Workshop 7.4a: Single factor ANOVA

Murray Logan

23 Nov 2016

Section 1

Revision

Estimation

Y	X
3	0
2.5	1
6	2
5.5	3
9	4
8.6	5
12	6

$$
\begin{aligned}
3.0 &= \beta_0 \times 1 + \beta_1 \times 0 + \varepsilon_1 \\
2.5 &= \beta_0 \times 1 + \beta_1 \times 1 + \varepsilon_1 \\
6.0 &= \beta_0 \times 1 + \beta_1 \times 2 + \varepsilon_2 \\
5.5 &= \beta_0 \times 1 + \beta_1 \times 3 + \varepsilon_3
\end{aligned}
$$

Estimation

$$\begin{aligned}
3.0 &= \beta_0 \times 1 + \beta_1 \times 0 + \varepsilon_1 \\
2.5 &= \beta_0 \times 1 + \beta_1 \times 1 + \varepsilon_1 \\
6.0 &= \beta_0 \times 1 + \beta_1 \times 2 + \varepsilon_2 \\
5.5 &= \beta_0 \times 1 + \beta_1 \times 3 + \varepsilon_3 \\
9.0 &= \beta_0 \times 1 + \beta_1 \times 4 + \varepsilon_4 \\
8.6 &= \beta_0 \times 1 + \beta_1 \times 5 + \varepsilon_5 \\
12.0 &= \beta_0 \times 1 + \beta_1 \times 6 + \varepsilon_6
\end{aligned}$$

$$\underbrace{\begin{pmatrix} 3.0 \\ 2.5 \\ 6.0 \\ 5.5 \\ 9.0 \\ 8.6 \\ 12.0 \end{pmatrix}}_{\text{Response values}} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{pmatrix}}_{\text{Model matrix}} \times \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\text{Parameter vector}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}}_{\text{Residual vector}}$$

Matrix algebra

$$\underbrace{\begin{pmatrix} 3.0 \\ 2.5 \\ 6.0 \\ 5.5 \\ 9.0 \\ 8.6 \\ 12.0 \end{pmatrix}}_{\text{Response values}} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{pmatrix}}_{\text{Model matrix}} \times \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\text{Parameter vector}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}}_{\text{Residual vector}}$$

$$\mathrm{Y} = \mathrm{X}\beta + \epsilon$$

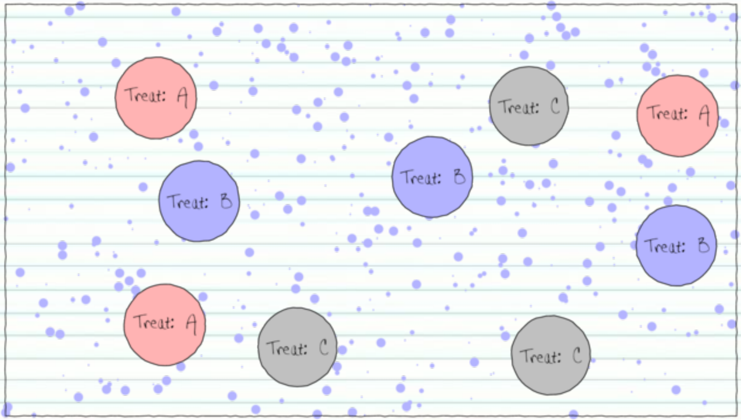
$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$$

Section 2

Anova Param-eterization

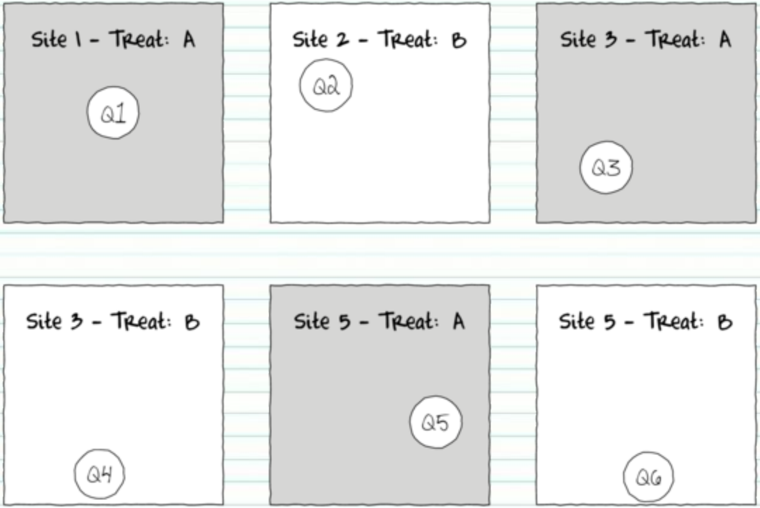
Simple ANOVA

Three treatments (One factor - 3 levels), three replicates



Simple ANOVA

Two treatments, three replicates



Categorical predictor

Y	A	dummy1	dummy2	dummy3
2	G1	1	0	0
3	G1	1	0	0
4	G1	1	0	0
6	G2	0	1	0
7	G2	0	1	0
8	G2	0	1	0
10	G3	0	0	1
11	G3	0	0	1
12	G3	0	0	1

$$y_{ij} = \mu + \beta_1(\mathtt{dummy}_1)_{ij} + \beta_2(\mathtt{dummy}_2)_{ij} + \beta_3(\mathtt{dummy}_3)_{ij} + \varepsilon_{ij}$$

Overparameterized

$$y_{ij} = \mu + \beta_1(\mathtt{dummy}_1)_{ij} + \beta_2(\mathtt{dummy}_2)_{ij} + \beta_3(\mathtt{dummy}_3)_{ij} + \varepsilon_{ij}$$

Y	A	Intercept	dummy1	dummy2	dummy3
2	G1	1	1	0	0
3	G1	1	1	0	0
4	G1	1	1	0	0
6	G2	1	0	1	0
7	G2	1	0	1	0
8	G2	1	0	1	0
10	G3	1	0	0	1
11	G3	1	0	0	1
12	G3	1	0	0	1

Overparameterized

$$\mathtt{y}_{\mathtt{ij}} = \mu + \beta_1(\mathtt{dummy}_1)_{\mathtt{ij}} + \beta_2(\mathtt{dummy}_2)_{\mathtt{ij}} + \beta_3(\mathtt{dummy}_3)_{\mathtt{ij}} + \varepsilon_{\mathtt{ij}}$$

- three treatment groups
- four parameters to estimate
- need to re-parameterize

Categorical predictor

$$y_i = \mu + \beta_1(\mathtt{dummy}_1)_i + \beta_2(\mathtt{dummy}_2) + \beta_3(\mathtt{dummy}_3)_i + \varepsilon_i$$

MEANS PARAMETERIZATION

$$y_i = \beta_1(\mathtt{dummy}_1)_i + \beta_2(\mathtt{dummy}_2)_i + \beta_3(\mathtt{dummy}_3)_i + \varepsilon_{ij}$$

$$y_{ij} = \alpha_i + \varepsilon_{ij} \quad i = p$$

Categorical predictor

MEANS PARAMETERIZATION

$$y_i = \beta_1(\text{dummy}_1)_i + \beta_2(\text{dummy}_2)_i + \beta_3(\text{dummy}_3)_i + \varepsilon_i$$

Y	A	dummy1	dummy2	dummy3
2	G1	1	0	0
3	G1	1	0	0
4	G1	1	0	0
6	G2	0	1	0
7	G2	0	1	0
8	G2	0	1	0
10	G3	0	0	1
11	G3	0	0	1
12	G3	0	0	1

Categorical predictorDD

MEANS PARAMETERIZATION

$$y_i = \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \varepsilon_i$$
$$y_i = \alpha_p + \varepsilon_i,$$

	Y	A
1	2.00	G1
2	3.00	G1
3	4.00	G1
4	6.00	G2
5	7.00	G2
6	8.00	G2
7	10.00	G3
8	11.00	G3
9	12.00	G3

where p = number of levels of the factor and D = dummy variables

$$\begin{pmatrix} 2 \\ 3 \\ 4 \\ 6 \\ 7 \\ 8 \\ 10 \\ 11 \\ 12 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix}$$

Categorical predictor

MEANS PARAMETERIZATION

Parameter	Estimates	Null Hypothesis
α_1^*	mean of group 1	H$_0$: $\alpha_1 = \alpha_1 = 0$
α_2^*	mean of group 2	H$_0$: $\alpha_2 = \alpha_2 = 0$
α_3^*	mean of group 3	H$_0$: $\alpha_3 = \alpha_3 = 0$

```
> summary(lm(Y~-1+A))$coef
```

```
    Estimate Std. Error   t value     Pr(>|t|)
AG1        3  0.5773503  5.196152 2.022368e-03
AG2        7  0.5773503 12.124356 1.913030e-05
AG3       11  0.5773503 19.052559 1.351732e-06
```

Categorical predictor

$$y_i = \mu + \beta_1(\mathtt{dummy}_1)_i + \beta_2(\mathtt{dummy}_2)_i + \beta_3(\mathtt{dummy}_3)_i + \varepsilon_i$$

EFFECTS PARAMETERIZATION

$$y_{ij} = \mu + \beta_2(\mathtt{dummy}_2)_{ij} + \beta_3(\mathtt{dummy}_3)_{ij} + \varepsilon_{ij}$$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = p - 1$$

Categorical predictor

EFFECTS PARAMETERIZATION

$$y_i = \alpha + \beta_2(\mathtt{dummy}_2)_i + \beta_3(\mathtt{dummy}_3)_i + \varepsilon_i$$

Y	A	alpha	dummy2	dummy3
2	G1	1	0	0
3	G1	1	0	0
4	G1	1	0	0
6	G2	1	1	0
7	G2	1	1	0
8	G2	1	1	0
10	G3	1	0	1
11	G3	1	0	1
12	G3	1	0	1

Categorical predictor

EFFECTS PARAMETERIZATION

	Y	A
1	2.00	G1
2	3.00	G1
3	4.00	G1
4	6.00	G2
5	7.00	G2
6	8.00	G2
7	10.00	G3
8	11.00	G3
9	12.00	G3

$$y_i = \alpha + \beta_2 D_{2i} + \beta_3 D_{3i} + \varepsilon_i$$

$$y_i = \alpha_p + \varepsilon_i,$$

where p = number of levels of the factor minus 1 and

D = dummy variables

$$\begin{pmatrix} 2 \\ 3 \\ 4 \\ 6 \\ 7 \\ 8 \\ 10 \\ 11 \\ 12 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix}$$

Categorical predictor

TREATMENT CONTRASTS

Parameter	Estimates	Null Hypothesis
Intercept	mean of control group	H_0: $\mu = \mu_1 = 0$
α_2^*	mean of group 2 minus mean of control group	H_0: $\alpha_2 = \alpha_2 = 0$
α_3^*	mean of group 3 minus mean of control group	H_0: $\alpha_3 = \alpha_3 = 0$

```
> contrasts(A) <-contr.treatment
> contrasts(A)
```

Categorical predictor

TREATMENT CONTRASTS

Parameter	Estimates	Null Hypothesis
Intercept	mean of control group	H$_0$: $\mu = \mu_1 = 0$
α_2^*	mean of group 2 minus mean of control group	H$_0$: $\alpha_2 = \alpha_2 = 0$
α_3^*	mean of group 3 minus mean of control group	H$_0$: $\alpha_3 = \alpha_3 = 0$

```
> summary(lm(Y~A))$coef
```

Categorical predictor

USER DEFINED CONTRASTS

User defined contrasts

Grp2 vs Grp3

Grp1 vs (Grp2 & Grp3)

	Grp1	Grp2	Grp3
α_2^*	0	1	-1
α_3^*	1	-0.5	-0.5

```
> contrasts(A) <- cbind(c(0,1,-1),c(1, -0.5, -0.5))
> contrasts(A)
```

```
   [,1] [,2]
G1    0  1.0
```

Categorical predictor

USER DEFINED CONTRASTS

- $p - 1$ comparisons (contrasts)
- all contrasts must be orthogonal

Categorical predictor

ORTHOGONALITY

Four groups (A, B, C, D)

$p - 1 = 3$ comparisons

1. A vs B :: $A > B$
2. B vs C :: $B > C$
3. A vs C ::

Categorical predictor

USER DEFINED CONTRASTS

```
> contrasts(A) <- cbind(c(0,1,-1),c(1, -0.5, -0.5))
> contrasts(A)
```

```
   [,1] [,2]
G1    0  1.0
G2    1 -0.5
G3   -1 -0.5
```

$$
\begin{aligned}
0 \times 1.0 &= 0 \\
1 \times -0.5 &= -0.5 \\
-1 \times -0.5 &= 0.5 \\
\text{sum} &= 0
\end{aligned}
$$

Categorical predictor

USER DEFINED CONTRASTS

```
> contrasts(A) <- cbind(c(0,1,-1),c(1, -0.5, -0.5))
> contrasts(A)
```

```
   [,1] [,2]
G1    0  1.0
G2    1 -0.5
G3   -1 -0.5
```

```
> crossprod(contrasts(A))
```

```
     [,1] [,2]
[1,]    2  0.0
[2,]    0  1.5
```

```
> summary(lm(Y~A))$coef
```

Categorical predictor

USER DEFINED CONTRASTS

```
> contrasts(A) <- cbind(c(1, -0.5, -0.5),c(1,-1,0))
> contrasts(A)
```

```
   [,1] [,2]
G1  1.0    1
G2 -0.5   -1
G3 -0.5    0
```

```
> crossprod(contrasts(A))
```

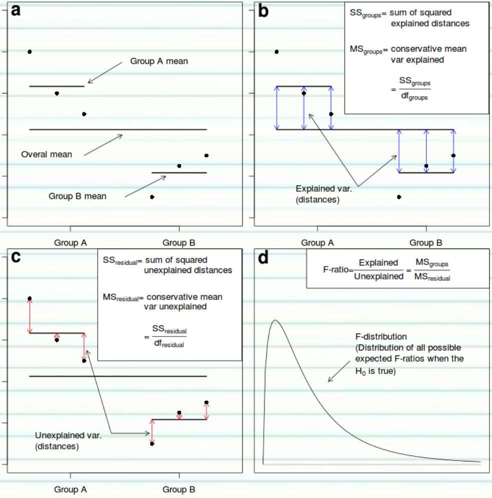
```
     [,1] [,2]
[1,]  1.5  1.5
[2,]  1.5  2.0
```

Section 3

Partitioning of variance (ANOVA)

ANOVA

PARTITIONING VARIANCE



ANOVA

PARTITIONING VARIANCE

```
> anova(lm(Y~A))
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)    
A          2     96      48      48 0.0002035 ***
Residuals  6      6       1                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Categorical predictor

POST-HOC COMPARISONS

No. of Groups	No. of comparisons	Familywise Type I error probability
3	3	0.14
5	10	0.40
10	45	0.90

Categorical predictor

POST-HOC COMPARISONS

Bonferoni

```
> summary(lm(Y~A))$coef
```

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)        7  0.3333333 21.000000 7.595904e-07
A1                -8  0.9428090 -8.485281 1.465426e-04
A2                 4  0.8164966  4.898979 2.713682e-03
```

```
> 0.05/3
```

```
[1] 0.01666667
```

Categorical predictor

POST-HOC COMPARISONS

Tukey?s test

```
> library(multcomp)
> data.lm<-lm(Y~A)
> summary(glht(data.lm, linfct=mcp(A="Tukey")))
```

```
	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = Y ~ A)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
G2 - G1 == 0   4.0000     0.8165   4.899  0.00653 **
G3 - G1 == 0   8.0000     0.8165   9.798  < 0.001 ***
G3 - G2 == 0   4.0000     0.8165   4.899  0.00679 **
---
```

Assumptions

- Normality
- Homogeneity of variance
- Independence

- As for regression

Section 4

Worked Examples

Worked Examples

```
> day <- read.csv('../data/day.csv', strip.white=T)
> head(day)
```

```
  TREAT BARNACLE
1  ALG1       27
2  ALG1       19
3  ALG1       18
4  ALG1       23
5  ALG1       25
6  ALG2       24
```

Worked Examples

Question: what effects do different substrate types have on barnacle recruitment

Linear model:

$$\text{Barnacle}_\text{i} = \mu + \alpha_\text{j} + \varepsilon_\text{i} \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Worked Examples

```
> partridge <- read.csv('../data/partridge.csv', strip.white=T)
> head(partridge)
```

```
  GROUP LONGEVITY
1 PREG8        35
2 PREG8        37
3 PREG8        49
4 PREG8        46
5 PREG8        63
6 PREG8        39
```

```
> str(partridge)
```

```
'data.frame':	125 obs. of  2 variables:
 $ GROUP    : Factor w/ 5 levels "NONE0","PREG1",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ LONGEVITY: int  35 37 49 46 63 39 46 56 63 65 ...
```

Worked Examples

Question: what effects does mating have on the longevity of male fruitflies

Linear model:

$$\text{Longevity}_i = \mu + \alpha_j + \varepsilon_i \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$