Lecture 4

- Assignment 1
- Regression modelling
- Extending the basic model
- Regression models in R, diagnostics

Presenter: Dr John Betts

Week-by-week outline

Week Starting	Seminar	Topic	App Ses	A1	A2	Q/P	A3	Due Date
2/3/2026	1	Introduction to Data Science, R, review of basic statistics	-					
9/3/2026	2	Data visualisation	S1					
16/3/2026	3	Data manipulation	S2					
23/3/2026	4	Regression modelling	S3					
30/3/2026	5	Clustering	S4					
6/4/2026	-	Mid-semester Break						
13/4/2026	6	Classification using decision trees	S5					17/4/2026
20/4/2026	7	Improving and evaluating classifiers. Naïve Bayes classification	S6					
27/4/2026	8	Ensemble methods, Artificial Neural Networks	S7					
4/5/2026	9	Network analysis	S8					
11/5/2026	10	Introduction to text analysis	S9					15/5/2026
18/5/2026	11	Text analysis applications	Quiz/Prac					22/5/2026
25/5/2026	12	Text Network Analysis, Review of the unit, Assignment 3	S10,11,12					
1/6/2026		SWOT VAC	-					
8/6/2026		EXAM PERIOD	-					12/6/2026

Assessment details

Assignment 1, Due 17th April, Weighting 25%

- Covers data manipulation, visualisation, and data analysis using a variety of techniques. Submission is a written report and short video explaining the key findings of your research.

Assignment 2, Due 15th May, Weighting 20%

- Covers machine learning/artificial intelligence models using R. Submission is a written report and short video.

Assignment 3, Due 12th June, Weighting 25%

- Covers text analysis, networks and clustering using R. Submission is a written report and short video.

Quiz + Practical Activity, Week 11 (Due 22nd May), Weighting 30%

- You will do practical activities and quiz style questions under supervision during your applied session. Content will cover topics from Weeks 1 – 9.

Consultations

Clayton students: see additional information and resources, under the “Learning” tile.

https://learning.monash.edu/course/view.php?id=41077&section=5

Clayton Week 5 A/S replacement

For Clayton students: Applied Session 02, Good Friday class has been replaced by online session, Monday 30th 12:00 – 2:00pm.

Details will be posted on Moodle under the Week 4 Real-time tile.

Assignment 1

Assignment 1

MONASH University

Faculty of Information Technology

FIT3152 Data analytics – 2026: Assignment 1

Your task	• Analyse the country level predictors of confidence in **social organisations and how these change over time** using data from the World Values Survey. • This is an individual assignment.
Value	• This assignment is worth **25%** of your total marks for the unit. • It has **40** marks in total.
Suggested Length	• 8 – 10 A4 pages, approximately 2,000 words (for your report) + extra pages as appendix for your R script and report on how Generative AI used, if required. • Font size 11 or 12pt, single spacing.

Assignment 1

Due Date	**11.55pm Friday 17th April 2026**
Submission	• Submit a single PDF file **and** single video file on Moodle. • Note that submission of a video report is a hurdle requirement. • Use the naming convention: *FirstnameSecondnameID.{pdf, mp4, mov etc.}* • Turnitin will be used for similarity checking of all written submissions.
Generative AI Use	• In this assessment, you can use generative artificial intelligence (AI) in order to search for R functions and examples to perform tasks that you specify only. Any use of generative AI must be appropriately acknowledged (see Learn HQ).
Late Penalties	• 5% (**2** mark) deduction per calendar day for up to one week. • Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.

Assignment 1

Instructions

Address each of the research questions below and report the results of your analysis and your interpretation of those results. Report any assumptions you've made in your analysis. Include your R code as an appendix. Your R code must be machine readable text as the university requires all student submissions to be processed by plagiarism detection software. See information on Generative AI below.

There are two options for compiling your written report:
(1) You can create your report using any word processor with your R code pasted in as machine-readable text as an appendix, and save as a pdf, or
(2) As an R Markdown document that contains the R code with results and discussion interleaved. Render this as a HTML file and save as a pdf.

Your video report should be less than 100MB in size. You may need to reduce the resolution of your original recording to achieve this. Use a standard file format such as .mp4, or mov for submission.

Assignment 1

Software

It is expected that you will use R for your data analysis, graphics and tables. You are free to use any R packages you need but must document these in your report and include in your R code.

Use of Generative AI

AI & Generative AI tools may be used in GUIDED ways within this assessment/task as per the guidelines provided.

In this assessment, you can use generative artificial intelligence (AI) in order to search for R functions and examples to perform tasks that you specify only. Any use of generative AI must be appropriately acknowledged (see Learn HQ).

If you do use Generative AI for your assignment, then you must include the statement "Generative AI was used in this assignment." in the introductory/first paragraph of your report. You must also include the following information as an appendix in your report: (1) the technology you used (e.g. ChatGPT), (2) the information that was generated (e.g. R code fragments), (3) the prompts used (i.e. the questions you asked), and (4) how the output was used in your work.

If you did not use generative AI in your assignment, then include the statement "Generative AI was not used in this assignment." in the introductory/first paragraph of your report.

Assignment 1

Questions

The World Values Survey (WVS) is an international research program that studies the social, political, economic, religious and cultural attitudes and values of people around the world. You can read more here: https://www.worldvaluessurvey.org/WVSContents.jsp.

For this assignment you will analyse data collected over Waves 1 - 7, from 1981 to 2022. The aim of this assignment is to understand country-level differences in participant responses and the predictors of confidence in social organisations, and how these responses and predictors of confidence have changed over time.

Social organisations include aspects of society such as religion, armed forces, the press, television, trade unions, police, the courts, government, banks, and international and environmental organisations etc. They are indicated in your data by column names having the prefix "C". Predictor variables (**attributes**) include personal information such as age and gender, happiness indicators, attitudes and values towards others, political and social views and participation.

Each student will be assigned a **different** subset of organisations and attributes to study. Your task is to analyse **all** the survey data assigned to you, with a **focus** on the country you have been allocated.

Assignment 1

1. Descriptive analysis. (5 Marks)

(a) Describe the data overall, including things such as dimension, data types, distribution of numerical responses, variety of non-numerical (text) responses, missing values, and anything else of interest or relevance.

Assignment 1

2. Focus country vs all other countries as a group (independent of time). (13 Marks)

For Question 2 ignore the effect of time. That is, do not separate your data by years or waves when answering the questions below.

(a) Identify your focus country from the accompanying list (**WVSFocusCountry.pdf**). How do participant responses for your focus country differ from the other countries in the survey (treating them as a group)?

(b) How well do participant responses (attributes) predict confidence in social organisations in your **focus country**? Which attributes seem to be the best predictors? Confidence in which social organisations can be more reliably predicted? Explain your reasoning.

(c) Repeat Question 2(b) for the **other countries** as a group. Which attributes are the strongest predictors? Confidence in which social organisations can be more reliably predicted? How do these results compare to those of your focus country?

Assignment 1

3. Focus country vs all other countries as a group (over time). (12 Marks)

For Question 3 study the effect of time by separating your data by years or waves when answering the questions below.

(a) How do participant responses for your **focus country** vary over time (using either years or successive waves)? Describe these changes over time and comment on whether they are significant or not. Perform the same analysis for the **other countries** (as a group) and compare the results with your focus country. What are the major differences between the two groups? Create a graphic enabling a reader to compare results (focus *vs* other countries) over time, for the most interesting results. Describe your reasoning for the design of the graphic.

(b) How does the ability of participant responses (attributes) to predict confidence in social organisations in your **focus country** change over time? Do the important attributes for predicting confidence change over time? Perform the same analysis for the **other countries** (as a group) and compare the results. What are the major differences between the two groups? Create a graphic enabling a reader to compare results (focus *vs* other countries) over time, for the strongest predictors. Describe your reasoning for the design of the graphic.

Assignment 1

4. Video Presentation: (Submission Hurdle and 4 Marks)

Record a short presentation using your smartphone, Zoom, or similar method. Your presentation should be approximately 5 minutes in length and summarise your main findings for Sections 1 – 3, as well as describing how you conducted your research, any assumptions made, and how you designed your graphics.

5 Overall considerations (6 Marks)

This includes: the quality and clarity of your reasoning and assumptions; the strength of support for your findings; the quality of your writing in general and communication of results;-the quality of your graphics throughout; the quality of your R coding.

Assignment 1

Data

The data for this assignment is a reduced version of the World Values Survey Waves 1 -7 data. The filename is "WVSExtract.csv". The data includes ordinal data coded on a numerical scale. For this assignment assume it is reasonable to treat these responses as numerical.

Create your individual data as follows:

```
rm(list = ls())
set.seed(12345678) # Your Student Number
VCData = read.csv("WVSExtract.csv")
VC = VCData[sample(1:nrow(VCData),100000, replace=FALSE),]
VC = VC[,c(1:3,sort(sample(4:50,25,replace=FALSE)),
sort(sample(51:65,8,replace=FALSE)))]
#write.csv(VC, "FIT3152A1Data_YourName.csv", row.names = FALSE)
```

You can save the “VC” file you created by uncommenting the last line above. You can then delete the WVSExtract.csv file you downloaded.

Locate your focus country using the accompanying document FocusCountryByID.pdf. A list matching country names with three letter code is in WVSCountryCodes.pdf.

Assignment 1

Data fields and brief descriptor

Most fields are on integer scales over varying range. The convention is that larger numbers generally indicate greater agreement with statement or frequency of occurrence. Some exceptions given below. Fields in bold indicate confidence in social organisations.

You can access more detail on each field in your data from the *WVS-7 Master Questionnaire 2017-2020 English.pdf*, linked from https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp. Use the question ID given in the **WVS Wave 7 Reference** in the table below.

Column Name	**Original Descriptor**	**WVS Wave 7 Reference**
Wave	Chronology of EVS-WVS waves	A_WAVE
Country	ISO 3166-1 alpha-3 country code	B_COUNTRY_ALPHA
Year	Year survey	A_YEAR
ILFam	Important in life: Family	Q1
ILFriends	Important in life: Friends	Q2
ILLeisure	Important in life: Leisure time	Q3
ILPolitics	Important in life: Politics	Q4
ILWork	Important in life: Work	Q5

Assignment 1

CChurches	Confidence: Churches	Q64
CArmedForces	Confidence: Armed Forces	Q65
CPress	Confidence: The Press	Q66
CUnions	Confidence: Labour Unions	Q68
CPolice	Confidence: The Police	Q69
CParliament	Confidence: Parliament	Q73
CCivilService	Confidence: The Civil Services	Q74
CTelevision	Confidence: Television	Q67
CGovernment	Confidence: The Government	Q71
CPolParties	Confidence: The Political Parties	Q72
CMajComp	Confidence: Major Companies	Q77
CEnvProt	Confidence: The Environmental Protection Movement	Q79
CWomensMvt	Confidence: The Womens' Movement	Q80
CCourts	Confidence: Justice System/Courts	Q70
CEU	Confidence: The European Union	Q82_EU

Assignment 1

JB EDA V00.R | VC

Filter

	Wave	Country	Year	ILFam	ILWork	ICQIndependence	ICQHardWork	ICQTolerance	ICQDetermination	ICQFaith	ACTReligion	ACTSport	ACTUnions
149999	5	JOR	2007	1	1	1	1	0	0	1	0	0	0
57968	5	UKR	2006	1	2	1	0	1	0	1	2	1	0
34203	7	AUS	2018	1	2	1	1	1	1	0	1	0	0
128452	6	BLR	2011	1	1	0	1	1	1	0	0	0	0
196160	7	NLD	2022	1	2	1	1	1	0	0	1	2	0
54028	7	USA	2017	2	1	0	1	1	1	0	0	0	0
45965	5	HUN	2009	2	2	1	0	1	1	0	0	0	0
102378	3	GEO	1996	1	1	1	1	1	0	0	0	0	0
84273	4	SGP	2002	1	3	1	0	1	0	1	-4	-4	-4
92075	5	FIN	2005	1	1	1	0	1	1	0	2	2	0
180253	6	MEX	2012	1	1	1	0	1	1	1	0	0	0
111049	3	NZL	1998	1	2	0	0	1	1	0	1	2	0
135207	5	GBR	2005	1	1	1	1	1	1	0	2	1	0
146977	7	MYS	2018	1	1	0	0	1	0	1	2	0	2
59242	7	CHN	2018	1	1	0	1	1	0	0	0	0	0
149092	5	CHE	2007	1	1	0	1	0	0	0	0	0	0
39163	4	ZAF	2001	1	1	0	0	1	0	1	-4	-4	-4
79507	7	GBR	2022	1	1	0	0	1	0	0	1	0	0
107645	3	SVK	1998	1	1	0	1	0	0	1	2	0	0
115588	7	ROU	2018	1	1	0	0	0	1	0	0	0	0
180425	6	ARM	2011	1	1	0	1	1	0	0	0	0	0
84826	2	NGA	1990	1	1	0	1	0	0	1	-4	-4	-4
53654	5	USA	2006	1	1	0	1	1	0	1	0	0	1
65630	5	UKR	2006	1	2	0	1	1	0	0	0	0	0
139335	5	GEO	2009	1	1	1	1	0	0	1	0	0	1
160800	6	MAR	2011	2	2	0	1	0	0	0	0	0	0
96936	3	FIN	1996	1	1	1	0	0	0	0	2	0	0

Assignment 1

Assignment 1 Notes

- Students who joined the unit late (and are not on the FocusCountryByID.pdf) need to email john.betts@monash.edu to be assigned a focus country.
- Data may contain missing/NA values. Check the survey documentation: WVS-7 Master Questionnaire 2017-2020 English.pdf
- It is likely many attributes will have low predictive power. The aim of the analysis is to find the "best" ones.

Assignment 1 Notes

Examples of bad summaries in the assignment report.

Data Types: ordinal data coded on a numerical scale and hence treated as numerical data
Distribution of Numerical Attributes: summary statistics of data, (mean, median, mode)
Variety of Non-Numerical (text) Attributes: discuss data that is not numbers
Missing Values: sum of missing values, number of missing values in each column

```
> str(cvbase)
'data.frame':     40000 obs. of  52 variables:
 $ employstatus_1     : int  NA NA NA NA 1 NA NA NA NA NA ...
 $ employstatus_2     : int  NA NA NA 1 NA 1 1 1 NA NA ...
 $ employstatus_3     : int  NA NA NA NA NA NA NA NA 1 1 ...
 $ employstatus_4     : int  NA NA NA NA 1 NA NA NA NA NA ...
 $ employstatus_5     : int  NA NA NA NA NA NA NA NA NA NA ...
 $ employstatus_6     : int  NA NA NA NA NA NA NA NA NA NA ...
 $ employstatus_7     : int  1 1 NA NA NA NA NA NA NA NA ...
 $ employstatus_8     : int  NA NA NA NA NA NA NA NA NA NA ...
 $ employstatus_9     : int  NA NA 1 NA 1 NA NA NA NA NA ...
 $ employstatus_10    : int  NA NA NA NA NA NA NA NA NA NA ...
 $ isoFriends_inPerson: int  0 0 0 0 1 5 0 4 2 1 ...
 $ isoOthPpl_inPerson : int  7 0 0 7 1 6 2 7 7 7 ...
 $ isoFriends_online  : int  0 5 6 7 7 6 7 7 0 2 ...
 $ isoOthPpl_online   : int  0 7 0 7 0 6 0 7 1 3 ...
 $ lone01             : int  1 2 3 3 3 4 1 4 2 2 ...
 $ lone02             : int  1 2 4 3 2 4 2 3 4 2 ...
 $ lone03             : int  1 2 1 3 2 5 1 3 2 2 ...
 $ happy              : int  6 10 6 3 6 9 6 5 6 4 ...
 $ lifeSat            : int  4 6 3 2 3 5 5 2 4 3 ...
 $ MLQ                : int  0 2 -3 1 1 2 2 -3 0 3 ...
 $ bor01              : int  0 1 3 0 0 1 3 -3 -1 -1 ...
 $ bor02              : int  -2 2 3 0 0 2 1 0 1 0 ...
 $ bor03              : int  0 -1 -1 0 -1 2 -3 -1 0 0 ...
 $ consp01            : int  5 3 10 5 7 8 8 10 7 5 ...
 $ consp02            : int  5 4 10 3 10 9 6 10 8 5 ...
 $ consp03            : int  5 3 6 3 5 8 5 10 5 5 ...
 $ rankOrdLife_1      : chr  "E" NA "E" "F" ...
 $ rankOrdLife_2      : chr  "A" NA "D" "B" ...
 $ rankOrdLife_3      : chr  "F" NA "F" "D" ...
 $ rankOrdLife_4      : chr  "B" NA "B" "A" ...
```

```
##  employstatus_9  employstatus_10 isoFriends_inPerson isoOthPpl_inPerson
##  Min.   :1       Min.   :1       Min.   :0.00        Min.   :0.000
##  1st Qu.:1       1st Qu.:1       1st Qu.:0.00        1st Qu.:0.000
##  Median :1       Median :1       Median :1.00        Median :1.000
##  Mean   :1       Mean   :1       Mean   :2.07        Mean   :1.955
##  3rd Qu.:1       3rd Qu.:1       3rd Qu.:4.00        3rd Qu.:3.000
##  Max.   :1       Max.   :1       Max.   :7.00        Max.   :7.000
##  NA's   :46138   NA's   :56601   NA's   :468         NA's   :749
##  isoFriends_online isoOthPpl_online     lone01          lone02
##  Min.   :0.000     Min.   :0.000    Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000     1st Qu.:0.000    1st Qu.:1.000   1st Qu.:2.000
##  Median :5.000     Median :2.000    Median :2.000   Median :3.000
##  Mean   :4.399     Mean   :2.855    Mean   :2.419   Mean   :2.665
##  3rd Qu.:7.000     3rd Qu.:5.000    3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   :7.000     Max.   :7.000    Max.   :5.000   Max.   :5.000
##  NA's   :1358      NA's   :1667     NA's   :120     NA's   :173
##      lone03          happy           lifeSat           MLQ
##  Min.   :1.000   Min.   : 1.000   Min.   :1.000   Min.   :-3.0000
##  1st Qu.:1.000   1st Qu.: 5.000   1st Qu.:3.000   1st Qu.: 0.0000
##  Median :2.000   Median : 7.000   Median :4.000   Median : 1.0000
##  Mean   :2.079   Mean   : 6.333   Mean   :4.138   Mean   : 0.8439
##  3rd Qu.:3.000   3rd Qu.: 8.000   3rd Qu.:5.000   3rd Qu.: 2.0000
##  Max.   :5.000   Max.   :10.000   Max.   :6.000   Max.   : 3.0000
##  NA's   :198     NA's   :732      NA's   :161     NA's   :167
##      bor01             bor02             bor03            consp01
##  Min.   :-3.0000   Min.   :-3.00000   Min.   :-3.0000   Min.   : 0.000
##  1st Qu.:-1.0000   1st Qu.:-2.00000   1st Qu.:-1.0000   1st Qu.: 5.000
##  Median : 0.0000   Median : 0.00000   Median : 0.0000   Median : 7.000
##  Mean   : 0.3271   Mean   : 0.04387   Mean   : 0.3101   Mean   : 6.835
##  3rd Qu.: 2.0000   3rd Qu.: 2.00000   3rd Qu.: 2.0000   3rd Qu.: 9.000
##  Max.   : 3.0000   Max.   : 3.00000   Max.   : 3.0000   Max.   :10.000
##  NA's   :226       NA's   :244        NA's   :249       NA's   :2187
##     consp02          consp03        c19perBeh01     c19perBeh02
##  Min.   : 0.000   Min.   : 0.000   Min.   :-3.00   Min.   :-3.000
##  1st Qu.: 5.000   1st Qu.: 4.000   1st Qu.: 2.00   1st Qu.: 2.000
##  Median : 8.000   Median : 5.000   Median : 3.00   Median : 3.000
##  Mean   : 7.151   Mean   : 5.585   Mean   : 2.31   Mean   : 2.426
##  3rd Qu.: 9.000   3rd Qu.: 8.000   3rd Qu.: 3.00   3rd Qu.: 3.000
##  Max.   :10.000   Max.   :10.000   Max.   : 3.00   Max.   : 3.000
```

Regression

COVID-19

https://www.nature.com/articles/s41598-020-68862-x

COVID-19

A question central to the Covid-19 pandemic is why the Covid-19 mortality rate varies so greatly across countries. This study aims to investigate factors associated with cross-country variation in Covid-19 mortality. Covid-19 mortality rate was calculated as number of deaths per 100 Covid-19 cases. To identify factors associated with Covid-19 mortality rate, linear regressions were applied to a cross-sectional dataset comprising 169 countries. We retrieved data from the Worldometer website, the Worldwide Governance Indicators, World Development Indicators, and Logistics Performance Indicators databases. Covid-19 mortality rate was negatively associated with Covid-19 test number per 100 people (RR = 0.92, *P* = 0.001), government effectiveness score (RR = 0.96, *P* = 0.017), and number of hospital beds (RR = 0.85, *P* < 0.001). Covid-19 mortality rate was positively associated with proportion of population aged 65 or older (RR = 1.12, *P* < 0.001) and transport infrastructure quality score (RR = 1.08, *P* = 0.002). Furthermore, the negative association between Covid-19 mortality and test number was stronger among low-income countries and countries with lower government effectiveness scores, younger populations and fewer hospital beds. Predicted mortality rates were highly associated with observed mortality rates (r = 0.77; *P* < 0.001). Increasing Covid-19 testing, improving government effectiveness and increasing hospital beds may have the potential to attenuate Covid-19 mortality.

https://www.nature.com/articles/s41598-020-68862-x

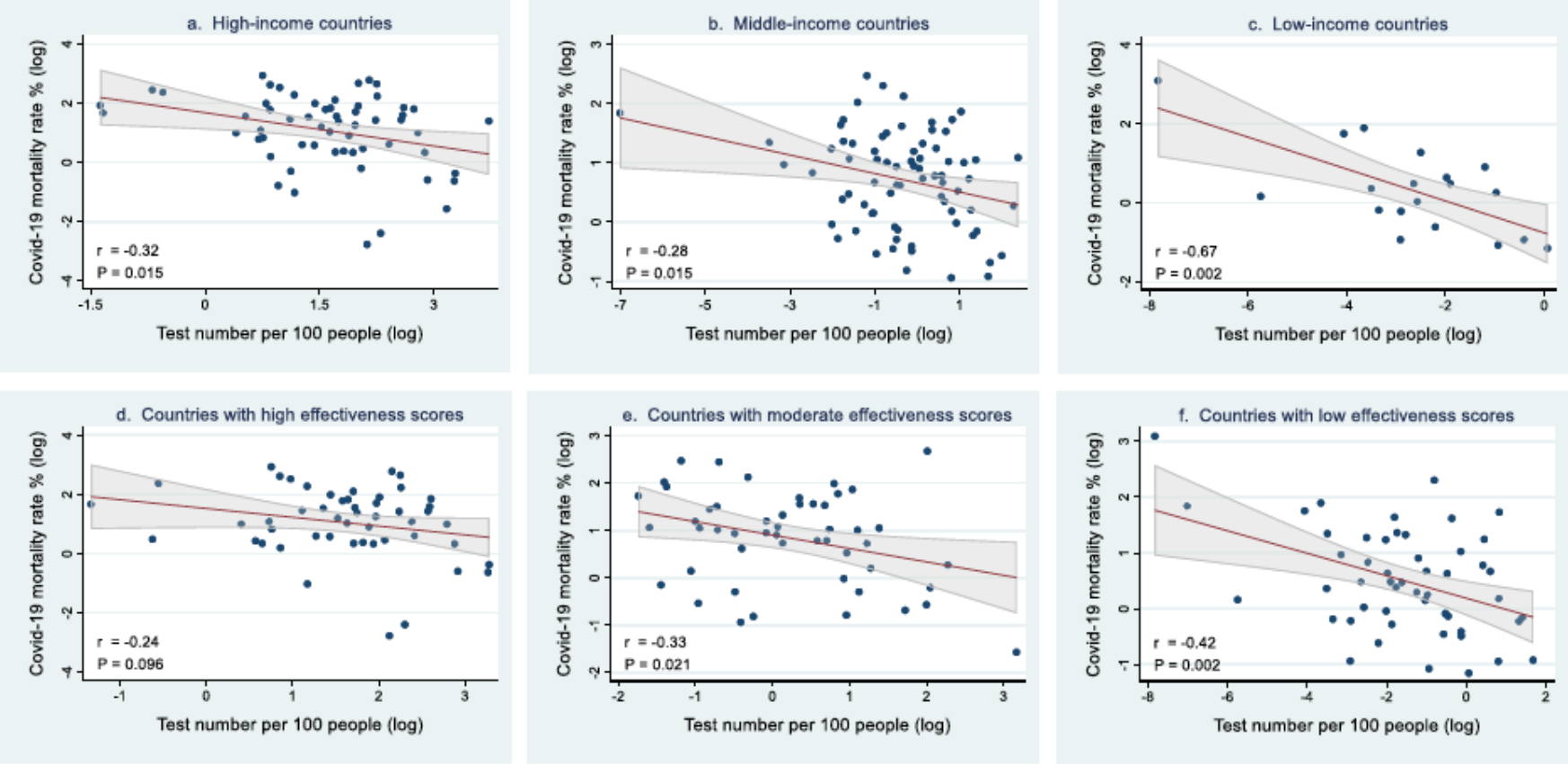
COVID-19

	N	Mean	SE	95% CI
Covid-19 mortality rate (%)	169	3.70	0.28	3.15–4.25
Covid-19 related factors				
Test number per 100 people	153	3.75	0.47	2.82–4.69
Case number per 1,000 people	169	1.69	0.25	1.20–2.18
Critical case rate (%)[a]	120	0.56	0.06	0.44–0.68
Country related factors				
Government effectiveness score[b]	167	−0.01	0.08	−0.17–0.16
Population aged 65 or older (%)	162	9.17	0.51	8.15–10.18
Bed number per 1,000 people	146	3.14	0.22	2.72–3.57
Communicable disease death rate (%)	159	31.04	1.79	27.50–34.58
Transport infrastructure quality score[c]	153	2.75	0.05	2.64–2.86

Table 1. Descriptive statistics of model variables. [a]Critical case rate = number of critical cases/total number of cases. [b]Range of data: from −2.5 (worst) to 2.5 (best). [c]Range of data: from 1 (worst) to 5 (best).

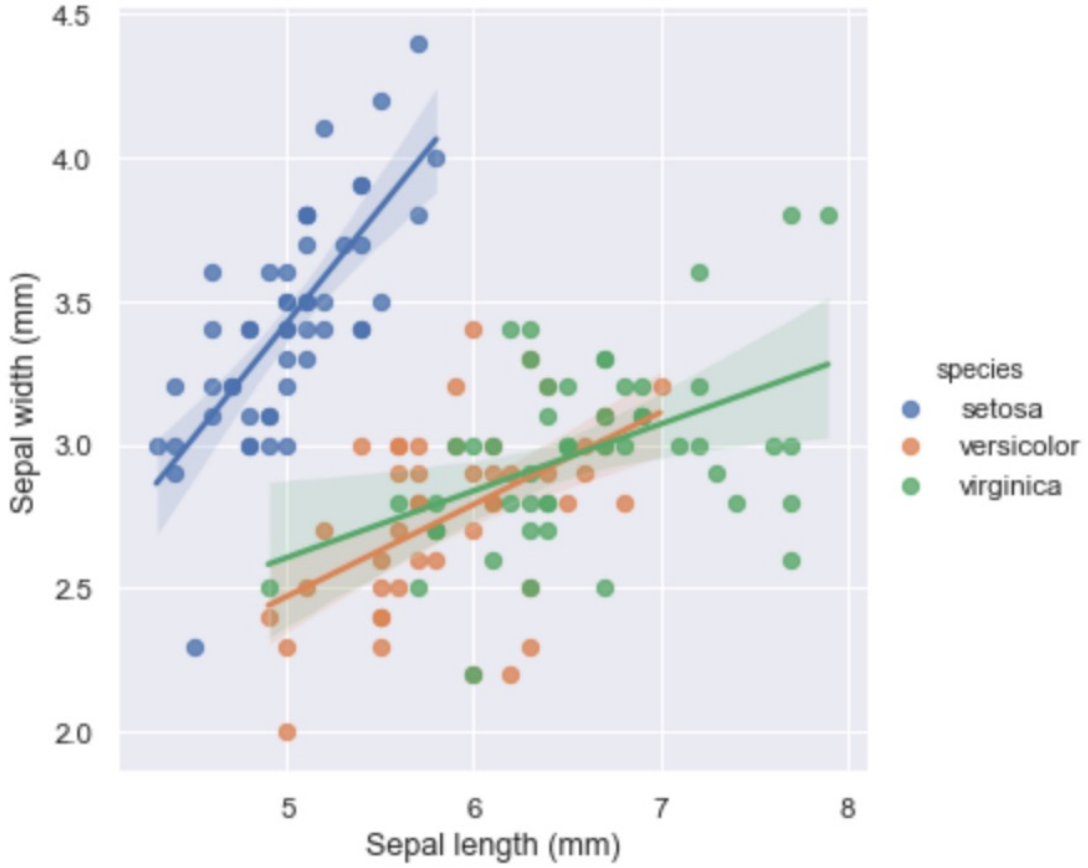
https://www.nature.com/articles/s41598-020-68862-x

COVID-19



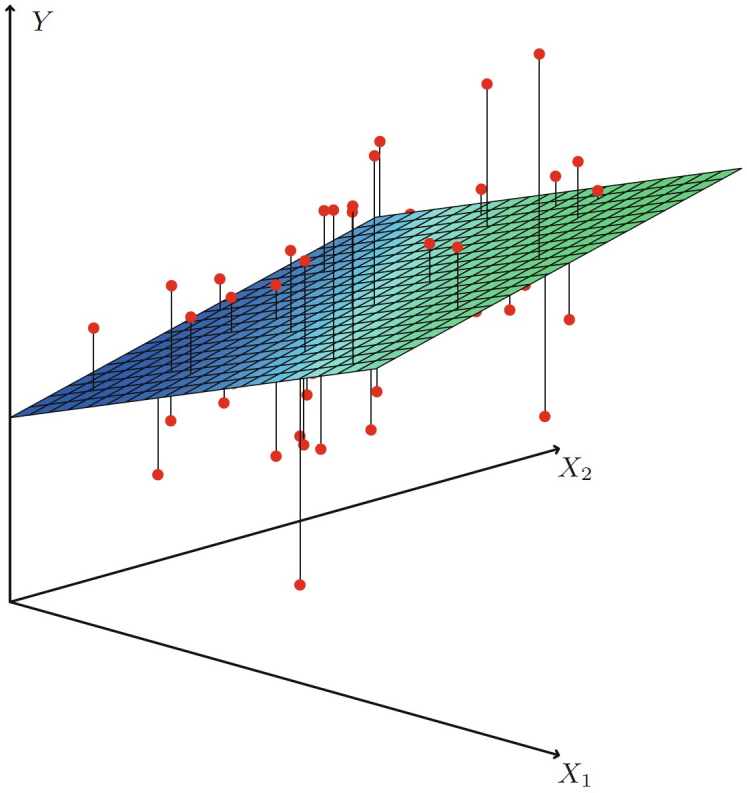
https://www.nature.com/articles/s41598-020-68862-x

Iris sepals: width *vs* length by species



https://hackernoon.com/types-of-linear-regression-w4o227s5

Multiple linear regression



From: G. James et al., An Introduction to Statistical Learning: with Applications in R (2021).

Regression

Regression models the relationship between two or more variables, from which we can:

- Observe the effect of independent variables (inputs) on the dependent variable (output),
- Predict the values for new data (e.g., forecasting),
- Determine the relative importance of variables the model,
- Linear regression assumes a straight-line relationship, but many other relationships can be modelled.

Regression

- Fitting a regression model is a form of supervised learning – that is, the model is 'learned' from data consisting of known inputs and outputs.

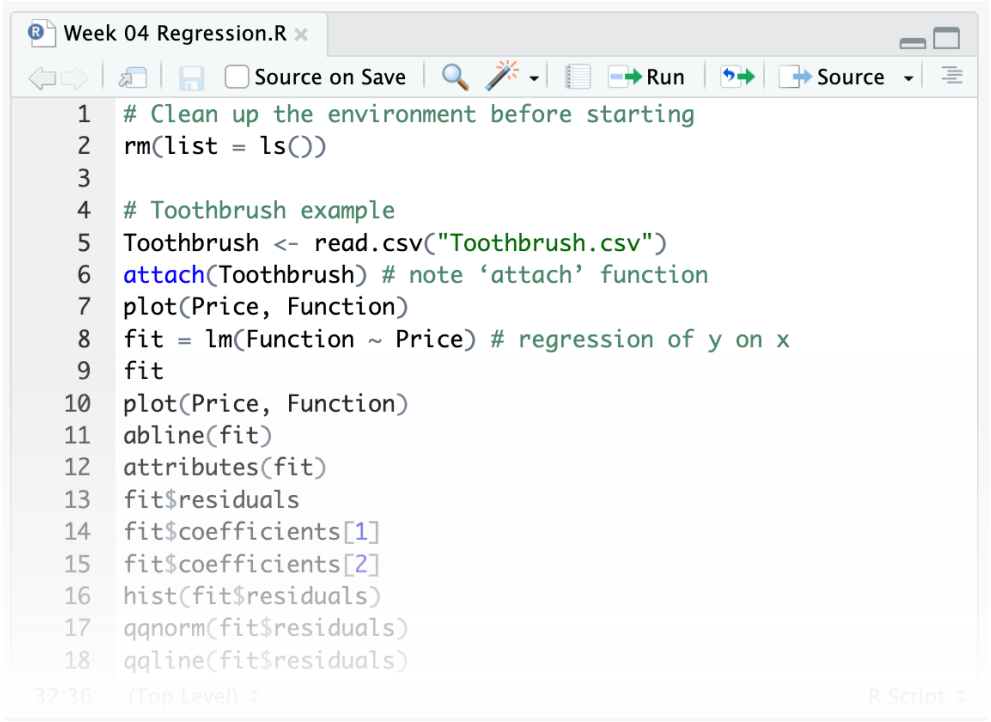
- The learned model can then be applied to unknown cases, this includes forecasting.

- *r-squared* and other regression diagnostics tell us the degree to which variability in the response is explained by the input variables...

Linear regression

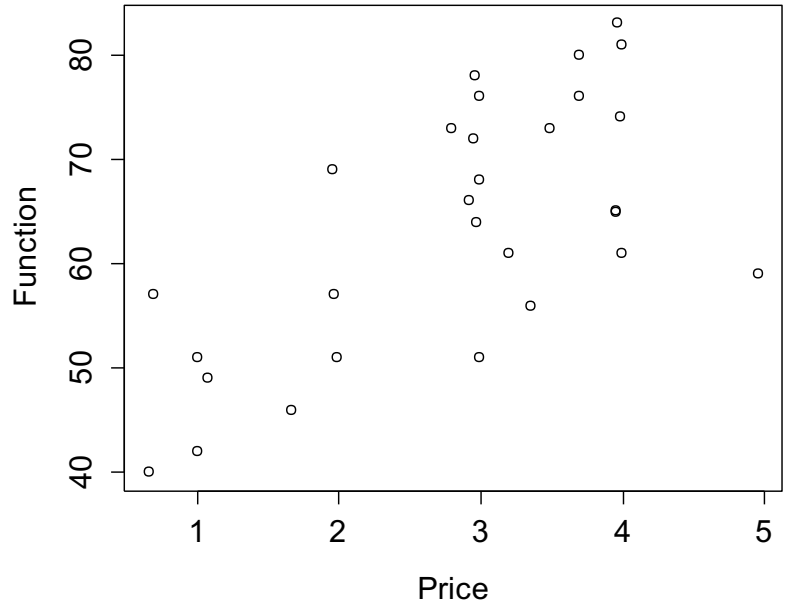
See R Script of lecture examples

> Week 04 Regression.R



Recall: Toothbrush – function *vs* price

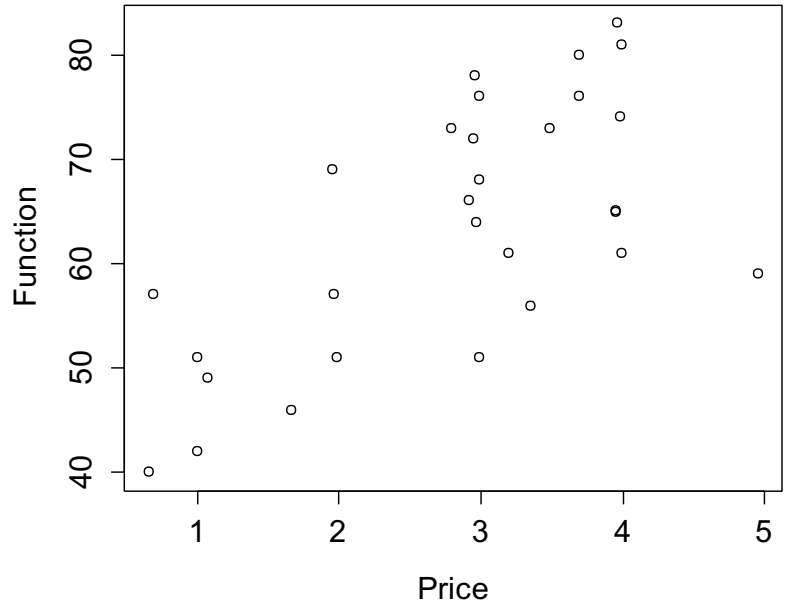
```
> Toothbrush <- read.csv("Toothbrush.csv")
> attach(Toothbrush) # note 'attach' function
> plot(Price, Function)
```



Linear regression – purpose

Tells the following:

- The linear relationship between Function and Price?
- The strength of the relationship (predictability).



Linear regression – assumptions

Simple least squares regression assumes that

- The relationship approximately linear, which is of the form: $y \approx ax + b$
- x and y are numerical variables, not categories for example.
- a and b are calculated to minimise the squared error between the observed values (the data) and the *fitted values* (i.e., those predicted by the model).
- Errors are (approximately) normally distributed.

Fitting the (linear model)

The lm() function performs a least squares regression and creates a linear model object:

```
> fit = lm(Function ~ Price) # regression of y on x
> fit
Call:
lm(formula = Function ~ Price)
Coefficients:
(Intercept)        Price
     44.020        6.942
```

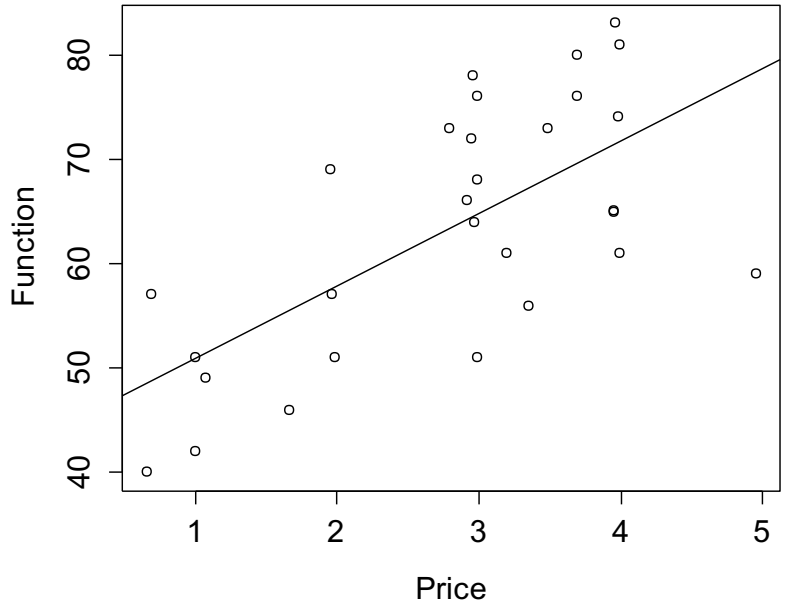
However, the linear model object contains much more information than just the coefficients!

Line of best fit

This has been covered but worth remembering

> plot(Price, Function)

> abline(fit) # Intercept and gradient are read directly from “fit”



Linear model object

To see the details of what the object contains use:

```
> attributes(fit)   # to see contents of an object
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"
```

- Thus, fields can be addressed by name or index. For example:

```
> fit$residuals   # to access elements by "column"
...
```

Linear model object

More details in the Environment inspector:

```
fit                    List of 12
  coefficients : Named num [1:2] 44.02 6.94
  ..- attr(*, "names")= chr [1:2] "(Intercept)" "Price"
  residuals : Named num [1:29] -6.34 13.43 7.5 -8.6 8.19 ...
  ..- attr(*, "names")= chr [1:29] "1" "2" "3" "4" ...
  effects : Named num [1:29] -342.44 42.45 8.39 -13.09 3.77 ...
  ..- attr(*, "names")= chr [1:29] "(Intercept)" "Price" "" "" ...
  rank : int 2
  fitted.values: Named num [1:29] 71.4 64.6 64.5 48.6 48.8 ...
  ..- attr(*, "names")= chr [1:29] "1" "2" "3" "4" ...
  assign : int [1:2] 0 1
  qr :List of 5
  ..$ qr : num [1:29, 1:2] -5.385 0.186 0.186 0.186 0.186 ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:29] "1" "2" "3" "4" ...
```

Addressing coefficients

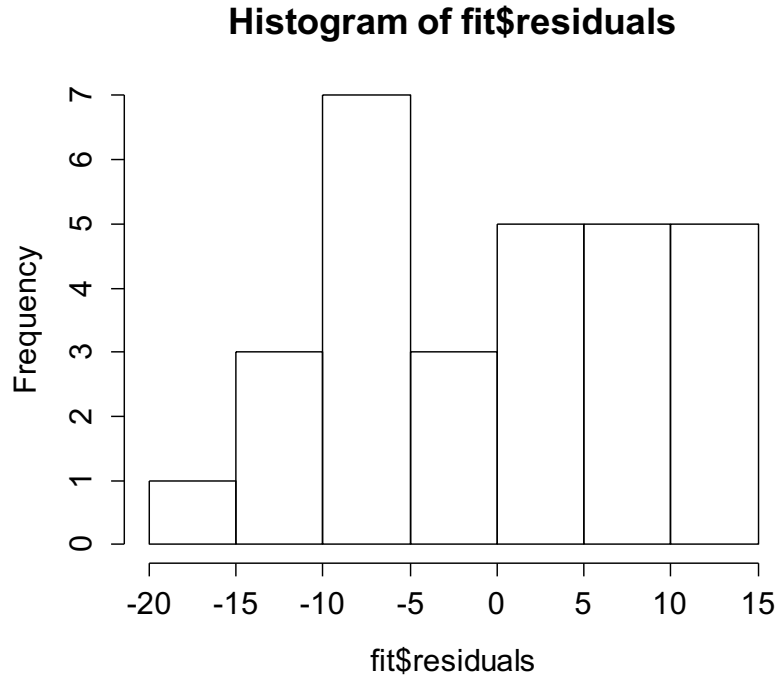
Intercept and slope can be addressed directly as:

```
> fit$coefficients[1]
(Intercept)
44.01954
> fit$coefficients[2] # index for specific element in “column”
Price
6.942303

# These were the parameters used to draw the abline.
```

Diagnostics – residuals

Ideally, residuals should be normally distributed.

> hist(fit$residuals)

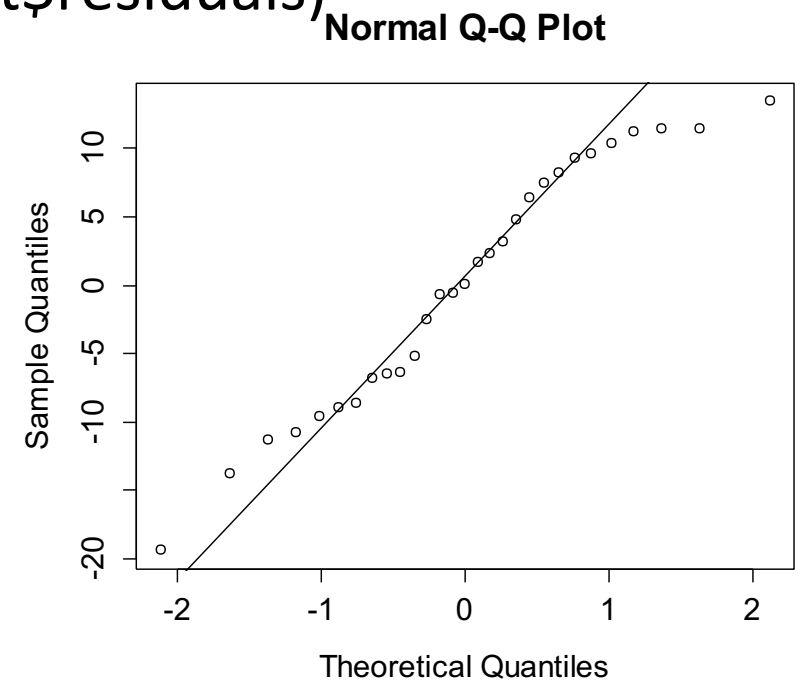


Not conclusive!

Diagnostics – residuals

A normal quantile plot is a better visual reference

- > qqnorm(fit$residuals)
- > qqline(fit$residuals)

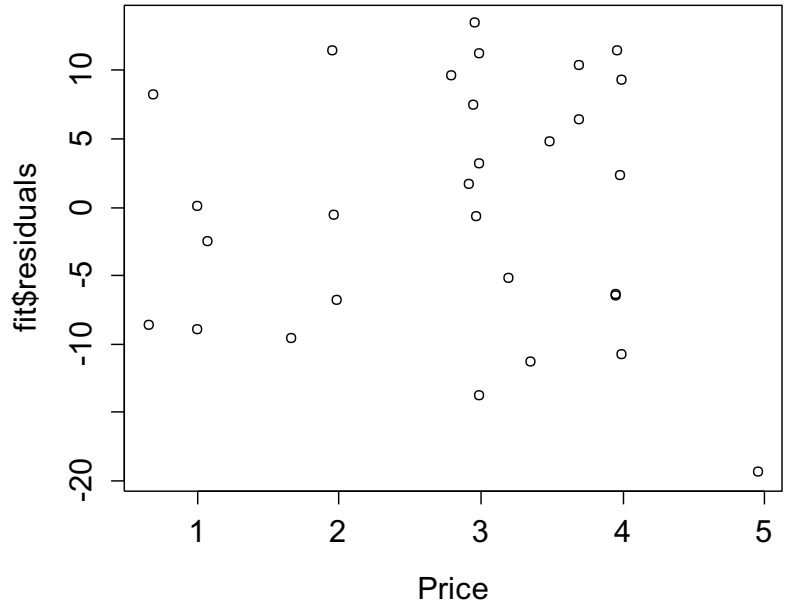


Good fit
for $-1 < z < 1$

Diagnostics – residuals

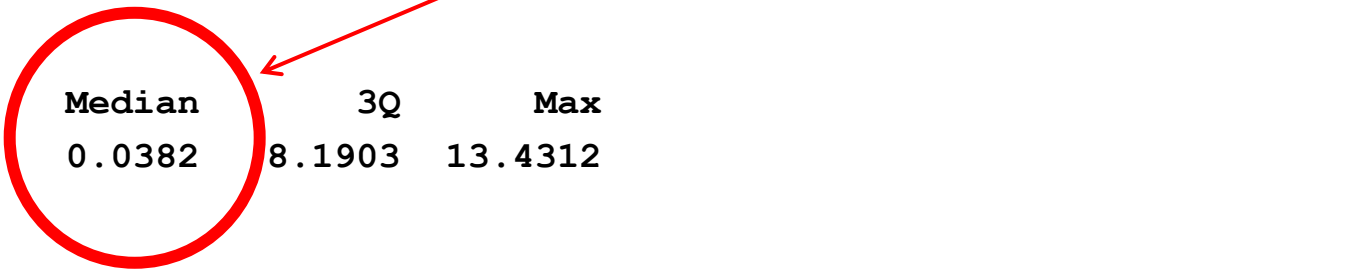
Residuals should be uncorrelated with input

> plot(Price, fit$residuals)



By eye $r \approx 0$

Diagnostics – summary




```
> summary(fit)
Call:
lm(formula = Function ~ Price)

Residuals:
     Min       1Q   Median       3Q      Max
-19.3839  -6.8347   0.0382   8.1903  13.4312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.020      4.565   9.642 3.09e-10 ***
Price          6.942      1.502   4.621 8.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.185 on 27 degrees of freedom
Multiple R-squared:  0.4416,      Adjusted R-squared:  0.421
F-statistic: 21.36 on 1 and 27 DF,  p-value: 8.428e-05
```

Diagnostics – summary

```
> summary(fit)
Call:
lm(formula = Function ~ Price)

Residuals:
     Min       1Q   Median       3Q      Max
-19.3839  -6.8347   0.0382   8.1903  13.4312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.020      4.565   9.642 3.09e-10 ***
Price          6.942      1.502   4.621 8.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.185 on 27 degrees of freedom
Multiple R-squared:  0.4416,      Adjusted R-squared:  0.421
F-statistic: 21.36 on 1 and 27 DF,  p-value: 8.428e-05
```

Coefficients: α, β

Diagnostics – summary

```
> summary(fit)
Call:
lm(formula = Function ~ Price)

Residuals:
     Min       1Q   Median       3Q      Max
-19.3839  -6.8347   0.0382   8.1903  13.4312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.020      4.565   9.642 3.09e-10 ***
Price          6.942      1.502   4.621 8.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.185 on 27 degrees of freedom
Multiple R-squared:  0.4416,      Adjusted R-squared:  0.421
F-statistic: 21.36 on 1 and 27 DF,  p-value: 8.428e-05
```

Hypothesis test that $\alpha, \beta = 0$ vs $\alpha, \beta \neq 0$

... Note on the p-value

The p-value is the probability of obtaining the value of the test statistic (coefficient) if null hypothesis was true (that is, the coefficient = 0 in this case).

Diagnostics – summary

```
> summary(fit)
Call:
lm(formula = Function ~ Price)

Residuals:
     Min       1Q   Median       3Q      Max
-19.3839  -6.8347   0.0382   8.1903  13.4312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.020      4.565   9.642 3.09e-10 ***
Price          6.942      1.502   4.621 8.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.185 on 27 degrees of freedom
Multiple R-squared:  0.4416,    Adjusted R-squared:  0.421
F-statistic: 21.36 on 1 and 27 DF,  p-value: 8.428e-05
```

This is the proportion of the variability in the data explained by the model

Coefficient of Determination: r^2

Diagnostics – summary

```
> summary(fit)
Call:
lm(formula = Function ~ Price)

Residuals:
     Min       1Q   Median       3Q      Max
-19.3839  -6.8347   0.0382   8.1903  13.4312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.020      4.565   9.642 3.09e-10 ***
Price          6.942      1.502   4.621 8.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.185 on 27 degrees of freedom
Multiple R-squared:  0.4416,      Adjusted R-squared:  0.421
F-statistic: 21.36 on 1 and 27 DF,  p-value: 8.428e-05
```

Overall significance of regression: that at least one coefficient ≠ 0

Diagnostics – summary

```
> summary(fit)
Call:
lm(formula = Function ~ Price)

Residuals:
     Min       1Q   Median       3Q      Max
-19.3839  -6.8347   0.0382   8.1903  13.4312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.020      4.565   9.642 3.09e-10 ***
Price          6.942      1.502   4.621 8.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.185 on 27 degrees of freedom
Multiple R-squared:  0.4416,    Adjusted R-squared:  0.421
F-statistic: 21.36 on 1 and 27 DF,  p-value: 8.428e-05
```

Median close to 0

Coefficients: α, β

Hypothesis test that α, β = 0 vs α, β ≠ 0

Coefficient of Determination: r^2

Overall significance of regression: that at least one coefficient ≠ 0

Correlation

Let's not forget that correlation and regression are intimately connected:

```
> cor(Price, Function) # Pearson's least squares r
0.6645614
> cor(Price, Function)^2
0.4416419
```

Which is the multiple R-squared reported on the previous slide.

Prediction

The linear model object can be used to calculate other fitted values such as forecasts as well as confidence and prediction intervals.

- For example, calculate the functionality of toothbrushes costing $6, $7 and $8:

```
> predict.lm(fit, newdata = data.frame(Price=c(6,7,8)),
  int="conf")
     fit   lwr    upr
1 85.67 75.26  96.08
2 92.62 79.26 105.97
3 99.56 83.21 115.91
```

?predict.lm

- Description

  Predicted values based on linear model object.

- Usage

```
predict(object, newdata, se.fit = FALSE, scale =
NULL, df = Inf, interval = c("none", "confidence",
"prediction"), level = 0.95, type = c("response",
"terms"), terms = NULL, na.action = na.pass,
pred.var = res.var/weights, weights = 1, ...)
```

- Arguments

  object : Object of class inheriting from "lm"

  newdata : An optional data frame of input variables. If omitted make fitted values.

  Interval : Type of interval calculation.

Summary

In this section we covered:

- Linear regression
- Interpreting the regression model
- Fitting a regression model in R and interpreting the output

Multiple linear regression

In this section we'll cover:

- Multiple linear regression
- Regression with qualitative variables and non-linear data
- Fitting a regression model in R and interpreting the output

Multiple linear regression

Applying least squares to multiple predictors:

- The relationship is now of the form:
  $y \approx a_1x_1 + a_2x_2 + a_3x_3 + \ldots + b$, *or*
  $y = a_1x_1 + a_2x_2 + a_3x_3 + \ldots + b + e$, *where* $e \sim N(\mu,\sigma^2)$
- x and y are numerical variables. We consider categories in x next.
- a_i and b are calculated to minimise the squared error between the observed values (the data) and the *fitted values* (i.e., those predicted by the model).
- Errors are (approximately) normally distributed.

Concrete compressive strength

Given the components and age of concrete, predict the resulting compressive strength.

- File: Concrete.csv

Cement	Slag	Ash	Water	Plas	CA	FA	Age	Strength
540	0	0	162	2.5	1040	676	28	79.99
540	0	0	162	2.5	1055	676	28	61.89
332.5	142.5	0	228	0	932	594	270	40.27
332.5	142.5	0	228	0	932	594	365	41.05
…	…	…	…	…	…	…	…	…

http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength

Variables

Inputs

- Cement kg/m^3
- Blast Furnace Slag kg/m^3
- Fly Ash kg/m^3
- Water kg/m^3
- Superplasticizer kg/m^3
- Coarse Aggregate kg/m^3
- Fine Aggregate kg/m^3
- Age Days

Output

- Concrete compressive strength MPa

Each predictor individually

Let's look at the ability of each input to predict the strength of concrete individually.

> Concrete <- read.csv("Concrete_regression.csv")

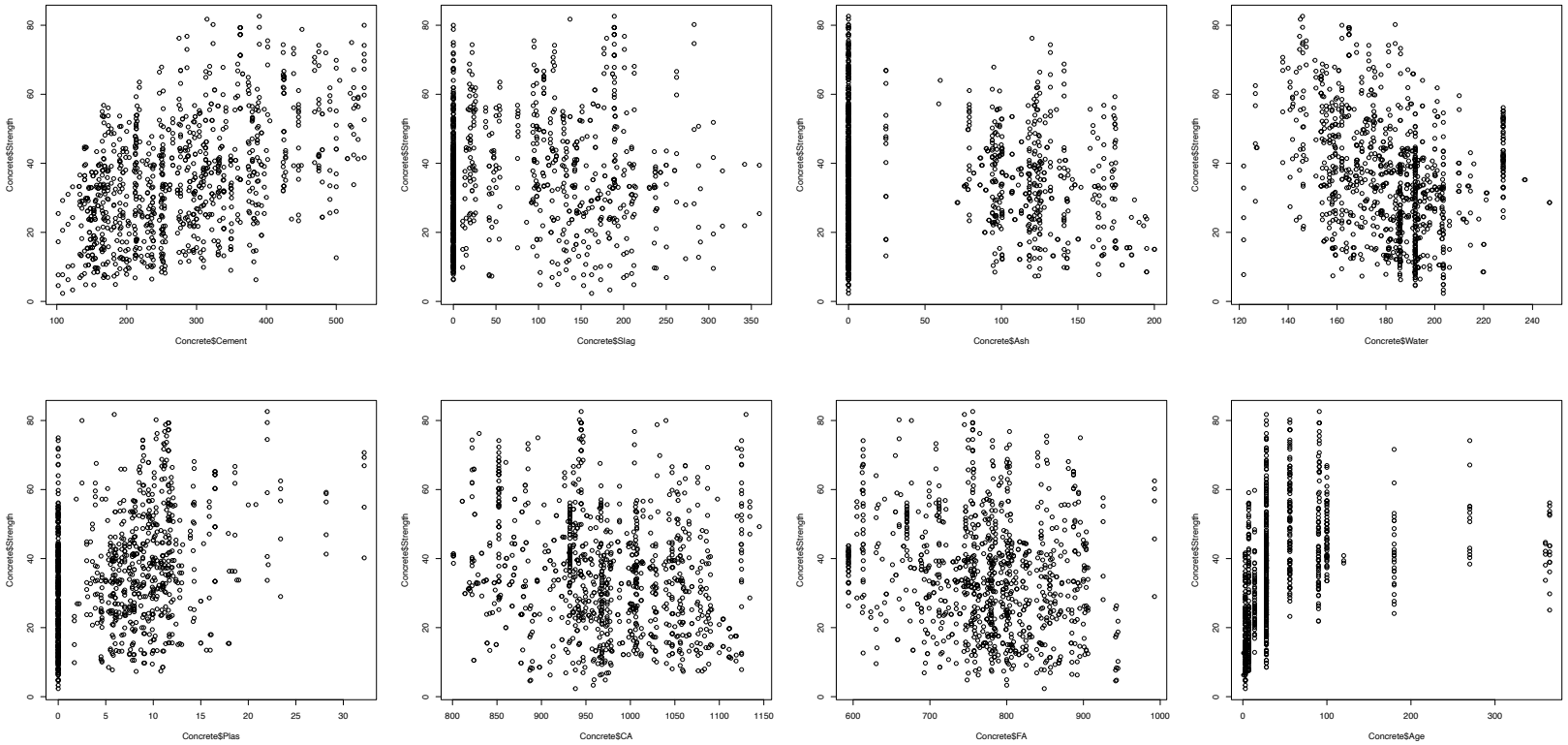
> > round(cor(Concrete[,1:8],Concrete[,9]), digits = 3)

```
          [,1]
Cement   0.498
Slag     0.135
Ash     -0.106
Water   -0.290
Plas     0.366
CA      -0.165
FA      -0.167
Age      0.329
```

But what is the limitation of this approach?

Each predictor individually

Or as a set of scatter plots:



To plot each predictor individually

```
> pdf("Concrete Plots.pdf", width=20, height=10)
> par(mfrow = c(2, 4))
> for (i in 1:8) {
>   plot(Concrete$Cement,Concrete$Strength)
>   plot(Concrete$Slag,Concrete$Strength)
>   plot(Concrete$Ash,Concrete$Strength)
>   plot(Concrete$Water,Concrete$Strength)
>   plot(Concrete$Plas,Concrete$Strength)
>   plot(Concrete$CA,Concrete$Strength)
>   plot(Concrete$FA,Concrete$Strength)
>   plot(Concrete$Age,Concrete$Strength)
> }
> dev.off()
```

Model: 2 predictors

Using only two input variables: cement and water:

```
> Concrete <- read.csv("Concrete_regression.csv")
> attach(Concrete)
> fit <- lm(Strength ~ Cement + Water)
> fit
```

```
Call:
lm(formula = Strength ~ Cement + Water)

Coefficients:
(Intercept)       Cement        Water
    49.9699       0.0763      -0.1961
```

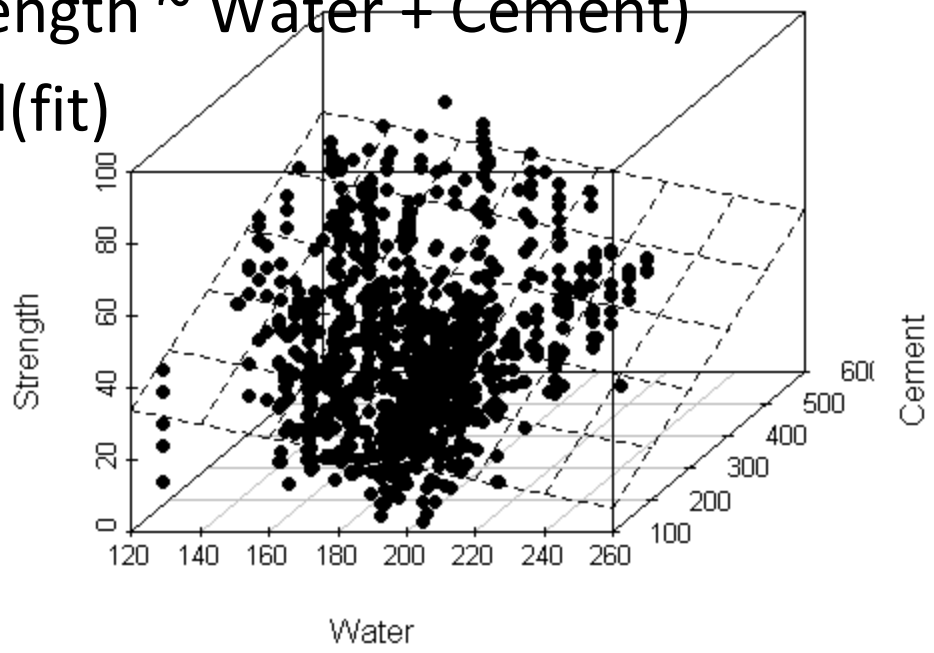
Summary

> summary(fit)

```
Call:
lm(formula = Strength ~ Cement + Water)

Residuals:
   Min     1Q Median     3Q    Max
-36.60 -10.76   0.00   9.46  41.57

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.96990    3.98731   12.53   <2e-16 ***
Cement       0.07631    0.00416   18.36   <2e-16 ***
Water       -0.19612    0.02034   -9.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.9 on 1027 degrees of freedom
Multiple R-squared: 0.31,       Adjusted R-squared: 0.309
F-statistic:  231 on 2 and 1027 DF,  p-value: <2e-16
```

3D scatterplot

```
> install.packages("scatterplot3d") # random find
> library(scatterplot3d)
> sur <-scatterplot3d(Water, Cement, Strength, pch=16)
> fit <- lm(Strength ~ Water + Cement)
> sur$plane3d(fit)
```



Model: all predictors

Using all input variables: cement and water:

```
> fit <- lm(Strength ~ . , data = Concrete) # note "." = all
> fit
```

Use "." to include all other columns

```
Call:
lm(formula = Strength ~ ., data = Concrete)

Coefficients:
(Intercept)       Cement         Slag          Ash
   -23.3312       0.1198       0.1039       0.0879
      Water         Plas           CA           FA
    -0.1499       0.2922       0.0181       0.0202
        Age
     0.1142
```

Summary (coefficients)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.33121   26.58550   -0.88   0.3804
Cement        0.11980    0.00849   14.11   <2e-16 ***
Slag          0.10387    0.01014   10.25   <2e-16 ***
Ash           0.08793    0.01258    6.99    5e-12 ***
Water        -0.14992    0.04018   -3.73   0.0002 ***
Plas          0.29222    0.09342    3.13   0.0018 **
CA            0.01809    0.00939    1.93   0.0544 .
FA            0.02019    0.01070    1.89   0.0595 .
Age           0.11422    0.00543   21.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

Summary (coefficients)

Accessing elements in the summary output directly, to put in a table, or make a graph for example:

```
> summary(fit)$coefficients[,4] # all p-values
 (Intercept)        Cement          Slag           Ash ...
3.803719e-01 1.897989e-41 1.598993e-23 5.019648e-12 ...

> summary(fit)$coefficients[7,3] # 7th t value
1.925656
```

Summary (residuals/model)

```
Call:
lm(formula = Strength ~ ., data = Concrete)

Residuals:
    Min      1Q  Median      3Q     Max
-28.65  -6.30   0.70    6.57   34.45


Residual standard error: 10.4 on 1021 degrees of
freedom
Multiple R-squared: 0.616, Adjusted R-squared: 0.613
F-statistic:  204 on 8 and 1021 DF,  p-value: <2e-16
```

Qualitative predictors

These are non-numerical, for example eye colour:

- When the variable has more than two levels, each level must be a separate variable in the regression equation. Indicator (0, 1) variables show the status of each observation at each factor level. This is also known as *one-hot encoding* among machine learners!`

Person	Eye.colour
A	Blue
B	Brown
C	Green
D	Blue
E	Blue

---->

Person	Eye.Blue	Eye.Brown	Eye.Green
A	1	0	0
B	0	1	0
C	0	0	1
D	1	0	0
E	1	0	0

Diamond data

From Applied Session 2:

```
> library(ggplot2)
> set.seed(9999) # Random seed
> dsmall <- diamonds[sample(nrow(diamonds), 1000), ]
  # sample of 1000 rows
> g = ggplot(data = dsmall, aes(x = carat, y = price,
  colour = color, size = clarity, alpha = cut)) +
  geom_point()
```

Diamond data

```
> dsmall
# A tibble: 1,000 x 10
   carat cut       color clarity depth table price     x     y
   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl>
 1  0.59 Very …    H     VVS2     61.1    57  1771  5.39  5.48
 2  0.3  Good      I     VS1      63.3    59   473  4.2   4.23
 3  0.42 Premi…    F     IF       62.2    56  1389  4.85  4.8
 4  0.95 Ideal     H     SI1      61.9    56  4958  6.31  6.35
 5  0.32 Premi…    D     VVS1     62      60   973  4.4   4.37
 6  0.52 Premi…    E     VS2      60.7    58  1689  5.17  5.21
 7  1.04 Ideal     H     SI1      62.3    57  5102  6.45  6.48
 8  0.5  Premi…    E     VS2      62.1    62  1559  5.1   5.08
 9  0.72 Ideal     F     SI1      62      55  2737  5.76  5.79
10  0.24 Good      F     VVS1     64.8    57   492  3.9   3.94
# ... with 990 more rows, and 1 more variable: z <dbl>
```

Basic plot: first observations

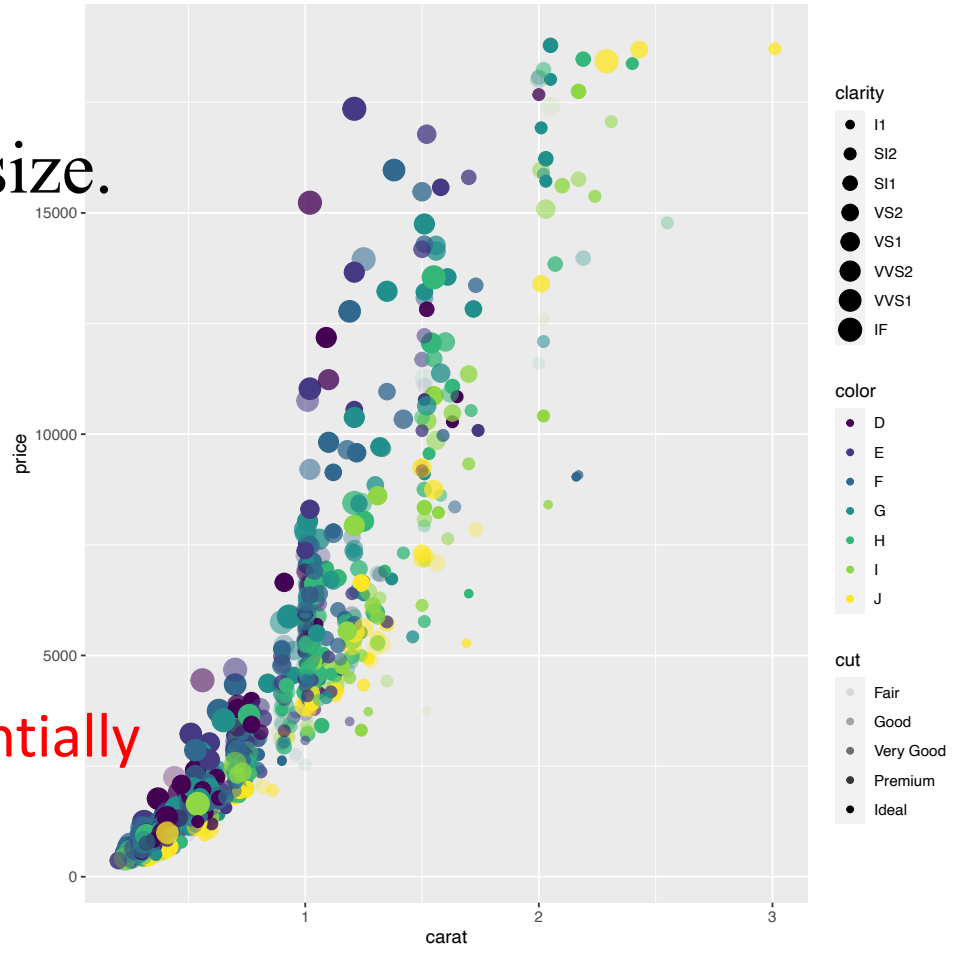
Non-linear:

- Take logs of price and size.

Categorical variables:

- Clarity
- Color
- Cut

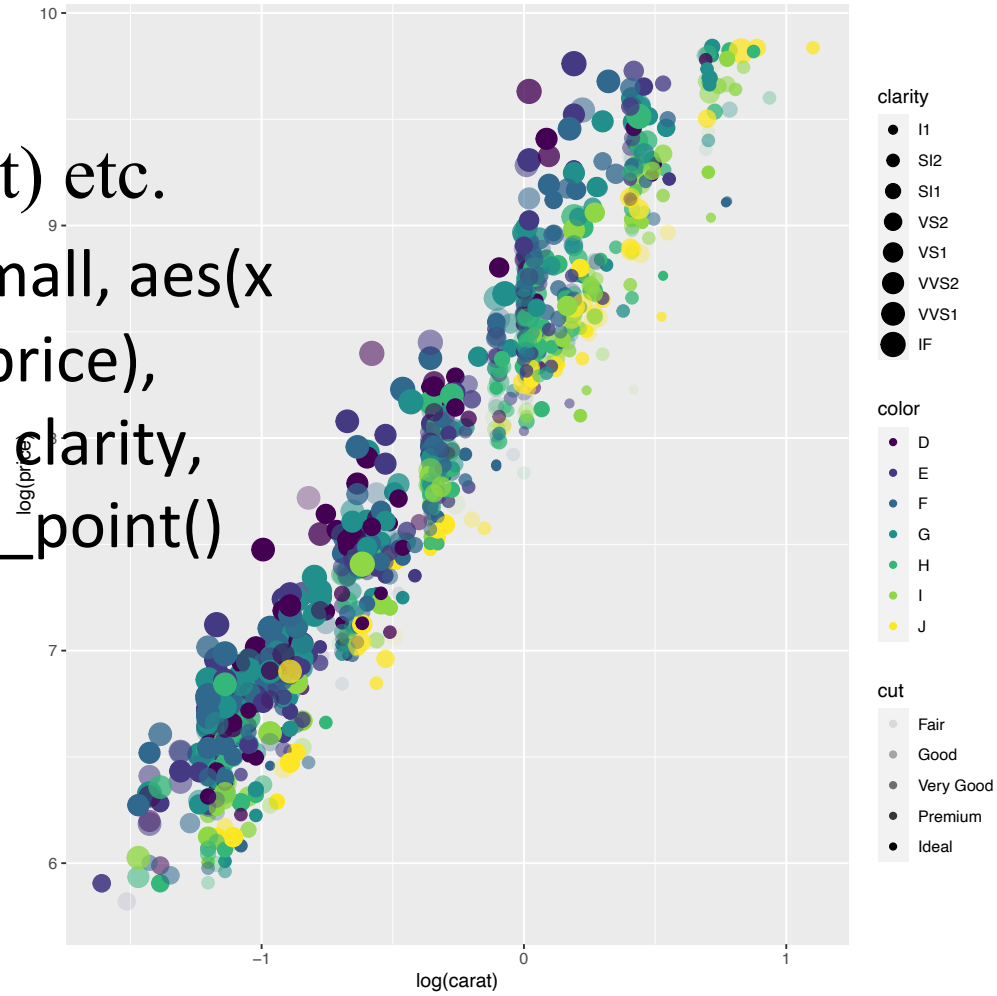
Note that data appears exponentially distributed in both x and y axes.



Plot using all variables on log scale

Linear relationship

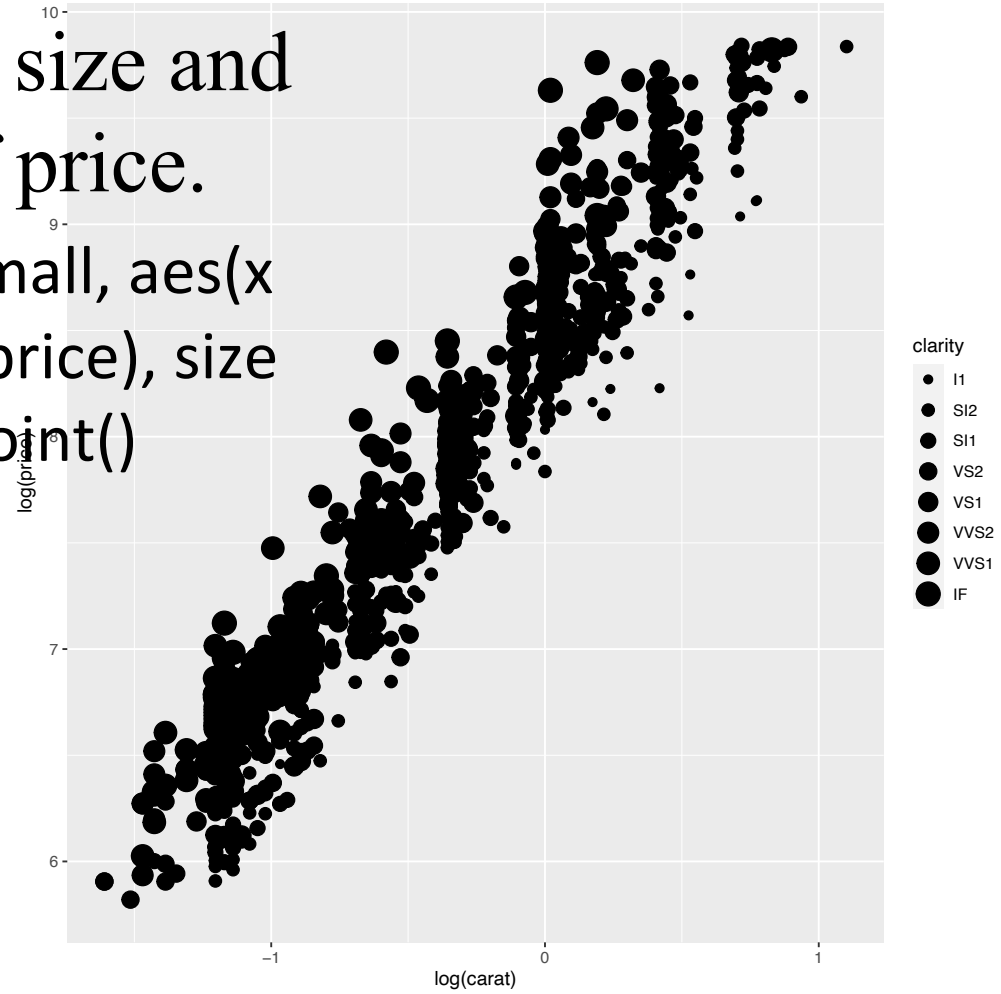
- Natural logs: $\log_e$(carat) etc.
  - > g = ggplot(data = dsmall, aes(x = log(carat), y = log(price), colour = color, size = clarity, alpha = cut)) + geom_point()
- Note, R uses:
  - > log to mean ln or $\log_e$
  - > log10 for log base 10
  - > Clarity has 8 levels



Plot using size and clarity only

Concentrating only on size and clarity as predictors of price.

> g = ggplot(data = dsmall, aes(x = log(carat), y = log(price), size = clarity)) + geom_point()



Regression with factors

Specify ‘clarity’ as a ‘treatment’ having 8 levels and perform the regression as usual.

- R implicitly creates an indicator matrix (0, 1 terms) for levels.

```
> attach(dsmall)
> contrasts(clarity) = contr.treatment(8) # 8 levels
> d.fit <- lm(log(price) ~ log(carat) + clarity)
> d.fit
```

Coefficients

> d.fit

```
Call:lm(formula = log(price) ~ log(carat) + clarity)

Coefficients:
(Intercept)   log(carat)      clarity2
     7.7884        1.8324        0.4506
   clarity3      clarity4      clarity5
     0.6052        0.7852        0.8264
   clarity6      clarity7      clarity8
     0.9675        1.0290        1.1138
```

> Note that the final model implicitly includes the lowest factor level of the treatment (I1 = clarity1) as the base case.

Summary

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.78844    0.04926 158.108   <2e-16 ***
log(carat)   1.83242    0.01108 165.319   <2e-16 ***
clarity2     0.45065    0.05137   8.772   <2e-16 ***
clarity3     0.60524    0.05086  11.900   <2e-16 ***
clarity4     0.78523    0.05099  15.398   <2e-16 ***
clarity5     0.82644    0.05200  15.893   <2e-16 ***
clarity6     0.96753    0.05321  18.184   <2e-16 ***
clarity7     1.02899    0.05410  19.019   <2e-16 ***
clarity8     1.11380    0.05809  19.173   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
etc.
```

Contrasts

To see which clarity level corresponds to each treatment look at the contrast matrix:

```
> contrasts(clarity)
     2 3 4 5 6 7 8
I1   0 0 0 0 0 0 0
SI2  1 0 0 0 0 0 0
SI1  0 1 0 0 0 0 0
VS2  0 0 1 0 0 0 0
VS1  0 0 0 1 0 0 0
VVS2 0 0 0 0 1 0 0
VVS1 0 0 0 0 0 1 0
IF   0 0 0 0 0 0 1
```

Summary (overall)

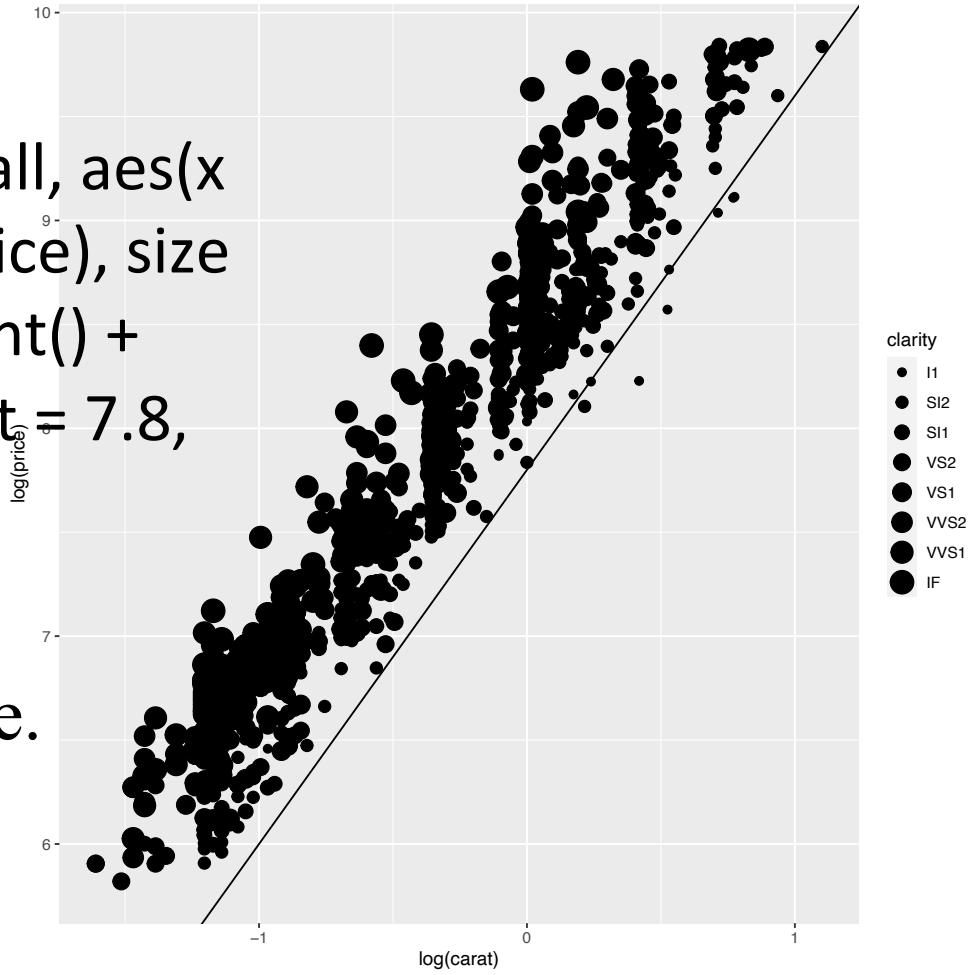
```
Residual standard error: 0.1843 on 991 degrees of
freedom
Multiple R-squared:0.9672,
Adjusted R-squared: 0.9669
F-statistic: 3652 on 8 and 991 DF,
p-value: < 2.2e-16
```

Fitted model

ln(price) v ln(carat)

> g = ggplot(data = dsmall, aes(x = log(carat), y = log(price), size = clarity)) + geom_point() + geom_abline(intercept = 7.8, slope = 1.8)

- Basic model fitted to I1.
- Quality increase additive.



Fitted values

Recall

> d.fit

```
Call:
lm(formula = log(price) ~ log(carat) + clarity)
Coefficients:
(Intercept)   log(carat)     clarity2     clarity3
     7.7884       1.8324       0.4506       0.6052
   clarity4     clarity5     clarity6     clarity7
     0.7852       0.8264       0.9675       1.0290
   clarity8
     1.1138
```

- What should a 1.5 carat, VVS1 diamond sell for?

Fitted values

- What should a 1.5 carat, VVS1 diamond sell for?

```
Log(y) = log(price) = log(carat) * log(x) (+ intercept) + clarity
         log(price) = 1.8324 * log(1.5) + 7.7884 + 1.0290
         log(price) = 1.8324 * 0.4055  + 7.7884 + 1.0290
         log(price) = 9.5603
             price  = $14,191
```

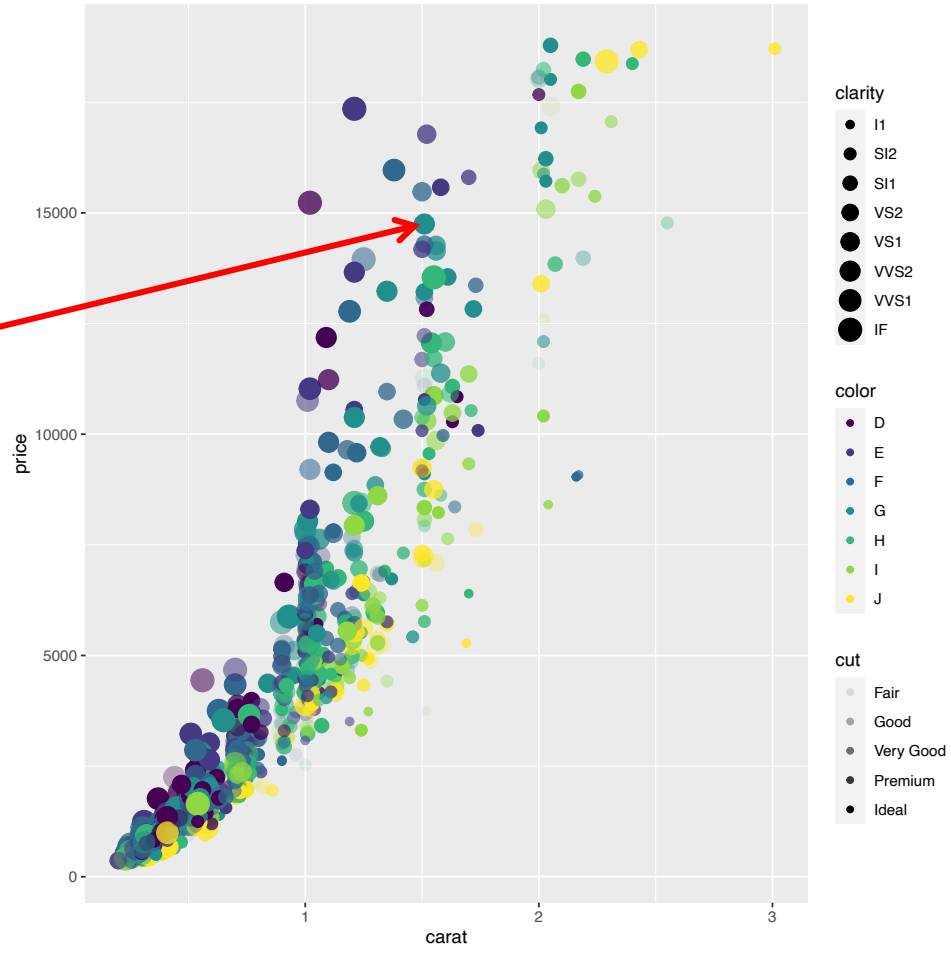
Raising each side to the power of e^x

```
Coefficients:
(Intercept)   log(carat)     clarity2     clarity3
     7.7884        1.8324       0.4506       0.6052
   clarity4     clarity5     clarity6     clarity7
     0.7852        0.8264       0.9675       1.0290
```

Fitted values

Going back to the original plot:

```
Size = 1.5
Clarity = VVS1
price  = $14,191
```



Summary

In this section we covered:

- Multiple linear regression
- Regression with qualitative variables and non-linear data
- Fitting a regression model in R and interpreting the output

Extension

Subsets and shrinkage methods:

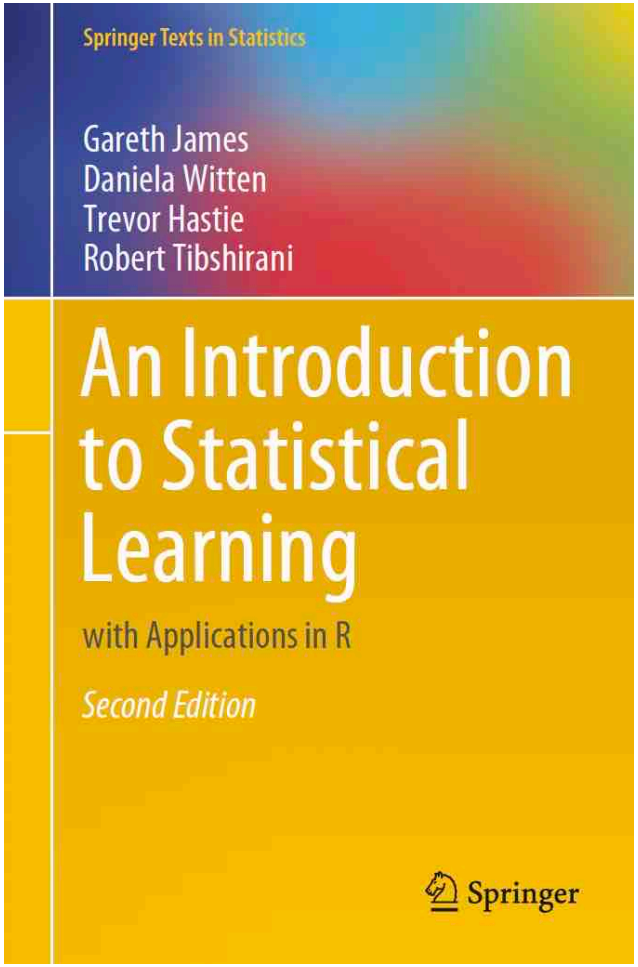
- When the number of input variables is large relative to the number of observations, a regression model may be improved by reducing the number of inputs.
- Improvements include: less variability in predictions, irrelevant variables removed, model more interpretable.
- Methods include identifying a subset of inputs, shrinkage and regularization, dimension reduction.

Extension

You can learn these methods in James et al., *An Introduction to Statistical Learning: with Applications in R*, 2nd Ed. (2021)

- Best subset regression 6.1,
- Ridge regression 6.2.1,
- Least Absolute Shrinkage and Selection Operator (LASSO) 6.2.2,
- Dimension reduction 6.3, and
- Examples of how to apply these methods in 6.5.

References


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer

References

Books available online from the Monash Library

Teetor, P., R Cookbook (2012)

- (pp 267 – 288 a good reference on regression and regression diagnostics)

G. James et al., *An Introduction to Statistical Learning: with Applications in R*, 2nd Ed. (2021)

- Chapter 3, Linear Regression, Sections 3.1 – 3.3, This is quite technical and statistically heavy!, 3.6 (Lab) has some good examples. “Advertising” data example is used in the tutorial, “carseats” data also.