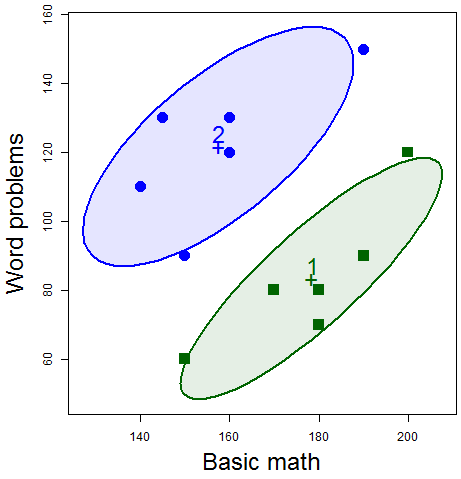
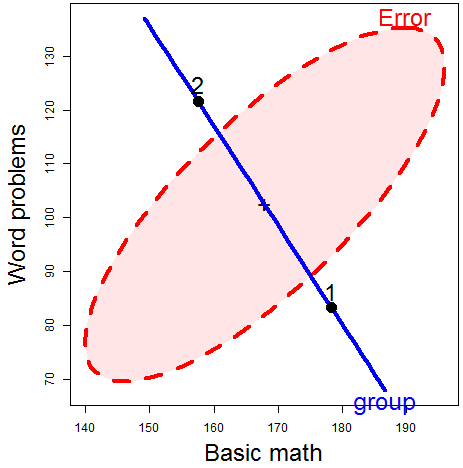
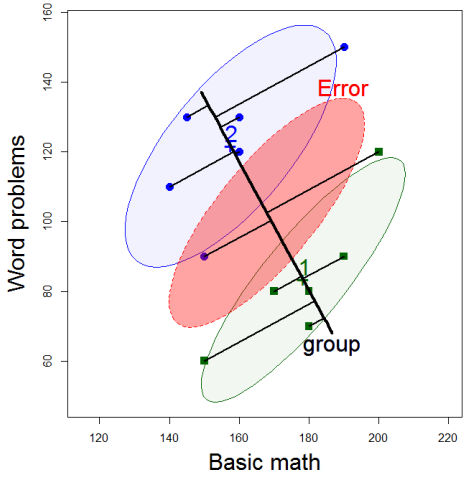
Data ellipses



HE plot





Visualizing Linear Models: An R Bag of Tricks Session 2: Multivariate Models

Michael Friendly
SCS Short Course
Oct-Nov 2021

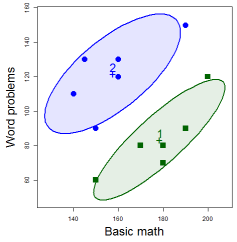
https://friendly.github.io/VisMLM-course/

Today's topics

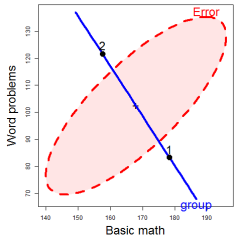
- Brief review of the GLM & MLM

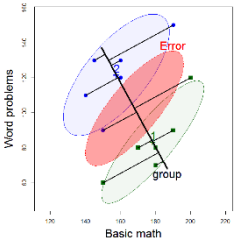
$$\underset{(n\times p)}{\mathbf{Y}} = \underset{(n\times q)}{\mathbf{X}}\ \underset{(q\times p)}{\mathbf{B}} + \underset{(n\times p)}{\mathcal{E}}$$

- Data ellipses
  - sufficient visual summaries

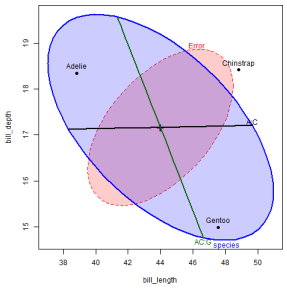


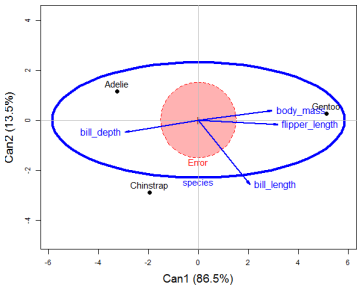
- HE plot framework
  - H & E matrices/ellipses
  - Discriminant/canonical views





- Example: Penguins data





- Checking assumptions

One-way MANOVA

- *p* responses, 1 "factor" (IV), *g* groups

  H_o: $\underline{\boldsymbol{\mu}}_1 = \underline{\boldsymbol{\mu}}_2 = \ldots \underline{\boldsymbol{\mu}}_g$

  H_1: at least one group centroid is different

- Assumptions:
  - Independent groups, independent observations
  - Responses are independent, multivariate normal w/in each group
  - Pop. within-group covariance matrices are **equal** across groups
    - H_0: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g$
    - ($\boldsymbol{\Sigma}$ estimated by $\mathbf{S} = \mathbf{E} / df_e$)
    - tested by e.g., Box's test, **heplots::boxM**
  - $\rightarrow \mathbf{y}_{ij\ (p \times 1)} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

One-way ANOVA vs. MANOVA

ANOVA

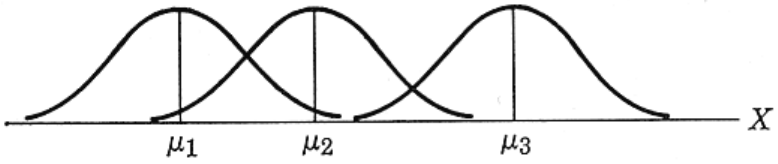


Figure 8.1. The simple anova situation, when the differences among the populations are "real."

source: Cooley & Lohnes ((1971)

Do means differ?

(Assume equal within-group variances)

MANOVA

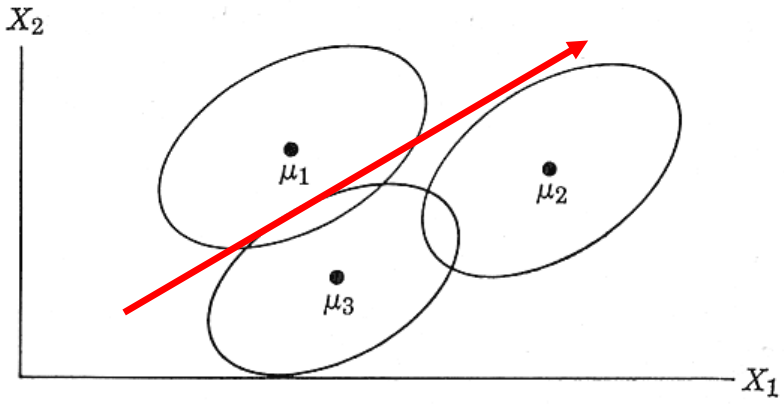


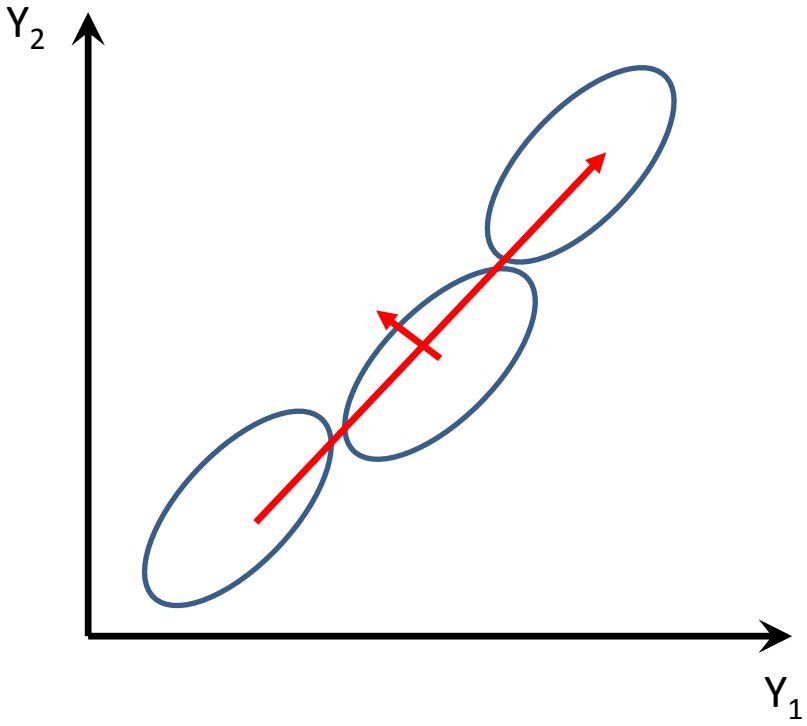
Figure 8.2. The simple manova situation, when the differences among the populations are "real."

How do centroids differ?
How many dimensions?

(Assume equal within-group variance-covariance matrices)

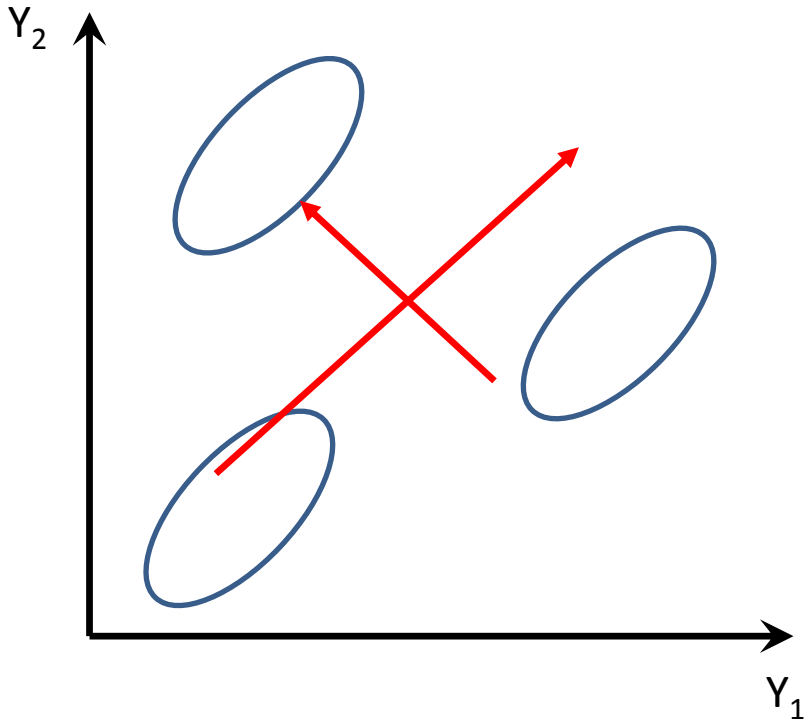
Response dimensions

Means on Y_1 and Y_2 are nearly perfectly correlated



Only 1 dimension required to understand the group effect

Means on Y_1 and Y_2 have a low correlation



Two different aspects are reflected in group means

GLM: the design matrix (X)

- In the full GLM, the design matrix (**X**) may consist of:
  - A constant, **1**, for the intercept (usually implicit)
  - Quantitative regressors: age, income, education
  - Transformed regressors: √age, log(income)
  - Polynomial terms: age^2, age^3, …
  - Categorical predictors ("factors", class variables): treatment (control, drug A, drug B), sex
  - Interactions: treatment * sex, age * sex

Model formulae in R define y ~ X:

```
prestige ~ income + education                    # 2 main effects
prestige ~ income * education                    #  + interaction
prestige ~ income + education + women + type  # 4 main effects
prestige ~ education + poly(women, 2) + log(income)*type
```

Univariate linear model

- Model $\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times q)}{\mathbf{X}}\ \underset{(1\times q)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\epsilon}}$

$$\underset{(n\times q)}{\mathbf{X}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q)$$

matrix of predictors, factors, …

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \epsilon_i$$

- Sums of squares

data fit residuals

$$SS_{\text{Tot}} = \sum_{i,j} (y_{i,j} - \bar{y}_i)^2 + \sum_{i,j} (y_{i,j} - \hat{y}_i)^2$$

$$= SS_H + SS_E$$

- Hypothesis tests

$$F = \frac{SS_H / df_H}{SS_E / df_E} = \frac{MS_H}{MS_E}$$

How big is hypothesis variation relative to error variation?

mean square is a variance estimate

Least squares: SS_T and SS_E

In simple linear regression,

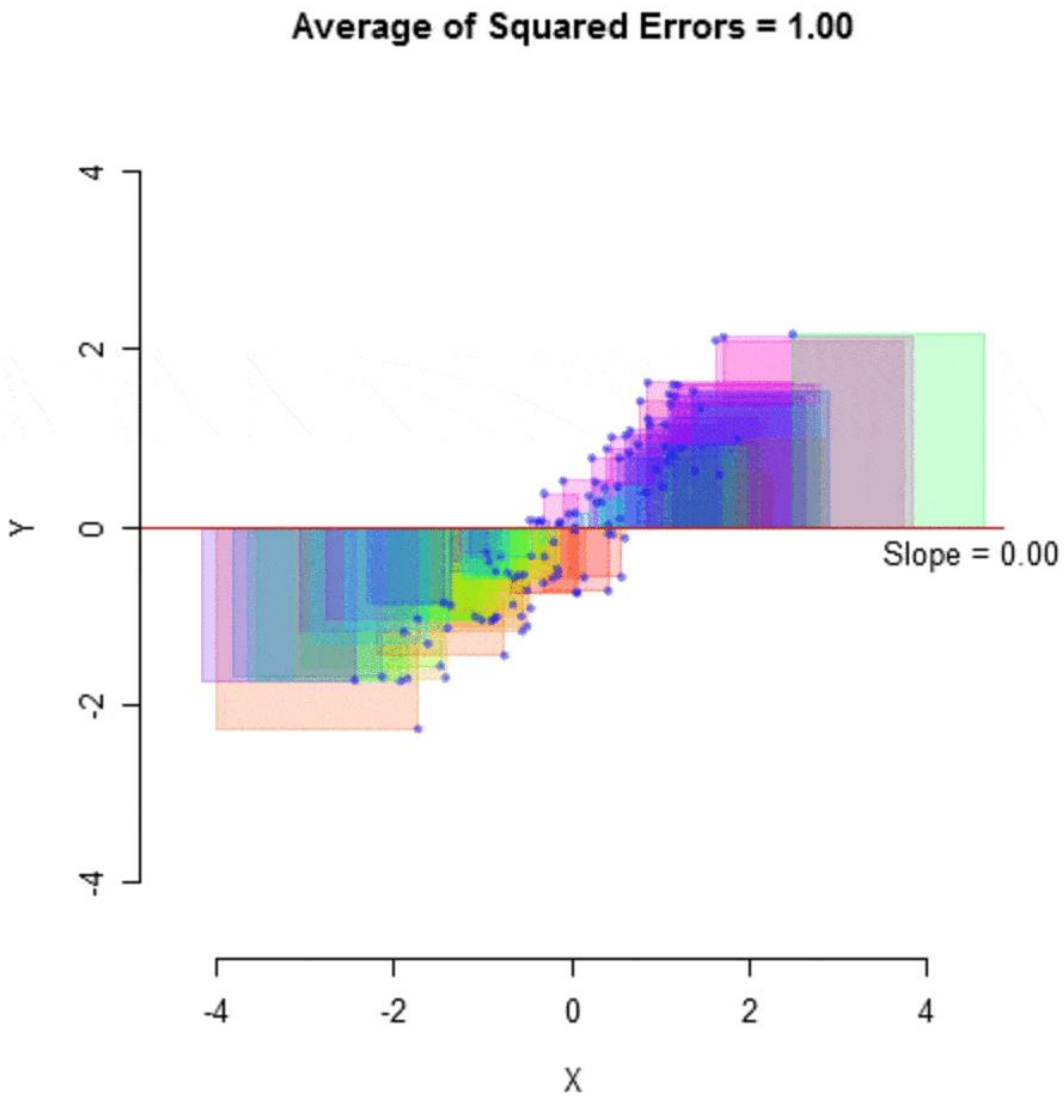
$$y_i = b_0 + b_1 \times x_i + e_i$$

the intercept b_0 & slope b_1 are values that minimize the SS_E (or MS_E)

$$SS_E = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

SS_T is that value when b_1=0

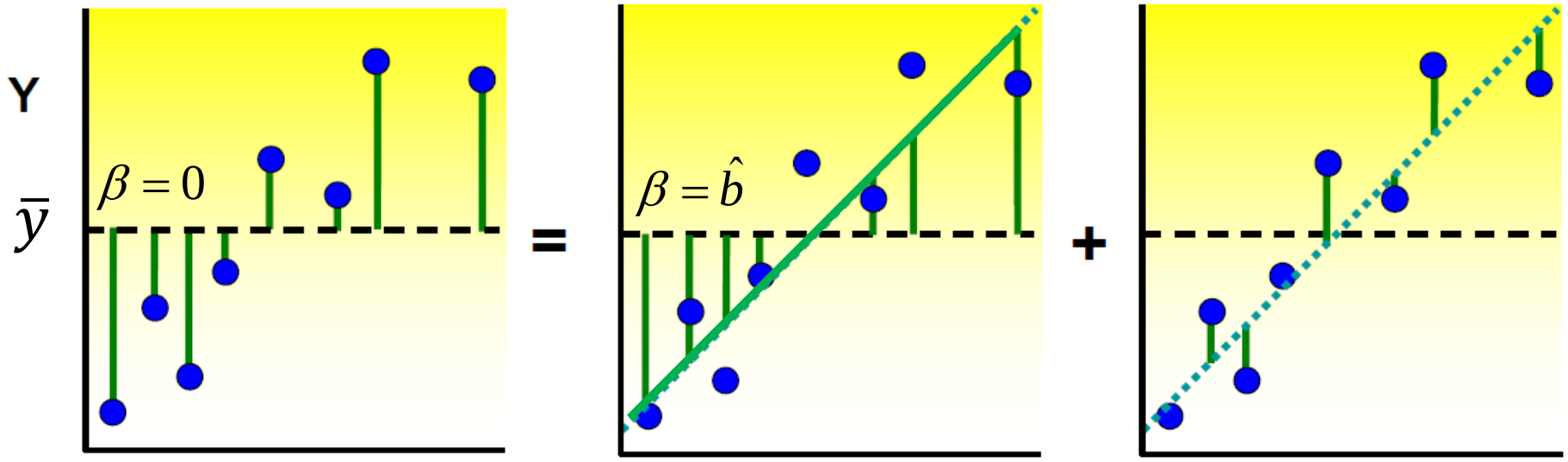
b_1	MS_E
.00	1.0
.89	0.2



Regression: Visualizing $SS_T = SS_H + SS_E$

Total variance (SS_T) = Regression variance (SS_H) + Residual variance (SS_E)

$$\Sigma_i (y_i - \bar{y})^2 \qquad \Sigma_i (\hat{y}_i - \bar{y})^2 \qquad \Sigma_i (y_i - \hat{y}_i)^2$$

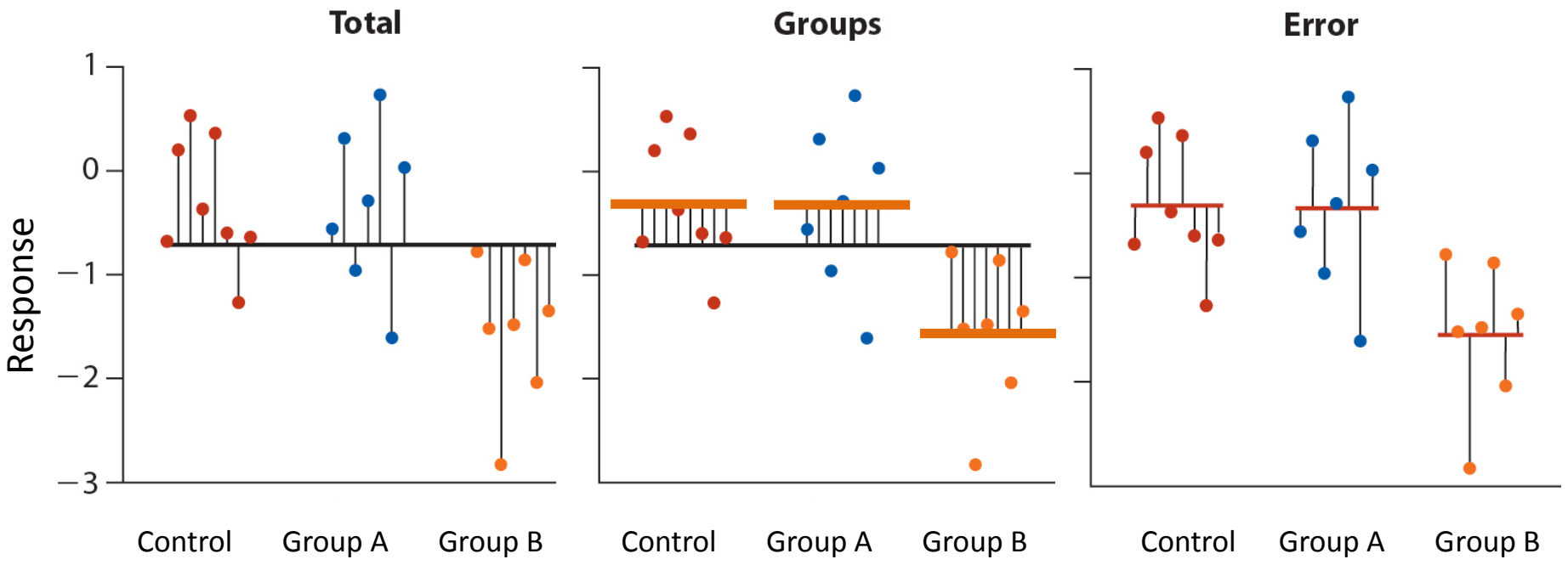


F test: How much better is the fitted regression line ($\beta = \hat{b}$) than the flat line **($\beta = 0$) ?**

ANOVA: Visualizing $SS_T = SS_H + SS_E$

Total variance = Between group variance + Within group variance

$$\Sigma_{ij}(y_{ij} - \bar{y}_{..})^2 \qquad \Sigma_{ij}(\bar{y}_{.j} - \bar{y}_{..})^2 \qquad \Sigma_{ij}(y_{ij} - \bar{y}_{.j})^2$$



F test: How much better is the groups model than the model ignoring groups?

Which means differ?

- In ANOVA, when a factor is significant, follow-up to find which means differ
- Post-hoc tests:
  - all-pairwise comparisons
  - all treatments vs. control group
- Need to correct for multiple testing– control family-wise error rate
  - Bonferroni: $\alpha_i = \alpha_{FW} / k$ [too conservative]
  - Tukey pairwise: "honestly significant difference"
  - many others: Dunnett's test, Sidak, FDR, …

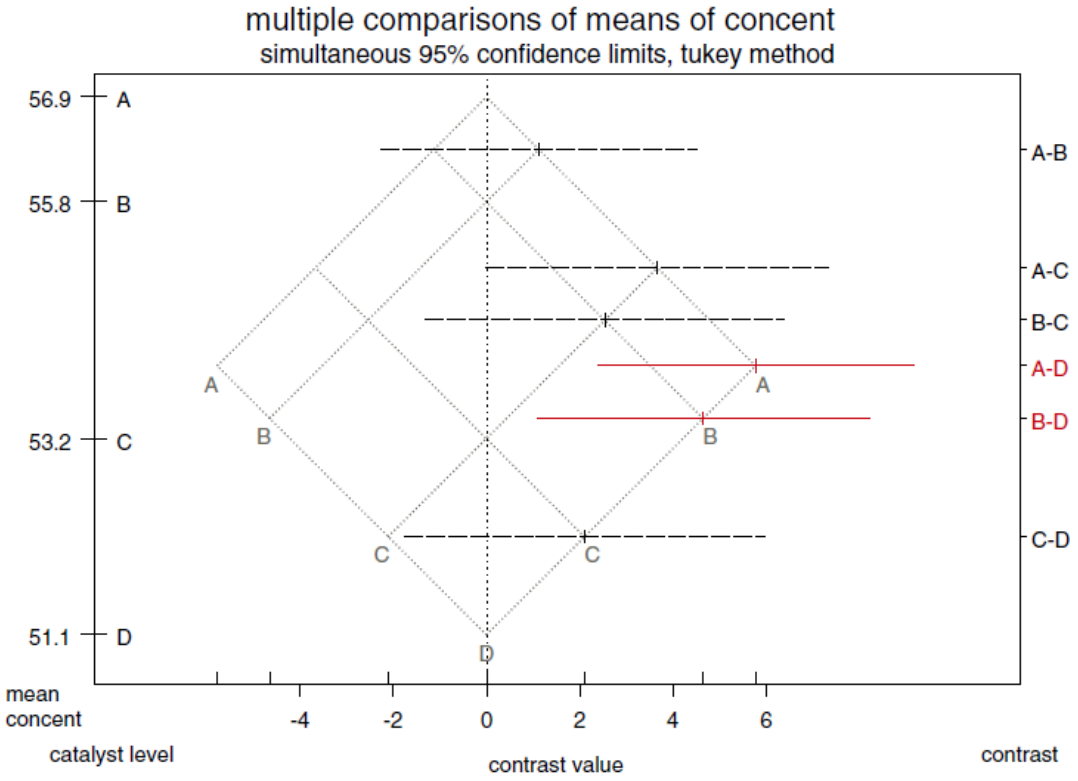
Plotting multiple comparisons

`HH::mmcplot()` – the mean-mean multiple comparison plot shows multiple comparisons or contrasts for any linear model

```
library(HH)
catalystm.lm <- lm(concent ~ catalyst, data=catalystm)
catalystm.mmc <-mmc(catalystm.lm, linfct = mcp(catalyst = "Tukey"))
mmcplot(catalystm.mmc)
```

Construction:

- plot means, $\overline{y}_i$, $\overline{y}_j$ on grid
- rotate 45°
- horizontal axis shows: $\overline{y}_l - \overline{y}_j$
- SE determined by MC method
- signif. comparisons highlighted



Contrasts: planned comparisons

- Better to test specific, planned comparisons, rather than all-pairwise
- A contrast is a weighted sum, *L*, of the means, with weights, **c**, that sum to zero

$$L = \mathbf{c}'\mu = \Sigma\, c_i \mu_i \quad \text{such that} \quad \Sigma\, c_i = 0$$

r=4 groups

$$L_1 = (\mu_1 + \mu_2) - (\mu_3 + \mu_4) \rightarrow \mathbf{c}_1 = (1 \quad 1 \quad -1 \quad -1)'$$
$$L_2 = \mu_1 - \mu_2 \rightarrow \mathbf{c}_2 = (1 \quad -1 \quad 0 \quad 0)'$$
$$L_3 = \mu_3 - \mu_4 \rightarrow \mathbf{c}_3 = (0 \quad 0 \quad 1 \quad -1)'$$

- In words: average of one subset of groups vs. another subset
- Any *r*-1 linearly independent contrasts → same overall test
- A priori contrasts can be tested w/o adjusting α

The **X** matrix for a factor can be represented by a set of *r-1* contrasts, combined with the unit vector

$$\mathbf{X}_{(r \times r)} = (\mathbf{1}, \mathbf{C})$$

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

c1 c2 c3

Some special contrasts:

Deviation contrasts

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix}$$

each treatment vs control or baseline [not orthogonal]

Helmert contrasts

$$\mathbf{C} = \begin{pmatrix} 3 & 0 & 0 \\ -1 & 2 & 0 \\ -1 & -1 & 1 \\ -1 & -1 & -1 \end{pmatrix}$$

ordered treatments: each vs all the rest [always orthogonal]

Polynomial contrasts

$$\mathbf{C} = \begin{pmatrix} -3 & 1 & -1 \\ -1 & -1 & 3 \\ 1 & -1 & -3 \\ 3 & 1 & 1 \end{pmatrix}$$

lin quad cubic

quantitative treatment levels [orthogonal]

Using contrasts in R

- R has 4 basic functions for generating contrasts for a factor
  - Dummy coding, aka "reference level", "treatment" contrasts
  - Deviation coding, aka "sum-to-zero" constraints
  - Polynomial contrasts for an ordered/quantitative factor
  - Helmert contrasts for ordered factor comparisons
- Defaults are set separately for unordered and ordered factors
- Define your own by assigning a matrix to `contrasts(myfactor)<- cmat`
- These affect the tests of coefficients, but not overall tests

```
> contr.treatment(4)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

```
> contr.sum(4)
  [,1] [,2] [,3]
1    1    0    0
2    0    1    0
3    0    0    1
4   -1   -1   -1
```

```
> contr.poly(4)
           .L   .Q      .C
[1,] -0.6708  0.5 -0.2236
[2,] -0.2236 -0.5  0.6708
[3,]  0.2236 -0.5 -0.6708
[4,]  0.6708  0.5  0.2236
```

```
> options("contrasts")
$contrasts
        unordered           ordered
"contr.treatment"      "contr.poly"
```

```
> contr.helmert(4)
  [,1] [,2] [,3]
1   -1   -1   -1
2    1   -1   -1
3    0    2   -1
4    0    0    3
```

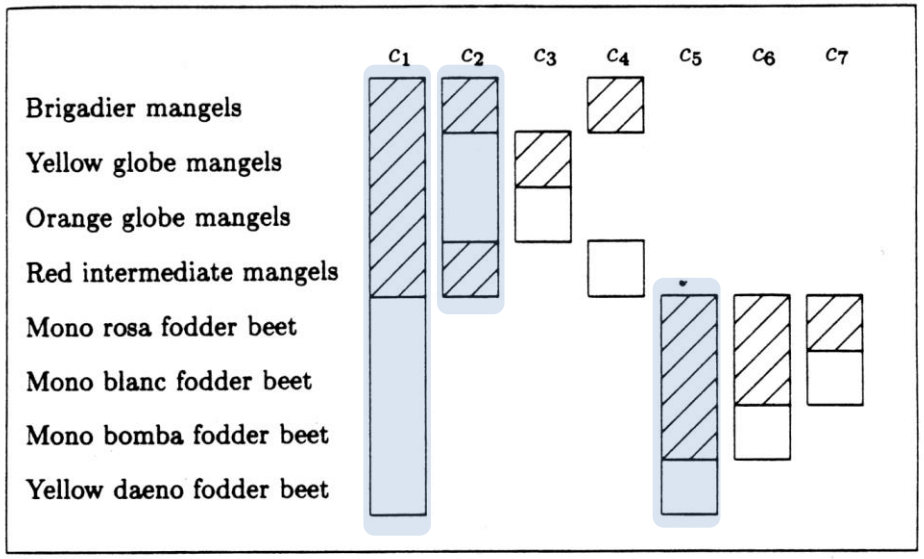
See: http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm

Nested dichotomies

- Orthogonal contrasts can always be generated as nested dichotomies
- They correspond to independent research questions
- Sums of squares decompose the overall effect

$$SS_A = SS_{c1} + SS_{c2} + \ldots + SS_{c(r-1)}$$

c_1 = mangles vs beets c_2 = globe mangles vs other c_5 = mono beets vs yellow



Treatment		c_1	c_2	c_3	c_4	c_5	c_6	c_7
Brigadier mangels	μ_1	1	1	0	1	0	0	0
York globe mangels	μ_2	1	−1	1	0	0	0	0
Orange globe mangels	μ_3	1	−1	−1	0	0	0	0
Red intermediate mangels	μ_4	1	1	0	−1	0	0	0
Mono rosa fodder beet	μ_5	−1	0	0	0	1	1	1
Mono blanc fodder beet	μ_6	−1	0	0	0	1	1	−1
Mono bomba fodder beet	μ_7	−1	0	0	0	1	−2	0
Yellow daeno fodder beet	μ_7	−1	0	0	0	−3	0	0

Multivariate linear model

- Model $\underset{(n\times p)}{\mathbf{Y}} = \underset{(n\times q)}{\mathbf{X}}\ \underset{(q\times p)}{\mathbf{B}} + \underset{(n\times p)}{\mathcal{E}}$

$\underset{(n\times p)}{\mathbf{Y}} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p)$

matrix of *p* responses

- Sums of squares & cross-products

$$\underset{(p\times p)}{\mathbf{SSP}_T} = \left(\hat{\mathbf{Y}}'\hat{\mathbf{Y}} - n\bar{\mathbf{y}}\bar{\mathbf{y}}'\right) + \mathcal{E}'\mathcal{E}$$

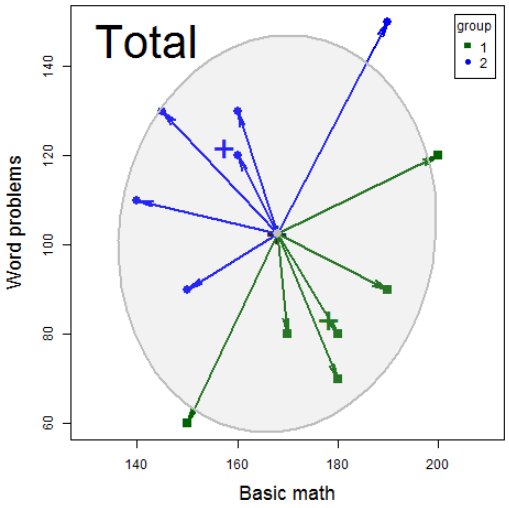
$$= \mathbf{SSP}_H + \mathbf{SSP}_E = \mathbf{H} + \mathbf{E}$$

- Hypothesis tests
  - Eigenvalues λ_i, *i=1:p* of H E^{-1}
  - Wilks' Λ, Pillai & Hotelling trace, Roy's test
  - how many dimensions (aspects of responses)?

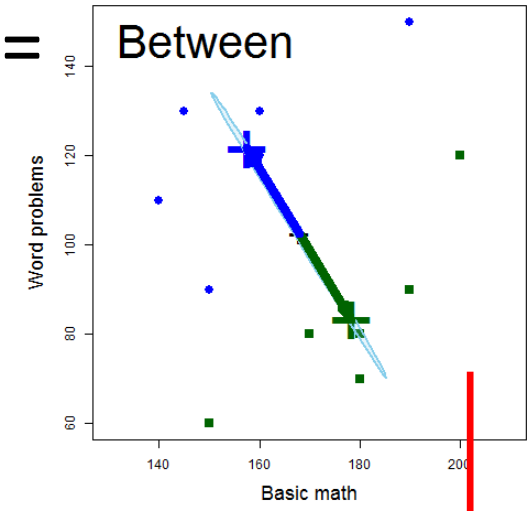
How big is hypothesis variation relative to error variation?

Ah, but there are up to $s = min(p, df_h)$ dimensions of size

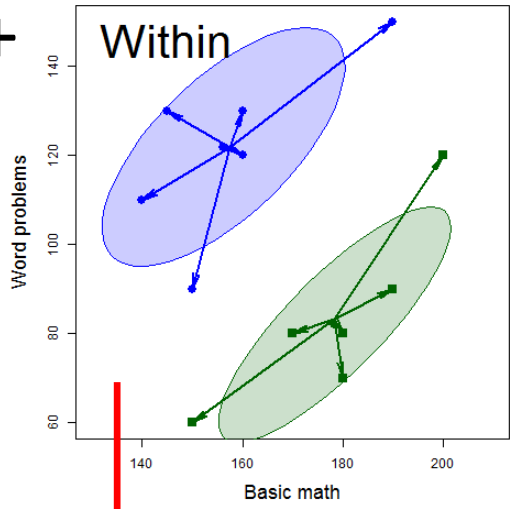
Visualizing $SSP_T = SSP_H + SSP_E$



=

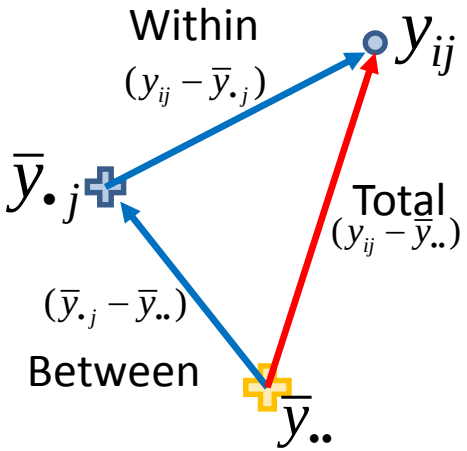


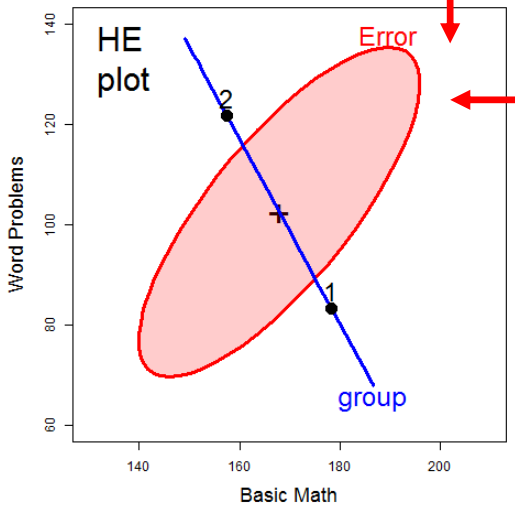
+

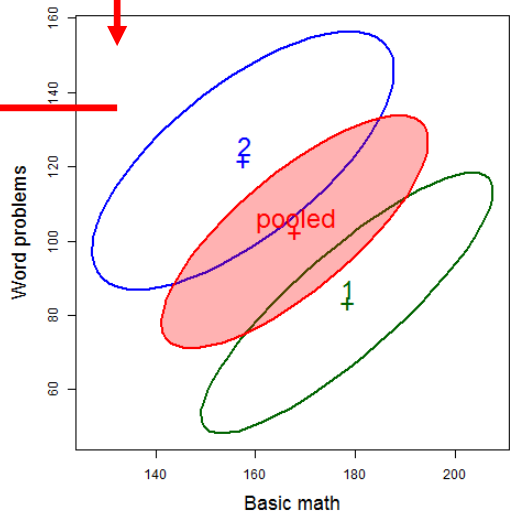


scale & overlay

pool→ SSP_E



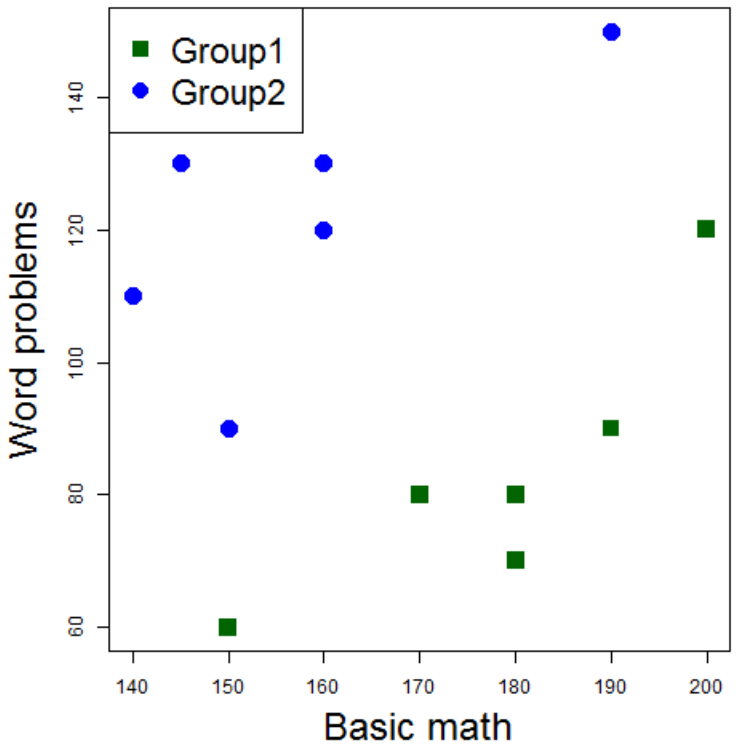




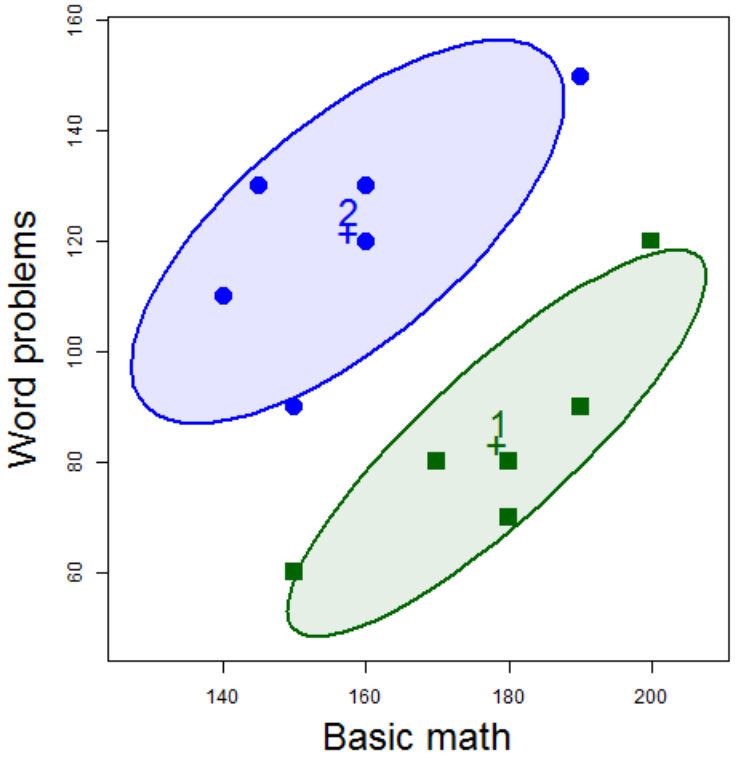
Data ellipsoids

The data ellipsoid is a sufficient visual summary for multivariate location & scatter, just as $(\bar{y}, \mathbf{S})$ are sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Data



Data ellipses



Data ellipsoids: definitions

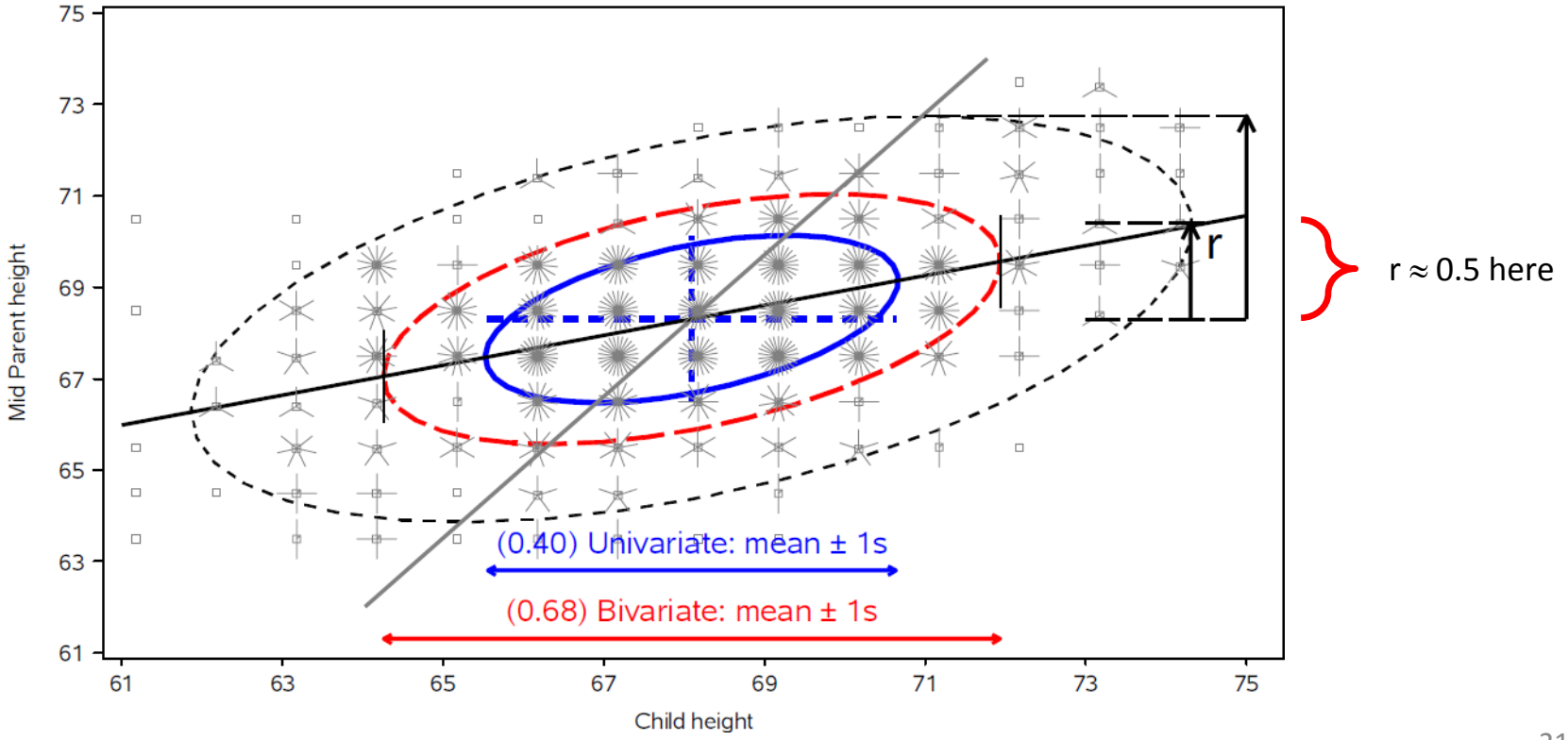
- For a *p*-dimensional multivariate sample, $\mathbf{Y}_{N\times p}$, the sample mean vector, $\bar{\mathbf{y}}$, and sample covariance matrix, **S**, are minimally sufficient statistics under classical (gaussian) assumptions.
- These can be represented visually by the *p*-dimensional data ellipsoid, $\mathcal{E}_c$ of size ("radius") *c* centered at $\bar{\mathbf{y}}$,

$$\mathcal{E}_c(\bar{\mathbf{y}}, \mathbf{S}) := \{\mathbf{y} : (\mathbf{y} - \bar{\mathbf{y}})^T \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \le c^2\} \qquad \text{or,} \quad D_M^2(\mathbf{y}) \le c^2$$

- $\rightarrow$ an ellipsoid centered at the means whose size & shape reflects variances & covariances
- We consider this a minimally sufficient visual summary of multivariate location and scatter.

Data ellipsoids: properties

- Ellipsoid boundary: Mahalanobis $D_M^2(y_i) \sim \chi_p^2$
  - *p*=2: shadows generalize univariate confidence intervals
  - eccentricity: precision; visual estimate of correlation



The HE plot framework

- Hypothesis-Error (HE) plots
  - Visualize multivariate tests in the MLM
  - Linear hypotheses--- lower-dimensional ellipsoids
  - Extension: HE plot matrices
- Canonical displays
  - low-dimensional multivariate juicers
  - shows data in the space of maximal effects
- Covariance ellipsoids
  - visualize tests of homogeneity of covariance matrices
- For all: robust methods are available or good research projects!

HE plot framework: Trivial example

Two groups of middle-school students are taught algebra by instructors using different methods, and then tested on:

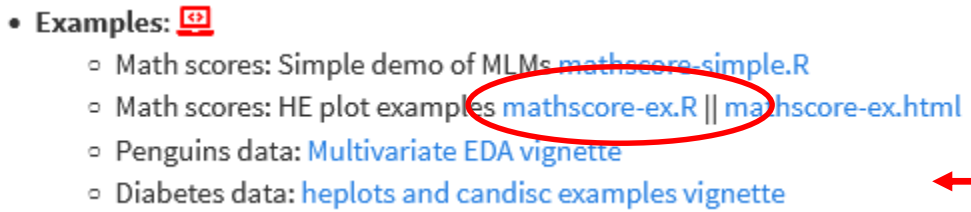
- **BM**: basic math problems (7 * 23 – 2 * 9 = ?)
- **WP**: word problems ("a train travels at 23 mph for 7 hours, but for 2 hours …")

Do the groups differ on (BM, WP) by a multivariate test?
If so, how ???

```
> data(mathscore, package="heplots")
> mod <- lm(cbind(BM, WP) ~ group, data=mathscore)
> Anova(mod)

Type II MANOVA Tests: Pillai test statistic
      Df test stat approx F num Df den Df    Pr(>F)
group  1   0.86518   28.878      2      9 0.0001213 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

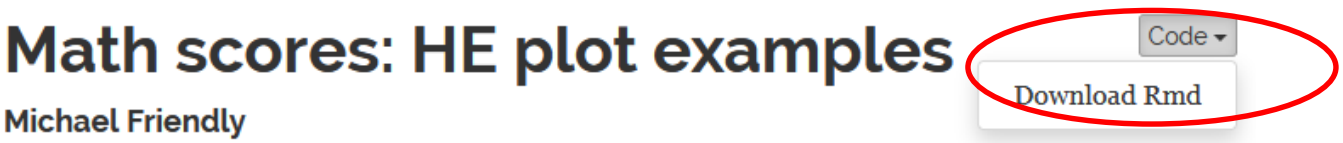
Follow along

The R script (mathscore-ex.R) for this example is linked on the course page. Download and open in R Studio to follow along.



HW: explore other examples

The script was run with `knitr` (ctrl+shift+K) in R Studio to create the HTML output (mathscore-ex.html)
The **Code** button there allows you do download the R code and comments



(R notebooks are a simple way to turn R scripts into finished documents)

Why do multivariate tests?

Could do univariate ANOVAs (or t-tests) on each response variable (BM, WP)

```
> Anova(lm(BM ~ group, data=mathscore))
Anova Table (Type II tests)

Response: BM
          Sum Sq Df F value Pr(>F)
group       1302  1    4.24  0.066 .
Residuals   3071 10
```

✘

```
> Anova(lm(WP ~ group, data=mathscore))
Anova Table (Type II tests)

Response: WP
          Sum Sq Df F value Pr(>F)
group       4408  1    10.4  0.009 **
Residuals   4217 10
```

✓

From this, might conclude that:

- Groups don't differ on Basic Math score ✘
- Groups are significantly different on Word problems ✓

Multivariate tests:

- Do not require correcting for multiple tests (e.g., Bonferroni)
- Combine evidence from multiple response variables ("pooling strength")
- Show how the multivariate responses are jointly related to the predictors
  - How many aspects (dimensions?)

Why do multivariate tests?

Overall test is highly significant:

- Combines the evidence for all predictors ✓
- Takes response correlations into account ✓

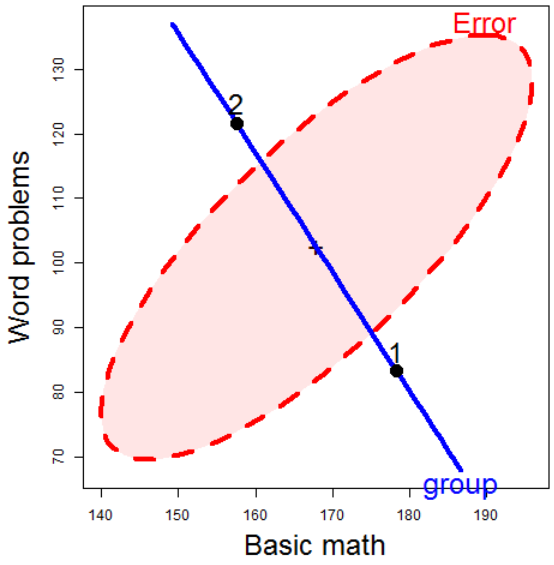
```
> mod <- lm(cbind(BM, WP) ~ group, data=mathscore)
> Anova(mod)

Type II MANOVA Tests: Pillai test statistic
      Df test stat approx F num Df den Df    Pr(>F)
group  1   0.86518   28.878      2      9 0.0001213 ***
```

Visual test of significance (Roy's test)

- The **H** ellipse projects outside the **E** ellipse iff the effect is significant.

HE plot provides an interpretation:

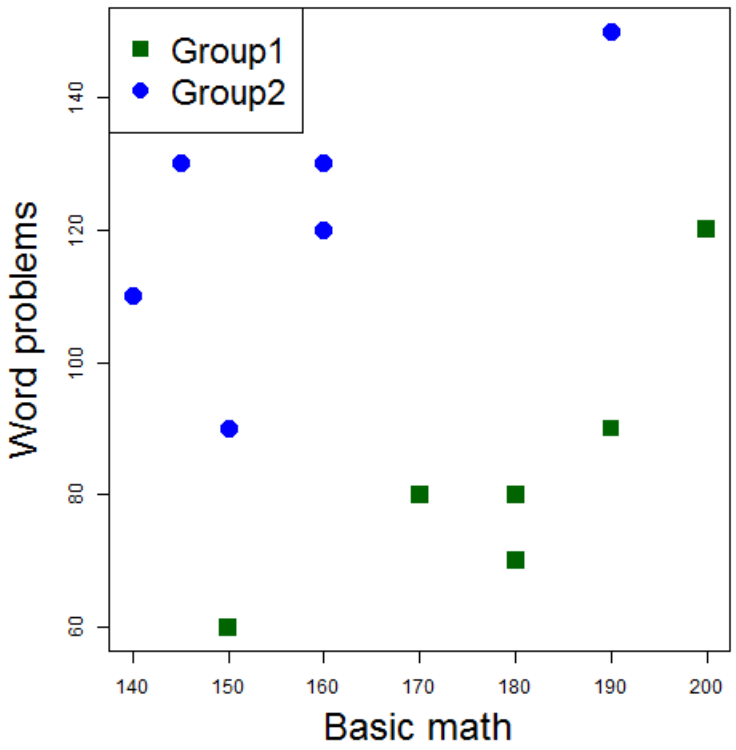
- Group 1 > Group 2 on Basic Math, but worse on Word Problems
- Group 2 > Group 1 on Word Problems, but worse on Basic Math
- BM & WP are + correlated w/in groups



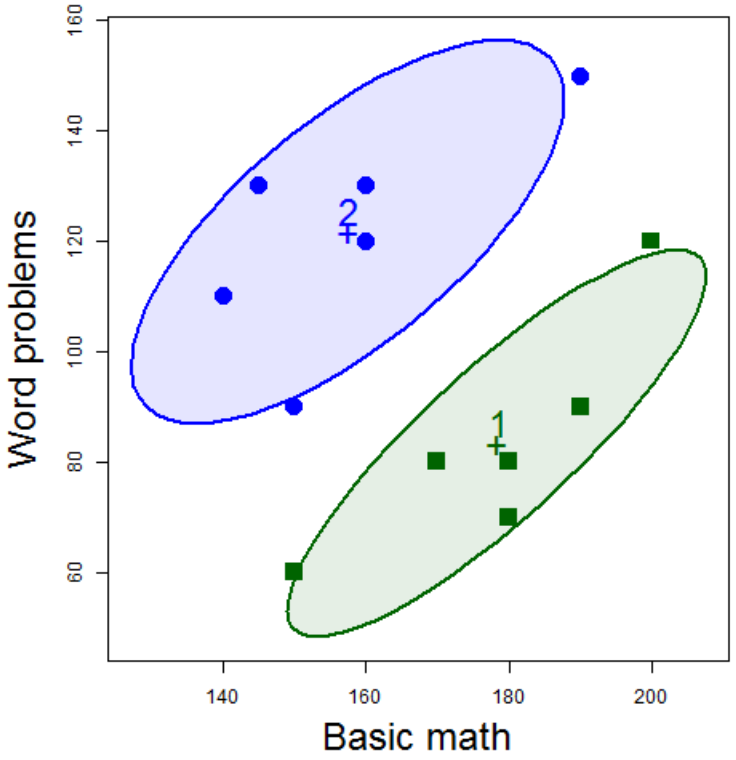
HE plot framework: Visual overview

The data ellipsoid is a sufficient visual summary for multivariate location & scatter, just as $(\bar{y}, \mathbf{S})$ are sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Data

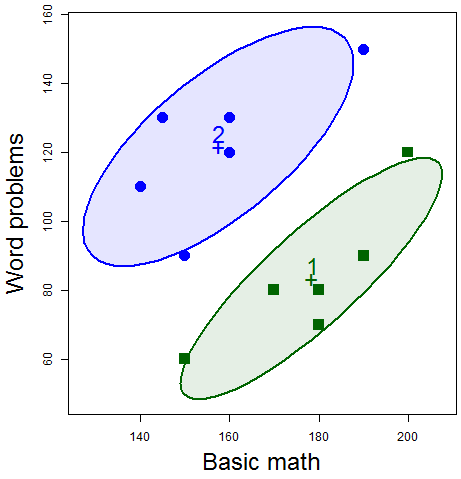


Data ellipses

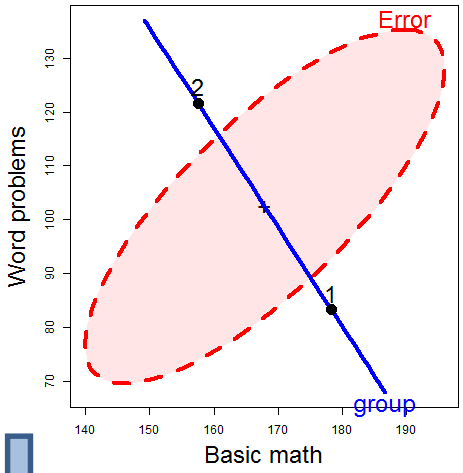


Visual overview

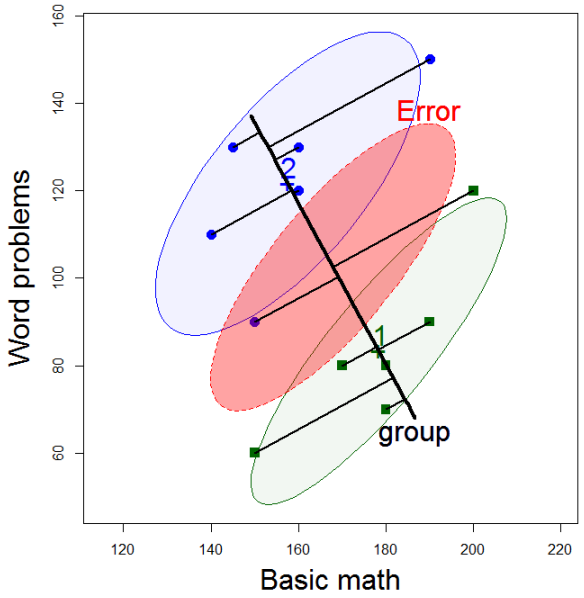
Data ellipses



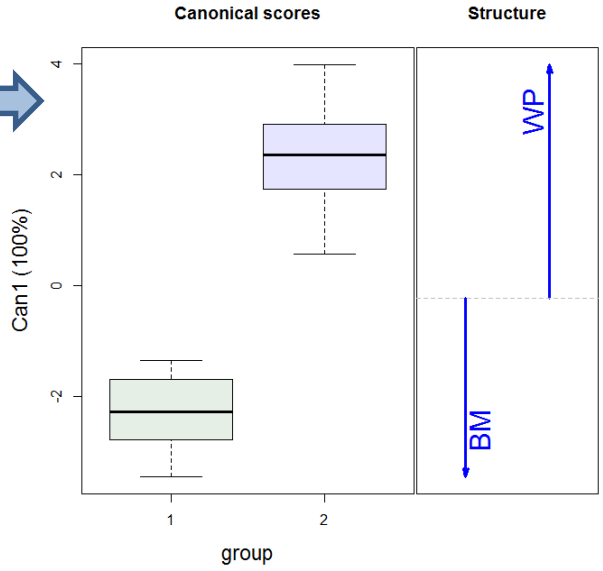
HE plot



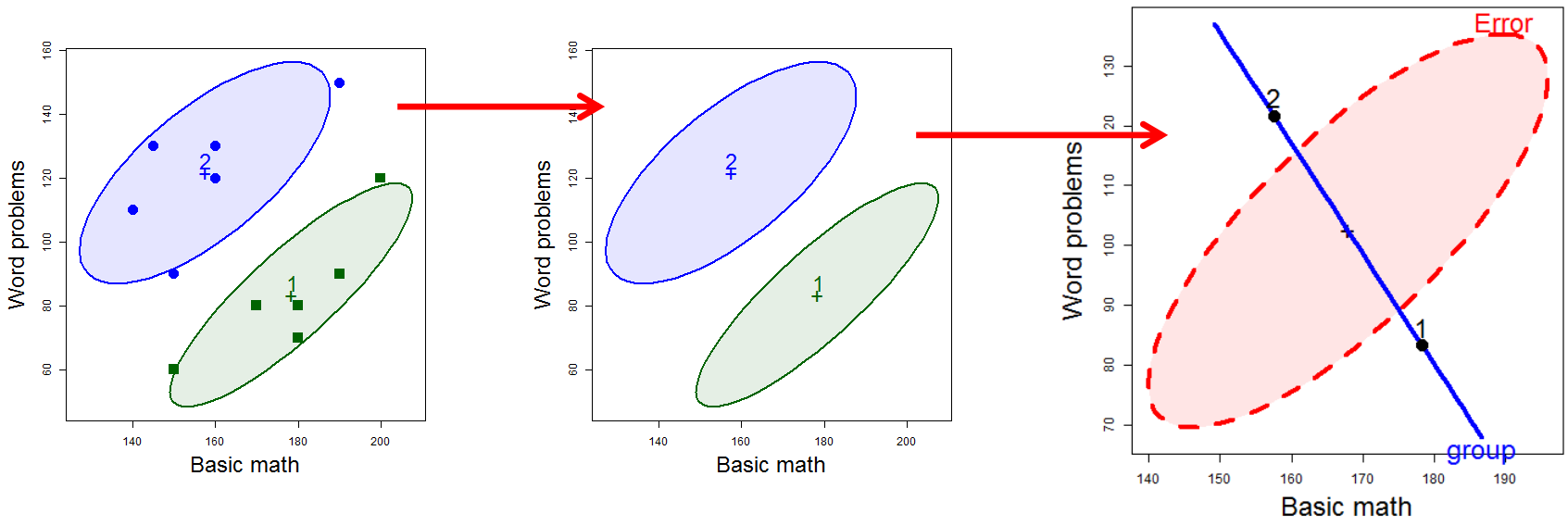
Discriminant scores



Canonical space

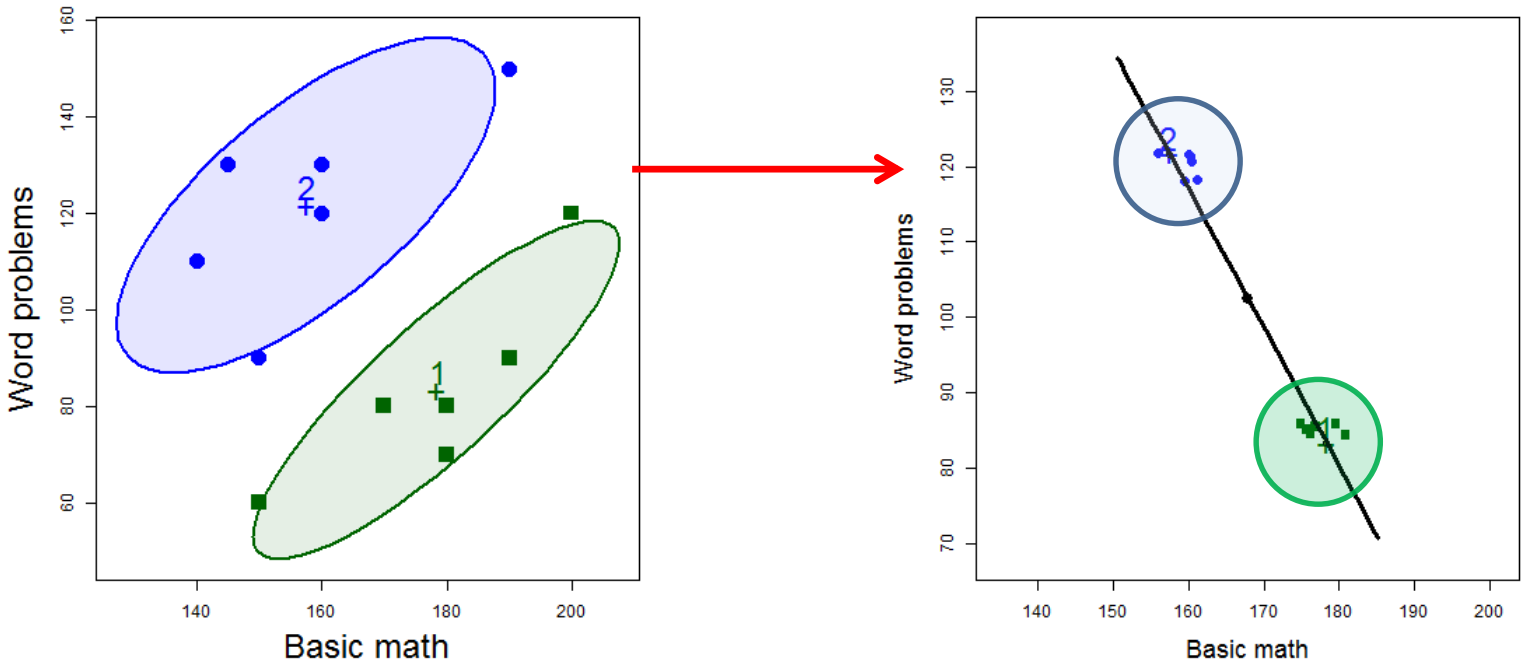


Data → Data ellipses → HE plot



- Differences between group means are shown by the **H** ellipsoid– data ellipsoid of the fitted values (w/ 1 df, degenerates to a line)
  - Direction shows relation of groups to response variables
  - Size shows "how big is H relative to E"
- Variation within groups is reflected in the **E** ellipsoid-- data ellipsoid of the residuals
  - Direction: residual (partial) correlation between BM & WP
  - Size/shape: residual variance

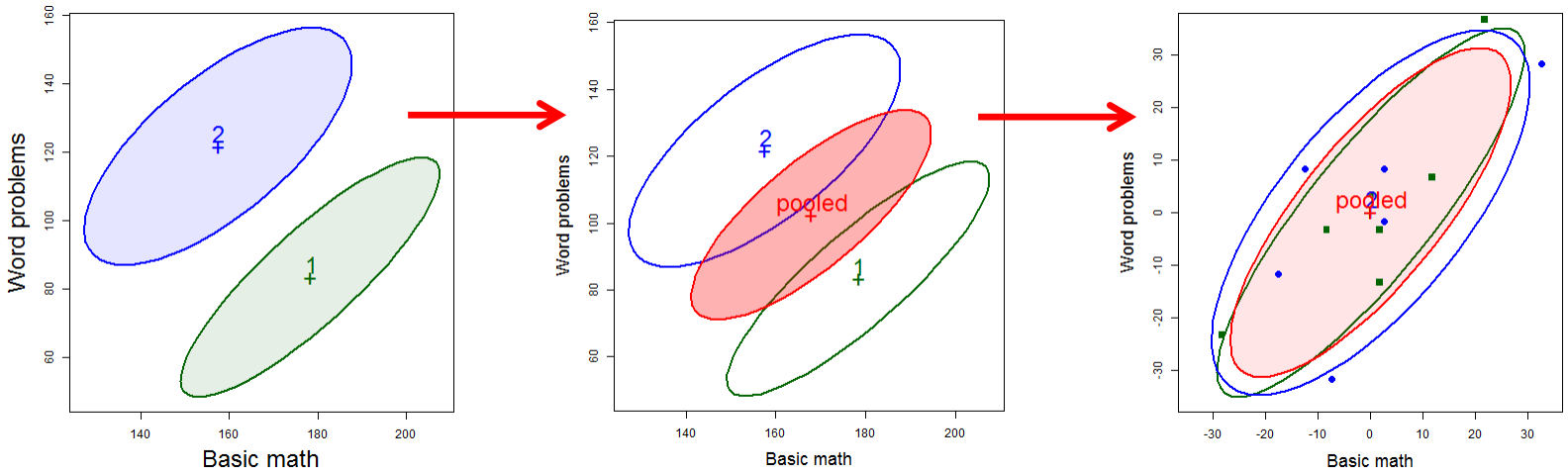
The H ellipse



- The **H** ellipse is the data ellipse of the fitted values (group means, here)
  - The **H** matrix is the sum of squares and crossproducts of the fitted values, corrected for the grand mean

$$\mathbf{H} = \left(\hat{\mathbf{Y}}' \hat{\mathbf{Y}} - n \bar{\mathbf{y}} \bar{\mathbf{y}}' \right)$$

The E ellipse



- The **E** ellipse is the data ellipse of the residuals
  - What you get when you subtract the group means from all observations, shifting them to the grand means.
  - **E** matrix called the "within-group **pooled** covariance matrix"

$$\mathbf{E} = \left(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right)'\left(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right) = \mathcal{E}'\mathcal{E}$$

H & E in numbers

The **H** and **E** matrices are calculated in the car::Anova() function and saved as the SSP and SSPE components, used in the statistical tests.

```
> math.aov <- Anova(math.mod)
> (H <- math.aov$SSP)
$group
         BM       WP
BM   1302.1 -2395.8
WP -2395.8  4408.3
```

Direct calculation: $\mathbf{H} = \left(\hat{\mathbf{Y}}' \hat{\mathbf{Y}} - n \bar{\mathbf{y}} \bar{\mathbf{y}}' \right)$

```
> fit <- fitted(math.mod)
> ybar <- colMeans(mathscore[,2:3])
> n <- nrow(mathscore)
> crossprod(fit) - n*outer(ybar, ybar)
         BM       WP
BM   1302.1 -2395.8
WP -2395.8  4408.3
```

```
> fit
         BM       WP
1   178.33   83.333
2   178.33   83.333
3   178.33   83.333
4   178.33   83.333
5   178.33   83.333
6   178.33   83.333
7   157.50  121.667
8   157.50  121.667
9   157.50  121.667
10  157.50  121.667
11  157.50  121.667
12  157.50  121.667
```

H & E in numbers

```
> (E <- math.aov$SSPE)
        BM     WP
BM 3070.8 2808.3
WP 2808.3 4216.7
```

Direct calculation: $\mathbf{E} = \left(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right)'\left(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right) = \mathcal{E}'\mathcal{E}$

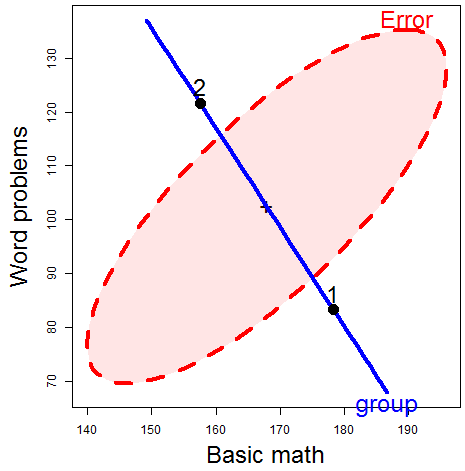
```
> resids <- residuals(math.mod)
> crossprod(resids)
        BM     WP
BM 3070.8 2808.3
WP 2808.3 4216.7
```

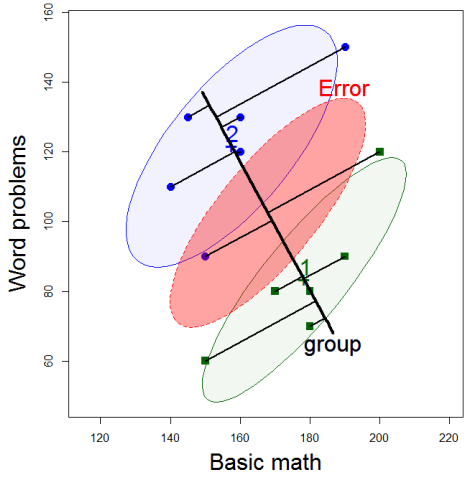
```
> cor(resids)
     BM   WP
BM 1.00 0.78
WP 0.78 1.00
```

```
> resids
          BM      WP
1    11.667   6.667
2    -8.333  -3.333
3     1.667  -3.333
4    21.667  36.667
5   -28.333 -23.333
6     1.667 -13.333
7     2.500  -1.667
8    32.500  28.333
9    -7.500 -31.667
10    2.500   8.333
11 -17.500 -11.667
12 -12.500   8.333
```

Discriminant analysis

- MANOVA and linear discriminant analysis (LDA) are intimately related and differ mainly in perspective:
  - MANOVA: Do means of groups on 2+ responses differ?
  - LDA: Find weighted sums of responses that best discriminate groups
- In both cases,
  - Group differences are represented by the **H** matrix; residuals: **E** matrix
  - Test statistics based on eigenvalues of HE^{-1}
  - Discriminant weights are eigenvectors of HE^{-1}

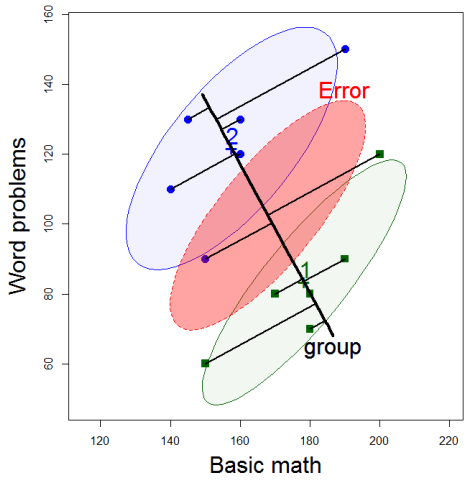




Discriminant analysis

- For 2 groups,
  - the discriminant axis is the line joining the two group centroids,
  - discriminant scores are the projections of observations on this line.
- MASS:lda() does this analysis

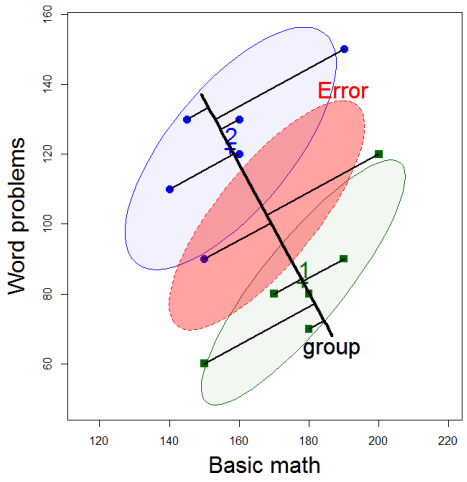
```
> (mod.lda <- MASS::lda(group ~ ., mathscore))

Group means:
     BM     WP
1 178.3  83.33
2 157.5 121.67

Coefficients of linear discriminants:
        LD1
BM -0.08350
WP  0.07527
```

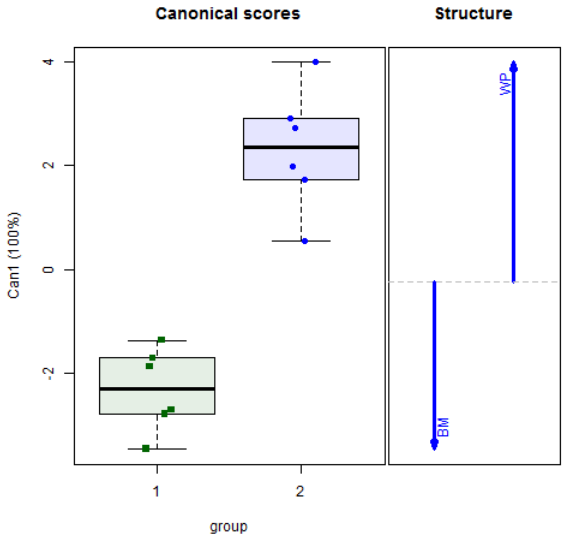


The canonical dimension is Can1 = 0.075 WP - 0.083 BM, a contrast between the two tests

Canonical space

- The HE plot view shows the data in **data** space
- Easier to see effects by projecting scores to **canonical** space – the best-discriminating axes.
- For a 1 df effect, there is only one canonical dimension
  - Arrows show the relative size & direction of discriminant weights



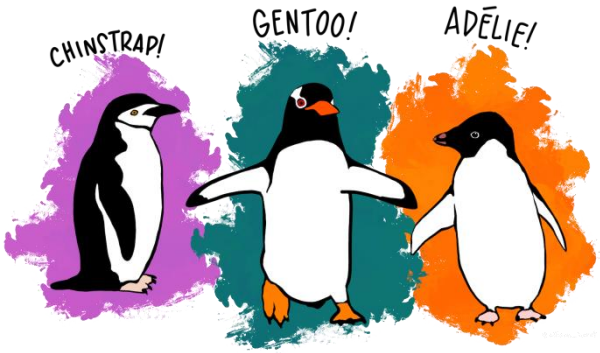


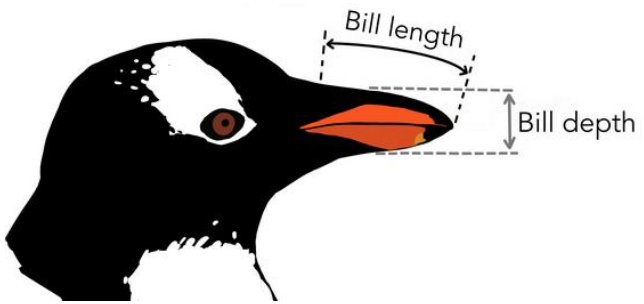

```
library(candisc)
mod.can <- candisc(math.mod)
plot(mod.can)
```

Penguin data

- Data on 3 species of penguins, measured on 3 Antarctic islands
  - How does penguin "size" differ by species, island, … ?


```
> library(palmerpengiuns)
> peng <- penguins %>% rename(...) %>% ...          # clean up names, etc.
> peng[sample(1:333, 5), ]
# A tibble: 5 x 8
  species   island    bill_length bill_depth flipper_length body_mass sex    year
  <fct>     <fct>           <dbl>      <dbl>          <int>     <int> <fct> <int>
1 Chinstrap Dream            58         17.8            181      3700 f      2007
2 Adelie    Torgersen        39.6       17.2            196      3550 f      2008
3 Gentoo    Biscoe           46.2       14.1            217      4375 f      2009
4 Chinstrap Dream            49         19.5            210      3950 m      2008
5 Gentoo    Biscoe           50.4       15.7            222      5750 m      2009
```

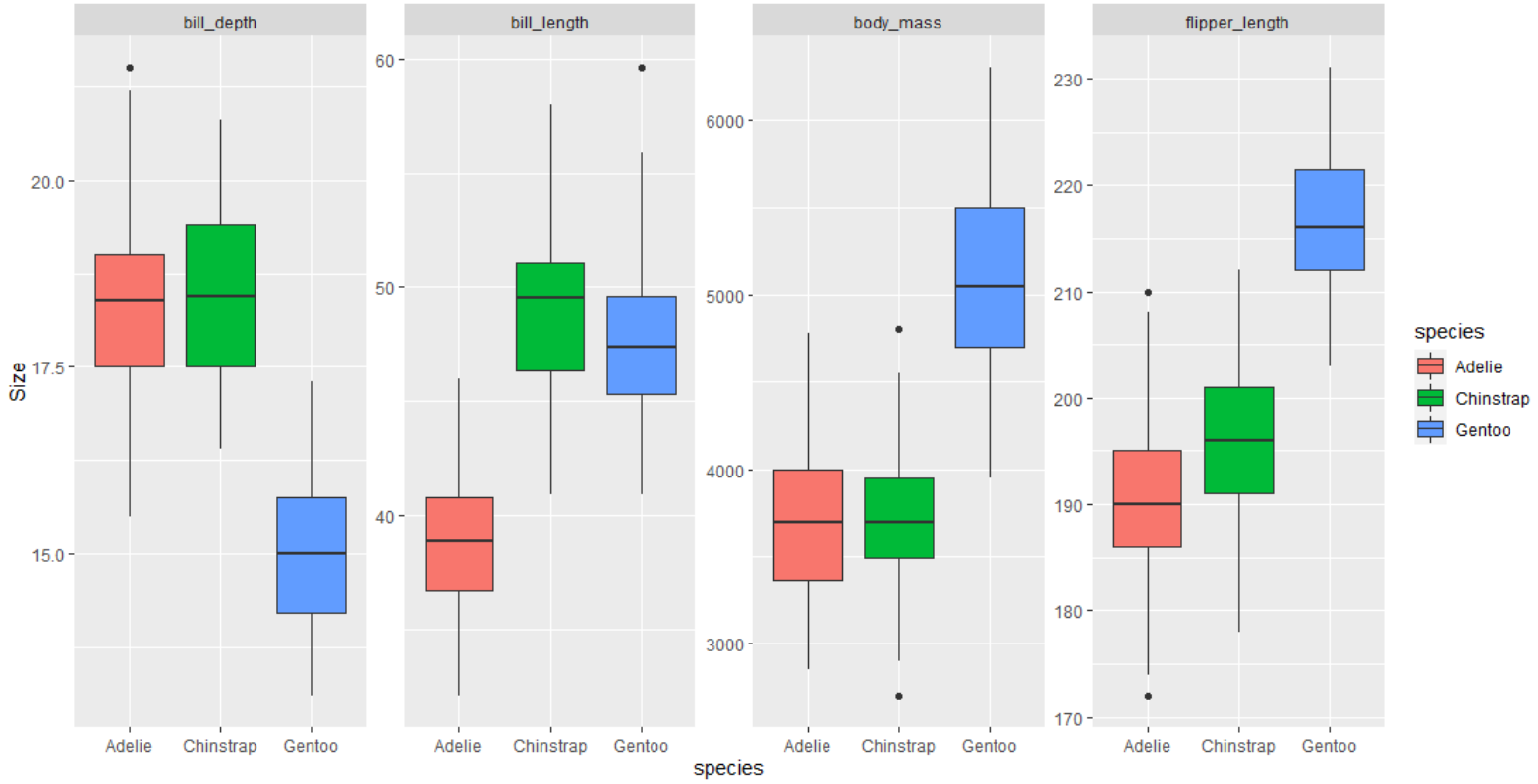




Penguins: Multivariate EDA

Boxplots by grouping variables (factors) are often useful for an initial overview

- Can show multiple variables, but hard for >1 factor.
- What is the pattern here?

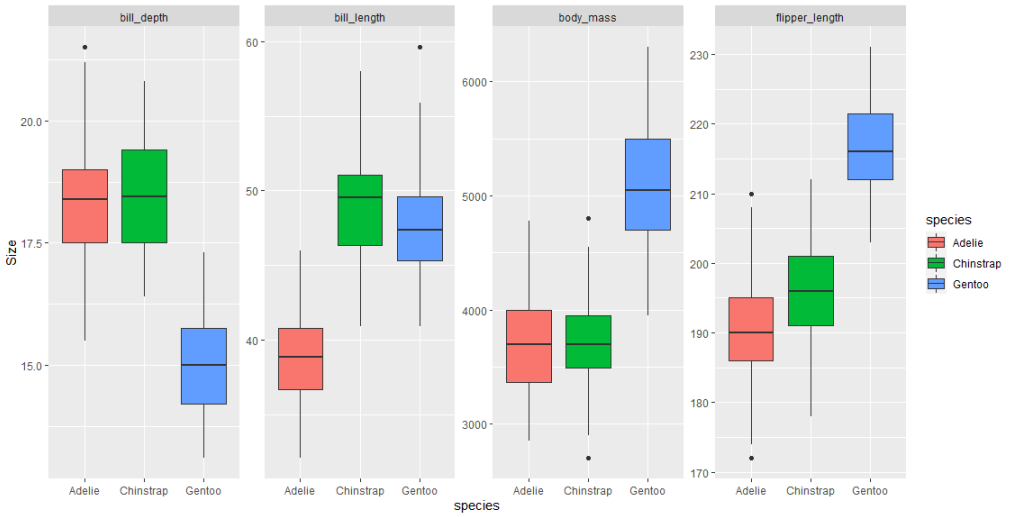


Penguins: Multivariate EDA

Boxplots by grouping variables (factors) are often useful for an initial overview

- Need to reshape data from wide to long format

```
peng_long <- peng  |>                                  # convert wide to long format
  tidyr::gather(Measure, Size, bill_length:body_mass)

ggplot(peng_long, aes(x=species, y=Size, fill=species)) +
  geom_boxplot() +
  facet_wrap(. ~ Measure, scales="free_y", nrow=1)
```



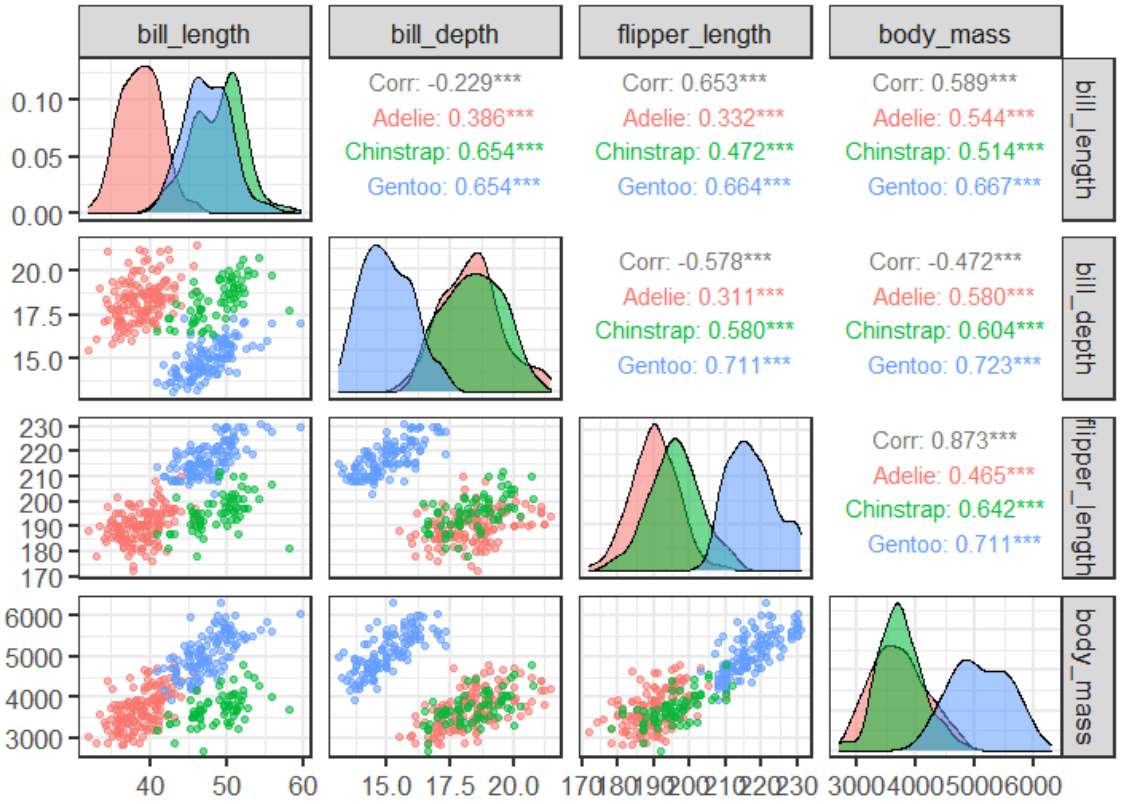
Scatterplot matrix with GGally

EDA should also include a scatterplot matrix
Here, plot the numeric variables, grouped by species

```
library(GGally)
ggpairs(peng, aes(color=species, alpha=0.5),columns=3:6)
```

Diagonals show univariate distributions

Lower Δ: scatterplots
Upper Δ: correlations

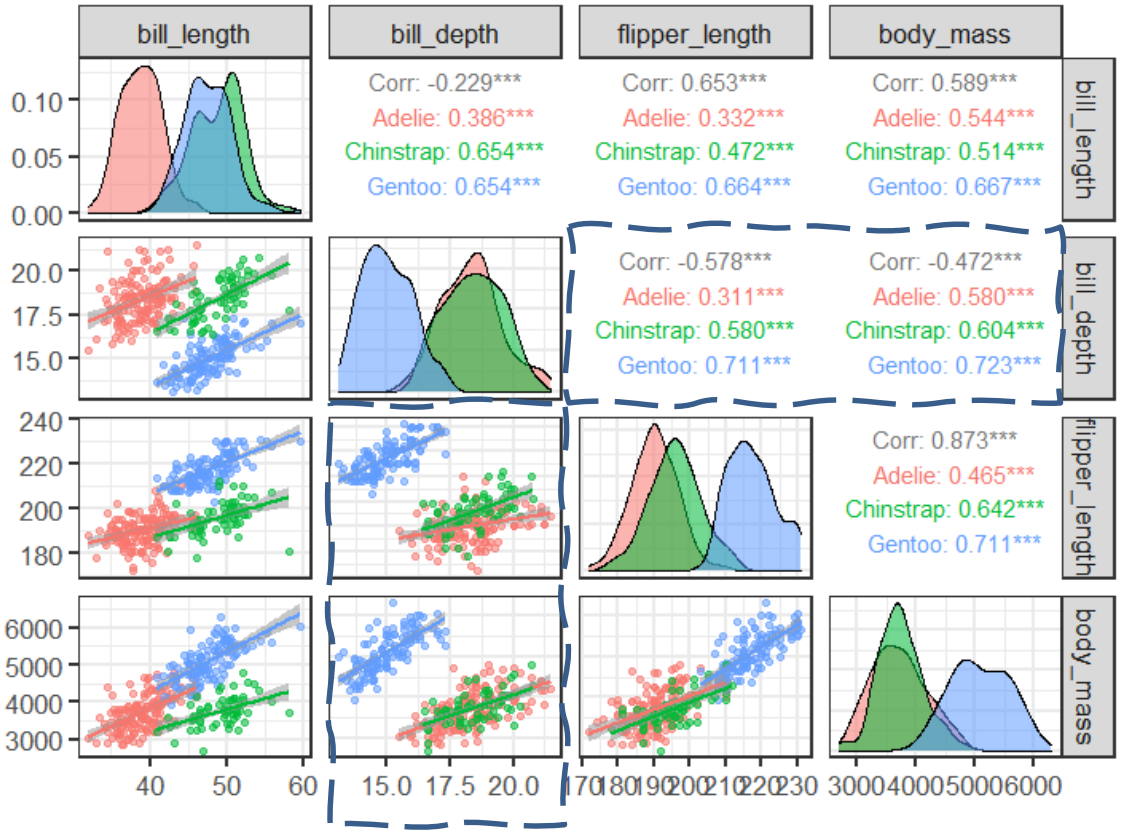


But wait: I want to see the regression lines!
No worries: define a custom function for scatterplot in the lower triangle

```
ggpanel <- function(data, mapping, ...){
  p <- ggplot(data = data, mapping = mapping) +
    geom_point() +
    geom_smooth(method=lm, formula = y~x, ...)
  p}
ggpairs(peng,  aes(color=species, alpha=0.5), columns=3:6,  lower = list(continuous = ggpanel))
```

All plots look reasonably linear ✓

But: there's something weird in the correlations
Can you spot it?

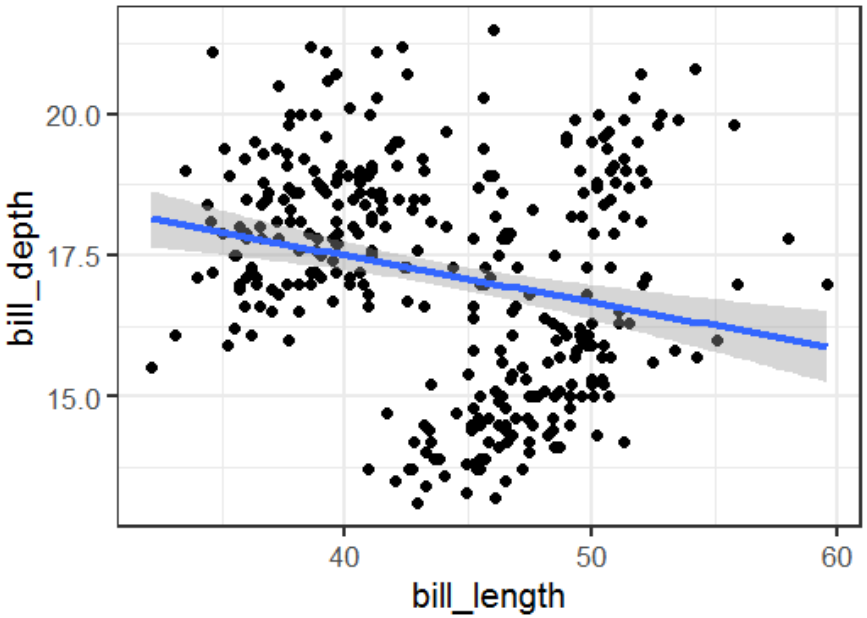


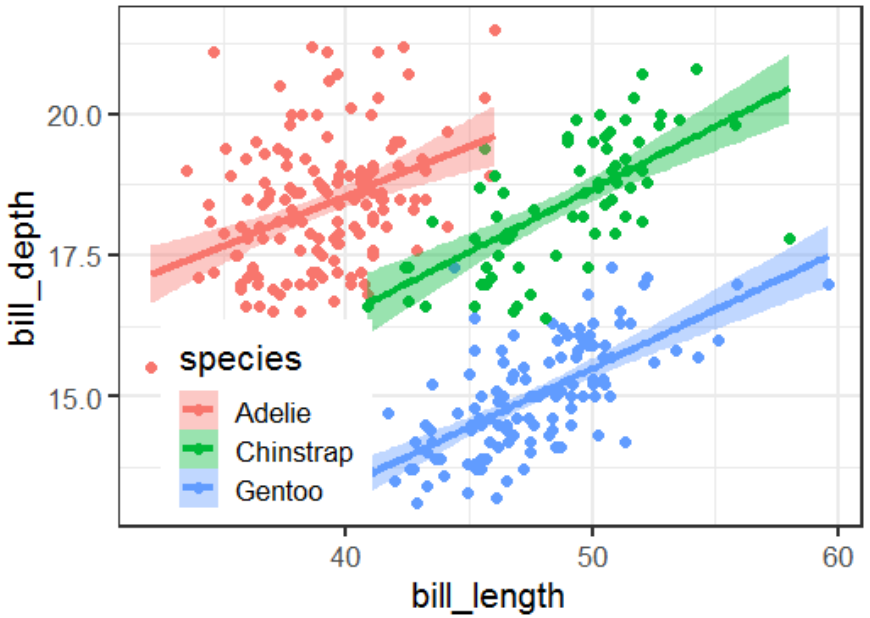
Simpson's Paradox

Simpson's paradox: within-group *r* s are reversed when the data are pooled
This happens here because the group means are negatively correlated

```
# Ignoring species
ggplot(peng, aes(x=bill_length, y=bill_depth)) +
  geom_point() +
  geom_smooth(method="lm", formula=y~x)
```

```
# Including species
ggplot(peng, aes(x=bill_length, y=bill_depth
                 color=species, fill=species)) +
  geom_point() +
  geom_smooth(method="lm", formula=y~x)
```





PCA & Biplots

- For multivariate data, often want to view the data in a low-D space that shows the most total variance
- PCA: finds weighted sums of variables which are:
  - Uncorrelated
  - Account for maximum variance
  - How many dimensions are necessary?
- A biplot is a 2D (or 3D) plot of the largest PCA dimensions
  - Vectors in this plot show the original data variables
  - Points in this plot show the observations
    - Data ellipses here show within group relations

PCA animation

PCA:

- PC1 is the direction along which points have max. variance
- Equivalently, the perp. deviations from the line have min. residual SS

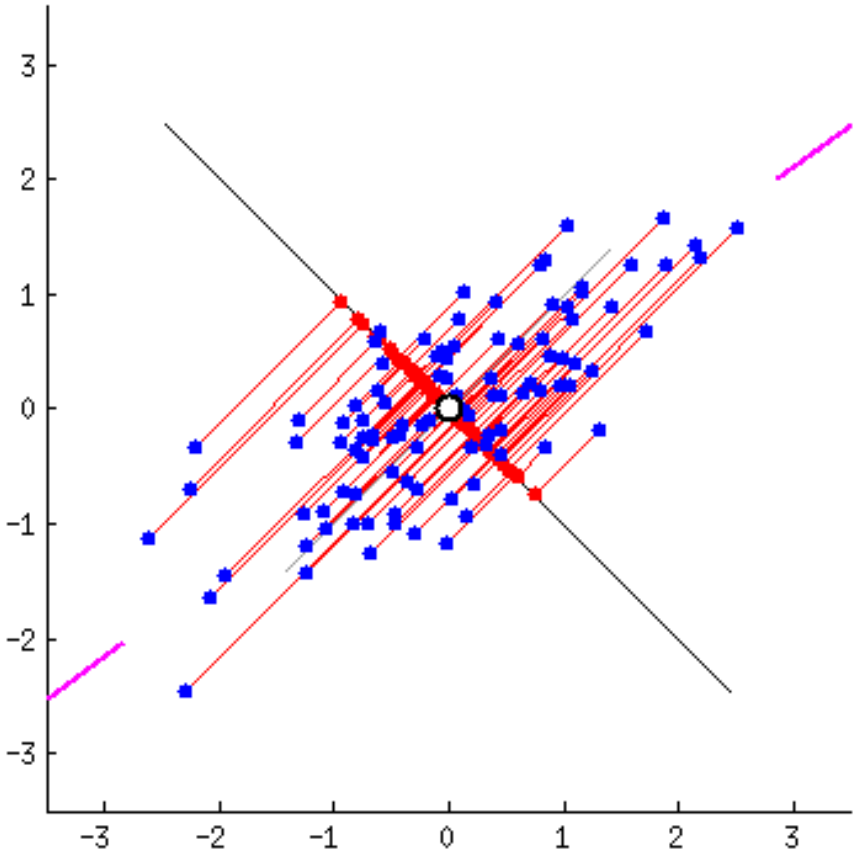
PCA by springs

- Imagine each pt connected to a possible PC1 line by springs
- Force ~ deviation^2

Forces balance, naturally seek the min. residual SS position.

Voila, QED!

- A visual proof



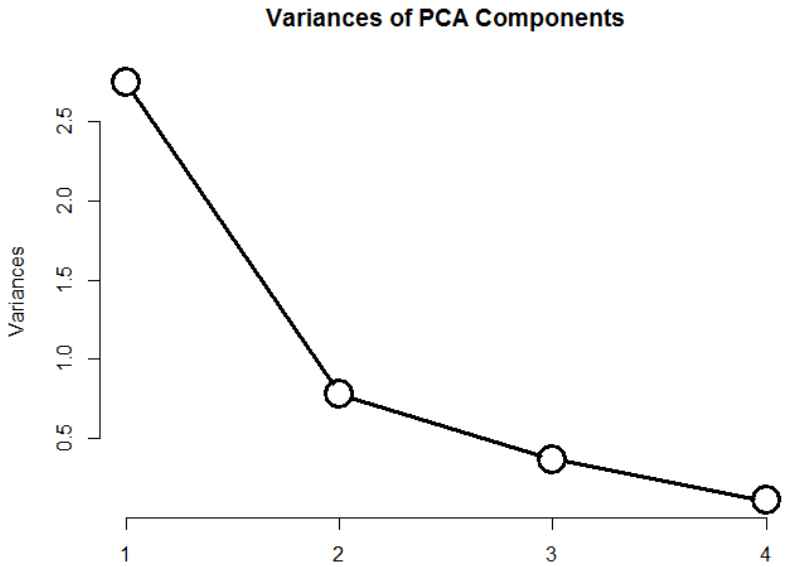
PCA

```
peng.pca <- prcomp (~ bill_length + bill_depth + flipper_length + body_mass,
                    data=peng,
                    na.action=na.omit,
                    scale. = TRUE)
screeplot(peng.pca, type = "line", lwd=3, cex=3,
                          main="Variances of PCA Components")
```

```
> summary(peng.pca)
Importance of components:
                          PC1   PC2    PC3   PC4
Standard deviation     1.657 0.882 0.6072 0.328
Proportion of Variance 0.686 0.195 0.0922 0.027
Cumulative Proportion  0.686 0.881 0.9730 1.000
```

2D: 88.1 %

3D: 97.3 %



See: https://rpubs.com/friendly/penguin-biplots for details

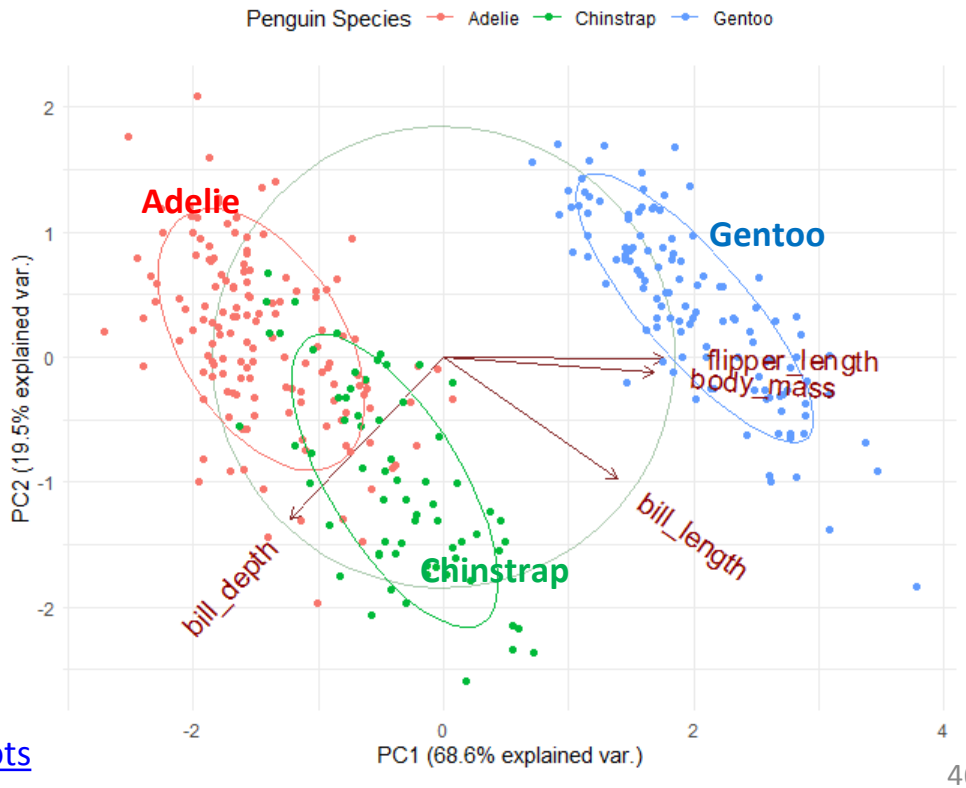
Biplot

```
library(ggbiplot)
ggbiplot(peng.pca, obs.scale = 1, var.scale = 1,
        groups = peng$species,
        ellipse = TRUE, circle = TRUE) +
        scale_color_discrete(name = 'Penguin Species')
```

PC1, PC2 ~ 88.1% of variance

- PC1: largely flipper length & body mass: “penguin size”
- PC2 (& PC1): relates to “bill shape”

Easy to characterize the species in terms of these variables



See: https://rpubs.com/friendly/penguin-biplots

Penguins: MANOVA

Assume the goal is to determine whether/how the penguins differ in "size" by species

- A MLM tests all 4 size variables together: ~ species
- Could also use other factors: ~ species + sex + island

```
> peng.mod0 <-lm(cbind(bill_length, bill_depth, flipper_length, body_mass) ~ species,
                 data=peng)
> Anova(peng.mod0)

Type II MANOVA Tests: Pillai test statistic
        Df test stat approx F num Df den Df Pr(>F)
species  2      1.64      371      8    656 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

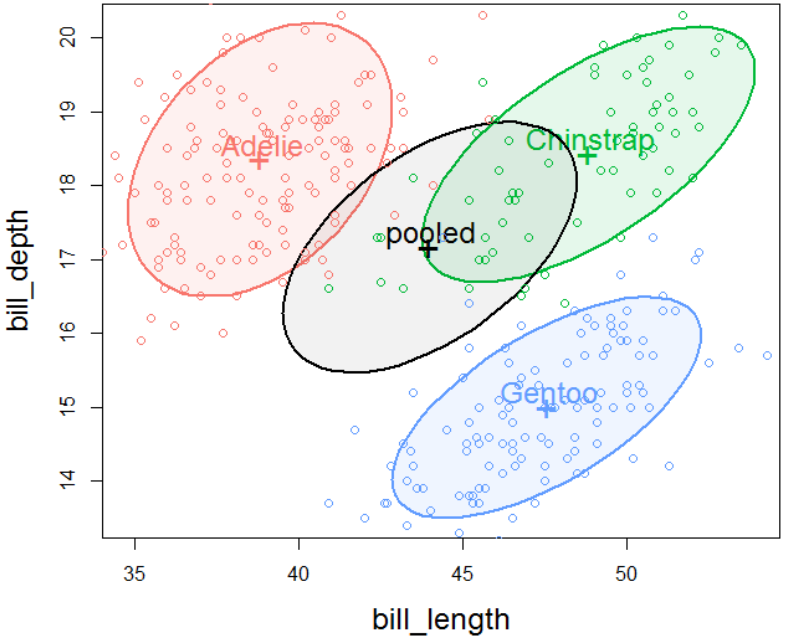
Yet, we are left to understand the nature of this effect wrt. the size variables.

See: https://rpubs.com/friendly/penguin-manova for details

Penguins: view data ellipses

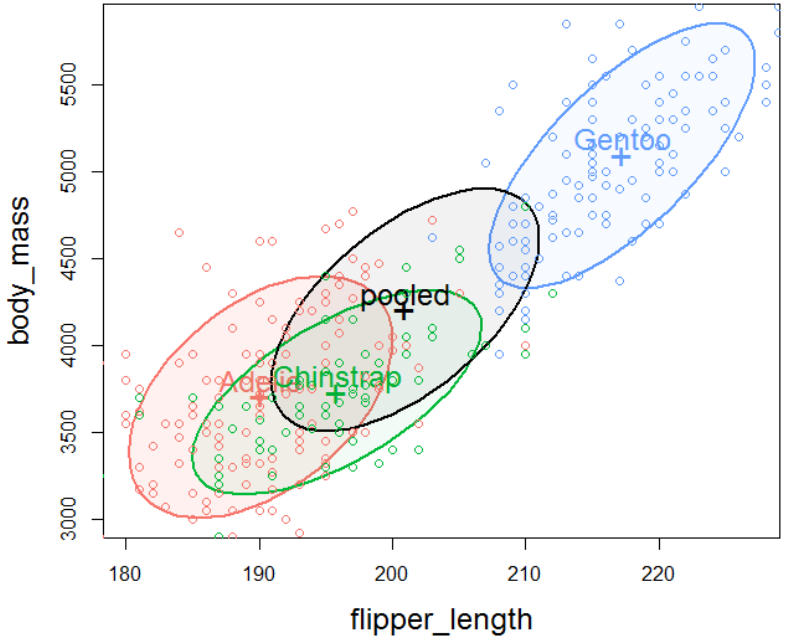
Data ellipses in 2D provide a good start for pairwise relations

bill depth & length



- group means negatively correlated
- within group correlation > 0

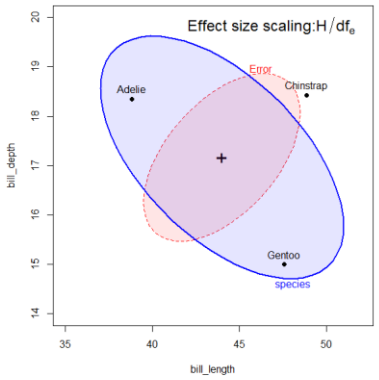
body mass & flipper length

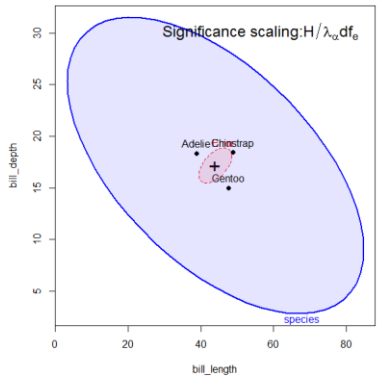


- group means positively correlated
- within group correlation > 0

HE plot details

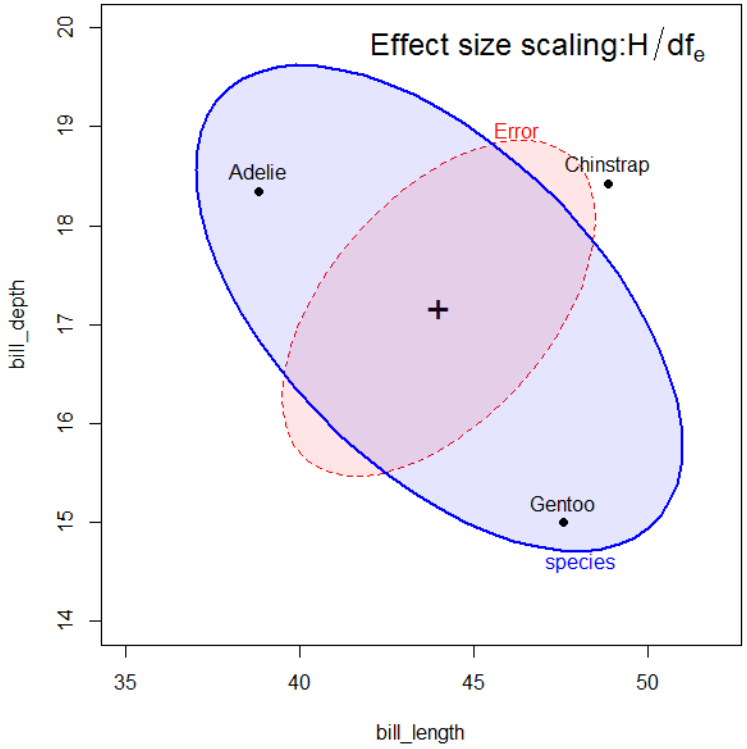
- **E** ellipse reflects within-group error (co)variation
  - Size: **E** / df_e set to cover 68%, an analog of $\bar{y} \pm 1$ std
  - Shift to grand mean for direct comparison with **H**
- **H** ellipse reflects (co)variation of group means
  - **effect size** scaling, uses **H**/df_e to put this on the same scale as the **E** ellipse. Analog of effect size in univariate designs.
  - **significance** ("evidence") scaling: uses **H**/λ_α df_e .
    - The **H** ellipse protrudes outside the **E** ellipse somewhere, *iff* an effect is significant (Roy's largest root test) at $p < \alpha$

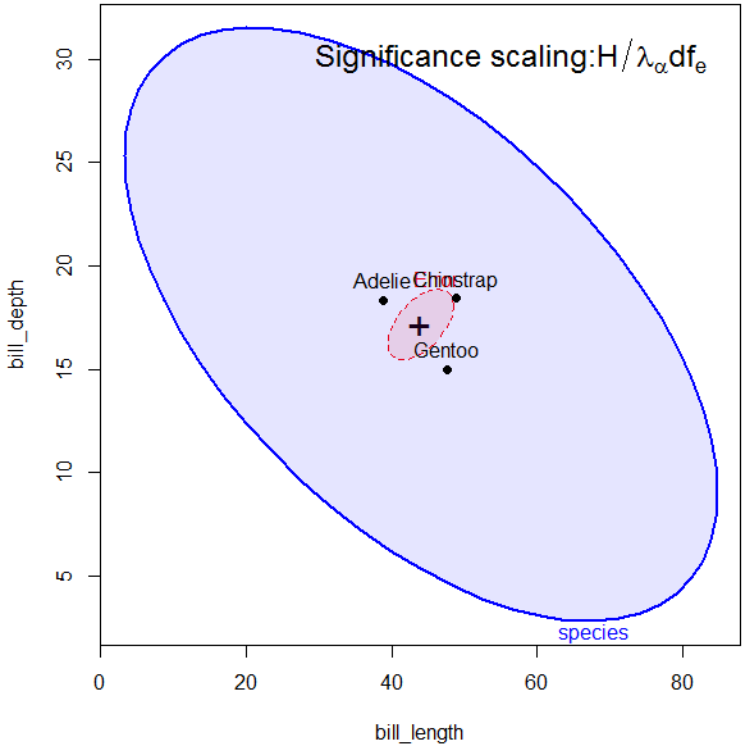




Penguins: HE plots

Orientation of the **H** ellipse reflects negative correlation of the species means: species with larger bill depth have smaller bill length (bill "shape"?)
E ellipse: within species, larger bill length → larger bill depth





heplot(peng.mod0, size="effect")

heplot(peng.mod0, size="evidence")

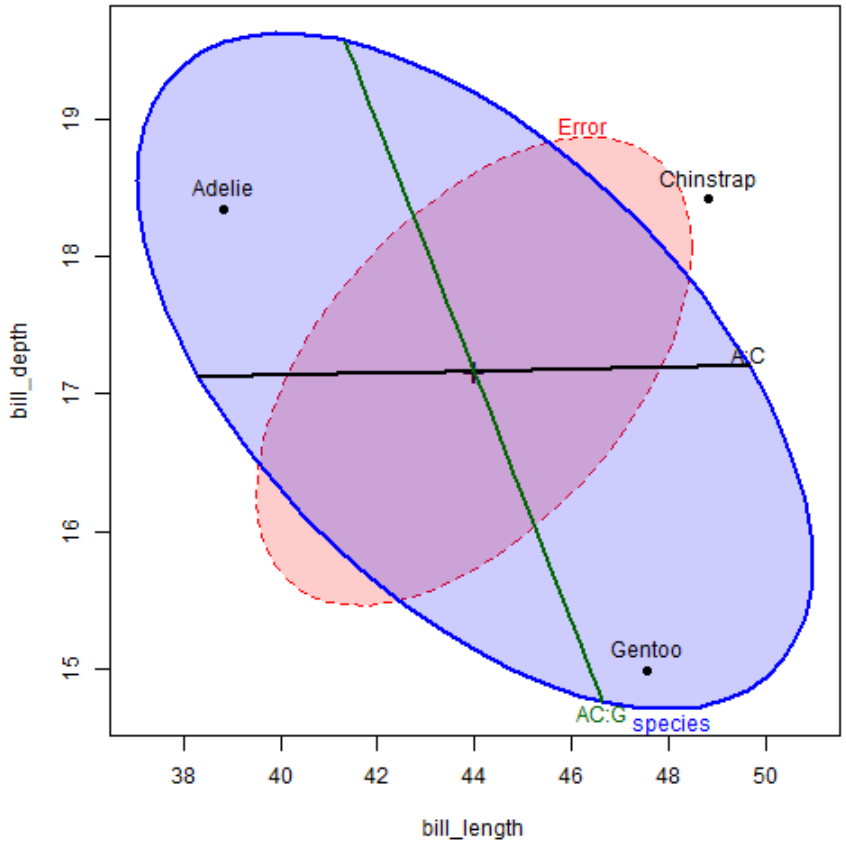
Contrasts

- In linear models, any effect of $df_h > 1$ can be partitioned into df_h separate 1 df tests of contrasts
  - If orthogonal, $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2 + \ldots \mathbf{H}_{dfh}$ -- accounts for total effect
  - Tested as a linear hypothesis, e.g., $x_1 - (x_2 + x_3)/2 = 0$
  - Each $\mathbf{H}_i$ has rank=1, so appears as a line in HE plots
- Assume we want to compare the species as two contrasts:
  - Do Adelie differ from Chinstrap?
  - Do Gentoo penguins differ from the other two?

```
> contrasts(peng$species)<-matrix(c(1,-1, 0,   -1, -1, -2), 3,2)
> contrasts(peng$species)
          [,1] [,2]
Adelie       1   -1
Chinstrap   -1   -1
Gentoo       0    2
```

Contrasts

```
hyp <- list("A:C"="species1","AC:G"="species2")  # give names to contrasts
heplot(peng.mod0, fill=TRUE, fill.alpha=0.2,
       hypotheses=hyp, size="effect")
```



Result is very clear:

- Adelie & Chinstrap differ only in bill length
- Gentoo differ from other two – longer, but less deep bills (bill shape)

Both of these are large effects!

Together, they are the entire species effect!

Other models

```
peng.mod2 <-lm(cbind(bill_length, bill_depth, flipper_length, body_mass) ~ species + sex, data=peng)
Anova(peng.mod2)
```

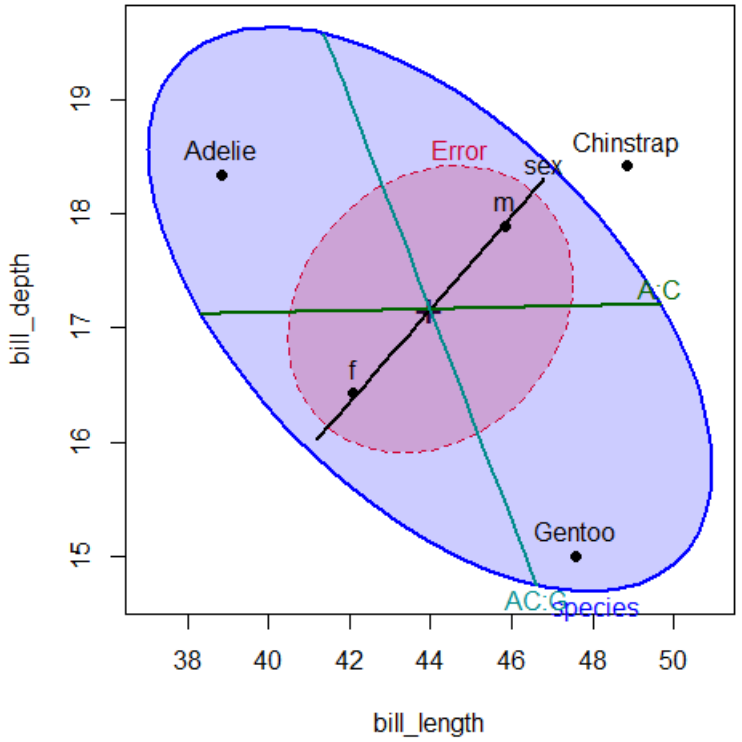
```
Type II MANOVA Tests: Pillai test statistic
        Df test stat approx F num Df den Df    Pr(>F)
species  2   1.65480   391.89      8    654 < 2.2e-16 ***
sex      1   0.64004   144.91      4    326 < 2.2e-16 ***
```

```
heplot(peng.mod2, fill=TRUE, fill.alpha=0.2,
        hypotheses=hyp)
```

Effect of sex: male penguins have larger bills

Adding sex reduces **E** variances
→Effect of species now more pronounced

Each 1 df effect plots as a line



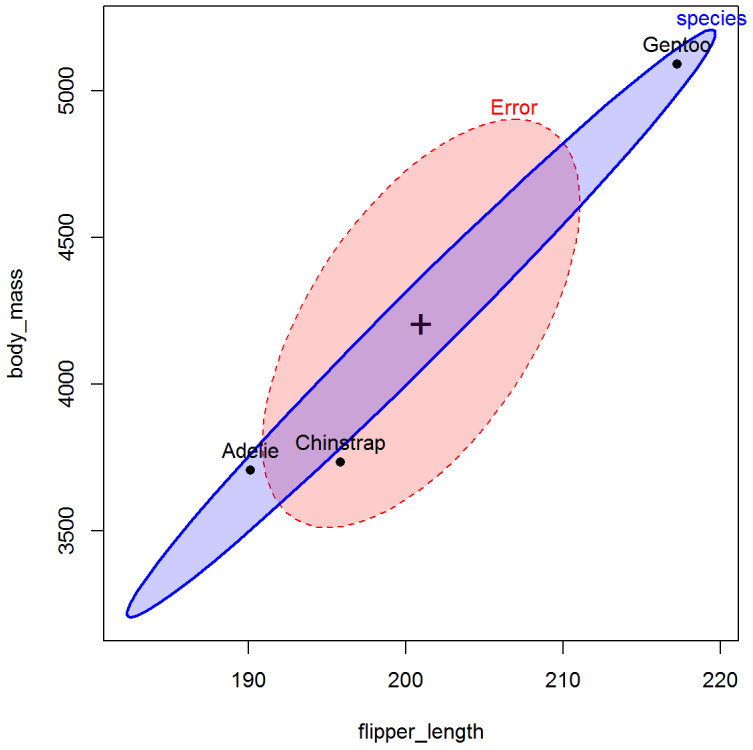
Other HE plots

- 2D: can plot any pair of responses in data space
- pairs.mlm(): all pairwise 2D views
- heplot3d(): plots in 3D, can rotate, spin, zoom, …

```
heplot(peng.mod0, variables=3:4,
      fill=TRUE, fill.alpha=0.2, size="effect")
```

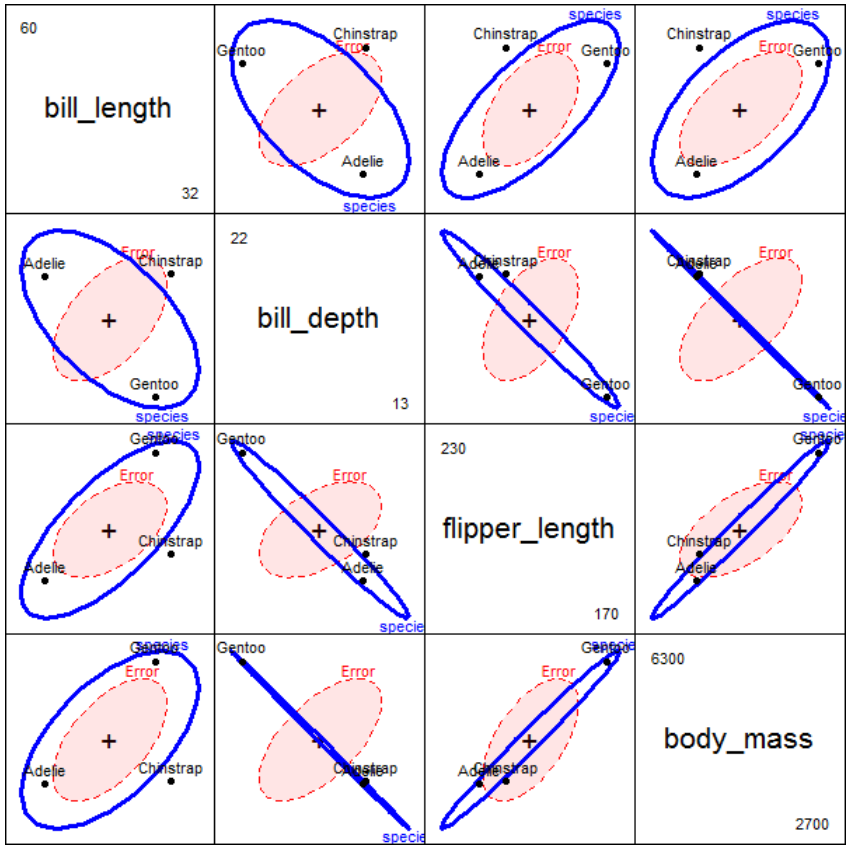
Interpretation:

- major axis of the **H** ellipse measures "penguin size"
- Gentoo are the Big Birds in this story!



HE Pairs plots

The pairs() method for mlm objects gives all pairwise HE plots in a scatterplot matrix format.




```
pairs(peng.mod0, size="effect",
      fill=c(TRUE, FALSE))
```

Something new here:

- avg. bill depth is negatively correlated with “size” variables – larger penguin species have smaller bill depths (curvature?)
- correlation of avg. bill depth with body mass nearly -1

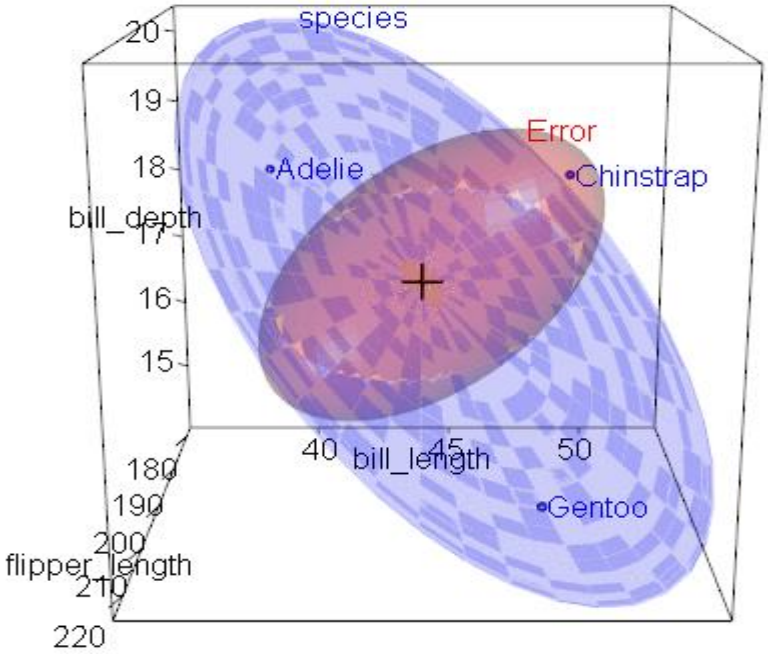
heplot3d()

3D HE plots can show other features

```
heplot3d(peng.mod0, size="effect")
```

The H ellipsoid here is flat (2D), because the species effect has 2 df

In this 3D view, the 3 species form a triangle, suggesting some further interpretation, not seen in 2D views



Canonical view

- 4 response variables, but only s=min(q, dfh)=2 dimensions.
  - Here, both dimensions are significant
  - Can1 accounts for 86.5% of between-species variance
  - Can 2 accounts for the rest: 13.5%

```
> library(candisc)
> (peng.can <- candisc(peng.mod0))

Canonical Discriminant Analysis for species:

  CanRsq Eigenvalue Difference Percent Cumulative
1  0.938      15.03       12.7    86.5       86.5
2  0.700       2.34       12.7    13.5      100.0

Test of H0: The canonical correlations in the
current row and all that follow are zero

  LR test stat approx F numDF denDF Pr(> F)
1       0.0187      516     8   654  <2e-16 ***
2       0.2997      255     3   328  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

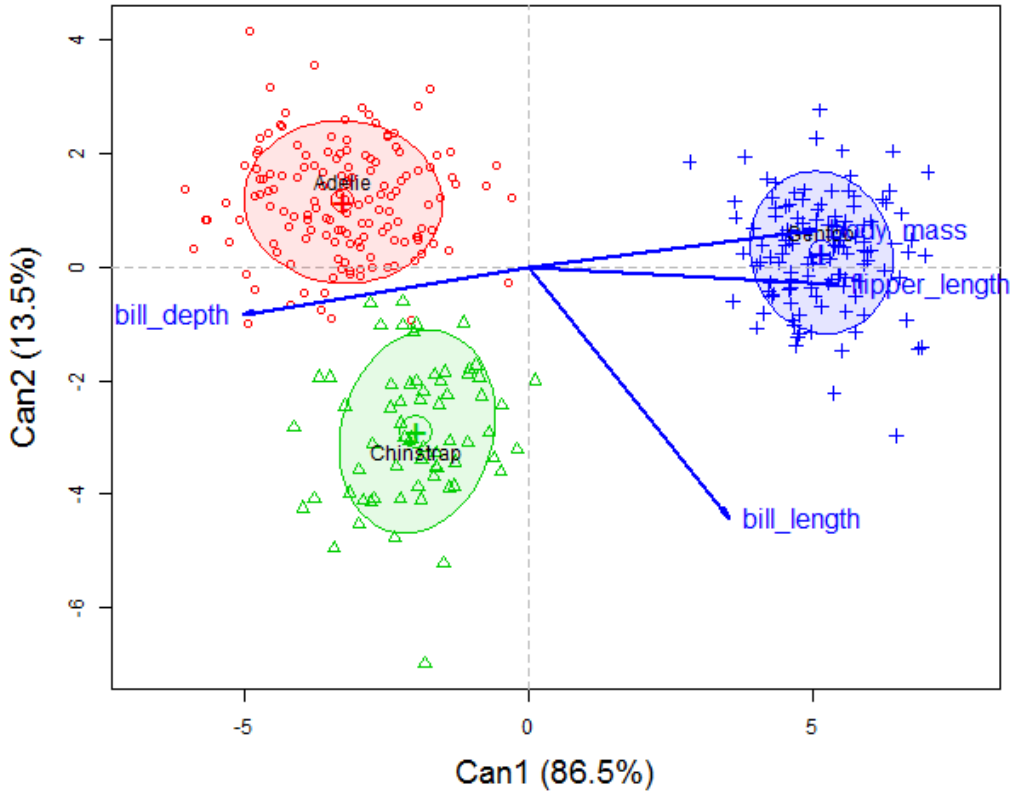
Canonical view

The plot() method for candisc objects shows points for observations and vector for variables

```
plot(peng.can, ellipse = TRUE … )  #plot CAN scores with ellipses
```

Can1: largely body mass & flipper length, that separate Gentoo from (Adelie, Chinstrap)

Can2: bill length distinguishes Chinstrap from others.



Canonical HE plot

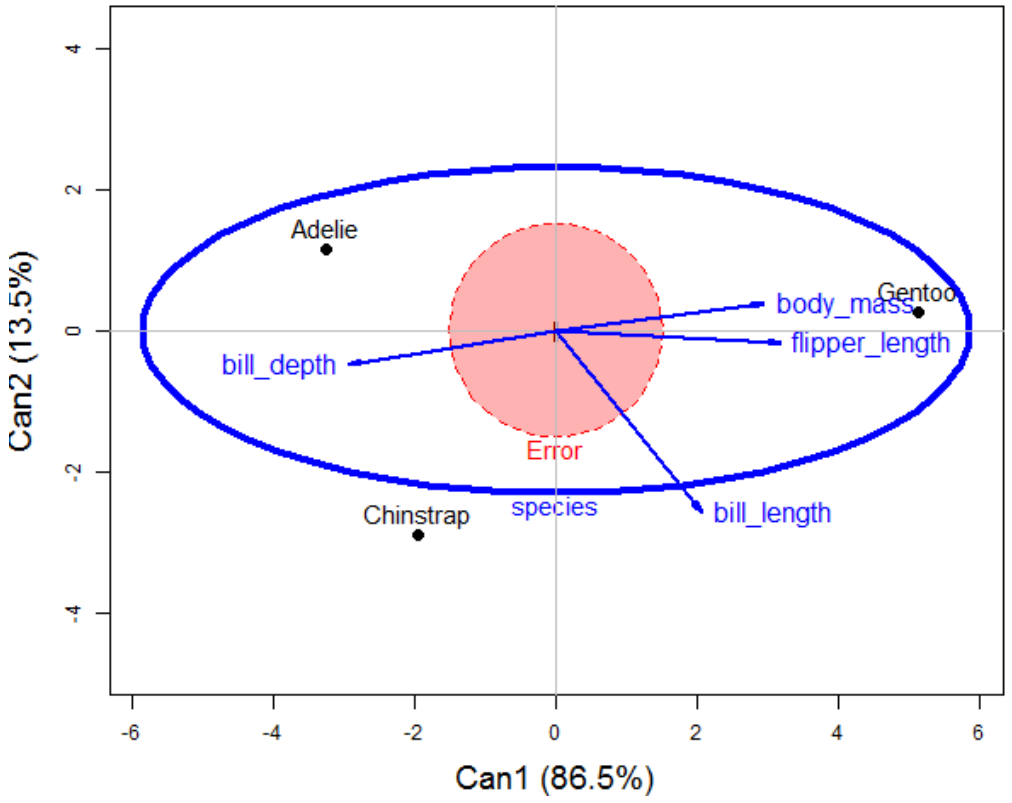
```
heplot(peng.can, size="effect", fill=c(TRUE, FALSE))
```

Here is the entire effect of species shown in one HE plot

In CAN space, residuals are uncorrelated: **E** = circle

Size of **H** shows the total effect of species

Variable vectors show how the groups are discriminated.



Checking assumptions

- Assumptions in the MLM extend those in univariate models
  - Linearity: Each y_i is linearly related to all xs
  - Constant variance matrices of residuals
  - Residuals are multivariate normal

$$\left.\begin{array}{l} \mathbf{S}_i = \mathbf{S}_2 = \cdots = \mathbf{S}_g \\ \text{Residuals are multivariate normal} \end{array}\right\} \epsilon_i \underset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$$

- In addition, need to check
  - No multivariate outliers
  - No multicollinearity among predictors

Checking assumptions

- Linearity: plot each $\mathbf{y}_i$ against each $\mathbf{x}_j$
  - quantitative $\mathbf{x}_j$: plot($\mathbf{y}_i$ ~ $\mathbf{x}_j$) + loess smooth
  - factor: boxplots
- Constant variance
  - visual: plot data ellipses for each group
    - `heplots:: covEllipses(data, group=, …)`
  - univariate-- levene test: `heplots::leveneTests()`
  - multivariate– Box M test: $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g$
    - `res <- heplots::boxM(); plot(res)`
- Multivariate outliers
  - Mahalanobis D^2 ($\mathbf{y}_i$) ~ χ^2_p : outlier if prob (χ^2_p) < .01
  - Chisquare QQ plot : plot D^2 ($\mathbf{y}_i$) vs. χ^2_p quantiles: cqplot()

Constant variance: Visual

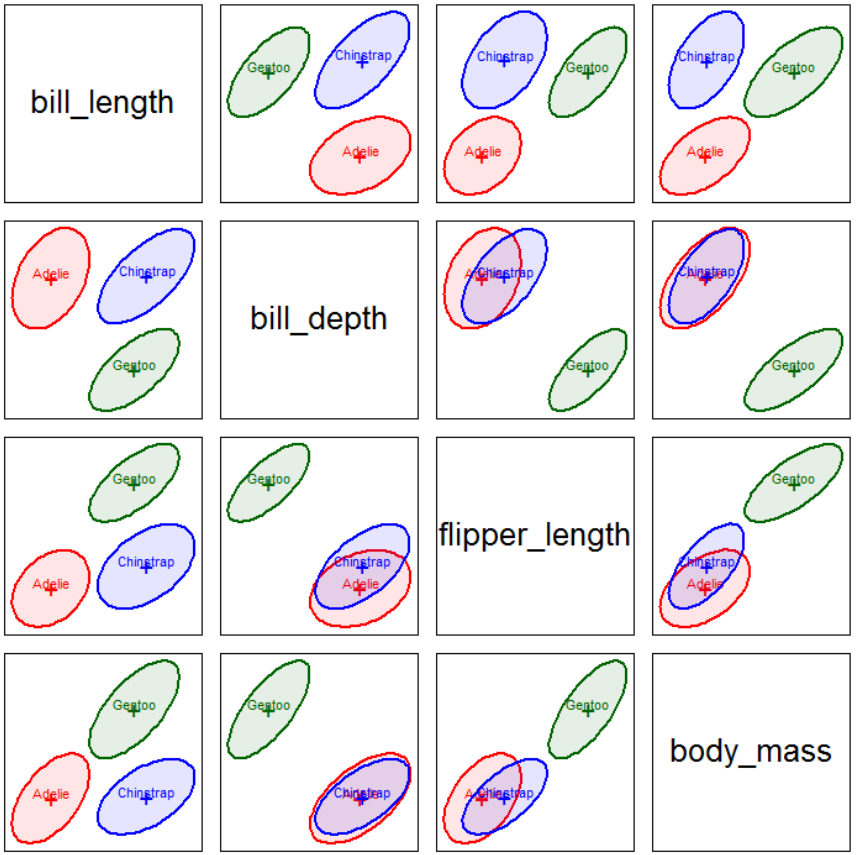
`heplots::covEllipses()` plots the data ellipses for each group, for 2+ variables
Are the sizes and shapes & orientations ≅ the same in all panels?

Approximately true, w/ some diff^{ces}

- Gentoo looks a bit smaller
- Adelie: correlations ~ differ?

This might be good enough

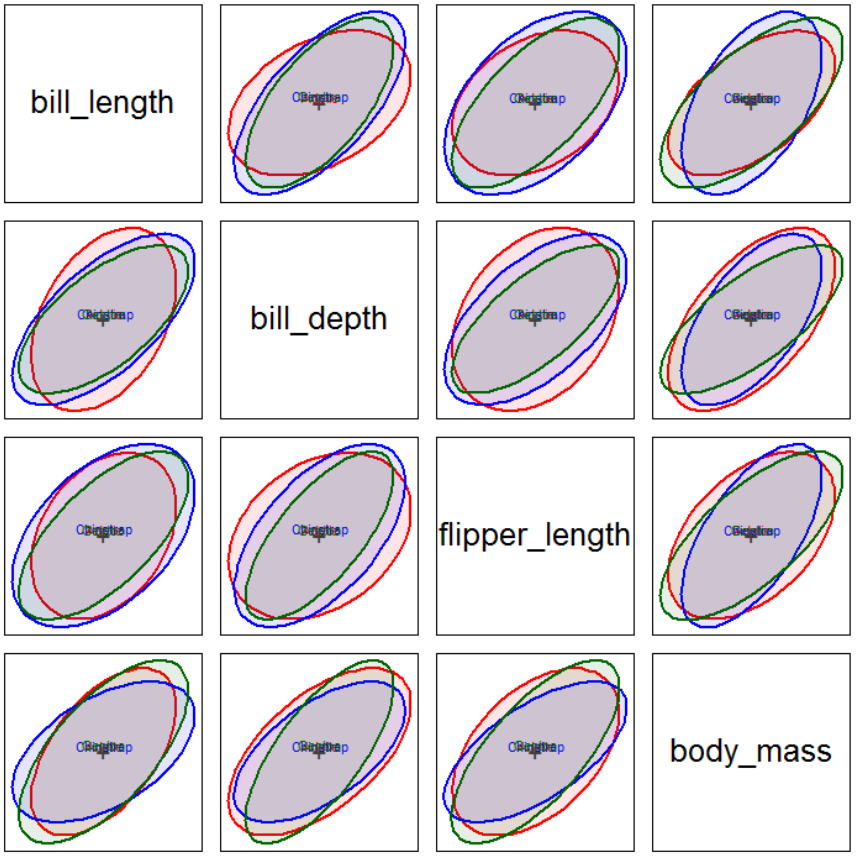
```
covEllipses(peng[,3:6],
        group = peng$species,
        variables=1:4,              # all pairs
        fill=TRUE, fill.alpha=0.1,
        pooled=FALSE)
```



Constant variance: Visual

This is easier to judge if all groups are centered at the grand mean in each panel

```
covEllipses(peng[,3:6],
        group = peng$species,
        variables=1:4,
        center=TRUE,
        fill=TRUE, fill.alpha=0.1,
        pooled=FALSE)
```



Constant variance: statistical tests

Levene tests for each response variable separately:

```
> heplots::leveneTests(peng[,3:6], group=peng$species)
Levene's Tests for Homogeneity of Variance (center = median)

                df1 df2 F value Pr(>F)
bill_length       2 330    2.29 0.1033
bill_depth        2 330    1.91 0.1494
flipper_length    2 330    0.44 0.6426
body_mass         2 330    5.13 0.0064 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Box’s M test: all responses together – equal variances & correlations !

```
> heplots::boxM(peng[,3:6], group = peng$species)

        Box's M-test for Homogeneity of Covariance Matrices

data:  peng[, 3:6]
Chi-Sq (approx.) = 75, df = 20, p-value = 3e-08
```

NB: Box’s M test is highly sensitive to small diff^{ces}; use α = 0.001

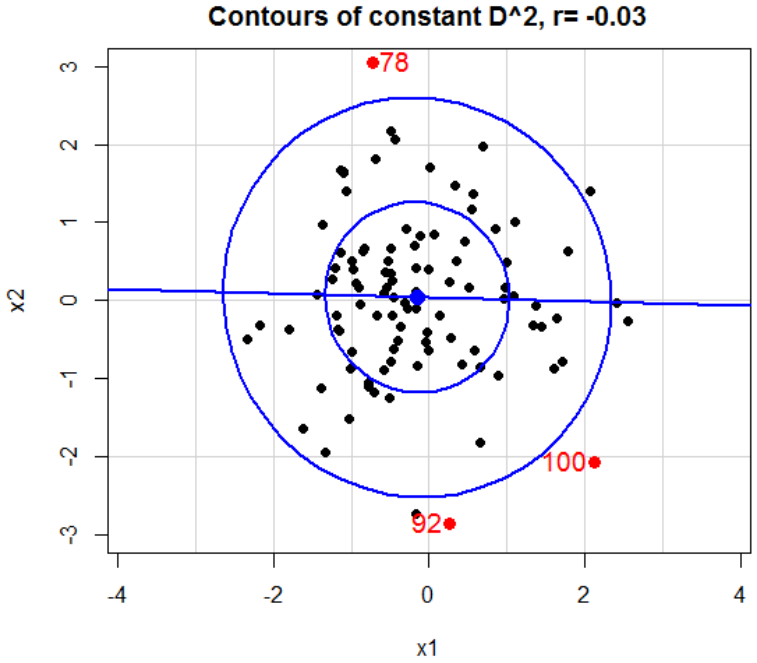
Multivariate normality: $z^2 \rightarrow D^2$

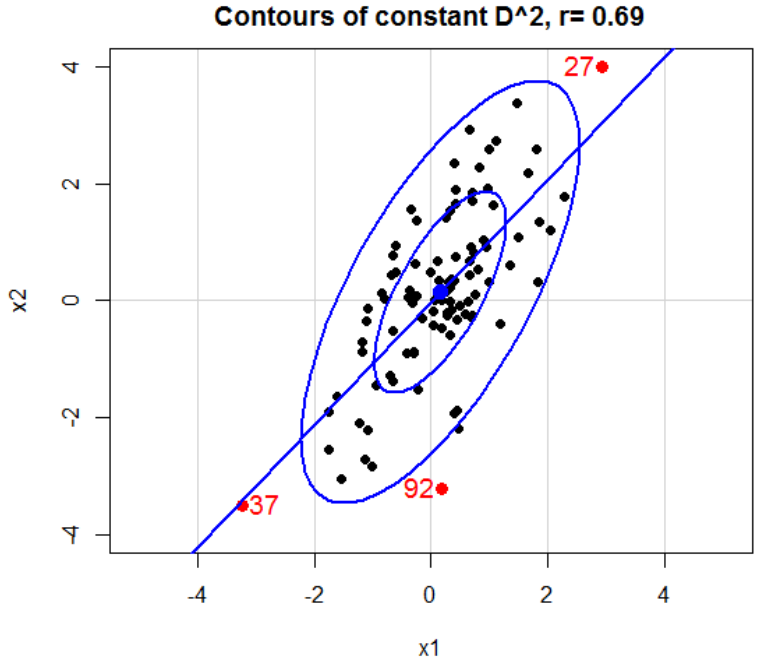
For MVN & outliers, Mahalanobis D^2 generalizes z scores

- 1 variable: $z_i = (x_i - \bar{x})/s \sim N(0,1)$ or, $z_i^2 \sim \chi^2_{(1)}$
- 2 variables, uncorrelated: squared distance from mean is

  $D_i^2 = z_{i1}^2 + z_{i2}^2 \sim \chi^2_{(2)}$
- p variables: D_i^2 = Mahalanobis squared distance of x_i from centroid

$$D_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \sim \chi^2_{(p)}$$

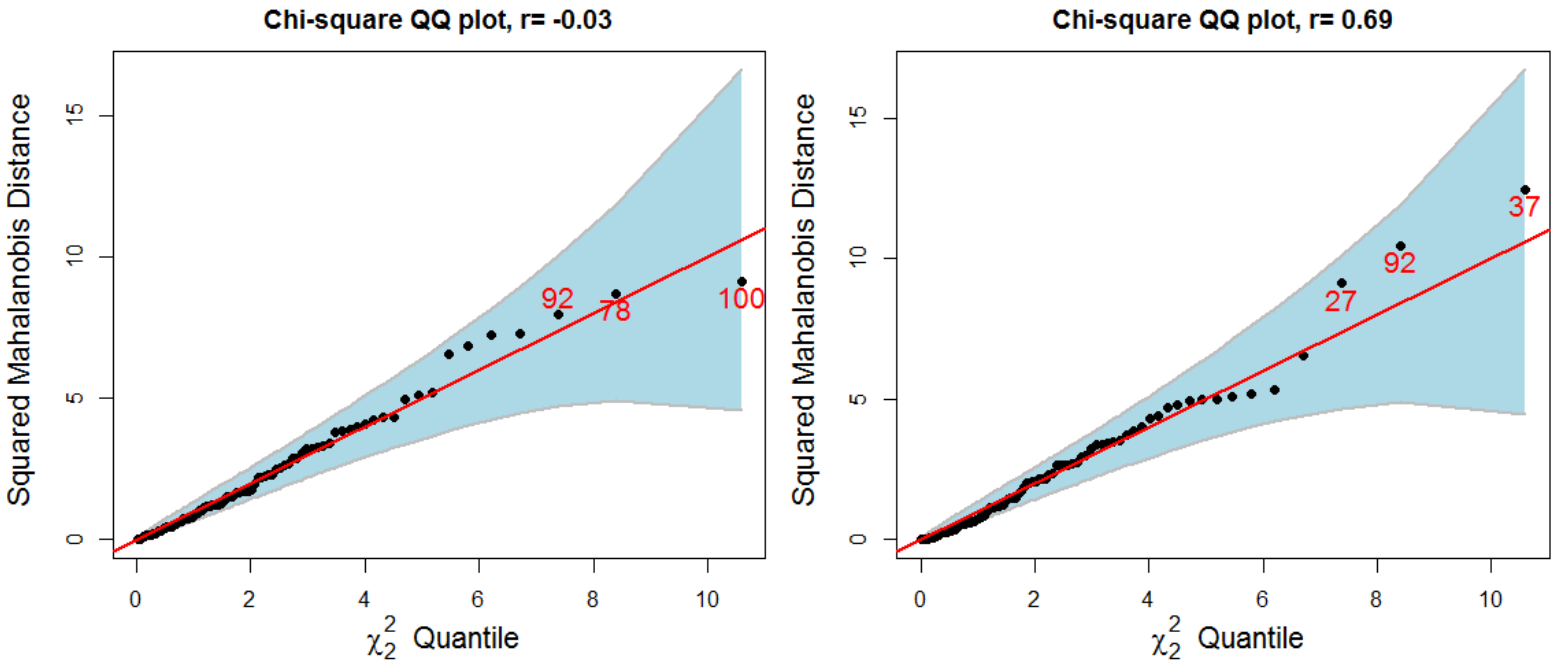




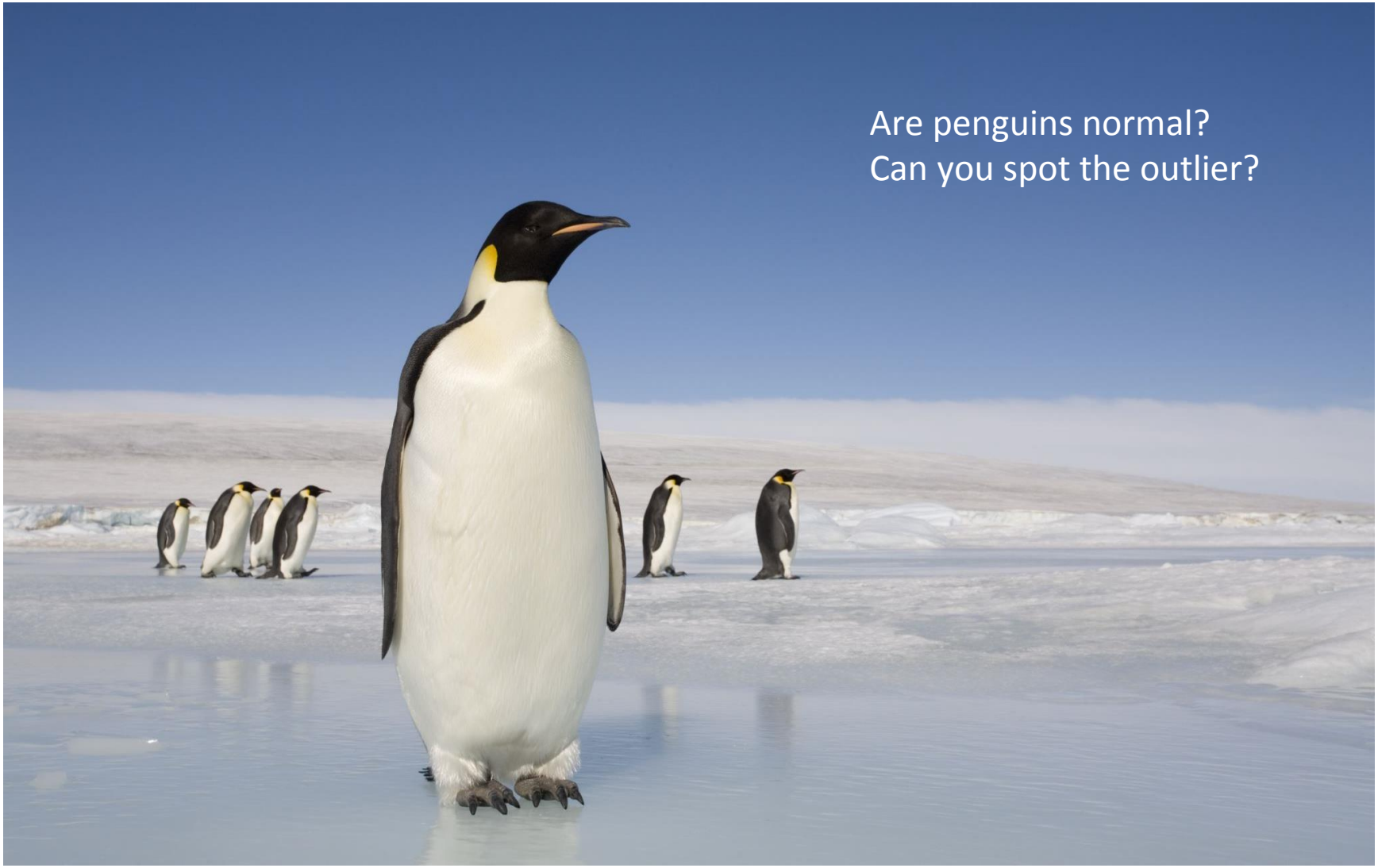
Chi-squared QQ plot

- QQ plot of ordered distances, $D^2_{(i)}$ vs $\chi^2_{(p)}$ quantiles should plot as a 45^o line through origin if MVN
- Multivariate outliers: outside the envelope
- Here: both cases check out as OK: no outliers, MVN ✓

```
heplots::cqplot(df, id.n=3)
```

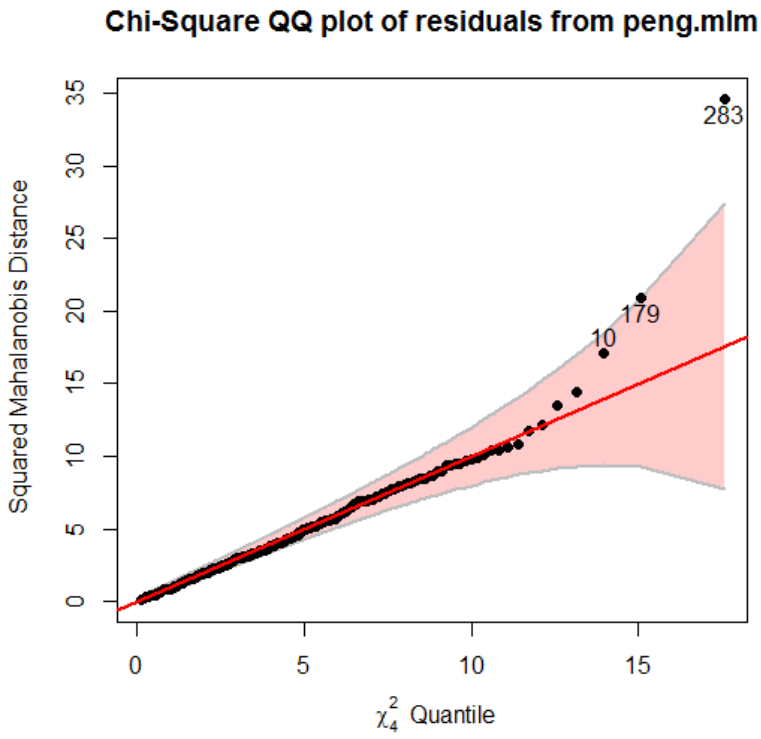


Penguins: MVN & outliers



Penguins: MVN & outliers

```
heplots::cqplot(peng.mlm,
                id.n = 3, conf=0.999)
```



Get D^2 values
with rstatix::mahalanobis_distance
Find z-scores
Select outliers (is.outlier==TRUE)

```
peng |>
 group_by(species) |>
 mahalanobis_distance(bill_length:body_mass) |>
 tibble::rownames_to_column() |>
 mutate(across(bill_length:body_mass,
                .fns= scale)) |>
 filter(is.outlier == TRUE) |>
 as.data.frame()
```

```
  rowname bill_length bill_depth flipper_length body_mass mahal.dist is.outlier
1     283       2.561     0.3225         -1.425   -0.6297      27.76       TRUE
```

MVN: Numerical tests

- Shapiro-Wilk test
  - Originally for univariate normality: stats::shapiro.test()
  - Multivariate version: rstatix::mshapiro_test()

```
peng |>
 select(bill_length : body_mass) |>
 rstatix::mshapiro_test()
```

```
# A tibble: 1 x 2
  statistic   p.value
      <dbl>     <dbl>
1     0.978 0.0000484
```

- Mardia test: multivariate skewness & kurtosis

```
res <- MVN::mvn(data = peng[,c(3:6)],
                mvnTest="mardia")
res$multivariateNormality
```

```
             Test Statistic   p value Result
1 Mardia Skewness    127.42   < 0.001     NO
2 Mardia Kurtosis     -2.51    0.0118     NO
3              MVN      <NA>      <NA>     NO
```

- But: these are overly-sensitive; MLM is relatively robust

Summary

- MLM just like univariate LM, but for multiple responses
  - Simultaneous tests – no need for p-value adjustment
  - Take correlations among responses into account
  - Indicates # of dimensions of responses
- Data ellipses
  - Summarize bivariate data to show means, variances, correlation
  - MANOVA: shows how groups differ in these
- HE framework
  - Visualize multivariate tests in the MLM
  - Canonical displays show these results in the 2D (or 3D) space that accounts for largest between-group variance.
- Checking assumptions: visual tests are often sufficient
  - homogeneity of variances: `heplots::covEllipses()`
  - outliers & MVN: `heplot::cqplot()`