STAT 3690 Lecture Note

Part II: Basics of statistical modelling

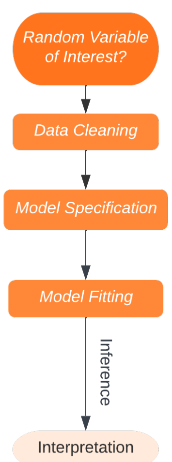
Zhiyang Zhou (zhiyang.zhou@umanitoba.ca, zhiyanggeezhou.github.io)

2023/Feb/26 15:26:58

"All models are wrong, but some are useful."

— G. E. P. Box. (1976). *Journal of the American Statistical Association*, 71:791–799

Statistical modelling



What is a statistical model?

- The (joint) distribution of the random variable(s) of interest
  - E.g., reformulate linear regression and logit regression models in terms of distributions

---

Recall the characterization of univariate distributions

- A random variable (RV), say X, is a real-valued function (defined on a sample space).
- The cumulative distribution function (cdf) of X, say $F_X(x) = \Pr(X \le x)$, $x \in \mathbb{R}$, if (right continuous) $\lim_{t\to x_0^+} F_X(x) = F_X(x_0)$, (non-decreasing) $F_X(x_0) \le F_X(x_1)$ for $x_0 < x_1$, and (ranging from 0 to 1) $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
  - Reversely, any function satisfying the three properties must be a cdf for certain RV.

- Discrete RV
  - RV X takes countable different values
  - Probability mass function (pmf): $p_X(x) = \Pr(X = x)$
- Continuous RV
  - RV X is continuous iff its cdf F_X is (absolutely) continuous, i.e., there exists f_X, s.t.

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u, \quad \forall x \in \mathbb{R}.$$

  - Probability density function (pdf): $f_X(x) = F'_X(x)$.
- Moment-generating function (mgf) $M_X(t) = \mathrm{E}\{\exp(tX)\}$ if $\mathrm{E}\{\exp(tX)\} < \infty$ for t in a neighbourhood of 0
  - If the mgf exists, then $\mathrm{E}(X^k) = M_X^{(k)}(t) \mid_{t=0}$.

Support of RV

- Support of X, say $\mathrm{supp}(X)$, is $\{x \in \mathbb{R} : p_X(x) \text{ (or } f_X(x)) > 0\}$
  - e.g., support of $Binom(n, p)$ is $\{0, \ldots, n\}$; support of $\mathcal{N}(0, 1)$ is $\mathbb{R}$.

Indicator function

- Given a set A, the indicator function of A is

$$\mathbf{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

  - Hence, e.g., if $X \sim Binom(n, p)$, then $p_X(x) = \binom{x}{n} p^x (1-p)^{1-x}$, $x \in \{0, \ldots, n\}$, $p \in (0, 1)$, or equivalently, $p_X(x) = \binom{x}{n} p^x (1-p)^{1-x} \mathbf{1}_{\{0,\ldots,n\}}(x) \mathbf{1}_{(0,1)}(p)$

Characterization of joint/multivariate distributions

- Random (column) vector/vector-valued RV
  - $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$
- Joint cdf: $F_{\boldsymbol{X}}(x_1, \ldots, x_p) = \Pr(X_1 \le x_1, \ldots, X_p \le x_p)$
- Joint distribution of continuous RVs
  - Joint pdf:

$$f_{\boldsymbol{X}}(x_1, \ldots, x_p) = \frac{\partial^p}{\partial x_1 \cdots \partial x_p} F_{\boldsymbol{X}}(x_1, \ldots, x_p)$$

  - E.g., multivariate normal (MVN) distribution
- Joint distribution of discrete RVs
  - Joint pmf:

$$p_{\boldsymbol{X}}(x_1, \ldots, x_p) = \Pr(X_1 = x_1, \ldots, X_p = x_p)$$

  - E.g., categorical distribution & multinomial distribution

---

- Exercise 2.1: Suppose that we independently observe an experiment that has m possible outcomes $O_1, \ldots, O_m$ for n times; e.g., sample n balls with replacement from a pool of balls of m colors. Let $p_1, \ldots, p_m$ denote probabilities of $O_1, \ldots, O_m$ in each experiment respectively. Let X_i denote the number of times that outcome O_i occurs in the n repetitions.
  - What is the distribution of X_i?
  - What is the joint pmf of $\boldsymbol{X} = [X_1, \ldots, X_m]^\top$?

```
xs = c(10, 4, 7, 9)
ps = c(.3, .4, .2, .1)
dmultinom(x = xs, prob = ps)

# verify that binomial is a special case of multinomial
xs = c(4, 6)
ps = c(.6, .4)
dmultinom(x = xs, prob = ps)
dbinom(x = xs[1], size = sum(xs), prob = ps[1])
```

- Moment-generating function (mgf) $M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}\{\exp(\boldsymbol{t}^\top \boldsymbol{X})\}$ if there exists $\delta > 0$ s.t. $\mathrm{E}\{\exp(\boldsymbol{t}^\top \boldsymbol{X})\} < \infty$ for all $\boldsymbol{t} \in \{\boldsymbol{t} : \boldsymbol{t}^\top \boldsymbol{t} < \delta\}$
  - If the mgf of $\boldsymbol{X}$ exists and X_i are independent of each other, then $M_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{i=1}^p M_{X_i}(t_i)$.

Marginalization

- $\boldsymbol{X} = [X_1, \ldots, X_m]^\top$,
- $\boldsymbol{Y} = [X_1, \ldots, X_q]^\top$, $p > q$, as part of $\boldsymbol{X}$
- Marginal cdf of $\boldsymbol{Y}$

$$F_{\boldsymbol{Y}}(x_1, \ldots, x_q) = \lim_{x_{q+1}, \ldots, x_m \to \infty} F_{\boldsymbol{X}}(x_1, \ldots, x_m)$$

- Marginal pdf of $\boldsymbol{Y}$ (when $X_1, \ldots, X_m$ are all continous)

$$f_{\boldsymbol{Y}}(x_1, \ldots, x_q) = \int_{\mathbb{R}^{m-q}} f_{\boldsymbol{X}}(x_1, \ldots, x_m) \mathrm{d}x_{q+1} \cdots x_m$$

- Marginal pmf of $\boldsymbol{Y}$ (when $X_1, \ldots, X_m$ are all discrete)

$$p_{\boldsymbol{Y}}(x_1, \ldots, x_q) = \sum_{x_{q+1}, \ldots, x_m} p_{\boldsymbol{X}}(x_1, \ldots, x_m)$$

Conditioning

- $\boldsymbol{X} = [X_1, \ldots, X_m]^\top$ and $\boldsymbol{Y} = [Y_1, \ldots, Y_q]^\top$
- Conditional pdf of $\boldsymbol{Y}$ given $\boldsymbol{X}$

$$f_{\boldsymbol{Y}|\boldsymbol{X}}(y_1, \ldots, y_q \mid x_1, \ldots, x_m) = \frac{f_{\boldsymbol{X},\boldsymbol{Y}}(x_1, \ldots, x_m, y_1, \ldots, y_q)}{f_{\boldsymbol{X}}(x_1, \ldots, x_m)}$$

- Conditional pmf of $\boldsymbol{Y}$ given $\boldsymbol{X}$

$$p_{\boldsymbol{Y}|\boldsymbol{X}}(y_1,\ldots,y_q \mid x_1,\ldots,x_m) = \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(x_1,\ldots,x_m,y_1,\ldots,y_q)}{p_{\boldsymbol{X}}(x_1,\ldots,x_m)}$$

Transformation of random vectors

- Derive the pdf of continuous $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ from the pdf of continuous $\boldsymbol{X}$
- Prerequisite
  - $\boldsymbol{X} = [X_1,\ldots,X_p]^\top$ and $\boldsymbol{Y} = [Y_1,\ldots,Y_p]^\top$
  - $\boldsymbol{g} = (g_1,\ldots,g_p)$: $\mathbb{R}^p \to \mathbb{R}^p$ is a continuous one-to-one map with inverse $\boldsymbol{g}^{-1} = (h_1,\ldots,h_p)$, i.e., $Y_i = g_i(\boldsymbol{X})$ and $X_i = h_i(\boldsymbol{Y})$
- Elaborate $\mathrm{supp}(\boldsymbol{Y}) = \{[y_1,\ldots,y_p]^\top : [h_1(y_1,\ldots,y_p),\ldots,h_p(y_1,\ldots,y_p)]^\top \in \mathrm{supp}(\boldsymbol{X})\}$
- Jacobian matrix of $\boldsymbol{g}^{-1}$ is $\mathbf{J}_{\boldsymbol{g}^{-1}} = [\partial x_i/\partial y_j]_{p\times p} = [\partial h_i(y_1,\ldots,y_p)/\partial y_j]_{p\times p}$
  - Also, $|\det(\mathbf{J}_{\boldsymbol{g}^{-1}})| = |\det([\partial y_i/\partial x_j]_{p\times p})|^{-1} = |\det([\partial g_i(x_1,\ldots,x_p)/\partial x_j]_{p\times p})|^{-1}$
- Then

$$f_{\boldsymbol{Y}}(y_1,\ldots,y_p) = f_{\boldsymbol{X}}(h_1(y_1,\ldots,y_p),\ldots,h_p(y_1,\ldots,y_p))|\det(\mathbf{J}_{\boldsymbol{g}^{-1}})|\mathbf{1}_{\mathrm{supp}(\boldsymbol{Y})}(y_1,\ldots,y_p)$$

- Exercise 2.2: Let $\boldsymbol{X} = [X_1, X_2]^\top$ follow the standard bivariate normal, i.e., its pdf is

$$f_{\boldsymbol{X}}(x_1,x_2) = (2\pi)^{-1}\exp\{-(x_1^2+x_2^2)/2\}\mathbf{1}_{\mathbb{R}^2}(x_1,x_2).$$

  Find out the joint pdf of $\boldsymbol{Y} = [Y_1, Y_2]^\top$, where $Y_1 = \sqrt{X_1^2+X_2^2}$ and $0 \le Y_2 < 2\pi$ is the angle from the positive x-axis to the ray from the origin to the point (X_1, X_2), that is, Y is X in the polar coordinate.

- Exercise 2.3: Given positive α, β and θ, $\boldsymbol{X} = [X_1, X_2]^\top$ follow

$$f_{\boldsymbol{X}}(x_1,x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)\theta^{\alpha+\beta}} x_1^{\alpha-1}x_2^{\beta-1}\exp\left(-\frac{x_1+x_2}{\theta}\right)\mathbf{1}_{\mathbb{R}^+\times\mathbb{R}^+}(x_1,x_2),$$

  where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$, e.g., $\Gamma(n) = (n-1)!$ for positive integer n. Find out the joint pdf of $\boldsymbol{Y} = [Y_1, Y_2]^\top$, where $Y_1 = X_1/(X_1+X_2)$ and $Y_2 = X_1 + X_2$.

Mean matrix

- $\mathrm{E}(\boldsymbol{X}) = [\mathrm{E}(X_{ij})]_{n\times p}$, where
  - Random $n \times p$ matrix $\boldsymbol{X} = [X_{ij}]_{n\times p}$
- (Linearity) $\mathrm{E}(\mathbf{A}\boldsymbol{X} + \mathbf{B}\boldsymbol{Y}) = \mathbf{A}\mathrm{E}(\boldsymbol{X}) + \mathbf{B}\mathrm{E}(\boldsymbol{Y})$, where
  - Fixed $\mathbf{A} \in \mathbb{R}^{\ell\times n}$ and $\mathbf{B} \in \mathbb{R}^{\ell\times m}$
  - Random matrices $\boldsymbol{X} = [X_{ij}]_{n\times p}$ and $\boldsymbol{Y} = [Y_{ij}]_{m\times p}$

Covariance matrix

- Random p-vector $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$ and random q-vector $\boldsymbol{Y} = [Y_1, \ldots, Y_q]^\top$
- Covariance matrix (defined via expectation) $\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \text{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \text{E}[\{\boldsymbol{X} - \text{E}(\boldsymbol{X})\}\{\boldsymbol{Y} - \text{E}(\boldsymbol{Y})\}^\top]$
  - Also, $\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \text{E}(\boldsymbol{XY}^\top) - \text{E}(\boldsymbol{X})\text{E}(\boldsymbol{Y}^\top)$
  - The (i, j)-entry of $\boldsymbol{\Sigma}_{\boldsymbol{XY}}$ is $\text{cov}(X_i, Y_j)$
- $\boldsymbol{\Sigma}_{\mathbf{A}\boldsymbol{X}+\boldsymbol{a}, \mathbf{B}\boldsymbol{Y}+\boldsymbol{b}} = \mathbf{A}\Sigma_{\boldsymbol{XY}}\mathbf{B}^\top$ for fixed $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\boldsymbol{a} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{\ell\times q}$ and $\boldsymbol{b} \in \mathbb{R}^\ell$
- $\boldsymbol{\Sigma}_{\boldsymbol{X}} \geq 0$, where $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \text{cov}(\boldsymbol{X})$ is short for $\boldsymbol{\Sigma}_{\boldsymbol{XX}} = \text{cov}(\boldsymbol{X}, \boldsymbol{X})$

Sample covariance matrix

- Samples $\boldsymbol{X}_k = [X_{k1}, \ldots, X_{kp}]^\top$ and $\boldsymbol{Y}_k = [Y_{k1}, \ldots, Y_{kq}]^\top$, $k = 1, \ldots, n$
- $(\boldsymbol{X}_k, \boldsymbol{Y}_k) \overset{\text{iid}}{\sim} (\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$ and $\boldsymbol{Y} = [Y_1, \ldots, Y_q]^\top$
- Sample mean vectors
  - $\bar{\boldsymbol{X}} = n^{-1}\sum_{k=1}^n \boldsymbol{X}_k = [\bar{X}_{\cdot 1}, \cdots, \bar{X}_{\cdot p}]^\top$
    - $\bar{X}_{\cdot i} = n^{-1}\sum_{k=1}^n X_{ki}$
  - $\bar{\boldsymbol{Y}} = n^{-1}\sum_{k=1}^n \boldsymbol{Y}_k = [\bar{Y}_{\cdot 1}, \cdots, \bar{Y}_{\cdot q}]^\top$
    - $\bar{Y}_{\cdot j} = n^{-1}\sum_{k=1}^n Y_{kj}$
- Sample covariance matrix:

$$\mathbf{S}_{\boldsymbol{XY}} = \frac{1}{n-1}\sum_{k=1}^n \{(\boldsymbol{X}_k - \bar{\boldsymbol{X}})(\boldsymbol{Y}_k - \bar{\boldsymbol{Y}})^\top\}$$

  - The (i, j)-entry of $\mathbf{S}_{\boldsymbol{XY}}$ is $(n-1)^{-1}\sum_{k=1}^n (X_{ki} - \bar{X}_{\cdot i})(Y_{kj} - \bar{Y}_{\cdot j})$, i.e., the sample covariance between X_i (the ith entry of $\boldsymbol{X}$) and Y_j (the jth entry of $\boldsymbol{Y}$)
  - Unbiasedness: $\text{E}(\mathbf{S}_{\boldsymbol{XY}}) = \boldsymbol{\Sigma}_{\boldsymbol{XY}}$
  - $\mathbf{S}_{\mathbf{A}\boldsymbol{X}+\boldsymbol{a}, \mathbf{B}\boldsymbol{Y}+\boldsymbol{b}} = \mathbf{A}\mathbf{S}_{\boldsymbol{XY}}\mathbf{B}^\top$ for $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\boldsymbol{a} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{\ell\times q}$ and $\boldsymbol{b} \in \mathbb{R}^\ell$
  - $\mathbf{S}_{\boldsymbol{X}} \geq 0$
  - Implementation in *R*: cov() (or var() if $\boldsymbol{X} = \boldsymbol{Y}$)

Computing sample mean vectors and sample covariance matrices via *R*

```
options(digits = 4)
set.seed(1)

# mean vector and covariance matrix
(Mu = runif(3))
A = matrix(runif(15), nrow = 3, ncol = 5)
(Sigma = A %*% t(A))

# generation of samples
n = 100
sample = MASS::mvrnorm(n, Mu, Sigma)
colnames(sample) = c('W', 'H', "BP")
head(sample)

# reference for various scatterplots https://www.statmethods.net/graphs/scatterplot.html
```

```r
# scatterplots for paired features
pairs(sample)
# (spinning) 3D scatterplot
rgl::plot3d(sample[,1], sample[,2], sample[,3], col = "red", size = 6)

# sample mean vector for [V1,V2,V3]^T
(MuHat = apply(sample, 2, mean))
(MuHat = colMeans(sample))
# sample covariance matrix for [W,H,BP]^T
## following the definition
S = 0; for (i in 1:n){S = S + 1/(n-1) * (sample[i,]-MuHat) %*% t(sample[i,]-MuHat)}; S
## via var()
(S = var(sample))
## via cov()
(S = cov(sample))

var(sample[,2], sample[,1])

# sample covariance matrix for W & [H,BP]^T
cov(sample[,1], sample[,2:3])
# sample covariance matrix for H & [BP,W]^T
cov(sample[,2], sample[,c(3,1)])

# another sample
(Mu2 = runif(2))
A2 = matrix(runif(10), nrow = 2, ncol = 5)
(Sigma2 = A2 %*% t(A2))
sample2 = MASS::mvrnorm(n, Mu2, Sigma2)
colnames(sample2) = c('CH', 'HR')
head(sample2)
cov(sample, sample2)
sample_c = cbind(sample, sample2)
cov(sample_c)
cov(sample, sample2)
```