Workshop 8.2a: Heterogeneity

Murray Logan

23 Jul 2016

Section 1

Linear modelling assumptions

Assumptions

$$\mathtt{y}_\mathtt{i} = \beta_0 + \beta_1 \times \mathtt{x}_\mathtt{i} + \varepsilon_\mathtt{i}$$
$$\epsilon_\mathtt{i} \sim \mathcal{N}(0, \sigma^2)$$

Linear modelling assumptions

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Homogeneity of variance

$$y_i = \underbrace{\beta_0 + \beta_1 \times x_i}_{\text{Linearity}} + \varepsilon_i \qquad \varepsilon_i \sim \underbrace{\mathcal{N}(0, \sigma^2)}_{\text{Normality}} \qquad \mathbf{V} = cov = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \cdots & \sigma^2 & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{pmatrix}$$

Zero covariance (=independence)

Dealing with Heterogeneity

y	x
41.9	1
48.5	2
43	3
51.4	4
51.2	5
37.7	6
50.7	7
65.1	8
51.7	9
38.9	10
70.6	11
51.4	12

Dealing with Heterogeneity

```
> data1 <- read.csv('../data/D1.csv')
> summary(data1)
```

```
       y               x        
 Min.   :34.90   Min.   : 1.00  
 1st Qu.:42.73   1st Qu.: 4.75  
 Median :51.30   Median : 8.50  
 Mean   :53.68   Mean   : 8.50  
 3rd Qu.:63.00   3rd Qu.:12.25  
 Max.   :95.30   Max.   :16.00  
```

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- estimate β_0, β_1 and σ^2

Dealing with Heterogeneity

Dealing with Heterogeneity

Dealing with Heterogeneity

$$
V = \text{cov} = \begin{pmatrix}
\sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0
\end{pmatrix}
$$

Dealing with Heterogeneity

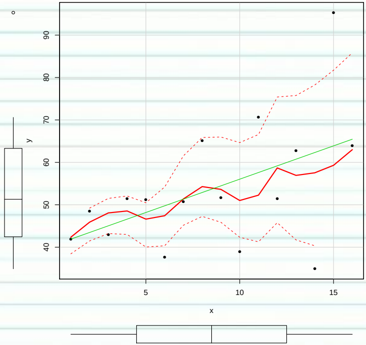
$$y_i = \underbrace{\beta_0 + \beta_1 \times x_i}_{\text{Linearity}} + \varepsilon_i \qquad \varepsilon_i \sim \underbrace{\mathcal{N}(0, \sigma^2)}_{\text{Normality}} \rightarrow \mathbf{V} = cov = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \cdots & \sigma^2 & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{pmatrix}$$

Homogeneity of variance (diagonal σ^2)

Zero covariance (=independence) (off-diagonal 0)

$$\mathbb{V} = \sigma^2 \times \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \vdots \\ \vdots & \cdots & 1 & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix}}_{\text{Identity matrix}} = \underbrace{\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \cdots & \sigma^2 & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{pmatrix}}_{\text{Variance-covariance matrix}}$$

Dealing with Heterogeneity



- variance proportional to X
- variance inversely proportional to X

Dealing with Heterogeneity

- variance inversely proportional to X

$$\mathrm{V} = \sigma^2 \times \mathrm{X} \times \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \vdots \\ \vdots & \cdots & 1 & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix}}_{\text{Identity matrix}} = \underbrace{\begin{pmatrix} \sigma^2 \times \frac{1}{\sqrt{\mathrm{X}_1}} & 0 & \cdots \\ 0 & \sigma^2 \times \frac{1}{\sqrt{\mathrm{X}_2}} & \cdots \\ \vdots & \cdots & \sigma^2 \times \frac{1}{\sqrt{\mathrm{X}_i}} \\ 0 & \cdots & \cdots & \sigma \end{pmatrix}}_{\text{Variance-covariance matrix}}$$

Dealing with Heterogeneity

$$\mathtt{V} = \sigma^2 \times \omega, \qquad \text{where } \omega = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{X_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{X_2}} & \cdots & \vdots \\ \vdots & \cdots & \frac{1}{\sqrt{X_i}} & \vdots \\ 0 & \cdots & \cdots & \frac{1}{\sqrt{X_n}} \end{pmatrix}}_{\text{Weights matrix}}$$

Dealing with Heterogeneity

Calculating weights

```
> 1/sqrt(data1$x)
```

```
 [1] 1.0000000 0.7071068 0.5773503 0.5000000 0.4472136 0.4082483 0.3779645 0.35
[10] 0.3162278 0.3015113 0.2886751 0.2773501 0.2672612 0.2581989 0.2500000
```

Generalized least squares (GLS)

1. use OLS to estimate fixed effects
2. use these estimates to estimate variances via ML
3. use these to re-estimate fixed effects (OLS)

Generalized least squares (GLS)

ML is biased (for variance) when N is small:

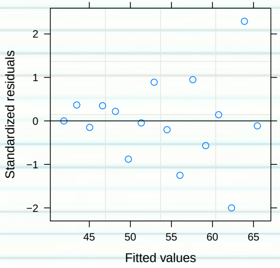
- use REML
- max. likelihood of residuals rather than data

Variance structures

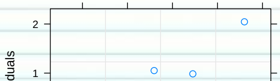
Variance function	Variance structure	Description
`varFixed( x)`	$V = \sigma^2 \times x$	variance proportional to ?x? (the covariate)
`varExp(form= x)`	$V = \sigma^2 \times e^{2\delta \times x}$	variance proportional to the exponential of ?x? raised to a constant power

Generalized least squares (GLS)

```
> library(nlme)
> data1.gls <- gls(y~x, data1,
+                  method='REML')
> plot(data1.gls)
```

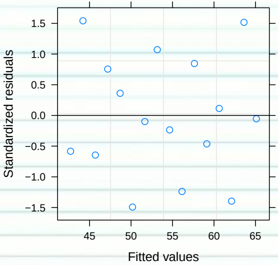



```
> library(nlme)
> data1.gls1 <- gls(y~x, data=data1, weights=varFixed(~x),
+                   method='REML')
> plot(data1.gls1)
```



Generalized least squares (GLS)

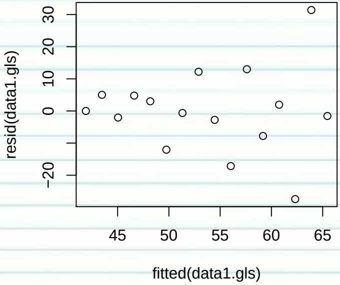
```
> library(nlme)
> data1.gls2 <- gls(y~x, data=data1, weights=varFixed(~x^2),
+                    method='REML')
> plot(data1.gls2)
```



Generalized least squares (GLS)

WRONG

```
> plot(resid(data1.gls) ~
+      fitted(data1.gls))
```

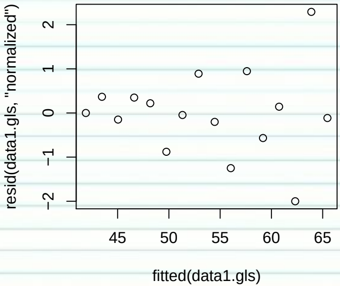



```
> plot(resid(data1.gls2) ~
+      fitted(data1.gls2))
```

Generalized least squares (GLS)

CORRECT

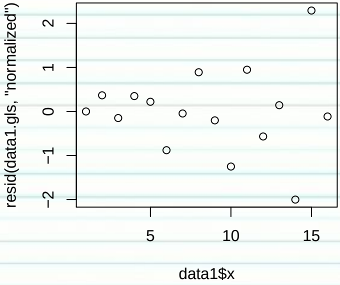
```
> plot(resid(data1.gls,'normalized') ~
+      fitted(data1.gls))
```



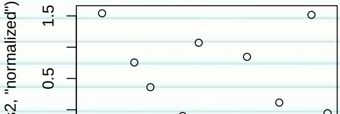

```
> plot(resid(data1.gls2,'normalized') ~
+      fitted(data1.gls2))
```

Generalized least squares (GLS)

```
> plot(resid(data1.gls,'normalized') ~ data1$x)
```




```
> plot(resid(data1.gls2,'normalized') ~ data1$x)
```



Generalized least squares (GLS)

```
> AIC(data1.gls, data1.gls1, data1.gls2)
```

```
            df      AIC
data1.gls    3 127.6388
data1.gls1   3 121.0828
data1.gls2   3 118.9904
```

```
> library(MuMIn)
> AICc(data1.gls, data1.gls1, data1.gls2)
```

```
            df     AICc
data1.gls    3 129.6388
data1.gls1   3 123.0828
data1.gls2   3 120.9904
```

```
> #OR
> anova(data1.gls, data1.gls1, data1.gls2)
```

```
           Model df      AIC      BIC    logLik
data1.gls      1  3 127.6388 129.5559 -60.81939
data1.gls1     2  3 121.0828 123.0000 -57.54142
data1.gls2     3  3 118.9904 120.9076 -56.49519
```

Generalized least squares (GLS)

```
> summary(data1.gls)
```

```
Generalized least squares fit by REML
  Model: y ~ x
  Data: data1
       AIC      BIC    logLik
  127.6388 129.5559 -60.81939

Coefficients:
               Value Std.Error  t-value p-value
(Intercept) 40.33000  7.189442 5.609615  0.0001
x            1.57074  0.743514 2.112582  0.0531

 Correlation:
  (Intr)
x -0.879

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.00006105 -0.29319830 -0.02282621  0.35357567  2.29099872

Residual standard error: 13.70973
Degrees of freedom: 16 total; 14 residual
```

```
> summary(data1.gls2)
```

Generalized least squares (GLS)

```
> data1$cx <- scale(data1$x, scale=FALSE)
> data1.gls <- gls(y~cx, data1,
+                  method='REML')
> summary(data1.gls)
```

```
Generalized least squares fit by REML
  Model: y ~ cx
  Data: data1
       AIC      BIC    logLik
  127.6388 129.5559 -60.81939

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 53.68125  3.427432 15.662236  0.0000
cx           1.57074  0.743514  2.112582  0.0531

 Correlation:
   (Intr)
cx 0

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.00006105 -0.29319830 -0.02282621  0.35357567  2.29099872

Residual standard error: 13.70973
```

Section 2

Heteroscadacity in ANOVA

Heteroscadacity in ANOVA

```
> data2 <- read.csv('../data/D2.csv')
> summary(data2)
```

```
       y              x   
 Min.   :29.29   A:10  
 1st Qu.:36.17   B:10  
 Median :40.89   C:10  
 Mean   :42.34   D:10  
 3rd Qu.:49.32   E:10  
 Max.   :56.84         
```

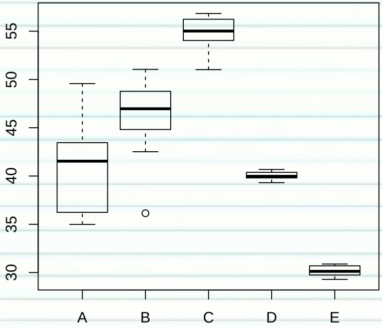
$$y_i = \mu + \alpha_i + \varepsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- estimate μ, α_i and σ^2

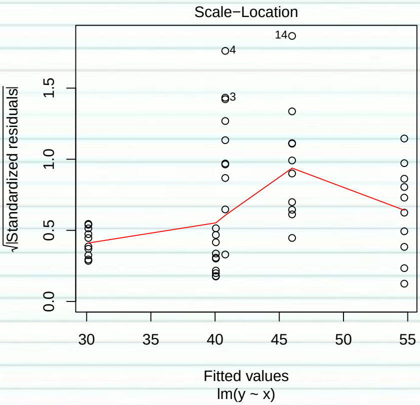
Heteroscadacity in ANOVA

```
> boxplot(y~x, data2)
```



Heteroscadacity in ANOVA

```
> plot(lm(y~x, data2), which=3)
```



Heteroscadacity in ANOVA

$$\varepsilon \sim \mathcal{N}(0, \sigma_{i}^{2} \times \omega)$$

Effect (α) 1 (i=1) $\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} \quad \varepsilon_i \sim \mathcal{N}(0,$

Effect (α) 2 (i=2) $\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} \quad \varepsilon_i \sim \mathcal{N}(0,$

Effect (α) 3 (i=3) $\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} \quad \varepsilon_i \sim \mathcal{N}(0,$

Heteroscadacity in ANOVA

$$\mathbb{V}_{ji} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_2^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_2^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_3^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_3^2 \end{pmatrix}$$

Heteroscadacity in ANOVA

```
> data2.sd <- with(data2, tapply(y,x,sd))
> 1/(data2.sd[1]/data2.sd)
```

```
         A          B          C          D          E
1.00000000 0.91342905 0.40807277 0.08632027 0.12720488
```

Variance structures

Variance function	Variance structure	Description
`varFixed( x)`	$V = \sigma^2 \times x$	variance proportional to □x□ (the covariate)
`varExp(form= x)`	$V = \sigma^2 \times e^{2\delta \times x}$	variance proportional to the exponential of □x□ raised to a constant power

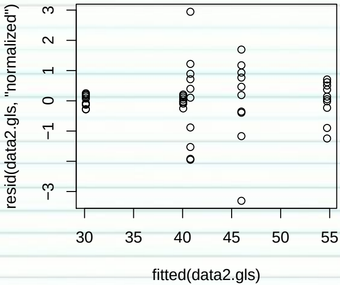
Heteroscadacity in ANOVA

```
> library(nlme)
> data2.gls <- gls(y~x, data=data2,
+                  method="REML")
```

```
> library(nlme)
> data2.gls1 <- gls(y~x, data=data2,
+    weights=varIdent(form=~1|x), method="REML")
```

Heteroscadacity in ANOVA

```
> library(nlme)
> data2.gls <- gls(y~x, data=data2,
+                method="REML")
> plot(resid(data2.gls,'normalized') ~
+      fitted(data2.gls))
```




```
> library(nlme)
> data2.gls1 <- gls(y~x, data=data2,
+    weights=varIdent(form=~1|x), method="REML")
> plot(resid(data2.gls1,'normalized') ~
+      fitted(data2.gls1))
```

Heteroscadacity in ANOVA

```
> AIC(data2.gls,data2.gls1)
```

```
            df      AIC
data2.gls    6 249.4968
data2.gls1  10 199.2087
```

```
> anova(data2.gls,data2.gls1)
```

```
           Model df      AIC      BIC     logLik   Test  L.Ratio p-value
data2.gls      1  6 249.4968 260.3368 -118.74841
data2.gls1     2 10 199.2087 217.2753  -89.60435 1 vs 2 58.28812  <.0001
```

- note: it costs d.f.

Heteroscadacity in ANOVA

```
> summary(data2.gls)

Generalized least squares fit by REML
  Model: y ~ x
  Data: data2
       AIC      BIC    logLik
  249.4968 260.3368 -118.7484

Coefficients:
                Value Std.Error  t-value p-value
(Intercept)  40.79322 0.9424249 43.28538  0.0000
xB            5.20216 1.3327901  3.90321  0.0003
xC           13.93944 1.3327901 10.45884  0.0000
xD           -0.73285 1.3327901 -0.54986  0.5851
xE          -10.65908 1.3327901 -7.99757  0.0000

 Correlation:
   (Intr) xB     xC     xD
xB -0.707
xC -0.707  0.500
xD -0.707  0.500  0.500
xE -0.707  0.500  0.500  0.500

Standardized residuals:
        Min          Q1          Med          Q3          Max
```

Section 3

Worked Examples

Worked Examples