

Data Analysis in R

Prediction

Ken Stiller

23rd November 2022



oooo

oooooo

oooooooooooooooooooo

o

Syllabus: Data Analysis in R

1. Introduction
2. Causality & Basics of Statistics
3. Sampling & Measurement
4. **Prediction**
5. Multivariate Regression
6. Probability & Uncertainty
7. Hypothesis Testing
8. Assumptions & Limits of OLS
9. Interactions & Non-Linear Effects

Table of Contents

Overview

The Big Picture

Linear Regression: The Basics

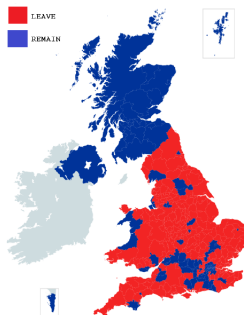
Regression Anatomy

Wrap Up

Predicting Brexit I

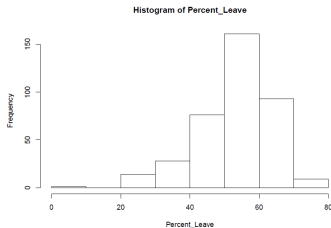
► Context:

- On the 23rd of June 2016 the UK voted to leave the EU
- 51.89% of voters who turned out (72.21%), voted to leave the EU



Predicting Brexit II

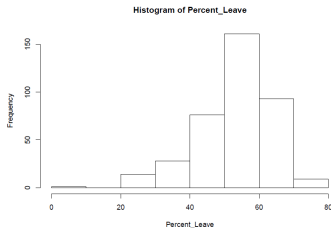
- ▶ We want to predict the % **Vote Leave** in a given constituency.
- ▶ We want to **minimise errors** (in stats generally) and be **parsimonious** (e.g. don't add variables unless we learn from them)
- ▶ All we know is the actual Leave Vote in each constituency.
 - ▶ We want to take one at random, predict its % **Vote leave**. How?



- ▶ Now let's say that we know one other variable that we can use as a predictor: *mean age in constituency*.
 - ▶ Does this help us? What would you do with it?

Predicting Brexit II

- ▶ We want to predict the % **Vote Leave** in a given constituency.
- ▶ We want to **minimise errors** (in stats generally) and be **parsimonious** (e.g. don't add variables unless we learn from them)
- ▶ All we know is the actual Leave Vote in each constituency.
 - ▶ We want to take one at random, predict its % **Vote leave**. How?

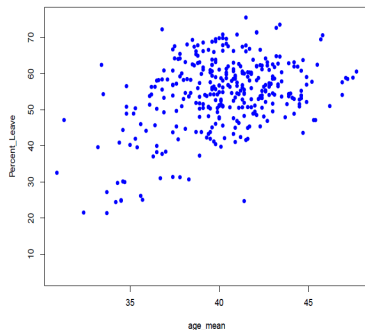


- ▶ Now let's say that we know one other variable that we can use as a predictor: *mean age in constituency*.
 - ▶ Does this help us? What would you do with it?

The Big Picture

- ▶ Our goal **always** is to make predictions about a *population*, minimising errors
- ▶ We can add increasingly more information about the data to help predict an outcome
- ▶ You might think of tools to do this as a continuum:
 1. Descriptive statistics: we draw upon one variable only
 2. **Bivariate regression**: we draw upon two variables
 3. Multivariate regression: we draw upon more than two variables

Linear Regression: Motivation



- ▶ How would you summarize the relationship between X and Y ?
- ▶ It seems that counties with older population also have a higher leave vote
- ▶ **What we are interested in today: By how much?**
(Note that correlation can't tell!)

Linear Regression: Intuition

What we often want to summarise is **conditional expectation** of a variable (Y) dependent on another variable (X)

- ▶ This is written as $E[Y|X]$, where E stands for *expectation*
- ▶ $E[Y]$ is the *population mean* and known as expectation only
- ▶ $E[Y|X = x]$ is the expectation of Y given a value of x

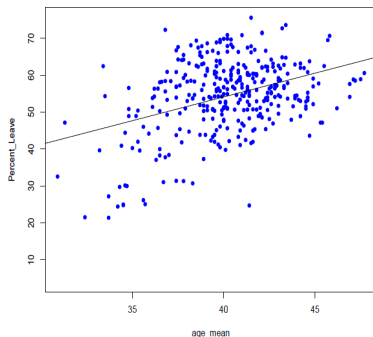
Regression allows us to provide overall estimates about how Y changes with X - without having to rely on specific values of X .

- ▶ Linear regression assumes Y varies in X in the same way through the range of values of X
- ▶ This allows us to predict the value of Y for each value of X
- ▶ It's a simple linear form of the conditional expectation function:

$$E[Y|X] = \beta_0 + \beta_1 X$$

Linear Regression: Equation

$$E[Y|X] = \beta_0 + \beta_1 X$$



- What does β_0 stand for?

$$E[Y|X] = 0$$

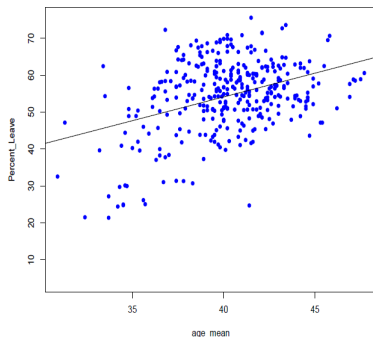
- And β_1 ?

$$E[Y|X = x] - E[Y|X = x - 1]$$

- Does the value of X matter?
No - there is a single, uniform slope.

Linear Regression: Equation

$$E[Y|X] = \beta_0 + \beta_1 X$$



- What does β_0 stand for?

$$E[Y|X] = 0$$

- And β_1 ?

$$E[Y|X = x] - E[Y|X = x - 1]$$

- Does the value of X matter?
No - there is a single, uniform slope.

Linear Regression: Notation

So, a linear regression model is a *linear* approximation of the relationship between explanatory variables X and a dependent variable Y

$$E[Y|X] = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} X$$

- ▶ Y : Dependent variable (outcome)
- ▶ X : Explanatory/independent variable
- ▶ β_0 : Intercept (or constant)
- ▶ β_1 : Slope coefficient (**association** between X and Y)

Quick interpretation:

- ▶ $\beta_0 + \beta_1 X$: Conditional mean of Y given a value of X
- ▶ β_0 : Value of Y when $X = 0$
- ▶ β_1 : Change in Y associated with a **one unit increase** in X across

Linear Regression: Notation II

$$E[Y|X] = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} X$$

- ▶ **Sign:** denotes the direction of the relationship:
 - ▶ $\beta_1 > 0$: Increase in X is associated with an **increase** in values of Y
 - ▶ $\beta_1 < 0$: Increase in X is associated with a **decrease** in values of Y
- ▶ **Magnitude:** β_1 tells us the extent to which Y changes with a **one unit** increase in X

Linear Regression: Prediction v Causality

Prediction

Regression allows us to *predict* the value of Y for any value of X , even if the specific x is not included in the sample

$$E[Y \mid X = x_{any}] = \hat{\beta}_0 + \hat{\beta}_1 \times x_{any}$$

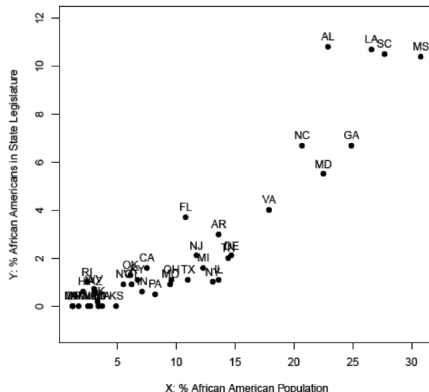
In a predictive model, $\hat{\beta}_1$ is interpreted as the expected difference in Y when there is **one unit increase** in X

Causality

Prediction \neq causality! Predicting Y on the basis of X does not imply that it is the change in X that **causes** Y !

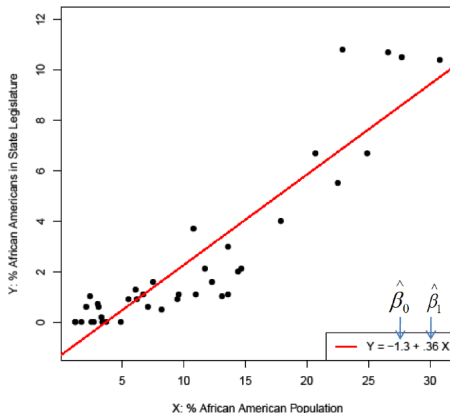
- ▶ Causality implies that we make sure other factors (**confounders**) that can cause a change in both X and Y are being accounted for
- ▶ Regression is helpful only because it provides a framework to account for other **observed** confounders
- ▶ It's a means to end - not more, not less

A Simple Bivariate Relationship



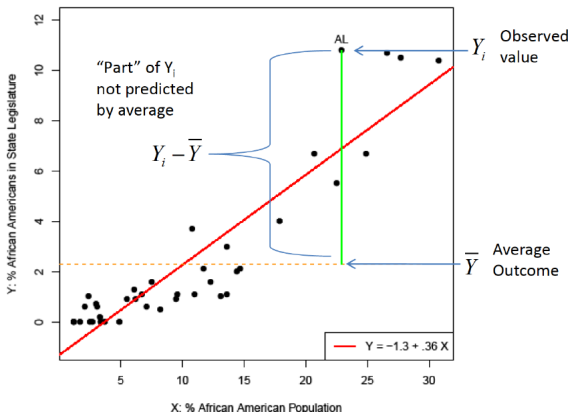
- ▶ How do we choose which line to fit to the data?
- ▶ In principle, infinite number of lines available
- ▶ Recall our aims as social scientists

Fit the Line



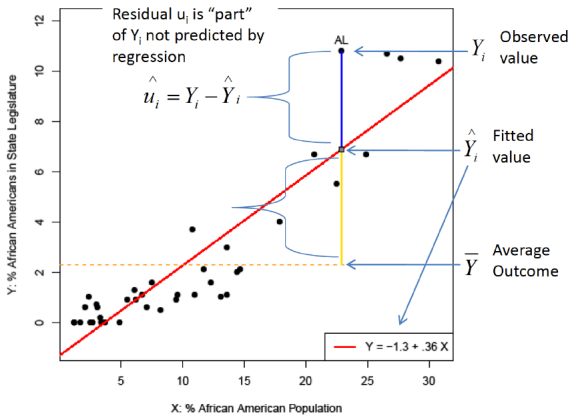
Why choose this line though?

Compare with Benchmark



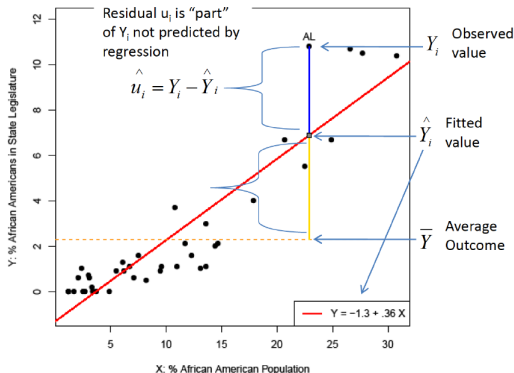
Our benchmark is the veil of ignorance: predicting Y without knowledge of x .

Compare with Benchmark II



Is this helpful? Let's decompose the distance between Y_i and \bar{Y} to find out.

Decompose Distance between Y_i and \bar{Y}



From \bar{Y} to \hat{Y}_i : Improvement in prediction from veil of ignorance:

Predicted \hat{Y}

From \hat{Y}_i to Y_i : Remaining mistakes we make in prediction: Residual

Linear Regression Model

► Model:

$$Y = \underbrace{\alpha}_{\text{Intercept}} + \underbrace{\beta}_{\text{Slope}} X + \underbrace{\epsilon}_{\text{Error Term}}$$

- (α, β) : coefficients (parameters of the model)
- ϵ : unobserved error/disturbance term (mean zero)

► Fitted model:

$\hat{Y} = \hat{\alpha} + \hat{\beta}x$: predicted/fitted values

$\hat{\mu} = Y - \hat{Y}$: residuals

- $(\hat{\alpha}, \hat{\beta})$: estimated coefficients

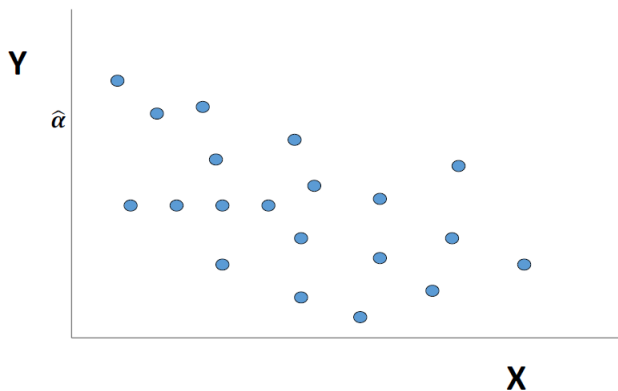
Ordinary Least Squares: OLS

- ▶ Estimating the model parameters from the data, we usually obtain these estimates via the *least squares method*
- ▶ Minimise the **sum of squared residuals (SSR)**:

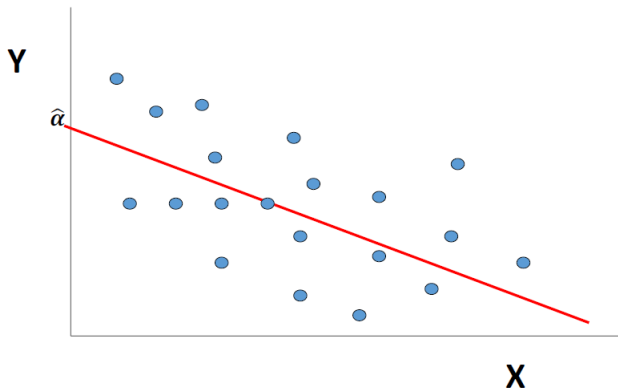
$$\text{SSR} = \sum_{i=1}^n \left(Y_i - \hat{Y} \right)^2 = \sum_{i=1}^n (\hat{\mu}_i)^2 = \sum_{i=1}^n \left(Y_i - \hat{\alpha} - \hat{\beta} X_i \right)^2$$

- ▶ There is only one line that satisfies this criteria: **Ordinary Least Squares (OLS)** regression. OLS estimates β_0 and β_1 in so that **SSR** are being minimised
- ▶ Simply speaking, OLS estimates a β_1 that on average minimises the (squared) errors between the dots and the line.

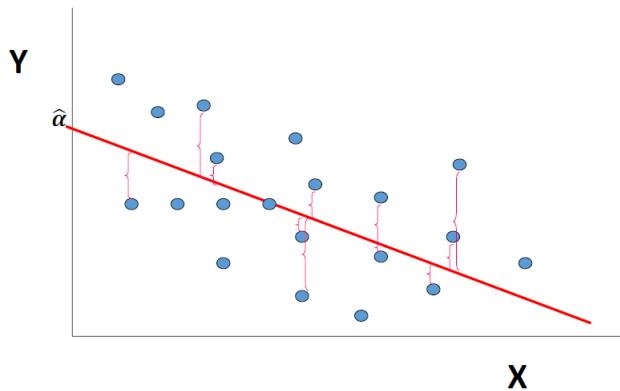
OLS II



OLS II



OLS II



OLS III

- In OLS mean of residuals is always zero (only one possible line satisfies the condition):

$$\text{mean of } \hat{\mu} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}X_i) = \bar{Y} - \hat{\alpha} - \hat{\beta}\bar{X} = 0$$

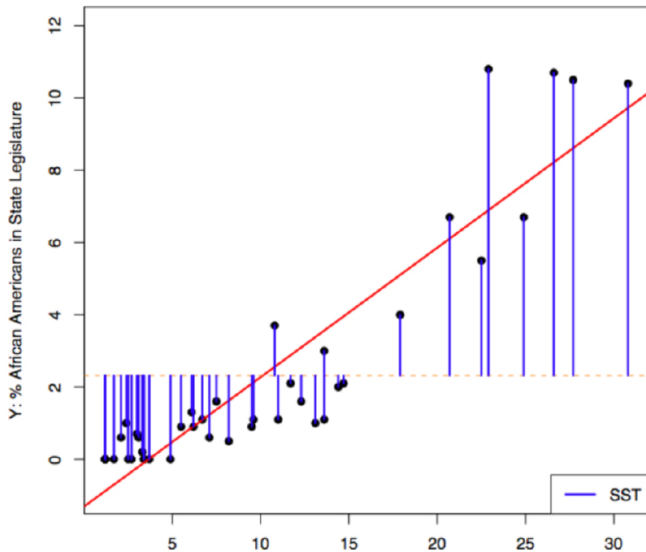
- How do we compute the OLS estimators? The slope....

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{COV}(X, Y)}{\text{VAR}(X)}$$

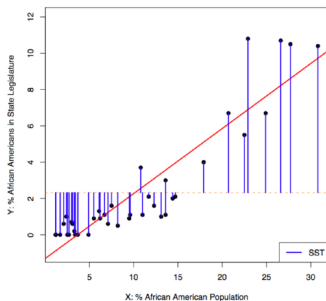
- Notice that, unlike with correlation, order matters
- ...and the intercept:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

Goodness of Fit



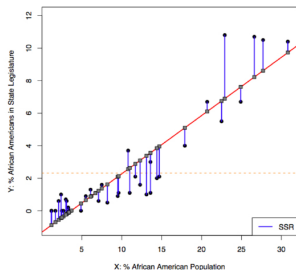
Goodness of Fit



$$\sum_{i=1}^n (y_i - \bar{y})^2 = \text{TSS}$$

- ▶ This represents the sum of squares in the *null model* - i.e., when we don't know anything but values of the dependent variable
- ▶ You can think about it as a measure of how "off" your prediction is from the real data, if all you rely on is the average

Goodness of Fit II



$$\sum_{i=1}^n (y_i - \hat{y})^2 = RSS$$

- ▶ This represents the sum of squares in the *model* i.e., when we do know something beyond values of the dependent variable
- ▶ You can think about it as a measure of how "off" your prediction is from the real data, if you rely on an *independent variable* to explain variation in Y

Model Fit - Goodness of Fit

- ▶ How well does our model perform? Do we learn anything from adding the independent variable (vis-a-vis the null model)?
- ▶ The R-squared gives us a measure of this

$$\text{TSS} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- ▶ R^2 represents the proportion of total variation in the outcome variable explained by the predictor(s) included in the model
- ▶ R^2 is bounded between 0 and 1
- ▶ Do we care?

Guidelines

- ▶ R^2 denotes the goodness of fit, but not relevance of the variable in explaining the outcome
- ▶ We are interested in β_1 , **statistical significance**, **slope** and its **magnitude**

$$\%VoteLeave|MeanAge = \underbrace{2.61}_{\text{intercept}} + \underbrace{1.28}_{\text{slope}} \text{MeanAge}$$

- ▶ **Magnitude:** Can you interpret what the slope means here?

Take Away

Always interpret the magnitude of the findings. A finding may be significant but too small for us to care; or vice-versa.

Wrap Up

- ▶ Key points from today:
 - ▶ Bivariate regression can help explain variation in an outcome
 - ▶ We learnt what the intercept, slope and SSR are.
 - ▶ Be cautious when discussing R^2 - it *can* tell us something about our model but isn't as relevant as some make it
 - ▶ Regression is a technique that is a *tool* in conducting science - always be aware of what you are doing/estimating
- ▶ Next time we'll be talking about:
 - ▶ Confounding - Omitted variable bias
 - ▶ Multiple Regression analysis and the ceteris paribus principle
 - ▶ Multivariate regression: OLS with several predictors