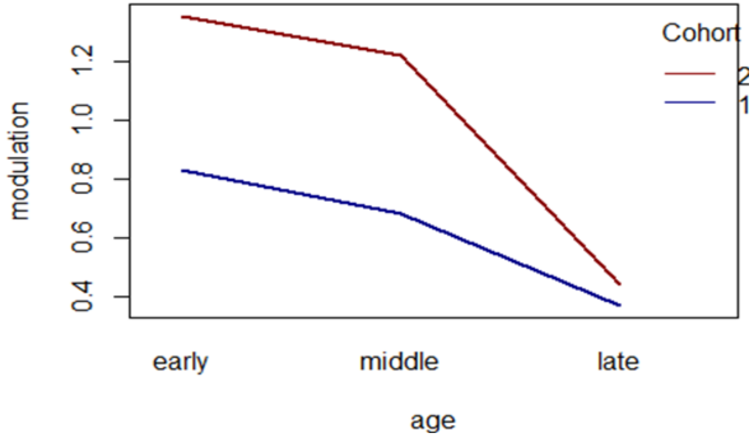
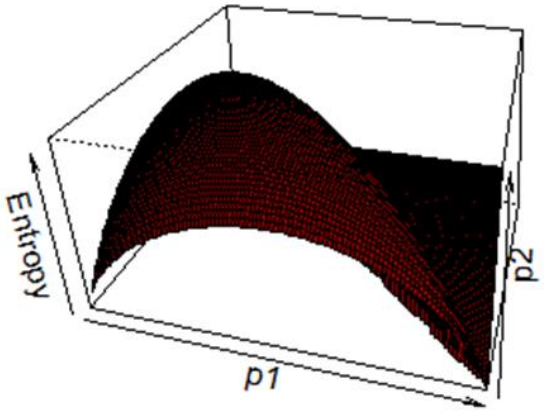


Statistics for linguists

2023-12-13

Confounds, model fitting, interactions, multivariate regression



Packages for today

```r
library(dagitty)
library(ggdag)
library(V8)
library(Rling)
library(AICcmodavg)
library(tidyverse)
library(gridExtra)
```

Multivariate regression

- We have thus far considered models with a single predictor

- A multivariate model has the following structure

$$y = a + b_1 \times\ x_1 + b_2 \times\ x_2\ +\ \ldots b_n \times\ x_n\ + e$$

$$e \sim N(0, s)$$

Multivariate regression

- We have thus far considered models with a single predictor
- A multivariate model has the following structure

$$y = a + b_1 \times x_1 + b_2 \times x_2 + \ldots b_n \times x_n + e$$

Dependent variable

$$e \sim N(0, s)$$

Independent variables

Coefficients

Normally distributed error term

Multivariate regression

- We have thus far considered models with a single predictor
- A multivariate model has the following structure

$$y = a + b_1 \times\ x_1 + b_2 \times\ x_2 +\ \dots b_n \times\ x_n\ + e \qquad e \sim N(0, s)$$

Dependent variable

Independent variables

Coefficients

Normally distributed error term

Correlations and confounds

- Multivariate regression is a powerful research because it can allow us to better distinguish between cause and effect.
- Why is this the case?
- What is the reason that causation cannot always be inferred from correlation?

Confounds or confounding

- 1. to perplex or amaze , esp. by a sudden disturbance or surprise; bewilder; confuse
  The complicated directions confounded him
- 2. to throw into confusion or disorder
  The revolution confounded the people
- 3. to throw into increased confusion or disorder
- 4. to treat or regard erroneously as identical; mix or associate by mistake
  truth confounded with error
- 5. to mingle so that the elements cannot be distinguished or separated
- 6. to damn (used in mild imprecations)
  Confound it!
- 7. to contradict or refute
  to confound their arguments

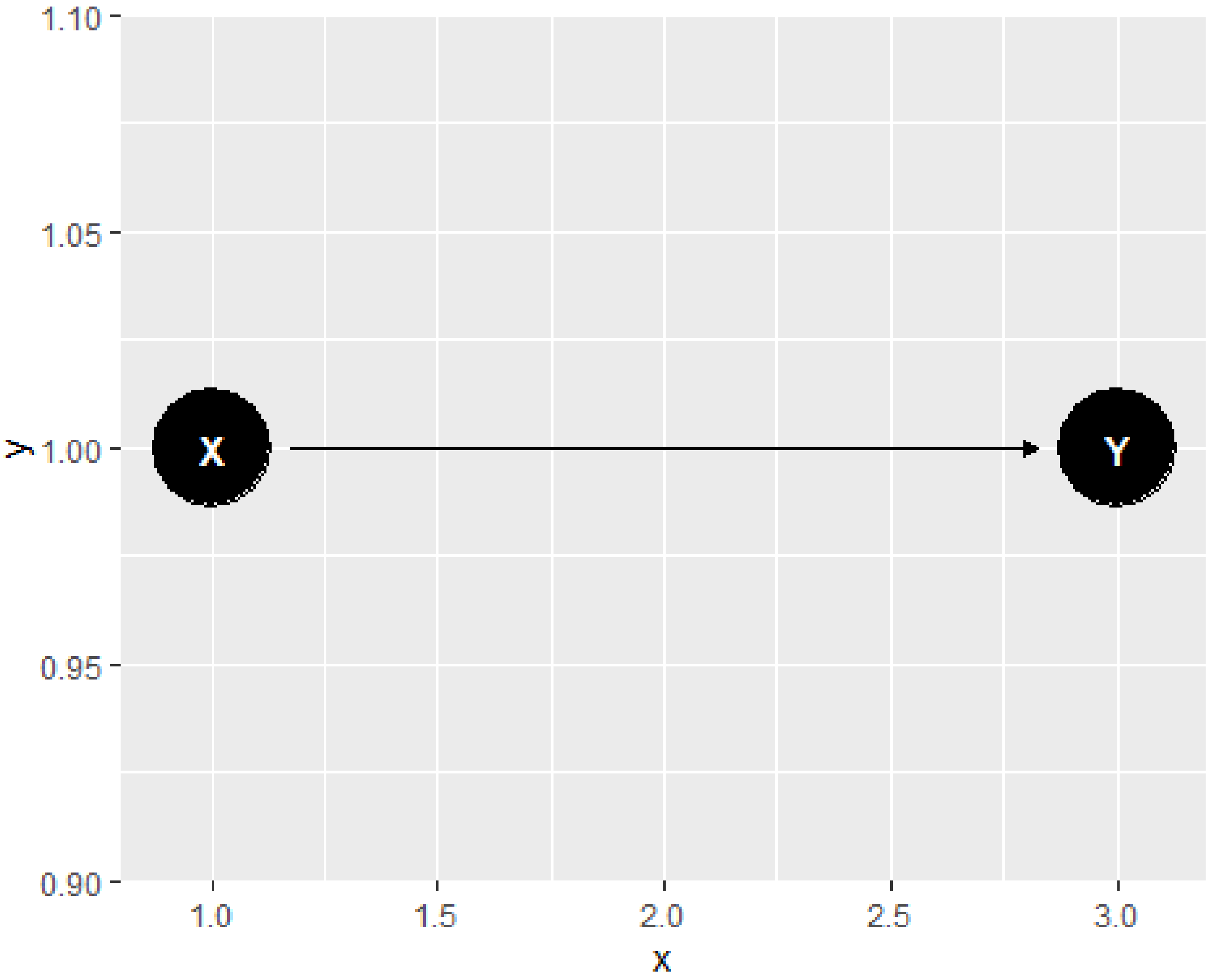
https://www.collinsdictionary.com/de/worterbuch/englisch/confound

Confounds or confounding

"If we undertake to estimate the effect of one variable (X) on another (Y) by examining the statistical association between the two, we ought to ensure that the association is not produced by factors other than the effect under study. The presence of spurious association – due, for example, to the influence of extraneous variables – is called *confounding* because it tends to confound our reading and to bias our estimate of the effect studied. Conceptually, therefore, we can say that *X* and *Y* are confounded when there is a third variable *Z* that influences both *X* and *Y;* such a variable is then called a *confounder* of *X* and *Y.*"
Pearl 2009: 183

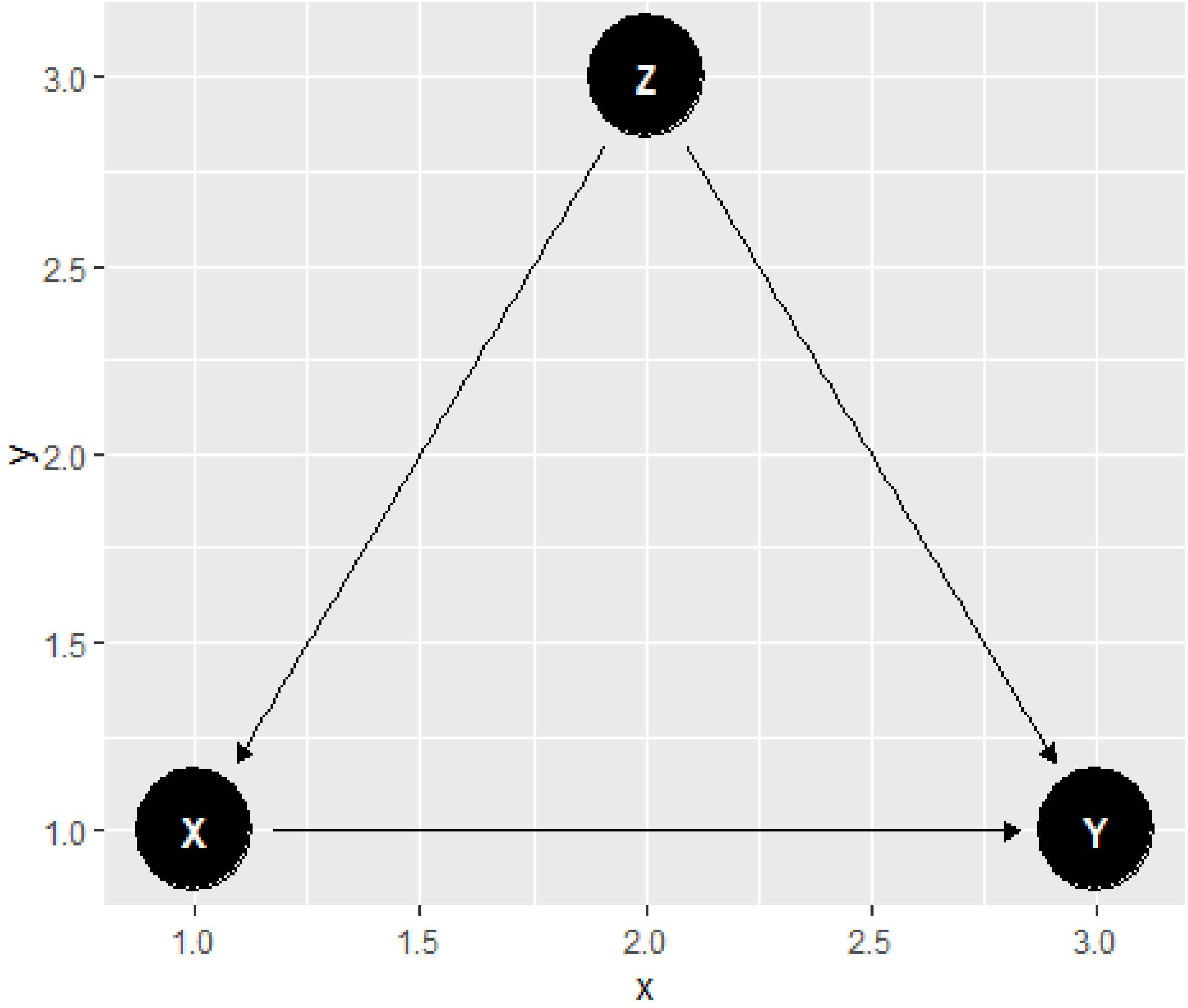
Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference (Second Edition).* Cambridge.

Fork




```r
coord_dag <- list(
  x = c(X=1, Y=3),
  y = c(X=1, Y=1)
)

our_dag <- ggdag::dagify(Y~X,
                         coords = coord_dag)
ggdag::ggdag(our_dag)
```

Fork




```
coord_dag <- list(
  x = c(X=1, Y=3, Z=2),
  y = c(X=1, Y=1, Z=3)
)

our_dag <- ggdag::dagify(Y~X,
                         X~Z,
                         Y~Z,
                         coords = coord_dag)
ggdag::ggdag(our_dag)
```

Fork

- To illustrate we can simulate the **Fork confounder**

```
set.seed(1234)
z <- rnorm(100, 10, 10)
b1 <- 2
b2 <- 3
a1 = 3
a2 = 4
y <- a1 + b1*z + rnorm(100, 0, 3)
x <- a2 + b2*z + rnorm(100, 0, 3)
d <- list(y,
          z,
          x)

summary(lm(y~x, data=d))
```

```r
summary(lm(y~x, data=d))

##
## Call:
## lm(formula = y ~ x, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5904 -2.1783  0.0315  2.5965  7.1595
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52548    0.48810   1.077    0.284
## x            0.65398    0.01148  56.953   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.485 on 98 degrees of freedom
## Multiple R-squared:  0.9707, Adjusted R-squared:  0.9704
## F-statistic:  3244 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
summary(lm(y~x, data=d))

##
## Call:
## lm(formula = y ~ x, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5904 -2.1783  0.0315  2.5965  7.1595
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52548    0.48810   1.077    0.284
## x            0.65398    0.01148  56.953   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.485 on 98 degrees of freedom
## Multiple R-squared:  0.9707, Adjusted R-squared:  0.9704
## F-statistic:  3244 on 1 and 98 DF,  p-value: < 2.2e-16
```

There appears to be a significant relationship even though we know there is none.

```r
summary(lm(y~x+z, data=d))

##
## Call:
## lm(formula = y ~ x + z, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6577 -1.8151  0.0409  1.8996  8.6932
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7516     0.6181   4.452 2.28e-05 ***
## x             0.1027     0.1090   0.942    0.348
## z             1.6817     0.3310   5.081 1.82e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.113 on 97 degrees of freedom
## Multiple R-squared:  0.9768, Adjusted R-squared:  0.9764
## F-statistic:  2045 on 2 and 97 DF,  p-value: < 2.2e-16
```

The relationship disappears when put in the actual causal factor

Colliders and causal salad

- It is not so simple as adding as many factors as possible to a model – we can create a bias by *adding* variables – something that is rarely considered when basic model fitting is taught.

- A **collider confound** is one which creates a spurious correlation between variables by being added to a model.
  - A **pipe** does something similar but removes a correlation where there would be one.

- The practice of unthinkingly adding all the variables one can think of to a model to find correlations, without thinking about causal relationships is referred to as **causal salad** – it is extremely wrong and widely practiced

"... there is **Causal Salad**: You put everything into a **regression equation**, toss with some creative story-telling, and hope the reviewers eat it. In general, this is not a valid approach, for well-known reasons. But it can get you published. Causal salad can discover causes too. But you have to get lucky. The Salad isn't only regression. Really any procedure that hopes to take a list of variables (features) and return causal inference is Causal Salad. No amount of data reliably turns salad into sense."
- McElreath

https://elevanth.org/blog/2021/06/15/regression-fire-and-dangerous-things-1-3/

https://bigthink.com/surprising-science/judea-pearls-the-book-of-why-brings-news-of-a-new-science-of-causes/

Pearl, Judea (& Dana Mackenzie). 2018. *The Book of Why: The New Science of Cause and Effect.* Penguin

Confounding types

- Fork
- Pipe
- Collider
- Descendant

Interactions

- Sometimes the strength, significance or even direction of an effect is related to

- An interaction is a term which allows the direction and structure (slope) of a term to vary with another term.

- It’s most easily illustrated and conceptualized with an ANOVA model

Interactions

- Download the data modref from the Rlang package

```
data("sharedref")
head(sharedref)

##     mod   age cohort
## 1 0.75 early      1
## 2 0.85 early      1
## 3 0.93 early      1
## 4 0.80 early      1
## 5 1.24 early      2
## 6 1.38 early      2
```

Interactions

- The data show how many modulations of a given sign occur according to age.

- These are data from Nicaraguan sign language – modulation should decrease by age and cohort (the age group that started learning)

- But cohort and age are not independent from one another.

Senghas, Richard J. et al. 2005. The Emergence of Nicaraguan Sign Language: Questions of Development, Acquisition, and Evolution. *Biology and Knowledge revisited: From neurogenesis to psychogenesis.* Lawrence Erlbaum Associates. 287-306.

Interactions

- **mod** = continuous: number of spatial modulations per verb
- **age** = categorical: early, middle, late
- **cohort** = categorical: 1, 2, 3

Interactions

```
model1 <- lm(mod~age, data=sharedref)
anova(model1)

## Analysis of Variance Table
##
## Response: mod
##            Df Sum Sq Mean Sq F value    Pr(>F)
## age         2 4.2243 2.11217  38.663 1.691e-10 ***
## Residuals 45 2.4584 0.05463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

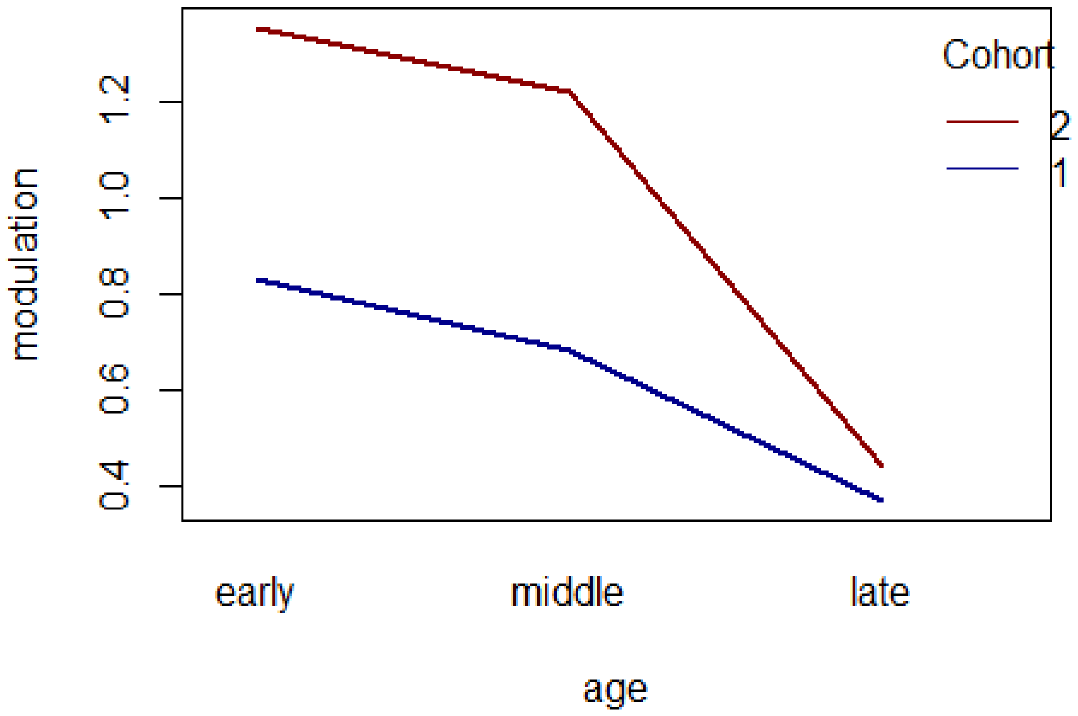
Interaction plot

```
ref <- aggregate(mod~age+cohort, data=sharedref, FUN = mean)
interaction.plot(ref$age,
                 ref$cohort,
                 ref$mod,
                 xlab="age",
                 ylab ="modulation",
                 lty = 1,
                 lwd = 2,
                 col = c("blue4", "red4"),
                 trace.label = "Cohort")
```

Interaction plot

- In both cohorts the amount of modulation goes down by age.

- But the slope is different depending on the age.

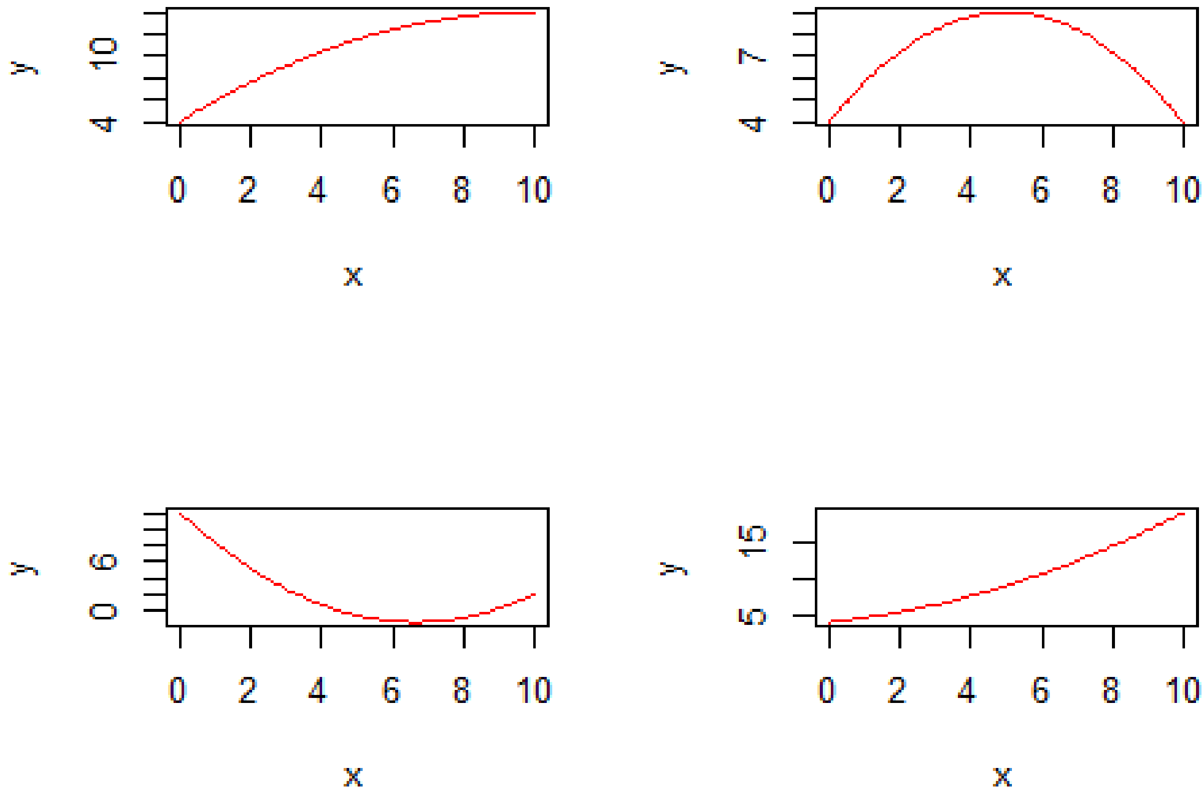


```
model2 <- lm(mod~age*cohort, data=sharedref)
anova(model2)

## Analysis of Variance Table
##
## Response: mod
##             Df Sum Sq Mean Sq F value    Pr(>F)
## age          2 4.2243 2.11217 491.884 < 2.2e-16 ***
## cohort       1 1.7101 1.71008 398.243 < 2.2e-16 ***
## age:cohort   2 0.5679 0.28397  66.132 1.054e-13 ***
## Residuals   42 0.1804 0.00429
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model2)

##
## Call:
## lm(formula = mod ~ age * cohort, data = sharedref)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16000 -0.03438  0.00000  0.04000  0.11750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.816250   0.009458   86.30  < 2e-16 ***
## age1        -0.138750   0.006688  -20.75  < 2e-16 ***
## age2        -0.272500   0.011584  -23.52  < 2e-16 ***
## cohort1      0.188750   0.009458   19.96  < 2e-16 ***
## age1:cohort1 -0.036250   0.006688   -5.42 2.70e-06 ***
## age2:cohort1 -0.117500   0.011584  -10.14 7.32e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06553 on 42 degrees of freedom
## Multiple R-squared:  0.973,  Adjusted R-squared:  0.9698
## F-statistic: 302.9 on 5 and 42 DF,  p-value: < 2.2e-16
```

Model fitting and overfitting

- There are different types of relationships we can construct by adding variables to our regression equation in different ways.



Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Code

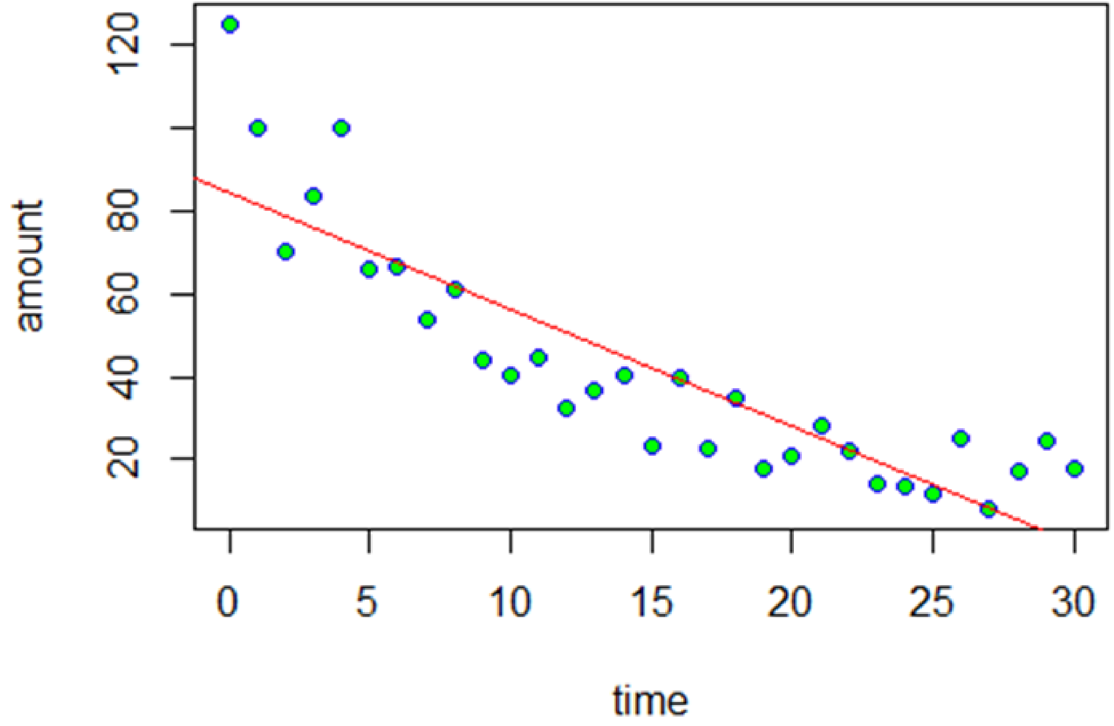
```
par(mfrow=c(2,2))
curve(4+2*x-0.1*x^2,0,10,col="red",ylab="y")
curve(4+2*x-0.2*x^2,0,10,col="red",ylab="y")
curve(12-4*x+0.3*x^2,0,10,col="red",ylab="y")
curve(4+0.5*x+0.1*x^2,0,10,col="red",ylab="y")
```

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Model fitting and overfitting

- Download the decay.csv data which shows nuclear waste decay over time.

```
par(mfrow=c(1,1))
data <- read.csv("YourPathway/decay.csv",
header=TRUE)
attach(data)
plot(time,amount,pch=21,col="blue",bg="gre
en")
abline(lm(amount~time),col="red")
```



Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Model fitting and overfitting

- We could develop a more complex model

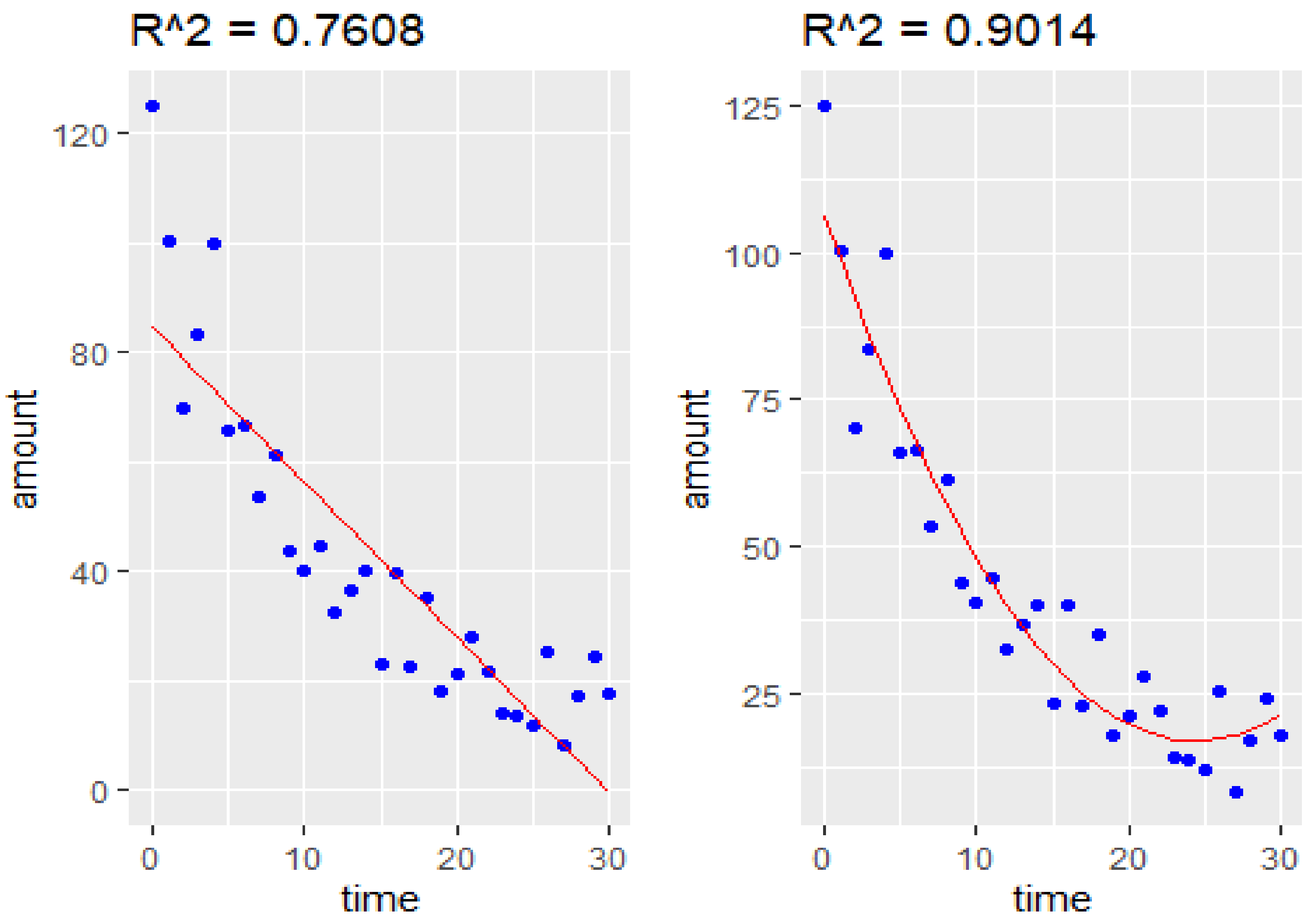
$$amount = a + b_1 * time + b_2 \; * time^2 + e$$

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

```r
model2 <- lm(amount~time)
model3 <- lm(amount~time+I(time^2))
```

```r
p1 <- ggplot(data=data, aes(x=time, y =amount))+
  geom_point(color='blue')+
  geom_line(color='red', data=predict_model3, aes(x=time, y=amount_pred))+
  ggtitle("R^2 = 0.9014")
predict_model2 <- data.frame(amount_pred = predict(model2, data), time= data$time)

p2 <- ggplot(data=data, aes(x=time, y =amount))+
  geom_point(color='blue')+
  geom_line(color='red', data=predict_model2, aes(x=time, y=amount_pred))+
  ggtitle("R^2 = 0.7608")

grid.arrange(p2, p1, nrow = 1, ncol =2)
```

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.



Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

```r
summary(model2)

##
## Call:
## lm(formula = amount ~ time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.065 -10.029  -2.058   5.107  40.447
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.5534     5.0277   16.82  < 2e-16 ***
## time         -2.8272     0.2879   -9.82 9.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.34 on 29 degrees of freedom
## Multiple R-squared:  0.7688, Adjusted R-squared:  0.7608
## F-statistic: 96.44 on 1 and 29 DF,  p-value: 9.939e-11
```

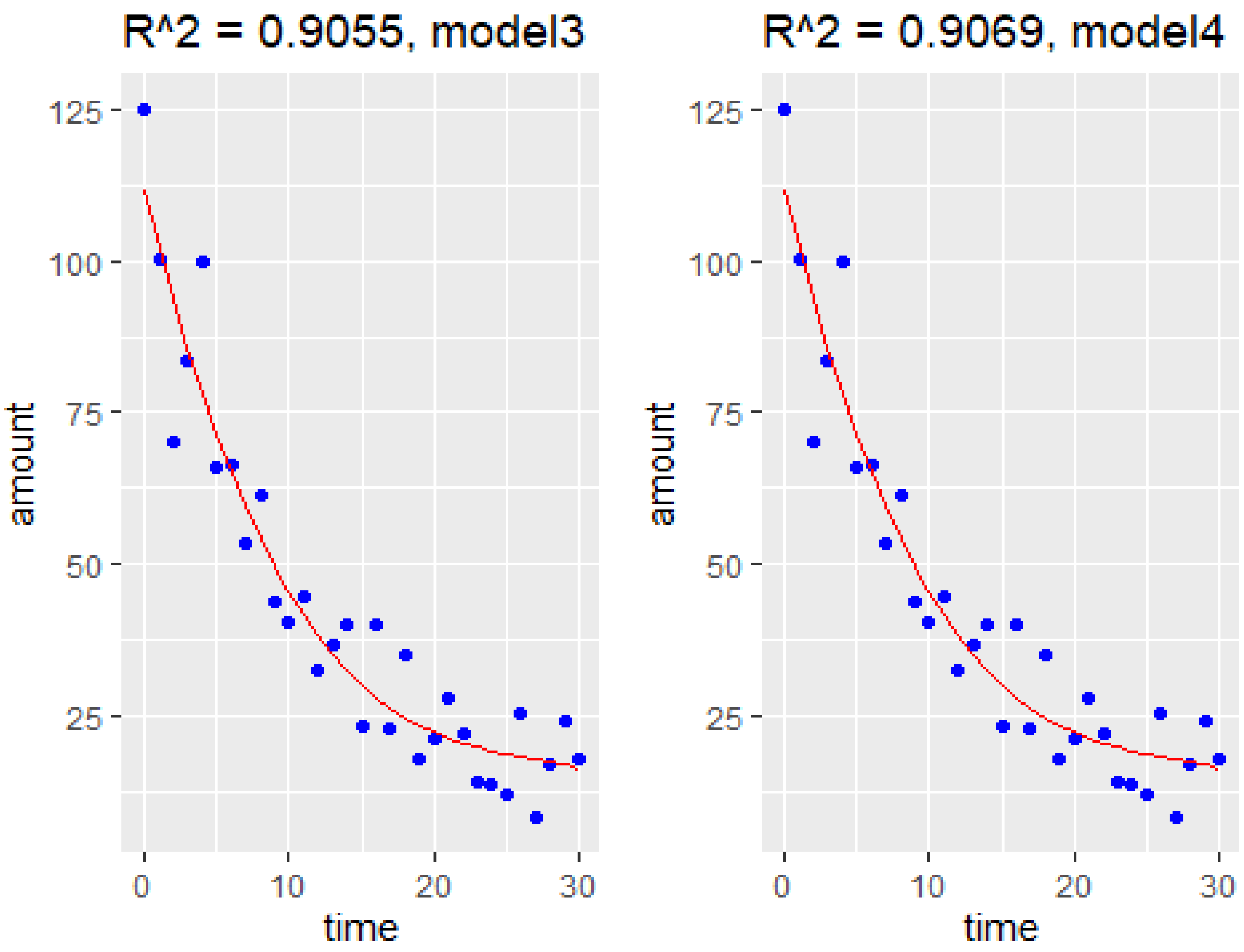
Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

```
summary(model3)

##
## Call:
## lm(formula = amount ~ time + I(time^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.302  -6.044  -1.603   4.224  20.581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 106.38880    4.65627  22.849  < 2e-16 ***
## time         -7.34485    0.71844 -10.223 5.90e-11 ***
## I(time^2)     0.15059    0.02314   6.507 4.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.205 on 28 degrees of freedom
## Multiple R-squared:  0.908,  Adjusted R-squared:  0.9014
## F-statistic: 138.1 on 2 and 28 DF,  p-value: 3.122e-15
```

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

```r
model4 <- lm(amount~time+I(time^2)+I(time^3))
predict_model4 <- data.frame(amount_pred = predict(model4, data), time= data$time)

p3 <- ggplot(data=data, aes(x=time, y =amount))+
geom_point(color='blue')+
    geom_line(color='red', data=predict_model4, aes(x=time, y=amount_pred))+
    ggtitle("R^2 = 0.9055, model3")
model5 <- lm(amount~time+I(time^2)+I(time^3)+I(time^4))
predict_model5 <- data.frame(amount_pred = predict(model4, data), time= data$time)

p4 <- ggplot(data=data, aes(x=time, y =amount))+
    geom_point(color='blue')+
    geom_line(color='red', data=predict_model5, aes(x=time, y=amount_pred))+
    ggtitle("R^2 = 0.9069, model4")
grid.arrange(p3, p4, ncol=2)
```

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.



Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Model fitting

- If we keep adding more complexity to the polynomial equation, we can make the model fit exactly the line

$$y = a + bx + cx^2 + dx^3 + ex^4 + fx^5 \ \ldots$$

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Akaike Information Criterion

- Complex models tend to not extend beyond the data they are modelling
- We should ask to what extent are we modelling noise (deviations due to error by adding a parameter.
- Akaike information criterion is a criterion for model selection that makes the model incur penalties for its complexity.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions of Automatic Control* 19: 6: 716-723.

Akaike Information Criterion

- k = the number of parameters

- *ln*($\hat{L}$) *=* the loglikelihood

$$AIC = 2K \ - 2\ln(\hat{L})$$

Akaike, H. A new look at the statistical model identification. *IEEE Transactions of Automatic Control* 19: 6: 716-723.

```
AIC(model2)
## [1] 257.0016
AIC(model3)
## [1] 230.4445
AIC(model4)
## [1] 229.9901
AIC(model5)
## [1] 230.3781
```

- The AIC is based on **information theory,** concerned with entropy, the degree of disorder in a system

Akaike, H. A new look at the statistical model identification. *IEEE Transactions of Automatic Control* 19: 6: 716-723.

‘Basic’ model fitting

- Once one understands the causal relationships we are interested in, we need some methodology for weighing simplicity against accuracy/fit.

- The simplest way of doing this is by starting with a maximal model and moving to a minimal adequate model

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Basic types of models

- **Saturated model:** There is one parameter for every data (perfect fit)

- **Maximal model**: Contains all factors, interactions and covariates that might be of interest (+ and that should be added considering their causal relations).

- **Minimal adequate model:** A simplified model which has removed superfluous variables.

- **Null model:** Just one parameter, the overall mean.

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Classic model selection process (simplified)

- **Fit maximal model:** Fit all the factors, interactions and covariates of interest. Note the Akaike Information Criterion

- **Begin model simplification:** Inspect the parameter estimates using summary(). Remove the least significant terms first, using update(), starting with highest order interactions.

- **What does the deletion do to the AIC?**
  - If it increases the AIC > Keep the interaction term and go back to step one looking at another term
  - If it decreases the AIC -> Leave the parameter deleted and continue to simplify the model
  - **Check assumptions:** Use plot() to check model assumptions making sure there is no heteroskedasticity (unequal scatter of residuals)

Crawley, Michael J. 2015. *Statistics: An Introduction using R.* Wiley.

Classic model selection process (simplified)

- Main problems / issues / questions for the trad-stat modelling practice:

- What should go in the maximal model? Does it incorporate confounds? By adding one variable do we confound our ability to assess the causal effect of another?

- Controversial: To what extent does the process of deleting variables result in an informative statistical model vis-a-vis causes? Are we really removing confounds?

- Can’t we fit more than one model to assess our causal relationships?

Iconicity data

- Let’s practice with the iconicity data of Winter (2017)

```
icon <- read.csv("/YourPathWay/perry_winter_2017_iconicity.csv")
head(icon)

##        Word         POS  SER CorteseImag Conc Syst    Freq  Iconicity
## 1         a Grammatical   NA          NA 1.46   NA 1041179  0.4615385
## 2     abide        Verb   NA          NA 1.68   NA     138  0.2500000
## 3      able   Adjective 1.73          NA 2.38   NA    8155  0.4666667
## 4     about Grammatical 1.20          NA 1.77   NA  185206 -0.1000000
## 5     above Grammatical 2.91          NA 3.33   NA    2493  1.0625000
## 6  abrasive   Adjective   NA          NA 3.03   NA      23  1.3125000
```

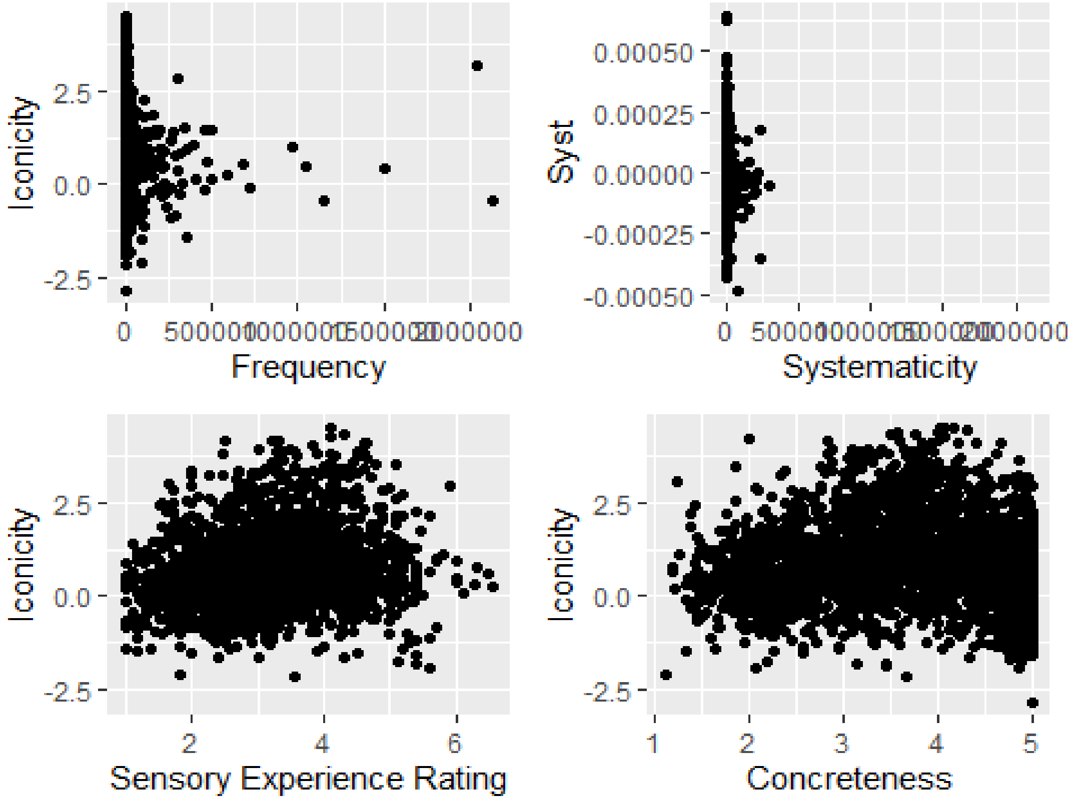
Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? Iconicity in English sensory words. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems, 18*(3), 443–464. https://doi.org/10.1075/is.18.3.07win

Iconicity data

- POS: part of speech
- SER: Sensory experience rating (does the word evoke a sensory experience)
- Conc: Concreteness
- Syst: Systematicity, overall contribution to form meaning correlation
- Freq: Frequency
- Iconicity: how much does the form sound like the word

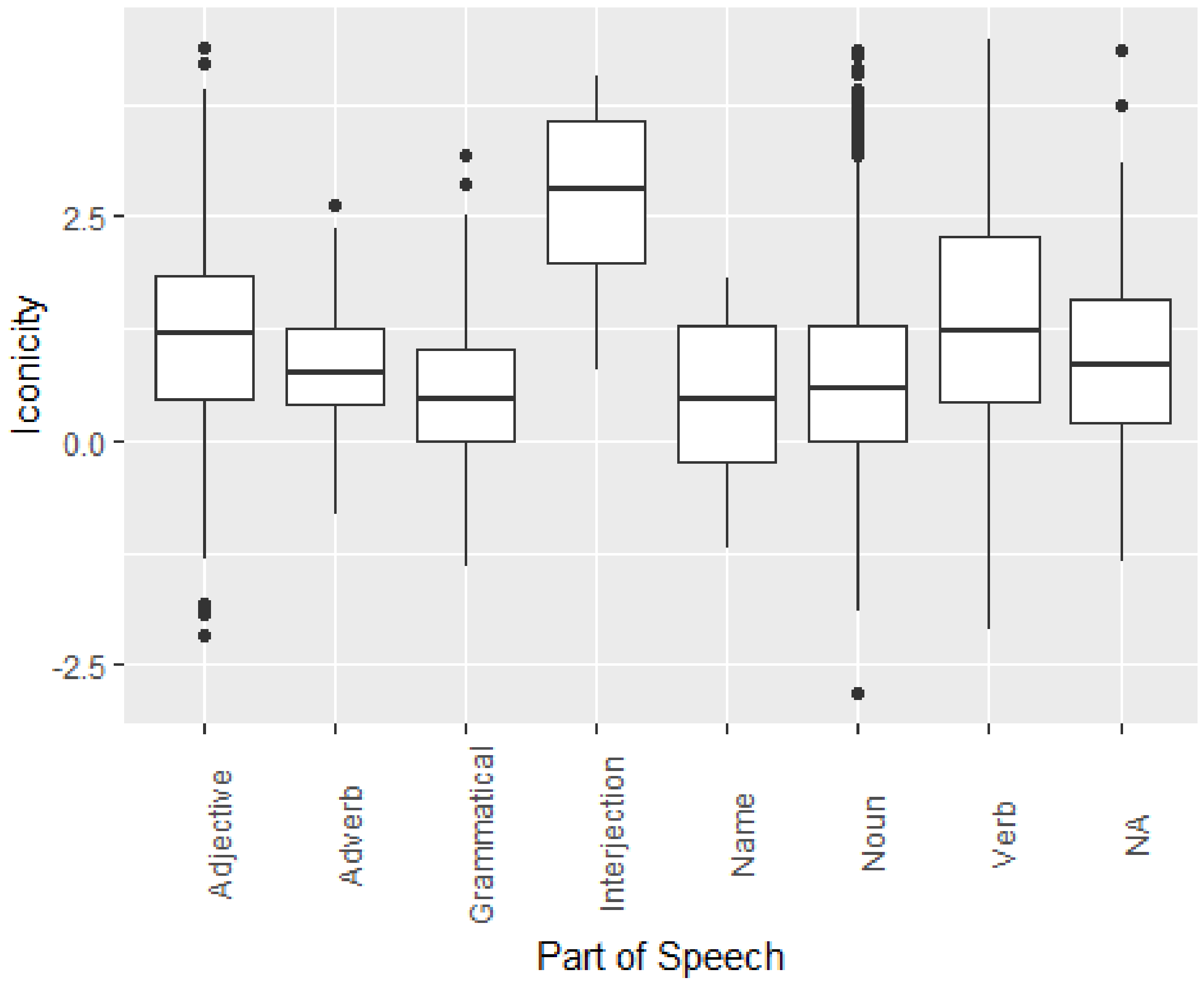
Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? Iconicity in English sensory words. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems, 18*(3), 443–464. https://doi.org/10.1075/is.18.3.07win

Data exploration




```
p1 <- ggplot(icon, aes(x=Freq, y = Iconicity))+
  geom_point()+
  xlab("Frequency")
p2 <- ggplot(icon, aes(x=Freq, y = Syst))+
  geom_point()+
  xlab("Systematicity")
p3 <- ggplot(icon, aes(x=SER, y = Iconicity))+
  geom_point()+
  xlab("Sensory Experience Rating")
p4 <- ggplot(icon, aes(x= Conc, y = Iconicity))+
  geom_point()+
  xlab("Concreteness")
grid.arrange(p1, p2, p3, p4, nrow = 2, ncol =2)
```

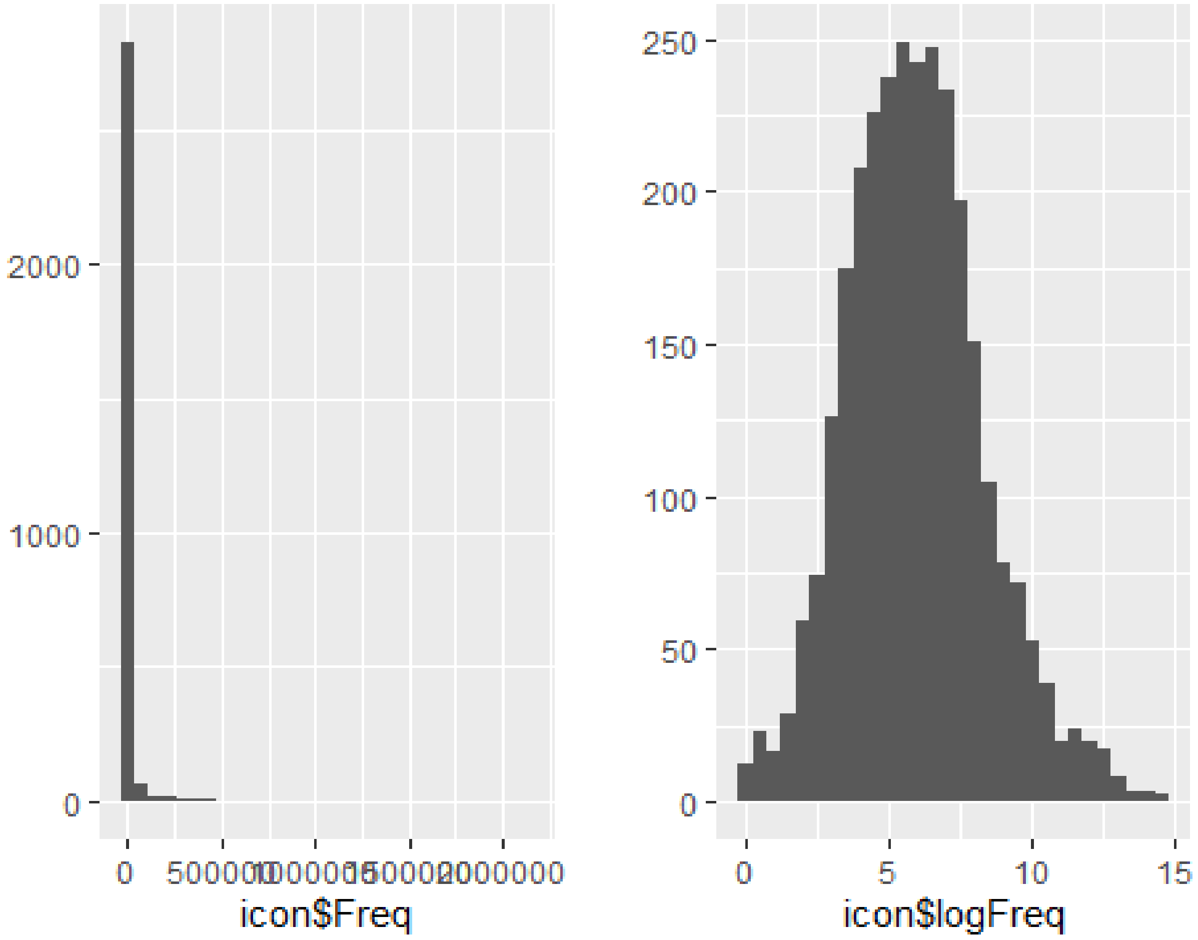
Data exploration




```
ggplot(icon, aes(x=POS, y = Iconicity))+
  geom_boxplot()+
  xlab("Part of Speech")+
  theme(axis.text.x = element_text(angle = 90))
```

Data transformation

```
p1 <- qplot(icon$Freq)

icon$logFreq <- log(icon$Freq)
p2 <- qplot(icon$logFreq)
grid.arrange(p1, p2, ncol=2)
```



```r
model.saturated <- lm(Iconicity~logFreq+Syst+Conc+SER, data=icon)
summary(model.saturated)

##
## Call:
## lm(formula = Iconicity ~ logFreq + Syst + Conc + SER, data = icon)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.12346 -0.73861 -0.07942  0.66380  2.82933
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.88197    0.22289   8.443  < 2e-16 ***
## logFreq      -0.13414    0.01717  -7.813 1.43e-14 ***
## Syst        376.62000  270.60854   1.392    0.164
## Conc         -0.34187    0.03967  -8.618  < 2e-16 ***
## SER           0.47043    0.04128  11.396  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 976 degrees of freedom
##   (2020 observations deleted due to missingness)
## Multiple R-squared:  0.1859, Adjusted R-squared:  0.1826
## F-statistic: 55.71 on 4 and 976 DF,  p-value: < 2.2e-16
```

```r
model2 <- lm(Iconicity~logFreq+Conc+SER, data=icon)
summary(model2)
```

```
##
## Call:
## lm(formula = Iconicity ~ logFreq + Conc + SER, data = icon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2650 -0.7107 -0.0936  0.6282  3.3881
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37487    0.15695   8.760  < 2e-16 ***
## logFreq     -0.09372    0.01244  -7.535 7.75e-14 ***
## Conc        -0.13750    0.02771  -4.962 7.65e-07 ***
## SER          0.19445    0.02764   7.035 2.84e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.051 on 1761 degrees of freedom
##   (1236 observations deleted due to missingness)
## Multiple R-squared:  0.06841,    Adjusted R-squared:  0.06683
## F-statistic: 43.11 on 3 and 1761 DF,  p-value: < 2.2e-16
```

AIC

```
AIC(model.saturated)
## [1] 2831.413
AIC(model2)
## [1] 5188.832
```