Linear regression, collinearity, splines and extensions

Morten Frydenberg ©
Department of Biostatistics, Aarhus Univ, Denmark

General things for regression models:

- **Collinearity** - correlated explanatory variables
- **Flexible modelling af response curves** - Cubic splines

Normal regression models – an extension

- **Clustered** data / data with several random components

Collinearity

Consider a subsample of the serum cholesterol data set and the **three** models:

```
model 0:   regress logscl sex sbp dbp
model 1:   regress logscl sex     dbp
model 2:   regress logscl sex sbp
```

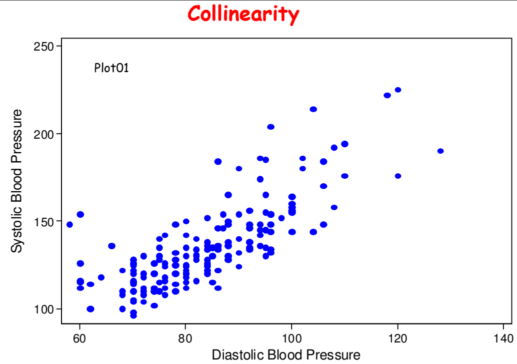
Variable	model0	model1	model2
sbp	**.00126448**		**.0014988**
	.00087992		**.0005548**
	0.1524		**0.0075**
dbp	**.00056517**	**.00239702**	
	.00164485	**.0010424**	
	0.7315	**0.0226**	
sex	.02080574	.02446746	.0197773
	.02636149	.02631111	.02613048
	0.4310	0.3536	0.4501
_cons	5.1444085	5.1555212	5.1615877
	.09912234	.09909537	.08539118
	0.0000	0.0000	0.0000
N	194	194	194

legend: b/se/p

Estimate
Se
p

Each BP-measure is statistical significant, when the other is removed!

Collinearity



SBP and DBP are **highly positively correlated**, that will lead to **highly negatively correlated estimates**!!!

Collinearity

This can be seen by listing the **correlation between the estimates.**
In Stata by the command: `vce, cor`

```
regress logscl sbp dbp sex
vce,cor
             |      sbp       dbp       sex     _cons
-------------+----------------------------------------
         sbp |   1.0000
         dbp |  -0.7750    1.0000
         sex |  -0.0967    0.1135    1.0000
       _cons |  -0.0780   -0.5044   -0.4665    1.0000
```

If two estimates are highly correlated, it indicates that it is very difficult to estimate the **"independent effect"** of the each of the two variables.

Often it is even **nonsense** to try to do it!

Often it is better to try to **reformulate the problem**.

Collinearity

One way to work around the problem of collinearity is to '**ortogonalize**' it:

Create two new variable:

one measures the **blood pressure**

and another that measure the **difference** in systolic and diastolic blood pressure.

Some **candidates**:

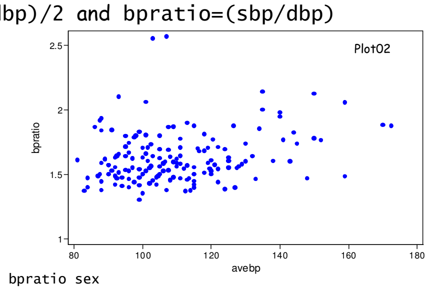
```
(sbp+dbp)/2     and   (sbp-dbp)
(sbp+dbp)/2     and   (sbp/dbp)
ln(sbp*dbp)/2   and   ln(sbp/dbp)
```

We will here consider the second pair.

Collinearity

`avebp=(sbp+dbp)/2 and bpratio=(sbp/dbp)`

Only weakly associated




```
regress logscl avebp bpratio sex
vce,cor
             |    avebp   bpratio       sex     _cons
-------------+----------------------------------------
       avebp |   1.0000
     bpratio |  -0.2456    1.0000
         sex |   0.0382   -0.1041    1.0000
       _cons |  -0.4542   -0.6874   -0.2585    1.0000
```

Collinearity

The serum cholesterol data set and the **three** models:

```
model 0:   regress logscl sex avebp bpratio
model 1:   regress logscl sex avebp
model 2:   regress logscl sex       bpratio
```

Variable	model0	model1	model2
avebp	.00198973	.00206564	
	.0007887	.00076285	
	0.0125	0.0074	
bpratio	.02769662		.07148118
	.07067134		.06946246
	0.6956		0.3048
sex	.02060675	.02168128	.01806662
	.02632924	.026128	.02667689
	0.4348	0.4077	0.4991
_cons	5.1003417	5.1351912	5.2485724
	.12936418	.09374803	.11685799
	0.0000	0.0000	0.0000
N	194	194	194

legend: b/se/p

Blood pressure seems to play a role,

The ratio between SBP and DBP might not.

Collinearity

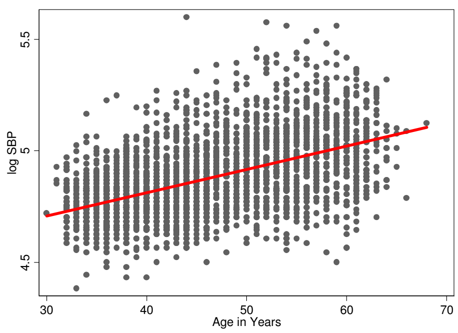
Look out for it:

- systolic and diastolic blood pressure
- 24 hour blood pressure and 'clinical' blood pressure
- weight and height
- age and parity
- age and time since menopause
- BMI and skinfold measure
- age , birth cohort and calendar time
- volume and concentration
- ……

Remember you will need **a huge amount** of data to disentangle the effects of correlated explanatory variables

Flexible modelling of response curves – cubic splines

Log SBP against age for 2650 women with fitted straight line.



Flexible modelling of response curves – cubic splines

We want to model the relationship between SBP and age more flexible.

There are several ways to do this, including fractional polynomial, splines and cubic splines.

We will here look at restricted cubic splines as they are implemented in Stata.

If one want to use the restricted cubic splines you start by generating a set of new independent variables:

```
mkspline sage=age, cubic nk(6) disp
```

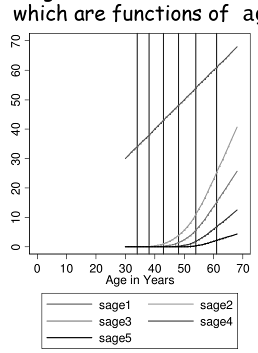
	knot1	knot2	knot3	knot4	knot5	knot6
age	34	38	43	48	54	61

Flexible modelling of response curves – cubic splines

The `mkspline` command will generate 5 new variables named `sage1` to `sage5`, which are functions of `age`.

Where `sage1=age.`

```
sage2=0 if age<34
sage3=0 if age<38
sage4=0 if age<43
sage5=0 if age<48
```



Flexible modelling of response curves – cubic splines

$$knots: \quad a_1, a_2, \ldots, a_k$$

$$sage_i = age$$

$$sage_{j+1} = \left(age - a_j\right)_+^3 - \left(age - a_{k-1}\right)_+^3 \frac{a_k - a_j}{a_k - a_{k-1}} + \left(age - a_k\right)_+^3 \frac{a_{k-1} - a_j}{a_k - a_{k-1}}$$

Flexible modelling of response curves – cubic splines

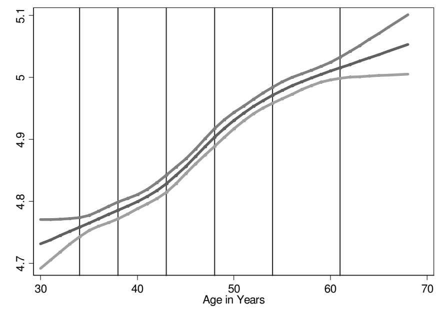
```
drop sage1
regress lsbp age sage?
------------------------------------------------------------------------------
        lsbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0067837   .0035322     1.92   0.055    -.0001425    .0137099
       sage2 |  -.0005598   .0525269    -0.01   0.991    -.1035577    .1024381
       sage3 |   .0553357   .1336906     0.41   0.679    -.2068131    .3174845
       sage4 |  -.1398205   .1547781    -0.90   0.366    -.4433189    .1636778
       sage5 |   .0932052   .1207685     0.77   0.440    -.1436051    .3300155
       _cons |   4.527844   .1253021    36.14   0.000     4.282144    4.773544
------------------------------------------------------------------------------
testparm sage?
 ( 1)  sage2 = 0
 ( 2)  sage3 = 0
 ( 3)  sage4 = 0
 ( 4)  sage5 = 0
       F(  4,  2644) =    3.81
            Prob > F =    0.0043
```

Test of linearity
The hypothesis is rejected

The relationship is not linear, but how does it look ?

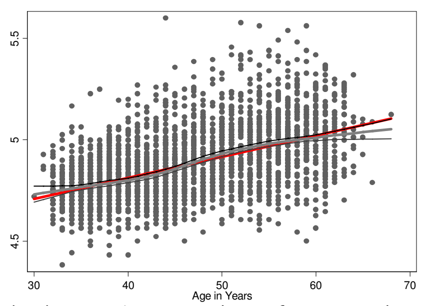
Flexible modelling of response curves – cubic splines

```
predict fit if e(sample)                     /// fit values
predict fitsd if e(sample),stdp              /// standard error
generate low=fit-1.96*fitsd                  /// lower ci-limit
generate hig=fit+1.96*fitsd                  /// upper ci-limit
line fit low hig  age                        /// plot
```



Flexible modelling of response curves – cubic splines

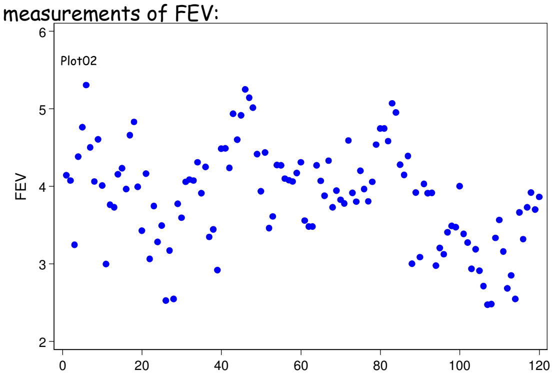
Compare with the straight line model:



Although, there is 'statistical significant' non-linearity, it has no practical implications- the straight line model is a valid approximation.

Clustered data / data with several random components

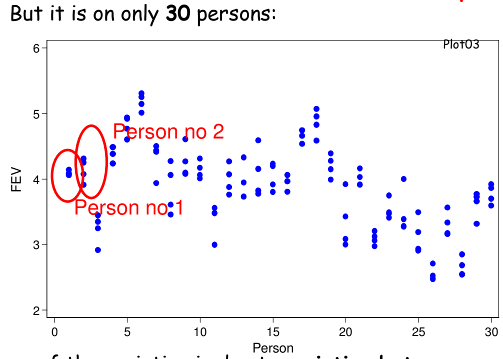
120 measurements of FEV:



Some variation in the data.

Clustered data / data with several random components

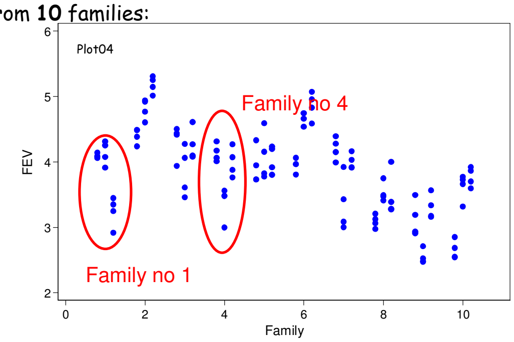
But it is on only **30** persons:



Some of the variation is due to **variation between persons** and some within person.

Clustered data / data with several random components

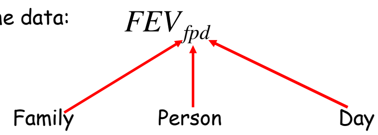
From **10** families:



Some of the variation between persons is due to **variation between families and some within family.**

Clustered data / data with several random components

Structure of the data:



Three sources of random variation:

Variation between **families**

Variation between **persons** (variation within family)

Variation between **days** (variation within person)

Clustered data / data with several random components

Factors of interest:

household **I**ncome	Constant within **family**
Urbanization	Constant within **family**
Age	Constant within **person;** varies within family
Sex	Constant within **person;** varies within family
Grass pollen	Constant within **day;** varies within person

A model:

$$FEV = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G$$

+random variation

Clustered data / data with several random components

$$FEV = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G$$

+random variation

If the **three** levels/sources of **random** variation **are not** taken into account :

- The **precision** of β_I and β_U are **highly overestimated**
- The **precision** of β_A and β_S are **overestimated**
- The **estimates** of β_I and β_U will be **biased** if the not all families are represented by the **same number of persons** and each person is measured the **same number of times**.
- The **estimates** of β_A and β_S will be **biased** if not all persons are measured the **same number of times**.

Clustered data / data with several random components

$$FEV = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G + F_f + P_{fp} + E_{fpd}$$

		variance
F_f	: Random family contribution	σ_F^2
P_{fp}	: Random person contribution	σ_P^2
E_{fpd}	: Random day contribution	σ_E^2

$$\text{var}\left(FEV_{fpd}\right) = \sigma_F^2 + \sigma_P^2 + \sigma_E^2$$

Variance components

Assumed to be normal distributed

Clustered data / data with several random components

Systematic part

$$FEV = \boxed{\beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G} + \boxed{F_f + P_{fp} + E_{fpd}}$$

Random part

$\beta_0, \beta_I, \beta_U, \beta_A, \beta_S$ and β_G Quantify the **systematic** variation

σ_F^2, σ_P^2 and σ_E^2 Quantify the **random** variation

This is a:

- **Variance component** model
- **Mixed** model (both systematic and random variation)
- **Multilevel** model

The theory behind and the understanding of such models is well **established!!!**