

MATH 60604A

Statistical modelling

§ 5c - Model formulation

Léo Belzile

HEC Montréal
Department of Decision Sciences

Linear regression for the revenge data

- Let's start by fitting an ordinary regression model, which will serve as a basis for the next analyses.
- This model ignores the possible within-person correlation, and proceeds as if these observations are independent.
 - The desire for revenge for a person at a certain time is likely correlated with the desire for revenge at other times, simply because these measurements came from the same person.
 - If this is true, the assumption that the error terms are independent is not valid; therefore, any inference made through this model is not valid.
- The linear model is

$$\text{revenge} = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{age} + \beta_3 \text{vc} + \beta_4 \text{wom} + \beta_5 \text{t} + \varepsilon,$$

where the error terms ε are assumed independent.

Modelling the time effect

- There are two natural ways of modeling the time variable:
 - We could assume a linear effect between `t` and `revenge` (continuous variable).
 - We could instead include `t` as a categorical variable.
- We will use `proc mixed` in order to familiarize you with this procedure.

SAS code to fit a linear model

```
proc mixed data=statmod.revenge method=reml;  
model revenge = sex age vc wom t / solution;  
run;
```

proc mixed output for linear regression

Data Set	INFE.REVENGE
Dependent Variable	revenge
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

Dimensions	
Covariance Parameters	1
Columns in X	6
Columns in Z	0
Subjects	1
Max Obs per Subject	400

Covariance Parameter Estimates	
Cov Parm	Estimate
Residual	0.3791

Fit Statistics	
-2 Res Log Likelihood	776.7
AIC (Smaller is Better)	778.7
AICC (Smaller is Better)	778.7
BIC (Smaller is Better)	782.6

The output of `proc mixed` is more complicated than that of `proc glm`.

Mean parameter estimates

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	-0.1689	0.2249	394	-0.75	0.4532
sex	0.1357	0.06748	394	2.01	0.0450
age	0.04586	0.004507	394	10.18	<.0001
vc	0.5225	0.01951	394	26.78	<.0001
wom	0.3989	0.02474	394	16.12	<.0001
t	-0.5675	0.02177	394	-26.07	<.0001

We see that all the variables are significant, though just barely for *sex*.

Interpretation of parameters for linear regression

- The more the person had initial behaviour of type `vc` or `wom`, the higher the desire for revenge.
- The effect of time is particularly interesting here. We see that the effect is negative. In each questionnaire, the value of revenge decreases by 0.568, on average, when all other variables remain constant. This is exactly what we saw in our earlier plots.
- **But can we be confident in our hypothesis tests?** The answer is no. Any kind of inference (tests and confidence intervals) will not be valid when we ignore the within-person correlation.

Notation

- Suppose that we collect observations from m groups such that:
 1. There are n_i observations within group i ($i = 1, \dots, m$).
 2. Any two observations from the same group are possibly correlated.
 3. Any two observations from different groups are assumed independent.
- Groups can be formed in several ways:
 - Several measures can be taken from the same subject (repeated measures) and each individual forms a group.
 - A group could also consist of individuals from the same school, department, or family.
- As before, we assume that we have a response variable and a collection of p explanatory variables.
- To simplify the notation, we'll call \mathbf{X}_i the set of all explanatory variables for all observations in group i .

Notation

- We use the index i to indicate the group, and j to indicate an observation within a group.
 - If the group is a business, then i represents the business, and j represents the subject.
 - For longitudinal data, i represents the subject and j represents an observation for that subject at a specific **time**.
- We call $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$ the **set of observations** of the outcome variable for group i .
- For the explanatory variables, we now need three indices, namely
 - i for the group,
 - j for the observation number within the group
 - k for the explanatory variable.
- We call $\mathbf{X}_{ij} = (1, X_{ij1}, \dots, X_{ijp})$ the set of p explanatory variables for observation j in group i .

Linear model with correlated errors

The linear regression model is

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} + \varepsilon_{ij}$$

for $i = 1, \dots, m$ and $j = 1, \dots, n_i$, where ε_{ij} is the error term for observation j in group i .

- As before, we assume that $E(\varepsilon_{ij} \mid \mathbf{X}_{ij}) = 0$ and therefore

$$E(Y_{ij} \mid \mathbf{X}_i) = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp}.$$

Covariance/correlation structure

- When we assume that the \mathbf{X} terms are fixed, correlation between error terms ε is equivalent to correlation among the responses \mathbf{Y} .
- We will allow dependence between observations within the same group.
- We assume the groups are independent from one another, so $\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$ if $i \neq i'$.
- We model the **within-group** correlation by assuming that the covariance matrix of \mathbf{Y} for group i is

$$\text{Cov}(\mathbf{Y}_i \mid \mathbf{X}_i) = \boldsymbol{\Sigma}_i,$$

or equivalently

$$\text{Cov}(\varepsilon_i \mid \mathbf{X}_i) = \boldsymbol{\Sigma}_i,$$

where $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})$ is the vector of errors for group i .

Block covariance structure for longitudinal data

- Assume for simplicity that data are ordered by group.
- We assume that observations for group i are correlated, but the observations for different groups are independent.
- The full covariance matrix of the **measurements** is therefore **block-diagonal**, i.e.,

$$\text{Cov}(\mathbf{Y}) = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Sigma}_m \end{pmatrix}.$$

- In our revenge example, we have $n = 80 \times 5 = 400$ observations.
- The **within-group covariance** matrix, $\mathbf{\Sigma}_i$, is 5×5 because we have a balanced sample ($n_1 = \cdots = n_m = 5$). The block $\mathbf{\Sigma}_i$ is thus identical for each group.
- The **between-group** covariance is **zero** ($\mathbf{0}$) because we assumed data for different individuals are independent from one another.

Covariance/correlation structure

- Generally, the covariance structure will depend on several parameters that will be estimated at the same time as the β parameters.
- The covariance structure is specified by the analyst. Sometimes, several covariance structures can be fitted to see which is most appropriate for the data at hand.