simon-5502-03-slides

Topics to be covered

- What you will learn
  - What is covariate imbalance?
  - Indicator variable coding
  - Mathematical model
  - Assumptions
  - Choosing your covariates

What is a covariate?

- Variable not of direct interest
  - Relationship to outcome is already established
  - Still must be accounted for
- Examples
  - Smoking in a cancer study
  - Gestational in a neonatology study
- A covariate can be continuous or categorical

Speaker notes

This is a repeat of what I said earlier. A covariate is not of direct interest. Testing the relationship of the covariate with the outcome variable is not of great interest. Often this is because the relationship between the covariate and the outcome has already been established.

Even though it is not of direct interest, you feel an obligation to account for the covariate. It plays an important role and failure to measure and adjust for the covariate makes your research appear naive.

A covariate can be continuous or categorical, but more often the former.

What is covariate imbalance?

- Difference in mean value between treatment and control group
  - Often a problem in observational studies
  - Sometimes a problem in randomized studies

Speaker notes

Covariate imbalance is an issue in many research studies. It occurs in a comparison of an outcome between a treatment group and a control group (or maybe an exposure group and a control group). You want the treatment group to be identical to the control group in every way except for the treatment itself. But sometimes a covariate also differs between the treatment group and the control group. If that happens then the outcome could be influenced not by the treatment but by the covariate.

This is often a problem in observational studies.

It can happen in randomized studies at times as well. In theory, randomization will insure balance between the treatment and control group. Patients with large values of the covariate appear in the treatment and control group with equal likelihood. Patients with small values of the covariate appear in the treatment and control group with equal likelihood.

Sometimes, though, randomization doesn’t work. It relies on the law of large numbers and this doesn’t hold for some sample sizes. In particular, studies with less than 20 observations are fairly likely to see covariate imbalance, even with randomization. Even with larger sample sizes, sometimes you just get a bad batch of random numbers.

Why is covariate imbalance an issue?

- Biased estimates
  - Comparing apples to oranges
- Harms study credibility

Speaker notes

If a covariate is imbalanced, it can produce biased estimates. If younger patients are more likely to be in the treatment group, for example, and younger patients tend to have better outcomes, then you don’t know if the outcome variable is influenced by age instead of treatment. The variables are tangled up. This is comparable to the issue of collinearity in multiple linear regression.

Now the bias can go in either direction. Sometimes covariate imbalance will produce an artificial difference in the outcome between treatment and control. But it is also possible that a covariate imbalance masks a difference between the treatment and control.

You will find many times that the covariate imbalance does not produce any serious bias, but you still need to account for it. Failure to control for or to adjust for covariate imbalance will hurt the credibility of your study.

Examples of covariate imbalance

- Age in a study of smoking and Down's syndrome
- Smoking in a study of artillery assignment and sperm count

Speaker notes

A study of Down’s syndrome births had a covariate imbalance between women who smoked during pregnancy, the exposure group, and women who did not smoke during pregnancy, the control group. The average age in the exposure group was much lower than the average age in the control group.

I was involved in a study of lead exposure on male fertility. The exposed group were soldiers who worked on an artillery crew. These big guns could shoot missiles out at great speed, but when the missiles flew out, a cloud of lead dust washed back into the crew. Now, I should note the linkage between lead exposure and fertility is not well established. The control group were soldiers working in an office setting, far away from the big guns. It turns out that the proportion of smokers varied quite a bit between the exposed group and the control. When you are out in the field with explosives all around, smoking was totally banned. Now the artillery crew could still smoke while off duty, but the office workers had more opportunities to smoke on and off duty. This was in the 1990s before workplace bans on smoking were very common.

Now, I also have to admit that the link between smoking and male fertility is also not well established. But there was enough evidence to at least raise some concern about the covariate imbalance.

Covariate imbalance versus confounding

- Covariate imbalance is simpler
- Confounder definition relies on causation arguments

Speaker notes

I’m in the minority here, because I like the term “covariate imbalance” and most other researchers talk about “confounding.”

I like talking about covariate imbalance because it is simpler. Calculate the mean of the covariate in the treatment group and compare it to the mean of the covariate in the control group.

The definition of confounder involves complex arguments about causation. When I find myself needing to use a term like “confounder” or “confounding”, I find myself qualifying it. I’ll say “potential confounder” or “possible confounder.”

Preventing covariate imbalance

- Randomization
- Matching
- Stratification

Speaker notes

If you can, you should try to prevent covariate imbalance during the design of a research study. Randomization, if you can do it, is a great way to reduce the risk of covariate imbalance. It doesn’t always work, but it does work quite often. One of the nicest things about randomization is that it prevents covariate imbalance among those variables that you have measured, but it also prevents covariate imbalance among covariates that you didn’t measure, either because it was difficult or impossible to measure them.

Matching is another research strategy that helps to prevent covariate imbalance. For every treatment subject of a certain age, find a control subject with a closely matched age. Make sure that males are matched with other males and females are matched with other females.

You might choose other variables to match on. Just make sure that the covariates that you match on are important influences on the outcome. And don’t choose so many variables to match on that you have difficulty finding good matches.

Stratification also can help. Divide your patients into broad categories such as young, middle-aged, and older. Then randomly assign to treatment and control within these broad categories.

Adjusting for covariate imbalance

- Propensity score models
- Analysis of covariance

Speaker notes

If you did not or could not design the study to minimize covariate imbalance, you still have the option of adjusting for covariate imbalance.

Later on in this semester, you will learn about propensity score models. This week, you will learn about analysis of covariance.

Variables cannot be in the causal pathway

- Fixed at time of randomization
- Temporally preceding exposure
- Example: bottles given during a breast feeding study

Speaker notes

Covariates must be fixed and in place at the time when you flip the coin to choose whether they get into the treatment group or the control group.

If you are studying an exposure, then the covariate must be a variable that precedes the exposure in time.

Variables that are intermediate between the treatment/exposure and the assessment of the outcome are said to be in the causal pathway.

If you adjust for variables in the causal pathway, that may reduce or even eliminate the effect of your treatment or exposure.

I saw an example of this in a study I helped with. It was an examination of breast feeding patterns in pre-term infants. It is difficult to maintain breast feeding in a pre-term infant because the mother goes home from the hospital before the baby does. A rule of thumb is that the number of weeks that a pre-term baby has to stay in the hospital is roughly equal to the number of weeks early that the baby arrived.

In this study, mothers of newborn infants were randomly assigned to a treatment or control group. In both groups, mothers were encouraged to breast feed when they were in the hospital visiting their baby. They were given breat pumps to collect milk when they were at home. The difference was that in the control group, infant were fed by bottle when the mothers were not around. In the treatment group, the infants were fed through an nasogastral (ng) tube. The intervention was designed to avoid having the infants becoming habituated to the artificial nipple of the bottle and then having trouble latching onto the mother’s nipple.

The intervention was quite successful. There were a few minor mistakes made during the experiement, though. Sometimes an infant in the treatment group was given a few bottles at the hospital instead of being fed exclusively through the ng tube. That’s not surprising and not too much of a cause for concern.

The researchers did measure the number of bottles given in the birth hospital in both the treatment and control group and the average number of bottles was quite a bit lower than the treatment group, even though it was still a bit above zero.

I decided, on a lark, to put the number of bottles received in as a covariate in my analysis. To my initial horror, the effect of the treatment disappeared when you adjusted for the number of bottles.

But I quickly realized that this, if anything, reinforced the conclusions of the study. The number of bottles given was part of the causal pathway. It occured after random assignment, and the intervention was deliberately designed to influence this intermediate variable. So the adjustment actually helped to explain why the intervention worked.

Adjusting for baseline measurements

- Baseline = measurements prior to intervention
  - Done to improve precision
  - Can use baseline as a covariate
- Change score is an alternative
  - Also known as difference in differences (DID) model
  - Possible regression to the mean

Speaker notes

Many research studies include baseline measurements, measurement of the same outcome measure that you plan to use to compare the treatment and control groups. It often helps to make an assessment of the outcome BEFORE you implement any intervention. This is done mostly to improve precision, but it can also help control bias. If the intervention and control groups differ on the baseline measure, that is an indiciation that one group is more seriously ill at the start of the research than the other group.

You can consider the baseline value to be a covariate. Your outcome at baseline certainly does have some influence on your outcome at the end of the study. But there is no direct interest in the baseline measure itself.

A common alternative to using the baseline measure as a covariate is to compute change scores, the difference between the baseline measure of the outcome and the measure of the outcome at the end of the study. This is measuring the relative decline or improvement in health.

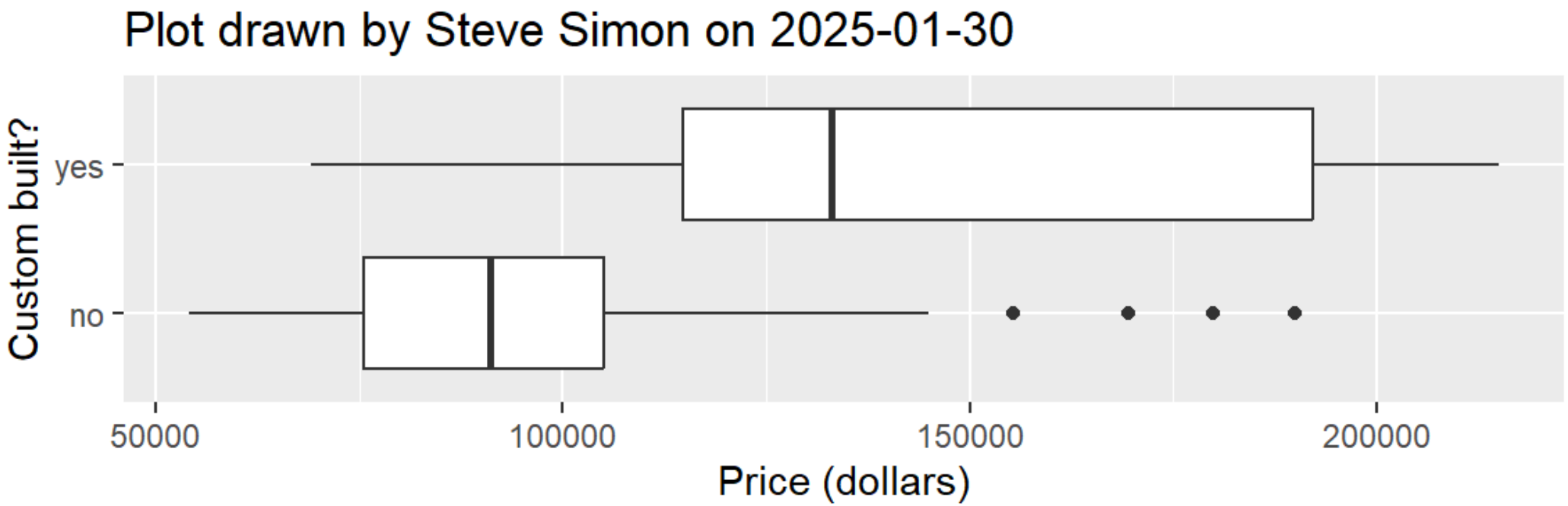
Use of change scores or difference in difference models is controversial. I like this approach but most of the research community is criticizes this choice. One criticism is that regression to the mean can possibly mess up the change score analysis.

Regression to the mean is the tendency for extremely low scores at one time point often are not quite as extreme when they are measured again, even if there is no change or intevention going on. Similarly extremely high scores at one time point often are not as extreme when measured again.

Break #1

- What you have learned
  - What is covariate imbalance?
- What’s coming next
  - Indicator variable coding

Traditional t-test, 1



Traditional t-test, 2

```
# A tibble: 2 × 4
  custom_build mean_price sd_price     n
  <chr>             <dbl>    <dbl> <int>
1 no               94752.   25366.    90
2 yes             144678.   47579.    27
```

Traditional t-test, 3

```
    Two Sample t-test

data:  alb$price[alb$custom_build == "yes"] and alb$price[alb$custom_build == "no"]
t = 7.1602, df = 115, p-value = 8.159e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 36114.14 63736.98
sample estimates:
mean of x mean of y
144677.78  94752.22
```

T-test using general linear model

```
# A tibble: 2 × 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       94752.     3350.     28.3  1.26e-53
2 custom_buildyes   49926.     6973.      7.16 8.16e-11
```

```
                   2.5 %    97.5 %
(Intercept)     88117.43 101387.01
custom_buildyes 36114.14  63736.98
```

Indicator variable coding

- Convert binary categories
  - 1 for “first” category
  - 0 for “second” category
    - Choice of “first” is arbitrary
- Can also be used for three or more categories

Why use the general linear model?

- Works for a variety of models
  - Independent samples t-test
  - Simple linear regression
  - One factor analysis of variance
  - Multiple linear regression
  - Analysis of covariance
  - Multi-factor analysis of variance

Why use indicator variables

- Simple interpretation
  - Average indicator is a probability
- Mix categorical and continuous independent variables

Break #2

- What you have learned
  - Indicator variable coding
- What’s coming next
  - Mathematical model

Mathematical model

- $Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + \epsilon_i$
  - X_i is the covariate
  - T_i is an indicator variable for treatment
- Estimate $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2$

Interpretation, 1

- $\hat{\beta}_0$
  - estimated average value of Y
    - when X=0, T=0

Interpretation, 2

- $\hat{\beta}_1$
  - estimated average change in Y
    - when X increases by one unit and
    - T is held constant

Interpretation, 3

- $\hat{\beta}_2$
  - estimated average change in Y
  - when T increases by one unit and
  - X is held constant

Interpretation, 4

- $\hat{\beta}_0$ is intercept for controls
- $\hat{\beta}_0 + \hat{\beta}_2$ is intercept for treatment

- $\hat{\beta}_2$ is the adjusted average difference between the treatment and the control

Diference in price adjusted for size

```
# A tibble: 3 × 5
  term            estimate std.error statistic  p.value
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)      11413.    6550.        1.74 8.41e- 2
2 custom_buildyes  14286.    5103.        2.80 6.01e- 3
3 sqft                55.4      4.12     13.4  3.15e-25
```

Live demo, parts 1, 2, and 3

Refer to the simon-5502-03-demo program.

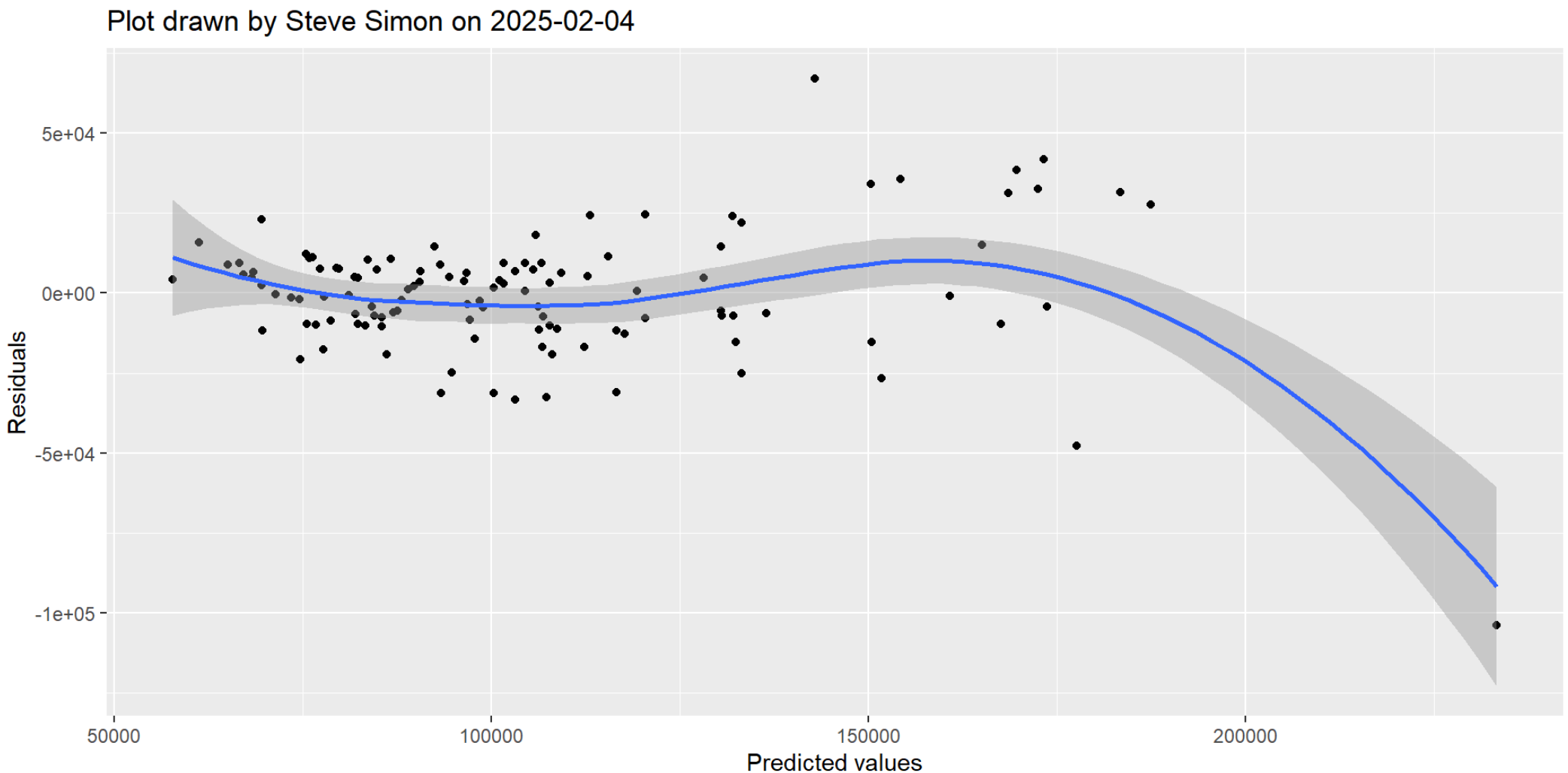
Break #3

- What you have learned
  - Mathematical model
- What’s coming next
  - Assumptions

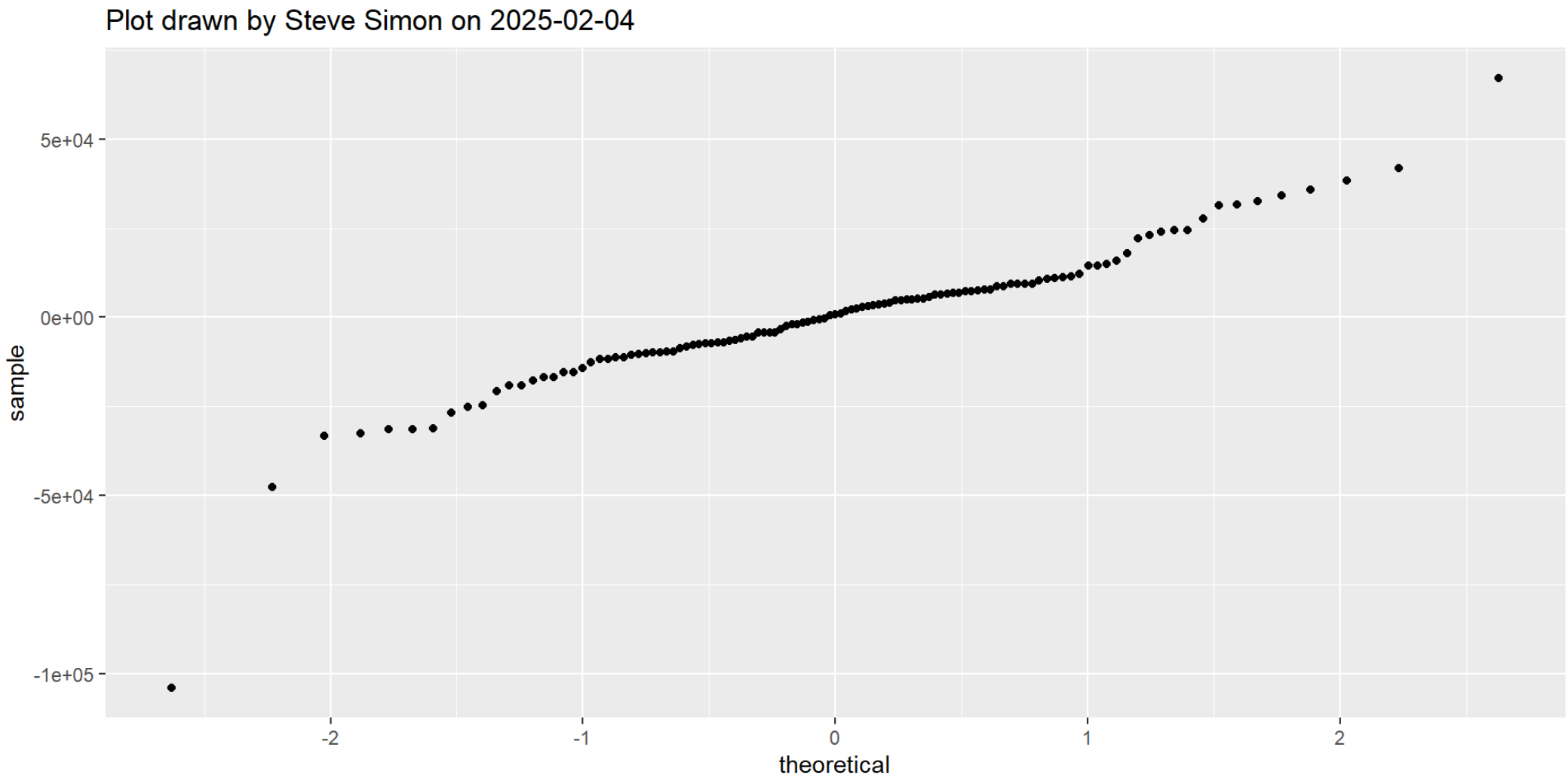
Assumptions of analysis of covariance

- Same as multiple linear regression
  - Equal variances
  - Linearity
  - Normality
- One additional assumption
  - No interaction

Residuals versus predicted



Q-Q plot of residuals



Testing for an interaction

Save this discussion for next week

Live demo, part 4

Refer to the simon-5502-03-demo program.

Break #4

- What you have learned
  - Assumptions
- What's coming next
  - Choosing your covariates

What variable(s) should you include as covariates

- Not recommended
  - Pre-screen for imbalance
  - Pre-screen for correlation
  - Use stepwise regression
- Recommended
  - Use medical/scientific judgement
  - Be expansive

Speaker notes

One of the more difficult questions is what variable or variables should you include as covariates. There are several approaches that are commonly used but which are frowned upon by the experts. First, you should not use any pre-screening. Do not include covariates because they look unbalanced with repsect to the treatment or exposure variable. Don’t include covariates because they appear to be correlated with the outcome variable. Above all do not use stepwise regression to select covariates.

The problem with all these approaches is that they produce residual confounding–confounding that remains even after you do all the adjustemts mandated by the pre-screen or stepwise regression.

What you should do is include covariates based on your medical and scientific expertise. That’s not to say that pre-screening should be totally ignored. Just be sure to include variables that you know are important, even if they don’t show serious imbalance and strong correlation with the outcome. And do not include a variable that shows imbalance or correlation if there is not a plausible scientific or medical reason to include it.

As a general rule, it is best to be expansive in the list of covariates that you adjust for. Keep in mind the rule of 15 (you should have 15 observations for each independent variable in your model).

Using matching instead of covariate adjustment

- The algorithms for matching are tricky
  - Greedy matching
- Allow a small amount of wiggle room
  - Example: age plus or minus 3 years
- Some data will be lost

Speaker notes

You can try to match your control group to the intervention/exposed group, but the actual algorithms are tricky. A simple approach called greedy matching finds the closest match between the a patient in the treatment/exposuregroup and a patient in the control group. That becomes your first pair. Then look for the closest match in the remaining patients. Continue until no more good matches are found. This seems reasonable, but sometimes an early match could be switched to a patient that was just barely not as good, but which allows for better matching in the future.

Matching works best when there is a large pool of control patients to choose from. In any case, you will often find a few patients in the intervention/exposure group that do not have a good match with any of the patients in the control group.

Propensity score models

- Very useful with a large number of covariates
- Logistic model of treatment/exposure versus control
  - Calculate predicted probabilities
  - Known as propensity scores
- Several ways to use propensity scores
  - Use to reweight observations
  - Use to match patients
  - Use as a single composite covariate

Speaker notes

If you have a very large number of covariates, you can use propensity score models. Develop a logistic regression model that ignores the outcome but instead tries to predict whether a patient is in the treatment/exposure group or in the control group based just on the covariates in the model.

The logistic model will produce predicted probabilities. These probabilities are called propensity scores. A high propensity score represents a covariate pattern that looks more like the covariate patterns seen in the treatment/exposure group. A low propensity score represents a covariate pattern that looks more like the covariate patterns seen in the control group.

You will find a few control patients with a high propensity score. These are very important patients. They are the ones which represent “fair” comparisons to the treatment/exposure group. If you give greater weight to these subjects and lesser weight to the control subjects with low propensity scores, you are likely to produce a control population whose weighted covariates look like the covariates in the treatment/exposure group.

You can also use the propensity score to create matches between the treatment/exposure group and the control group. Matching on a single propensity score is a lot simpler than trying to match on several covariates simultaneously.

You can also use the propensity score as a single composite covariate.

The propensity score approach usually works quite well, but it is important to look at various diagnostic tables and graphs that show how much covariate imbalance remains after applying the propensity score through weighting or matching.

If I can find room in the next few modules, I will try to show some actual propensity score models in action.

Break #5

- What you have learned
  - Choosing your covariates
- What’s coming next
  - Your homework

Summary

- What you have learned
  - What is covariate imbalance?
  - Indicator variable coding
  - Mathematical model
  - Assumptions
  - Choosing your covariates