Session 7: Introduction to Regression

Kostis Christodoulou
London Business School

Contents

- Modelling and Correlation
- Simple regression
- Multiple regression
- Colinearity
- Categorical variables

Session overview

- Why understanding relationships is important

- Tools for analysing relationships

- Scatter plots
- Correlation
  - Interpretation
  - Pitfalls

- Regression
  - Building models
  - Interpreting and evaluating models
  - Assessing model validity
  - Using the model for prediction

Modelling

- So far we have concentrated on analysing individual variables
  - confidence intervals / hypothesis tests on means / proportions
  - testing for differences in mean value across samples
- More interesting (and useful) case is to explore relationships between variables
  - build explanatory or predictive **models** and analyse performance
- Applications in all areas of business:
  - targeting customers for mailshot in direct marketing, credit scoring
  - understand factors that drive market share / brand preference
  - forecast sales / demand / market share / investment return……
  - limits are: quality data (improving), computer power (not now!), skills

Why analysing relationships is important

- Development of theory in the social sciences and empirical testing
- Finance e.g.
  - How are stock prices affected by market movements?
  - What is the impact of mergers on stockholder value?
- Marketing e.g.
  - How effective are different types of advertising?
  - Do promotions simply shift sales without affecting overall volume?
- Economics e.g.
  - How do interest rates affect consumer behaviour?
  - How do exchange rates influence imports and exports?

Correlation

- Correlation between ***X*** and ***Y***

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Correlation measures how closely two variables are related and direction: do they move in the same or opposite direction?
- As the value of ***X*** increases, does ***Y*** tend to also increase (positive` relationship) or does it tend to go down (negative relationship)
- It is always between -1 and +1 and

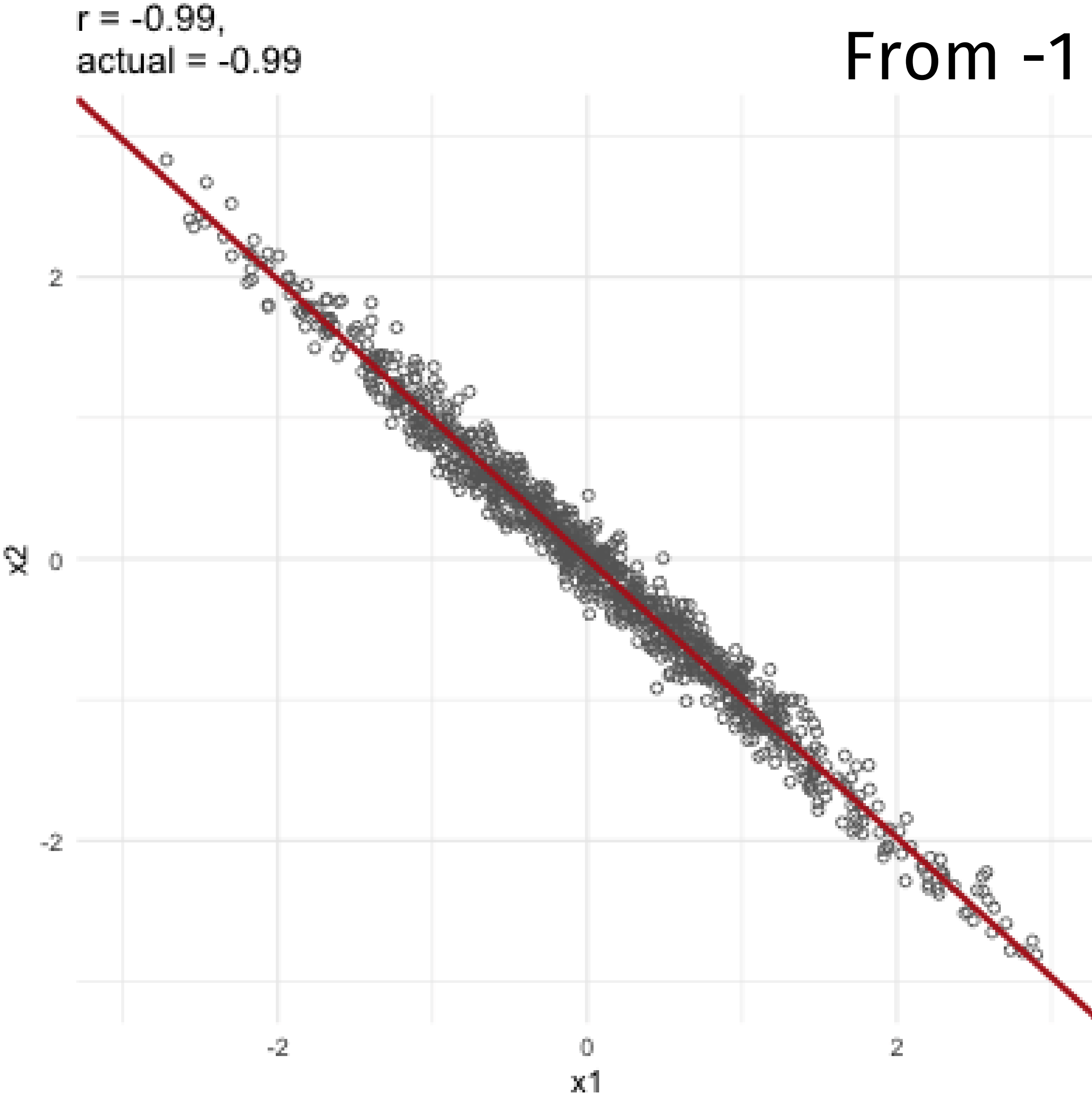
The <u>maximum possible</u> correlation is **+1** (perfect positive correlation) means the two variables move together in the same direction

The <u>minimum possible</u> correlation is **-1** (perfect negative correlation) means that the two variables move in opposite directions

A correlation of zero implies that there is no linear relationship between the variables

Correlation shows direction and magnitude of relationship

From -1 to +1



General Guidelines

0	No relationship
0.01 – 0.19	Little to no relationship
0.20 – 0.29	Weak relationship
0.30 – 0.39	Moderate relationship
0.40 – 0.69	Strong relationship
0.70 – 0.99	Very strong relationship
1	Perfect relationship

Correlation can be positive or negative

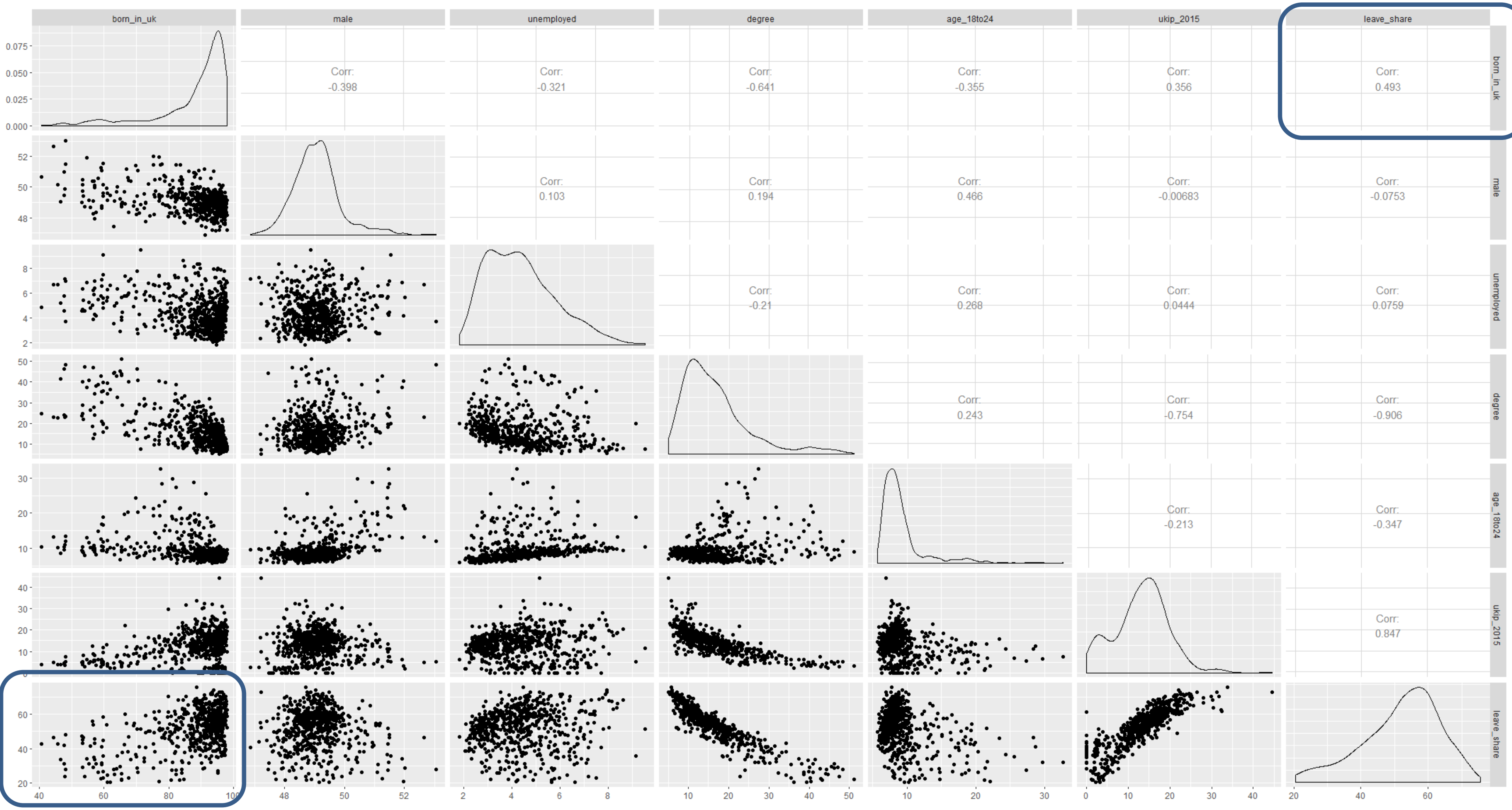
Some basic terminology

Y	~	X (or lots of Xs)
Variable you want to explain or predict		Variable to help you explain changes in Y
Outcome variable		Explanatory variable
Response variable		Predictor variable
Dependent variable		Independent variable
Target variable		Regressor

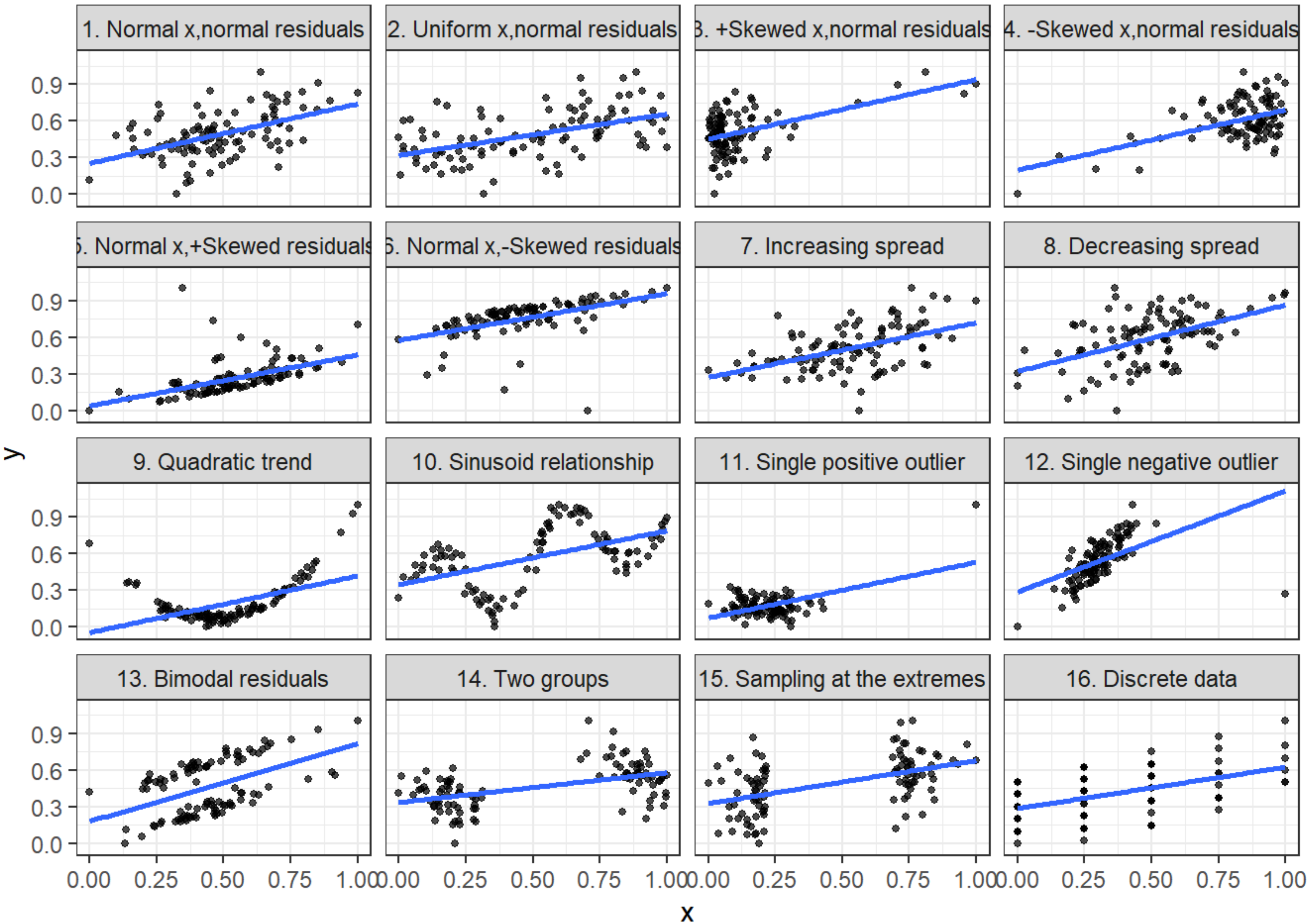
Two main purposes of regression

Prediction	Explanation
Predict the future	Explain effect of X on Y
Focus is on Y	Focus is on X
Netflix trying to guess your next show	Netflix looking at the effect of time of day on show selection
Predicting the price of a used Prius	Look at the effect of mileage on the price of a used Prius
You try to make the best prediction of Y.	Try to explain the effect that specific variables Xs have on Y
Include basically as many variables as you can	Need to have some theoretical reason to include each variable.

Brexit Correlations, using ***GGally::ggpairs()***



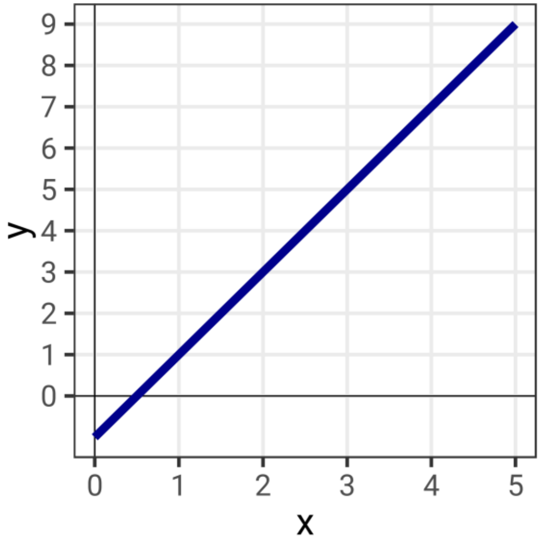
Plot your data– all of these correlations = 0.50



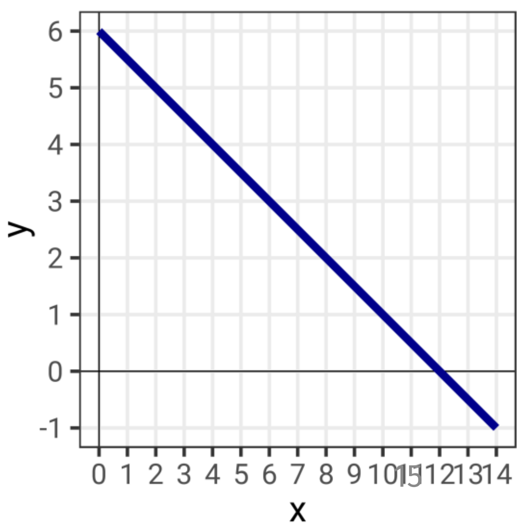
Drawing lines

$y = mx + b$	
y	A number
x	A number
m	Slope, Gradient, Rise/Run
b	y intercept

$$y = 2x - 1$$



$$y = -0.5x + 6$$



Contents

- Modelling and Correlation
- Simple regression
- Multiple regression
- Colinearity
- Categorical variables

The Need to Understand Relationships (1/2)

In 1856, the Reverend John Clay felt that it was time to figure out what factors were playing a role in the incidence of criminal behaviour in Britain. He stated that:

It is a mere truism to say that the progress of popular education, and the formation of religious habits, are fatally opposed by the temptations to animal pleasures, which abound wherever BEER-HOUSES and low ALE-HOUSES abound.

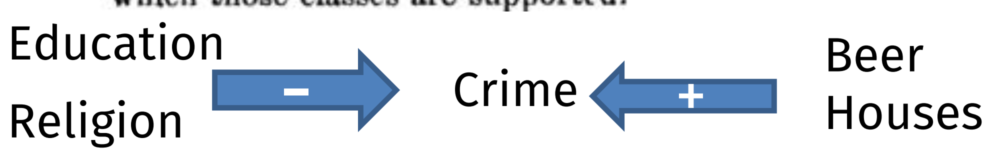
22 [Mar.

On the Relation between Crime, Popular Instruction, Attendance on Religious Worship, and Beer-houses. By The Rev. John Clay, B.D., *Chaplain to the Preston House of Correction.*

[Read before the Statistical Society, 18th November, 1856.]

It is obvious that inquiries into the causes and encouragements of crime must lead to considerations touching the state of Popular Education, attention to Religious Observances, and the influence of Ale and Beer-houses in promoting drunkenness, and its consequent evils.

The five years ending with 1853 are well suited to inquiries of this nature, inasmuch as, during that period, there was little to disturb the ordinary course of existence among the labouring class; no political or social excitement; no cessation of the employments by which those classes are supported.



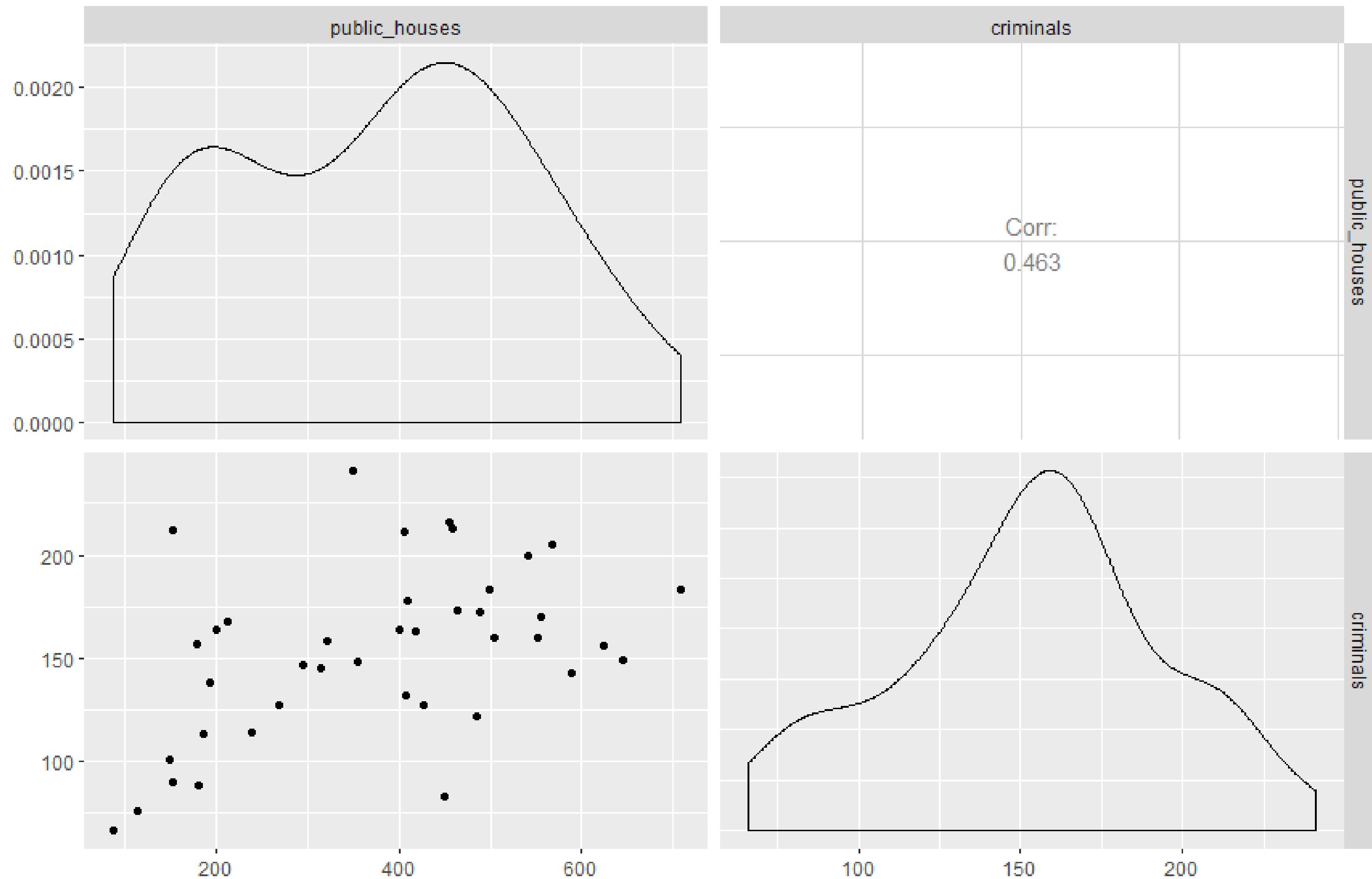
Source: John Clay (1856). "On the Relation Between Crime, Popular Instruction, Attendance on Religious Worship, and Beer-Houses", Journal of the Statistical Society of London, Vol. 20, No 1, pp 22-32.

English Data on Criminals, 1856

	county	region_name	region_code	criminals	public_houses	school_attendance	worship_attendance
1	Middlesex	South Eastern	1	200	541	560	434
2	Surrey	South Eastern	1	160	504	630	482
3	Kent	South Eastern	1	160	552	790	680
4	Sussex	South Eastern	1	147	295	820	678
5	Hants	South Eastern	1	178	409	990	798
6	Berks	South Eastern	1	205	568	930	698
7	Herts	South Midland	1	183	708	1020	888
8	Bucks	South Midland	1	156	624	1130	970
9	Oxford	South Midland	1	173	463	950	848
10	Northampton	South Midland	1	132	408	1090	976
11	Huntingdon	South Midland	1	149	646	1110	1104
12	Beds	South Midland	1	143	588	1250	1136
13	Cambridge	South Midland	1	170	555	960	926
14	Essex	Eastern	2	163	418	890	852
15	Suffolk	Eastern	2	164	200	880	988
16	Norfolk	Eastern	2	158	321	890	816
17	Wilts	South Western	3	157	178	1170	1018
18	Dorset	South Western	3	113	186	1150	938
19	Devon	South Western	3	138	192	760	804

EDA on Criminals/100K

```
> favstats(~criminals, data = crime)
 min  Q1 median  Q3 max mean   sd  n
  66 127    158 174 241  153 41.4 40
```

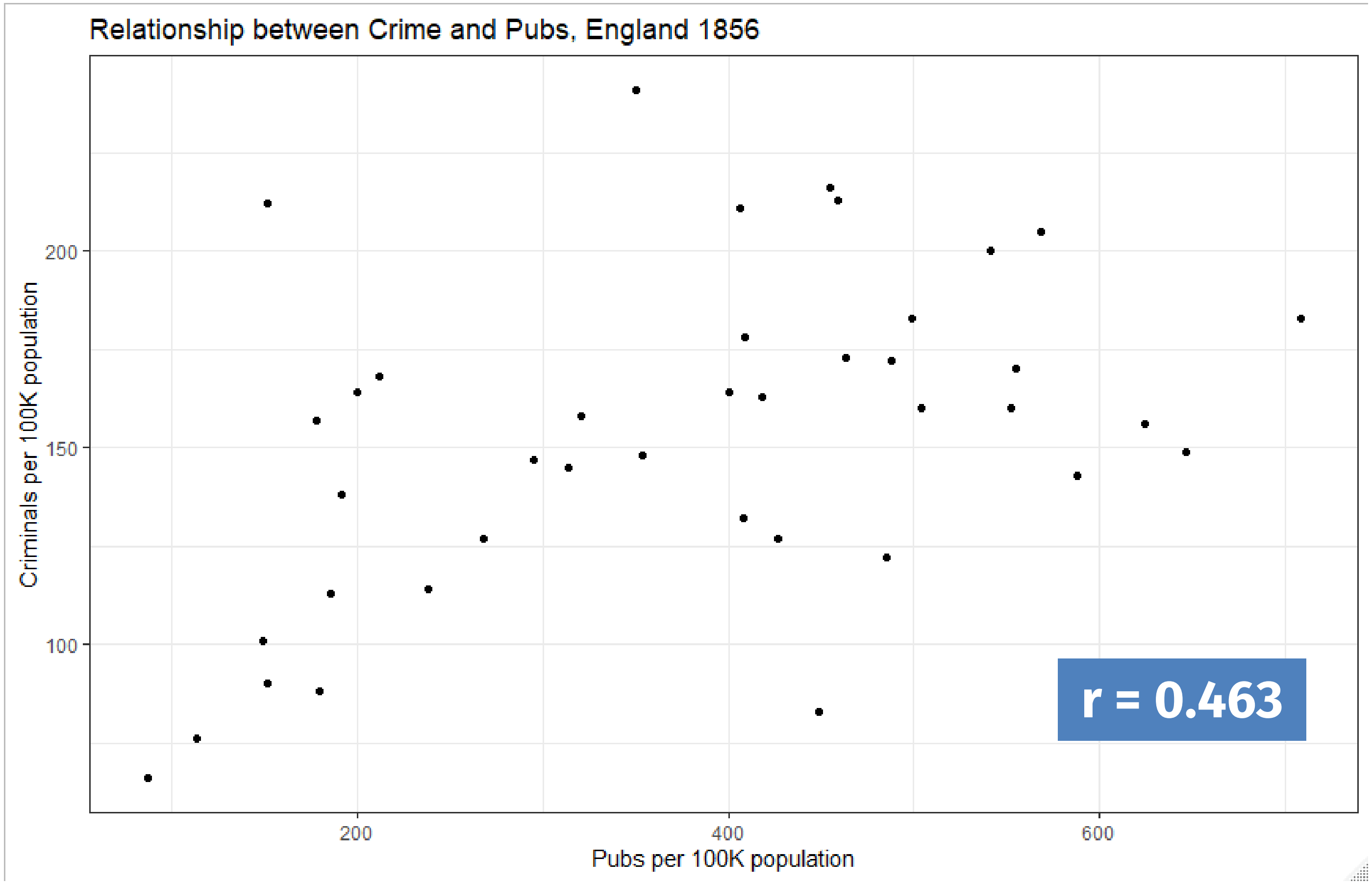


How would you predict criminals (per 100k population) for, e.g., Scottish regions?

The Need to Understand Relationships (2/2)

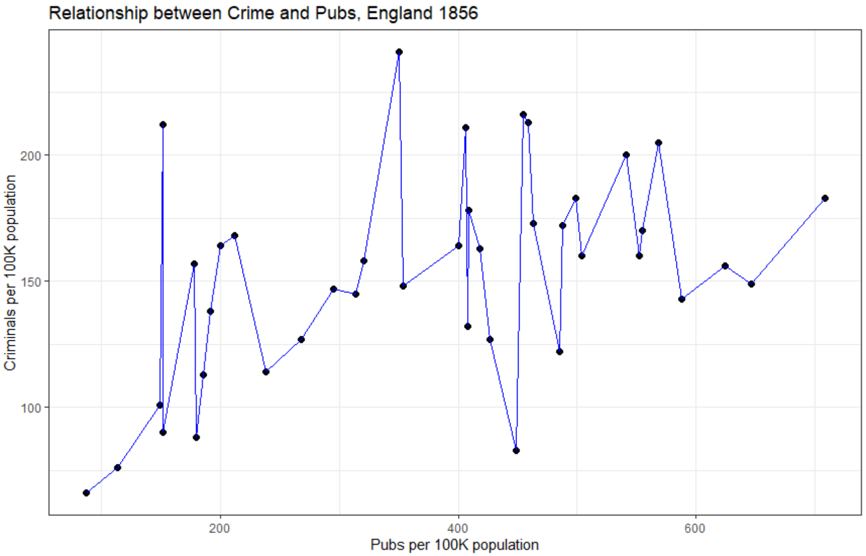
Clearly, the reverend considered public houses in Britain to be a scourge on society, namely that they "*promote drunkenness and its consequent evil*" (i.e., crime).

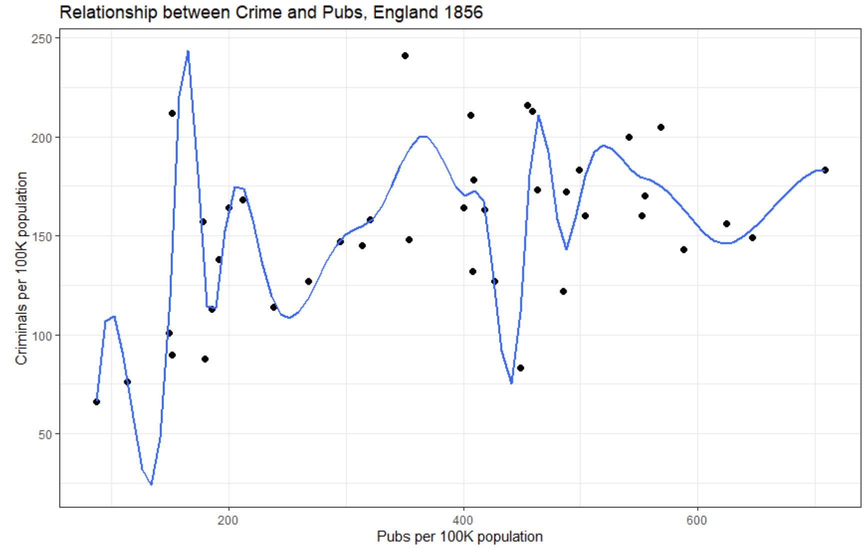
How well we can predict criminals (per 100k population) from the number of public houses (ale/beer houses per 100k population) using simple linear regression?

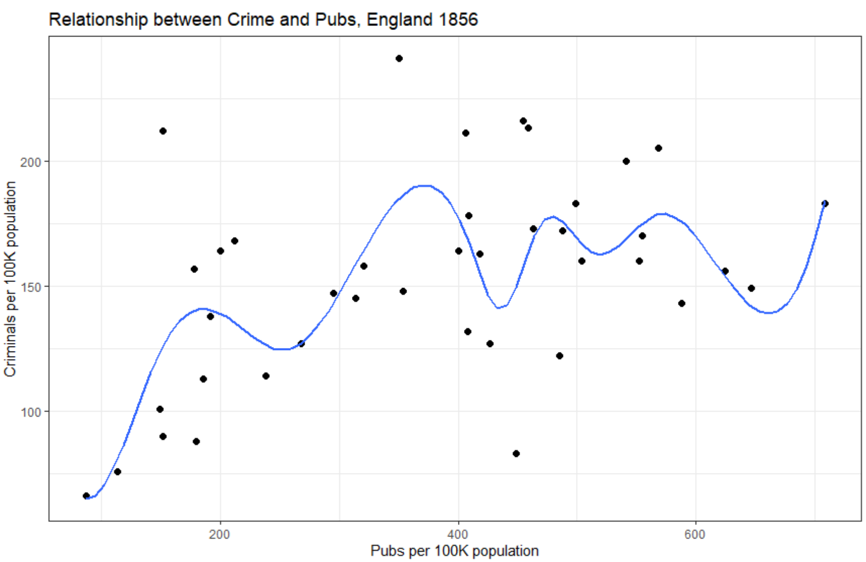


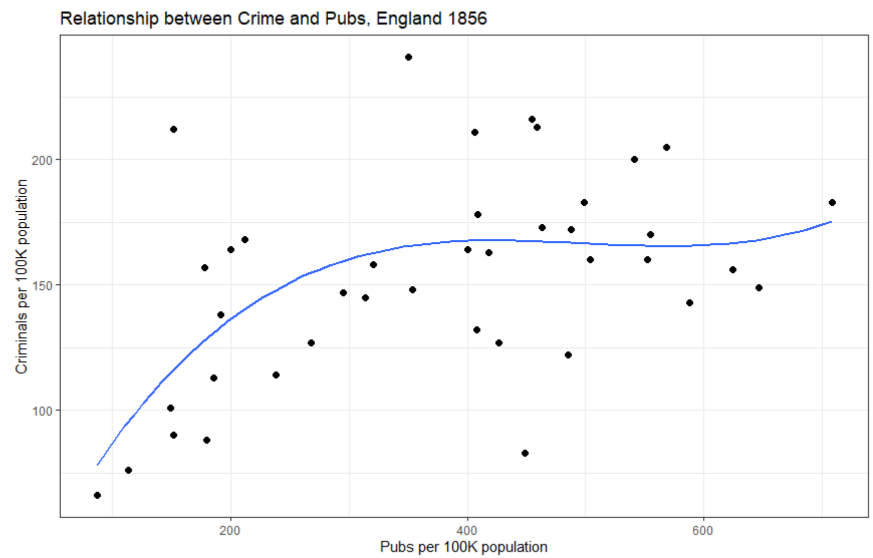
Statistics is the explanation of variation in the context of what remains unexplained, D Kaplan (2009)

Fitting a line through the points



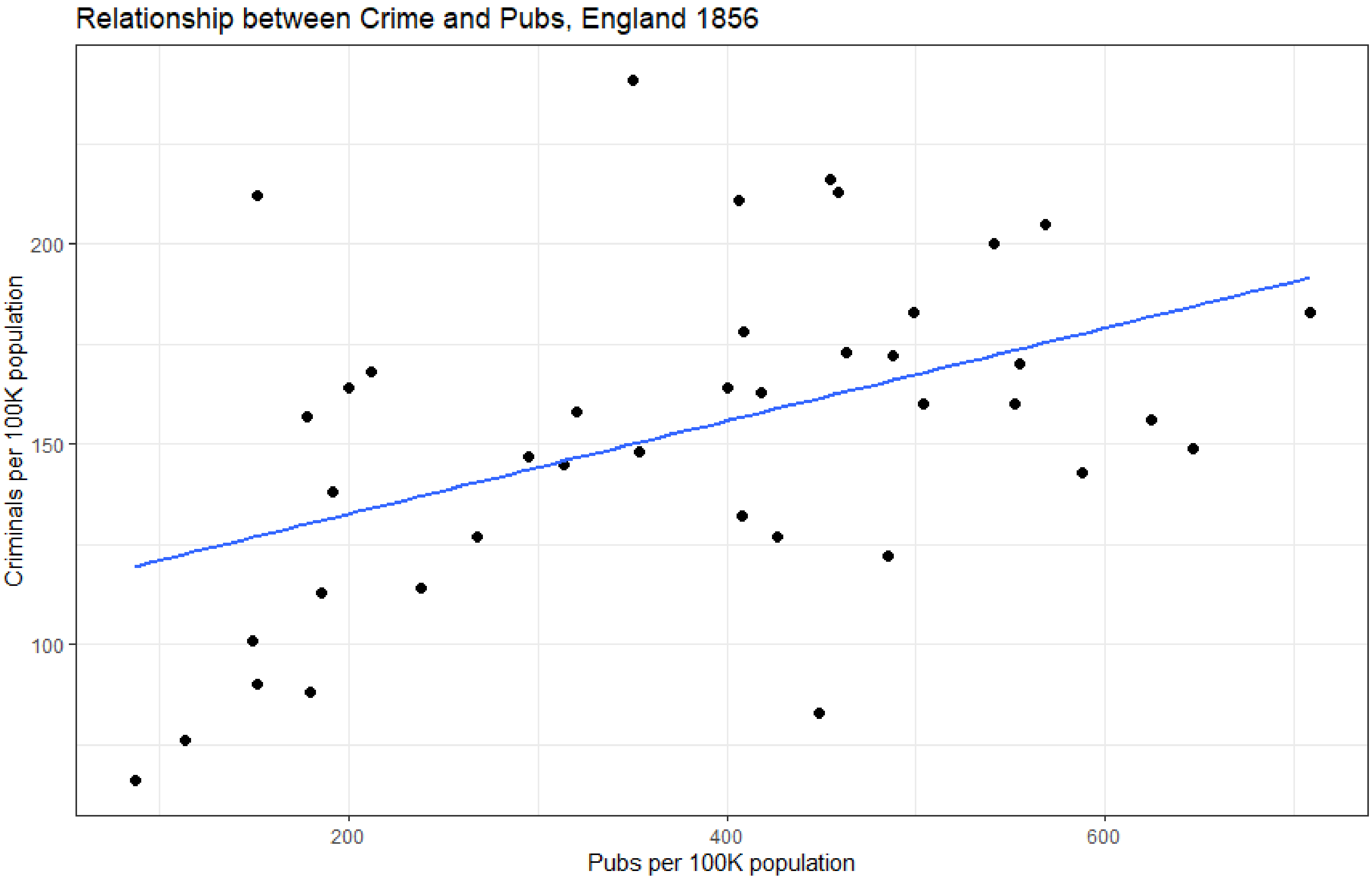




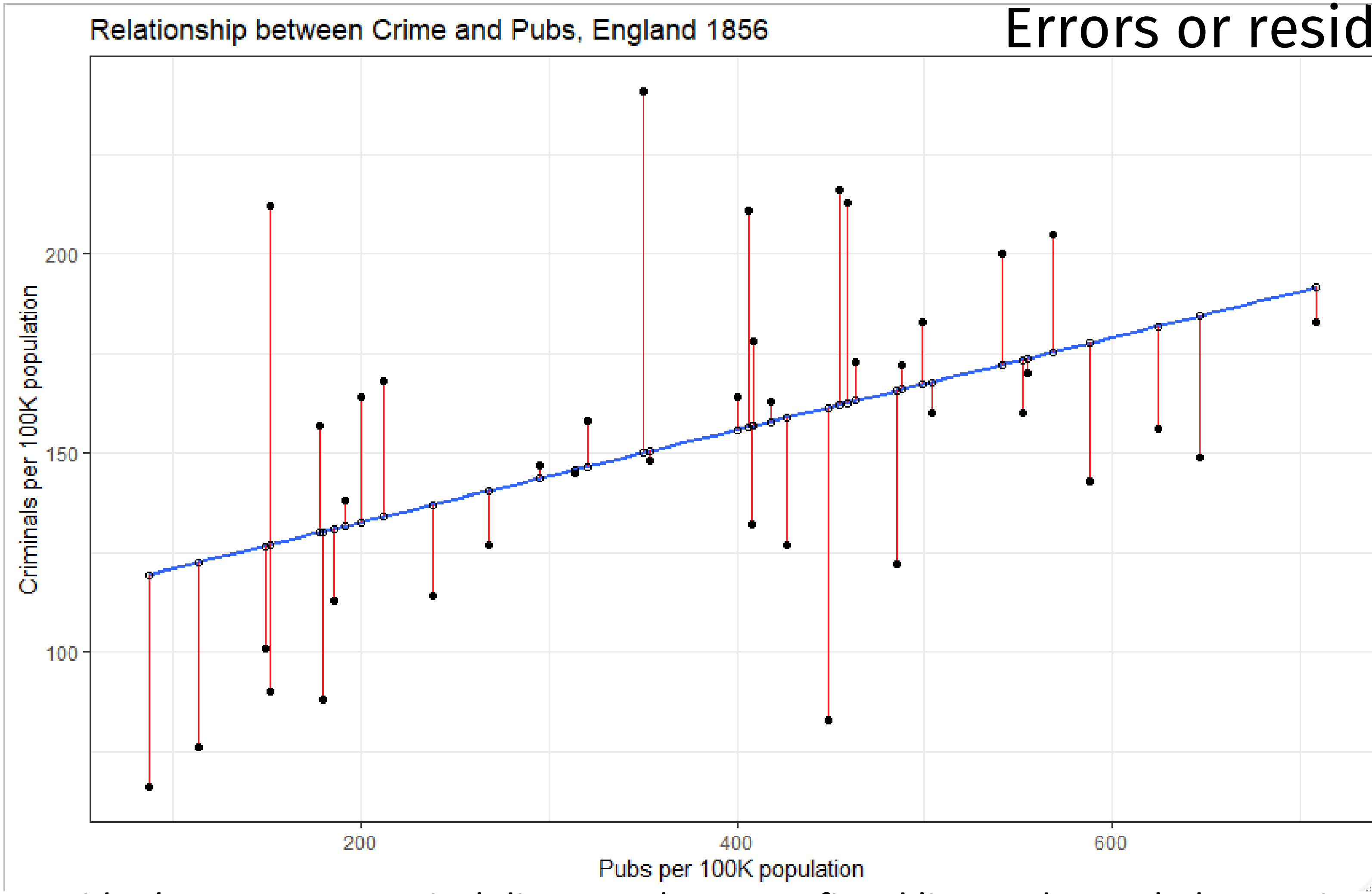


Statistics is the explanation of variation in the context of what remains unexplained, D Kaplan (2009)

Linear Models: fitting a **straight** line



Errors or residuals



Residuals or *errors*: vertical distances between fitted line and actual observations

We want to make these errors

- Have an average of zero, and
- Make them "as small as possible" (technically, minimize the squares of the errors)

Finding the best fit

The **regression algorithm** (technically known as ***Ordinary Least Squares***) finds the <u>parameters</u> of the line (its slope and intercept) such that:

1) **the AVERAGE error is zero**
(under-estimates and over-estimates cancel)

this is equivalent to saying that there should be no BIAS
also has the effect that the line passes through point (m_x, m_y)
i.e. the fitted value for the average x-value is the average y-value

2) **the AVERAGE SQUARED ERROR is as small as possible**
(want the scatter about the line to be as small as possible)

this is equivalent to saying we want to minimise the standard deviation of the residual errors

TRICK: for doing this by hand, if the two variables are <u>standardised</u> then the intercept is zero and the slope is simply the correlation, i.e.,

$$(\mathrm{Y} - \mathrm{m_Y}) / \mathrm{s_Y} = 0 + correl(X, Y) * (\mathrm{X} - \mathrm{m_x}) / \mathrm{s_X}$$

But of all these principles, least squares is the most simple: by the others we would be led into the most complicated calculations. --K.F. Gauss, 1809

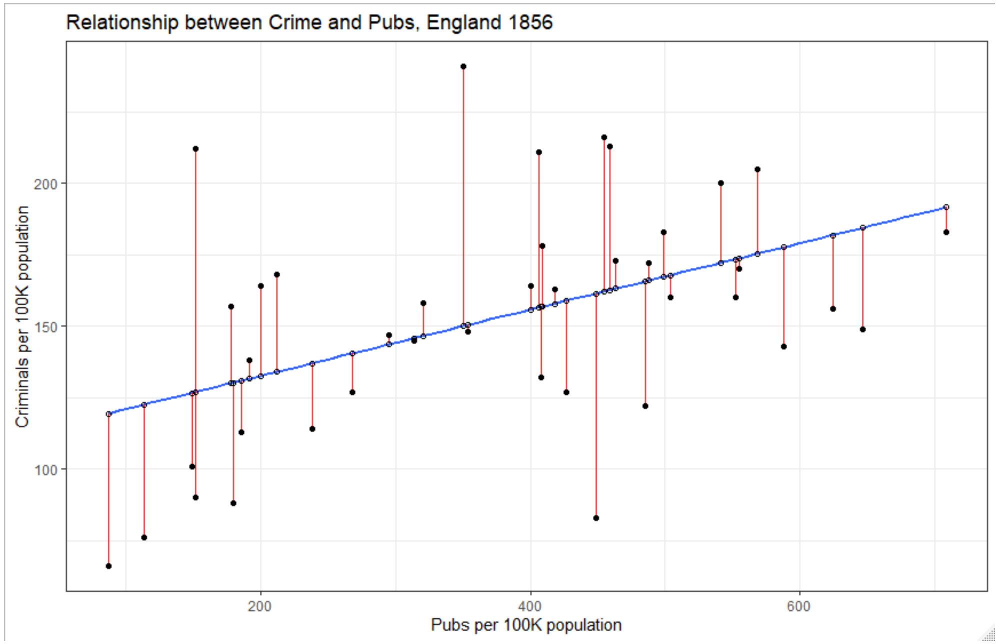
Drawing regression lines

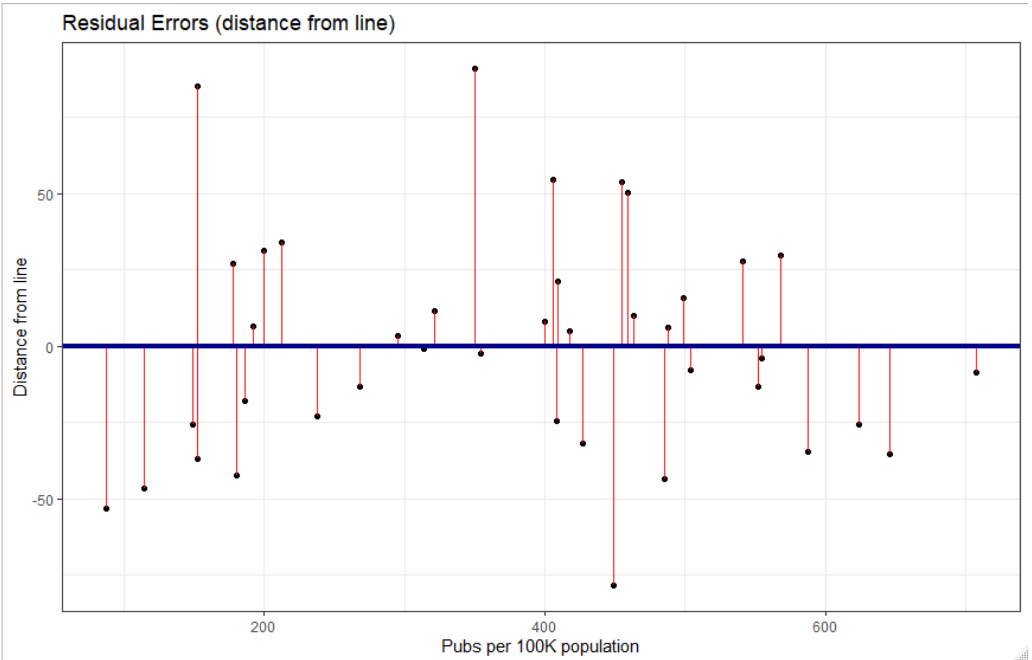
$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

$y = mx + b$

y	$\hat{y}$	Outcome variable
x	x_1	Explanatory variable
m	β_1	Slope
b	β_0	y intercept
	ε	Error (residuals)

Ordinary Least Squares: Find **best** line through the points



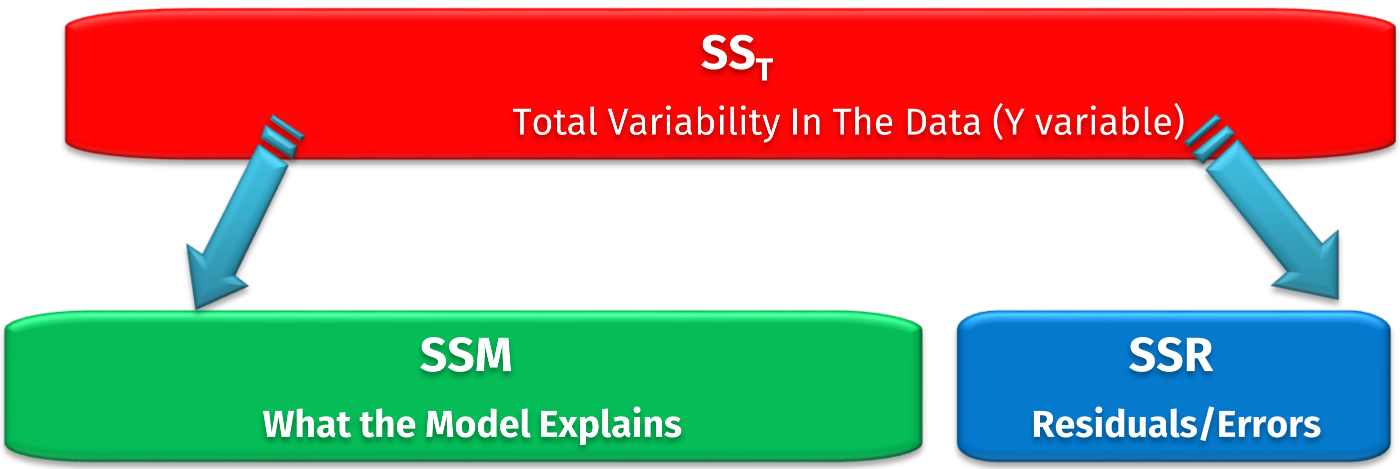


The regression model is never perfect, so it is more correct to think that it captures part of the variability:

systematic component + random component (errors/residuals)

Splitting Variability into *Model* and *Residual*

- SS_T : Total variability between Y variable values and the mean value of Y.
- SS_R : Residual/Error variability (variability between the regression model and the actual data).
- SS_M : Model variability (difference in variability between the model and the mean).

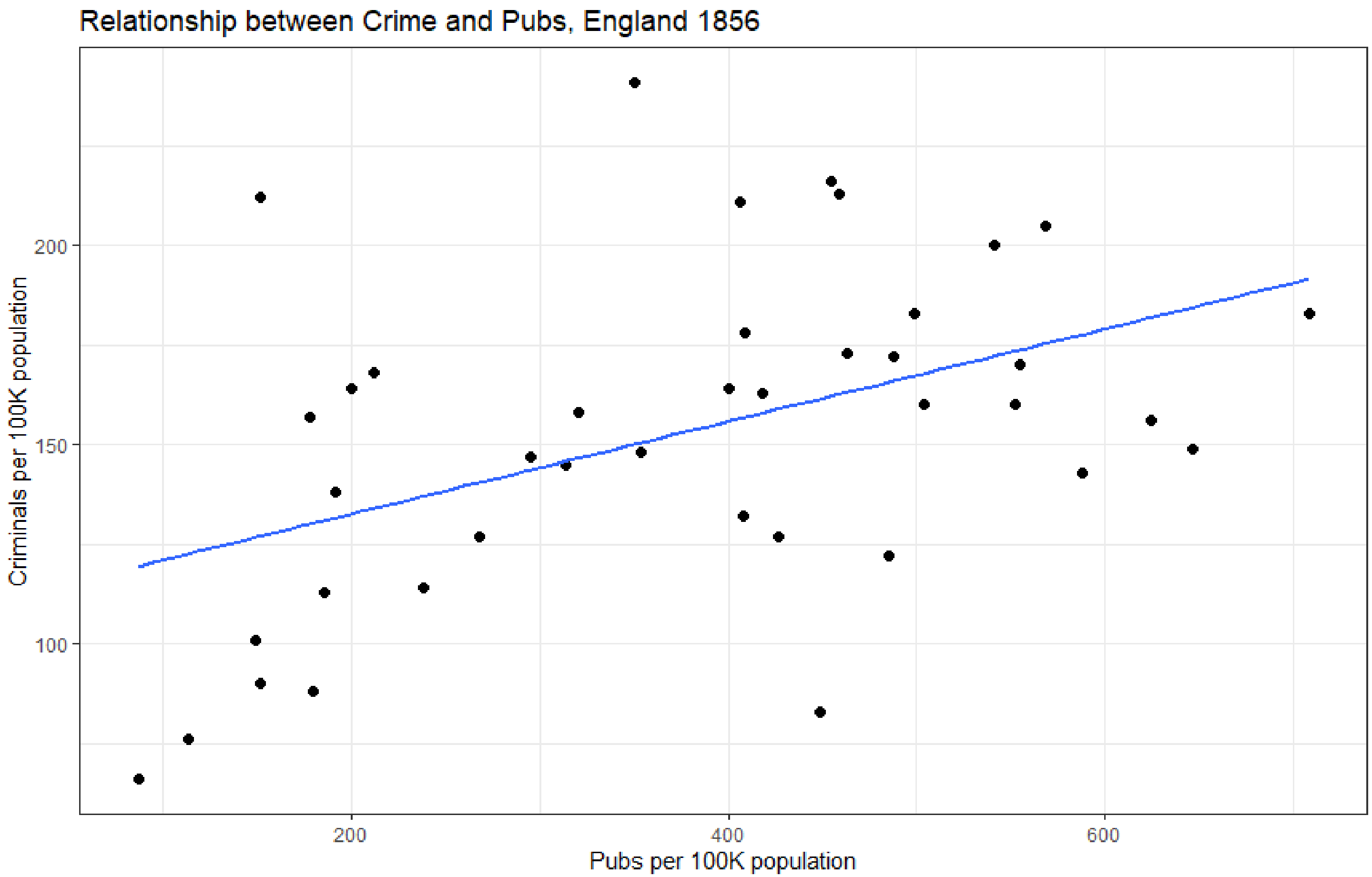


If the regression model results in better prediction than using the mean, then we expect SS_M to be much greater than SS_R

$$R^2 = \frac{SS_M}{SS_T}$$

R^2 is the proportion of the total variability of Y which is explained by the model

Modelling relationship between crime and pubs



$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$\widehat{criminals} = \beta_0 + \beta_1 * pubs + \varepsilon$$

Linear Models

Let us break down the regression line

$$\hat{y} = b_0 + b_1 \cdot x$$

$$\downarrow$$

$$\overbrace{\hat{y}}^{fitted} \;=\; \overbrace{b_0}^{intercept} \;+\; \overbrace{b_1}^{slope} \cdot x$$

$$\downarrow$$

$$y = \overbrace{b_0 + b_1 \cdot x}^{\hat{y}} + \text{error}$$

$$\downarrow$$

$$\text{outcome}_i = (\text{Model}_i) + \text{error}_i$$

The reason we call this last equation a ***linear model***, is that we are trying to predict *y* with a linear equation.

We also realise that our model is not perfect, and there will be errors between what actually happened in the data set (***outcome**$_i$*) vs what we predicted would happen (***model`**$_i$*)

Modelling linear models in R

```
lm(y ~ x1 + x2 +x3, data = dataframe)
```

outcome

predictors
or
exposure

your
data

We will also use the broom package

broom: **tidy models**

tidy() **Model coefficients**

glance() **Model fit**

augment() **Model predictions**

Modelling linear models in R

We build models and test how much of the variability can be explained by our ‘model’ (systematic/model variance) versus the noise, the residual/random ‘error’ (unsystematic/random variance)

data = (model) + error

The first, and easiest, model is the good old average, or arithmetic mean

We get a model for the mean y by using

```
lm(y ~ 1, data = dataframe)
```

Modelling crime: model 0 , the mean

```
> crime %>% select(criminals) %>% skim()
-- Data Summary ------------------------
                           Values
Name                       Piped data
Number of rows             40
Number of columns          1
_______________________
Column type frequency:
  numeric                  1
________________________
Group variables            None

-- Variable type: numeric ------------------------------------------------------------------
# A tibble: 1 x 11
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>             <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 criminals             0             1  153.  41.4    66   127  158.  174.   241
> model0 <- lm(criminals ~ 1, data= crime)
> model0 %>% broom::tidy()
# A tibble: 1 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     153.      6.55      23.3 1.57e-24
```

$$\widehat{criminals} = 153 + \varepsilon$$

- What is the estimate of 153? The same value as the mean of crime.
- Where does the std.error of 6.55 come from? How about $\frac{41.4}{\sqrt{40}} = 6.55$

Modelling crime and pubs

```
> model1 <- lm(criminals ~ public_houses, data= crime)
> model1 %>%  broom::tidy()
# A tibble: 2 x 5
  term          estimate std.error statistic     p.value
  <chr>            <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)    109.       14.8        7.41 0.00000000690
2 public_houses    0.116     0.0361     3.22 0.00263
```

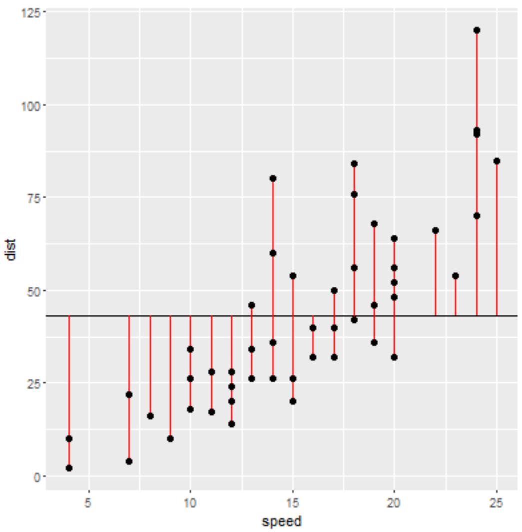
$$\widehat{\boldsymbol{criminals}} = \mathbf{109} + \mathbf{0.116} * \boldsymbol{pubs} + \boldsymbol{\varepsilon}$$

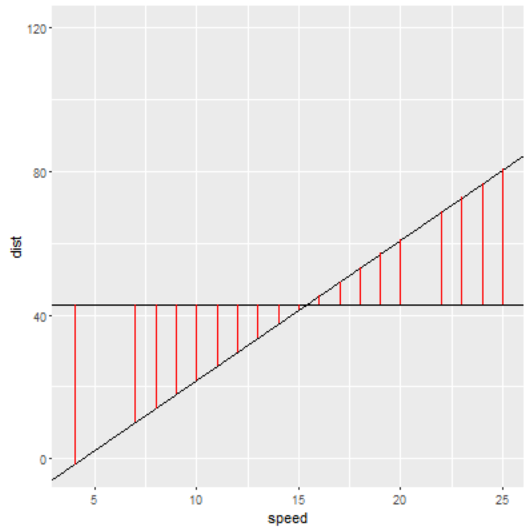
- On average, a one unit increase in X is associated with a $\boldsymbol{\beta_1}$ change in Y.
- In our case, if number of pubs increases by 1 (per 100K), we expect criminals to increase by 0.116 (per 100K)
- We never really worry about the intercept. In this case, the value of 109 makes no sense in this context, as there are no regions with zero pubs/ 100k

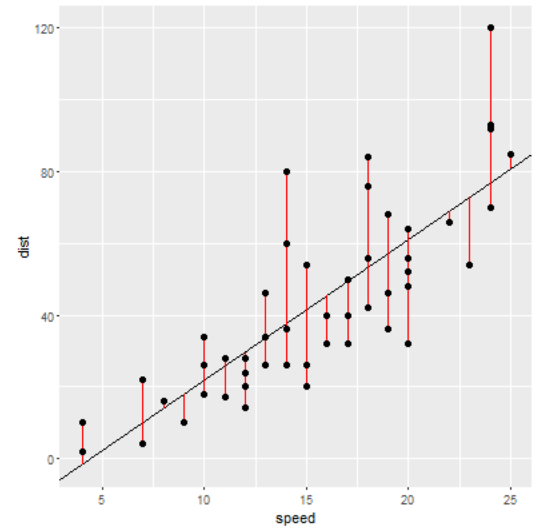
Splitting Variability into *Model* and *Residual*

SSTotal = SSModel + SSResidual

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$







SSTotal how well the mean fits the data. The mean is the simplest model we can fit and hence serves as the model to which the least-squares regression line is compared to.

SSModel how much better the regression line is compared to the mean (i.e. the difference between the SStotal and the SSresidual).

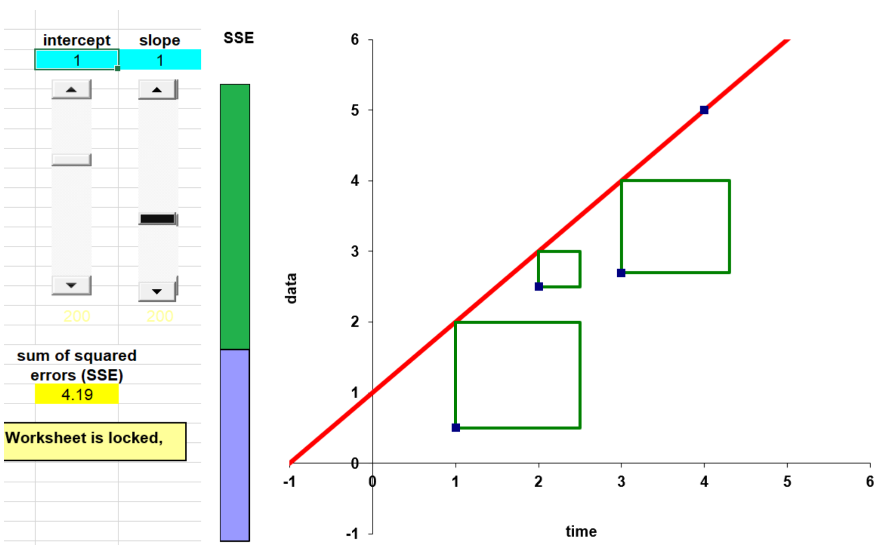
SSResidual how well the regression line fits the data.

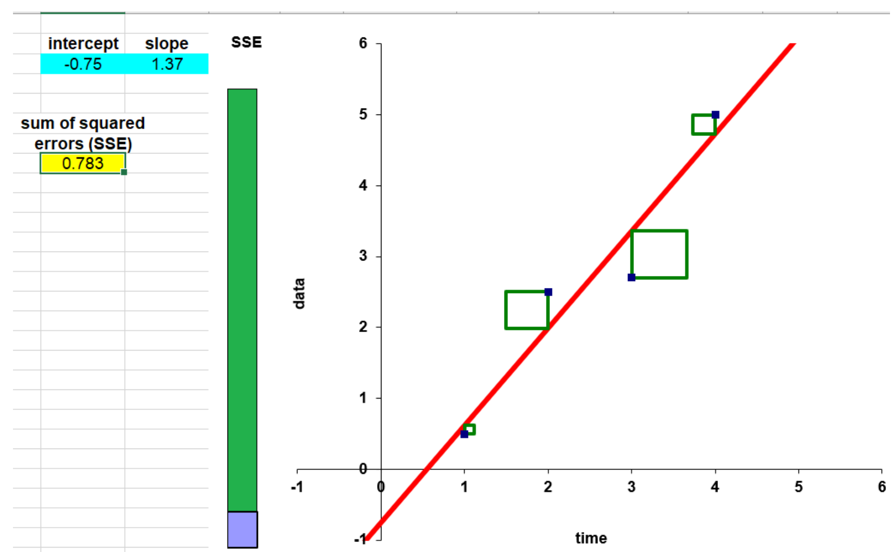
R^2: Measure of fit

SSTotal = SSModel + SSResidual

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$R^2 = \frac{SS_{Total} - SS_{Residual}}{SS_{Total}}$$





Is the model a good fit?

- R-square (explained variance / total variance)
  - How much variation in Y is explained by X.
  - The higher the better; No magic threshold; depends on domain
  - In the absence of a model you would just use the naïve model of mean(Y) to predict; You can think of R^2 as how much better your model is compared to the naïve model of the mean
  - Template: *This model explains X% of the variation in Y*

```
> model1 %>%  broom::glance()
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC deviance df.residual  nobs
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
1     0.214         0.194  37.2      10.4 0.00263     1  -200.  407.  412.   52565.          38    40
```

model1 explains about 21% of the total variation in Criminals...

... while the typical SE is **37.2**

Correlation, R^2, Adjusted R^2

- Letter we use for correlation coefficient is r
- **R^2 = correlation²**
  - It only works if you have one explanatory variable X
- What happens when a model has multiple Xs?
  - We can't use the regular R^2, we use the adjusted R^2

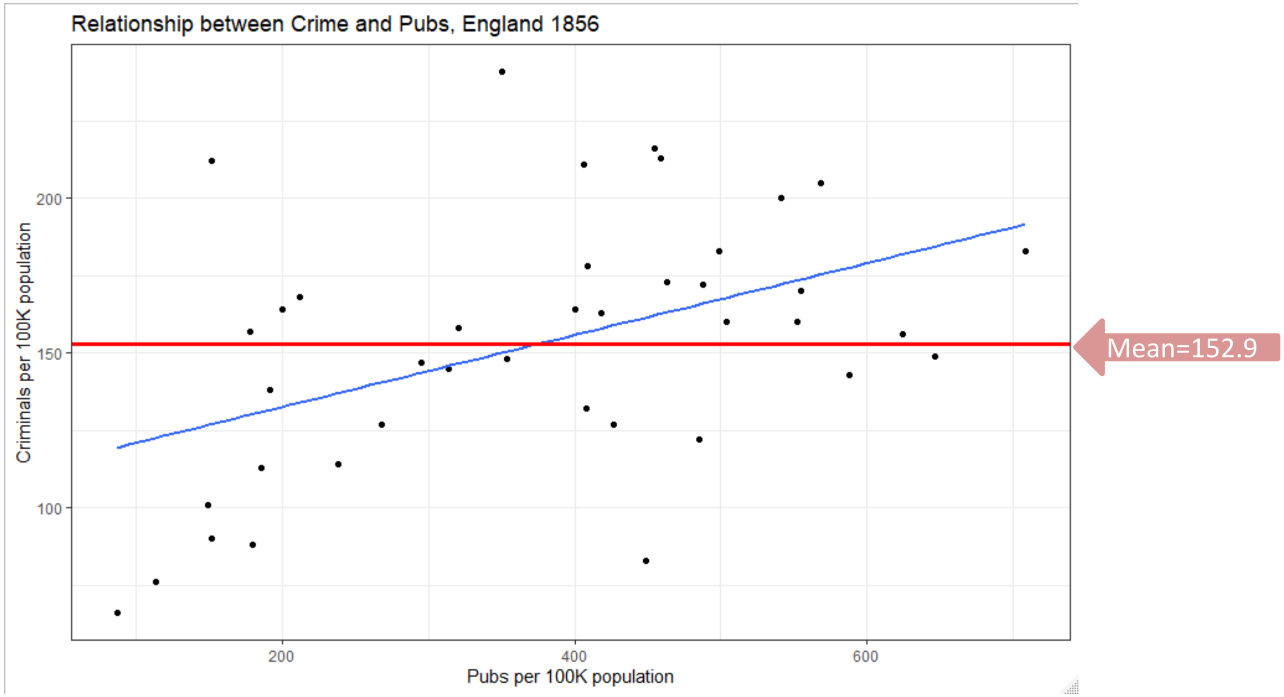
$$R^2_{adj} = R^2 \times \frac{\text{number of observations} - 1}{\text{number of observations} - \text{number of variables in model} - 1}$$

- Penalizes for small data sets and lots of explanatory variables

```
#  r_squared adj_r_squared   mse  rmse sigma statistic p_value    df
       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl>
1      0.214         0.194 1314.  36.3  37.2      10.4   0.003     2
```

Is slope *b* significant?

- Is there a (linear) relationship between ***criminals*** and ***pubs***?
  - Yes $\Rightarrow$ slope (*b*) is different from zero. The *mean = 152.9* has a slope of zero
  - No $\Rightarrow$ slope (*b*) is zero (we obtained a non-zero slope by chance only)
- Every regression coefficient is a δ*, like in hypothesis testing
  - Something that reflects the population and might be zero or not



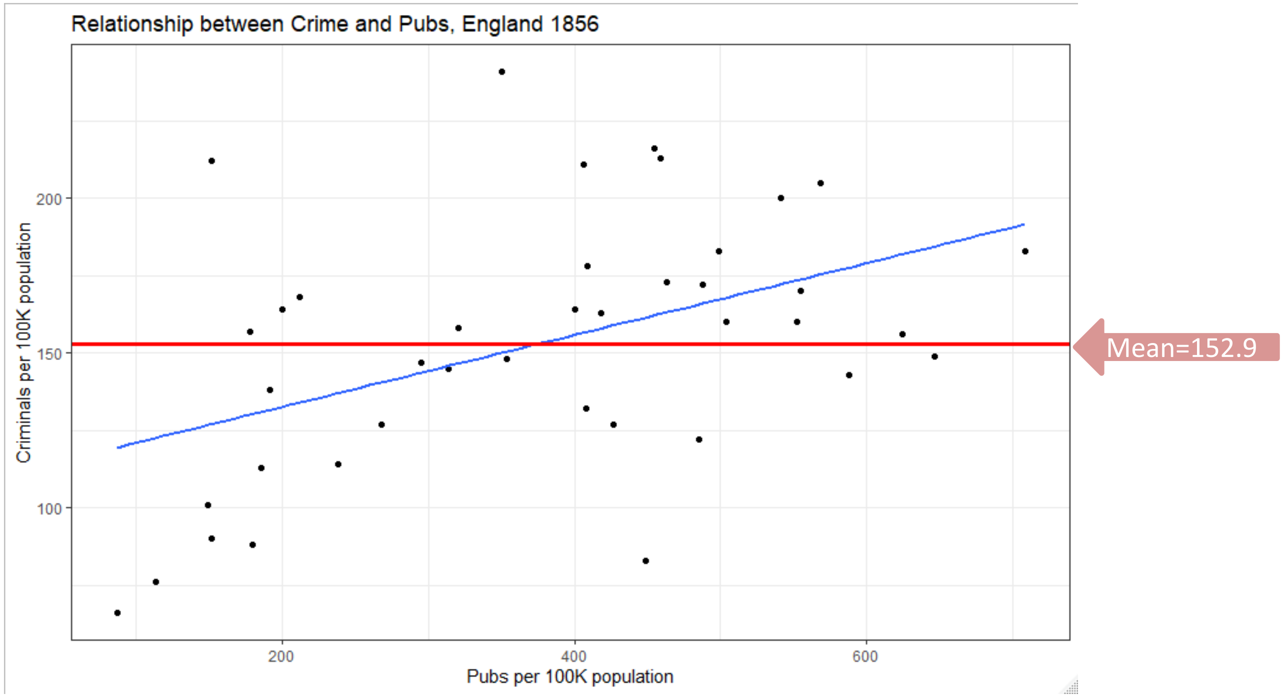
- Test: Could *b* be equal to 0 ? (Could it be that there is no relationship ?)

  compute t-statistic (*b* / SE of *b*)

  if absolute value(test statistic) > 2, *b* is probably not zero (95% confident)

  p-value = probability that b could be zero (if <5%, confident that $b \neq 0$)

Is slope *b* significant?




```
# A tibble: 2 x 7
  term           estimate std_error statistic p_value lower_ci upper_ci
  <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept       109.       14.8        7.41   0         79.5    139.
2 public_houses     0.116     0.036      3.22   0.003      0.043    0.189
```

In class exercise
world_happiness_2021.R

World Happiness Report

From Wikipedia, the free encyclopedia

The **World Happiness Report** is a publication of the United Nations Sustainable Development Solutions Network. It contains articles and rankings of national happiness, based on respondent ratings of their own lives,[1] which the report also correlates with various (quality of) life factors.[2] As of March 2021, Finland had been ranked the happiest country in the world four times in a row.[3][4][5]

The report primarily uses data from the Gallup World Poll. Each annual report is available to the public to download on the World Happiness Report website.[6] The Editors of the 2020 report are John F. Helliwell, Richard Layard, Jeffrey D. Sachs, and Jan-Emmanuel De Neve. Associate Editors are Lara Aknin, Shun Wang, and Haifang Huang.

In this exercise we investigate whether there is a connection between happiness score *life_ladder* in 2019 and freedom. *life_ladder* will be the Y variable, what we are trying to understand/explain, and *freedom_to_make_life_choices*, the explanatory, X variable

1. Plot a scatterplot of all numerical variables using ggpairs().
   1. What explanatory variables (X's) have the highest relationship with Y?
   2. Are there any high correlations among the explanatory variables

2. Run two regressions
   - `model1 <- lm(happiness_score ~ 1, data = world_happiness_19)`
   - `model2 <- lm(happiness_score ~ freedom_to_make_life_choices, data= world_happiness_19)`
3. Write down the equation for model2 and check whether freedom's effect (slope) is different from zero
4. What % of the variability in people's happiness does freedom alone explain?

https://worldhappiness.report/
https://en.wikipedia.org/wiki/World_Happiness_Report

Simple regression and CAPM

- A fundamental idea in finance is that investors need financials incentives to take on risk. Thus, the expected return *R* on a risky investment, e.g., a stock, should exceed the risk-free return R_f, or the excess return $(R - R_f)$ should be positive
- Capital Asset Market Pricing Model (CAPM):

  *Return Stock = **α** + **β** * Return Market + error*

- ***α*** = "excess return" of a stock

  according to the CAPM, the "excess return" is the reward for taking on the "specific risk" of the stock. As this risk can be eliminated through holding a diversified portfolio, CAPM says the excess return should be close to zero

- ***β*** = "market risk" of a stock

  this is an indication of how sensitive the stock is to movements in the market as a whole; for a 1% market movement, b = 1 implies the stock tends to move 1%, b < 1 means the stock tends to move by less than 1%,   b > 1 means the stock tends to move more than 1%

- We measure the relationship between monthly returns on a stock against returns on the market index, e.g., the S&P500 and use regression to estimate the a and b parameters

S&P500 Market Index

SPDR® S&P 500® ETF Trust

SPY

Overview | Performance | Holdings | Document | Purchase Information

Top Holdings

Fund Top Holdings as of Sep 10 2020

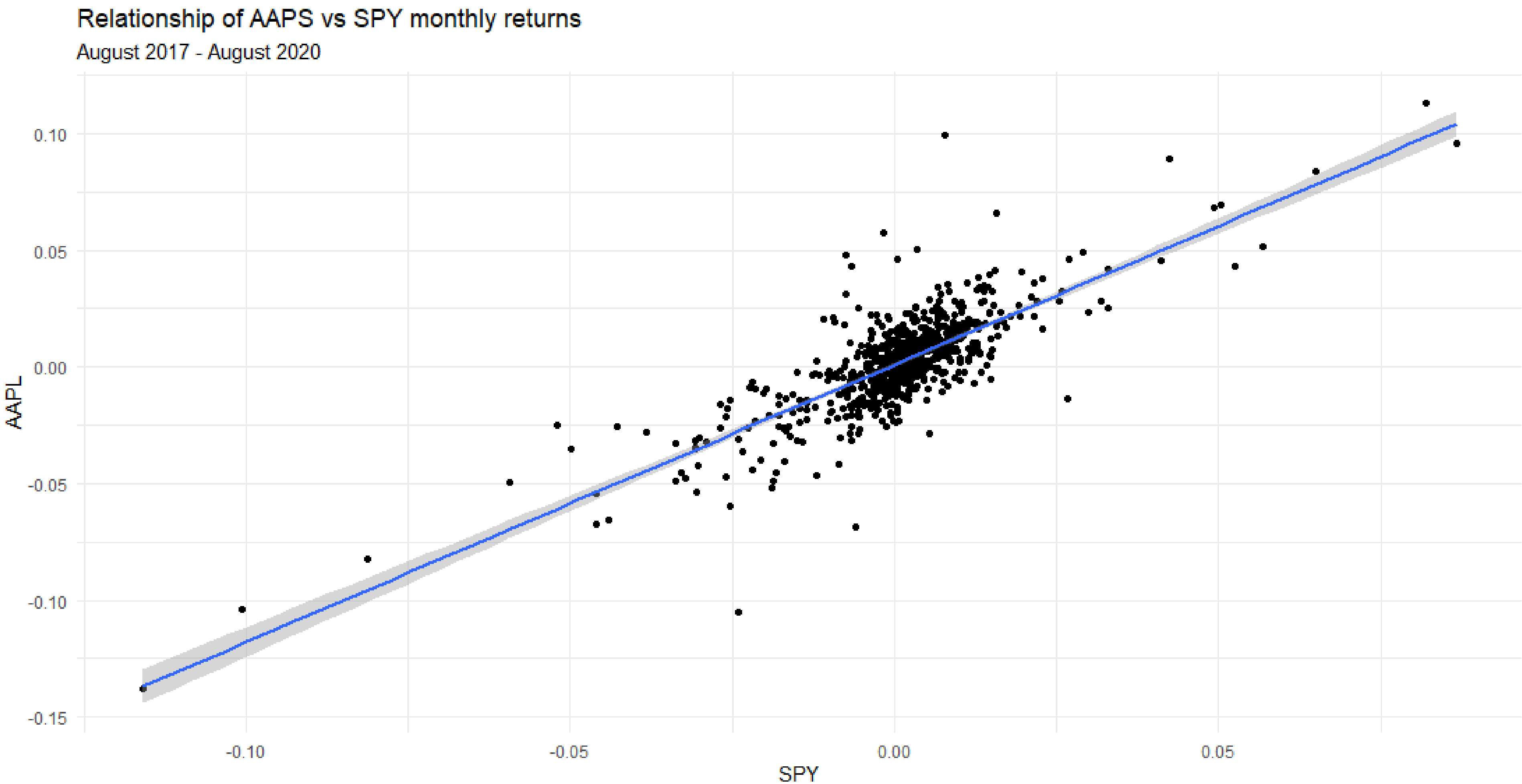
Name	Shares Held	Weight
Apple Inc.	173,845,010	6.70%
Microsoft Corporation	80,899,370	5.64%
Amazon.com Inc.	4,469,340	4.82%
Facebook Inc. Class A	25,649,900	2.33%
Alphabet Inc. Class A	3,203,468	1.66%
Alphabet Inc. Class C	3,119,612	1.62%
Berkshire Hathaway Inc. Class B	20,757,784	1.53%
Johnson & Johnson	28,128,524	1.40%
Visa Inc. Class A	18,012,536	1.23%
Procter & Gamble Company	26,423,544	1.23%

Index Top Holdings as of Sep 10 2020

Name	Weight
Apple Inc.	6.70%
Microsoft Corporation	5.64%
Amazon.com Inc.	4.82%
Facebook Inc. Class A	2.33%
Alphabet Inc. Class A	1.66%
Alphabet Inc. Class C	1.62%
Berkshire Hathaway Inc. Class B	1.53%
Johnson & Johnson	1.40%
Visa Inc. Class A	1.23%
Procter & Gamble Company	1.23%

https://www.ssga.com/us/en/individual/etfs/funds/spdr-sp-500-etf-trust-spy

AAPL (Y-axis) versus the SP500 (X-axis)



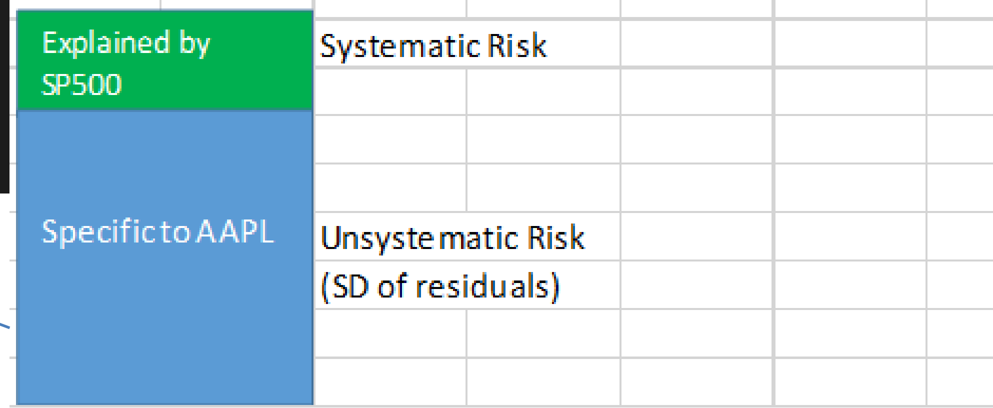
term <chr>	estimate <dbl>	std_error <dbl>	statistic <dbl>	p_value <dbl>	lower_ci <dbl>	upper_ci <dbl>
intercept	0.001	0.000	2.199	0.028	0.000	0.002
SPY	1.188	0.032	36.755	0.000	1.125	1.252

r_squared <dbl>	adj_r_squared <dbl>	mse <dbl>	rmse <dbl>	sigma <dbl>	statistic <dbl>	p_value <dbl>	df <dbl>	nobs <dbl>
0.636	0.635	0.0001545551	0.01243202	0.012	1350.946	0	1	776

Splitting volatility in MARKET (systematic) and SPECIFIC (unsystematic) risk

```
> anova(aapl_capm)
Analysis of Variance Table

Response: AAPL
           Df  Sum Sq  Mean Sq F value    Pr(>F)
SPY         1 0.20933 0.209335  1350.9 < 2.2e-16 ***
Residuals 774 0.11993 0.000155
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

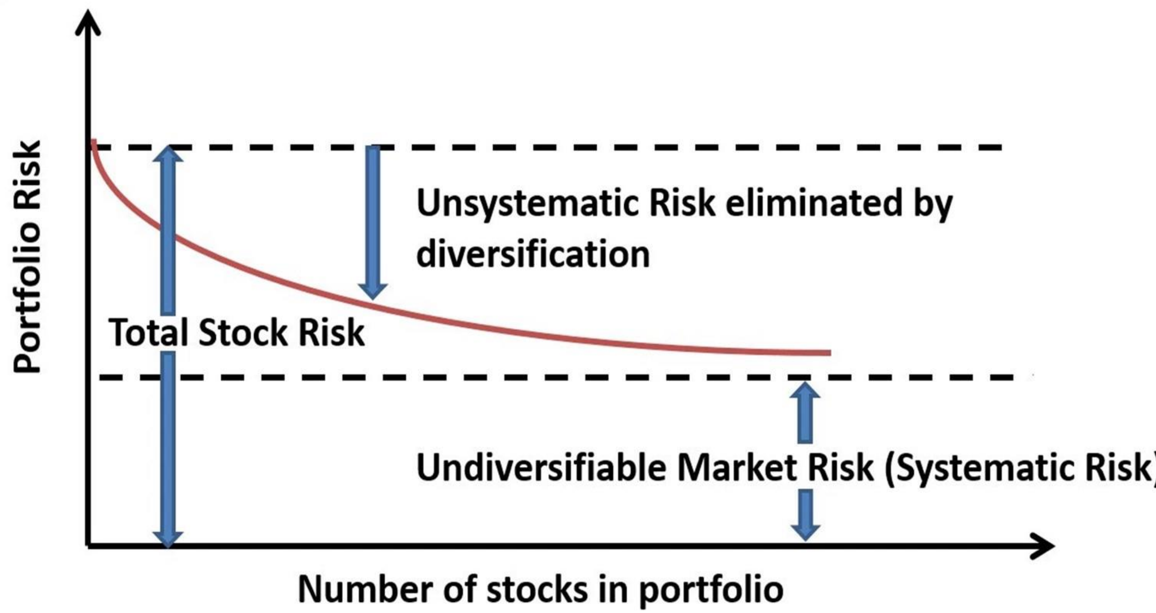
$$Variance(AAPL) = b^2 * Variance(SP500) + (SD\ of\ Residuals)^2$$



- The variability of AAPL which is explained by the market (SPY) is 0.20933
- The residual variability, i.e., that which is **not** explained by the market is 0.11993

$$R^2 = \frac{0.2093}{0.2093 + 0.1199}$$

$$= \frac{0.2093}{0.3292} = 0.636$$



CAPM: AAPL

```
> mosaic::msummary(aapl_capm)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0009832  0.0004472   2.199   0.0282 *
SPY         1.1882598  0.0323290  36.755   <2e-16 ***

Residual standard error: 0.01245 on 774 degrees of freedom
Multiple R-squared:  0.6358,    Adjusted R-squared:  0.6353
F-statistic:  1351 on 1 and 774 DF,  p-value: < 2.2e-16
```

AAPL monthly returns = 0.0009832 + 1.1883 * SPY monthly returns + *error*

- The “excess return” alpha= 0.0009832 suggesting excess return over market of 0.09% per month
- The beta of the stock is 1.189, and characterises the overall relationship, i.e., this is an 'average' beta over 36 months.
- R square 63.6% means that the market (the SP500) explains about two thirds of Apple’s volatility
- The Residual Standard Error of 0.01245 means that the specific risk of Apple is 1.245%

Reuters, from where Google get their data, calculate beta over a 5-year horizon, whereas Yahoo define beta as a 3-year calculation

Contents

Multiple Regression

- Generally there is more than one “explanatory variable” in which we are interested

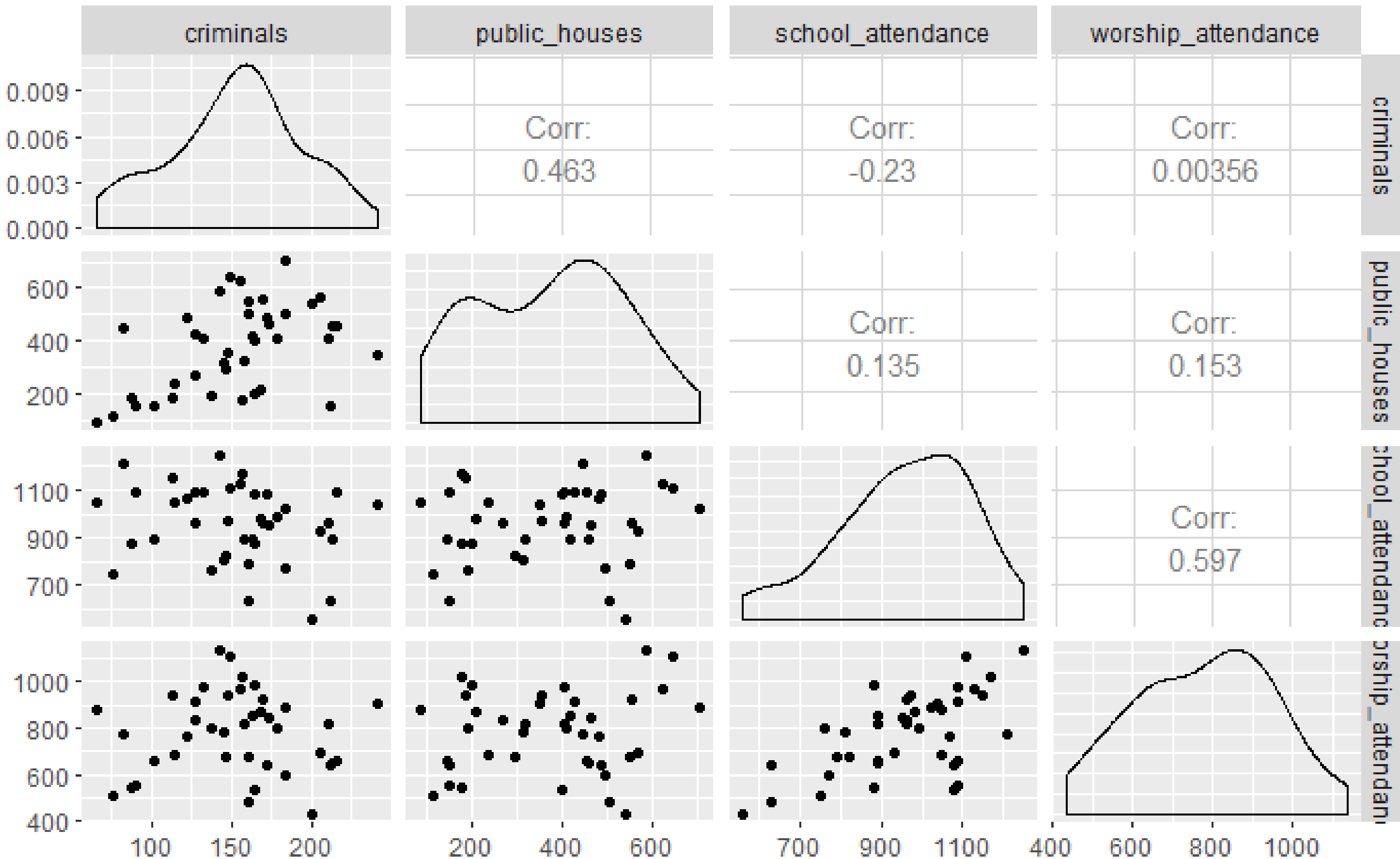
- We can generalise “simple” regression to deal with a set of explanatory (independent) variables - “multiple regression”
- We hope we can build more powerful models by taking other relevant factors into account
- Better forecasts $\Rightarrow$ tighter Confidence Intervals (smaller errors)
- We would like to be able to see which variables have a significant impact on our “target” variable (hypothesis testing)

- This raises a number of additional issues:
  - How do we interpret the models
  - How do we represent the data
  - We now have a number of different possible models - how to choose “the best”
  - What can go wrong
  - How can we avoid the pitfalls

Multiple Regression

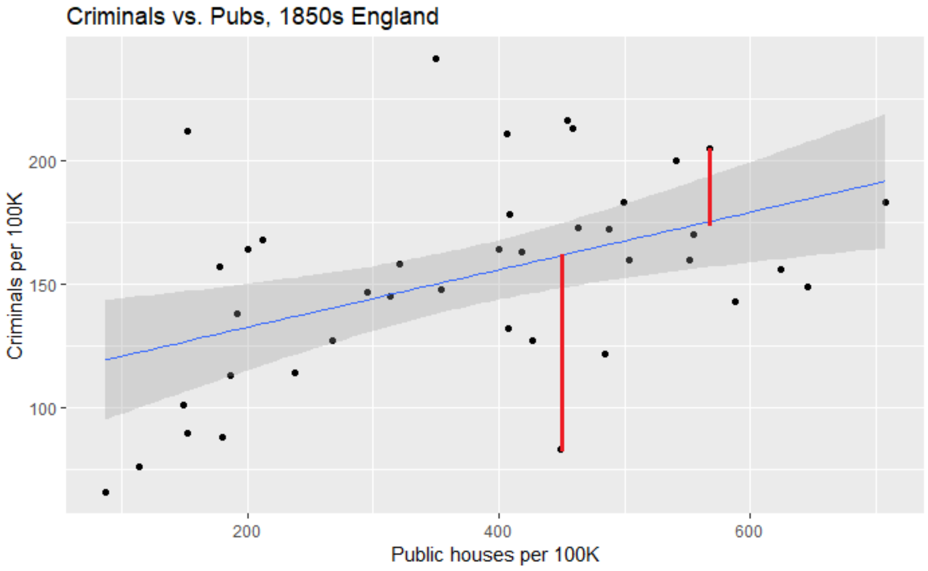
Simple Regression	**Multiple Regression**
$Y = b_0 + b_1 \cdot X_1 + \text{error}$	$Y = a + b_1 X_1 + b_2 X_2 + ... + b_n X_n + \text{error}$
One dependent/target variable (Y)	One dependent/target variable (Y)
One independent/explanatory variable (X_1)	A set of 'n' independent/explanatory variables ($X_1, X_2, \ldots X_n$)
The intercept $\mathbf{b_0}$ can be thought of as a "baseline"	The intercept $\mathbf{b_0}$ can be thought of as a "baseline"
The slope $\mathbf{b_1}$ is the increase in Y per unit increase in X	The slope 'b_i' is the increase in Y per unit increase in variable X_i

Crime and Pubs (1/2)



Model 1

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      109.3399    14.7553    7.41  6.9e-09 ***
public_houses      0.1162     0.0361    3.22   0.0026 **

Residual standard error: 37.2 on 38 degrees of freedom
Multiple R-squared:  0.214,	Adjusted R-squared:  0.194
F-statistic: 10.4 on 1 and 38 DF,  p-value: 0.00263
```



Intercept meaning?

Model 2


```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         172.8861    35.8385    4.82  2.6e-05 ***
public_houses         0.1233     0.0350    3.52   0.0012 ** 
school_attendance    -0.1011     0.0441   -2.30   0.0276 *  
worship_attendance    0.0393     0.0413    0.95   0.3482    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.6 on 36 degrees of freedom
Multiple R-squared:  0.319,     Adjusted R-squared:  0.262 
F-statistic: 5.61 on 3 and 36 DF,  p-value: 0.00291
```

$$\widehat{\text{criminals}} = 172.89 + 0.123\ \text{public\_houses} - 0.101\ \text{school\_attendance} + 0.039\ \text{worship\_attendance} + \mathit{error}$$

Criminals per 100K

Public_houses per 100K   Significant? Effect?

School_attendance per 10K   Significant? Effect?

Worship_attendance per 2K   Significant? Effect?

Compare both models

```
# A tibble: 2 x 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept        109.       14.8        7.41   0        79.5    139.
2 public_houses      0.116     0.036      3.22   0.003     0.043    0.189
```

```
# A tibble: 4 x 7
  term                 estimate std_error statistic p_value lower_ci upper_ci
  <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept             173.       35.8        4.82   0       100.     246.
2 public_houses           0.123     0.035      3.52   0.001     0.052    0.194
3 school_attendance      -0.101     0.044     -2.30   0.028    -0.19    -0.012
4 worship_attendance      0.039     0.041      0.951  0.348    -0.045    0.123
```

	(1)	(2)
(Intercept)	109.34 ***	172.89 ***
	(14.76)	(35.84)
public_houses	0.12 **	0.12 **
	(0.04)	(0.04)
school_attendance		-0.10 *
		(0.04)
worship_attendance		0.04
		(0.04)
N	40	40
R2	0.21	0.32
logLik	-200.38	-197.52
AIC	406.75	405.05

*** p < 0.001; ** p < 0.01; * p < 0.05.

Column names: names, model1, model2

Contents

- Modelling and Correlation
- Simple regression
- Multiple regression
- Colinearity
- Categorical variables

Multi-collinearity

- High degree of correlation between two or more explanatory variables
- Example: Suppose *y* is function of 'very similar' *x* and *z*

$$\boldsymbol{y = 2 + 7x + 3z + e}$$

Software is not able to identify if relationship is

- $y = 2 + 7x + 3z + e$
- or $y = 2 + 3x + 7z + e$
- or $y = 2 + 12x - 2z + e$
- or . . . . . . etc . . . .

- Result is unreliable coefficient estimates, reflected in big uncertainty on coefficients, i.e. large standard errors, and thus small t-stats
- Prediction may be OK but description/explanatory will not be

- Moral: Check correlations between explanatory variables before regression

Multi-Collinearity

- High correlation between independent (explanatory) variables causes problems in the regression calculations

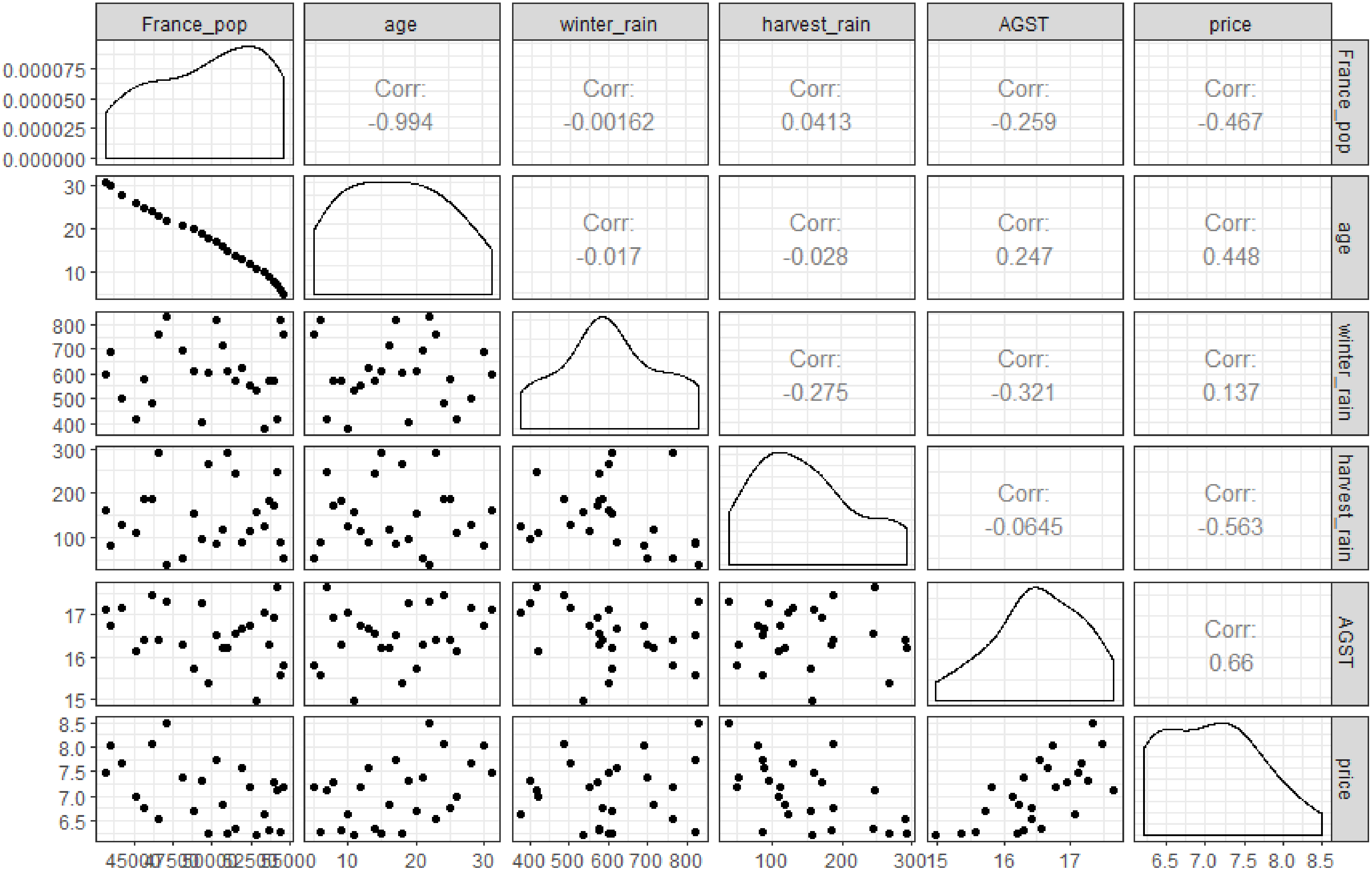
- Ideally, we'd like each new explanatory variable to have zero correlation with other explanatory variables and to bring in a lot of new information

- In the case of multi-collinearity the independent variables rob one another of explanatory power

- Signs of multi-collinearity
  - Magnitude/signs of regression coefficients different from expected
  - Standard error of coefficients high – lack of significance
  - VIF: Variance Inflation Factor > 5. Use ***car::vif(model1)***

- Solution:
  - Identify correlated variables
  - Remove one of them and repeat the regression

Predicting Wine Prices



Regression Models (1/2)

```
> model1 <- lm(price ~ AGST, data=wine)
> msummary(model1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.418      2.494   -1.37  0.18371
AGST           0.635      0.151    4.21  0.00034

Residual standard error: 0.499 on 23 degrees of freedom
Multiple R-squared:  0.435,	Adjusted R-squared:  0.41
F-statistic: 17.7 on 1 and 23 DF,  p-value: 0.000335
>
> model2 <- lm(price ~ AGST + harvest_rain, data=wine)
> msummary(model2)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.20265    1.85443   -1.19  0.24759
AGST          0.60262    0.11128    5.42 0.000019
harvest_rain -0.00457    0.00101   -4.52  0.00017

Residual standard error: 0.367 on 22 degrees of freedom
Multiple R-squared:  0.707,	Adjusted R-squared:  0.681
F-statistic: 26.6 on 2 and 22 DF,  p-value: 0.00000135
>
> model3 <- lm(price ~ ., data=wine)
> msummary(model3)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4503989 10.1888839   -0.04  0.96520
France_pop   -0.0000495  0.0001667   -0.30  0.76958
age           0.0005847  0.0790031    0.01  0.99417
winter_rain   0.0010425  0.0005310    1.96  0.06442
harvest_rain -0.0039581  0.0008751   -4.52  0.00023
AGST          0.6012239  0.1030203    5.84 0.000013

Residual standard error: 0.302 on 19 degrees of freedom
Multiple R-squared:  0.829,	Adjusted R-squared:  0.784
F-statistic: 18.5 on 5 and 19 DF,  p-value: 0.00000104
>
> car::vif(model3)
  France_pop          age  winter_rain harvest_rain         AGST
      98.253       97.220        1.299        1.117        1.275
```

```
> model4 <- lm(price ~ . - age, data=wine)
> msummary(model4)
                Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -0.3768251  2.1804321   -0.17   0.86453
France_pop   -0.0000508  0.0000170   -2.98   0.00743
winter_rain   0.0010417  0.0005070    2.05   0.05320
harvest_rain -0.0039578  0.0008518   -4.65   0.00016
AGST          0.6010955  0.0989776    6.07 0.0000062

Residual standard error: 0.294 on 20 degrees of freedom
Multiple R-squared:  0.829,     Adjusted R-squared:  0.795
F-statistic: 24.3 on 4 and 20 DF,  p-value: 0.000000195
>
> model5 <- lm(price ~ . - France_pop, data=wine)
> msummary(model5)
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -3.429980   1.765898   -1.94   0.06631
age           0.023931   0.008097    2.96   0.00782
winter_rain   0.001076   0.000507    2.12   0.04669
harvest_rain -0.003972   0.000854   -4.65   0.00015
AGST          0.607209   0.098702    6.15 0.0000052

Residual standard error: 0.295 on 20 degrees of freedom
Multiple R-squared:  0.829,     Adjusted R-squared:  0.794
F-statistic: 24.2 on 4 and 20 DF,  p-value: 0.000000204
```

World Happiness 2019


```
> # produce summary table comparing models using huxtable::huxreg()
> huxreg(model1, model2, model3, model4,model5,
+        statistics = c('#observations' = 'nobs',
+                       'R squared' = 'r.squared',
+                       'Adj. R Squared' = 'adj.r.squared',
+                       'Residual SE' = 'sigma'),
+        bold_signif = 0.05,
+        stars = NULL
+ ) %>%
+   set_caption('Comparison of models')
```

Comparison of models

	(1)	(2)	(3)	(4)	(5)
(Intercept)	**5.571**	**1.123**	**-2.716**	**-2.683**	**-3.414**
	(0.093)	**(0.523)**	**(0.523)**	**(0.494)**	**(0.530)**
freedom_to_make_life_choices		**5.598**	**3.111**	**2.654**	**2.261**
		(0.652)	**(0.538)**	**(0.520)**	**(0.517)**
log_gdp_per_capita			**0.614**	**0.391**	0.185
			(0.055)	**(0.074)**	(0.101)
social_support				**2.997**	**2.808**
				(0.720)	**(0.700)**
healthy_life_expectancy_at_birth					**0.048**
					(0.015)
#observations	144	143	137	137	135
R squared	0.000	0.344	0.656	0.696	0.720
Adj. R Squared	0.000	0.339	0.651	0.689	0.712
Residual SE	1.112	0.907	0.664	0.627	0.607

```
Column names: names, model1, model2, model3, model4, model5
>
> # Check whether any model has a VIF (Variance Inflation Factor) greater than 5
> car::vif(model3)
          log_gdp_per_capita freedom_to_make_life_choices
                    1.205776                     1.205776
> car::vif(model4)
          log_gdp_per_capita               social_support freedom_to_make_life_choices
                    2.495287                     2.568193                     1.261996
> car::vif(model5)
            log_gdp_per_capita healthy_life_expectancy_at_birth               social_support     freedom_to_make_life_choices
                      4.822825                         3.896442                     2.589634                         1.327391
```

World Happiness 2019

```
> # produce summary table comparing models using huxtable::huxreg()
> huxreg(model1, model2, model3, model4,model5,
+        statistics = c('#observations' = 'nobs',
+                       'R squared' = 'r.squared',
+                       'Adj. R Squared' = 'adj.r.squared',
+                       'Residual SE' = 'sigma'),
+        bold_signif = 0.05,
+        stars = NULL
+ ) %>%
+   set_caption('Comparison of models')
```

Comparison of models

	(1)	(2)	(3)	(4)	(5)
(Intercept)	**5.571**	**1.123**	**-2.716**	**-2.683**	**-3.414**
	(0.093)	**(0.523)**	**(0.523)**	**(0.494)**	**(0.530)**
freedom_to_make_life_choices		**5.598**	**3.111**	**2.654**	**2.261**
		(0.652)	**(0.538)**	**(0.520)**	**(0.517)**
log_gdp_per_capita			**0.614**	**0.391**	0.185
			(0.055)	**(0.074)**	(0.101)
social_support				**2.997**	**2.808**
				(0.720)	**(0.700)**
healthy_life_expectancy_at_birth					**0.048**
					(0.015)
#observations	144	143	137	137	135
R squared	0.000	0.344	0.656	0.696	0.720
Adj. R Squared	0.000	0.339	0.651	0.689	0.712
Residual SE	1.112	0.907	0.664	0.627	0.607

```
Column names: names, model1, model2, model3, model4, model5
>
> # Check whether any model has a VIF (Variance Inflation Factor) greater than 5
> car::vif(model3)
        log_gdp_per_capita freedom_to_make_life_choices
                  1.205776                     1.205776
> car::vif(model4)
        log_gdp_per_capita               social_support freedom_to_make_life_choices
                  2.495287                     2.568193                     1.261996
> car::vif(model5)
            log_gdp_per_capita healthy_life_expectancy_at_birth                   social_support     freedom_to_make_life_choices
                      4.822825                         3.896442                         2.589634                         1.327391
```

Contents

- Modelling and Correlation
- Simple regression
- Multiple regression
- Colinearity
- Categorical variables

Gapminder

- gapminder contains data on life expectancy, GDP, and population for all countries between 1952 and 2007

```
> install.packages("gapminder")
> library(gapminder)
> str(gapminder)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	1704 obs. of  6 variables:
 $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
 $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
 $ pop      : int  8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22227415 ...
 $ gdpPercap: num  779 821 853 836 740 ...
```

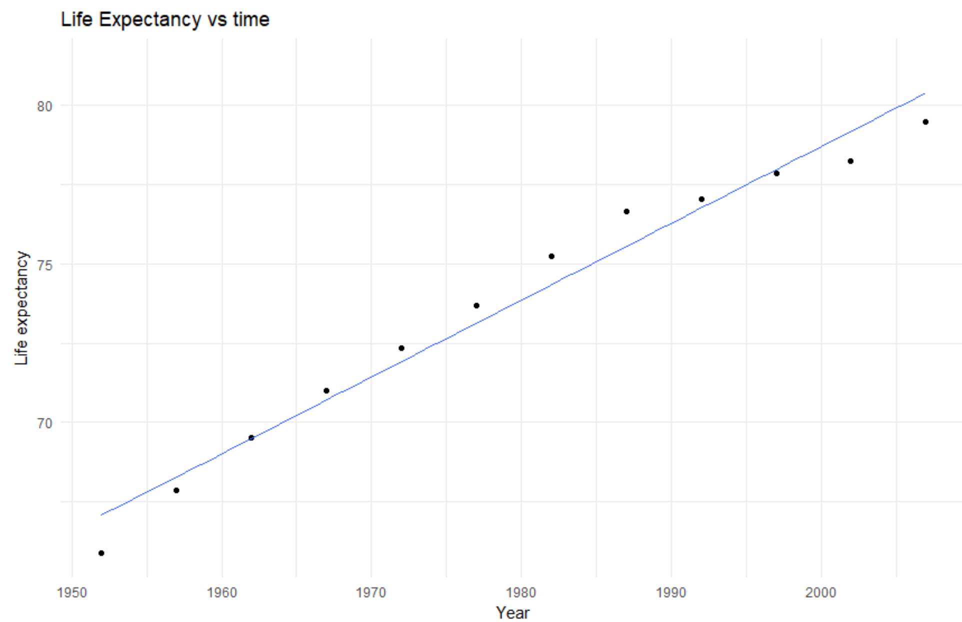
Filter

	country	continent	year	lifeExp	pop	gdpPercap
1	Afghanistan	Asia	1952	28.8	8.43e+06	779
2	Afghanistan	Asia	1957	30.3	9.24e+06	821
3	Afghanistan	Asia	1962	32.0	1.03e+07	853
4	Afghanistan	Asia	1967	34.0	1.15e+07	836
5	Afghanistan	Asia	1972	36.1	1.31e+07	740
6	Afghanistan	Asia	1977	38.4	1.49e+07	786
7	Afghanistan	Asia	1982	39.9	1.29e+07	978
8	Afghanistan	Asia	1987	40.8	1.39e+07	852
9	Afghanistan	Asia	1992	41.7	1.63e+07	649
10	Afghanistan	Asia	1997	41.8	2.22e+07	635
11	Afghanistan	Asia	2002	42.1	2.53e+07	727
12	Afghanistan	Asia	2007	43.8	3.19e+07	975

Life expectancy on year (1/2)

Pick one country and explore relationship between life expectancy and time

```
> tempCountry <- "Greece"  # Just a random example
> tempData <- subset(gapminder, country == tempCountry)  # temporary data file with country selected
> tempData
# A tibble: 12 x 6
   country continent  year lifeExp      pop gdpPercap
   <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
 1 Greece  Europe     1952    65.9  7733250     3531.
 2 Greece  Europe     1957    67.9  8096218     4916.
 3 Greece  Europe     1962    69.5  8448233     6017.
 4 Greece  Europe     1967    71    8716441     8513.
 5 Greece  Europe     1972    72.3  8888628    12725.
 6 Greece  Europe     1977    73.7  9308479    14196.
 7 Greece  Europe     1982    75.2  9786480    15268.
 8 Greece  Europe     1987    76.7  9974490    16121.
 9 Greece  Europe     1992    77.0 10325429    17541.
10 Greece  Europe     1997    77.9 10502372    18748.
11 Greece  Europe     2002    78.3 10603863    22514.
12 Greece  Europe     2007    79.5 10706290    27538.

> tempModel1 <- lm(lifeExp~year,data=tempData)
> msummary(tempModel1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -406.095     25.773   -15.8  2.2e-08
year           0.242      0.013    18.6  4.3e-09

Residual standard error: 0.778 on 10 degrees of freedom
Multiple R-squared:  0.972,     Adjusted R-squared:  0.969
F-statistic:  347 on 1 and 10 DF,  p-value: 4.32e-09
```

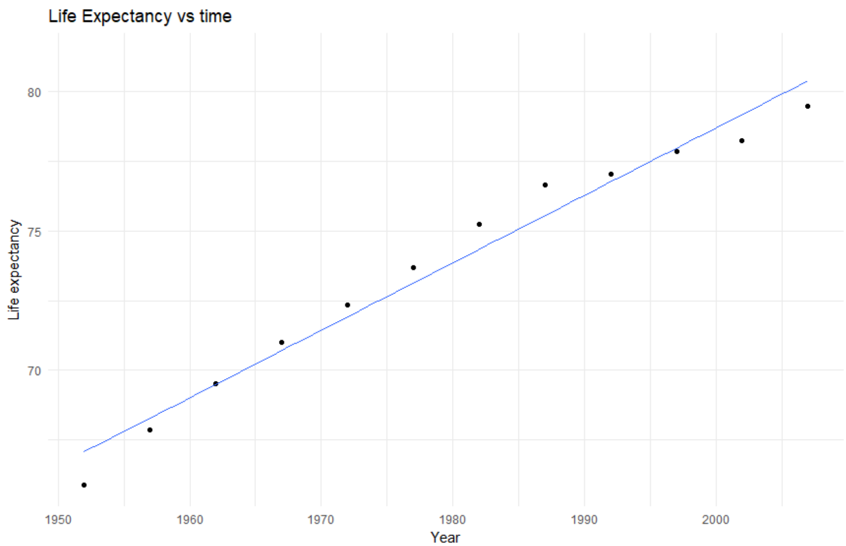


Value of intercept: Did Greek people have a life expectancy of **-406** years in year 0?

Life expectancy on year (2/2)

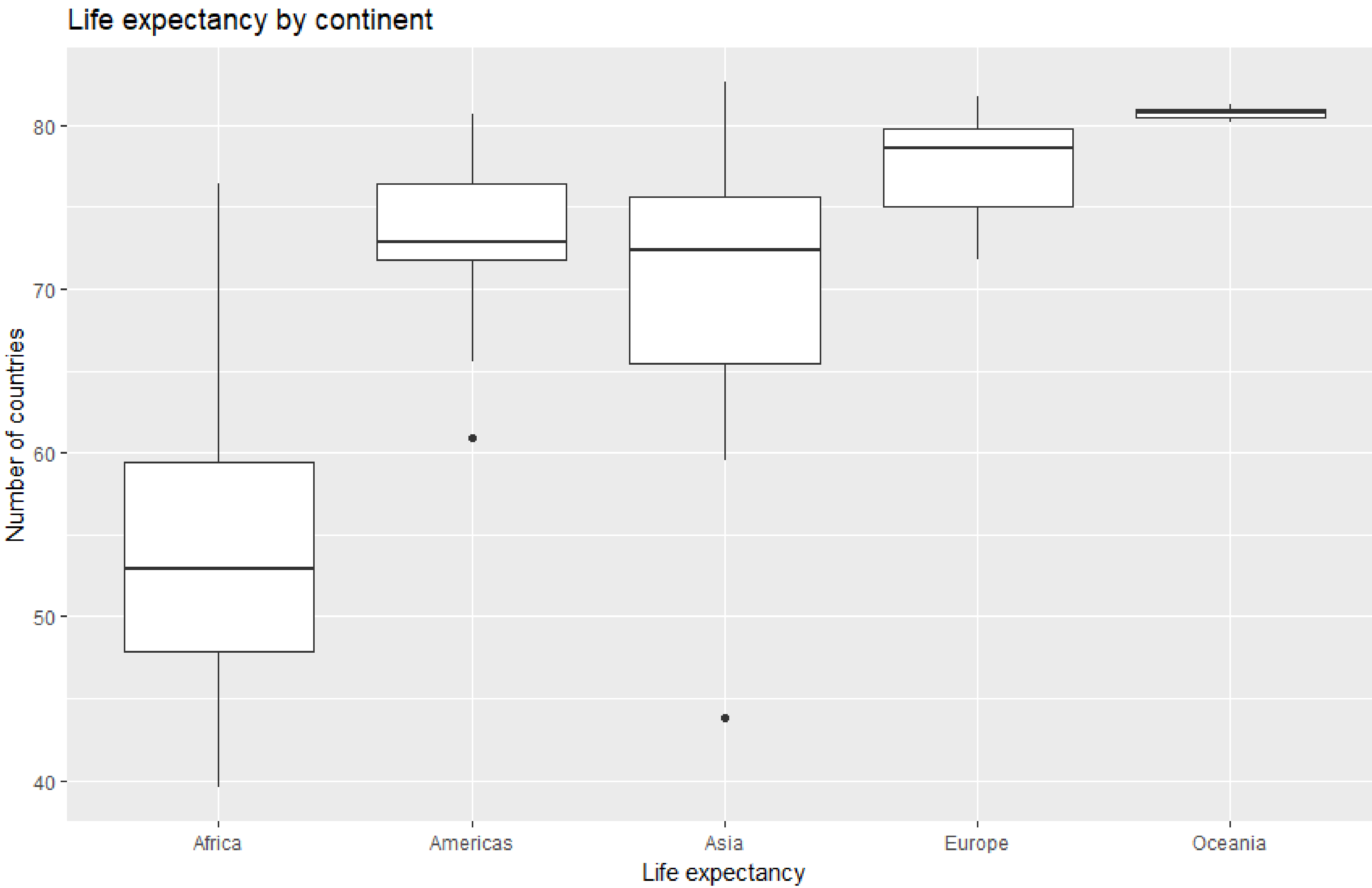
- Sanity check of model fit. It makes more sense for the intercept to correspond to life expectancy in 1952, the earliest date in our dataset, rather than year 0.
- Find the minimum year in the dataset and rerun the regression

```
> yearMin <- min(gapminder$year)
>
> tempModel2 <- lm(lifeExp ~ I(year - yearMin), data=tempDa
> summary(tempModel2)

Call:
lm(formula = lifeExp ~ I(year - yearMin), data = tempData)

Residuals:
   Min      1Q Median      3Q     Max
-1.207 -0.543  0.143  0.457  1.119

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         67.067      0.423   158.7  < 2e-16
I(year - yearMin)    0.242      0.013    18.6  4.3e-09

Residual standard error: 0.778 on 10 degrees of freedom
Multiple R-squared:  0.972,     Adjusted R-squared:  0.969
F-statistic:  347 on 1 and 10 DF,  p-value: 4.32e-09
```



67 (value of intercept) was the life expectancy in 1952

Life expectancy in 2007



Regression with categorical variables (1/2)

Summary statistics on life expectancy by continent

```
> favstats(~lifeExp | continent, data=gapminder2007)
  continent   min    Q1 median    Q3   max  mean    sd  n missing
1    Africa 39.61 47.83  52.93 59.44 76.44 54.81 9.631 52       0
2  Americas 60.92 71.75  72.90 76.38 80.65 73.61 4.441 25       0
3      Asia 43.83 65.48  72.40 75.64 82.60 70.73 7.964 33       0
4    Europe 71.78 75.03  78.61 79.81 81.76 77.65 2.980 30       0
5   Oceania 80.20 80.46  80.72 80.98 81.23 80.72 0.729  2       0
```

Could we use the categorical variable **continent** as an explanatory variable in regression?

```
> lifeExp_model1 <- lm(lifeExp ~ continent, data = gapminder2007)
> msummary(lifeExp_model1)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         54.81       1.03   53.45  < 2e-16
continentAmericas   18.80       1.80   10.45  < 2e-16
continentAsia       15.92       1.65    9.67  < 2e-16
continentEurope     22.84       1.70   13.47  < 2e-16
continentOceania    25.91       5.33    4.86  3.1e-06

Residual standard error: 7.39 on 137 degrees of freedom
Multiple R-squared:  0.635,	Adjusted R-squared:  0.625
F-statistic: 59.7 on 4 and 137 DF,  p-value: <2e-16
```

When a categorical variable has k levels, we include (k-1) in the regression model and the one left outside acts as our baseline (or zero). In this example, **continent** has 5 levels, but only 4 continents are included in the model. We have left out **continentAfrica** which will be our baseline.

The intercept of our model is 54.81– the mean life expectancy for Africa.

The slope of **continentAmericas** is 18.80– people in Americas live on average 18.80 years longer than the baseline continent of Africa.

- Is this what the summary stats tells us, too?

Regression with categorical variables (2/2)

We can run another model where we include year and continent as explanatory variables

```
> lifeExp_model2 <- lm(lifeExp ~ continent + I(year - yearMin), data = gapminder)
> msummary(lifeExp_model2)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        39.9030     0.4068    98.1   <2e-16
continentAmericas  15.7934     0.5140    30.7   <2e-16
continentAsia      11.1996     0.4700    23.8   <2e-16
continentEurope    23.0384     0.4842    47.6   <2e-16
continentOceania   25.4609     1.5218    16.7   <2e-16
I(year - yearMin)   0.3259     0.0103    31.7   <2e-16

Residual standard error: 7.32 on 1698 degrees of freedom
Multiple R-squared:  0.68,	Adjusted R-squared:  0.679
F-statistic:  722 on 5 and 1698 DF,  p-value: <2e-16
```

- What is the meaning of the slope for **I(year-yearMin)**?
- What is the meaning of the slope for **contintentEurope**?
- Are all variables significant?

Overview: correlation and regression

- Use scatterplots to examine data
  - identify possible patterns (and non-linearities!)
- Measure strength of linear relationship by correlation
  - Correlation is always lies between +/- 1
- Model relationship using regression: **$Y = b_0 + b_1*X_1 + b_2*X_2 + \textit{error}$**
  - Intercept and slope are fitted so as to minimise the average squared error
- Regression diagnostics
  - check t-values (or p-values) of coefficients, leave out insignificant ones (with absolute t-value <2 or p-value > 5%)
  - R^2 measures "percentage of variance" which is explained by the model
  - regression relies on the assumption that errors are uncorrelated
  - look out for: influential observations, outliers, non-linear relationships

Session Summary

We covered

- Correlation and regression
- Building regression models
    - Check whether the effect (estimated slope) of an explanatory variable is different from zero
    - 95% interval for the effect of explanatory variables X_1, X_2 etc

    - What proportion of the overall variability does our model explain
    - What is the regression residual SE?

- ***Readings:*** ModernDive chapters 5-6