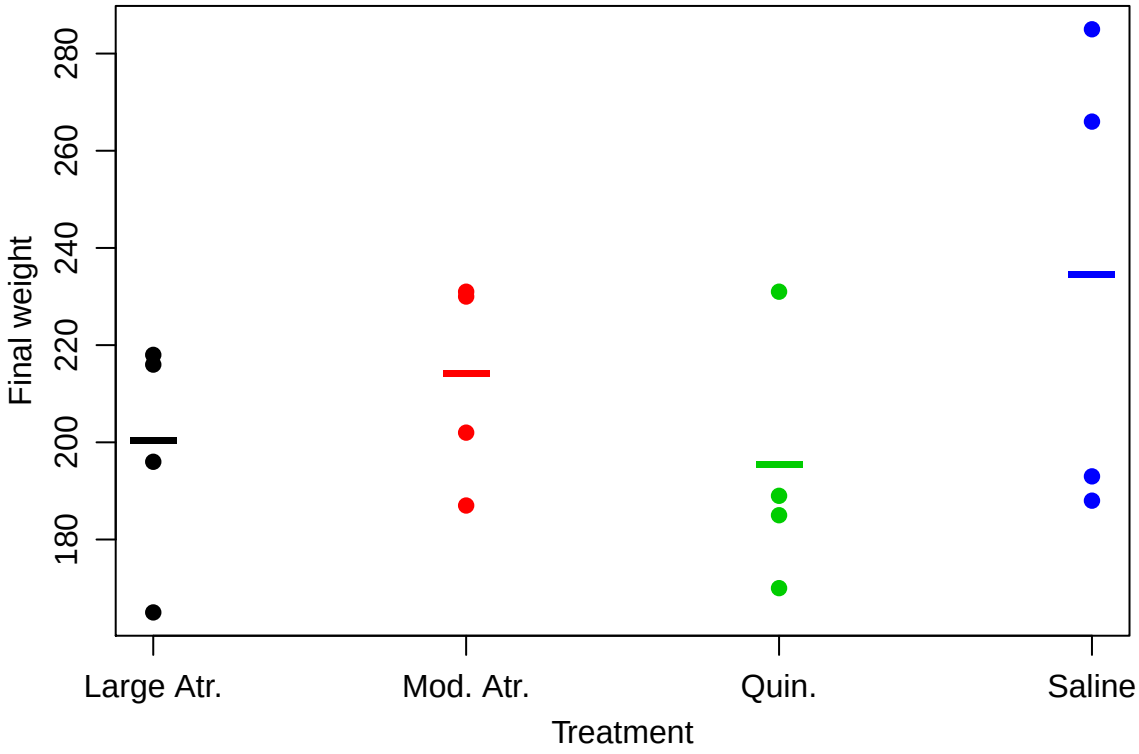
Stat 471/571: Analysis of covariance

Q: I measured baseline values for each of my subjects. How can I use that information?

Baseline values: Response variables measured **before** treatments applied.

Example data set: (DeLury 1946),
one of the earliest examples of use of analysis of covariance (ANCOVA)
rats exposed to toxicants.
4 treatments:
large dose atropine, moderate dose atropine, quinidine, saline solution (control/placebo)
Response is body weight after 4 days exposure
Also measured initial body weight = baseline value
CRD, 4 rats per treatment

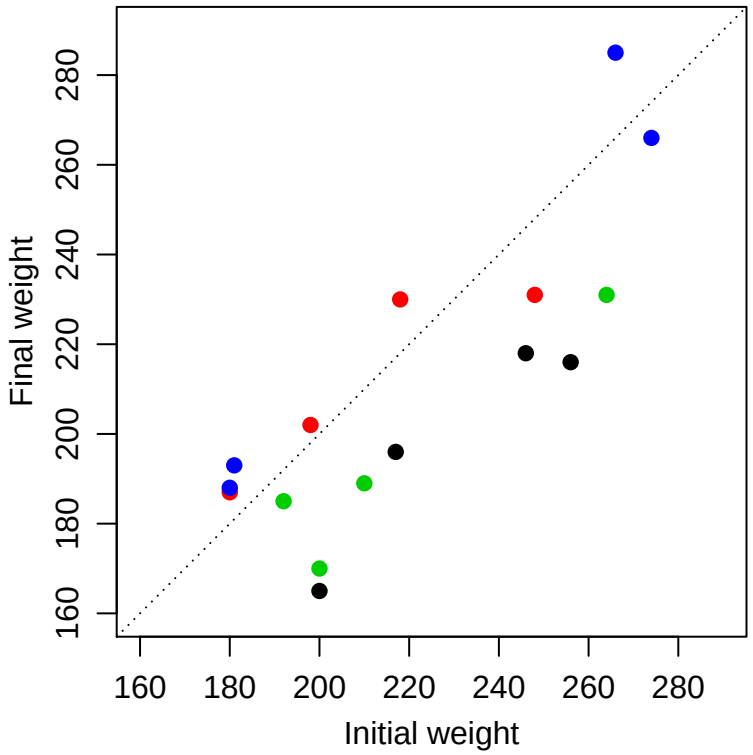


Analysis of final weight, 1 way ANOVA:
error sd: 32.5, 12 df
trt p-value: 0.36 (3 df)
contrast comparing saline to average of rest:
estimate = 31.3, se = 18.8, p = 0.12

Rats, at least in this study, have very different final weights
At least two explanations:
  Rats started out at same initial size, respond differently to the treatments
  Rats started out at different initial sizes

Study design issue:
  Could only use rats within a narrow range of initial sizes
  Not done here. Range of initial weights is 180 - 274!

Relationship between initial and final weight:



Study design issue:
  Could construct "initial" weight blocks
    find 4 small animals = a block, randomly assign treatments within that block
    4 "next smallest" = 2nd block, etc.
  Consequences are:
    Averaged over the blocks, each trt has about the same average initial weight
    Treatment comparisons are within a block, so between animals of similar initial weight
    Variability between blocks (i.e. initial weights) removed from error variance
  Could have done that here, but didn't

Analysis of covariance uses a model to achieve the same goals

Compare treatments at the same value of initial weight
Remove variability between initial weights from the comparison of treatments

Basic ANCOVA model:

$$Y_{ij} = \mu + \alpha_i + \beta X_{ij} + \varepsilon_{ij}$$

X_{ij} is the covariate, e.g. initial weight for observation j in treatment i
Assumed linearly related to Y_{ij}
β is the regression slope relating X_{ij} to Y_{ij}
$\mu + \alpha_i + \varepsilon_{ij}$ is the model for a 1 way ANOVA in a CRD.
Can replace that part with any other model: blocks, factorial treatments, both, $\cdots$

Compare the ANCOVA model to the ANOVA model (no covariate)

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \varepsilon^*_{ij} \\ Y_{ij} &= \mu + \alpha_i + \beta X_{ij} + \varepsilon_{ij} \end{aligned}$$

The ANCOVA model partitions the ANOVA error ε^*_{ij} into two components:
The part predictable from the covariate, βX_{ij}
and the random unpredictable part, ε_{ij}

Makes intuitive sense to compare two treatments at the same value of X_{ij}
When you do that, the difference in treatment means does not depend on β or X_{ij}
Uncertainty in that difference only depends on the variability in ε_{ij}
Which is always $\leq$ (almost always $<$) than the variability in ε^*_{ij}
Blocking does something similar: $Y_{ij} = \mu + \alpha_i + \text{block}_j + \varepsilon_{ij}$
Blocking makes no assumptions about differences among blocks, block_j can be anything
ANCOVA assumes a linear relationship with the covariate, $\beta\, X_{ij}$

Analysis of final weight, 1 way ANCOVA:
error sd: 10.3, 11 df (ca. 1/3 of ANVOA sd)
trt p-value: 0.0007 (3 df)
contrast comparing saline to average of rest:
estimate = 26.1, se = 6.0, p = 0.0011
estimated β: 0.855, se = 0.082

What's going on with the estimate? 26.1 (ANCOVA) or 31.3 (ANOVA)
The treatment means also differ!

Model	Large	Mod.	Quin.	Saline
ANOVA	199	212	194	233
ANCOVA	191	221	197	229
Mean X	229.8	211.0	216.5	225.2
$-\beta(\overline{X}_{i} - \overline{X}_{\cdot})$	-7.8	8.2	3.5	-3.9

Treatments have different average initial weights.
  Because of randomization, don't expect large differences
  but only 4 rats per treatment, so expect some random variation in initial weight

Treatment means in ANCOVA are "adjusted" to same value of X
  Common to adjust to overall average of X

$$\text{Adj. } \overline{Y}_i = \overline{Y}_i - \hat{\beta}(\overline{X}_i - \overline{X}_\cdot)$$

So estimated treatment differences also change:

$$\hat{\alpha}_i - \hat{\alpha_k} = \text{Adj. } \overline{Y}_i - \text{Adj. } \overline{Y}_k = \overline{Y}_i - \overline{Y}_k - \hat{\beta}(\overline{X}_i - \overline{X}_k)$$

  Two things to note:
    Adjustment will be small when $\overline{X}_i \approx \overline{X}_k$
    or when $\hat{\beta}$ close to 0
  se of Adj. $\overline{Y}_i$ or Adj. $\overline{Y}_i$ − Adj. $\overline{Y}_k$ can be computed

Other ways to use baseline values: change scores
  Compute $Y_{ij} - X_{ij}$: change in weight from initial to final, for each rat
    use this as the response variable
  Big advantage: you get to choose how to quantify change.
    Difference very much the most common
    But could use ratio (Y_{ij}/X_{ij}) or log ratio $\log(Y_{ij}/X_{ij})$
  Use your subject matter knowledge to decide

Analysis of difference (final - initial):
  error sd: 11.2, 12 df (slightly > than ANCOVA
  trt p-value: 0.0009 (3 df)
  contrast comparing saline to average of rest:
    estimate = 25.2, se = 6.4, p = 0.0021

Connection between the difference analysis and an ANCOVA model:

$$\begin{array}{lcll} Y_{ij} - X_{ij} & = & \mu + \alpha_i + \varepsilon_{ij} & \text{difference model} \\ Y_{ij} & = & \mu + \alpha_i + (1)X_{ij} + \varepsilon_{ij} & \text{ANCOVA model with slope} = 1 \end{array}$$

What about combining the ideas?: difference with the baseline covariate:

$$\begin{array}{lcll} Y_{ij} - X_{ij} & = & \mu + \alpha_i + \beta X_{ij} + \varepsilon_{ij} & \text{difference model with a covariate} \\ Y_{ij} & = & \mu + \alpha_i + (1+\beta)X_{ij} + \varepsilon_{ij} & \text{is just an ANCOVA model w/diff. slope} \end{array}$$

Analysis of differences can fail, one example:

	Analysis using ANOVA for final wt.	ANOVA for difference	ANCOVA
error sd	32.5	48.0	33.9
trt p-value	0.36	0.34	0.42
se of est.	18.8	27.7	19.8
$\hat{\beta}$			0.022

For these data, ANCOVA is similar to ignoring the baseline value.

Analysis of differences is worse than ignoring the covariate.

Difference in trt p-value mostly because 1 fewer df for error in the ANCOVA

What's going on?

Estimated slope is essentially zero

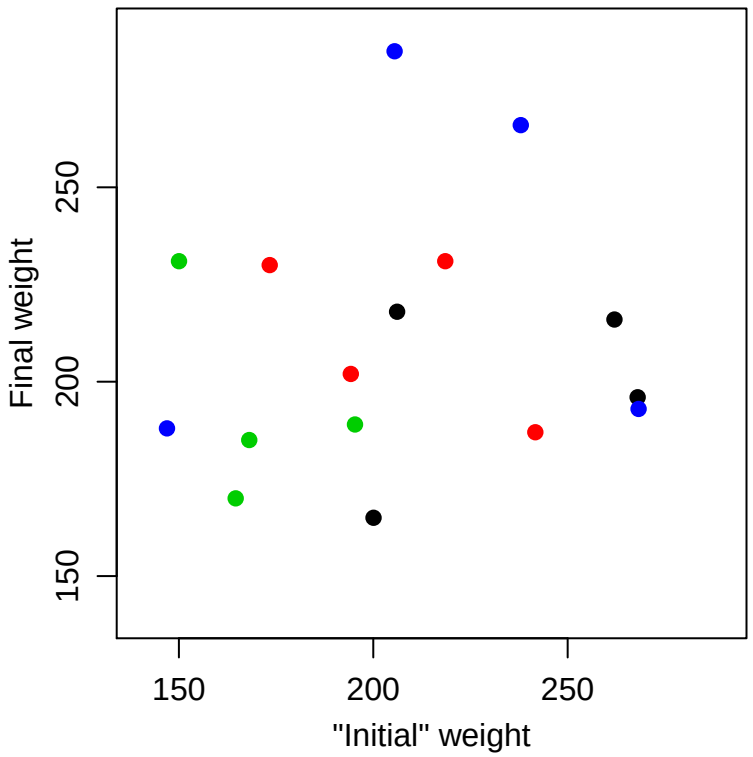
Covariate is essentially uncorrelated with the response.

When Var X = Var $Y = \sigma^2$, Var $Y - X = (2 - 2\rho)\sigma^2$

Var difference > Var Y when correlation, $\rho < 0.5$.

All that happens to an ANCOVA is that $\beta \to 0$

and you lose one error df.



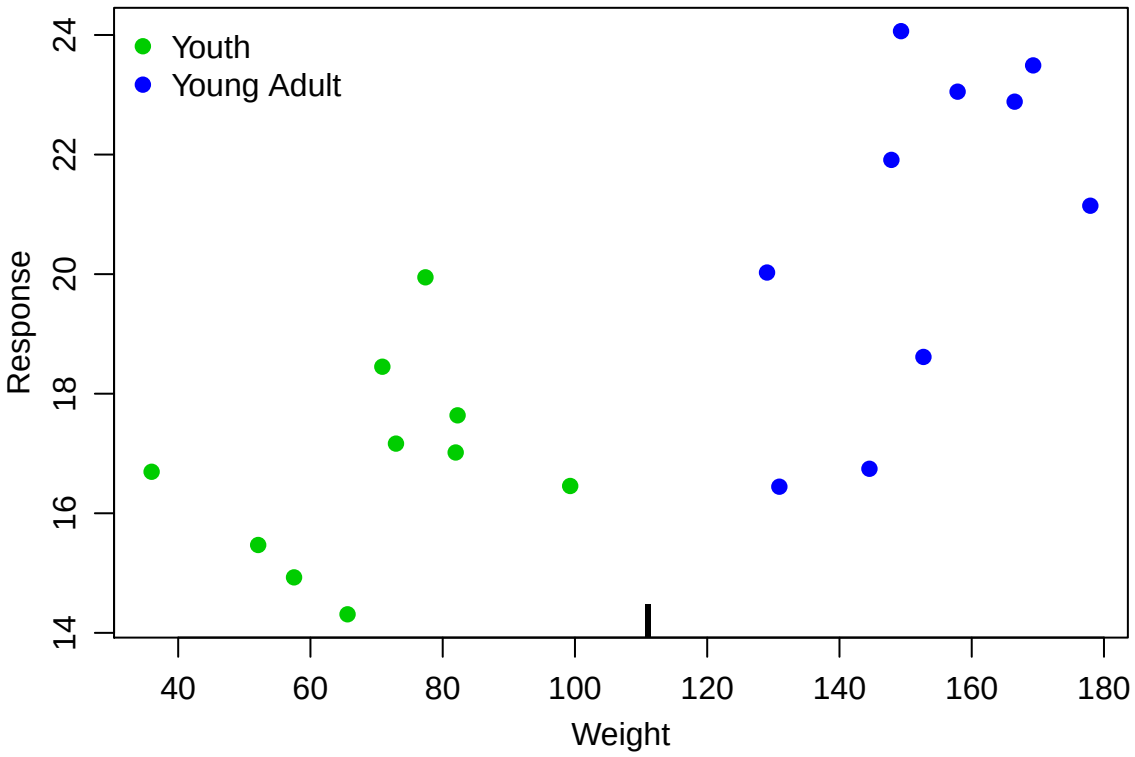
Should you use differences or ANCOVA?

Argued multiple times in many literatures

In favor of differences:
easy to interpret,
can choose appropriate change measure
Against differences:
If correlation between response and covariate small (e.g. < 0.5),
differences have larger sd than ignoring initial value
In favor of ANCOVA:
never worse than ANOVA ignoring baseline values
often more precise than using the difference
Against differences:
Assumes a straight-line relationship between covariate and response
Especially dangerous assumption when treatments have very different covariate values

More uses of ANCOVA, not just baseline value adjustment

1. Adjust for other continuous "nuisance" variables, e.g. age
   - so comparisons of treatments are for "same age"
   - MUST be variables not influenced by the treatments
   - easiest to justify when measured before treatment imposed
2. Especially common in observational studies
   - No randomization, so groups not necessarily "alike on average"
   - Compares groups at same measured X values.
   - Sometimes called "regression matching"
   - Be very careful of extrapolation
     - If groups have very different X values, "in the middle' X value may not describe any realistic individual



3. To estimate the regression slope
4. Mediation analysis
   - Treatment has an effect on Y (measured at end of study)
   - Treatment also has an effect on some X (measured at end of study)
   - Is the effect on Y "mediated" by the effect on X?
   - Include both treatment and X in a model
     - Is the effect of treatment now close to 0?
   - Very popular in social science applications.
     - Lots of details. This is just a very quick intro

Extensions of ANCOVA:

1. Heterogeneous regression lines models
   - ANCOVA model assumes same slope for all treatments / groups
   - Very convenient - difference between two treatments is same for any X
   - What if we want to allow the slopes to differ?

$$Y_{ij} = \mu + \alpha_i + \beta_i X_{ij} + \varepsilon_{ij}$$

- Often parameterized as an effects model for both intercepts and slopes:

$$Y_{ij} = \mu + \alpha_i + \beta X_{ij} + \gamma_i X_{ij} + \varepsilon_{ij}$$

- Same slopes $\Rightarrow$ all $\gamma_i = 0$
- Can estimate the intercept and slope for each treatment / group

Group	Intercept	Slope
1	$\mu + \alpha_1$	$\beta + \gamma_1$
2	$\mu + \alpha_2$	$\beta + \gamma_2$
$\vdots$		
k	$\mu + \alpha_k$	$\beta + \gamma_k$

- And test hypothesis that slopes are the same
  - model comparison between het. reg. model (with γ_i) and ANCOVA model (without)
- Difference between treatments depends on where you assess it (what X?)
  - Two lines always cross, so treatment difference can be positive, close to 0, or negative
  - Interpretation depends on what happens for X values in the range of the data
  - Software allows you to estimate trt differences at specified X values
  - Johnson-Neyman technique tells you the range of X values where the trt diff is not significantly different from 0.
- Need to be very careful interpreting tests of the intercepts.
  - Intercept is $X = 0$. Not relevant if X values from 180 to 280
  - Want tests of treatments at relevant X values
  - Either construct the appropriate estimate statements
  - Or shift X so the intercept is relevant, i.e. use $X^* = X - 180$ or $X^* = X - 230$

2. More complicated models for the relationship between X and Y

- What if the relationship isn't a straight line?
  - Transform Y and/or X.
  - Use a polynomial, e.g. quadratic:

$$Y_{ij} = \mu + \alpha_i + \beta_1 X_{ij} + \beta_2 X_{ij}^2 + \varepsilon_{ij}$$

  - much easier to interpret if the β's are constants, not varying by treatment

3. Not sure what covariates to use in the model?

- Quite common in observational studies.
- My suggestion:
  - Remove the treatment variable from the model
  - Do model selection (probably using AIC or BIC) to choose a set of covariates that predict Y
  - Add treatment back to the model with that set of covariates
- And report both the adjusted (using the covariates) and unadjusted (no covariates) results