Tutorial 11

Biological Data Analysis

Spring 2023

Outline

- Two –way ANOVA
- Classical approach
- Regression approach

2-way ANOVA
Example 1

Sodium can affect the levels of the blood pressure. One of the kidney functions is to regulate the level of sodium in the body, using enzyme called Na-K-ATPase. If the enzyme does not function properly the patient might suffer from the high blood pressure.

The kidneys have different parts that have different functions. Garg conducted research on the activity of the enzyme in the various parts of the kidneys.

Two types of rats (4 of each type) participated in the experiment: healthy rats and the rats with high blood pressure.

Enzyme activity was measured in 3 different parts: DCT(distal collecting tubule), CCT(cortical collecting tubule), OMCD(outer medulary collecting tubule)

2-way ANOVA
Example 1

Research questions: is the enzyme activity different in healthy and sick rats? Is there an area in the kidneys with abnormal enzyme activity in rats with high blood pressure?

Factors:

- Rat type: healthy or sick (with high blood pressure)
- Kidney area: OMCD, CCD, DCT

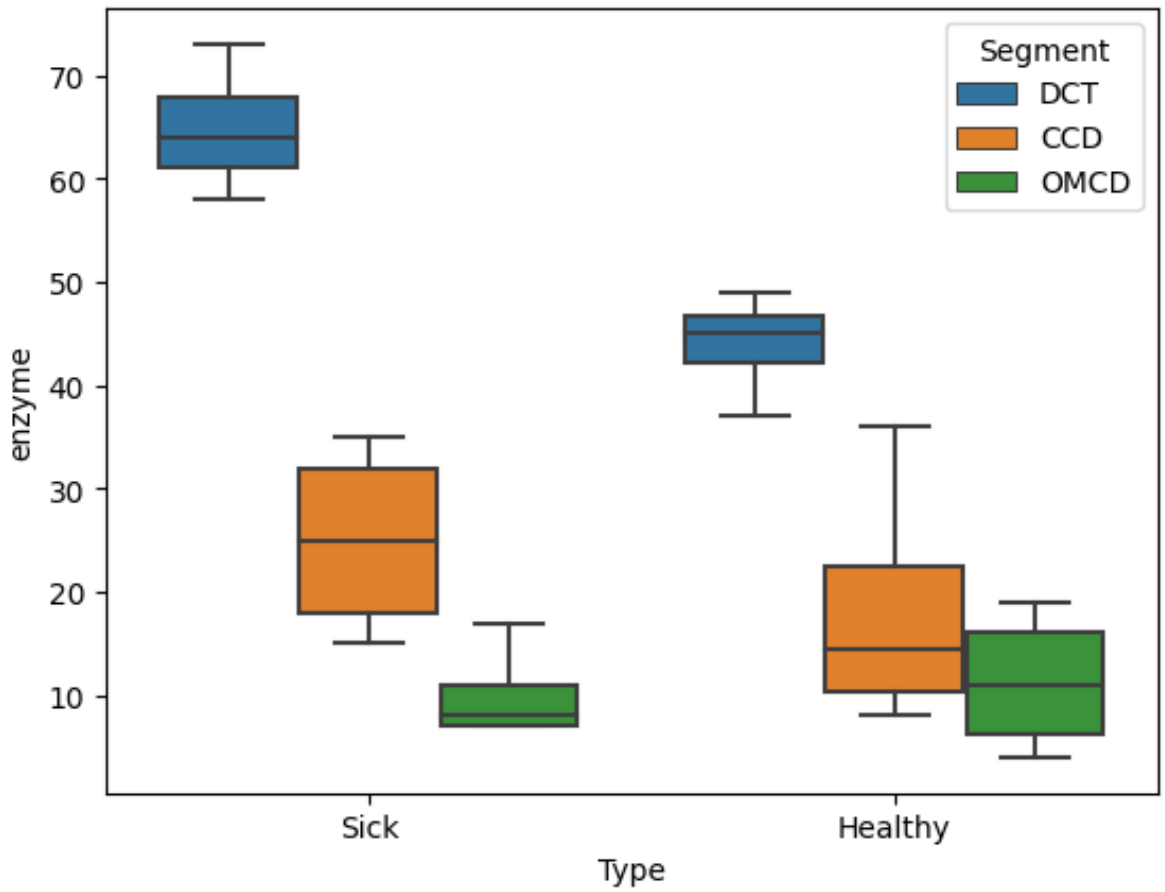
Dependent variable: Na-K-ATPase enzyme activity

Example 1

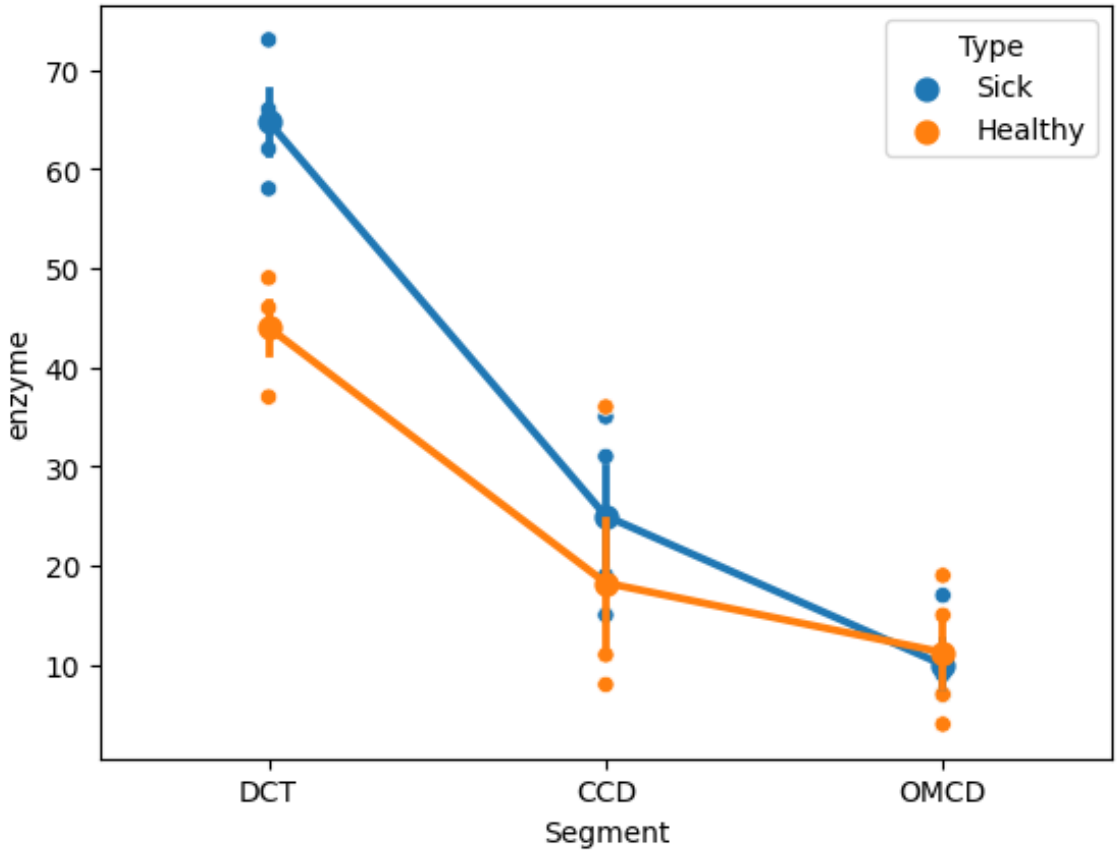
	Sick rats	Healthy rats
DCT	62	44
	73	49
	58	46
	66	37
CCD	15	8
	31	36
	19	11
	35	18
OMCD	7	19
	7	7
	9	15
	17	4

2-way ANOVA

enzyme	Type	Segment
62	Sick	DCT
73	Sick	DCT
58	Sick	DCT
66	Sick	DCT
15	Sick	CCD
31	Sick	CCD
19	Sick	CCD
35	Sick	CCD
7	Sick	OMCD
7	Sick	OMCD
9	Sick	OMCD
17	Sick	OMCD
44	Healthy	DCT
49	Healthy	DCT
46	Healthy	DCT
37	Healthy	DCT
8	Healthy	CCD
36	Healthy	CCD
11	Healthy	CCD
18	Healthy	CCD
19	Healthy	OMCD
7	Healthy	OMCD
15	Healthy	OMCD
4	Healthy	OMCD



2-way ANOVA


Type
Sick
Healthy
enzyme
70
60
50
40
30
20
10
DCT
CCD
OMCD
Segment

2way ANOVA

We can explore three questions here:

- Is there an effect in factor 1 if we neutralize the effect of factor 2?
- Is there an effect in factor 2 if we neutralize the effect of factor 1?
- Does the effect of one factor depends on the second factor? - Interactions

Classical approach – F test

We have two factors:

- Factor A with a levels
- Factor B with b levels

We have n subject in every group:

Sample size (number of measurements) = abn = N

Indices in the analysis:

- i - for levels of factor A
- j - for levels of factor B
- k – for subjects within the group

Classical approach – F test

$$SStot = SS_A + SS_B + SS_{AB} + SS_{res}$$

$$SStot = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x})^2 \qquad df_{tot} = N - 1$$

$$SS_{res} = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{ij})^2 \qquad df_{res} = N - ab$$

$$SS_A = bn \sum_i (\bar{x}_i - \bar{x})^2 \qquad df_A = a - 1$$

$$SS_B = an \sum_j (\bar{x}_j - \bar{x})^2 \qquad df_B = b - 1$$

$$SS_{AB} = n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{x}_{ij} - \bar{x}_j - \bar{x}_i + \bar{x})^2 \qquad df_{AB} = ab - a - b + 1 = (a-1)(b-1)$$

$$SS_{AB} = SStot - SS_A - SS_B - SS_{res} \qquad df_{AB} = (a-1)(b-1)$$

	Sick	Healthy
DCT	62	44
	73	49
	58	46
	66	37
CCD	15	8
	31	36
	19	11
	35	18
OMCD	7	19
	7	7
	9	15
	17	4

Classical approach – F test

Now we can do the F-test for all factors

$$MS = \frac{SS}{df}$$

$$F_A = \frac{MS_A}{MS_{res}} \qquad F_{statA} \sim F(df_A, df_{res})$$

$$F_B = \frac{MS_B}{MS_{res}} \qquad F_{statB} \sim F(df_B, df_{res})$$

$$F_{AB} = \frac{MS_{AB}}{MS_{res}} \qquad F_{statAB} \sim F(df_{AB}, df_{res})$$

Example 1

- Factors
- A – rat type: healthy or sick; a = 2
- B – kidney area: DCT, CCT, OMCD; b = 3
- N = 4*2*3 = 24

- Assumptions
  - Factors:
    1. There is a difference between healthy and sick rats
    2. There is a difference between different kidney areas
  - Interactions:
    1. There is a difference in enzyme activity between healthy and sick rats in specific kidney areas

Two-way ANOVA - significance

$$MS = \frac{SS}{df}$$

$$F_A = \frac{MS_A}{MS_{res}} \qquad F_{statA} \sim F(df_A, df_{res})$$

$$F_B = \frac{MS_B}{MS_{res}} \qquad F_{statB} \sim F(df_B, df_{res})$$

$$F_{AB} = \frac{MS_{AB}}{MS_{res}} \qquad F_{statAB} \sim F(df_{AB}, df_{res})$$

```
Fa = 7.139002805957262,  pa = 0.015551678465012486
Fb = 64.39283401683574,  pb = 6.270463814850302e-09
Fab = 3.854090222318152, pab = 0.04043993883338026
```

Two-way ANOVA – effect size

$$\eta_A^2 = \frac{SS_A}{SS_{\text{tot}}}$$

Small effect - 0.01

Medium effect - 0.06

Large effect - 0.14

```
η2a = 0.0442
η2b = 0.7968
η2ab = 0.0477
```

‘statsmodels’ library

	df	sum_sq	mean_sq	F	PR(>F)
Type	1.0	459.375	459.375000	7.139003	1.555168e-02
Segment	2.0	8287.000	4143.500000	64.392834	6.270464e-09
Type:Segment	2.0	496.000	248.000000	3.854090	4.043994e-02
Residual	18.0	1158.250	64.347222	NaN	NaN

Conclusions

Effect of the rat type factor is significant

P = 0.0156

Effect of the kidney area factor is significant

P = 6.27*10(-9)

Effect of the interactions is significant

P = 0.04

Solution using regression

We can use dummy variables for the group coding.

Number of dummy vectors:

- For factor A: a-1
- For factor B: b-1
- For interactions: (a-1)(b-1)

Coding
reference or effects coding

The coding matters only with respect to the interactions.

In reference coding adding interactions creates structural multicollinearity – better use effects coding

If we are not adding interactions – we can use both reference and effects coding (but we rarely have prior knowledge about the interaction effect in a model)

DUMMY coding
reference coding

$$\hat{y} = b_0 + b_1 D_H + b_2 D_{A2} + b_3 D_{A3} + b_4 D_{HA2} + b_5 D_{HA3}$$

*factor*1 - 2 categories

*factor*2 - 3 categories

$$D_H = \begin{cases} 1 & \textit{factor}1 \text{ category2} \\ 0 & \textit{else} \end{cases}$$

$$D_{A2} = \begin{cases} 1 & \textit{factor}2 \text{ category2} \\ 0 & \textit{else} \end{cases}$$

$$D_{A3} = \begin{cases} 1 & \textit{factor}2 \text{ category3} \\ 0 & \textit{else} \end{cases}$$

$$D_{HA2} = D_H * D_{A2}$$

$$D_{HA3} = D_H * D_{A3}$$

$$\hat{y} = b_0 + b_1 D_H + b_2 D_{A2} + b_3 D_{A3} + b_4 D_{HA2} + b_5 D_{HA3}$$

enzyme	Type	Segment
62	Sick	DCT
73	Sick	DCT
58	Sick	DCT
66	Sick	DCT
15	Sick	CCD
31	Sick	CCD
19	Sick	CCD
35	Sick	CCD
7	Sick	OMCD
7	Sick	OMCD
9	Sick	OMCD
17	Sick	OMCD
44	Healthy	DCT
49	Healthy	DCT
46	Healthy	DCT
37	Healthy	DCT
8	Healthy	CCD
36	Healthy	CCD
11	Healthy	CCD
18	Healthy	CCD
19	Healthy	OMCD
7	Healthy	OMCD
15	Healthy	OMCD
4	Healthy	OMCD

y

D_0	D_H	D_{A2}	D_{A3}	D_{HA2}	D_{HA3}
1.	0.	0.	0.	0.	0.
1.	0.	0.	0.	0.	0.
1.	0.	0.	0.	0.	0.
1.	0.	0.	0.	0.	0.
1.	0.	1.	0.	0.	0.
1.	0.	1.	0.	0.	0.
1.	0.	1.	0.	0.	0.
1.	0.	1.	0.	0.	0.
1.	0.	0.	1.	0.	0.
1.	0.	0.	1.	0.	0.
1.	0.	0.	1.	0.	0.
1.	0.	0.	1.	0.	0.
1.	1.	0.	0.	0.	0.
1.	1.	0.	0.	0.	0.
1.	1.	0.	0.	0.	0.
1.	1.	0.	0.	0.	0.
1.	1.	1.	0.	1.	0.
1.	1.	1.	0.	1.	0.
1.	1.	1.	0.	1.	0.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.

X

$$\hat{y} = b_0 + b_1 D_H + b_2 D_{A2} + b_3 D_{A3} + b_4 D_{HA2} + b_5 D_{HA3}$$

b

[64.75
-20.75
-39.75
-54.75
14.
22.]

b_0 - mean for sick DCT

b_1 - difference between sick and healthy in DCT

b_2 - difference between DCT and CCD in sick

b_3 - difference between DCT and OMCD in sick

$$\hat{y} = b_0 + b_1 D_H + b_2 D_{A2} + b_3 D_{A3} + b_4 D_{HA2} + b_5 D_{HA3}$$

b

[64.75
-20.75
-39.75
-54.75
14.
22.]

$b_0 + b_1 + b_2 + b_4$ - mean for healthy CCD

$b_0 + b_1 + b_3 + b_5$ - mean for healthy OMCD

$b_0 + b_1$ - mean for healthy DCT

$b_0 + b_3$ - mean for sick OMCD

$b_0 + b_1 + b_2 / 3 + b_3 / 3 + b_4 / 3 + b_5 / 3 -$ mean for healthy

$b_0 + b_2 / 3 + b_3 / 3 -$ mean for sick

$b_0 + b_1 / 2$ - mean for DCT

$b_0 + b_1 / 2 + b_2 + b_4 / 2$ - mean for CCD

DUMMY coding
effects coding

The variables a coded with -1, 0 or 1. We have k groups

$$E_1 = \begin{cases} 1 & group \quad 1 \\ -1 & group \quad k \\ 0 & else \end{cases}$$

$$\vdots$$

$$E_{k-1} = \begin{cases} 1 & group \quad k-1 \\ -1 & group \quad k \\ 0 & else \end{cases}$$

- b(0) – average value of a whole sample
- A coefficient of a variable (b(1) – b(k-1)) – the difference between the group average and the whole sample average
- The difference between the average of group k and the whole sample average is minus sum of all the coefficients

$$\hat{y} = b_0 + b_H D_H + b_{A1} D_{A1} + b_{A2} D_{A2} + b_{HA1} D_{HA1} + b_{HA2} D_{HA2}$$

Effects coding:

1. Rat type:

   2 groups – 1 dummy vector

H	
1	sick
-1	healthy

2. Kidney area:

   3 groups – 2 dummy vectors

A2	A1	
0	1	DCT
1	0	CCD
-1	-1	OMCD

3. Interactions: 2 dummy vectors

multiplication of dummy vectors for type and area

$$\hat{y} = b_0 + b_H D_H + b_{A1} D_{A1} + b_{A2} D_{A2} + b_{HA1} D_{HA1} + b_{HA2} D_{HA2}$$

	Sick	Healthy
DCT	H=1 A1=1 A2=0	H= -1 A1=1 A2=0
CCD	H=1 A1=0 A2=1	H= -1 A1=0 A2=1
OMCD	H=1 A1= -1 A2= -1	H= -1 A1= -1 A2= -1

enzyme	Type	Segment
62	Sick	DCT
73	Sick	DCT
58	Sick	DCT
66	Sick	DCT
15	Sick	CCD
31	Sick	CCD
19	Sick	CCD
35	Sick	CCD
7	Sick	OMCD
7	Sick	OMCD
9	Sick	OMCD
17	Sick	OMCD
44	Healthy	DCT
49	Healthy	DCT
46	Healthy	DCT
37	Healthy	DCT
8	Healthy	CCD
36	Healthy	CCD
11	Healthy	CCD
18	Healthy	CCD
19	Healthy	OMCD
7	Healthy	OMCD
15	Healthy	OMCD
4	Healthy	OMCD

y

D_0	D_H	D_{A1}	D_{A2}	D_{HA1}	D_{HA2}
1.	1.	1.	0.	1.	0.
1.	1.	1.	0.	1.	0.
1.	1.	1.	0.	1.	0.
1.	1.	1.	0.	1.	0.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.
1.	1.	-1.	-1.	-1.	-1.
1.	1.	-1.	-1.	-1.	-1.
1.	1.	-1.	-1.	-1.	-1.
1.	1.	-1.	-1.	-1.	-1.
1.	-1.	1.	0.	-1.	-0.
1.	-1.	1.	0.	-1.	-0.
1.	-1.	1.	0.	-1.	-0.
1.	-1.	1.	0.	-1.	-0.
1.	-1.	0.	1.	0.	-1.
1.	-1.	0.	1.	0.	-1.
1.	-1.	0.	1.	0.	-1.
1.	-1.	0.	1.	0.	-1.
1.	-1.	-1.	-1.	1.	1.
1.	-1.	-1.	-1.	1.	1.
1.	-1.	-1.	-1.	1.	1.
1.	-1.	-1.	-1.	1.	1.

X

$$\hat{y} = b_0 + b_H D_H + b_{A1} D_{A1} + b_{A2} D_{A2} + b_{HA1} D_{HA1} + b_{HA2} D_{HA2}$$

```
Model significance:
F-statistic = 28.7266 , p-value = 0.0

Model effect size: R2 = 0.8886
```

```
b =
    [28.875
     4.375
     25.5
    -7.25
     6.
    -1. ]
```

```
model coefficient significance:
b0: 0.0
b1: 0.0049
b2: 0.0
b3: 0.0017
b4: 0.0059
b5: 0.3232
```

Separate factors contribution addition test

Full model vs model without factor

Significance for factors

```
p type = 0.0156
p area = 0.0
p interactions = 0.0404
```

2-way ANOVA

Effect size for factors

$$\eta_A^2 = \frac{SS_A}{SS_{\text{tot}}}$$

0.01 – Small effect

0.06 – Median effect

0.14 – Large effect

```
factors effect size:
η2 type = 0.0442
η2 area = 0.7968
η2 interactions = 0.0477
```

Results summary

$$\hat{y} = b_0 + b_H D_H + b_{A1} D_{A1} + b_{A2} D_{A2} + b_{HA1} D_{HA1} + b_{HA2} D_{HA2}$$

b = 28.8750 4.3750 25.5000 -7.2500 6.0000 -1.0000

- Factor of rat type: enzyme activity in significantly higher in sick rats than in healthy ones.
- Factor of kidney area: There is a significance difference in enzyme activity in different areas. In DCT – higher than average and in CCD and OMCD – lower than average.
- Factor of interactions: in some areas the difference in enzyme activity is bigger than in the others.

Meaning of coefficients in a model

$$\hat{y} = b_0 + b_H D_H + b_{A1} D_{A1} + b_{A2} D_{A2} + b_{HA1} D_{HA1} + b_{HA2} D_{HA2}$$

b0 – overall average
bH – difference between the healthy rats and overall average
-bH – difference between the sick rats and overall average
bA1 – difference between the DCT area (for both types of rates) and overall average
bA2 – difference between the CCD (for both types of rates) and overall average
-(bA1+bA2) – difference between the OMCD (for both types) and the overall average

Meaning of coefficients in a model

$$\hat{y} = b_0 + b_H D_H + b_{A1} D_{A1} + b_{A2} D_{A2} + b_{HA1} D_{HA1} + b_{HA2} D_{HA2}$$

		1 0	0 1	-1 -1	
		DCT	CCD	OMCD	
1	S	$b_0+b_H+b_{A1}+b_{HA1}$	$b_0+b_H+b_{A2}+b_{HA2}$	$b_0+b_H-b_{A1}-b_{A2}-b_{HA1}-b_{HA2}$	b_0+b_H
-1	H	$b_0-b_H+b_{A1}-b_{HA1}$	$b_0-b_H+b_{A2}-b_{HA2}$	$b_0-b_H-b_{A1}-b_{A2}+b_{HA1}+b_{HA2}$	b_0-b_H
		b_0+b_{A1}	b_0+b_{A2}	$b_0-b_{A1}-b_{A2}$	b_0

enzyme	Type	Segment
62	Sick	DCT
73	Sick	DCT
58	Sick	DCT
66	Sick	DCT
15	Sick	CCD
31	Sick	CCD
19	Sick	CCD
35	Sick	CCD
7	Sick	OMCD
7	Sick	OMCD
9	Sick	OMCD
17	Sick	OMCD
44	Healthy	DCT
49	Healthy	DCT
46	Healthy	DCT
37	Healthy	DCT
8	Healthy	CCD
36	Healthy	CCD
11	Healthy	CCD
18	Healthy	CCD
19	Healthy	OMCD
7	Healthy	OMCD
15	Healthy	OMCD
4	Healthy	OMCD

y

D_0	D_H	D_{A1}	D_{A2}	D_{HA1}	D_{HA2}
1.	1.	1.	0.	1.	0.
1.	1.	1.	0.	1.	0.
1.	1.	1.	0.	1.	0.
1.	1.	1.	0.	1.	0.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.
1.	1.	0.	1.	0.	1.
1.	1.	-1.	-1.	-1.	-1.
1.	1.	-1.	-1.	-1.	-1.
1.	1.	-1.	-1.	-1.	-1.
1.	1.	-1.	-1.	-1.	-1.
1.	-1.	1.	0.	-1.	-0.
1.	-1.	1.	0.	-1.	-0.
1.	-1.	1.	0.	-1.	-0.
1.	-1.	1.	0.	-1.	-0.
1.	-1.	0.	1.	0.	-1.
1.	-1.	0.	1.	0.	-1.
1.	-1.	0.	1.	0.	-1.
1.	-1.	0.	1.	0.	-1.
1.	-1.	-1.	-1.	1.	1.
1.	-1.	-1.	-1.	1.	1.
1.	-1.	-1.	-1.	1.	1.
1.	-1.	-1.	-1.	1.	1.

X

Comparing groups - significance

$$\hat{y} = b_0 + b_H D_H + b_{A1} D_{A1} + b_{A2} D_{A2} + b_{HA1} D_{HA1} + b_{HA2} D_{HA2}$$

$$t_{mean1-mean2} = \frac{mean1 - mean2}{\sqrt{\text{var}(mean1 - mean2)}} \sim T_{vres}$$

$$\begin{aligned} \text{Var}\left(\sum_i A_i\right) &= \text{Cov}\left(\sum_i A_i, \sum_j A_j\right) \\ &= \sum_i \sum_j \text{Cov}(A_i, A_j) \\ &= \sum_i \text{Var}(A_i) + 2 \sum_{i<j} \text{Cov}(A_i, A_j) \end{aligned}$$

$$Var(aX) = a^2 Var(X)$$

Comparing groups - significance

Sick and healthy in DCT

$$t_{sick-healthy \text{ in DCT}} = \frac{2b_H + 2b_{HA1}}{\sqrt{\text{var}(2b_H + 2b_{HA1})}}$$

$$\text{var}(2b_H + 2b_{HA1}) = \text{var}(2b_H) + \text{var}(2b_{HA1}) + 2\,\text{cov}(2b_H, 2b_{HA1}) =$$
$$4\,\text{var}(b_H) + 4\,\text{var}(b_{HA1}) + 8\,\text{cov}(b_H, b_{HA1})$$

sick vs healthy in DCT p value = 0.0005

Comparing groups – effect size

Sick and healthy in DCT

$$\frac{2b_H + 2b_{HA1}}{\sqrt{MSres}}$$

```
sick vs healthy in DCT effect size = 2.5867
```

Comparing groups significance and effect size

DCT vs OMCD

$$t_{DCT-OMCD} = \frac{2b_{A1} + b_{A2}}{\sqrt{\mathrm{var}(2b_{A1} + b_{A2})}}$$

$$\mathrm{var}(2b_{A1} + b_{A2}) = 4\,\mathrm{var}(b_{A1}) + \mathrm{var}(b_{A2}) + 4\,\mathrm{cov}(b_{A1}, b_{A2})$$

DCT vs OMCD p value = 3.38793548770866e-10

Effect size

$$\frac{2b_{A1} + b_{A2}}{\sqrt{MSres}}$$

DCT vs OMCD effect size = 5.454

Comparing groups significance and effect size

DCT vs CCD in healthy

$$t_{DCT-CCD \text{ in healthy}} = \frac{b_{A1} - b_{A2} - b_{HA1} + b_{HA2}}{\sqrt{\text{var}(b_{A1} - b_{A2} - b_{HA1} + b_{HA2})}}$$

$$\begin{aligned}\text{var}(b_{A1} - b_{A2} - b_{HA1} + b_{HA2}) &= \text{var}(b_{A1}) + \text{var}(b_{A2}) + \text{var}(b_{HA1}) + \text{var}(b_{HA2}) \\ &- 2\,\text{cov}(b_{A1}, b_{A2}) - 2\,\text{cov}(b_{A1}, b_{HA1}) + 2\,\text{cov}(b_{A1}, b_{HA2}) \\ &+ 2\,\text{cov}(b_{A2}, b_{HA1}) - 2\,\text{cov}(b_{A2}, b_{HA2}) - 2\,\text{cov}(b_{HA1}, b_{HA2})\end{aligned}$$

DCT CCD in healthy p value = 5.723560529313776e-0

Effect size

$$\frac{b_{A1} - b_{A2} - b_{HA1} + b_{HA2}}{\sqrt{MSres}}$$

DCT vs CCD in healthy effect size = 3.2101

Comparing groups significance and effect size

Sick vs healthy

$$t_{sick-healthy} = \frac{2b_H}{\sqrt{\text{var}(2b_H)}}$$

$$\text{var}(2b_H) = 4\,\text{var}(b_H)$$

sick vs healthy p value = 0.0049194280143939 3

Effect size

$$\frac{2b_H}{\sqrt{MSres}}$$

sick vs healthy effect size = 1.0908