Linear Regression

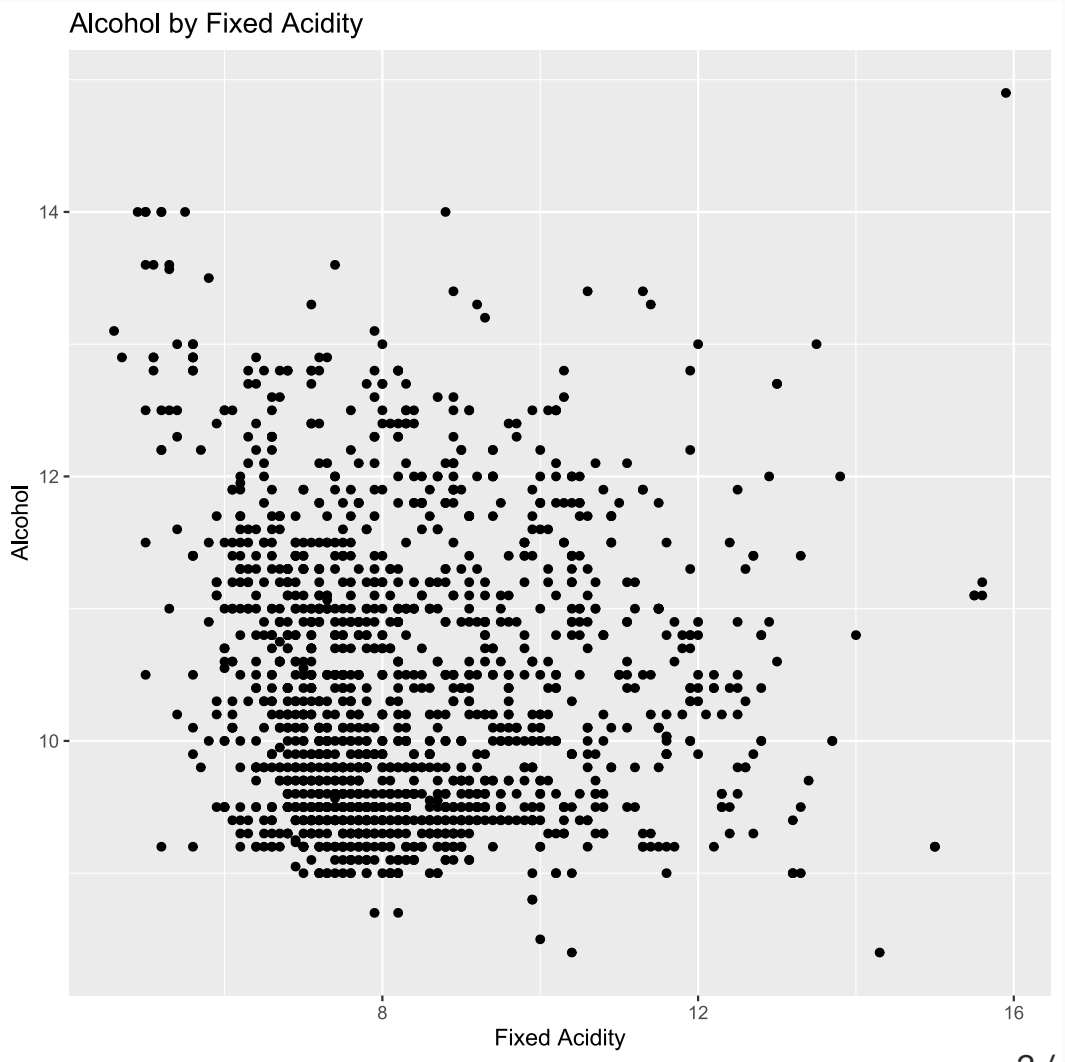
Jeroen Mahieu
jeroen.mahieu@vu.nl
Updated 2022-03-14

Today - Linear Regression and OLS

- I assume you know the theory behind linear regression and ***Ordinary Least Squares (OLS)***
- Here, we will focus on estimating and reporting linear models using OLS

The Relationship between Fixed Acidity and Alcohol

```
ggplot(data=wine)+
  geom_point(mapping=aes(x = `fixed acidity`,
                         y = alcohol)) +
  labs(title = "Alcohol by Fixed Acidity",
       x = "Fixed Acidity",
       y = "Alcohol")
```

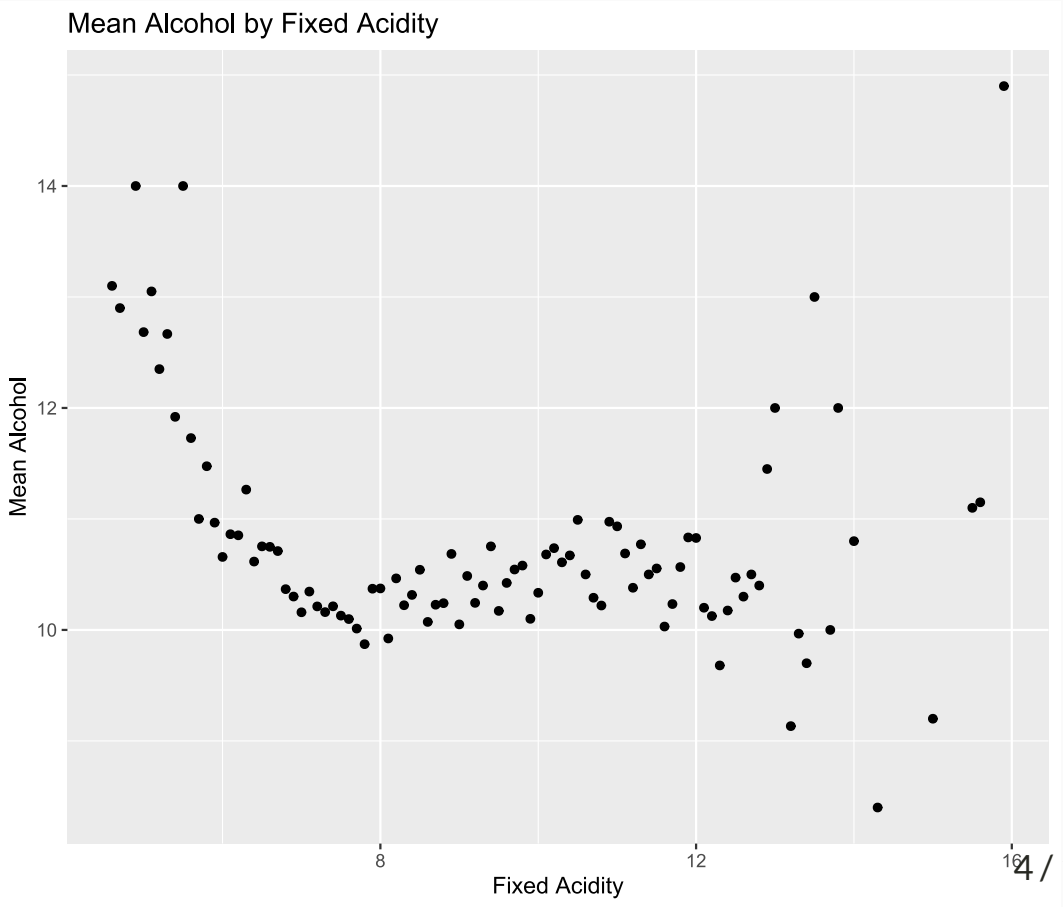


The Relationship between Fixed Acidity and Alcohol

- Seems like these two variables are negatively related
- Let's check by computing the mean alcohol level for each value of fixed acidity

```
acidity_avg_alcohol <- wine %>%
  group_by(`fixed acidity`) %>%
  summarise(avg_alcohol = mean(alcohol))

ggplot(data=acidity_avg_alcohol)+
  geom_point(mapping=aes(x = `fixed acidity`,
                         y = avg_alcohol)) +
  labs(title = 'Mean Alcohol by Fixed Acidity',
       x = 'Fixed Acidity',
       y = 'Mean Alcohol')
```



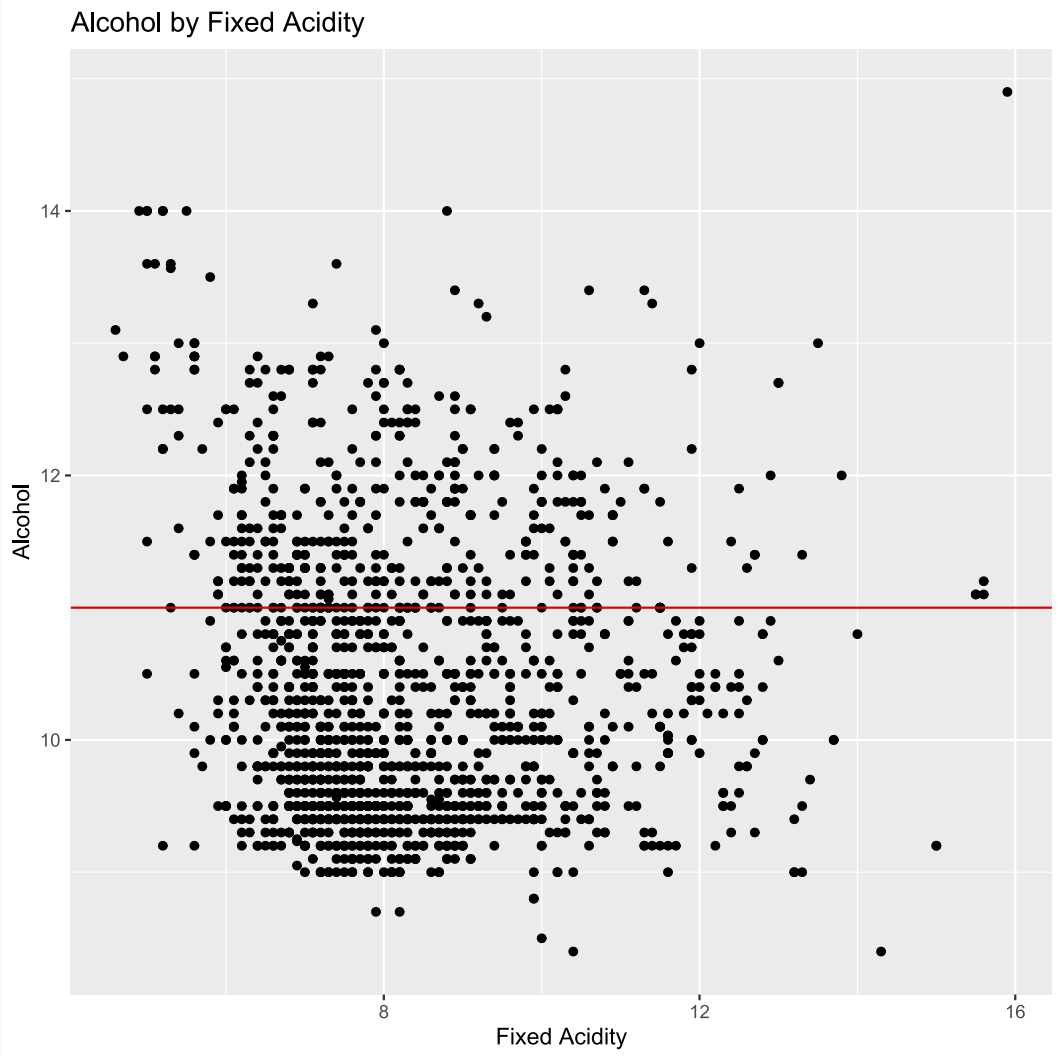
Fixed Acidity and Alcohol: Regression Line

- Instead of showing the mean of alcohol for all levels of fixed acidity, we can assume their underlying relationship is represented by a *shape*
- Enter *regression* aka *line-fitting*
- In basic forms of regression, this will be a *straight* line
- The equation for such a line with an intercept b_0 and a slope b_1 is:

$$\hat{y}_i = b_0 + b_1 x_i$$

  - an **outcome variable** (also called **dependent variable**):
    average alcohol (y)
  - an **explanatory variable** (also called **independent variable** or **regressor**):
    fixed acidity (x)

Fixed Acidity and Alcohol: Regression Line



- A *line*! Great. But **which** line? This one?
- That's a *flat* line. But average alcohol is somewhat *decreasing* with fixed acidity 😩
- We need a rule to decide!

Ordinary Least Squares (OLS)

- OLS gives us the line that minimizes *the sum of the squared residuals*
- The residual is the 'error' or *vertical distance* between your line and the actual observation

Intercept

-4 0.5 4

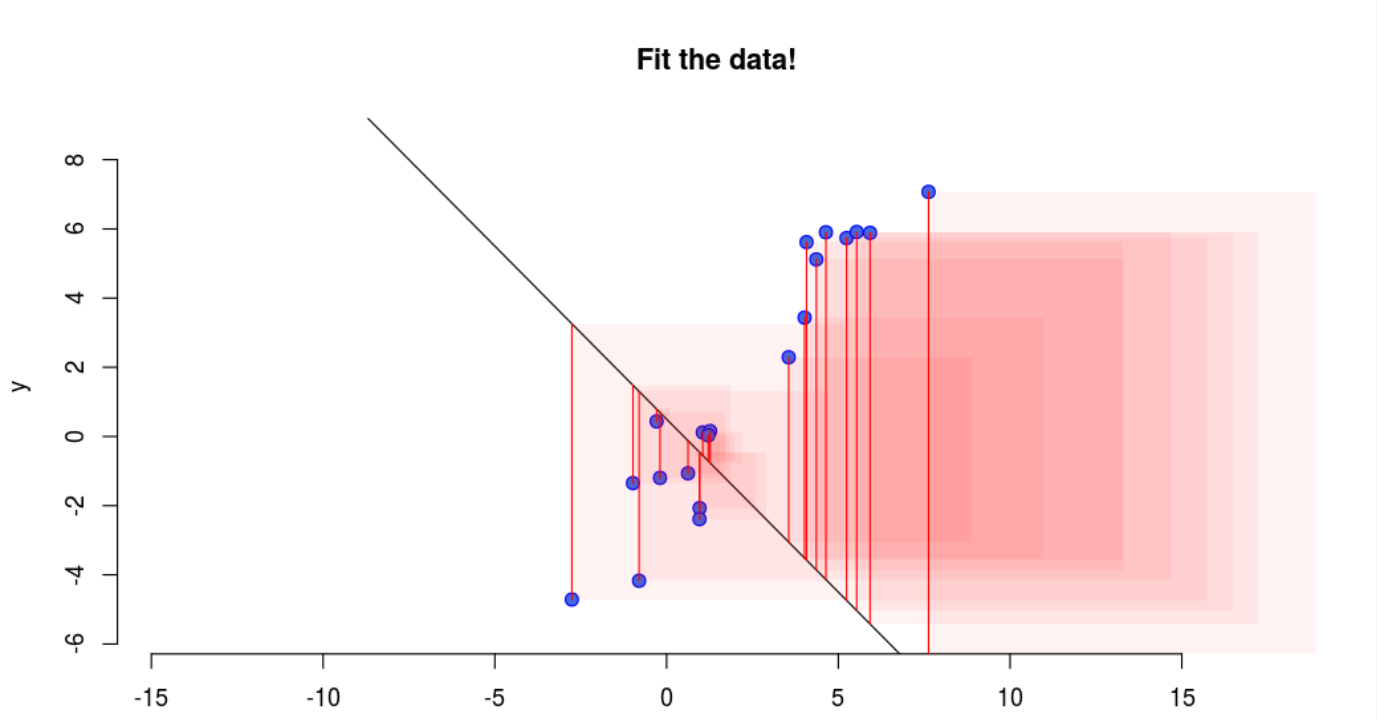
-4 -3.2 -2.4 -1.6 -0.8 0 0.8 1.6 2.4 3.2 4

Slope

-2 -1 2

-2 -1.6 -1.2 -0.8 -0.4 0 0.4 0.8 1.2 1.6 2

Your guess: y = 0.5 + -1x



Ordinary Least Squares (OLS): Interpretation

For now assume both the dependent variable (y) and the independent variable (x) are numeric.

> Intercept (b_0): **The predicted value of y $(\hat{y})$ if $x = 0$.**
>
> Slope (b_1): **The predicted change, on average, in the value of y *associated* to a one-unit increase in x.**

- ⚠️ Note that we use the term *associated*, **clearly avoiding interpreting b_1 as the causal impact of x on y**. To make such a claim, we need some specific conditions to be met.
- Also notice that the units of x will matter for the interpretation (and magnitude!) of b_1.
- **You need to be explicit about what the unit of x is!**

OLS with R

- In `R`, OLS regressions are estimated using the `lm` function.
- This is how it works:

```
lm(formula = dependent variable ~  independent variable, data = data.frame containing the data)
```

Alcohol and Fixed Acidity

Let's estimate the following model by OLS: $\text{alcohol}_i = b_0 + b_1 \text{fixed acidity}_i + e_i$

```
# OLS regression of alcohol on fixed acidity
lm(alcohol ~ `fixed acidity`, wine)
```

```
##
## Call:
## lm(formula = alcohol ~ `fixed acidity`, data = wine)
##
## Coefficients:
##     (Intercept)  `fixed acidity`
##         10.73701          -0.03775
```

Ordinary Least Squares (OLS): Prediction

```
##
## Call:
## lm(formula = alcohol ~ `fixed acidity`, data = wine)
##
## Coefficients:
##     (Intercept)  `fixed acidity`
##        10.73701         -0.03775
```

This implies (abstracting the $_i$ subscript for simplicity):

$$\begin{aligned}
\hat{y} &= b_0 + b_1 x \\
\widehat{\text{alcohol}} &= b_0 + b_1 \cdot \text{fixed acidity} \\
\widehat{\text{alcohol}} &= 10.73701 + (-0.03775) \cdot \text{fixed acidity}
\end{aligned}$$

What's the predicted level of alcohol for a wine with an acidity of 10? (Using the *exact* coefficients.)

$$\begin{aligned}
\widehat{\text{alcohol}} &= 10.73701 + (-0.03775) \cdot 10 \\
\widehat{\text{alcohol}} &= 10.35951
\end{aligned}$$

Task 1: Simple OLS Regression

05:00

1. Regress `pH` on `residual sugar` (so `pH` is the *dependent* variable)

2. Using the exact coefficients, calculate the predicted pH value when residual sugar = 2

Exporting Regression Output with Stargazer

- We can again use `stargazer` to create publication-ready regression tables

```
# OLS regression of alcohol on fixed acidity
reg <- lm(alcohol ~ `fixed acidity`, wine)
stargazer(reg,
      out = "regression.html",
      type = "html")
```

	Dependent variable:
	alcohol
`fixed acidity`	-0.038**
	(0.015)
Constant	10.737***
	(0.130)
Observations	1,599
R^2	0.004
Adjusted R^2	0.003
Residual Std. Error	1.064 (df = 1597)
F Statistic	6.097** (df = 1; 1597)
Note:	*p<0.1; **p<0.05; ***p<0.01

Task 2: Exporting Simple Regression Output

1. Regress `pH` on `residual sugar` (so `pH` is the *dependent* variable) and assign the output to an object with a name you prefer. Export the output to an html file using the `stargazer` function

05:00

Multiple Regression

- The prior examples only had one predictor. However, you can easily add more to your model.
- This may be necessary when you need to control for confounding factors
- Or when you want to test for interaction (moderating) effects, including polynomials

Alcohol and Fixed Acidity: Controlling for Density

Let's estimate the following model by OLS:

$$\text{alcohol}_i = b_0 + b_1 \text{fixed acidity}_i + b_2 \text{density}_i + e_i$$

```
# OLS regression of alcohol on fixed acidity
lm(alcohol ~ `fixed acidity` + density, wine)
```

```
##
## Call:
## lm(formula = alcohol ~ `fixed acidity` + density, data = wine)
##
## Coefficients:
##     (Intercept)  `fixed acidity`          density
##        470.3950           0.2982        -463.9627
```

Alcohol and Fixed Acidity: Interacting with Density

Let's estimate the following model by OLS:

$$\text{alcohol}_i = b_0 + b_1 \text{fixed acidity}_i + b_2 \text{density}_i + b_3 \text{fixed acidity} \cdot \text{density}_i + e_i$$

```
# OLS regression of alcohol on fixed acidity
lm(alcohol ~ `fixed acidity`*density, wine)
```

```
##
## Call:
## lm(formula = alcohol ~ `fixed acidity` * density, data = wine)
##
## Coefficients:
##             (Intercept)          `fixed acidity`                  density
##                  907.49                   -53.56                  -902.07
## `fixed acidity`:density
##                   53.97
```

Task 3: Multiple Regression

1. Regress `quality` on `alcohol`, `sulphates`, their interaction, and control for `chlorides`
2. Export the output using `stargazer`

05:00