TCSM
The General Linear Model

CAUSALITY

- Causality
- General linear model
- ANCOVA
- Assumptions of ANCOVA

Hypothesis, theory and causality

Deductive scientific research is about **testing hypotheses**.

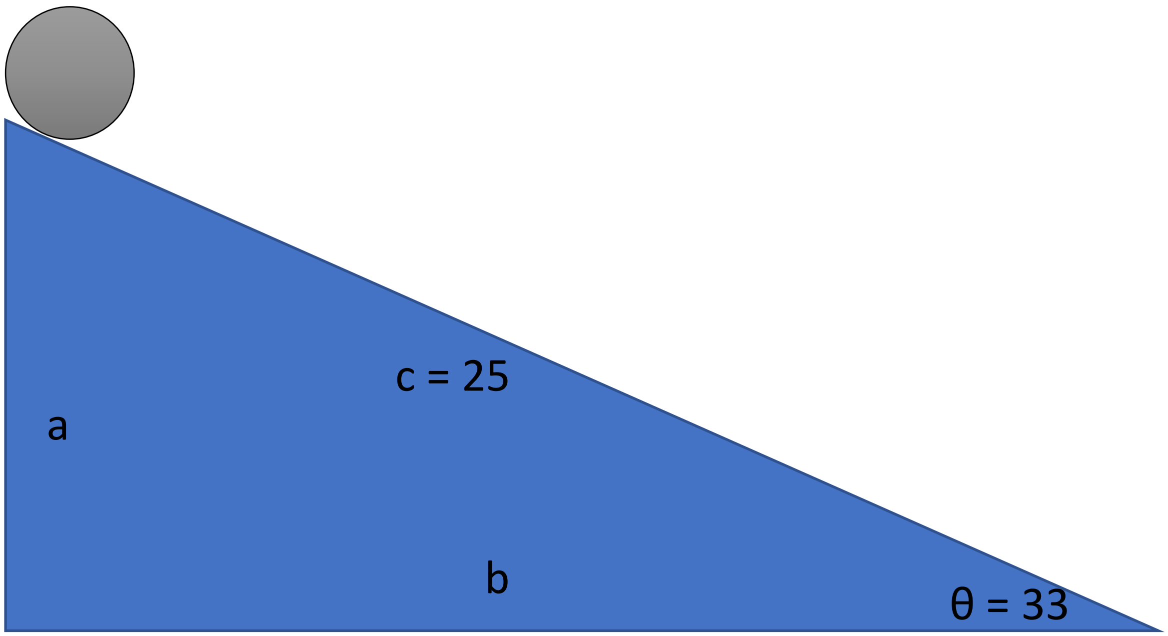
Hypothesis: specific, testable **prediction**

- Typically based on **theory**

Theory: A well-established **principle** (model) to explain some aspect of the natural world. It typically involves **causal relationships**.

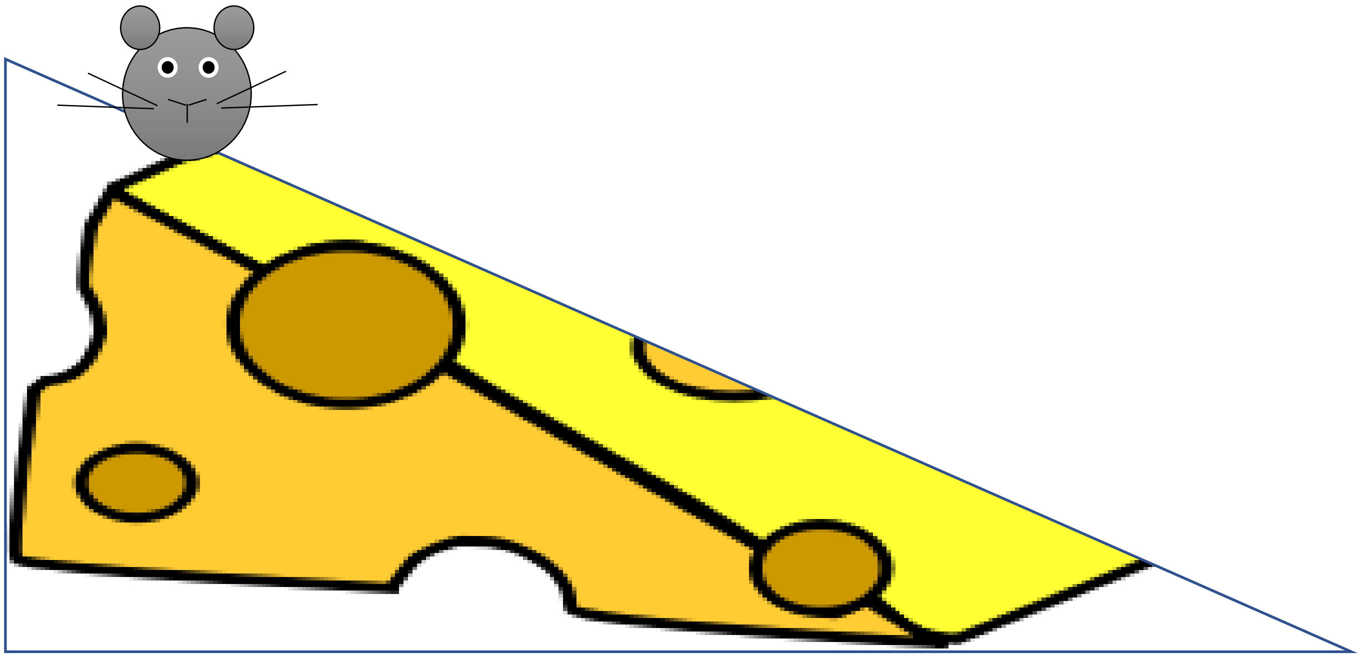
Hypothesis, theory and causality

Hypothesis, theory and causality



Hypothesis, theory and causality

Hypothesis, theory and causality



SEM Models and Theory

SEM models:

- **Statistical** model
- Represents **theoretical** model
- Includes all **causal relationships** and **assumptions**
- Tests all **hypotheses** at once

Definition of causality

X can be a **cause** of Y if there is:

1. Relationship
2. temporal precedence
3. Nonspurious

Of course, we need a *theoretically plausible mechanism* for the effect of X on Y.

Prediction VS causation

- **Prediction** is not the same as **causation**!
- Prediction only requires “relationship” between X and Y
- Statistical models can reveal relationships; conclusions about causality are based on **methods** and **theory**

More on causality

Investigating causality

1. **Relationship:** change in X accompanied by change in Y;

   $\rightarrow$ e.g., nonzero correlation

2. **Temporal precedence:** X precedes Y in time;

3. **Nonspuriousness:** X and Y are associated even if other relevant predictors are eliminated;

Two research traditions

Experimental VS **Correlational**

Two research traditions

Experimental research

- Meets all requirements for causality
  - 1. If you observe a relationship
  - 2. Your manipulation occurs prior to the outcome
  - 3. Other relevant variables are canceled out through
    - **Random assignment:**
    - To ensure any differences between groups are random and cancel out

- *Different from random sampling:*
  - *Best way to get <u>representative</u> sample; ensures generalizability*

Two research traditions

Correlational research

- **Random sample**: ensures generalizability
- **Measure** other relevant variables
  - Include these in your analysis
  - Their effects are “partialed out”
- Causal statements remain **problematic**/dangerous/wrong

Why?

Investigating causality

1. **Relationship:** change in X accompanied by change in Y;

   → e.g., nonzero correlation

2. **Temporal precedence:** X precedes Y in time;

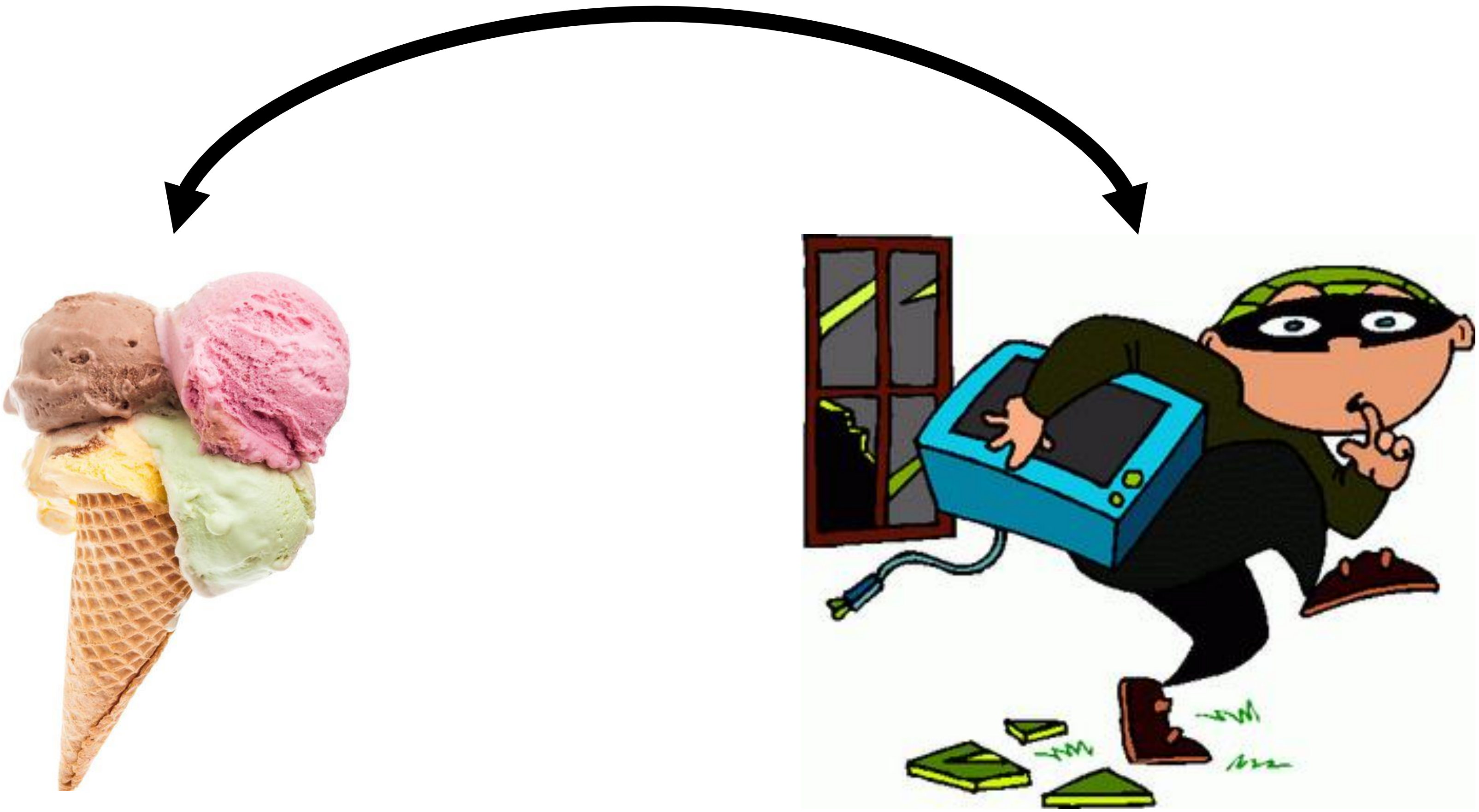
   → Longitudinal research

   → Logic

3. **Nonspuriousness:** the relationship between X and Y holds even if the influences of other variables are eliminated;

   → Remove the influence of other variables that may influence the outcome variable

1. Relationship



Isolating the effect

Isolate the effect of X by removing effect of other relevant variables on Y

We "**control**" for these variable(s).

Two popular ways of controlling:

I. **Experimentally** controlling
   random assignment in an experiment

II. **Statistically** controlling
   adjusting for third variable in correlational research

GENERAL LINEAR MODEL

- Causality
- General linear model
- ANCOVA
- Assumptions of ANCOVA

Statistical techniques

Experimental research

- *(factorial) ANOVA*:

  DV continuous

  IVs categorical (“factors”)

- *ANCOVA*:

  DV continuous

  IVs categorical (“factors”) and continuous (“covariates”)

Correlational research

- *Multiple regression analysis*:

  DV continous

  IVs interval or ratio **measurement level**

Insert Web Page

This app allows you to insert secure web pages starting with https:// into the slide deck. Non-secure web pages are not supported for security reasons.

Please enter the URL below.

https://	utrecht-university.shinyapps.io/mva_2019_embed_anova/

Note: Many popular websites allow secure access. Please click on the preview button to ensure the web page is accessible.

ANOVA vs. multiple regression

ANOVA specification estimates all 3 means:

$$y_i = b_0 D_{1i} + b_1 D_{2i} + b_2 D_{3i} + e_i$$

Regression specification estimates an intercept, and two differences-with-the-intercept:

$$y_i = b_0 + b_1 D_{1i} + b_2 D_{2i} + e_i$$

y_i: y-value of individual *i*

b: Regression coefficient (slope)

D: Dummy variables coding for group membership (1: member, 0: not a member) for all individuals *i*

e_i: Prediction error for individual *i*

Conclusion

ANOVA is a special case/different presentation of multiple regression analysis.

Example analysis: rejection

What is the effect of **rejection** on **strategic shopping**?

(shopping to enhance chances of inclusion)

Set up

1. Participant sees video message of their research "partner"
2. Participant records video message for the partner
3. Participant is told that the partner has watched the message, and then quit the experiment.

Rejection manipulation (3 conditions, random assignment):

- **rejection**: no reason given for partner's departure
- **neutral**: "Partner forgot an appointment"
- **confirming**: "Partner forgot an appointment and is really sorry"

Set up

4. Participant is told that they have to wait for a new partner
5. While waiting for new partner, participant is asked to spend 10 dollars in a fake webshop with University-branded products, and neutral products

DV = money spent on University products

IV = condition (factor with 3 levels)

Hypothesis:

Participants in the rejection condition will spend more on University products than those in the neutral or confirming condition, in order to ensure inclusion with the next partner.

ANOVA in R

```
> fit <- aov(Spent ~ Condition)
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
Condition    2  188.7   94.36   10.21 0.000166 ***
Residuals   56  517.7    9.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 
0.1 ' ' 1
```

ANOVA in R

```
> fit <- lm(Spent ~ Condition -1)
> summary(fit)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
Conditionconfirming  -0.5858     0.6799  -0.862    0.393
Conditionneutral     -0.7846     0.6799  -1.154    0.253
Conditionrejection    3.1385     0.6976   4.499 3.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 3.041 on 56 degrees of freedom
Multiple R-squared:  0.285,   Adjusted R-squared:  0.2467
F-statistic: 7.439 on 3 and 56 DF,  p-value: 0.0002807
```

ANOVA in R

```
> fit <- lm(Spent ~ Condition)
> summary(fit)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.5858     0.6799  -0.862 0.392548
Conditionneutral    -0.1987     0.9615  -0.207 0.837011
Conditionrejection   3.7243     0.9741   3.823 0.000333 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 3.041 on 56 degrees of freedom
Multiple R-squared:  0.2671,  Adjusted R-squared:  0.241
F-statistic: 10.21 on 2 and 56 DF,  p-value: 0.0001662
```

ANOVA in R

```
> fit <- lm(Spent ~ Condition)
> anova(fit)

Analysis of Variance Table

Response: Spent
          Df Sum Sq Mean Sq F value    Pr(>F)
Condition  2 188.72  94.362  10.207 0.0001662 ***
Residuals 56 517.73   9.245
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

ANCOVA

- Causality
- General linear model
- ANCOVA
- Assumptions of ANCOVA

Introducing ANCOVA

Used when IVs are **categorical** and **continuous** (=covariate).

Also a special case of the **GLM**.

Developed for **experimental data**

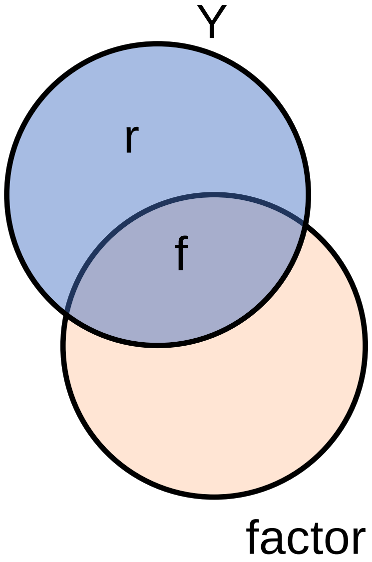
- **Purpose:** interpret **group differences**
- **Random assignment**: groups do not differ on covariates.
- Controlling for covariates **reduces unexplained variance**
- Relationship with covariate usually not of interest

F-test in ANOVA

Test statistic for variances: F

- The **ratio** of the explained part and the unexplained variance

$$F = \frac{MS_{model}}{MS_{residual}} = \frac{SS_{model} / df_{model}}{SS_{residual} / df_{residual}}$$




```
> fit <- aov(Spent ~ Condition)
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)    
Condition    2  175.4   87.70   8.132 0.000794 ***
Residuals   56  603.9   10.78                    
---
```

Sums of squares in ANOVA

Shiny app: https://utrecht-university.shinyapps.io/cj_anova/

Between groups sum of squares and df:

AKA model/explained SS

$$SS_{between} = \sum n_k (Y_k - Y_{grand})^2 \qquad df_{between} = k - 1$$

Within groups sum of squares and df:

AKA error/residual SS

$$SS_{within} = \sum (Y_{ik} - Y_k)^2 \qquad df_{within} = k(n-1)$$

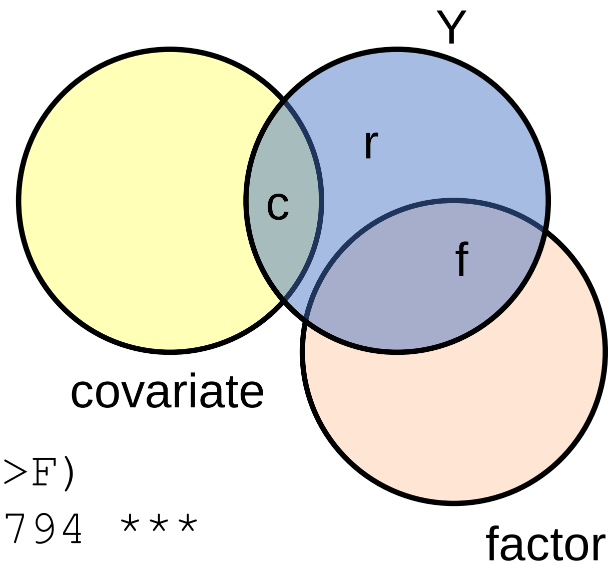
Disconnected from the server.
Reload

F-test in ANCOVA

Including a **covariate** reduces MS of the residual

Thus, *F* for the factor becomes larger:

$$F = \frac{MS_{factor}}{MS_{residual}}$$




```
> fit <- aov(Spent ~ Condition)
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
Condition    2  175.4   87.70   8.132 0.000794 ***
Residuals   56  603.9   10.78
---
> fit <- aov(Spent ~ Condition + SelfEst)
> summary(fit)
            Df Sum Sq Mean Sq F value  Pr(>F)
Condition    2  175.4   87.70   9.118 0.00038 ***
SelfEst      1   74.9   74.94   7.791 0.00721 **
Residuals   55  529.0    9.62
```

ASSUMPTIONS OF ANCOVA

- Causality
- General linear model
- ANCOVA
- Assumptions of ANCOVA

Assumptions

- No interaction between factor and covariate
- Homogeneity of residual variances
- Take care with interpretation of ANCOVA when
  - correlation between factor and covariate

ANCOVA vs. multiple regression

ANCOVA = regression with dummies and continuous predictor:

$$\boxed{y_i = b_0 + b_1 \text{Reject}_i + b_2 \text{Conf}_i + b_3 \text{SelfEst}_i + e_i}$$

Assumption: **no interaction** between factor and covariate (= regression lines are **parallel** in the groups).

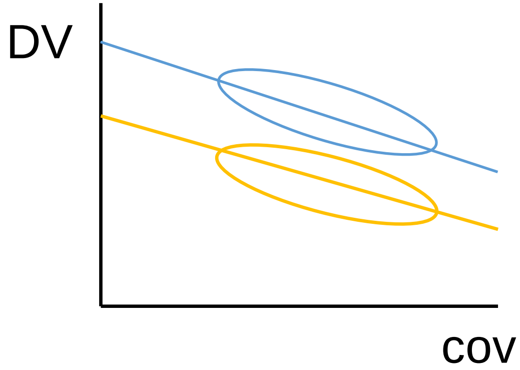
Assumption can be **tested**

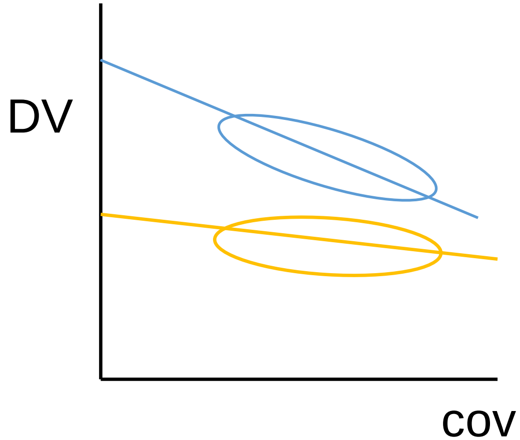
If the assumption is not met, you could include the interaction in the model

Interaction

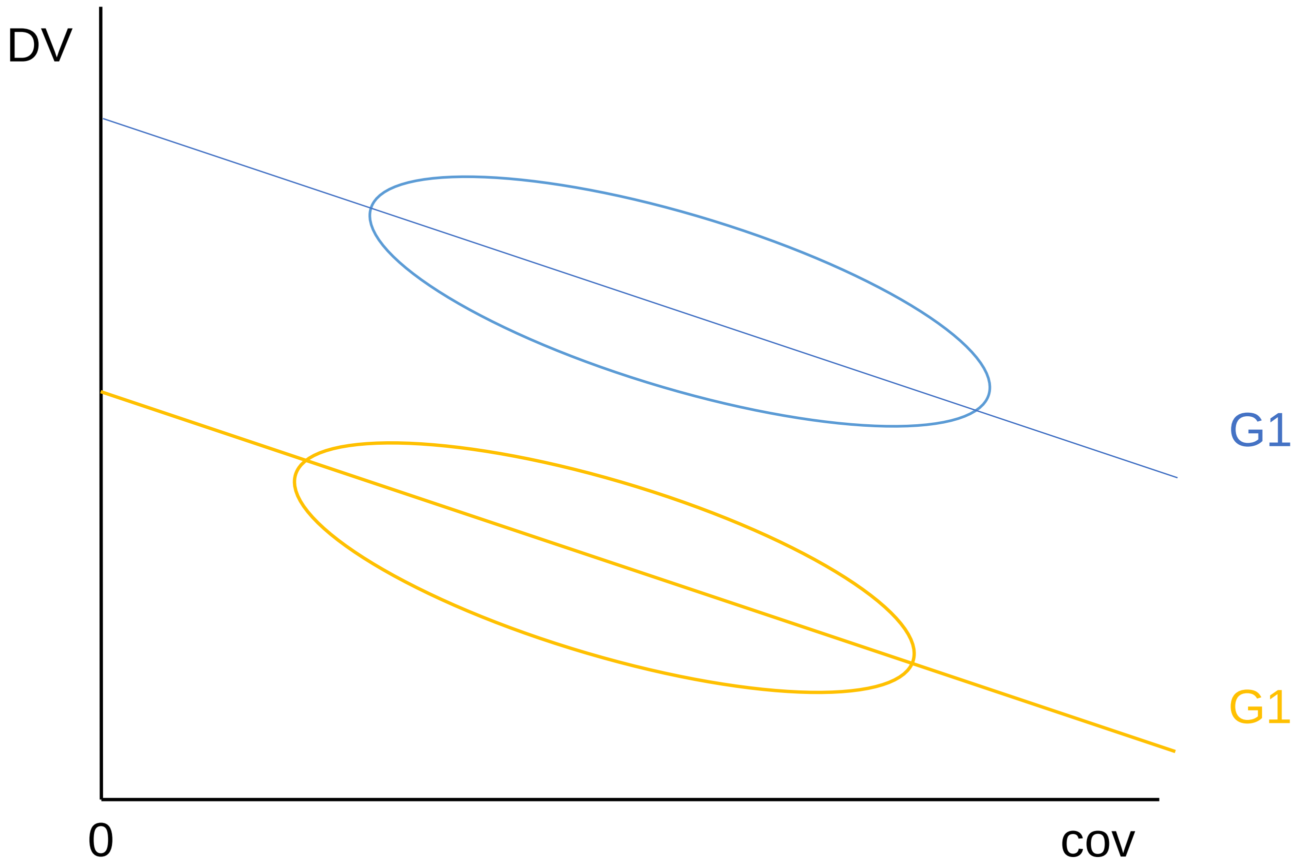
Is there an interaction? Shiny app:

https://utrecht-university.shinyapps.io/ANOVA_ANCOVA/

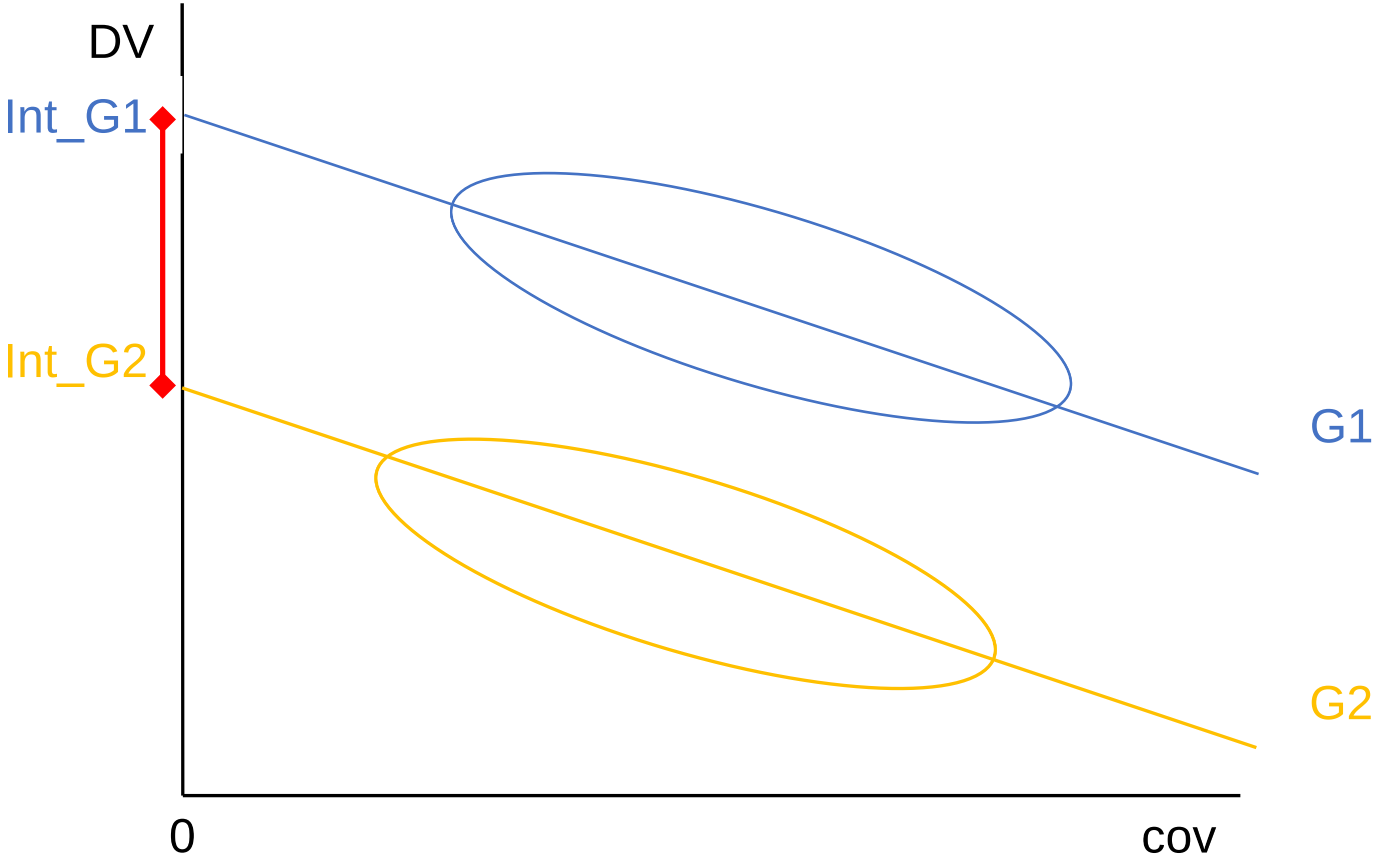




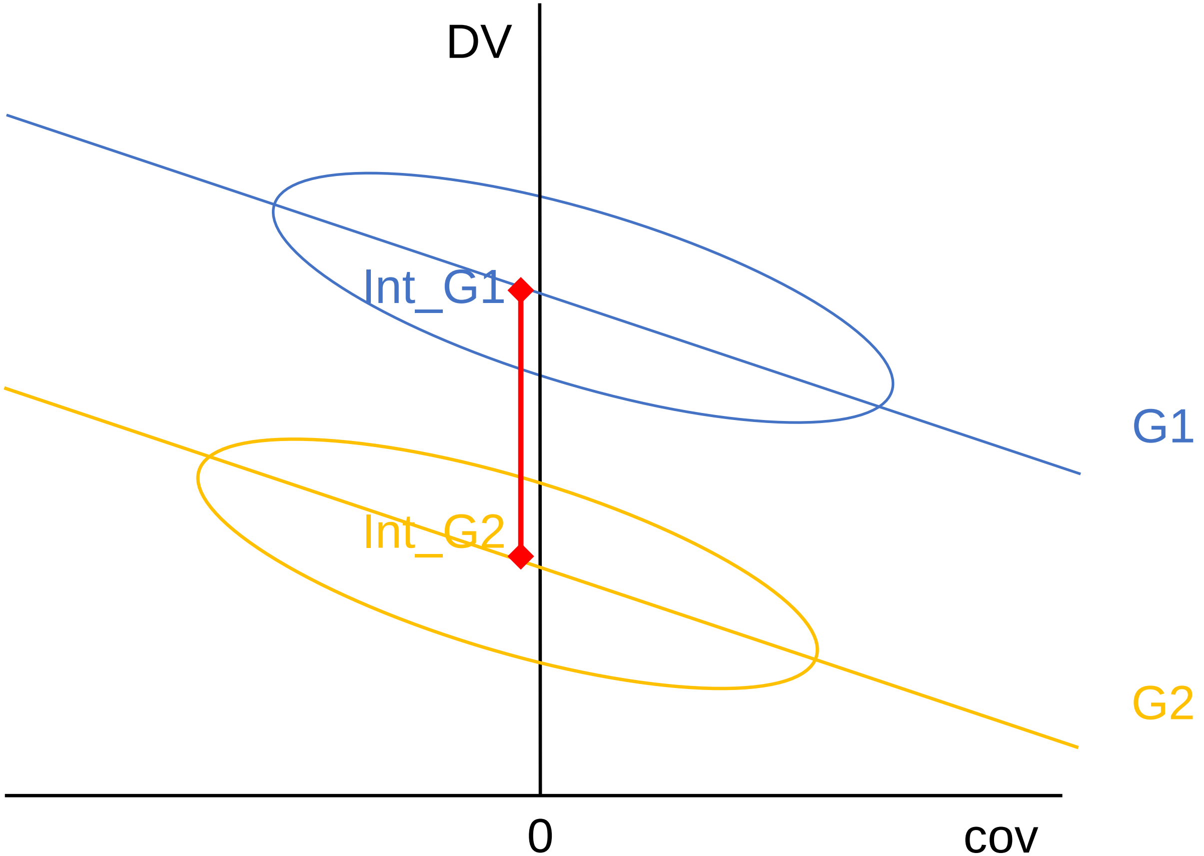
ANCOVA without Interaction



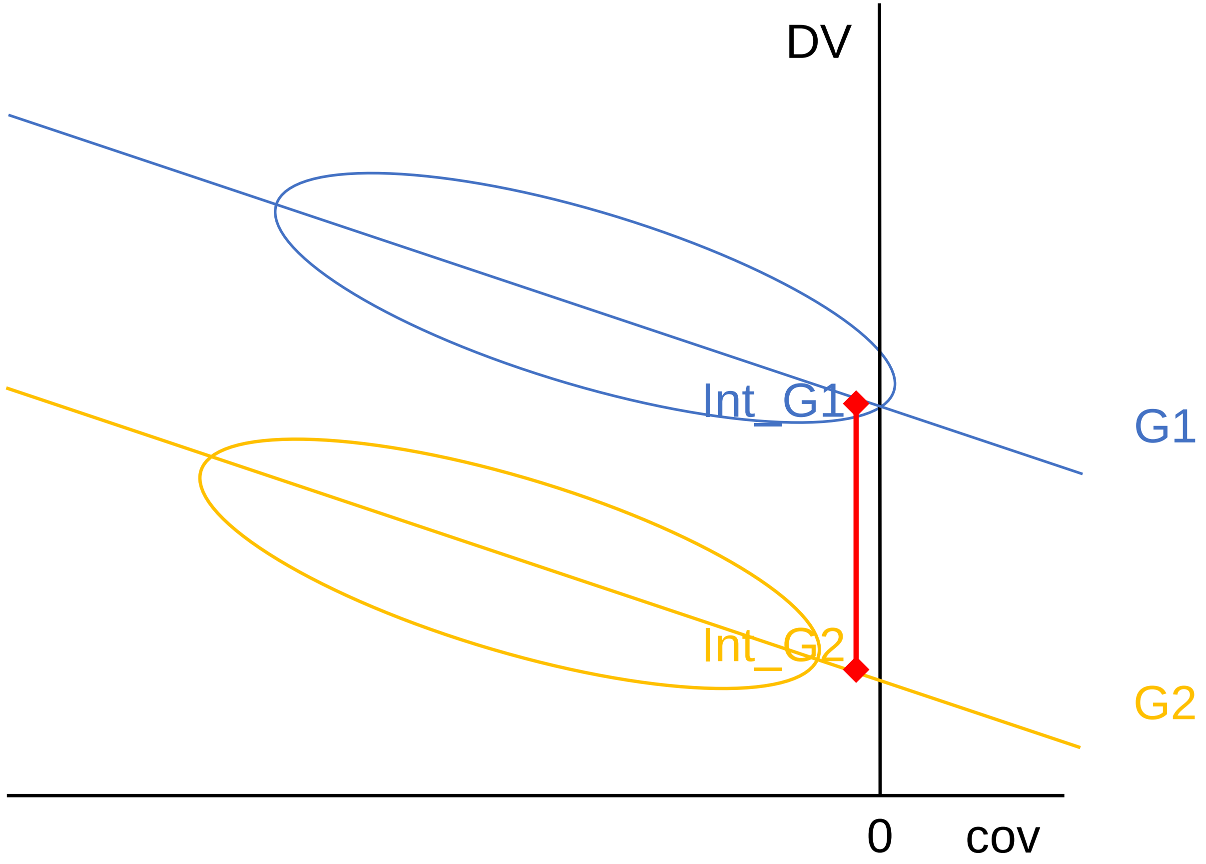
ANCOVA without Interaction



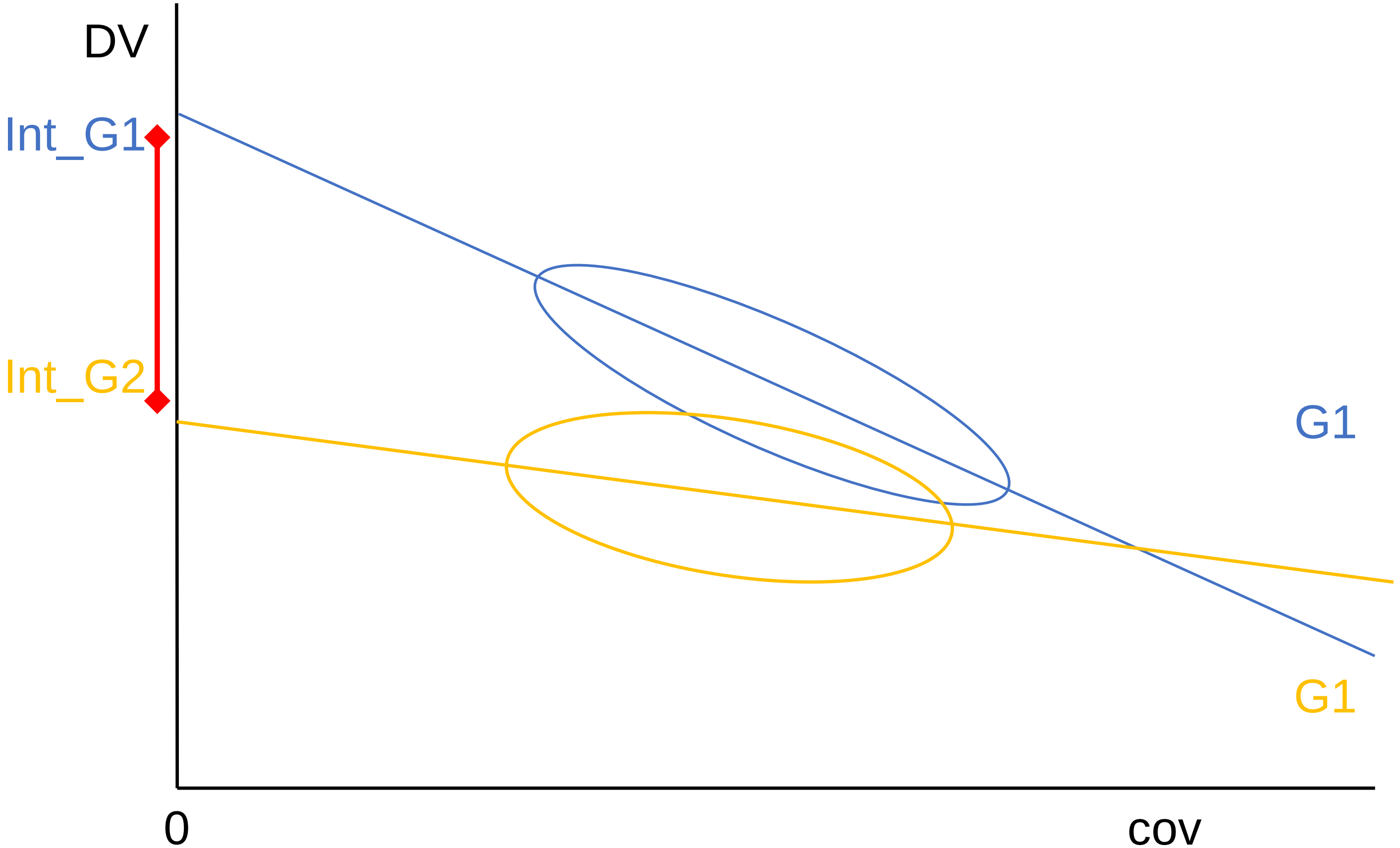
ANCOVA without Interaction



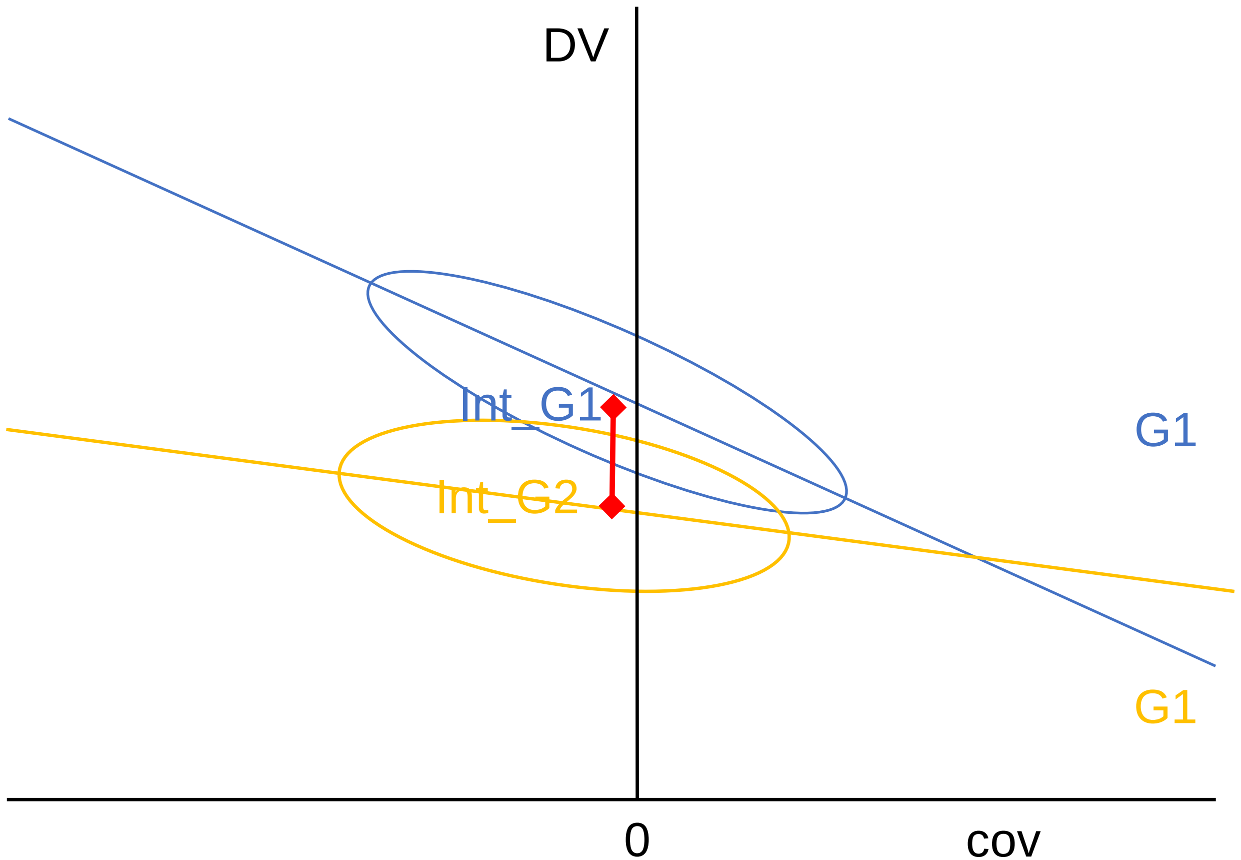
ANCOVA without Interaction



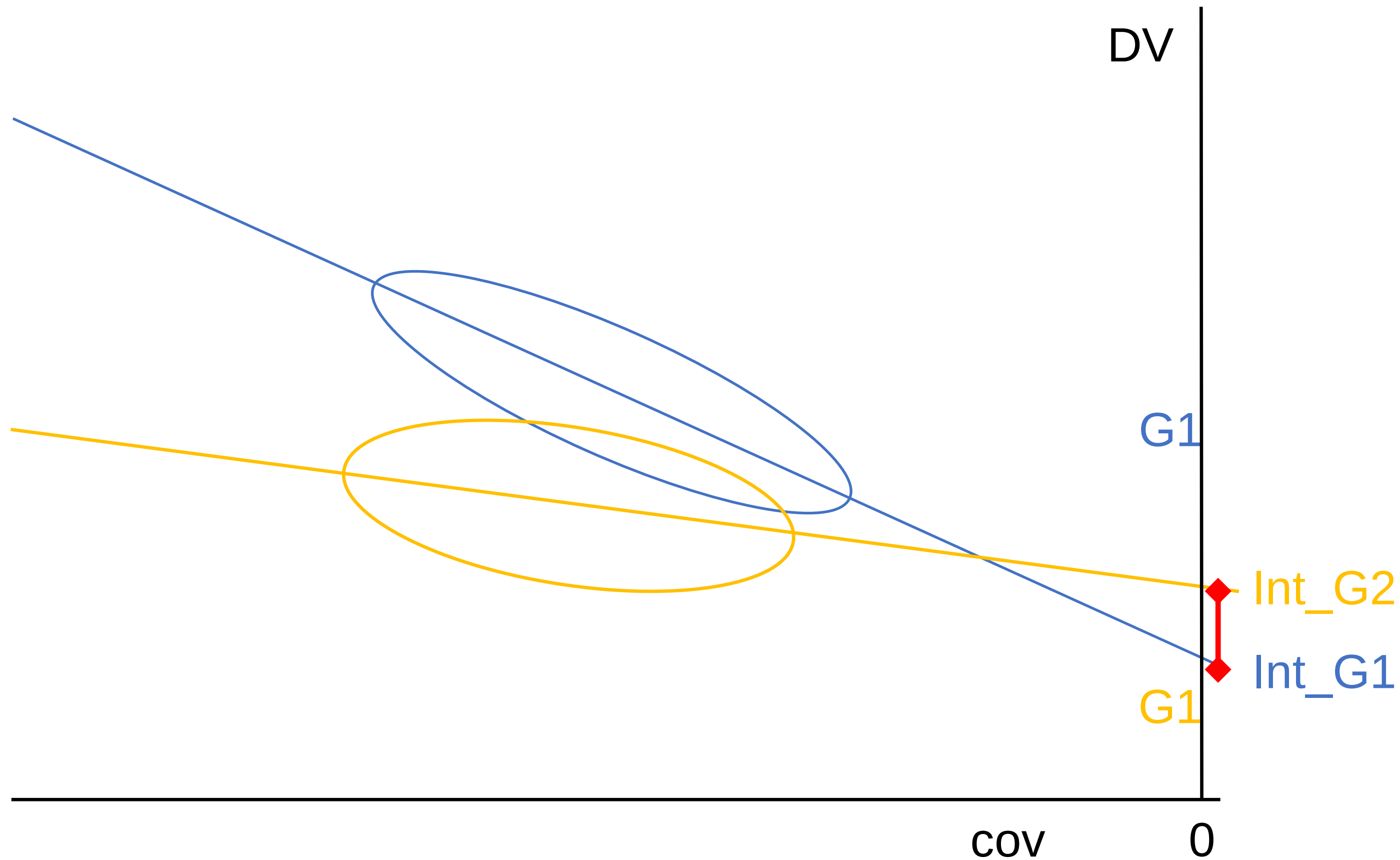
ANCOVA with Interaction



ANCOVA with Interaction



ANCOVA with Interaction



ANCOVA vs Multiple regression

- ANCOVA is a special case / presentation of multiple regression

- E.g., when you find a significant interaction in ANCOVA, run regression with dummy variables:

$$y_i = b_0 + b_1 * D_{Reject_i} + b_2 * D_{Conf_i} + b_3 * SelfEst_i + b_4 * D_{Reject_i} * SelfEst_i + b_5 * D_{Conf_i} * SelfEst_i + e_i$$

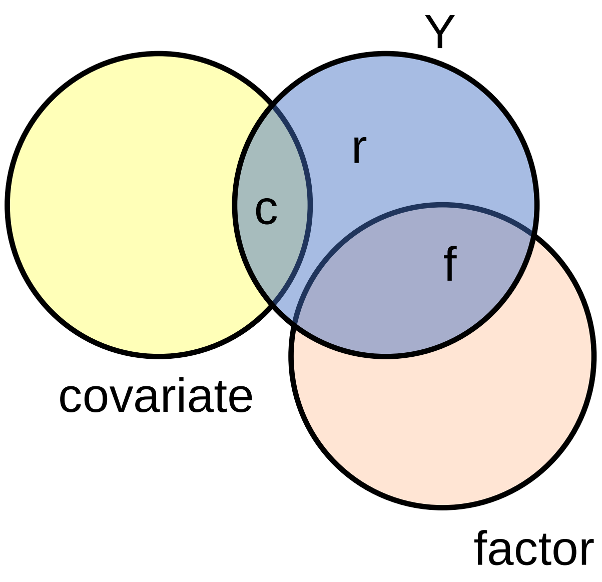
- NOTE: Parameters (b_1, b_2) still represent group differences in intercepts for Self-Esteem = 0.

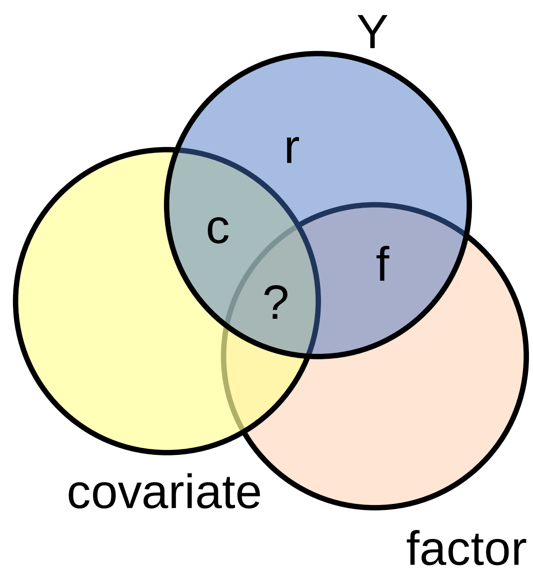
ANCOVA and existing groups

ANCOVA often used to compare **existing groups** that differ on the covariate

Researchers hope to **control** for these differences (= as if the groups are the same on covariate).

Other researchers (e.g., Miller & Chapman) indicate that ANCOVA **cannot** be used to investigate **existing groups**.

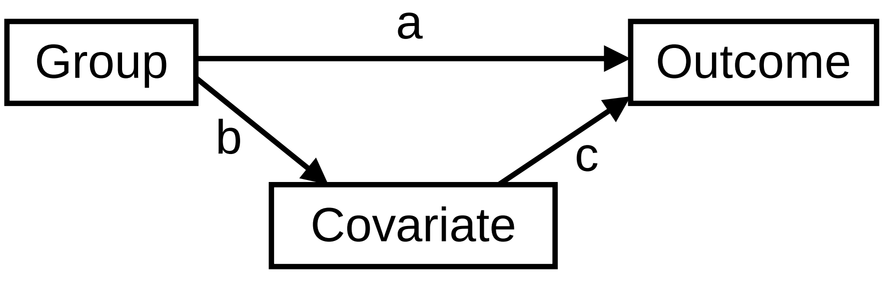




Possible problems: 1

Grouping variable may have **caused differences** on the covariate (= **mediation**)

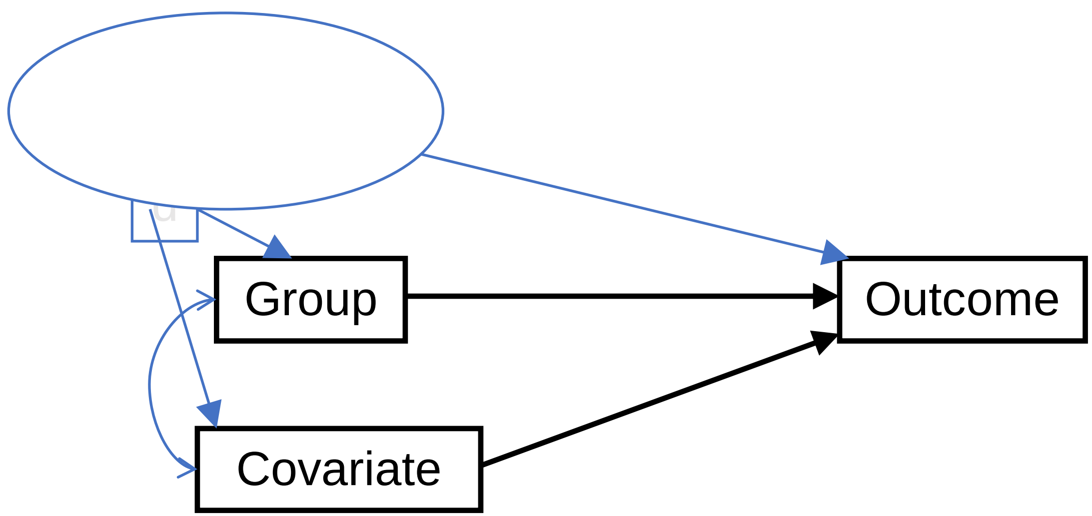
"Controlling" for covariate → underestimate total effect of Group on Outcome.



Possible problems: 2

Other (non-observed) **variables** account for differences in Outcome

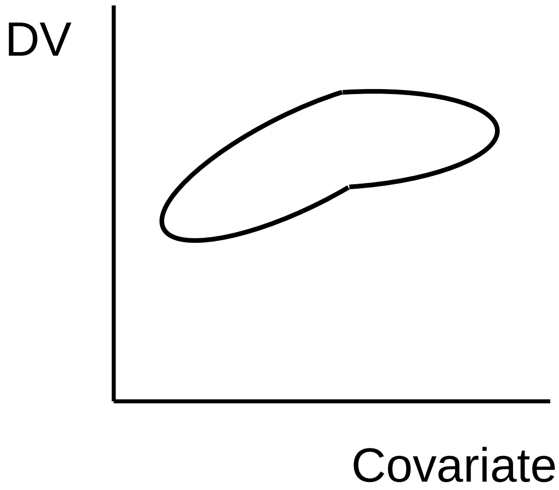
Controlling for an observed variable does not help.

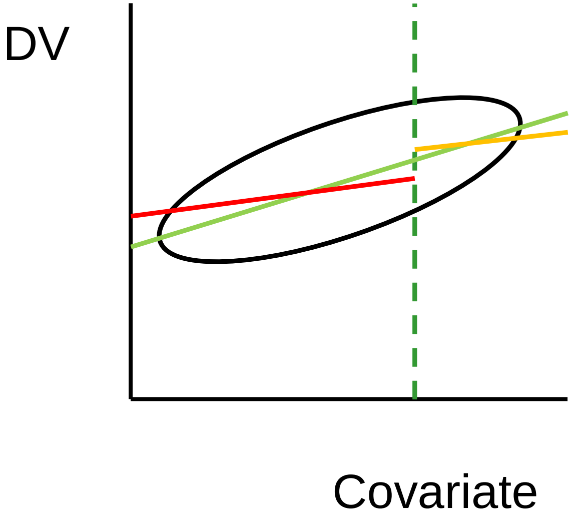


Possible problems: 3

Extrapolating beyond range of data is risky/impossible:

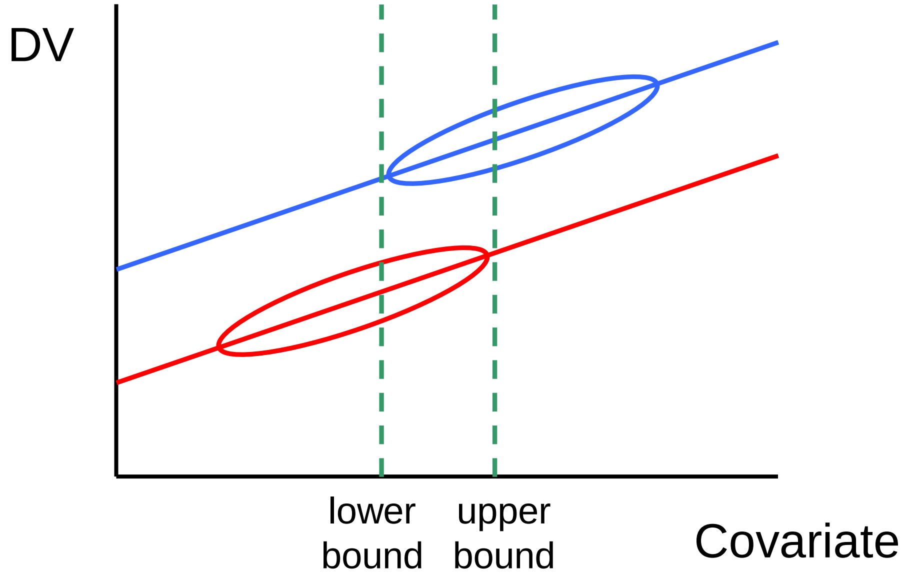
Can you draw conclusions about a relationship you did not observe?





Possible problems: 4

Selecting subgroups that do not differ on the covariate (i.e., form of matching) introduces the problem of **regression towards the mean**.



Existing groups

How to compare existing groups?

Use **multiple regression analysis**; instead of claiming that we control for the covariate, consider the covariate as another relevant variable.

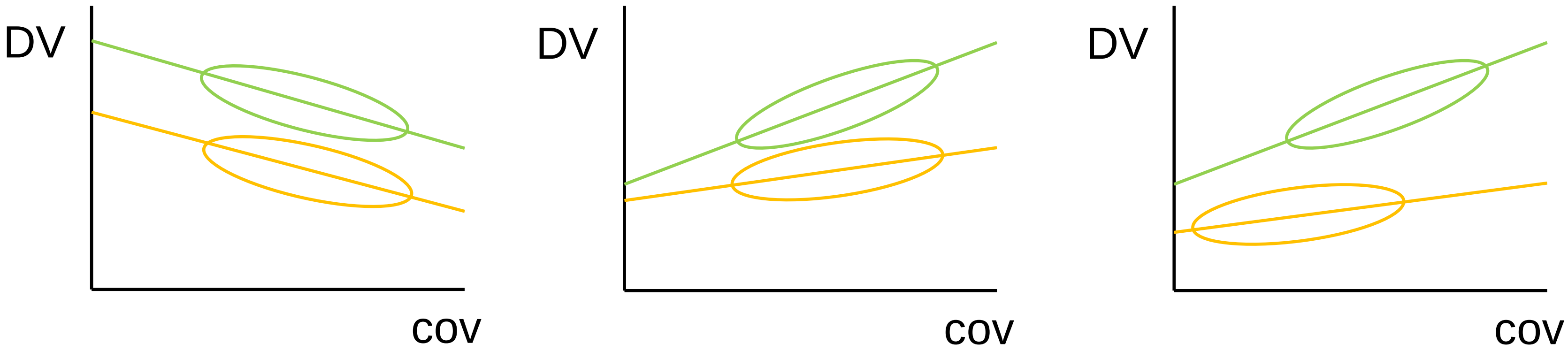
More sophisticated: Model the (hypothesized) causal relations between the Ivs

We can use **structural equation modeling** to compare alternative "data stories" that explain why variables share variance (= are correlated).

Very important!!!

If the factor and the covariate are **correlated**, this does not mean that there is an **interaction**;

When there is an **interaction**, this does not mean the factor and covariate are **correlated**.



Additional reading

Useful website:

http://www.statsoft.com/Textbook/General-Linear-Models

http://www.statsoft.com/Textbook/Basic-Statistics