RMSC4002 Data Analysis in Finance and Risk Management Science 2018 – 19 Final Examination Suggested Solution

Benjamin Chun Ho Chan*†‡§

December 20, 2018

Question 1 [15%]

Let $x_1, x_2, \ldots, x_{13}$ be zero-coupon rates with 1m, 3m, 6m, 9m, **1y**, 18m, **2y**, **3y**, **4y**, **5y**, 7y, 10y and 15y maturities respectively, where m stands for month and y stands for year. Suppose that Principle Component Analysis (PCA) is applied to the covariance matrix S. The loadings of the first 3 PC's are given by

$$\boldsymbol{h_1} = \begin{bmatrix} 0.2732 \\ 0.2758 \\ 0.2770 \\ 0.2783 \\ \mathbf{0.2786} \\ 0.2798 \\ \mathbf{0.2799} \\ \mathbf{0.2798} \\ \mathbf{0.2791} \\ \mathbf{0.2785} \\ 0.2772 \\ 0.2754 \\ 0.2729 \end{bmatrix}, \quad \boldsymbol{h_2} = \begin{bmatrix} -0.4043 \\ -0.3446 \\ -0.3027 \\ -0.2361 \\ \mathbf{-0.2023} \\ -0.0915 \\ \mathbf{-0.0353} \\ \mathbf{0.0998} \\ \mathbf{0.1682} \\ \mathbf{0.2298} \\ 0.3015 \\ 0.3754 \\ 0.4438 \end{bmatrix}, \quad \boldsymbol{h_3} = \begin{bmatrix} 0.5976 \\ 0.2823 \\ -0.0285 \\ -0.1771 \\ \mathbf{-0.2520} \\ -0.2765 \\ \mathbf{-0.2885} \\ \mathbf{-0.2226} \\ \mathbf{-0.1881} \\ \mathbf{-0.0836} \\ 0.0391 \\ 0.2187 \\ 0.4082 \end{bmatrix}$$

with corresponding eigenvalues being $\lambda_1 = 12.7239$, $\lambda_2 = 0.2377$ and $\lambda_3 = 0.0249$ respectively.

We are going to apply PCA to calculate Value-at-Risk (VaR) of portfolio depending on interest rate. Suppose that we have a portfolio with the exposures to 1 basis point (1 b.p. = 0.01%) increase in interest rate as follows: 1 b.p. increase in 1y, 2y, 3y, 4y and 5y interest rate would cause change (in million \$) of $+4$, $+9$, -6, -8 and $+1$ respectively.

(a) Use the first PC to calculate the 10-day 95% VaR of this portfolio using normal model.

(b) Use the first 2 PC's to calculate the 20-day 98% VaR of this portfolio using normal model.

(c) Use the first 3 PC's to calculate the 30-day 99% VaR of this portfolio using normal model.

*For enquiry, please email to 1155049861@link.cuhk.edu.hk.
†Personal profile: www.linkedin.com/in/benjamin-chan-chun-ho
‡GitHub repository: https://github.com/BenjaminChanChunHo/CUHK-STAT-or-RMSC-Tutorial-Note
§RPubs: http://rpubs.com/Benjamin_Chan_Chun_Ho

Solution: Assume the mean of loss of portfolio is 0. Let $\boldsymbol{y} = \boldsymbol{H}'\boldsymbol{x}$ or $\boldsymbol{x} = \boldsymbol{H}\boldsymbol{y}$.
(a) [5%] Consider the first PC only,

$$\begin{aligned}\boldsymbol{h_1'y} &= \begin{bmatrix}0.2786 & 0.2799 & 0.2798 & 0.2791 & 0.2785\end{bmatrix}\begin{bmatrix}+4\\+9\\-6\\-8\\+1\end{bmatrix}\\ &= (0.2786)(4) + (0.2799)(9) + (0.2798)(-6) + (0.2791)(-8) + (0.2785)(1)\\ &= 0.0004.\end{aligned}$$

That means change in 1 b.p. would result in 0.0004 (in million \$) increase in the value of portfolio. Then

$$s_1 \triangleq \mathrm{SD}(\Delta P) = \sqrt{\mathrm{Var}(0.0004y_1)} = 0.0004\sqrt{12.7239} = 0.001426823$$

or $\mathrm{Var}(\Delta P) = 2.0358 \times 10^{-6}$. Therefore, the 10-day 95% VaR is

$$\sqrt{10} \cdot z_{0.95} \cdot s_1 = \sqrt{10} \cdot 1.645 \cdot 0.001426823 = 0.0074222575.$$

(b) [5%] Consider the second PC only,

$$\begin{aligned}\boldsymbol{h_2'y} &= \begin{bmatrix}-0.2023 & -0.0353 & 0.0998 & 0.1682 & 0.2298\end{bmatrix}\begin{bmatrix}+4\\+9\\-6\\-8\\+1\end{bmatrix}\\ &= (-0.2023)(4) + (-0.0353)(9) + (0.0998)(-6) + (0.1682)(-8) + (0.2298)(1)\\ &= -2.8415.\end{aligned}$$

Then

$$\begin{aligned}s_2 \triangleq \mathrm{SD}(\Delta P) &= \sqrt{\mathrm{Var}(0.0004y_1) + \mathrm{Var}(-2.8415y_2)}\\ &= \sqrt{(0.0004)^2(12.7239) + (-2.8415)^2(0.2377)}\\ &= 1.385359482\end{aligned}$$

or $\mathrm{Var}(\Delta P) = 1.919220895$. Therefore, the 20-day 98% VaR is

$$\sqrt{20} \cdot z_{0.98} \cdot s_2 = \sqrt{20} \cdot 2.054 \cdot 1.385359482 = 12.72558976.$$

(c) [5%] Consider the third PC only,

$$\begin{aligned}\boldsymbol{h_3'y} &= (-0.2520)(4) + (-0.2885)(9) + (-0.2226)(-6) + (-0.1881)(-8) + (-0.0836)(1)\\ &= -0.8477.\end{aligned}$$

Then

$$\begin{aligned}s_3 \triangleq \mathrm{SD}(\Delta P) &= \sqrt{\mathrm{Var}(0.0004y_1) + \mathrm{Var}(-2.8415y_2) + \mathrm{Var}(-0.8477y_3)}\\ &= \sqrt{(0.0004)^2(12.7239) + (-2.8415)^2(0.2377) + (-0.8477)^2(0.0249)}\\ &= 1.391802399\end{aligned}$$

or $\mathrm{Var}(\Delta P) = 1.937113917$. Therefore, the 30-day 99% VaR is

$$\sqrt{30} \cdot z_{0.99} \cdot s_3 = \sqrt{30} \cdot 2.326 \cdot 1.391802399 = 17.73159971.$$

The following information refers to Question 2 to 4.
Define the relative return of stock price in percentage as $u_i = 100 \times (P_t - P_{t-1})/P_{t-1}, t = 1, ..., n$. Let $\boldsymbol{u} = (u_1, u_2, u_3)'$ be the daily relative return in percentage of three stocks: A, B and C. Suppose that the sample mean of $\boldsymbol{u}$ is $\bar{\boldsymbol{u}} = (0.1546, 0.1282, 0.0824)'$ and the last updated value of $\boldsymbol{u}$ is $\boldsymbol{u_n} = (-0.2060, -0.9592, -0.5076)'$ and the last stock price is $\boldsymbol{P_n} = (68.5, 78.7, 94.7)'$. Let $\boldsymbol{S}$ be the covariance matrix of $\boldsymbol{u}$ and $\boldsymbol{C}$ be the Cholesky decomposition of $\boldsymbol{S}$, i.e., $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{S}$.

$$\boldsymbol{S} = \begin{bmatrix} 1.0497 & 0.0384 & 0.0837 \\ 0.0384 & 0.5225 & 0.1426 \\ 0.0837 & 0.1426 & 0.5739 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ 0 & 0.7219 & 0.1933 \\ 0 & 0 & 0.7279 \end{bmatrix}.$$

Let $\boldsymbol{S} = \boldsymbol{HDH'}$ be the spectral decomposition of $\boldsymbol{S}$, where $\boldsymbol{D} = diag(\lambda_1, \lambda_2, \lambda_3)$ and $\lambda_1 > \lambda_2 > \lambda_3$,

$$\boldsymbol{H} = \begin{bmatrix} h_{11} & 0.2279 & -0.0390 \\ 0.1195 & -0.6401 & -0.7590 \\ 0.1979 & -0.7338 & 0.6500 \end{bmatrix}.$$

Question 2 [20%]

(a) Find the values of c_{11}, c_{12} and c_{13} and the correlation matrix $\boldsymbol{R}$ of $\boldsymbol{u}$.

(b) Compute the Cholesky decomposition of $\boldsymbol{R}$ in part (a). Show your calculation in detail.

(c) A random vector $\boldsymbol{z} = (-0.42, -1.55, -0.06)'$ is generated from $N_3(\boldsymbol{0}, \boldsymbol{I_3})$. Based on this $\boldsymbol{z}$, treat it as a realization of the next simulated $\boldsymbol{u_{n+1}} \sim N_3(\bar{\boldsymbol{u}}, \boldsymbol{S})$ and hence compute the next simulated stock price of A, B and C respectively.

(d) Given $\lambda_3 = 0.4024$, find the eigenvalues λ_1, λ_2 and hence find the square root of $\boldsymbol{S}$. [Hint: what are the determinant and trace of $\boldsymbol{S}$?]

Solution: (a) [5%] Note that $c_{11} = 1.0245$, $c_{12} = 0.0375$ and $c_{13} = 0.0817$ since

$$\begin{aligned} c_{11} &= \sqrt{a_{11}} = \sqrt{1.0497} = 1.024548681, \\ c_{12} &= \frac{a_{12}}{c_{11}} = \frac{0.0384}{1.024548681} = 0.037479917, \\ c_{13} &= \frac{a_{13}}{c_{11}} = \frac{0.0837}{1.024548681} = 0.081694507. \end{aligned}$$

Then

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{D^{-1/2} S D^{-1/2}} \\ &= \begin{bmatrix} \frac{1}{\sqrt{1.0497}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.5225}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.5739}} \end{bmatrix} \begin{bmatrix} 1.0497 & 0.0384 & 0.0837 \\ 0.0384 & 0.5225 & 0.1426 \\ 0.0837 & 0.1426 & 0.5739 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1.0497}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.5225}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.5739}} \end{bmatrix} \\ &= \begin{bmatrix} 1.024548681 & 0.037479917 & 0.081694507 \\ 0.05312367 & 0.722841614 & 0.197276965 \\ 0.110486024 & 0.188235448 & 0.757561878 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1.0497}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.5225}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.5739}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0.051850801 & 0.107838724 \\ 0.051850801 & 1 & 0.260410364 \\ 0.107838724 & 0.260410364 & 1 \end{bmatrix}. \end{aligned}$$

(b) [5%] Note that

$$\begin{aligned}\boldsymbol{R} &= \boldsymbol{D^{-1/2}SD^{-1/2}} \\ &= \boldsymbol{D^{-1/2}C'CD^{-1/2}} \\ &= (\boldsymbol{CD^{-1/2}})'(\boldsymbol{CD^{-1/2}}) \\ &\triangleq \boldsymbol{B'B},\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{B} = \boldsymbol{CD^{-1/2}} &= \begin{bmatrix} 1.024548681 & 0.037479917 & 0.081694507 \\ 0 & 0.7219 & 0.1933 \\ 0 & 0 & 0.7279 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1.0497}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.5225}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.5739}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0.051850801 & 0.107838724 \\ 0 & 0.998697342 & 0.255160674 \\ 0 & 0 & 0.960845602 \end{bmatrix}.\end{aligned}$$

(c) [5%] Note that $\boldsymbol{u_{n+1}} = \bar{\boldsymbol{u}} + \boldsymbol{C'z} \sim N_3(\bar{\boldsymbol{u}}, \boldsymbol{S})$ since

$$\begin{aligned}E(\boldsymbol{u_{n+1}}) &= \bar{\boldsymbol{u}} + \boldsymbol{C'}E(\boldsymbol{z}) = \bar{\boldsymbol{u}} + \boldsymbol{C'0} = \bar{\boldsymbol{u}}, \\ \mathrm{Cov}(\boldsymbol{u_{n+1}}) &= \boldsymbol{C'}\mathrm{Cov}(\boldsymbol{z})\boldsymbol{C} = \boldsymbol{C'C} = \boldsymbol{S}.\end{aligned}$$

Thus

$$\begin{aligned}\boldsymbol{u_{n+1}} &= \begin{bmatrix} 0.1546 \\ 0.1282 \\ 0.0824 \end{bmatrix} + \begin{bmatrix} 1.024548681 & 0 & 0 \\ 0.037479917 & 0.7219 & 0 \\ 0.081694507 & 0.1933 & 0.7279 \end{bmatrix} \begin{bmatrix} -0.42 \\ -1.55 \\ -0.06 \end{bmatrix} \\ &= \begin{bmatrix} 0.1546 \\ 0.1282 \\ 0.0824 \end{bmatrix} + \begin{bmatrix} -0.430310446 \\ -1.134686565 \\ -0.377600692 \end{bmatrix} \\ &= \begin{bmatrix} -0.275710446 \\ -1.006486565 \\ -0.295200692 \end{bmatrix}.\end{aligned}$$

Therefore, the next simulated stock prices are

$$\begin{bmatrix} 68.5 \\ 78.7 \\ 94.7 \end{bmatrix} \begin{bmatrix} 1 - 0.275710446/100 \\ 1 - 1.006486565/100 \\ 1 - 0.295200692/100 \end{bmatrix} = \begin{bmatrix} 68.31 \\ 77.91 \\ 94.42 \end{bmatrix}.$$

(d) [5%] By $\det(\boldsymbol{S}) = \prod_{i=1}^{3} \lambda_i$, we have

$$0.289830463 = \lambda_1\lambda_2 \cdot 0.4024 \;\Rightarrow\; \lambda_1\lambda_2 = 0.720254629.$$

By $\mathrm{tr}(\boldsymbol{S}) = \sum_{i=1}^{3} \lambda_i$, we have

$$2.1461 = \lambda_1 + \lambda_2 + 0.4024 \;\Rightarrow\; \lambda_1 + \lambda_2 = 1.7437.$$

Combining the results,

$$\begin{aligned}&\lambda_1(1.7437 - \lambda_1) = 0.720254629 \\ \Rightarrow\quad &1.7437\lambda_1 - \lambda_1^2 = 0.720254629\end{aligned}$$

and hence

$$\lambda_1^2 - 1.7437\lambda_1 + 0.720254629 = 0.$$

Solving the quadratic equation and given that $\lambda_1 > \lambda_2 > \lambda_3$, we get

$$\lambda_1 = 1.07151921 \quad \text{and} \quad \lambda_2 = 0.672180789.$$

To solve h_{11}, note that the eigenvectors are orthogonal to each other, i.e.

$$0.2279h_{11} + (-0.6401)(0.1195) + (-0.7338)(0.1979) = 0.2279h_{11} - 0.22171097 = 0.$$

Thus $h_{11} = 0.97284322$. The square root of $\boldsymbol{S}$ is

$$\begin{aligned}
\boldsymbol{S^{1/2}} &= \boldsymbol{H D^{1/2} H'} \\
&= \begin{bmatrix} 0.9728 & 0.2279 & -0.0390 \\ 0.1195 & -0.6401 & -0.7590 \\ 0.1979 & -0.7338 & 0.6500 \end{bmatrix} \begin{bmatrix} \sqrt{1.072} & 0 & 0 \\ 0 & \sqrt{0.672} & 0 \\ 0 & 0 & \sqrt{0.402} \end{bmatrix} \begin{bmatrix} 0.9728 & 0.1195 & 0.1979 \\ 0.2279 & -0.6401 & -0.7338 \\ -0.0390 & -0.7590 & 0.6500 \end{bmatrix} \\
&= \begin{bmatrix} 1.007030994 & 0.186847535 & -0.024739652 \\ 0.123699483 & -0.524796433 & -0.481471696 \\ 0.204854625 & -0.601617908 & 0.412327529 \end{bmatrix} \begin{bmatrix} 0.9728 & 0.1195 & 0.1979 \\ 0.2279 & -0.6401 & -0.7338 \\ -0.0390 & -0.7590 & 0.6500 \end{bmatrix} \\
&= \begin{bmatrix} 1.023230674 & 0.019516492 & 0.046101938 \\ 0.019516492 & 0.716141302 & 0.096619147 \\ 0.046101938 & 0.096619156 & 0.750020845 \end{bmatrix}.
\end{aligned}$$

Question 3 [10%]

Suppose that we form a portfolio Q by spending \$6000 on buying each stock A, B and C on the last day n. Assume that the 1-day loss distribution of Q is $L \sim N(\mu, \sigma^2)$.

(a) Compute reasonable estimates of μ and σ, and hence compute the 22-day 99% VaR of Q.

(b) Do you expect the calculated VaR in part (a) is smaller than that obtained via extreme value theory (EVT)? Why?

Solution: (a) [8%]

$$\hat{\mu} = -\frac{1}{100}\boldsymbol{w}'\bar{\boldsymbol{u}} = -\frac{1}{100}\begin{bmatrix} 6000 & 6000 & 6000 \end{bmatrix}\begin{bmatrix} 0.1546 \\ 0.1282 \\ 0.0824 \end{bmatrix} = -\frac{6000}{100} \cdot 0.3652 = -21.912.$$

$$\begin{aligned}
\widehat{\sigma^2} = \frac{1}{100^2}\boldsymbol{w}'\boldsymbol{S}\boldsymbol{w} &= \frac{1}{100^2}\begin{bmatrix} 6000 & 6000 & 6000 \end{bmatrix}\begin{bmatrix} 1.0497 & 0.0384 & 0.0837 \\ 0.0384 & 0.5225 & 0.1426 \\ 0.0837 & 0.1426 & 0.5739 \end{bmatrix}\begin{bmatrix} 6000 \\ 6000 \\ 6000 \end{bmatrix} \\
&= \frac{1}{100^2}\begin{bmatrix} 6000 & 6000 & 6000 \end{bmatrix}\begin{bmatrix} 7030.8 \\ 4221 \\ 4801.2 \end{bmatrix} \\
&= 9631.8.
\end{aligned}$$

Thus the reasonable estimates of μ and σ are

$$\hat{\mu} = -21.912 \quad \text{and} \quad \hat{\sigma} = \sqrt{9631.8} = 98.14173424.$$

Assume that the daily returns are i.i.d. and returns are additive. Then the 22-day loss follows $N(22 \cdot \hat{\mu},\ \{\sqrt{22} \cdot \hat{\sigma}\}^2)$ and hence the 22-day 99% VaR is

$$22\hat{\mu} + \sqrt{22} \cdot z_{0.99} \cdot \hat{\sigma} = (22)(-21.912) + \sqrt{22} \cdot 2.326 \cdot 98.14173424 = 588.653199.$$

(b) [2%] Yes, it is expected to be smaller because extreme value distribution has a heavier tail than normal distribution does.

Question 4 [15%]

(a) Find a reasonable estimate of the current variance rate σ_n^2 of stock A based on S.

(b) A GARCH(1,1) model $\sigma_{n+1}^2 = \omega + \beta\sigma_n^2 + \alpha u_n^2$ is fitted to stock A with $\hat{\omega} = 0.1142, \hat{\alpha} = 0.0837, \hat{\beta} = 0.8347$. Find the long-run variance rate V_L.

(c) Why is GARCH(1,1) model mean-reverted? Give an explanation in your own words.

(d) What is $\mathbb{E}(\sigma_{n+10}^2|\sigma_n^2)$?

Solution: (a) [2%] A reasonable estimate is $\widehat{\sigma_n^2} = 1.0497$ because it is the variance of returns of stock A based on the past n days.

(b) [4%] By $\sigma_{n+1}^2 = \gamma V_L + \beta\sigma_n^2 + \alpha u_n^2$, where $\gamma + \alpha + \beta = 1$. So $\gamma = 1 - \alpha - \beta$ and $\hat{\omega} = (1 - \hat{\alpha} - \hat{\beta})V_L$. The long-run variance rate is

$$V_L = \frac{\hat{\omega}}{1 - \hat{\alpha} - \hat{\beta}} = \frac{0.1142}{1 - 0.0837 - 0.8347} = 1.399509804.$$

(c) [5%] Note that $E(\sigma_{n+k}^2|\sigma_n^2) = V_L + (\alpha+\beta)^k(\sigma_n^2 - V_L)$. The expected variance rate exhibits a mean reversion property with reversion level V_L. If the current variance rate is above V_L, the estimated future variance rate would be pushed down to V_L. Similarly, if the current variance rate is below V_L, the estimate future variance rate would be pulled up to V_L.

(d) [4%]

$$\begin{aligned} E(\sigma_{n+10}^2|\sigma_n^2) &= V_L + (\alpha+\beta)^{10}(\sigma_n^2 - V_L) \\ &= 1.399509804 + (0.0837 + 0.8347)^{10}(1.0497 - 1.399509804) \\ &= 1.250178545. \end{aligned}$$

Dataset for Questions 5 to 7.
The dataset is random subset from a customer database of a telephone company.
Note that the column 7 is the target variable: Vmail_Plan is a binary variable. For the next 3 questions, we want to ask for classification problems in relation to the other 6 variables.

Colomn	Name	Description
1	Vmail_Plan	Binary, Voice mail plan (1=yes or 0=no)
2	Day_Mins	Continuous, minutes used during daytime
3	Day_Charge	Continuous, charge fee for calls in daytime
4	Eve_Mins	Continuous, minutes used during evening
5	Eve_Charge	Continuous, charge fee for calls in evening
6	CustServ_Calls	Integer-valued, number of calls to Customer Service
7	Change	Binary target variable: 0=stay, 1=change to other company

Question 5 [20%]

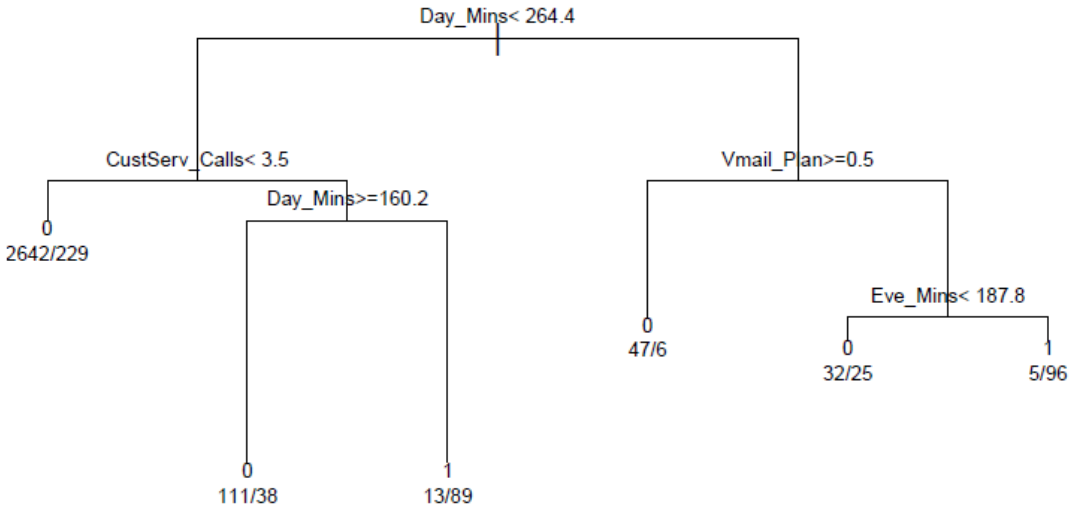
Classification tree is built using all data. The result is saved in *ctree* with the following output:

```
n= 3333
node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 3333 483 0 (0.85508551 0.14491449)
   2) Day_Mins< 264.45 #  # 0 (#  #)
     4) CustServ_Calls< 3.5 2871 229 0 (0.92023685 0.07976315) *
     5) CustServ_Calls>=3.5 #  # 1 (#  #)
      10) Day_Mins>=160.2 149  38 0 (0.74496644 0.25503356) *
      11) Day_Mins< 160.2 102  13 1 (0.12745098 0.87254902) *
   3) Day_Mins>=264.45 #  # 1 (#  #)
     6) Vmail_Plan>=0.5 53   6 0 (0.88679245 0.11320755) *
     7) Vmail_Plan< 0.5 #  # 1 (#  #)
      14) Eve_Mins< 187.75 57  25 0 (0.56140351 0.43859649) *
      15) Eve_Mins>=187.75 101   5 1 (0.04950495 0.95049505) *
```

(a) Plot the classification tree and provide the number of observations for each class in each terminal node in it.

Solution: [5%]



(b) The terminal node 6) means that if (Day_Mins>=264.45 and Vmail_Plan>=0.5) then Change=0. Additional information is given here. Under node 6), there are 15 observations such that Eve_Mins<187.75, where all of them (i.e. 15) have true label 0. Under node 6), there are another 38 observations such that Eve_Mins>=187.75, where 32 of them have true label 0 while 6 of them have true label 1.

i Using Gini index, what is the change in impurity measure when we use only one rule Vmail_Plan>=0.5 right after node 3)? Also using Gini index, what is the change in impurity measure when we use only one rule Eve_Mins<187.75 right after node 3)?

ii Which of the two rules should you use if you can only choose one rule right after node 3)? Give the reason.

Solution: (b) i [10%] (1) When we use only one rule Vmail_Plan>=0.5,

At the left node,

$$\text{Gini} = 1 - \left(\frac{47}{53}\right)^2 - \left(\frac{6}{53}\right)^2 = 0.200783196.$$

At the right node,

$$\text{Gini} = 1 - \left(\frac{37}{158}\right)^2 - \left(\frac{121}{158}\right)^2 = 0.358676494.$$

The change in impurity measure is

$$0.479234518 - \frac{53}{211} \cdot 0.200783196 - \frac{158}{211} \cdot 0.358676494 = 0.160218425.$$

(2) When we use only one rule Eve_Mins<187.75,

At the left node,

$$\text{Gini} = 1 - \left(\frac{47}{72}\right)^2 - \left(\frac{25}{72}\right)^2 = 0.453317901.$$

At the right node,

$$\text{Gini} = 1 - \left(\frac{37}{139}\right)^2 - \left(\frac{102}{139}\right)^2 = 0.390663009.$$

The change in impurity measure is

$$0.479234518 - \frac{72}{211} \cdot 0.453317901 - \frac{139}{211} \cdot 0.390663009 = 0.06719164.$$

(b) ii [5%] I should use the first rule Vmail_Plan>=0.5 right after node 3) because the change in impurity measure (0.160218425) is larger. It is the reason why the rule is chosen there for splitting.

Question 6 [15%]

There are 3333 observations in total. Around 70% of them (exactly 2333) are used as training data while around 30% of them (exactly 1000) are used as testing data. A logistic regression is fitted with backward elimination using the training data. The final model is as follows:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.182694   0.477553 -15.041  < 2e-16 ***
Vmail_Plan          4.813856   0.931470   5.168 2.37e-07 ***
Day_Mins            0.015923   0.001453  10.957  < 2e-16 ***
Eve_Mins            0.008211   0.001488   5.520 3.39e-08 ***
CustServ_Calls      0.435194   0.044858   9.702  < 2e-16 ***
Vmail_Plan:Day_Mins -0.017321   0.003283  -5.276 1.32e-07 ***
Vmail_Plan:Eve_Mins -0.011452   0.003398  -3.370 0.000752 ***
```

(a) Write down the logistic regression model for Vmail_Plan=l and Vmail_Plan=0 separately.

(b) Suppose that classification tables for training and testing data are as follows:

pred_train \ label_train	0	1
0	1991	283
1	16	43

pred_test \ label_test	0	1
0	833	138
1	10	19

Note that rows refer to prediction outcomes and columns refer to true labels. Calculate precision, recall and F1 score for training and testing data respectively. Justify the differences, i.e. why result in one batch is bigger in general.

(c) Compare the similarities and differences between linear regression and logistic regression.

Solution: (a) [5%] Let $P(\text{Change} = 1|\boldsymbol{x}) \triangleq \pi$.

The logistic regression model for Vmail_Plan=1 is

$$\ln\left(\frac{\pi}{1-\pi}\right) = -2.368838 - 0.001398\text{Day\_Mins} - 0.003241\text{Eve\_Mins} + 0.435194\text{CustServ\_Calls}.$$

The logistic regression model for Vmail_Plan=0 is

$$\ln\left(\frac{\pi}{1-\pi}\right) = -7.182694 + 0.015923\text{Day\_Mins} + 0.008211\text{Eve\_Mins} + 0.435194\text{CustServ\_Calls}.$$

(b) [8%] Define Change=1 as positive since we are interested in whether a customer would change to other company or not. For the training data,

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{43}{43+16} = \frac{43}{59} = 0.728813559,$$
$$\text{Recall} = \frac{TP}{TP+FN} = \frac{43}{43+283} = \frac{43}{326} = 0.13190184,$$
$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot 0.728813559 \cdot 0.13190184}{0.728813559 + 0.13190184} = 0.223376623.$$

Similarly, for the testing data,

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{19}{19+10} = \frac{19}{29} = 0.655172413,$$
$$\text{Recall} = \frac{TP}{TP+FN} = \frac{19}{19+138} = \frac{19}{157} = 0.121019108,$$
$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot 0.655172413 \cdot 0.121019108}{0.655172413 + 0.121019108} = 0.204301074.$$

The precision, recall and F1 score are bigger in the training data in general since the model is built using the training data and hence the performance should be better in training data.

(c) [2%] For the similarities, the systematic component of the models is linear, i.e. $\sum_{j=1}^{p} \beta_j X_j$. Actually, they belong to a class called generalized linear models. For the differences, the response in linear regression is the target outcome Y while that in logistic regression is the log-odd ratio of probability of success, i.e. $\text{logit}(P(Y = 1))$.

Question 7 [5%]

Artitical neural network with (i1, i2, i3, i4)= (Vmail_Plan, Day_Mins, Eve_Mins, CustServ_Cal1s) as input and *Result* as target variable is fitted while *size* = *3* and *linout* = *T* option is used.

```
a 4-3-1 network with 19 weights
options were - linear output units
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1
 213.97   31.40 -171.65   51.45  529.72
  b->h2  i1->h2  i2->h2  i3->h2  i4->h2
-232.58  181.95   52.66  -16.25   24.08
  b->h3  i1->h3  i2->h3  i3->h3  i4->h3
  -0.50   -0.54   -0.77   -0.98   -0.32
   b->o   h1->o   h2->o   h3->o
  -0.69    0.79    0.84   -0.42
```

Suppose that we have a record x_0=(i1, i2, i3, i4)= (1, 161.6, 195.5, 1), how should we classify x_0 according to this ANN model?

Solution:

$$
\begin{aligned}
h_1 &= 213.97 + 31.40(1) - 171.65(161.6) + 51.45(195.5) + 529.72(1) = -16905.075, \\
h_2 &= -232.58 + 181.95(1) + 52.66(161.6) - 16.25(195.5) + 24.08(1) = 5306.431, \\
h_3 &= -0.50 - 0.54(1) - 0.77(161.6) - 0.98(195.5) - 0.32(1) = -317.382, \\
h_1' &= \frac{\exp(h_1)}{1 + \exp(h_1)} = \frac{\exp(-16905.075)}{1 + \exp(-16905.075)} \approx 0, \\
h_2' &= \frac{\exp(h_2)}{1 + \exp(h_2)} = \frac{\exp(5306.431)}{1 + \exp(5306.431)} \approx 1, \\
h_3' &= \frac{\exp(h_3)}{1 + \exp(h_3)} = \frac{\exp(-317.382)}{1 + \exp(-317.382)} \approx 0, \\
v &= -0.69 + 0.79(0) + 0.84(1) - 0.42(0) = 0.15.
\end{aligned}
$$

So the prediction for x_0 is 0. In fact, the true label is also 0.