Stat 412/512

Two Way Anova

Feb 6 2015

Charlotte Wickham stat512.cwick.co.nz

Quiz #2

Study guide posted.

You get two attempts, keep highest score.

Opens today at noon, closes Monday at noon, 60 minutes to complete.

Study guide is on website

Roadmap

DONE:

Understand what a multiple regression model is.

Know how to do inference on single and multiple parameters.

Some extra tools for checking models.

Our general strategy.

TO DO:

Some special cases, two way ANOVA, multi factor studies, no replication.

Model selection.

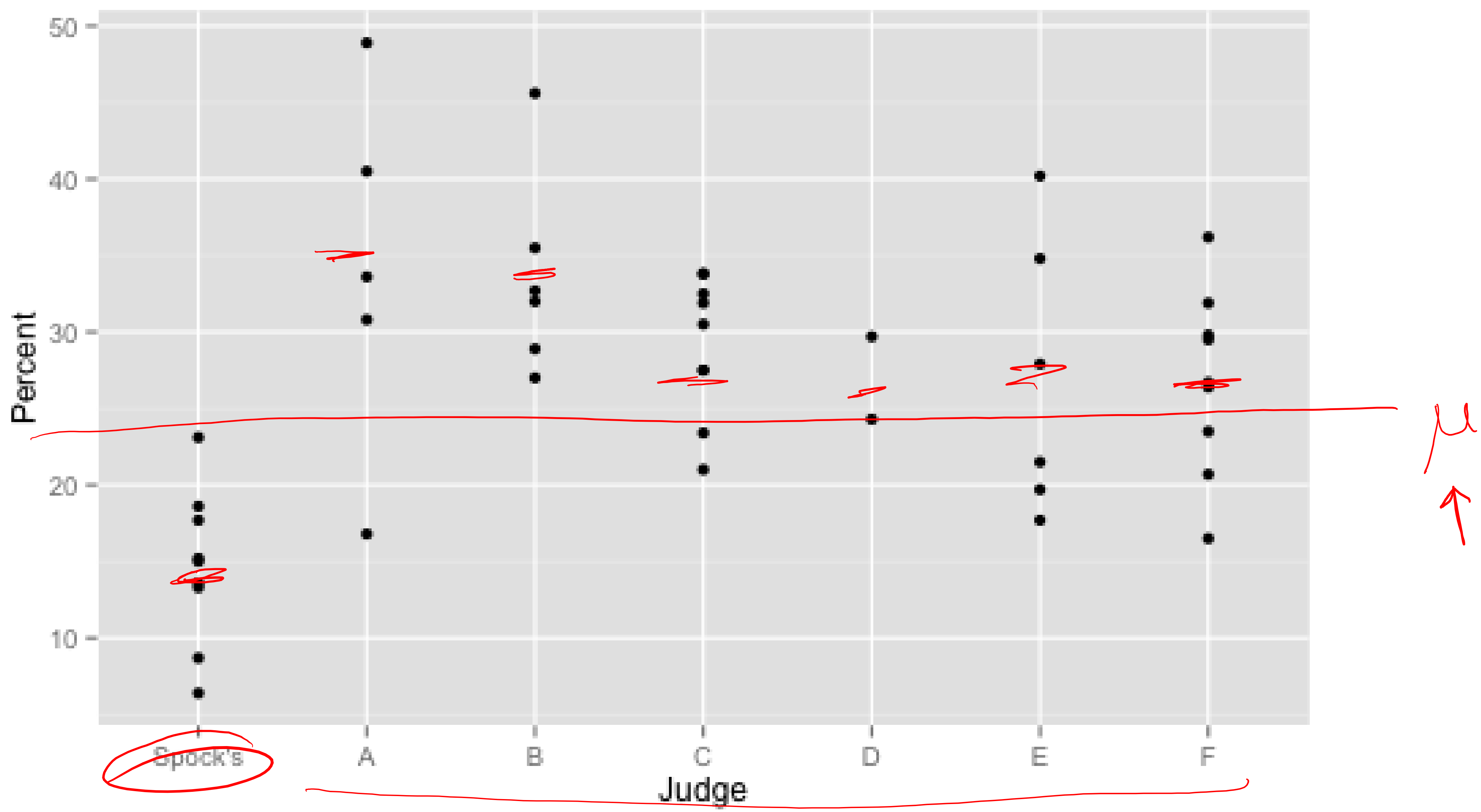
Serial correlation.

Multivariate responses.

Strategy

1. Explore with plots, summaries etc.
2. Fit a tentative model
3. Check residual plots for problems and outliers
4. Investigate influential points
5. Find a simple good fitting model
6. Answer questions of interest

Remember the Spock case study?



Do the judges all have the same population mean percent of women on their venires?

A one-way ANOVA

One response variable.

One grouping variable with many levels.

Null: All the groups have the same mean response

Alternative: At least one group has a different mean response

Full model: separate means

Reduced model: equal means

Compare with an extra sum of squares F-test

afterwards answer particular questions about the means

One way ANOVA is just a special case of multiple regression

The full and reduced models are examples of multiple regression models with a single categorical variable.

Full model: $\mu\{\,\textit{\% women} \mid \textit{Judge}\,\} = \text{JUDGE} =$

$$\beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D + \beta_5 E + \beta_6 F$$

where A, B, C, D, E and F are indicator variables for each judge. (Spock is the baseline)

$A = \begin{cases} 1 & \text{Judge was Judge A} \\ 0 & \end{cases}$

Reduced model: $\mu\{\,\textit{\% women} \mid \textit{Judge}\,\} = \underline{\beta_0}$

Two way ANOVA

One response variable.

Two grouping variables with many levels.

A multiple regression models with two categorical variables.

$\mu\{\, \mathit{Response} \mid \underbrace{\mathit{Factor1},\ \mathit{Factor2}}_{\text{categorical}} \}$

Case1302: Pygmalion Effect

Pygmalion Effect = High expectations translate to better performance

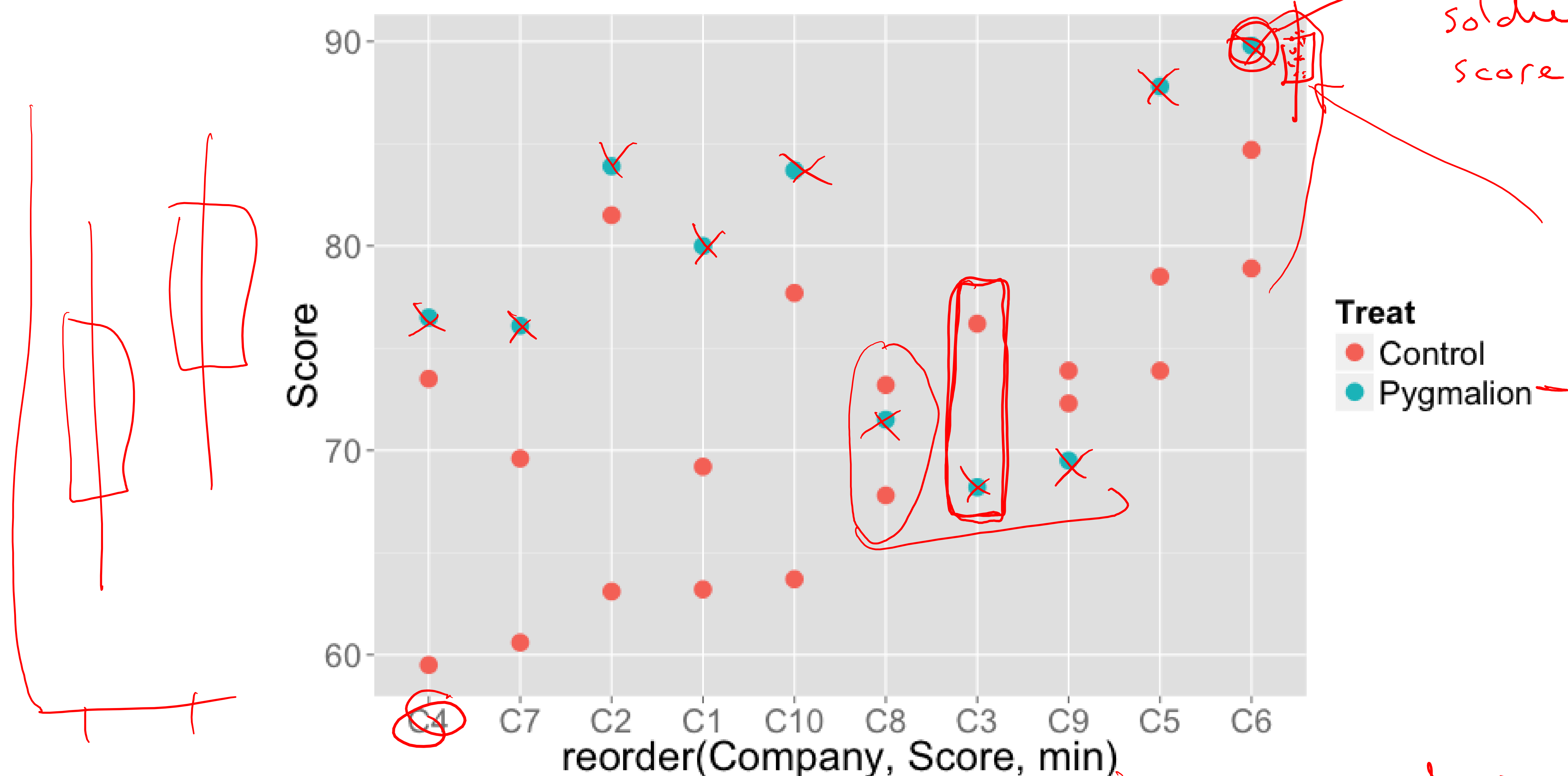
10 army companies each with 3 platoons,

one platoon randomly picked for pygmalion treatment,

platoon leader is told his platoon has scored highly on tests that indicate they are superior.

After basic training platoons are given a Practical Specialty Test.

Response = "average score on PST of platoon" qplot(reorder(Company, Score, min), Score, data = case1302, colour = Treat)



averaging soldiers scores

C P

increasing values of worst performing company

Two factors: Company (10 levels)

Treatment (2 levels)

Another way to think of the data, is as a **two-way** table.

Rows = one factor

Columns = the other factor

Cell = Response(s)

Balance - the data is **balanced** if each cell has the same number of observations.

Display 13.3 J columns p.378

Average scores of soldiers on the Practical Specialty Test, for platoons given the Pygmalion treatment and for control platoons

I rows

	Treatments		
Company	*Pygmalion*	*Control*	
1	80.0	63.2	69.2
2	83.9	63.1	81.5
3	68.2	76.2	
4	76.5	59.5	73.5
5	87.8	73.9	78.5
6	89.8	78.9	84.7
7	76.1	60.6	69.6
8	71.5	67.8	73.2
9	69.5	72.3	73.9
10	83.7	63.7	77.7

this data is unbalanced

I x J cells

Your turn

different mean for every combination of company & treatment

What are five possible models?

(in shorthand)

μ{ *Score* | *Company, Treatment*} =

① " " = COMPANY + TREATMENT + COMPANY×TREATMENT

② " " = COMPANY + TREATMENT

③ " " = COMPANY

④ " " = TREAT

⑤ " " = constant

The additive model

μ{ *Response* | *Factor1, Factor2*} = FACTOR1 + FACTOR2

The effect of either factor is the same regardless of the other factor.

$= \beta_0 + \beta_1 Pyg + \beta_2 C2 + \ldots + \beta C10$

μ{ *Score* | *Company, Treatment*} = COMPANY + TREATMENT

Mean scores on the Practical Specialty Test according to the additive model, in terms of coefficients in a multiple regression model with indicators

Company	Treatments: *Pygmalion*	Treatments: *Control*	Treatment Effects *(Pygmalion - Control)*
1	$\beta_0 + \beta_1$	β_0	β_1
2	$\beta_0 + \beta_2 + \beta_1$	$\beta_0 + \beta_2$	β_1
3	$\beta_0 + \beta_3 + \beta_1$	$\beta_0 + \beta_3$	β_1
4	$\beta_0 + \beta_4 + \beta_1$	$\beta_0 + \beta_4$	β_1
5	$\beta_0 + \beta_5 + \beta_1$	$\beta_0 + \beta_5$	β_1
6	$\beta_0 + \beta_6 + \beta_1$	$\beta_0 + \beta_6$	β_1
7	$\beta_0 + \beta_7 + \beta_1$	$\beta_0 + \beta_7$	β_1
8	$\beta_0 + \beta_8 + \beta_1$	$\beta_0 + \beta_8$	β_1
9	$\beta_0 + \beta_9 + \beta_1$	$\beta_0 + \beta_9$	β_1
10	$\beta_0 + \beta_{10} + \beta_1$	$\beta_0 + \beta_{10}$	β_1

How many parameters ?

11

$(I-1)$ 9

$+(J-1)$ 1

$+1$ 1

$= 11$

or non-additive

The **saturated** model

μ{ *Response* | *Factor1, Factor2*} = FACTOR1 + FACTOR2 + FACTOR1 x FACTOR2

The effect of either factor the depends on the other factor.

μ{ *Score* | *Company, Treatment*} = COMPANY + TREATMENT +

COMPANY x TREATMENT

Mean scores on the Practical Specialty Test, in terms of the parameters in a saturated multiple linear regression model with interaction

Company	Treatments: *Pygmalion*	Treatments: *Control*	Treatment Effects (*Pygmalion - Control*)
1	$\beta_0 + \beta_1$	β_0	β_1
2	$\beta_0 + \beta_2 + \beta_1 + \beta_{11}$	$\beta_0 + \beta_2$	$\beta_1 + \beta_{11}$
3	$\beta_0 + \beta_3 + \beta_1 + \beta_{12}$	$\beta_0 + \beta_3$	$\beta_1 + \beta_{12}$
4	$\beta_0 + \beta_4 + \beta_1 + \beta_{13}$	$\beta_0 + \beta_4$	$\beta_1 + \beta_{13}$
5	$\beta_0 + \beta_5 + \beta_1 + \beta_{14}$	$\beta_0 + \beta_5$	$\beta_1 + \beta_{14}$
6	$\beta_0 + \beta_6 + \beta_1 + \beta_{15}$	$\beta_0 + \beta_6$	$\beta_1 + \beta_{15}$
7	$\beta_0 + \beta_7 + \beta_1 + \beta_{16}$	$\beta_0 + \beta_7$	$\beta_1 + \beta_{16}$
8	$\beta_0 + \beta_8 + \beta_1 + \beta_{17}$	$\beta_0 + \beta_8$	$\beta_1 + \beta_{17}$
9	$\beta_0 + \beta_9 + \beta_1 + \beta_{18}$	$\beta_0 + \beta_9$	$\beta_1 + \beta_{18}$
10	$\beta_0 + \beta_{10} + \beta_1 + \beta_{19}$	$\beta_0 + \beta_{10}$	$\beta_1 + \beta_{19}$

How many parameters ?

20

$1 + 1$

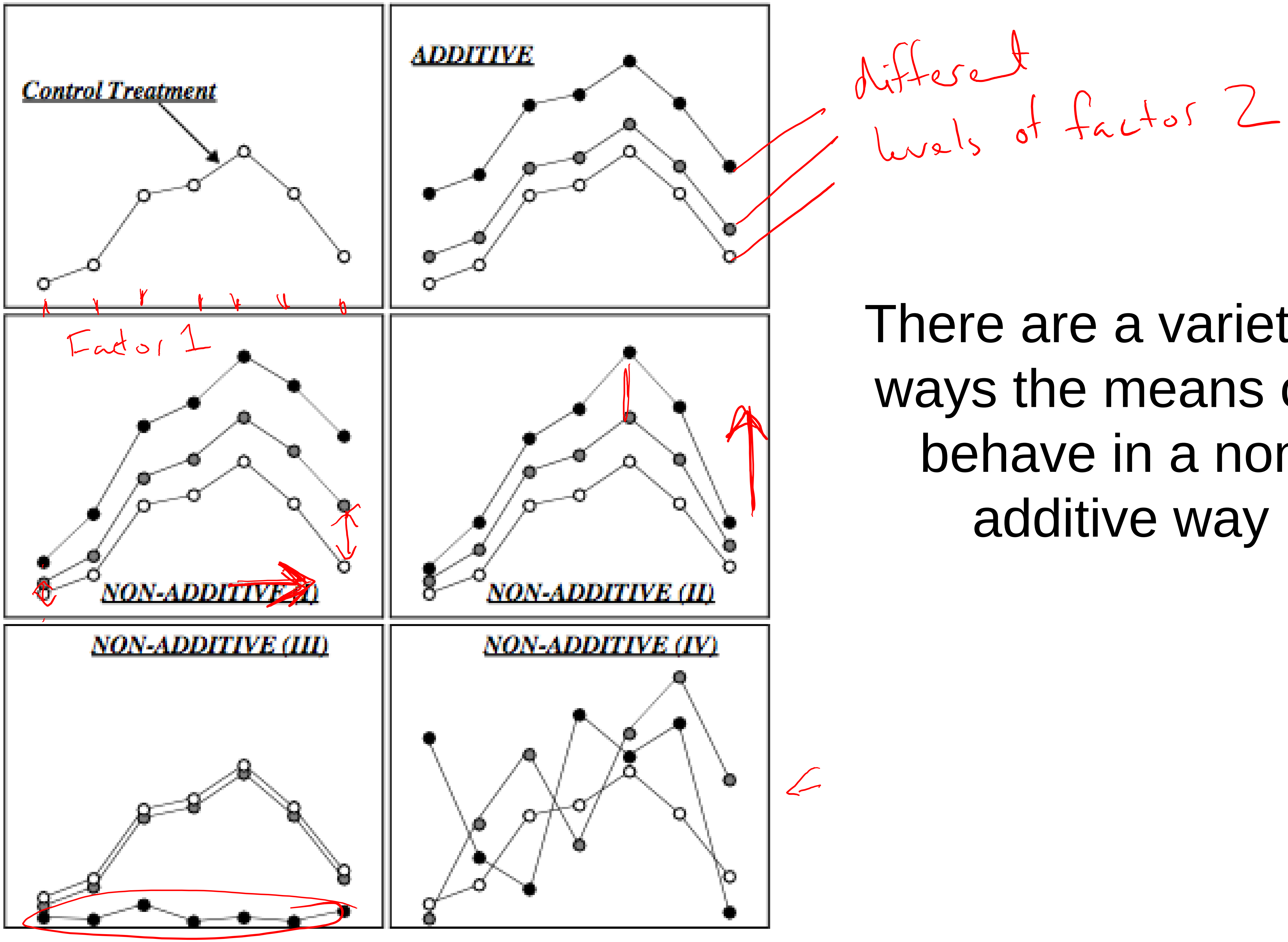
$(I-1) + 9$

$(J-1) + 1$

$(I-1)(J-1)\ 9$

$= 20$

Hypothetical treatment curves plotted against another factor, illustrating additive and some non-additive conditions



There are a variety of ways the means can behave in a non-additive way

A two-way ANOVA

Sometimes only one factor is of interest, sometimes both are, sometimes the interaction is the primary interest.

The general approach is the same:

Start with the saturated model

Use F-tools to simplify

Then answer specific questions about means

$$\mu\{\, Score \mid Company, Treatment\} = \beta_0 + \beta_1 pyg + \beta_2 cmp2 + \beta_3 cmp3 + \beta_4 cmp4 + \beta_5 cmp5 + \beta_6 cmp6 + \beta_7 cmp7 + \beta_8 cmp8 + \beta_9 cmp9 + \beta_{10} cmp10$$

The Pygmalion data with indicator variables defining treatment and companies, in an additive model

Case	score	pyg	cmp2	cmp3	cmp4	cmp5	cmp6	cmp7	cmp8	cmp9	cmp10
1	80.0	1	0	0	0	0	0	0	0	0	0
2	63.2	0	0	0	0	0	0	0	0	0	0
3	69.2	0	0	0	0	0	0	0	0	0	0
4	83.9	1	1	0	0	0	0	0	0	0	0
5	63.1	0	1	0	0	0	0	0	0	0	0
6	81.5	0	1	0	0	0	0	0	0	0	0
7	68.2	1	0	1	0	0	0	0	0	0	0
8	76.2	0	0	1	0	0	0	0	0	0	0
9	76.5	1	0	0	1	0	0	0	0	0	0
10	59.5	0	0	0	1	0	0	0	0	0	0
11	73.5	0	0	0	1	0	0	0	0	0	0
12	87.8	1	0	0	0	1	0	0	0	0	0
13	73.9	0	0	0	0	1	0	0	0	0	0
14	78.5	0	0	0	0	1	0	0	0	0	0
15	89.8	1	0	0	0	0	1	0	0	0	0
16	78.9	0	0	0	0	0	1	0	0	0	0
17	84.7	0	0	0	0	0	1	0	0	0	0
18	76.1	1	0	0	0	0	0	1	0	0	0
19	60.6	0	0	0	0	0	0	1	0	0	0
20	69.6	0	0	0	0	0	0	1	0	0	0
21	71.5	1	0	0	0	0	0	0	1	0	0
22	67.8	0	0	0	0	0	0	0	1	0	0
23	73.2	0	0	0	0	0	0	0	1	0	0
24	69.5	1	0	0	0	0	0	0	0	1	0
25	72.3	0	0	0	0	0	0	0	0	1	0
26	73.9	0	0	0	0	0	0	0	0	1	0
27	83.7	1	0	0	0	0	0	0	0	0	1
28	63.7	0	0	0	0	0	0	0	0	0	1
29	77.7	0	0	0	0	0	0	0	0	0	1

Extra SS F-test

Full model: (I x J) parameters

μ{ *Score* | *Company, Treatment*} =

COMPANY + TREATMENT +

COMPANY x TREATMENT

1 +
(I − 1) +
(J − 1) +
(I − 1)(J − 1)

Reduced model: 1 + (I - 1) + (J - 1) parameters

μ{ *Score* | *Company, Treatment*} =

COMPANY + TREATMENT

F-test with, (I - 1)(J - 1) and n - (I x J) d.f.

```
> fit_pyg_mult <- lm(Score ~ Company + Treat + Company:Treat,
+   data = case1302)
> qplot(.fitted, .resid, data = fit_pyg_add)
> fit_pyg_add <-  lm(Score ~ Company + Treat,
+   data = case1302)
> anova(fit_pyg_add, fit_pyg_mult)
Analysis of Variance Table

Model 1: Score ~ Company + Treat
Model 2: Score ~ Company + Treat + Company:Treat
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     18 778.50
2      9 467.04  9    311.46 0.6669 0.7221
```

There is no evidence the treatment effect differs depending on the company (extra SS F-test on interaction term, p-value = 0.72).

No evidence against additive model.

```
Call:
lm(formula = Score ~ Company + Treat, data = case1302)

Residuals:
    Min      1Q  Median      3Q     Max
-10.660  -4.147   1.853   3.853   7.740

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    68.39316    3.89308  17.568 8.92e-13 ***
CompanyC2       5.36667    5.36968   0.999   0.3308
CompanyC3       0.19658    6.01886   0.033   0.9743
CompanyC4      -0.96667    5.36968  -0.180   0.8591
CompanyC5       9.26667    5.36968   1.726   0.1015
CompanyC6      13.66667    5.36968   2.545   0.0203 *
CompanyC7      -2.03333    5.36968  -0.379   0.7094
CompanyC8       0.03333    5.36968   0.006   0.9951
CompanyC9       1.10000    5.36968   0.205   0.8400
CompanyC10      4.23333    5.36968   0.788   0.4407
TreatPygmalion  7.22051    2.57951   2.799   0.0119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.576 on 18 degrees of freedom
Multiple R-squared: 0.5647,    Adjusted R-squared: 0.3228
F-statistic: 2.335 on 10 and 18 DF,  p-value: 0.0564
```

It is estimated the pygmalion treatment adds 7.2 points to a platoon's score.

causal inference