

Linear and Generalized Linear Models (4433LGLM6Y)

ANCOVA

Meeting 3

Dr. Jos Hageman



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Analysis of covariance (ANCOVA)

ANCOVA is due to R. A. Fisher.

It is described in his book “Statistical methods for research workers” in 1930 and improved in 1935.

First applied to increase precision by H. G. Sanders (advised by Fisher) in 1930.



Example: Seed yield of peanut plants -1

y = seed yield of peanut plants

x = height of plant (measure of level of development or health)

Compare three types of fertilizer, i.e. control, fast and slow release (C, F, and S), with respect to seed yield.

30 plants, randomly assigned to C, F and S.



C	F	S
10 plants	10 plants	10 plants
Per plant: yield y height x	y x	y x

Example: Seed yield of peanut plants -2

C	F	S
10 plants	10 plants	10 plants
Per plant: yield y height x	y x	y x



We could consider making 10 blocks of 3 plants each with similar values of x :

CFS	CFS	CFS	CFS	CFS	CFS	CFS	CFS	CFS	CFS
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

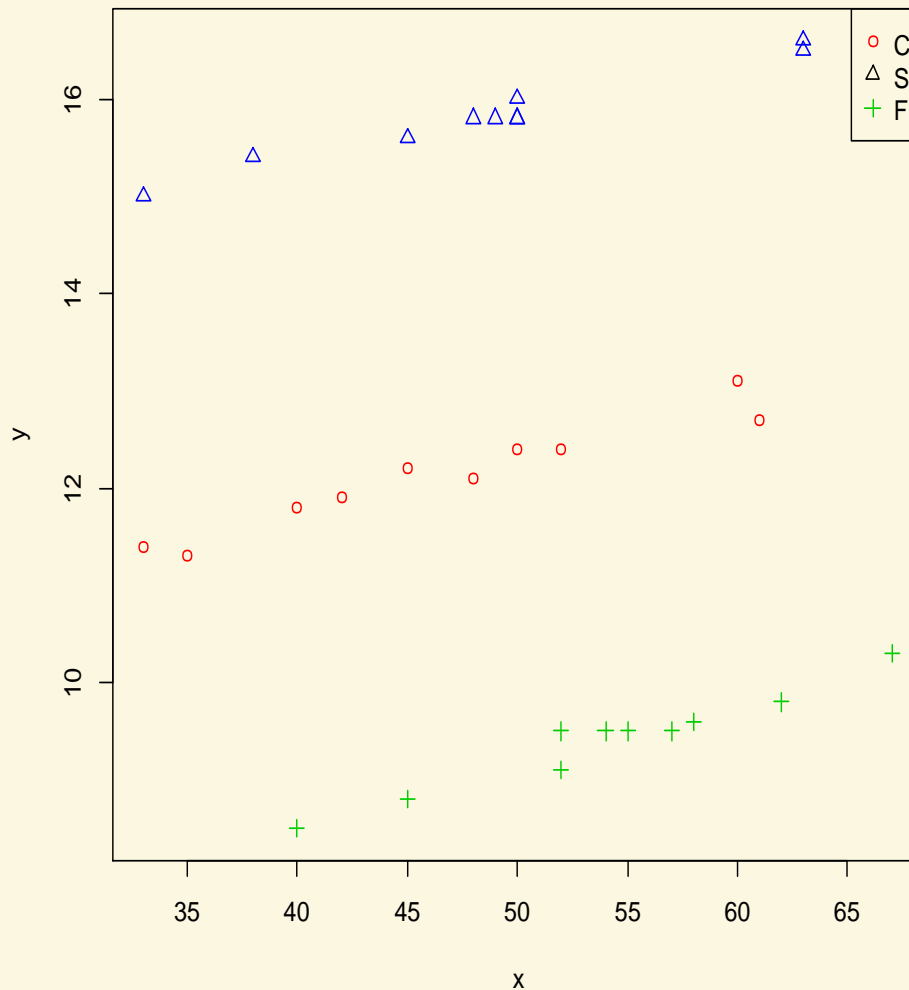
But the extra information in x is not qualitative but **quantitative**.

With some assumptions about the relationship between y and x , a more sophisticated approach is possible, called **analysis of covariance**.

Seed yield of peanut plants



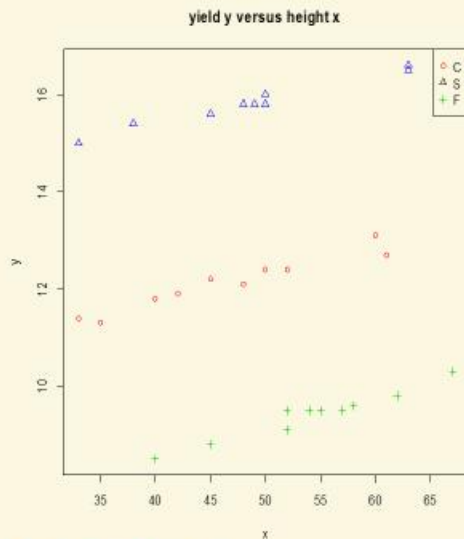
yield y versus height x



Fert	x	y
1	C	45 12.2
2	C	52 12.4
3	C	42 11.9
:	:	:
10	C	33 11.4
11	S	63 16.6
12	S	50 15.8
:	:	:
20	S	49 15.8
21	F	52 9.5
:	:	:
28	F	67 10.3
29	F	55 9.5
30	F	40 8.5

Plot of y against x suggests data are scattered roughly around 3 parallel lines.

Seed yield of peanut plants



Fert	x	y
1	C	45 12.2
2	C	52 12.4
3	C	42 11.9
:	:	:
10	C	33 11.4
11	S	63 16.6
12	S	50 15.8
:	:	:
20	S	49 15.8
21	F	52 9.5
:	:	:
28	F	67 10.3
29	F	55 9.5
30	F	40 8.5

Plot of y against x suggests data are scattered roughly around 3 parallel lines.



Meeting 5.2, ANCOVA, example

Biometris Wageningen University

52

The relation between yield (y) and height of the plant is

- A. a linear relation as seen in regression models **Raise your right hand**
- B. an example of differences between populations as seen in ANOVA models **Raise your left hand**

Peanut plants, one-way ANOVA (ignoring x)

```
> oneway.peanuts <- lm(y ~ Fert, data=peanuts)
> anova(oneway.peanuts)
```

Output from R

Analysis of Variance Table

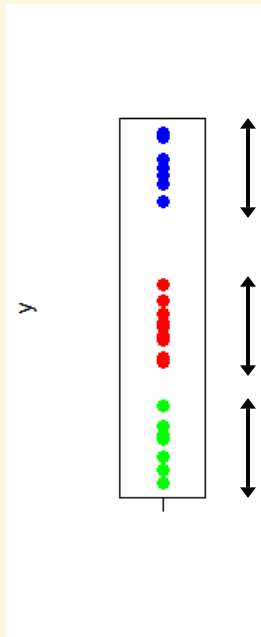
Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fert	2	207.683	103.841	394.28	< 2.2e-16
Residuals	27	7.111	0.263		

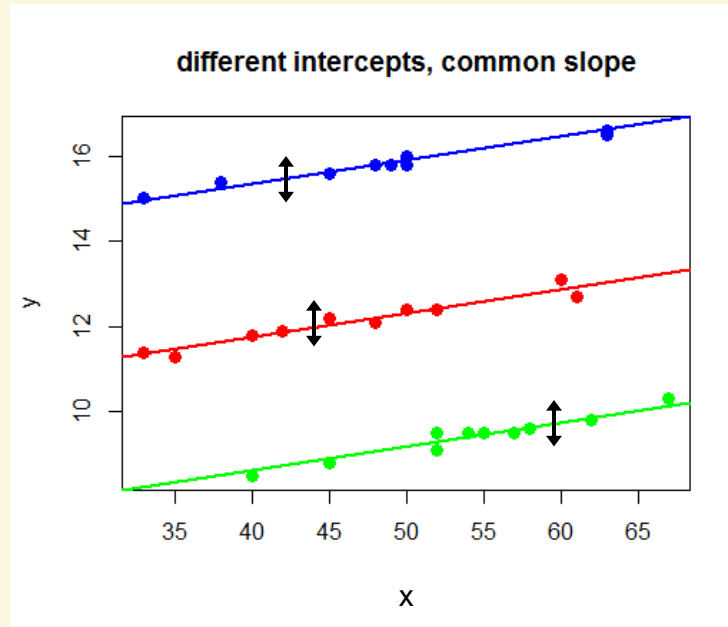


Note for later use: $SSE = 7.11$, $\hat{\sigma}_\epsilon^2 = MSE = 0.263$

Motivation to include covariate x



variability ignoring x



variability including x



C=standard
F=fast release
S=slow release

- Variability in y (ignoring x) around each mean on the left is higher than variability in y (employing x) around each line on the right.
- Height x is unevenly distributed over fertilizers. This may favour some fertilizers over others. Maybe we can correct for this.

Analysis of Covariance (ANCOVA) model, peanut plants

Height x is called a covariate and included as an explanatory variable.

Often a covariate is centred: $z = x - \bar{x}$, here \bar{x} = mean height = 49.90.

ANCOVA model:

$$y_{ij} = \beta_0 + \tau_i + \beta_1 z_{ij} + \epsilon_{ij} = \beta_0 + \tau_i + \beta_1 (x_{ij} - \bar{x}) + \epsilon_{ij}$$

$i = 1, 2, 3$ for fertilizer C, F and S; $j = 1 \dots 10$ for plants per fertilizer.

Analysis of Covariance (ANCOVA) model, peanut plants

$$y_{ij} = \beta_0 + \tau_i + \beta_1 z_{ij} + \epsilon_{ij} = \beta_0 + \tau_i + \beta_1(x_{ij} - \bar{x}) + \epsilon_{ij}$$

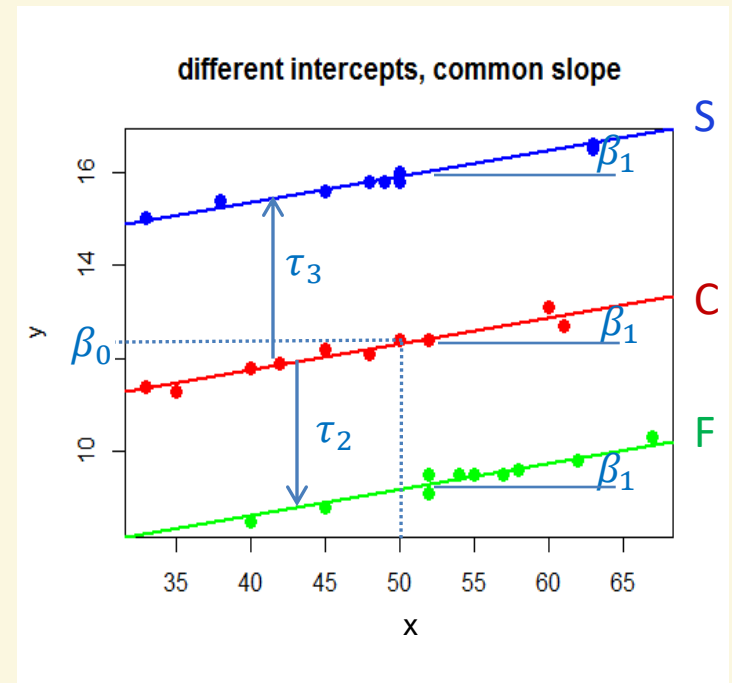
Let **C** be the reference, so $\tau_1 = 0$

β_0 = mean yield **C** at $\bar{x} = 49.90$

$\beta_0 + \tau_3$ = mean yield **F** at $\bar{x} = 49.90$

$\beta_0 + \tau_2$ = mean yield **S** at $\bar{x} = 49.90$

β_1 = common slope (parallel lines)



Adjusted treatment means, peanut plants



$$y_{ij} = \beta_0 + \tau_i + \beta_1 z_{ij} + \epsilon_{ij} = \beta_0 + \tau_i + \beta_1 (x_{ij} - \bar{x}) + \epsilon_{ij}$$

β_0 , $\beta_0 + \tau_2$ and $\beta_0 + \tau_3$ are expected yields for C, S and F, for plants of height $\bar{x} = 49.90$.

They are called **adjusted treatment means**.

E.g. τ_2 is difference in expected yield between F and C, for plants of the same height ($\bar{x} = 49.90$, or any other height, since lines are parallel).

We “correct” for difference in height between plants when we compare treatments.

ANCOVA versus one-way ANOVA, peanut plants

```
> z <- peanuts$x-mean(peanuts$x)
> ancova.peanuts <- lm(y ~ z + Fert, data=peanuts)
> anova(ancova.peanuts)
```

R output

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
z	1	0.4722	0.4722	29.39	1.11e-05
Fert	2	213.9038	106.952	6657.08	2.22e-16
Residuals	26	0.4177	0.016		



One-way ANOVA versus ANCOVA

<i>SSE</i>	7.11	0.42
<i>MSE</i>	0.263	0.016

Substantial part of error variation in one-way ANOVA explained by x :
SSE changes from 7.11 to 0.42 and *MSE* from 0.263 into 0.016.

ANCOVA offers more accurate comparisons between types of fertilizer.

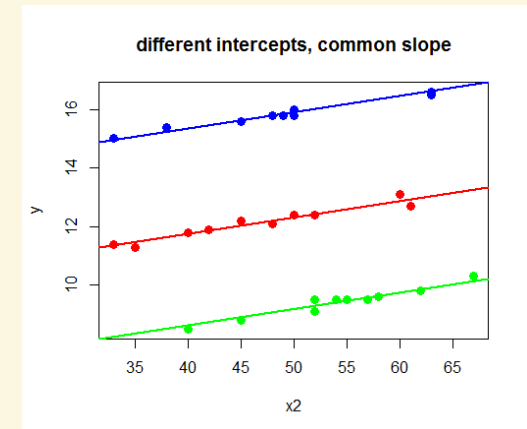
Three parallel lines, peanut plants



```
> coef(summary(ancova.peanuts))
```

R output

	Estimate	Std. Error	t value	Pr(> t)
(Intcpt)	12.314	0.04109	299.7	1.54e-47
z	0.056	0.00273	20.4	1.58e-17
FertF	-3.144	0.06037	-52.1	7.94e-28
FertS	3.572	0.05703	62.6	6.82e-30



The three estimated parallel lines for the three types of fertilizer:

Control $\hat{y} = 12.314 + 0.056 z = 12.314 + 0.056 (x - 49.90)$

Fast $\hat{y} = (12.314 - 3.144) + 0.056 z = 9.170 + 0.056 (x - 49.90)$

Slow $\hat{y} = (12.314 + 3.572) + 0.056 z = 15.886 + 0.056 (x - 49.90)$

To centre or not to centre?

C: $\hat{y} = 12.314 + 0.056(x - 49.90)$

F: $\hat{y} = 9.170 + 0.056(x - 49.90)$

S: $\hat{y} = 15.886 + 0.056(x - 49.90)$



x **centred**: (estimated) adj. treatment mean: $\hat{\beta}_0 + \hat{\tau}_i$

x **not centred**: (estimated) adj. treatment mean: $\hat{\beta}_0 + \hat{\tau}_i + \hat{\beta}_1 \bar{x}$

With 1st option β_0 is expected seed yield at $x = \bar{x}$.

With 2nd option β_0 is expected seed yield at $x = 0$.

The last one is rather silly for the peanut plants (plants of height 0).

But same results for adjusted treatment means with both options.

Adjusted treatment mean & standard error

$$adj. mean = \bar{y}_{i.} - \hat{\beta}_1(\bar{x}_{i.} - \bar{x}_{..})$$

$$estimated \ se = \sqrt{\frac{\hat{\sigma}_\epsilon^2}{n_i} + (\bar{x}_{i.} - \bar{x}_{..})^2 se(\hat{\beta}_1)^2}$$

Note that the first parts on the right hand side are the expressions for an estimated mean and se in one-way ANOVA.

Note that (hopefully) σ_ϵ^2 is (much) smaller in ANCOVA than in one-way ANOVA. Remember $\hat{\sigma}_\epsilon^2 = MSE$ from ANCOVA.

Difference between adjusted treatment means & standard error

Difference between e.g. adj. means for treatments 1 and 2:

$$\bar{y}_{1.} - \bar{y}_{2.} - \hat{\beta}_1 (\bar{x}_{1.} - \bar{x}_{2.})$$
$$\text{estimated se} = \sqrt{\hat{\sigma}_\varepsilon^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) + (\bar{x}_{1.} - \bar{x}_{2.})^2 \text{se}(\hat{\beta}_1)^2}$$

Again, first parts on the right hand sides are the expressions for a difference and se in one-way ANOVA.

F-tests, R

```
> z <- peanuts$x-mean(peanuts$x)
> ancova.peanuts <- lm(y ~ z + Fert, data=peanuts)
> anova(ancova.peanuts)
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
z	1	0.4722	0.4722	29.39	1.11e-05
Fert	2	213.9038	106.952	6657.08	2.22e-16
Residuals	26	0.4177	0.016		



These are sequential (type I) sums of squares.

In ANCOVA (like in regression) the **order of model terms matters!**

Here, the F-test for height (z) may be misleading, since height is before fertilizer in the model.

The F-test for fertilizers (Fert) is OK, because fertilizers are after height in the model.

Adjusted treatment means, R

```
> coef(summary(ancova.peanuts))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intcpt)	12.314	0.04109	299.7	1.54e-47
z	0.056	0.00273	20.4	1.58e-17
FertF	-3.144	0.06037	-52.1	7.94e-28
FertS	3.572	0.05703	62.6	6.82e-30

Here, C is the reference.

Adjusted treatment means are:

C: 12.31

F: $12.314 - 3.144 = 9.17$

S: $12.314 + 3.572 = 15.89$



'Intcpt'

'Intcpt' + 'FertF'

'Intcpt' + 'FertS'

Pairwise comparisons between adjusted treatment means, R

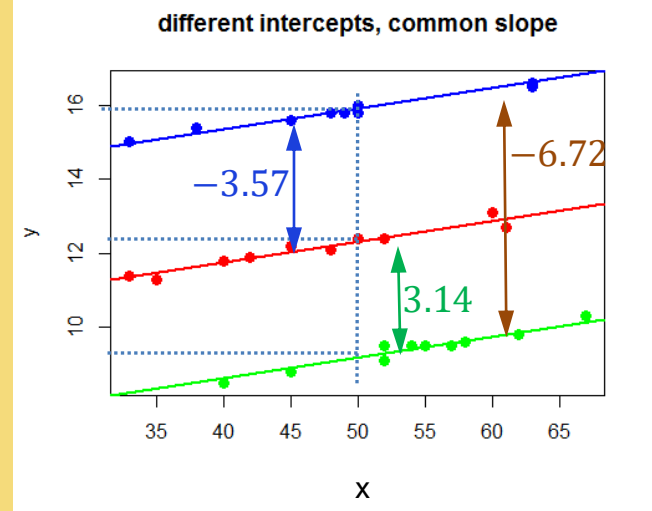
```
> library(emmeans)
> lsm <- emmeans(ancova.peanuts, pairwise ~ Fert)
> summary(lsm, adjust="none")
```



```
$lsmeans
Fert lsmean      SE df lower.CL upper.CL
C      12.31 0.0411 26    12.23    12.40
F       9.17 0.0418 26     9.08     9.26
S      15.89 0.0402 26    15.80    15.97
```

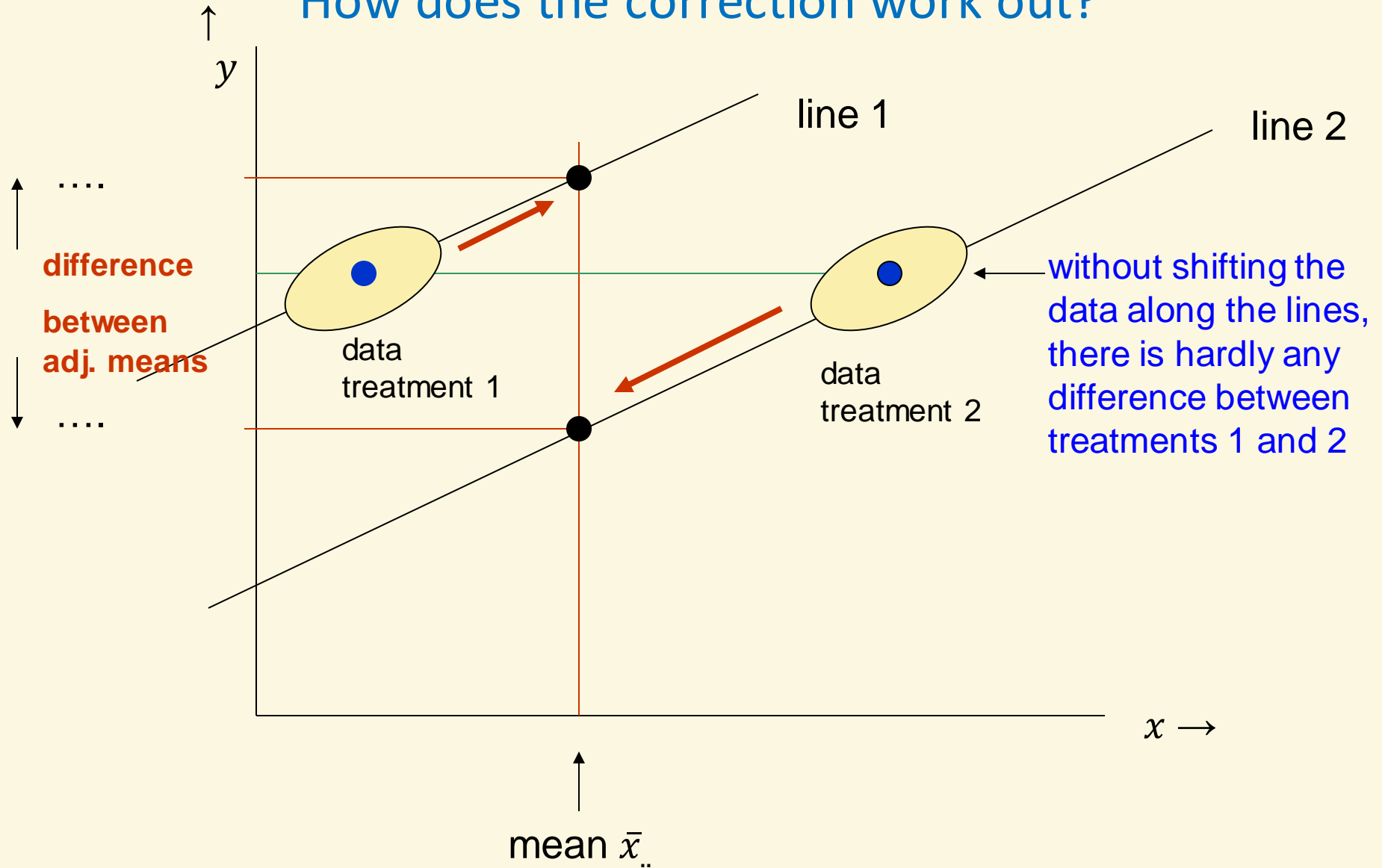
Confidence level used: 0.95

```
$contrasts
contrast estimate      SE df t.ratio p.value
C - F          3.14 0.0604 26    52.1  <.0001
C - S         -3.57 0.0570 26   -62.6  <.0001
F - S         -6.72 0.0585 26  -114.8  <.0001
```



This is Fisher's LSD
method:
adjust = "none"

How does the correction work out?



Ordinary & adjusted treatment means

adjusted means

Control : $\hat{y} = 12.314 + 0.056 (x - 49.9)$

Fast : $\hat{y} = 9.170 + 0.056 (x - 49.9)$

Slow : $\hat{y} = 15.886 + 0.056 (x - 49.9)$



	y.length	y.mean	yield y
Control	10	12.13	
Fast	10	9.41	
Slow	10	15.83	

sample means


	x.length	x.mean	height x
Control	10	46.6	
Fast	10	54.2	
Slow	10	48.9	

Why is the adjusted mean for C (= 12.314) higher than the sample mean (= 12.13)?

Assumptions in ANCOVA

In addition to the usual assumptions about error terms ϵ (independence, normality, equal variance), we need to verify:

- relationship is linear between the response y and
- slope is the same for all treatments (parallel lines)
- covariate x does not depend on the treatments



Any idea
about how we can
test this?

The last assumption will certainly hold when x is observed prior to random assignment of the treatments.

Test for non-parallel lines



Test whether lines are parallel by including extra interaction terms between factor and covariate,

e.g. interaction between fertilizer and height for peanut plants will give three lines with separate intercepts **and separate slopes**.

Compare with an F-test:

Complete model (= model with interaction):

separate slope and separate intercept for each fertilizer,
so three arbitrary lines

Reduced model (= analysis of covariance model):

common slope and separate intercept for each fertilizer,
so three parallel lines

Test for non-linearity

As a simple test for non-linearity, add a quadratic term to the model for e.g. height in the peanut plants example.

For instance put $x_1 = x$ and $x_2 = x^2$ in the model.

Test whether the coefficient of the quadratic term (x_2) significantly differs from 0 with the t-test (or F-test).

We might also consider adding both the quadratic term and the aforementioned interaction and test for significance of each.

Another example: verbalization skills

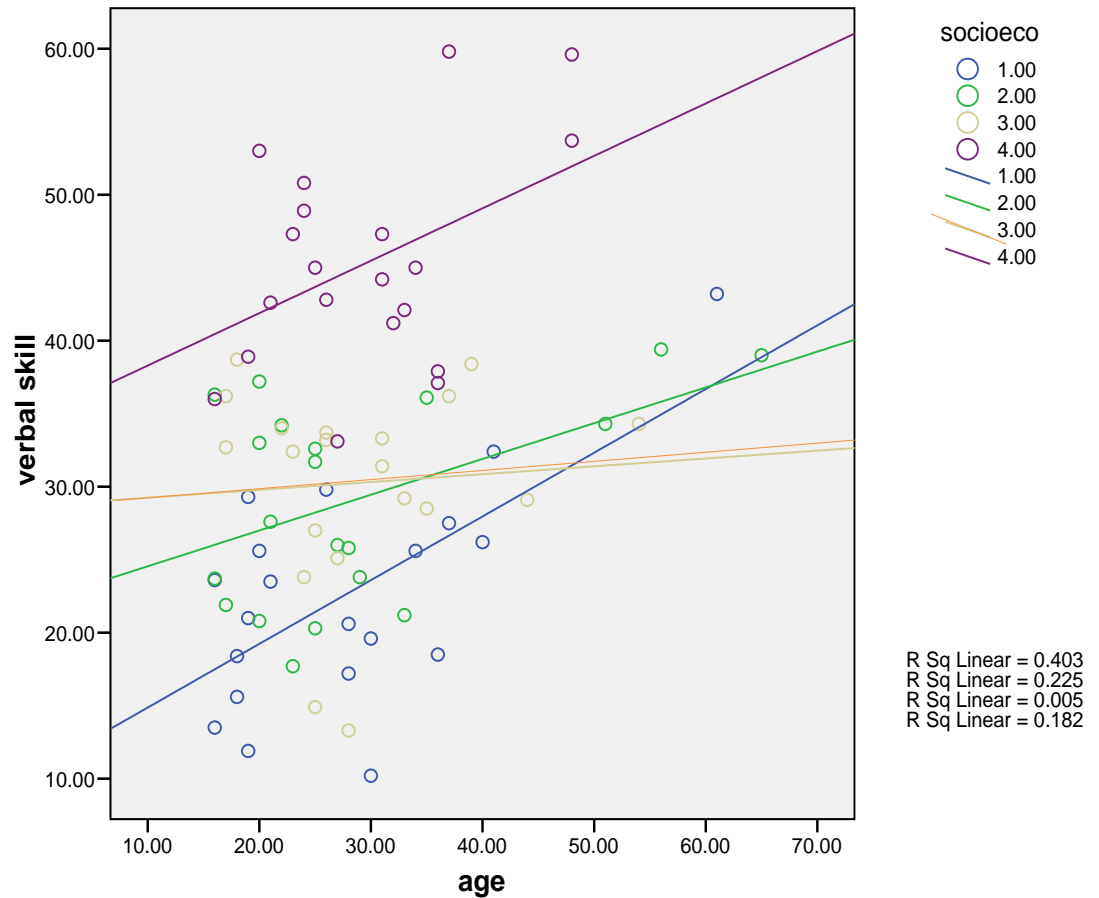
y = verbalization skill

x = age (months)

A = socioeconomic class
(factor at 4 levels)

20 children under age of
six from each class

Interest in effects of A .



Verbalization skills, parallel lines?



Analysis of Variance Table

Response: vs

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fclass	3	5269.6	1756.53	37.3221	1.16e-14
age	1	731.3	731.32	15.5388	0.0001853
fclass:age	3	185.9	61.97	1.3167	0.2756080
Residuals	72	3388.6	47.06		

Check for parallel lines: P-value for interaction = 0.28 > 0.05.
Omit interaction and adopt additive ANCOVA model.
This is the model with parallel lines.

Verbalisation skills, linear relationship?

Check by adding the square of the scaled covariate z .

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.087998	1.662620	13.285	< 2e-16
agecent	0.147946	0.099430	1.488	0.141018
agecent2	0.009872	0.005014	1.969	0.052705
fclass2	5.835949	2.152520	2.711	0.008329
fclass3	7.473394	2.163661	3.454	0.000917
fclass4	22.387407	2.176442	10.286	6.47e-16

agecent is
centred age

agecent2 is its
square



Near indication for curvature (P-value = 0.053).

We will ignore it here, but it may warrant further attention.

Testing, verbalization skills,

Anova Table (Type II tests)

Response: vs

	Sum Sq	Df	F value	Pr(>F)
age	731.3	1	15.344	0.000196
fclass	5051.8	3	35.332	2.42e-14
Residuals	3574.5	75		



Type II SS (obtained with routine 'Anova' from 'car' library) , so OK.

Covariate age adds significantly to the model (P-value = 0.000...).

Significant differences among socio economic classes (P-value = 0.000196).

Adjusted means & pairwise comparisons, verbalization skills

\$`emmeans`

fclass	emmean	SE	df	lower.CL	upper.CL
1	23.4	1.55	75	20.4	26.5
2	29.7	1.54	75	26.6	32.7
3	30.3	1.54	75	27.3	33.4
4	45.1	1.54	75	42.0	48.2



Confidence level used: 0.95

\$`contrasts`

contrast	estimate	SE	df	t.ratio	p.value
1 - 2	-6.228	2.18	75	-2.852	0.0056
1 - 3	-6.904	2.19	75	-3.160	0.0023
1 - 4	-21.671	2.19	75	-9.911	<.0001
2 - 3	-0.676	2.18	75	-0.310	0.7576
2 - 4	-15.443	2.18	75	-7.071	<.0001
3 - 4	-14.767	2.18	75	-6.763	<.0001

Multiple
testing
with
Fisher's
LSD

Relative efficiency, verbalization skills

Is it worthwhile to include the covariate?

This is one-way ANOVA versus ANCOVA.



Rough and ready approach, comparing *MSEs*:

$$RE = MSE(\text{one-way ANOVA}) / MSE(\text{ANCOVA}) = 56.66 / 47.06 = 1.2$$

Analysis of Variance Table

output from one-way ANOVA

Response: vs

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fclass	3	5269.6	1756.53	31.003	3.401e-13
Residuals	76	4305.8	56.66		

Should correction for covariate be made at all?

Correction for covariate may unwittingly remove (part of) treatment effects, when **covariate is influenced by treatment**.

Correction for covariate may erroneously rely on extrapolation of regression lines, when covariate values are very different between groups due to poor study design or unfortunate configuration in observational study; there **should be sufficient overlap**.

Correction for covariate may create **unrealistic adjusted means** for some groups, when covariate values differ systematically between groups in the underlying population.

E.g. comparing males and females, do not correct to common value for (initial) body weight.

The examples - 1



Peanut yield

Covariate plant height x measured before treatment was assigned, thus cannot be influenced by treatment.

Safe as a covariate if there is sufficient overlap

Question

what if x was measured a week after start of the experiment?

Answer

Plant height will likely be influenced by treatment.

Correction for plant height would remove part of the treatment effect on peanut yield.

Do not use this measure of height as a covariate.

The examples - 2



Verbalization skill of children

Covariate age is not influenced by social class.
Safe as a covariate if there is sufficient overlap

Question

what if age was not measured at the start of the study, but at the end?

Answer

Age of a child is, of course, not influenced by social class.
Irrelevant when age is measured, as long as it is measured at the same moment for all children.
So, age at the end may be used as covariate.

Generally, do not use potential covariate when:

- covariate is influenced by treatment (part of treatment effects will be removed)
- covariate values have little overlap between treatments (tricky linear extrapolation of regression lines)
- covariate differs systematically between treatments in the population (leads to unrealistic adjusted means)