Level M Regression Models Examples

Blood Pressure

A clinical trial has been carried out to compare three drug treatments which are intended to lower blood pressure in hypertensive patients. Table below gives the initial values of systolic blood pressure (bp) in mmHg for each patient, and the ensuing reduction achieved during the course of the trial. For each patient, allocation to treatment was carried out randomly and conditions such as the length of the treatment and dose of the drug were standardized as far as possible.

Initial bp	**Reduction**	**Drug**
158	4	1
176	21	1
174	36	1
168	14	1
174	34	1
186	37	1
150	25	1
146	15	1
142	7	2
155	24	2
168	44	2
170	52	2
175	43	2
144	26	2
166	36	2
136	-6	2
146	24	2
140	10	2
163	2	3
163	0	3
173	35	3
167	32	3
174	29	3
191	27	3
160	10	3

Table 1: Blood pressure drug treatments

For these data the model

$$Y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.}) + \epsilon_{ij}$$

is appropriate where y_{ij} denotes the reduction in blood pressure and x_{ij} denotes the initial blood pressure. Three drug treatments are compared and hence $i = 1, 2, 3; j = 1, ..., n_i$.

Write this model in vector-matrix form, $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and use standard vector-matrix formulae for linear models (i.e. $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$) to show that the estimators for the parameter estimates are:

$$\hat{\alpha}_i = \bar{y}_{i.}$$
$$\hat{\beta}_i = \frac{S_{x_i y_i}}{S_{x_i x_i}}$$

Now consider the model with a common slope parameter

$$E(Y_{ij}) = \alpha_i + \beta(x_{ij} - \bar{x}_{i.})$$

and show by similar means that the estimators for the parameter estimates are:

$$\hat{\alpha}_i = \bar{y}_{i.}$$
$$\hat{\beta} = \frac{S_{x_1 y_1} + S_{x_2 y_2} + S_{x_3 y_3}}{S_{x_1 x_1} + S_{x_2 x_2} + S_{x_3 x_3}}$$

Trout

When the behaviour of a group of trout is studied, some fish are observed to become dominant and others to become subordinate. Dominant fish (1) have freedom of movement whereas subordinate fish (2) tend to congregate in the periphery of the waterway to avoid crossing the path of the dominant fish. Energy expenditure and ration of food obtained were collected as part of a laboratory experiment for 10 trout in each group (grp 1 dominant, grp 2 subordinate). Interest lies in how the ration of food obtained depends on the energy expenditure and group the fish belongs to.

Initially a model with two completely separate regression lines was fitted and the `R` output obtained is displayed below:

```
trout <- read.csv("TROUT.csv")
trout$Group=factor(trout$Group)
anova(lm(Ration~Energy*Group, data=trout))
```

```
## Analysis of Variance Table
##
## Response: Ration
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Energy        1 2444.01 2444.01 42.8622 6.724e-06 ***
## Group         1 1415.22 1415.22 24.8197 0.0001357 ***
## Energy:Group  1  162.92  162.92  2.8573 0.1103471
## Residuals    16  912.32   57.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(Ration~Energy+Group, data=trout))
```

```
## Analysis of Variance Table
##
## Response: Ration
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Energy     1 2444.0 2444.01  38.641 9.395e-06 ***
```

```
## Group       1 1415.2 1415.22  22.375 0.0001934 ***
## Residuals 17 1075.2   63.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Explain what the `R` output from fitting these two models tells you about the relationship between ration of food obtained and energy expenditure for the two groups of fish.

Agriculture

In an agricultural experiment, the yield (y) in gm/plant of a crop was studied for a variety of densities of planting (x), in plants/m^2. The experiment was carried out in two different localities. Twenty plots of ground were used in each locality. You may assume that the model described below, which fits two separate regression lines to the data from the two localities, provides a good description of the data,

$$Y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.}) + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2),\ \epsilon_{ij}'\text{s independent}$$

where $i = 1, 2$ refers to locality and $j = 1, ..., 20$ refers to observations within locality.

1. Write this model in vector-matrix form, $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the parameter vector defined by $\boldsymbol{\beta}^T = (\alpha_1, \beta_1, \alpha_2, \beta_2)$. Identify clearly the elements of the matrix $\mathbf{X}$.

2. Use standard vector-matrix formulae for linear models (i.e. for $\hat{\boldsymbol{\beta}}$) to show that the least-squares estimators for the parameters of this model are:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y}_{1.} \\ S_{x_1y_1}/S_{x_1x_1} \\ \bar{y}_{2.} \\ S_{x_2y_2}/S_{x_2x_2} \end{pmatrix}$$

3. There is interest in assessing whether there are differences between the slopes of the two regression lines. The following estimates and summary statistics are relevant.

$$\begin{array}{ll} \hat{\beta}_1 = -0.1365 & \sum_{j=1}^{20}(x_{1j} - \bar{x}_{1.})^2 = S_{x_1x_1} = 1090.533 \\ \hat{\beta}_2 = -0.1894 & \sum_{j=1}^{20}(x_{2j} - \bar{x}_{2.})^2 = S_{x_2x_2} = 813.903 \\ \multicolumn{2}{l}{RSS/(n-p) = (RSS_1 + RSS_2)/(n_1 + n_2 - 4) = 1.264} \end{array}$$

   The quantity $(\beta_1 - \beta_2)$ can be written as $\mathbf{b}^T\boldsymbol{\beta}$, where $\mathbf{b}^T = (0, 1, 0, -1)$. Show that $\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}$ can be written as $\{1/S_{x_1x_1} + 1/S_{x_2x_2}\}$. Hence calculate a 95% confidence interval for $(\beta_1 - \beta_2)$. Interpret this interval.

4. Write down the model which corresponds to fitting two parallel lines to the data. Assuming that this model provides a good description of the data, state the quantity for which you would construct an interval estimate in order to assess whether there are any differences at all between the two localities. (An expression for this interval is not required).

Treatment

In order to study the side effects of a drug used in the treatment of a psychiatric illness, 30 patients were studied. These patients had received the drug at different doses (x) in mgml^{-1}. A visual detection test was given to each patient and the score achieved (y) was recorded. 15 of the patients were male, and 15 were female. The data can therefore be represented as, $i = 1, 2; j = 1, ..., 15$, where i denotes sex and j denotes the patients. You may assume that the following model provides a good description of the data:

$$\text{Model 1} : Y_{ij} = \alpha_i + \beta(x_{ij} - \bar{x}_{i.}) + \epsilon_{ij} \tag{1}$$

$$\epsilon_{ij} \sim N(0, \sigma^2),\ \epsilon_{ij}\text{'s independent}$$

1. Explain what assumption Model (1) makes about the differences between the regression line of y on x for males and females. Write down a model which does not have this assumption. Write down the quantity for which an interval estimate would allow you to check this assumption.

2. Write Model (1) in the vector-matrix form $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the parameter vector defined by $\boldsymbol{\beta}^T = (\alpha_1, \alpha_2, \beta)$. Identify clearly the elements of the matrix $\mathbf{X}$.

3. Use standard vector-matrix formulae for linear models (i.e. for $\hat{\boldsymbol{\beta}}$) to show that the least-squares estimators for the parameters of Model (1) are:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y}_{1.} \\ \bar{y}_{2.} \\ \frac{S_{x_1y_1} + S_{x_2y_2}}{S_{x_1x_1} + S_{x_2x_2}} \end{pmatrix}$$

4. Write down an expression for the difference between the two regression lines in model (1). Use the summary statistics and other information below to construct a 95% confidence interval for this quantity. Comment on what the confidence interval tells you about the difference between the regression lines for males and females.

$$\begin{aligned} \bar{y}_{1.} &= 3.9180 & \bar{y}_{2.} &= 3.9872 \\ \bar{x}_{1.} &= 10.6873 & \bar{x}_{2.} &= 11.8261 \\ S_{x_1x_1} &= 502.6836 & S_{x_2x_2} &= 312.1848 \\ S_{x_1y_1} &= 102.8271 & S_{x_2y_2} &= 92.7053 \\ RSS &= 56.1783 \end{aligned}$$

$$\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} = 0.1349 \quad \text{where } \mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ \{\bar{x}_{2.} - \bar{x}_{1.}\} \end{pmatrix}$$

Anti-hypertensive drugs

In an investigation of two anti-hypertensive drugs, 10 patients were assigned to drug A and 10 patients assigned to drug B. The final blood pressure of patient j on drug i is denoted Y_{ij} and their initial blood pressure is denoted by x_{ij}.

Consider the Normal Linear Model.

$$Y_{ij} = \alpha_i + \beta(x_{ij} - \bar{x}_{i.}) + \epsilon_{ij};\ i = 1,2; j = 1,...,10,\ \epsilon_{ij} \sim N(0,\sigma^2) \text{ and } \epsilon_{ij} \text{ independent.}$$

1. In words briefly describe this model.
2. Write down a vector-matrix expression for this model clearly identifying $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$. Hence find the least squares estimates for $\boldsymbol{\beta}$.
3. The difference between the regression lines corresponding to the two groups can be written as

$$\alpha_1 - \alpha_2 + \beta(\bar{x}_{2.} - \bar{x}_{1.})$$

   Identify the vector $\mathbf{b}$ which allows this expression to be written as $\mathbf{b}^T\boldsymbol{\beta}$ and hence derive the specific form of a 95% confidence interval for the difference between the two regression lines.