MAT20306 - Advanced Statistics

Lecture 11: Analysis of Covariance: ANCOVA

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Topics @ Advanced Level

GENERAL LINEAR MODEL

Confidence intervals

Hypothesis testing

Advancing in the grammar of the language

Map of ANCOVA

An illustrative example

ANCOVA model

Adjusted treatment means

Testing & pairwise comparisons

Assumptions

Reservations

Another example

Example: Seed yield of peanut plants

y = seed yield of peanut plants
3 types of fertilizer; control, slow, fast release (C, S or F)
x = height of plant
(measure of level of development or health)

30 plants, randomly assigned to C, S and F

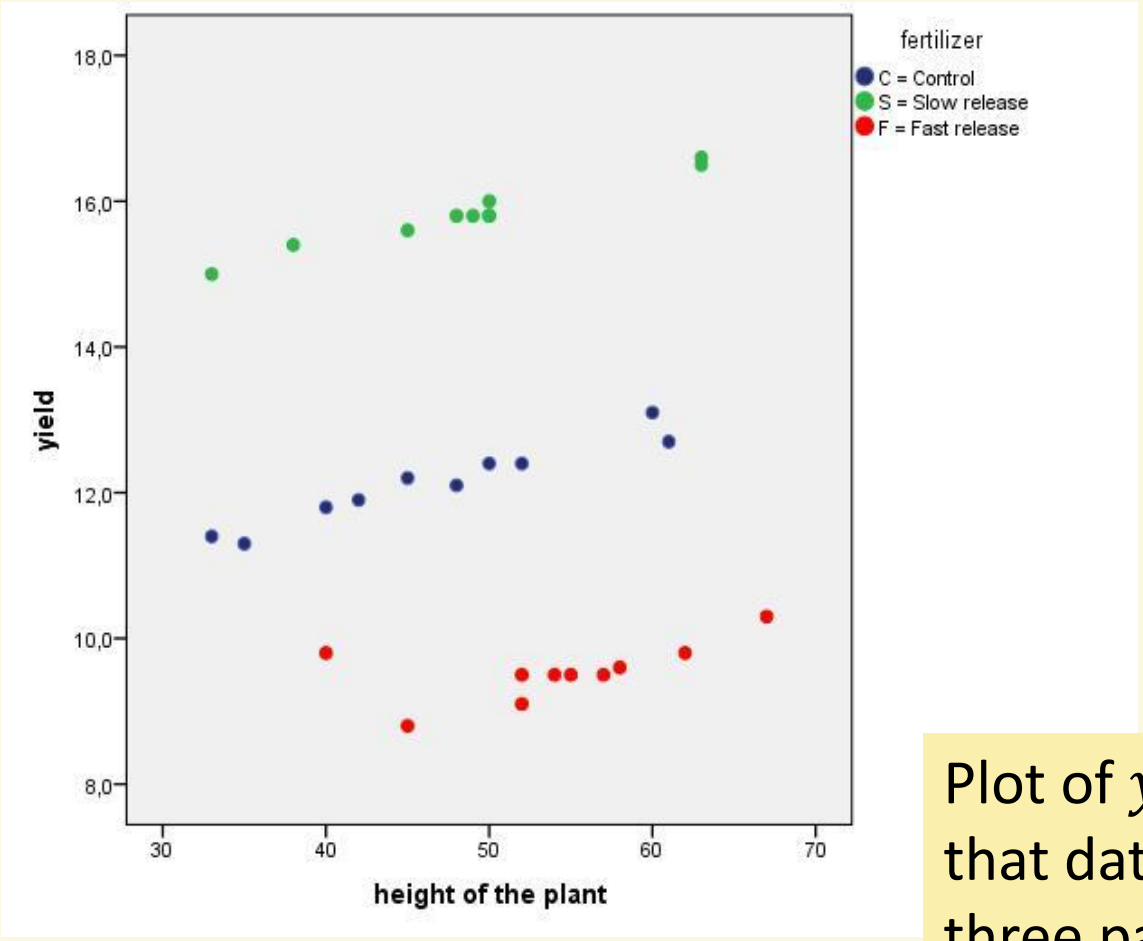
the aim is to compare types of fertilizer C, S and F

We could consider forming blocks based on values of x.

But the extra information (height) 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 (ANCOVA).

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 that data are roughly around three parallel lines.

Peanut plants, one-way ANOVA

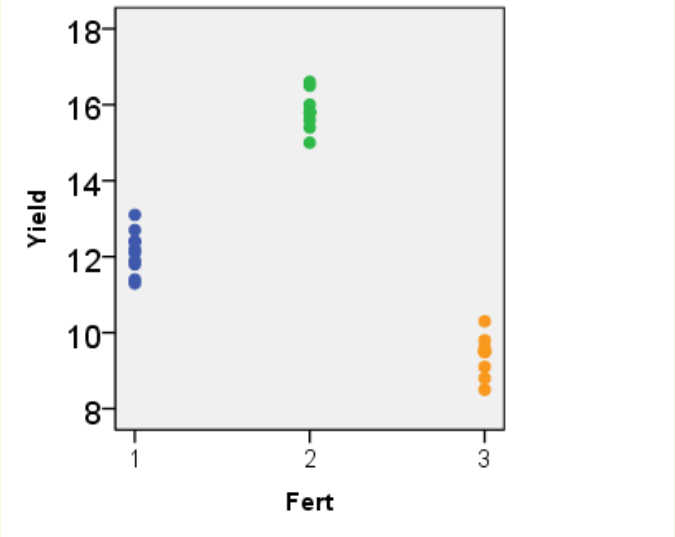
Note (for later use): $SSE = 7.11$, $\hat{\sigma}_\epsilon^2 = MSE = 0.263$

F-test: significant differences in mean yield among fertilizers

ANOVA

Yield

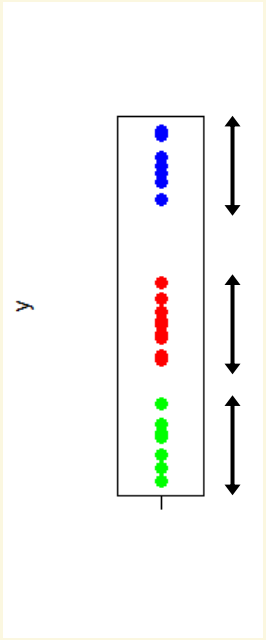
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	207.683	2	103.841	394.279	.000
Within Groups	7.111	27	.263		
Total	214.794	29			

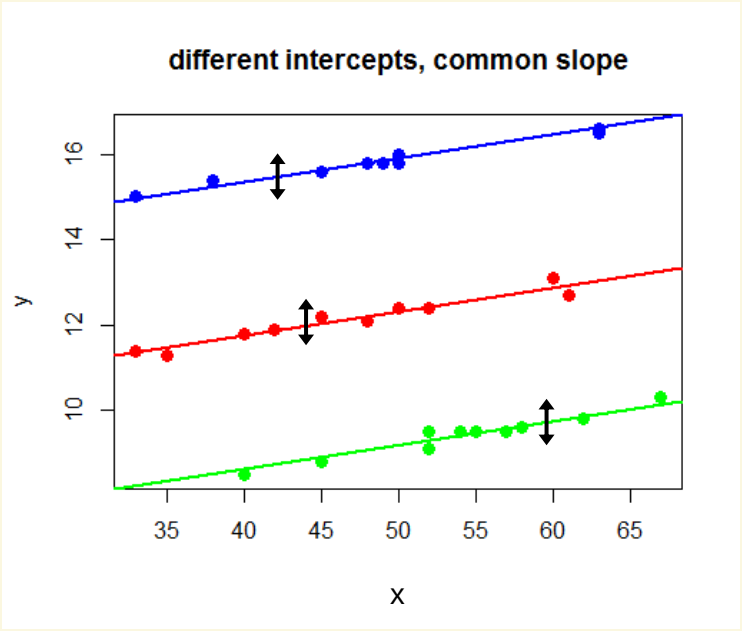


Note (for later use): $SSE = 7.11$, $\hat{\sigma}_\epsilon^2 = MSE = 0.263$

F-test: significant differences in mean yield among fertilizers

Motivation to include covariate x





C=standard
F=fast release
S=slow release

variability ignoring x

variability including x

Variability of y around each line smaller than variability ignoring x.

Height is unevenly distributed over fertilizers: this may favour some fertilizers over others, maybe we can correct for this.

Map of ANCOVA

An illustrative example

ANCOVA model

Adjusted treatment means

Testing & pairwise comparisons

Assumptions

Reservations

Another example

ANCOVA, peanut plants

Parallel lines model:

$$y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \epsilon_{ij}$$ - assumptions later!!!

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

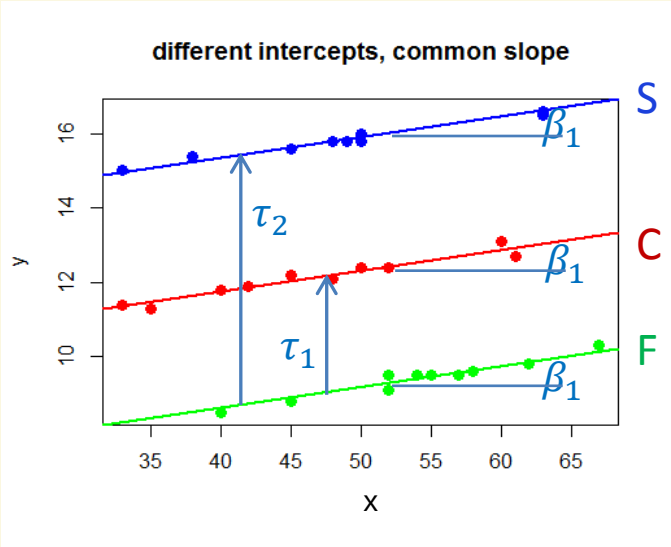
Suppose F is the reference (SPSS), so $\tau_3 = 0$

β_0 = mean yield for F at x=0 !!!

β_0

τ_i = difference in mean yield between fertilizer i ($i = 1, 2$ for C or S) and F

β_1 = common slope



ANCOVA vs. one-way ANOVA

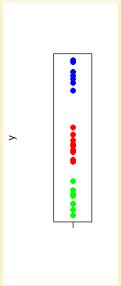
Tests of Between-Subjects Effects

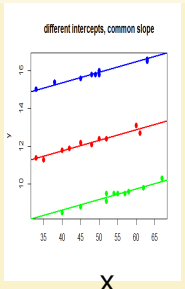
Dependent Variable: yield

Source	Type II Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	214.376[a]	3	71.459	4447.853	.000
Intercept	164.892	1	164.892	10263.470	.000
fert	213.904	2	106.952	6657.085	.000
height	6.693	1	6.693	416.615	.000
Error	.418	26	.016		
Total	4869.850	30			
Corrected Total	214.794	29			

a. R Squared = .998 (Adjusted R Squared = .998)

	One-way ANOVA *vs.*	ANCOVA
SSE	7.11	0.42
MSE	0.263	0.016





Large part error variation one-way ANOVA explained by covariate x in ANCOVA: *MSE* changes from 0.263 into 0.016

ANCOVA offers more accurate comparisons of types of fertilizer.

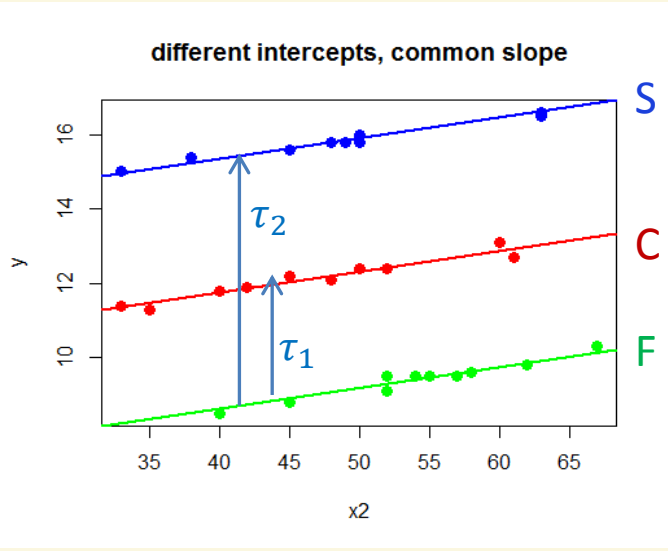
Three parallel lines, peanut plants

Parameter Estimates

Dependent Variable: yield

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	6.385	.1535	41.590	.000	6.070	6.701
[fertilizer=1]	3.144	.0604	52.078	.000	3.020	3.268
[fertilizer=2]	6.716	.0585	114.784	.000	6.596	6.836
[fertilizer=3]	0[a]	.	.	.	.	.
height	.056	.0027	20.411	.000	.050	.061

a. This parameter is set to zero because it is redundant.



fitted model: $\hat{y} = \hat{\beta}_0 + \hat{\tau}_i + \hat{\beta}_1 x$

three parallel lines for the three types of fertilizer:

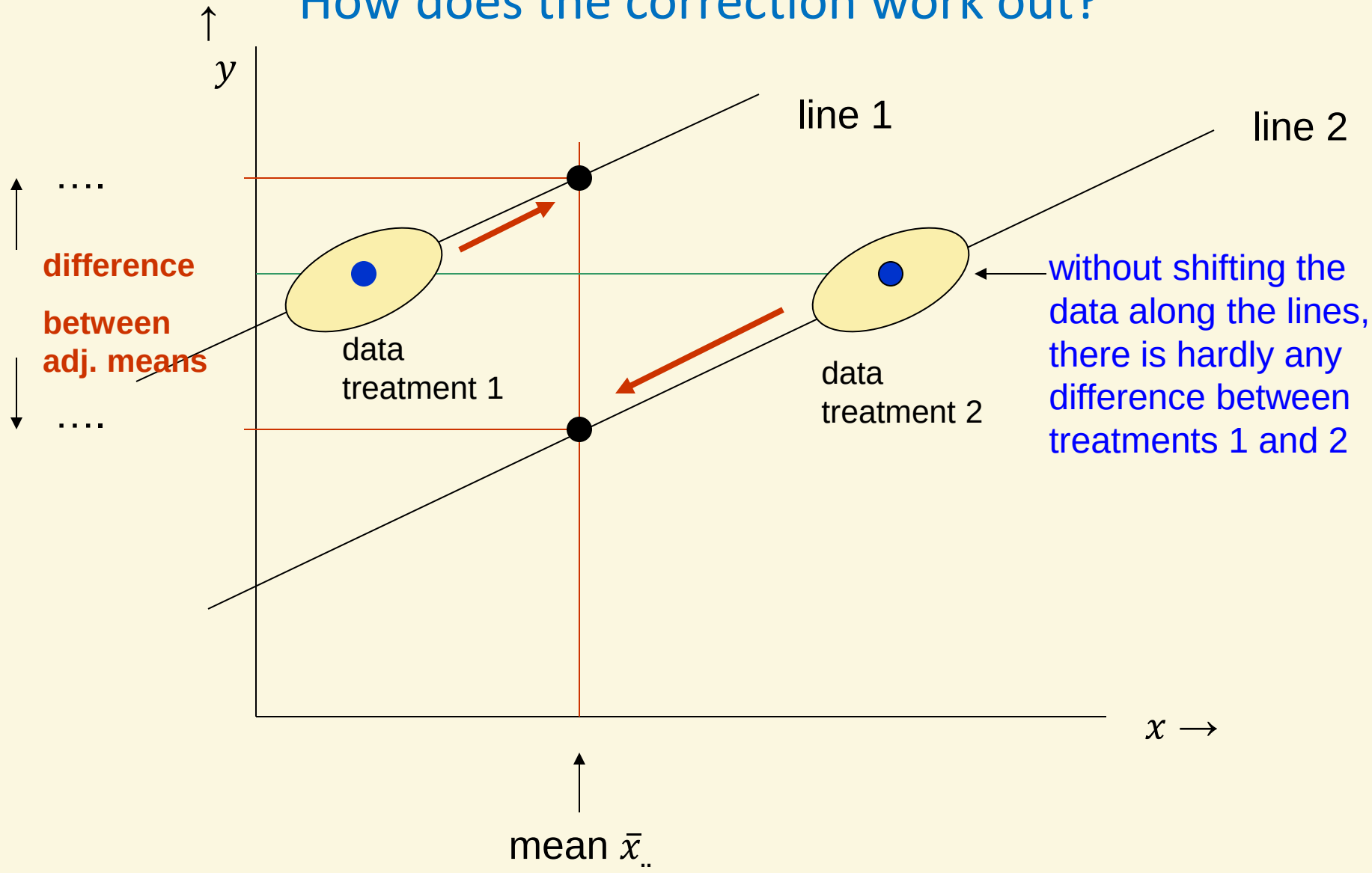
Fast $\hat{y} = 6.385 + 0.056 x$

Control $\hat{y} = (6.385 + 3.144) + 0.056 x$

Slow $\hat{y} = (6.385 + 6.716) + 0.056 x$

What is a smart choice of x ?

How does the correction work out?



Map of ANCOVA

An illustrative example

ANCOVA model

Adjusted treatment means

Testing & pairwise comparisons

Assumptions

Reservations

Another example

Estimation: Adjusted treatment means

$$\bar{y}_{i.} = \hat{\beta}_0 + \hat{\tau}_i + \hat{\beta}_1 \bar{x}_{i.}$$

The observed treatment means are not only different because of treatment differences but also because of differences in the mean of x for the units in the treatment group.

Interest is in the treatment effects, i.e. the (differences between) the τ_i 's. The differences between these are the differences in adjusted treatment means, not with the observed treatment means.

$$\bar{y}_{i.adj} = \hat{\beta}_0 + \hat{\tau}_i + \hat{\beta}_1 \bar{x}..$$

Adjusted treatment means
Estimated means of y, when $x = \bar{x}..$

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

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

Seed yield of peanut plants, continued

Means output
(Descriptive Statistics)

Report

fertilizer		yield	height
C	Mean	12.130	46.60
	N	10	10
S	Mean	15.830	48.90
	N	10	10
F	Mean	9.410	54.20
	N	10	10
Total	Mean	12.457	49.90
	N	30	30

Adjusted Treatment means
method 1:

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

C: 12.13 – 0.056*(46.60-49.90) = 12.315
S: 15.83 – 0.056*(48.90-49.90) = 15.886
F: 9.41 – 0.056*(54.20-49.90) = 9.169

Parameter Estimates

Dependent Variable: yield

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	6.385	.1535	41.590	.000	6.070	6.701
[fertilizer=1]	3.144	.0604	52.078	.000	3.020	3.268
[fertilizer=2]	6.716	.0585	114.784	.000	6.596	6.836
[fertilizer=3]	0[a]	.	.	.	.	.
height	.056	.0027	20.411	.000	.050	.061

a. This parameter is set to zero because it is redundant.

Seed yield of peanut plants, continued

Estimated adjusted treatment means (method 2):

$$\bar{y}_{i.,adj} = \hat{\beta}_0 + \hat{\tau}_i + \hat{\beta}_1 \bar{x}..$$

SPSS uses $x = 0$ as a reference in the model:

but the mean of x in the adjusted means:

Adj. mean C: 6.385 + 3.144 + 0.056 * 49.9 = 12.32

Adj. mean S: 6.385 + 6.716 + 0.056 * 49.9 = 15.90

Adj. mean F: 6.385 + 0.056 * 49.9 = 9.18

Give estimated adjusted treatment means method 3 (Easy), using **Estimates.**

- Note that observed means and adjusted means are not the same

Report

fertilizer		yield	height
C	Mean	12.130	46.60
	N	10	10
S	Mean	15.830	48.90
	N	10	10
F	Mean	9.410	54.20
	N	10	10
Total	Mean	12.457	49.90
	N	30	30

Estimates

Dependent Variable: yield

fertilizer	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
C	12.314[a]	.041	12.230	12.399
S	15.886[a]	.040	15.803	15.968
F	9.170[a]	.042	9.084	9.256

a. Covariates appearing in the model are evaluated at the following values: height = 49.90.

Map of ANCOVA

An illustrative example

ANCOVA model

Adjusted treatment means

Testing & pairwise comparisons

Assumptions

Reservations

Another example

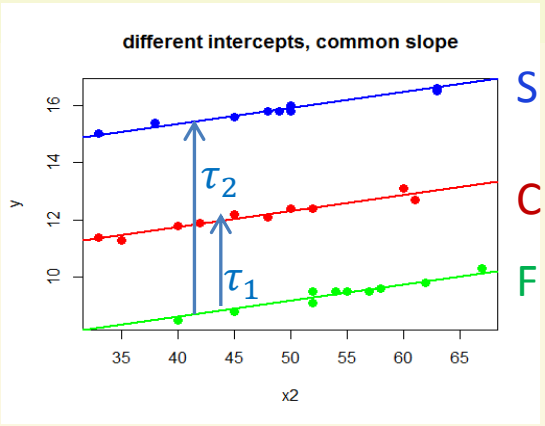
F-tests for treatment effects / covariate effect

Tests of Between-Subjects Effects

Dependent Variable: yield

Source	Type II Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	214.376[a]	3	71.459	4447.853	.000
Intercept	164.892	1	164.892	10263.470	.000
fert	213.904	2	106.952	6657.085	.000
height	6.693	1	6.693	416.615	.000
Error	.418	26	.016		
Total	4869.850	30			
Corrected Total	214.794	29			

a. R Squared = .998 (Adjusted R Squared = .998)



So the test statistics used are: (use Type II SS)

For variety: F=MSfert / MSE with null-distribution $F(df_{fert}, dfE)$

For height: F = MSheight / MSE (for a two-sided H_a) or

$t = \hat{\beta}_1 / se(\hat{\beta}_1)$ (for a one-sided or a two-sided H_a)

with null distributions F(1, dfE) or t_{dfE}

The test for the slope β_1 compares a model with 3 parallel (not horizontal) lines to a model with 3 horizontal lines(i.e. 1-way ANOVA model).

The test on the factor compares the model with 3 parallel lines to a model with only 1 line (where the 3 intercepts are the same).

ANCOVA, F- and t-tests

Tests of Between-Subjects Effects

Dependent Variable: yield

Source	Type II Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	214.376[a]	3	71.459	4447.853	.000
Intercept	164.892	1	164.892	10263.470	.000
fert	213.904	2	106.952	6657.085	.000
height	6.693	1	6.693	416.615	.000
Error	.418	26	.016		
Total	4869.850	30			
Corrected Total	214.794	29			

a. R Squared = .998 (Adjusted R Squared = .998)

Parameter Estimates

Dependent Variable: yield

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	6.385	.1535	41.590	.000	6.070	6.701
[fertilizer=1]	3.144	.0604	52.078	.000	3.020	3.268
[fertilizer=2]	6.716	.0585	114.784	.000	6.596	6.836
[fertilizer=3]	0[a]	.	.	.	.	.
height	.056	.0027	20.411	.000	.050	.061

a. This parameter is set to zero because it is redundant.

F- or t-test for height significant (P-value 0.000): seems good idea to have covariate in the model.

F-test for treatments significant (P-value = 0.000).

t-tests for pairwise comparisons significant (both P-values = 0.00)

Difference between adjusted treatment means, formal expression

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

$$\bar{y}_{1.,adj} - \bar{y}_{2.,adj} = \hat{\tau}_1 - \hat{\tau}_2 =$$

$$= \bar{y}_{1.} - \bar{y}_{2.} - \hat{\beta}_1(\bar{x}_{1.} - \bar{x}_{2.})$$

adjustment for differences in mean height

Pairwise Comparisons

Dependent Variable: yield

(I) fertilizer	(J) fertilizer	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
C	S	-3.572*	.0570	.000	-3.689	-3.454
	F	3.144*	.0604	.000	3.020	3.268
S	C	3.572*	.0570	.000	3.454	3.689
	F	6.716*	.0585	.000	6.596	6.836
F	C	-3.144*	.0604	.000	-3.268	-3.020
	S	-6.716*	.0585	.000	-6.836	-6.596

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

3. Pairwise comparisons by Fisher's LSD show that all 3 fertilizers differ significantly from each other.

Why are s.e.diff.'s different?

Seed yield of peanut plants, continued

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

Pairwise Comparisons

Dependent Variable: yield

(I) fertilizer	(J) fertilizer	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
C	S	-3.572*	.0570	.000	-3.689	-3.454
	F	3.144*	.0604	.000	3.020	3.268
S	C	3.572*	.0570	.000	3.454	3.689
	F	6.716*	.0585	.000	6.596	6.836
F	C	-3.144*	.0604	.000	-3.268	-3.020
	S	-6.716*	.0585	.000	-6.836	-6.596

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

3. Pairwise comparisons by Fisher's LSD show that all 3 fertilizers differ significantly from each other.

Why are s.e.d.'s different?

Map of ANCOVA

An illustrative example

ANCOVA model

Adjusted treatment means

Testing & pairwise comparisons

Assumptions

Reservations

Another example

Assumptions in ANCOVA

In addition to usual assumptions about error terms ϵ, we need to verify:

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

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 between factor and covariate,

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

$$y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \lambda_i x_{ij} + \varepsilon_{ij}$$

Compare complete and reduced model 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.

$$y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \beta_2 x_{ij}^{\ 2} + \varepsilon_{ij}$$

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

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

$$y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \lambda_i x_{ij} + \beta_2 x_{ij}^{\ 2} + \varepsilon_{ij}$$

Map of ANCOVA

An illustrative example

ANCOVA model

Adjusted treatment means

Testing & pairwise comparisons

Assumptions

Reservations

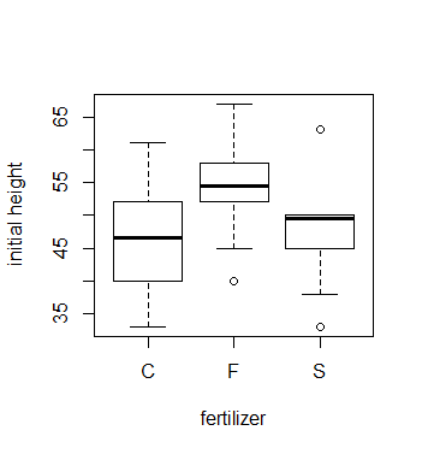
Another example

The example

Example: Peanut yield

- experimental study on effect of fertilizers
- covariate = initial plant height
  measured before treatment was assigned
  cannot be influenced by treatment
- Safe as a covariate

Question: what if plant height was measured a week after the start of the experiment?



Plant height will likely be influenced by treatment.

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

Don't use this measure of height as a covariate.

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.

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

Map of ANCOVA

An illustrative example

ANCOVA model

Adjusted treatment means

Testing & pairwise comparisons

Assumptions

Reservations

The very last Example: GLM

Two points of view

The same model can be used for two distinct situations

- Interest is in the treatments; x is introduced to

1. increase the accuracy of comparison between treatments. Technically, we filter out part of the residual variation (similar to blocking).
2. "correct" for differences in the mean of the covariate between the treatments.

This is usually what we have in mind for analysis of covariance.

- There is a specific interest in the relationship between μ_y and x.
  We want to know whether this relationship is affected by the treatments/ is different for different sub-populations.

Regression in the presence of a qualitative factor.

The model is the same, but
goals and interpretation are different.

Example: anxiety of rats

Study of relationship between anxiety score and dose for two types of drugs (A, B).

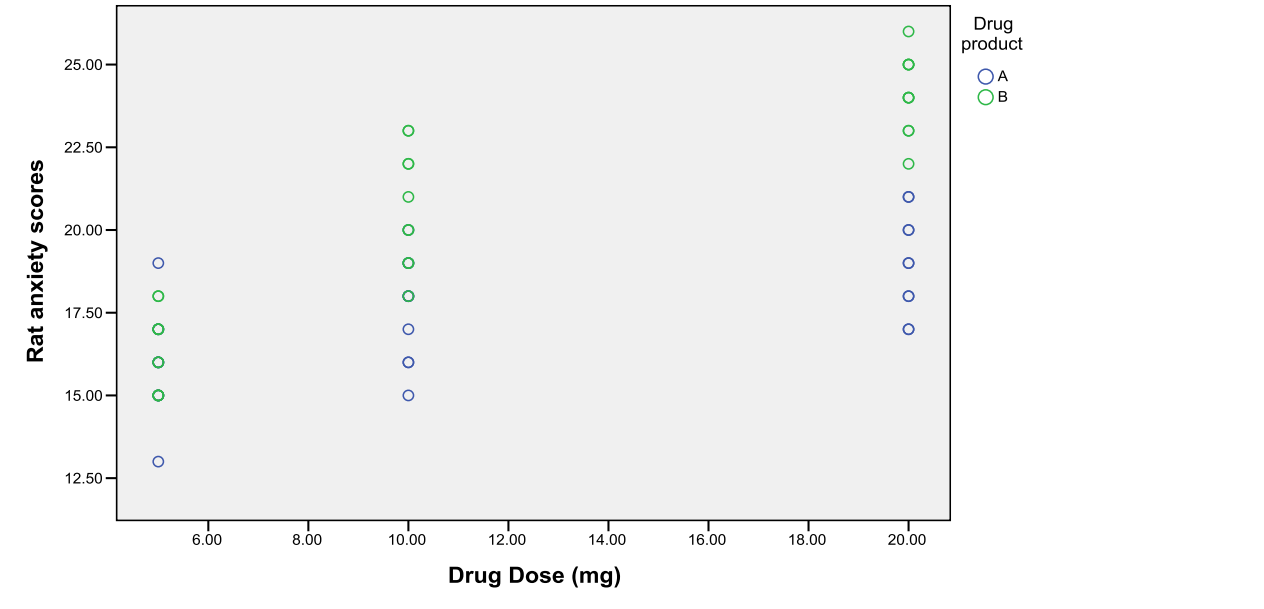
Drugs administered to 60 rats, in a number of doses for each drug.

y = anxiety score,

x_1 = dose (possible covariate),

x_2 = 1 for drug B, 0 for drug A (dummy variable)

We assume a linear relationship, i.e. a linear effect of dose on expected anxiety.



Anxiety of rats, 1st try: parallel lines

First assume that the effect of dose on anxiety is the same for drugs A and B: the slopes of the regression lines for A and B are the same, the lines are parallel.

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ so $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

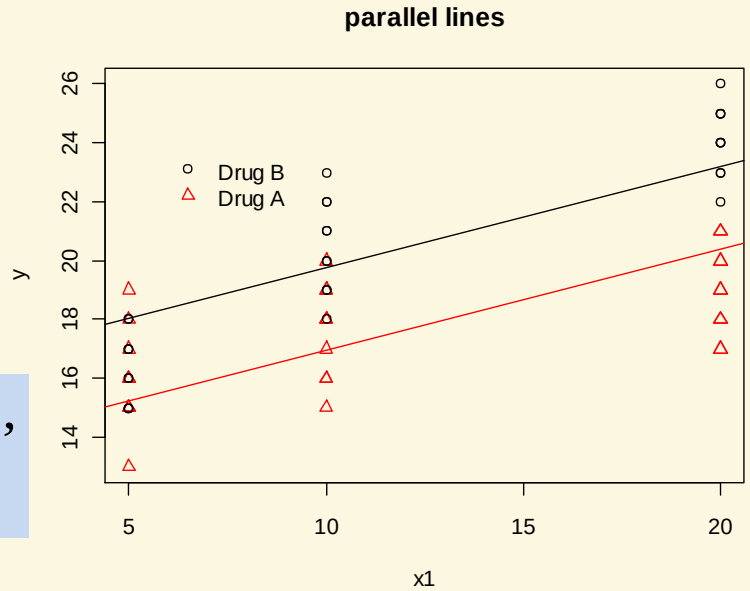
For drug A $x_2 = 0$, so $\mu_{yA} = \beta_0 + \beta_1 x_1$

For drug B $x_2 = 1$, so $\mu_{yB} = (\beta_0 + \beta_2) + \beta_1 x_1$

- intercept for A: β_0
  intercept for B: $\beta_0 + \beta_2$
- common slope for A and B: β_1

Equivalent model form: (i=1,2)

$$y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \varepsilon_{ij}, \quad \text{with } \tau_2 = 0$$



Model form is as the ANCOVA model. But effect of dose (x_1 / x) is also of interest in itself. So, the aims of the analysis are different.

Anxiety of rats, 2nd try: different slopes

- Now, we allow for a different effect of dose on anxiety for drugs A and B: the slopes of the regression lines may be different.

- Add product x_1x_2 as a third variable ($x_3 = x_1 * x_2$) to model:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- for drug A: $x_2 = 0$, $x_3 = 0$, so $\mu_y = \beta_0 + \beta_1 x_1$

  for drug B: $x_2 = 1$, $x_3 = x_1$, so $\mu_y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$

- intercept for A: β_0

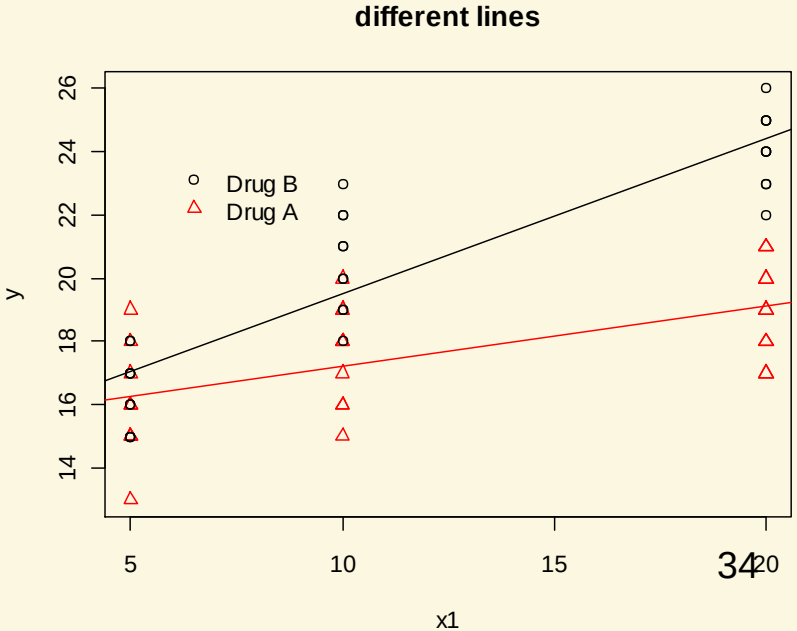
  intercept for B: $\beta_0 + \beta_2$

- slope for A: β_1

  slope for B: $\beta_1 + \beta_3$

drug A is the reference

- β_3: difference in slope between B and A
- β_2 : difference in intercept between B and A



Anxiety of rats, different slopes: regression output SPSS

- Regression of anxiety on dose, dummy coding for drug B, and product x_3.
- ANOVA table

SSR

SSE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	442.105	3	147.368	61.758	.000[a]
	Residual	133.629	56	2.386		
	Total	575.733	59			

a. Predictors: (Constant), x3, dose, dummy (drug B)

b. Dependent Variable: anxiety score

F-test "no drug- or x-effect"

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

- Parameter estimates

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	15.300	.598		25.573	.000
	dose	.191	.045	.385	4.233	.000
	dummy (drug B)	-.700	.846	-.113	-.827	.412
	x3	.300	.064	.708	4.691	.000

a. Dependent Variable: anxiety score

Estimated line drug A: $\hat{y} = 15.3 + 0.191 x_1$

Estimated line drug B: $\hat{y} = (15.3 - 0.700) + (0.191 + 0.3)\, x_1 = 14.6 + 0.491 x_1$

Model in GLM form

The interaction model in GLM form looks as follows:

$$y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \lambda_i x_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij}\text{'s indep from } N(0, \sigma)$$

$$\text{with } \tau_2 = 0 \quad \lambda_2 = 0 \qquad i = 1,2; \quad j = 1,2,\ldots,30$$

Two intercepts: β_0 and $\beta_0 + \tau_1$ (for groups 2 and 1, resp.)

Two slopes: β_1 and $\beta_1 + \lambda_1$ (for groups 2 and 1, resp.)

Parameter Estimates

Dependent Variable: anxiety score

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	14.600	.598	24.403	.000	13.402	15.798
[drug=A]	.700	.846	.827	.412	-.995	2.395
[drug=B]	0[a]	.	.	.	.	.
x1	.491	.045	10.866	.000	.401	.582
[drug=A] * x1	-.300	.064	-4.691	.000	-.428	-.172
[drug=B] * x1	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

GLM – output for rat example

Are Slopes signi-ficantly different?

Estimated line drug A: $\hat{y} = (14.6 + 0.7) + (0.491 - 0.31)x_1 = 15.3 + 0.191x_1$

Estimated line drug B: $\hat{y} = 14.6 + 0.491x_1$

SPSS output, General Linear Model (GLM – Univariate)

- See ANOVA table. Interaction is significant. So the two lines have different slopes.
- Suppose we want to know if x1 is useful in the model. Compare:

FM: drug, x1 and interaction

RM: only drug

$$F = \frac{\Delta SSE / \Delta dfE}{MSE_{FM}}, \; under \; H_0 \;\; F \sim F(2,56)$$

Tests of Between-Subjects Effects

Dependent Variable: anxiety score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	442.105[a]	3	147.368	61.758	.000
Intercept	2980.033	1	2980.033	1248.849	.000
drug	1.633	1	1.633	.684	.412
x1	272.005	1	272.005	113.990	.000
drug * x1	52.500	1	52.500	22.001	.000
Error	133.629	56	2.386		
Total	22084.000	60			
Corrected Total	575.733	59			

a. R Squared = .768 (Adjusted R Squared = .755)


THAT'S ALL FOLKS!
memegenerator.net