1 st Semester 2019/20 - 1st EXAM - 7/01/2020

GI a) $\hat{Y} = 2.3412 + 1.6159 x_1 + 0.0144 x_2$; $\hat{Var}(Y_1) = \hat{\sigma}^2 = MSE = 10.6$

b) $r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{233.7}{5382.4+168.4+233.7} \simeq 0.9596$

i)

$r^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{233.7/22}{\frac{5382.4+168.4+233.7}{24}} \simeq 0.9559$

Both coefficients are very high, suggesting that a high proportion of the total observed variability is explained by this regression model ($r^2 \simeq 0.9596$), leading to a very good model to this data.

ii) $H_0: \beta_1 = \beta_2 = 0$ vs $H_1: \beta_1 \neq 0 \vee \beta_2 \neq 0$

Under H_0, we have the test statistic: $F_0 = \frac{MSR}{MSE} \sim F_{(2,22)}$.

For the significant level $\alpha = 0.05$, reject H_0 if $F_0 > F^{-1}_{F(2,22)}(0.95) = 3.44$.

Thus, the critical region is $C.R =]3.44, +\infty[$

Observed value of the test statistic: $f_0 = \frac{(5382.4+168.4)/2}{233.7/22} \simeq 261.27 \in C.R \Rightarrow$

$\Rightarrow$ reject H_0 for $\alpha = 0.05$, ie, seems that there is evidence of a linear association between the delivery time (Y) and the two explanatory variables.

Assumption to obtain the test statistic: $\varepsilon_i \underset{iid}{\sim} N(0, \sigma^2)$

c) $H_{i0}: \beta_i = 0$ vs $H_{i1}: \beta_i \neq 0$, for $i = 1, 2$

Under H_{i0}, we have the test statistic: $T_{i0} = \frac{\hat{\beta}_i}{sd(\hat{\beta}_i)} \sim t_{(22)}$

reject H_0 if $|T_{i0}| > c$

R output

i	t_{i0}	p-value
1	9.464	3.25×10^{-9}
2	3.981	0.000631

Decision: For all usual significant levels (1%, 5%, and 10%) we reject the null hypothesis $\beta_1 = 0$ and $\beta_2 = 0$

d) $CI_{95\%}(\beta_0) = ?$

Pivotal quantity: $T = \frac{\hat{\beta}_0 - \beta_0}{sd(\hat{\beta}_0)} \sim t_{(22)}$

taking the symmetrical confidence interval, we have that

$P(-c \leq T \leq c) = 0.95 \Rightarrow c = F^{-1}_{t(22)}(0.975) = 2.074$

$P(-2.074 \leq \frac{\hat{\beta}_0 - \beta_0}{sd(\hat{\beta}_0)} \leq 2.074) = 0.95 \Leftrightarrow$

$P(\hat{\beta}_0 - 2.074 \times sd(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + 2.074 \times sd(\hat{\beta}_0)) = 0.95$

$C.I_{95\%}(\beta_0) = [\hat{\beta}_0 \pm 2.074 \times sd(\hat{\beta}_0)] \simeq [2.341 \pm 2.074 \times 1.097] =$

$= [0.0658; 4.6162]$

G II

a) $a = ?$; since $\hat{\gamma}_2^T \hat{\gamma}_1 = 0$, we have that

$-0.19534679 \times (-0.6917168) - 0.09741748 \times (-0.6917123) + a \times (-0.2075139) = 0$

$\Leftrightarrow -0.2075139a = -0.2025095 \Leftrightarrow a = +0.975884$

and $\|\hat{\gamma}_2\| = 1$

total variance $= tr(S) = 16.0001 = \sum_{i=1}^{3} \hat{\lambda}_i$

Generalized variance $= |S| = \prod_{i=1}^{3} \hat{\lambda}_i = 38.157466609$

b)

component	variance	% explained $= \frac{\lambda_i}{\sum_{i=1}^{3} \lambda_i}$
1	$\simeq 12.5$	$\simeq 0.781$
2	$\simeq 1.85$	$\simeq 0.116$
3	$\simeq 1.65$	$\simeq 0.103$

$\Rightarrow$ the 1st P.C. explains a very large proportion of the total variance. We may consider dropping components 2 and 3.

c) $\hat{Y}_1 = -0.6917 x_1 - 0.6917 x_2 - 0.2075 x_3$

Large values of Y_1 tend to occure if x_1, x_2 and x_3 are negative
$\Rightarrow w_1, w_2$ and w_3 smaller than the corresponding means.
On the opposite side, smaller values of Y_1 tend to occure if variables x_1, x_2 and x_3 are positive $\Rightarrow w_1, w_2$ and w_3 are all greater than their means.
Yes, the P.C. obtain from the correlation matrix are different.

d) $x_i = \underset{\sim}{e}_i^T \underset{\sim}{x}$ and $\hat{Y}_1 = \hat{\underset{\sim}{\gamma}}_1^T \underset{\sim}{x}$

$\hat{cov}(x_i, \hat{Y}_1) = cov(\underset{\sim}{e}_i^T \underset{\sim}{x}, \hat{\underset{\sim}{\gamma}}_1^T \underset{\sim}{x}) = \underset{\sim}{e}_i^T \hat{\Sigma} \hat{\underset{\sim}{\gamma}}_1 = \underset{\sim}{e}_i^T \hat{\lambda}_1 \hat{\underset{\sim}{\gamma}}_1 = \hat{\lambda}_1 \underset{\sim}{e}_i^T \hat{\underset{\sim}{\gamma}}_1 = \hat{\lambda}_1 \hat{\gamma}_{1i}$

$\hat{Var}(\hat{Y}_1) = \hat{Var}(\hat{\gamma}_1^T \underset{\sim}{x}) = \hat{\underset{\sim}{\gamma}}_1^T \hat{Var}(\underset{\sim}{x}) \hat{\underset{\sim}{\gamma}}_1 = \hat{\underset{\sim}{\gamma}}_1^T \hat{\Sigma} \hat{\underset{\sim}{\gamma}}_1 = \hat{\underset{\sim}{\gamma}}_1^T \hat{\lambda}_1 \hat{\underset{\sim}{\gamma}}_1 = \hat{\lambda}_1 \hat{\underset{\sim}{\gamma}}_1^T \hat{\underset{\sim}{\gamma}}_1$

$= \hat{\lambda}_1$ since $\hat{\underset{\sim}{\gamma}}_1^T \hat{\underset{\sim}{\gamma}}_1 = \|\hat{\underset{\sim}{\gamma}}_1\| = 1$

$$\hat{\rho}_{x_1, \hat{Y}_1} = \frac{\hat{\lambda}_1 \hat{\gamma}_{11}}{\hat{\sigma}_1 \sqrt{\hat{\lambda}_1}} = \frac{\sqrt{\hat{\lambda}_1}\, \hat{\gamma}_{11}}{\hat{\sigma}_1} = \frac{\sqrt{12.500054} \times (-0.6917168)}{\sqrt{6.8491}} \simeq -0.9345$$

$\rho_{x_2, \hat{Y}_1} \simeq -0.9349$

$\rho_{x_3, \hat{Y}_1} \simeq -0.4830$.

$\hat{Y}_1$ is negative correlated with all variables (more correlated with x_1 and x_2). The same interpretation of part c).

G III

a) communalities: $h_i^2 = \sum_{j=1}^{2} l_{ij}^2 = \begin{cases} 1^2+1^2 = 2 \\ 1^2+0^2 = 1 \\ 2^2+2^2 = 8 \\ 2^2+0^2 = 4 \\ 3^2+1^2 = 10 \end{cases}$

unique variances : $\Psi_i = Var(x_i) - h_i^2 = \begin{cases} 4-2 = 2 \\ 3-1 = 2 \\ 10-8 = 2 \\ 6-4 = 2 \\ 12-10 = 2 \end{cases}$

total variance : $tr(\Sigma) = 35$

proportion of variance explained by each common factor:

$f_1 : \frac{1^2+1^2+2^2+2^2+3^2}{35} = 19/35 = 0.5429$

$f_2 : \frac{1^2+0^2+2^2+0^2+1^2}{35} = 6/35 = 0.1714$

both f_1 and f_2 explained 0.7143 of the total variance.
However, the contribution of f_1 is more than 3x greater than f_2 in the explanation of the total variance.

b) $\underset{\sim}{\Lambda}^* = \underset{\sim}{\Lambda} \times \underset{\sim}{C} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 2 & -2 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 \\ 0 & 2\sqrt{2} \\ \sqrt{2} & \sqrt{2} \\ 2\sqrt{2} & \sqrt{2} \end{bmatrix}$ (2)

New communalities: $h_i^{*2} = \begin{cases} 2 + 0 = 2 \\ 1/2 + 1/2 = 1 \\ 0 + 8 = 8 \\ 2 + 2 = 4 \\ 8 + 2 = 10 \end{cases}$ $\quad \psi_i^* = 2, \forall i$

Proportion of total variance explained by: $f_1 = \frac{2 + 1/2 + 0 + 2 + 8}{35} = \frac{12.5}{35} = 0.3571$

$f_2 = \frac{0 + 1/2 + 8 + 2 + 2}{35} = \frac{12.5}{35} = 0.3571$

$f_1 + f_2$ explain $25/35 = 5/7 = 0.7142$

∴ there is no vantage in this rotation.

G IV

a) $\underset{\sim}{D}^{(1)} = \begin{array}{c|ccccc} A & 0 \\ B & 16.84 & 0 \\ C & (4.90) & 13.59 & 0 \\ Da & 9.54 & 25.10 & 11.51 & 0 \\ De & 11.45 & 24.07 & 11.09 & 5.32 & 0 \end{array}$

single-Linkage: $d_{AB} = \min\{d_{ij}, \forall i \in A, j \in B\}$

h_1 = threshold distance = 4.9 ⇒ {A, C} = c_1

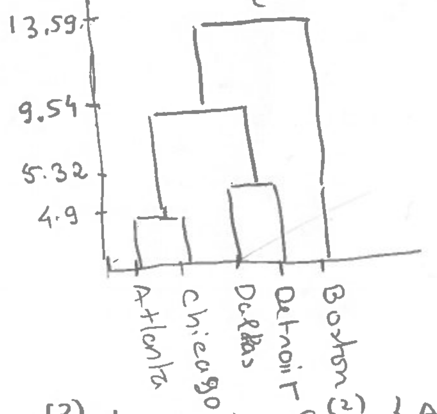
$D^{(2)} = \begin{array}{c|cccc} \{A,C\} & 0 \\ B & 13.59 & 0 \\ Da & 9.54 & 25.1 & 0 \\ De & 11.09 & 24.07 & [5.32] & 0 \end{array}$

$h_2 = 5.32$ ⇒ {Da, De}

$D^{(3)} = \begin{array}{c|ccc} & \{A,C\} & B & \{Da,De\} \\ \{A,C\} & 0 \\ B & 13.59 & 0 \\ \{Da,De\} & (9.54) & 24.07 & 0 \end{array}$

$h_3 = 9.54$ ⇒ {{A,C}, {Da, De}}

$D^{(4)} = \begin{array}{c|cc} \{\{A,C\},\{Da,DE\}\} & 0 \\ B & 13.59 & 0 \end{array}$



b)

centroid	A	B	C	Da	De
$c_1^{(1)}$ (4.2; 13.3)	16.84	(0)	13.59	–	24.07
		25.10	(11.51)	(0)	(5.32)
$c_2^{(1)}$ (18.1; 34.2)	(9.54)				

$c_1^{(2)} = \{Boston\}$; $c_2^{(2)} = \{At; C, Da, De\}$ $\quad \bar{c}_2 = \left(\frac{16.5 + 11.6 + 18.1 + 13}{4}; \frac{24.8 + 24.7 + 34.2 + 35.07}{4}\right) = (14.8; 29.85)$

centroid	A	B	C	Da	De
$c_1^{(2)}$: (4.2; 13.3)	16.85	(0)	13.59	25.10	24.07
$c_2^{(2)}$: (14.8; 29.85)	(5.33)	–	(6.063)	(5.460)	(6.12)

End

Final partition:

C_1 = {Boston}; C_2 = {Atlanta; Chicago; Dallas; Detroit}