EXERCISE 5

26TH November 2024

Luca Danese - l.danese1@campus.unimib.it

EXERCISE 1

1.1)

Among $n = 15$ poisoned lab rats, data about survival time (Y) and antidote/treatment (T) were collected. The researchers consider three kind of treatment: A, B and C, where

Group	n_j	$\bar{y}_j$	s_j
A ($j = 1$)	4	0.347	0.2055
B ($j = 2$)	3	3.067	0.4618
C ($j = 3$)	8	1.764	0.5007

Compute the total deviance (total sum of squares - SST) and explain its components.

Perform a statistical test to evaluate if the means (of each group) are homogeneous. Specify the system of hypothesis, the test statistic and the p-value.

$$\underbrace{\sum_{j=1}^{3}\sum_{i=1}^{n_j}(y_i - \bar{y})^2}_{\text{TOTAL SUM OF SQUARES}} = \underbrace{\sum_{j=1}^{3}(n_j - 1)s_j^2}_{\text{WITHIN GROUP VARIABILITY}} + \underbrace{\sum_{j=1}^{3} n_j(\bar{y}_j - \bar{y})^2}_{\text{BETWEEN GROUP VARIABILITY}}$$

$$\bar{y} = \frac{4 \cdot 0.347 + 3 \cdot 3.067 + 8 \cdot 1.764}{15} = 1.646733$$

$$\sum_{j=1}^{3}(n_j - 1)s_j^2 = (4-1)\, 0.2055^2 + (3-1) \cdot 0.4618^2 + (8-1) \cdot 0.5007^2 = 2.308113$$

$$\sum n_j(\bar{y}_j - \bar{y})^2 = 4(0.347 - 1.646733)^2 + 3(3.067 - 1.646733)^2 + 8(1.764 - 1.646733)^2$$

$$= 12.91871$$

$$SST = 2.308113 + 12.91871 = 15.22682$$

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \exists_j \text{ s.t. } \mu_j \neq \mu_i \ (i \neq j,\ i,j \in \{1,2,3\}) \end{cases}$$

$$F = \frac{SSR/(s-1)}{SSE/(n-s)} \overset{H_0}{\sim} F_{s-1,\,n-s}$$

$$f^{obs} = \frac{12.92/2}{2.308/12} = 33.59$$

$$\alpha^{obs} = P\left(F_{2,12} > 33.59\right) \approx 0 \implies \text{we reject } H_0$$

	0.6	0.7	0.8	0.85	0.9	0.95	0.98	0.99	0.995
$f_{1,8;p}$	0.7901	1.228	1.9511	2.5352	3.4579	5.3177	8.3895	11.2586	14.6882
$f_{2,8;p}$	1.0297	1.4048	1.9814	2.4274	3.1131	4.459	6.6366	8.6491	11.0424
$f_{1,9;p}$	0.7804	1.2094	1.9128	2.4766	3.3603	5.1174	7.9605	10.5614	13.6136
$f_{2,9;p}$	1.0162	1.3804	1.9349	2.3597	3.0065	4.2565	6.234	8.0215	10.1067
$f_{1,10;p}$	0.7727	1.1948	1.8829	2.4312	3.285	4.9646	7.6384	10.0443	12.8265
$f_{2,10;p}$	1.0056	1.3613	1.8986	2.3072	2.9245	4.1028	5.9336	7.5594	9.427
$f_{1,11;p}$	0.7666	1.183	1.8589	2.3949	3.2252	4.8443	7.388	9.646	12.2263
$f_{2,11;p}$	0.997	1.3459	1.8697	2.2654	2.8595	3.9823	5.7012	7.2057	8.9122
$f_{1,12;p}$	0.7614	1.1733	1.8393	2.3653	3.1765	4.7472	7.1878	9.3302	11.7542
$f_{2,12;p}$	0.99	1.3333	1.846	2.2313	2.8068	3.8853	5.5163	6.9266	8.5096
$f_{1,13;p}$	0.7572	1.1653	1.823	2.3407	3.1362	4.6672	7.0241	9.0738	11.3735
$f_{2,13;p}$	0.984	1.3227	1.8262	2.203	2.7632	3.8056	5.3657	6.701	8.1865
$f_{1,14;p}$	0.7535	1.1584	1.8091	2.3198	3.1022	4.6001	6.888	8.8616	11.0603
$f_{2,14;p}$	0.979	1.3137	1.8095	2.1791	2.7265	3.7389	5.2408	6.5149	7.9216
$f_{1,15;p}$	0.7504	1.1525	1.7972	2.302	3.0732	4.5431	6.7729	8.6831	10.798
$f_{2,15;p}$	0.9746	1.306	1.7952	2.1586	2.6952	3.6823	5.1354	6.3589	7.7008
$f_{1,16;p}$	0.7476	1.1473	1.7869	2.2865	3.0481	4.494	6.6744	8.531	10.5755
$f_{2,16;p}$	0.9708	1.2993	1.7828	2.1409	2.6682	3.6337	5.0455	6.2262	7.5138
$f_{1,17;p}$	0.7453	1.1428	1.7779	2.273	3.0262	4.4513	6.5892	8.3997	10.3842
$f_{2,17;p}$	0.9675	1.2934	1.7719	2.1255	2.6446	3.5915	4.9678	6.1121	7.3536
$f_{1,18;p}$	0.7431	1.1389	1.7699	2.2611	3.007	4.4139	6.5146	8.2854	10.2181
$f_{2,18;p}$	0.9646	1.2882	1.7623	2.1119	2.6239	3.5546	4.9001	6.0129	7.2148

1.2)

Specify an appropriate linear regression model. Given the previous data, find the estimates for each coefficient and interpret them.

$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, where

- $X = [\underline{1} \quad \underline{X}_2 \quad \underline{X}_3] = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ \vdots & 0 & \vdots \\ \vdots & 1 & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & 1 & 0 \\ \vdots & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}_{15\times 3}$ $\qquad$ - $\underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$

- $\underline{\varepsilon} \sim N_{15}(\underline{0}, I_{15}\cdot\sigma^2)$

$Y_i \sim N(\beta_1, \sigma^2) \quad i = 1, \dots, 4 \quad$ (group A)

$Y_i \sim N(\beta_1 + \beta_2, \sigma^2) \quad i = 5, \dots, 7 \quad$ (group B)

$Y_i \sim N(\beta_1 + \beta_3, \sigma^2) \quad i = 8, \dots, 15 \quad$ (group C)

$$\mathbb{E}[Y_{iA}] = \beta_1 \qquad \beta_1 = \mathbb{E}[Y_{iA}] = \bar{y}_1$$

$$\mathbb{E}[Y_{iB}] = \beta_1 + \beta_2 \quad \Rightarrow \quad \beta_2 = \mathbb{E}[Y_{iB}] - \mathbb{E}[Y_{iA}] = \bar{y}_2 - \bar{y}_1$$

$$\mathbb{E}[Y_{iC}] = \beta_1 + \beta_3 \qquad \beta_3 = \mathbb{E}[Y_{iC}] - \mathbb{E}[Y_{iA}] = \bar{y}_3 - \bar{y}_1$$

$$\hat{\beta}_1 = \bar{Y}_1 = 0.347 \quad \hat{\beta}_2 = \bar{Y}_2 - \bar{Y}_1 = 2.72 \quad \hat{\beta}_3 = \bar{Y}_3 - \bar{Y}_1 = 1.417$$

$$\Rightarrow \hat{Y}_i = 0.347 + 2.72\, X_{i2} + 1.417\, X_{i3}$$

where

$$X_{i2} = \begin{cases} 1 & \text{if } i \text{ in group B} \\ 0 & \text{otherwise} \end{cases} \qquad X_{i3} = \begin{cases} 1 & \text{if } i \text{ in group C} \\ 0 & \text{otherwise} \end{cases}$$

- $\hat{\beta}_1 = 0.347 \Rightarrow$ The average survival time in group A is equal to 0.347

- $\hat{\beta}_2 = 2.72 \Rightarrow$ The average increase in survival time when moving from group A to group B is 2.72

- $\hat{\beta}_3 = 1.417 \Rightarrow$ The average increase in survival time when moving from group A to group C is 1.417

1.3)

Propose an alternative test for ex. 1.1 specifying the system of hypothesis. Then, compute R^2 of our model.

$$\begin{cases} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \text{at least one } \beta_j \neq 0 \end{cases}$$

$$M_0: \; Y_i = \beta_1 + \varepsilon_i$$

$$M_2: \; Y_i = \beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + \varepsilon_i$$

$\tilde{\sigma}^2$: estimated variance in H_0 $= 1/n \sum (Y_i - \bar{Y})^2 = SST/n = 15.23/15 = 1.015$

$\hat{\sigma}^2$: estimated variance in M_2 $= 1/n \sum (Y_i - \hat{Y})^2 = SSE/n = 2.308/15 = 0.1539$

$$F = \frac{(\tilde{\Sigma}^2 - \hat{\Sigma}^2)}{\hat{\Sigma}^2} \cdot \frac{N-5}{5-1} = \frac{\overbrace{(SST - SSE)}^{SSR}}{SSE} \cdot \frac{N-5}{5-1} = \left(\frac{1.015 - 0.1539}{0.1539}\right) \frac{15-3}{3-1} = 33.57 = f^{obs}$$

$\Rightarrow$ we reject H_0 (see ex. 1.1)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{2.308}{15.23} = 0.84$$

The model explains 84% of the variability

EXERCISE 2

Lab *Precise* got some measurements to see whether the tar contents (in milligrams) for three different brands of cigarettes are different. The measurements are showed in the following table.

Sample	Brand A	Brand B	Brand C
1	10.21	11.32	11.60
2	10.25	11.20	11.90
3	10.24	11.40	11.80
4	9.80	10.50	12.30
5	9.77	10.68	12.20
6	9.73	10.90	12.20

Write the equation of the linear regression model. Find and interpret the estimates of regression coefficients.

2.1)

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2) \quad i = 1, \dots, 18$$

$$x_{i2} = \begin{cases} 1 & \text{if brand = "B"} \\ 0 & \text{otherwise} \end{cases} \qquad x_{i3} = \begin{cases} 1 & \text{if brand = "C"} \\ 0 & \text{otherwise} \end{cases}$$

$\bar{y}_A = 10 \quad \bar{y}_B = 11 \quad \bar{y}_C = 12$

$\hat{\beta}_1 = \bar{y}_A = 10 \quad \hat{\beta}_2 = \bar{y}_B - \bar{y}_A = 11 - 10 = 1 \quad \hat{\beta}_3 = \bar{y}_C - \bar{y}_A = 12 - 10 = 2$

- The average tar content in group A is equal to 10
- The average increase from group A to group B is equal to 1
- The average increase from group A to group C is equal to 2

2.2)

Complete the following table.

Sources of variation	**D.o.f.**	**Deviance**
Total	17	13.37
Regression	2	12
Residual	15	1.37

Perform the statistical test to evaluate if the means (of each group) are homogeneous.

$$SSR = \sum_{j=1}^{3} n_j (\bar{y}_j - \bar{y})^2 = \dots = 12$$

$$SSE = \sum_{j=1}^{3} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 = \dots = 1.37$$

$$SST = 12 + 1.37 = 13.37$$

$Dof(SST) = n-1 = 17$ $\quad Dof(SSE) = n-J = 15$ $\quad Dof(SSR) = J-1 = 2$

$$\begin{cases} H_0: \mu_A = \mu_B = \mu_C \\ H_1: \exists_j \text{ s.t. } \mu_j = \mu_i \ (i \neq j, \ i,j \in \{1,2,3\}) \end{cases}$$

$$F = \frac{SSR \setminus (J-1)}{SSE \setminus (n-J)} \qquad f^{obs} = \frac{12 \setminus 2}{1.3748 \setminus 15} = 65.46407$$

$\alpha^{obs} = P(F_{2,15} > 65.464407) \approx 0$ => we reject the null hypothesis

	0.6	0.7	0.8	0.85	0.9	0.95	0.98	0.99	0.995
$f_{1,8;p}$	0.7901	1.228	1.9511	2.5352	3.4579	5.3177	8.3895	11.2586	14.6882
$f_{2,8;p}$	1.0297	1.4048	1.9814	2.4274	3.1131	4.459	6.6366	8.6491	11.0424
$f_{1,9;p}$	0.7804	1.2094	1.9128	2.4766	3.3603	5.1174	7.9605	10.5614	13.6136
$f_{2,9;p}$	1.0162	1.3804	1.9349	2.3597	3.0065	4.2565	6.234	8.0215	10.1067
$f_{1,10;p}$	0.7727	1.1948	1.8829	2.4312	3.285	4.9646	7.6384	10.0443	12.8265
$f_{2,10;p}$	1.0056	1.3613	1.8986	2.3072	2.9245	4.1028	5.9336	7.5594	9.427
$f_{1,11;p}$	0.7666	1.183	1.8589	2.3949	3.2252	4.8443	7.388	9.646	12.2263
$f_{2,11;p}$	0.997	1.3459	1.8697	2.2654	2.8595	3.9823	5.7012	7.2057	8.9122
$f_{1,12;p}$	0.7614	1.1733	1.8393	2.3653	3.1765	4.7472	7.1878	9.3302	11.7542
$f_{2,12;p}$	0.99	1.3333	1.846	2.2313	2.8068	3.8853	5.5163	6.9266	8.5096
$f_{1,13;p}$	0.7572	1.1653	1.823	2.3407	3.1362	4.6672	7.0241	9.0738	11.3735
$f_{2,13;p}$	0.984	1.3227	1.8262	2.203	2.7632	3.8056	5.3657	6.701	8.1865
$f_{1,14;p}$	0.7535	1.1584	1.8091	2.3198	3.1022	4.6001	6.888	8.8616	11.0603
$f_{2,14;p}$	0.979	1.3137	1.8095	2.1791	2.7265	3.7389	5.2408	6.5149	7.9216
$f_{1,15;p}$	0.7504	1.1525	1.7972	2.302	3.0732	4.5431	6.7729	8.6831	10.798
$f_{2,15;p}$	0.9746	1.306	1.7952	2.1586	2.6952	3.6823	5.1354	6.3589	7.7008
$f_{1,16;p}$	0.7476	1.1473	1.7869	2.2865	3.0481	4.494	6.6744	8.531	10.5755
$f_{2,16;p}$	0.9708	1.2993	1.7828	2.1409	2.6682	3.6337	5.0455	6.2262	7.5138
$f_{1,17;p}$	0.7453	1.1428	1.7779	2.273	3.0262	4.4513	6.5892	8.3997	10.3842
$f_{2,17;p}$	0.9675	1.2934	1.7719	2.1255	2.6446	3.5915	4.9678	6.1121	7.3536
$f_{1,18;p}$	0.7431	1.1389	1.7699	2.2611	3.007	4.4139	6.5146	8.2854	10.2181
$f_{2,18;p}$	0.9646	1.2882	1.7623	2.1119	2.6239	3.5546	4.9001	6.0129	7.2148