Exercises: Multiple Gaussian Linear Model (part 2)

Exercise 1: Exam 27/06/2024

Assume that $y_1, \dots, y_{200}$ are realizations of independent Gaussian random variables with variance equal to 1 and and mean $\beta_1 + \beta_2 \exp\{z_i\}$ for $i = 1, \dots, 120$, and mean $\beta_1 + \beta_3 \exp\{z_i^2\}$ for $i = 121, \dots, 200$; where the z_i are known constants and $(\beta_1, \beta_2, \beta_3)$ are unknown real parameters.

a) Are the assumptions of a Gaussian linear model satisfied in the above formulation? Motivate the answer.

b) State the parameter space and sample space.

c) Express the model in matrix form: $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, explicitly stating how $\boldsymbol{Y}$, X, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ are defined and their dimensions. Write the distribution of $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$.

d) Obtain the expression of the matrix $X^T X$ and the vector $X^T \boldsymbol{y}$; state how these elements should be used to obtain the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

e) Write the distribution of the maximum likelihood estimator $\hat{\boldsymbol{B}}$.

f) Let $\boldsymbol{e} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ be the vector of the residuals. State which of the following identities are satisfied and motivate the answer:

$$\sum_{i=1}^{200} e_i = 0 \qquad \sum_{i=1}^{200} e_i z_i = 0 \qquad \sum_{i=1}^{200} e_i z_i^2 = 0$$

$$\sum_{i=1}^{200} e_i \exp\{z_i\} = 0 \qquad \sum_{i=1}^{200} e_i \exp\{z_i^2\} = 0 \qquad \sum_{i=1}^{120} e_i \exp\{z_i\} = 0$$

(hint: read the indices in the sum!)

Exercise 2

It is measured the weight (in kg) of several trees subjected to 4 different treatments:
(1) no treatment,
(2) fertilizer,
(3) watering,
(4) fertilizer and watering.

Each treatment was applied to 5 trees. The interest is understanding how each treatment affects the trees' weight using a Gaussian linear regression model.

It is known that the sample mean and sample variances of the trees' weight, for each treatment, are:

(1) $\bar{y}_1 = 0.184 \quad (s_1^2 = 0.01613)$,
(2) $\bar{y}_2 = 0.332 \quad (s_2^2 = 0.31922)$,
(3) $\bar{y}_3 = 0.164 \quad (s_3^2 = 0.01587)$,
(4) $\bar{y}_4 = 1.334 \quad (s_4^2 = 0.73788)$.

a) Assume that the observations are sorted according to the treatment and that dummy variables are used to encode the treatments.
Write the statistical model corresponding to the analysis (assuming that the model includes the intercept). Express the model in matrix form:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

explicitly stating $\boldsymbol{Y}$ (and its distribution), X, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ (and its distribution).

b) Write the parameter and sample space.

c) Define and explain the sum of squares decomposition, compute the individual terms.

d) Perform a statistical test for the hypothesis that the type of treatment does not have an effect on the tree weight using a 5% significance level.

e) Obtain the maximum likelihood estimate of $\boldsymbol{\beta}$, and write the expression of the estimated model.

f) Knowing that

$$(X^T X)^{-1} = \begin{bmatrix} 0.20 & -0.20 & -0.20 & -0.20 \\ -0.20 & 0.40 & 0.20 & 0.20 \\ -0.20 & 0.20 & 0.40 & 0.20 \\ -0.20 & 0.20 & 0.20 & 0.40 \end{bmatrix}$$

write the distribution of the maximum likelihood estimator $\hat{\boldsymbol{B}}$ and the marginal distribution of $\hat{B}_2$.

g) Perform a statistical test about the significance of β_2 using a 5% significance level. What is the meaning of this test in the context of the study about the treatments?

h) Compute the coefficient of determination R^2 and explain it.

i) Let $\boldsymbol{e} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ be the vector of the residuals. State which of the following identities are satisfied and motivate the answer:

$$\sum_{i=1}^{20} e_i = 0 \qquad \sum_{i=1}^{5} e_i = 0 \qquad \sum_{i=1}^{10} e_i = 0 \qquad \sum_{i=4}^{12} e_i = 0$$

Exercise 3

The `mtcars` dataset comprises the fuel consumption (`mpg`: Miles/US gallon) and 8 aspects of automobile design and performance for 32 automobiles. Specifically, the covariates are

- `wt`: Weight (1000 lbs)
- `am`: Transmission (0 = automatic, 1 = manual)
- `cyl`: Number of cylinders
- `disp`: Displacement (cu.in.)
- `hp`: Gross horsepower
- `drat`: Rear axle ratio
- `qsec`: 1/4 mile time
- `vs`: Engine (0 = V-shaped, 1 = straight)

Fitting a Gaussian linear model in R produces the following output

	Estimate	Std. Error	t value	Pr(>|t|)
`(Intercept)`	15.5731	16.3817	0.951	0.3517
`wt`	-3.9437	1.2874	-3.063	0.0055
`am1`	2.7937	1.8682	1.495	0.1484
`cyl`	-0.2786	0.9348	-0.298	0.7683
`disp`	0.0147	0.0120	1.223	0.2338
`hp`	-0.0214	0.0162	?	0.1995
`drat`	0.8151	1.5101	0.540	0.5946
`qsec`	?	0.6587	1.229	0.2314
`vs1`	0.3684	2.0116	0.183	?
Residual standard error = 2.544				
$R^2 = 0.8678$				

a) Write the statistical model corresponding to the analysis. Denote this model as "model A".

b) Write the parameter space and sample space.

c) Complete the missing values in the table.

d) Perform a test of overall significance of the model using a 5% significance level.

e) On the same dataset, it is then estimated a reduced model ("model B") that only includes the variables `wt` and `am`. The software estimates the following quantities:

Residual standard error = 3.098
$R^2 = 0.7528$

What procedure would you use to compare model A and model B? Following your chosen procedure, which model do you prefer?

f) Starting from model B, it is then introduced as an additional covariate the interaction between `wt` and `am`. Write the model formulation and provide the interpretation of the parameters.

	p						
	0.90	0.95	0.975	0.99	0.995	0.9975	0.999
z_p	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902

Table 1: Some quantiles of the Gaussian distribution: $p = \mathbb{P}(Z \leq z_p)$. Columns correspond to probabilities p.

	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
$t_{16;p}$	1.3368	1.7459	2.1199	2.5835	2.9208	3.252	3.6862
$t_{17;p}$	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458
$t_{18;p}$	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105
$t_{19;p}$	1.3277	1.7291	2.093	2.5395	2.8609	3.1737	3.5794
$t_{20;p}$	1.3253	1.7247	2.086	2.528	2.8453	3.1534	3.5518
$t_{21;p}$	1.3232	1.7207	2.0796	2.5176	2.8314	3.1352	3.5272
$t_{22;p}$	1.3212	1.7171	2.0739	2.5083	2.8188	3.1188	3.505
$t_{23;p}$	1.3195	1.7139	2.0687	2.4999	2.8073	3.104	3.485
$t_{24;p}$	1.3178	1.7109	2.0639	2.4922	2.7969	3.0905	3.4668

Table 2: Some quantiles of the t distribution: $p = \mathbb{P}(T \leq t_{\alpha;p})$ with $T \sim t_\alpha$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom α.

	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
$f_{3,16;p}$	2.4618	3.2389	4.0768	5.2922	6.3034	7.4027	9.0059
$f_{6,16;p}$	2.1783	2.7413	3.3406	4.2016	4.9134	5.6843	6.8049
$f_{8,16;p}$	2.088	2.5911	3.1248	3.8896	4.5207	5.2034	6.195
$f_{9,16;p}$	2.0553	2.5377	3.0488	3.7804	4.3838	5.0364	5.9839
$f_{3,22;p}$	2.3512	3.0491	3.7829	4.8166	5.6524	6.5391	7.796
$f_{6,22;p}$	2.0605	2.5491	3.0546	3.7583	4.3225	4.9178	5.758
$f_{8,22;p}$	1.9668	2.3965	2.8392	3.453	3.944	4.4612	5.1901
$f_{9,22;p}$	1.9327	2.3419	2.7628	3.3458	3.8116	4.3021	4.9929
$f_{3,23;p}$	2.3387	3.028	3.7505	4.7649	5.5823	6.447	7.6688
$f_{6,23;p}$	2.0472	2.5277	3.0232	3.7102	4.2591	4.8366	5.6486
$f_{8,23;p}$	1.9531	2.3748	2.8077	3.4057	3.8822	4.3826	5.0853
$f_{9,23;p}$	1.9189	2.3201	2.7313	3.2986	3.7502	4.2243	4.8896

Table 3: Some quantiles of the F distribution: $p = \mathbb{P}(F \leq f_{df_1,df_2;p})$ with $F \sim F_{df_1,df_2}$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom (df_1, df_2).