ETC5521: Exploratory Data Analysis

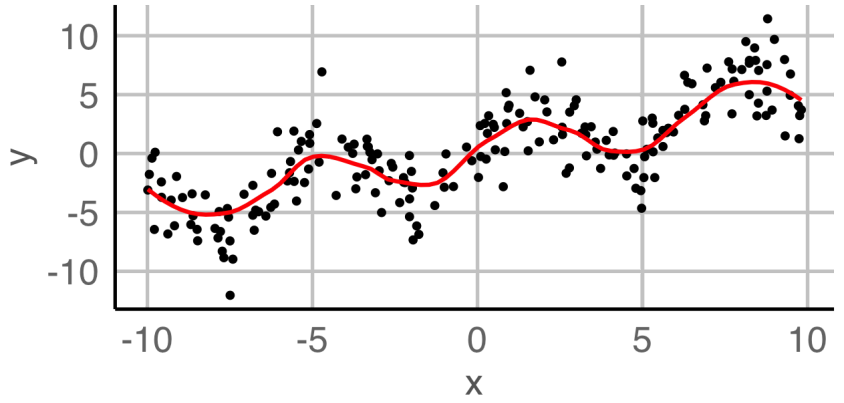
Sculpting data using models, checking assumptions, co-dependency and performing diagnostics

Lecturer: *Emi Tanaka*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 8 - Session 2

Non-parametric regression

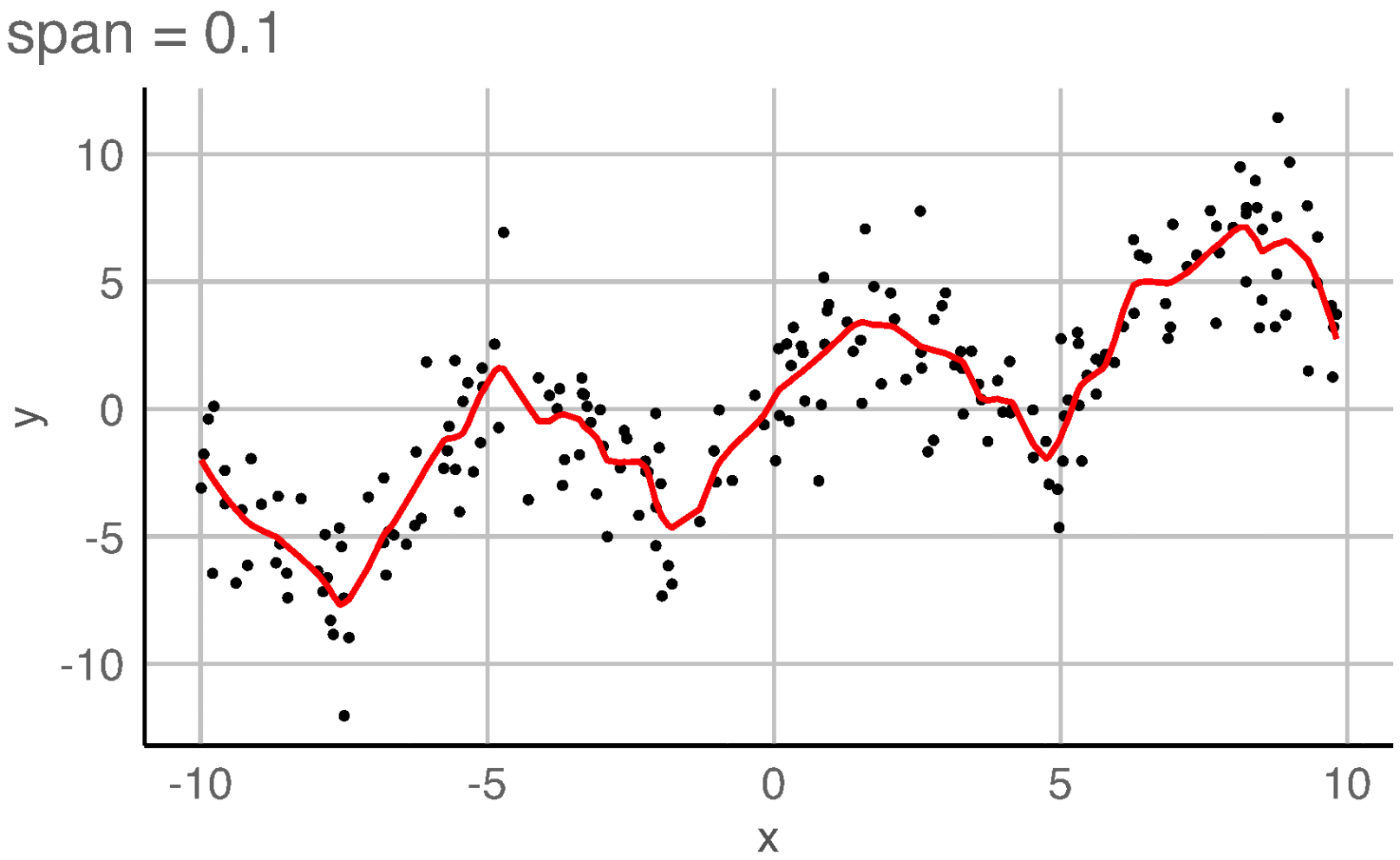
LOESS

- LOESS (LOcal regrESSion) and LOWESS (LOcally WEighted Scatterplot Smoothing) are **non-parametric regression** methods (LOESS is a generalisation of LOWESS)
- **LOESS fits a low order polynomial to a subset of neighbouring data** and can be fitted using `loess` function in R
- a user specified "bandwidth" or "smoothing parameter" α determines how much of the data is used to fit each local polynomial.



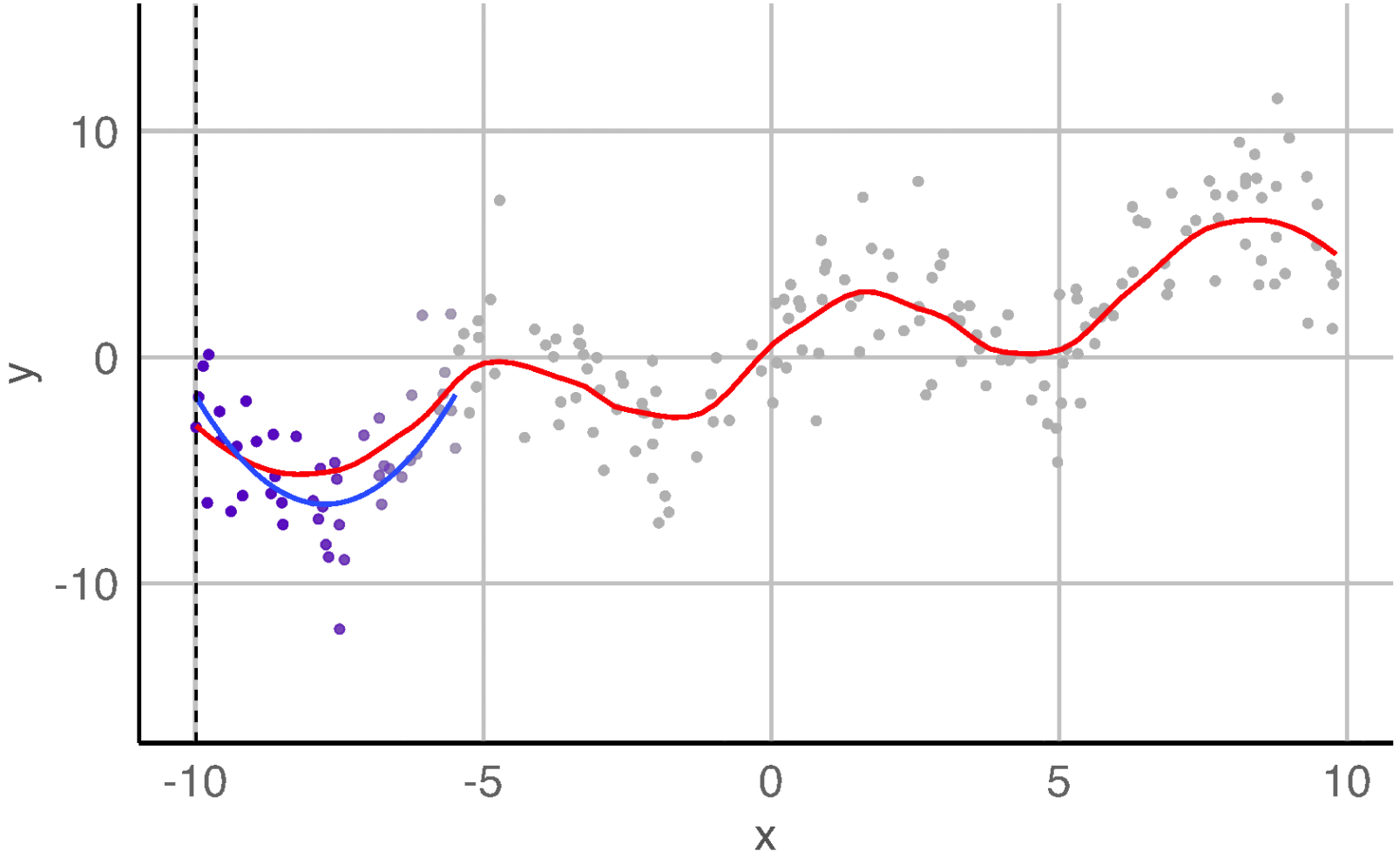
- $\alpha \in \left(\frac{\lambda+1}{n}, 1\right)$ (default `span=0.75`) where λ is the degree of the local polynomial (default `degree=2`) and n is the number of observations.
- Large α produce a smoother fit.
- Small α overfits the data with the fitted regression capturing the random error in the data.

How span changes the loess fit



Code inspired by http://varianceexplained.org/files/loess.html

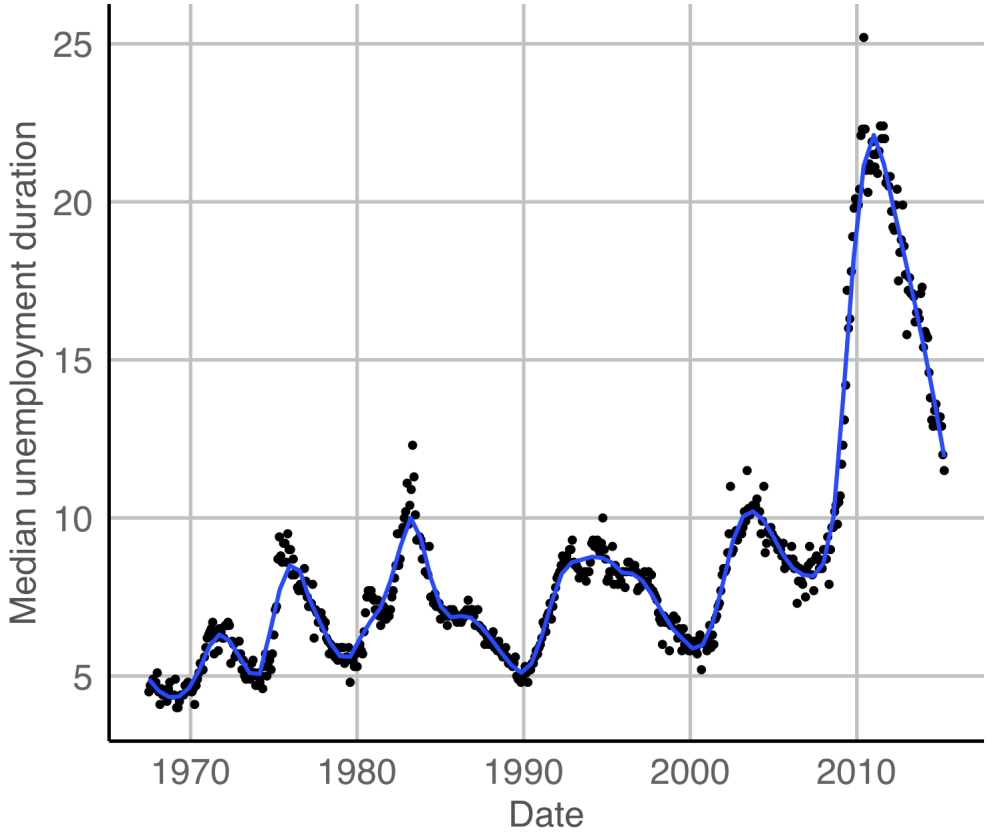
How loess works



Case study 3 US economic time series

This dataset was produced from US economic time series data available from http://research.stlouisfed.org/fred2.

data R



How to fit LOESS curves in R?

Model fitting

The model can be fitted using the `loess` function where

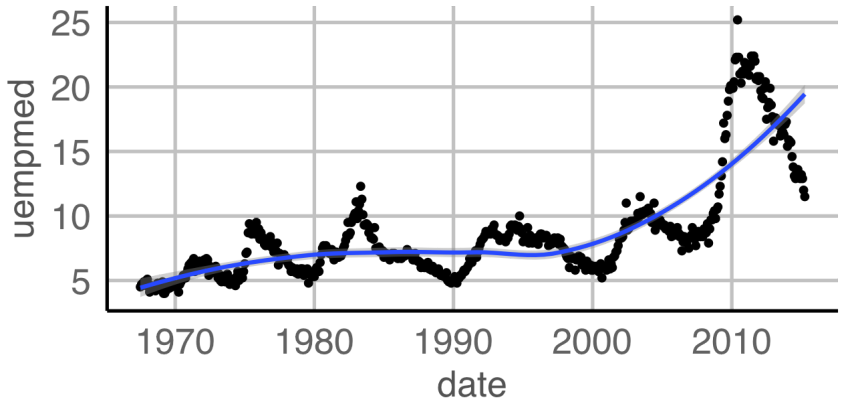
- the default span is 0.75 and
- the default local polynomial degree is 2.

```
fit <- economics %>%
          mutate(index = 1:n()) %>%
          loess(uempmed ~ index,
                data = .,
                span = 0.75,
                degree = 2)
```

Showing it on the plot

In `ggplot`, you can add the loess using `geom_smooth` with `method = loess` and method arguments passed as list:

```
ggplot(economics, aes(date, uempmed)) +
  geom_point() +
  geom_smooth(method = loess,
              method.args = list(span = 0.75,
                                 degree = 2))
```



Why non-parametric regression?

- Fitting a line to a scatter plot where noisy data values, sparse data points or weak inter-relationships interfere with your ability to see a line of best fit.
- Linear regression where least squares fitting doesn't create a line of good fit or is too labour intensive to use.
- Data exploration and analysis.
- Recall: In a parametric regression, some type of distribution is assumed in advance; therefore fitted model can lead to fitting a smooth curve that misrepresents the data.
- In those cases, non-parametric regression may be a better choice.
- *Can you think of where it might be useful?*

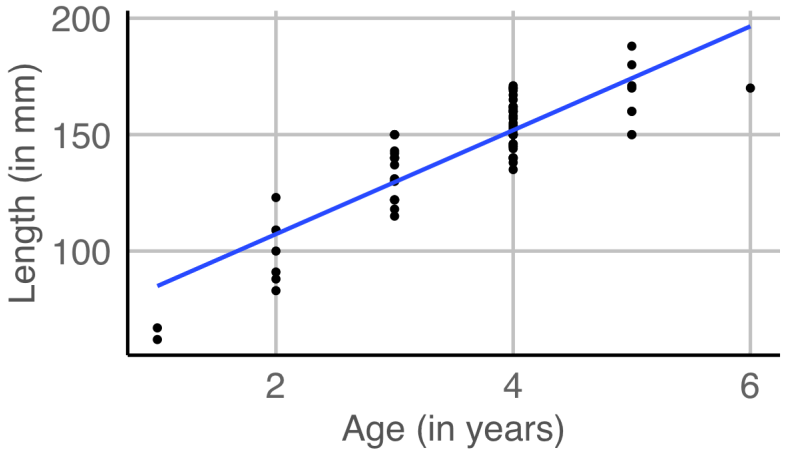
Case study 4 Bluegills Part 1/3

Data were collected on length (in mm) and the age (in years) of 78 bluegills captured from Lake Mary, Minnesota in 1981.

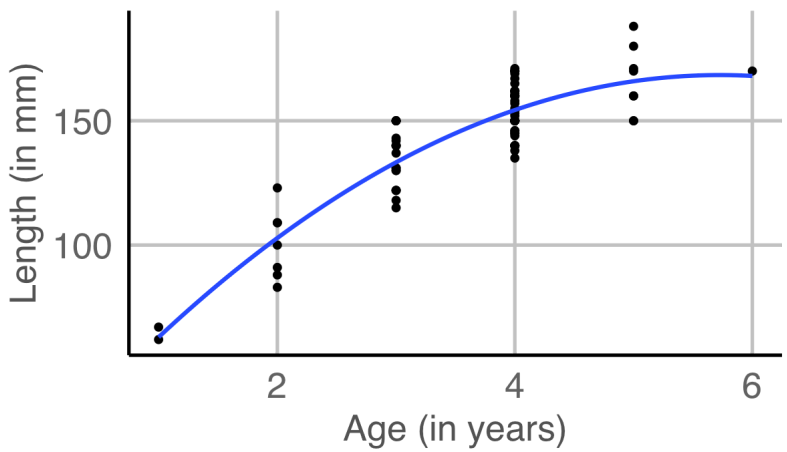
data R

Which fit looks better?

(A)



(B)

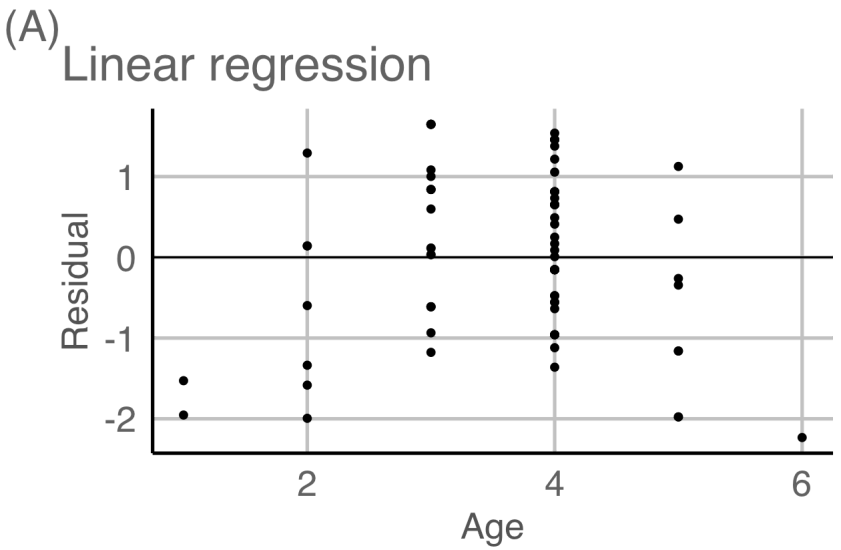


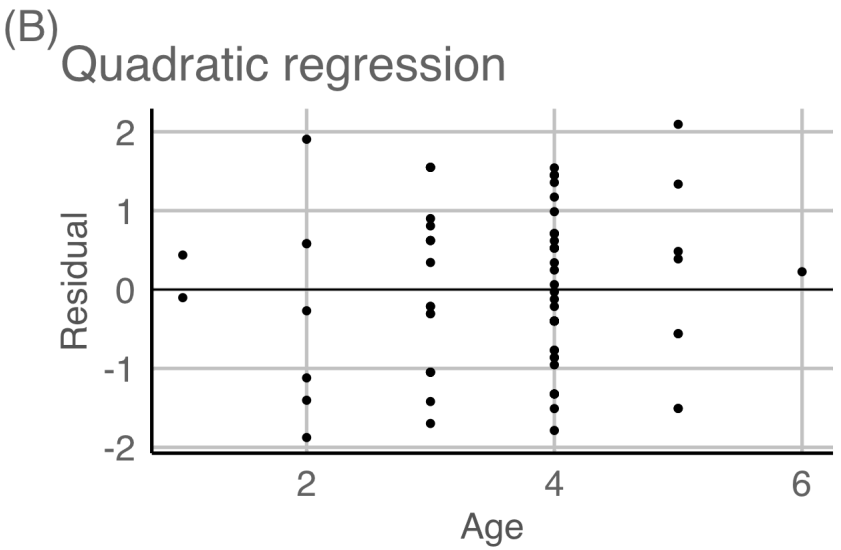
Weisberg (1986) A linear model approach to backcalculation of fish length, *Journal of the American Statistical Association* **81** (196) 922-929

Case study 4 Bluegills Part 2/3

- Let's have a look at the residual plots.
- Do you see any patterns on either residual plot?

data R



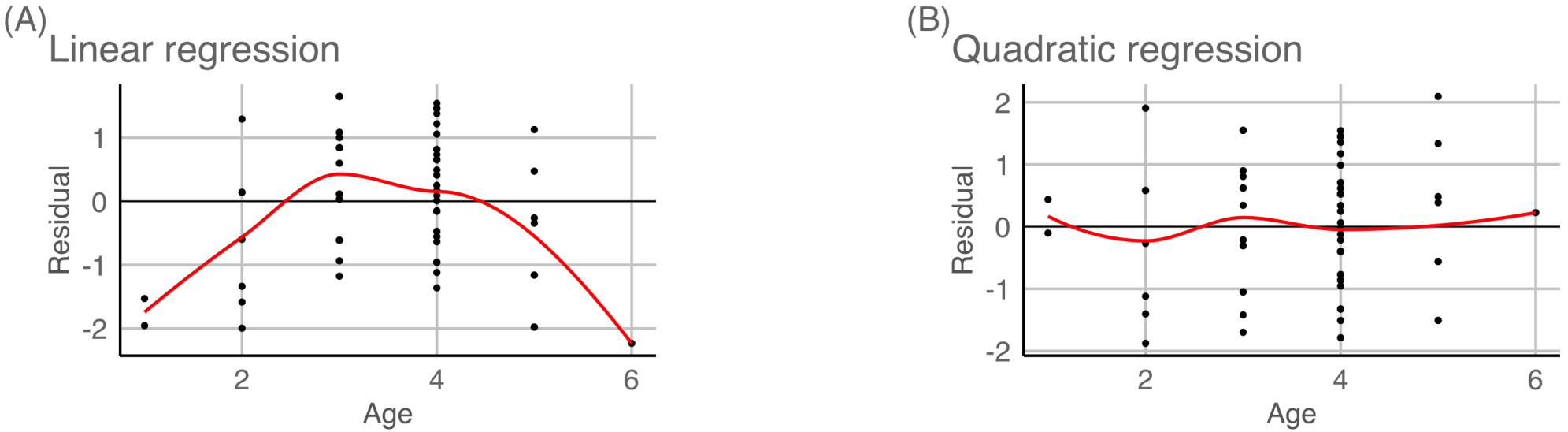


Weisberg (1986) A linear model approach to backcalculation of fish length, *Journal of the American Statistical Association* **81** (196) 922-929

Case study 4 Bluegills Part 3/3

The structure is easily visible with the LOESS curve:

data R

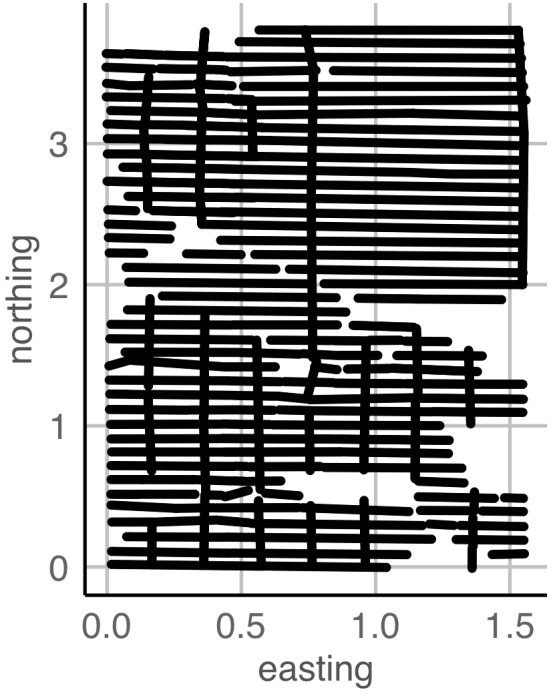


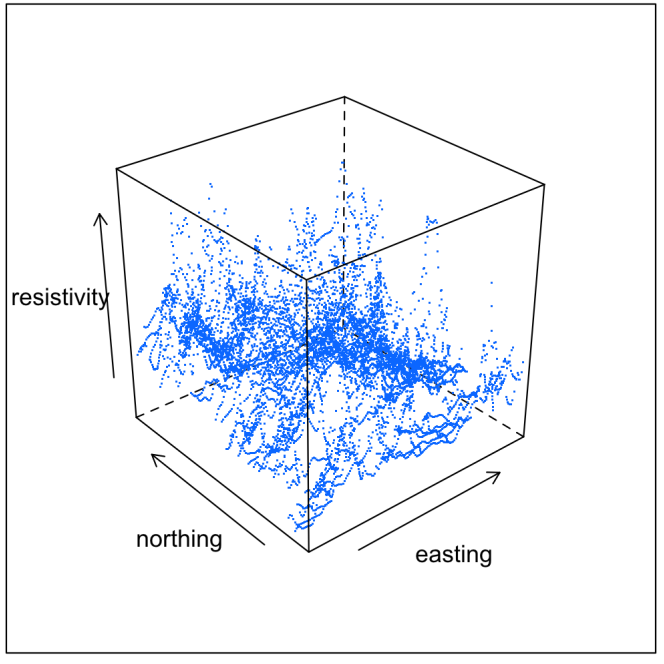
Weisberg (1986) A linear model approach to backcalculation of fish length, *Journal of the American Statistical Association* **81** (196) 922-929

Case study 5 Soil resistivity in a field

This data contains measurement of soil resistivity of an agricultural field.

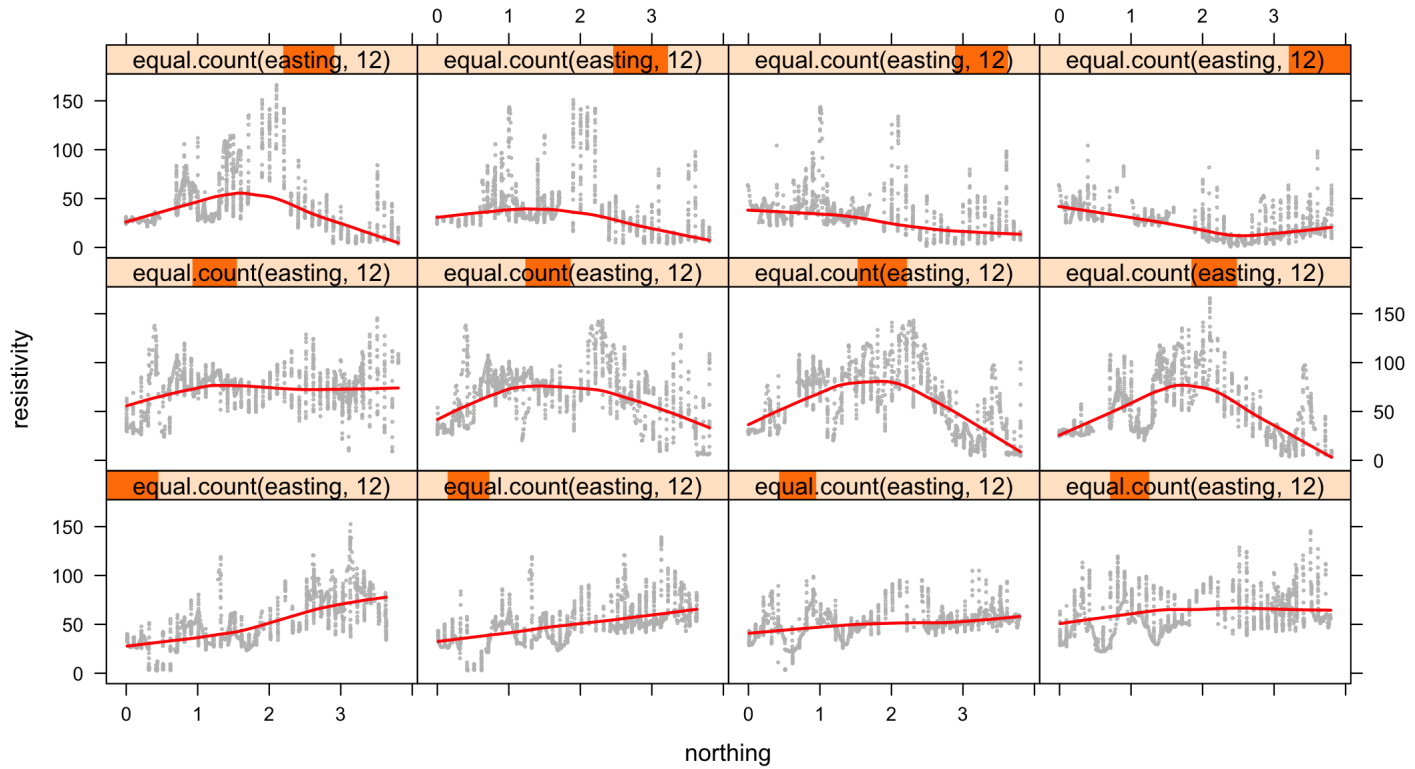
data R





Conditioning plots (Coplots)

```
library(lattice)
xyplot(resistivity ~ northing | equal.count(easting, 12),
       data = cleveland.soil, cex = 0.2,
       type = c("p", "smooth"), col.line = "red",
       col = "gray", lwd = 2)
```



See also: https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/threenum.html

Coplots via ggplot2

- Coplots with ggplot2 where the panels have overlapping observations is tricky.
- Below creates a plot for non-overlapping intervals of easting:

```
ggplot(cleveland.soil, aes(northing, resistivity)) +
  geom_point(color = "gray") +
  geom_smooth(method = "loess", color = "red", se = FALSE) +
  facet_wrap(~ cut_number(easting, 12))
```

That's it, for this lecture!

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu