Analysis of Covariance

Analysis of covariance

- ANOVA: explanatory variables categorical (divide data into groups)

- traditionally, analysis of covariance has categorical x's plus one numerical x ("covariate") to be adjusted for.

- `lm` handles this too.

- Simple example: two treatments (drugs) (a and b), with before and after scores.

- Does knowing before score and/or treatment help to predict after score?

- Is after score different by treatment/before score?

Data

Treatment, before, after:

```
a 5 20
a 10 23
a 12 30
a 9 25
a 23 34
a 21 40
a 14 27
a 18 38
a 6 24
a 13 31
b 7 19
b 12 26
b 27 33
b 24 35
b 18 30
b 22 31
b 26 34
b 21 28
b 14 23
b 9 22
```

Packages

tidyverse and broom:

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- ti
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(broom)
library(marginaleffects)
```

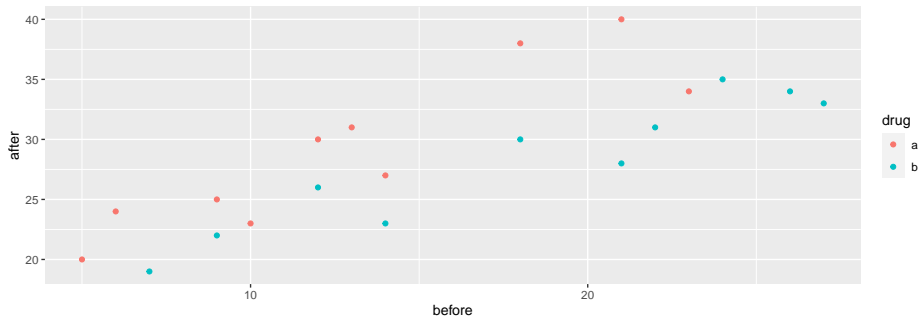# Read in data

```
url <- "http://ritsokiguess.site/datafiles/ancova.txt"
prepost <- read_delim(url, " ")
prepost %>% sample_n(9) # randomly chosen rows
```

drug	before	after
b	24	35
a	21	40
b	26	34
a	9	25
b	21	28
a	12	30
b	27	33
a	10	23
b	22	31

Making a plot

```
ggplot(prepost, aes(x = before, y = after, colour = drug)) +
  geom_point()
```

Comments

- As before score goes up, after score goes up.

- Red points (drug A) generally above blue points (drug B), for comparable before score.

- Suggests before score effect *and* drug effect.

The means

```
prepost %>%
  group_by(drug) %>%
  summarize(
    before_mean = mean(before),
    after_mean = mean(after)
  )
```

drug	before_mean	after_mean
a	13.1	29.2
b	18.0	28.1

- Mean "after" score slightly higher for treatment A.

- Mean "before" score much higher for treatment B.

- Greater *improvement* on treatment A.

Testing for interaction

```
prepost.1 <- lm(after ~ before * drug, data = prepost)
anova(prepost.1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
before	1	430.92384	430.923838	62.68945	0.0000006
drug	1	115.30596	115.305957	16.77435	0.0008442
before:drug	1	12.33708	12.337080	1.79476	0.1990662
Residuals	16	109.98313	6.873945	NA	NA

- Interaction not significant. Will remove later.

Predictions

```
predictions(prepost.1, variables = c("before", "drug"))
```

rowid	type	predicted	std.error	conf.low	conf.high	after	before	drug
1	response	21.29948	1.4347929	18.25786	24.34111	28.65	5.0	a
2	response	18.71739	1.7924347	14.91760	22.51718	28.65	5.0	b
3	response	25.68866	0.9789055	23.61347	27.76385	28.65	9.5	a
4	response	21.96522	1.3293052	19.14722	24.78322	28.65	9.5	b
5	response	30.07784	0.8392398	28.29873	31.85694	28.65	14.0	a
6	response	25.21304	0.9625426	23.17254	27.25354	28.65	14.0	b
7	response	37.39313	1.4704029	34.27601	40.51024	28.65	21.5	a
8	response	30.62609	0.9329795	28.64826	32.60392	28.65	21.5	b
9	response	42.75768	2.1738090	38.14941	47.36595	28.65	27.0	a
10	response	34.59565	1.3776090	31.67525	37.51605	28.65	27.0	b

Predictions (with interaction included), plotted

```
plot_cap(model = prepost.1, condition = c("before", "drug"))
```
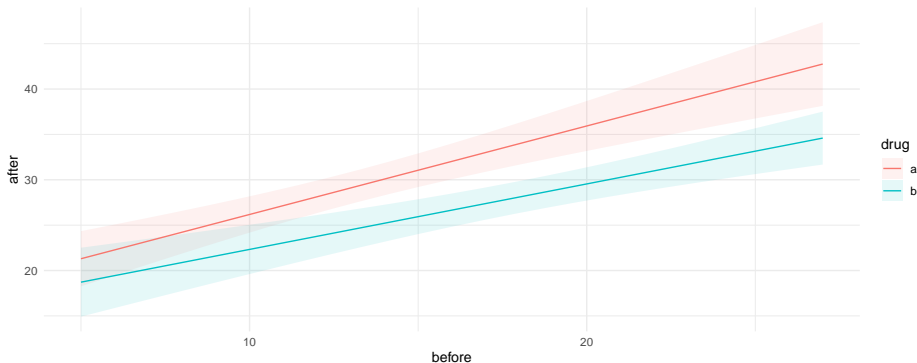

Lines almost parallel, but not quite.

Taking out interaction

```
prepost.2 <- update(prepost.1, . ~ . - before:drug)
anova(prepost.2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
before	1	430.9238	430.923838	59.88958	0.0000006
drug	1	115.3060	115.305957	16.02516	0.0009209
Residuals	17	122.3202	7.195306	NA	NA

- Take out non-significant interaction.

- `before` and `drug` strongly significant.

- Do predictions again and plot them.

Predictions

```
predictions(prepost.2, variables = c("before", "drug"))
```

rowid	type	predicted	std.error	conf.low	conf.high	after	before	drug
1	response	22.49740	1.1480151	20.07530	24.91950	28.65	5.0	a
2	response	17.34274	1.5036376	14.17035	20.51514	28.65	5.0	b
3	response	26.22107	0.9152785	24.29000	28.15214	28.65	9.5	a
4	response	21.06641	1.1740954	18.58928	23.54353	28.65	9.5	b
5	response	29.94473	0.8525951	28.14591	31.74355	28.65	14.0	a
6	response	24.79007	0.9303019	22.82731	26.75284	28.65	14.0	b
7	response	36.15084	1.1675127	33.68761	38.61408	28.65	21.5	a
8	response	30.99618	0.9117340	29.07259	32.91978	28.65	21.5	b
9	response	40.70199	1.5753532	37.37829	44.02569	28.65	27.0	a
10	response	35.54733	1.2076034	32.99951	38.09515	28.65	27.0	b

Plot of predicted values

```
plot_cap(prepost.2, condition = c("before", "drug"))
```
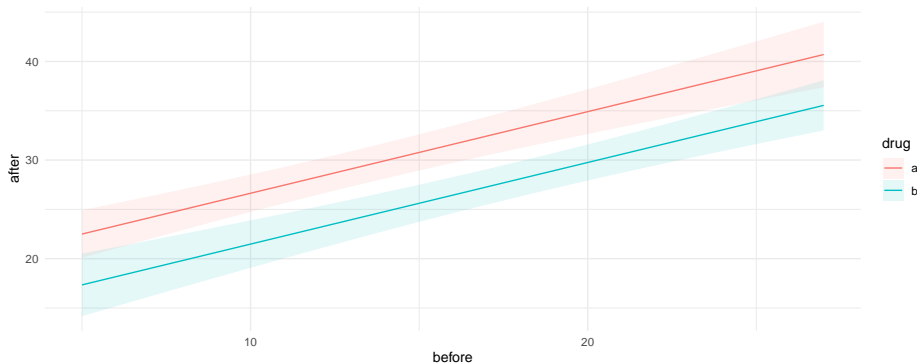

This time the lines are *exactly* parallel. No-interaction model forces them to have the same slope.

Different look at model output

- anova(prepost.2) tests for significant effect of before score and of drug, but doesn't help with interpretation.

- summary(prepost.2) views as regression with slopes:

```
summary(prepost.2)
```

```
##
## Call:
## lm(formula = after ~ before + drug, data = prepost)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6348 -2.5099 -0.2038  1.8871  4.7453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.3600     1.5115  12.147 8.35e-10 ***
## before        0.8275     0.0955   8.665 1.21e-07 ***
## drugb        -5.1547     1.2876  -4.003 0.000921 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Understanding those slopes

`tidy(prepost.2)`

term	estimate	std.error	statistic	p.value
(Intercept)	18.3599949	1.5115326	12.146608	0.0000000
before	0.8274813	0.0955023	8.664520	0.0000001
drugb	-5.1546584	1.2876524	-4.003144	0.0009209

- `before` ordinary numerical variable; `drug` categorical.

- `lm` uses first category `druga` as baseline.

- Intercept is prediction of after score for before score 0 and *drug A*.

- `before` slope is predicted change in after score when before score increases by 1 (usual slope)

- Slope for `drugb` is *change* in predicted after score for being on drug B rather than drug A. Same for *any* before score (no interaction).

Summary

- ANCOVA model: fits different regression line for each group, predicting response from covariate.

- ANCOVA model with interaction between factor and covariate allows different slopes for each line.

- Sometimes those lines can cross over!

- If interaction not significant, take out. Lines then parallel.

- With parallel lines, groups have consistent effect regardless of value of covariate.