STAT3015/7030:

Generalised Linear Modelling

One Way Anova

Semester 2 2016

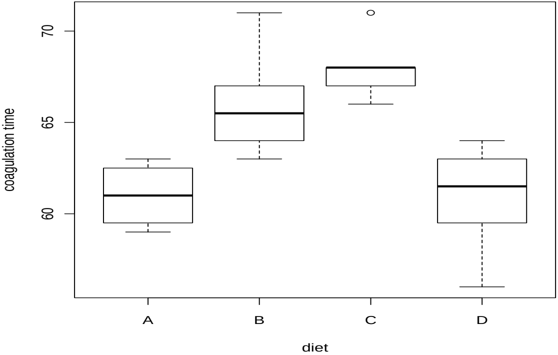
Originally prepared by Bronwyn Loong

Reference

Faraway J. Linear Models with R. Ch 14

Case Study I - Coagulation

To study the influence of different diets on blood coagulation times, 24 mice were randomly assigned to four different diets (A,B,C,D) and the samples were taken in a random order. Coagulation times were recorded for each animal. A boxplot of the data by diet is shown below.



Case Study I - Coagulation

Question: Is there a significant difference in coagulation times between different diets? Which diet would you recommend to produce the fastest coagulation time?

The box plot shows diets A and D recorded the lowest average coagulation times. Are the average times for diets A and D significantly different from the average times for diets B and C?

Is diet a significant treatment effect on coagulation time? Can we influence coagulation time by prescribing different diets?

GOAL: partition variance in coagulation times into that due to each diet (A,B,C,D) plus error. → use ANOVA!!

One-Way Analysis of Variance

ANOVA type problems traditionally arose from randomised experiments where we want to estimate the treatment effect on the response.

Terminology: predictors are now all qualitative and we call them *factors*. Parameter estimates are termed *effects*

One-Way Analysis of Variance

The Model I

Suppose we have a factor occurring at $i = 1, ..., \mathrm{I}$ levels, with $j = 1, .., n_i$ observations at level i. y_{ij} is the j^{th} observed outcome at factor level i. ϵ_{ij} is the underlying error variable, and we assume $\epsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$. We use the model:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

The treatment effects are the group means μ_i

(note: if all n_i's are equal - this is a **balanced** design).
$(\sum_{i=1}^{I} n_i = N) =$ total sample size

One-Way Analysis of Variance

The Model II - (alternative parametisation)

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Is model formulation identifiable? That is, can we uniquely determine parameter values μ and α_i?

One-Way Analysis of Variance

The Model II - (alternative parametisation)

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Is model formulation identifiable? That is, can we uniquely determine parameter values μ and α_i?

(If we add a constant c to μ and subtract the same constant from α_i, the model is unchanged → need some constraints to reduce the number of parameters to estimate!!)

One-Way Analysis of Variance

The Model II

Some possible constraints:

first group as control group

1. Set $\alpha_1 = 0$, the μ represents the expected mean response for the first level and α_i for $i \neq 1$ represents the difference between level i and level one. Level 1 is the *reference level* or *baseline level* (baseline/ treatment contrasts constraint)
2. Set $\sum_{i=1}^{\mathrm{I}} \alpha_i = 0$. Now μ represents the mean response over all levels and α_i is the difference from that mean (sum constrast constraint)

The first constraint demonstrates the connection between the ANOVA model and multiple regression.

Parameter estimation

Model I

We want unbiased estimates for $\mu_i \rightarrow$ use least squares!

Distance function to minimise is

$$d(\mu_1, \mu_2, ..., \mu_I) = \sum_{i=1}^{\mathrm{I}} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$

... and we can derive the estimates

$$\hat{\mu}_i = \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

Exercise: show that $\hat{\mu}_i$ is an unbiased estimate. Find $Var(\hat{\mu}_i)$.

[Note - this exercise , whilst worthwhile, will not be examinable in this course]

Parameter estimation

Model II

Distance function to minimise is

$$d(\mu, \alpha_1, \alpha_2, \alpha_3, ..., \alpha_{\mathrm{I}}) = \sum_{i=1}^{\mathrm{I}} \sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2$$

For the baseline constraint, $(\alpha_1 = 0)$, we can show that the least squares estimates are:

$$\hat{\mu} = \bar{Y}_1$$

$$\hat{\alpha}_i = \bar{Y}_i - \bar{Y}_1$$

Exercise: show that these estimates are unbiased.

What is the unbiased estimate for $\mu_i =$? What is the variance estimate for $Var(\hat{\alpha}_i)$?

Parameter estimation

To estimate σ^2 note the breakdown

$$\sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n_i}(\bar{Y}_i-\bar{Y})^2 + \sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$$

$$= \sum_{i=1}^{\mathrm{I}} n_i(\bar{Y}_i-\bar{Y})^2 + \sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$$

$$SST = SSR + SSE$$

- SST: total variability of response
- SSR: variability *between* factor levels, $MSR = \frac{SSR}{\mathrm{I}-1}$
- SSE: variability *within* factor levels, $MSE = \frac{SSE}{N-\mathrm{I}} = \hat{\sigma}^2$

Exercise: show that $\hat{\sigma}^2$ is an unbiased estimate of σ^2.
[Again, not examinable - but used to be an old tutorial q. if you want the solution]

Checking assumptions

What are the fitted values in the one-way ANOVA? → the cell means $\bar{Y}_i$.

What are the residuals in the one-way ANOVA? → $e_{ij} = Y_{ij} - \bar{Y}_i$.

To check assumptions........

- plot residuals vs fitted values - check constant variance, independence
- qq plot of residuals - check normality
- compare "within group" variance estimates to check for homoscedasticity. That is compare, $s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$ Rule of thumb: ratio of largest to smallest "within group" variance is no more than about I (the number of treatments).

Hypothesis testing

In an ANOVA model - what is our primary research question? To test whether the treatment is significant. That is, are there significant differences in the levels of the factor?

What is the null hypothesis? What is the alternative hypothesis?

$$H_0 : Y_{ij} = \mu + \epsilon_{ij}$$

i.e. there's no level effect

$$H_A : Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \text{ for some } \alpha_i, i = 2...., \mathrm{I}$$

How to perform the test? → compare residual sums of squares and conduct an F-test as used in regression. What is the F-test statistic??

$$F = \frac{(SSE_{red} - SSE_{full})/(\mathrm{I} - 1)}{SSE_{full}/N - \mathrm{I}} = \frac{MSR}{MSE}$$

What is the p-value?

- If significant, find out which levels drive the significance.

Hypothesis testing

$$H_0 : \alpha_1 ..... = \alpha_{\mathrm{I}}$$

$$H_A : \alpha_i \neq \alpha_j, \text{ for some pair } i \neq j,$$

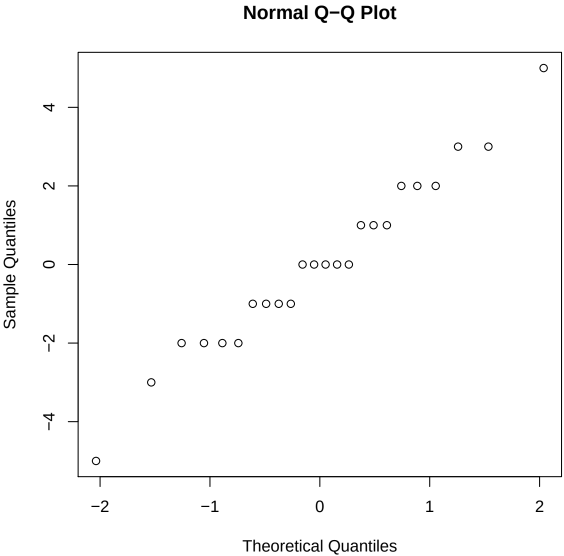
Is this equivalent to testing the significance of all the α_i's individually? → not quite, because my null requires I set all my α_i's equal to α_1 (jointly, not individually). For the individual tests, $H_0 : \alpha_i = 0$ (a single parameter test).

Case Study - Coagulation Data

(See R Code pdf file)

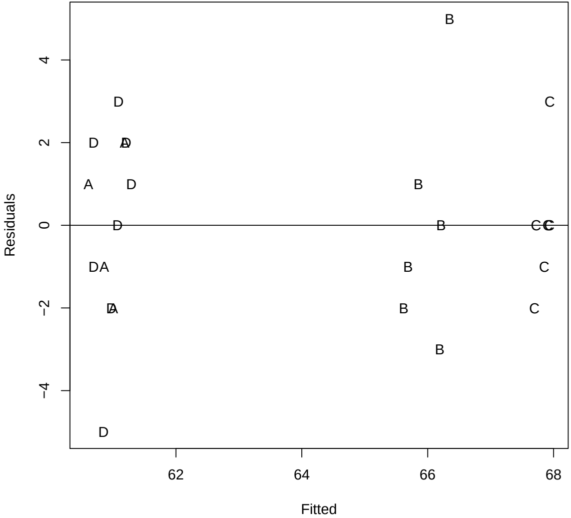
Case Study - Coagulation Data

Check diagnostics - normality



Case Study - Coagulation Data

Check diagnostics -constant variance



Case Study - Coagulation Data

Check diagnostics - constant variance

```
> with(coagulation, var(coag[diet=="A"]))
[1] 3.333333
> with(coagulation, var(coag[diet=="B"]))
[1] 8
> with(coagulation, var(coag[diet=="C"]))
[1] 2.8
> with(coagulation, var(coag[diet=="D"]))
[1] 6.857143
> 8/2.8
[1] 2.857143 (< 4)
```

Recap

The ANOVA model - test for differences between 'treatment' group means.

The statistical model to describe the observed outcomes given treatment effects:

$$\text{Model I: } Y_{ij} = \mu_i + \epsilon_{ij}$$

$$\text{Model II: } Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Why is the statistical model useful?

Because under a certain set of assumptions, we can perform valid statistical tests to test the significance of the treatment effect. (vs subjective judgement).

We can also quantify the uncertainty in our estimates. How??

Recap

The ANOVA model - test for differences between 'treatment' group means.

The statistical model to describe the observed outcomes given treatment effects:

$$\text{Model I: } Y_{ij} = \mu_i + \epsilon_{ij}$$

$$\text{Model II: } Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Why is the statistical model useful?

Because under a certain set of assumptions, we can perform valid statistical tests to test the significance of the treatment effect. (vs subjective judgement).

We can also quantify the uncertainty in our estimates. How?? through the random error component ϵ_{ij}.

Pairwise comparisons and confidence intervals

After detecting a significant treatment effect, what is the 'logical' next step? we want to find out which combinations of levels or combinations of levels are different.

A pairwise comparison of level i and level k? - construct a confidence interval for $\mu_i - \mu_k$

$$\bar{Y}_i - \bar{Y}_k \pm t_{n-\mathrm{I}}(1-\alpha/2)\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}$$

where $Var(\bar{Y}_i - \bar{Y}_k) = Var(\bar{Y}_i) + Var(\bar{Y}_k) = \hat{\sigma}^2\left(\frac{1}{n_i} + \frac{1}{n_k}\right)$

why is $Cov(\bar{Y}_i - \bar{Y}_k) = 0$? because, $\bar{Y}_i$ and $\bar{Y}_k$ are estimates from disjoint and therefore <u>independent</u> subsets of the observed data

Pairwise comparisons and confidence intervals

If we simultaneously want to test all possible pairs of treatment effects, we need to adjust our significance level for multiple comparisons to achieve an overall type I level error $\alpha = 0.05$.

Familywise confidence level - success rate of a procedure for constructing a family of confidence intervals, where a "successful" usage is one in which all intervals in the family capture their parameters

Suppose we want to test 10 confidence intervals simultaneously, each with individual confidence level 95%.
What is an upper bound on the familywise confidence level? 95%
(consider if none of the intervals overlap)
What is a lower bound on the familywise confidence level?
$100(1 - 0.05 \times 10) = 50\% \rightarrow$
(Bonferroni inequality)
$Pr(A_g \cap A_2 .... \cap A_g) \geq 1 - \sum_{i=1}^{g} Pr(A_i^c)$ (proof not required)

Pairwise comparisons and confidence intervals

Bonferroni method - if we want the simultaneous confidence of g confidence intervals to be at least 1 - α, then we can guarantee this is the case if we use individual confidence intervals which each have confidence level 1-α/g.

Pairwise comparisons and confidence intervals

We can find a $100(1-\alpha)\%$ confidence interval for any linear combination of the treatment effects.

A **contrast** among the effects $\mu_1, ..., \mu_{\mathrm{I}}$ is a linear combination $\sum_i c_i \mu_i$ where the c_i are known and $\sum_i c_i = 0$.

For example, $\mu_1 - \mu_2$ is a contrast with $c_1 = 1$, $c_2 = -1$ and the other $c_i = 0$. All pairwise differences are contrasts.

The desired confidence interval is

$$\sum_{i=1}^{\mathrm{I}} c_i \bar{Y}_i \pm t_{n-\mathrm{I}}(1-\alpha/2)\hat{\sigma}\sqrt{\sum_{i=1}^{\mathrm{I}} \frac{c_i^2}{n_i}}$$

Case Study - Coagulation Data

4 diets - how many possible pairwise comparions? ${}^4C_2 = 6$. (see R code)

Case Study - Coagulation Data

Suppose Diet A and Diet D contain a special protein which Diets B and C do not. So we might want to test whether diets with the special protein significantly decrease the coagulation time. What linear combination of parameters are we interested in?

$$\frac{\mu_A + \mu_D}{2} - \frac{\mu_B + \mu_C}{2}$$

The confidence interval estimate for the above linear combination is (-8.08,-3.92). This interval does not contain zero.

Conclusion: We conclude that the special protein will significantly decrease coagulation times (see R code pdf file)

Indicator variables and regression

We can write the ANOVA model under the treatment contrast constraint as a multiple regression. Write down the equivalent model for the coagulation data example.

Define $\mathrm{I} - 1$ indicator/dummy variables.

$$z_{p,ij} = \begin{cases} 1 & \text{if i} = \text{p} + 1 \\ 0 & \text{otherwise} \end{cases}$$

For the coagulation data example, the equivalent multiple regression model is of the form

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

Indicator variables and regression

(Rcode)

```
diet1 < − with(coagulation,ifelse(diet=="B",1,0))
diet2 < − with(coagulation,ifelse(diet=="C",1,0))
diet3 < − with(coagulation,ifelse(diet=="D",1,0))
diets < − cbind(diet1,diet2,diet3)
coag.lm < − lm(coag ∼ diets,coagulation)
anova(coag.lm)
Analysis of Variance Table

Response: coag

          Df Sum Sq Mean Sq F value Pr(>F)
diet       3 228     76.0    13.571 4.658e-05 ***
Residuals 20 112     5.6
```

Indicator variables and regression

Compare to results using aov() command in R.

```
> coag.aov < − aov(coag∼diet, data=coagulation)
> summary(coag.aov)
          Df Sum Sq Mean Sq F value Pr(>F)
diet       3 228     76.0    13.571 4.658e-05 ***
Residuals 20 112     5.6
```

Random effects

(Chapter 8, Faraway, Extending the Linear Model with R)
(Chapter 5.6, Sleuth)

Fixed effect: unknown constant we try to estimate from the data

Random effect: is a random variable, estimate the parameters which describe the distribution of this random effect

Random effects

Why might we consider random effects?

- Account for grouped structure of data - correlation between observations within the same group. Relax assumption of independence between observations.
- Levels of a factor or units of observation are seen as a random sample from a larger population. Inference is desired to a larger set from which the groups are a sample.

Random effects

Example: The manager of an industrial plant wanted to determine whether workers with the same skill level have any discernible differences in the number of units of an automobile part that are manufactured during a fixed period of time. Five workers were randomly selected and the number of units produced by each worker for six equal time length periods was recorded.

The workers are randomly selected from a population of workers, and thus a random effects model seems appropriate for the data. Let's consider the following model for the data:

Random effects

$$
\begin{aligned}
y_{ij}|\mu_i &\overset{\text{indep.}}{\sim} \mathrm{Normal}(\mu_i, \sigma^2) \\
\mu_i &\overset{\text{iid}}{\sim} \mathrm{Normal}(\mu, \sigma_\mu^2) \\
& i = 1, \ldots, \mathrm{I} = 5; j = 1, \ldots, n_i = 6
\end{aligned}
$$

Here we see:

$$
\begin{aligned}
E[y_{ij}|\mu_i] &= \mu_i \\
E[y_{ij}] &= \mu
\end{aligned}
$$

Random effects

We can also consider the following model:

$$\begin{aligned} y_{ij}|\alpha_i &\overset{\text{indep.}}{\sim} \mathrm{Normal}(\mu+\alpha_i,\sigma^2) \\ \alpha_i &\overset{\text{iid}}{\sim} \mathrm{Normal}(0,\sigma_\alpha^2) \\ & i=1,\ldots,\mathrm{I}=5; j=1,\ldots,n_i=6 \end{aligned}$$

σ^2: measurement variance; σ_α^2= population variance of the effect (variation in worker productivity)
Here we see:

$$\begin{aligned} E[y_{ij}|\alpha_i] &= \mu+\alpha_i \\ E[y_{ij}] &= \mu \end{aligned}$$

Random effects

We can show:

- $Cov(y_{ij}, y_{il}) = \sigma_a^2$ (observations within the same group are no longer independent)
- $V(y_{ij}) = \sigma_a^2 + \sigma^2$
- $Cov(y_{ij}, y_{kr}) = 0$
- The intraclass correlation coefficient $Cor(y_{ij}, y_{il}) = \frac{\sigma_a^2}{(\sqrt{(\sigma_a^2+\sigma^2)})(\sqrt{(\sigma_a^2+\sigma^2)})} = \frac{\sigma_a^2}{\sigma_a^2+\sigma^2}$

Random effects - balanced design

- Fixed effects model:

Source	df	SS	E(MS)
Treatments	$\mathrm{I}-1$	$SSR=\sum_{i=1}^{\mathrm{I}} n(\bar{Y}_{i\cdot}-\bar{Y})^2$	$\sigma^2+n\frac{\sum_{i=1}^{\mathrm{I}}\alpha_i^2}{\mathrm{I}-1}$
Error	$(n-1)\mathrm{I}$	$SSE=\sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i\cdot})^2$	σ^2

- Random effects model:

Source	df	SS	E(MS)
Treatments	$\mathrm{I}-1$	$SSR=\sum_{i=1}^{\mathrm{I}} n(\bar{Y}_{i\cdot}-\bar{Y})^2$	$\sigma^2+n\sigma_\alpha^2$
Error	$(n-1)\mathrm{I}$	$SSE=\sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i\cdot})^2$	σ^2

Random effects - balanced design

How would you test for a significant worker effect? → for a balanced one-way ANOVA the same F statistics for the fixed effects models will work here.

Test for random effect: $H_0 : \sigma_\alpha^2 = 0$ based on

$$F = \frac{MSR}{MSE} \sim F_{\mathrm{I}-1,N-\mathrm{I}} \text{ under } H_0$$

Reason: Under H_0, $SSR \sim \sigma^2\chi^2_{\mathrm{I}-1}$ and $SSE \sim \sigma^2\chi^2_{N-\mathrm{I}}$. Therefore the F-test has the distribution $F_{\mathrm{I}-1,(n-1)\mathrm{I}}$ under H_0

Random effects - balanced design

We can apply the same ANOVA and F-test in the fixed effects case for analyzing the data. However we need to compute the expected mean squares under the alternative of $\sigma_\alpha^2 > 0$ to estimate the variance components

Random effects - balanced design

$$\sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2=\sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n_i}(\bar{Y}_i-\bar{Y})^2+\sum_{i=1}^{\mathrm{I}}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$$

$$SST = SSR + SSE$$

(assume $n_i = n$) (== balanced design) We can show: $E[SSE] = \mathrm{I}(n-1)\sigma^2$ and $E[SSR] = (\mathrm{I}-1)(n\sigma_\alpha^2 + \sigma^2)$. Which suggests using the estimators:

$$\hat{\sigma}^2 = SSE/(\mathrm{I}(n-1)) = MSE$$

$$\hat{\sigma}_\alpha^2 = \frac{SSR/(\mathrm{I}-1) - \sigma^2}{n} = \frac{MSR - MSE}{n}$$

Problems with $\hat{\sigma}_\alpha^2$?

Random effects - Inference for μ

When cell sizes are the same (balanced)

$$\hat{\mu} = \bar{Y} = \frac{1}{n\mathrm{I}} \sum_{i,j} Y_{ij}$$

We can check that:

$$E[\bar{Y}] = \mu$$

$$Var[\bar{Y}] = \frac{n\sigma_\alpha^2 + \sigma^2}{n\mathrm{I}}$$

Now $E[MSR] = n\sigma_\alpha^2 + \sigma^2$. Therefore,

$$\frac{\bar{Y} - \mu}{\sqrt{\frac{SSR}{(\mathrm{I}-1)n\mathrm{I}}}} \sim t_{\mathrm{I}\text{-}1}$$

Example

(See R code example)

Random effects

For unbalanced designs, we can use the likelihood ratio test to compare the two models with and without the random effect:

$$2(logL(\hat{\theta}_L|y) - logL(\hat{\theta}_S|y))$$

S: smaller model without random effect; L: larger model with random effect
Under H_0 the likelihood ratio test statistic is approximately chi-squared with degrees of freedom equal to the difference in the number of parameters between the two models.
(see R code)

Predicting the Random effects

Want to estimate the α_i's given the data y_i ..
If we write out the density
$p(\alpha_i|y) \propto \prod_i \prod_j p(y_{ij}|\sigma^2, \alpha_i, \mu)p(\alpha_i|\sigma_\alpha^2)$
Then with some algebra (not required).....we can show that

$$\alpha_i|y \sim N(m, v)$$

where

$$v = \left(\frac{n_i}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}_\alpha^2}\right)^{-1}$$

$$m = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}^2/n_i}(\bar{y}_{i\cdot} - \hat{\mu})$$