

Chapter 3 Other Issues in Multiple regression

(Part 5. Application: Single-factor study)

1 Overview

- factors: variables (usually discrete or categorical variable) that possibly affect the response
- treatments: factor levels (values of explanatory variable) and interaction of different factors

2 Single-factor ANOVA model

Basic ideas

For each factor level, there is a probability distribution of the response. We need the following assumptions

- Each probability distribution is normal
- each probability distribution have the same variance
- the response for each factor level are random selection

We hope to be able to

- Determine whether or nor the factor level means (the mean in each probability distribution) are the same
- if the factor level means are different, examine how they differ and what implications of the difference are.

Notations

- number of levels/groups r ,
- number of cases for the i th factor level/group, $n_i, i = 1, \dots, r$

- total number of cases/individuals n ,

$$n = n_1 + \dots + n_r$$

- Y_{ij} denote the response of j th case (individual) in i th treatment (factor level/group).
Since we assume the individuals from i th treatment are IID normal, we have

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

for $j = 1, \dots, n_i$, or

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where μ_i is the mean in the i th factor level/group.

Important feature of the model

- Repeated observations: for each value of the factor, denoted also by X , we have a number of observations (they form a group: Y_{i1}, \dots, Y_{in_i})
- $Y_{ij} \sim N(\mu_i, \sigma^2)$

The ANOVA model is a linear model

Denote the factor by X , our NONlinear model is

$$Y = g(X) + \varepsilon$$

Suppose the factor X ONLY takes values " a_1 ", ..., " a_r ". So we need to estimate the function values EY

$$g("a_1"), \dots, g("a_r")$$

denoted by

$$\mu_1, \dots, \mu_r$$

By introducing Dummy variables

$$D_i = \begin{cases} 1, & \text{if an observation have value "a}_i\text{"} \\ 0, & \text{otherwise} \end{cases}$$

$$i = 1, \dots, r$$

We have a linear regression model

$$Y = \beta_1 D_1 + \dots + \beta_r D_r + \varepsilon$$

or

$$\begin{aligned}
 Y_{11} &= \beta_1 D_{11,1} + \dots + \beta_r D_{11,r} + \varepsilon_{11} \\
 &\vdots \\
 Y_{1n_1} &= \beta_1 D_{1n_1,1} + \dots + \beta_r D_{1n_1,r} + \varepsilon_{1n_1} \\
 Y_{21} &= \beta_1 D_{21,1} + \dots + \beta_r D_{21,r} + \varepsilon_{21} \\
 &\vdots \\
 Y_{2n_2} &= \beta_1 D_{2n_2,1} + \dots + \beta_r D_{2n_2,r} + \varepsilon_{2n_2} \\
 &\vdots \\
 Y_{r1} &= \beta_1 D_{r1,1} + \dots + \beta_r D_{r1,r} + \varepsilon_{r1} \\
 &\vdots \\
 Y_{rn_r} &= \beta_1 D_{rn_r,1} + \dots + \beta_r D_{rn_r,r} + \varepsilon_{rn_r}
 \end{aligned}$$

Then

$$g("a_i") = \beta_i, \quad i = 1, \dots, r.$$

The design matrix

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{r1} \\ \vdots \\ Y_{rn_r} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_r \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{r1} \\ \vdots \\ \varepsilon_{rn_r} \end{pmatrix},$$

The model can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \mathcal{E}.$$

How about using r-1 dummy variables? the model is

$$Y = \beta_0 + \beta_1 D_1 + \dots + \beta_{r-1} D_{r-1} + \varepsilon$$

Then

$$g("a_i") = \beta_0 + \beta_i, \quad i = 1, \dots, r-1.$$

and

$$g("a_r") = \beta_0.$$

3 notations and estimation

- For a group (observations with the same factor level, treatment)

$$Y_{i.} = Y_{i1} + \dots + Y_{in_i} = \sum_{j=1}^{n_i} Y_{ij}$$

and group mean

$$\bar{Y}_{i.} = (Y_{i1} + \dots + Y_{in_i})/n_i$$

- all observations,

$$Y_{..} = \sum_{i=1}^r \sum_{j=1}^{n_i} Y_{ij}$$

and the overall mean

$$\bar{Y}_{..} = Y_{..}/n = \sum_{i=1}^r \frac{n_i}{n} \bar{Y}_{i.}$$

- For each group (treatment), the mean is estimated as

$$\hat{\mu}_i = \bar{Y}_{i.}$$

- the prediction/fitted error is

$$e_{ij} = Y_{ij} - \bar{Y}_{i.}$$

We have

$$\sum_{j=1}^{n_i} e_{ij} = 0$$

and

$$\sum_{i=1}^r \sum_{j=1}^{n_i} e_{ij} = 0$$

4 Analysis of Variance

Decomposition

$$\underbrace{(Y_{ij} - \bar{Y}_{..})}_{\text{Total deviation}} = \underbrace{(\bar{Y}_{i.} - \bar{Y}_{..})}_{\text{Deviation of estimated factor level mean around overall mean}} + \underbrace{(Y_{ij} - \bar{Y}_{i.})}_{\text{Deviation around estimated factor level mean}}$$

Square both sides above, we have

$$\underbrace{\sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2}_{\text{Total sum of squares}} = \underbrace{\sum_{i=1}^r n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{\text{treatment sum of squares or between treatment}} + \underbrace{\sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2}_{\text{error sum of squares or within treatment}}$$

$$SST = SSTR + SSE$$

Interpretation

- SSE: A measure of the random variation of the observations around the respective estimated factor level means. The less variation among the observations for each factor level, the smaller is SSE . If $SSE = 0$, the observations for any given factor level are all the same, and this hold for all the factor levels.
- SSTR: A measure of the extend of differences between the estimated factor level means, based on the deviation of the estimated factor level means \bar{Y}_i around the overall mean $\bar{Y}_{..}$. If all the estimated level means \bar{Y}_i are the same, then $SSTR = 0$.

A Proof for the decomposition

Please try to prove it.

Breakdown of degrees of freedom

$$SST = \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 \quad : \quad n - 1$$

$$SSTR = \sum_{i=1}^r n_i (\bar{Y}_i - \bar{Y}_{..})^2 \quad : \quad r - 1.$$

$$SSE = \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \quad : \quad (n_1 - 1) + \dots + (n_r - 1) = n - r.$$

Mean squares

$$MSTR = \frac{SSTR}{r - 1}$$

$$MSE = \frac{SSE}{n - r}$$

Expected mean Square

$$E(MSE) = \sigma^2$$

$$E(MSTR) = \sigma^2 + \frac{\sum_{i=1}^r n_i (\mu_i - \mu_{..})^2}{r - 1}$$

where

$$\mu_{..} = \frac{\sum_{i=1}^r n_i \mu_i}{n}$$

5 F-test for equality of Factor level means

Hypothesis

$$H_0 : \mu_1 = \dots = \mu_r, \quad H_a : \text{not all } \mu_i \text{ are equal}$$

Consider

$$F = \frac{MSTR}{MSE}$$

Under H_0 , F should be small. F follows $F(r-1, n-r)$

with significant level α , if

If $F^* \leq F(1 - \alpha, r - 1, n - r)$, accept H_0

If $F^* > F(1 - \alpha, r - 1, n - r)$, accept H_a

ANOVA

Source of variation	SS	df	MS	F-value
between treatment	$SSTR = \sum_{i=1}^r n_i (\bar{Y}_i - \bar{Y}_{..})^2$	r-1	$MSTR = \frac{SSTR}{r-1}$	$F^* = \frac{MSTR}{MSE}$
Error (within treatment)	$SSE = \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	n-r	$MSE = \frac{SSE}{n-r}$	
Total	$\sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$	n-1		

6 Alternative Model: factor effect model

Let

$$\mu_{.} = \sum_{i=1}^r \frac{n_i}{n} \mu_i$$

and

$$\mu_i = \mu_{.} + (\mu_i - \mu_{.}) \equiv \mu_{.} + \tau_i$$

τ_i is called the i th factor level/treatment/group effect. Then the ANOVA model is

$$Y_{ij} = \mu_{.} + \tau_i + \varepsilon_{ij}$$

$$Y_{ij} = \mu_{.} + \tau_i + \varepsilon_{ij}$$

where

$\mu_{.}$ is a constant component common to all observations

τ_i is the effect of the i th factor level/treatment/group

ε_{ij} are IID $N(0, \sigma^2)$

$i = 1, \dots, r; j = 1, \dots, n_i$

It is easy to see that

$$\sum_{i=1}^r \frac{n_i}{n} \tau_i = 0 \quad \text{and} \quad \tau_r = - \sum_{i=1}^{r-1} \frac{n_i}{n_r} \tau_i$$

In other words

$$\mu_i = \mu. + \tau_i, \quad i = 1, \dots, r - 1$$

and

$$\mu_r = \mu. + \tau_r = \mu. - \sum_{i=1}^{r-1} \frac{n_i}{n_r} \tau_i$$

The above test is equivalent to

$$H_0 : \tau_1 = \dots = \tau_r = 0, \quad H_a : \text{not all } \tau_i \text{ equal } 0$$

or

$$H_0 : \tau_1 = \dots = \tau_{r-1} = 0, \quad H_a : \text{not all } \tau_i \text{ equal } 0$$

7 regression approach to the ANOVA

Based on the above model, consider

$$Y_{ij} = \mu. + \tau_1 X_{ij,1} + \dots + \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij}$$

where

$$X_{ij,1} = \begin{cases} 1 & \text{if the case is from factor level 1} \\ -\frac{n_1}{n_r}, & \text{if the case is from factor level r} \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$X_{ij,r-1} = \begin{cases} 1 & \text{if the case is from factor level r-1} \\ -\frac{n_{r-1}}{n_r}, & \text{if the case is from factor level r} \\ 0 & \text{otherwise} \end{cases}$$

Then, the design matrix is

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -\frac{n_1}{n_r} & -\frac{n_2}{n_r} & \dots & -\frac{n_{r-1}}{n_r} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -\frac{n_1}{n_r} & -\frac{n_2}{n_r} & \dots & -\frac{n_{r-1}}{n_r} \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu. \\ \tau_1 \\ \vdots \\ \tau_{r-1} \end{pmatrix},$$

and

$$\mathbf{X}\beta = \begin{pmatrix} \mu. + \tau_1 \\ \vdots \\ \mu. + \tau_1 \\ \mu. + \tau_2 \\ \vdots \\ \mu. + \tau_2 \\ \vdots \\ \mu. - \sum_{i=1}^{r-1} \frac{n_i}{n_r} \tau_i \\ \vdots \\ \mu. - \sum_{i=1}^{r-1} \frac{n_i}{n_r} \tau_i \end{pmatrix} = \begin{pmatrix} \mu. + \tau_1 \\ \vdots \\ \mu. + \tau_1 \\ \mu. + \tau_2 \\ \vdots \\ \mu. + \tau_2 \\ \vdots \\ \mu. + \tau_r \\ \vdots \\ \mu. + \tau_r \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \vdots \\ \mu_r \\ \vdots \\ \mu_r \end{pmatrix}$$

The model is $\mathbf{Y} = \mathbf{X}\beta + \mathcal{E}$

For the test of

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_{r-1} (= \tau_r)$$

We can use the F-test for linear regression model.

ANother way to define the model is by dummy variable

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{r1} \\ \vdots \\ Y_{rn_r} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu_r \\ \mu_1 - \mu_r \\ \vdots \\ \mu_{r-1} - \mu_r \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{r1} \\ \vdots \\ \varepsilon_{rn_r} \end{pmatrix},$$

The model can be written as $\mathbf{Y} = \mathbf{X}\beta + \mathcal{E}$. The hypothesis becomes

$$H_0 : \mu_1 - \mu_r = \mu_2 - \mu_r = \dots = \mu_{r-1} - \mu_r = 0$$

We can use the F-test for linear regression model.

Example 7.1 *To check whether different package designs may affect the sales. The following sales were observed in different stores*

package design <i>i</i>	Store <i>j</i>				
	1	2	3	4	5
1	11	17	16	14	15
2	12	10	15	19	11
3	23	20	18	17	
4	27	33	22	26	28

Let μ_i be the mean for package design i .

H_0 : package design's effect is not big, $\mu_1 = \dots = \mu_4$

<i>Source of variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F-value</i>
between treatment	588.221	3	196.0737	18.59
Error (within treatment)	158.2	15	10.54667	
Total	746.421	18		

$F(1 - 0.05, 3, 15) = 3.29$. Since $F\text{-value } 18.59 > 3.29$, we reject H_0 , that is the factor level means are different.

Regression approach 1

$$Y_{ij} = \beta_0 + \beta_1 D_{ij,1} + \beta_2 D_{ij,2} + \beta_3 D_{ij,3} + \varepsilon_{ij}$$

test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

we have

$$F^* = 18.59 > 3.29$$

reject H_0

Regression approach 2

$$Y_{ij} = \mu. + \tau_1 X_{ij,1} + \tau_2 X_{ij,2} + \tau_3 X_{ij,3} + \varepsilon_{ij}$$

test

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0$$

we have

$$F^* = 18.59 > 3.29$$

reject H_0

(code)