Multivariate analysis of variance

- Standard ANOVA has just one response variable.
- What if you have more than one response?
- Try an ANOVA on each response separately.
- But might miss some kinds of interesting dependence between the responses that distinguish the groups.
- SAS can run MANOVA using an option on PROC GLM.

Small example

- Measure yield and seed weight of plants grown under 2 conditions: low and high amounts of fertilizer.
- Data (fertilizer, yield, seed weight):

```
low 34 10
low 29 14
low 35 11
low 32 13
high 33 14
high 38 12
high 34 13
high 35 14
```

- 2 responses, yield and seed weight.
- First get means by fertilizer amount.
- Then run 1-way ANOVA for each of yield and seed weight, using fertilizer type as explanatory.

Code

```
data manova1;
  infile "manova1.dat";
  input fertilizer $ yield weight;

proc means;
  var yield weight;
  class fertilizer;

proc glm;
  class fertilizer;
  model yield=fertilizer;

proc glm;
  class fertilizer;
  model weight=fertilizer;
```

The means

The MEANS Procedure

fertilizer	N Obs	Variable	N	Mean	Std Dev	Minimum
high	4	yield	4	35.0000000	2.1602469	33.0000000
		weight	4	13.2500000	0.9574271	12.0000000
low	4	yield	4	32.5000000	2.6457513	29.0000000
		weight	4	12.0000000	1.8257419	10.0000000

Means on both variables are slightly higher for high fertilizer. Are those differences significant? Look at ANOVAs (2-sample t-tests would also have worked.)

The ANOVAs

Only one x (fertilizer amount) so look at "model" line.

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.50000000	12.50000000	2.14	0.1936
Error	6	35.00000000	5.83333333		
Corrected Total	7	47.50000000			

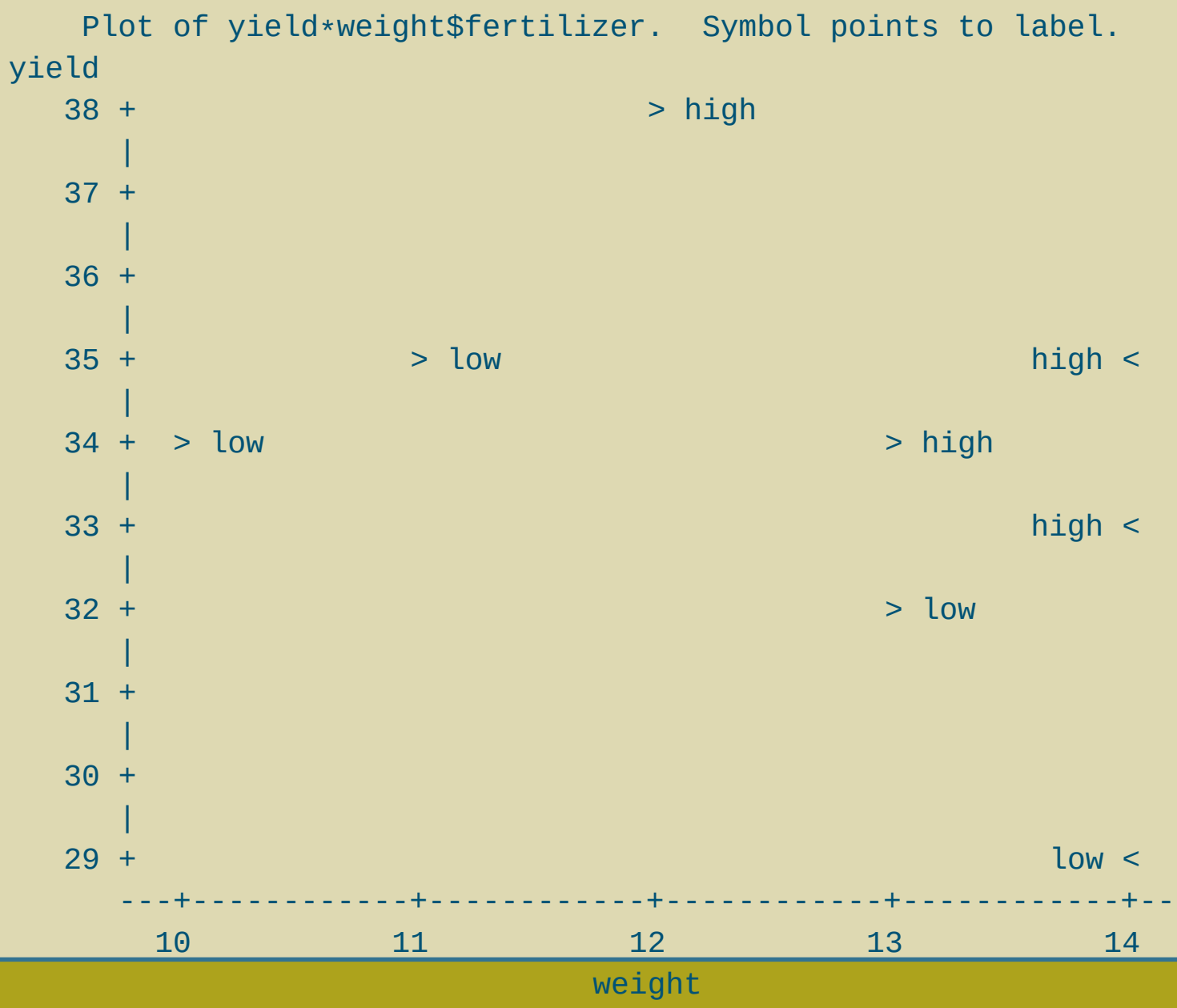
Dependent Variable: weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.12500000	3.12500000	1.47	0.2708
Error	6	12.75000000	2.12500000		
Corrected Total	7	15.87500000			

Neither mean yield nor mean weight depends on the amount of fertilizer. But: look at plot of yield vs. weight labelled by fertilizer, using this code:

```
proc plot;
  plot yield*weight $ fertilizer;
```

Plot of yield vs. weight



MANOVA code

- High-fertilizer plants have both yield and weight high.
- True even though no sig difference in yield or weight individually.
- Could draw a line separating highs from lows on graph.
- Is *that* significant? MANOVA finds out.
- Code:

```
proc glm;
  class fertilizer;
  model yield weight=fertilizer;
  manova h=_all_;
```

Output

Includes this:

The GLM Procedure
Multivariate Analysis of Variance

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.19845779	10.10	2	5	0.0175
Pillai's Trace	0.80154221	10.10	2	5	0.0175
Hotelling-Lawley Trace	4.03885481	10.10	2	5	0.0175
Roy's Greatest Root	4.03885481	10.10	2	5	0.0175

- Four versions of ANOVA F-test, here all agree: the multivariate difference seen on graph *is* significant.
- With more than 2 responses, cannot draw graph. What then?
- Use *discriminant analysis* (of which more later).

A discriminant analysis

Treat this as “magic” for now, but: obtain output data set and look at it.

```
proc discrim can out=fred;
  class fertilizer;
  var yield weight;

proc print data=fred;
```

Ignore most of output from PROC DISCRIM, then look at output data set.

Output

Linear Discriminant Function for fertilizer

Variable	high	low
Constant	-943.76534	-798.70399
yield	33.60736	30.93865
weight	53.68098	49.32515

- For an observation's yield and weight, calculate discriminant functions, classify into fertilizer group with higher one.
- SAS does this for us (see Can1 below).

Output data set

Obs	fertilizer	yield	weight	Can1	Can2	high	low	_INTO_
1	low	34	10	-3.09314	.	0.00002	0.99998	low
2	low	29	14	-1.92110	.	0.00125	0.99875	low
3	low	35	11	-1.07511	.	0.02315	0.97685	low
4	low	32	13	-0.87242	.	0.04579	0.95421	low
5	high	33	14	1.14561	.	0.98180	0.01820	high
6	high	38	12	2.47628	.	0.99982	0.00018	high
7	high	34	13	0.66093	.	0.90893	0.09107	high
8	high	35	14	2.67896	.	0.99991	0.00009	high

- In `Can1`, low value suggests low fertilizer, high suggests high.
- "high" and "low" are estimated probabilities that observation with that yield and weight was high or low fertilizer.
- Last column is SAS's guess at which group it comes from (higher est prob). Got them all right.
- Distinction between high and low quite clear when looked at the right way.
- Procedure works no matter what combination of responses best divides data into groups by x.