

## Lecture 20

1. Adjusting and imbalance
2. Longitudinal data and plots
3. Within-subject correlation
4. Area under the curve (AUC)

1

### Adjusting and imbalance

We often need to adjust the comparison of group responses  $Y$  for covariate  $X$  that

- is associated with the outcome ( $Y$  is associated with  $X$ )
- differs between groups ( $\bar{X}_1 \neq \bar{X}_2$ )

Under these circumstances,  $X$  may cause part or all of difference between groups.

**Adjusting** or **controlling** for  $X$  means: *remove effects of  $X$  from group comparison.*

Compare groups restricted to subjects with the same (or similar) values of  $X$ .

2

Compare groups 1 and 2. Main approaches for controlling  $X$ :

- Use regression ( $Y$  on  $X$ ) to estimate “effect” of  $X$  on  $Y$ .

Compare predicted group means of  $Y$  at average  $X$

LSmeans group

- Cut  $X$  into strata. Find group difference ( $\bar{Y}_{1i} - \bar{Y}_{2i}$ ) within each stratum  $i$ .

*If* these appear to be estimates of a common difference

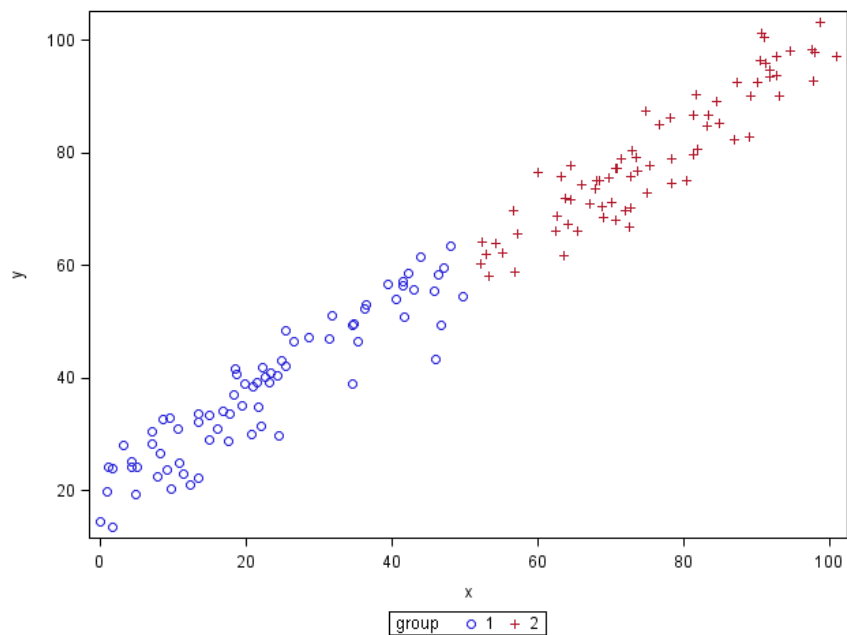
(ie. no group  $\times$   $X$  interaction) then pool differences.

LSmeans group

*If there's an interaction:* report the interaction

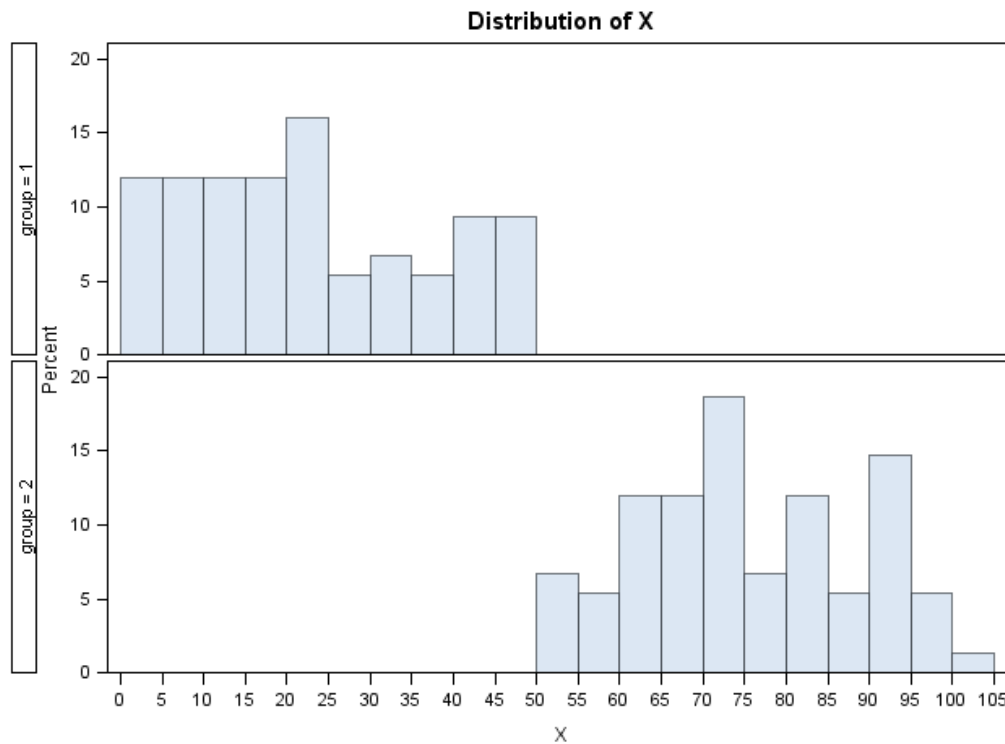
3

Imbalance of  $X$  between groups: how do we control  $X$ ?



**Regression discontinuity**

4



No subjects in group 1 with same  $X$  value as a subject in group 2.

5

No subjects in group 1 with same  $X$  value as a subject in group 2.

Can we hold  $X$  fixed and compare the groups? Perhaps just in the middle?

If  $X$  divides the groups, how can we remove  $X$  from the comparison?

What happens to conventional adjustment?

```
Proc GLM data=discont;
  class Group;
  model Y = X Group / solution;
  LSmeans Group / stderr E;
```

Which adjustment is this?

6

Parameter		Estimate	Standard Error	t Value	Pr >  t
Intercept		17.02386252 B	2.15169005	7.91	<.0001
X		0.82569032	0.02758120	29.94	<.0001
group	1	2.55129055 B	1.64488893	1.55	0.1230
group	2	0.00000000 B	.	.	.

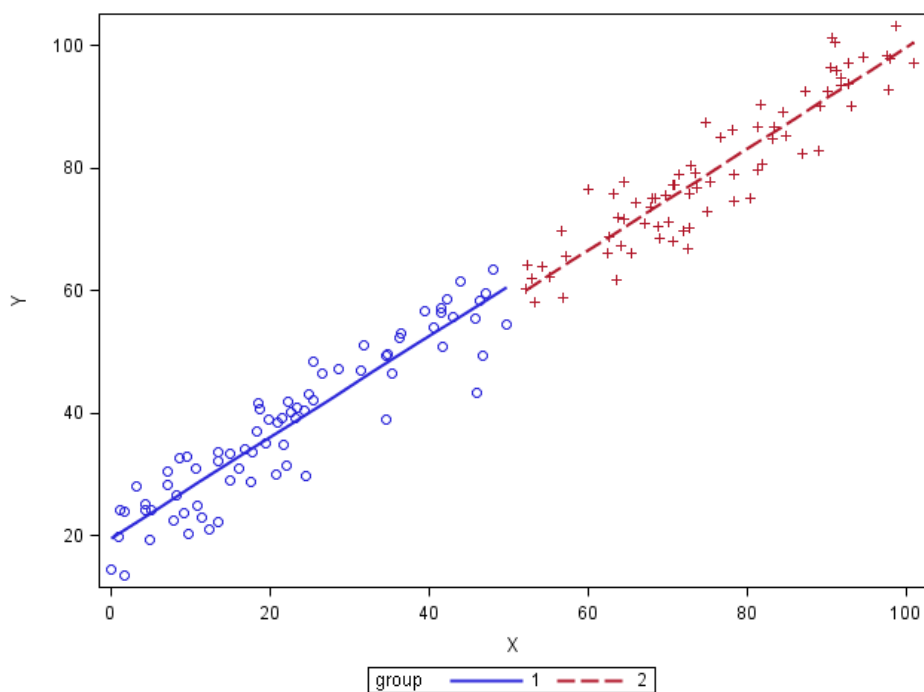
Coefficients for group Least Square Means

Effect	group Level	
	1	2
Intercept	1	1
X	49.074533	49.074533
group	1	0
group	2	1

group	Y LSMEAN	Standard Error	Pr >  t
1	60.0955199	0.9044473	<.0001
2	57.5442293	0.9044473	<.0001

7

Where are the LSmeans?



8

Regression adjustment compares *extrapolated* predicted means, near edge or outside data.

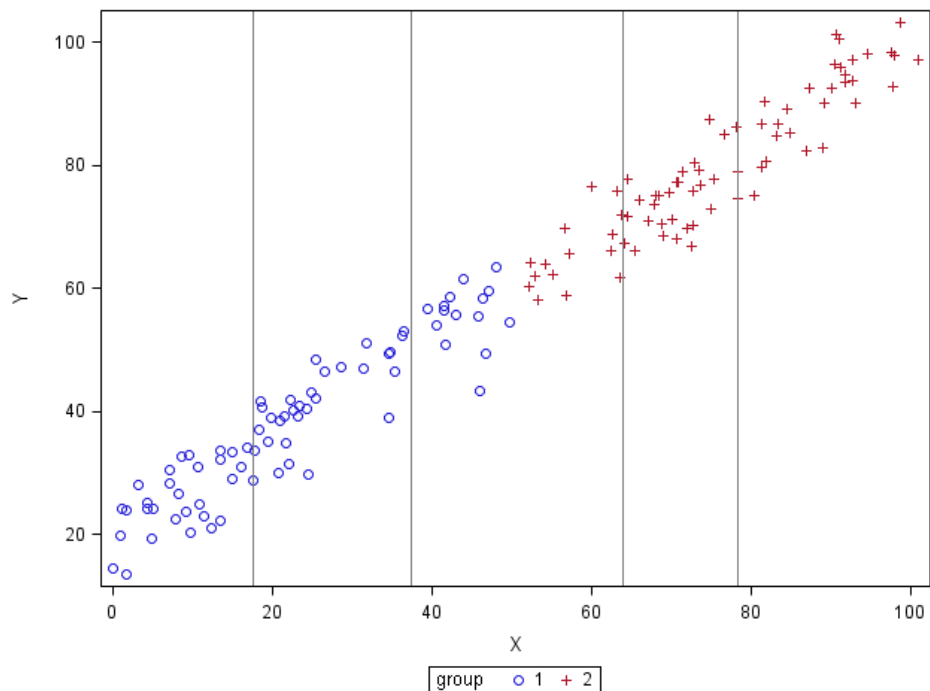
Regression “creates” group means at same value of  $X$ .

Can regression really change subjects' values of  $X$ ?

Any sign of problems in the SAS output?

9

Stratification adjustment: cut  $X$  into quintiles



Look at interaction first:

```
Proc GLM data=discont;
  class Group Quintile_X;
  model Y = Quintile_X Group Quintile_X*Group / solution;
  LSmeans Quintile_X*Group / stderr; 10 means
  LSmeans Group / stderr E; 2 pooled means
```

Need Quintile\_X\*Group means for interaction plot

11

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Quintile_X	4	16675.90561	4168.97640	109.48	<.0001
group	1	767.49901	767.49901	20.16	<.0001
group*Quintile_X	0	0.00000	.	.	.

Parameter		Estimate	Standard Error	t Value	Pr >  t
Intercept		90.78888403 B	1.12663049	80.58	<.0001
Quintile_X	0	-54.66803695 B	2.75966984	-19.81	<.0001
Quintile_X	1	-39.00549269 B	2.75966984	-14.13	<.0001
Quintile_X	2	-25.00392411 B	1.95138126	-12.81	<.0001
Quintile_X	3	-15.87762490 B	1.59329612	-9.97	<.0001
Quintile_X	4	0.00000000 B	.	.	.
group	1	-10.11598740 B	2.25326099	-4.49	<.0001
group	2	0.00000000 B	.	.	.
group*Quintile_X	1 0	0.00000000 B	.	.	.
group*Quintile_X	1 1	0.00000000 B	.	.	.
group*Quintile_X	1 2	0.00000000 B	.	.	.
group*Quintile_X	2 2	0.00000000 B	.	.	.
group*Quintile_X	2 3	0.00000000 B	.	.	.
group*Quintile_X	2 4	0.00000000 B	.	.	.

12

How does this reveal the imbalance?

Least Squares Means

group	Quintile_ X	Y LSMEAN	Standard Error	Pr >  t
1	0	26.0048597	1.1266305	<.0001
1	1	41.6674039	1.1266305	<.0001
1	2	55.6689725	1.5932961	<.0001
2	2	65.7849599	1.5932961	<.0001
2	3	74.9112591	1.1266305	<.0001
2	4	90.7888840	1.1266305	<.0001

group	Y LSMEAN
1	Non-est
2	Non-est

13

What if we don't check the interaction, but just fit the main-effects adjustment:

```
Proc GLM data=discont;  
  class Group Quintile_X;  
  model Y = Quintile_X Group / solution;  
  LSmeans Group / stderr E;
```

14

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	90.78888403 B	1.12663049	80.58	<.0001
Quintile_X 0	-54.66803695 B	2.75966984	-19.81	<.0001
Quintile_X 1	-39.00549269 B	2.75966984	-14.13	<.0001
Quintile_X 2	-25.00392411 B	1.95138126	-12.81	<.0001
Quintile_X 3	-15.87762490 B	1.59329612	-9.97	<.0001
Quintile_X 4	0.00000000 B	.	.	.
group 1	-10.11598740 B	2.25326099	-4.49	<.0001
group 2	0.00000000 B	.	.	.

How is Proc GLM estimating these?

Any signs of trouble?

15

The GLM Procedure  
Least Squares Means

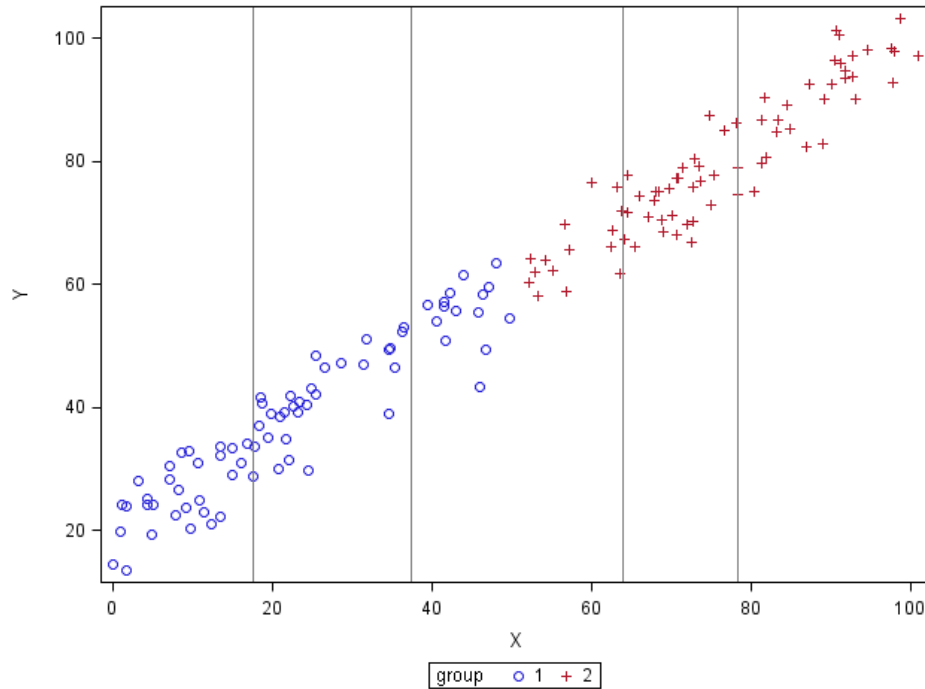
Coefficients for group Least Square Means

Effect	group Level	
	1	2
Intercept	1	1
Quintile_X 0	0.2	0.2
Quintile_X 1	0.2	0.2
Quintile_X 2	0.2	0.2
Quintile_X 3	0.2	0.2
Quintile_X 4	0.2	0.2
group 1	1	0
group 2	0	1

group	Y LSMEAN	Standard Error	Pr >  t
1	53.7618809	1.2341619	<.0001
2	63.8778683	1.2341619	<.0001

16

Where are the LSmeans?



17

LSmeans use quintile means plus difference in group means derived *only from strata with subjects from both groups*.

What happens if we make the center stratum wider?

narrower?

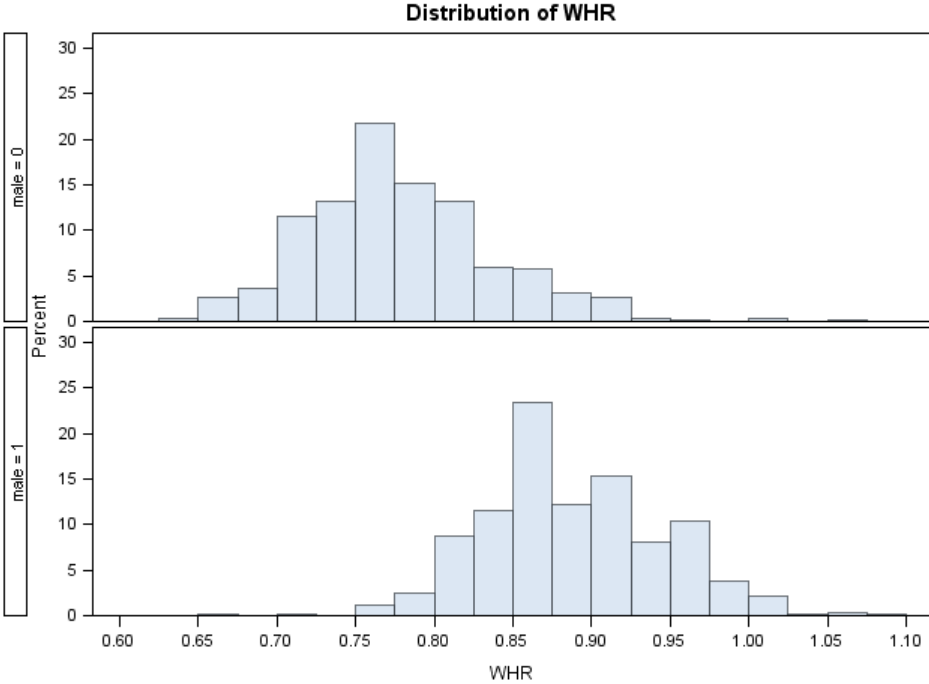
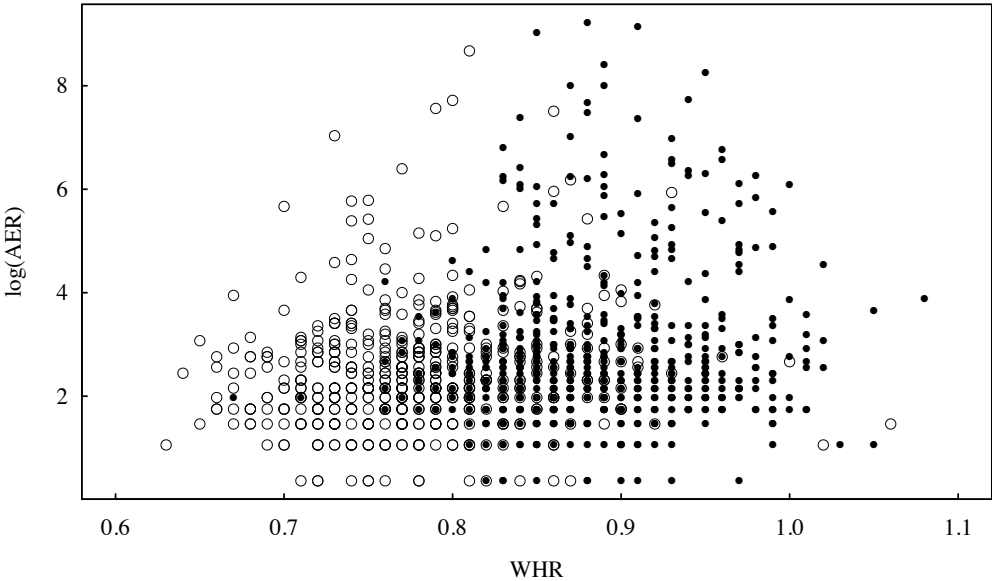
Adjusted means depend on selection of strata.

No indication of this in the SAS output from the main-effect model.

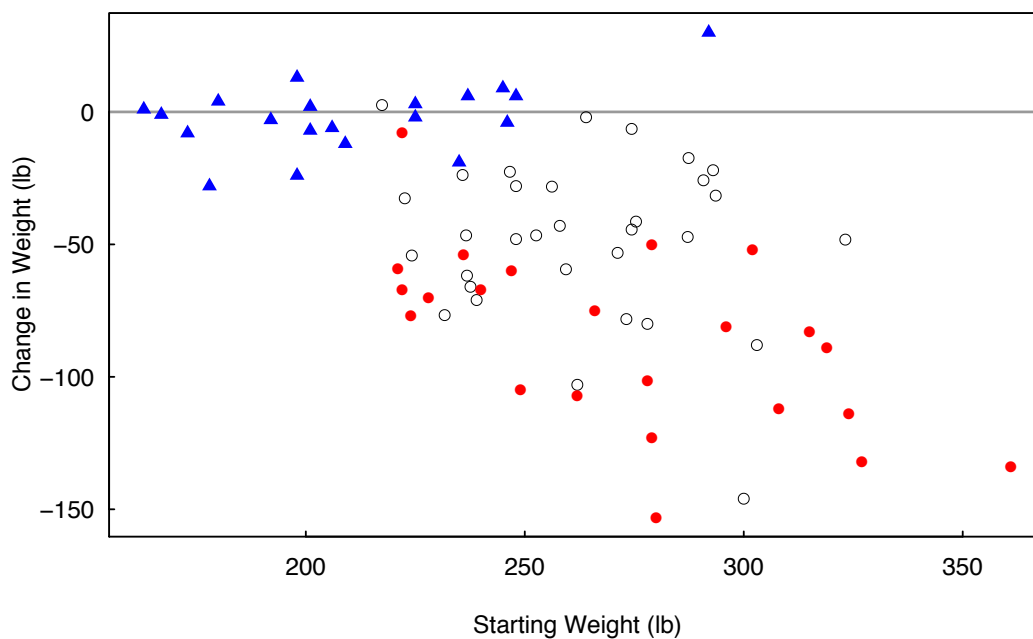
18

Complete separation is unusual, but imbalance from limited overlap is not.

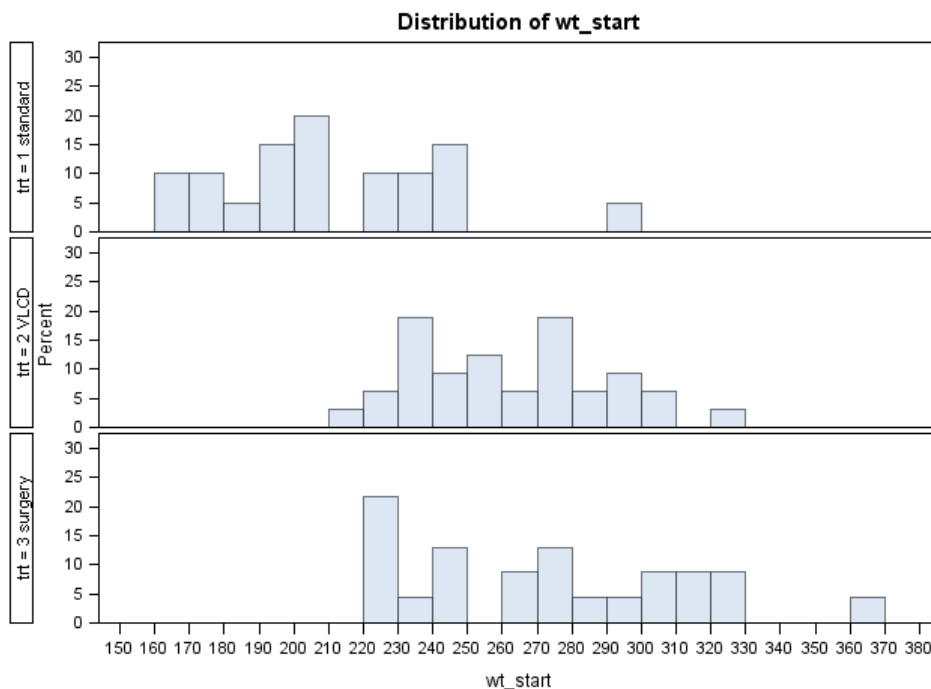
Compare albumin excretion rate (AER) between genders, adjusting for waist-hip ratio (WHR).



Treatments: standard medical care ( $\blacktriangle$ ), very-low-calorie liquid diet ( $\circ$ ), or gastric-bypass surgery ( $\bullet$ ).



21



This imbalance can occur in several covariates simultaneously,

22

Imbalance in an adjusting covariate affects comparison of study groups.

Estimated group differences are based on the limited portion of data in overlap between groups. This data may not be typical of the rest.

⇒ estimated differences may be biased

Estimated differences may depend on the arbitrary boundaries of strata:

⇒ estimated differences may be unreliable

### Longitudinal data

When people or experimental units are measured more than once over time, we have *longitudinal data*, also called *repeated measures* or *time series* data.

**Family economics data:** total family income, expenditures, debt status for 50 families in two groups (A and B), annual records from 1990–1995.

Records for family 1. One observation for each year.

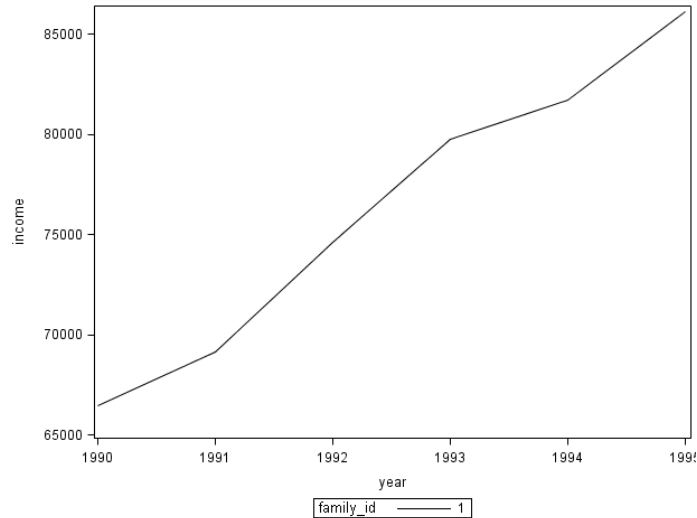
Obs	family_ id	income	year	expenses	debt	group
1	1	66483	1990	49804	no	A
2	1	69146	1991	65634	no	A
3	1	74643	1992	61820	no	A
4	1	79783	1993	68387	no	A
5	1	81710	1994	85504	yes	A
6	1	86143	1995	75640	no	A

## Plotting longitudinal data

Want to plot the income against year for each family:

x = year y = income      need year and income as variables.

Family 1.



25

Obs	family_ id	income	year	expenses	debt	group
1	1	66483	1990	49804	no	A
2	1	69146	1991	65634	no	A
3	1	74643	1992	61820	no	A
4	1	79783	1993	68387	no	A
5	1	81710	1994	85504	yes	A
6	1	86143	1995	75640	no	A
7	2	17510	1990	21609	yes	B
8	2	19484	1992	18180	no	B
9	2	20979	1993	22985	yes	B
10	2	21268	1994	11097	no	B
11	2	22998	1995	21768	no	B

```
Proc SGplot data=a;
```

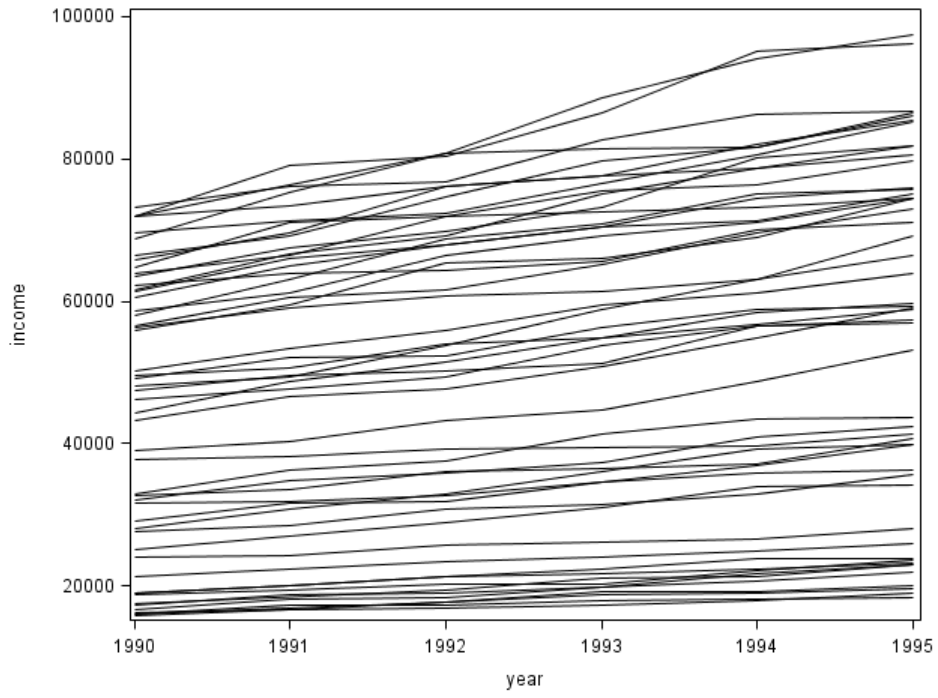
```
  series x=year y=income / group =family_id
```

```
  LineAttrs= (pattern=1 color="black");
```

series – draws a line connecting sequential observations

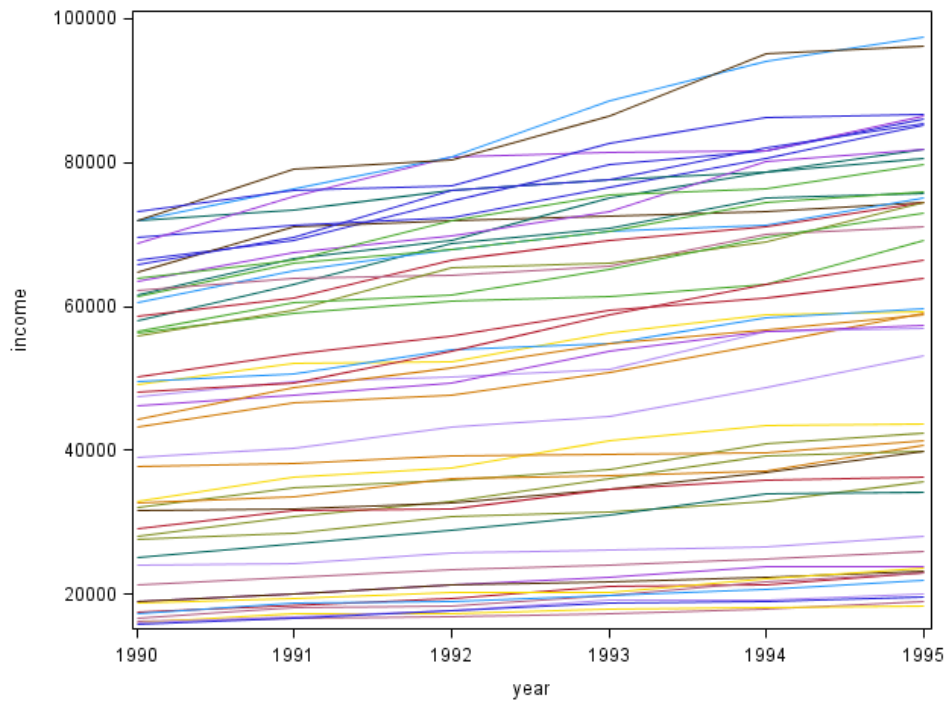
LineAttrs – draw solid, black lines

26

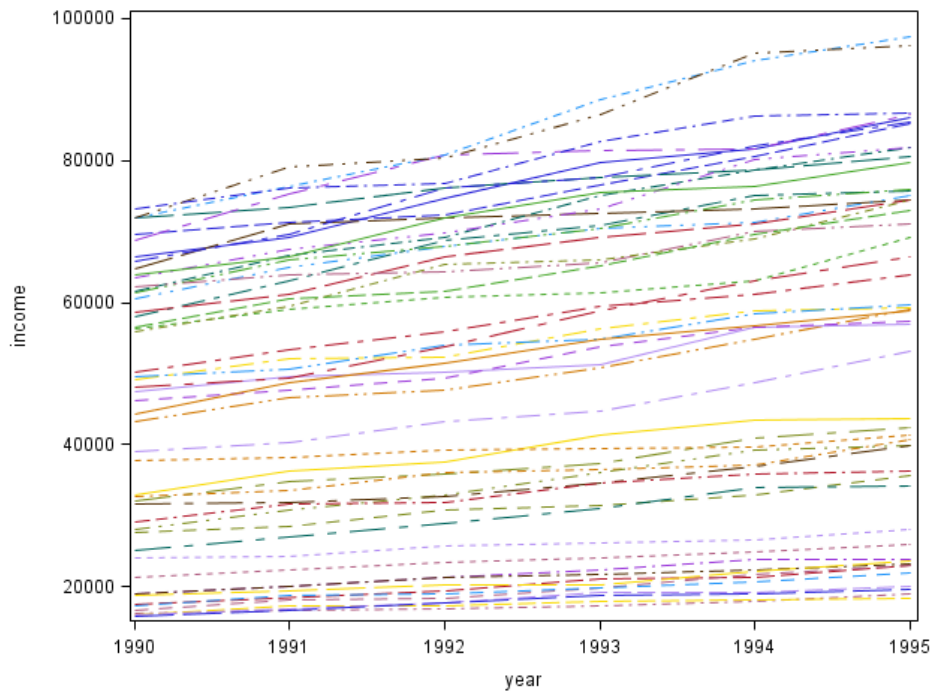


Often called “spaghetti plot.”

Without specifying `color="black"`:



Without specifying (pattern=1 color="black"):



29

Families are in 2 groups, *A* or *B*.

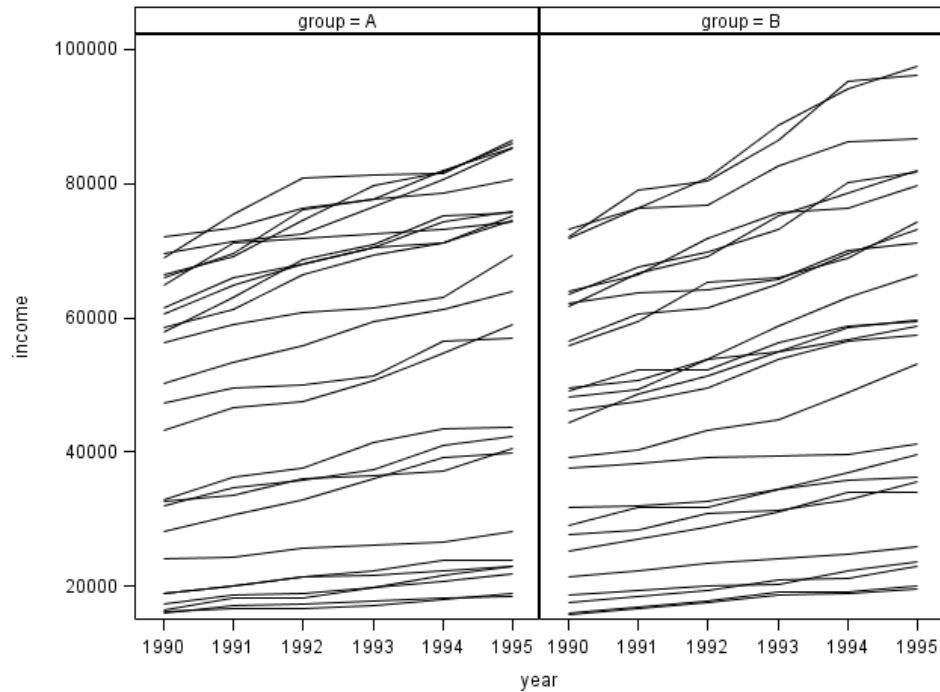
To get separate plots for each group, we must group by family id within each group.

SGplot allows only one grouping variable.

```
Proc SGpanel data=a; produces multiple plots on one page  
  PanelBy group / columns=2;  
  series x=year y=income / group =family_id  
  LineAttrs= (pattern=1 color="black");
```

30

SGpanel plots by group:



31

### Correlation within subjects

**Alzheimer's disease trial.** Alzheimer's disease is a progressive incurable deterioration of intellect and memory. A clinical trial compared lecithin (dietary supplement) against placebo, both given as daily for 4 months; 22 patients in lecithin group, 25 in placebo group.

Participant took a memory test at baseline (first visit), and end of each month. Score is number of words recalled from a list, so higher scores are better.

idno	lecithin	score1	score2	score3	score4	score5
1	0	20	15	14	13	13
2	0	14	12	12	10	10
3	0	7	5	5	6	5
4	0	6	10	9	8	7

(Source: Der and Everitt, Ch. 11)

32

Plot individual profiles and correlation scatterplot matrix.

```
Proc SGpanel data=alz_long;  
  PanelBy lecithin / columns=2;  
  series x=visit y=score / group=idno  
  LINEATTRS= (pattern=1 color="black");
```

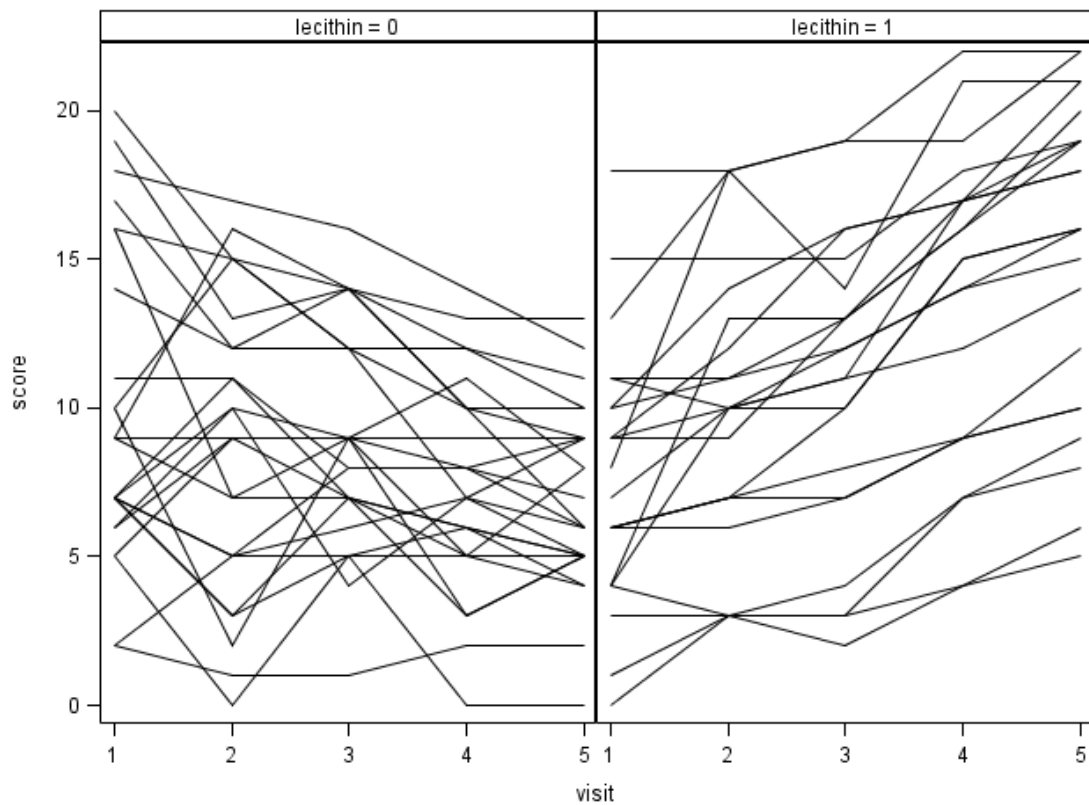
ODS graphics on;

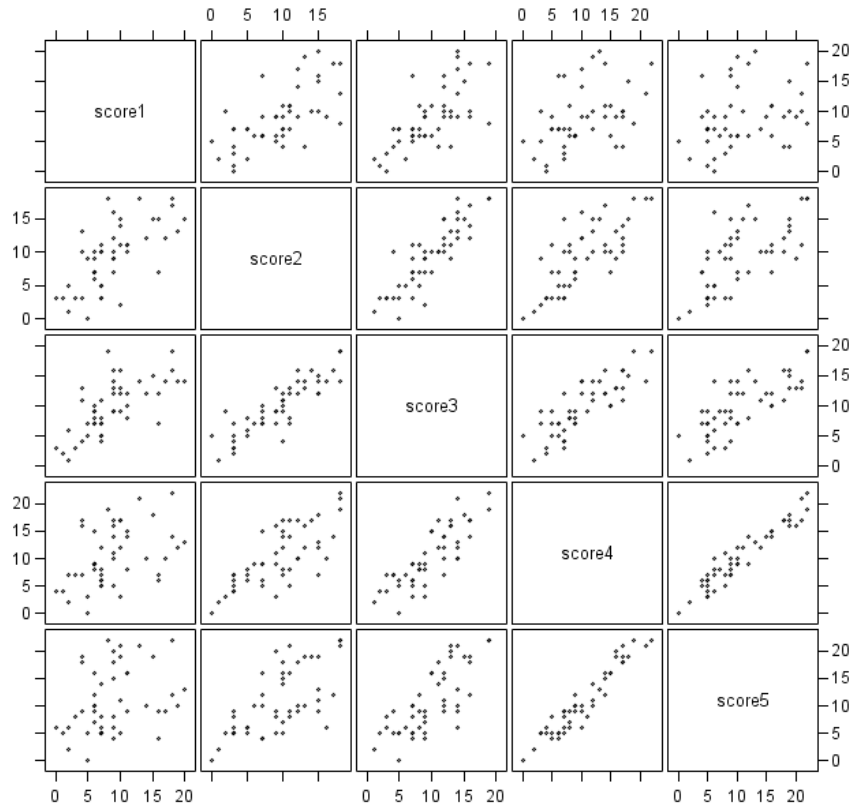
```
Proc Corr data=alzheimer plots=matrix;
```

```
  var score1-score5;
```

```
  run;
```

ODS graphics off;





35

Decrease in correlation over longer time intervals:

Pearson Correlation Coefficients, N = 47  
 Prob > |r| under H0: Rho=0

	score1	score2	score3	score4	score5
score1	1.00000	0.66267 <.0001	0.67951 <.0001	0.42892 0.0026	0.30906 0.0345
score2	0.66267 <.0001	1.00000	0.86712 <.0001	0.75344 <.0001	0.66498 <.0001
score3	0.67951 <.0001	0.86712 <.0001	1.00000	0.82909 <.0001	0.76285 <.0001
score4	0.42892 0.0026	0.75344 <.0001	0.82909 <.0001	1.00000	0.95437 <.0001
score5	0.30906 0.0345	0.66498 <.0001	0.76285 <.0001	0.95437 <.0001	1.00000

36

## Consequence of within-subject correlation

Repeated longitudinal observations from the same subject are correlated = **within-subject observations are not independent.**

ANOVA and regression assume independent observations, hence don't apply correctly to correlated data.

Model for longitudinal observations must include within-subject correlation.

Greater within-subject correlation  $\implies$  smaller within-subject sample.

What if within-subject correlation = 1?

37

## Longitudinal data: Response Feature Analysis

**Response feature analysis** replaces repeated measurements with one outcome: no more longitudinal data, apply simpler analysis method: ANOVA, regression, *t*-test. Common response features:

- mean
- area under the curve (AUC)
- for peaked data: maximum or minimum value
- for peaked data: *time* to maximum or minimum value
- for growth data: slope of regression line

More than one feature can be used, with multiple analyses to compare groups.

38

## Visual Analog Scale (VAS) example

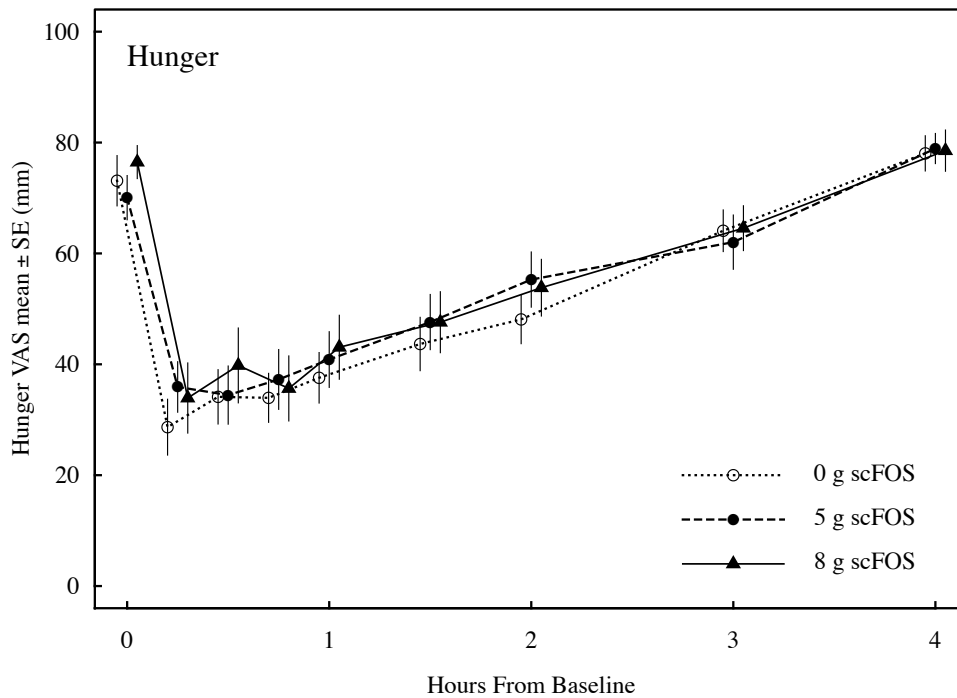
A nutrition study compared the immediate effect on feelings of hunger after a breakfast muffin containing 0, 5, or 8 g of short-chain fructooligosaccharides (scFOS).

To measure hunger, participants marked a visual analog scale (VAS) to indicate how hungry they felt:



Distance from zero on scale was numeric response. Participants completed the VAS at 0, 15, 30, 45, 60, 90, 120, 180, and 240 minutes after eating the muffin.

39

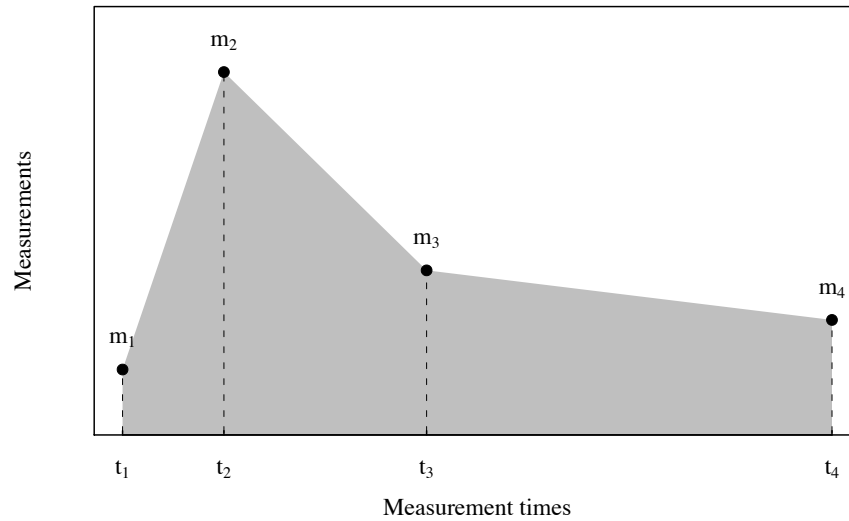


Mean curves in response to each treatment. Differences?

40

## Finding the area under a curve: trapezoid rule

Sequence of individual's measurements  $m_i$ , taken at times  $t_i$

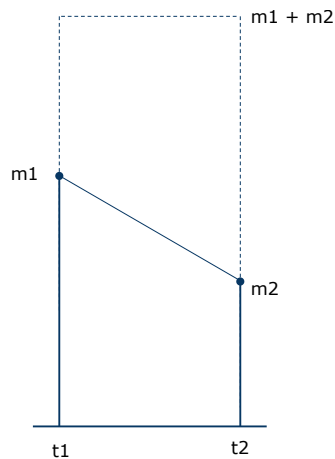


**Trapezoid rule:** connect measurements with line segments, find area below in gray.

41

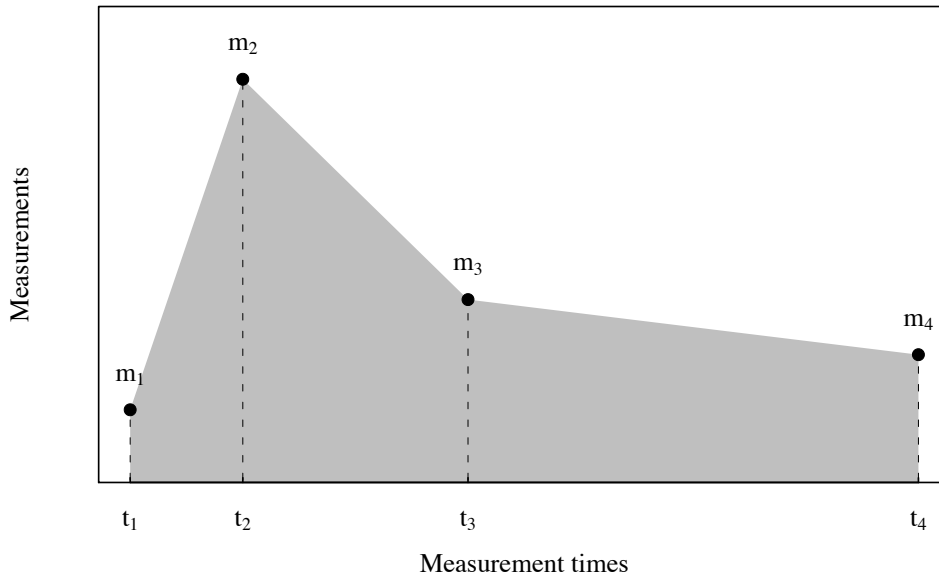
Trapezoid: 4-sided plane figure with 2 parallel sides.

Duplicate trapezoid on top gives rectangle that has twice the area.



$$\text{Trapezoid area} = \frac{1}{2} \left\{ (t_2 - t_1)(m_1 + m_2) \right\}$$

42



$$\text{Area under the curve} = \frac{1}{2} \left\{ (t_2 - t_1)(m_1 + m_2) + (t_3 - t_2)(m_2 + m_3) + (t_4 - t_3)(m_3 + m_4) \right\}$$

Approximates area under true curve of measured quantity  $m$ .

43

### Calculating AUC (Area Under Curve)

In example, VAS hunger measured 9 times: at 0, 15, 30, 45, 60, 90, 120, 180, and 240 minutes after eating the muffin.

Convert times to hours:  $t_i = 0, .25, .5, .75, 1, 1.5, 2, 3, 4$ .

$$\text{AUC} = \frac{1}{2} \left\{ (t_2 - t_1)(m_1 + m_2) + (t_3 - t_2)(m_2 + m_3) + \dots + (t_9 - t_8)(m_8 + m_9) \right\}$$

How many trapezoids?

Use one array for times, one for measurements.

44

```

data AUC;
  set VAS_hunger;
  array m[9] hunger1-hunger9;
  array t[9] time1-time9;
  time1=0; time2=0.25; time3=0.5; time4=0.75; time5=1;
  time6=1.5; time7=2; time8=3; time9=4;
  AUC = 0;
  do j=1 to 8; why 8 instead of 9?
    next_trapezoid = 0.5 * (t[j+1] - t[j])*(m[j] + m[j+1]);
    AUC = sum(AUC, next_trapezoid);
  end;

```

How should we adapt this to find maximum hunger score?

45

### What if some observations are missing?

Suppose a subject is missing VAS hunger measurement  $m_2$  at time 0.25 hours.

What happens to the AUC calculation in SAS?

$$AUC = \frac{1}{2} \left\{ (t_2 - t_1)(m_1 + m_2) + (t_3 - t_2)(m_2 + m_3) + (t_4 - t_3)(m_3 + m_4) + \cdots + (t_9 - t_8)(m_8 + m_9) \right\}$$

46

Write code to alert you to problems: write observations with missing data to a separate data set.

To create 2 datasets, give 2 names, and separate output statements.

```
data hunger_AUC missing; create 2 data sets
  set VAS_hunger;
  array m[9] hunger1-hunger9;
  array t[9] time1-time9;
  time1=0; time2=0.25; time3=0.5; time4=0.75; time5=1;
  time6=1.5; time7=2; time8=3; time9=4;
```

47

```
AUC = 0;
do j=1 to 8;
  next_trapezoid = 0.5 * (t[j+1] - t[j])*(m[j] + m[j+1]);
  if (next_trapezoid = .) then do; deal with a missing value
    output missing;
    GOTO Duluth; jump to label 'Duluth'
  end;
  AUC= sum(AUC,next_trapezoid);
end;
output hunger_AUC;
Duluth: SAS label ends with full colon, not semicolon
```

48

```
proc print data=missing;
```

Obs	subject	hunger1	hunger2	hunger3	hunger4	hunger5	...
1	1	81	.	10	15	37	...

Common practice: replace missing  $k$ -th value at  $t_k$  by linear interpolation from measurements  $m_{k-1}$ ,  $m_{k+1}$  on either side.

Solve for  $x$ :

$$\frac{m_{k-1} - x}{m_{k-1} - m_{k+1}} = \frac{t_{k-1} - t_k}{t_{k-1} - t_{k+1}}$$

Use imputation for missing measurements at the ends.