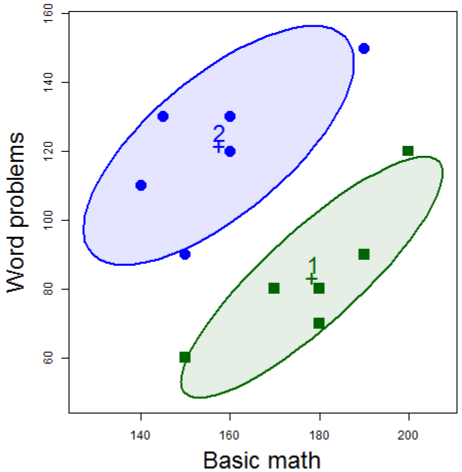
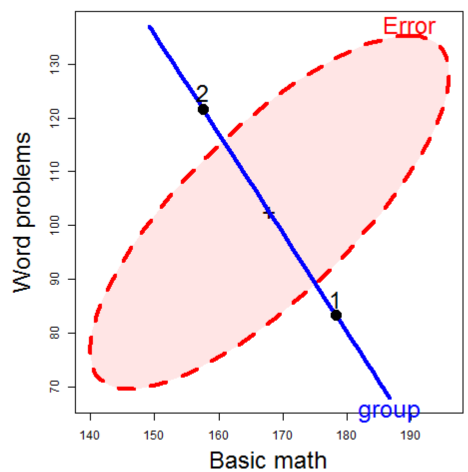
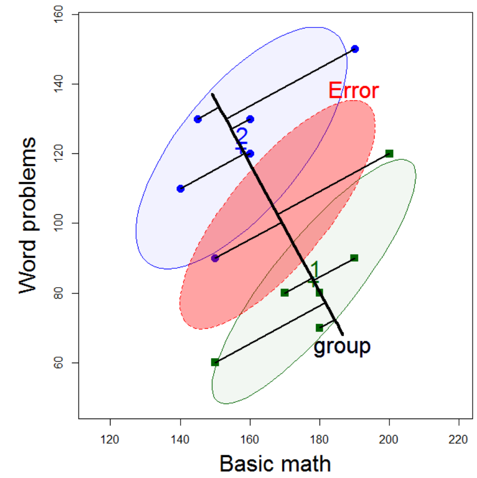
## Data ellipses




## HE plot




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

Michael Friendly

SCS Short Course

March, 2021

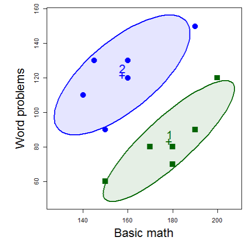
https://friendly.github.io/VisMLM-course/

# Today's topics

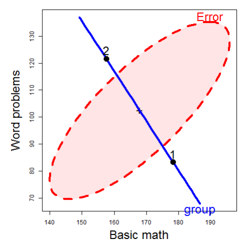
- Brief review of the GLM & MLM

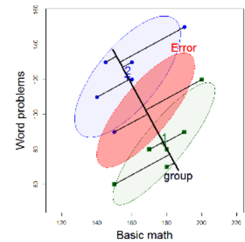
$$\underset{(n\times p)}{\mathbf{Y}} = \underset{(n\times q)}{\mathbf{X}}\ \underset{(q\times p)}{\mathbf{B}} + \underset{(n\times p)}{\mathcal{E}}$$

- Data ellipses
  - sufficient visual summaries

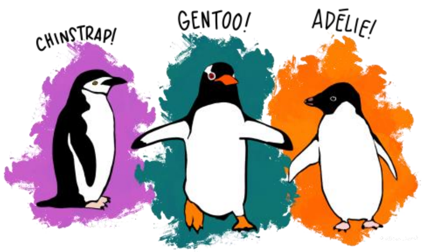


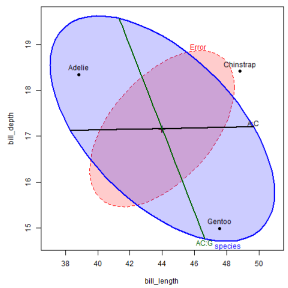

- HE plot framework
  - H & E matrices/ellipses
  - Discriminant/canonical views

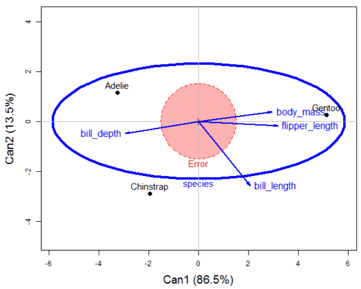



- Example: Penguins data



# One-way ANOVA vs. MANOVA

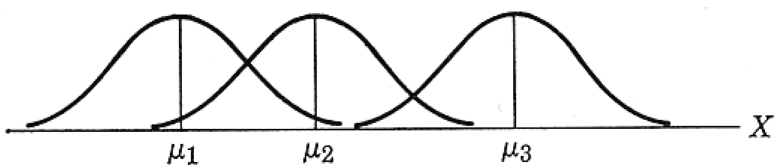



**Figure 8.1.** The simple anova situation, when the differences among the populations are "real."

source: Cooley & Lohnes ((1971)

**How do means differ?**
(Assume equal within-group variances)

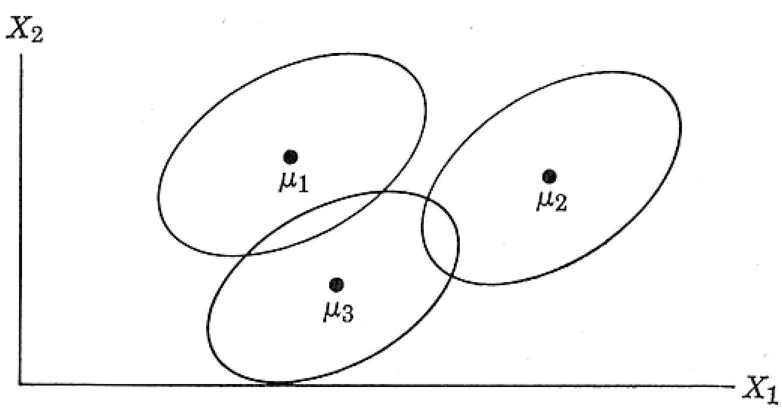



**Figure 8.2.** The simple manova situation, when the differences among the populations are "real."

**How do centroids differ?**
How many dimensions?

(Assume equal within-group variance-covariance matrices)

# 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, …

- 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

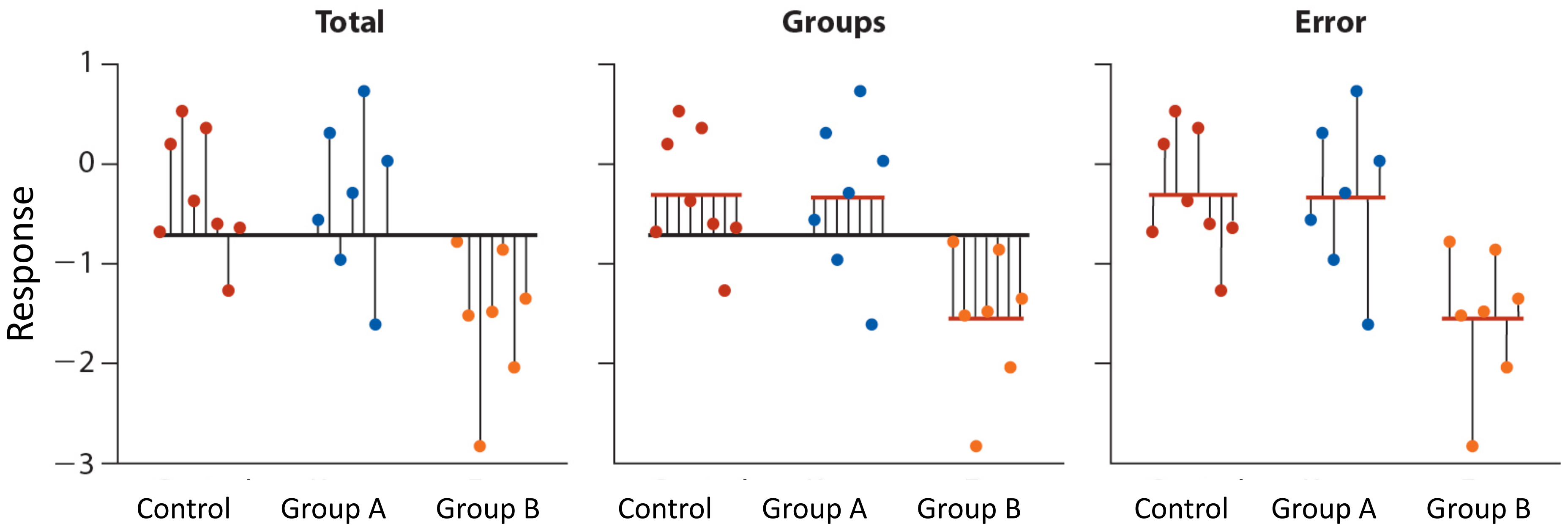
$$F = \frac{SS_H / df_H}{SS_E / df_E} = \frac{MS_H}{MS_E}$$

How big is hypothesis variation relative to error variation?

# Visualizing $SS_T = SS_H + SS_E$

Total variance = Between group variance + Within group variance

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



# 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

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

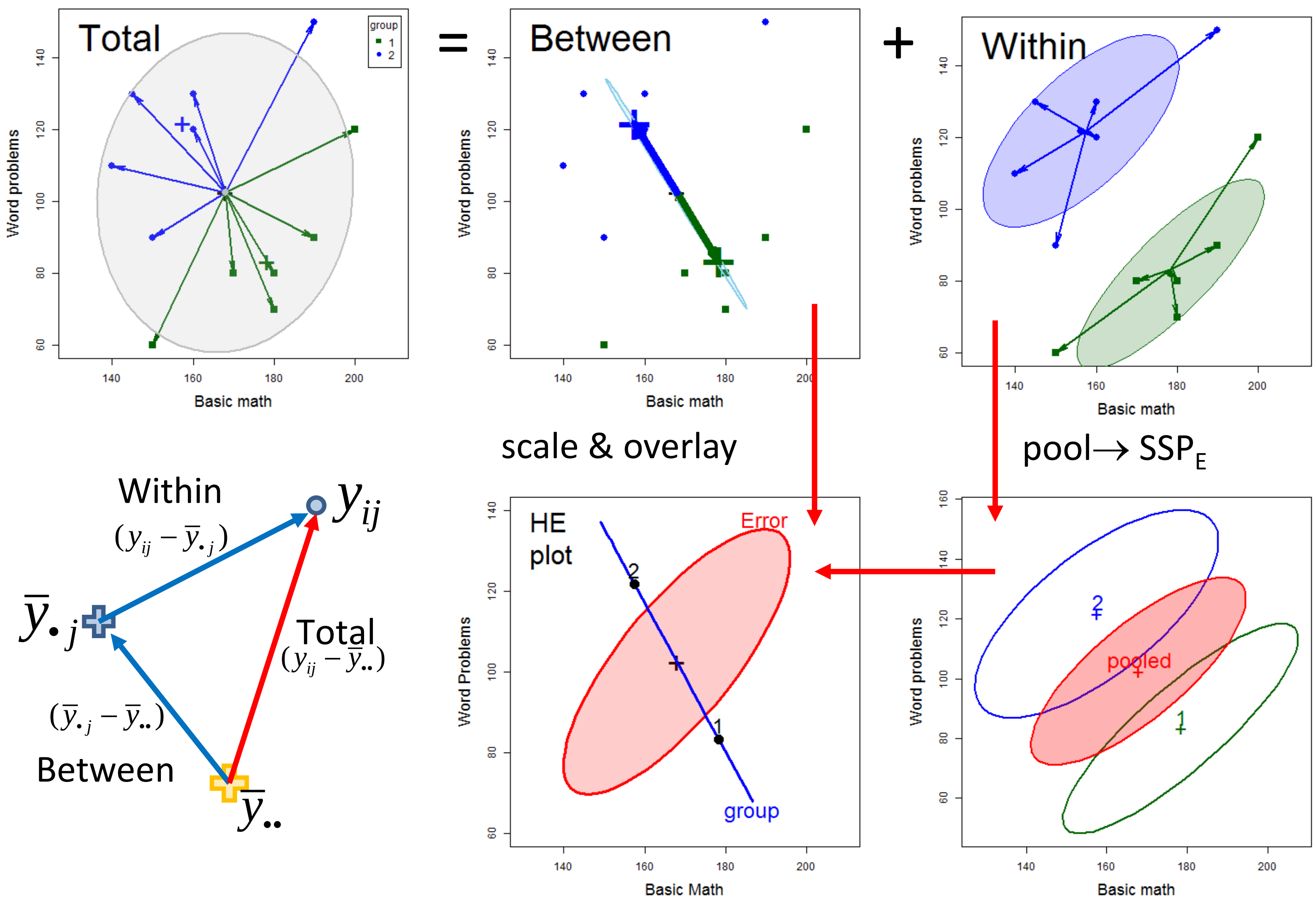
$$= \mathbf{SSP}_H + \mathbf{SSP}_E = \mathbf{H} + \mathbf{E}$$

- Hypothesis tests
  - Eigenvalues $\lambda_i$, $i=1{:}p$ 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

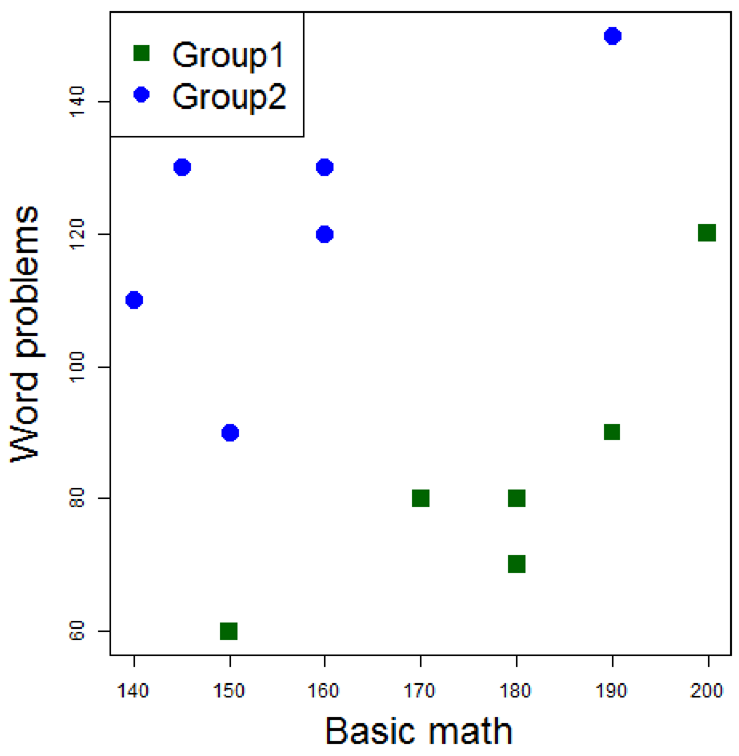
# Visualizing $SSP_T = SSP_H + SSP_E$



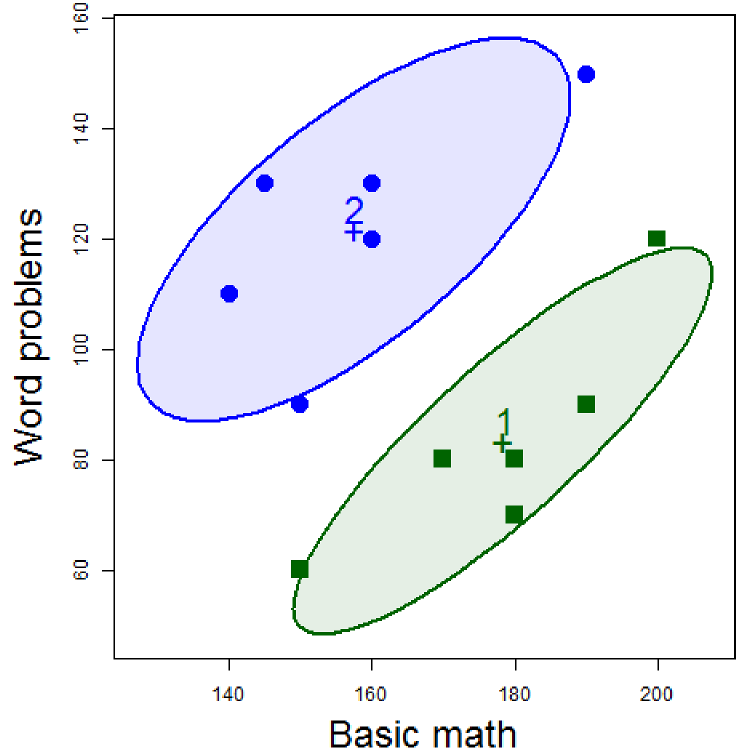
# Data ellipsoids

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

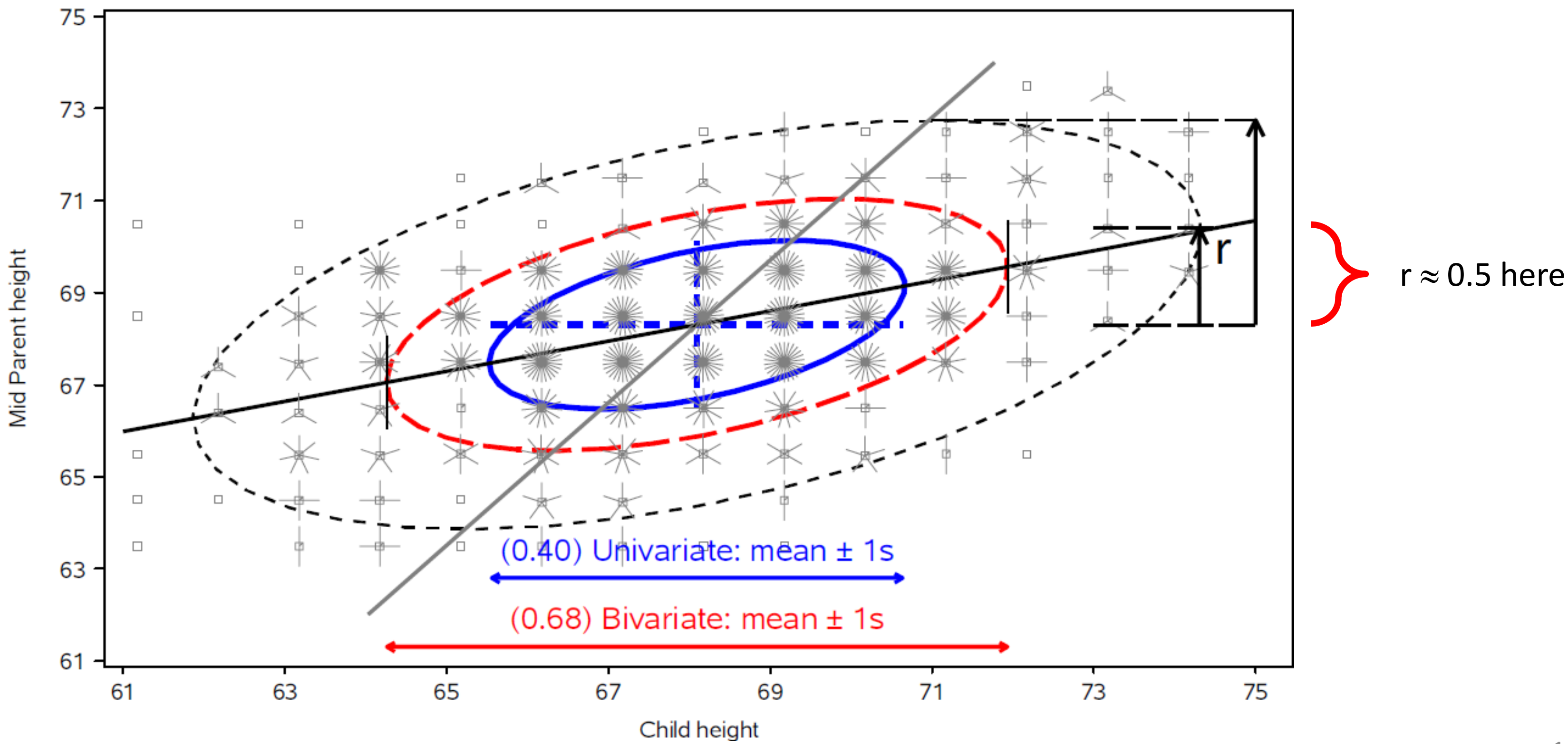
- For a *p*-dimensional multivariate sample, $\mathbf{Y}_{N \times p}$ , the sample mean vector, $\bar{\mathbf{y}}$, and sample covariance matrix, **S**, are minimally sufficient statistics under classical (gaussian) assumptions.
- These can be represented visually by the *p*-dimensional data ellipsoid,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

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

- $\rightarrow$ 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
```

# 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

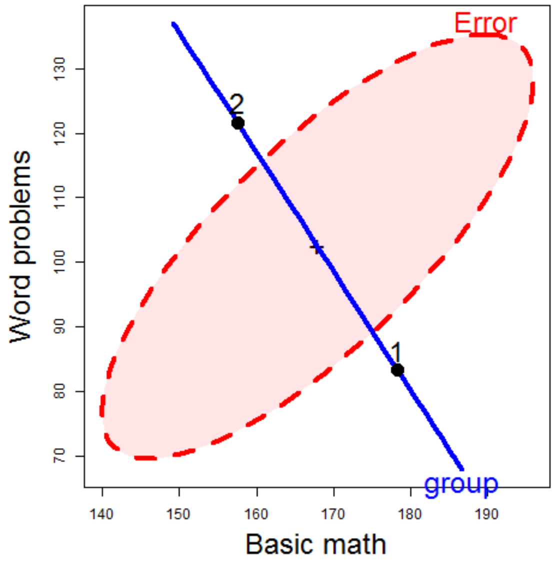
```
> mod <- lm(cbind(BM, WP) ~ group, data=mathscore)
> Anova(mod)

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:

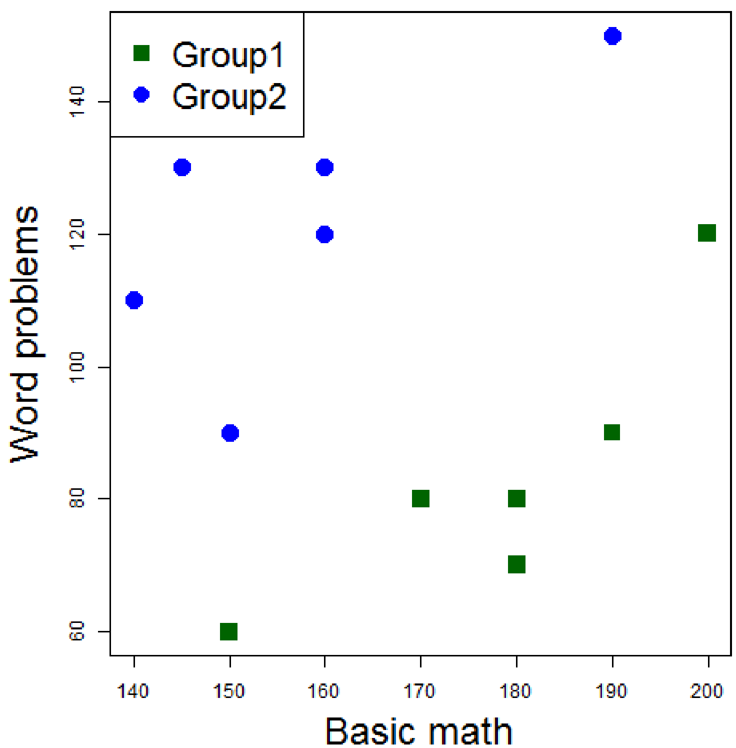
- Group 1 > Group 2 on Basic Math, but worse on Word Problems
- Group 2 > Group 1 on Word Problems, but worse on Basic Math
- BM & WP are + correlated w/in groups



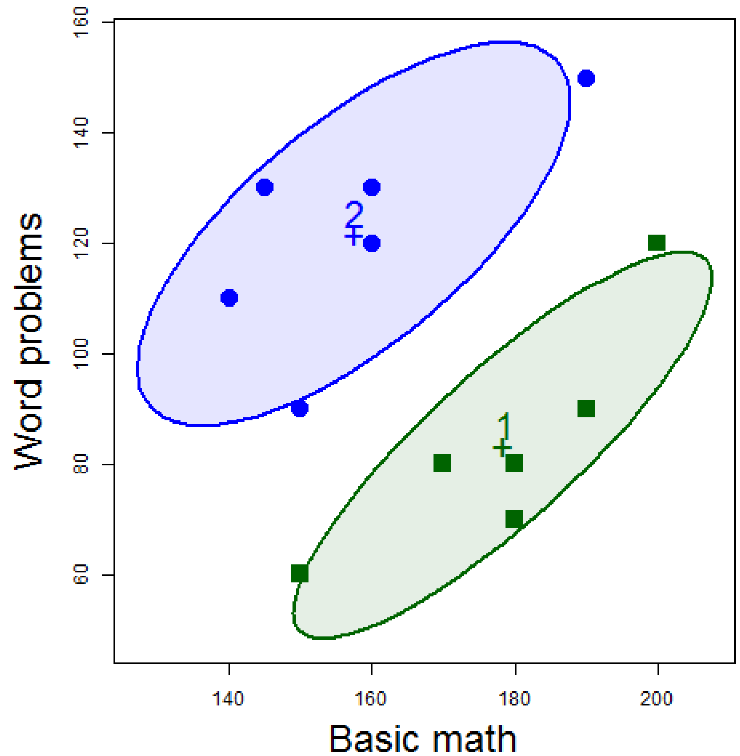
# HE plot framework: Visual overview

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

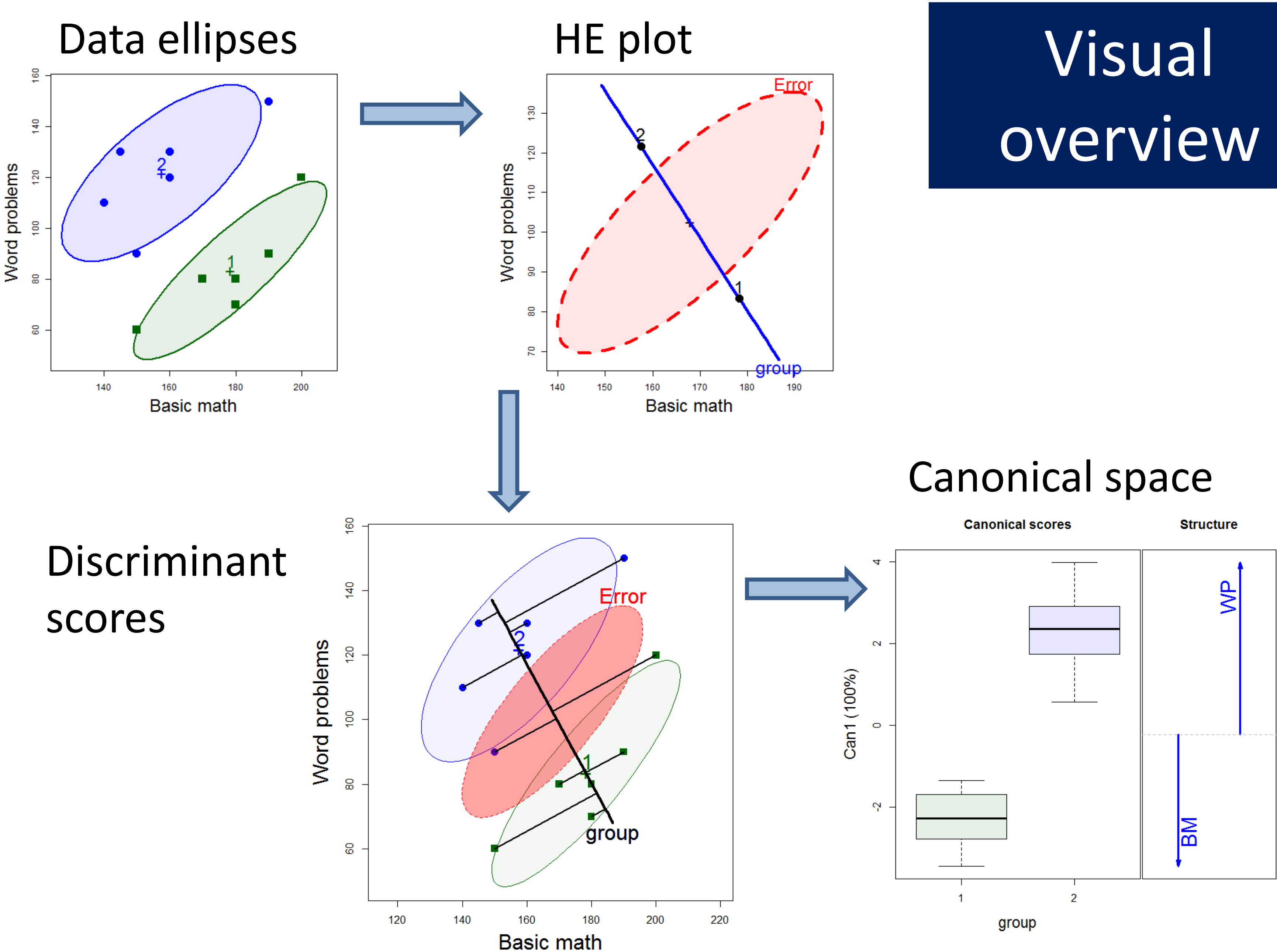
Data



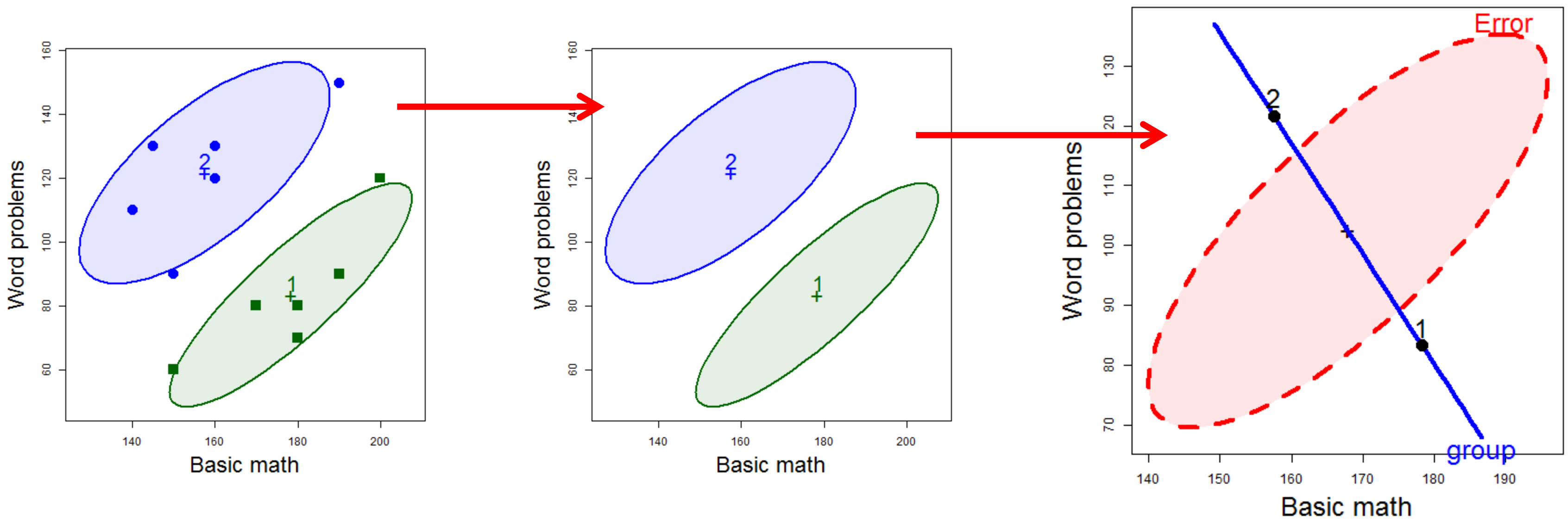

Data ellipses



# Visual overview

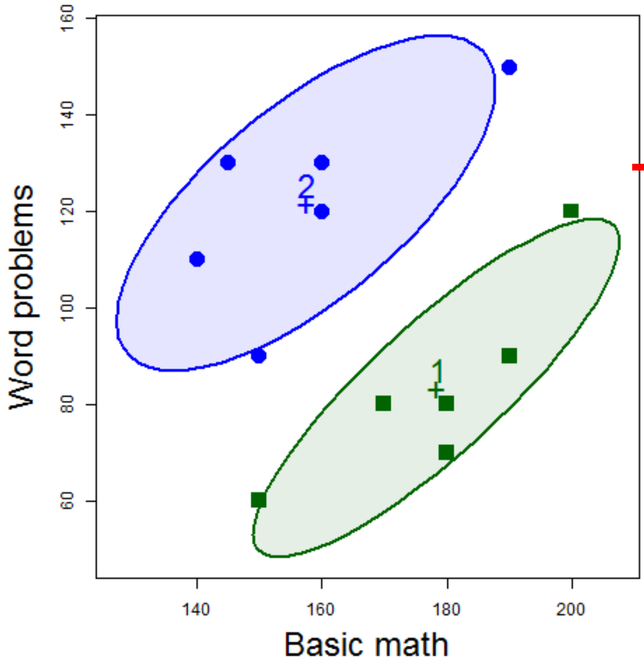


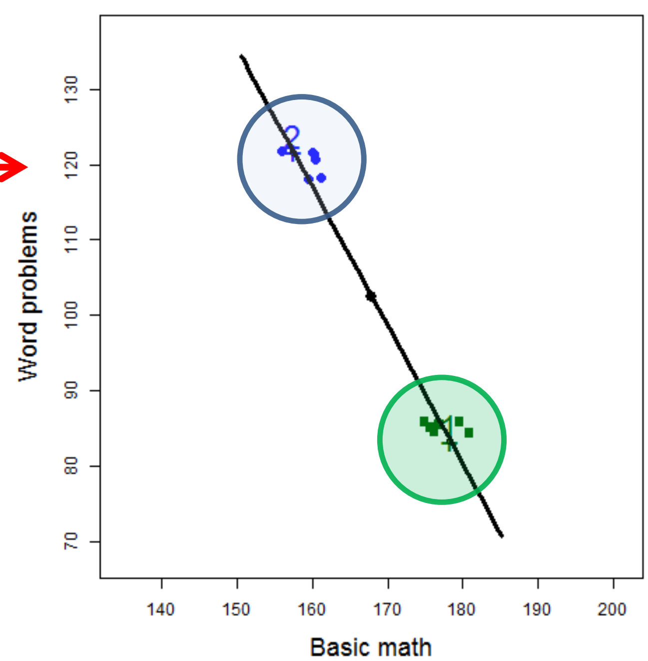
# Data → Data ellipses → HE plot




- 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



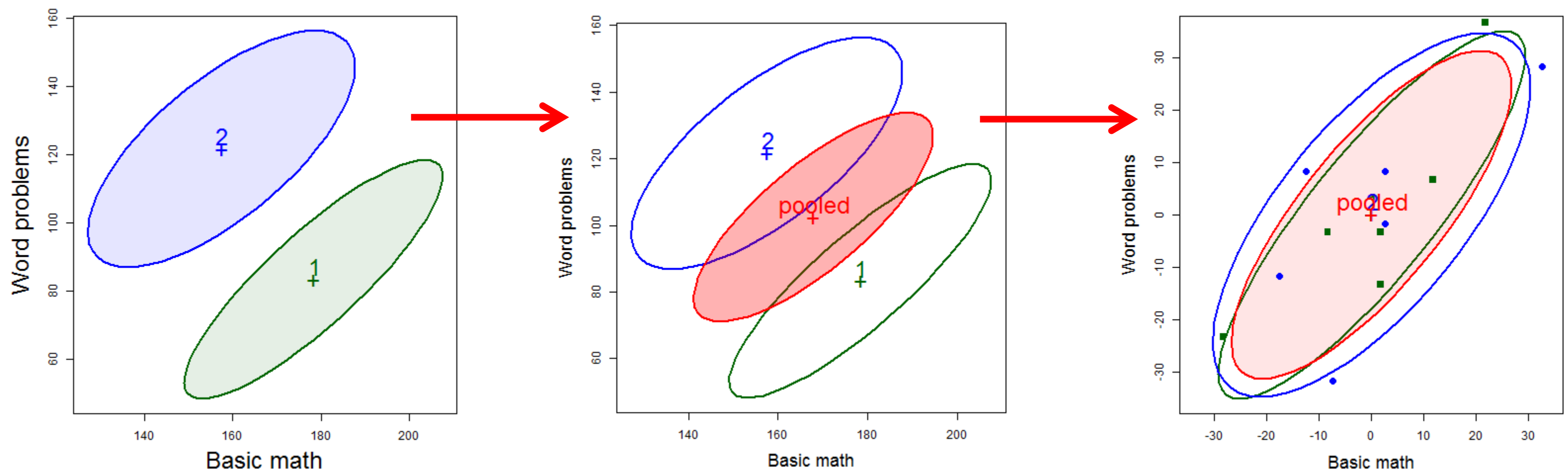




Jittered fitted values (gp means)

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

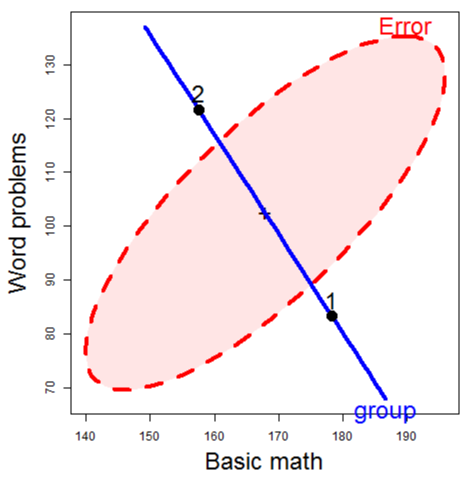
```
> resids <- residuals(math.mod)
> crossprod(resids)
       BM     WP
BM 3070.8 2808.3
WP 2808.3 4216.7
```

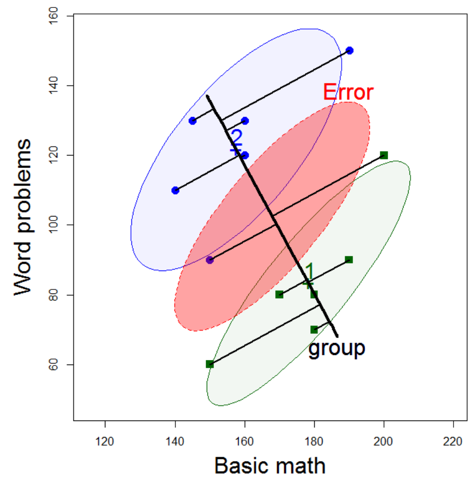
```
> cor(resids)
     BM   WP
BM 1.00 0.78
WP 0.78 1.00
```

```
> 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 $HE^{-1}$

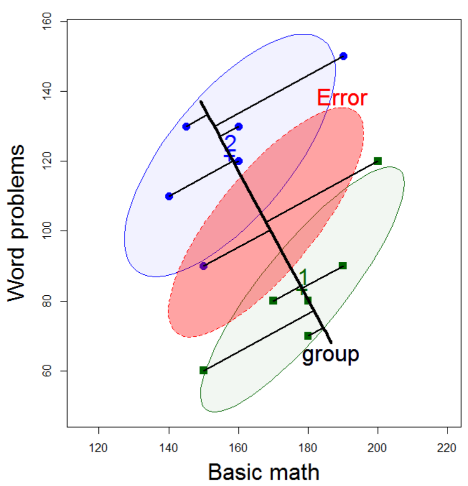


# Discriminant analysis

- 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

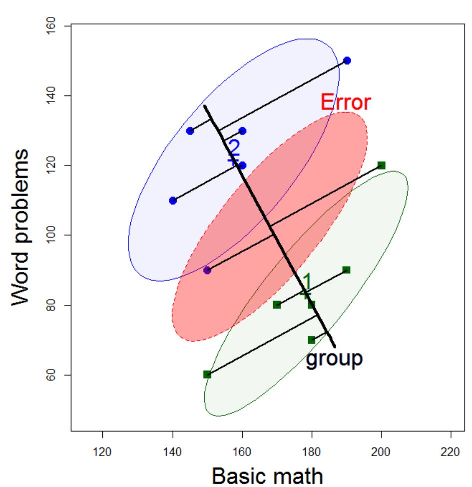
```
> (mod.lda <- MASS::lda(group ~ ., mathscore))

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
```

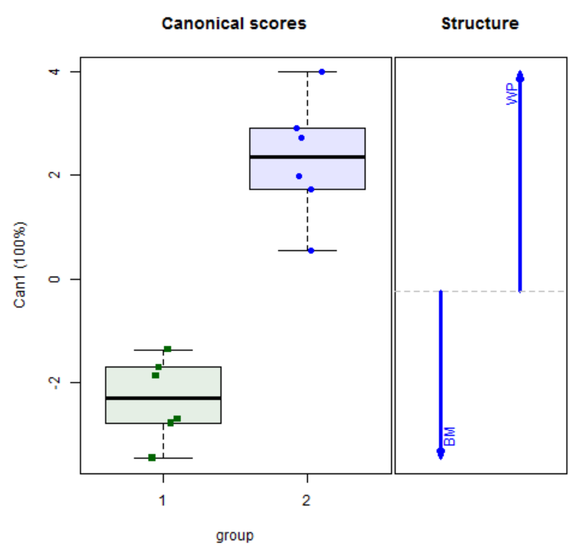



The canonical dimension is Can1 = 0.075 WP - 0.083 BM, a contrast between the two tests

# 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



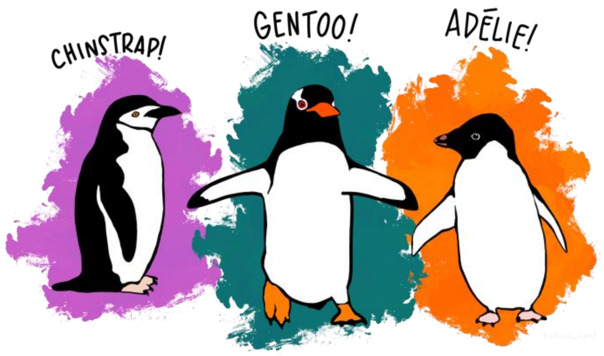



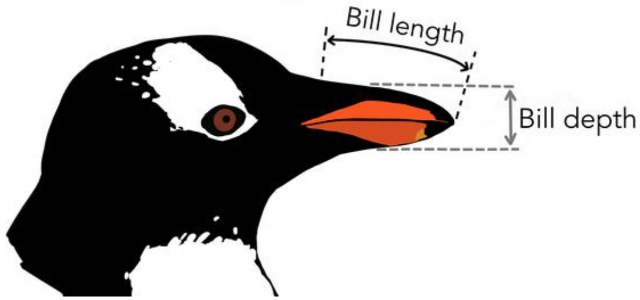

```
library(candisc)
mod.can <- candisc(math.mod)
plot(mod.can)
```

# 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
```

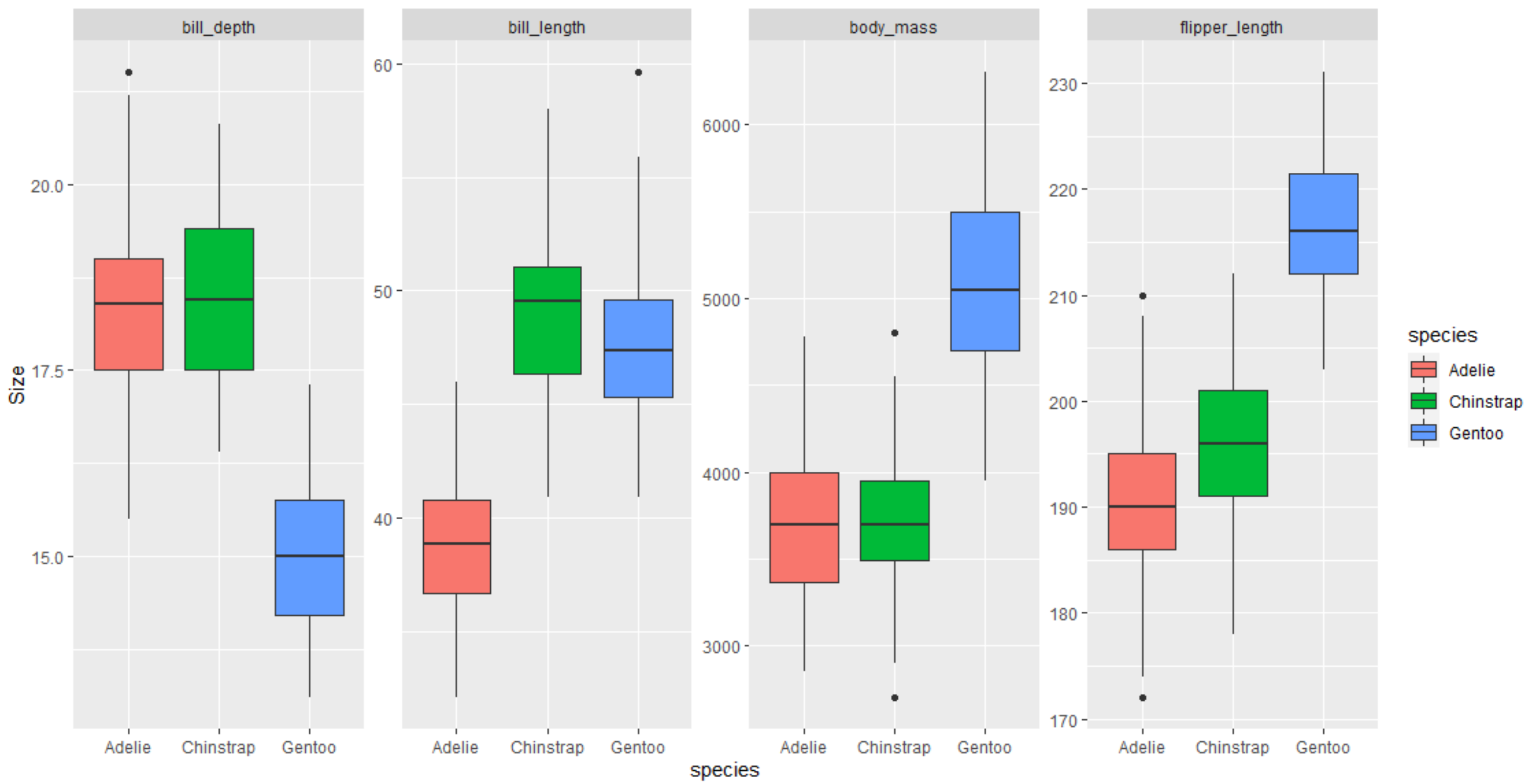


# Penguins: Multivariate EDA

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?

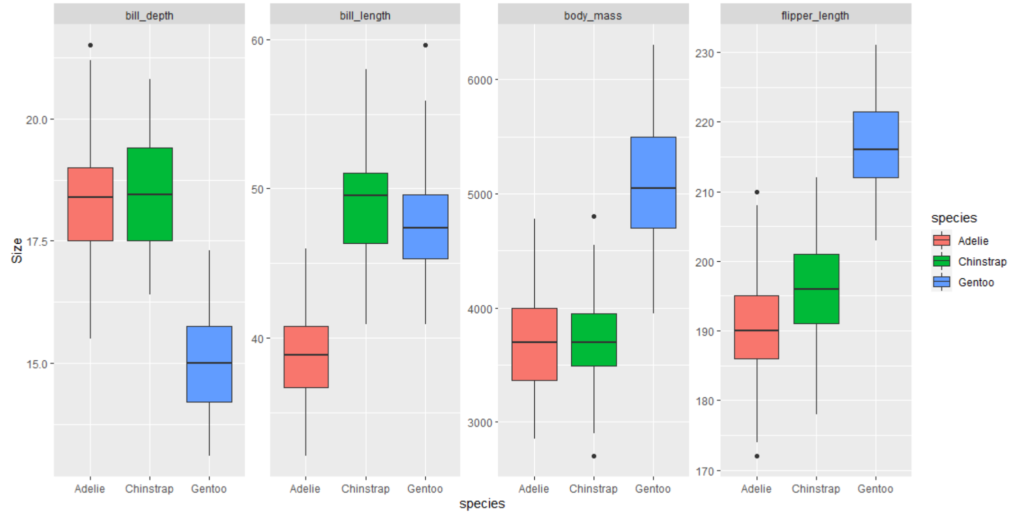


# Penguins: Multivariate EDA

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

- Need to reshape data from wide to long format

```
peng_long <- peng %>%
  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)
```



# 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

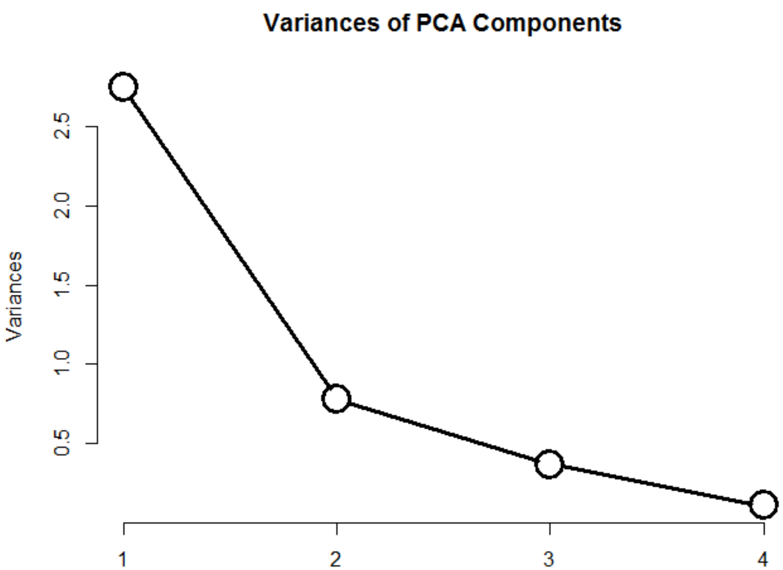
# PCA

```
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](https://rpubs.com/friendly/penguin-biplots) for details

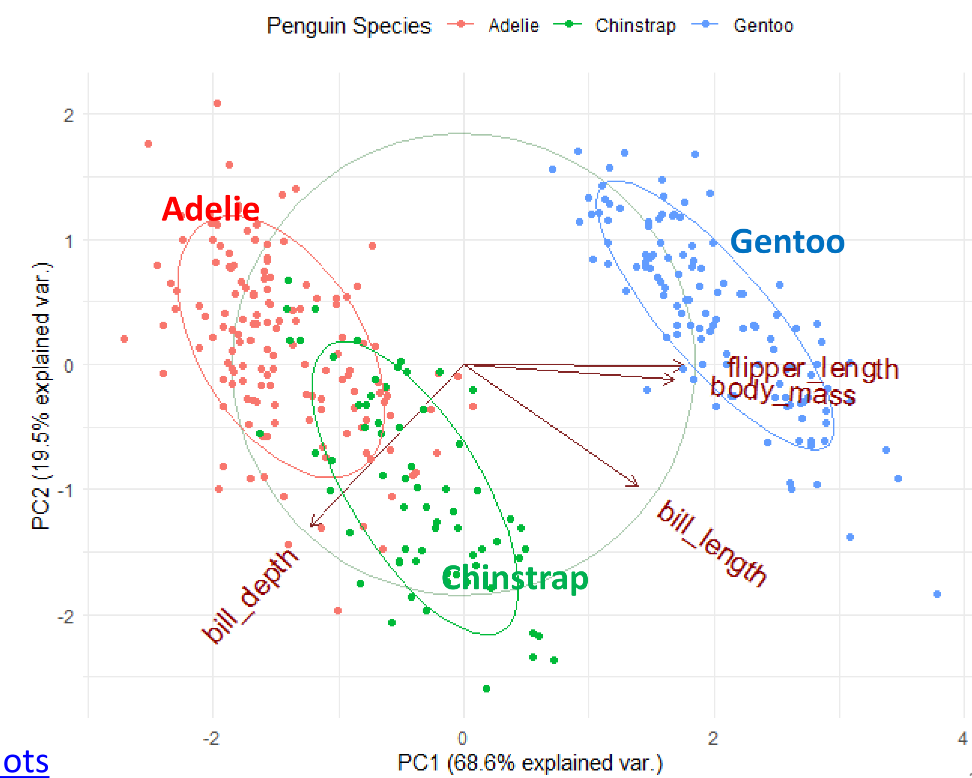
# Biplot

```
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 + island + sex

```
> 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
```

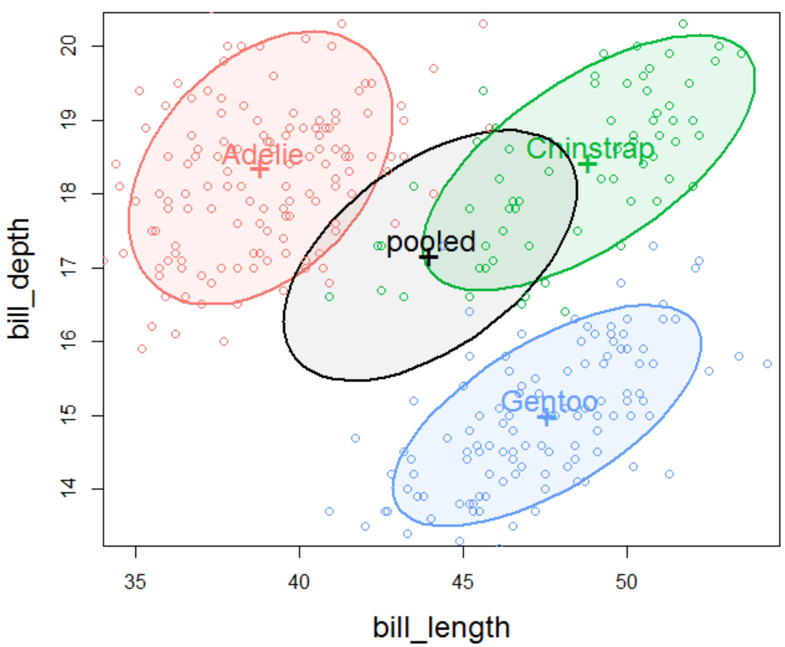
Yet, we are left to understand the nature of this effect wrt. the size variables.

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

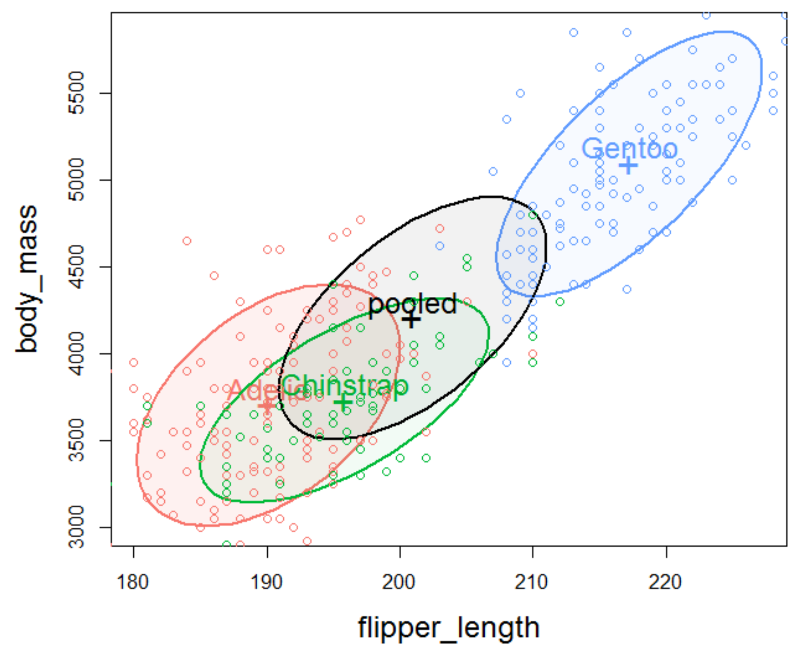
# Penguins: view data ellipses

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

bill depth & length




body mass & flipper length

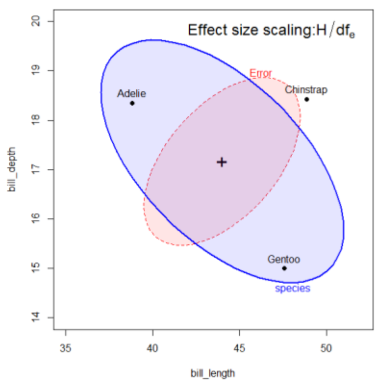


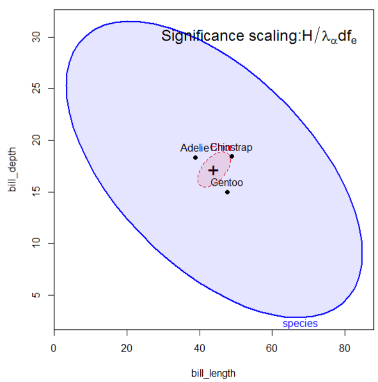

- group means negatively correlated
- within group correlation > 0

- 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 $\mathbf{H}/df_e$ to put this on the same scale as the **E** ellipse. Analog of effect size in univariate designs.
  - **significance** ("evidence") scaling: uses $\mathbf{H}/\lambda_\alpha \, df_e$ .
    - The **H** ellipse protrudes outside the **E** ellipse somewhere, *iff* an effect is significant (Roy's largest root test) at $p < \alpha$

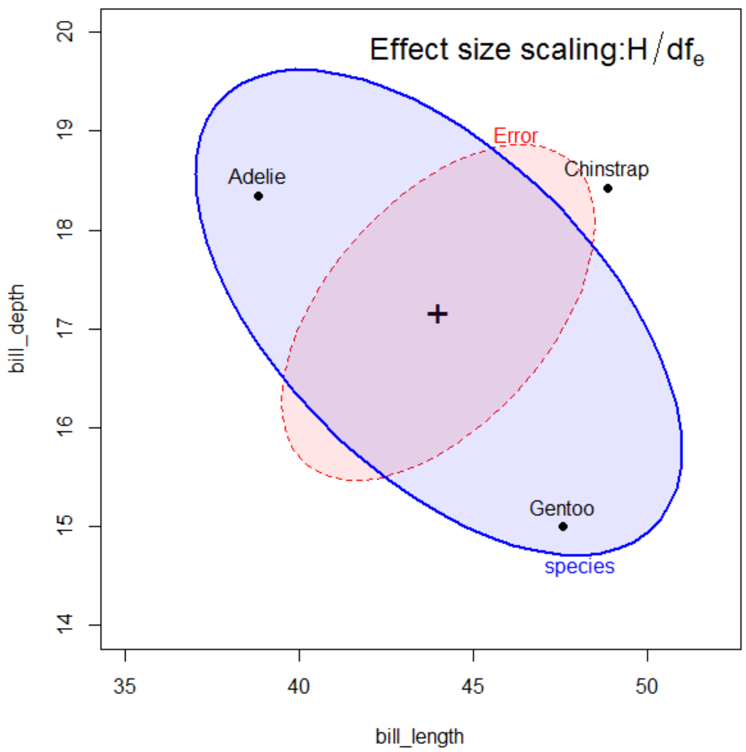


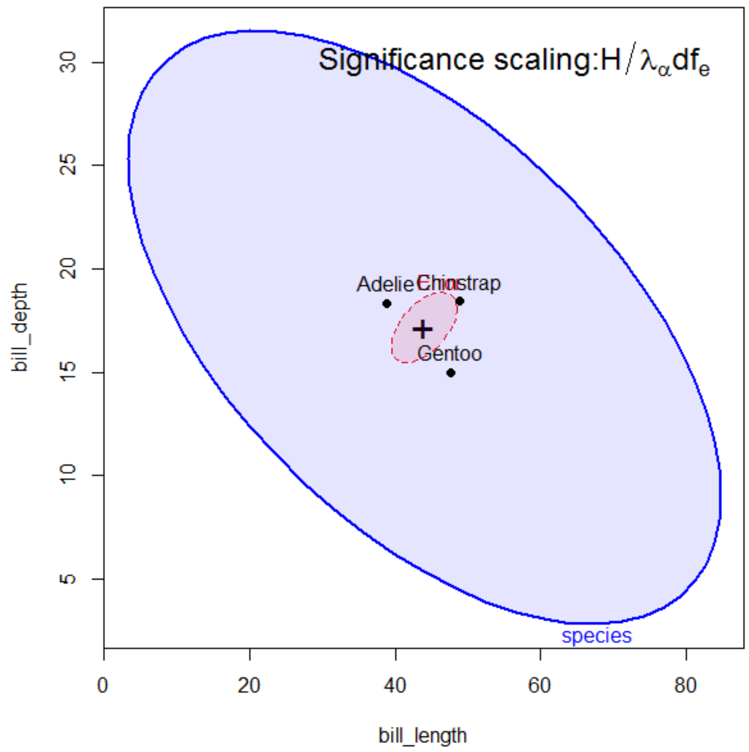



# Penguins: HE plots

Orientation of the **H** ellipse reflects negative correlation of the species means: species with larger bill depth have smaller bill length.

**E** ellipse: within species, larger bill length $\rightarrow$ larger bill depth




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

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

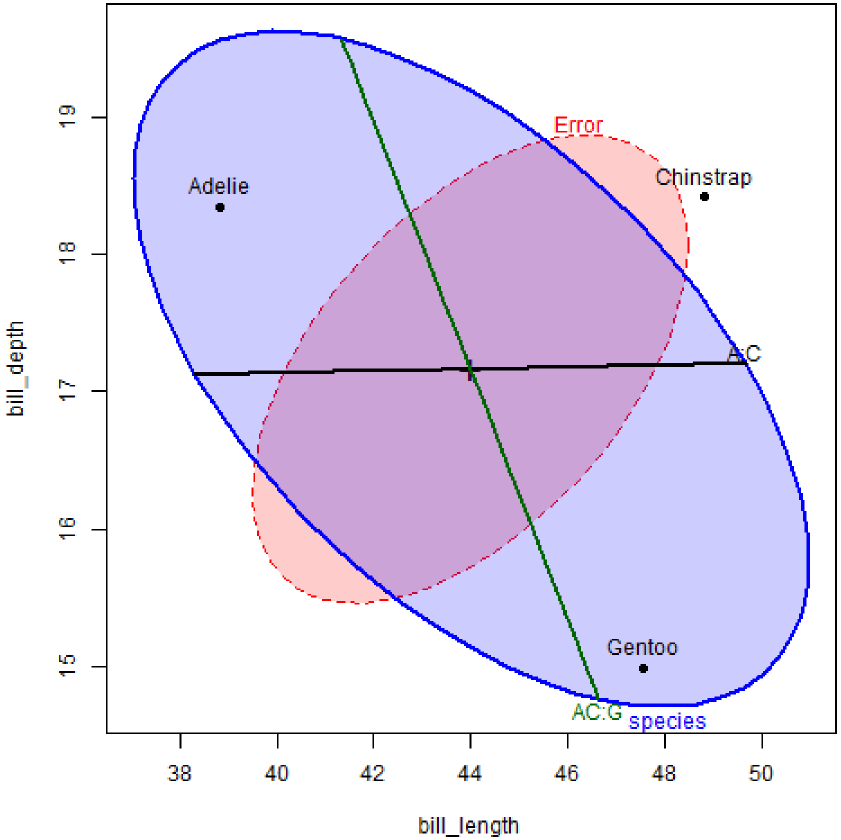
# Contrasts

- 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.

Both of these are large effects!

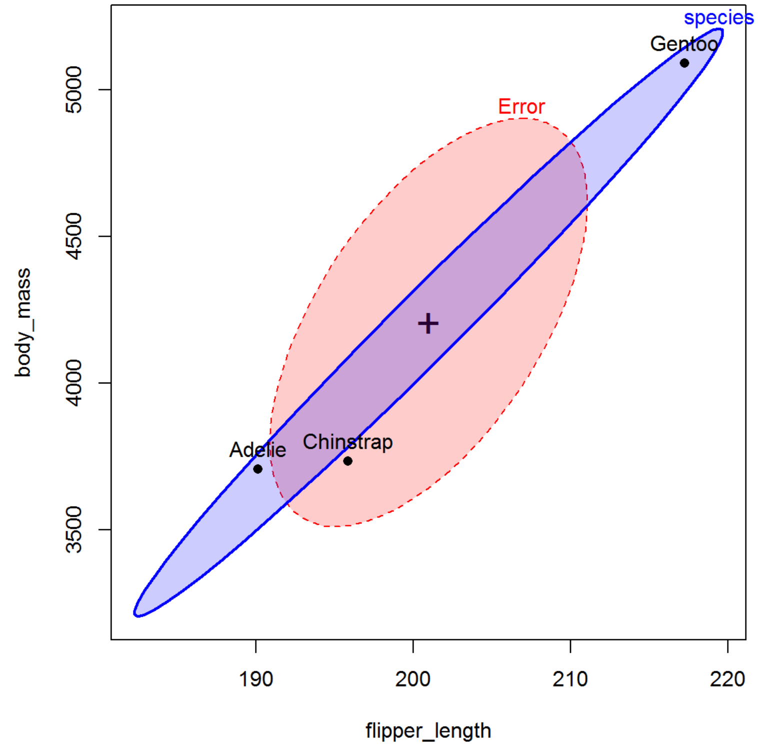
# Other HE plots

- 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")
```

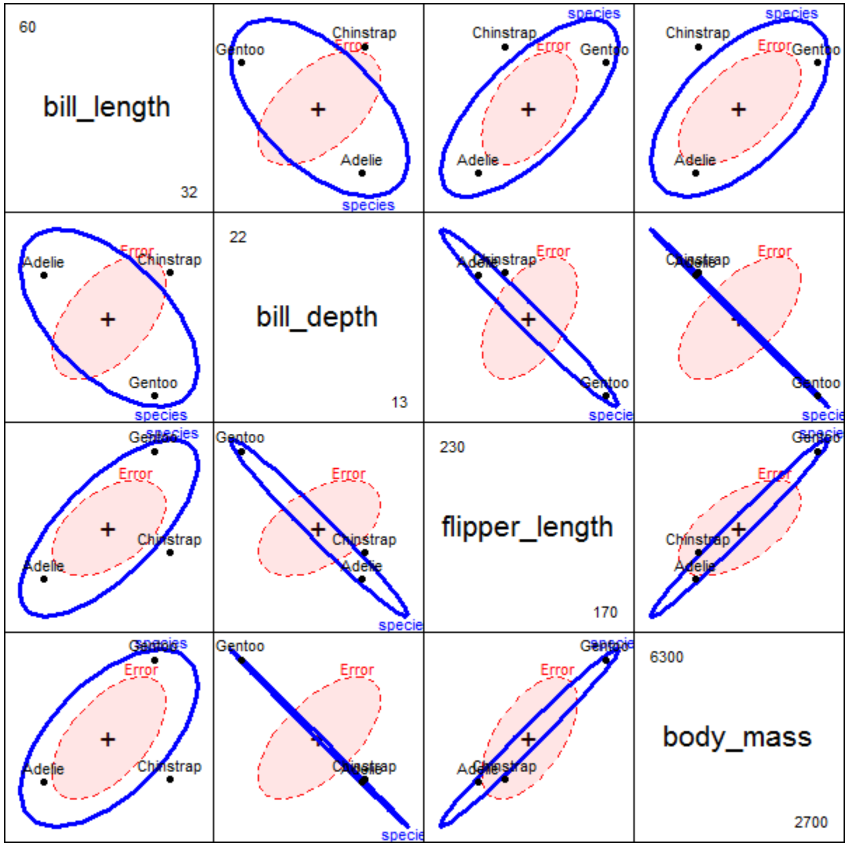
Interpretation:

- 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 a 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

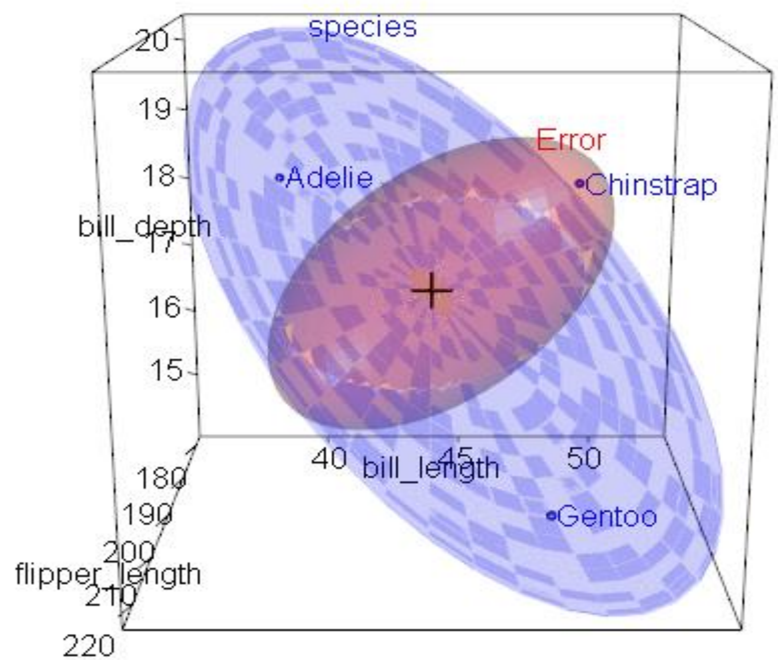
# heplot3d()

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
```

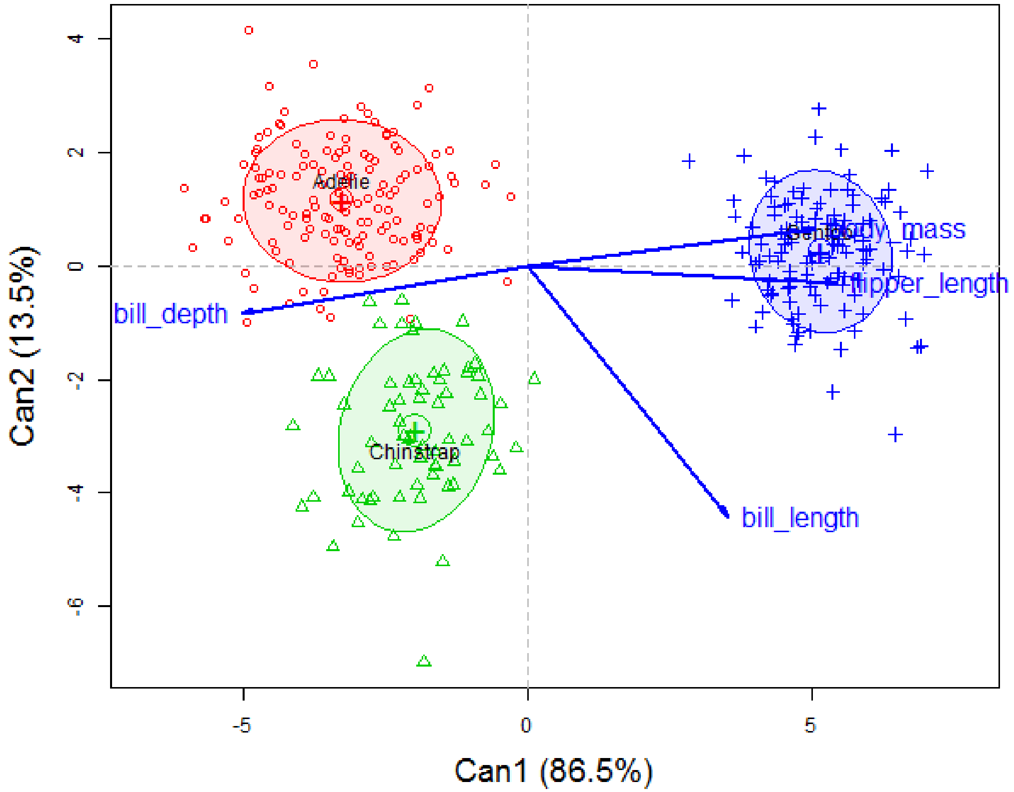
# Canonical view

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

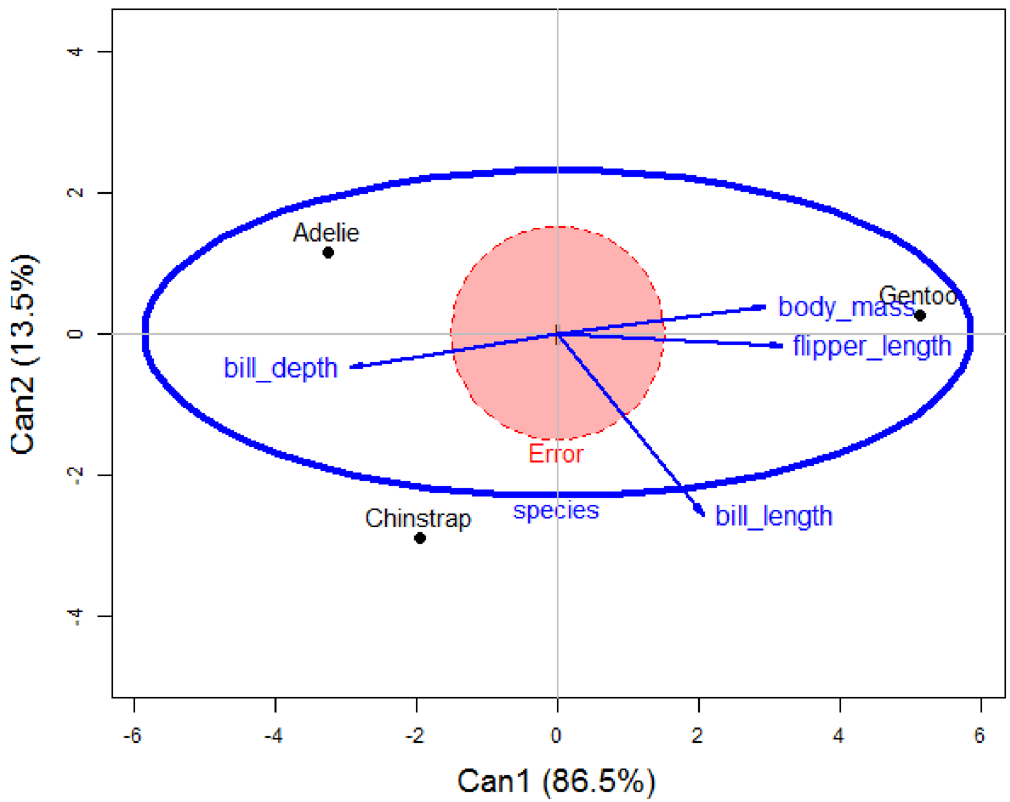
```
heplot(peng.can, size="effect", fill=c(TRUE, FALSE))
```

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.



# Summary

- MLM just like univariate LM, but for multiple responses
  - Simultaneous tests – no need for p-value adjustment
  - Take correlations among responses into account
- Data ellipses
  - Summarize bivariate data to show means, variances, correlation
- 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.