FIT3152 Data analytics– Lecture 6

Regression

- Assignment Q&A
- Network review questions
- Linear regression
- Regression diagnostics
- Multiple linear regression
- Regression with qualitative variables

Consultations on Zoom

Clayton consultations have commenced:

- Any student can attend any consultation.
- Schedule on Moodle, https://lms.monash.edu/
- Current days/times:
- Monday 9:30-10:30AM, 2:00-3:00PM, 6:00-7:00PM,
- Tuesday 9:00-10:00AM, 12:00PM-1:00PM,
- Wednesday 10:00AM-11:00, 11:00-12:00PM,
- Thursday 1:00PM-02:00PM, 6:00PM-7:00PM.
- Please check the schedule for any changes.

Week-by-week

Week Starting	Lecture	Topic	Tutorial	A1	A2
28/2/22	1	Intro to Data Science, review of basic statistics using R	...		
7/3/22	2	Exploring data using graphics in R	T1		
14/3/22	3	Data manipulation in R	T2	Released	
21/3/22	4	Data Science methodologies, dirty/clean/tidy data, data manipulation	T3		
28/3/22	5	Network analysis	T4		
4/4/22	6	Regression modelling	T5		
11/4/22	7	Classification using decision trees	T6		
		Mid-semester Break		Submitted	
25/4/22	8	Naïve Bayes, evaluating classifiers	T7		Released
2/5/22	9	Ensemble methods, artificial neural networks	T8		
9/5/22	10	Clustering	T9		
16/5/22	11	Text analysis	T10		Submitted
23/5/22	12	Review of course, Exam preparation	T11		

Assignment 1

Assignment 1: Summary

FIT3152 Data analytics – 2022: Assignment 1

Your task	• Analyse the activity, language use and social interactions of an on-line community using metadata and linguistic summary from a real on-line forum and submit a report of your findings. • This is an individual assignment.
Value	• This assignment is worth **20%** of your total marks for the unit. • It has 30 marks in total.
Suggested Length	• 6 – 8 A4 pages (for your report) + extra pages as appendix (for your code) • Font size 11 or 12pt, single spacing
Due Date	**11.55pm Friday 22nd April 2022**
Submission	• PDF file only. Naming convention: *FirstnameSecondnameID.pdf* • Via Moodle Assignment Submission. • Turnitin will be used for similarity checking of all submissions.
Late Penalties	• 10% (3 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

Instructions

Submit the results of your analysis, answering the research questions and report anything else you discover of relevance. If you choose to analyse only a subset of your data, you should explain why.

You are expected to include at least one multivariate graphic summarising key results. You may also include simpler graphs and tables. Report any assumptions you've made in modelling, and include your R code as an appendix.

There are two options for compiling your report:
(1) You can submit a single pdf with R code pasted in as machine-readable text as an appendix, or
(2) As an R Markup document that contains the R code with the discussion/text interleaved. Render this as an HTML file and print off as a pdf and submit.

Regardless of which method you choose, you will submit a single pdf, and your R code will be machine readable text. We need to conform to this format as the university now requires all student submission to be processed by plagiarism detection software.

Submit your report as a single PDF with the file name ***FirstnameSecondnameID.pdf*** on Moodle.

Assignment 1: Software

Software

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

Assignment 1: Questions a & b

Questions

Activity, language use and social interactions in an on-line community. Analyse the metadata and linguistic summary from a real on-line forum and submit a report of your findings. Do the following:

(a) Analyse activity and language on the forum over time:

1. How active are participants over the longer term (that is, over months and/or years)? Are there periods where activity increases or decreases? Is there a trend over time? **(3 Marks)**
2. Looking at the linguistic variables, do the levels of these change over the duration of the forum? Is there a relationship between linguistic variables over the longer term? **(3 Marks)**

(b) Analyse the language used by threads:

We can think of threads as groups of participants posting on the same topic.

1. Using the relevant linguistic variables, is it possible to see whether or not particular threads are happier or more optimistic than other threads, or the forum in general, at different periods in time. **(3 Marks)**

Assignment 1: Question c

(c) <u>Analyse social networks online:</u>

We can think of authors posting to the same thread at similar times (for example during the same month) as having a connection to each other, forming a social network. This is called a two-mode network. When an author posts to more than one network during the same time period their social network extends to include authors from both networks, and so on. We will cover social network analysis in Lecture 5.

1. Create a non-trivial social network of all authors who are posting over a particular time period. For example, over one month. To create this, your social network should include at least 30 authors, some of whom will have posted to multiple (2 or more) threads during this period. Your social network should be connected, although some authors may be disconnected from the main group. Present your result as a network graph. **(3 Marks)**
2. Identify the most important author in the social network you created. Looking at the language they use, can you observe any difference between them and other members of their social network? **(3 Marks)**

Assignment 1: Overall considerations

(d) Overall considerations:

- The quality and clarity of your reasoning and assumptions. **(3 Marks)**
- The strength of support for your findings. **(3 Marks)**
- The quality of your writing in general and communication of results. **(3 Marks)**
- The quality of your graphics throughout, including at least one high-quality multivariate graphic. **(3 Marks)**
- The quality of your R coding. **(3 Marks)**

Assignment 1: Data generation

Data

The data is contained in the file webforum.csv and consists of the metadata and linguistic analysis of posts over the years 2002 to 2011. You will each work with 20,000 posts, randomly selected from the original file. The linguistic analysis was conducted using Linguistic Inquiry and Word Count (LIWC), which assesses the prevalence of certain thoughts, feelings and motivations by calculating the proportion of key words used in communication. See http://liwc.wpengine.com/ for more information, including the language manual http://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_LanguageManual.pdf

Create your individual data as follows:

```
rm(list = ls())
set.seed(XXXXXXXX) # XXXXXXXX = your student ID
webforum <- read.csv("webforum.csv")
webforum <- webforum [sample(nrow(webforum), 20000), ] # 20000 rows
```

Assignment 1: Data fields

Data fields given. (see the language manual for more detail and examples):

Column	Brief Descriptor	Column	Brief Descriptor
ThreadID	Unique ID for each thread	we	"We, us, our" words
AuthorID	Unique ID for each author	you	"You" words
Date	Date	shehe	"She, her "him words
Time	Time	they	"They" words
WC	Word count of the text of the post	posemo	Expressing positive emotions
Analytic	Summary: Analytical thinking	negemo	Expressing negative emotions
Clout	Summary: Power, force, impact	anx	Indicating anxiety
Authentic	Summary: Authentic tone of voice	anger	Indicating anger
Tone	Summary: Emotional tone	sad	Indicating sadness
ppron	"I, we, you" words	focuspast	Expressing a focus on the past
i	"I, me, mine" words	focuspresent	Expressing a focus on the present
~~focusfuture~~	~~Expressing a focus on the future~~	focusfuture	Expressing a focus on the future

Assignment 1: Data extract

ThreadID	AuthorID	Date	Time	WC	Analytic	Clout	Authentic	Tone	ppron	i	we	you	shehe	they	...
144564	41084	9/8/04	4:46	134	55.23	69.94	63.91	68.05	7.46	2.99	2.24	1.49	0	0.75	...
404119	128515	21/7/07	22:27	12	1	79.76	74.76	25.77	33.33	8.33	0	0	0	25	...
395992	93243	19/6/07	1:02	28	13.85	76.25	1.06	99	7.14	3.57	0	3.57	0	0	...
405421	99958	24/7/07	1:40	16	84.57	89.42	35.37	1	6.25	0	0	6.25	0	0	...
662470	185647	5/12/09	16:05	37	32.06	79.13	21.26	75.85	18.92	8.11	0	0	5.41	5.41	...
420058	53655	13/9/07	22:59	17	26.21	3.89	99	1	11.76	5.88	0	0	0	5.88	...
13933	1740	9/3/02	2:01	61	22.35	37.15	72.51	25.77	11.48	6.56	1.64	0	0	3.28	...
245087	80190	9/11/05	15:06	94	82.45	66.48	44.79	25.77	4.26	2.13	1.06	0	0	1.06	...
442550	47686	6/12/07	5:06	80	61.95	54.96	59.88	96.76	7.5	5	0	1.25	0	1.25	...
352716	26979	5/1/07	21:33	10	8.19	84.14	1	25.77	0	0	0	0	0	0	...
463617	104430	29/2/08	8:02	249	98.57	78.92	15.3	83.06	3.61	0.8	1.61	0	0.8	0.4	...
363541	-1	15/2/07	11:30	26	53.63	87.57	38.39	99	11.54	3.85	0	7.69	0	0	...
258941	44297	1/1/06	13:47	59	94.34	91.23	10.76	6.73	8.47	1.69	1.69	5.08	0	0	...
765163	54960	17/12/10	21:06	139	26.01	58.53	13.52	66.61	7.91	1.44	0.72	2.88	0	2.88	...
263152	79878	18/1/06	7:34	114	48.42	73.03	9.58	1	10.53	4.39	0	2.63	0	3.51	...
228773	166362	6/9/09	4:52	14	13.85	98.33	89.63	25.77	14.29	0	0	14.29	0	0	...
254482	83344	6/1/06	0:17	107	80.6	77.26	24.3	1	2.8	0.93	0	0.93	0	0.93	...
255544	81721	17/12/05	21:46	166	98.84	45.21	34.91	17.07	1.2	0	0.6	0.6	0	0	...
218880	22130	18/7/05	5:07	11	12.85	81.84	99	1	18.18	9.09	0	9.09	0	0	...
244912	41084	8/11/05	2:46	35	99	38.74	13.15	98.56	0	0	0	0	0	0	...
273089	-1	25/2/06	4:22	92	90.46	58.59	68.63	11.64	8.7	2.17	1.09	0	5.43	0	...
265715	38794	2/2/06	0:57	275	81.4	69.47	29.78	20.28	6.55	2.91	0.73	0.73	1.09	1.09	...
198321	21367	17/4/05	22:23	110	54.02	89.83	14.1	94.75	10.91	5.45	0	1.82	0.91	2.73	...
45244	13359	21/12/02	18:01	45	92.84	81.29	10.08	67.75	8.89	4.44	0	0	0	4.44	...
233103	70832	1/10/05	9:19	77	95.05	69.84	65.41	97.38	2.6	0	0	1.3	0	1.3	...
566748	109818	25/3/09	5:25	77	89.94	74.2	9.09	99	2.6	0	1.3	0	0	1.3	...
146671	116703	24/1/07	7:25	38	33.88	1.81	98.54	74.74	7.89	7.89	0	0	0	0	...
745917	105443	1/11/10	6:46	242	27.37	38.61	93.65	6.99	12.81	8.26	1.24	2.48	0	0.83	...
618782	165386	11/7/09	2:46	119	55.71	50	10.42	1	3.36	0.84	0	0	0.84	1.68	...
55689	19796	10/2/03	2:07	12	1	20.24	98.01	25.77	16.67	16.67	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Assignment 1: Draft rubric

Question	Part	Mark 1	Mark 2	Mark 3
a	1	Activity over time (posts and/or threads) calculated.	Suitable graphical presentation is created.	Visual inspection and descriptive analysis performed.
	2	Suitable time series summary calculations of LIWC variables performed.	Suitable graphical presentation is created.	Descriptive analysis of relationship between variables over time.
b	1	Relevant LIWC variables identified with justification.	Summary of relevant LIWC variables presented.	Descriptive or graphical comparison of threads performed.
c	1	Suitable time interval for calculating social network identified.	Social network with Author IDs as nodes created.	Suitable network plot created.
	2	Vertex importance measures created.	Most important author identified with reasons given.	Comparison of most important authors with others.
d	1	Data pre-processing (or not) is justified and performed.	Suitable time division identified and time coded accordingly.	Data analysis and conclusions are logical overall.
	2	Higher level of justification for Part a (hyp tests other statistical models etc.)	Higher level of justification for Part b (hyp tests other statistical models etc.).	Higher level of justification for Part c (hyp tests other statistical models etc.).
	3	Report has good structure and flow.	Quality of writing is good throughout.	Quality of writing is excellent throughout.
	4	One high quality multivarate graphic included.	Graph choice is appropriate throughout.	Graphs are high quality throughout.
	5	R coding looks sensible and has good readability.	Coding is used to automate analysis across multiple fields - Parts a and b.	Coding used to automate network construction - Part c.

Response to student questions

- I was attempting assignment 1 and was trying to get the year and month separated from the webforum$Date. However, this ended up with a new column year and all values are N/A. It's still the same even if I do not create a new column. Is there a mistake I'm making?
  - > You need to have date set to the correct format.
  - > View your data in R and set filter accordingly.
  - > See the Example in Lecture 4.

Response to student questions

- Do we need to include data cleaning in our report too?
  - > Your data should be clean, but you may pre-process your data before doing your analysis, which you should document in your report.
- Does -1 value for AuthorID stand for a valid user?
  - > Perhaps look through the forum to see when it appears. Does it seem to be a referring to a single person with a valid ID? Make your decision on whether or not to use these posts and note this in your report.

Review questions from last lecture

Please respond via Zoom chat if you want!

Question 1

For the graph below, *diameter* is:

a. 1

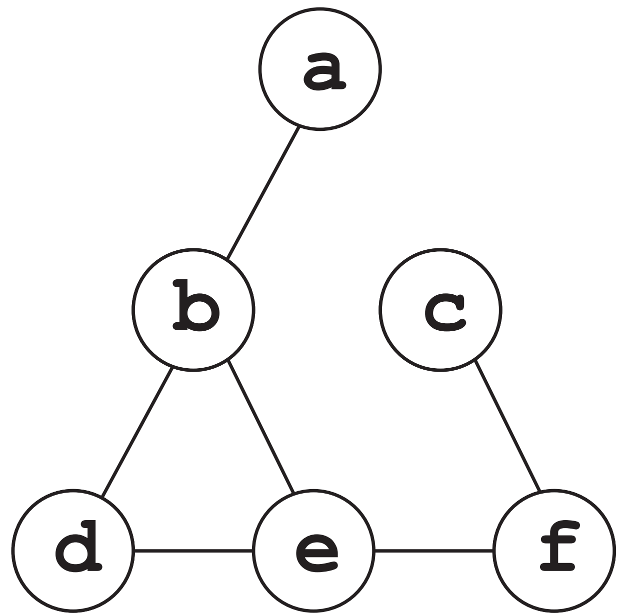
b. 2

c. 3

d. 4

e. 5

f. 6



Question 2

For the graph below, d_b is:

a. 1

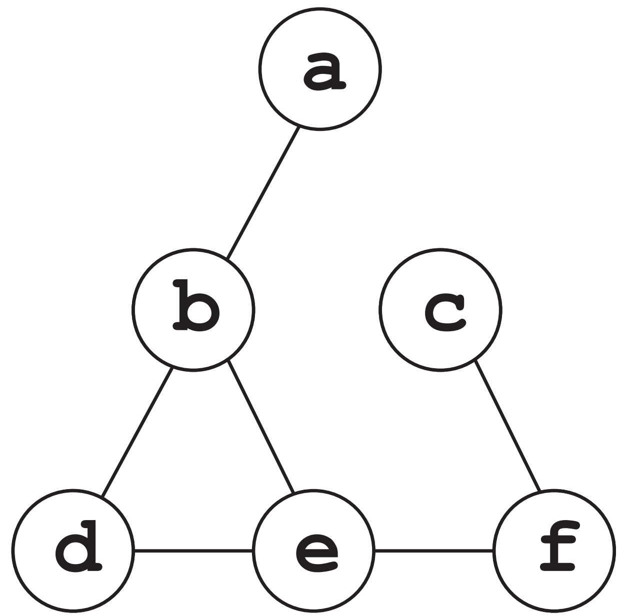
b. 2

c. 3

d. 4

e. 5

f. 6



Question 3

For the graph below, $c_B(b)$ is:

a. 1

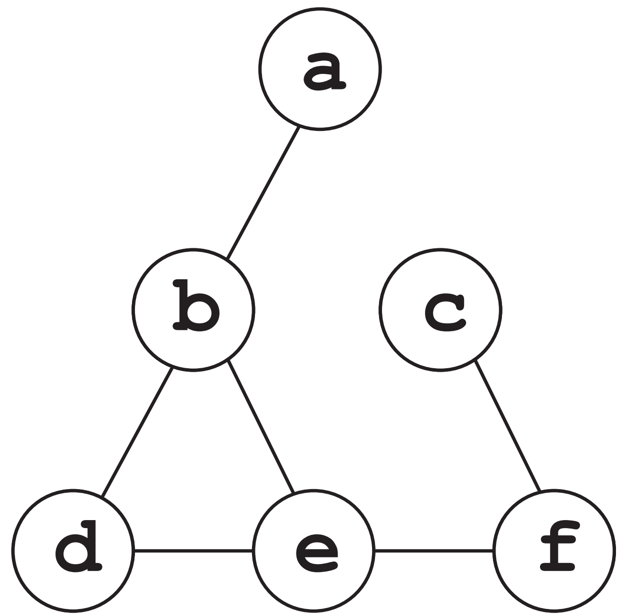
b. 2

c. 3

d. 4

e. 5

f. 6



Question 4

For the graph below, $c_{CL}(b)$ is:

a. 1/1

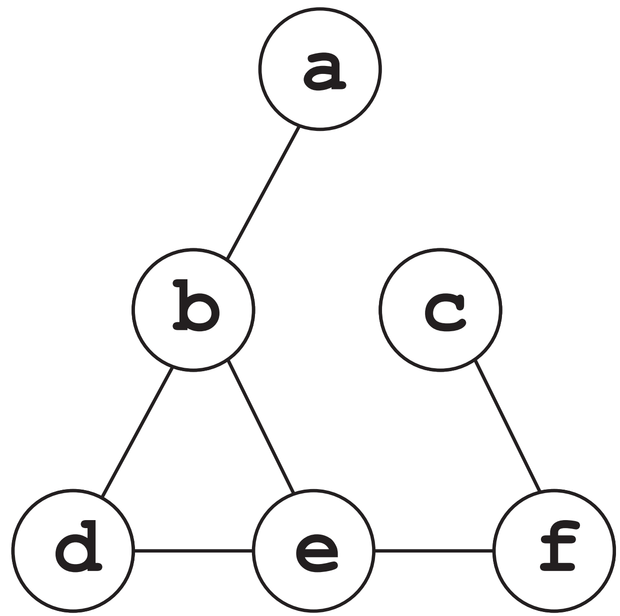
b. 1/2

c. 1/3

d. 1/4

e. 1/6

f. 1/8



Question 5

For the graph below, *largest clique size* is:

a. 1

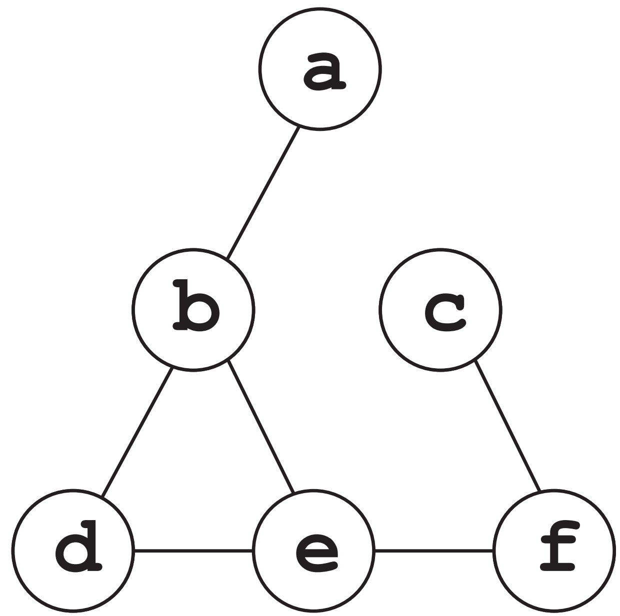
b. 2

c. 3

d. 4

e. 5

f. 6



Regression

COVID-19

SCIENTIFIC REPORTS

natureresearch

Check for updates

Covid-19 mortality is negatively associated with test number and government effectiveness

Li-Lin Liang[1,7], Ching-Hung Tseng[2], Hsiu J. Ho[3] & Chun-Ying Wu[4,5,6,7]✉

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

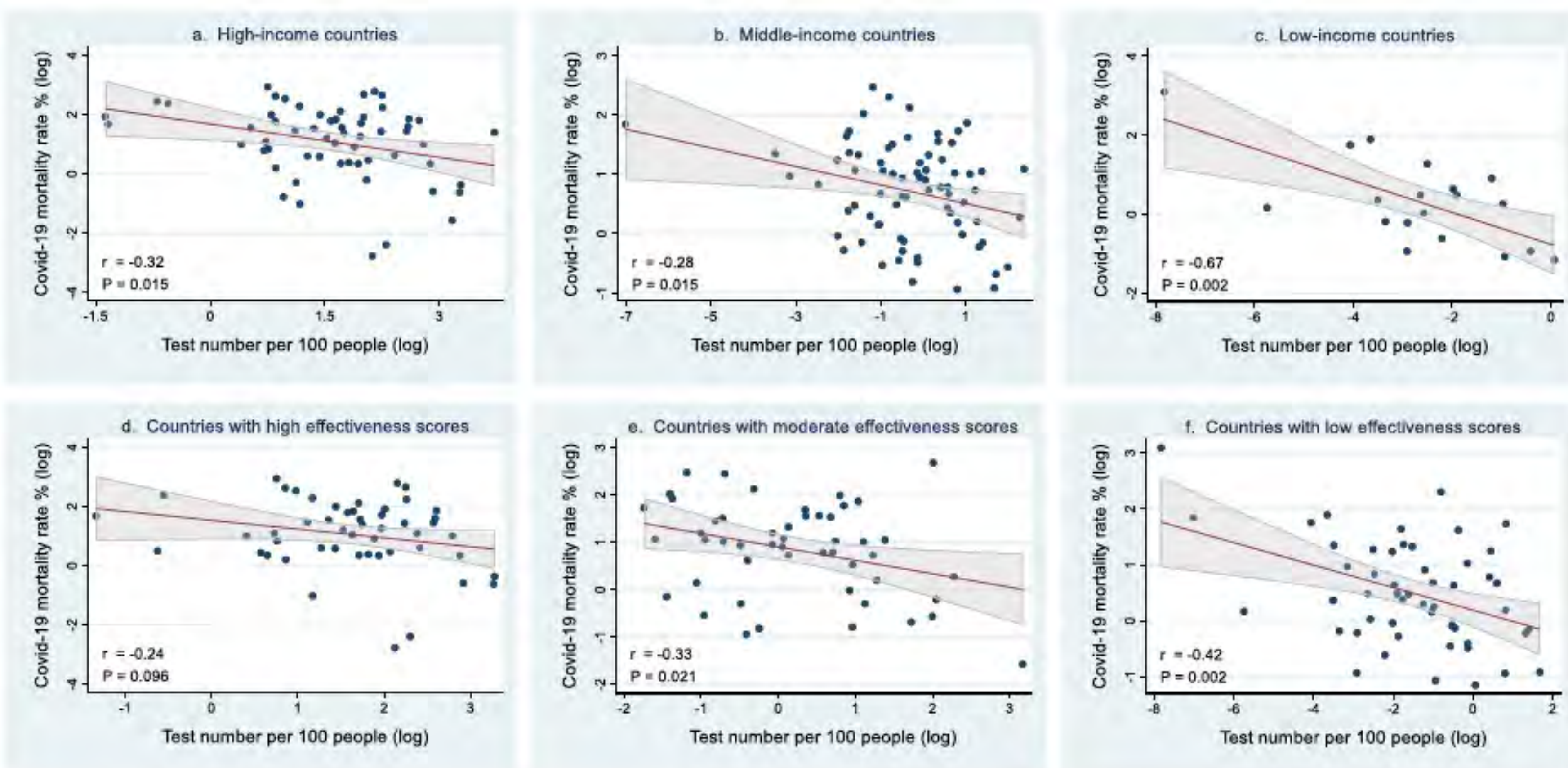
COVID-19

	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).

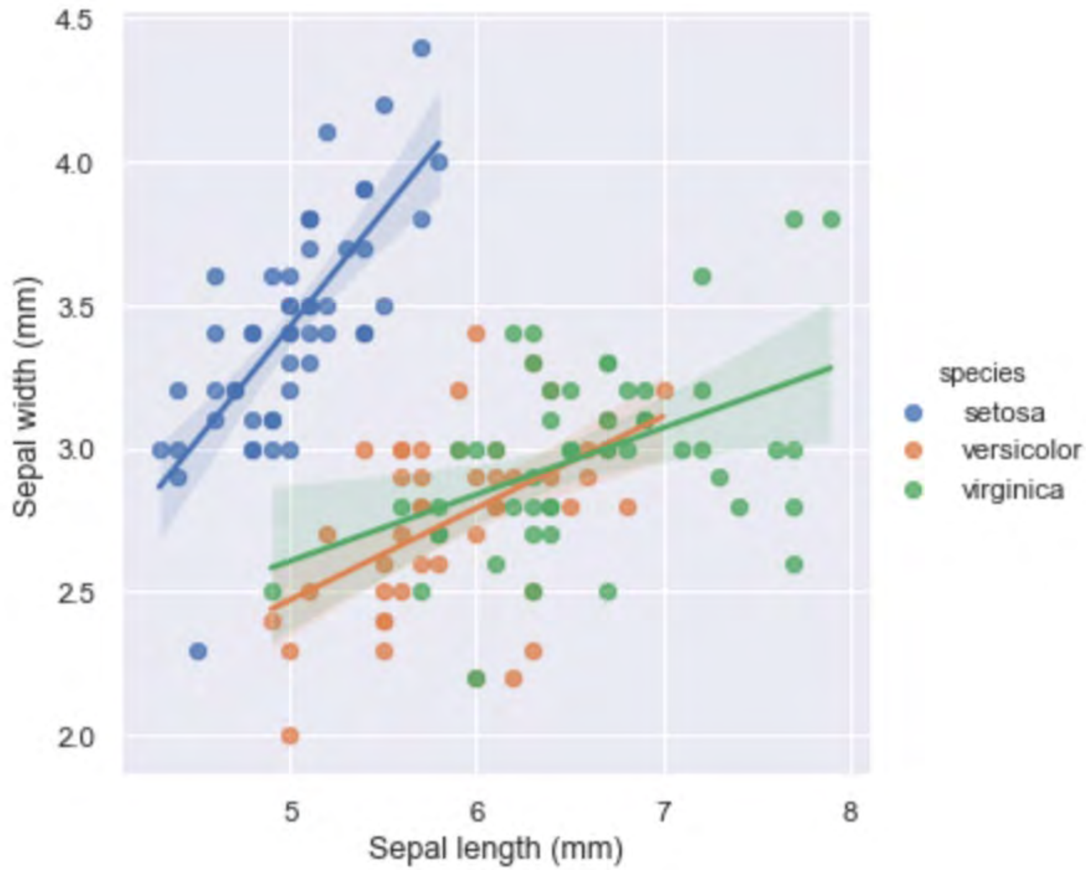
https://www.nature.com/articles/s41598-020-68862-x

COVID-19



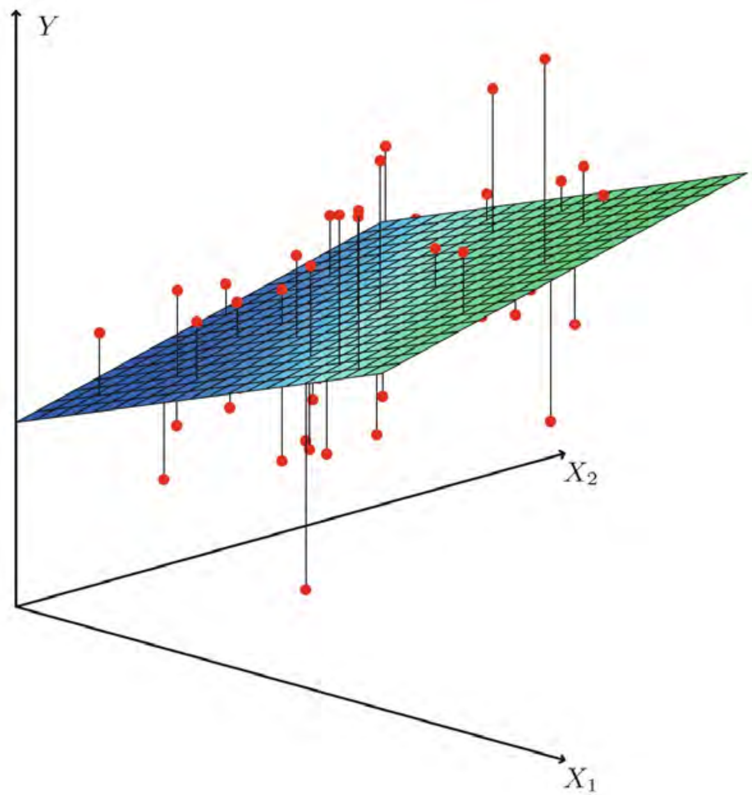
https://www.nature.com/articles/s41598-020-68862-x

Linear regression – 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.

Linear regression

See R Script of lecture examples

> Lecture 6 Regression.R

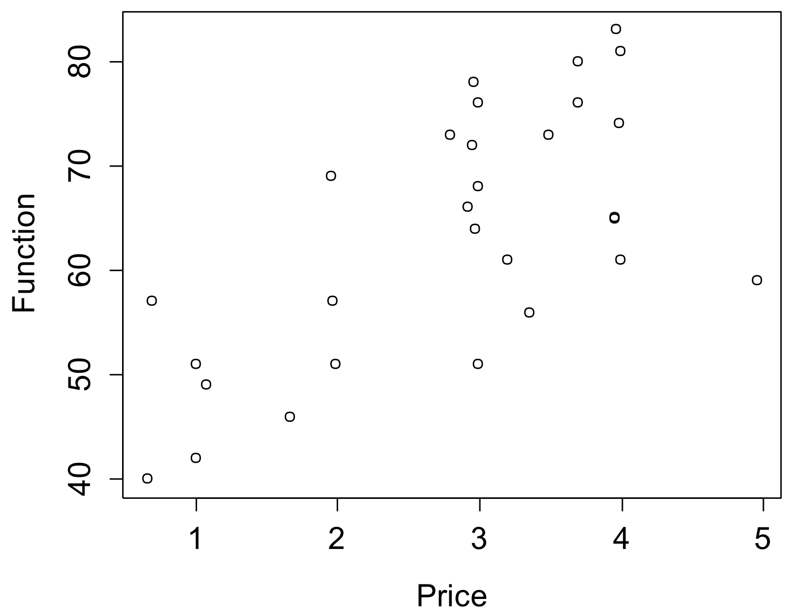
Lecture 4 Regression.R

Source on Save

```r
# clean up the environment before starting
rm(list = ls())
Toothbrush <- read.csv("Toothbrush.csv")
attach(Toothbrush) # note 'attach' function
plot(Price, Function)
fit = lm(Function ~ Price) # regression of y on x
fit
plot(Price, Function)
abline(fit)
attributes(fit)
fit$residuals
fit$coefficients[1]
fit$coefficients[2]
hist(fit$residuals)
```

Recall: Toothbrush – function v price

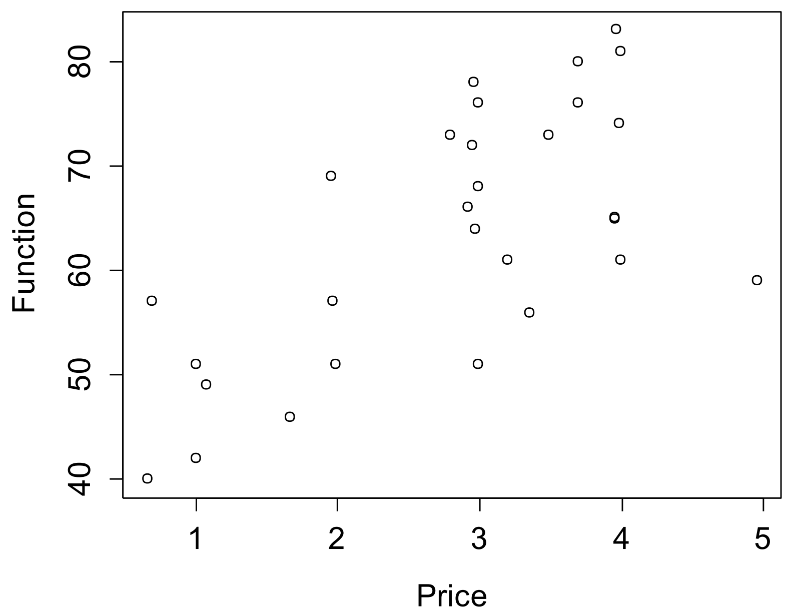
```
> Toothbrush <- read.csv("Toothbrush.csv")
> attach(Toothbrush) # note 'attach' function
> plot(Price, Function)
```



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
```

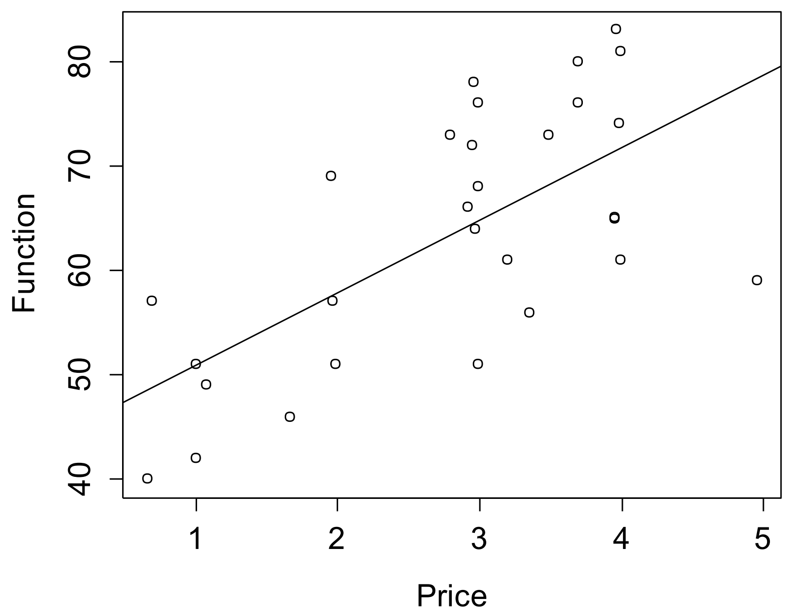
However, the linear model object contains much more information than just the coefficients!

Line of best fit

This has been covered but worth remembering

```
>   plot(Price, Function)
>   abline(fit)
```

Intercept and gradient 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:

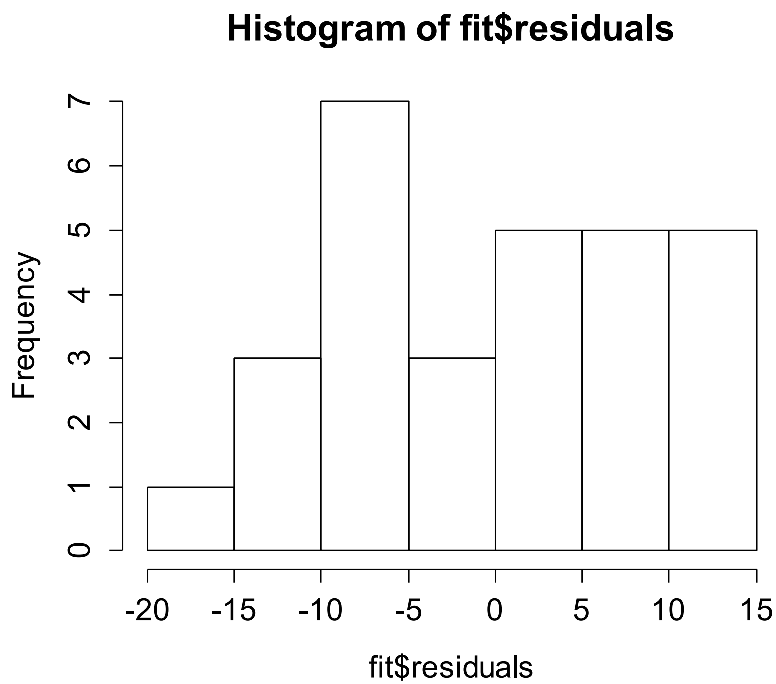
```
> fit$coefficients[1]
(Intercept)
44.01954
> fit$coefficients[2]
Price
6.942303
```

Index refers to specific element in “column”

Diagnostics – residuals

Ideally, residuals should be normally distributed.

```
> hist(fit$residuals)
```

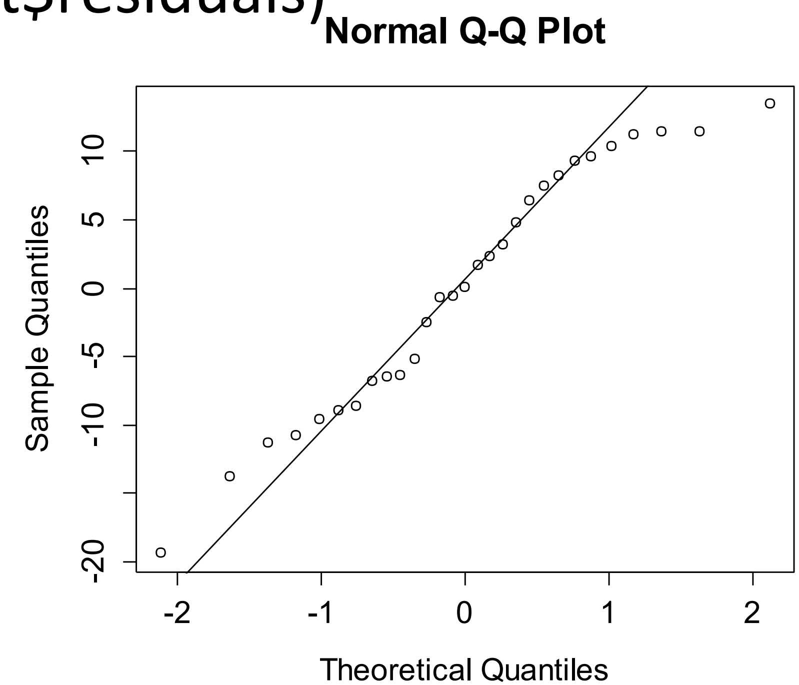


Not conclusive!

Diagnostics – residuals

A normal quantile plot is a better visual reference

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

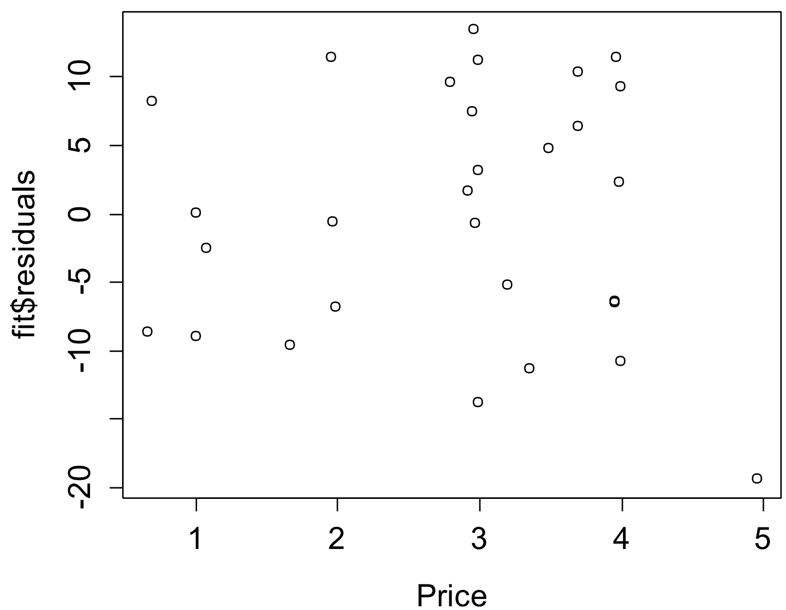


Good fit
for *-1 < z < 1*

Diagnostics – residuals

Residuals should be uncorrelated with input

```
> plot(Price, fit$residuals)
```

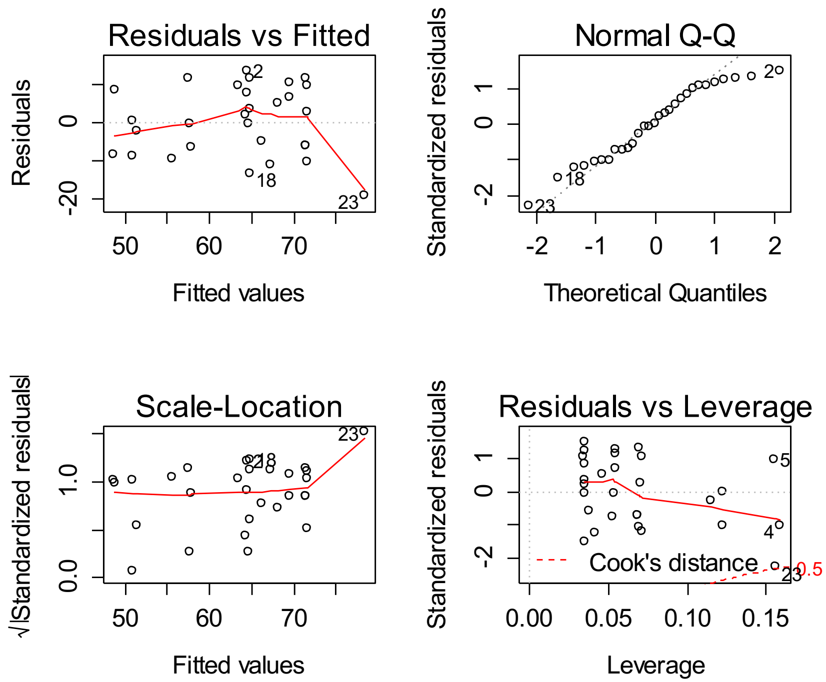


By eye $r \approx 0$

Diagnostics – residuals

R gives 4 default plots as a summary:

```
> par(mfrow =c(2,2)) # creates a 2 x 2 matrix for plots
> plot(fit)
```



Diagnostics – summary

Median close to 0

```
> 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

Coefficients: α, β

```
> 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
```

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, 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 $\alpha, \beta = 0$ vs $\alpha, \beta \neq 0$

Coefficient of Determination: r^2

Overall significance of regression: that at least one coefficient $\neq 0$

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.

Multiple linear regression

OLS applied to multiple predictors, assumptions:

- The relationship is now of the form:
  $y \approx a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + b$, *or*
  $y = a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + b + e$, *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

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
```

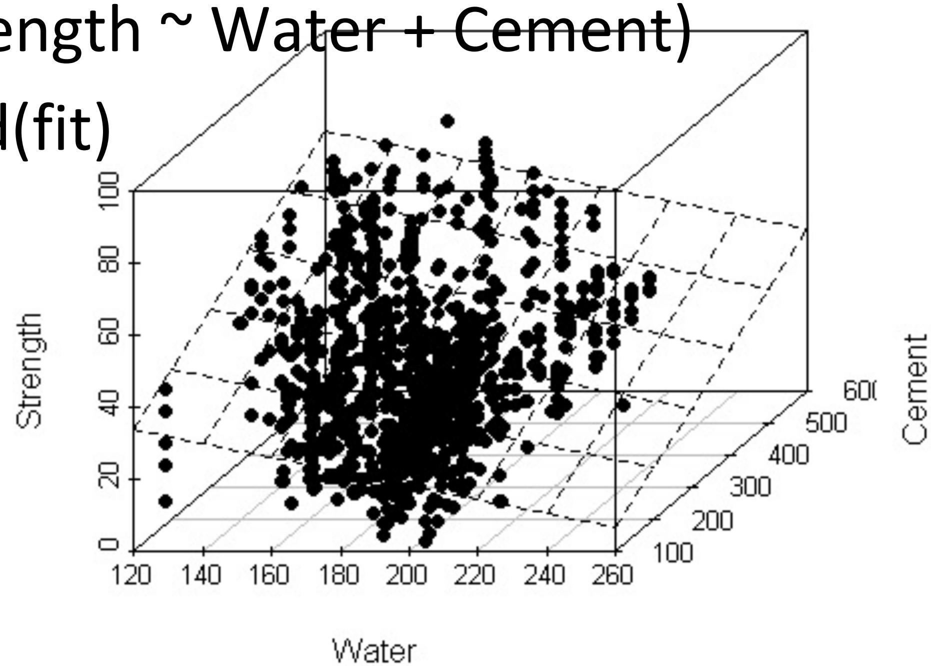
Summary

\> 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 mean 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 (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

Qualitative (or categorical) predictors include: gender, hair/eye colour, season, job type etc.

- When the variable has more than two factor levels, each factor level is included as a variable in the regression equation. Indicator (0, 1) variables show the status of each observation at each factor level. See below:

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 Tutorial 2:

```
> library(ggplot2)
> set.seed(9999) # Random seed
> dsmall <- diamonds[sample(nrow(diamonds), 1000), ]
  # sample of 1000 rows
> qplot(carat, price, data = dsmall, color = color, size =
  clarity, alpha = cut)
```

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

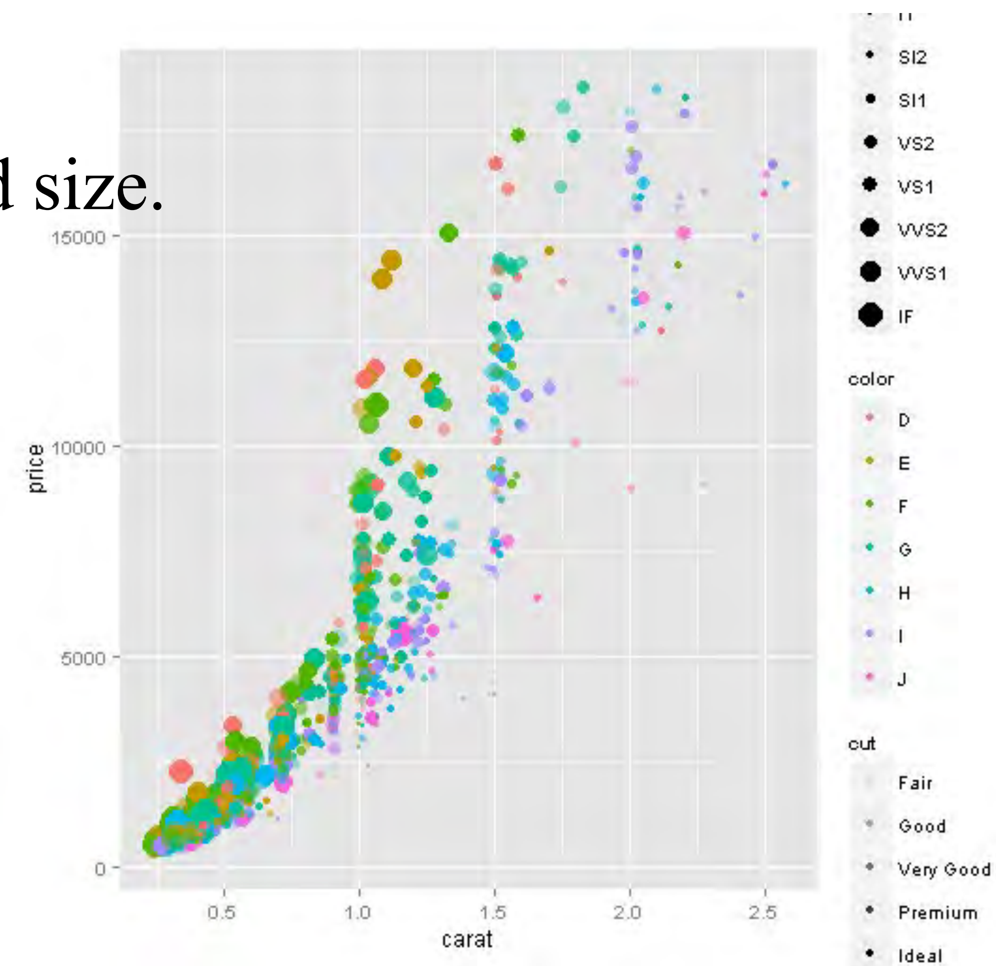
Non-linear:

- Take logs of price and size.

Categorical variables:

- Clarity
- Color
- Cut

Note that data appears exponential in both x and y

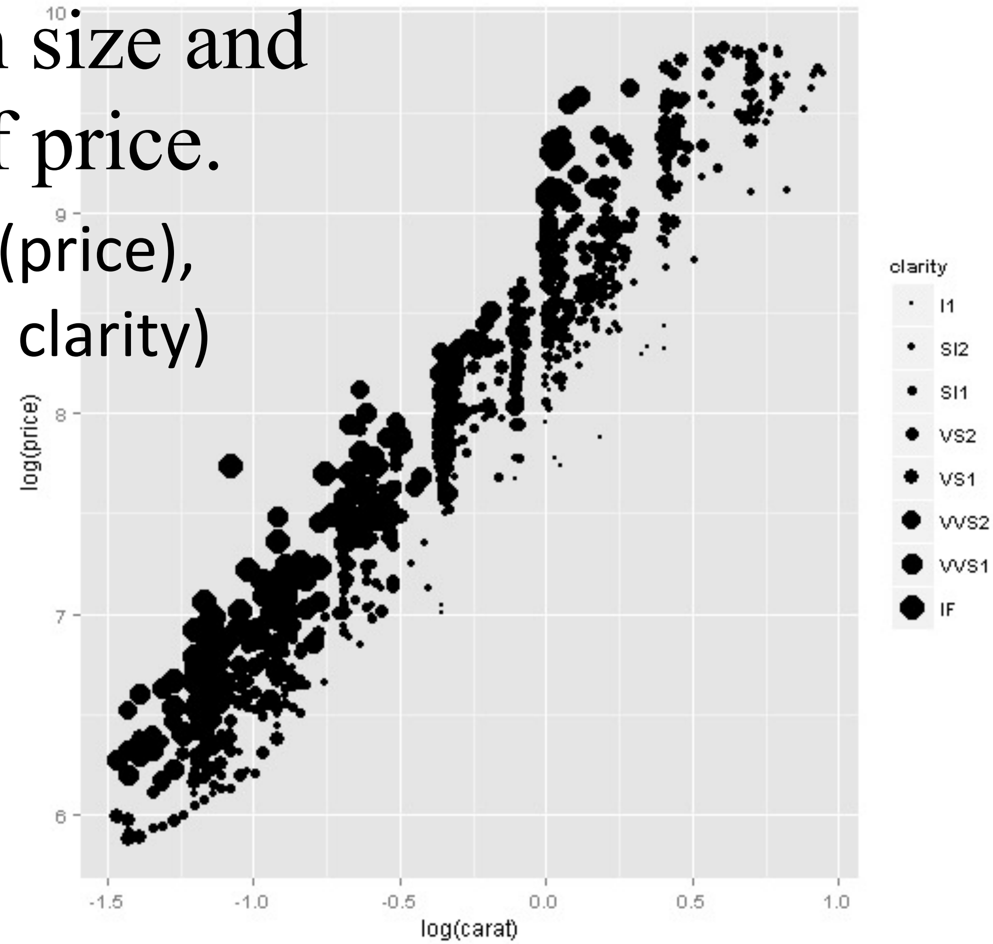


Plot using log scale

Concentrating only on size and clarity as predictors of price.

> qplot(log(carat), log(price), data = dsmall, size = clarity)

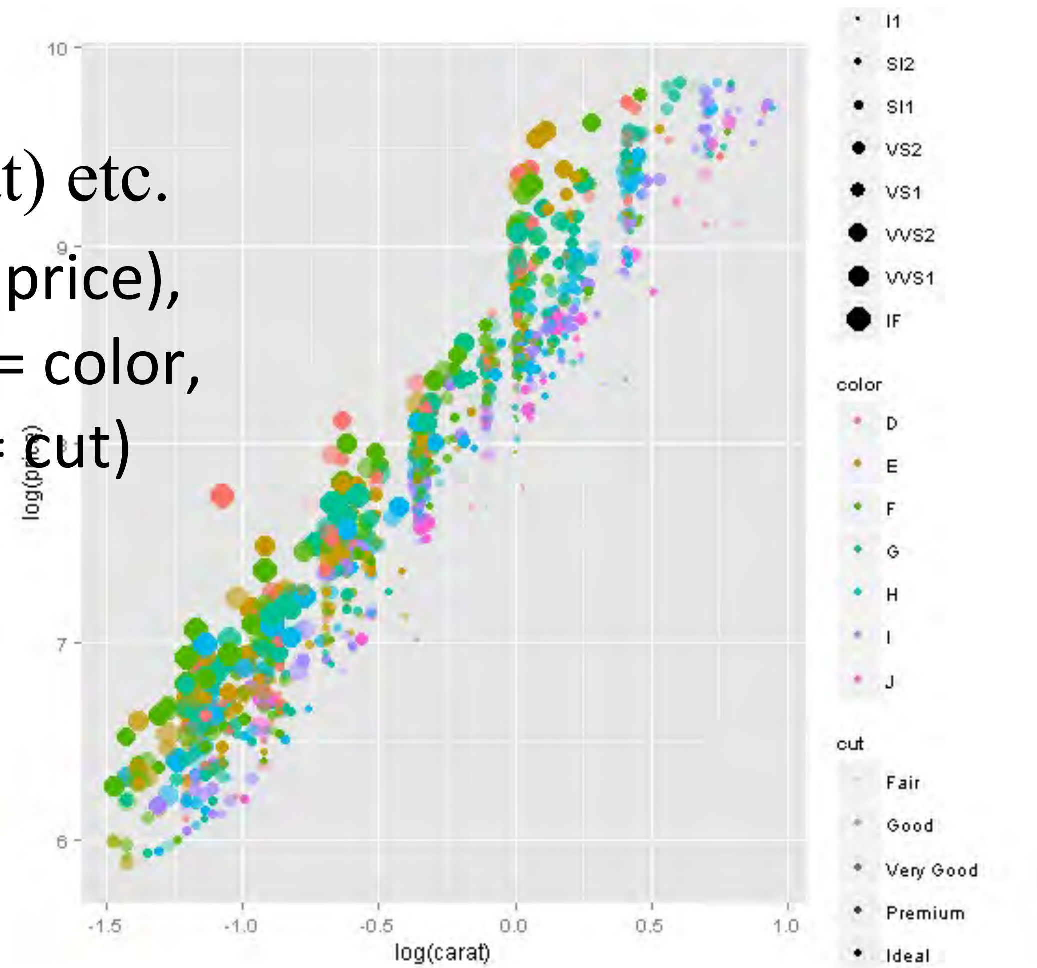
- Note, R uses:
  > log to mean ln or $\log_e$
  > log10 for log base 10
  > Clarity has 8 levels



Plot using all variables

Linear relationship

- Natural logs: $\log_e$(carat) etc.
  - > qplot(log(carat), log(price), data = dsmall, color = color, size = clarity, alpha = cut)



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
```

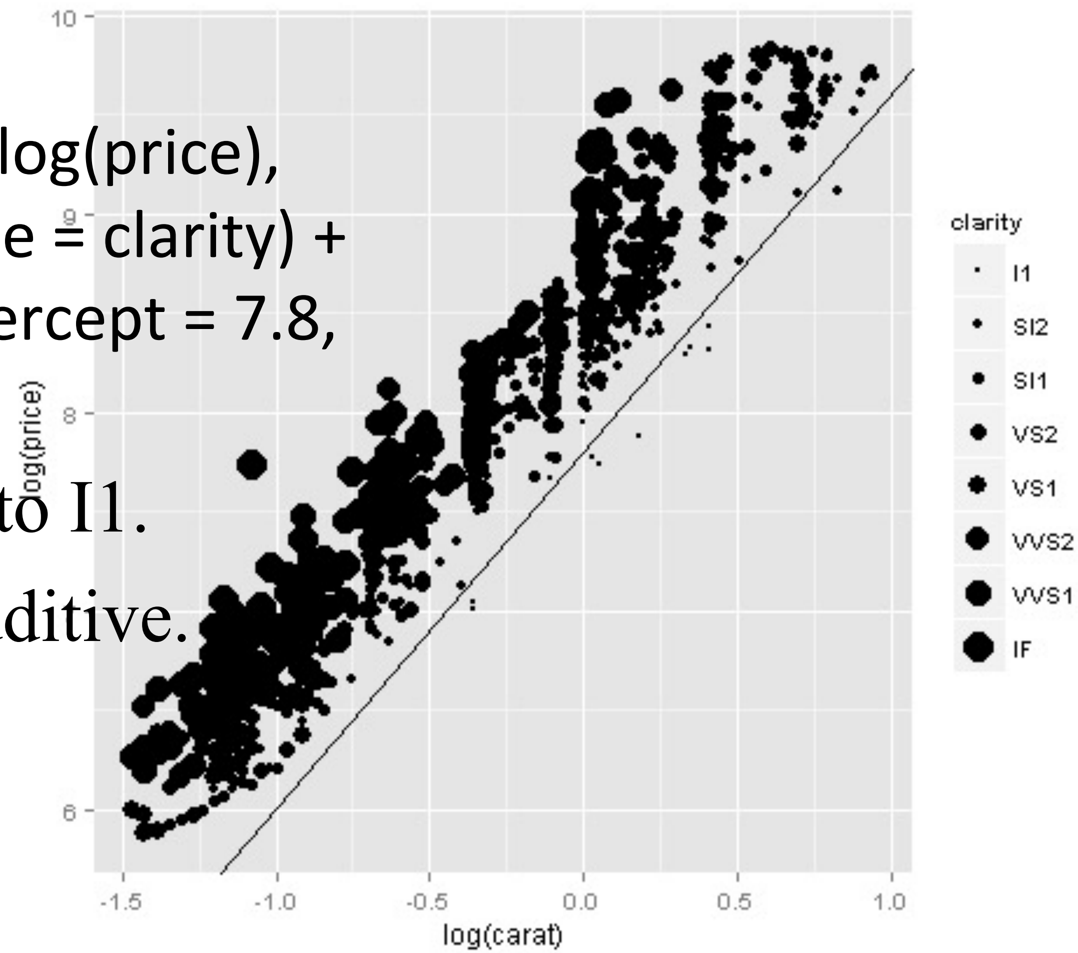
Summary (overall)

```
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)

```
> qplot(log(carat), log(price),
  data = dsmall, size = clarity) +
  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
```

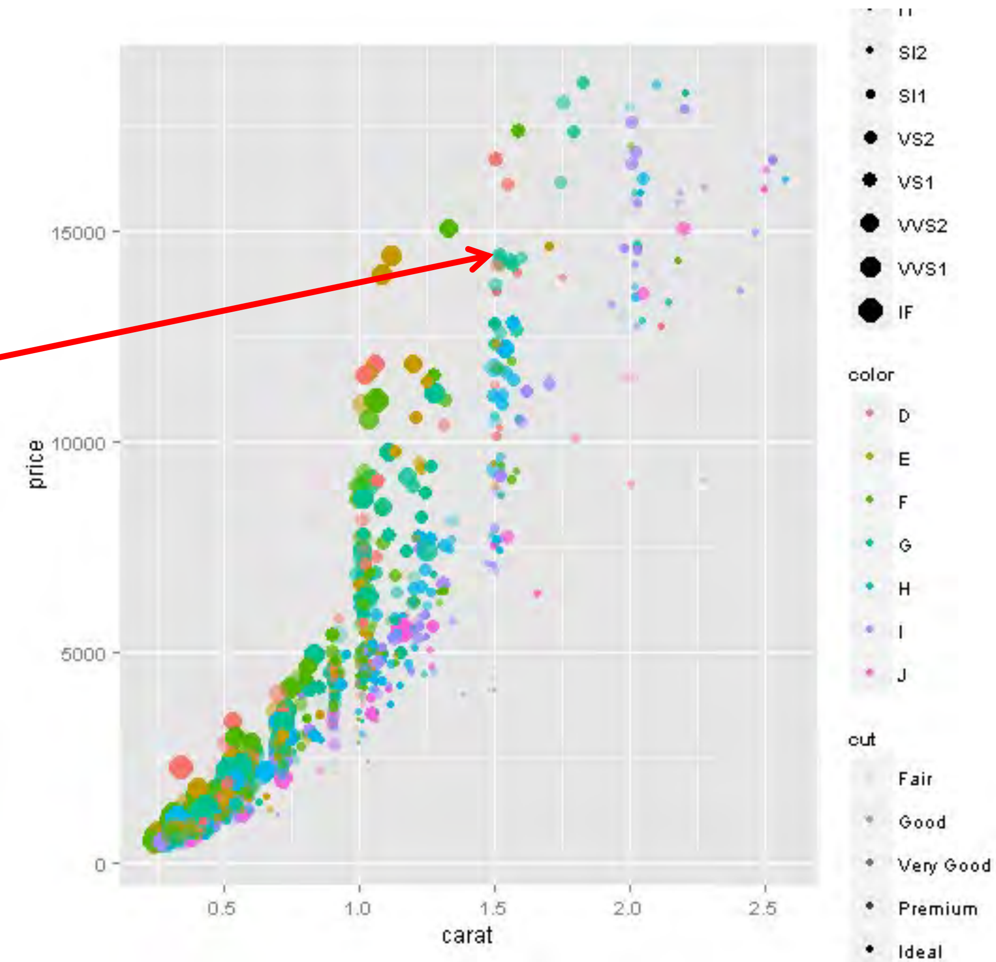
Raising each side to the power of e^x

```
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
```



Other types of regression

There are many other regression models in addition to those covered today. Some examples from ATHR P65.

Model	Formula	
$y = \beta_0 + \beta_1 x + e$	`y ~ x`	Simple regression
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$	`y ~ x1+x2`	Multiple regression
$y = \beta_0 + e$	`y ~ 1`	Intercept only (null) model
$y = \beta_1 x + e$	`y ~ 0+x`	Slope only
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$	`y ~ x1*x2`	Main effects and products
	`y ~ x1+x2+x1:x2`	
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$	`y ~ x+I(x^2)`	Quadratic term
$ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$	`log(y) ~ x1+x2`	Log dependent

Solutions to review questions

1. D
2. C
3. D
4. F
5. C

Summary

OLS regression

Regression diagnostics

Multiple regression

Indicator variables

Next week: Supervised learning: Decision trees

Following weeks: improving the basic tree:

- Classification, testing and fitting a model

Unsupervised techniques:

- Clustering, Text mining
- Comparison of techniques

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.