R-lecture-notes-2023-12-13-multivariate-models-causation-interactions

Adam Tallman

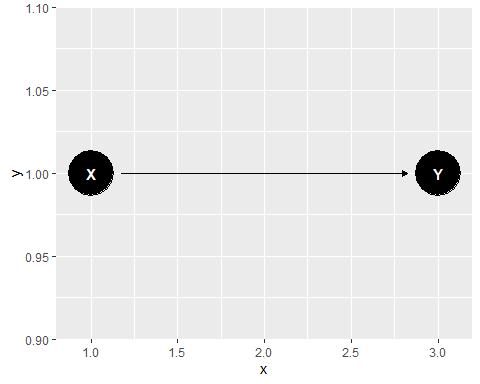
2023-12-11

knitr::opts\_chunk$set(echo = TRUE)  
library(dagitty)  
library(ggdag)  
library(V8)  
library(Rling)  
library(AICcmodavg)  
library(tidyverse)  
library(gridExtra)

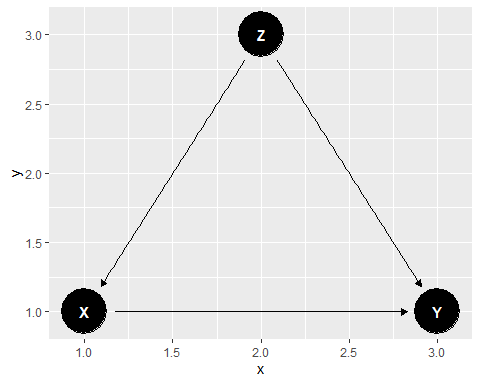
## Multivariate regression

* We have dealt with cases where there is one predictor
* The power of regression comes for multivariate regression
* This is the main tool that allows us to distinguish between cause and effect

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)



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)



Let’s simulate a confound which shows why its important to have more than one variable.

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))

##   
## 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+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

## Interactions

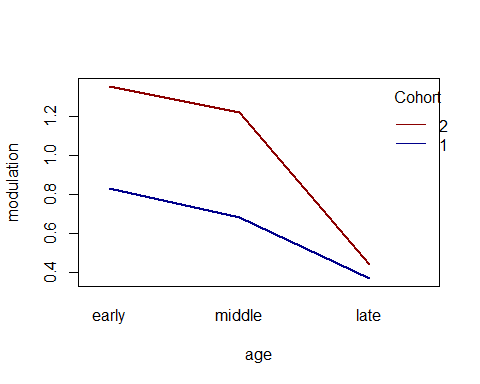
data("sharedref")  
head(sharedref)

## 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

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

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")



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

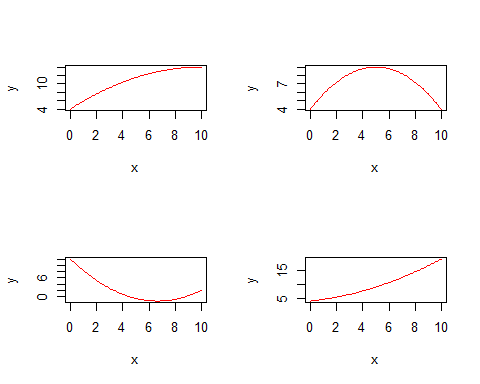
## Analysis of Variance Table  
##   
## Response: mod  
## Df Sum Sq Mean Sq F value Pr(>F)   
## age 2 4.2243 2.11217 491.884 < 2.2e-16 \*\*\*  
## cohort 1 1.7101 1.71008 398.243 < 2.2e-16 \*\*\*  
## age:cohort 2 0.5679 0.28397 66.132 1.054e-13 \*\*\*  
## Residuals 42 0.1804 0.00429   
## ---  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

summary(model2)

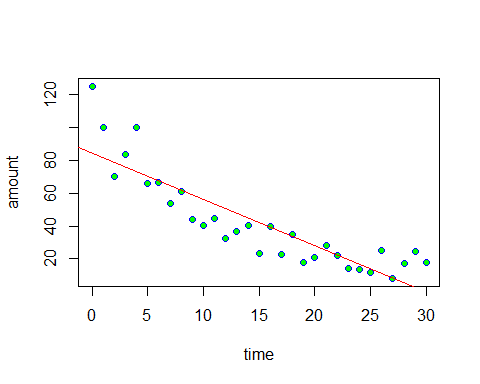
##   
## Call:  
## lm(formula = mod ~ age \* cohort, data = sharedref)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.16000 -0.03438 0.00000 0.04000 0.11750   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.816250 0.009458 86.30 < 2e-16 \*\*\*  
## age1 -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

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")



par(mfrow=c(1,1))  
data <- read.csv("/Users/User/Documents/GitHub/StatisticsinLinguistics\_FSU\_2023\_2024/07\_data/decay.csv", header=TRUE)  
attach(data)  
plot(time,amount,pch=21,col="blue",bg="green")  
abline(lm(amount~time),col="red")



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

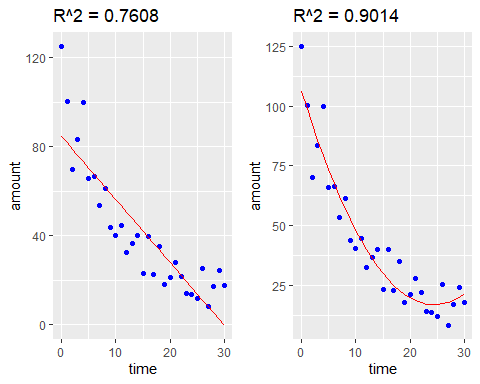
summary(model2)

##   
## 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

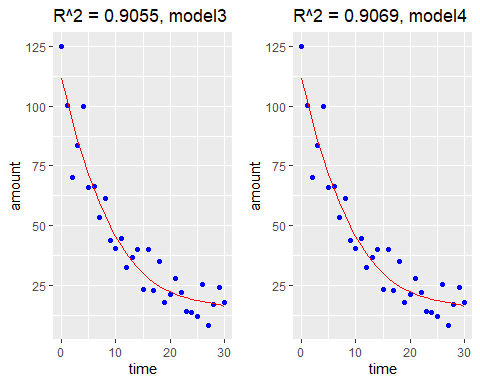
summary(model3)

##   
## 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

predict\_model3 <- data.frame(amount\_pred = predict(model3, data),  
 time= data$time)  
p1 <- ggplot(data=data, aes(x=time, y =amount))+  
 geom\_point(color='blue')+  
 geom\_line(color='red', data=predict\_model3, aes(x=time, y=amount\_pred))+  
 ggtitle("R^2 = 0.9014")  
predict\_model2 <- data.frame(amount\_pred = predict(model2, data), time= data$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)



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)

 ## Akaike Information Criterion

AIC(model2)

## [1] 257.0016

AIC(model3)

## [1] 230.4445

AIC(model4)

## [1] 229.9901

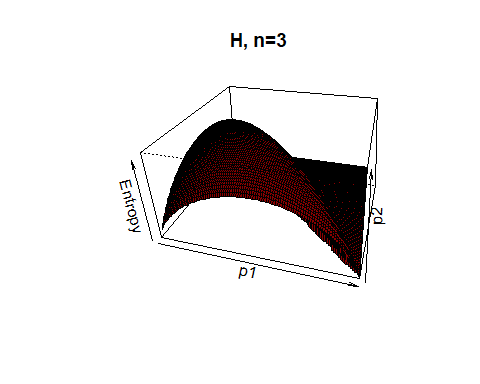
AIC(model5)

## [1] 230.3781

entropytwo<-function(x,y)  
{z<- 1-x-y  
ifelse(z>0,-x\*log2(x)-y\*log2(y)-z\*log2(z),0)  
}  
x<-(0:100)/100  
y<-(0:100)/100  
z<-outer(x,y,entropytwo)

## Warning in ifelse(z > 0, -x \* log2(x) - y \* log2(y) - z \* log2(z), 0): NaNs  
## produced

persp(x, y, z, theta = 15, phi = 30, expand =  
0.5, col = "darkred", xlab="p1", ylab="p2",  
zlab="Entropy", main="H, n=3")



## Multivariate regression

icon <- read.csv("/Users/User/Documents/GitHub/StatisticsinLinguistics\_FSU\_2023\_2024/07\_data/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

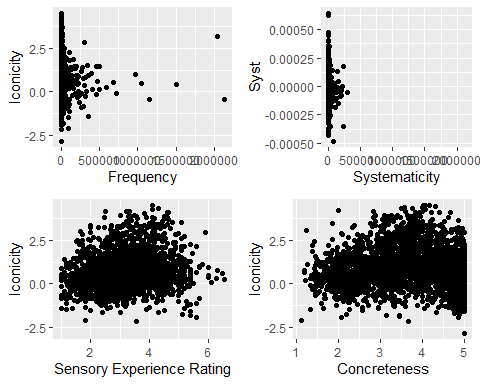
p1 <- ggplot(icon, aes(x=Freq, y = Iconicity))+  
 geom\_point()+  
 xlab("Frequency")  
p2 <- ggplot(icon, aes(x=Freq, y = Syst))+  
 geom\_point()+  
 xlab("Systematicity")  
p3 <- ggplot(icon, aes(x=SER, y = Iconicity))+  
 geom\_point()+  
 xlab("Sensory Experience Rating")  
p4 <- ggplot(icon, aes(x=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

## Warning: Removed 53 rows containing missing values (`geom\_point()`).

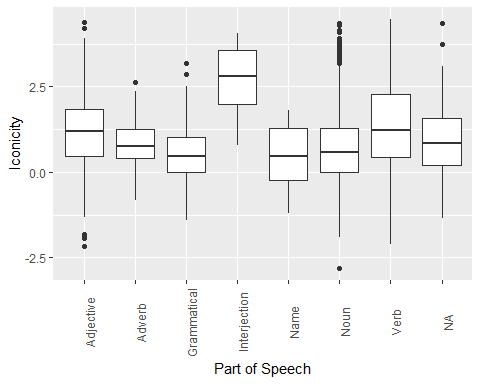
## Warning: Removed 1898 rows containing missing values (`geom\_point()`).

## Warning: Removed 1222 rows containing missing values (`geom\_point()`).

## Warning: Removed 181 rows containing missing values (`geom\_point()`).



ggplot(icon, aes(x=POS, y = Iconicity))+  
 geom\_boxplot()+  
 xlab("Part of Speech")+  
 theme(axis.text.x = element\_text(angle = 90))



p1 <- qplot(icon$Freq)

## Warning: `qplot()` was deprecated in ggplot2 3.4.0.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

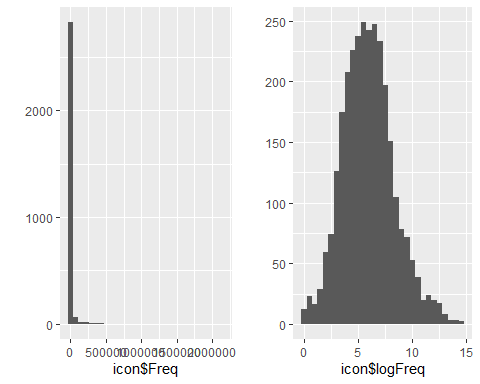
icon$logFreq <- log(icon$Freq)  
p2 <- qplot(icon$logFreq)  
grid.arrange(p1, p2, ncol=2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 53 rows containing non-finite values (`stat\_bin()`).

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 53 rows containing non-finite values (`stat\_bin()`).



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

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(model.saturated)

## [1] 2831.413

AIC(model2)

## [1] 5188.832