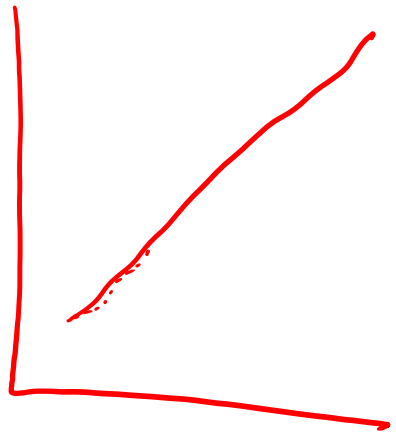
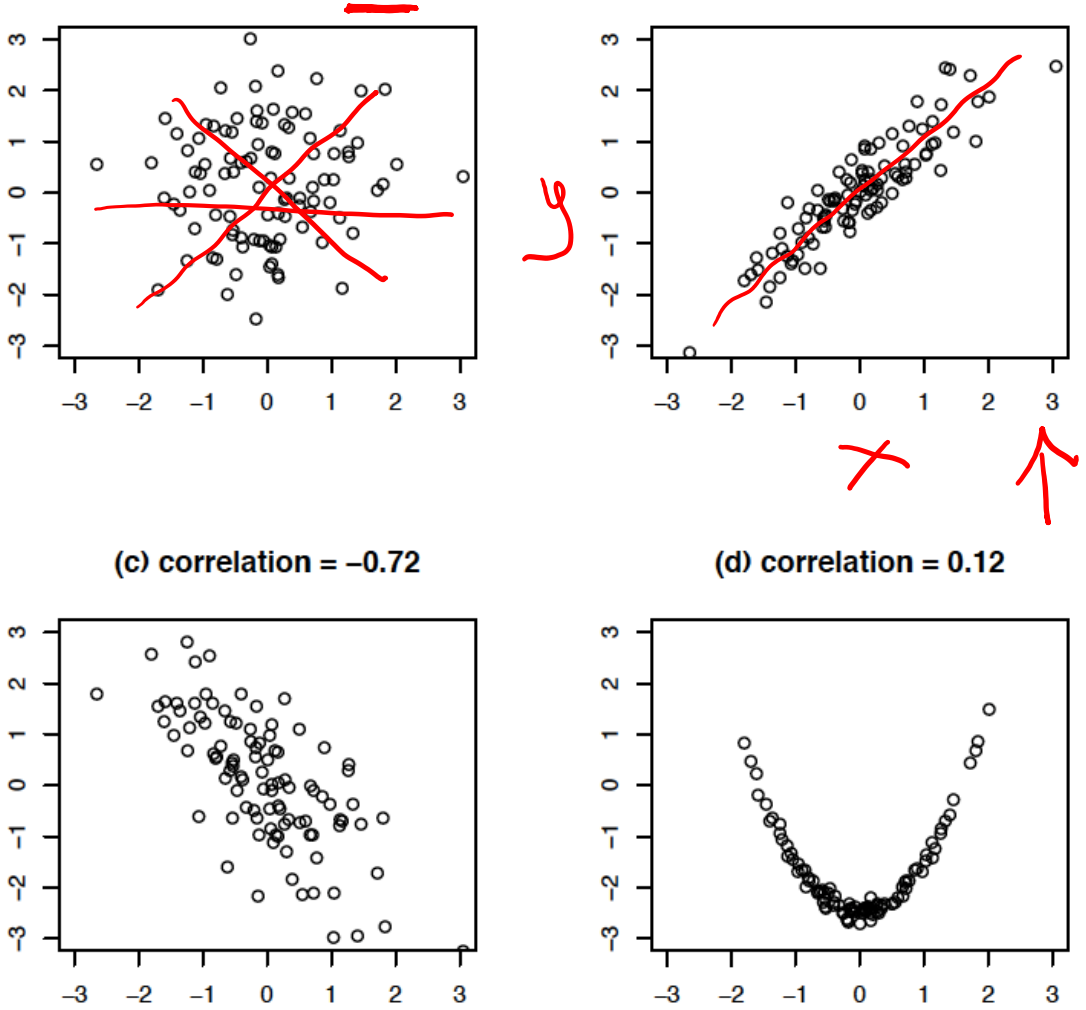
Regression

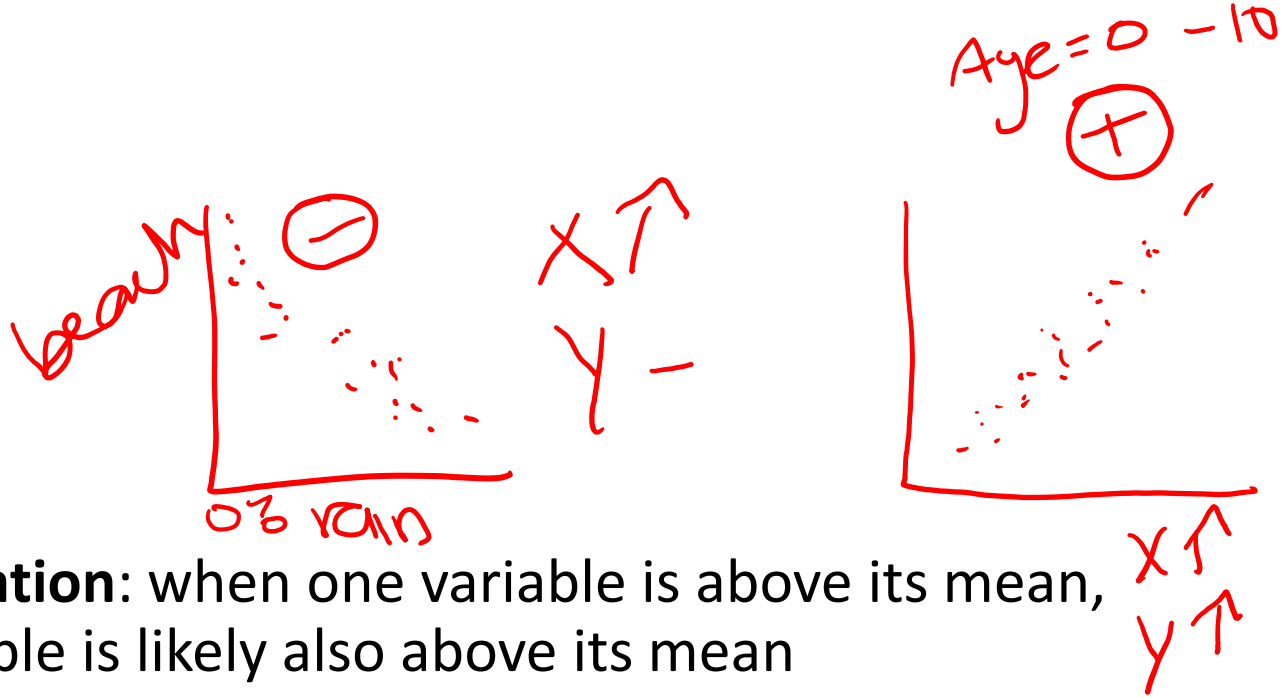
Correlation

- Recall that up to this point, we have mostly examined the *correlation* between two variables

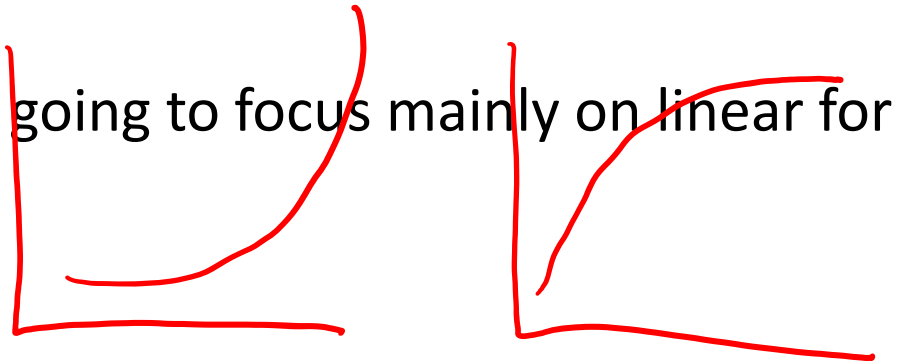




Correlation



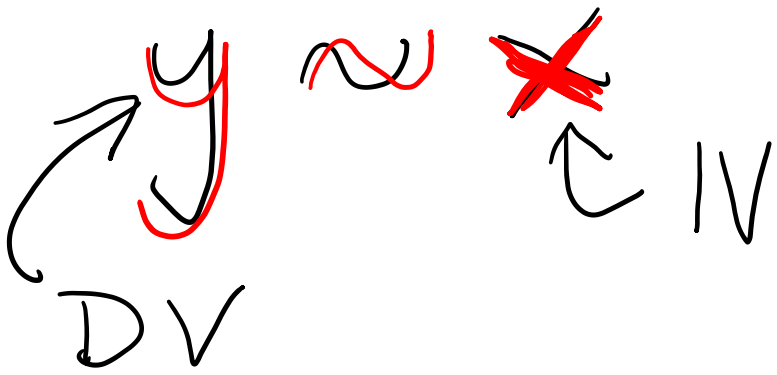
- **Positive correlation**: when one variable is above its mean, the other variable is likely also above its mean
- **Negative correlation**: when one variable is above its mean, the other variable is likely below its mean
- Correlation is often not useful for nonlinear relationships between variables
- But that's okay, we are going to focus mainly on linear for now



Correlation and Causation

≠

- We know correlation does not equal causation
- By the way: what is the fundamental problem of causal inference?
- How can we get causation?
- We looked into causation so far with experiments (mostly) and designs like diff in diff… but these are not the most common methods
- **Linear regression with control variables**

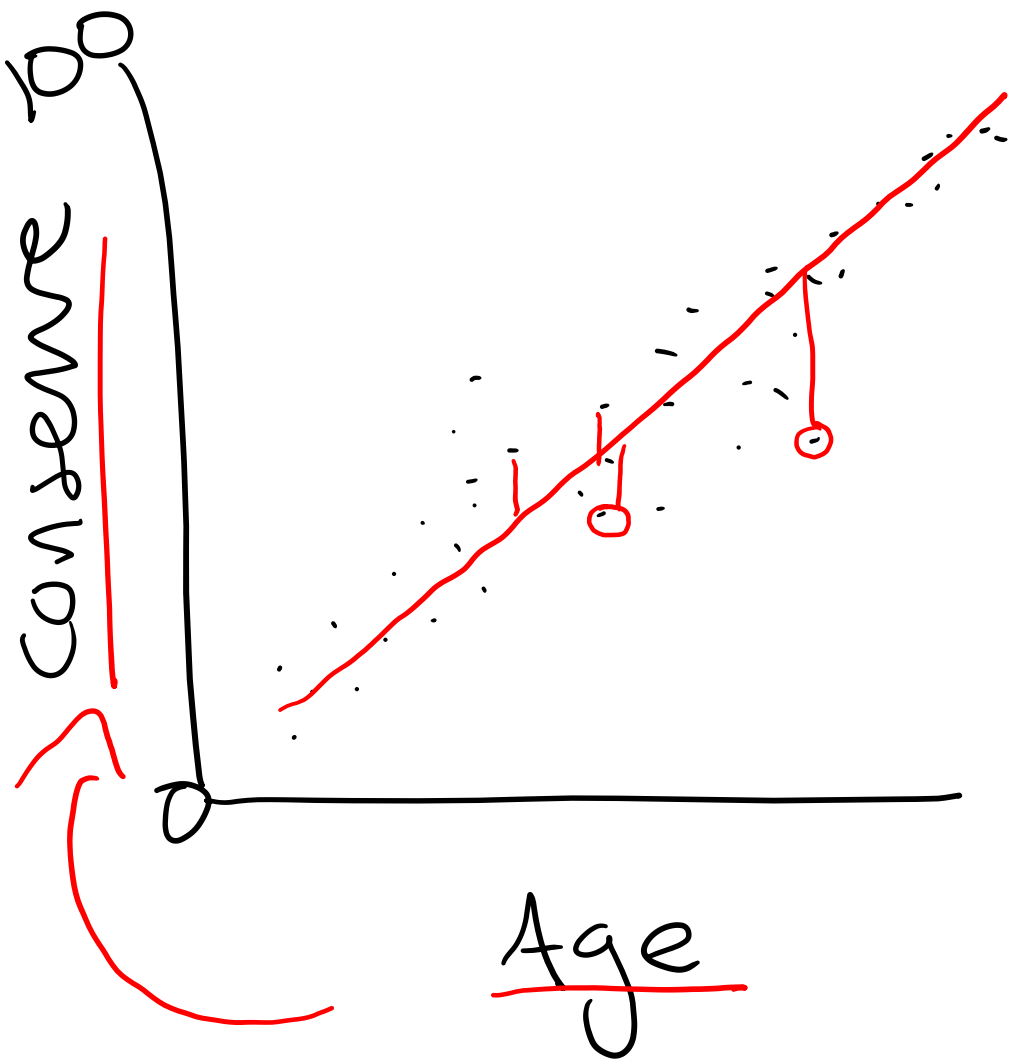


← causality

IV x = Age

DV y = 0 - 100 conservatism

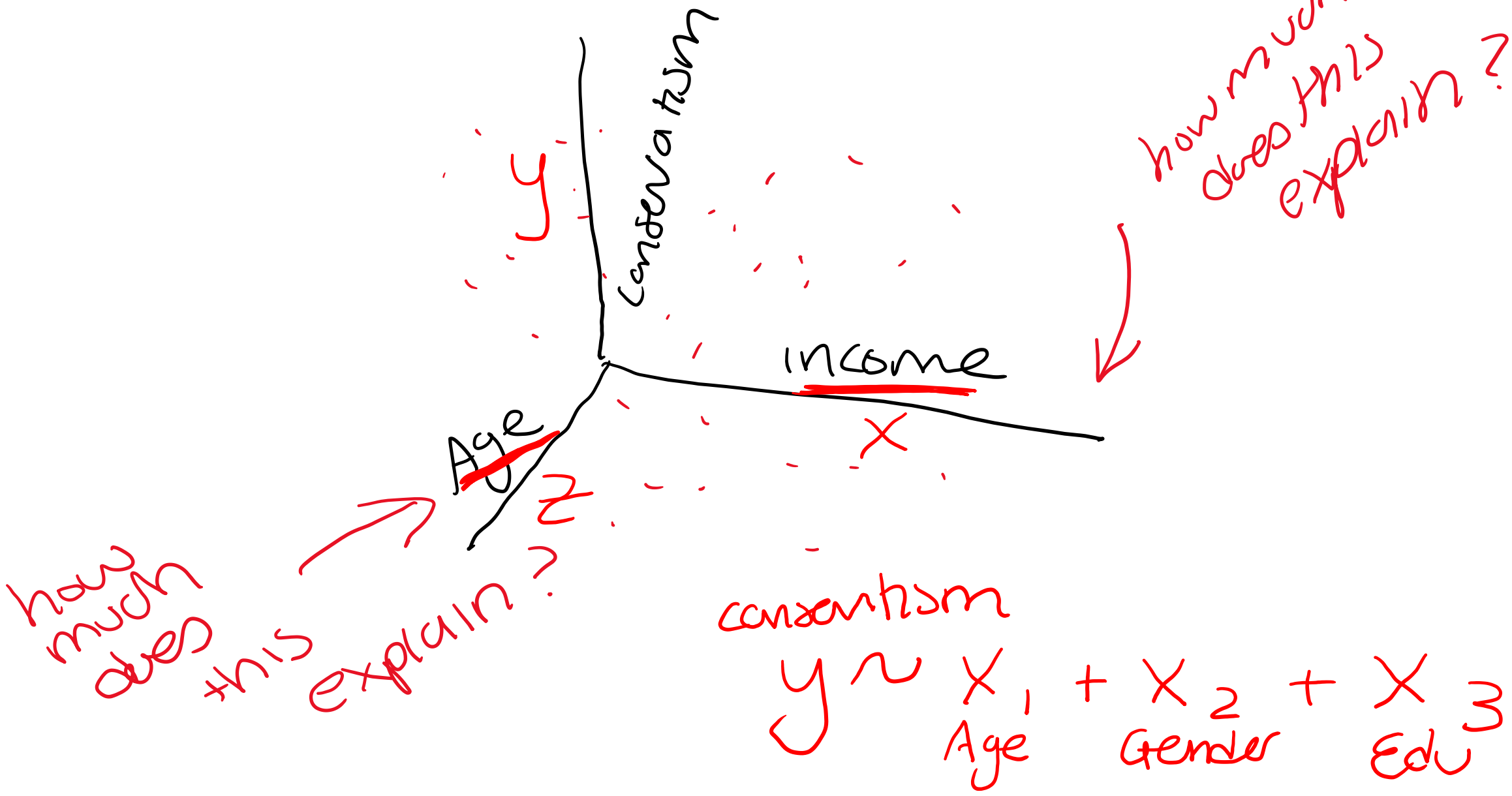
Can someone's age predict how conservative they will be?



① imposing a theory on relationship

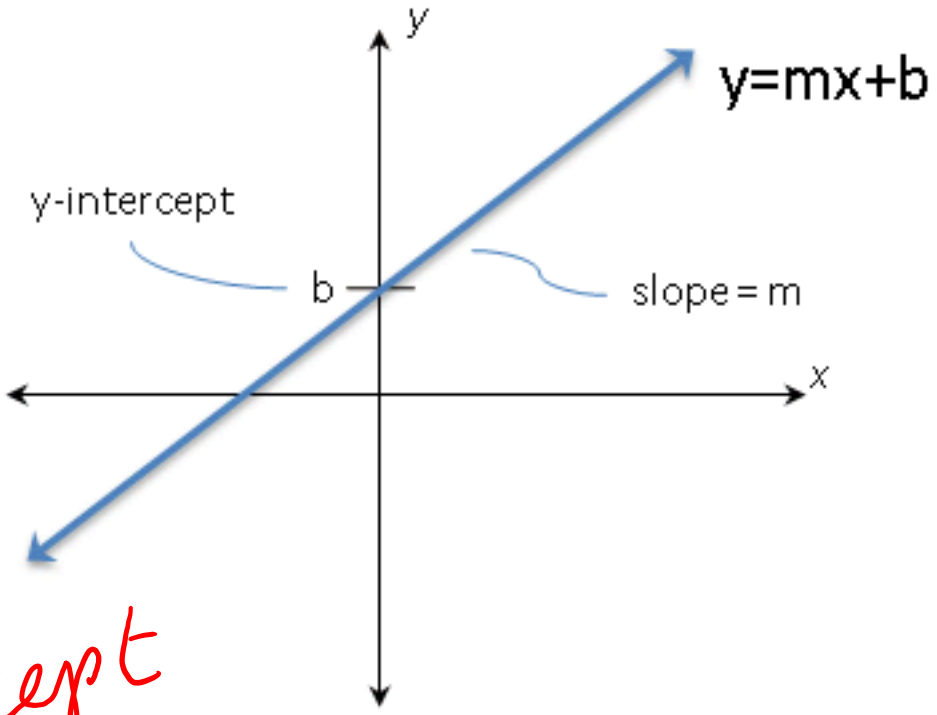
② do this w/ line
→ linear relationship

③ incorporating error


conservatism
y
income
X
Age
Z
how much does this explain?
how much does this explain?
conservatism
$y \sim X_1 + X_2 + X_3$
Age
Gender
Edu

Linear Model

- Recall from algebra:
- Y = mx + b
- We adapt this model for statistics…



y = mx + b

slope

?

Aye

intercept

Linear regression

Ordinary Least Squares (OLS) model

$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$

- Y = outcome variable (response variable) – DV
- X = predictor or independent variable – IV
- Alpha (α) is the intercept
- Beta (β) is the slope
- Epsilon (ε) is the error term (also called the residual)
- *Note for OLS, the DV (Y) needs to be continuous.

1 X
0.5

X → [Y] 0 1
classification

OLS

α intercept β slope ε error

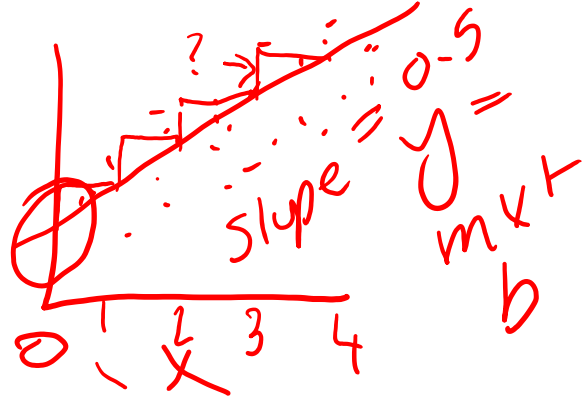
- Assumes a data generating process (DGP):
  - The relationship between X and Y is linear
- May not be true, but can still be useful

OLS

- Moving to prediction
- Predicting the outcome variable, Y
- Past correlation and on to causation
- What *causes* Y? Does X have a role in *causing* Y?

How much can age tell us about conservatism

$$Y = \alpha + \beta X + \varepsilon$$

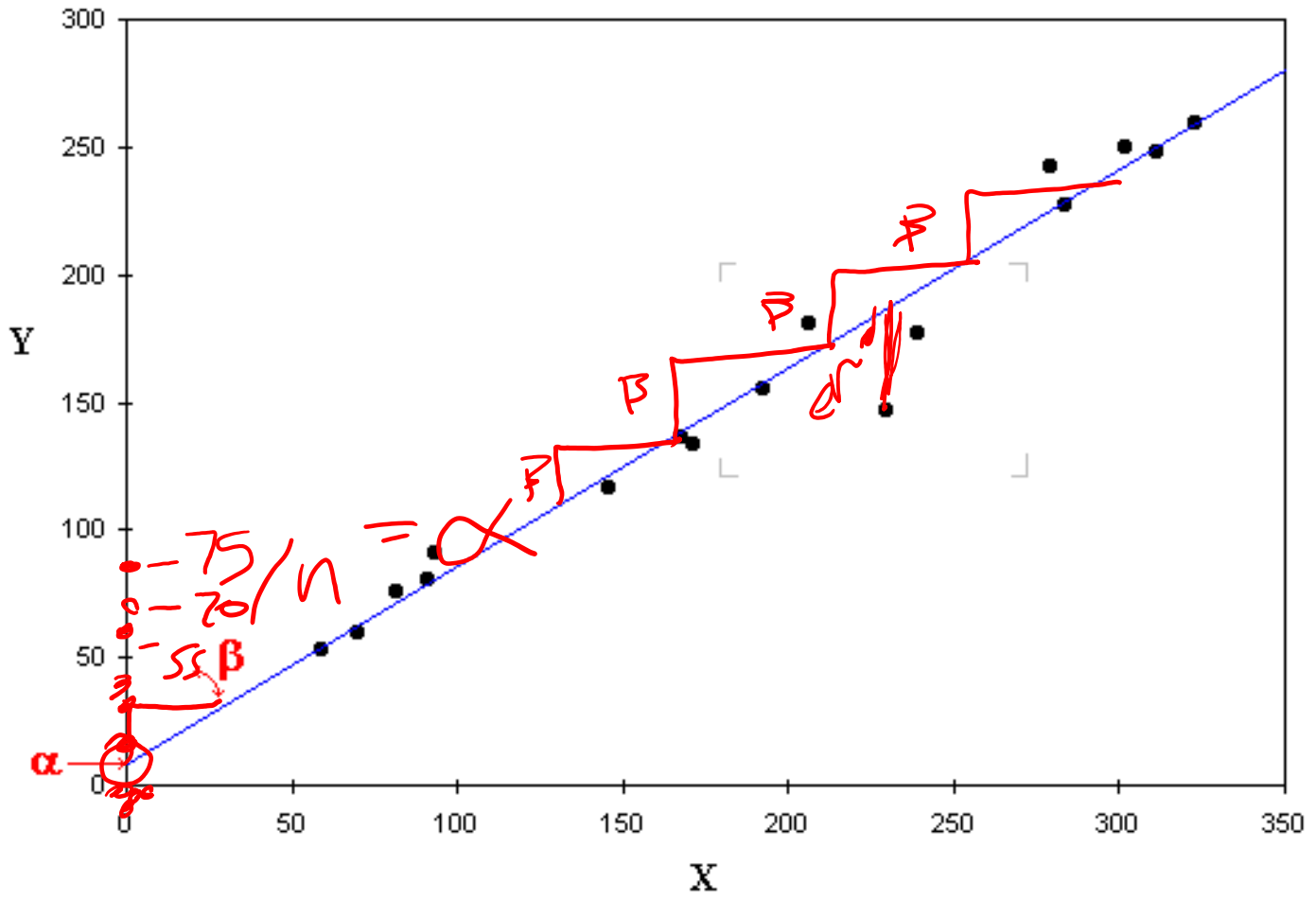


mean @ X = 0

- The intercept (α) is the average value of Y when X is zero
- The slope (β) is the average increase in Y that corresponds to a one unit increase in X
  - Unit will vary based on how the variable is coded

0.5 (X) ← 2

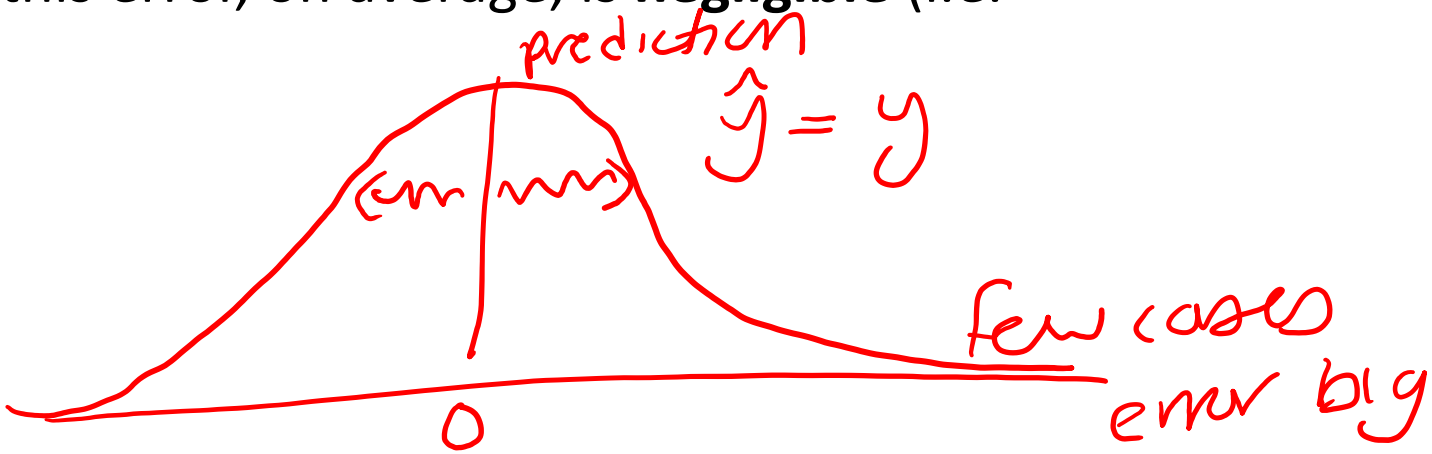
- Together, the intercept and slope are called ***coefficients***
- The error term allows for some deviation from perfection in our model
- Main difference between algebra and statistics: introduction of the error term



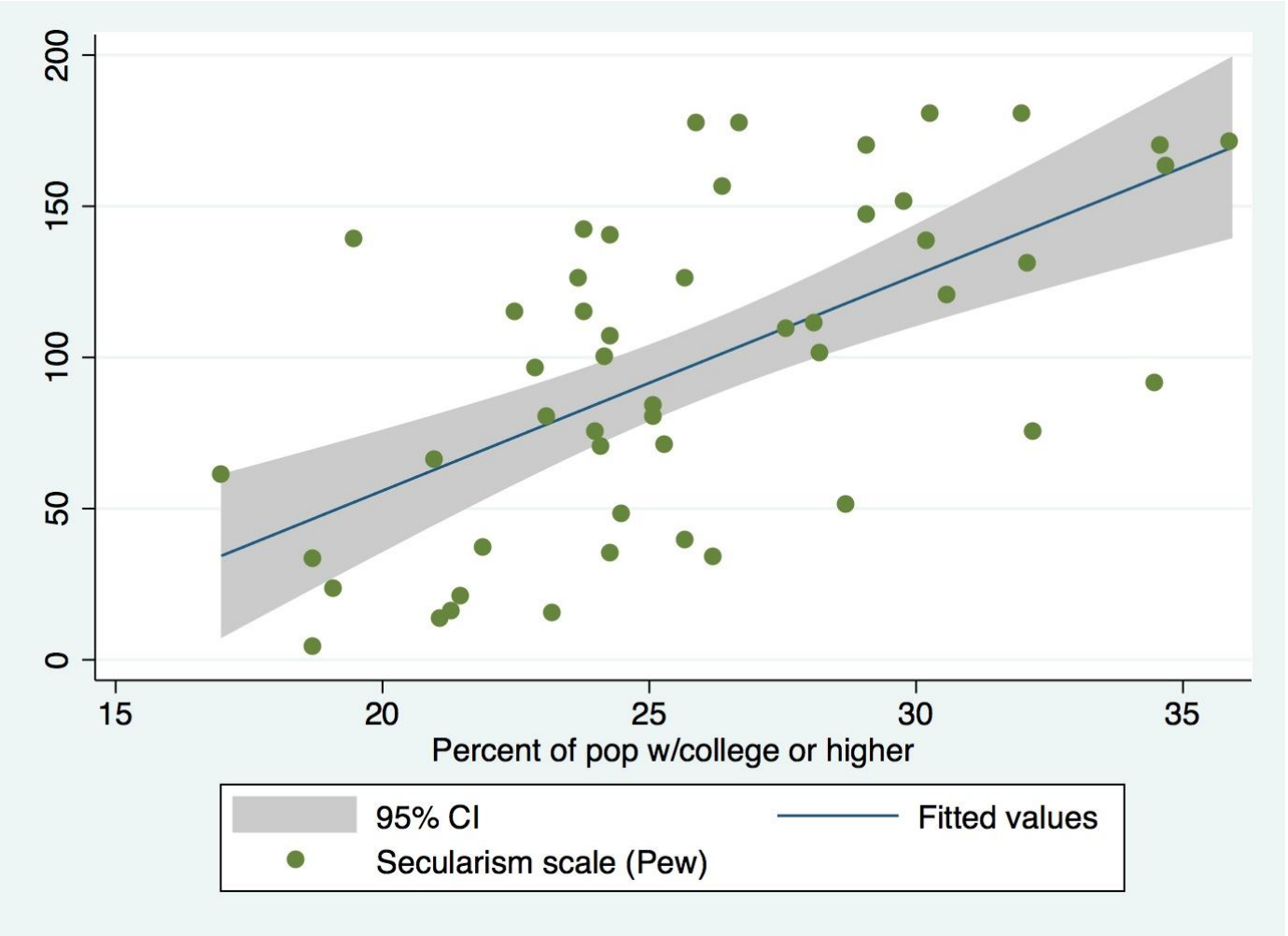
- Y is the dependent variable (outcome variable we want to explain)
- X is the independent variable (variable that seeks to explain Y)
- Alpha (α) is the intercept
- Beta (β) is the slope of the best fitting line
- The dots are the data points
- Note that the data points are not perfectly on the line
- This difference, represents some error (ε)

About this error…

- "all models are wrong, but some are useful" – George Box (statistician)
- There's a fundamental random component to human activity
- This error will vary with each observation
- We assume that this error, on average, is **negligible** (i.e. has a mean of 0)

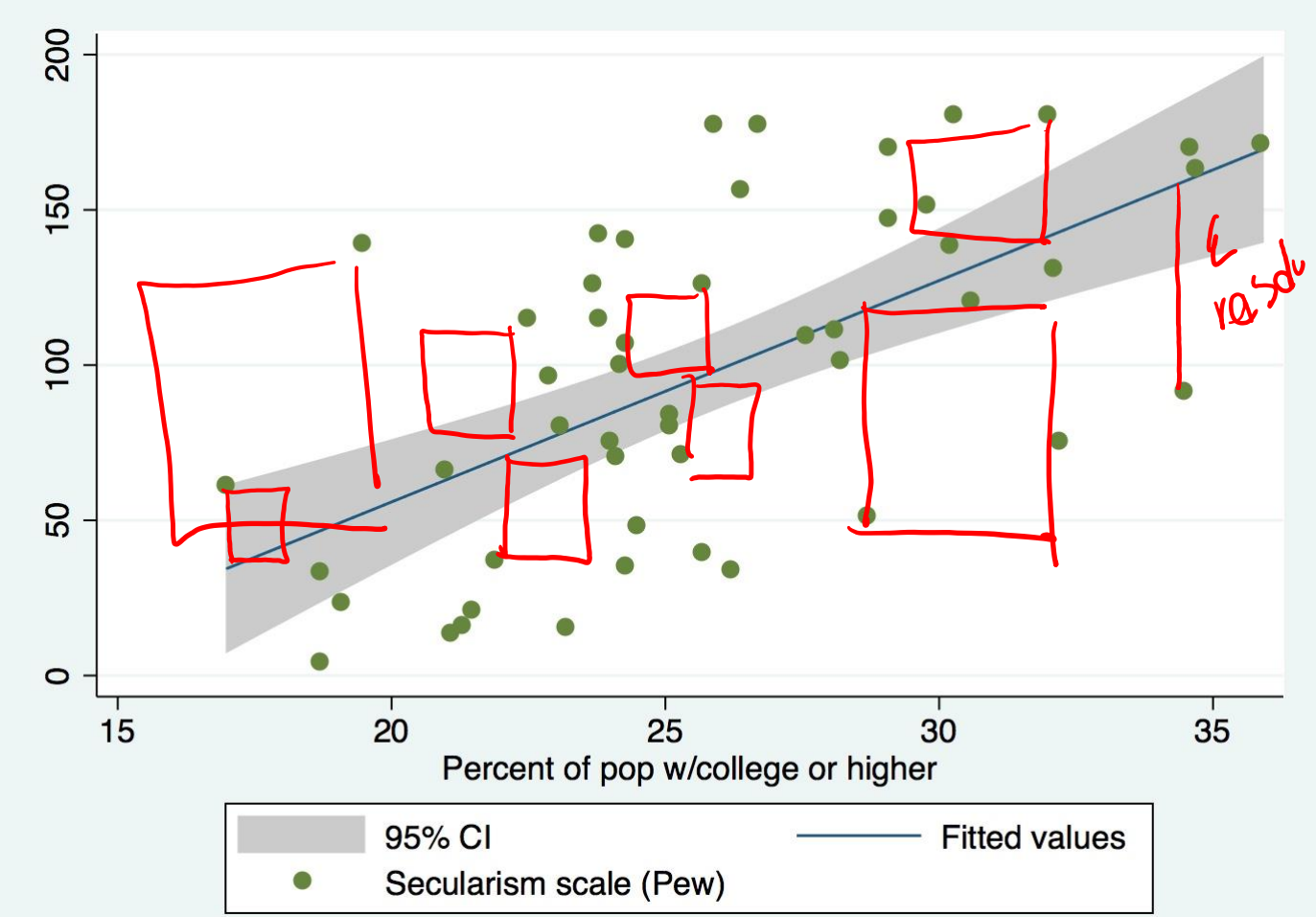


OLS



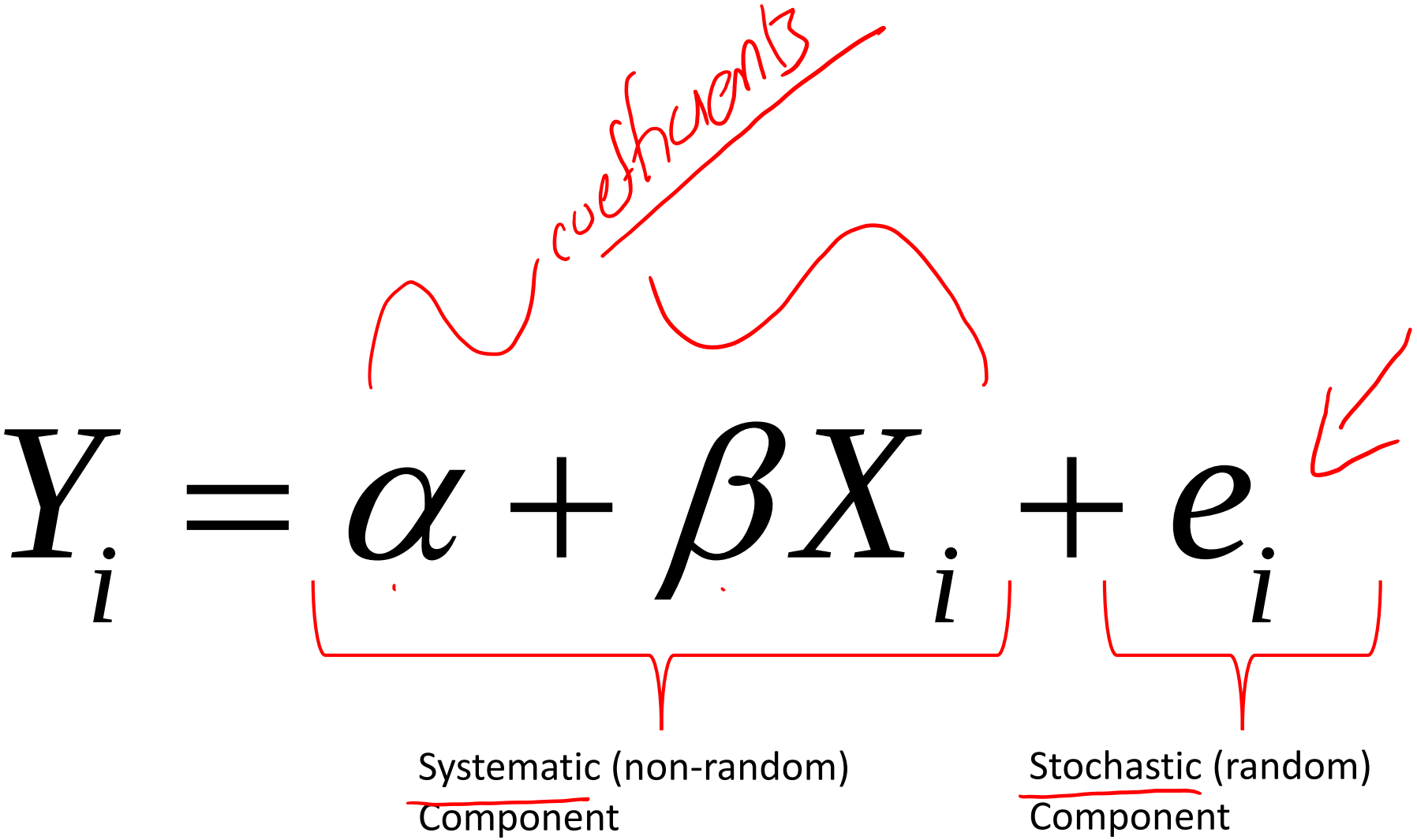
- If there were a truly *deterministic* relationship between two variables, then knowing the slope and intercept would tell you what value of Y goes with each value of X.
- BUT remember there is *random variation* or *randomness*

OLS

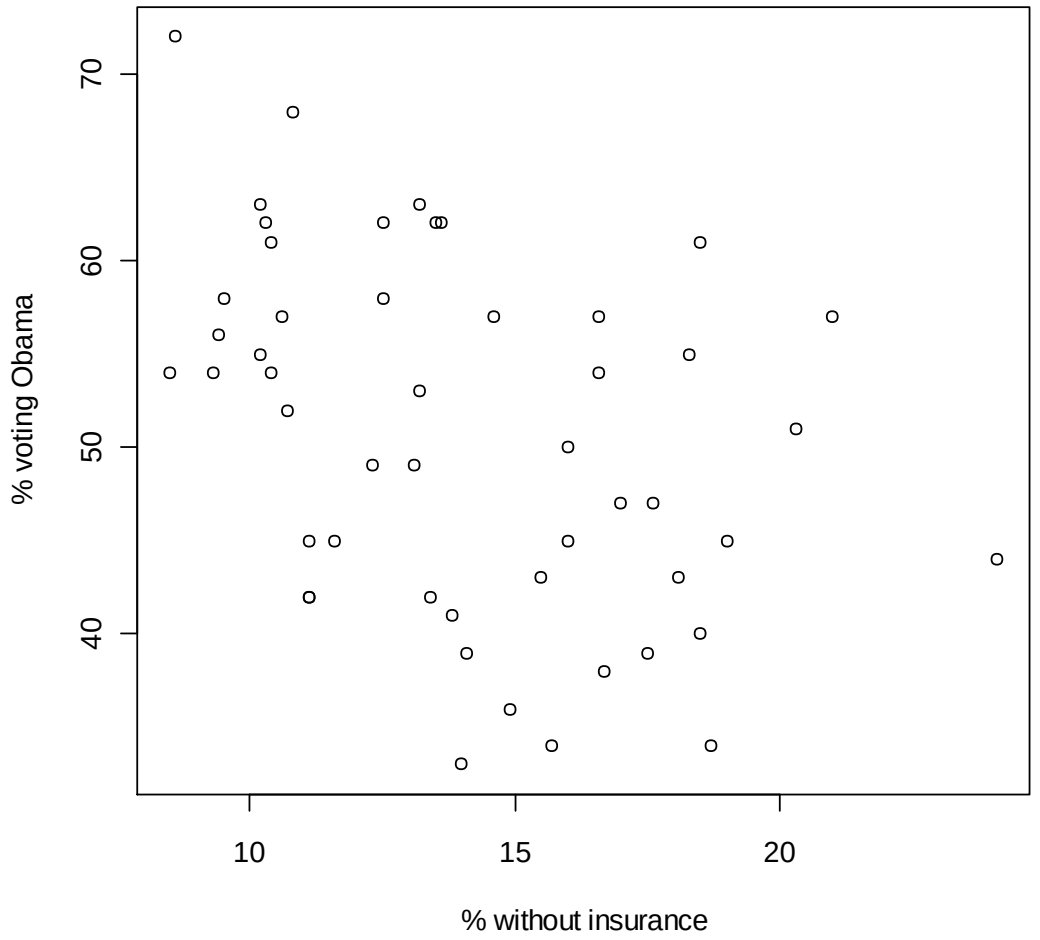


To incorporate randomness, we add an "error term" or "residual" that acknowledges the line will not perfectly go through each point.

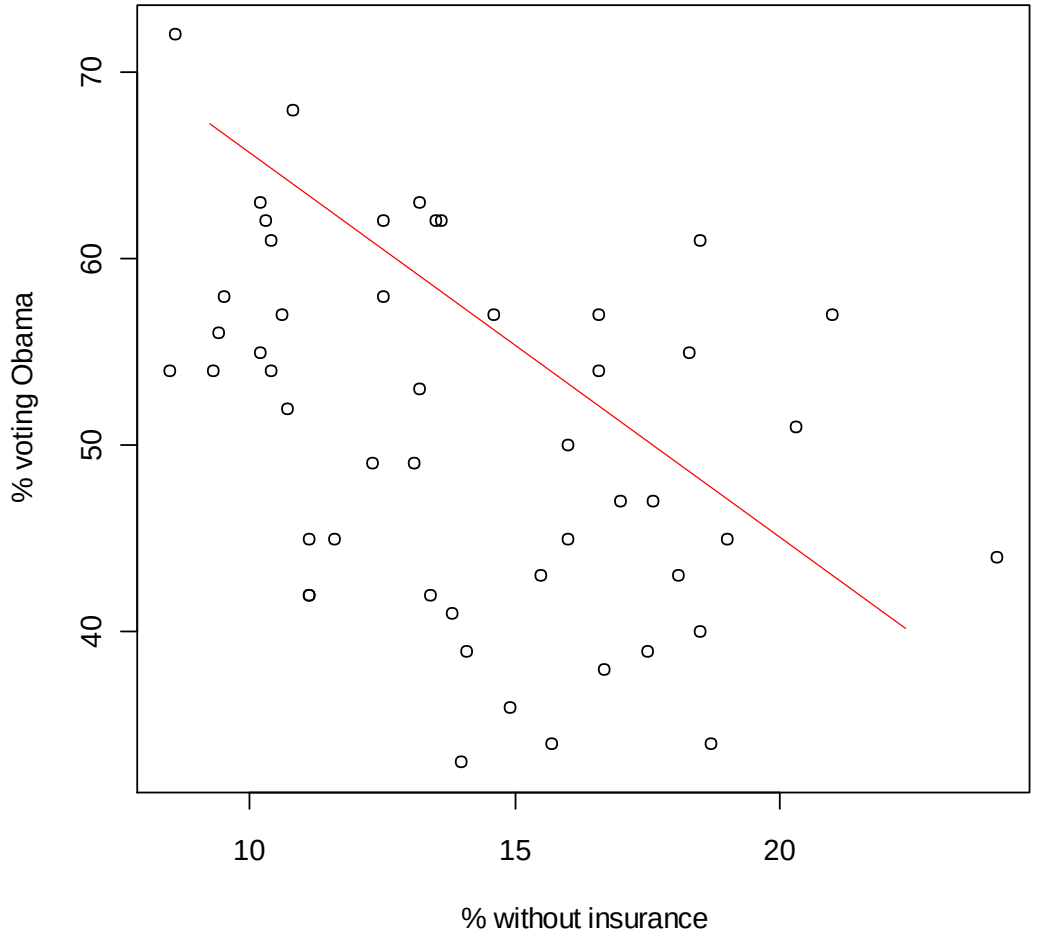
close to 0 as possible


coefficients
$Y_i = \alpha + \beta X_i + e_i$
Systematic (non-random) Component
Stochastic (random) Component

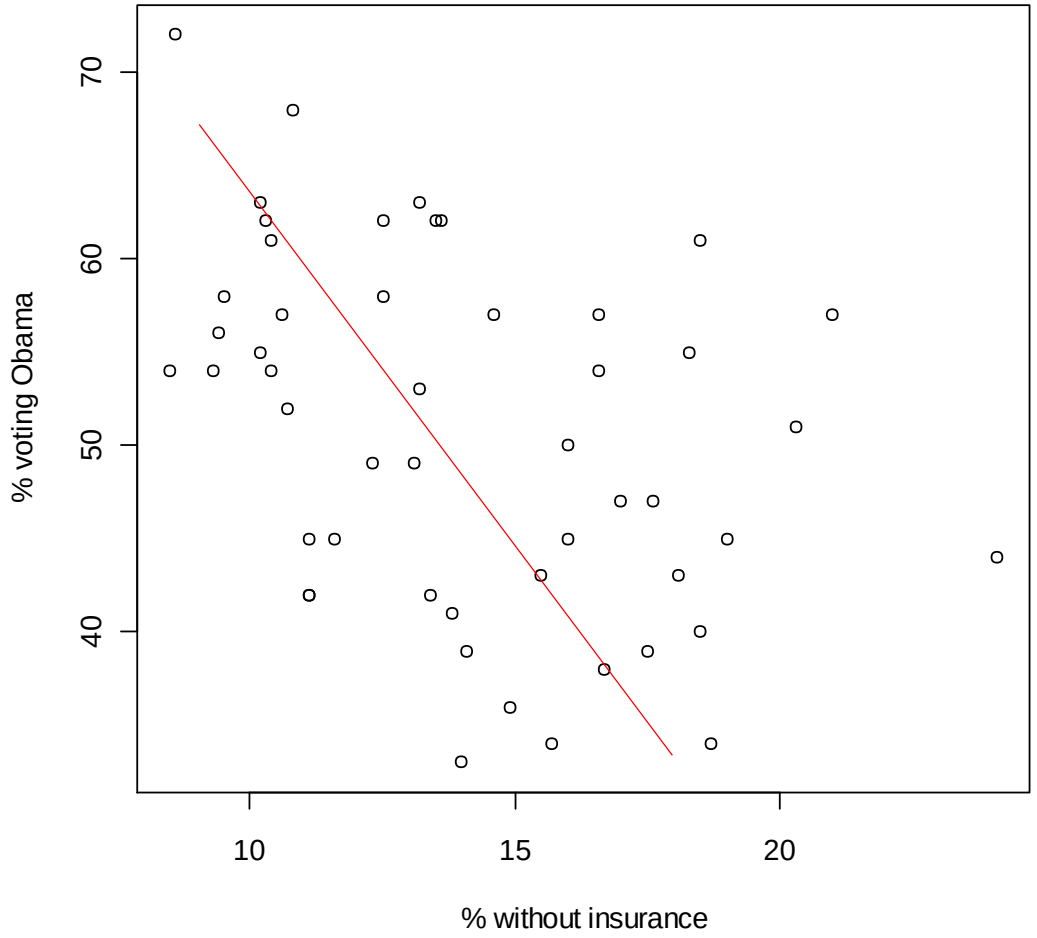
Best-Fit Line:
Percent w/out Insurance & % Voting Obama



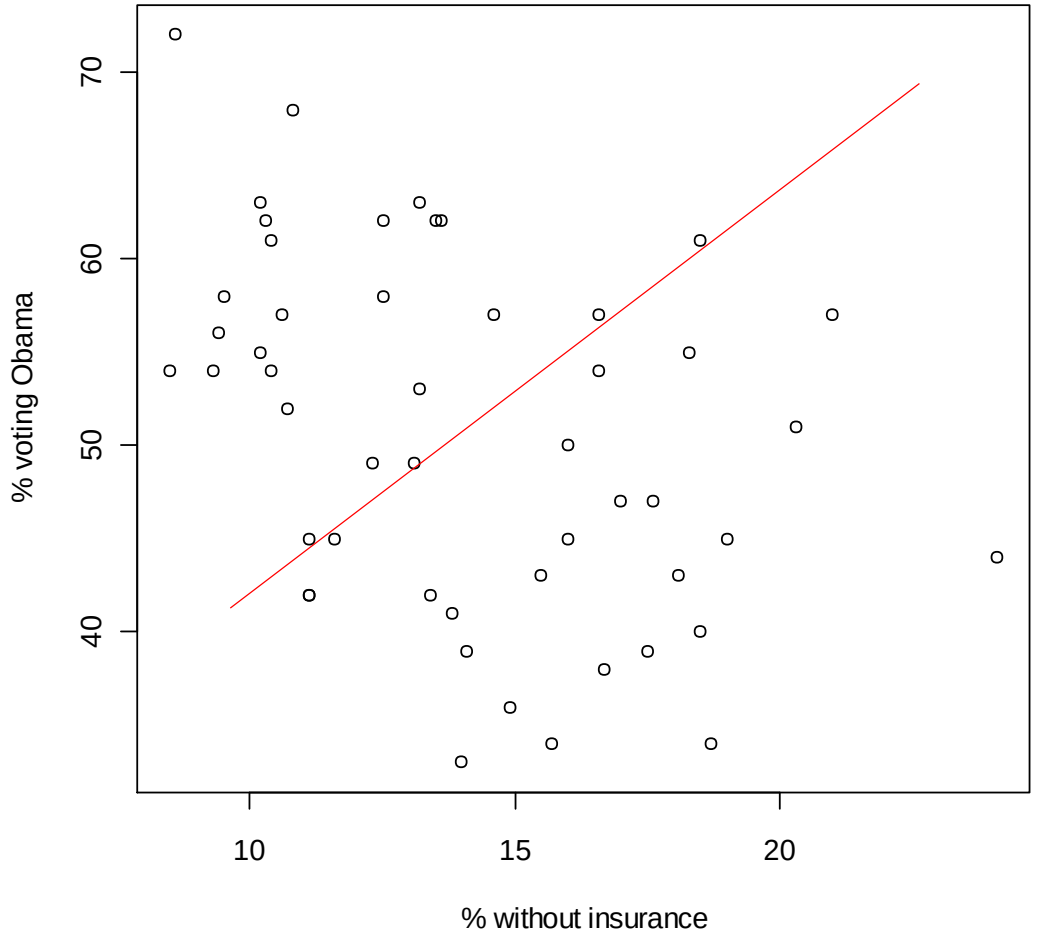
Best-Fit Line:
Percent w/out Insurance & % Voting Obama



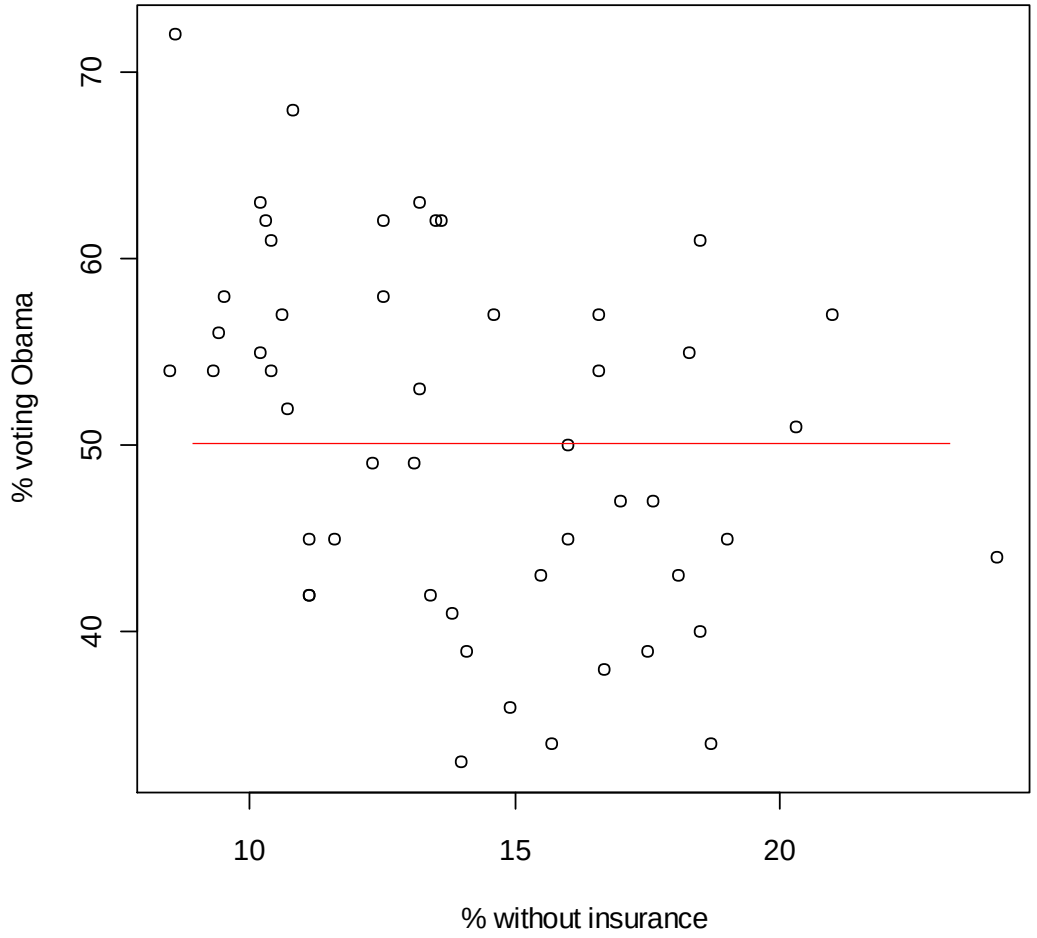
Best-Fit Line:
Percent w/out Insurance & % Voting Obama



Best-Fit Line:
Percent w/out Insurance & % Voting Obama



Best-Fit Line:
Percent w/out Insurance & % Voting Obama



Best-Fit Line….

…is the one that minimize the total error.

$$\sum (e_i)^2$$

- This is what we calculate, statistically
- This is why R can be really helpful to us!
- Rather than calculating the errors, squaring them, and summing them up by hand, we can use the computer

Estimates

- Once we obtain these estimates, the notation changes slightly
- We add hats, to demonstrate that they are our estimates, that we are no longer talking theoretically, but instead about specific estimates we obtained from our model
- Estimated error is the actual value of Y minus its predicted value, obtained from the statistical estimate

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x$$

intercept slope

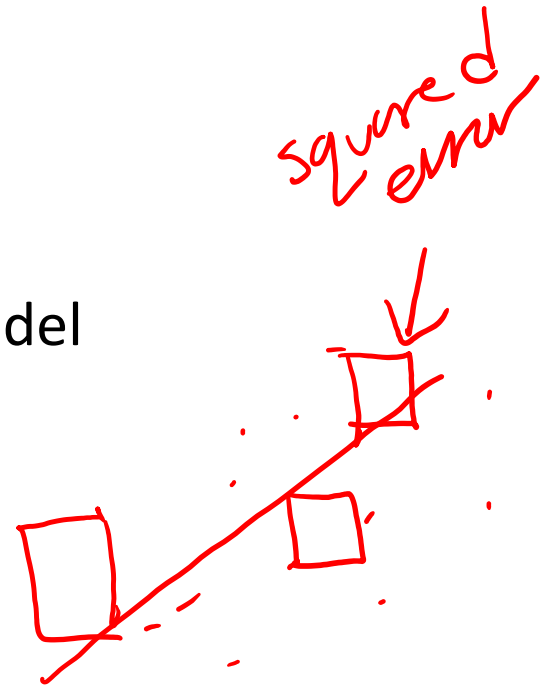
$$\hat{\epsilon} = Y - \hat{Y}$$

→ sampled data

predicted values

SSR

- When we use linear regression (OLS), the statistical model chooses estimates that minimize the sum of squared residuals (which are errors)
- Ordinary Least Squares – minimizing squared errors



$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

minimize error

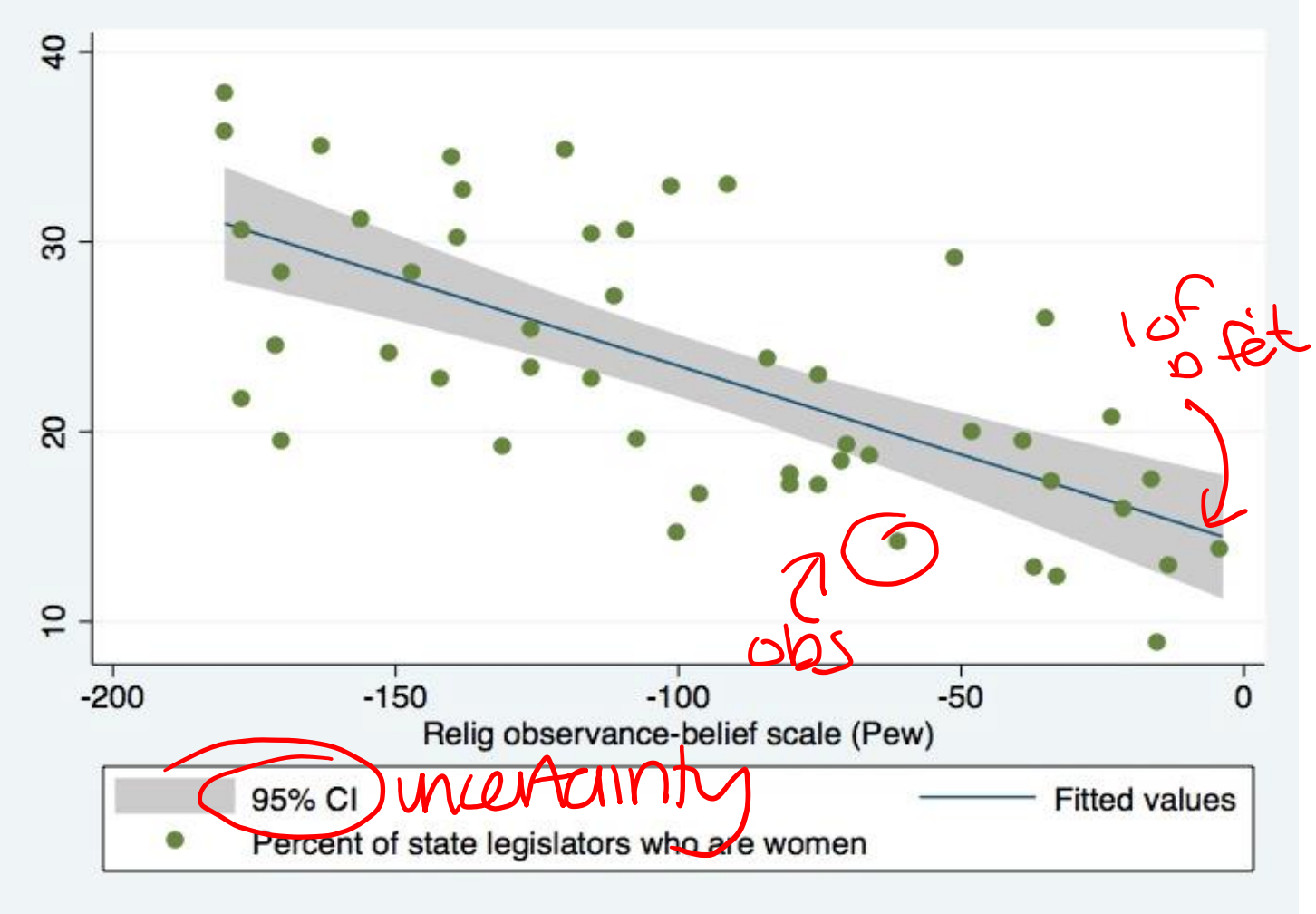
Let's see an Example

religiosity ↑
female leg ↓

- State Religiosity & Women in State Legislatures
- Hypothesis: states with higher religiosity scores are less likely to have female legislators
- Independent variable?
- State religiosity (IV)
- Dependent variable?
- Women in state legislatures (DV)

How would we write this model?

- IV: State religiosity
  - This is X
- DV: Female legislators
  - This is Y
- $Y = \alpha + \beta X + \varepsilon$
- Female legislators = $\alpha + \beta$ (state religiosity) + ε
- β will tell us the effect that state religiosity has on the presence of female legislators



State religiosity (measure from Pew) is the independent variable

Percentage of state legislators who are women is the independent variable

Dots are the observations

Blue line is the best fit line, the line that minimizes squared errors

Regression Table:
State Religiosity & Women in State Leg.

- This is what an output from a regression looks like
- The intercept is α (the effect of state religiosity on female legislators when the value of X (state religiosity) is 0
- "Religiosity" is the X variable you gave R
- X (religiosity) has a negative effect on the presence of female state legislators
  - What does this mean?
- N = 50.
  - What does this mean?
- R2 is a measure of model fit, we will talk about it later

	Model 1
Intercept	14.13*
	(1.64)
Religiosity	-0.09*
	(0.01)
N	*50*
R2	*0.45*

14.13 − 0.09(1) =

How to Interpret OLS Regression Coefficients

- Identify the unit of analysis (ex: States in the US)
- Identify the independent variable and its unit (ex: Religiosity Scale (-175, 0))
- Describe a one-unit increase in the independent variable in everyday language
  - Ex: One point higher on the religiosity scale
- Identify the dependent variable and its units
  - Ex: % of women in the state legislature
- Interpret the regression coefficient (slope) as the average change in the dependent variable given a one-unit increase in the independent variable
  - Ex: Increasing the religiosity index by 1 point decreases the % of women in the state legislature by .09% **on average**.

	Model 1
Intercept	14.13*
	(1.64)
Religiosity	-0.09*
	(0.01)
N	*50*
R2	*0.45*

Writing out your model

- Now that you've run your model, you can write it out in more detail
- You know alpha, the intercept, and you know beta, the slope
- You also already knew X and Y (your independent and dependent variables)
- So – you can translate that into a more specific linear model

Predicting Values of Y:
State Religiosity & Women in State Leg.

	Model 1
Intercept	14.13*
	(1.64)
Religiosity	-0.09*
	(0.01)
N	*50*
R2	*0.45*

% women in legislature = 14.13 + (- 0.09*Religiosity) + error

Predicting Values of Y: State Religiosity & Women in State Leg.

$y = \alpha + \beta x + \varepsilon$

	Model 1
Intercept	14.13*
	(1.64)
Religiosity	-0.09*
	(0.01)
N	*50*
R2	*0.45*

% women in legislature = 14.13 - 0.09*Religiosity (+ error)

- You can put values for X in to find out what the predicted value of Y, your outcome, would be at that value of X
  - *Of course it will not necessarily be equal to the observed value because of **randomness**!*
- What is the expected % of women in a state's legislature if the level of religiosity in that state is -50?

% women = 14.13 – (0.09*-50) = 18.63%

Practice