

Visualizing Linear Models: An R Bag of Tricks Session 3: Examples & Extensions

Michael Friendly
SCS Short Course
Oct-Nov, 2022

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

Today's topics

- Manipulating & visualizing models
- MANOVA examples
 - Distinguishing among psychiatric groups
 - Robust MLMs: down-weighting outliers
- Multivariate regression
 - PA tests & ability
 - Canonical correlation
 - MANCOVA & homogeneity of regression
- Homogeneity of (co)variance
 - Visualizing Box's M test

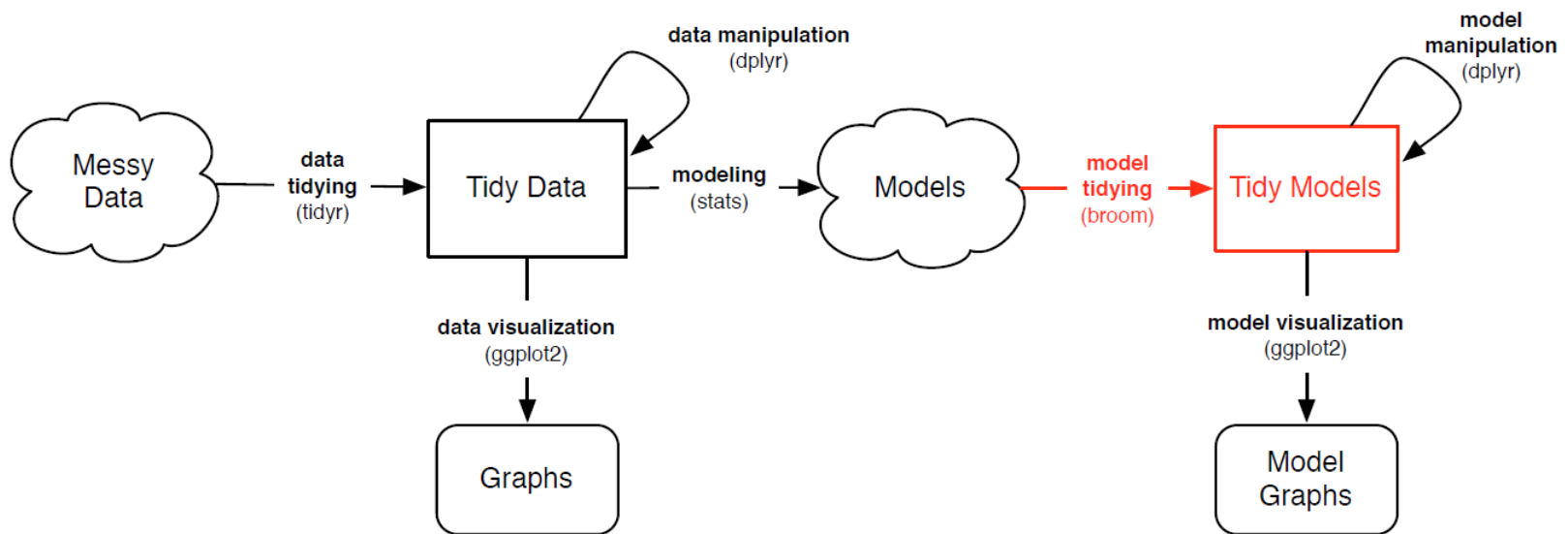
Manipulating & visualizing models

- R modeling functions [`lm()`, `glm()`, ...] return model objects, but these are “messy”
 - extracting coefficients takes several steps: `data.frame(coef(mymod))`
 - some info (R^2 , F , p .value) is computed in `print()` method, not stored
 - can't easily combine models
- Some have associated plotting functions
 - `plot(model)`: diagnostic plots
 - `car` package: many model plot methods
 - `effects` package: plot effects for model terms
- But what if you want to:
 - make a table of model summary statistics
 - fit a **collection** of models, compare, summarize or visualize them?



broom: manipulating models

- The broom package turns model objects into tidy data frames
 - **glance**(models) extracts model-level summary statistics (R^2 , df, AIC, BIC)
 - **tidy**(models) extracts coefficients, SE, p-values
 - **augment**(models) extracts observation-level info (residuals, ...)





Example: Penguin data

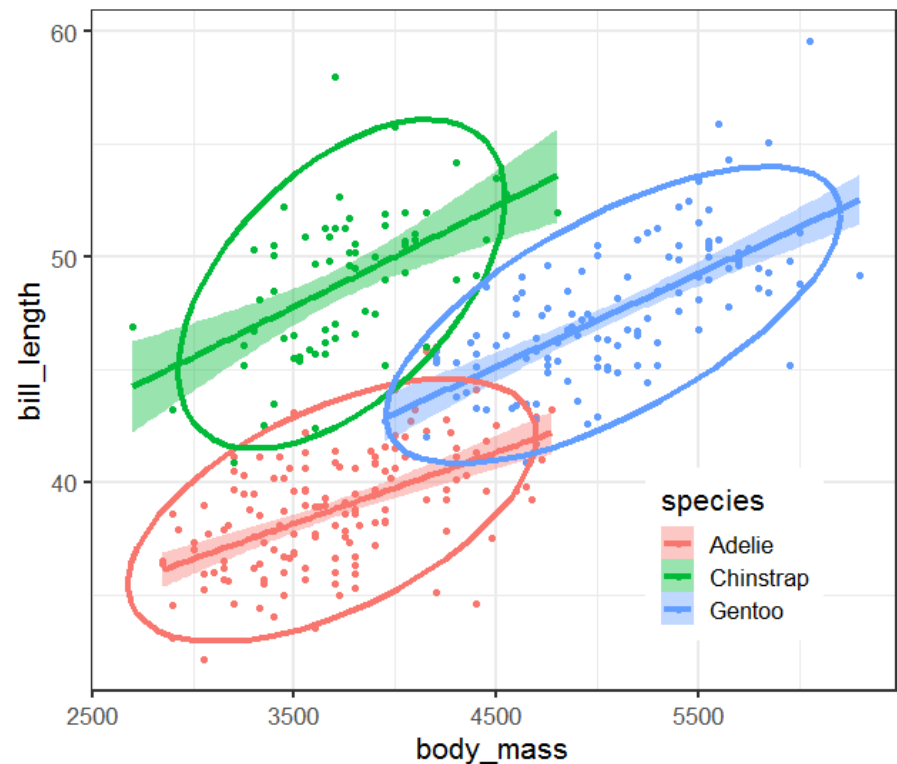
Say we want to model the length of penguin bills

- Potential predictors: body_mass, species, sex, island, ...

```
ggplot(data=peng, aes(x=body_mass, y=bill_length, color=species, fill=species)) +  
  geom_point() +  
  geom_smooth(method = "lm", size=1.5) +  
  stat_ellipse(size=1.5)
```

How to extract & plot model statistics?

How to fit multiple models for subsets (by: sex, island)?



Predict bill_length from body_mass, species, sex, island

```
peng.mod0 <- lm(bill_length ~ body_mass + species + sex + island, data=peng)
summary(peng.mod0)
```

Call:

```
lm(formula = bill_length ~ body_mass + sex + species + island, data = peng)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.938	-1.473	0.002	1.257	10.490

observation level

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	31.237676	1.378201	22.67	< 2e-16	***
body_mass	0.001744	0.000394	4.43	1.3e-05	***
sexm	2.532799	0.361396	7.01	1.4e-11	***
speciesChinstrap	10.281510	0.410391	25.05	< 2e-16	***
speciesGentoo	6.149100	0.672481	9.14	< 2e-16	***
islandDream	-0.463555	0.457443	-1.01	0.31	
islandTorgersen	0.092246	0.474401	0.19	0.85	

component level
(coefficients)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.26 on 326 degrees of freedom

Multiple R-squared: 0.832, Adjusted R-squared: 0.829

F-statistic: 269 on 6 and 326 DF, p-value: <2e-16

model level

glance() gives the **model level** summary statistics

```
> glance(peng.mod0)
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC    BIC deviance df.residual
1     0.832       0.829   2.26     269. 4.60e-123     6  -741. 1497. 1528.    1667.         326
```

tidy() gives the **model component** (term) statistics

```
> tidy(peng.mod0)
  term                estimate std.error statistic  p.value
1 (Intercept)         31.2      1.38      22.7    5.90e-69
2 body_mass           0.00174 0.000394      4.43    1.29e- 5
3 sexm                2.53      0.361      7.01    1.39e-11
4 speciesChinstrap   10.3      0.410     25.1    5.61e-78
5 speciesGentoo       6.15      0.672      9.14    6.65e-18
6 islandDream        -0.464     0.457     -1.01    3.12e- 1
7 islandTorgersen     0.0922    0.474      0.194   8.46e- 1
```

augment() gives the **observation level** statistics

```
> augment(peng.mod0) %>% sample_n(5)
  bill_length body_mass sex species island .fitted .resid   .hat .sigma   .cooksd .std.resid
  <dbl>      <int> <fct> <fct>   <fct>   <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>
1     35.7     3150 f    Adelie  Biscoe  36.7 -1.03  0.0273  2.26 0.000858 -0.463
2     39.7     3550 m    Adelie  Biscoe  40.0 -0.263 0.0331  2.26 0.0000684 -0.118
3     41.1     4050 m    Adelie  Biscoe  40.8  0.265 0.0257  2.26 0.0000532  0.119
4     49.3     4050 m    Chinstrap Dream  50.7 -1.35 0.0177  2.26 0.000939 -0.604
5     48.4     5850 m    Gentoo   Biscoe  50.1 -1.72 0.0169  2.26 0.00145 -0.769
```

Fitting multiple models

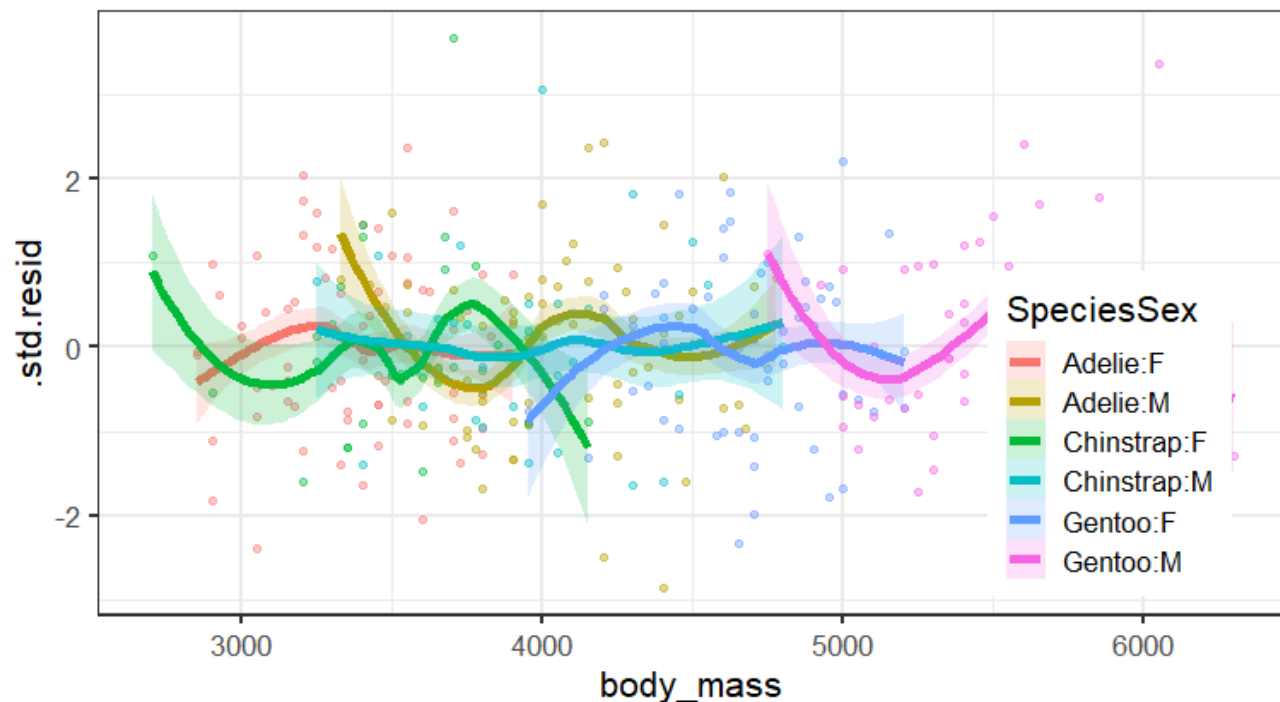
With `dplyr::nest_by()`, you can fit one model for each subset of the data
Feed this to `broom::glance()` to get model-level statistics

```
models <- peng |>  
  nest_by(species, sex) |>  
  mutate(model = list(lm(bill_length ~ body_mass, data = data)))  
models |>  
  summarise(broom::glance(model), .groups = "keep") |>  
  select(species, sex, r.squared, sigma, statistic, p.value, nobs)
```

```
# A tibble: 6 x 7  
# Groups:   species, sex [6]  
  species sex  r.squared sigma statistic p.value  nobs  
  <fct>   <fct>    <dbl> <dbl>    <dbl>   <dbl> <int>  
1 Adelie  f      0.0289  2.01      2.12  0.150    73  
2 Adelie  m      0.0486  2.24      3.62  0.0610    73  
3 Chinstrap f      0.0760  3.03      2.63  0.115    34  
4 Chinstrap m      0.0568  1.54      1.93  0.175    34  
5 Gentoo  f      0.0739  1.99      4.47  0.0389    58  
6 Gentoo  m      0.153   2.52     10.7  0.00182    61
```


Plotting via broom::augment

```
models |>
  summarise(broom::augment(model), .groups = "keep") |>
  mutate(
    sex = stringr::str_to_upper(sex),
    SpeciesSex = forcats::fct_cross(species, sex)) |>      # Combine species and sex
  ggplot(aes(x=body_mass, y=.std.res, color=SpeciesSex, fill=SpeciesSex)) +
  geom_point(alpha=0.4) +
  geom_smooth(method = "lm", size = 2, alpha=0.3)
```



Ex: Neuro- & Social-Cognitive measures in psychiatric groups

- A study by Leah Hartman @York examined whether patients classified as 'schizophrenic' or 'schizoaffective' (on DSM-IV) could be distinguished from a normal, control sample on standardized tests in the following domains:
 - **Neuro-Cognitive**: processing speed, attention, verbal learning, visual learning, problem solving
 - **Social-cognitive**: managing emotions, theory of mind, externalizing bias, personalizing bias
- Research questions → MANOVA contrasts
 - Analyze neuro-cog (NC) and social-cog (SC) separately
 - Do the two psychiatric groups differ from the controls?
 - Do the psychiatric groups differ from each other?

Schizophrenia symptoms: Hallucinations, disorganized thinking, delusions, ...

Schizoaffective disorder combines symptoms of schizophrenia with mood disorder (bipolar or depression)



Neuro-cognitive measures

Data

```
> data(NeuroCog, package="heplots")  
> car::some(NeuroCog)
```

	Dx	Speed	Attention	Memory	Verbal	Visual	ProbSolv	SocialCog	Age	Sex
14	Schizophrenia	19	9	19	33	24	39	28	44	Female
88	Schizoaffective	27	44	24	33	26	31	36	53	Female
105	Schizoaffective	23	41	42	48	46	29	50	52	Female
114	Schizoaffective	41	53	47	39	30	50	63	32	Female
130	Control	44	25	21	37	32	43	29	43	Female
165	Control	35	35	43	53	57	37	40	30	Male
194	Control	47	40	53	53	40	45	49	55	Female

diagnostic group ----- neuro-cognitive measures ----- ----- ignored -----

Questions:

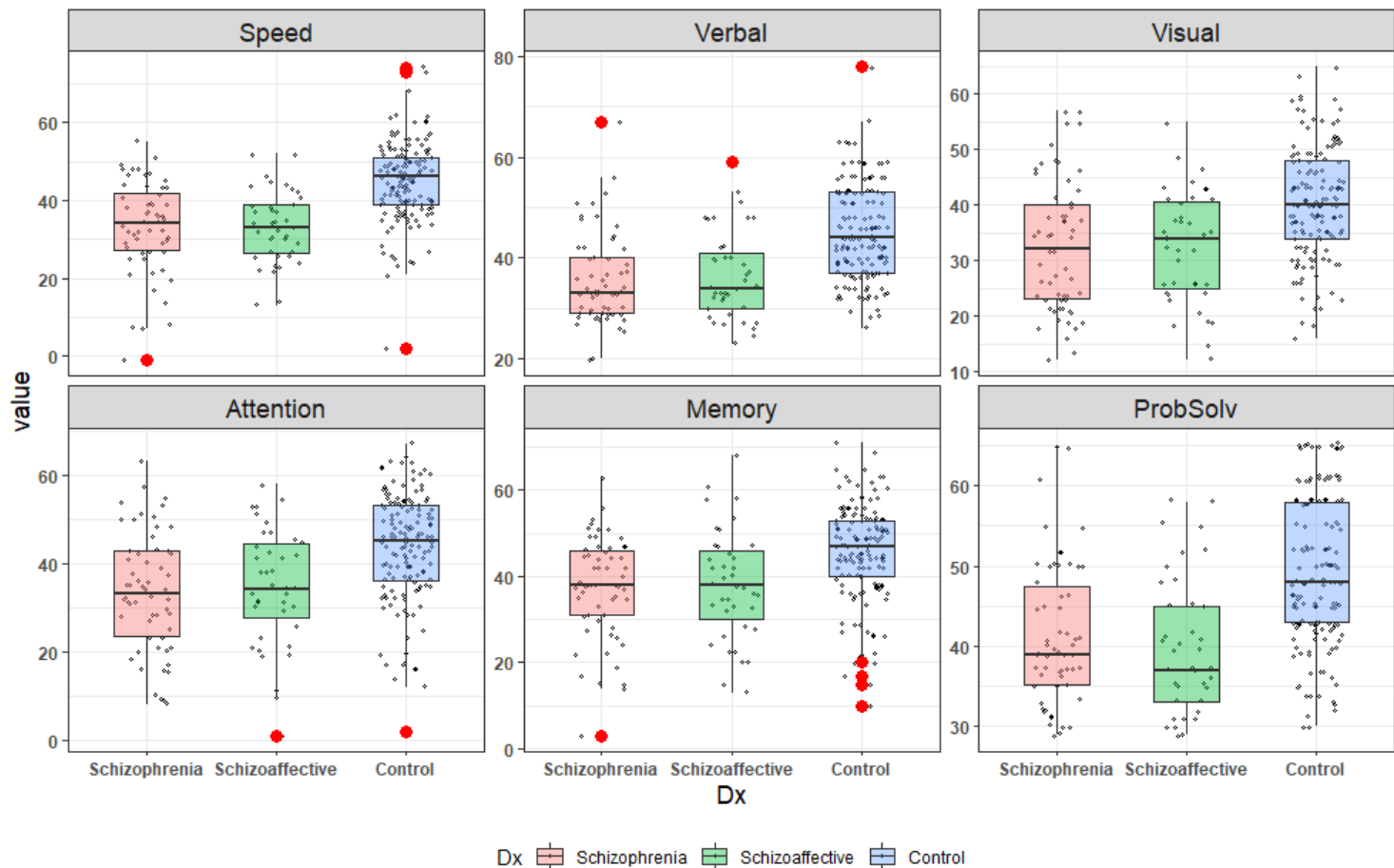
- Do the diagnostic groups differ **collectively** on the neuro-cognitive measures?
- How do group differences relate to **research hypotheses**?
- How many **dimensions** (aspects) are reflected in the differences among means?

NeuroCog: EDA

Always plot the data first!

With boxplots, often useful to show more: jittered points show the data

What can
we see
here?



How did I make this plot?

1. Reshape the data wide → long
2. `ggplot() + geom_jitter() + geom_boxplot() + facet_wrap(~response)`

```
NC_long <- NeuroCog |>  
  select(-SocialCog, -Age, -Sex) |>  
  tidyr::gather(key = response, value = "value", Speed:ProbSolv)
```

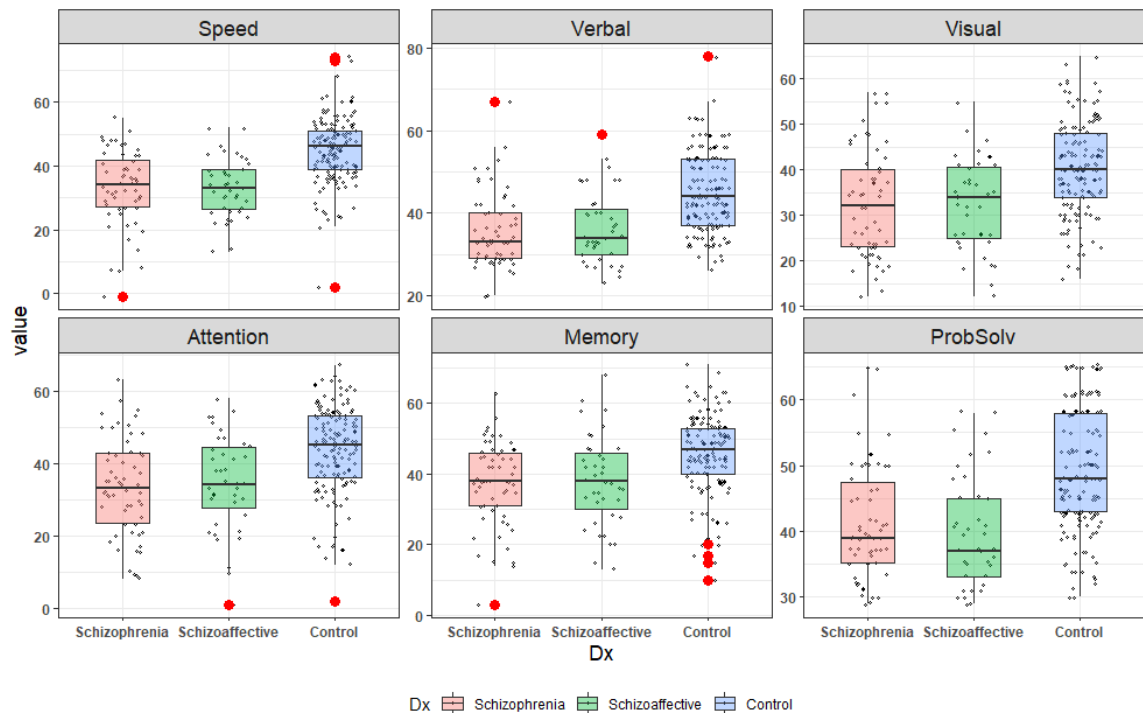
```
# show a few observations  
library(dplyr)  
NC_long |>  
  group_by(Dx) |>  
  sample_n(4) |>  
  ungroup()
```

	Dx	response	value
1	Schizophrenia	Verbal	37
2	Schizophrenia	Speed	49
3	Schizophrenia	ProbSolv	61
4	Schizophrenia	Visual	23
5	Schizoaffective	Attention	11
6	Schizoaffective	Speed	39
7	Schizoaffective	Visual	41
8	Schizoaffective	Memory	54
9	Control	Speed	57
10	Control	Memory	45
11	Control	Speed	41
12	Control	ProbSolv	61

How did I make this plot?

1. Reshape the data wide → long
2. `ggplot() + geom_jitter() + geom_boxplot() + facet_wrap(~response)`

```
ggplot(NC_long, aes(x=Dx, y=value, fill=Dx)) +  
  geom_jitter() +  
  geom_boxplot() +  
  facet_wrap(~response, scales = "free_y") +  
  theme_bw()
```



Neuro-cognitive measures: Analyses

- Could do a series of univariate ANOVAs

- `car::Anova(NC.mlm1 <- lm (Speed ~ DX, data=NeuroCog))`
 - `car::Anova(NC.mlm2 <- lm (Attention ~ DX, data=NeuroCog))`
 - `car::Anova(NC.mlm3 <- lm (Memory ~ DX, data=NeuroCog))`
 - ...

- Better: MANOVA test for **all 6 responses together**

```
Anova(NC.mlm = lm(cbind(Speed, Attention, Memory, ...) ~ Dx, data=NeuroCog )
```

- Research hypotheses: test contrasts

- Dx1: Control – (Schiz + SchizAff)/2 = 0 → `linearHypotheses(NC.mlm, "Dx1")`
 - Dx2: Schiz – SchAff = 0 → `linearHypotheses(NC.mlm, "Dx2")`

- Dimensions:

- Canonical analysis → `NC.can <- candisc(NC.mlm); plot(NC.can)`
 - Canonical HE plot → `heplot(NC.can)`

- Tech note: `anova()` in base R vs. `car::Anova`

- `anova()` uses only Type 1 (sequential) tests, rarely useful; doesn't handle MLM well
 - `car::Anova()` provides Type 2, 3 (partial) tests; give sensible results for MLMs
 - `car::linearHypotheses()` gives univariate and multivariate tests of contrasts

✗

✓

✓

Tidy univariate analyses: *F*-tests

Using `broom::tidy()` we can fit a separate univariate ANOVA model for each response
I've used some tidyverse magic to make a nicely formatted table of *F* values

```
NC_long |>
  mutate(response = factor(response, levels=unique(response))) |>      # keep variable order
  group_by(response) |>
  do(tidy(anova(lm(value ~ Dx, .)))) |>                                  # run anova on each submodel
  filter(term != "Residuals") |>
  select(-term) |>
  rename(F = statistic, df1 = df,                                       # use more specific names
         SS = sumsq, MS = meansq) |>
  mutate(df2 = 239) |>                                                  # kludge: extract dfe from object?
  relocate(df2, .after = df1) |>
  mutate(signif = noquote(gtools::stars.pval(p.value))) |>            # add significance stars
  mutate(p.value = noquote(scales::pvalue(p.value)))                   # format p.values nicely
```

response	df1	df2	SS	MS	F	p.value	signif
<fct>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<noquote>	<noquote>
1 Speed	2	239	8360.	4180.	37.1	<0.001	***
2 Attention	2	239	5579.	2790.	17.4	<0.001	***
3 Memory	2	239	3764.	1882.	13.9	<0.001	***
4 Verbal	2	239	4672.	2336.	27.3	<0.001	***
5 Visual	2	239	3692.	1846.	16.6	<0.001	***
6 ProbSolv	2	239	4165.	2083.	25.1	<0.001	***

**All variables
are signif.**

Tidy univariate analyses: *t*-tests

Can do the same to extract the *t*-tests for coefficients or contrasts in the univariate models for each response

```
NC_long |>
  mutate(response = factor(response, levels=unique(response))) |>
  group_by(response) |>
  do(tidy(lm(value ~ Dx, .))) |>
  filter(term != "(Intercept)") |>
  mutate(signif = noquote(gtools::stars.pval(p.value))) |>
  mutate(p.value = noquote(scales::pvalue(p.value)))
```

	response	term	estimate	std.error	statistic	p.value	signif
	<fct>	<chr>	<dbl>	<dbl>	<dbl>	<noquote>	<noquote>
1	Speed	Dx1	7.99	0.940	8.51	<0.001	***
2	Speed	Dx2	-0.0252	1.10	-0.0229	0.982	
3	Attention	Dx1	6.44	1.12	5.75	<0.001	***
4	Attention	Dx2	-0.558	1.31	-0.426	0.671	
5	Memory	Dx1	5.31	1.03	5.16	<0.001	***
6	Memory	Dx2	-0.375	1.20	-0.311	0.756	
7	Verbal	Dx1	5.91	0.818	7.22	<0.001	***
8	Verbal	Dx2	-0.442	0.957	-0.461	0.645	
9	Visual	Dx1	5.31	0.933	5.69	<0.001	***
10	Visual	Dx2	-0.0371	1.09	-0.0340	0.973	
11	ProbSolv	Dx1	5.70	0.807	7.07	<0.001	***
12	ProbSolv	Dx2	0.719	0.944	0.762	0.447	

Neuro-cognitive measures: MANOVA

Multivariate tests are more straight-forward & comprehensive

```
# fit the MANOVA model
NC.mlm <- lm(cbind(Speed, Attention, Memory, Verbal, Visual, ProbSolv) ~ Dx,
             data=NeuroCog)
Anova(NC.mlm)

Type II MANOVA Tests: Pillai test statistic
  Df test stat approx F num Df den Df      Pr(>F)
Dx  2      0.2992    6.8902     12    470 1.562e-11 *** ✓
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the groups differ. But how?

What about the research hypotheses?

```
> contrasts(NeuroCog$Dx)
      [, 1] [, 2]
Schizophrenia -0.5  1
Schizoaffective -0.5 -1
Control       1.0  0
```

Dx1: Control – (Schiz + SchizAff)/2 = 0

Neuro-cognitive measures: Contrasts

A simple result: Control \neq (Schizophrenia \approx Schizoaffective)

```
> print(car::linearHypothesis(NC.mlm, "Dx1"), SSP=FALSE)
```

Multivariate Tests:

	Df	test	stat	approx	F	num	Df	den	Df	Pr(>F)
Pillai	1		0.289		15.9	6		234		2.8e-15 ***
Wilks	1		0.711		15.9	6		234		2.8e-15 ***
Hotelling-Lawley	1		0.407		15.9	6		234		2.8e-15 ***
Roy	1		0.407		15.9	6		234		2.8e-15 ***



Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> print(car::linearHypothesis(NC.mlm, "Dx2"), SSP=FALSE)
```

Multivariate Tests:

	Df	test	stat	approx	F	num	Df	den	Df	Pr(>F)
Pillai	1		0.006		0.249	6		234		0.96
Wilks	1		0.994		0.249	6		234		0.96
Hotelling-Lawley	1		0.006		0.249	6		234		0.96
Roy	1		0.006		0.249	6		234		0.96



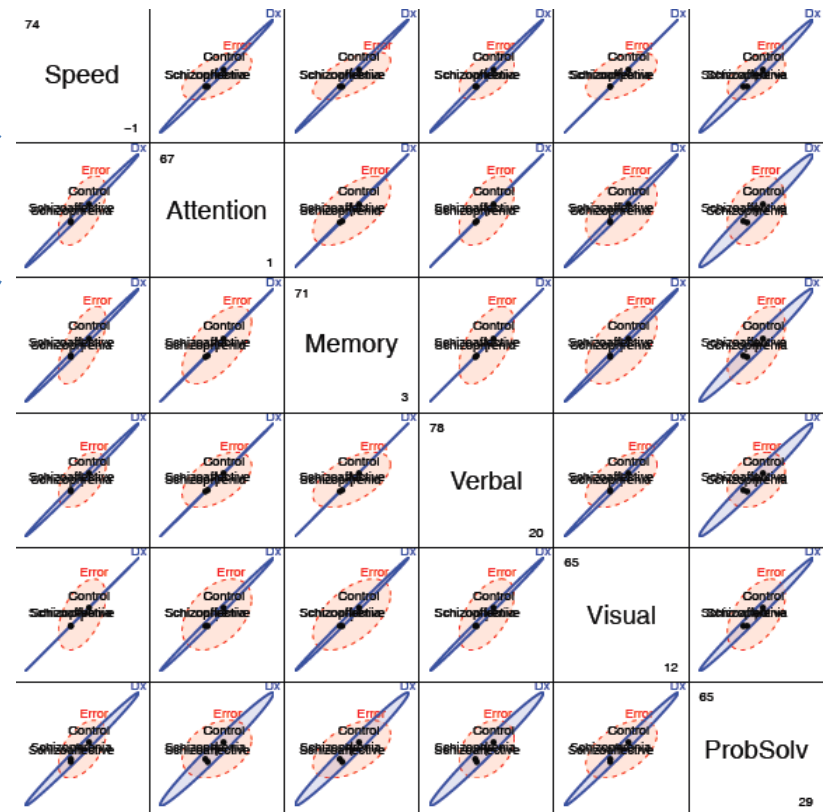
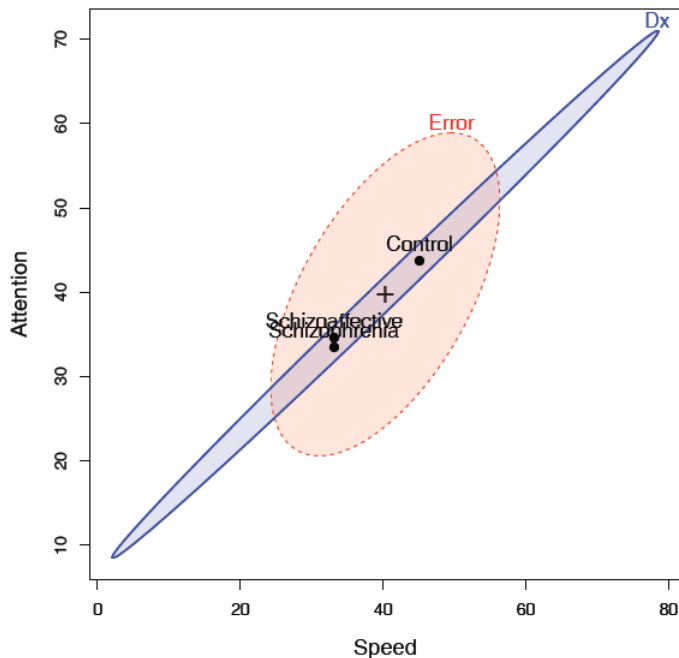
Visualize me: in data space

Bivariate view for any 2 responses:

```
heplot (NC. ml m, var=1:2, ...)
```

HE plot matrix: for all responses

```
pairs (NC. ml m, ...)
```

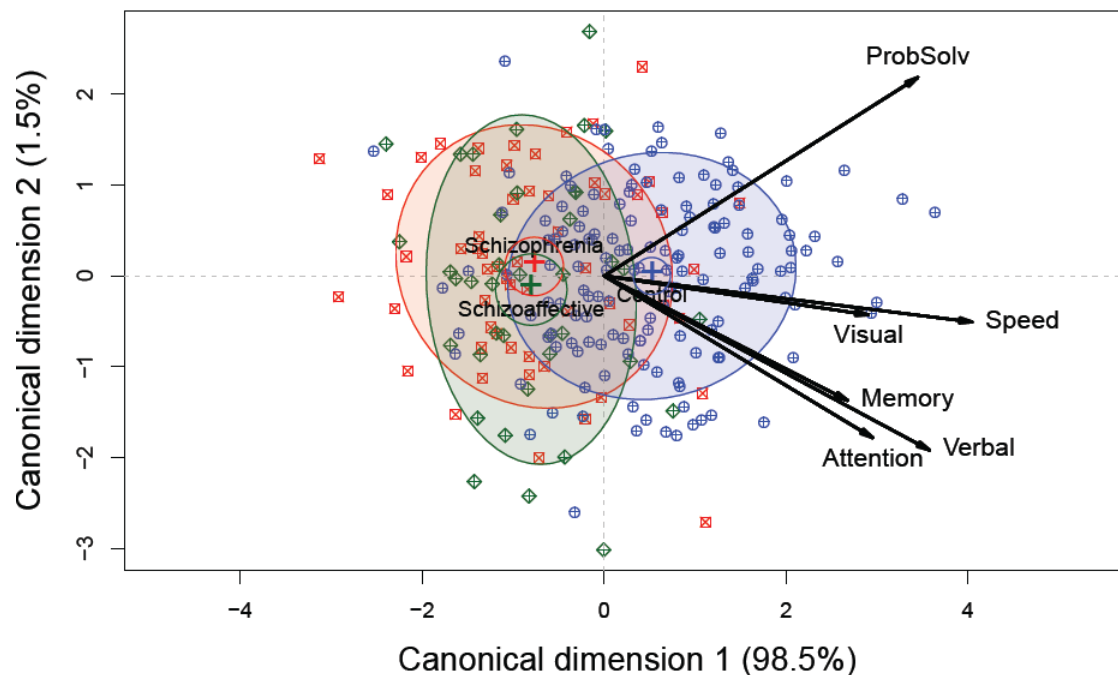


Wow! All neuro-cog measures highly correlated in group means!

Only 1 dim. of **H** variation

Visualize me: in canonical space

- As with biplot, we can visualize MLM hypothesis variation for *all* responses by projecting \mathbf{H} and \mathbf{E} into low-rank space.
- Canonical projection:** $\mathbf{Y}_{n \times p} \mapsto \mathbf{Z}_{n \times s} = \mathbf{Y}\mathbf{E}^{-1/2}\mathbf{V}$, where \mathbf{V} = eigenvectors of $\mathbf{H}\mathbf{E}^{-1}$.
- This is the view that maximally discriminates among groups, ie max. \mathbf{H} wrt \mathbf{E} !



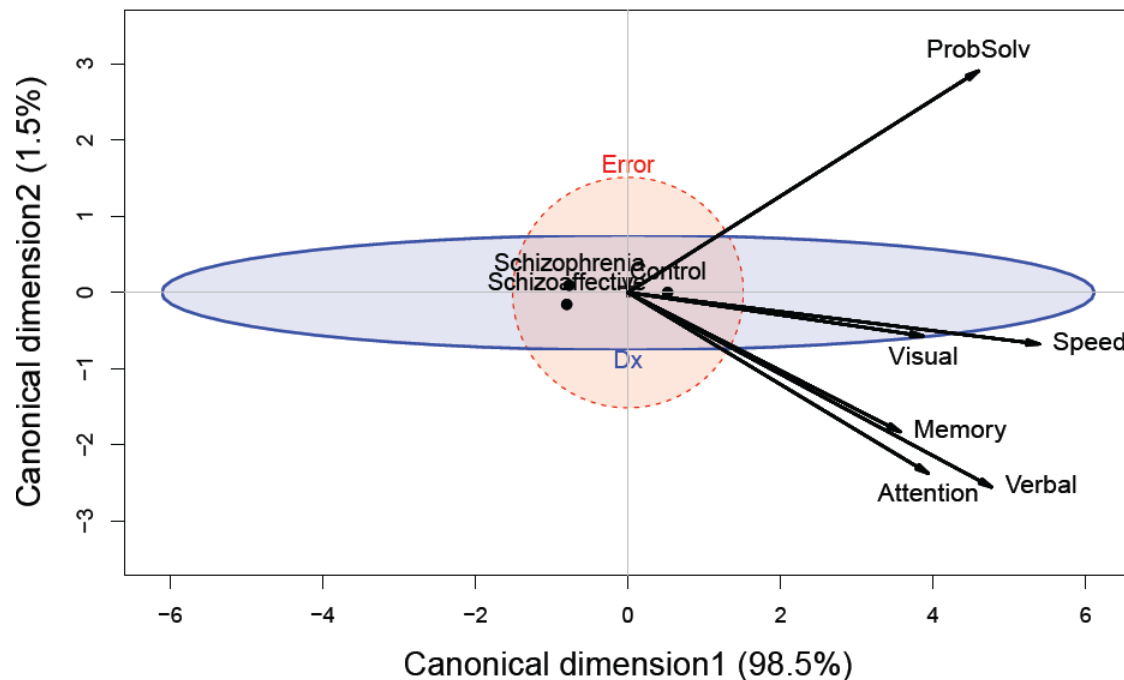
Very simple interpretation

Can1: normal vs. others
All vars highly + correlated;

Can2: only 1.5%, NS; but
perhaps suggestive (ProbSolv
vs. Attention)

Visualize me: canonical HE plots

- Canonical HE plot is just the HE plot of canonical scores, $(\mathbf{z}_1, \mathbf{z}_2)$ in 2D,
- or, $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$, in 3D.
- As in biplot, we add vectors to show relations of the \mathbf{y}_j response variables to the canonical variates.
- variable vectors here are **structure coefficients** = correlations of variables with canonical scores.



The multivariate “juicer”

Shows just group means, H ellipse & E ellipse

Variable vectors offer interpretation of Can dimensions.

Social cognitive measures

- These measures deal with the person's perception and cognitive processing of emotions of others
 - Scales: managing emotions, theory of mind, externalizing bias, personalizing bias
- Questions:
 - Do these differentiate normal from patient groups?
 - Can they distinguish between schizophrenic & schizoaffective
 - If so, this could be a major finding.

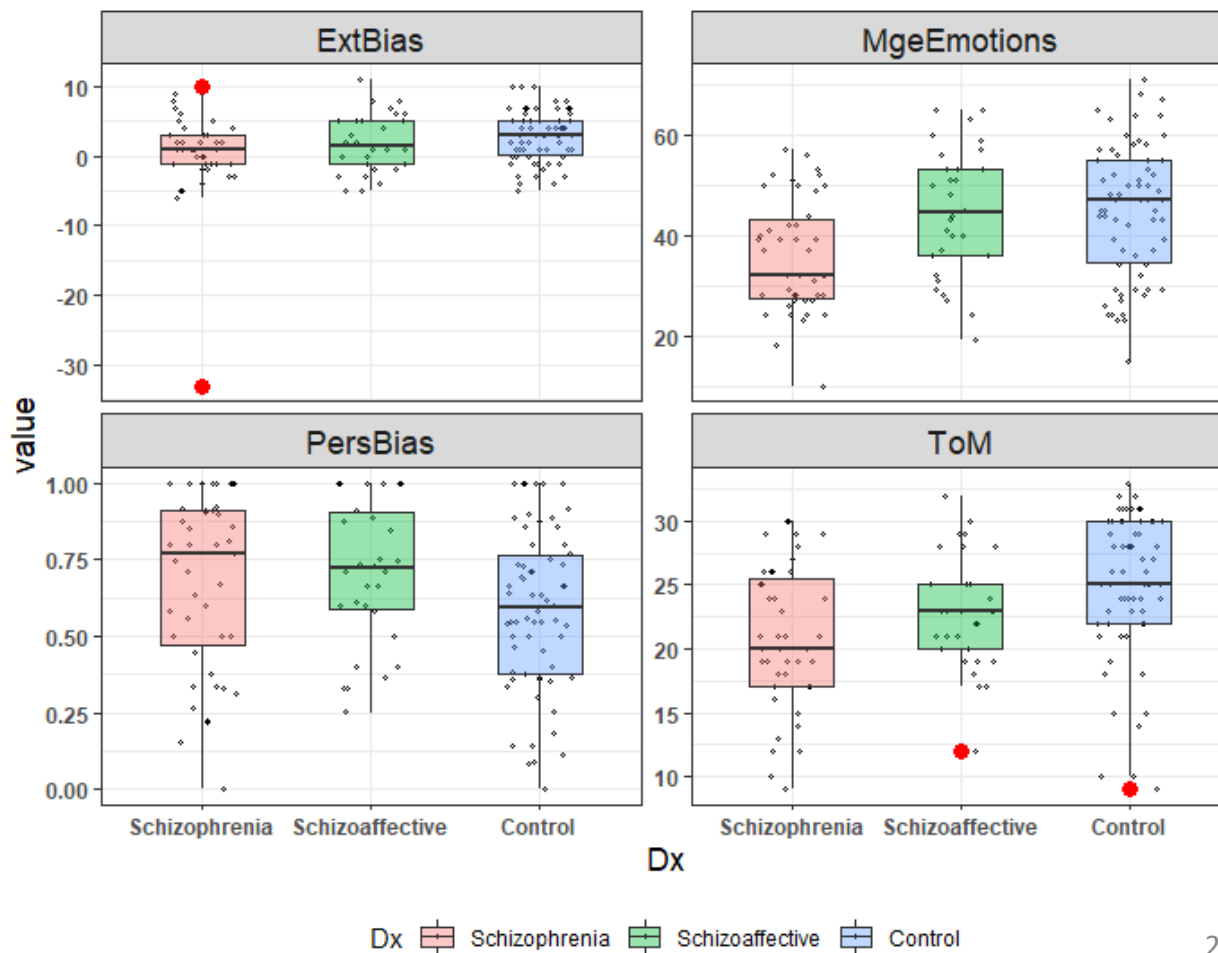
SocialCog: EDA

Again, it is helpful to plot the data!

Same steps: (a) Reshape to long; (b) `ggplot() + ... + facet_wrap(~response)`

What can we see here?

How do groups differ?
Unusual features?



Social cognitive measures

Data

```
> car::some(SocialCog)
```

	Dx	MgeEmotions	ToM	ExtBias	PersBias
24	Schizophrenia	32	18	-2	1.00
58	Schizophrenia	32	17	4	0.92
69	Schizophrenia	52	23	2	0.80
85	Schizoaffective	28	12	-5	1.00
99	Schizoaffective	40	22	1	0.36
111	Schizoaffective	51	23	6	0.91
124	Control	51	24	-5	1.00
155	Control	57	28	1	0.30
168	Control	47	32	0	0.77
216	Control	64	31	3	0.69

diagnostic group ----- social-cognitive measures -----

Analyses

Standard MANOVA test:

```
Anova(SC.mlm = lm(cbind(...) ~ DX, data=SocialCog)
```

Test contrasts:

```
linearHypotheses(SC.mlm, "Dx1")
```

```
linearHypotheses(SC.mlm, "Dx2")
```

```
> contrasts(      Soci al Cog$Dx)
      [, 1] [, 2]
Schizophrenia -0.5  1
Schizoaffective -0.5 -1
Control        1.0  0
```

Social cognitive measures

```
> data( Social Cog, package="heplots")
> SC.mlm <- lm(cbind(MgeEmotions, ToM Ext Bias, Pers Bias) ~ Dx,
               data=Social Cog)
> Anova(SC.mlm)
```

Type II MANOVA Tests: Pillai test statistic

	Df test	stat	approx F	num Df	den Df	Pr(>F)
Dx	2	0.212	3.97	8	268	0.00018 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Test contrasts: Dx1 = Normal vs. Patient; Dx2 = Schizo vs. Schizoaffective

```
> print(linearHypothesis(SC.mlm, "Dx1"), SSP=FALSE)
```

Multivariate Tests:

	Df test	stat	approx F	num Df	den Df	Pr(>F)
Pillai	1	0.1355	5.212	4	133	0.000624 ***



```
> print(linearHypothesis(SC.mlm, "Dx2"), SSP=FALSE)
```

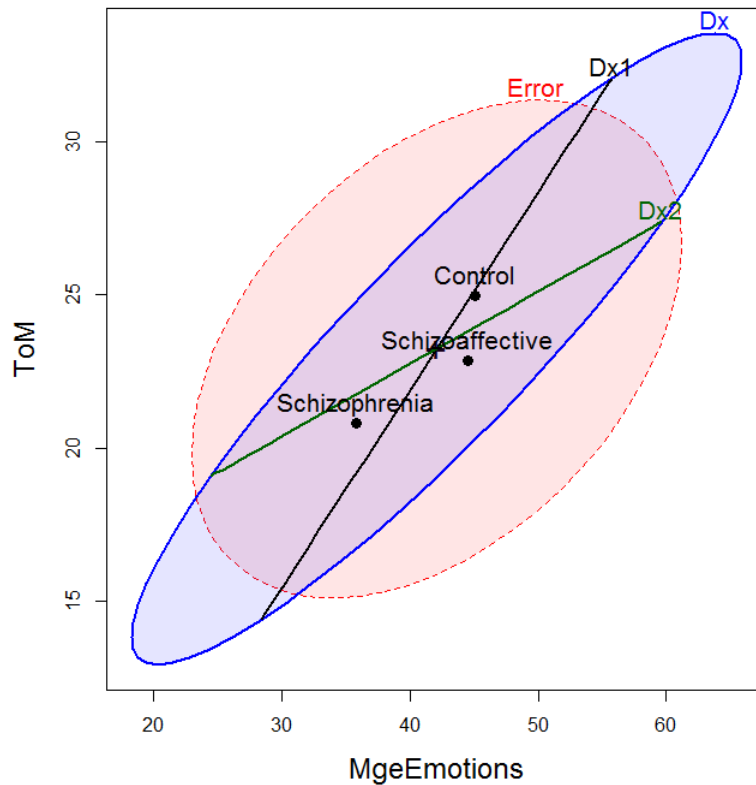
Multivariate Tests:

	Df test	stat	approx F	num Df	den Df	Pr(>F)
Pillai	1	0.0697	2.493	4	133	0.0461 *

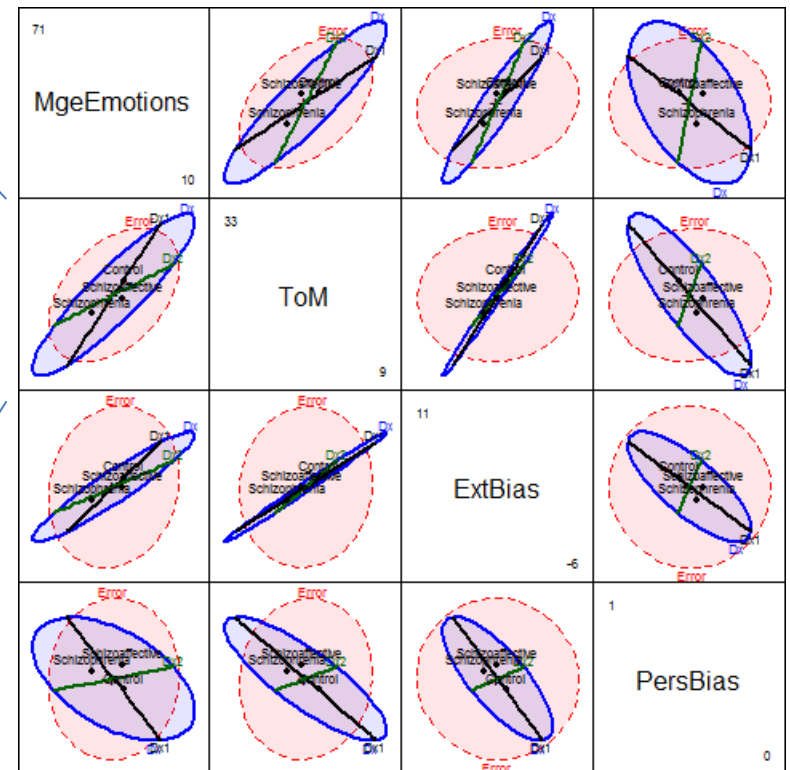


Visualize me: data space

```
heplot (SC.mlm,
hypot heses=list("Dx1", "Dx2"),...)
```



```
pairs(SC.mlm,
hypot heses=list("Dx1", "Dx2"),...)
```



The groups are ordered Schizo < ScAffective < Control on these measures

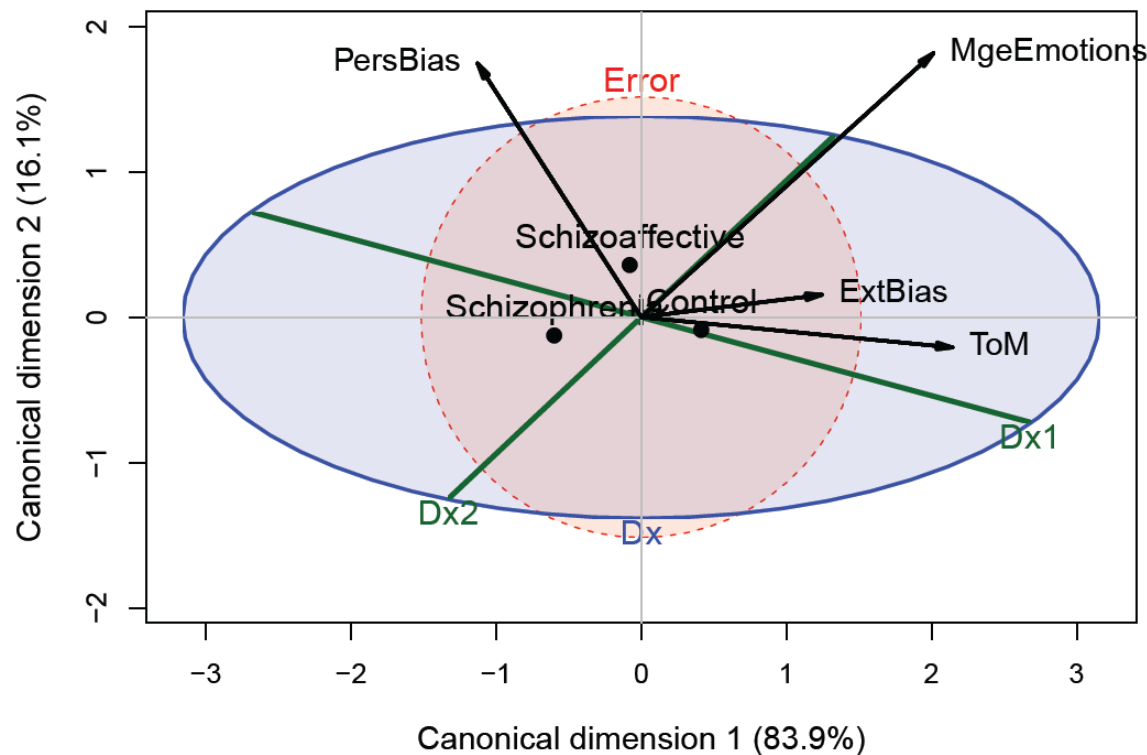
One of these vars is not like the others!

Visualize me: canonical space

Contrasts:

Dx1 : Control vs. patients. Controls > patients on MgeEmotions, ExtBias, ToM

Dx2 : Schizo vs. schizoaffective.



Can1: group order

Can2: Schizoaffective
vs. others

Dx1 suggests \approx spacing
Schizo < ScAff <
Control

Model checking & remedies

- The MLM assumes residuals are multivariate normal

- → Squared Mahalanobis distances

$$D^2_M(\mathbf{y}_i) = (\mathbf{y}_i - \bar{\mathbf{y}})^T \mathbf{S}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}) \sim \chi^2_p \text{ with } p \text{ d.f.}$$

- → a quantile – quantile plot of ordered $D^2_M(\mathbf{y}_i)$ vs. quantiles of χ^2_p should plot as **straight line**

- Outliers are readily apparent

- plots: `heplots::cqplot()`

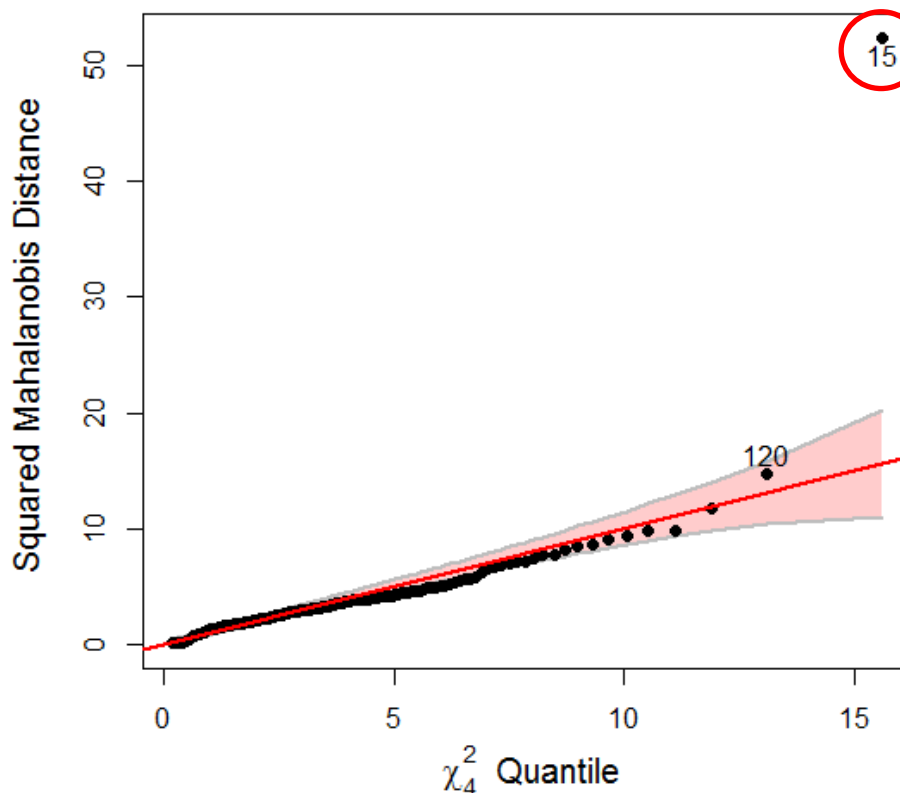
- Influence plots

- `mvinfluence::influence()` calculates multivariate analogs of influence measures

Social cog: cqplot

`heplots::cqplot()` creates a chi-square QQ plot from a MLM

```
cqplot(SC.mlm, id.n=2)
```



One observation appears as an extreme outlier.

This was a case w/ ExtBias = -33, but valid range = (-10, +10)

Refitting w/o case 15:

Overall & DX1 tests still OK 😊

Dx2 test: $p=0.074$, now NS 😞

Social cog: Influence

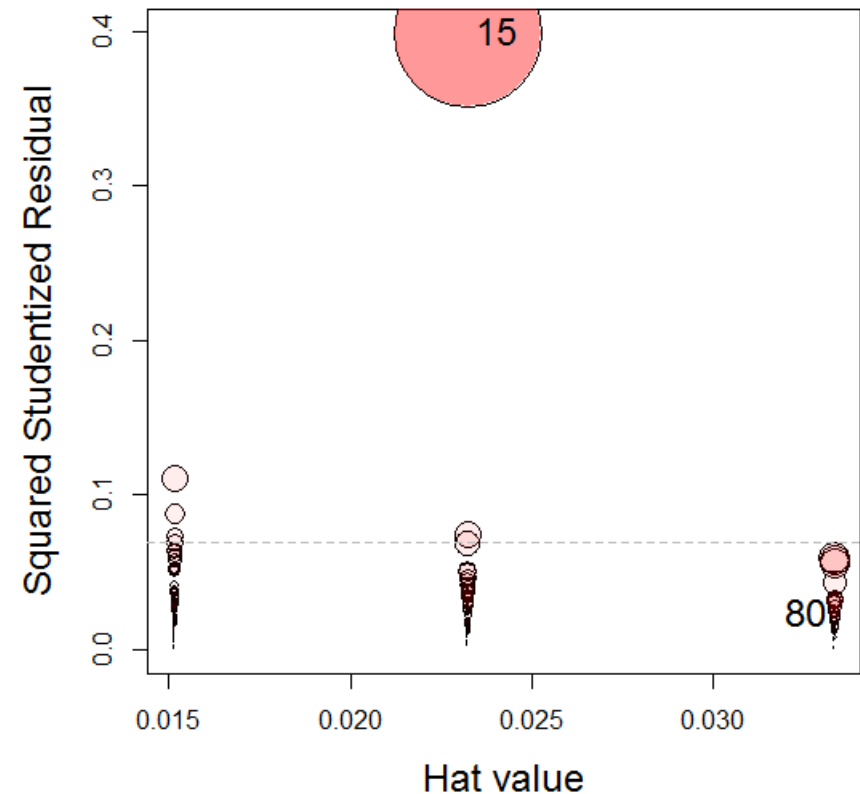
`mvinfluence::influencePlot()` creates a multivariate analog of an influence plot

```
>library(mvinfluence)
>influencePlot(SC.mlm, ...)
```

	H	Q	CookD	L	R
15	0.0233	0.3985	0.4202	0.0238	0.4080
80	0.0333	0.0233	0.0352	0.0345	0.0241

Case 15 stands out as hugely influential (a data error)

The 3 columns of circles correspond to the 3 groups



Robust MLMs

- Robust methods for univariate LMs are now well-developed and implemented
 - → proper SEs, CIs and hypothesis tests
- Analogous methods for multivariate LMs are a current hot research topic
- The heplots package now provides `robmlm()` for the fully general MLM (MANOVA, MMReg)
 - Uses simple M-estimator via IRLS
 - Weights: calculated from Mahalanobis D^2 , a robust covariance estimator and weight function, $\psi(D^2)$

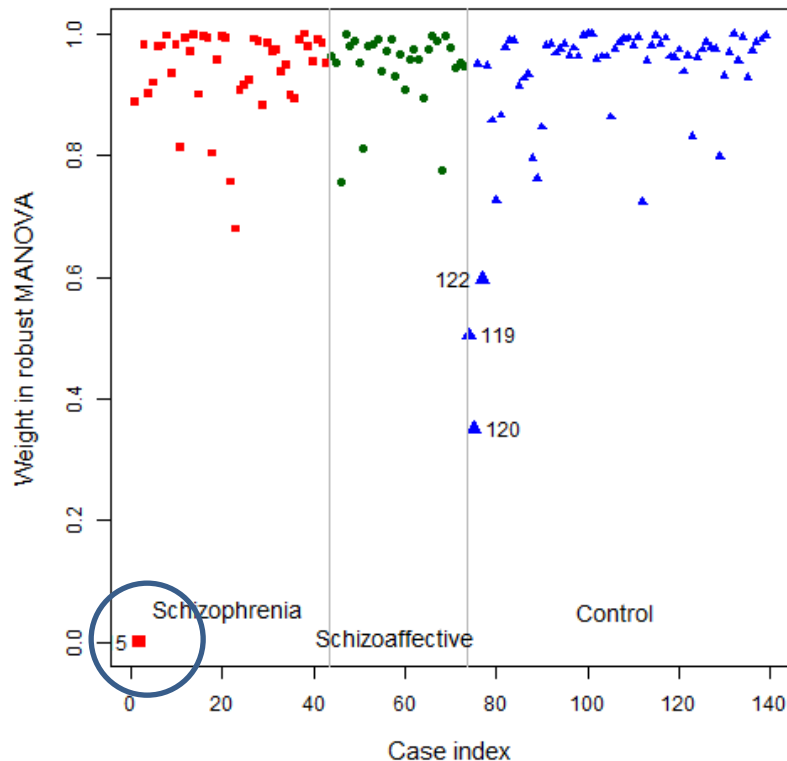
$$D^2 = (\mathbf{Y} - \bar{\mathbf{Y}})^T \mathbf{S}_{\text{robust}}^{-1} (\mathbf{Y} - \bar{\mathbf{Y}}) \sim \chi_p^2$$

- Downside: SEs, p -values only approximate

Robust MLMs

```
SC.rlm <- robmlm(cbind( MgeEmotions, ToM, ExtBias, PersBias) ~ Dx,  
data=SocialCog)
```

Observation weights



Approx test of Dx2 in robust model

```
> print(linearHypothesis(SC.rlm, "Dx2"), SSP=FALSE)
```

Multivariate Tests:

	Df	test stat	approx F	num Df	den Df	Pr(>F)
Pillai	1	0.069	2.44	4	132	0.05 *
wilks	1	0.931	2.44	4	132	0.05 *
...						

Robust MLMs: Pottery data

- Chemical composition of ancient pottery found at four sites in Great Britain
 - Sites: AshleyRails, Caldicot, IsleThorns, Llanedyrn
 - Measures: Al Fe Mg Ca Na
- Analyses:
 - standard MANOVA `lm(cbind(Al, Fe, Mg, ...) ~ Site, data=Pottery)`
 - robust MANOVA `robmlm(cbind(Al, Fe, Mg, ...) ~ Site, data=Pottery)`

- Data:

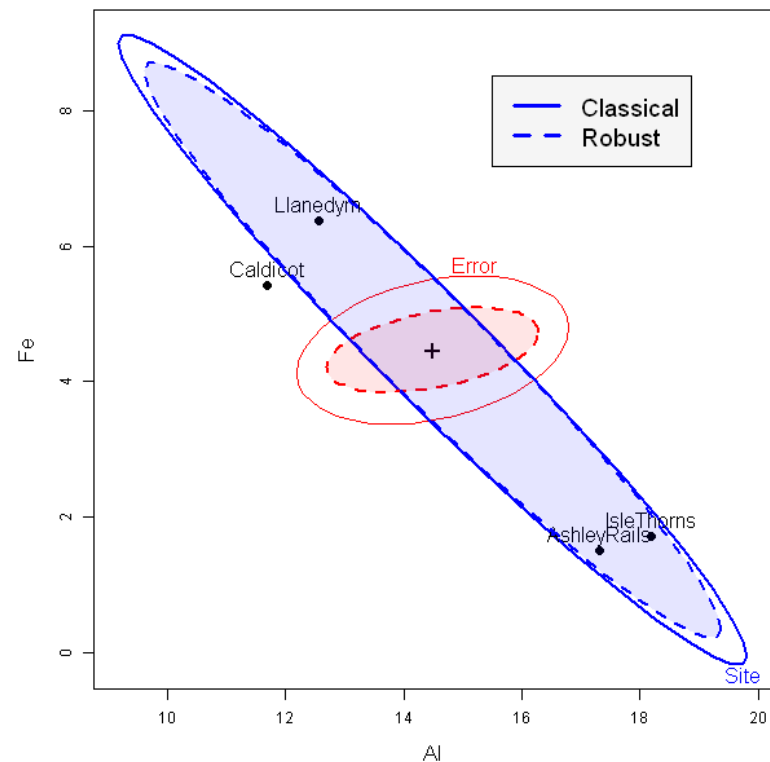
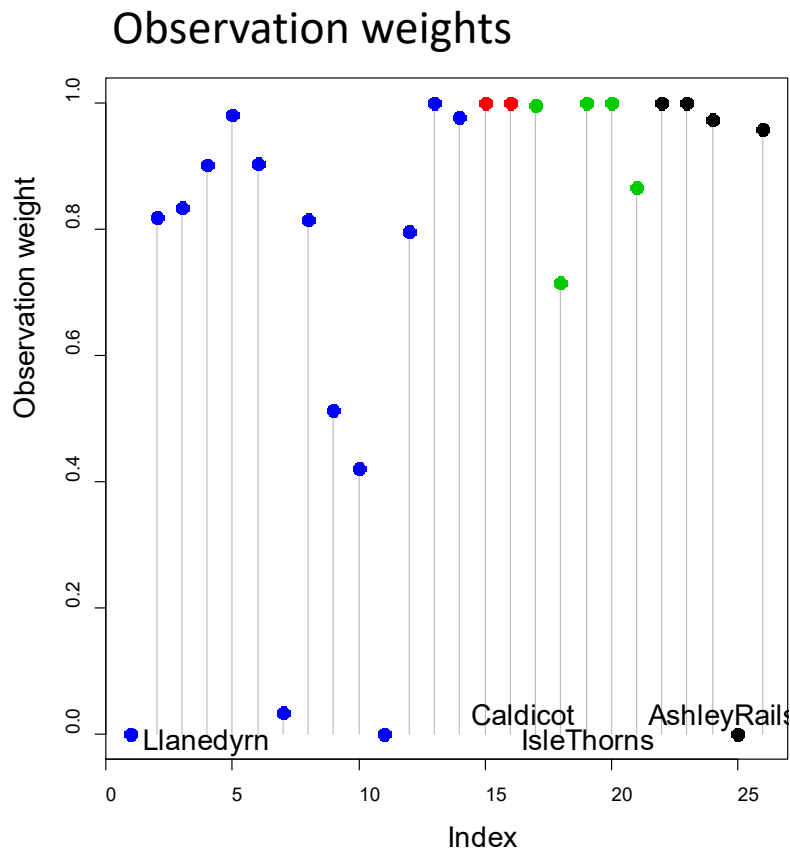
```
> car::some(Pottery)
```

	Site	Al	Fe	Mg	Ca	Na
2	Llanedyrn	13.8	7.08	3.43	0.12	0.17
14	Llanedyrn	12.5	6.44	3.94	0.22	0.23
17	IsleThorns	18.3	1.28	0.67	0.03	0.03
18	IsleThorns	15.8	2.39	0.63	0.01	0.04
21	IsleThorns	20.8	1.51	0.72	0.07	0.10
22	AshleyRails	17.7	1.12	0.56	0.06	0.06
23	AshleyRails	18.3	1.14	0.67	0.06	0.05

Robust MLMs: Pottery data

```
> pottery.mod <- lm( cbind ( Al,Fe,Mg,Ca,Na )~Site, data=Pottery)
> pottery.rmod <- robmlm ( cbind ( Al,Fe,Mg,Ca,Na )~Site, data=Pottery)
```

residual E ellipse shrinks a lot
overlaid HE plots



MMRA example: PA tasks & ability

- Rohwer data from Timm (1975)
- How well do paired associate (PA) tasks predict performance on measures of aptitude & achievement in kindergarten children?
 - Samples: 69 children in two groups (schools): 'Lo' | 'Hi' SES
 - Outcomes (Y):
 - Scholastic aptitude test (SAT)
 - Peabody picture vocabulary test (PPVT)
 - Raven progressive matrices (Raven)
 - Predictors (X): Scores (0—40) on PA tasks where the stimuli were:
 - named (n), still (s), named-still (ns), named-action (na), sentence-still (ss)

```
> data("Rohwer", package="heplots")
> car::some(Rohwer, n=5)
  group SES SAT PPVT Raven n  s ns na ss
8      1  Lo   8   68    8  0  0 10 19 14
9      1  Lo  49   74   11  0  0  7 16 13
17     1  Lo  19   66   13  7 12 21 35 27
52     2  Hi  38   66   14  0  0  3 16 11
66     2  Hi   8   55   16  4  7 19 20 13
```

```
> Rohwer2 <- subset(Rohwer, subset=SES=="HI")
```

Having a group factor makes the analysis more complicated (MANCOVA)

Start with analysis of the Hi SES group

Why not univariate models?

```
rohwer.mod1 <- lm(SAT ~ n + s + ns + na + ss, data = Rohwer2)
rohwer.mod2 <- lm(PPVT ~ n + s + ns + na + ss, data = Rohwer2)
rohwer.mod3 <- lm(Raven ~ n + s + ns + na + ss, data = Rohwer2)

library(stargazer)
stargazer(rohwer.mod1, rohwer.mod2, rohwer.mod3, type="text", ...)
```

Univariate regression models for Rohwer data

Dependent variable:			
	SAT	PPVT	Raven
n	3.26*	0.07	0.06
s	3.00	0.37	0.49**
ns	-5.86***	-0.37	-0.16
na	5.67***	1.52*	0.12
ss	-0.62	0.41	-0.12
R ²	0.56	0.35	0.31
F Statistic (df = 5; 26)	6.54***	2.85*	2.32
Note: *p<0.05; **p<0.01; ***p<0.001			

Results are disappointing

- Only model for SAT highly signif.
- Only a few coefs. signif. $\neq 0$

MANOVA tests

```
> rohwer.mlm <- lm(cbind(SAT, PPVT, Raven) ~ n + s + ns + na + ss, data=Rohwer2)
> Anova(rohwer.mlm)
```

Type II MANOVA Tests: Pillai test statistic

	Df	test	stat	approx	F	num	Df	den	Df	Pr(>F)
n	1	0.202	2.02	3	24	0.1376				
s	1	0.310	3.59	3	24	0.0284 *				
ns	1	0.358	4.46	3	24	0.0126 *				
na	1	0.465	6.96	3	24	0.0016 **				
ss	1	0.089	0.78	3	24	0.5173				

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Much better!

Multivariate tests:

- pool evidence for all Ys
- take correlations of Ys into account

Can also test overall hypothesis, $H_0: \mathbf{B} = \mathbf{0}$ (all coefs = 0)

```
> print(linearHypothesis(rohwer.mlm,
+                         c("n", "s", "ns", "na", "ss")), SSP=FALSE)
```

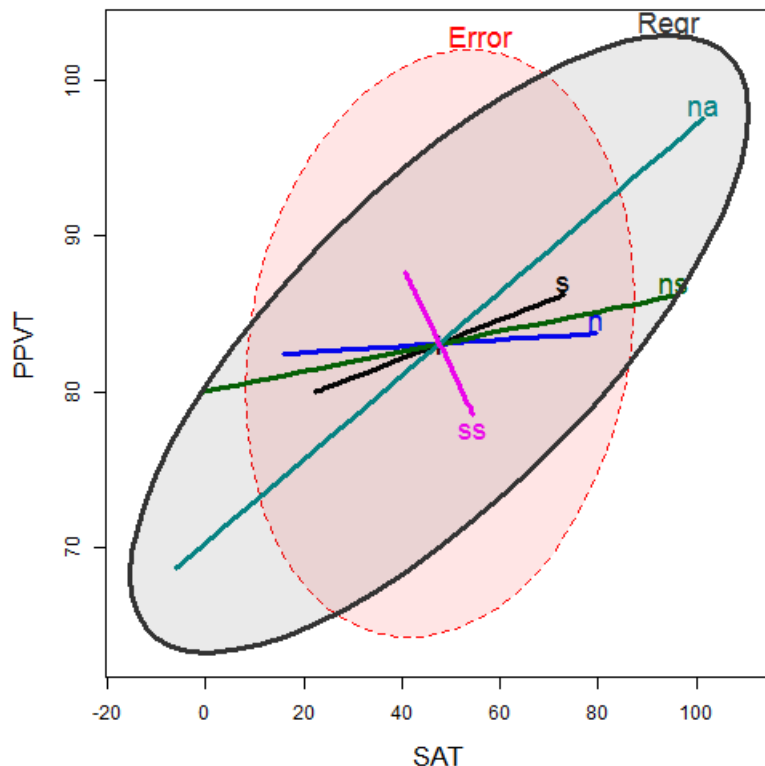
Multivariate Tests:

	Df	test	stat	approx	F	num	Df	den	Df	Pr(>F)
Pillai	5	1.0386	2.753	15	78.00	0.001912	**			
Wilks	5	0.2431	2.974	15	66.65	0.001154	**			
Hotelling-Lawley	5	2.0615	3.115	15	68.00	0.000697	***			
Roy	5	1.4654	7.620	5	26.00	0.000160	***			

Strongly reject H_0
by all criteria

Visualize me!

```
cols <- c("red", "blue", "black", "darkgreen", "darkcyan", "magenta", "gray20")
hyp <- list("Regr" = c("n", "s", "ns", "na", "ss")) # Test of B = 0
heplot(rohwer.mlm,
       hypotheses = hyp,
       fill=TRUE, fill.alpha=0.1, col=cols, lwd=c(1,3))
```



Each predictor gives a 1 df test $\rightarrow H_i$ ellipse is a line

E here is a 3D ellipsoid ($\text{rank}(E) = \min(p, q)$)

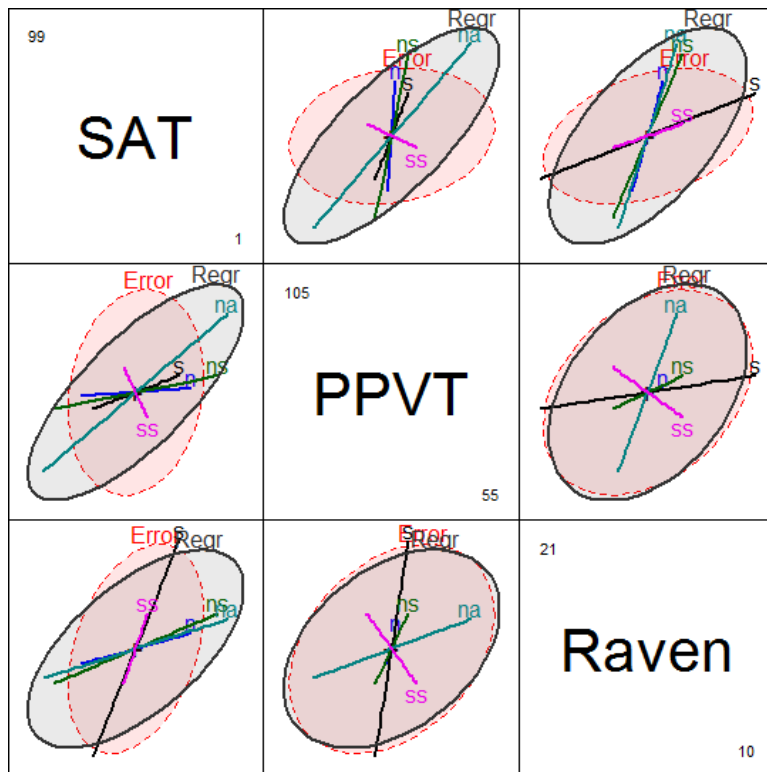
Interpretation:

- Any H ellipse that protrudes outside E ellipse is significant by Roy's test
- Length of each H_i line \approx strength of association
- Orientation of each H_i line shows relation of X_i to the two Y_j responses shown.

pairs.mlm() plot

View all pairwise HE plots

```
pairs(rohwer.mlm, hypotheses=hyp,  
      col=cols, fill=TRUE, fill.alpha=0.1, ...)
```



Can now understand more subtle aspects

SAT is best predicted overall, but relation with PA tests varies

The na & ns tasks are strongest for SAT

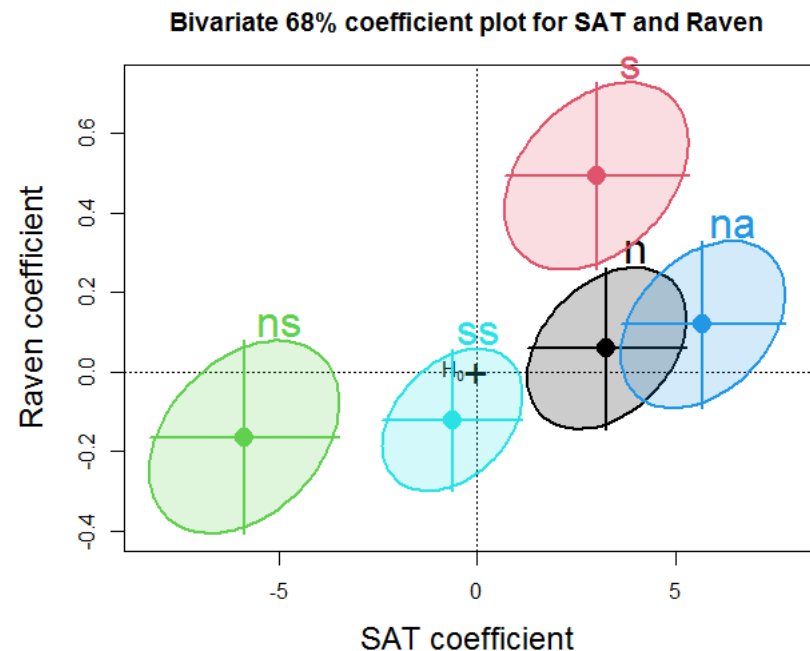
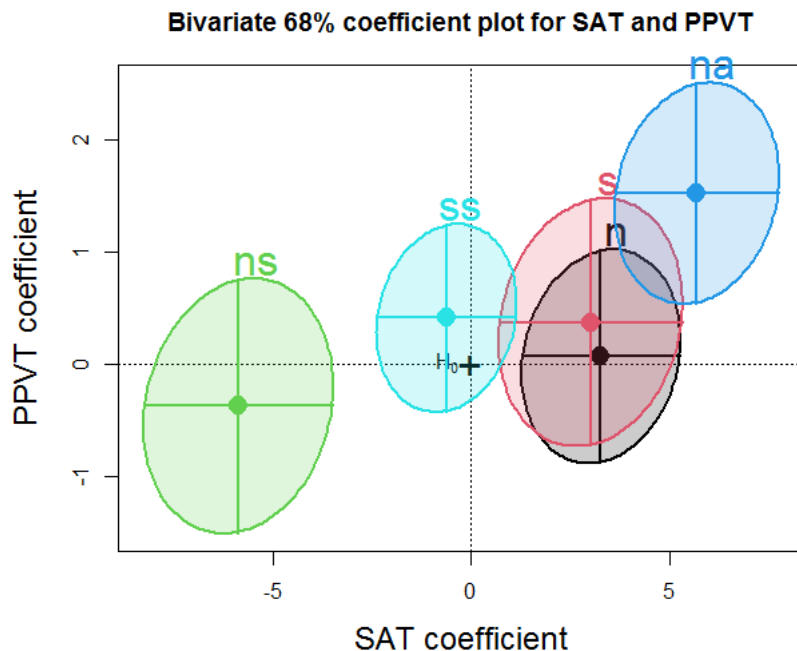
Raven is weakly predicted

Coefficient plots for MLMs

In multivariate models, confidence intervals for coefficients become ellipsoids
Two coefficients are **jointly significant** if their conf. ellipse $\mathbb{E}(b_1, b_2)$ excludes (0,0)
These plots are provided by `heplots::coefplot(mlm)`

```
coefplot(rohwer.mlm, var=c(1,2))
```

```
coefplot(rohwer.mlm, var=c(1,3))
```

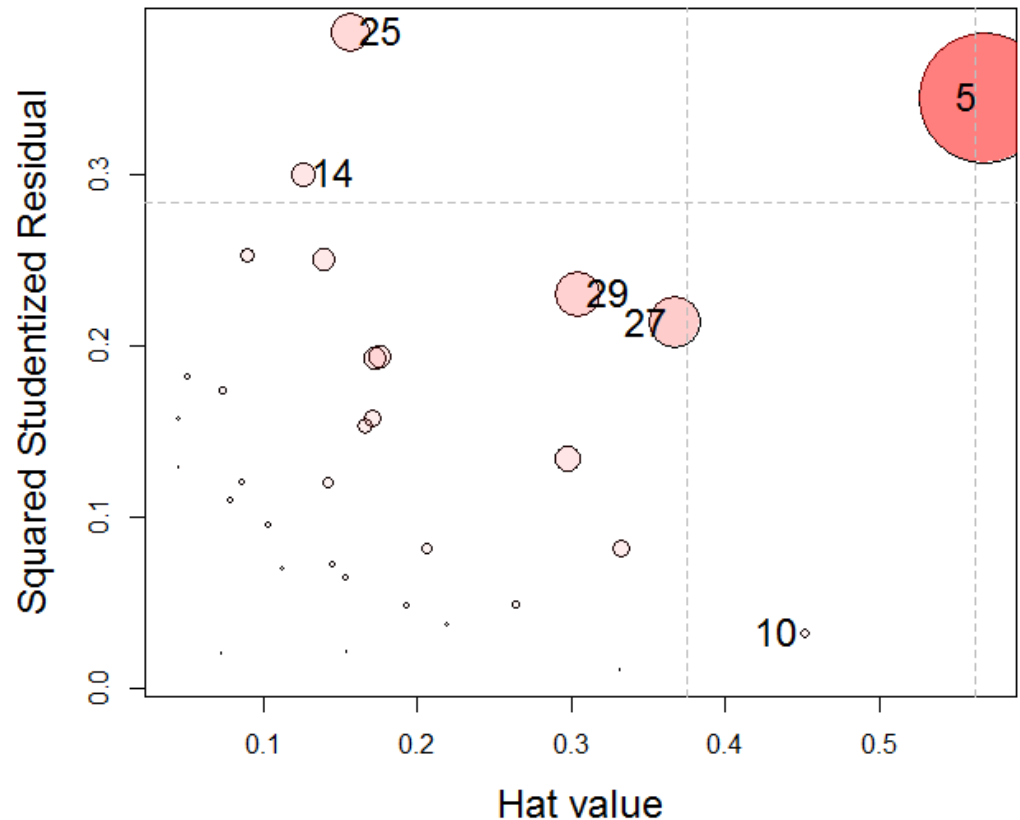


Influence plots

As always, we should check for influential observations and multivariate outliers

```
mvinfluence::influencePlot(Rohwer.mlm, id.n=4, ...)
```

One point (5) deserves further attention



Canonical correlations

For quantitative (X, Y) data, canonical correlation analysis is an alternative to MMRA
It finds the weighted sums of the Y variables most highly correlated with the Xs

```
> X <- Rohwer2[, 6:10] # X variables for High SES students
> Y <- Rohwer2[, 3:5]  # Y variables for High SES students
> (cc <- cancor(X, Y, set.names=c("PA", "Ability")))
```

Canonical correlation analysis of:

5 PA variables: n, s, ns, na, ss
with 3 Ability variables: SAT, PPVT, Raven

	CanR	CanRSQ	Eigen	percent	cum	scree
1	0.7710	0.5944	1.4654	71.080	71.08	*****
2	0.5465	0.2987	0.4259	20.659	91.74	*****
3	0.3815	0.1455	0.1703	8.261	100.00	***

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

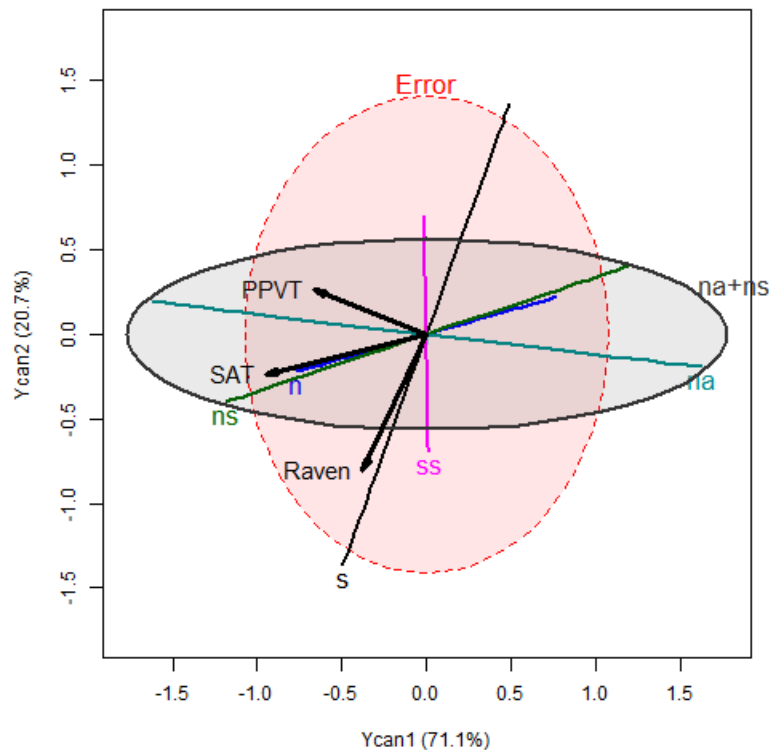
	CanR	LR test stat	approx F	numDF	denDF	Pr(> F)	
1	0.77096	0.24307	2.9738	15	66.655	0.001154	** ✓
2	0.54652	0.59926	1.8237	8	50.000	0.094538	. ✗
3	0.38147	0.85448	1.4759	3	26.000	0.244178	

Two dimensions
acct for 91.7% of
(X,Y) association

Only Can1 is
significant

Visualize CCA in HE plot

```
cols <- c("red", "blue", "black", "darkgreen", "darkcyan",  
          "magenta", "gray20")  
heplot(cc, hypotheses=list("na+ns"=c("na", "ns")),  
       fill = TRUE, fill.alpha=0.1, col=cols,  
       label.pos = c(3, rep(1,5), .1),  
       cex=1.4, var.cex=1.25, var.lwd=3, var.col="black")
```



Residuals are uncorrelated in canonical space

H ellipses for X terms same as in ordinary HE plots – outside **E** ellipse *iff* signif. by Roy's test

Variable vectors for Ys: correlations with canonical variables Ycan1, Ycan2

- SAT & PPVT: mainly Ycan1
- Raven: more aligned with Ycan2

MANCOVA & homogeneity of regression

- With a group variable (SES) can test differences in means (intercepts)
 - `rohwer.mod <- lm(cbind(SAT, PPVT, Raven) ~ SES + n + s + ns + na + ss, data=Rohwer)`
 - This assumes that slopes (B) are the **same** for both groups (homogeneity of regression)
- Can test for equal slopes by adding interactions of SES with Xs
 - `rohwer.mod1 <- lm(cbind(SAT, PPVT, Raven) ~ SES * (n + s + ns + na + ss))`
- Or, fit separate models for each group

```
rohwer.ses1 <- lm(cbind(SAT, PPVT, Raven) ~ n + s + ns + na + ss,  
                  data = Rohwer, subset = SES == "Hi")  
rohwer.ses2 <- lm(cbind(SAT, PPVT, Raven) ~ n + s + ns + na + ss,  
                  data = Rohwer, subset = SES == "Lo")
```

MANCOVA

Fit the MANCOVA model & test hypotheses

```
> rohwer.mod <- lm(cbind(SAT, PPVT, Raven) ~ SES + n + s + ns + na + ss,  
+                   data=Rohwer)  
> Anova(rohwer.mod)
```

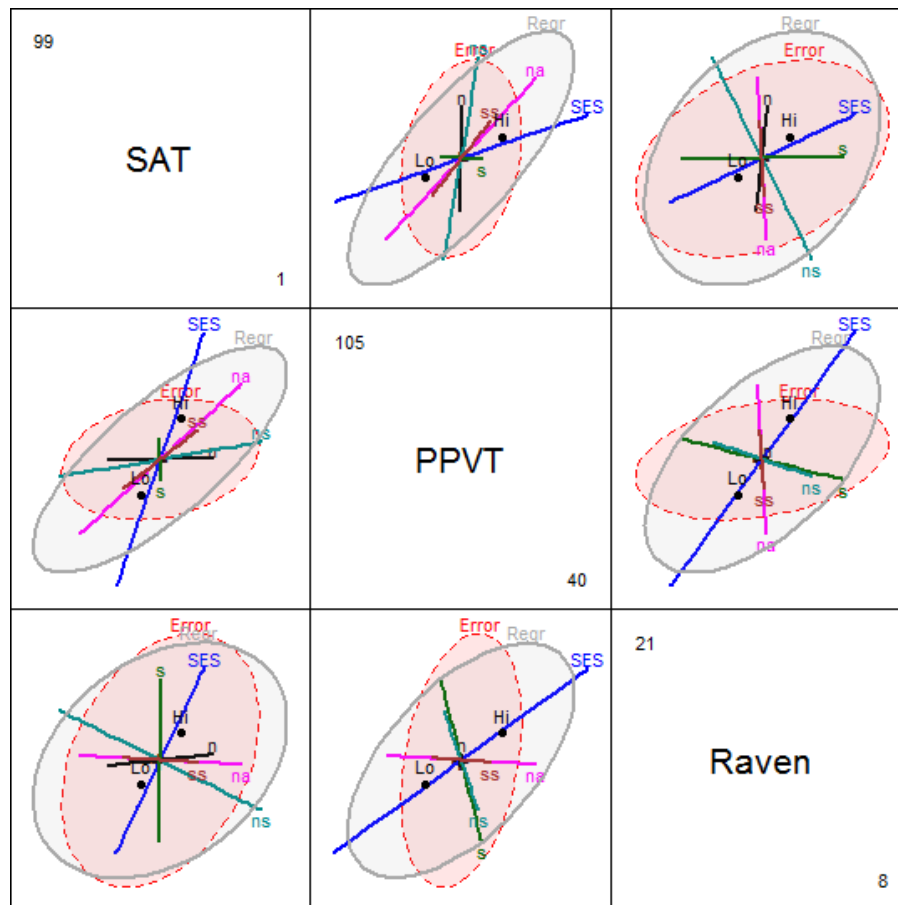
Type II MANOVA Tests: Pillai test statistic

	Df	test	stat	approx	F	num	Df	den	Df	Pr(>F)	
SES	1	0.379	12.18	3	60	2.5e-06	***				SES: diff in means
n	1	0.040	0.84	3	60	0.4773					
s	1	0.093	2.04	3	60	0.1173					
ns	1	0.193	4.78	3	60	0.0047	**				
na	1	0.231	6.02	3	60	0.0012	**				
ss	1	0.050	1.05	3	60	0.3770					

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Visualize effects

```
pairs(rohwer.mod,
      hypotheses=list("Regr" = c("n", "s", "ns", "na", "ss")),
      fill=TRUE, fill.alpha=0.1)
```



The SES effect is positive for all Y variables
Hi SES group > Lo SES group

Fit model with interactions

Fit heterogeneous regression model with SES interactions

```
> rohwer.mod1 <- lm(cbind(SAT, PPVT, Raven) ~ SES * (n + s + ns + na + ss),  
+ data=Rohwer)  
> Anova(rohwer.mod1)
```

Type II MANOVA Tests: Pillai test statistic

	Df	test stat	approx F	num Df	den Df	Pr(>F)	
SES	1	0.391	11.78	3	55	4.5e-06	***
n	1	0.079	1.57	3	55	0.20638	
s	1	0.125	2.62	3	55	0.05952	.
ns	1	0.254	6.25	3	55	0.00100	***
na	1	0.307	8.11	3	55	0.00015	***
ss	1	0.060	1.17	3	55	0.32813	
SES:n	1	0.072	1.43	3	55	0.24417	
SES:s	1	0.099	2.02	3	55	0.12117	
SES:ns	1	0.118	2.44	3	55	0.07383	.
SES:na	1	0.148	3.18	3	55	0.03081	*
SES:ss	1	0.057	1.12	3	55	0.35094	

OK, as expected

Hmm ???

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Test interactions

Can test all interactions **simultaneously** with linearHypothesis()
Do I need any interaction terms?

I use a 'grep' trick here to find the names of coefficients like 'SES:' containing a ':'

```
> coefs <- rownames(coef(rohwer.mod1)) # store coefficient names in a vector
> print(linearHypothesis(rohwer.mod1, # only test for interaction effects
+                          coefs[grep(":", coefs)]), SSP=FALSE)
```

Multivariate Tests:

	Df	test stat	approx F	num Df	den Df	Pr(>F)	
Pillai	5	0.4179	1.845	15	171.0	0.03209	*
Wilks	5	0.6236	1.894	15	152.2	0.02769	*
Hotelling-Lawley	5	0.5387	1.927	15	161.0	0.02396	*
Roy	5	0.3846	4.385	5	57.0	0.00191	**

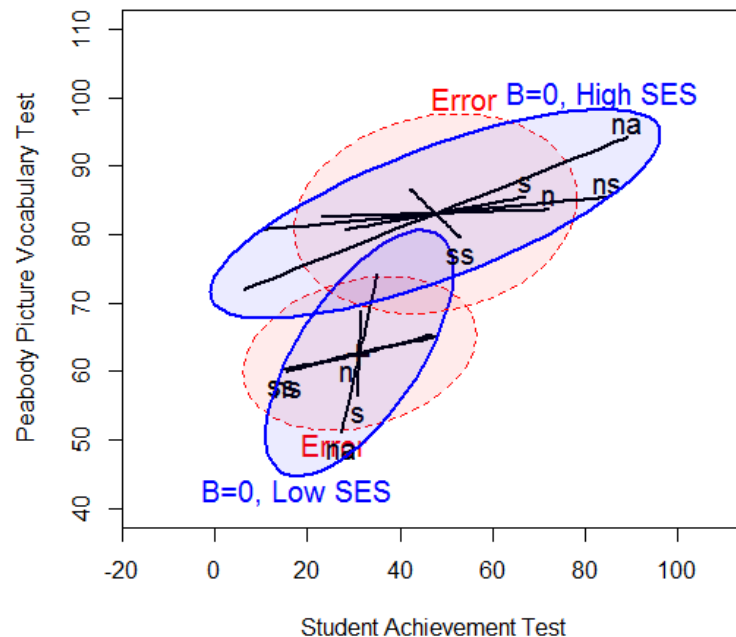
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Evidence shows that **some** slopes differ for Hi/Lo SES

Fit separate models

Fitting a model for each group allows all slopes to differ
Also allows within-group covariances to differ

```
rohwer.ses1 <- lm(cbind(SAT, PPVT, Raven) ~ n + s + ns + na + ss,  
                  data = Rohwer, subset = SES == "Hi")  
rohwer.ses2 <- lm(cbind(SAT, PPVT, Raven) ~ n + s + ns + na + ss,  
                  data = Rohwer, subset = SES == "Lo")
```



For SAT & PPVT:

- means higher for Hi SES
- within-group covariance larger for Hi SES
- slopes of predictors smaller for Hi SES
→ SAT more important for this group.

Homogeneity of (co)variances

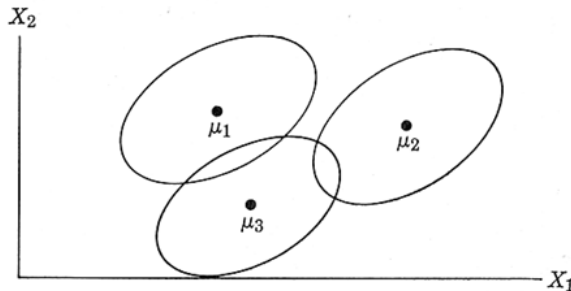
- ANOVA assumes equality of residual variances



$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_g^2$$

Levine's test: ANOVA of $z_{ij} = |y_{ij} - \bar{y}_i|$

- MANOVA: assumes equality of covariance matrices



$$\Sigma_1 = \Sigma_2 = \dots = \Sigma_g$$

Box's M test:

$$M = (N - g) \ln |\mathbf{S}_p| - \sum_{i=1}^g (n_i - 1) \ln |\mathbf{S}_i|$$

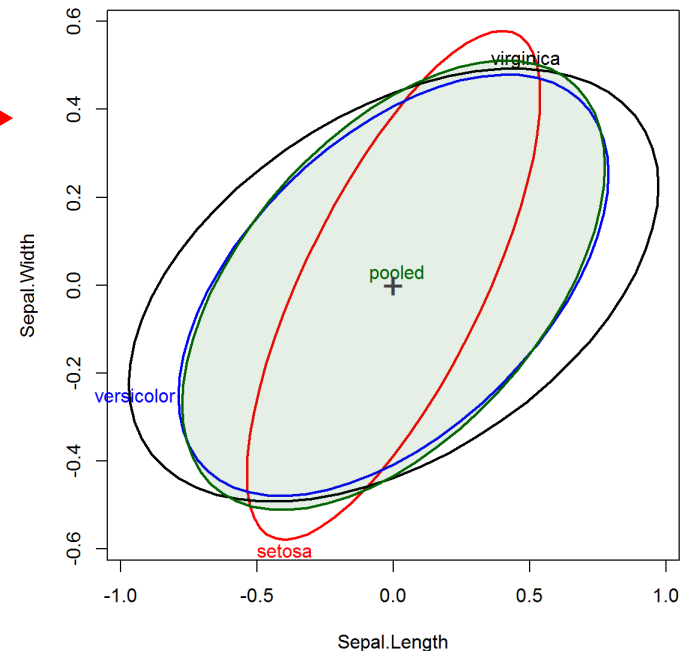
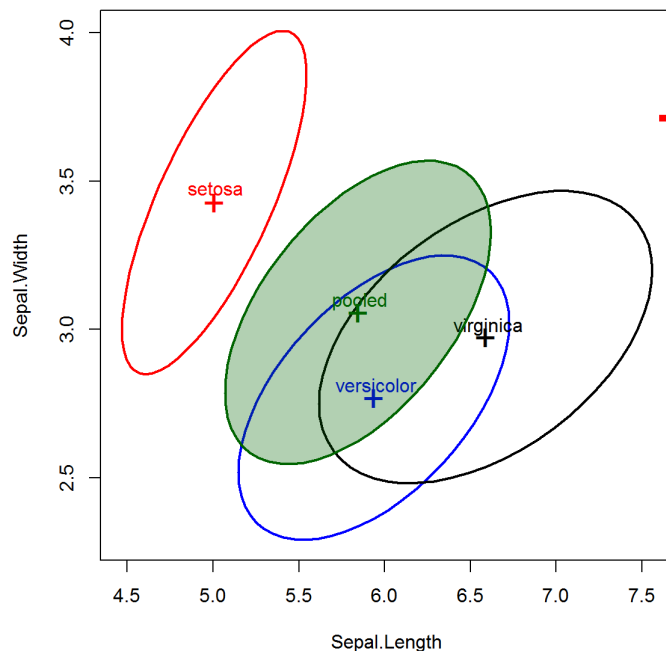
$$-2 \ln(M) \sim \chi^2_{df}$$

See: <http://www.datavis.ca/papers/EqCov-TAS.pdf>

Visualizing covariance matrices

Visualize covariance ellipses in data space
Center to see pure differences in size & shape

```
covEllipses(iris[,1:2], iris$Species, ...)  
covEllipses(iris[,1:2], iris$Species, center=TRUE, ...)
```

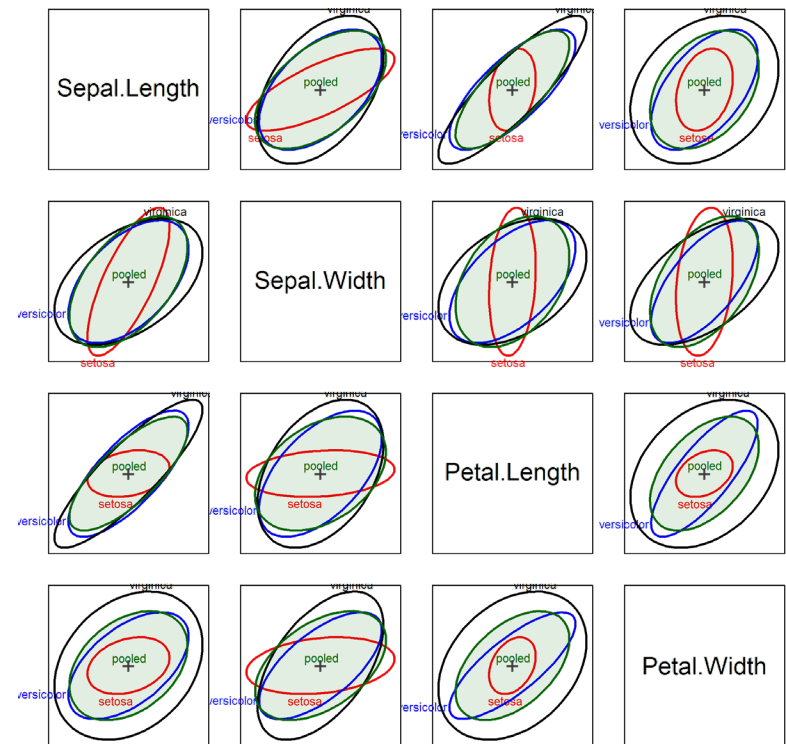


View all pairs

```
covEllipses(iris[,1:4], iris$Species,  
            center=TRUE, variables=1:4, ...)
```

In all cases, **setosa** stands out as different from the others

- sometimes correlation differs
- sometimes smaller variance(s)

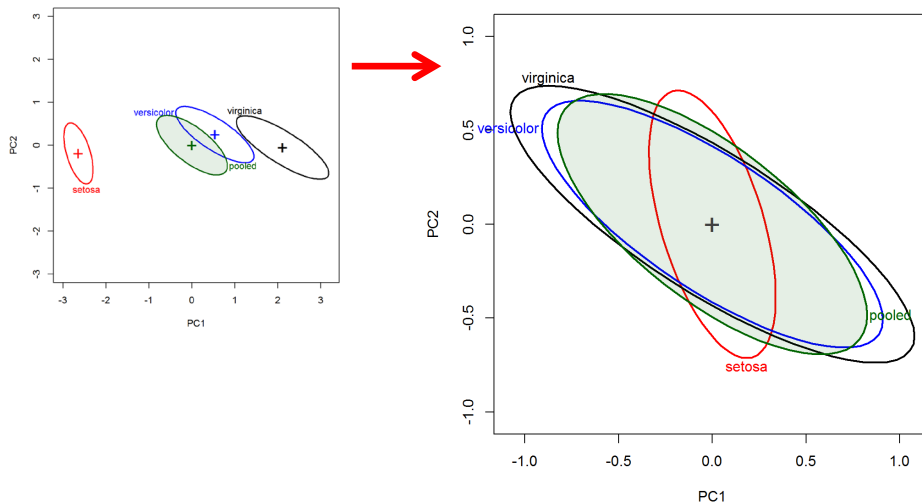


Visualize in PCA space

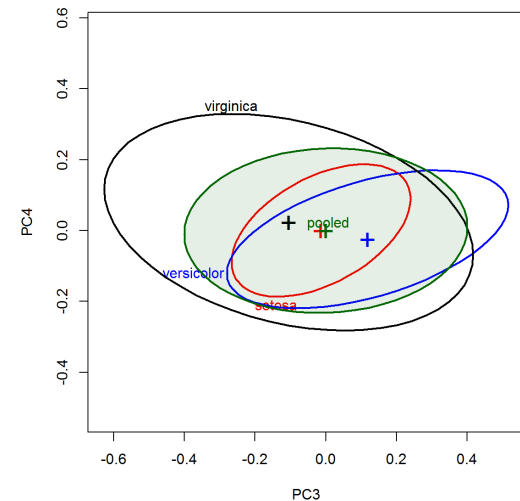
PCA projects the data into an orthogonal space accounting for maximum variance
Covariance ellipses show the differences among groups in this space

Surprisingly, the **small** dimensions contribute largely to Box's *M* test.

PC1 & PC2



PC3 & PC4



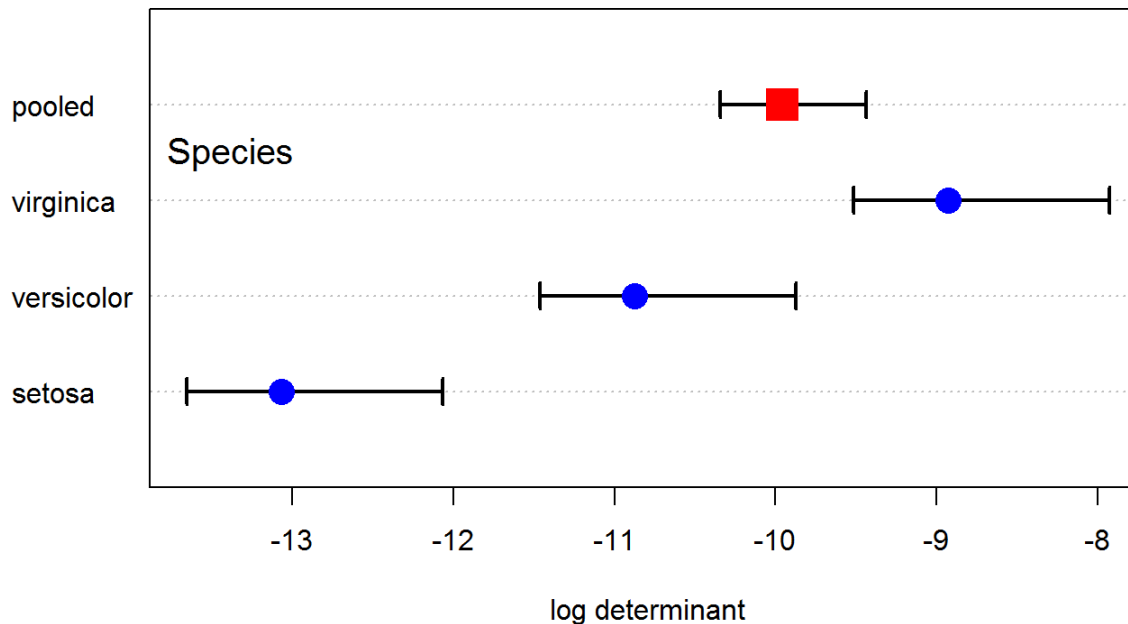
```
iris.pca <- prcomp(iris[,1:4])  
covEllipses(iris.pca$x, iris$Species, ...)  
covEllipses(iris.pca$x, iris$Species, center=TRUE, ...)
```

```
covEllipses(iris.pca$x, iris$Species, center=TRUE,  
            variables=3:4, ...)
```

Visualizing Box's M test

Box's test is based on a comparison of the $\log |S_i|$ relative to $\log |S_p|$: **plot them!**

```
iris.boxm <- boxM(iris[, 1:4], iris[, "Species"])  
plot(iris.boxm, glabel="Species")
```

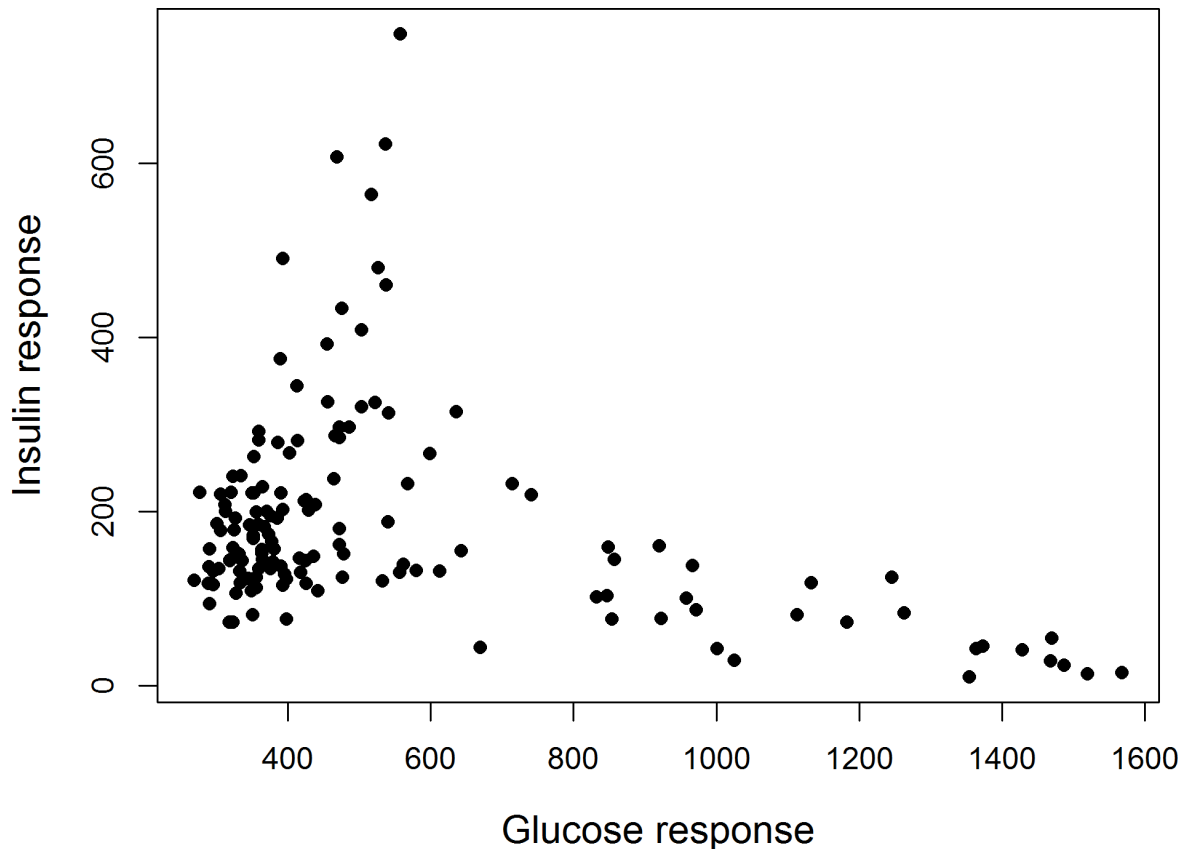


CI's based on an asymptotic CLT \approx distribution of $\ln |S|$ (Cai, Liang, and Zhou 2016)
(Thx: Augustine Wong)

Unsolved: Bootstrap CI

Diabetes data: 2D mystery

Reaven & Miller (1968) found a peculiar “horse shoe” result in analysis of data on the relationship of blood glucose levels and production of insulin in patients with varying degrees of hyperglycemia

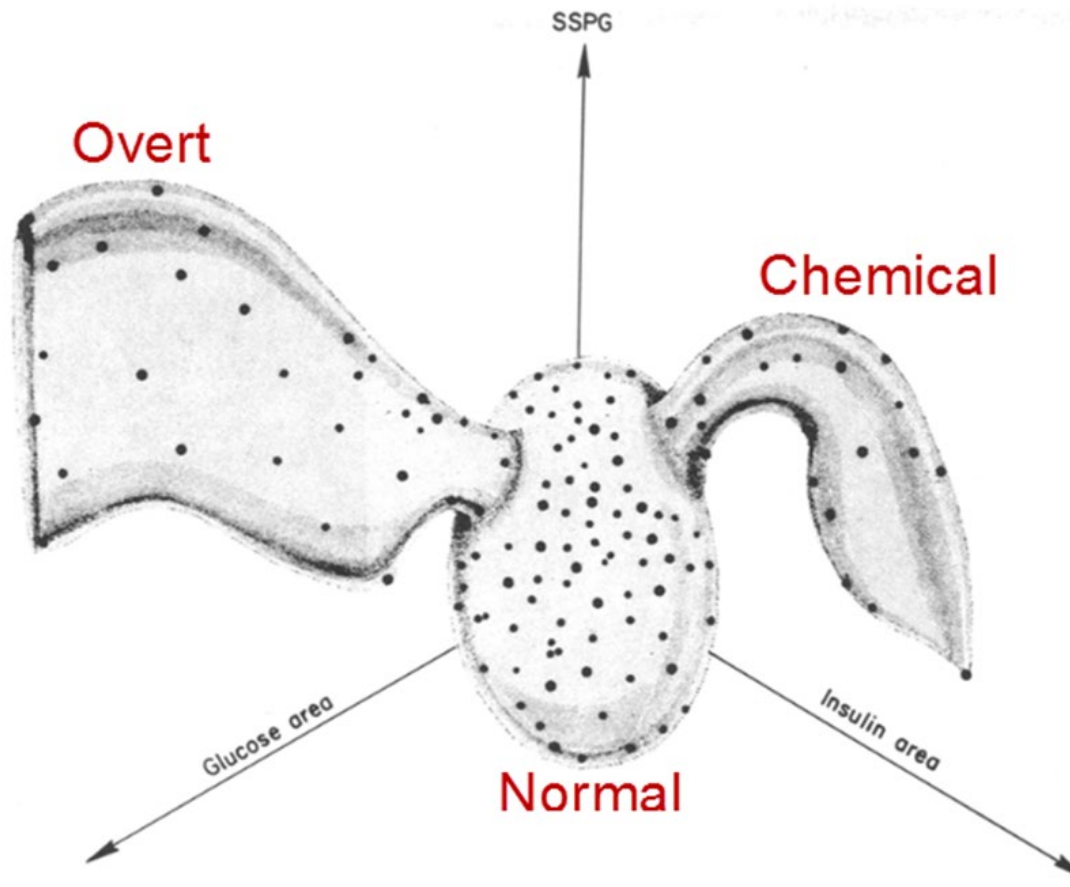


In a 2D plot this was a medical mystery.

What could be the explanation?

Diabetes data: 3D clarity

Using the first 3D computer graphics system (PRIM-9) they rotated the data in 3-space until a hypothesis was suggested.



Artist's view of the data suggests there were actually three groups in the data.

Two categories of Type 2 diabetes:

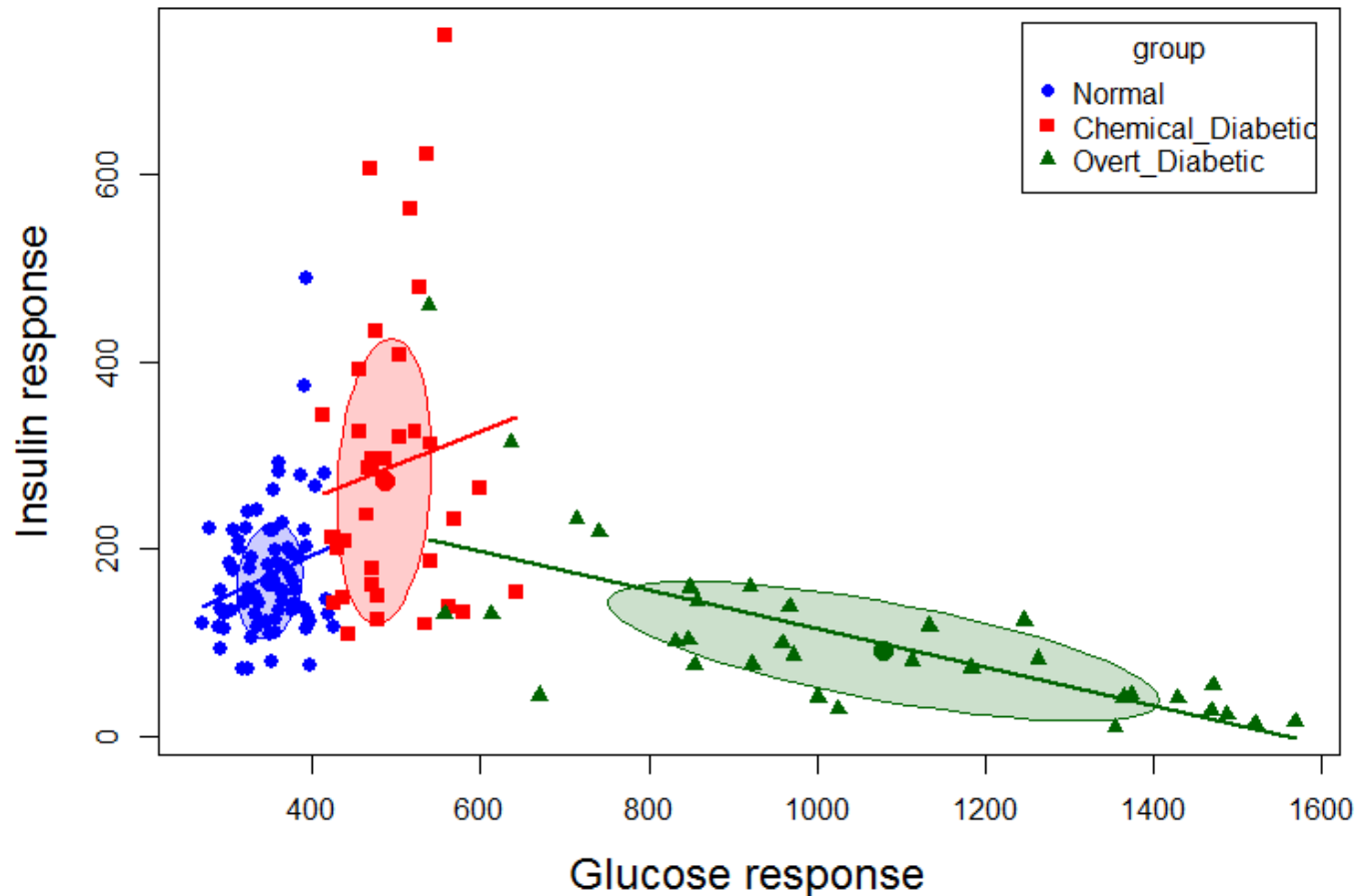
- Overt (advanced)
- Chemical (latent)

But, these were NOT stages in a progression!

Diabetes data: scatterplots

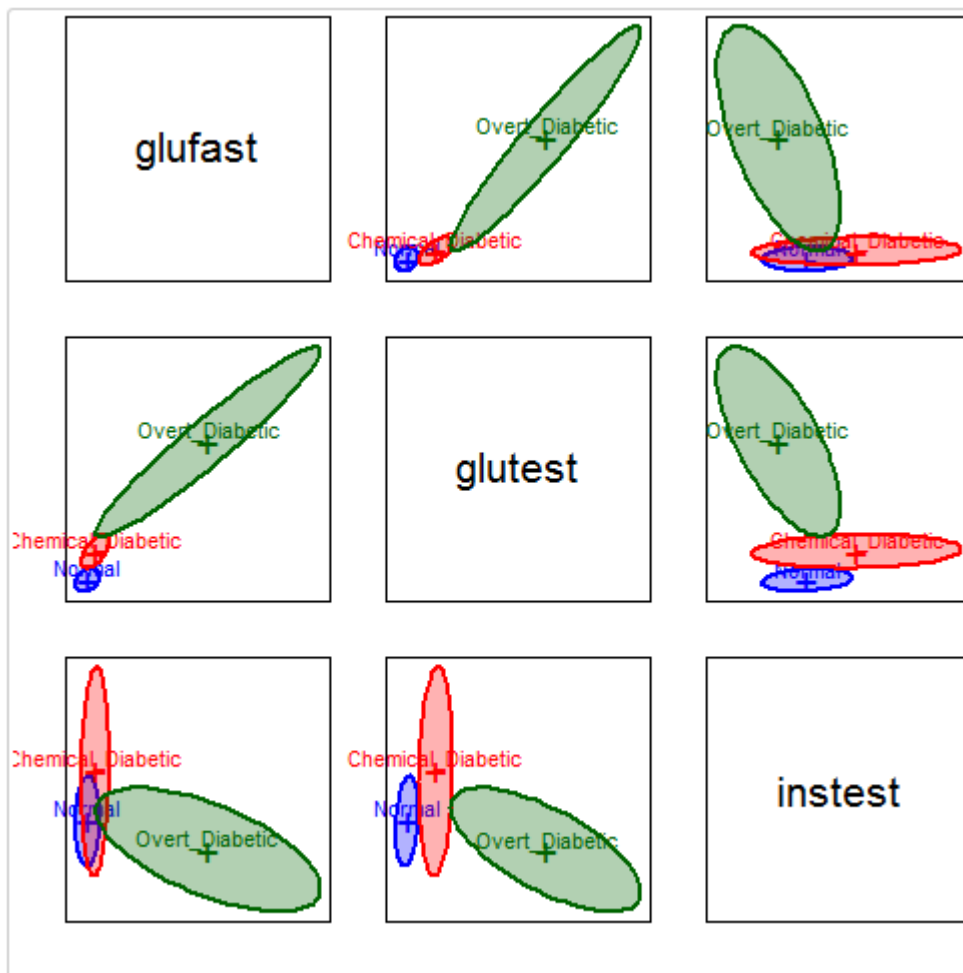
A grouped scatterplot reveals the truer story

```
scatterplot( instest ~ glutest | group, data=Diabetes, ellipse=TRUE, ... )
```



Diabetes data: Ellipses

```
covEllipses(Diabetes[,2:5], Diabetes$group, fill=TRUE, pooled=FALSE,  
col=c("blue", "red", "darkgreen"), variables=1:3)
```



The pairwise data ellipses show visual summaries of the data

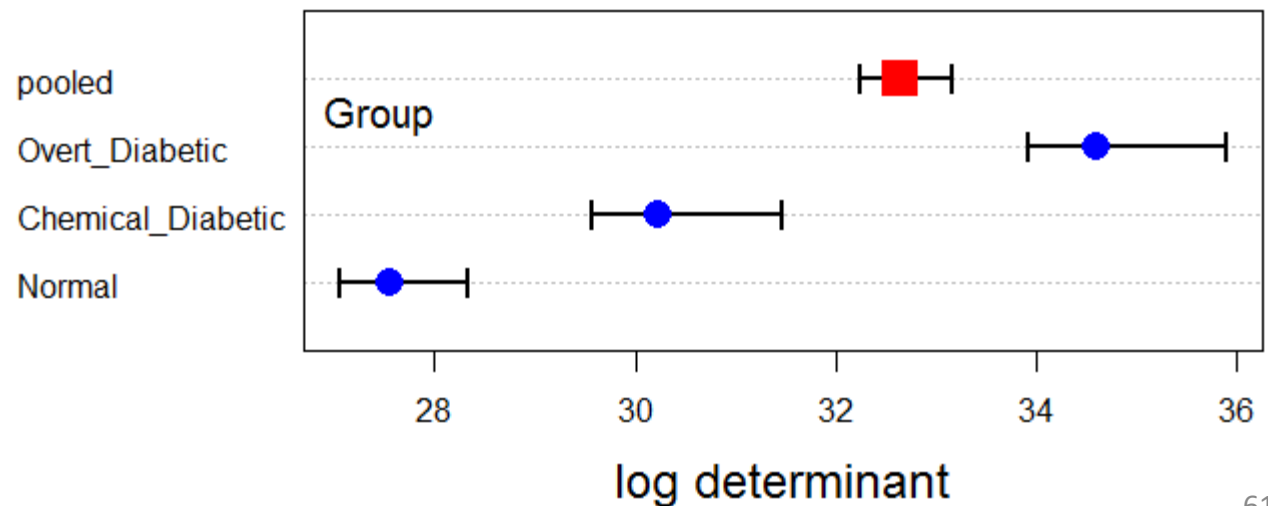
The differences in correlation and variances are dramatic here

Box's M test

```
diab.boxm <- boxM(Diabetes[,2:5], Diabetes$group)
diab.boxm
##
## Box's M-test for Homogeneity of Covariance Matrices
##
## data: Diabetes[, 2:5]
## Chi-Sq (approx.) = 383, df = 20, p-value <2e-16

plot(diab.boxm)
```

All pairs differ!



MANOVA & HE plots

```
diab.mlm <- lm(cbind(glufast, glutest, instest, sspg) ~ group, data=Diabetes)
```

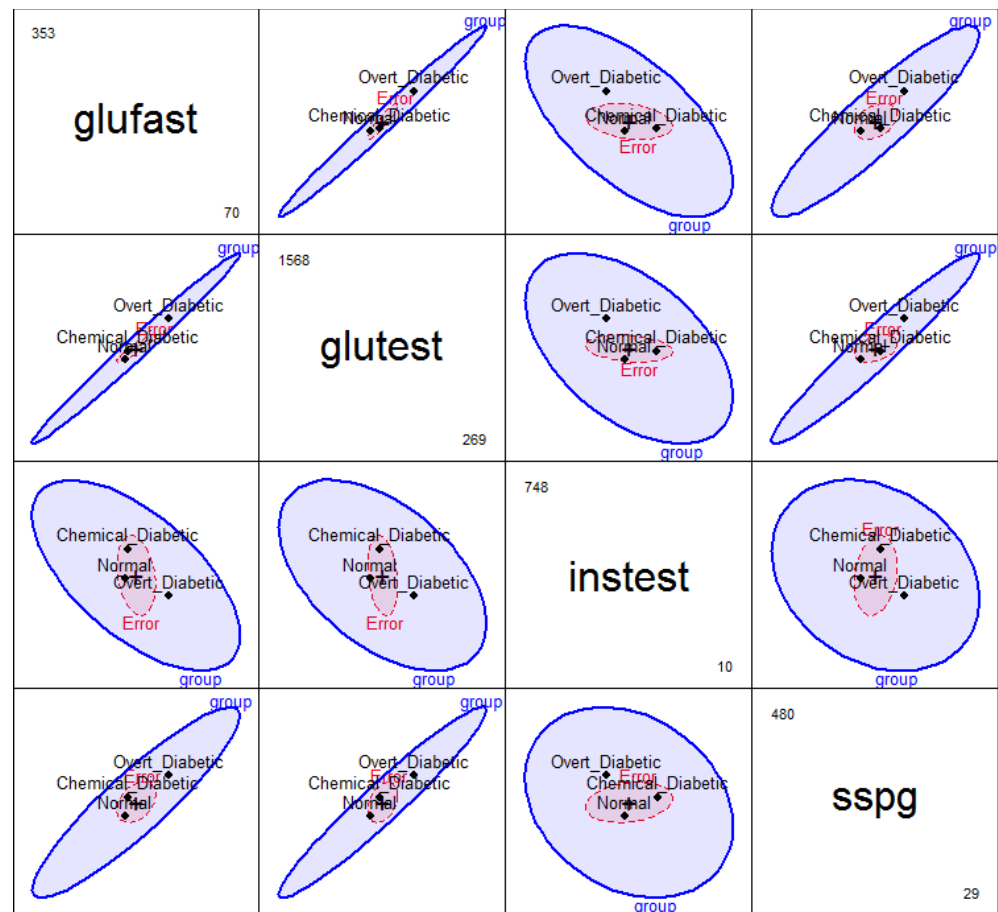
```
# HE pairs plot
```

```
pairs(diab.mlm, fill=TRUE, fill.alpha=0.1, var.cex=3, cex=1.25)
```

The groups are well-separated in all views

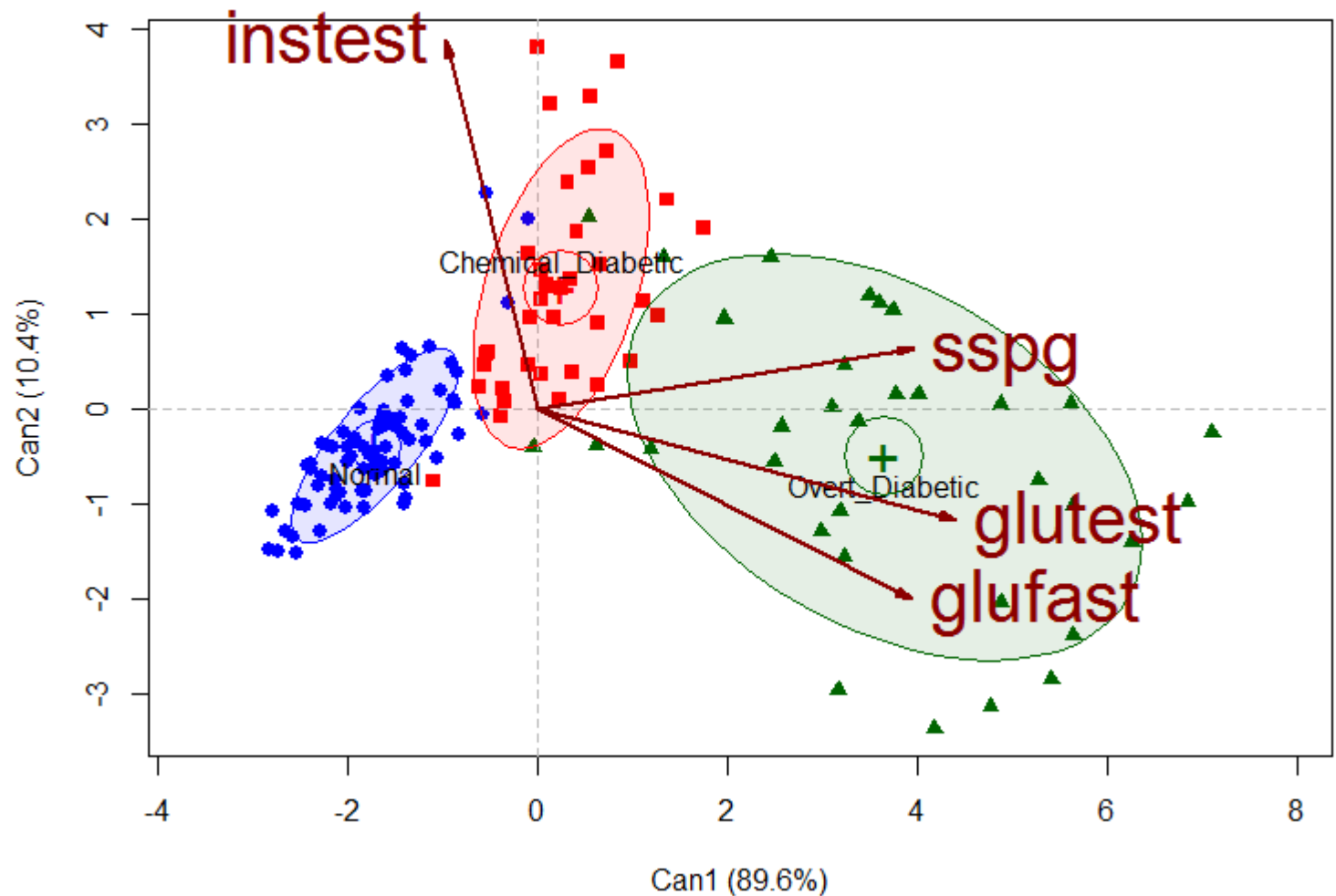
Mostly positive correlation among group means

But: one of these plots is not like the others!



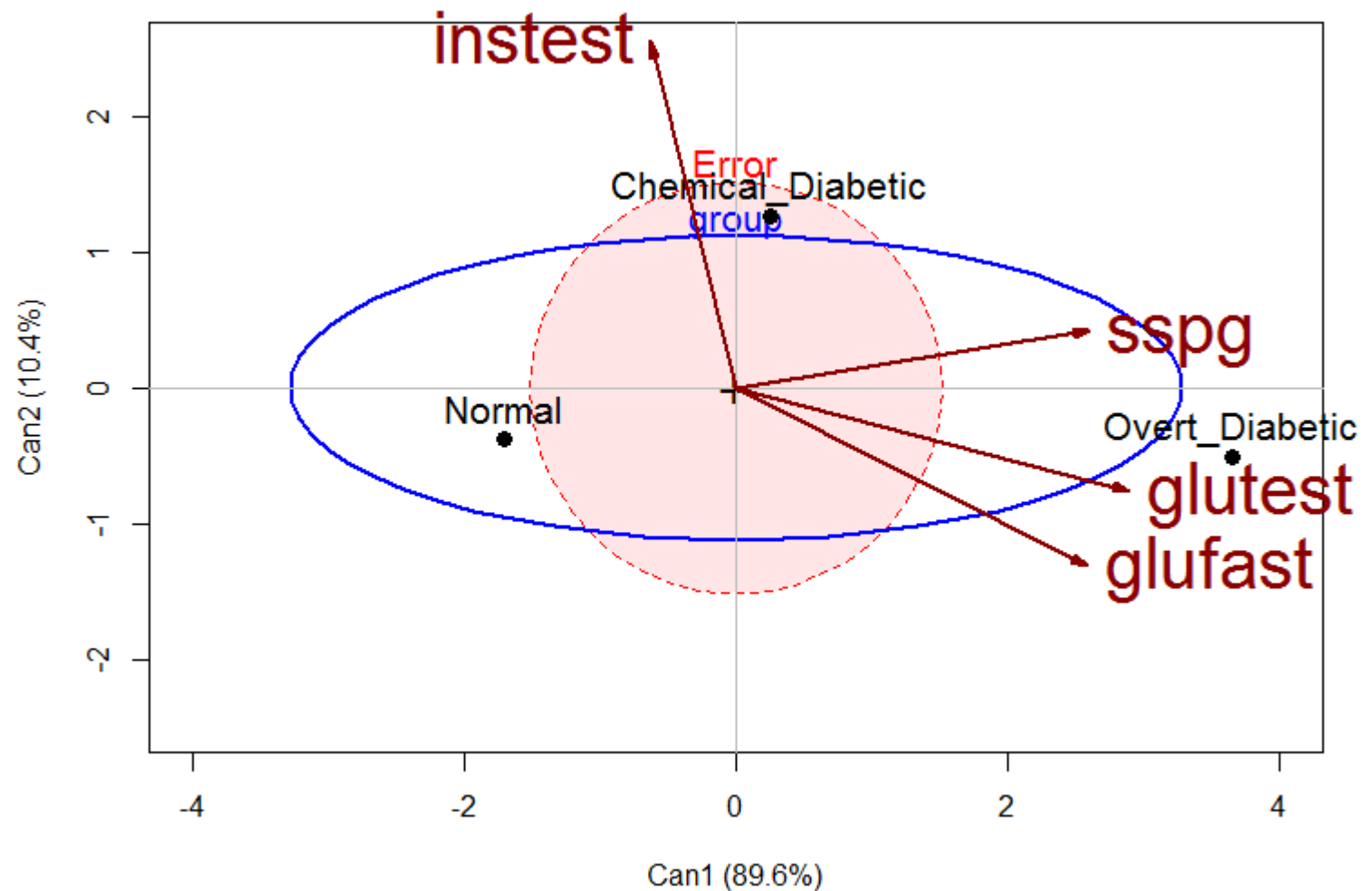
Run the juicer

```
diab.can <- candisc(diab.mlm)  
plot(diab.can, ellipse=TRUE, var.lwd=2, ...)
```



Filter the juice

```
heplot(diab.can, fill=c(TRUE, FALSE), fill.alpha=0.1, var.lwd=2, size="effect")
```



Summary

- MANOVA tests of MLMs are easily visualized in HE plots
 - Contrasts among groups can be easily shown
 - Canonical plots show data in 2D/3D space of max. group differences
 - Robust methods can help guard against outliers
- MMRA models
 - Visualize effects of quant. predictors as lines in data space
 - Test & visualize any linear hypothesis
 - Canonical correlations: visualize in 2D/3D of max. (X, Y) correlations
- Homogeneity of covariances
 - Visualize within-group \mathbf{S}_i and pooled \mathbf{S}_p by data ellipses
 - Visualize Box's M test by simple dot plot of $|\mathbf{S}_p|$ and $|\mathbf{S}_i|$