Zuur Ch 02 part 1 slides

Ed Harris

26/03/2020

Ch 02 outline

-Learn as you go philosophy

-Data exploration

-Linear regression modelling

-Linear regression assumptions

2.1 Data exploration

Nereis data

here I think:
concentration (numeric) nutrient concentration
biomass (numeric) polychaete biomass
nutrient (factor) nutrient type - reads in as numeric but is actually a cetegorical factor

```
##   concentration biomass nutrient
## 1         0.050     0.0        1
## 2         0.105     0.0        1
## 3         0.105     0.0        1
## 4         0.790     0.5        1
## 5         0.210     0.5        1
## 6         2.100     0.5        1
```

Fig 2.1 (modified)

```
par(mfrow = c(1,2))
Nereis <- read.table(file = "Nereis.txt", header = T)
dotchart(Nereis$concentration,   groups = factor(Nereis$nutrient),
  ylab = "Nutrient",xlab = "Concentration",
  main = "Clevelanddotplot",  pch = Nereis$nutrient)

plot(jitter(nutrient) ~ concentration, data = Nereis,
     pch = Nereis$nutrient)
```

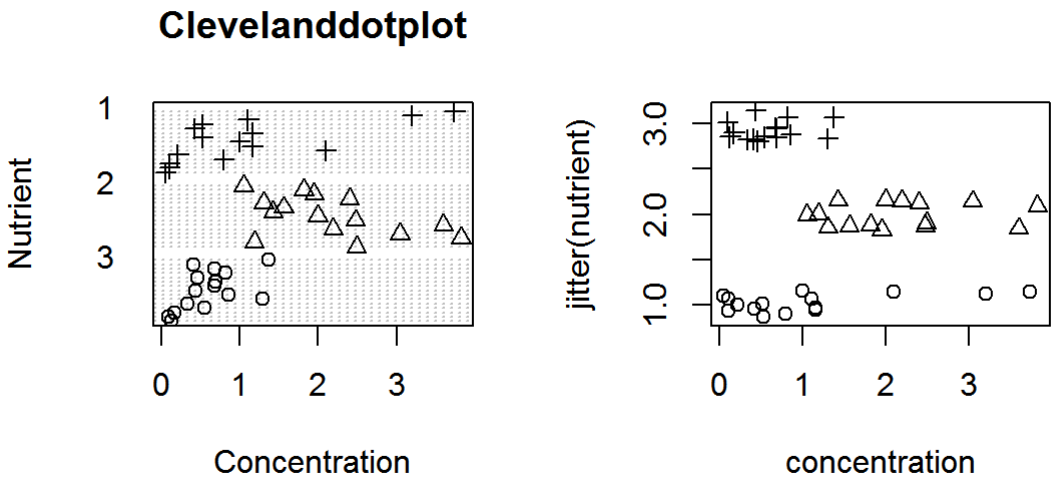


Fig 2.2 pairs plot

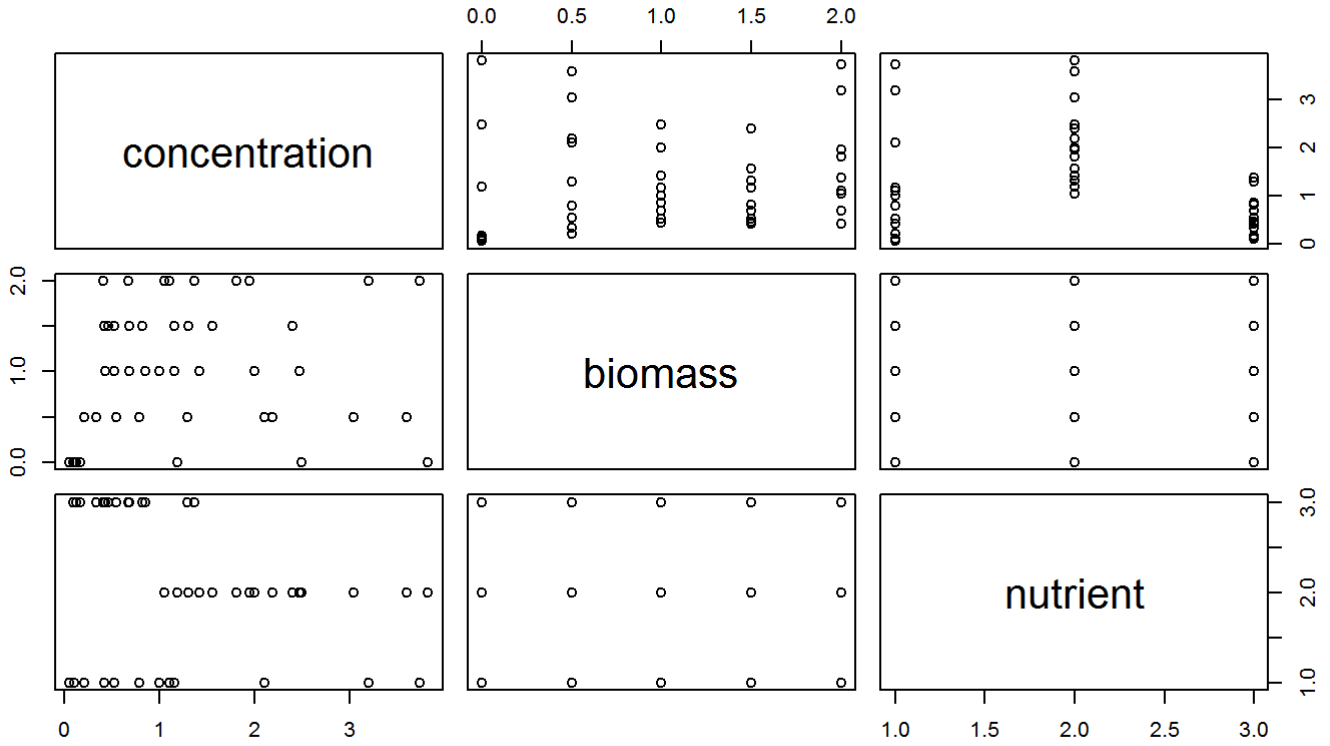
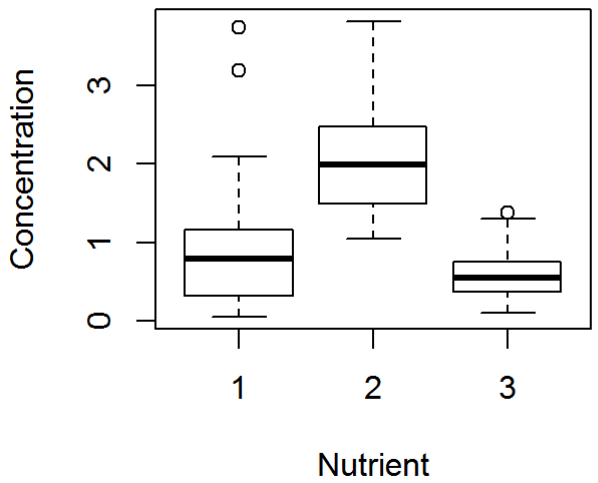


Fig 2.3 Boxplot

```
# NB possible effect (bigger mean Nutrient 2),
# Also variance seems lower on nutrient level 3
boxplot(concentration ~ factor(nutrient),
        data = Nereis,
        ylab = "Concentration",
        xlab = "Nutrient")
```



2.2 Your old buddy the linear model

$$Y_i = \alpha + \beta \times X_i + \epsilon_i$$

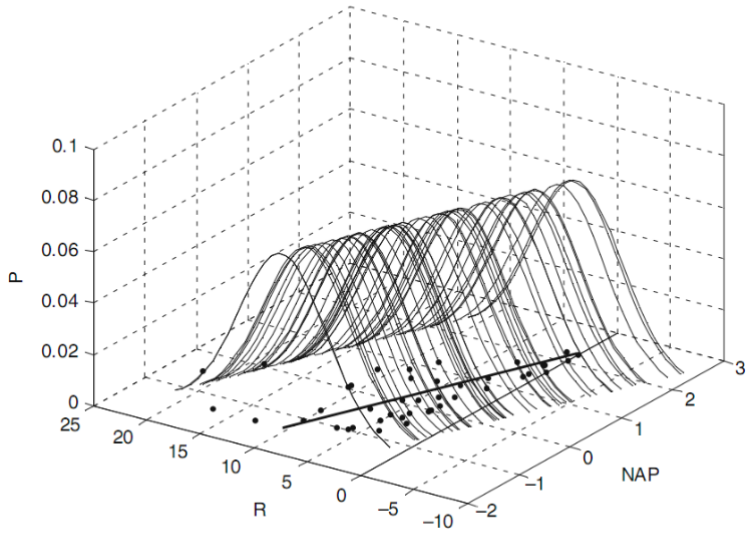
Y_i = dependent var
X_i = explanatory var
α and β = intercept and slope
ϵ_i = residual error

$$\epsilon_i \sim N(0, \sigma^2)$$

Assumption the residual error is Gaussian with expected value = 0, variance = σ^2

The “magic of assumptions”

-assume Gaussian residual error
-homoscedasity of error
-no weird values
RIKZ data, R = spp richness, NAP = tide height
P = probability density



2.3 Violate those assumptions

Vanilla linear model full assumptions

-Gaussian residuals

-Homogeneous variance

-“fixed” X (discuss briefly)

-Independence

-Correct model specification…

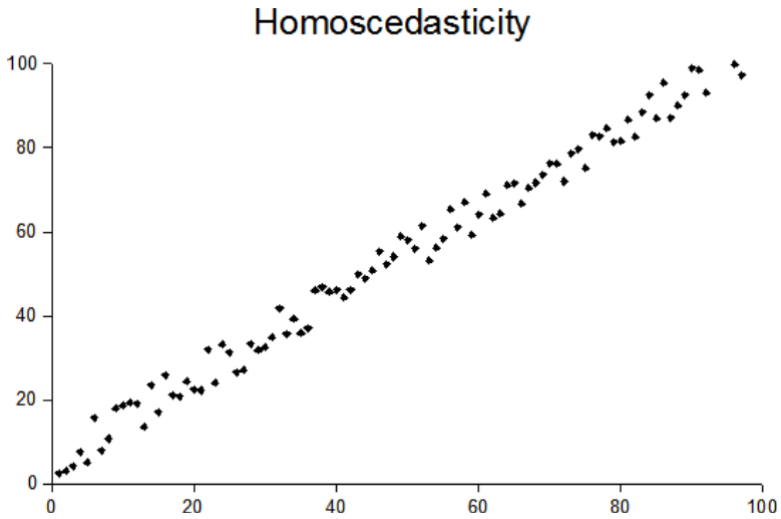
2.3.2 Gaussian residuals

“the underlying concept of normality is grossly misunderstood by many researchers. The linear regression model requires normality of the data, and therefore of the residuals at *each X* value”

Important but is it black and white? (Sokal and Rohlf, 1995; Zar, 1999)

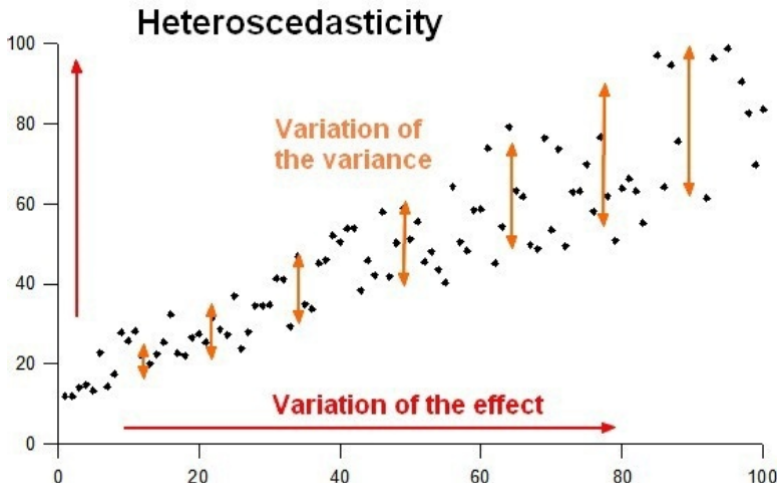
2.3.3 Heterogeneity

“heterogeneity (violation of homogeneity), also called heteroscedasticy, happens if the spread of the data is not the same at each X value, and this can be checked by comparing the spread of the residuals for the different X values”



2.3.3 Heterogeneity

“heterogeneity (violation of homogeneity), also called heteroscedasticy, happens if the spread of the data is not the same at each X value, and this can be checked by comparing the spread of the residuals for the different X values”



Fixed X

-Concept of fixed versus random “effects”

-Explanatory variables are fixed if:
1) experimentally assigned
2) low error in sample estimate relative to pop’n

-Can be serious (ref to Faraway 2005)

Independence

This is the most serious of violated assumptions in linear models and is very, very common too.

2 related causes:

-Dependence structure inherent in the model
(e.g. multiple samples in a plot)

-Other dependence in the data
(e.g. measuring growth at multiple points in time)

NB we fix this with a mixed effects model…

2.3.6 wedge clams

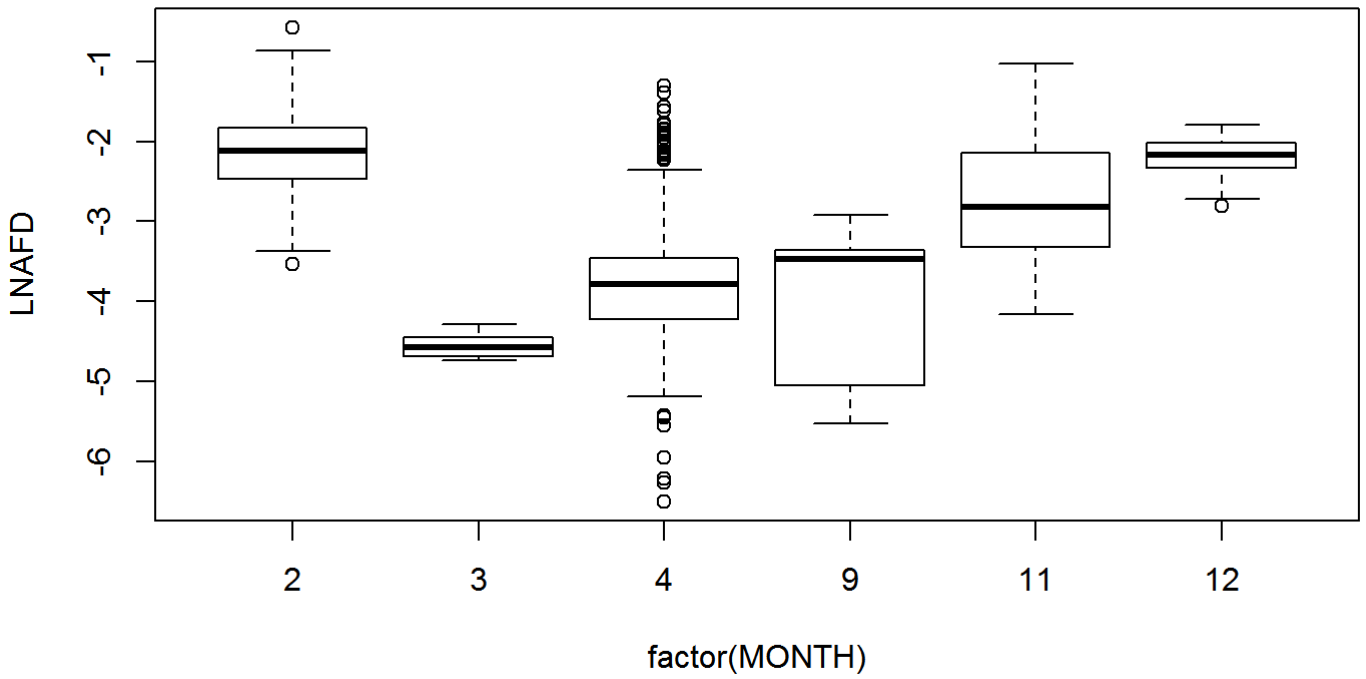
2.3.6 wedge clams

```
Clams <- read.table("Clams.txt", header = T)
str(Clams)
```

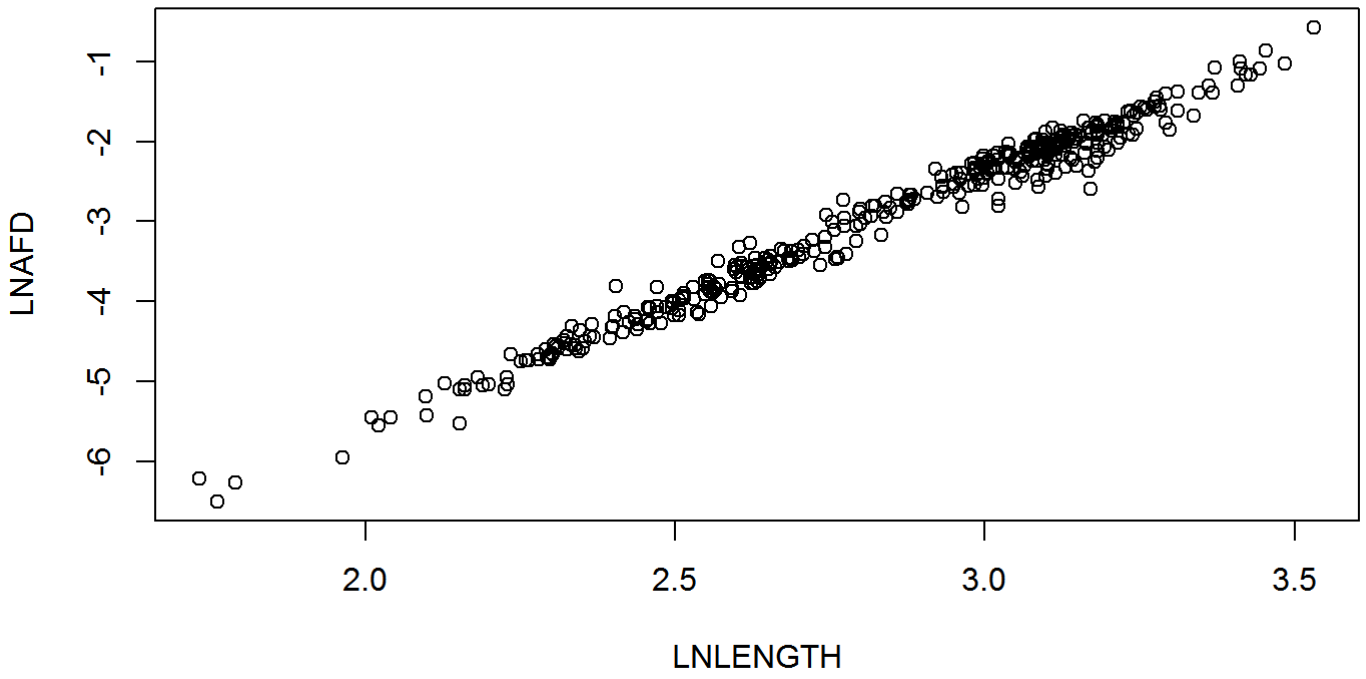
```
## 'data.frame':    398 obs. of  5 variables:
##  $ MONTH   : num  11 11 11 11 11 11 11 11 11 11 ...
##  $ LENGTH  : num  28.4 16.6 13.7 17.4 11.8 ...
##  $ AFD     : num  0.248 0.052 0.028 0.07 0.022 0.187 0.361 0.05 0.087 0.128 ...
##  $ LNLENGTH: num  3.35 2.81 2.62 2.86 2.47 ...
##  $ LNAFD   : num  -1.39 -2.96 -3.57 -2.65 -3.83 ...
```

Month - month of measurement
Length - length (mm?)
AFD - weight
LNLENGTH - log(Length)
LMAFD - log(AFD)

2.3.6 wedge clams



2.3.6 wedge clams



2.3.6 wedge clams

models: LNAFN ~ LNLENGTH + MONTH LNAFN ~ LNLENGTH * MONTH

```
Clams$MONTH <- factor(Clams$MONTH)
M1 <-lm(LNAFD ~ LNLENGTH * MONTH, data = Clams)
drop1(M1, test = "F")
```

```
## Single term deletions
##
## Model:
## LNAFD ~ LNLENGTH * MONTH
##                Df Sum of Sq    RSS     AIC F value  Pr(>F)
## <none>                      6.4490 -1616.8
## LNLENGTH:MONTH  5   0.20328 6.6523 -1614.4  2.4334 0.03444 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.3.6 wedge clam model validation