FIT3152 Data analytics– Lecture 6

Regression

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

Week-by-week

Week Starting	Lecture	Topic	Tutorial	A1	A2
2/3/21	1	Intro to Data Science, review of basic statistics using R	...		
9/3/21	2	Exploring data using graphics in R	T1		
16/3/21	3	Data manipulation in R	T2	Released	
23/3/21	4	Data Science methodologies, dirty/clean/tidy data, data manipulation	T3		
30/3/21	5	Network analysis	T4		
6/4/21		Mid-semester Break			
13/4/21	6	Regression modelling	T5		
20/4/21	7	Classification using decision trees	T6	Submitted	
27/4/21	8	Naïve Bayes, evaluating classifiers	T7		Released
4/5/21	9	Ensemble methods, artificial neural networks	T8		
11/5/21	10	Clustering	T9		
18/5/21	11	Text analysis	T10		Submitted
25/5/21	12	Review of course, Exam preparation	T11		

Assignment 1

FIT3152 Data analytics: Assignment 1

This assignment is worth 20% of your final marks in FIT3152. Due: Friday 23rd April 2021.

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:

Assignment 1

a. Analyse activity and language on the forum over time. Some starting points:

- Describe your data: How active are participants, and are there periods where this increases or decreases? Is there a trend over time?
- Looking at the linguistic variables, do these change over time? Is there a relationship between variables?

b. Analyse the language used by groups. Some starting points:

- Threads indicate groups of participants communicating on the same topic. Describe the threads present in your data.
- By analysing the linguistic variables for all or some of the threads, is it possible to see a difference in the language used by different groups?
- Does the language used within threads (or between threads) change over time? How consistent or variable is the language used within threads?

Assignment 1

c. Challenge: Social networks online. We can think of participants posting to the same thread at similar times (for example during the same month) as forming a social network. When these participants also post to other threads over the same period, their social network extends.

- Can you define, graph and describe the social network that exists at a particular point in time, for example over one month? How does this change in the following months?
- Note: you only need to analyse a small portion of the social network over a short time period. We will cover social network analysis in Lecture 5.

d. Reflection on your investigation. What did you first investigate? How did you then modify your research based on the results of your first investigation?

- Using one of the data science methodologies in Lecture 4, illustrate your research process.

Assignment 1

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

ThreadID	AuthorID	Date	Time	WC	Analytic	Clout	Authentic	Tone	WPS	i	we	you	they	number	affect	posemo	negemo	anx
659289	193537	24/11/2009	5:36	53	82.26	71.43	25.14	25.77	26.5	0	1.89	0	3.77	3.77	3.77	1.89	1.89	0
432269	136196	26/11/2007	23:42	216	25.71	94.73	45.81	33.77	24	1.85	6.48	0.46	5.09	0.46	6.02	3.24	2.78	0
572531	170305	17/02/2009	7:31	136	31.61	67.04	28.81	79.41	13.6	3.68	0	5.15	2.94	0.74	9.56	5.88	2.94	0.74
230003	32359	7/09/2005	21:25	29	39.74	91.6	3.81	85.87	14.5	3.45	0	6.9	0	6.9	3.45	3.45	0	0
459059	47875	19/02/2008	5:23	108	80.75	60.95	23.51	88.52	13.5	2.78	0	0	0	0.93	9.26	6.48	2.78	0
635953	181593	28/09/2009	8:40	86	64.98	45.37	57.24	1	43	1.16	0	0	5.81	3.49	3.49	0	3.49	0
235116	51993	29/09/2005	15:59	49	33.33	20.71	13.15	25.77	16.33	6.12	0	0	2.04	0	8.16	4.08	4.08	0
593767	169459	23/04/2009	19:21	368	85.91	63.82	19.13	7.15	24.53	1.36	2.17	0	0.54	0.54	5.43	1.9	3.53	0.54
532649	248548	25/12/2011	8:28	13	92.84	50	1	25.77	13	0	0	0	0	61.54	0	0	0	0
517685	65	20/02/2005	10:50	65	91.21	62.1	33.6	81.28	13	7.69	0	0	0	0	9.23	6.15	3.08	0
588291	158329	23/04/2009	23:40	265	55.7	73.95	45.85	11.21	44.17	1.89	1.13	0.38	3.4	5.66	3.4	1.13	2.26	0
29936	194	25/07/2002	4:29	106	80.44	80.2	20.42	98.46	15.14	1.89	0	4.72	0	0.94	7.55	6.6	0.94	0.94
199787	47875	20/05/2005	16:48	160	94.48	73.4	2.07	5.64	22.86	1.25	0	0	0	5.62	8.12	3.12	5	1.88
545552	143229	24/11/2008	23:39	33	79.25	18.16	98.01	80.64	8.25	6.06	0	0	0	3.03	3.03	3.03	0	0
303058	88912	25/07/2006	23:57	244	44.21	65.92	33.49	7.09	27.11	2.87	0.82	0.41	4.51	1.64	6.56	2.46	4.1	0
772248	75628	16/01/2011	2:24	108	39.91	57.35	45.81	25.77	13.5	5.56	0	2.78	0	0.93	1.85	0.93	0.93	0
761807	227011	4/12/2010	23:48	104	73.9	57.63	74.76	62.24	34.67	0.96	0	2.88	3.85	2.88	5.77	3.85	1.92	0
110837	34501	24/01/2004	2:53	49	90.62	20.71	46.05	1	24.5	2.04	0	0	0	0	6.12	0	6.12	0
636255	180475	3/09/2009	22:25	2	92.84	99	1	99	2	0	0	0	0	0	50	50	0	0
178736	43291	18/01/2005	2:40	75	69.57	92.87	1	1	15	0	0	2.67	6.67	0	10.67	1.33	9.33	1.33
275754	-1	6/03/2006	18:01	56	92.84	70.4	41.07	6.15	18.67	1.79	0	1.79	0	1.79	1.79	0	1.79	0
833308	231141	21/09/2011	21:39	32	78.67	82.58	74.76	25.77	16	0	0	6.25	0	0	0	0	0	0
642657	180098	13/11/2009	16:34	13	92.84	6.21	99	1	13	23.08	0	0	0	0	7.69	0	7.69	7.69
365246	116735	17/02/2007	9:48	48	49.05	33.83	62.53	1	48	2.08	0	2.08	2.08	0	10.42	2.08	8.33	4.17
279233	84070	21/03/2006	1:59	51	77.76	50	66.34	25.77	51	3.92	0	1.96	0	1.96	7.84	3.92	3.92	0
300539	-1	8/06/2006	22:43	24	49.05	33.83	23.51	92.4	6	8.33	0	0	4.17	8.33	4.17	4.17	0	0
277955	32925	14/03/2006	23:45	87	55.99	78.96	62.98	3.63	43.5	0	0	1.15	4.6	2.3	2.3	0	2.3	1.15
90325	32485	25/09/2003	3:30	48	94.65	79.76	3.9	25.77	12	0	0	0	2.08	2.08	12.5	6.25	6.25	0
321495	90627	12/09/2006	1:40	42	40.66	68.29	37.24	70.57	21	4.76	4.76	2.38	2.38	0	2.38	2.38	0	0
281667	79878	28/03/2006	2:45	60	32.98	56.63	65.14	1.03	20	1.67	1.67	0	3.33	0	3.33	0	3.33	0
294983	75902	21/05/2006	0:07	60	56.15	25.24	32.84	25.77	60	3.33	0	0	0	0	6.67	3.33	3.33	0
397699	125170	21/06/2007	21:41	34	92.84	92.92	14.7	25.77	17	0	2.94	2.94	0	0	5.88	2.94	2.94	0
313191	101368	30/07/2006	17:53	25	81.4	2.31	43.37	25.77	25	0	0	0	0	12	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Assignment 1

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

Column	Brief Descriptor
ThreadID	Unique ID for each thread
AuthorID	Unique ID for each author
Date	Date
Time	Time
WC	Word count of the text of the post
Analytic	LIWC Summary (Analytical thinking)
Clout	LIWC Summary (Power, force, impact)
Authentic	LIWC Summary (Using an authentic tone of voice)
Tone	LIWC Summary (Emotional tone)
WPS	LIWC (Words per sentence)
i	LIWC ("I, me, mine" words) First person singular
we	LIWC ("We, us, our" words) First person plural
you	LIWC ("You" words) Second person
they	LIWC ("They" words) Third person plural
number	LIWC(Quantities and ranks)
affect	LIWC (Expressing sentiment)
posemo	LIWC (Positive emotions)
negemo	LIWC (Negative emotions)
anx	LIWC (Indicating anxiety)

Assignment 1

Submission. Due Friday 23rd April 2021 11:55pm GMT+10.
Suggested length: 6–8 A4 pages + appendix.
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. Submit your report as a single PDF with the file name *FirstnameSecondnameID.pdf* on Moodle.

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

Assessment criteria will include:

The quality of your analysis and description of your analytical process; Graphics and tables supporting your analysis; The quality of graphics used in the report. Justification of your findings and the degree of proof you can offer (for example statistical tests); Readability and quality of your written report; Insights gained from the data; Novelty of your approach.

Factors you should consider (starting points, not a complete list):

Techniques: summary/descriptive statistics, identification of important variables, networks, etc.
Major grouping variables: author, thread, date and/or time, or a combination of these.
Time window (days, weeks, months, years…); Subsets of the data to be analysed.
Graphics to communicate your analysis and insights (histograms, scatterplots, heat maps, time series are some basic starting points, but see https://datavizproject.com/ for inspiration.

Response to student questions

- Can I filter my thread IDs based on frequency to only analyse those with a larger number of posts. I’m thinking there just too many threads otherwise.
  - > You could focus on threads having more than a certain number of posts, or you choose to analyse the top 10 or 20 etc. threads having the most posts.
- Do we need a cover page for the report? If so, is it included in the page limits?
  - > We don't require a cover page for your report. If you use one it will not count towards the page limit.

Network review questions

Feel free to type your responses in the chat…

Question 1

For the graph below, *diameter* is:

a. 1

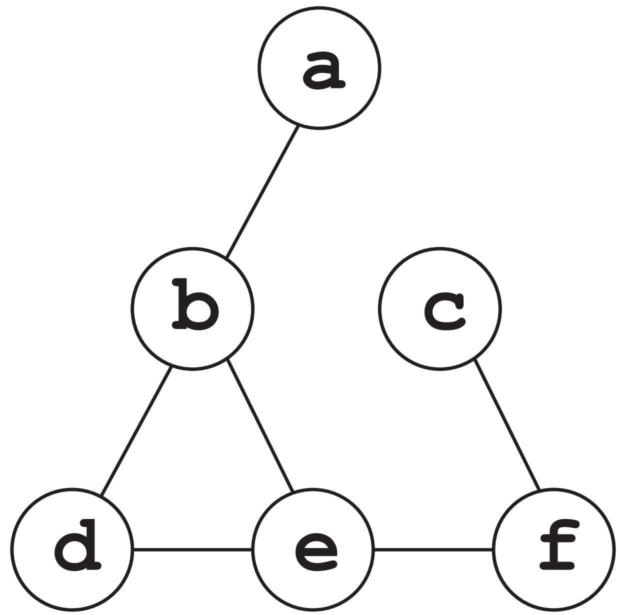
b. 2

c. 3

d. 4

e. 5

f. 6



Question 2

For the graph below, d_b is:

a. 1

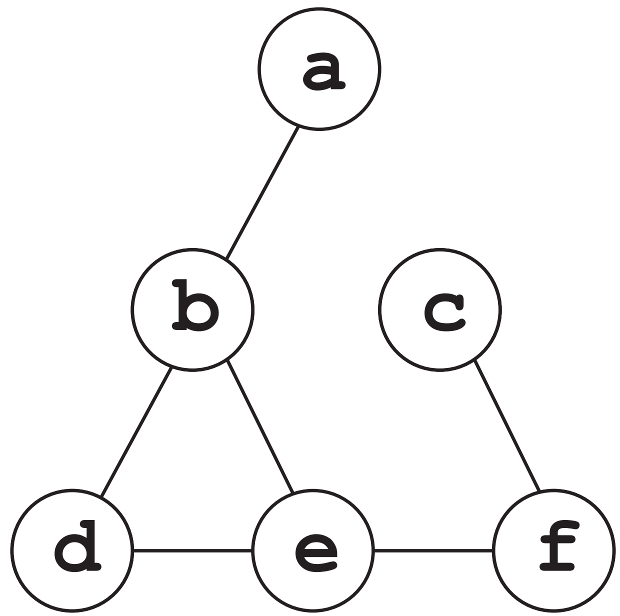
b. 2

c. 3

d. 4

e. 5

f. 6



Question 3

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

a. 1

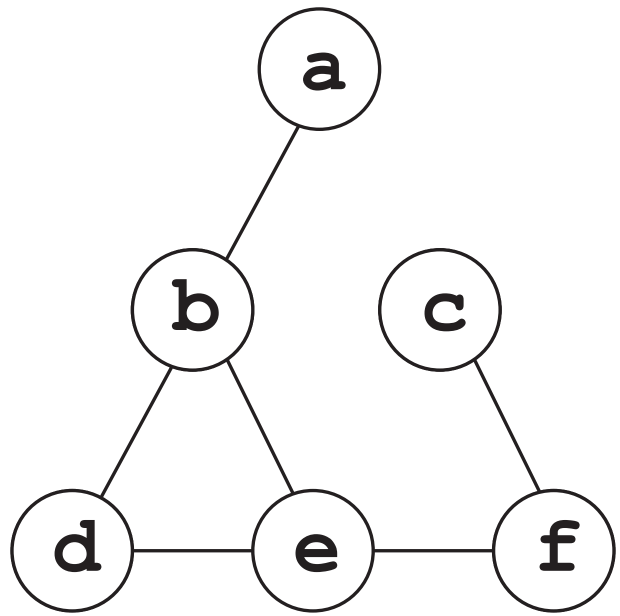
b. 2

c. 3

d. 4

e. 5

f. 6



Question 4

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

a. 1/1

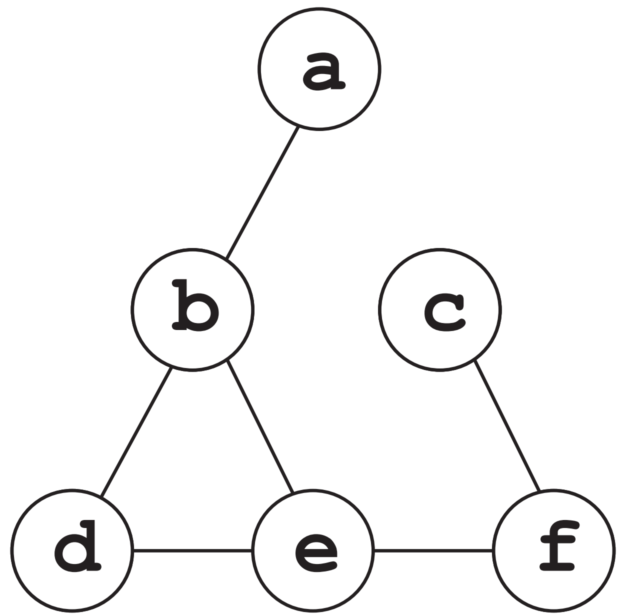
b. 1/2

c. 1/3

d. 1/4

e. 1/6

f. 1/8



Question 5

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

a. 1

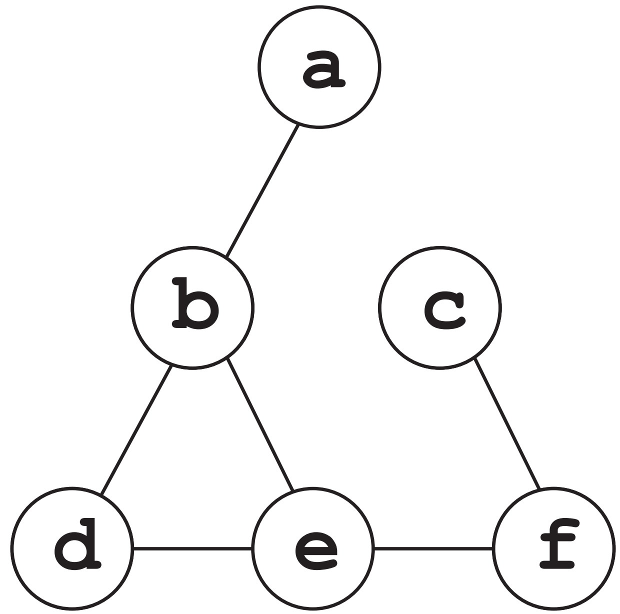
b. 2

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

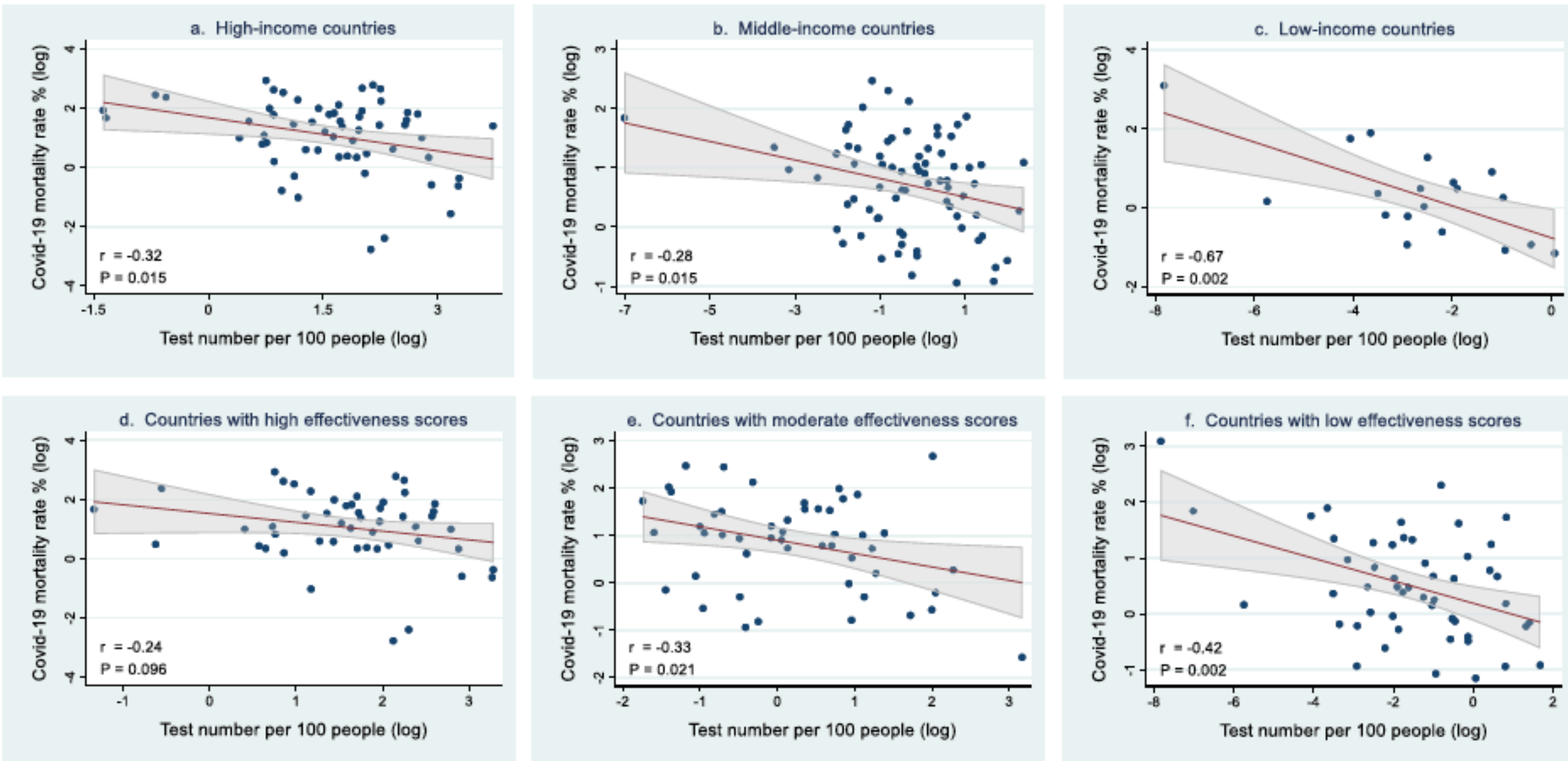
COVID-19

	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).

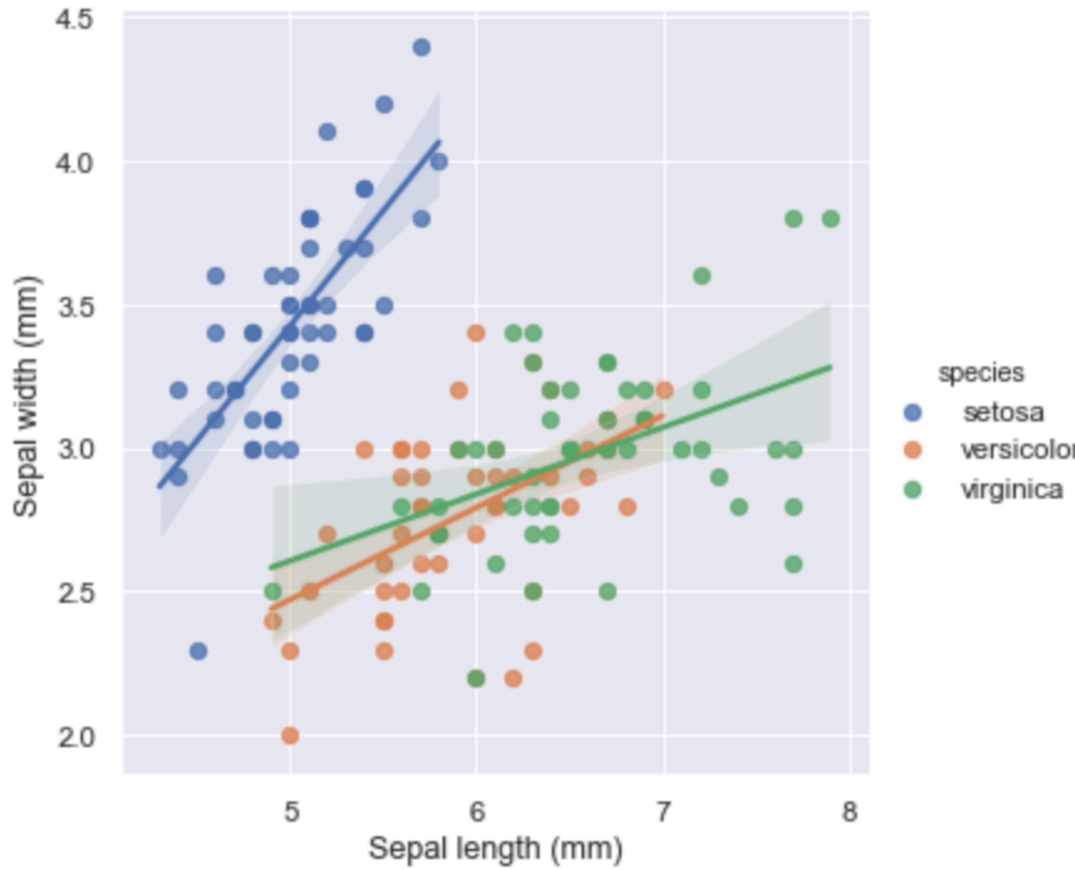
https://www.nature.com/articles/s41598-020-68862-x

COVID-19



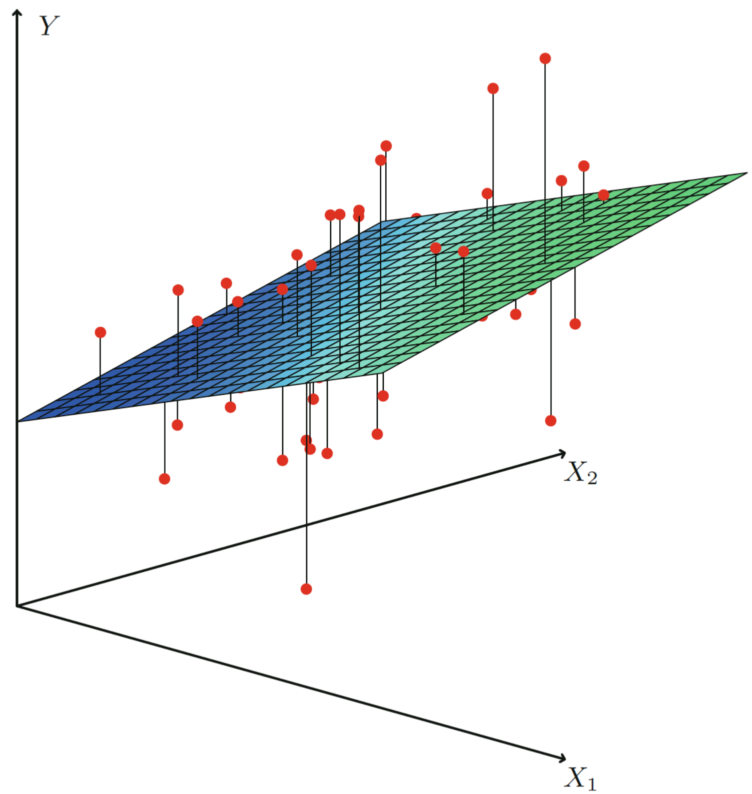
https://www.nature.com/articles/s41598-020-68862-x

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 (2013).

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

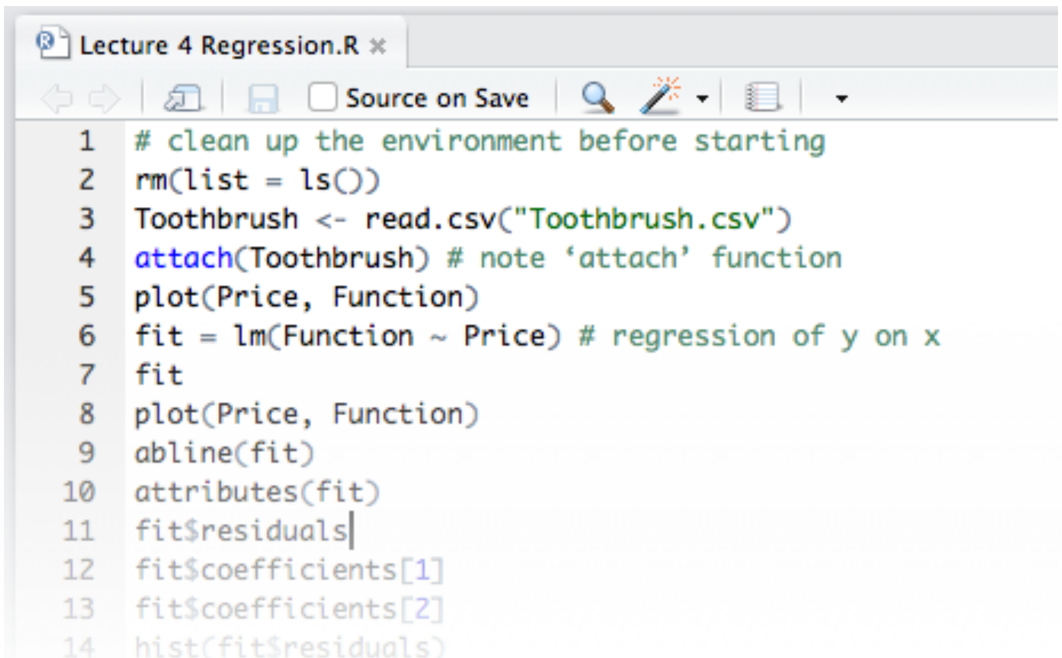
- Fitting a regression model is a form of supervised learning – that is, the model is 'learned' from data consisting of known inputs and outputs.

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

Linear regression

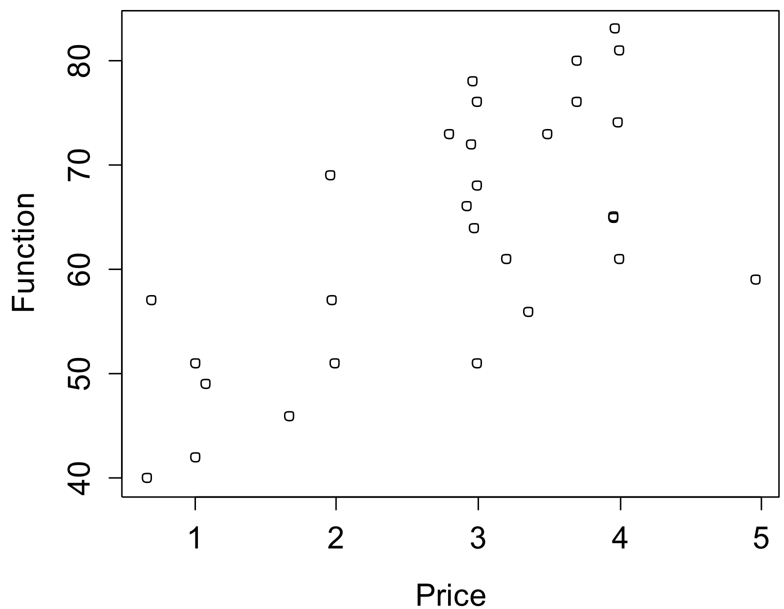
See R Script of lecture examples

> Lecture 6 Regression.R



Recall: Toothbrush – function v price

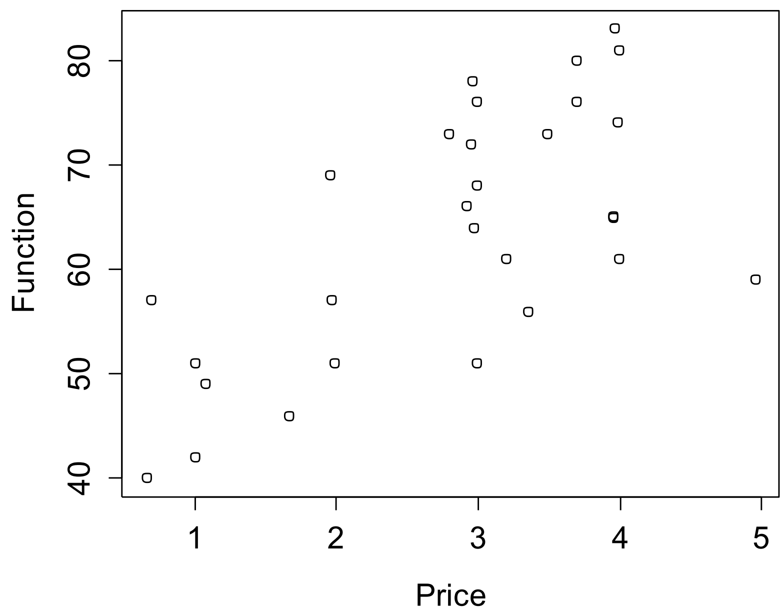
```
> Toothbrush <- read.csv("Toothbrush.csv")
> attach(Toothbrush) # note 'attach' function
> plot(Price, Function)
```



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:

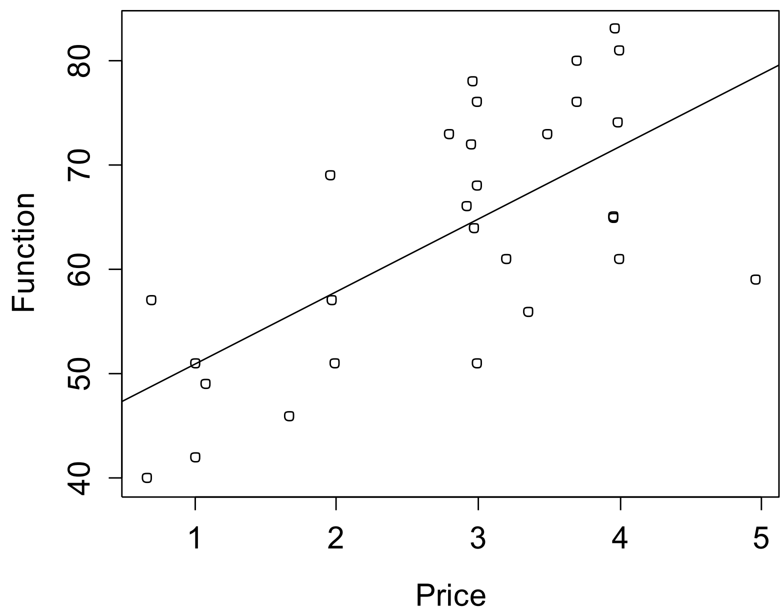
```
> fit = lm(Function ~ Price) # regression of y on x
> fit
Call:
lm(formula = Function ~ Price)
Coefficients:
(Intercept)        Price
     44.020        6.942
```

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)
```



Linear model object

To see the details of what the object contains use:

```
>   attributes(fit)
$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
...
```

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

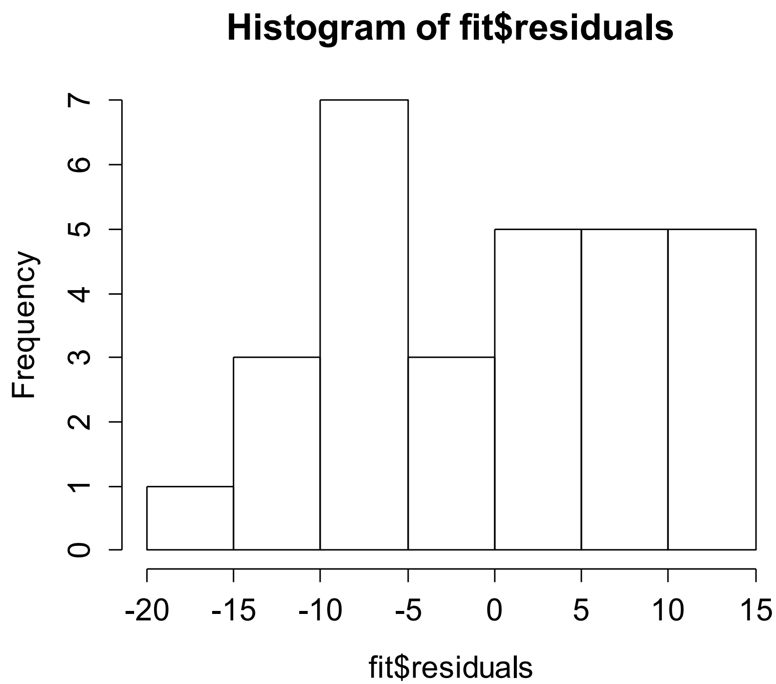
Intercept and slope can be addressed directly as:

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

Diagnostics – residuals

Ideally, residuals should be normally distributed.

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

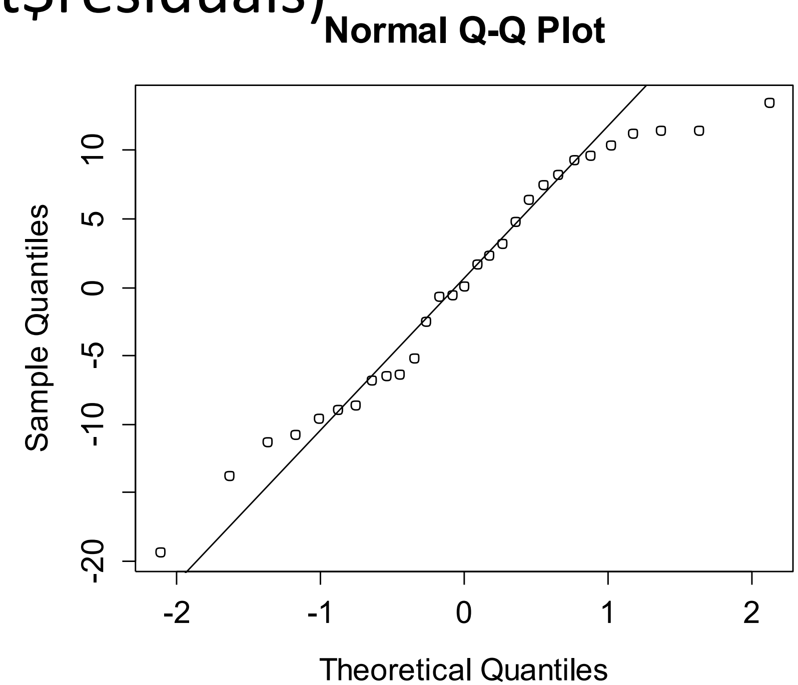


Not conclusive!

Diagnostics – residuals

A normal quantile plot is a better visual reference

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

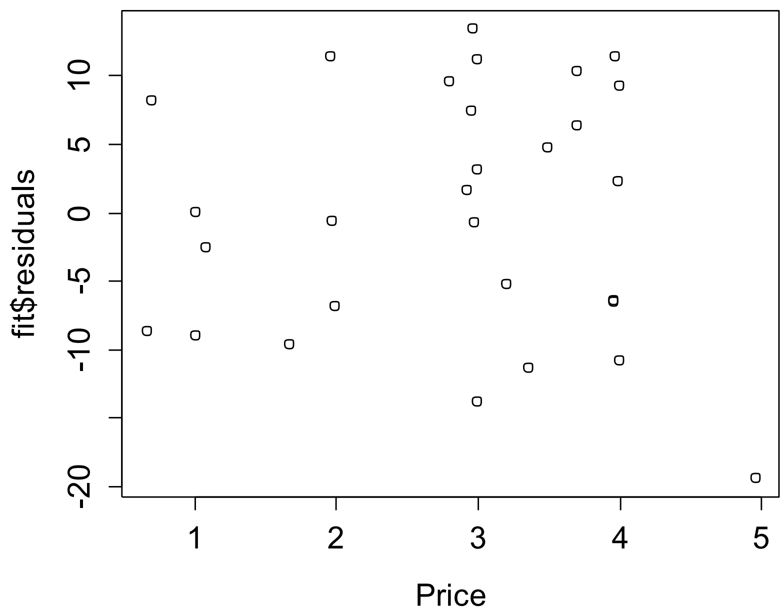


Good fit
for *-1 < z < 1*

Diagnostics – residuals

Residuals should be uncorrelated with input

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

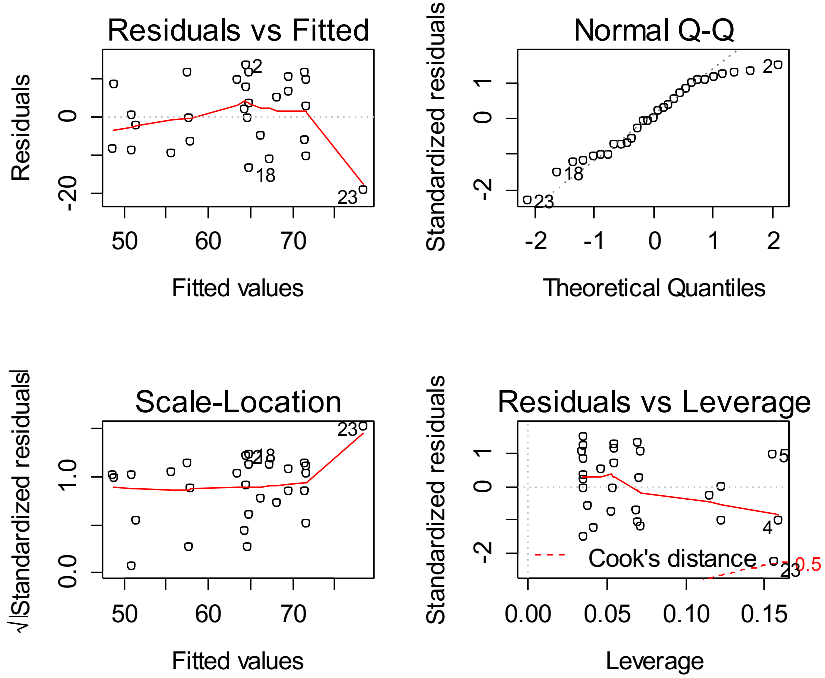


By eye $r \approx 0$

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

```
> 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, 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
```

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

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:
  $y \approx a_1x_1 + a_2x_2 + a_3x_3 + \ldots + b$, *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

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
```

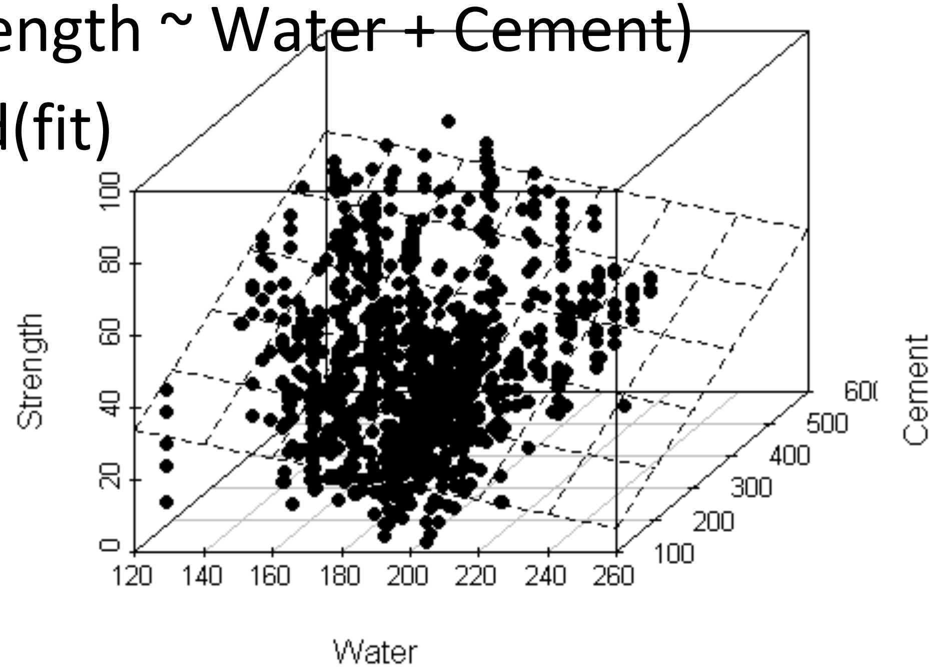
Summary

> 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

```
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>
```

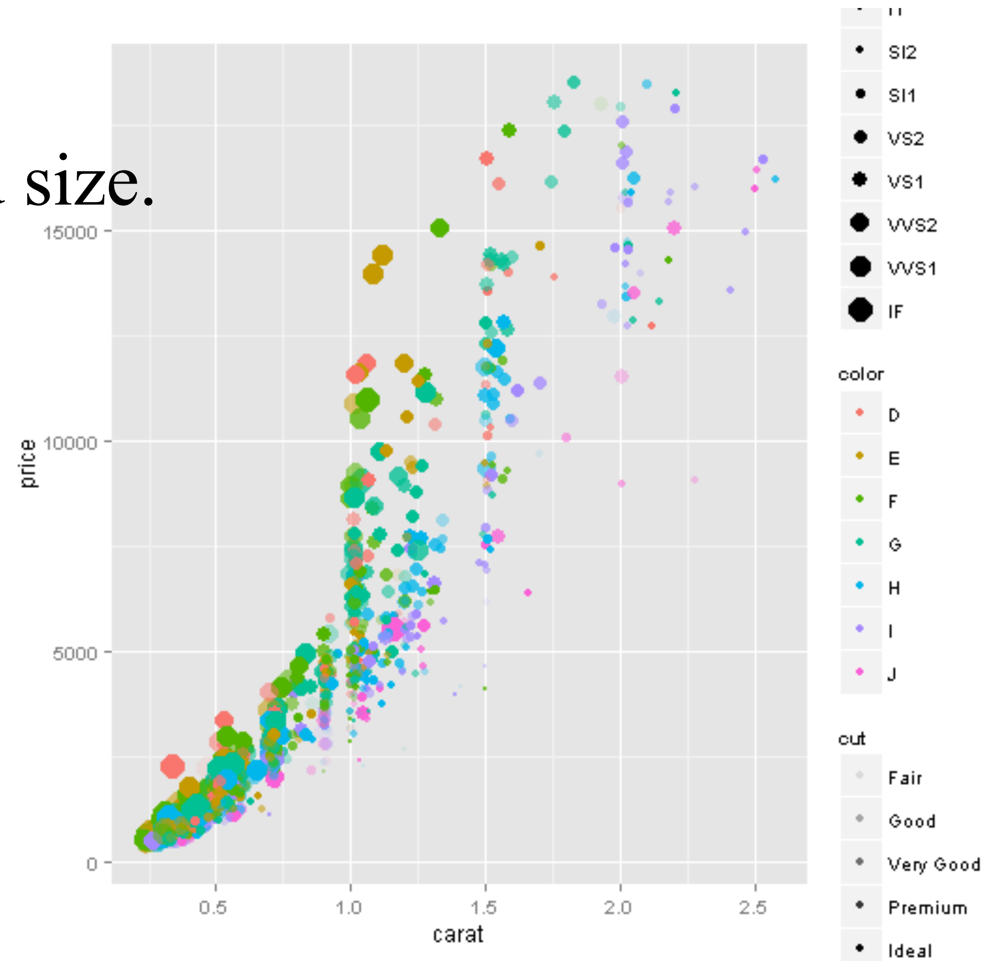
Basic plot: first observations

Non-linear:

- Take logs of price and size.

Categorical variables:

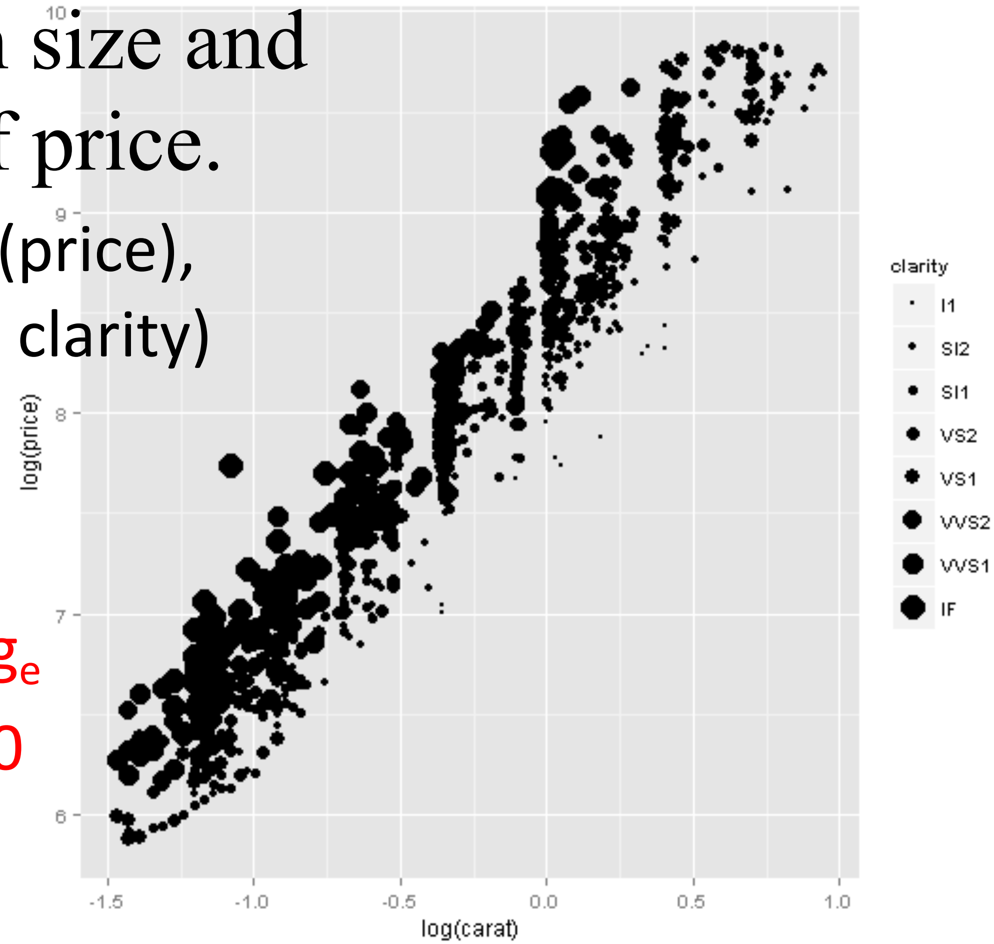
- Clarity
- Color
- Cut



Plot using log scale

Concentrating only on size and clarity as predictors of price.

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



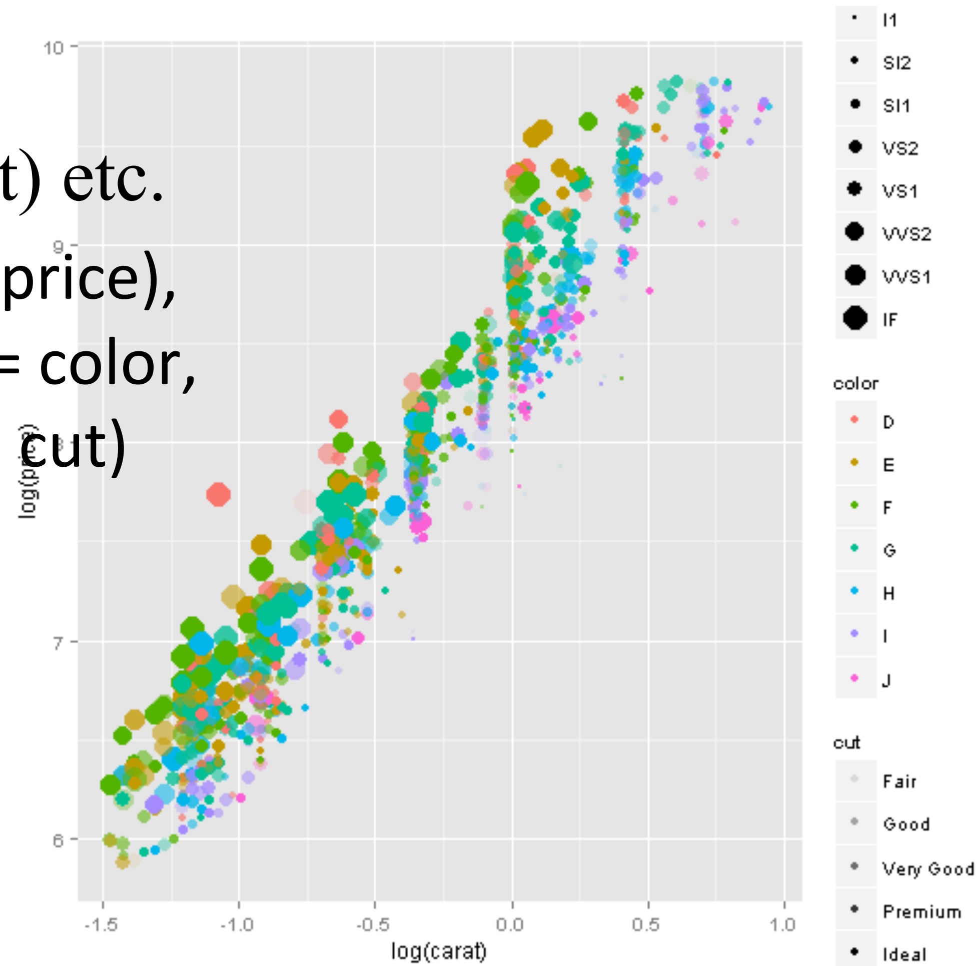
- Note, R uses:
  > log to mean ln or log_e

  > log10 for log base 10

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
```

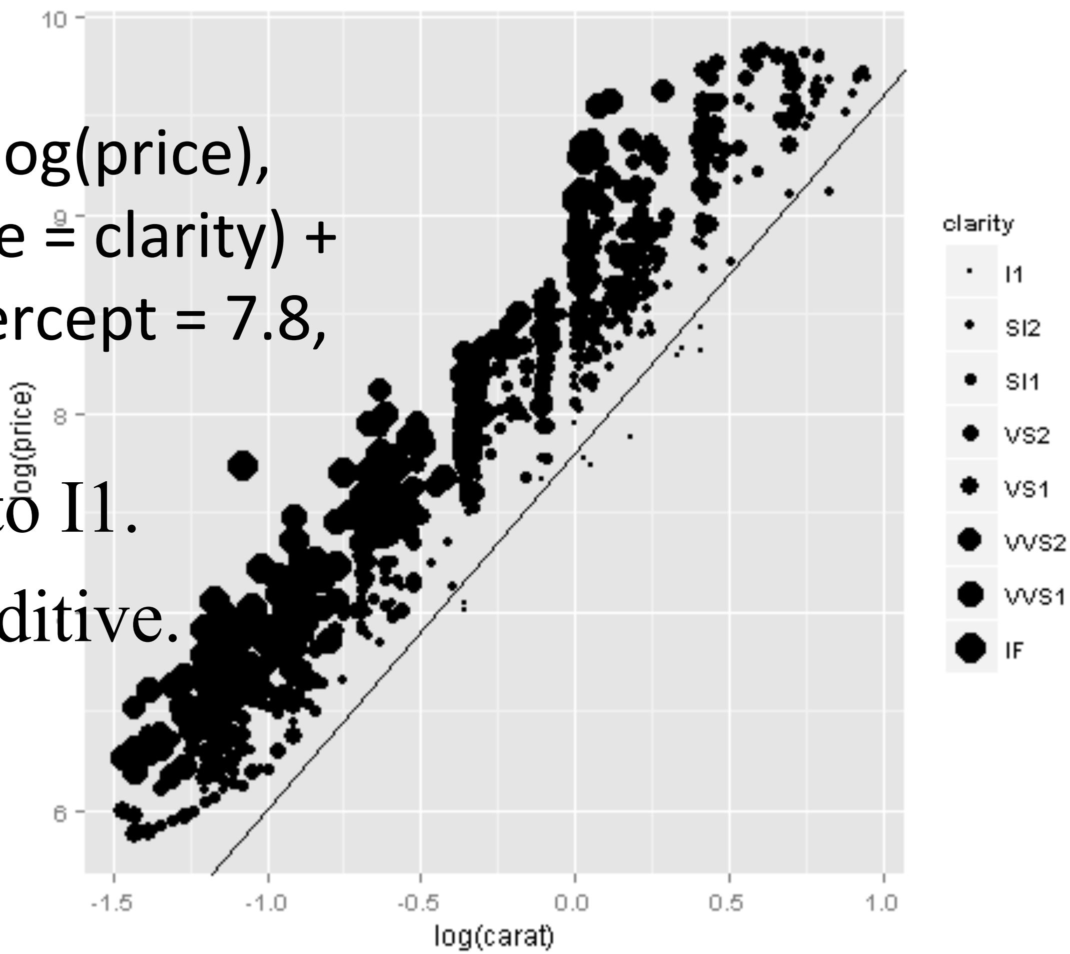
Summary (overall)

```
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?

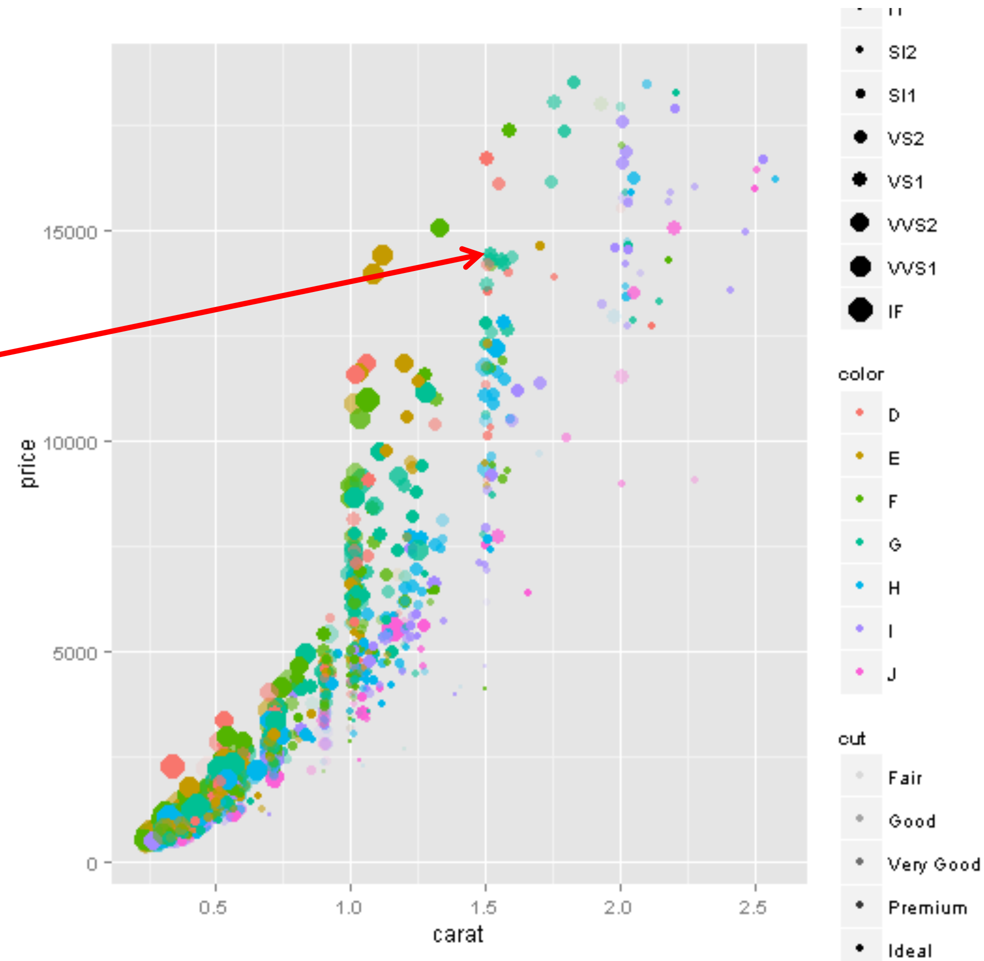
```
Log(y) = log(price) = log(carat) * log(x) (+ intercept) + clarity
         log(price) = 1.8324 * log(1.5) + 7.7884 + 1.0290
         log(price) = 1.8324 * 0.4055  + 7.7884 + 1.0290
         log(price) = 9.5603
              price  = $14,191
```

```
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 (2013)

- 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.