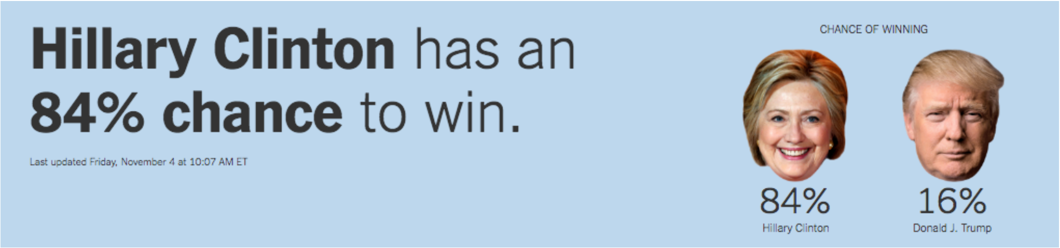
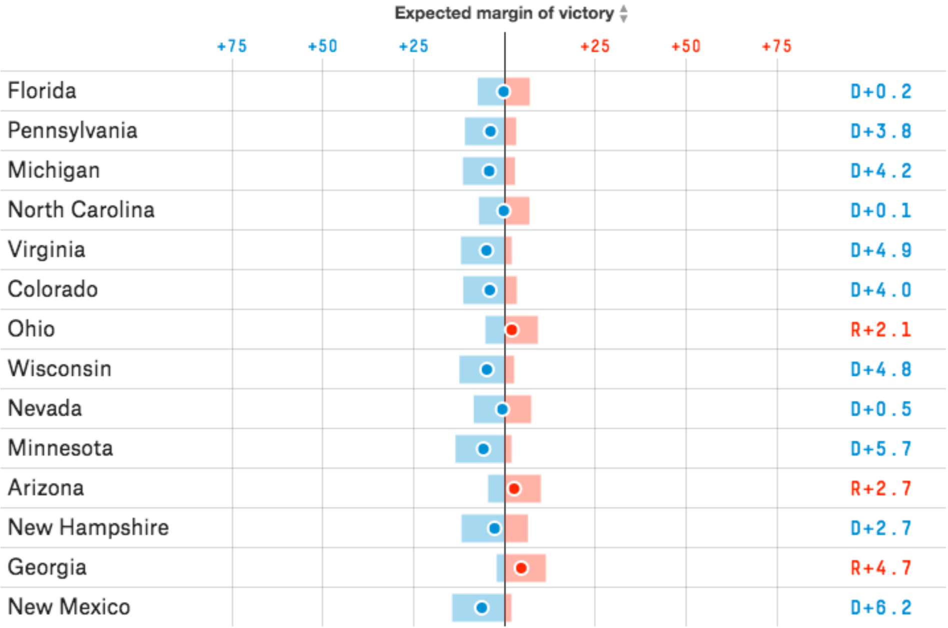
Uncertainty and Prediction

Prediction

Prediction


Chance of winning
Hillary Clinton
67.7%
Donald Trump
32.3%


Hillary Clinton has an
84% chance to win.
Last updated Friday, November 4 at 10:07 AM ET
CHANCE OF WINNING
84%
Hillary Clinton
16%
Donald J. Trump


Expected margin of victory
+75
+50
+25
+25
+50
+75
Florida
D+0.2
Pennsylvania
D+3.8
Michigan
D+4.2
North Carolina
D+0.1
Virginia
D+4.9
Colorado
D+4.0
Ohio
R+2.1
Wisconsin
D+4.8
Nevada
D+0.5
Minnesota
D+5.7
Arizona
R+2.7
New Hampshire
D+2.7
Georgia
R+4.7
New Mexico
D+6.2

Error in a Regression Model

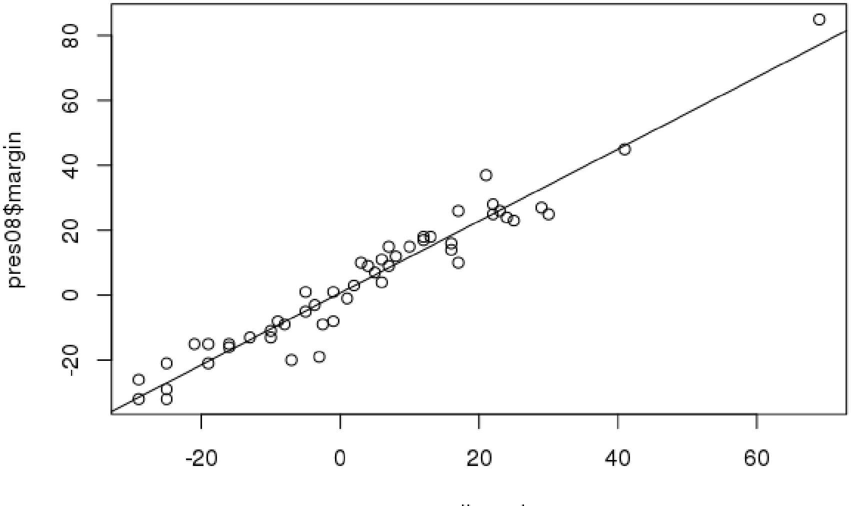
$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

Error in a Regression Model

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

Obama's Vote Share 2008 = 0.71 + 1.11 * Poll Margin

Error in a Regression Model



Standard Error of Coefficients

Error comes from:

Standard Error of Coefficients

Error comes from:

1. The model itself (RMSE)

Standard Error of Coefficients

Error comes from:

1. The model itself (RMSE)
2. Variation in the independent variable

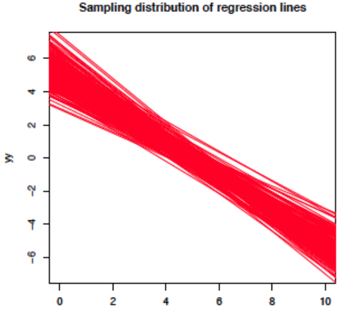
Standard Error of Coefficients

Error comes from:

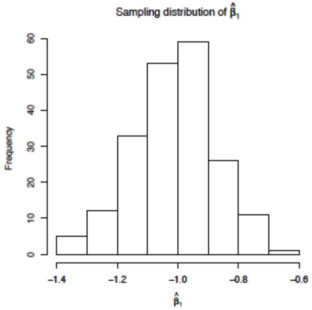
1. The model itself (RMSE)
2. Variation in the independent variable

To understand the second part, imagine sampling the underlying distribution used to build the model:

Standard Error of Coefficients



$\rightarrow$



Regression as Causal Analysis

We can use regression to test hypotheses!

Regression as Causal Analysis

We can use regression to test hypotheses! What is the effect of pipe smoking on mortality?

Regression as Causal Analysis

We can use regression to test hypotheses! What is the effect of pipe smoking on mortality?

Pipe Smoking $= 0$ or 1 (Binary variable)

Regression as Causal Analysis

We can use regression to test hypotheses! What is the effect of pipe smoking on mortality?

Pipe Smoking = 0 or 1 (Binary variable)

$$Death_i = \alpha + \beta_i Pipe\ Smoking_i + \epsilon_i$$

Regression as Causal Analysis

We can use regression to test hypotheses! What is the effect of pipe smoking on mortality?

Pipe Smoking = 0 or 1 (Binary variable)

$$Death_i = \alpha + \beta_i Pipe\ Smoking_i + \epsilon_i$$

Can we interpret the slope coefficient β_i as a measure of the causal effect of Pipe Smoking?

Regression as Causal Analysis

$$Treatment\ Effect = avg(Death_{Pipe}) - avg(Death_{No\ Pipe})$$

Regression as Causal Analysis

$$Treatment\ Effect = avg(Death_{Pipe}) - avg(Death_{No\ Pipe})$$

Control:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (0)$$

Regression as Causal Analysis

$$Treatment\ Effect = avg(Death_{Pipe}) - avg(Death_{No\ Pipe})$$

Control:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (0)$$

$$Death = \alpha + \epsilon_i$$

Regression as Causal Analysis

$$Treatment\ Effect = avg(Death_{Pipe}) - avg(Death_{No\ Pipe})$$

Control:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (0)$$

$$Death = \alpha + \epsilon_i$$

Treated:

Regression as Causal Analysis

$$Treatment\ Effect = avg(Death_{Pipe}) - avg(Death_{No\ Pipe})$$

Control:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (0)$$

$$Death = \alpha + \epsilon_i$$

Treated:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (1)$$

Regression as Causal Analysis

$$Treatment\ Effect = avg(Death_{Pipe}) - avg(Death_{No\ Pipe})$$

Control:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (0)$$

$$Death = \alpha + \epsilon_i$$

Treated:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (1)$$

$$Death = (\alpha + \beta_i) + \epsilon_i$$

Regression as Causal Analysis

$$Treatment\ Effect = avg(Death_{Pipe}) - avg(Death_{No\ Pipe})$$

$$Death_i = \alpha + \beta_i Pipe\ Smoking_i + \epsilon_i$$

Control:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (0)$$

$$Death = \alpha + \epsilon_i$$

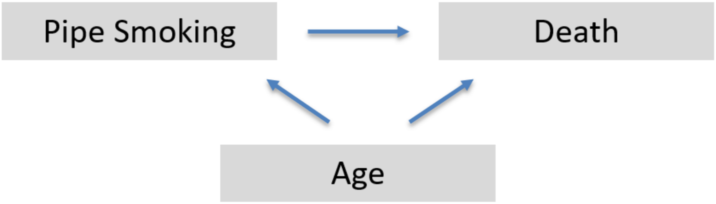
Treated:

$$Death = \alpha + \beta_i Pipe\ Smoking\ (1)$$

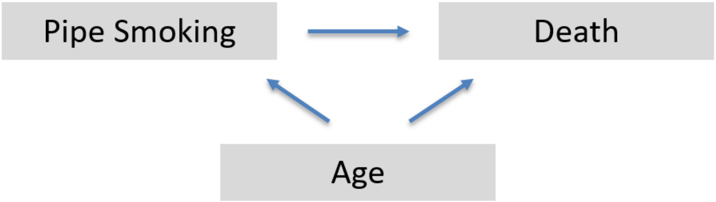
$$Death = (\alpha + \beta_i) + \epsilon_i$$

Treatment Effect = Treated – Control = $((\alpha + \beta_i) + \epsilon_i \breve{}\,(\alpha \breve{}\,\epsilon_i) = \beta_i)$

Regression as Causal Analysis

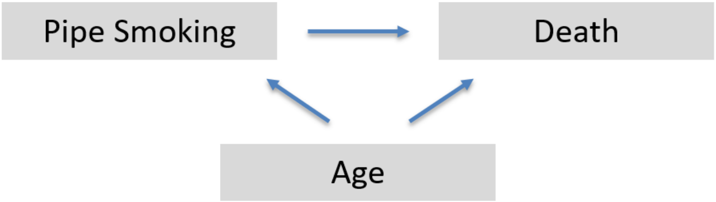


Regression as Causal Analysis



If 'age' is omitted, and age is correlated with pipe smoking **and** death, then we can prove that β_1 is biased.

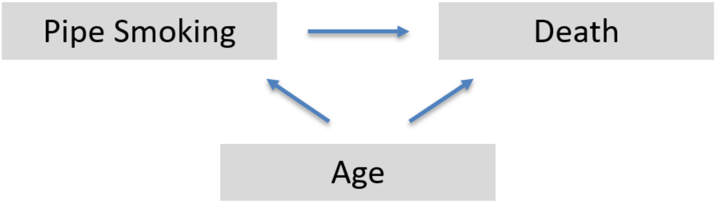
Regression as Causal Analysis



If ‘age’ is omitted, and age is correlated with pipe smoking **and** death, then we can prove that β_1 is biased.

$$Death_i = \alpha + \beta_1 Pipe\ Smoking_i + \beta_2 Age_i + \epsilon_i$$

Regression as Causal Analysis

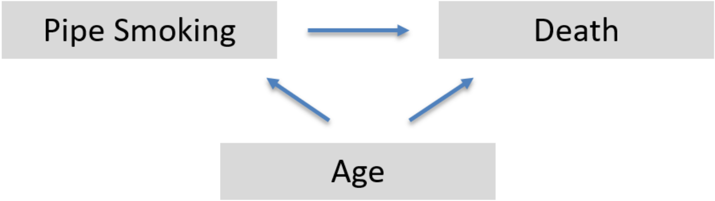


If 'age' is omitted, and age is correlated with pipe smoking **and** death, then we can prove that β_1 is biased.

$$Death_i = \alpha + \beta_1 Pipe\ Smoking_i + \beta_2 Age_i + \epsilon_i$$

This approach is better! But our estimate of β_1 may still be biased by other confounders; in regression this possibility is known as *"omitted variable bias"*

Regression as Causal Analysis



Regression as Causal Analysis

$$Death_i = \alpha + \beta_1 Pipe\ Smoking_i + \epsilon_i$$

Regression as Causal Analysis

$$Death_i = \alpha + \beta_1 Pipe\ Smoking_i + \epsilon_i$$

Assuming no other confounders; is β_1 equivalent to a treatment effect?

Regression as Causal Analysis

$$Death_i = \alpha + \beta_1 Pipe\ Smoking_i + \epsilon_i$$

Assuming no other confounders; is β_1 equivalent to a treatment effect?

- If there are no other confounders, then we can show that β_1 is equivalent to the causal effect of pipe smoking on death.

Regression as Causal Analysis

$$Death_i = \alpha + \beta_1 Pipe\ Smoking_i + \epsilon_i$$

Assuming no other confounders; is β_1 equivalent to a treatment effect?

- If there are no other confounders, then we can show that β_1 is equivalent to the causal effect of pipe smoking on death.
- In practice, the claim of "no confounders" is only credible in a randomized experiment or a natural experiment, in which the treatment condition has been randomly assigned.

Regression as Causal Analysis

$$Death_i = \alpha + \beta_1 Pipe\ Smoking_i + \epsilon_i$$

Assuming no other confounders; is β_1 equivalent to a treatment effect?

- If there are no other confounders, then we can show that β_1 is equivalent to the causal effect of pipe smoking on death.
- In practice, the claim of “no confounders” is only credible in a randomized experiment or a natural experiment, in which the treatment condition has been randomly assigned.
- **Thus, omitted variable bias can only be fully removed by *careful research design.***

Regression as Causal Analysis

$$Death_i = \alpha + \beta_1 Pipe\ Smoking_i + \epsilon_i$$

Assuming no other confounders; is β_1 equivalent to a treatment effect?

- If there are no other confounders, then we can show that β_1 is equivalent to the causal effect of pipe smoking on death.
- In practice, the claim of "no confounders" is only credible in a randomized experiment or a natural experiment, in which the treatment condition has been randomly assigned.
- **Thus, omitted variable bias can only be fully removed by *careful research design.***
- (Note that biased regression coefficients can still be theoretically informative)

Multiple Treatment Effects

- Regression models offer a useful way to simultaneously measure the effect of multiple treatments

GOTV Experiment

Treatment Group 1 – “Neighbor Shaming”

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	

GOTV Experiment

Treatment Group 2 – Hawthorne Effect

- Received a letter in the mail saying that it is a civic duty to vote.
- Letter indicates that researchers are watching whether they voted

GOTV Experiment

Treatment Group 2 – Hawthorne Effect
- Received a letter in the mail saying that it is a civic duty to vote.
- Letter indicates that researchers are watching whether they voted

Treatment Group 3 – Civic Duty
- Received a letter in the mail saying that it is a civic duty to vote.

GOTV Experiment

Treatment Group 2 – Hawthorne Effect
- Received a letter in the mail saying that it is a civic duty to vote.
- Letter indicates that researchers are watching whether they voted

Treatment Group 3 – Civic Duty
- Received a letter in the mail saying that it is a civic duty to vote.

Control Group – No Letter

GOTV Experiment

Treatment Group 2 – Hawthorne Effect

- Received a letter in the mail saying that it is a civic duty to vote.
- Letter indicates that researchers are watching whether they voted

Treatment Group 3 – Civic Duty

- Received a letter in the mail saying that it is a civic duty to vote.

Control Group – No Letter

Registered voters were randomly assigned to a group

GOTV Experiment

$$Vote_i = \alpha + \beta_1 Neighbors + \epsilon_(i)$$

GOTV Experiment

$$Vote_i = \alpha + \beta_1 Neighbors + \epsilon_(i)$$

$\beta_1 \rightarrow$ Expected Difference in Vote, above α

GOTV Experiment

$$Vote_i = \alpha + \beta_1 Neighbors + \epsilon_(i)$$

$\beta_1 \rightarrow$ Expected Difference in Vote, above α

$\alpha \rightarrow$ Expected baseline level of Vote for: ?

GOTV Experiment

$$Vote_i = \alpha + \beta_1 Neighbors + \beta_2 Hawthorne + \beta_3 Civic + \beta_4 Control + \epsilon_i$$

GOTV Experiment

$$Vote_i = \alpha + \beta_1 Neighbors + \beta_2 Hawthorne + \beta_3 Civic + \beta_4 Control + \epsilon_i$$

This regression **fails**. There is no possible set of observations to use to fit the intercept (α)

GOTV Experiment

$$Vote_i = \alpha + \beta_1 Neighbors + \beta_2 Hawthorne + \beta_3 Civic + \varepsilon_i$$

Control

Neighbors - Control

Hawthorne - Control

Civic - Control

Non-binary variables

- Usually we use binary variables (indicating a treatment group) to calculate causal effects.

Non-binary variables

- Usually we use binary variables (indicating a treatment group) to calculate causal effects.
- But if the level of a continuous variable is randomly assigned (for instance, a medical dose; the number of refugees assigned to a city), then the same logic holds.

Non-binary variables

- Usually we use binary variables (indicating a treatment group) to calculate causal effects.
- But if the level of a continuous variable is randomly assigned (for instance, a medical dose; the number of refugees assigned to a city), then the same logic holds.

$$\text{Opposition}_i = \alpha + \beta_1\% \text{ Refugees} + \varepsilon_i$$

Opposition if assigned 0 refugees (control)

Effect of influx; measured
in unit change (each additional
1% has effect β_1)