STAT3015/7030:
Generalised Linear Modelling
Two Way Anova

Bronwyn Loong

Semester 2 2014

References

Ch 13 - Ramsey and Schafer, The Statistical Sleuth

Ch 15 - Faraway, Linear models with R

Case Study - The Pygamalion effect

The Pygmalion effect in psychology refers to a situation where the high expectations of a supervisor or teacher translate into improved performance by students. The Pygmalion group are told their performance is exceptional. If the Pygmalion effect is present, the Pygmalion group should perform better than students outside the group.

In addition, if a student perceives the reduced expectations of a supervisor on them, they may oblige with a reduced performance.

The research project: To study the Pygmalion effect a researcher set up an experiment at an army training camp. Ten companies of soliders were selected. Each company had three platoons (army units). Each platoon had its own platoon leader. Using a random assignment mechanism, one of the three platoons in each company was selected to be the Pygmalion platoon.

Case Study - The Pygamalion effect

The randomization was conducted separately for each company to create a **randomized block experiment** with companies as blocks.

By blocking we can control for the effect of different companies (a nuisance parameter), and isolate the treatment effect within a block.

Selecting a wide range of companies broadens the scope of inference. If the effects of treatments are similar in all blocks, one can make a more universal claim about them.

Case Study - The Pygamalion effect

Implementing the Pygmalion effect - prior to assuming command of a platoon, each leader met with an army psychologist. The psychologist described a nonexistent (imaginary) set of tests that had predicted superior performance from his/her platoon.

The data: At the conclusion of basic training, soldiers took a number of tests to evaluate their ability to operate weapons and answer questions about their use.

Case Study - The Pygamalion effect

Company	Pygmalion	Control	
1	80.0	63.2	69.2
2	83.9	63.1	81.5
3	68.2	76.2	
4	76.5	59.5	73.5
5	87.8	73.9	78.5
6	89.8	78.9	84.7
7	76.1	60.6	69.6
8	71.5	67.8	73.2
9	69.5	72.3	73.9
10	83.7	63.7	77.7

Notes

- Each solider took the test, but the data are recorded at the platoon level (that is, the platoon's average), because treatments were assigned at the platoon level.
- Company 3 had only 2 platoons, so only one control.

Case Study - The Pygamalion effect

Results: The Pygmalion treatment adds an estimated 7.22 points to a platoon's score (95% interval: 1.80 to 12.64 points). The evidence strongly suggests the effect is real (one-sided p-value = 0.0060), and the experimental design allows for a causal inference.

Q: In building a statistical model to analyse the Pygmalion data, what are the factor(s) that we need to allow for to explain the variation in response?

The two-way anova model

Suppose we have two treatment factors A at I levels and B at J levels. Let n_{ij} be the number of observations at level i of factor A and level j of factor B and let those observations be $y_{ij1}, y_{ij2}, \ldots.$. A **complete** layout has $n_{ij} \geq 1$ for all i, j. A **balanced** layout requires that $n_{ij} = n$. $\epsilon_{ijk} \overset{i.i.d}{\sim} N(0, \sigma^2)$.

THE ADDITIVE MODEL

$$y_{ijk} = \mu_i + \gamma_j + \epsilon_{ijk}$$

(apply constraint $\sum_{j=1}^{J} \gamma_j = 0$) OR reparametise as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

(apply constraint $\alpha_1 = \beta_1 = 0$; OR $\sum_{i=1}^{\mathrm{I}} \alpha_i = 0 = \sum_{j=1}^{J} \beta_j = 0$)

Two-way anova additive model

What does the additive model imply about the effect of each factor on the response?

Two-way anova additive model

What does the additive model imply about the effect of each factor on the response? $\rightarrow$ the effects of one factor are the same at all levels of the other factor.

Example: The Pygmalion effect adds the same value to the score of a treated platoon regardless of the company (no interactions between treatment and company).

Two-way anova additive model - parameter estimation

For a balanced design and the parametisation $y_{ijk} = \mu_i + \gamma_j + \epsilon_{ijk}$ (with constraint $\sum_{j=1}^{J} \gamma_j = 0$): we can show that the least squares estimates are

$$\hat{\mu}_i = \frac{1}{nJ} \sum_{j=1}^{J} \sum_{k=1}^{n} Y_{ijk} = \bar{Y}_{i\bullet}$$

$$\hat{\gamma}_j = \frac{1}{n\mathrm{I}} \sum_{i=1}^{\mathrm{I}} \sum_{k=1}^{n} (Y_{ijk} - \hat{\mu}_i) = \bar{Y}_{\bullet j} - \frac{1}{\mathrm{I}} \sum_{i=1}^{\mathrm{I}} \bar{Y}_{i\bullet} = \bar{Y}_{\bullet j} - \bar{Y}$$

and we can show these are unbiased estimates:

$$E(\hat{\mu}_i) = \mu_i$$

$$E(\hat{\gamma}_j) = \gamma_j$$

Two-way anova additive model - parameter estimation

For the parametisation $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$, for the baseline constraint, the least squares estimates are:

$$\hat{\mu} = \hat{\mu}_1 + \hat{\gamma}_1 = \bar{Y}_{1\bullet} + \bar{Y}_{\bullet 1} - \bar{Y}$$

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}_1 = \bar{Y}_{i\bullet} - \bar{Y}_{1\bullet}$$

$$\hat{\beta}_j = \hat{\gamma}_j - \hat{\gamma}_1 = \bar{Y}_{\bullet j} - \bar{Y}_{\bullet 1}$$

The two-way anova model - regression parameterization for additive model

pyg - indicator if platoon received Pygmalion treatment

$cmp2, ..., cmp10$ - indicator variables for companies 2 through 10.

$$y_{ijk} = \beta_0 + \beta_1 pyg_{i1k} + \beta_2 cmp_{2jk} + ... + \beta_{10} cmp_{10jk} + \epsilon_{ijk}$$

The mean effects of each (treatment, company) combination are given by a function of the regression coefficients. For example,

$$\mu \{score | pyg = 1, cmp = 3\} = \beta_0 + \beta_1 + \beta_3$$

What is the Pygmalion treatment effect in any company?

The two-way anova model - regression parameterization for additive model

pyg - indicator if platoon received Pygmalion treatment

$cmp2, ..., cmp10$ - indicator variables for companies 2 through 10.

$$y_{ijk} = \beta_0 + \beta_1 pyg_{i1k} + \beta_2 cmp_{2jk} + ... + \beta_{10} cmp_{10jk} + \epsilon_{ijk}$$

The mean effects of each (treatment, company) combination are given by a function of the regression coefficients. For example,

$$\mu \{score | pyg = 1, cmp = 3\} = \beta_0 + \beta_1 + \beta_3$$

What is the Pygmalion treatment effect in any company? $\rightarrow \beta_1$.

What is the estimated difference in mean score between company 5 and company 4?

The two-way anova model - regression parameterization for additive model

pyg - indicator if platoon received Pygmalion treatment

$cmp2, ..., cmp10$ - indicator variables for companies 2 through 10.

$$y_{ijk} = \beta_0 + \beta_1 pyg_{i1k} + \beta_2 cmp_{2jk} + ... + \beta_{10} cmp_{10jk} + \epsilon_{ijk}$$

The mean effects of each (treatment, company) combination are given by a function of the regression coefficients. For example,

$$\mu \{score | pyg = 1, cmp = 3\} = \beta_0 + \beta_1 + \beta_3$$

What is the Pygmalion treatment effect in any company? $\rightarrow \beta_1$.

What is the estimated difference in mean score between company 5 and company 4? $\rightarrow \beta_5 - \beta_4$.

How many regression coefficients do we need to estimate?

The two-way anova model - regression parameterization for additive model

pyg - indicator if platoon received Pygmalion treatment

$cmp2, ..., cmp10$ - indicator variables for companies 2 through 10.

$$y_{ijk} = \beta_0 + \beta_1 pyg_{i1k} + \beta_2 cmp_{2jk} + ... + \beta_{10} cmp_{10jk} + \epsilon_{ijk}$$

The mean effects of each (treatment, company) combination are given by a function of the regression coefficients. For example,

$$\mu \{score | pyg = 1, cmp = 3\} = \beta_0 + \beta_1 + \beta_3$$

What is the Pygmalion treatment effect in any company? $\rightarrow \beta_1$.

What is the estimated difference in mean score between company 5 and company 4? $\rightarrow \beta_5 - \beta_4$.

How many regression coefficients do we need to estimate?
$(\mathrm{I} - 1) + (J - 1) + 1 = \mathrm{I} + J - 1$

Two-way anova - with interactions (nonadditive model)

Include interactions between the two factors.

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

OR

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Regression parameterization - pygmalion study

$$\begin{aligned} y_{ijk} = {} & \beta_0 + \beta_1 pyg_{i1k} + \beta_2 cmp_{2jk} + ... + \beta_{10} cmp_{10jk} + \\ & \beta_{11}(pyg_{i1k} \times cmp2jk) + ... + \beta_{19}(pyg_{i1k} \times cmp10jk) + \epsilon_{ijk} \end{aligned}$$

What is the treatment effect in company 1?

Two-way anova - with interactions (nonadditive model)

Include interactions between the two factors.

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

OR

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Regression parameterization - pygmalion study

$$y_{ijk} = \beta_0 + \beta_1 pyg_{i1k} + \beta_2 cmp_{2jk} + ... + \beta_{10} cmp_{10jk} + \\ \beta_{11}(pyg_{i1k} \times cmp2jk) + ... + \beta_{19}(pyg_{i1k} \times cmp10jk) + \epsilon_{ijk}$$

What is the treatment effect in company 1? $\rightarrow \beta_1$
What is the treatment effect in company 2?

Two-way anova - with interactions (nonadditive model)

Include interactions between the two factors.

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

OR

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Regression parameterization - pygmalion study

$$y_{ijk} = \beta_0 + \beta_1 pyg_{i1k} + \beta_2 cmp_{2jk} + \ldots + \beta_{10} cmp_{10jk} +$$
$$\beta_{11}(pyg_{i1k} \times cmp2jk) + \ldots + \beta_{19}(pyg_{i1k} \times cmp10jk) + \epsilon_{ijk}$$

What is the treatment effect in company 1? $\rightarrow \beta_1$
What is the treatment effect in company 2? $\rightarrow \beta_1 + \beta_{11}$
Treatment effect depends on company.

Two-way anova - with interactions (nonadditive model)

How many regression coefficients to estimate?

Two-way anova - with interactions (nonadditive model)

How many regression coefficients to estimate?
$1 + (\mathrm{I} - 1) + (J - 1) + (\mathrm{I} - 1) \times (J - 1) = \mathrm{I} \times J$. This is a *saturated* model because there are as many cells in the table as coefficients.

Parameter estimation in a balanced design, (**equal numbers of units in each factor combination**)

• (Model 1) What are the least squares estimates of the cell means, μ_{ij}?
In the saturated model, the cell means are completely unrelated. Therefore, the least squares estimates of the mean in any cell is the sample average of responses in that cell. ($\hat{\mu}_{ij} = \bar{Y}_{ij}$)

Two-way anova - the saturated, nonadditive model

- (Model 2)

(baseline constraint) μ, α_i, β_j are estimated as before (as per the additive model without interactions) .

$$\widehat{\alpha\beta}_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$$

What are the residuals in the two way anova?

Two-way anova - the saturated, nonadditive model

- (Model 2)

(baseline constraint) μ, α_i, β_j are estimated as before (as per the additive model without interactions) .

$$\widehat{\alpha\beta}_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$$

What are the residuals in the two way anova? $\rightarrow$ difference between responses and cell averages.

What is the estimate of the residual variance?

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{\mathrm{I}} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij})^2}{N - \mathrm{I} \times J}$$

Two-way anova - Testing for additivity

How do we decide on whether to include interaction terms or not ?

Two-way anova - Testing for additivity

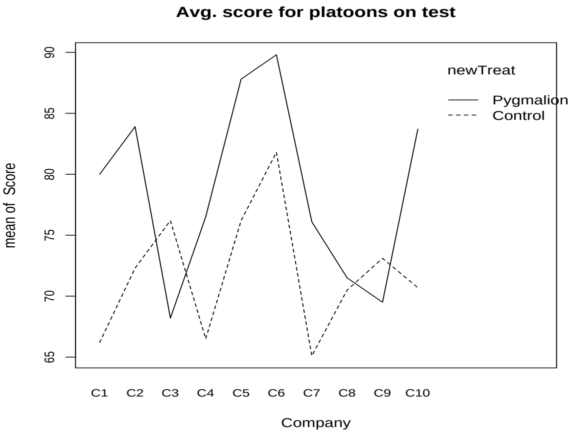
How do we decide on whether to include interaction terms or not ?
$\rightarrow$ conduct an F-test

$$H_0 : y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

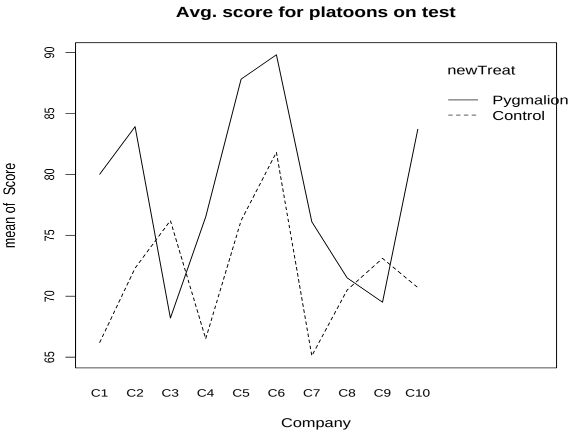
$$H_A : y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

$$F = \frac{(SSE_{red} - SSE_{full})/((\mathrm{I} - 1) \times (J - 1))}{SSE_{full}/(n - \mathrm{I} \times J)}$$

Analysis of the Pygmalion Data



Analysis of the Pygmalion Data



Test for significance of interactions

```
m1<-lm(SCORE ~ COMPANY * TREAT,data=pyg)
summary(m1)
anova(m1)

Analysis of Variance Table

Response: SCORE
               Df Sum Sq Mean Sq F value Pr(>F)
COMPANY         9 671     75      1.44     0.299
TREAT           1 339     339     6.53     0.031
COMPANY:TREAT   9 311     35      0.67     0.722
Residuals       9 467     52
```

Test for significance of interactions

```
m2<-lm(SCORE~COMPANY + TREAT, data=pyg)
summary(m2)

anova(m2)
Analysis of Variance Table

Response: SCORE
          Df  Sum Sq  Mean Sq  F value  Pr(>F)
COMPANY    9   671     75       1.72     0.156
TREAT      1   339     339      7.84     0.012
Residuals 18   779     43
```

Test for significance of interactions

```
> anova(m2,m1)
Analysis of Variance Table

Model 1: SCORE ~ COMPANY + TREAT
Model 2: SCORE ~ COMPANY * TREAT
  Res.Df RSS Df Sum of Sq    F    Pr(>F)
1 18     779
2 9      467 9   312       0.67   0.72
```

The p-value of the F-test is 0.72. We conclude the interaction terms are not significant and it is adequate to use the additive model.

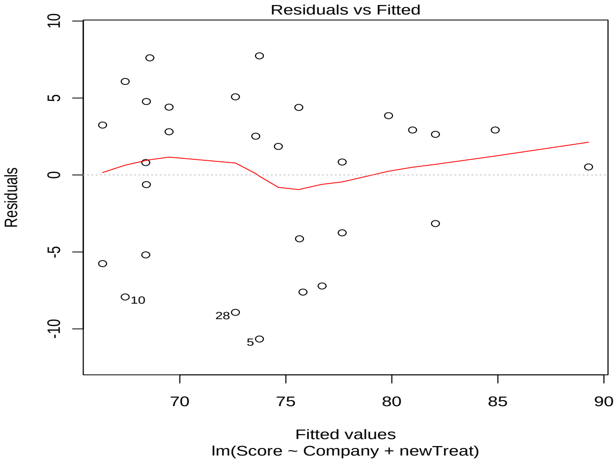
Test for significance of interactions

Calculating the F-test statistic

$$F - statistic = \frac{(779 - 467)/(18 - 9)}{52} = 0.667$$

$$p - value = Pr(F_{9,9} > 0.667) = 0.72$$

Residual diagnostic check



What is the treatment effect?

Obtain coefficient estimate for *pyg*.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
TREATPYGMALION    7.2205 2.5795       2.80   0.012
```

The estimate treatment effect is 7.2205. That is we expect on average the Pygmalion effect to increase the average test score by 7.2205 points, after accounting for the effects of company.

The appropriate multiplier to produce confidence intervals is $t_{18}(0.975) = 2.101$. The interval width is $2.101 \times 2.5795 = 5.4195$. Therefore the interval is $7.2205 \pm 5.4195 = (1.8, 12.6)$