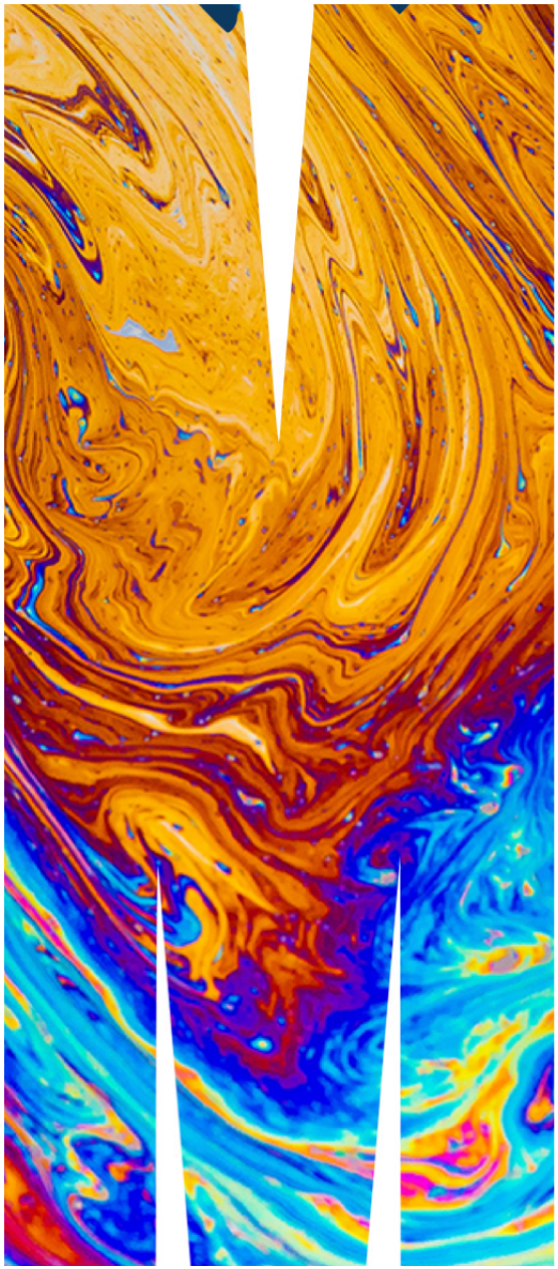
MONASH University

ETC5521: Exploratory Data Analysis

Initial data analysis and model diagnostics



Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 3 - Session 2

❷ Hypothesis Testing and Predictive Modeling Part 3/3

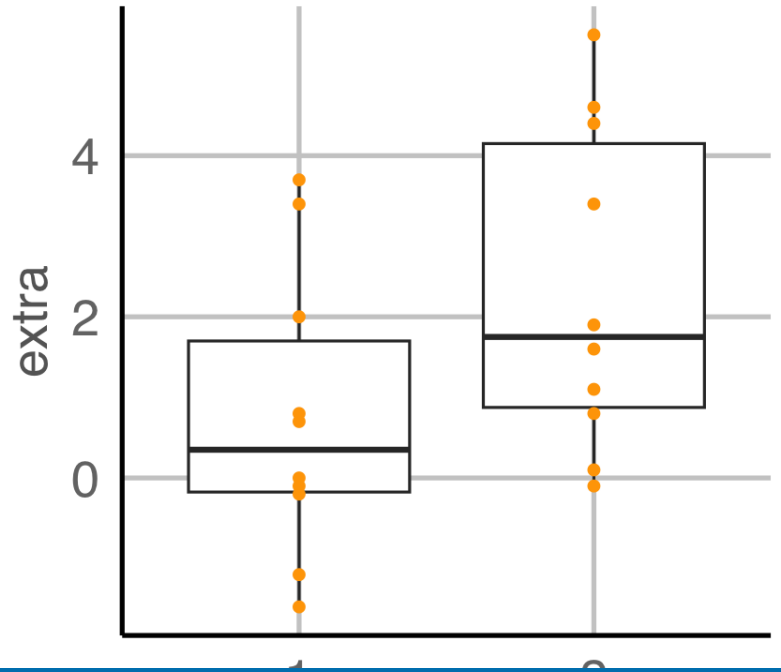
- Hypothesis testing: usually make assumptions about the distribution of the data, and are formed relative to a parameter.
- Predictive modeling: form of the relationship, distribution of the errors.

Hypothesis testing in R REVIEW Part 1/3

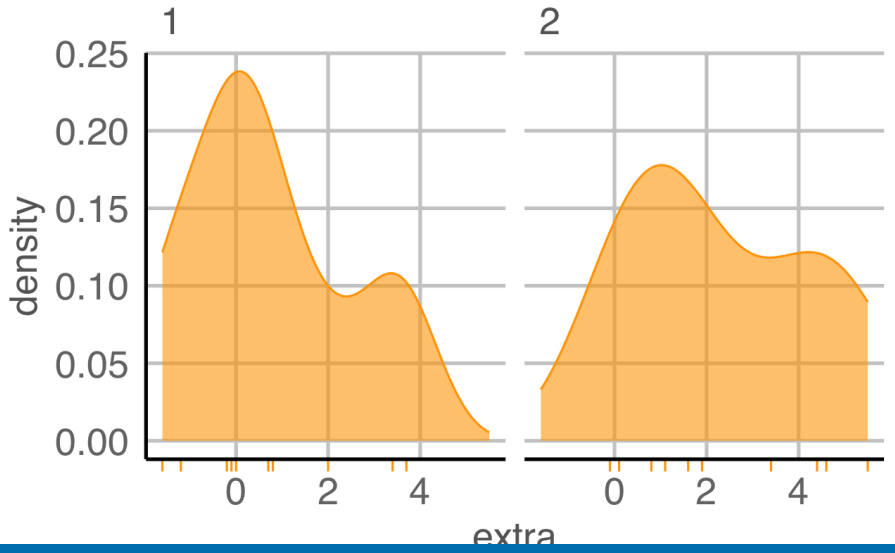
- State the hypothesis (pair), e.g. $H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 < \mu_2$.
- Test statistic depends on *assumption* about the distribution, e.g.
  - t-test will assume that distributions are *normal*, or small departures from if we have a large sample.
  - two-sample might assume both groups have the *same variance*
- Steps to complete:
  - Compute the test statistic
  - Measure it against a standard distribution
  - If it is extreme, p-value is small, decision is to reject H_o
  - p-value is the probability of observing a value as large as this, or large, assuming H_o is true.

Example 1 Checking variance and distribution assumption Part 1/2

```
data(sleep)
ggplot(sleep, aes(x=group, y=extra)) +
  geom_boxplot() +
  geom_point(colour="orange")
```




```
ggplot(sleep, aes(x=extra)) +
  geom_density(fill="orange", colour=
  geom_rug(outside = TRUE, colour="or
  coord_cartesian(clip = "off") +
  facet_wrap(~group)
```



Cushny, A. R. and Peebles, A. R. (1905) The action of optical isomers: II hyoscines. The Journal of Physiology 32, 501–510.

Example 1 Hypothesis test Part 2/2

```
tt <- with(sleep,
     t.test(extra[group == 1],
            extra[group == 2],
            paired = TRUE))

tt$estimate

## mean difference
##           -1.58

tt$null.value

## mean difference
##               0
```

```
tt$statistic

    t
```

-4.062128

```
tt$p.value
```

[1] 0.00283289

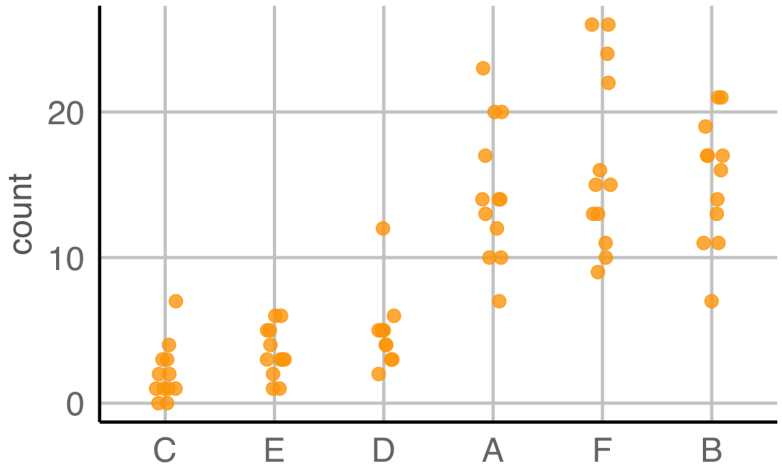
```
tt$conf.int
```

[1] -2.4598858 -0.7001142 attr(,"conf.level") [1] 0.95

Cushny, A. R. and Peebles, A. R. (1905) The action of optical isomers: II hyoscines. The Journal of Physiology 32, 501–510.

Example 2 Checking distribution assumption

```
InsectSprays %>%
  ggplot(aes(x=fct_reorder(spray, cou
             y=count)) +
  geom_jitter(width=0.1, height=0, co
  xlab("")
```



Can you see any violations of normality? Or equal variance?

```
fm1 <- aov(count ~ spray, data = Inse
summary(fm1)

##             Df Sum Sq Mean Sq F va
## spray        5   2669   533.8    3
## Residuals   66   1015    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**'
```

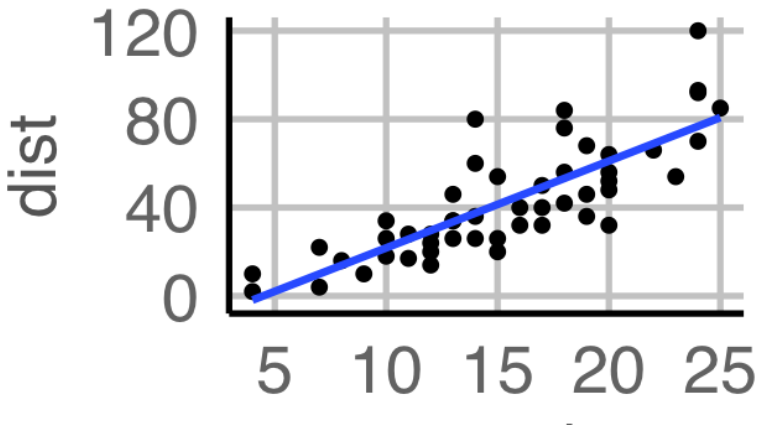
Write down the hypothesis being tested. What would the decision be?

Linear models in R REVIEW Part 1/3

```
library(tidyverse)
library(broom)
glimpse(cars)
```

```
## Rows: 50
## Columns: 2
## $ speed <dbl> 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 13, 13, 14, 14, 1
## $ dist  <dbl> 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28, 26, 34, 34, 46, 26, 3
```

```
ggplot(cars, aes(speed, dist)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Linear models in R REVIEW Part 2/3

- We can fit linear models in R with the `lm` function:

```
lm(dist ~ speed, data = cars)
```

  is the same as

```
lm(dist ~ 1 + speed, data = cars)
```

- The above model is mathematically written as

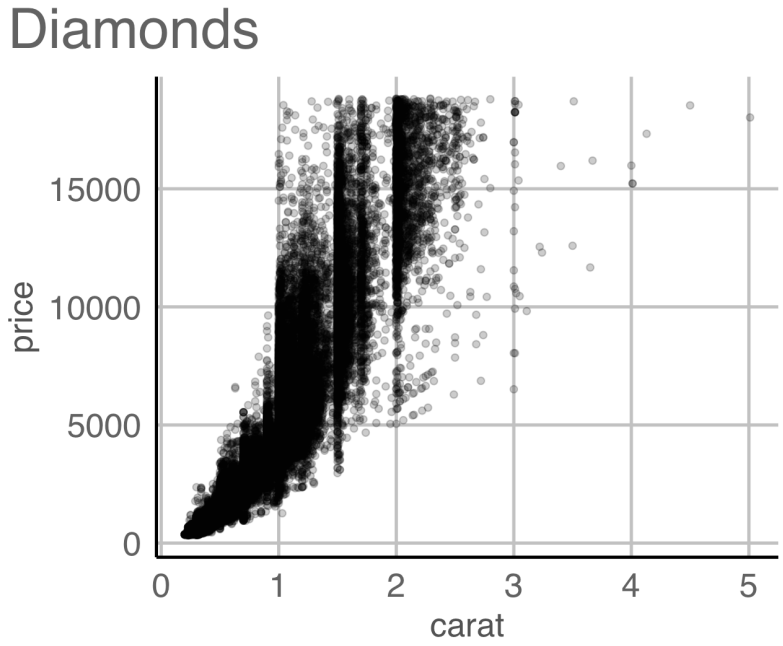
$$y_i = \beta_0 + \beta_1 x_i + e_i$$

  where

  - y_i and x_i are the stopping distance (in ft) and speed (in mph), respectively, of the i-th car;
  - β_0 and β_1 are intercept and slope, respectively; and
  - e_i is the random error; usually assuming $e_i \sim \text{NID}(0, \sigma^2)$.

❷ Model form Part 1/2

- Say, we are interested in characterising the price of the diamond in terms of its carat.

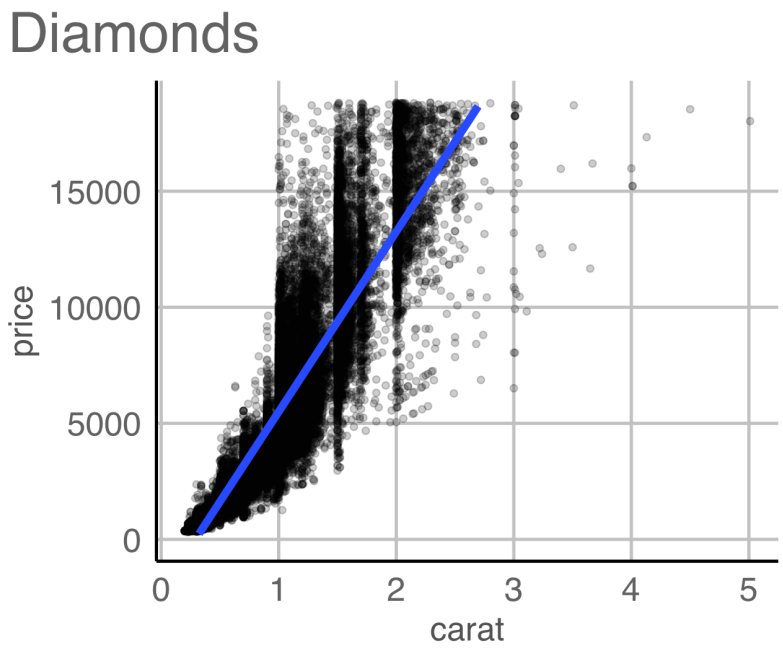


- Looking at this plot, would you fit a linear model with formula

`price ~ 1 + carat`?

❷ Model form Part 1/2

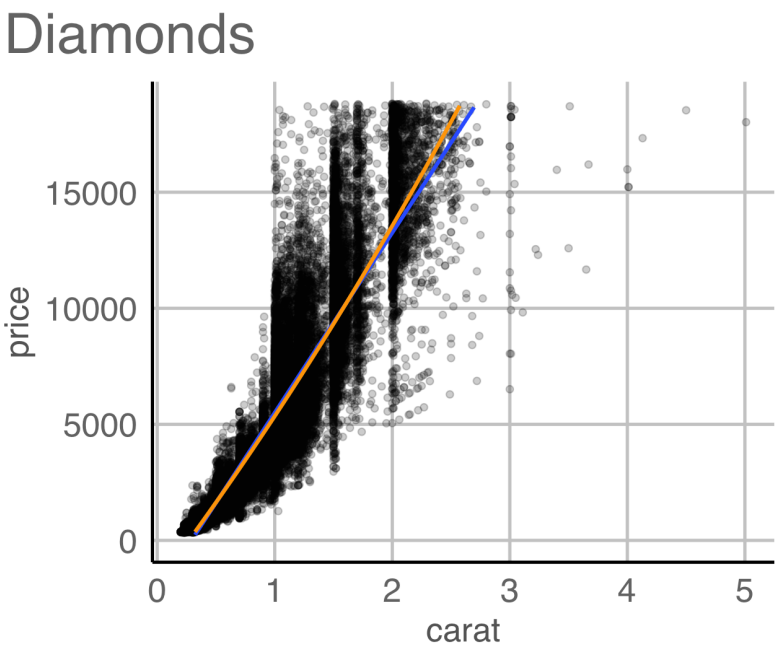
- Say, we are interested in characterising the price of the diamond in terms of its carat.



- Looking at this plot, would you fit a linear model with formula

`price ~ 1 + carat`?

❷ Model form Part 2/2



- What about

  `price ~ poly(carat, 2)`?

  which is the same as fitting:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i .$$

- Should the assumption for error distribution be modified if so?
- Should we make some transformation before modelling?
- Are there other candidate models?

❷ Model form Part 2/2

- Notice that there was ***no formal statistical inference*** when trying to determine an appropriate model form.
- The goal of the main analysis is to characterise the price of a diamond by its carat. This may involve:
  - formal inference for model selection;
  - justification of the selected "final" model; and
  - fitting the final model.
- There may be in fact many, many models considered but discarded at the IDA stage.
- These discarded models are hardly ever reported. Consequently, majority of reported statistics give a distorted view and it's important to remind yourself what might ***not*** be reported.

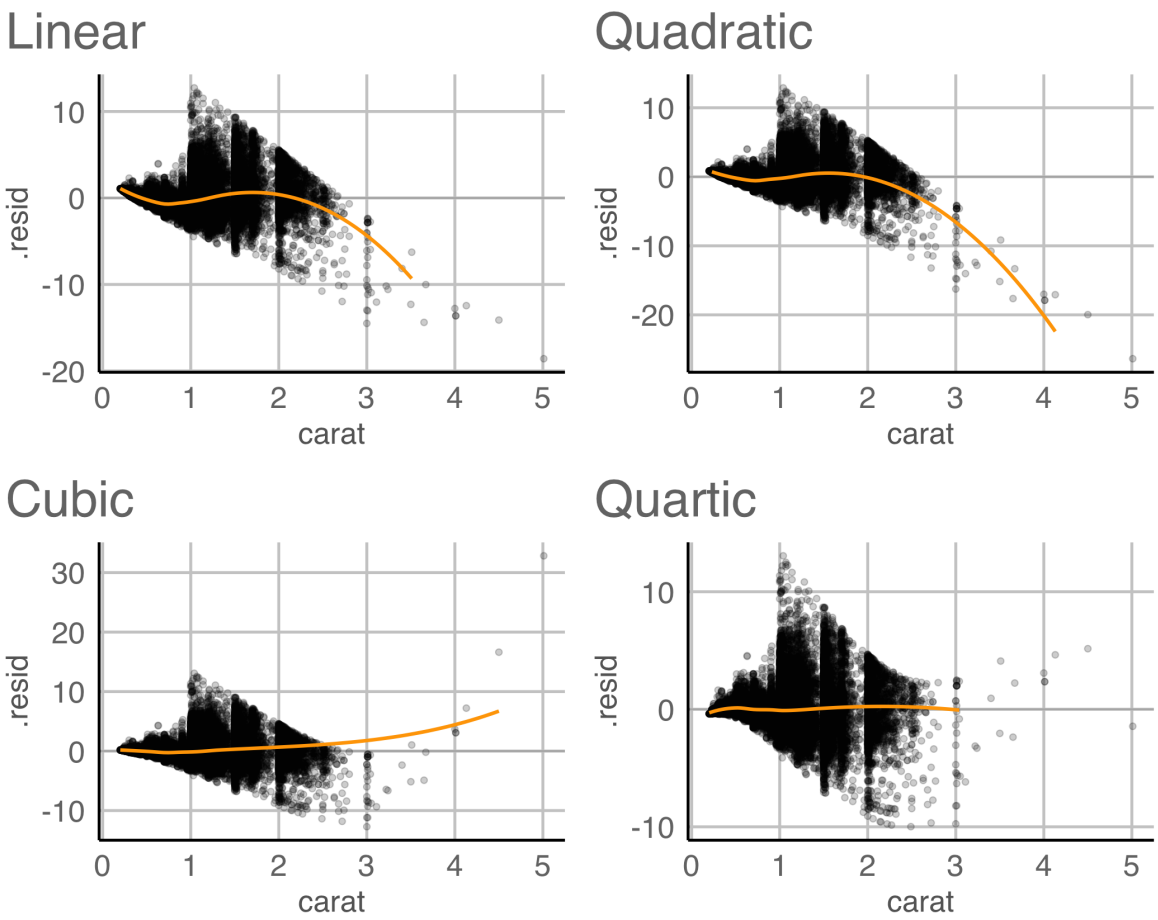
Model selection

> *All models are approximate and tentative; approximate in the sense that no model is exactly true and tentative in that they may be modified in the light of further data*
>
> —*Chatfield (1985)*

> *All models are wrong but some are useful*
>
> —*George Box*

Model diagnostics

Residuals 1/2

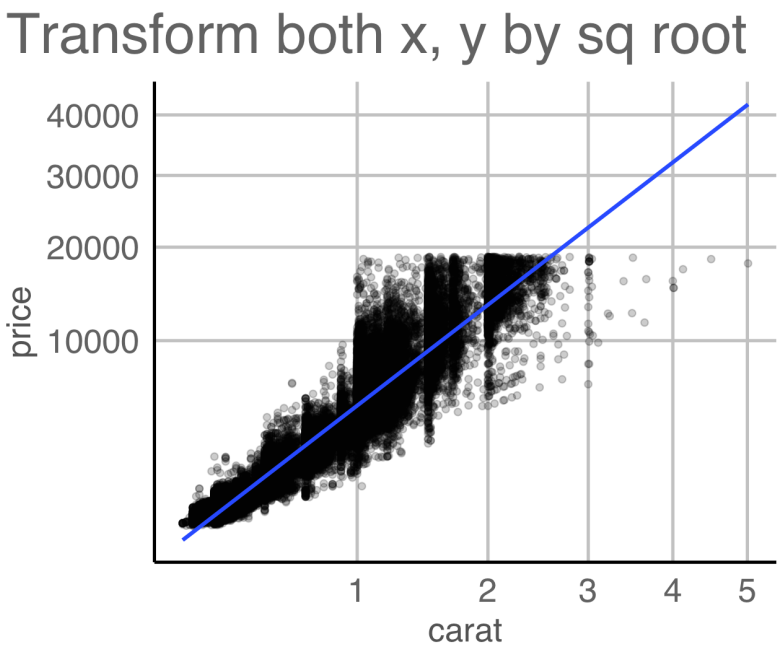


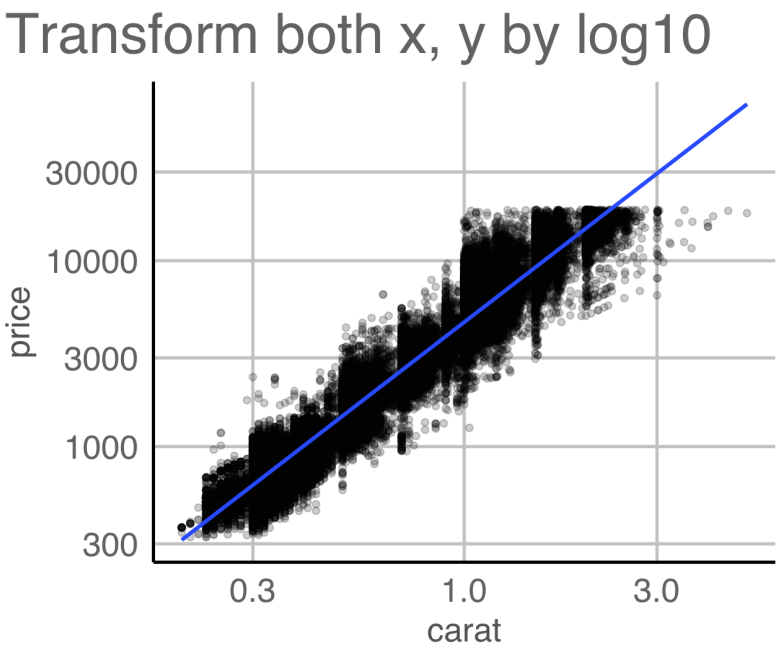
Residual = Observed - Fitted

Residual plot: Plot the residual against explanatory variable (or Fitted value)

Best residual plot has not obvious pattern.

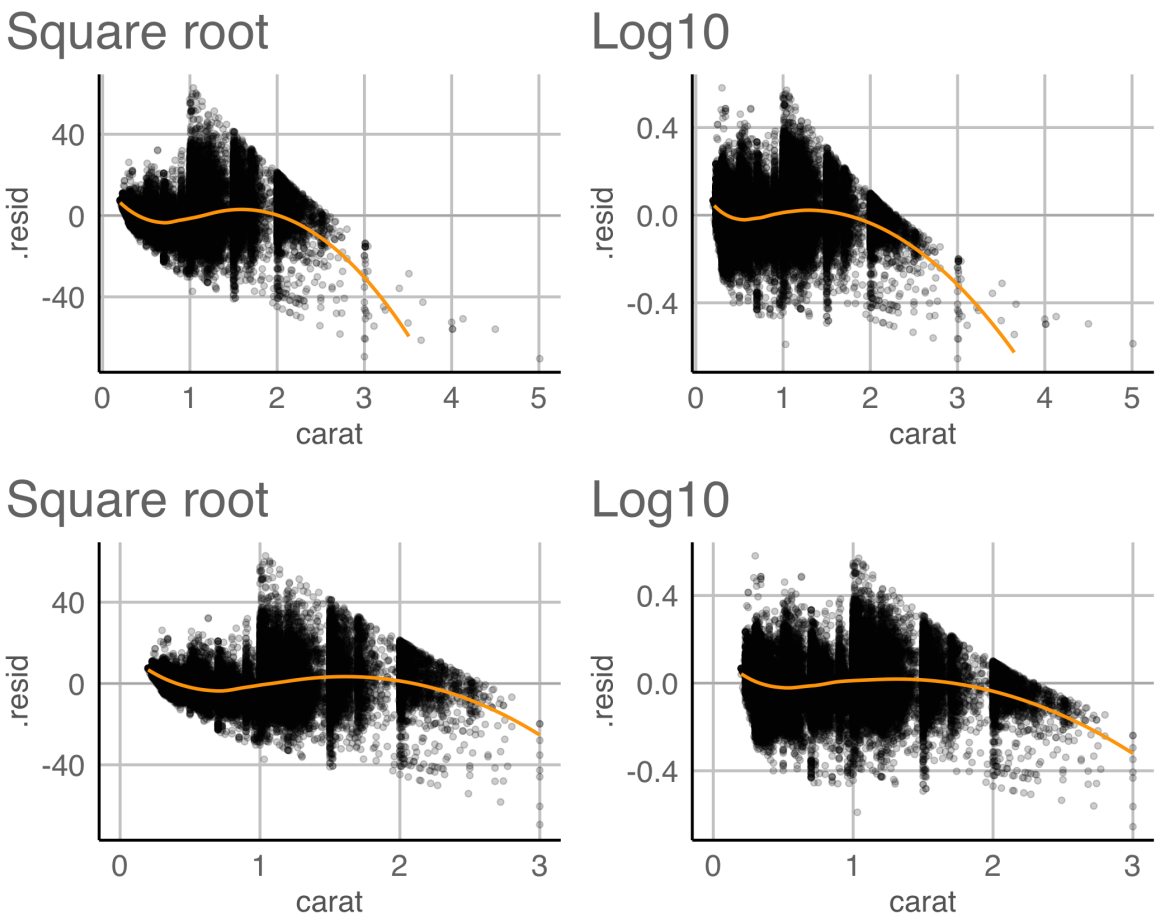
Alternative approach: linearise relationship





The log transformation of both variables linearises the relationship, so that a simple linear model can be used, and also corrects the heteroskedasticity.

Residuals 2/2



Which has the best residual plot?

> *"Teaching of Statistics should provide a more balanced blend of IDA and inference"*
>
> *Chatfield (1985)*

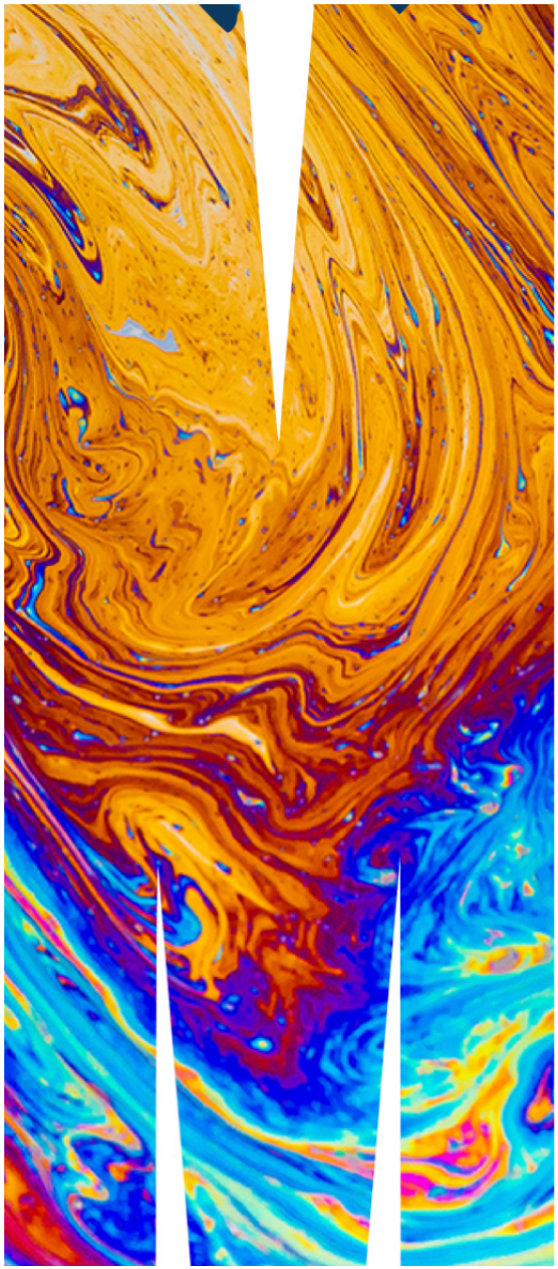
Yet there is still very little emphasis of it in teaching and also at times in practice.

So don't forget to do IDA!

Take away messages

- ***Initial data analysis*** (IDA) is a model-focused exploration to support a confirmatory analysis with:
  - ***data description and collection***
  - ***data quality checking, and***
  - ***checking assumptions***
  - ***model fit*** without any formal statistical inference.
- IDA may never see the limelight BUT it forms the foundation that the main analysis is built upon. Do it well!

Lecture materials originally developed by Dr Emi Tanaka



Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 3 - Session 2