DATA ANALYTICS

Unit 5: Advanced Techniques

Swati Pratap Jagdale
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5: Confounding Variables

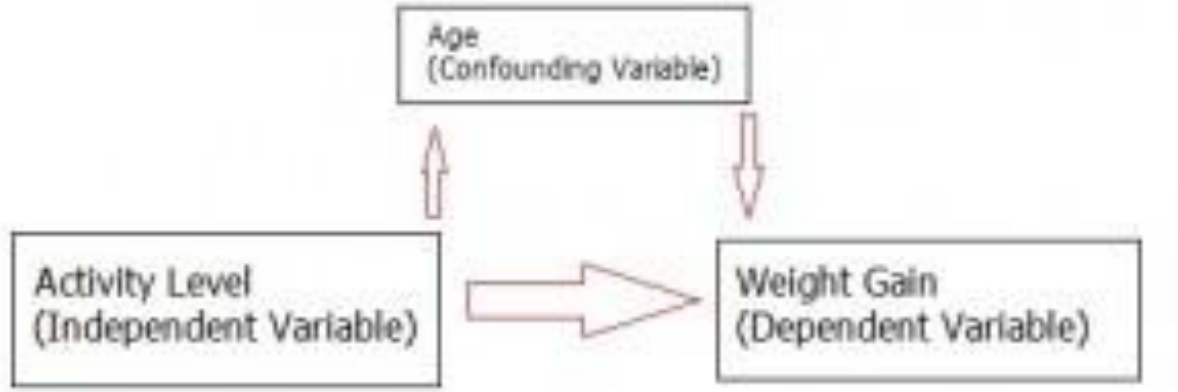
Swati Pratap Jagdale

Department of Computer Science and Engineering

Confounding Variables

- **What is a Confounding Variable?**

- A confounding variable is an “extra” variable that you didn’t account for. They can ruin an experiment and give you useless results. They can suggest there is correlation when in fact there isn’t.
- They can even introduce bias. That’s why it’s important to know what one is, and how to avoid getting them into your experiment in the first place.



A confounding variable can have a hidden effect on your experiment’s outcome.

Confounding Variables

- In an experiment, the independent variable typically has an effect on your dependent variable.
- For example, if you are researching whether lack of exercise leads to weight gain, then
  lack of exercise -- independent variable
  weight gain -- dependent variable.
- Confounding variables are any other variable that also has an effect on your dependent variable. They are like extra independent variables that are having a hidden effect on your dependent variables.

- Confounding variables can cause two major problems:
  - Increase variance
  - Introduce bias.

Confounding Variables

- **Example:**
- You test 200 volunteers (100 men and 100 women). You find that lack of exercise leads to weight gain.
- One problem with your experiment is that is lacks any control variables. For example, the use of placebos, or random assignment to groups.
- So you really can't say for sure whether lack of exercise leads to weight gain. One confounding variable is how much people eat. It's also possible that men eat more than women; this could also make sex a confounding variable.

- A poor study design like this could lead to bias.
- For example, if all of the women in the study were middle-aged, and all of the men were aged 16, age would have a direct effect on weight gain. That makes age a confounding variable.

Confounding Bias

- **Confounding Bias**

- Bias is usually a result of errors in data collection or measurement.
- However, one definition of bias is "***…the tendency of a statistic to overestimate or underestimate a parameter***", so in this sense, confounding is a type of bias.

- Confounding bias is the result of having confounding variables in your model. It has a direction, depending on if it over- or underestimates the effects of your model:
- Positive confounding is when the observed association is biased away from the null. In other words, it overestimates the effect.
- Negative confounding is when the observed association is biased toward the null. In other words, it underestimates the effect.

Confounding Bias

- **How to Reduce Confounding Variables?**

- Make sure you identify all of the possible confounding variables in your study.
- Make a list of everything you can think of and one by one, consider whether those listed items might influence the outcome of your study. Usually, someone has done a similar study before you. So check the academic databases for ideas about what to include on your list.

- Once you have figured out the variables, techniques to reduce the effect of those confounding variables:
  - Bias can be eliminated with random samples.
  - Introduce control variables to control for confounding variables. For example, you could control for age by only measuring 30 year olds.
  - Within subjects designs test the same subjects each time. Anything could happen to the test subject in the “between” period so this doesn’t make for perfect immunity from confounding variables.
  - Counterbalancing can be used if you have paired designs. In counterbalancing, half of the group is measured under condition 1 and half is measured under condition 2.

Confounding Bias

- Omitting confounding variables from your regression model can bias the coefficient estimates.
- When you're assessing the effects of the independent variables in the regression output, this bias can produce the following problems:
  - Overestimate the strength of an effect.
  - Underestimate the strength of an effect.
  - Change the sign of an effect.
  - Mask an effect that actually exists
- Synonyms for Confounding Variables and Omitted Variable Bias
  - confounding variables, confounders, and lurking variables.

Confounding Bias

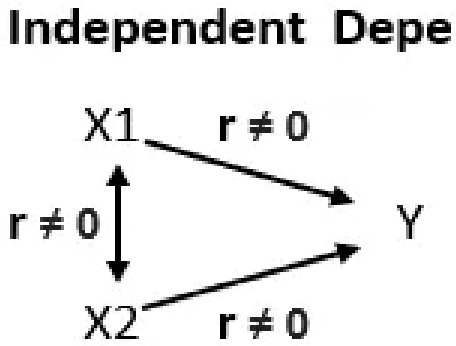
- **What Conditions Cause Omitted Variable Bias?**

- How does this bias occur? How can variables you leave out of the model affect the variables that you include in the model?

- For omitted variable bias to occur, the following two conditions must exist:
- The omitted variable must correlate with the dependent variable.
- The omitted variable must correlate with at least one independent variable that is in the regression model.

Confounding Bias

- There must be non-zero correlations (r) on all three sides of the triangle.
- This correlation structure causes confounding variables that are not in the model to bias the estimates that appear in your regression results. For example, removing either X variable will bias the other X variable.



- The amount of bias depends on the strength of these correlations.
- Strong correlations produce greater bias.
- If the relationships are weak, the bias might not be severe.
- And, if the omitted variable is not correlated with another independent variable at all, excluding it does not produce bias.

Confounding Bias

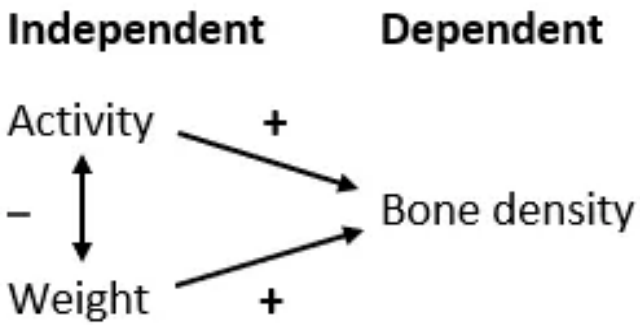
- **Example of How Confounding Variables Can Produce Bias**

Example:

- In a biomechanics lab, One study assessed the effects of physical activity on bone density.
- They measured various characteristics including the subjects' activity levels, their weights, and bone densities among many others.
- Theories about how our bodies build bone suggest that there should be a positive correlation between activity level and bone density. In other words, higher activity produces greater bone density.
- simple regression analysis to determine whether there is a relationship between activity and bone density.....there was no relationship at all!!!

Confounding Bias

- They included activity level as the only independent variable, but it turns out there is another variable that correlates with both activity and bone density—the subject's weight.

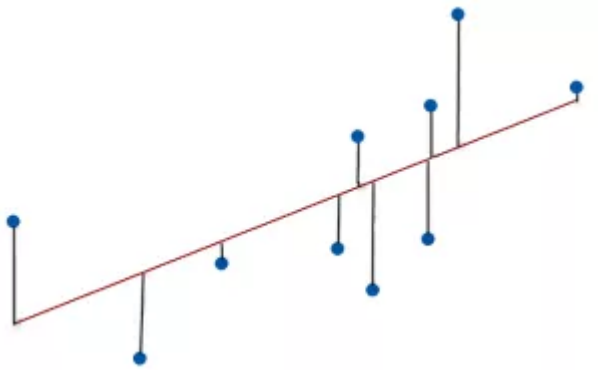


The diagram below shows the signs of the correlations between the variables.

Confounding Bias

- Correlations, Residuals, and OLS Assumptions

- When you satisfy the ordinary least squares (OLS) assumptions, the Gauss-Markov theorem states that your estimates will be unbiased and have minimum variance.



Residuals = Observed value – Fitted value

Omitted variable bias occurs because the residuals violate one of the assumptions.

Confounding Bias

- Consider, regression model with two significant independent variables, X1 and X2. These independent variables correlate with each other and the dependent variable—which are the requirements for omitted variable bias.

- Now, imagine that we take variable X2 out of the model. It is the confounding variable. Here's what happens:

- The model fits the data less well because we've removed a significant explanatory variable. Consequently, the gap between the observed values and the fitted values increases. These gaps are the residuals.
- The degree to which each residual increases depends on the relationship between X2 and the dependent variable. Consequently, the residuals correlate with X2.
- X1 correlates with X2, and X2 correlates with the residuals. Ergo, variable X1 correlates with the residuals.
- This condition violates the ordinary least squares assumption that independent variables in the model do not correlate with the residuals. Violations of this assumption produce biased estimates.

Confounding Bias

	Included and Omitted: Negative Correlation	Included and Omitted: Positive Correlation
Included and Dependent: Negative Correlation	Positive bias: coefficient is overestimated.	Negative bias: coefficient is underestimated.
Included and Dependent: Positive Correlation	Negative bias: coefficient is underestimated.	Positive bias: coefficient is overestimated.

The table summarizes these relationships and the direction of bias.

References

https://www.statisticshowto.com/experimental-design/confounding-variable/

https://statisticsbyjim.com/regression/confounding-variables-bias/

THANK YOU

Swati Pratap Jagdale
Department of Computer Science
swatigambhire@pes.edu