n^{-1}\sum_{t=1}^n y_t$. The degree of freedom of the full model is $n - 4 = 46$, and the degree of freedom of the null model is $n - 1 = 49$.

AIC (Akaike's Information Criterion) is defined by

$$-2\sum_{t=1}^n \log f(y_t; \hat{\mu}_t) + 2p,$$

where p is the number of the parameters. In this data, $p = 4$.

Finally, "Number of Fisher Scoring iterations" is the number of steps for the optimization procedure used in the package `glm`. The Fisher scoring algorithm is a version of the Newton-Raphson method. □

12 Information criterion

Q12-1

The following table shows the monthly average values of the daily maximum temperature in Tokyo for two years 2015 and 2016.

Month	1	2	3	4	5	6	7	8	9	10	11	12
2015	10.4	10.4	15.5	19.3	26.4	26.4	30.1	30.5	26.4	22.7	17.8	13.4
2016	10.6	12.2	14.9	20.3	25.2	26.3	29.7	31.6	27.7	22.6	15.5	13.8

We fit a normal linear model

$$y_t = a_0 + \sum_{j=1}^{k} \{a_j \cos(2\pi jt/12) + b_j \sin(2\pi jt/12)\} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma^2),$$

to the 2015 data $\{y_t\}_{t=1}^{12}$, and use it to predict the 2016 data $\{\tilde{y}_t\}_{t=1}^{12}$. Find k that minimizes the squared prediction error, where $0 \le k \le 5$. Also determine k that minimizes AIC.

Solution. Let $\hat{y}_t^{(k)}$ be the fitted values (predicted values) of y_t for each model $k = 0, 1, \ldots, 5$. Define the squared prediction error by

$$\mathrm{error}(k) = \frac{1}{n} \sum_{t=1}^{n} (\tilde{y}_t - \hat{y}_t^{(k)})^2,$$

where $n = 12$. The AIC of the model k is given by

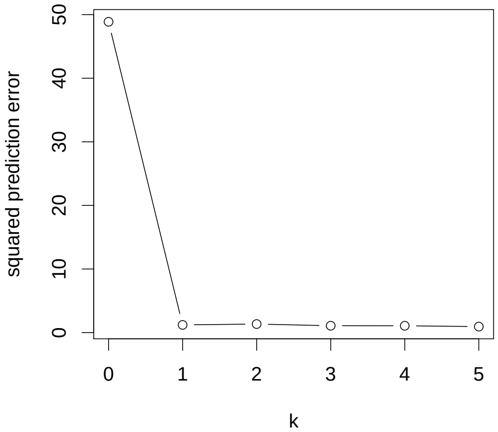
$$\mathrm{AIC}(k) = n \log \hat{\sigma}_k^2 + 2(2k + 2),$$

where $\hat{\sigma}_k^2 = n^{-1} \sum_{t=1}^{n} (y_t - \hat{y}_t^{(k)})^2$ is the MLE of the variance parameter σ^2. By numerical computation, we obtain the following table of the prediction error and AIC.

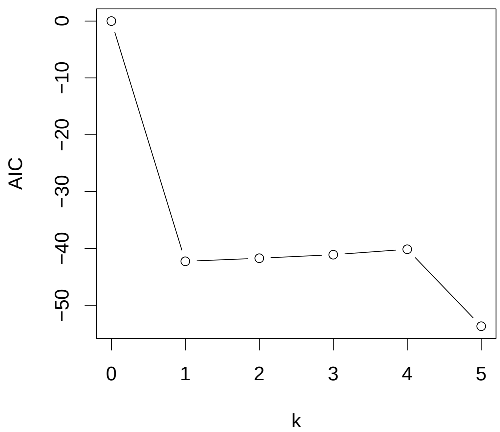
k	0	1	2	3	4	5
prediction error	48.88	1.21	1.34	1.08	1.07	0.94
AIC	50.72	8.45	8.99	9.61	10.57	−2.97

The number k which minimizes the prediction error is 5, and k which minimizes AIC is also 5. However, in both quantities, there is a large gap between the two models $k = 0$ and $k = 1$. In practice, the number of parameters in minimizing AIC is recommended to be at most $n/2$, where n is the sample size. Then we may select the model $k = 1$.

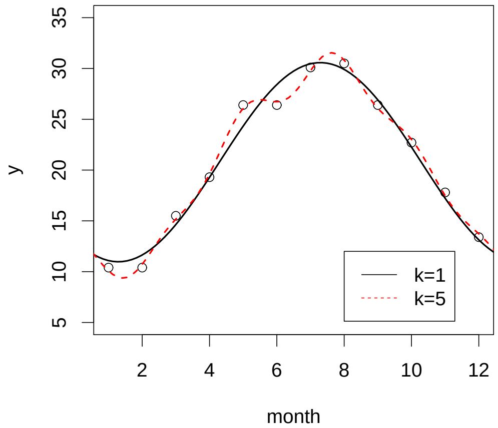
The following figure shows (a) the squared prediction errors, (b) the AIC values (up to an additive constant), (c) the fitted curves for $k = 1$ and $k = 5$ with the observed values, and (d) the same curves with the future values.



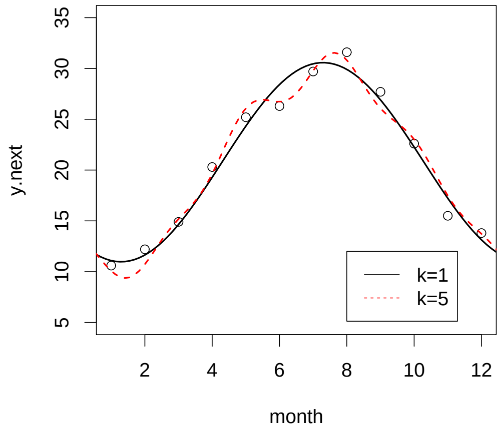
(a) Squared prediction errors.



(b) AIC values.



(c) Fitted curves and $\{y_t\}$.



(d) Fitted curves and $\{\tilde{y}_t\}$.

□

Q12-2

The data `nations.csv` in the course web site consists of GDP, GDP per capita, population density, and average lifespan of 68 countries. They were obtained from Tables 3-2, 3-3, 2-5, and 2-17 in the following site (in Japanese):

`http://www.stat.go.jp/data/sekai/0116.htm`

Let us make a model that determines the country is in Asia or not. Fit the logistic regression model. Furthermore, select a model by the backward selection method.

Solution. Use the following R code to perform the backward selection method.

```
> Xori = read.csv("nations.csv")
> X = Xori[,3:6]
> y = ifelse(Xori[,1] == "asia", 1, 0)
```

```
> glm.nations = glm(y ~ X[,1]+X[,2]+X[,3]+X[,4], family=binomial())
> step(glm.nations, y ~ X[,1]+X[,2]+X[,3]+X[,4])
```

The selected model is

$$\log \frac{\mu}{1-\mu} = -1.424 - 6.751\times 10^{-5}\times(\text{GDP per capita}) + 1.22\times 10^{-2}\times(\text{population density}),$$

where μ denotes the probability that the country is in Asia. The following table is the result of classification, where True means "the country is actually in Asia" and Positive means "$\mu > 0.5$".

	Positive	Negative
True	Japan, Israel, India, Korea, Singapore, Sri Lanka, Nepal, Pakistan, Bangladesh, Philippines, Vietnum. (11)	Iran, Indonesia, Oman, Kuwait, Saudi Arabia, Thailand, China, Turkey, Malaysia. (9)
False	Netherlands, Nigeria. (2)	USA, Canada, Guatemala, Panama, Mexico, Argentina, Ecuador, Columbia, Chile, Brazil, Venezuela, Peru, Iceland, Ireland, UK, Italy, Ukraine, Austria, Greece, Switzerland, Sweden, Spain, Slovakia, Czech, Denmark, Germany, Norway, Hungary, Finland, France, Belgium, Poland, Portugal, Romania, Luxembourg, Russia, Algeria, Egypt, Ethiopia, Kenya, DR Congo, Tunisia, South Africa, Morocco, Australia, New Zealand. (46)

For reference, we give the AIC values (up to an additive constant) of all submodels:

model	1234		123	124	134	234
AIC	61.71		61.07	83.37	71.12	60.60

model	12	13	14	23	24	34
AIC	82.66	71.56	87.66	60.10	81.91	69.39

model	1	2	3	4		$\emptyset$
AIC	86.33	81.19	69.63	85.79		84.39

□

Q12-3

Let random vectors $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}}$ be independent and follow $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{I}_n)$. Here $\boldsymbol{\mu} \in \mathbb{R}^n$ is an unknown parameter and $\boldsymbol{I}_n$ is the identity matrix of order n. Let M be a p-dimensional linear subspace of $\mathbb{R}^n$ and P be the orthogonal projection matrix from $\mathbb{R}^n$ onto M.

(i) Prove that for any $\boldsymbol{\mu} \in \mathbb{R}^n$, the following identity holds:

$$\mathrm{E}\left[\|\tilde{\boldsymbol{Y}} - \boldsymbol{P}\boldsymbol{Y}\|^2\right] = \|\boldsymbol{\mu} - \boldsymbol{P}\boldsymbol{\mu}\|^2 + n + p.$$

This quantity represents the risk of prediction when the future data $\tilde{\boldsymbol{Y}}$ is predicted by $\boldsymbol{P}\boldsymbol{Y}$.

(ii) Prove that for any $\boldsymbol{\mu} \in \mathbb{R}^n$, the following identity holds:

$$\mathrm{E}\left[\|\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y}\|^2\right] = \|\boldsymbol{\mu} - \boldsymbol{P}\boldsymbol{\mu}\|^2 + n - p.$$

(iii) Show that AIC of the model M is, except a constant term (i.e., a term independent of M,

$$\mathrm{AIC} = \|\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y}\|^2 + 2p.$$

In addition, show that AIC is an unbiased estimator of the risk.

Note that the above result is correct even if $\boldsymbol{\mu}$ does not belong to M. This is a point completely different from the hypothesis testing.

Solution. We first show that $\mathrm{E}[\|\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2] = p$ for any orthogonal projection matrix $\boldsymbol{P}$ onto a p-dimensional subspace. Indeed,

$$\begin{aligned}
\mathrm{E}\left[\|\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2\right] &= \mathrm{E}\left[(\boldsymbol{Y} - \boldsymbol{\mu})'\boldsymbol{P}'\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\mu})\right] \\
&= \mathrm{E}\left[\mathrm{tr}\left(\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})'\boldsymbol{P}'\right)\right] \quad (\because \mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})) \\
&= \mathrm{tr}\left(\boldsymbol{P}\,\mathrm{E}[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})']\boldsymbol{P}'\right) \\
&= \mathrm{tr}(\boldsymbol{P}\boldsymbol{P}') \quad (\because \boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)) \\
&= \mathrm{tr}(\boldsymbol{P}^2) = \mathrm{tr}(\boldsymbol{P}) = p.
\end{aligned}$$

(i) Since $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}}$ are i.i.d., we have

$$\begin{aligned}
\mathrm{E}\left[\|\tilde{\boldsymbol{Y}} - \boldsymbol{P}\boldsymbol{Y}\|^2\right] &= \mathrm{E}\left[\|(\tilde{\boldsymbol{Y}} - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{P}\boldsymbol{\mu}) + (\boldsymbol{P}\boldsymbol{\mu} - \boldsymbol{P}\boldsymbol{Y})\|^2\right] \\
&= \mathrm{E}\left[\|\tilde{\boldsymbol{Y}} - \boldsymbol{\mu}\|^2\right] + \|\boldsymbol{\mu} - \boldsymbol{P}\boldsymbol{\mu}\|^2 + \mathrm{E}\left[\|\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\mu})\|^2\right] \\
&= n + \|\boldsymbol{\mu} - \boldsymbol{P}\boldsymbol{\mu}\|^2 + p.
\end{aligned}$$

(ii) In a similar way, we obatin

$$\begin{aligned}
\mathrm{E}\left[\|\boldsymbol{Y}-\boldsymbol{P}\boldsymbol{Y}\|^2\right] &= \mathrm{E}\left[\|(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{Y}\|^2\right] \\
&= \mathrm{E}\left[\|(\boldsymbol{I}_n-\boldsymbol{P})(\boldsymbol{Y}-\boldsymbol{\mu}+\boldsymbol{\mu})\|^2\right] \\
&= \mathrm{E}\left[\|(\boldsymbol{I}_n-\boldsymbol{P})(\boldsymbol{Y}-\boldsymbol{\mu})\|^2\right] + \|(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{\mu}\|^2 \\
&= n-p+\|\boldsymbol{\mu}-\boldsymbol{P}\boldsymbol{\mu}\|^2.
\end{aligned}$$

(iii) The log-likelihood function is

$$\log L(\boldsymbol{\mu}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{\mu}\|^2.$$

The MLE of $\boldsymbol{\mu}$ in the subspace M is $\hat{\boldsymbol{\mu}} = \boldsymbol{P}\boldsymbol{Y}$. Therefore AIC of the model M is

$$\begin{aligned}
\mathrm{AIC} &= -2\log L(\hat{\boldsymbol{\mu}}) + 2p \\
&= n\log(2\pi) + \|\boldsymbol{Y}-\boldsymbol{P}\boldsymbol{Y}\|^2 + 2p,
\end{aligned}$$

which is the same as $\|\boldsymbol{Y}-\boldsymbol{P}\boldsymbol{Y}\|^2 + 2p$ except for a constant term $n\log(2\pi)$. Finally, we obtain

$$\begin{aligned}
\mathrm{E}[\|\boldsymbol{Y}-\boldsymbol{P}\boldsymbol{Y}\|^2 + 2p] &= \big(\|\boldsymbol{\mu}-\boldsymbol{P}\boldsymbol{\mu}\|^2 + n - p\big) + 2p \\
&= \mathrm{E}[\|\tilde{\boldsymbol{Y}}-\boldsymbol{P}\boldsymbol{Y}\|^2]
\end{aligned}$$

from the result of (ii) and (i). $\square$