GM – Probabilités et Statistique – Exercise Set 10

In class

Exercise 1 Literacy rate is a reflection of the educational facilities and quality of education available in a country, and mass communication plays a large part in the educational process. In an effort to relate the literacy rate of a country to various mass communication outlets, a demographer has proposed to relate literacy rate to the following variables : number of daily newspaper copies (per 1000 population), number of radios (per 1000 population), and number of TV sets (per 1000 population). Here are the data for a sample of 10 countries (you may assume that the countries are a random sample), along with the regression output :

Country	newspapers	radios	tv sets	literacy rate
Czech Republic / Slovakia	280	266	228	0.98
Italy	142	230	201	0.93
Kenya	10	114	2	0.25
Norway	391	313	227	0.99
Panama	86	329	82	0.79
Philippines	17	42	11	0.72
Tunisia	21	49	16	0.32
USA	314	1695	472	0.99
Russia	333	430	185	0.99
Venezuela	91	182	89	0.82

Predictor	Coef	SE Coef	T	P
Constant	0.51486	0.09368	5.50	0.002
newspaper copies	0.0005421	0.0008653	0.63	0.554
radios	-0.0003535	0.0003285	-1.08	0.323
television sets	0.001988	0.001550	1.28	0.247

S = 0.186455 R-Sq = 69.9% R-Sq(adj) = 54.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.48397	0.16132	4.64	0.053
Residual Error	6	0.20859	0.03477		
Total	9	0.69256			

(a) What is the response variable ? What are the explanatory variables ?

(b) Write the least-squares regression equation for this problem. Explain what each term in the regression equation represents in terms of the problem.

(c) Interpret the *newspaper copies* coefficient.

(d) Predict literacy rate for a country that has 200 daily newspaper copies (per 1000 in the population), 800 radios (per 1000 in the population), and 250 TV sets (per 1000 in the population).

(e) Determine the lower and upper bounds for the 95% confidence interval for β_3, the coefficient for the variable *television sets*.

Exercise 2 There are data during 25 years on wheat yield y (in tons per hectare) as a function of the annual quantity of rain x_1 (in centimeters) and the mean annual temperature x_2 (in degrees Fahrenheit).

The parameter estimates :

$$\hat{\beta}_0 = 32.051 \quad \hat{\beta}_1 = 0.480 \quad \hat{\beta}_2 = -0.555$$

The diagonal values of the matrix $\mathbf{Z} = (\mathbf{X}'\mathbf{X})^{-1}$:

$$\text{diag}\,(\mathbf{X}'\mathbf{X})^{-1} = (z_{11},\, z_{22},\, z_{33}) = (3.437,\, 0.0014,\, 0.0021)$$

For the full model Ω :

source	df	SS	MS	F	p-value
regression	2	184.2	?	?	0.0020
error	22	243.3	?		
total	?	?			

(a) Write out the complete linear statistical model (Ω).

(b) What is the matrix expression of the ordinary least squares estimator of the vector $\boldsymbol{\beta}$ for the parameters ?

(c) Find $Var(\hat{\boldsymbol{\beta}})$.

(d) Complete the ANOVA table.

(e) Find R^2 and adjusted R^2.
(f) Test $H : \beta_1 = \beta_2 = 0$ at level $\alpha = 5\%$ (make sure to specify A).
(g) Calculate a 95% onfidence interval for the parameter β_2.

Exercise 3 Scores have been collected for four independent groups of ten subjects each. The ANOVA table :

Source	df	SC	CM	F	p
(h)	(a)	(d)	116	(g)	1.8×10^{-5}
Error	(b)	360	(f)		
Total	(c)	(e)			

(a)-(h) Complete the table.

(i) Test the hypothesis that the means of the groups are equal (you may assume that the ANOVA assumptions are satisfied). Use $\alpha = 0.05$.

(j) Under H, how many degrees of freedom does the test statistic have ?

At home

Exercise 1 A data set demonstrates the oxidation of ammonia to produce nitric acid. The matter loss (y = 'stackloss') should be explained by air volume (x_1 = 'Air Flow'), the temperature of the cooling water (x_2 = 'Water Temp'), and the acid concentration (x_2 = 'Acid Conc.').

Consider the two R outputs (next page) representing the estimations for the following two models :

(A) $\Omega : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$,
(B) $\omega : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$

(a) Comment on all the numbers and give detailed interpretations. For the 'p-values', state the hypotheses H and A of the test and the conclusion.
(b) How many observations are there ?
(c) For the model Ω, verify the calculation of the estimated σ, the standard deviation of the model error.
(d) Construct a 95% confidence interval for β_2 in the model Ω.
(e) Carry out the test $H : \omega$ is the true model, using an F-test at level 5% ; what are your conclusions ?

```
Coefficients:
            Estimate Std.  t value Pr(>|t|)
                     Error
(Intercept) -39.9     11.9  -3.4   0.0038
Air Flow      0.7      0.1   5.3   5.8e-05
Water Temp    1.3      0.4   3.5   0.0026
Acid Conc.   -0.2      0.2  -0.97  0.3440
---
Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared: 0.9136,     Adjusted R-squared: 0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09


Analysis of Variance Table

Response: stack.loss
           Df  Sum Sq Mean Sq F value  Pr(>F)
stack.x     3 1890.4  630.1     59.5    3.01633e-09
Residuals  17  178.8   10.5
```

```
Coefficients:
            Estimate Std.  t value Pr(>|t|)
                     Error
(Intercept) -50.4    5.1    -9.8    1.22e-08
Air Flow      0.7    0.1     5.3    4.90e-05
Water Temp    1.3    0.4     3.5    0.0024
---
Residual standard error: 3.239 on 18 degrees of freedom
Multiple R-squared: 0.9088,     Adjusted R-squared: 0.8986
F-statistic: 89.64 on 2 and 18 DF,  p-value: 4.382e-10


Analysis of Variance Table

Response: stack.loss
           Df  Sum Sq Mean Sq F value  Pr(>F)
stack.x     2 1880.4   940.2  89.6     4.38154e-10
Residuals  18  188.8   10.5
```

Exercise 2 The number of grams of fat per serving for three different varieties of pizza from several manufacturers is measured and a partial ANOVA table is provided below :

ANOVA table :

Source	df	SS	MS	F	p
		23.0			0.53
Error					
Total	20	339.9			

(a) Assume that the ANOVA assumptions for a valid test are satisfied. Complete the ANOVA table.

(b) At the $\alpha = 0.05$ significance level, is there a significant difference in mean fat content for the three pizza varieties ?

(c) Does it make sense to make post-hoc (i.e., after the global test) pair-wise comparisons ? Explain.

Exercise 3 Twenty-two patients undergoing aortocoronary bypass have been randomized to one of three ventilation groups :

(i) a mix of 50% nitrous oxide and 50% oxygen, continuously for 24 heures

(ii) a mix of 50% nitrous oxide and 50% oxygen, only during the operation

(iii) no nitrous oxide, but 35-50% oxygen, continuously for 24 hours.

The question of interest is to find out if the average folate of the red blood cells are different for the three ventilation methods. The data are analyzed by ANOVA.

(a) What assumptions must be satisfied to obtain a valid p-value using ANOVA ?

(b) What is the null hypothesis for the ANOVA test ?

(c) Use the ANOVA table below (obtained from the `R` software) in order to determine if the null hypothesis is rejected at a level of 5%. Interpret the result.

(d) Explain why carrying out a single joint test (ANOVA) is preferable to carrying out multiple pairwise tests.

```
> redcell.aov<-aov(Folate~Group)
> summary(redcell.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
Group        2  15516    7758  3.7113 0.04359 *
Residuals   19  39716    2090
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```