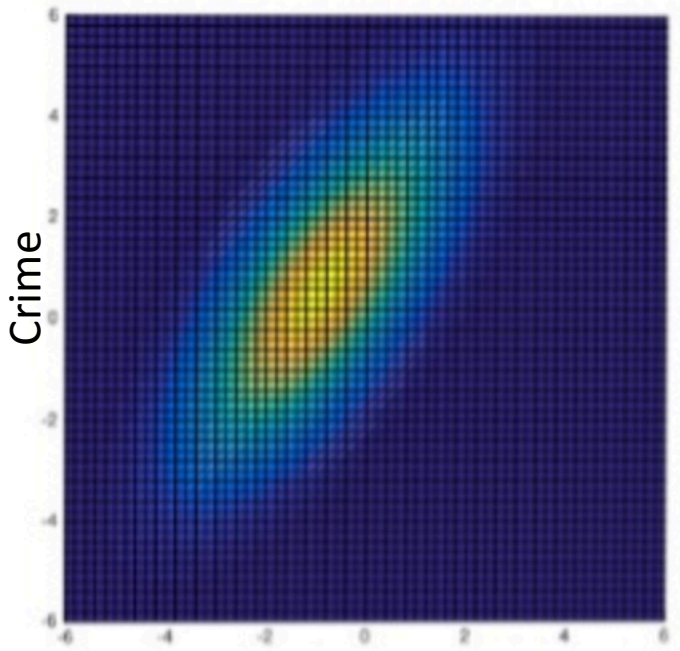
Causality and the Error Term

Crime and Ice Cream

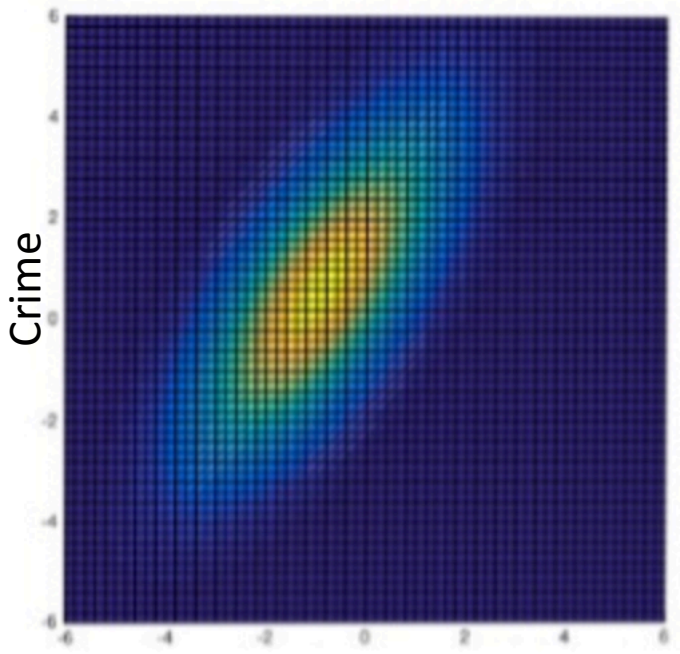
- You may have heard that there is a positive correlation between ice cream sales and crime.
- Suppose this heat map represents the joint distribution between the two variables.
- Now we collect a random sample from this distribution, and fit a linear regression:

  E(Crime) = .1 + 2.1 Sales
- Is this the causal effect of ice cream sales on crime?
  - If I buy another ice cream cone, does that result in 2.1 more crimes?
- You would say: probably not!
  - In fact, some research suggests that this pattern is caused by temperature.
  - Temperature is what we call an omitted variable.
- A crucial point: You can't know that from the joint distribution.



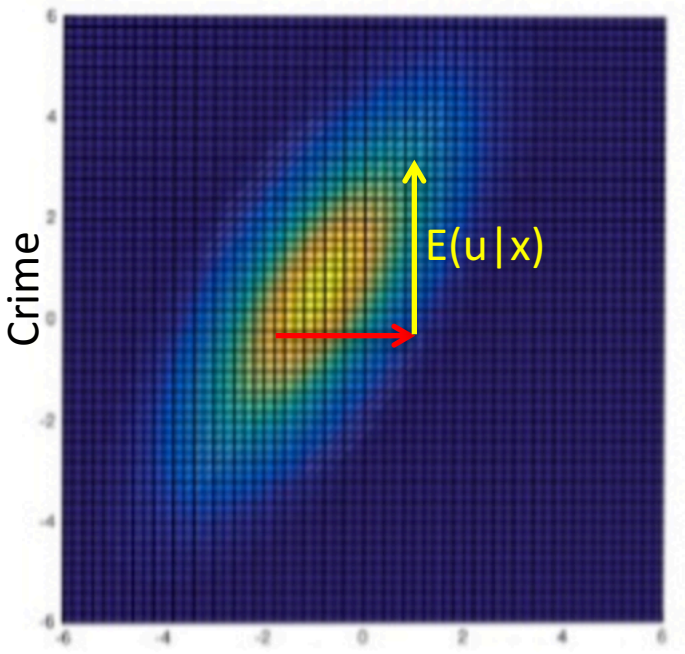
Crime and Ice Cream

- If I didn't tell you what the axes represent, you can't know whether x causes y, y causes x, or if some other variable causes both.
- A joint distribution tells you about the relative occurrences of x and y – it says nothing about manipulations to x.
- The same is true of data drawn from a joint distribution
- To make a causal argument you absolutely have to use background knowledge in some way.
- Again, if all you had was the joint distribution, or the data that comes from it, you can't know that there's an omitted variable.
  - You need background knowledge.



Crime and Ice Cream

- Causality is closely related to the error term.
- From a causal modeling perspective, the error term isn't just random noise – it represents the effect of all unmeasured factors.
- Let's represent that in this picture.
- The red arrow represents what would happen if I went out and bought a bunch of ice cream today.
- The yellow arrow represents a typical error away from this model.
  - This represents the expected effect of warmer temperatures on days with high ice cream sales.
- Really, the causal model is
- Crime = β_0 + 0 sales + u
- where u includes the effect of temperature.



Crime and Ice Cream

- What does this mean for linear regression?
- To have consistent estimates, we have to meet the exogeneity assumption.
- $cov(x_i,u) = 0$
- I added a couple more yellow arrows representing typical errors in our model.
- You can estimate that $cov(x_i,u) > 0$
- We have violated exogeneity.
  - We call x an *endogenous* variable. This is a variable correlated with the error term.
- The implication is that OLS is NOT consistent.
  - OLS cannot estimate the red line.
  - OLS doesn't know about omitted variables, all it can do is find a line that fits the joint distribution closely, but that's not the line that we want.