ADVANCED STATISTICAL MODELING

Version 1, 2018

Jacolien van Rij [j.c.van.rij@rug.nl]
Hermine Berberyan, and Stefan Huijser

Today's topic:

GAMMS & MODEL CRITICISM

TESTING FOR SIGNIFICANCE

- Three methods:
  1. Model comparisons
     - `method="REML"` for random effects
     - `method="ML"` for fixed effects
  2. Summary statistics
     - not for continuous x categorical interactions
  3. Plots
     - `plot_diff` (package `itsadug`)

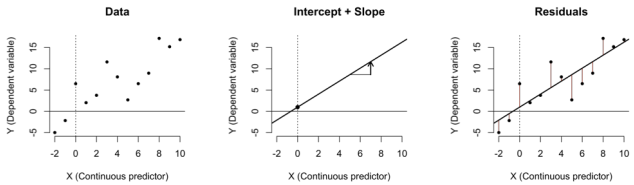
too slow for big data!

TESTING FOR SIGNIFICANCE

RELIABILITY OF CONCLUSIONS IS DEPENDENT ON MODEL VALIDITY

REGRESSION ANALYSIS

DATA = MODEL + RESIDUALS



REGRESSION ANALYSIS

- Assumptions:
  - normality
  - linearity
  - homogeneity of variance / no heteroscedasticity
  - **uncorrelated predictors**
  - residuals are independent / no structure in residuals
  - no bad outliers

QQ-plot
GAMMs
plot resid by predictor/fitted

COLLINEARITY

- Correlation between multiple predictors
  - co-occurring events, cause-effect relations, …
  - ➔ Predictors account for overlapping variance
    - it is not possible to identify the unique contribution of each predictor
- Partial solutions: centering, scaling, re-organizing predictors using principle components analysis (PCA)

  More info: Van Rij et al, NetWordS chapter (nestor)

REGRESSION ANALYSIS

- Assumptions:
  - normality
  - linearity
  - homogeneity of variance / no heteroscedasticity
  - uncorrelated predictors
  - **residuals are independent / no structure in residuals**
  - no bad outliers

QQ-plot
GAMMs
plot resid by predictor/fitted

INDEPENDENT RESIDUALS

- Observations are not independent:
  - Repeated-measures design
    - Subjects / items produce multiple responses / measurements
  - Time series data
    - Multiple measures per trial
- If unaccounted for, this results in structure in residuals

TIME SERIES DATA

- The value of each measure is (partly) determined by the previous samples
- Examples:
  - weather & environment: temperature, precipitation, CO2
  - financial: stock market analysis
  - psycholinguistic data:
    - EEG, eye tracking, articulography, …
    - RTs, …

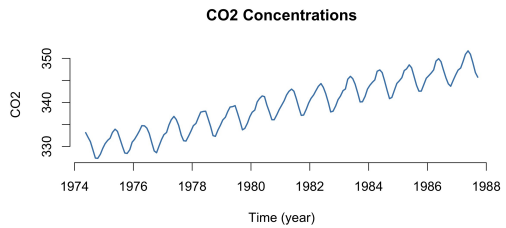
CO2 EXAMPLE

```
> dat <- read.table('co2.txt', header=T, sep='\t')
> head(dat)

     CO2 YearMonth Year Month
1 333.13   1974.38 1974     5
2 332.09   1974.46 1974     6
3 331.10   1974.54 1974     7
4 329.14   1974.63 1974     8
5 327.36   1974.71 1974     9
6 327.29   1974.79 1974    10
```


12

CO2 EXAMPLE

```
> plot(dat$YearMonth, dat$CO2, type='l',
      xlab='Time (year)', ylab='CO2',
      main='CO2 Concentrations')
```



13

CO2 EXAMPLE

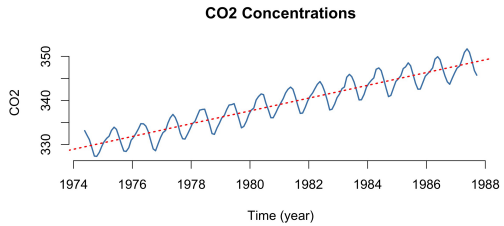
```
# linear regression:
> lm1 <- lm(CO2 ~ YearMonth, data=dat)

> summary(lm1)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2537.189     88.117 -28.793        0
YearMonth       1.452      0.044  32.642        0
```


CO2 EXAMPLE

```
# linear regression:
> dat$Time <- dat$YearMonth - 1974
> lm1 <- lm(CO2 ~ Time, data=dat)

> summary(lm1)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  328.906      0.358 919.941        0
Time           1.452      0.044  32.642        0

# CO2 = 328.906 + 1.452*(YearMonth - 1974)
```

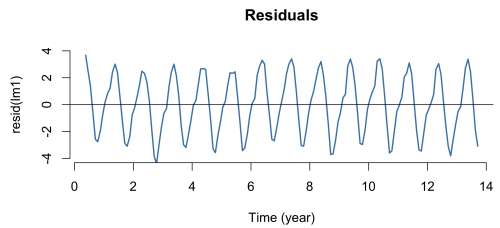

CO2 EXAMPLE

```
# CO2 = 328.906 + 1.452*(YearMonth - 1974)
```



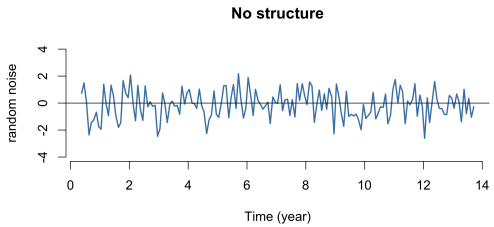
CO2 EXAMPLE

```
> plot(dat$Time, resid(lm1), type='l',
      xlab='Time (year)', ylab='resid(lm1)',
      main='Residuals')
```



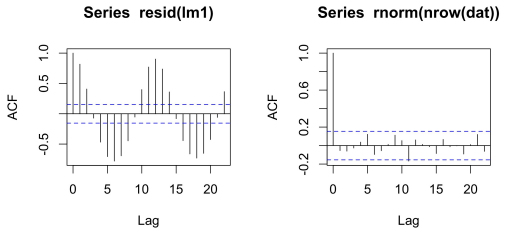
CO2 EXAMPLE

```
> plot(dat$Time, rnorm(nrow(dat)), type='l',
      xlab='Time (year)', ylab='resid(lm1)',
      main='Residuals')
```

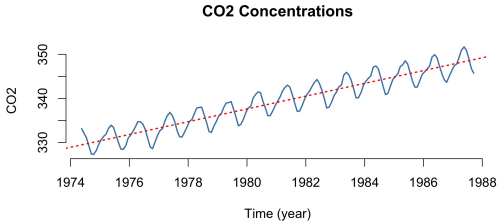


ACF PLOT

```
> par(mfrow=c(1,2), cex=1.2)
> acf(resid(lm1))
> acf(rnorm(nrow(dat)))
```



IMPROVING MODEL FIT



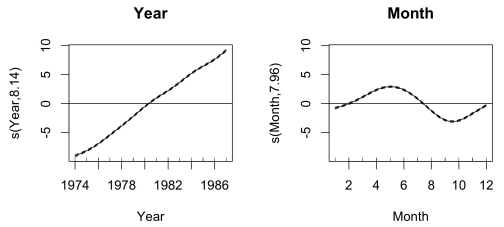
IMPROVING MODEL FIT

```
> head(dat)
     CO2 YearMonth Year Month
1 333.13   1974.38 1974     5
2 332.09   1974.46 1974     6
3 331.10   1974.54 1974     7
4 329.14   1974.63 1974     8
5 327.36   1974.71 1974     9
6 327.29   1974.79 1974    10

> gam1 <- gam(CO2 ~ s(Year) + s(Month), data=dat)
```

21

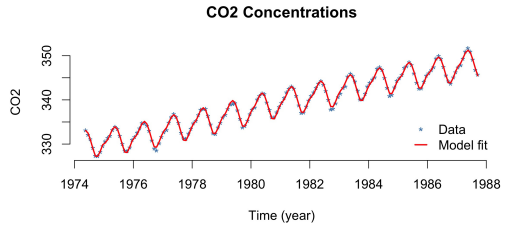
IMPROVING MODEL FIT

```
> plot( gam1 )
```



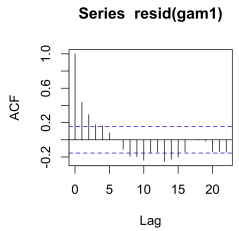
IMPROVING MODEL FIT

```
> dat$fit <- fitted( gam1 )
> plot(dat$YearMonth, dat$fit, type='l')
```



IMPROVING MODEL FIT

```
> acf( resid( gam1 ) )
```



INCLUDING AR1 MODEL

- Taking into account that the residuals are correlated
- AR(ρ) model: autoregressive model of order ρ
  - AR(1) : $X_t = \phi X_{t-1} + \varepsilon_t$
  - AR(ρ) : $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_\rho X_{t-\rho} + \varepsilon_t$

$$y_i = \beta_0 + f(x_i) + \varepsilon_i$$
$$\varepsilon_i = \rho \varepsilon_{i-1} + \text{noise}$$

- Including AR1 model:

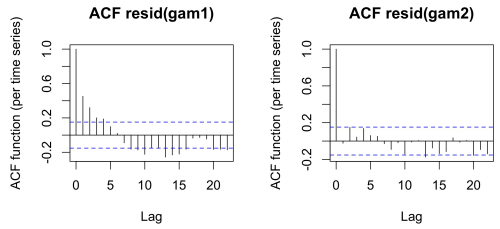
INCLUDING AR1 MODEL

```
gam1 <- gam(CO2 ~ s(Year) + s(Month), data=dat)

# find the autocorrelation value at Lag 1:
> rho1 <- start_value_rho(gam1)
[1] 0.4515055

# run new model with autocorrelation parameter:
> dat$start <- c(TRUE, rep(FALSE, nrow(dat)-1) )

> gam3 <- bam(CO2 ~ s(Year) + s(Month), data=dat,
      rho=rho1, AR.start=dat$start )
```


26

INCLUDING AR1 MODEL

```
> library(itsadug)
> acf_resid(gam1, main='ACF resid(gam1)')
> acf_resid(gam2, main='ACF resid(gam2)')
```



REGRESSION ANALYSIS

- Assumptions:
  - normality — QQ-plot
  - linearity — GAMMs, plot resid by predictor/fitted
  - homogeneity of variance / no heteroscedasticity — plot resid by predictor/fitted
  - uncorrelated predictors
  - **residuals are independent / no structure in residuals** — check with ACF plot + resid by predictor/fitted plot
  - no bad outliers — check with ACF plot + resid by predictor/fitted plot

1) improve model fit, 2) include AR(1)