Regression in 10 Minutes

Rob Williams

Washington University in St. Louis

January 27, 2020

Social science?

- What is the difference between historical and social scientific understandings of civil war?

Social science?

- What is the difference between historical and social scientific understandings of civil war?
  - History: how can be best understand a single specific civil war?

Social science?

- What is the difference between historical and social scientific understandings of civil war?
  - History: how can be best understand a single specific civil war?
  - Social science: what are general patterns of cause and effect in civil war?

Social science?

- What is the difference between historical and social scientific understandings of civil war?
  - History: how can be best understand a single specific civil war?
  - Social science: what are general patterns of cause and effect in civil war?
- The social scientific perspective implies the ability to make *predictions*

How do we make predictions?

- Look at what we know, try to summarize out

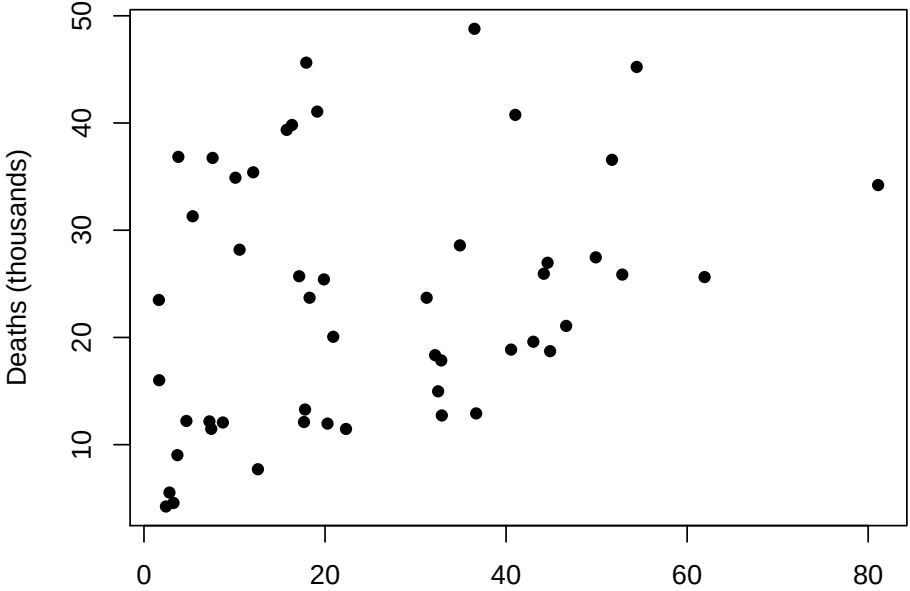
How do we make predictions?

- Look at what we know, try to summarize out
- Create a rule we can use to explain what we see

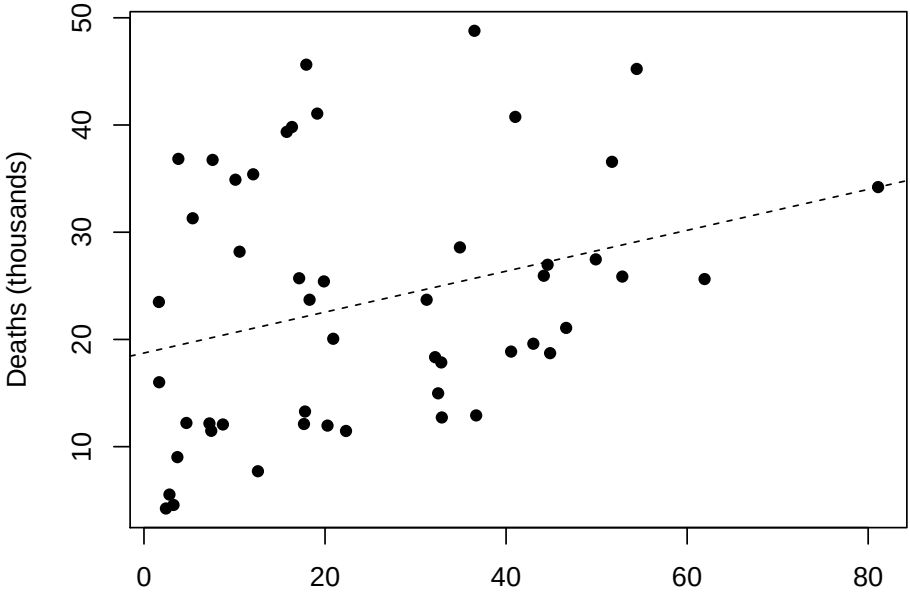
How do we make predictions?

- Look at what we know, try to summarize out
- Create a rule we can use to explain what we see
- Apply that rule to new information

Get the data



Line it up



What's in a line?

- If you think back to middle school geometry

$$y = mx + b$$

What's in a line?

- If you think back to middle school geometry

$$y = mx + b$$

- A line is defined by its *slope* and *intercept*

What's in a line?

- If you think back to middle school geometry

$$y = mx + b$$

- A line is defined by its *slope* and *intercept*
  - Slope: change in x associated with a one unit change in y

What's in a line?

- If you think back to middle school geometry

$$y = mx + b$$

- A line is defined by its *slope* and *intercept*
  - Slope: change in x associated with a one unit change in y
    - Rise over run

What's in a line?

- If you think back to middle school geometry

$$y = mx + b$$

- A line is defined by its *slope* and *intercept*
  - Slope: change in x associated with a one unit change in y
    - Rise over run
  - Intercept: where does the line intersect the y axis

Regression

- Linear regression, OLS (ordinary least squares), the linear model

Regression

- Linear regression, OLS (ordinary least squares), the linear model
- Best thought of (in two dimensions) as fitting a line to a cloud of data

Regression

- Linear regression, OLS (ordinary least squares), the linear model
- Best thought of (in two dimensions) as fitting a line to a cloud of data
- Equation:

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

Breaking it down

- Y: dependent/outcome/response variable

Breaking it down

- Y: dependent/outcome/response variable
- X: independent/explanatory/predictor variable

Breaking it down

- Y: dependent/outcome/response variable
- X: independent/explanatory/predictor variable
- α: intercept coefficient

Breaking it down

- Y: dependent/outcome/response variable
- X: independent/explanatory/predictor variable
- α: intercept coefficient
- β: slope coefficient

Breaking it down

- Y: dependent/outcome/response variable
- X: independent/explanatory/predictor variable
- α: intercept coefficient
- β: slope coefficient
- ϵ: unobserved error

Breaking it down

- Y: dependent/outcome/response variable
- X: independent/explanatory/predictor variable
- α: intercept coefficient
- β: slope coefficient
- ϵ: unobserved error
- $\alpha + \beta X$: mean of Y given the value of X

Breaking it down

- Y: dependent/outcome/response variable
- X: independent/explanatory/predictor variable
- α: intercept coefficient
- β: slope coefficient
- ϵ: unobserved error
- $\alpha + \beta X$: mean of Y given the value of X
- α: mean of Y when X is zero

Breaking it down

- Y: dependent/outcome/response variable
- X: independent/explanatory/predictor variable
- α: intercept coefficient
- β: slope coefficient
- ϵ: unobserved error
- $\alpha + \beta X$: mean of Y given the value of X
- α: mean of Y when X is zero
- β: increase in Y associated with a one unit increase in X

Breaking it down: α

- α is the **intercept**

Breaking it down: α

- α is the **intercept**
- We can think of it as where the line intersects the y-axis

Breaking it down: α

- α is the **intercept**
- We can think of it as where the line intersects the y-axis
- It is also the value of Y when $X = 0$

Breaking it down: α

- α is the **intercept**
- We can think of it as where the line intersects the y-axis
- It is also the value of Y when $X = 0$
  - This doesn't happen in every set of data

Breaking it down: α

- α is the **intercept**
- We can think of it as where the line intersects the y-axis
- It is also the value of Y when $X = 0$
  - This doesn't happen in every set of data
  - This also doesn't necessarily make sense in every set of data

Breaking it down: α

- α is the **intercept**
- We can think of it as where the line intersects the y-axis
- It is also the value of Y when $X = 0$
  - This doesn't happen in every set of data
  - This also doesn't necessarily make sense in every set of data
    - No one can be zero inches tall

Breaking it down: α

- α is the **intercept**
- We can think of it as where the line intersects the y-axis
- It is also the value of Y when $X = 0$
  - This doesn't happen in every set of data
  - This also doesn't necessarily make sense in every set of data
    - No one can be zero inches tall
    - No country can have zero population

Breaking it down: β

- β is the **slope**

Breaking it down: β

- β is the **slope**
- It is the *average* increase in Y when X increases by one unit

Breaking it down: β

- β is the **slope**
- It is the *average* increase in Y when X increases by one unit
  - Moving from 11 to 12 years of education is associated with a 2 point decrease in support for the death penalty (on a 100 point scale)

Breaking it down: β

- β is the **slope**
- It is the *average* increase in Y when X increases by one unit
  - Moving from 11 to 12 years of education is associated with a 2 point decrease in support for the death penalty (on a 100 point scale)
- What a one unit increase in X means is determined how you measure X

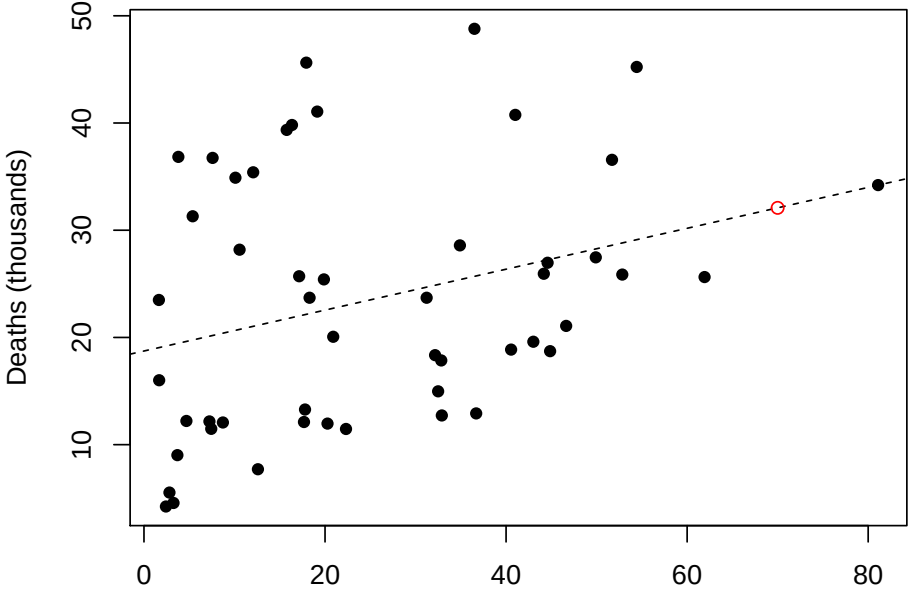
Breaking it down: β

- β is the **slope**
- It is the *average* increase in Y when X increases by one unit
  - Moving from 11 to 12 years of education is associated with a 2 point decrease in support for the death penalty (on a 100 point scale)
- What a one unit increase in X means is determined how you measure X
  - β is a function of your X e.g. it will be different from thousands of USD vs. millions of USD

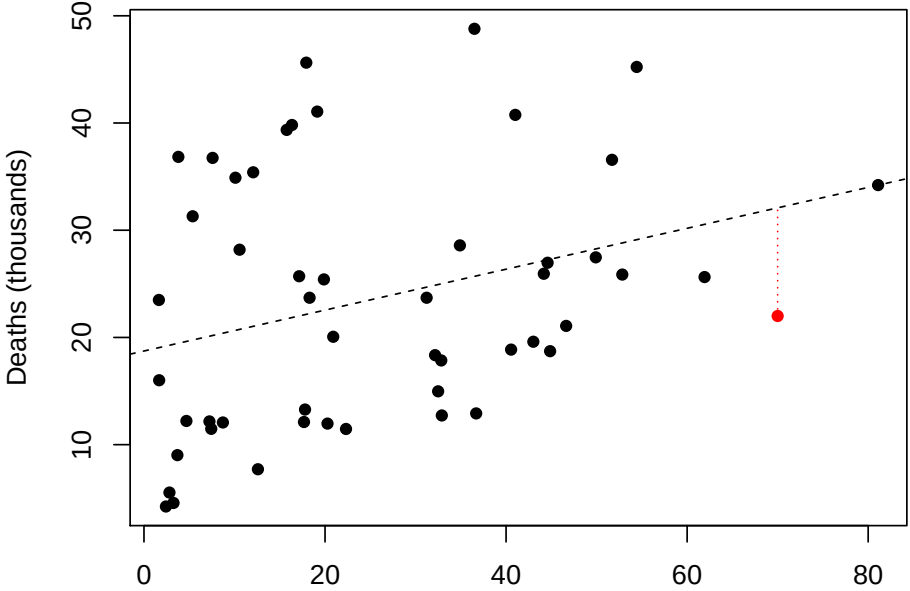
Breaking it down: β

- β is the **slope**
- It is the *average* increase in Y when X increases by one unit
  - Moving from 11 to 12 years of education is associated with a 2 point decrease in support for the death penalty (on a 100 point scale)
- What a one unit increase in X means is determined how you measure X
  - β is a function of your X e.g. it will be different from thousands of USD vs. millions of USD
- Taken together, α and β let us make predictions for Y for a given X value

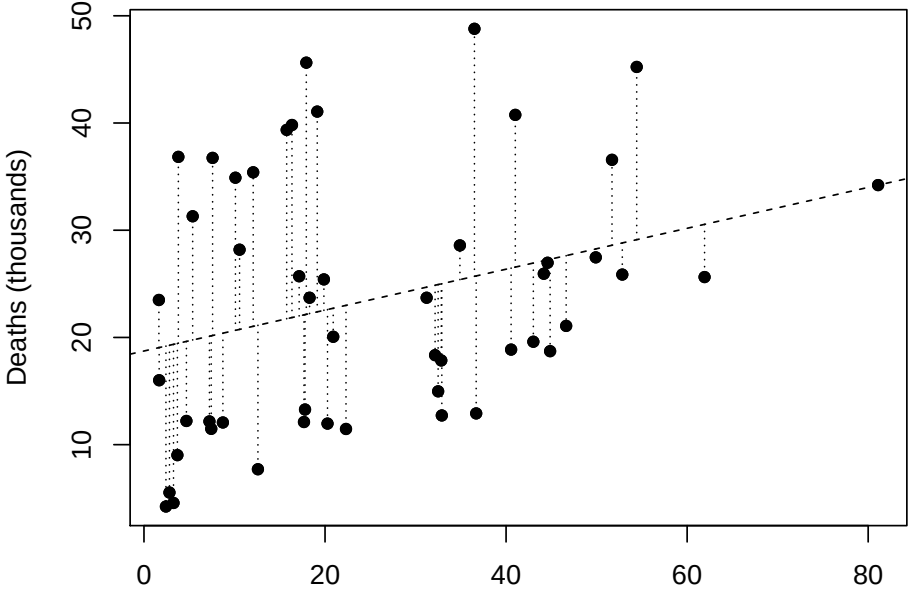
Predicting



Evaluating



Overall error



A picture is worth a thousand words

- You'll often have to read a regression table instead of looking at a plot

A picture is worth a thousand words

- You'll often have to read a regression table instead of looking at a plot

- This is common because we're usually dealing with more than two variables

A picture is worth a thousand words

- You'll often have to read a regression table instead of looking at a plot
- This is common because we're usually dealing with more than two variables
- Let's give it a try!

Get familiar with this

	Model 1
Years	0.19*
	(0.09)
(Intercept)	18.73***
	(2.70)
R^2	0.09
Adj. R^2	0.07
Num. obs.	50
RMSE	11.38

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Seeing stars

- What are those stars above each coefficient?

Seeing stars

- What are those stars above each coefficient?
- They represent the level of *uncertainty*

Seeing stars

- What are those stars above each coefficient?
- They represent the level of *uncertainty*
- The p-value is (very) roughly an estimate of how likely it is that we get the results we did when if is no relationship between x and y

Seeing stars

- What are those stars above each coefficient?
- They represent the level of *uncertainty*
- The p-value is (very) roughly an estimate of how likely it is that we get the results we did when if is no relationship between x and y
- This is a gross oversimplification, and you should really take QPM

Seeing stars

- What are those stars above each coefficient?
- They represent the level of *uncertainty*
- The p-value is (very) roughly an estimate of how likely it is that we get the results we did when if is no relationship between x and y
- This is a gross oversimplification, and you should really take QPM
- Political scientists have decided that a p-value *below* 0.05 is sufficiently safe to think about an actual relationship between x and y

Seeing stars

- What are those stars above each coefficient?
- They represent the level of *uncertainty*
- The p-value is (very) roughly an estimate of how likely it is that we get the results we did when if is no relationship between x and y
- This is a gross oversimplification, and you should really take QPM
- Political scientists have decided that a p-value *below* 0.05 is sufficiently safe to think about an actual relationship between x and y
  - Yes, this is incredibly arbitrary