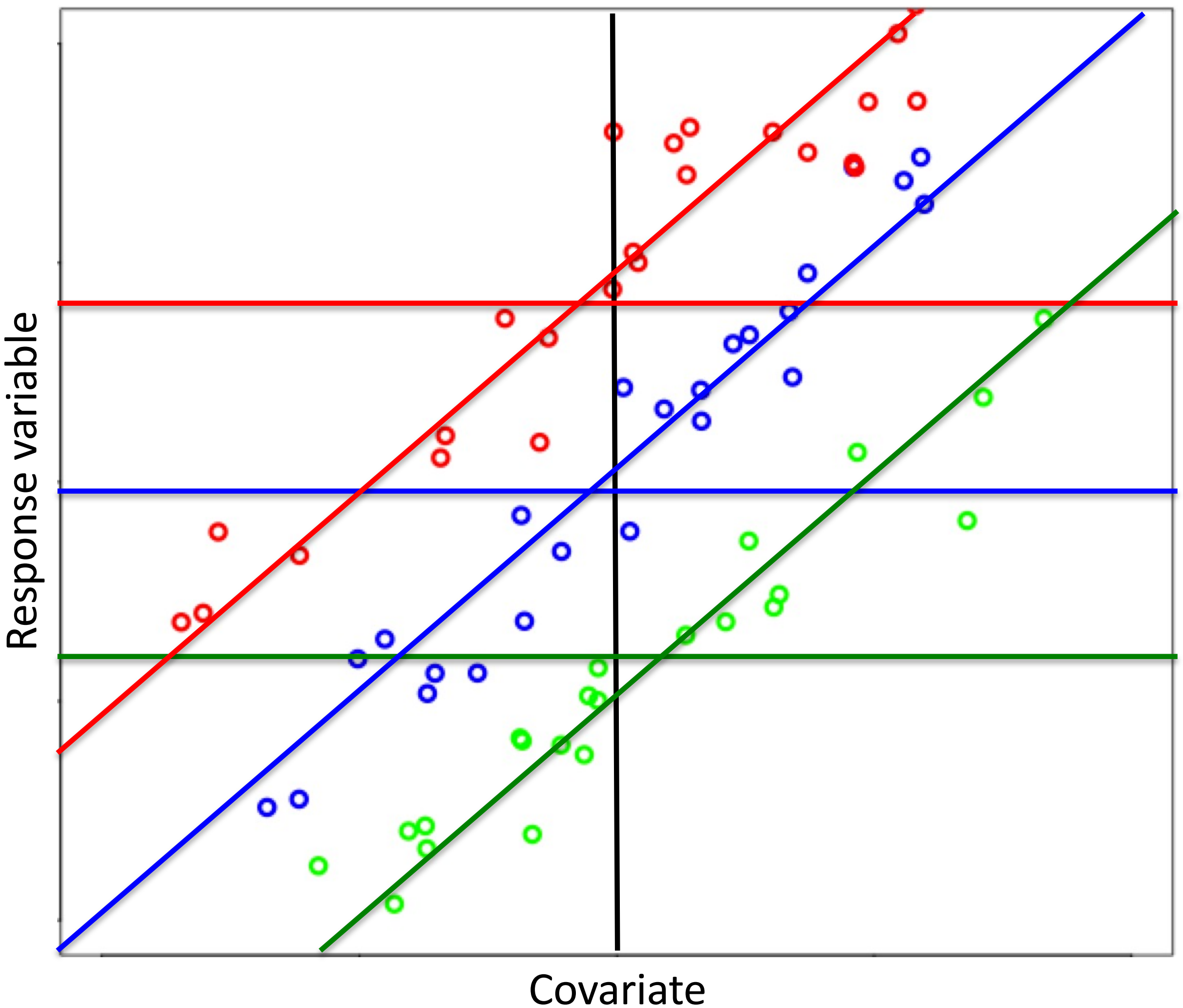
201ab Quantitative methods
ANCOVA

Ed Vul | UCSD Psychology

What does ANCOVA do?



In an **ANOVA**, we compare the variation in means of the response/dependent variable across factor levels to the remaining variability around the means.

In an **ANCOVA**, we compare the variation in intercepts across factor levels of the regression of the response/dependent variable as a function of the covariate. Thus, we can potentially greatly reduce residual error, if the covariate accounts for lots of it.

Setting up an ANCOVA analysis

```
anova(lm(data=dat, logwealth~sat+major))
```

```
              Df  Sum Sq Mean Sq F-value    Pr(>F)
sat            1 114.341 114.341 146.649 9.313e-16 ***
major          3 209.582  69.861  89.601 < 2.2e-16 ***
Residuals 45  35.086   0.780
```

Notes:

1) The model includes the covariate first, to factor out its effects before ascertaining effects of major (for sequential sums of squares).
2) The covariate takes 1 degree of freedom
   (extra covariates would take one each – a covariate is just a single numerical predictor which requires one coefficient as in ordinary regression)
3) We do NOT include the interaction between covariate:factor
4) The rest of the ANOVA proceeds as normal: F = MS[factor]/MS[error]

Why / When to use an ANCOVA

- You have some measure taken *before* your manipulation, and you think it might influence your response variable and contribute to variability.
  - E.g., parents' height will predict child's height, and you can measure parents' heights before manipulating nutrition.
  - E.g., IQ will influence response times, and you can measure it before administering your implicit attitudes test.
  - E.g., Word frequency will influence completion rates, and you can measure word frequency from a corpus beforehand.
- So you add this measure as a covariate to explain some variability in the response, and hopefully reduce residual error.

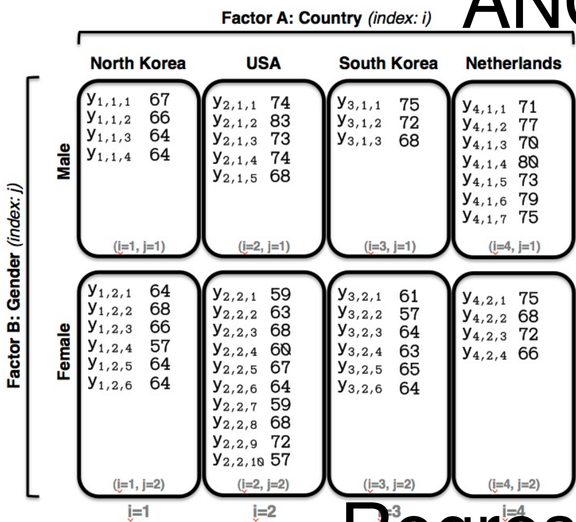
Why / When to use an ANCOVA

- You have some non-randomly assigned study, and want to argue that factor X influences response Y even after you '*control for*' all these other things that might relate to X and Y.
  - E.g., does religion predicts voting preference even when you control for income.
  - E.g., do gun control laws reduce crime even when you control for countries' economy.
  - E.g., do women get paid less even when you control for work hours?
- So you add these potential explanatory variables to factor out their effects, and 'control' for these variables.

When NOT to use ANCOVA

- When your covariate was measured *after* your manipulation, and your manipulation might influence the covariate.
- When your ANOVA doesn't work, and you get desperate, and try various covariates in hopes of getting $p<0.05$.
- When the covariate-response relationship changes with factor level (large factor:covariate interaction).
- When accounting for pre-test performance on the same task. (Repeated measures, take difference!)

ANCOVA and the general linear model



ANOVA: categorical explanatory variable(s)

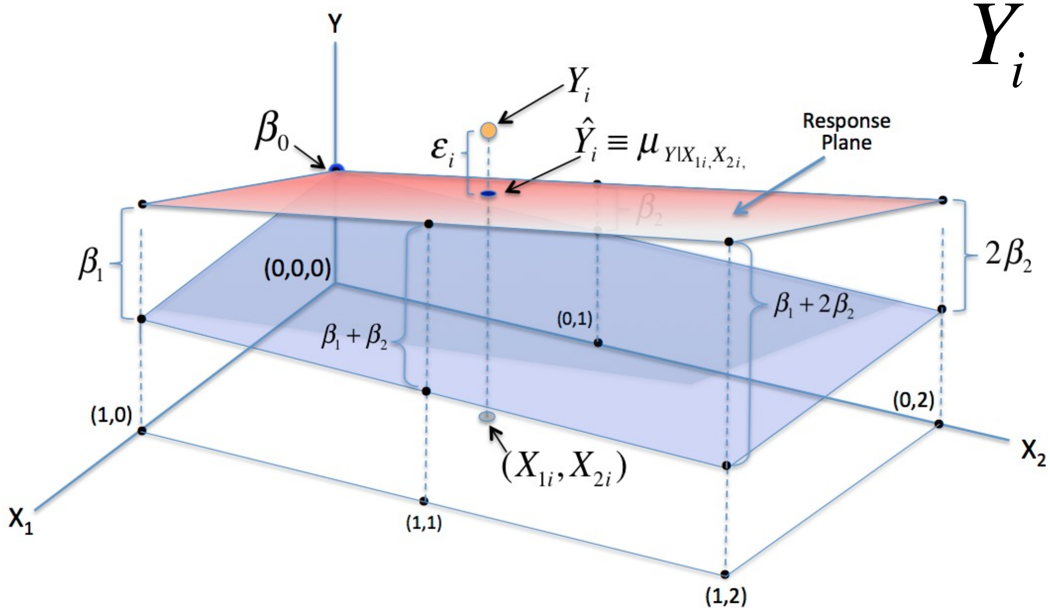
$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

Regressors are indicator / dummy variables used to code various factor levels

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

Regression: continuous explanatory variable(s)

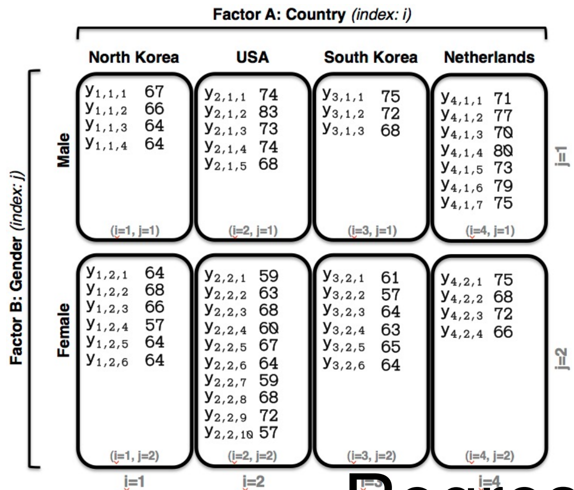
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$



Regressors are continuous variables.

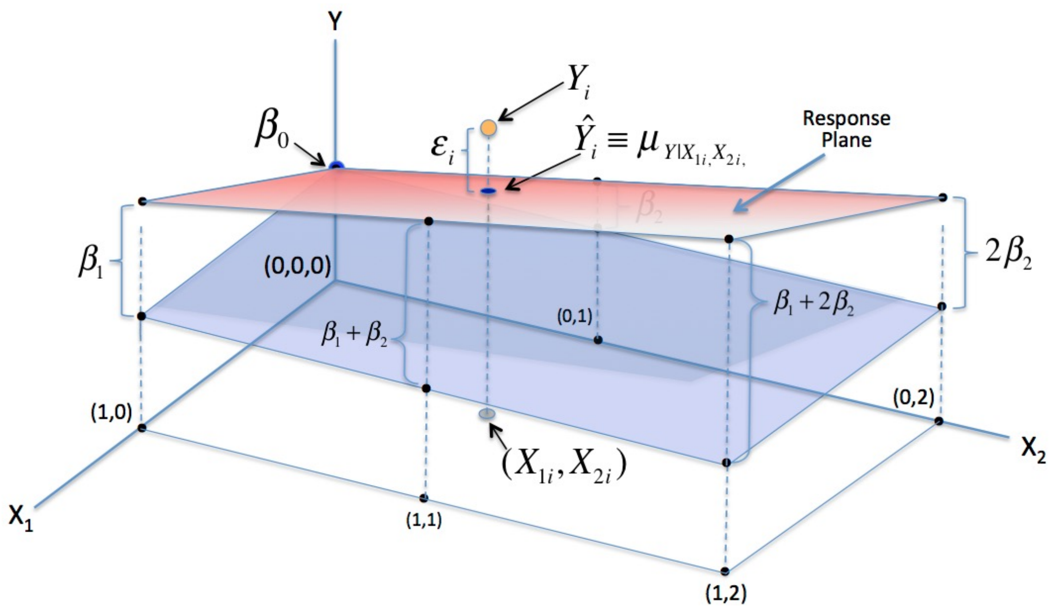
ANCOVA and the general linear model

ANOVA: categorical explanatory variable(s)



$$\begin{bmatrix} 65 \\ 72 \\ 70 \\ 58 \\ 63 \\ \dots \\ 69 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

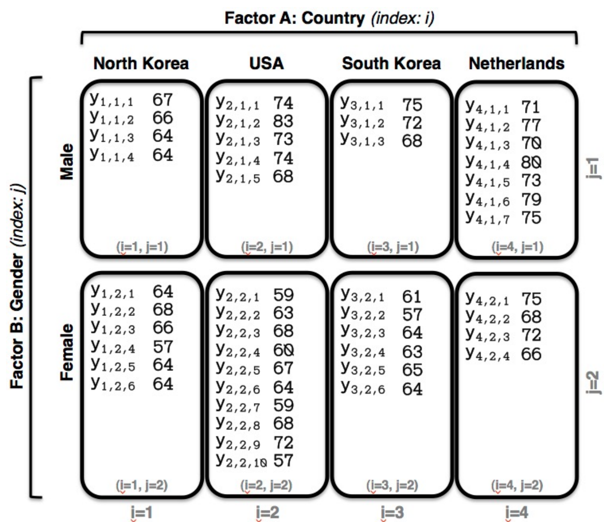
Regression: continuous explanatory variable(s)



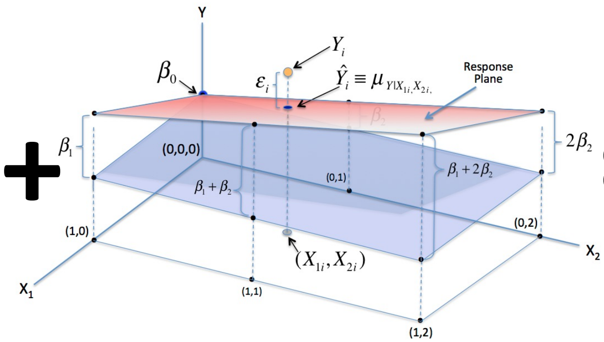
$$\begin{bmatrix} 65 \\ 72 \\ 70 \\ 58 \\ 63 \\ \dots \\ 69 \end{bmatrix} = \begin{bmatrix} 1 & 40 & 4.1 \\ 1 & 42 & 2.5 \\ 1 & 50 & 1.8 \\ 1 & 37 & 6.1 \\ 1 & 31 & -4.3 \\ \dots & \dots & \dots \\ 1 & 34 & -3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

ANCOVA and the general linear model

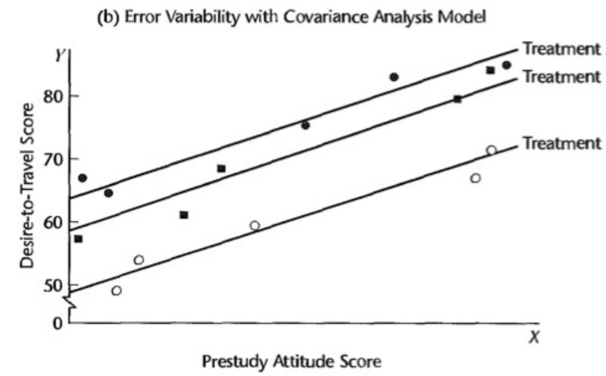
ANOVA + Regression = ANCOVA



+



=



$$\begin{bmatrix}65\\72\\70\\58\\63\\\dots\\69\end{bmatrix}=\begin{bmatrix}1&1&0&0&1\\1&1&0&0&0\\1&0&1&0&1\\1&0&1&0&0\\1&0&0&0&1\\\dots&\dots&\dots&\dots&\dots\\1&0&0&1&0\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\\\beta_3\\\beta_4\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\\\dots\\\varepsilon_n\end{bmatrix}$$

+

$$\begin{bmatrix}65\\72\\70\\58\\63\\\dots\\69\end{bmatrix}=\begin{bmatrix}1&40&4.1\\1&42&2.5\\1&50&1.8\\1&37&6.1\\1&31&-4.3\\\dots&\dots&\dots\\1&34&-3\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\\\dots\\\varepsilon_n\end{bmatrix}$$

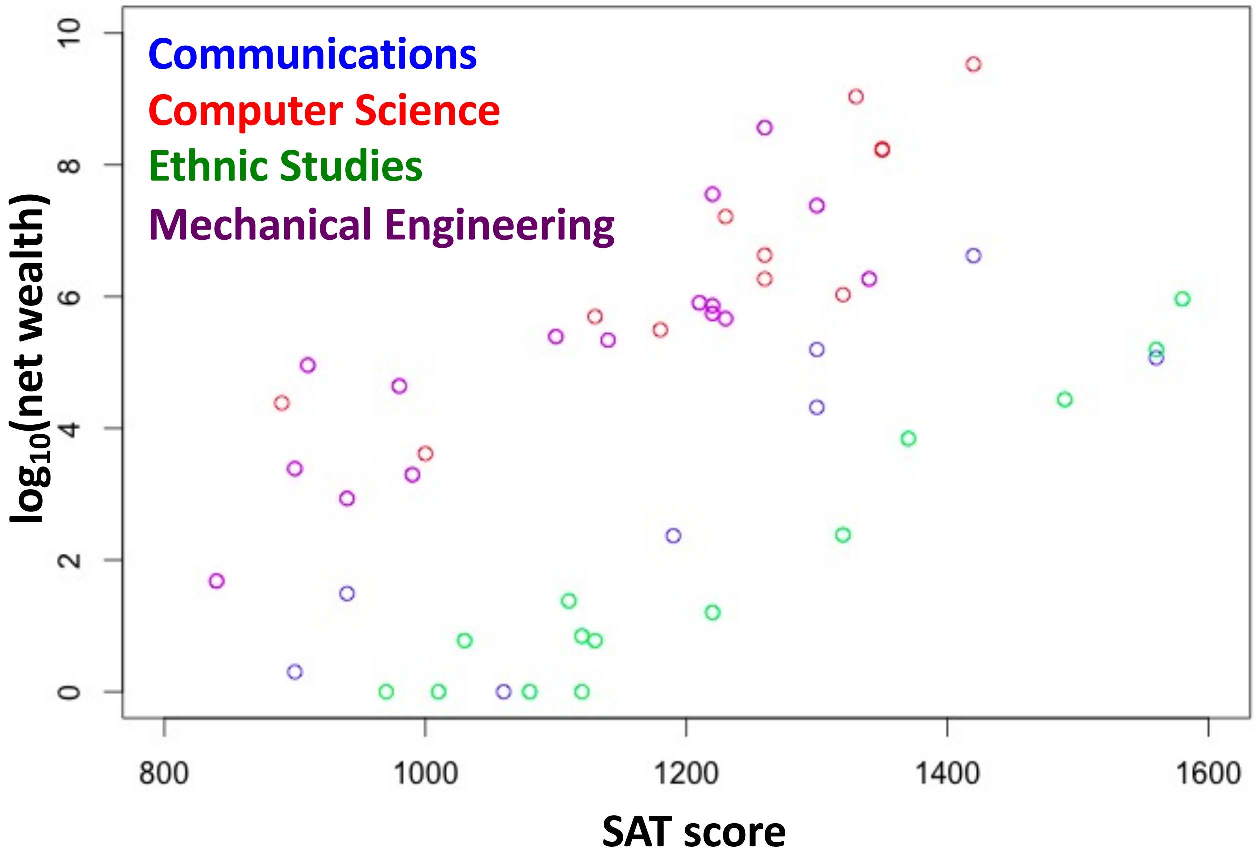
=

$$\begin{bmatrix}65\\72\\70\\58\\63\\\dots\\69\end{bmatrix}=\begin{bmatrix}1&1&0&0&1&40&4.1\\1&1&0&0&0&42&2.5\\1&0&1&0&1&50&1.8\\1&0&1&0&0&37&6.1\\1&0&0&0&1&31&-4.3\\\dots&\dots&\dots&\dots&\dots&\dots&\dots\\1&0&0&1&0&34&-3\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\\\beta_3\\\beta_4\\\beta_5\\\beta_6\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\\\dots\\\varepsilon_n\end{bmatrix}$$

ANCOVA example

	wealth	major	sat
1	1853675	Computer Science	1260
2	555228	Mechanical Engineering	1220
3	24098788	Mechanical Engineering	1300
4	35821392	Mechanical Engineering	1220
5	730253	Mechanical Engineering	1220
6	858	Mechanical Engineering	940
7	3381613071	Computer Science	1420
8	803771	Mechanical Engineering	1210
9	0	Ethnic Studies	1010
10	47	Mechanical Engineering	840
11	1	Communications	900
12	0	Ethnic Studies	970
13	1087200128	Computer Science	1330
14	0	Ethnic Studies	1120
15	246737	Mechanical Engineering	1100
16	463904	Mechanical Engineering	1230
17	368096210	Mechanical Engineering	1260
18	497842	Computer Science	1130
19	27483	Ethnic Studies	1490
20	20879	Communications	1300
21	157541	Ethnic Studies	1560
22	2436	Mechanical Engineering	900
23	0	Ethnic Studies	1080
24	90659	Mechanical Engineering	910
25	23	Ethnic Studies	1110
26	0	Communications	1060
27	5	Ethnic Studies	1130
28	1975	Mechanical Engineering	990
29	5	Ethnic Studies	1030
30	6963	Ethnic Studies	1370
31	4119	Computer Science	1000
32	117315	Communications	1560
33	4269880	Computer Science	1260
34	167620906	Computer Science	1350
35	16402426	Computer Science	1230
36	1852979	Mechanical Engineering	1340
37	4194607	Communications	1420
38	6	Ethnic Studies	1120
39	15	Ethnic Studies	1220
40	218646	Mechanical Engineering	1140
41	233	Communications	1190
42	240	Ethnic Studies	1320
43	43827	Mechanical Engineering	980
44	312956	Computer Science	1180
45	30	Communications	940
46	24235	Computer Science	890
47	919366	Ethnic Studies	1580
48	157185	Communications	1300
49	1072256	Computer Science	1320

What is the effect of major on future wealth?



There are big effects of SAT score. Over and above that there are some intercept differences of major: the ideal setting for an ANCOVA.

ANCOVA example

	wealth	major	sat
1	1853675	Computer Science	1260
2	555228	Mechanical Engineering	1220
3	24098788	Mechanical Engineering	1300
4	35821392	Mechanical Engineering	1220
5	730253	Mechanical Engineering	1220
6	858	Mechanical Engineering	940
7	3381613071	Computer Science	1420
8	803771	Mechanical Engineering	1210
9	0	Ethnic Studies	1010
10	47	Mechanical Engineering	840
11	1	Communications	900
12	0	Ethnic Studies	970
13	1087200128	Computer Science	1330
14	0	Ethnic Studies	1120
15	246737	Mechanical Engineering	1100
16	463904	Mechanical Engineering	1230
17	368096210	Mechanical Engineering	1260
18	497842	Computer Science	1130
19	27483	Ethnic Studies	1490
20	20879	Communications	1300
21	157541	Ethnic Studies	1560
22	2436	Mechanical Engineering	900
23	0	Ethnic Studies	1080
24	90659	Mechanical Engineering	910
25	23	Ethnic Studies	1110
26	0	Communications	1060
27	5	Ethnic Studies	1130
28	1975	Mechanical Engineering	990
29	5	Ethnic Studies	1030
30	6963	Ethnic Studies	1370
31	4119	Computer Science	1000
32	117315	Communications	1560
33	4269880	Computer Science	1260
34	167620906	Computer Science	1350
35	16402426	Computer Science	1230
36	1852979	Mechanical Engineering	1340
37	4194607	Communications	1420
38	6	Ethnic Studies	1120
39	15	Ethnic Studies	1220
40	218646	Mechanical Engineering	1140
41	233	Communications	1190
42	240	Ethnic Studies	1320
43	43827	Mechanical Engineering	980
44	312956	Computer Science	1180
45	30	Communications	940
46	24235	Computer Science	890
47	919366	Ethnic Studies	1580
48	157185	Communications	1300
49	1072256	Computer Science	1320

```
anova(lm(data=dat, logwealth~major))
```

```
           Df Sum Sq Mean Sq F-value    Pr(>F)
major       3 174.28  58.092  14.465 9.033e-07 ***
Residuals  46 184.73   4.016
```

There are big effects of SAT score. ANCOVA factors those out.

```
anova(lm(data=dat, logwealth~sat+major))
```

```
           Df  Sum Sq Mean Sq F-value    Pr(>F)
sat         1 114.341 114.341 146.649 9.313e-16 ***
major       3 209.582  69.861  89.601 < 2.2e-16 ***
Residuals  45  35.086   0.780
```

(1) **We add the covariate (SAT) first.**
This way we interpret the main effect *after* factoring out the covariate. This is the standard approach (esp. for observational studies, where the goal is to *control* for the covariate).

(2) **Our residual sum of squares / variance drops a lot!**

(3) **Consequently the F value for major goes up a lot.**

(4) **SS[factor] shouldn't change much**
Here, SS[major] increased a bit – generally we expect it not to change (or maybe to drop if factoring out confounds).

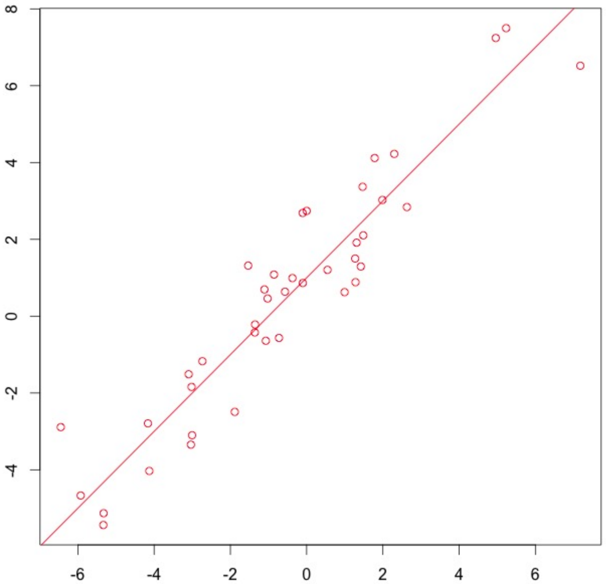
Test for the interaction

- Check for homogenous regression slopes by looking for the interaction.

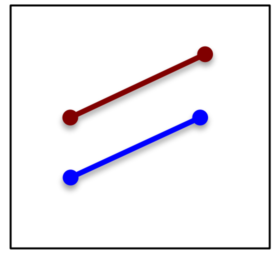
```
anova(lm(data=dat, logwealth~sat*major))
```

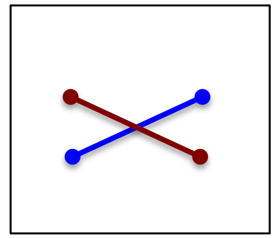
```
            Df  Sum Sq Mean Sq  F value    Pr(>F)
sat          1 114.341 114.341 137.3731  7.98e-15 ***
major        3 209.582  69.861  83.9333 < 2.2e-16 ***
sat:major    3   0.128   0.043   0.0512    0.9845
Residuals 42  34.958   0.832
```

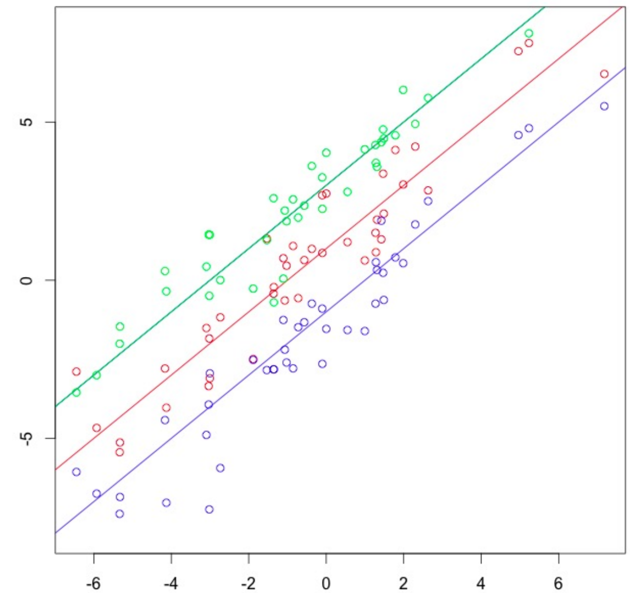
- Interaction between factor and continuous variables means: different slope as a function of factor level.
- Generally: check for interaction, but do not include it in the ANCOVA model (because if you include it, it is no longer ANCOVA, and significance of factor loses its meaning!)



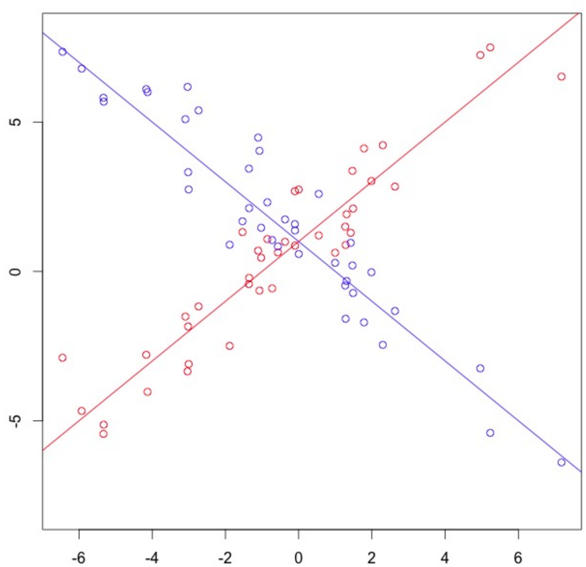
Main effect of continuous variable x: slope of y as a function of x is not 0.

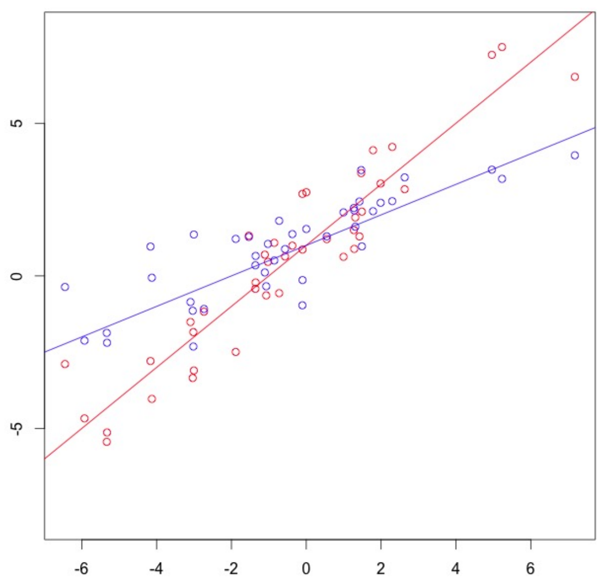


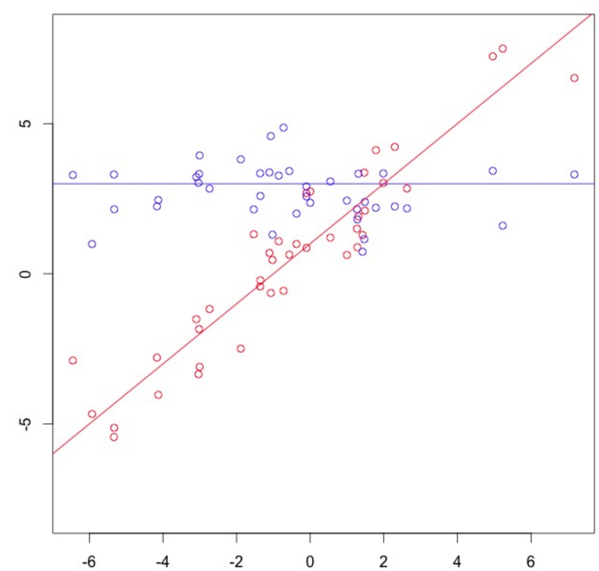




Main effect of qualitative variable (color): intercepts differ across colors.

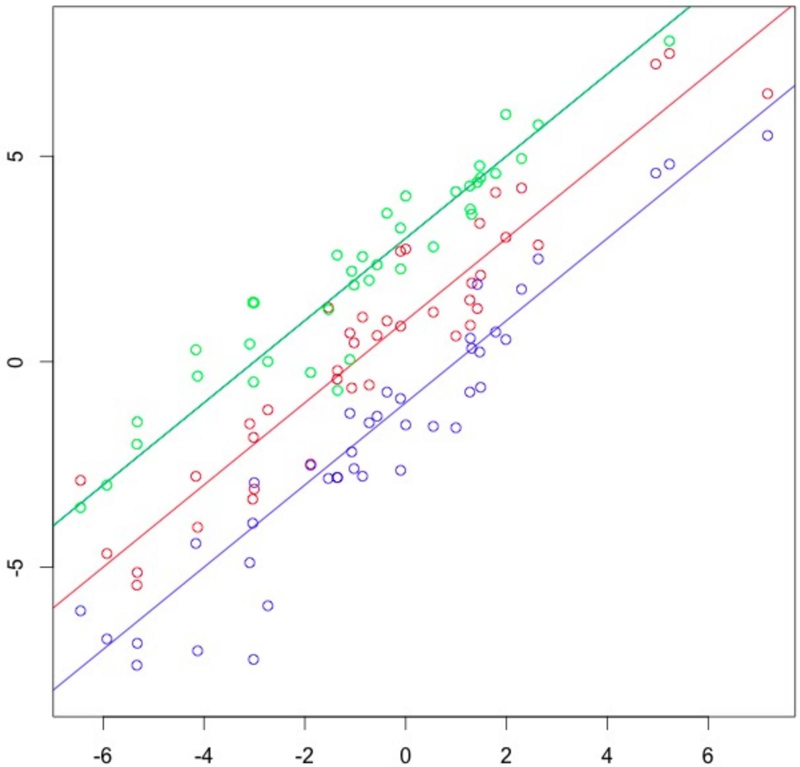




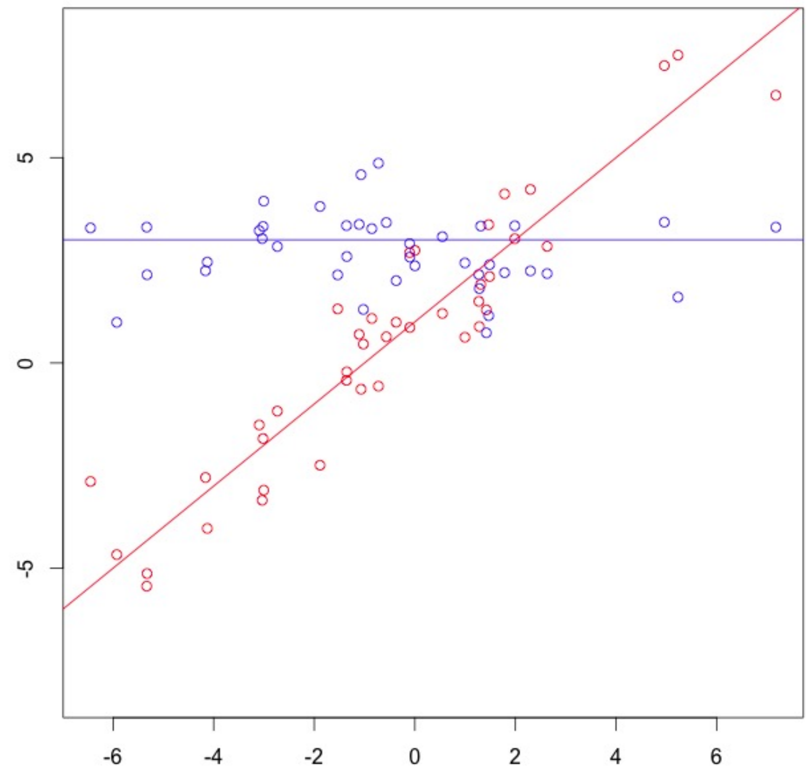


Interaction of continuous *x* and qualitative *color* variable: slope of y as a function of x differs across colors.

ANCOVA: varying intercepts.



ANCOVA: a constant slope on the covariate, and the intercept varies with factor level. Main effect of factor interpreted as differences in additive offsets for factors levels.



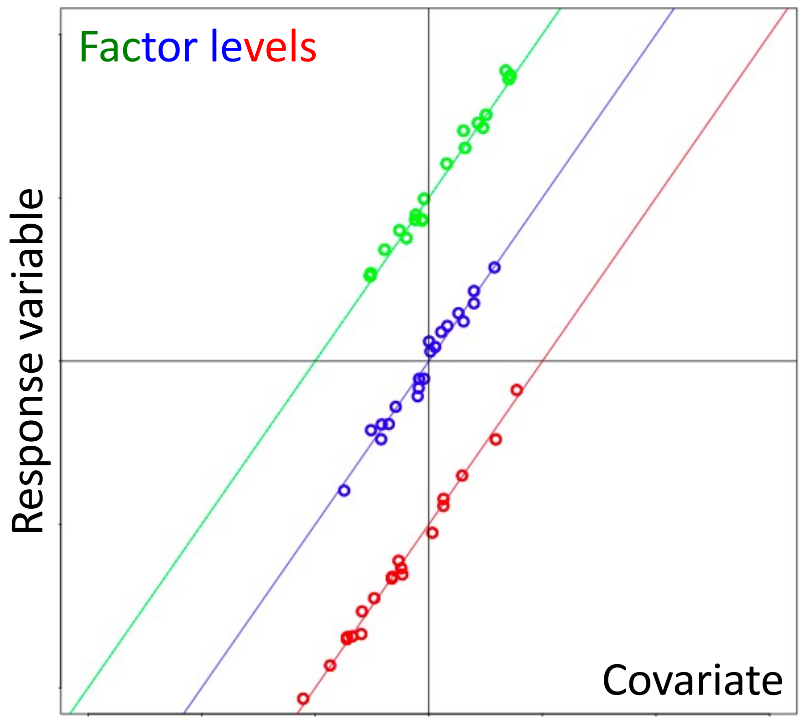
A factor*covariate interaction: slopes van vary as a function of factor level. Main effect of factor is still the difference in intercepts, but those are no longer meaningful.

This is NOT an ANCOVA!

Ideal ANOVA/ANCOVA result pattern

ANCOVA compared to ANOVA:

SS[error] drops, SS[factors] about the same

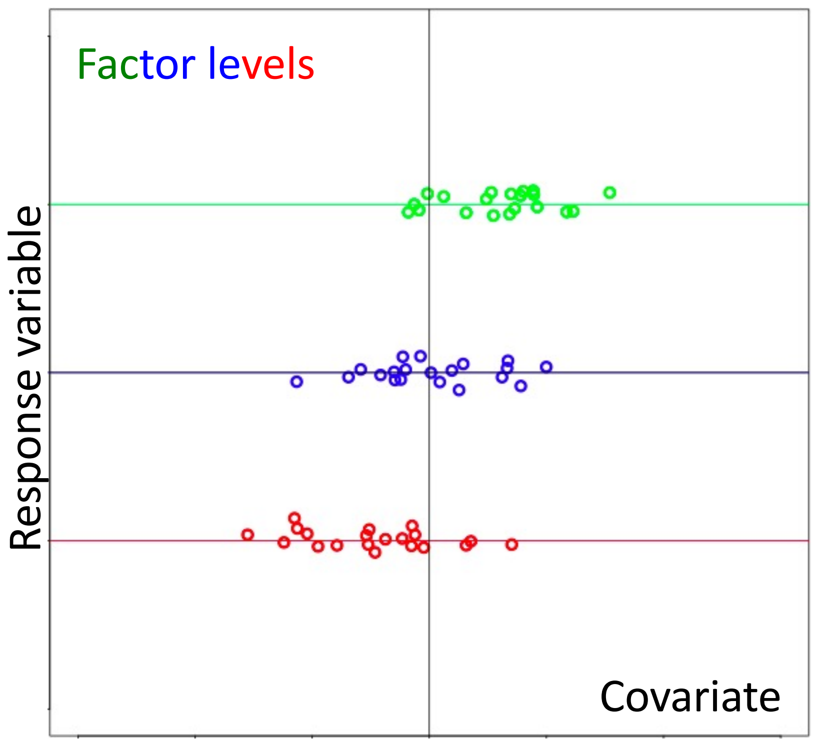


Covariate is constant with factor, and response variable changes with covariate. Thus, adding the covariate just factors out what would look like noise.

Bland ANOVA/ANCOVA result pattern

ANCOVA compared to ANOVA:

Nothing really changes.

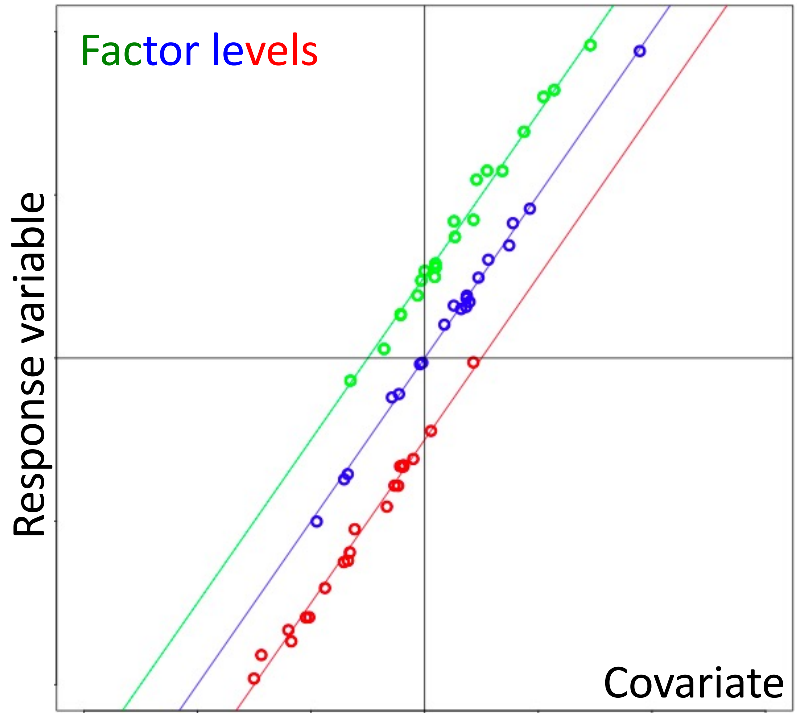


Covariate has no relationship with response variable.

Unfortunate ANOVA/ANCOVA results

ANCOVA compared to ANOVA:

SS[factor] drops

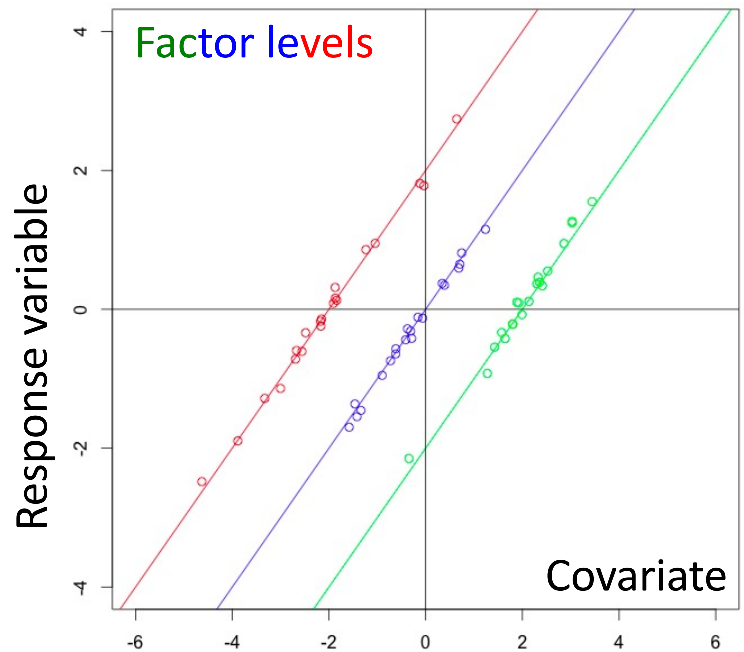


Covariate has relationship with response, and with factor, in the same direction. Thus, 'controlling' for covariate reduces apparent factor effect.

Weird ANOVA/ANCOVA results pattern

ANCOVA compared to ANOVA:

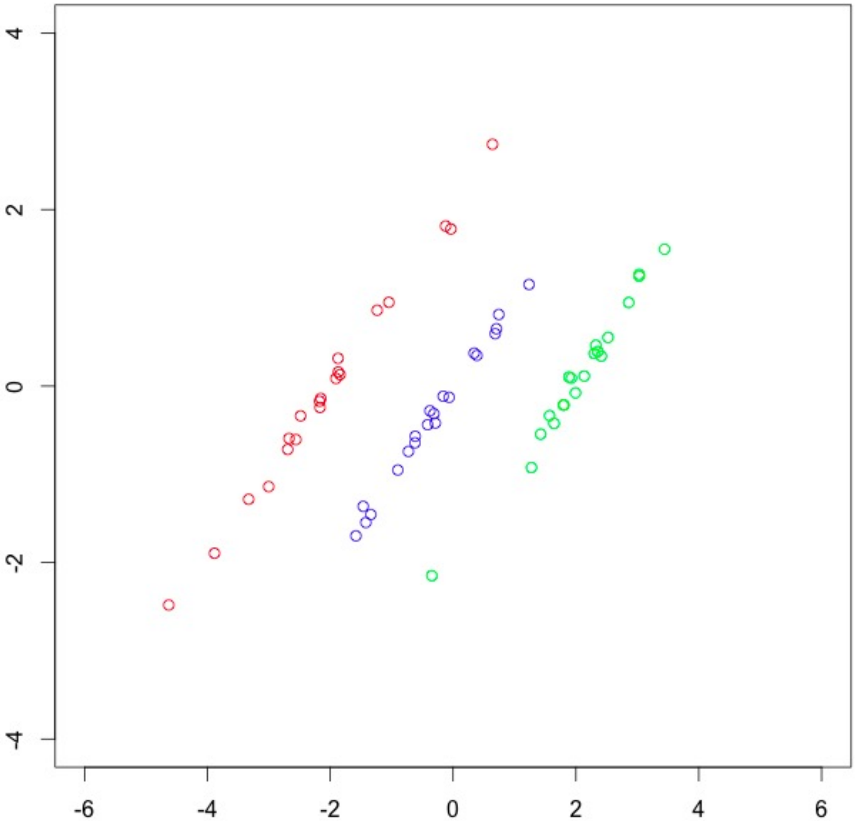
SS[factors] goes up!



Covariate has relationship with response variable and with factor, but in a different direction than the factor-response relationship. Thus they cancel each other out in the ANOVA, but not the ANCOVA.

Weird patterns: SS[factor] goes up.

When covariates are correlated with factor.



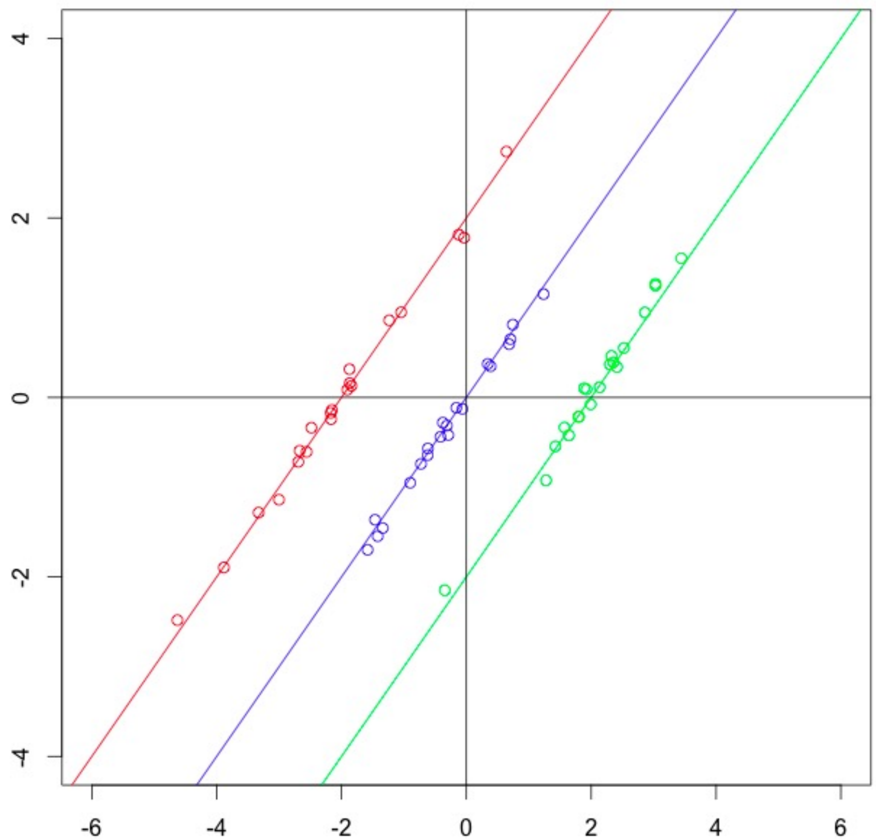
```
anova(lm(ys~cats))

Response: ys
          Df Sum Sq Mean Sq F value Pr(>F)
cats       2  2.211 1.10552  1.1278 0.3309
Residuals 57 55.876 0.98027
```

```
anova(lm(ys~xs+cats))

Response: ys
          Df Sum Sq Mean Sq F value    Pr(>F)
xs         1 18.015 18.0153  1906.7 < 2.2e-16 ***
cats       2 39.542 19.7711  2092.5 < 2.2e-16 ***
Residuals 56  0.529  0.0094
```

SS[factor] went up a lot

Weird patterns: SS[factor] goes up.

When covariates are correlated with factor.




```
anova(lm(ys~cats))

Response: ys
          Df Sum Sq Mean Sq F value Pr(>F)
cats       2  2.211 1.10552  1.1278 0.3309
Residuals 57 55.876 0.98027
```

```
anova(lm(ys~xs+cats))

Response: ys
          Df Sum Sq Mean Sq F value    Pr(>F)
xs         1 18.015 18.0153  1906.7 < 2.2e-16 ***
cats       2 39.542 19.7711  2092.5 < 2.2e-16 ***
Residuals 56  0.529  0.0094
```

SS[factor] went up a lot

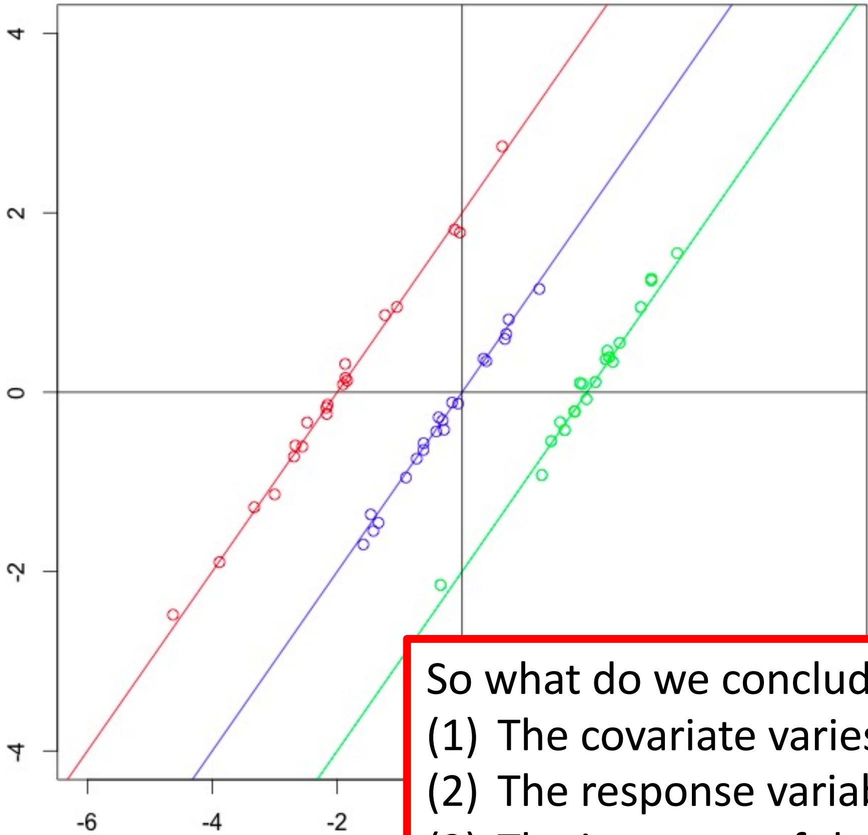
```
anova(lm(xs~cats))

Response: xs
          Df  Sum Sq Mean Sq F value    Pr(>F)
cats       2 171.145  85.573  87.821 < 2.2e-16 ***
Residuals 57  55.541   0.974
```

Covariate varies substantially with factor.

Means of the categories (in y) don't differ.
Means of the categories (in covariate) differ a lot.
Consequently, y-intercepts from ANCOVA differ a lot.

Weird patterns: SS[factor] goes up.

When covariates are correlated with factor.




```
anova(lm(ys~cats))

Response: ys
          Df Sum Sq Mean Sq F value Pr(>F)
cats       2  2.211 1.10552  1.1278 0.3309
Residuals 57 55.876 0.98027
```

```
anova(lm(ys~xs+cats))

Response: ys
          Df Sum Sq Mean Sq F value    Pr(>F)
xs         1 18.015 18.0153  1906.7 < 2.2e-16 ***
cats       2 39.542 19.7711  2092.5 < 2.2e-16 ***
Residuals 56  0.529  0.0094
```

SS[factor] went up a lot

```
F)
-16 ***
```

actor.

So what do we conclude?

(1) The covariate varies across factor levels.

(2) The response variable varies with the covariate.

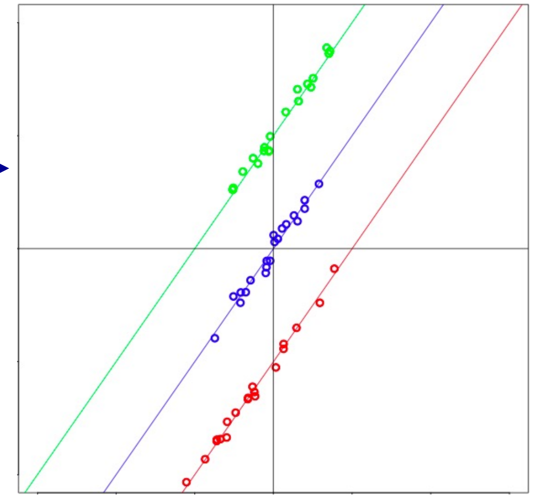
(3) The intercept of the covariate-response relationship varies across factor levels in such a way as to cancel out the factor -> covariate -> response relationship.

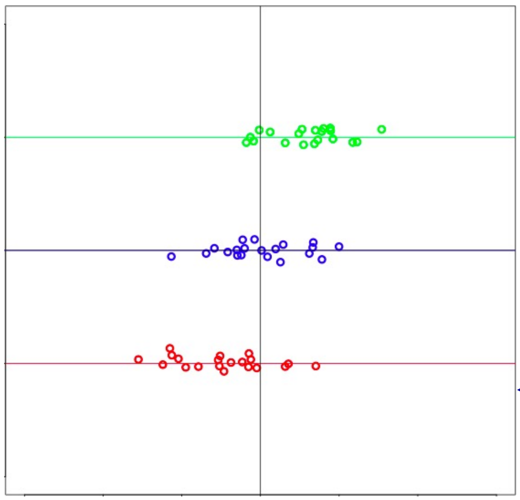
This is weird, and hard to interpret.

Interpreting ANCOVA results

- How might ANOVA and ANCOVA results differ?
  - SS[error] drops; SS[factors] ~ the same: Great! This is what ANCOVA is supposed to do!
  - SS[factors] drops: Bound to happen (esp. when using covariate as control) – means that covariate and factors are correlated.
  - Nothing changes much: covariate not correlated with factors or response variable.
    (literally nothing changes: very unlikely)
  - SS[factors] goes up: uh oh! (esp if null at first): covariate is correlated with factors, and correlated with response variable, but these correlations are in different directions than the factors-response variable correlation

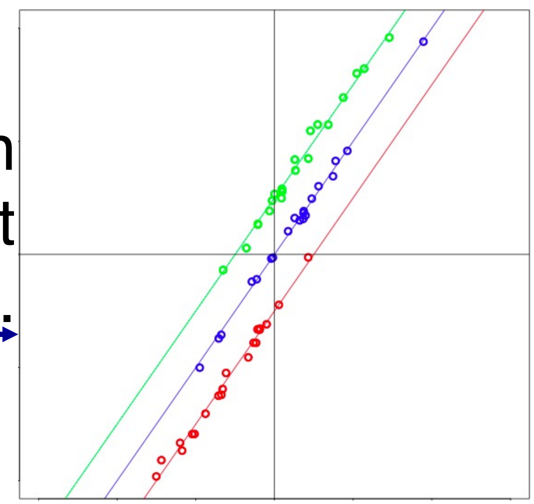
SS[error] drops; SS[factors] ~ the same: Great! This is what ANCOVA is supposed to do!

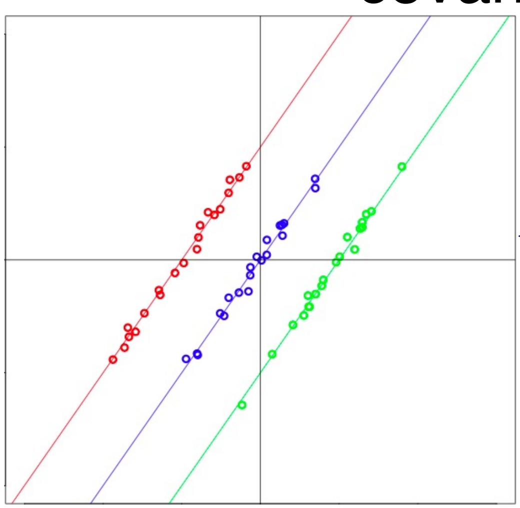




Nothing changes much: covariate not correlated with factors or response variable.
(literally nothing changes: very unlikely)

SS[factors] drops: Bound to happen (esp. when using covariate as control) – means that covariate and factors are correlated.





SS[factors] goes up: uh oh! (esp if null at first):
covariate is correlated with factors, and correlated with response variable, but these correlations are in different directions than the factors-response variable correlation

ANCOVA pointers.

- Rescale covariates.
  - If covariate x’ = (x-mean(x))/sd(x), the coefficients are easier to interpret.
- Measure covariates before treatment.
  - Interpretation of results is easier.
- Pre-test as a covariate of post test? Easier to just calculate the difference score.
- Covariates as control for confounds?
  - Strength of inference varies case by case.

ANCOVA reasoning.

2) A paper reports the following results when assessing the effect of different remedial mathematical education programs for high school students.

"There was a significant detrimental effect of class size on average standardized test improvement $F(1, 131)=6$, $p<0.05$. After factoring out class size effects, we found a significant main effect of textbook $F(3, 131)=4$, $p<0.05$, and a significant interaction between textbook and pedagogical style $F(6, 131)=3$, $p<0.05$, indicating that some textbooks are better suited for some pedagogical styles."

______ How many classrooms were evaluated in this design?

______ How many different textbooks were compared?

______ How many pedagogical styles were compared?

______ Assuming the design was balanced, how many classes were in each cell of the design?

What was the model (in R formula syntax) that the authors used?

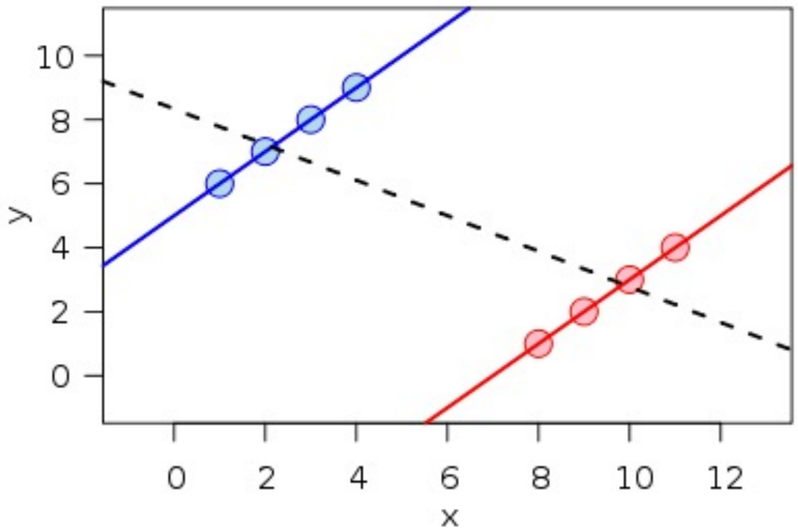
ANCOVA reasoning.

6) In an ANOVA, factor A is significant, factor B is not, and neither is the AxB interaction. However, when covariate C is taken into account, the ANCOVA shows that factor A is no longer significant, while factor B is (interaction still is not). Plot how this situation could come to pass.

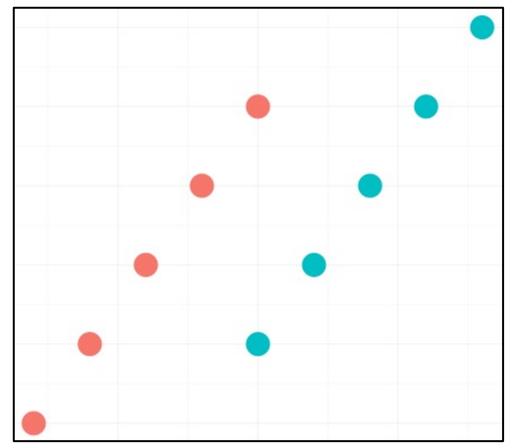
Extra credit: Make up a plausible scenario for which the situation in 6 would hold, indicating what the response variable [y] is, and what manipulations/measures A, B, and C correspond to.
Extra, extra credit: answer in the form of a limerick.

Simpson's paradox.

- Direction of apparent effect reverses when data are blindly aggregated disregarding latent variable.



y~x trend appears negative if we disregard difference between red/blue, but is really positive within categories.



Red appears lower on y than blue if we disregard effect of x. If we control for x, red has a higher intercept than blue

Simpson's paradox

- E.g., asian vs black undergraduate admissions.
- E.g., 1973 case against Berkeley admissions by sex:

Controls? Mechanisms?

Vox TOPICS TRENDING

Statistical controls tell us how the gender pay gap works, not that it isn't real

Updated by Matthew Yglesias | @mattyglesias | matt@vox.com | Feb 23, 2015, 1.

How racial discrimination in law enforcement actually works

EST

RACE AND JUSTICE: MUCH MORE THAN YOU WANTED TO KNOW

POSTED ON NOVEMBER 25, 2014 BY SCOTT ALEXANDER

The Gender Wage Gap And Wage Discrimination: Illusion or Reality?

Howard J. Wall

FRAMING FOR LIGHT INSTEAD OF HEAT

POSTED ON DECEMBER 3, 2014 BY SCOTT ALEXANDER

I.

Ezra Klein uses my analysis of race and justice as a starting point to offer a thoughtful and intelligent discussion of what exactly it means to control for something in a study.