GM – Probabilités et Statistique – Corrections 10

Please note : the **reasoning/justifications** for the steps in your solution are also important (not only the final result).

In class

Exercise 1 (a) Response variable : literacy rate.
Explanatory variables : number of daily newspaper copies, number of radios, and number of TV sets (all per 1000 people in the population of the country).

(b) $\widehat{y} = 0.51486 + 0.00054\,x_1 - 0.00035\,x_2 + 0.00199\,x_3$, where $\widehat{y}$ = predicted literacy rate,
x_1 = the number of daily newspaper copies in the country (per 1000 people),
x_2 = the number of radios in the country (per 1000 people),
x_3 = the number of TV sets in the country (per 1000 people).

(c) For countries with the same number of radios and same number of TV sets per 1000 people in the population, literacy rate is predicted to be 0.00054 higher for every additional daily newspaper copy per 1000 people in the population.

(d) First, we should check that we're not extrapolating : all values of the explanatory variables are within the range of the data collected, so we're okay. Second, make sure you have the right values in for the right x's.

Then $\hat{y} = 0.51486 + 0.00054 \times 200 - 0.00035\, times\, 800 + 0.00199\, times\, 250 \approx 0.84$
$\Rightarrow$ we predict about 84% of the residents to be literate in such a country.

(e) $\hat{b}_3 \pm \boxed{t_{n-p-1,0.975}}\, SE(\hat{b}_3)$; $n = 10, p = 3$; $\Rightarrow$ $\boxed{0.00199 \pm 2.447 \times 0.00155}$
($\Rightarrow$ $(-0.00180, 0.00578)$)

Exercise 2 (a) $\Omega : y = \beta_0 + \beta_1\,x_1 + \beta_2\,x_2 + \epsilon$

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

(c) $Var(\hat{\beta}_i) = \sigma^2\ [(\mathbf{X}'\mathbf{X})^{-1}]_{i+1,\,i+1}$, thus $Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \times (z_{11},\ z_{22},\ z_{33})$
$= 11.06 \times (3.437,\ 0.0014,\ 0.0021) = (38.01, 0.015, 0.023)$.

(d)

source	df	SS	MS	F	p-value
regression	2	184.2	$184.2/2 =$ **92.1**	$92.1/11.06 =$ **8.33**	0.0020
error	22	243.3	$243.3/22 =$ **11.06**		
total	2+22 = **24**	184.2+243.3 = **427.5**			

(e) $R^2 = 1 - SSE/SST = 1 - 243.3/427.5$ $\boxed{\approx 0.43}$

$$R^2_{aj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - (1 - 0.43)\left(\frac{24}{22}\right) \quad \boxed{\approx 0.38}$$

(f) $H : \beta_1 = \beta_2 = 0$; A : at least one $\beta_i \neq 0$. According to the anova table, $F_{obs} = 8.33 \sim F_{2,22}$ (under H) ; $p_{obs} = 0.002 < 0.05$, therefore we REJECT H.

(g) $$\boxed{\hat{\beta}_2 \pm \hat{\sigma} \cdot \sqrt{z_{33}} \cdot \boxed{t_{n-p-1,0.975}}}$$

$\Rightarrow -0.555 \pm \sqrt{11.06 \times 0.0021}\ \boxed{t_{22,0.975}} \Rightarrow \boxed{-0.555 \pm 0.15 \times 2.074}$

(or -0.555 ± 0.31 or $[-0.865, -0.245]$)

(h) $SSE(\omega) = 336.5$, $SSE(\Omega) = 243.3$.

$$F_{obs} = \frac{[SSE(\omega) - SSE(\Omega)]/(p-q)}{SSE(\Omega)/(n-p-1)} = \frac{(336.5 - 243.3)/(2-1)}{243.3/22} = \frac{93.2}{11.06} = 8.43 \sim F_{1,22}.$$

Numerator degrees of freedom = 1, therefore $F_{1,22} = t_{22}^2$, then we have $|t| = \sqrt{8.43} = 2.9$. $t_{22,0.975} = 2.074 < |t_{obs} = 2.9|$, and thus we REJECT H.

Source	df	SS	MS	F	p
(h) Groups	(a) # groups − 1 $\Longrightarrow$ $4-1=\mathbf{3}$	(d) MS group × df group 116×3 (= **348**)	116	(g) $\frac{\text{MS group}}{\text{MS error}}$ $\Longrightarrow$ $116/10 = \mathbf{11.6}$	1.8×10^{-5}
Error	(b) df total - df group $\Longrightarrow$ $39-3=\mathbf{36}$	360	$(f)\,\frac{\text{SS error}}{\text{df error}}$ $\Longrightarrow$ $360/36 = \mathbf{10}$		
Total	(c) # obs − 1 $\Longrightarrow$ $4 \times 10 - 1 = \mathbf{39}$	(e) SS group + SS error $\Longrightarrow$ $116 \times 3 + 360$ (=**708**)			

Exercise 3 (a)-(h) (see the table)

(i) 1. The parameters are $\mu_1, \mu_2, \mu_3, \mu_4$, the means of the 4 populations (the 4 groups)
2. $H : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. $A : \exists \mu_i \neq \mu_j, i \neq j$
3. $F_{obs} = 11.6$
4. $p_{obs} = 1.8 \times 10^{-5}$
5. $p_{obs} < \alpha = 0.05 \Longrightarrow$ REJECT H.

(j) df numerateur = 3, df denominateur = 36

At home

Exercise 1 (a) (see the annotated output)

(b) $n - 1 = 17 + 3 = 18 + 2 = 20$, thus $n = 20 + 1 = 21$

(c) $SSE = 178.8$, $MSE = 178.8/17 = 10.5$; $\hat{\sigma} = \sqrt{10.5} = 3.24$

(d) $\hat{\beta}_2 \pm SE(\hat{\beta}_2) \cdot t_{n-p-1,0.975} \Rightarrow \boxed{1.3 \pm 0.4\, t_{17,0.975}} \Rightarrow 1.3 \pm 0.4 \times 2.11$ (ou 1.3 ± 0.84 ou $[0.46, 2.14]$)

(e) $H : \omega$ is the true model ($\beta_3 = 0$)
$A : \Omega$ is the true model ($\beta_3 \neq 0$)

$$F_{obs} = \frac{[SSE(\omega) - SSE(\Omega)]/(p-q)}{SSE(\Omega)/(n-p-1)} = \frac{(188.8 - 178.8)/(3-2)}{178.8/17} = \frac{10}{10.5} = 0.95 \sim F_{1,17}.$$

Numerator degrees of freedom = 1, thus $F_{1,17} = t_{17}^2$, then we have $|t| = \sqrt{0.95} = 0.97$. $t_{17,0.975} = 2.11 > |t_{obs} = 0.97|$, thus we DO NOT REJECT H.

Exercise 2 (a) ANOVA table :

Source	df	SS	MS	F	p
Pizza variety	$3 - 1 = \mathbf{2}$	23.0	$\mathbf{23.0/2 = 11.5}$	$\mathbf{11.5/17.6 = 0.65}$	0.53
Error	$20 - 2 = \mathbf{18}$	$339.9 - 23.0 = \mathbf{316.9}$	$\mathbf{316.9/18 = 17.6}$		
Total	20	339.9			

(b) NO : $p = 0.53 > \alpha = 0.05 \Rightarrow$ DO NOT REJECT H that the mean fat contents of the three varieties are equal.

(c) NO : post-hoc tests are to assess WHICH means are different – this makes no sense when the global test has not detected a (significant) difference.

Exercise 3 (a) Independent samples (random attribution to groups), equal variances for each group, and either the measures are normally distributed in each group or the sample sizes are sufficiently large that the means can be assumed to be Normally distributed (CLT).

(b) ANOVA table :

Source	df	SS	MS	F	p
Treatments	$5 - 1 = \mathbf{4}$	$11681.4 - 4800 = \mathbf{6881.4}$	$6881.4/4 = \mathbf{1720.4}$	$1720.4/80 = \mathbf{21.5}$	4.8×10^{-11}
Error	60	$60 \times 80 = \mathbf{4800}$	80		
Total	$60 + 4 = \mathbf{64}$	11681.4			

(c) $H : \mu_1 = \mu_2 = \mu_3$
$A : \exists\, \mu_i \neq \mu_j$ (at least one mean is different from the others)
$p_{obs} = 0.04359$, thus we REJECT the NULL hypothesis. At least one group has a mean that is significantly different from the others.

(d) The more tests we carry out, the higher the global probability of a Type I error (falsely rejecting a null hypothesis that is true).