Lecture 15

1. Longitudinal data: plots of individual profiles and mean
2. Within-subject correlation
3. Response-feature analysis
4. Summary slopes
5. Area under the curve (AUC)

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 cohorts (A and B), annual records from 1990–1995.

Records for family 1. One observation for each year = *long form*.

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

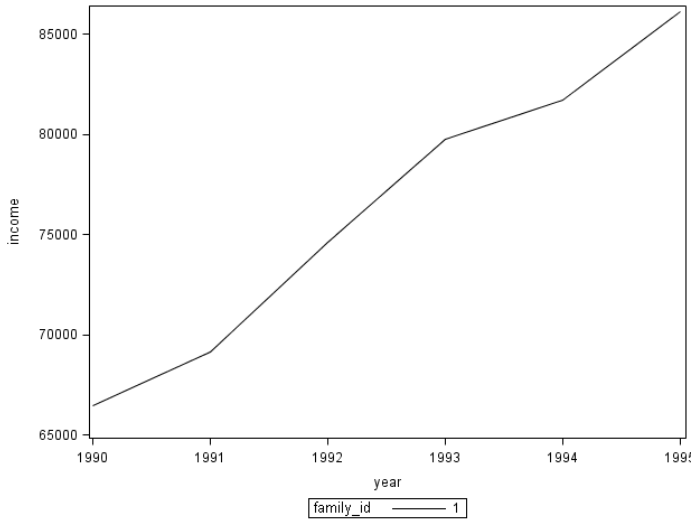
(Example data adapted from UCLA Academic Technology Services, `www.ats.ucla.edu/stat/`)

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.

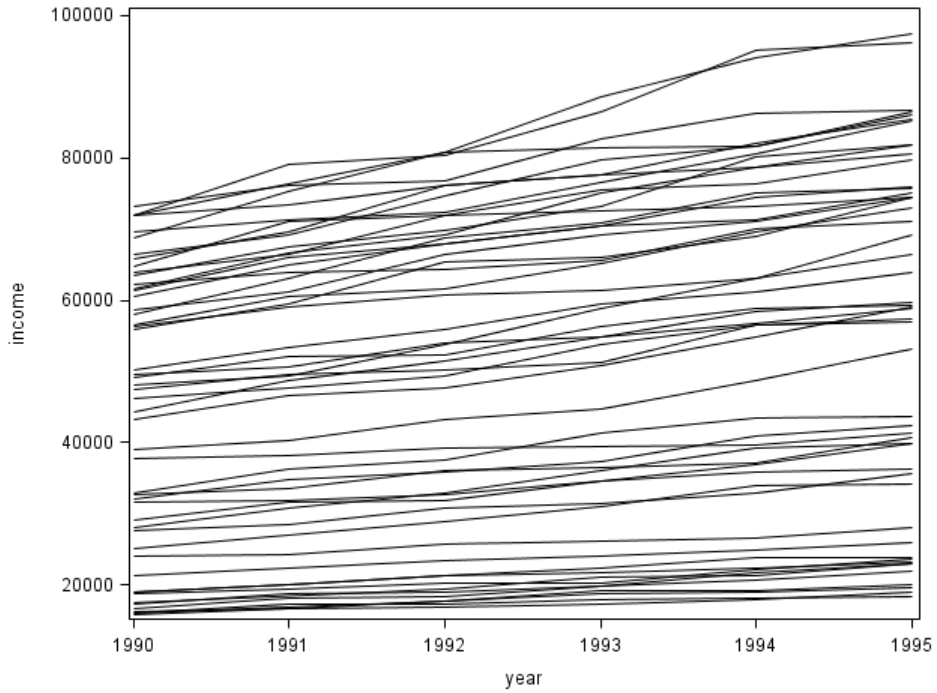


Obs	family_ id	income	year	expenses	debt	group
1	1	66483	1990	49804	no	A
2	1	69146	1991	65634	no	A
3	1	74643	1992	61820	no	A
4	1	79783	1993	68387	no	A
5	1	81710	1994	85504	yes	A
6	1	86143	1995	75640	no	A
7	2	17510	1990	21609	yes	B
8	2	19484	1992	18180	no	B
9	2	20979	1993	22985	yes	B
10	2	21268	1994	11097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

```
Proc SGplot  data = econ_longform;
   series   x=year y=income / group =family_id
          LineAttrs= (pattern=1 color="black");
```

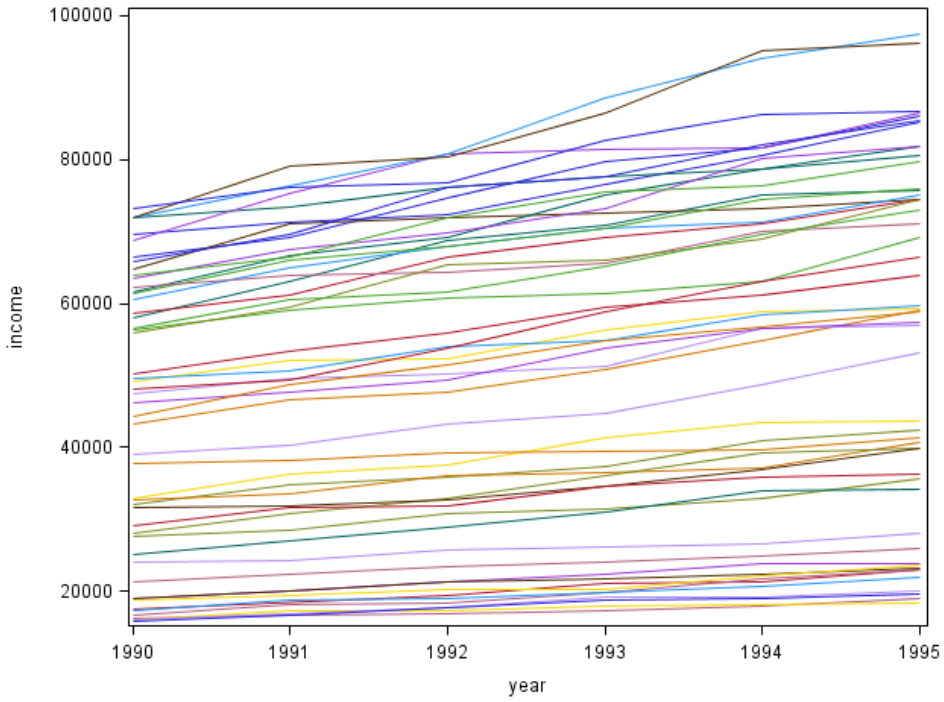
series – draws a line connecting sequential observations

LineAttrs – draw solid, black lines

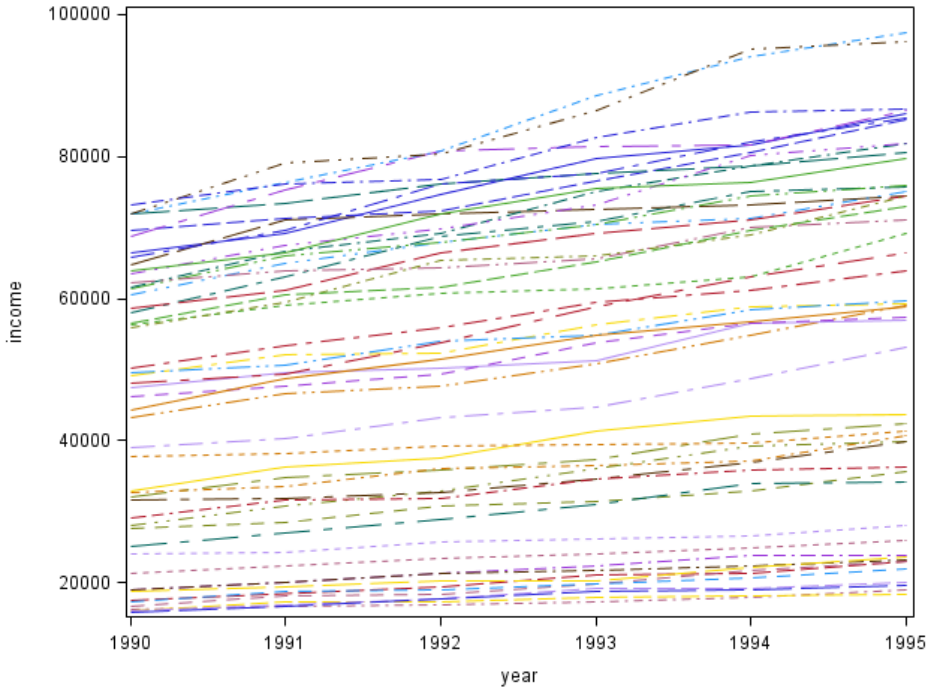


Often called “spaghetti plot.”

Without specifying color="black":



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



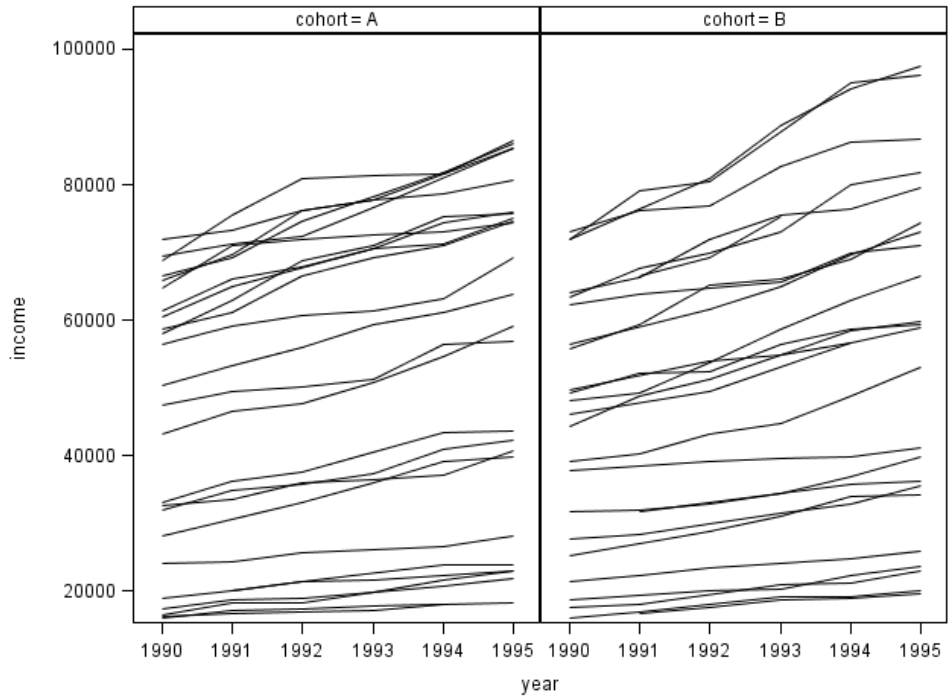
Families are in 2 cohorts, *A* or *B*.

Separate plots (panels) for each cohort, group by family id within each cohort.

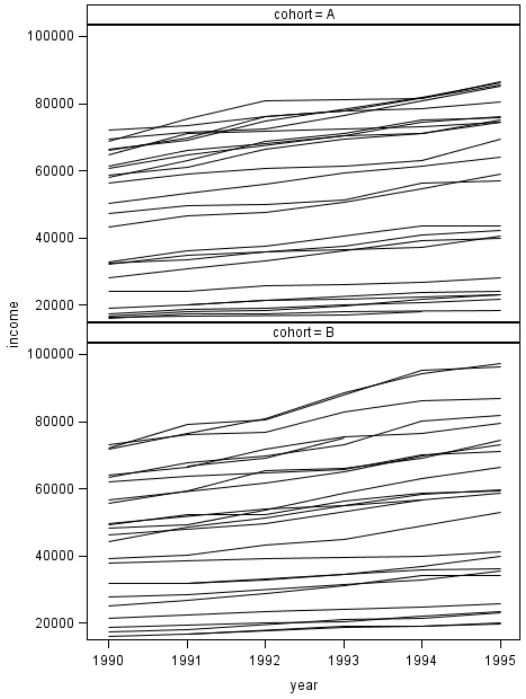
SGplot allows only one grouping variable.

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

SGpanel plots by group in columns:



SGpanel plots by group in rows: `PanelBy group / rows=2;`



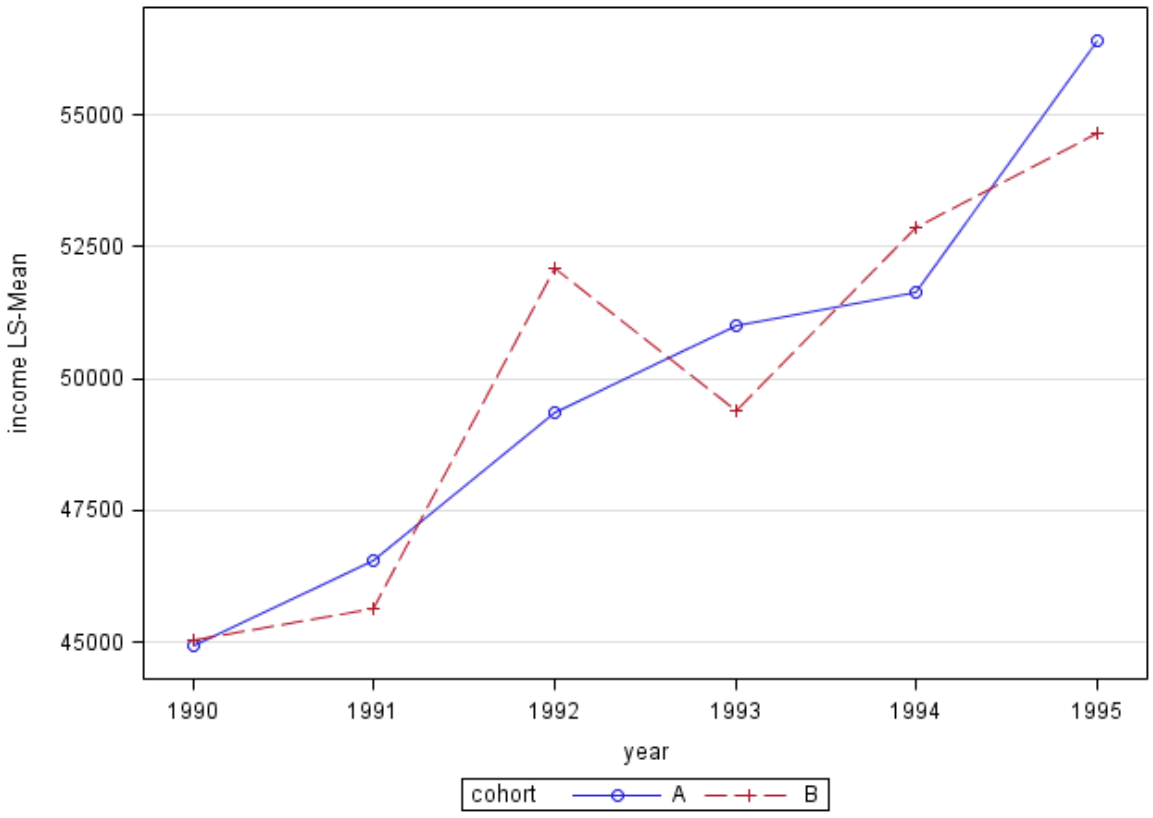
Plotting means over time

For longitudinal data, a plot of means over time is an interaction plot:

`group * time`

```
ODS graphics on;
Proc Glimmix data=econ_longform;
  class  year cohort;
  model income =year cohort year*cohort;
  lsmeans year*cohort
     / plots=(meanplot( join sliceby=cohort)); no CI or SE bar yet
run;
ODS graphics off;
```

No SEs yet, because we must include within-family correlation across years.

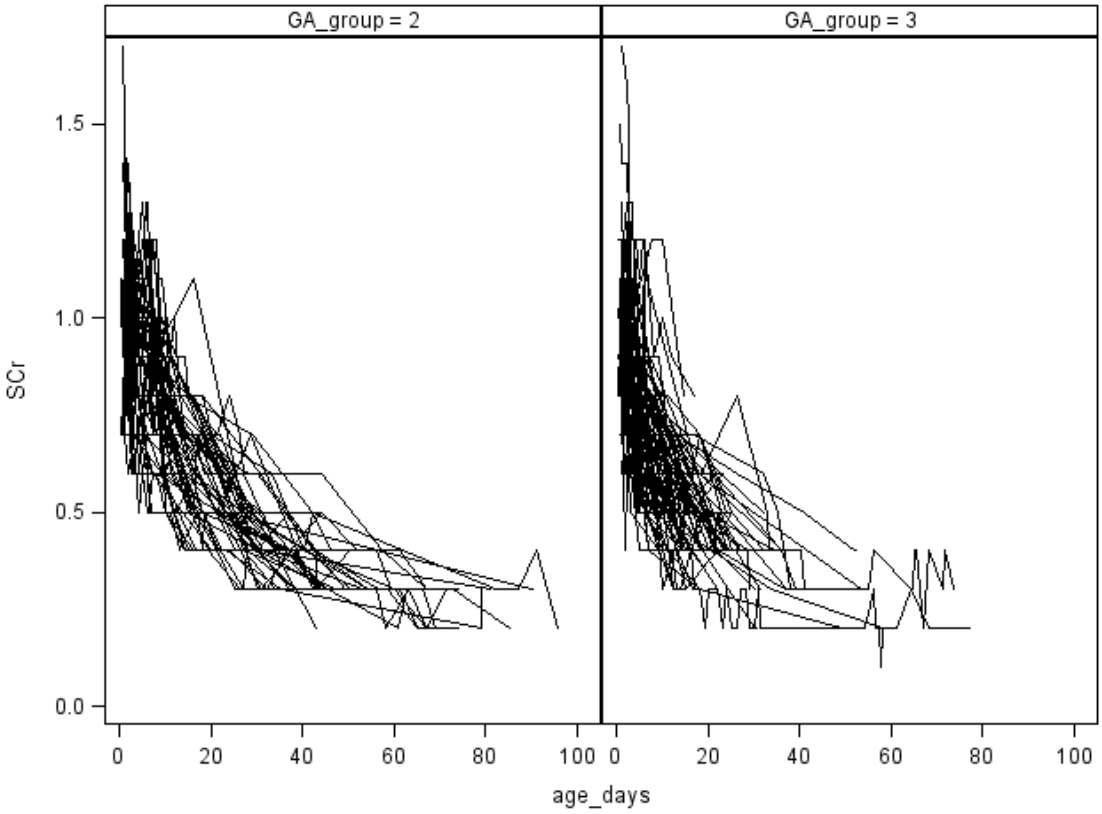


Example 2: Serum creatinine (SCr) data from infants (HW 3).

Researchers recorded serum creatinine, a measurement of kidney function, in newborn infants during the first months of life. Each infant was measured several times.

```
Proc SGpanel data=ph6470.infant_SCr;
  PanelBy GA_group / rows=2;
  series x=age_days y=SCr / group=id
     lineattrs= (pattern=1 color="black");
```

13



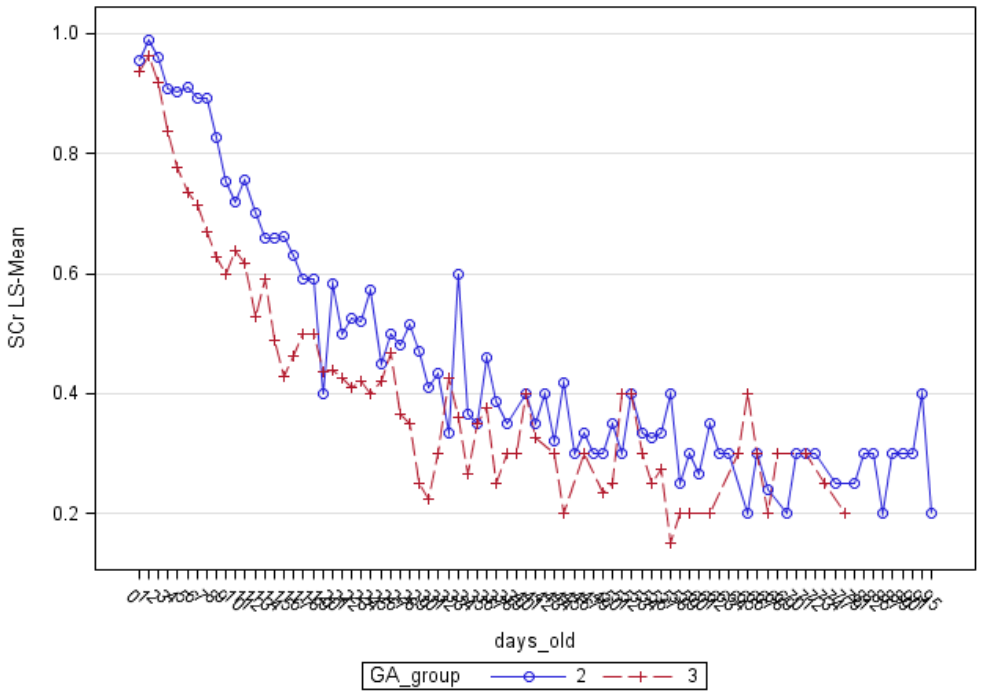
14

Next, plot means for infant SCr. Means by what time unit?

Obs	GA_group	id	SCr	age_days
1	3	1	0.8	0.9688
2	3	1	0.9	2.1958
3	3	1	0.7	3.1736
4	3	1	0.7	4.1875
5	3	1	0.6	6.1479
6	3	1	0.5	9.2431
7	3	1	0.4	14.5417
8	3	1	0.4	21.6160
9	3	1	0.3	30.2396
10	2	2	0.7	0.8194
11	2	2	0.8	0.8785
12	2	2	0.8	0.9583
13	2	2	0.8	1.1944
14	2	2	0.9	1.3708
15	2	2	0.8	2.1458
16	2	2	0.8	3.2708
17	2	2	0.8	4.2222
18	2	2	0.8	5.2049
19	2	2	0.7	11.2500
20	2	2	0.6	14.2361


```
data daily_SCr;
  set ph6470.infant_SCr;
  days_old = floor (age_days);   round down to integer nearer zero

ODS graphics on;
proc glimmix data=daily_SCr;
 class days_old GA_group;
 model SCr = days_old GA_group days_old*GA_group;
 lsmeans days_old*GA_group
    / plots=(meanplot( join sliceby=GA_group));
run;
ODS graphics off;
```

Again, no SE yet. We'll add SE bars later when we know how to calculate them.



Better choice of time unit?

Example 3: 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.

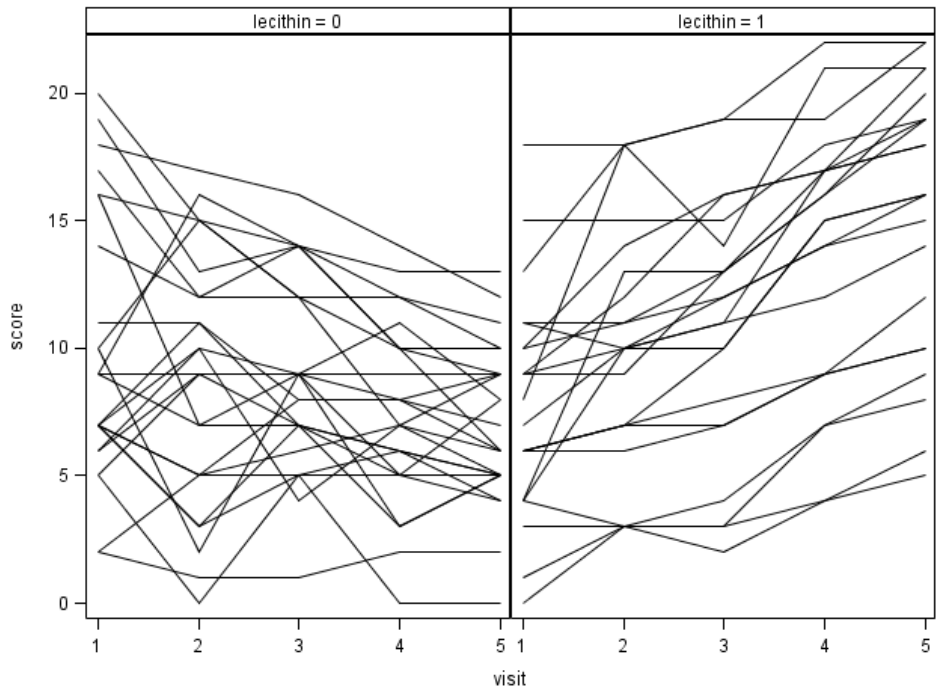
Participant took a memory test at baseline (first visit), and end of each month. Score is number of words recalled from a list, so higher scores are better.

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)

Plot individual profiles and means by treatment group.

```
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 glimmix data=ph6470.alz_long;
  class  visit lecithin;
  model score =visit lecithin visit*lecithin;
  lsmeans visit*lecithin / plots=(meanplot( join sliceby=lecithin));
 run;
 ODS graphics off;
```

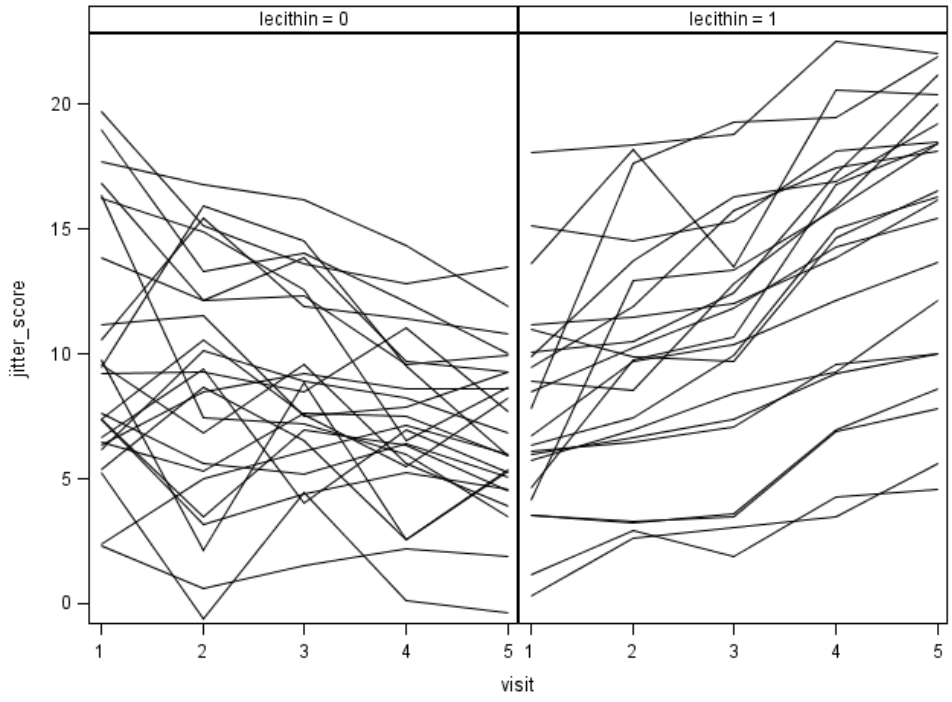
Scores are integers. Clarify plot by adding small random noise to scores.

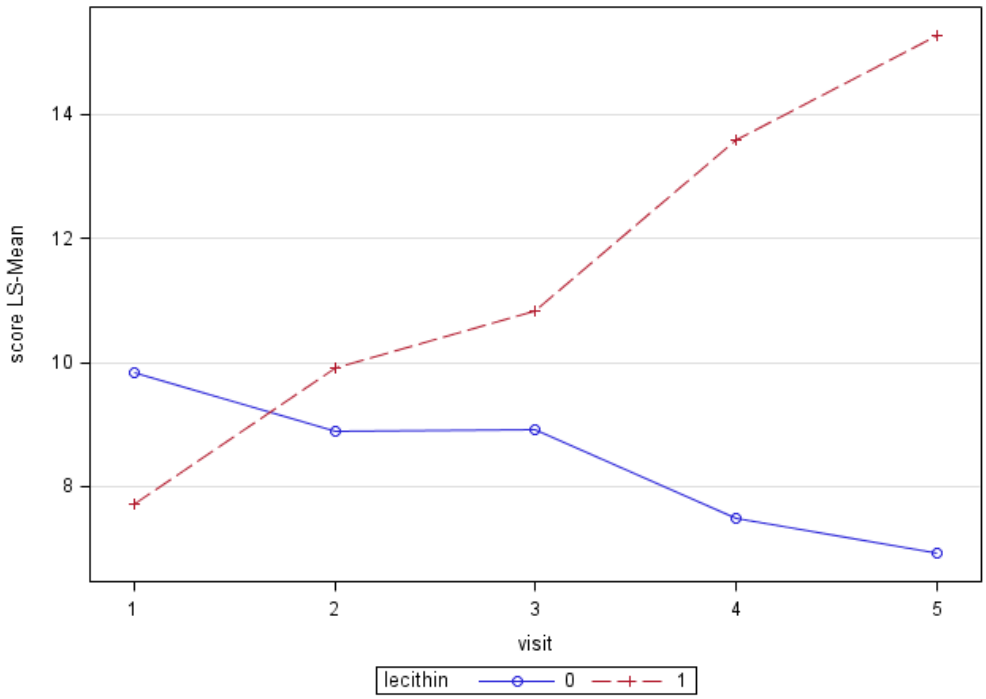
Jittering to clarify coincident points

When profiles coincide, add vertical or horizontal noise to spread out points. This can help display the data more clearly.

```
data Alz_jitter;
   set ph6470.alz_long;
   jitter_score = score + 1.25*(ranuni(6495521) - .5);
Proc SGpanel data=Alz_jitter;
   PanelBy lecithin / columns=2;
   series x=visit y=jitter_score
      / group=idno lineattrs= (pattern=1 color="black");
```

ranuni SAS function that generates pseudo-random numbers with uniform distribution on $[0,1]$





Longituinal data: correlation within subjects

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

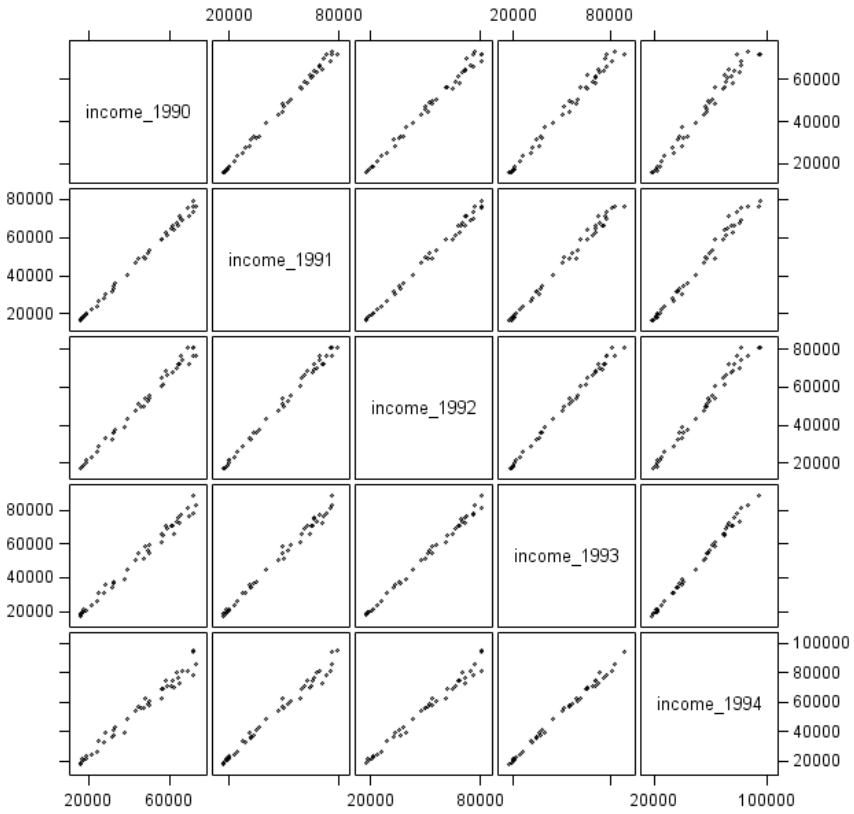
Examine correlation between observations from the same subject using Proc Corr.

Need to have data in *wide form*.

Example 1: Family economic data, incomes 1990–1995.

First, make wide-form data.

```
Proc Transpose data=econ_longform out=econ_wideform prefix=income_;
  ID year;
  VAR income;
  BY family_id cohort;

ODS graphics on;
Proc Corr data=econ_wideform  plots=matrix;
  var income_1990-income_1995;
run;
ODS graphics off;
```

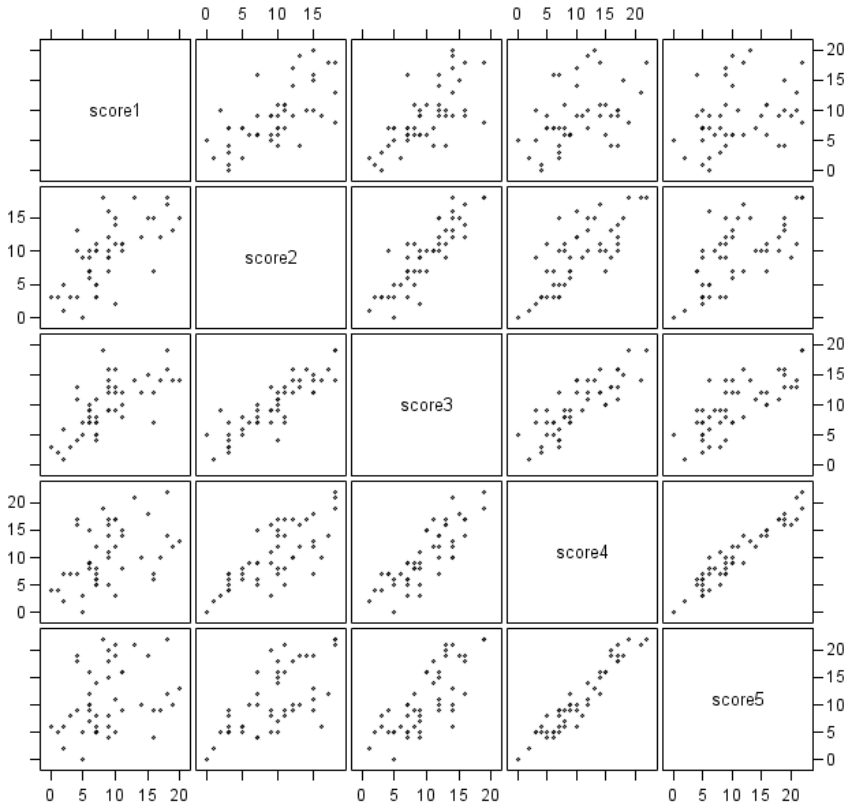



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	income_ 1990	income_ 1991	income_ 1992	income_ 1993	income_ 1994	income_ 1995
income_1990	1.00000	0.99817	0.99506	0.99344	0.98895	0.98883
		<.0001	<.0001	<.0001	<.0001	<.0001
	46	42	40	41	43	44
income_1991	0.99817	1.00000	0.99735	0.99604	0.99282	0.99194
	<.0001		<.0001	<.0001	<.0001	<.0001
	42	46	39	42	42	44
income_1992	0.99506	0.99735	1.00000	0.99739	0.99195	0.99331
	<.0001	<.0001		<.0001	<.0001	<.0001
	40	39	43	38	39	41
income_1993	0.99344	0.99604	0.99739	1.00000	0.99766	0.99674
	<.0001	<.0001	<.0001		<.0001	<.0001
	41	42	38	45	41	43
income_1994	0.98895	0.99282	0.99195	0.99766	1.00000	0.99817
	<.0001	<.0001	<.0001	<.0001		<.0001
	43	42	39	41	46	44
income_1995	0.98883	0.99194	0.99331	0.99674	0.99817	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	
	44	44	41	43	44	47

Example 3: Alzheimer's disease trial

```
ODS graphics on;
Proc Corr data=alzheimer_wide  plots=matrix;
   var score1-score5;
 run;
ODS graphics off;
```



Decrease in correlation over longer time intervals, but increase in correlation over trial:

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

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

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 $\Longrightarrow$ "smaller" sample from each subject.

What if within-subject correlation = 1?

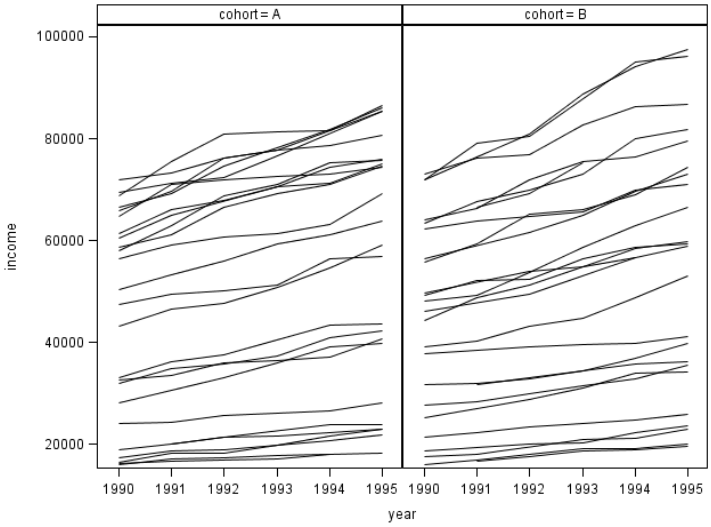
Longitudinal data: Response Feature Analysis

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

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

Individual slopes: family economic data



Summarize each family by its linear regression slope.

Interpretation?

Calculate linear regression of income on year, for each family separately.

```
proc sort data=econ_longform;
  by cohort family_id;

Proc GLM data=econ_longform;
  by cohort family_id;   both cohort and id kept in output data
  model income = year / solution ;  needed for reg coef output
  ODS output ParameterEstimates=income_slopes;

proc print data=income_slopes(obs=10);
```

Obs	cohort	family_ id	Dependent	Parameter	Estimate	StdErr	tValue	Probt
1	A	1	income	Intercept	-7857717.95	336491.138	-23.35	0.0002
2	A	1	income	year	3981.80	168.887	23.58	0.0002
3	A	3	income	Intercept	-7207855.05	814131.811	-8.85	0.0009
4	A	3	income	year	3651.91	408.598	8.94	0.0009
5	A	4	income	Intercept	-3028801.95	894816.500	-3.38	0.0277
6	A	4	income	year	1555.89	449.092	3.46	0.0257

Keep only the slopes:

```
data income_slopes1;
  set income_slopes;
  if parameter = "year";

 proc print data=income_slopes1(obs=5);
```


Obs	cohort	family_ id	Dependent	Parameter	Estimate	StdErr	tValue	Probt
1	A	1	income	year	3981.80	168.887	23.58	0.0002
2	A	3	income	year	3651.91	408.598	8.94	0.0009
3	A	4	income	year	1555.89	449.092	3.46	0.0257
4	A	5	income	year	1073.72	103.466	10.38	0.0019
5	A	6	income	year	2043.29	152.283	13.42	0.0002

Compare slopes (mean annual change in family income) between cohorts:

```
 Proc GLM data=income_slopes1;
   class cohort;
   model estimate = cohort;
   lsmeans cohort / stderr pdiff;
```

cohort	Estimate LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1= LSMean2 Pr > |t|
A	2142.53985	248.02041	<.0001	0.3913
B	2445.96258	248.02041	<.0001	

Annual increase in family income averaged \$2140 ± 250 in cohort A (mean ± SE), and \$2450 ± 250 in cohort B; the difference between cohorts was not significant.

Calculating 50 regressions produces 100 pages of output. However `noprint` option doesn't help.

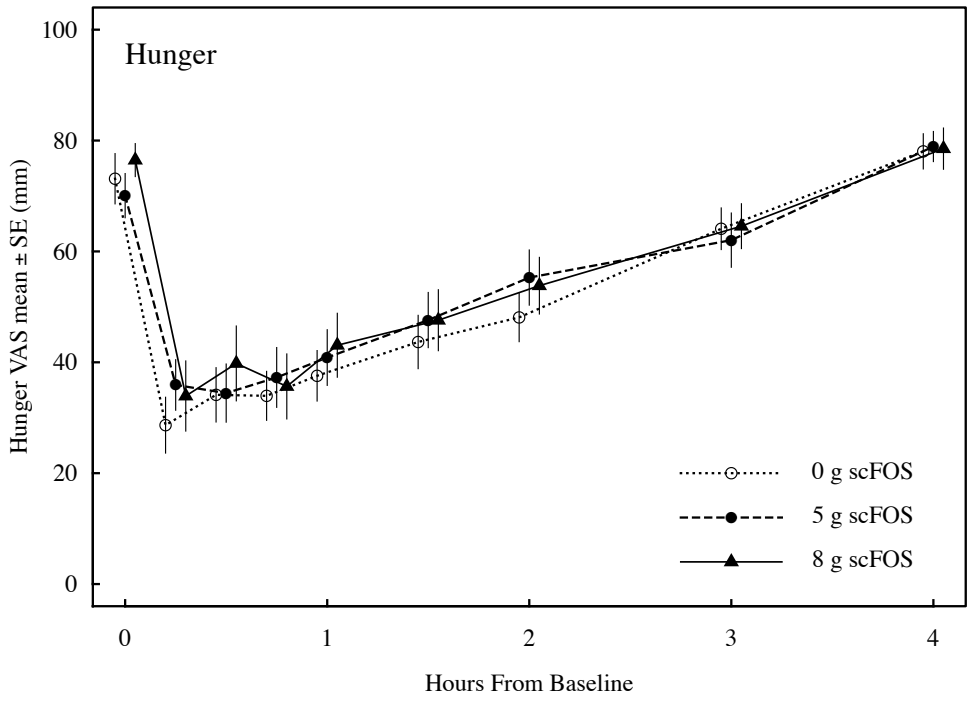
```
2312   Proc GLM data=family_econ noprint;
2313     by cohort family_id;
2314     model income = year / solution;
2315     ODS output ParameterEstimates=income_slopes;
2316

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds

WARNING: Output 'ParameterEstimates' was not created.  Make sure that the
         output object name, label, or path is spelled correctly.  Also, verify
         that the appropriate procedure options are used to produce the
         requested output object.  For example, verify that the NOPRINT option
         is not used.
```

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:

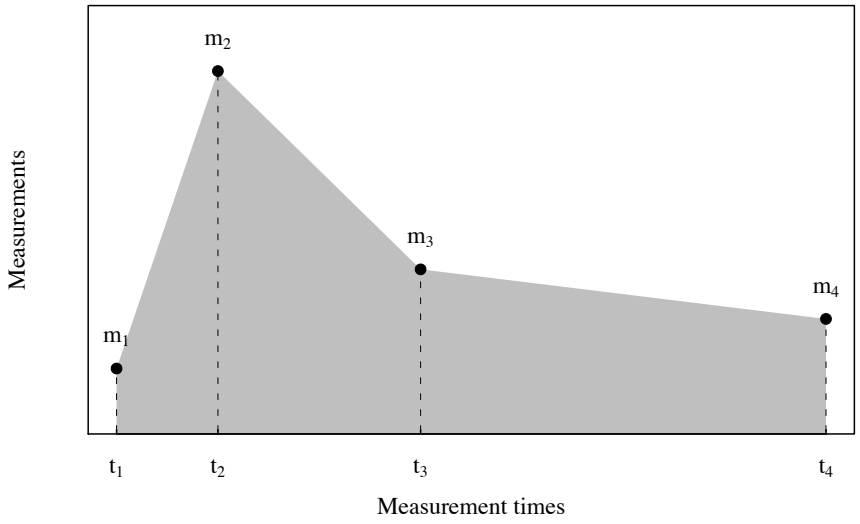
Distance from zero on scale was numeric response. Participants completed the VAS at 0, 15, 30, 45, 60, 90, 120, 180, and 240 minutes after eating the muffin.



Mean curves in response to each treatment. Differences?

Finding the area under a curve: trapezoid rule

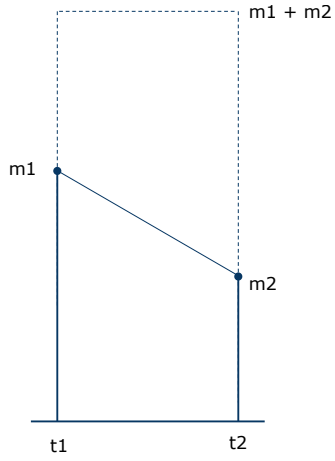
Sequence of individual's measurements m_i, taken at times t_i



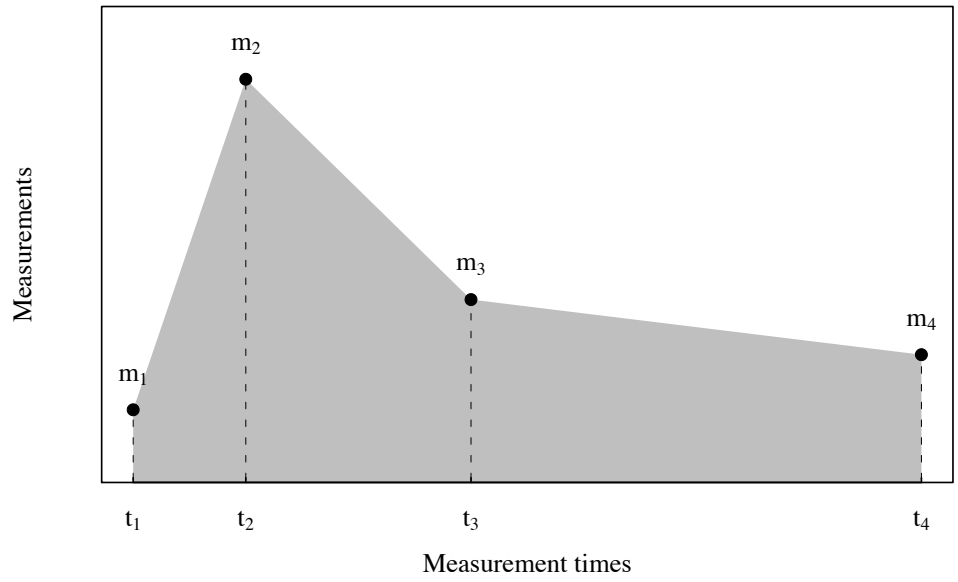
Trapezoid rule: connect measurements with line segments, find area below in gray.

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

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



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



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

Approximates area under true curve of measured quantity m.

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.

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

How many trapezoids?

Use one array for times, one for measurements.

```
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?

Suppose a subject is missing VAS hunger measurement m_2 at time 0.25 hours. What happens to the AUC calculation in SAS?

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;
```

```
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
```



```
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.