

Medizinische Hochschule Hannover
Institut für Biometrie

Anwendung des GEE-Modells in
Verlaufsanalysen

Eine kurze Einführung

H. Hecker

Contents

1	Allgemeiner Ansatz der GEE-Modelle (General Estimating Equation)	1
2	Beispiele: Verlaufsanalysen	2
2.1	1 Gruppe, parametrische Verlaufskurve	2
2.2	Mehrere Gruppen, additive Gruppeneffekte	6
2.3	Mehrere Gruppen, gruppenspezifische Verlaufparameter	7
2.4	Verläufe mit zeitabhängigen Kovariablen	7
2.5	Messwiederholungen ohne Struktur	8
3	SAS-Beispiele	8
3.1	Beispiel 1	8
3.2	Beispiel 2	10

Anwendung des GEE-Modells in Verlaufsanalysen

H.Hecker

1. Allgemeiner Ansatz der GEE-Modelle (General Estimating Equation)

GEE-Modelle stellen Verallgemeinerungen des *linearen Modells* und des "einfachen" *gemischten Modells* dar. Sie finden Anwendung in Situationen, in denen Beobachtungen einer Variablen Y innerhalb verschiedenener unabhängiger "Cluster" mehrfach auftreten und dabei mit Abhängigkeiten untereinander versehen sind. Wichtiges Beispiel hierfür ist die Analyse des zeitlichen Verlaufs einer Variablen durch Messwiederholungen an jedem Patienten.

Allgemeiner Ansatz

1. Es gibt n unabhängige Beobachtungseinheiten (Cluster, Subjects, Patienten)
2. Zu jeder Einheit i liegen Beobachtungen $y_{i1}, y_{i2}, \dots, y_{iT_i}$ vor, wobei die Anzahl T_i von Einheit zu Einheit variieren kann.

Beispiele:

1. Zeitliche Verläufe
 2. PCB-Werte in unterschiedlichen Gewebeproben (Tumor/Fibrose/Fett/Haut) von mehreren Patienten
 3. Körpergewicht aller Nachkommen eines Wurfes im Tierversuch
3. Für jede Einheit i sind die Beobachtungen y_{ij} bis auf einen Zufallsfehler durch ein *lineares Modell* mit einem Satz von *festen* (für alle Einheiten identischen) Koeffizienten und einem Satz von *patientenabhängigen* ("zufälligen") Koeffizienten bestimmt:

$$\begin{aligned} Y_{ij} & \quad \text{(Messung } j \text{ in Einheit } i \text{)} \\ = & \beta_0 + \beta_1 X_{i1j} + \beta_2 X_{i2j} + \dots \quad \text{(feste Effekte)} \\ & + \gamma_{i0} + \gamma_{i1} Z_{i1j} + \gamma_{i2} Z_{i2j} + \dots \quad \text{(zufällige Effekte)} \\ & + E_{ij} \quad \text{(Zufallsfehler)} \end{aligned} \quad (1.1)$$

Die mit den festen Koeffizienten β_k (Effekten) verbundenen Faktoren X_{ik} werden die "festen Faktoren" genannt, die mit den patienten (Einheiten-)spezifischen, "zufälligen" Koeffizienten ("Effekten") γ_{ik} verbundenen Faktoren Z_{ik} die "zufälligen Faktoren" bzw. "zufälligen Effekte".

In Matrix-Schreibweise:

$$Y_i = X_i b + Z_i \gamma_i + E_i \quad (i = 1, \dots, n) \quad (1.2)$$

4. Auch innerhalb jeder Einheit sind die zufälligen Effekte γ_{ik} und die Zufallsfehler E_{ij} voneinander unabhängig. Ihre Verteilungen sind

$$E_i \sim N(0, R) \tag{1.3}$$

$$\beta_i \sim N(0, G) \tag{1.4}$$

also multivariate Normalverteilungen mit Erwartungsvektoren 0 und den Kovarianzmatrizen R und G .

Man beachte: Die Dimension von R hängt von der Anzahl der Messungen in Einheit i ab (eigentlich also: R_i), die Dimension von G ist dagegen fest, nämlich gleich der Anzahl zufälliger Effekte des Modells.

Über die *Struktur* der Kovarianz-Matrizen können zusätzliche Voraussetzungen gemacht werden.

Folgerung für die Kovarianzen Σ_i der Beobachtungen $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iT_i})'$ innerhalb der Beobachtungseinheiten i :

$$\Sigma = cov(Y_i) = Z_i G Z_i' + R_i \tag{1.5}$$

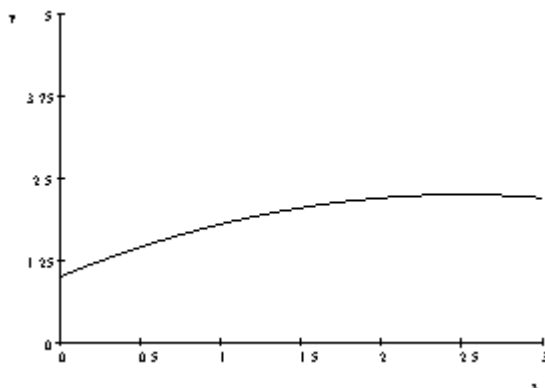
d.h. die Messungen sind korreliert, wobei die Abhängigkeiten sich aus den zufälligen Effekten und ggf. weiteren "bedingten" Abhängigkeiten je Beobachtungseinheit zusammensetzen. Das wird an späterem Beispiel erläutert

2. Beispiele: Verlaufsanalysen

2.1. 1 Gruppe, parametrische Verlaufskurve

Beispiel: Polynom 2.Grades

$$f(t) = b_0 + b_1 t + b_2 t^2 \tag{2.1}$$



Aber: Unterschiedliche Zeitpunkte und Anzahl Messungen je Patient:

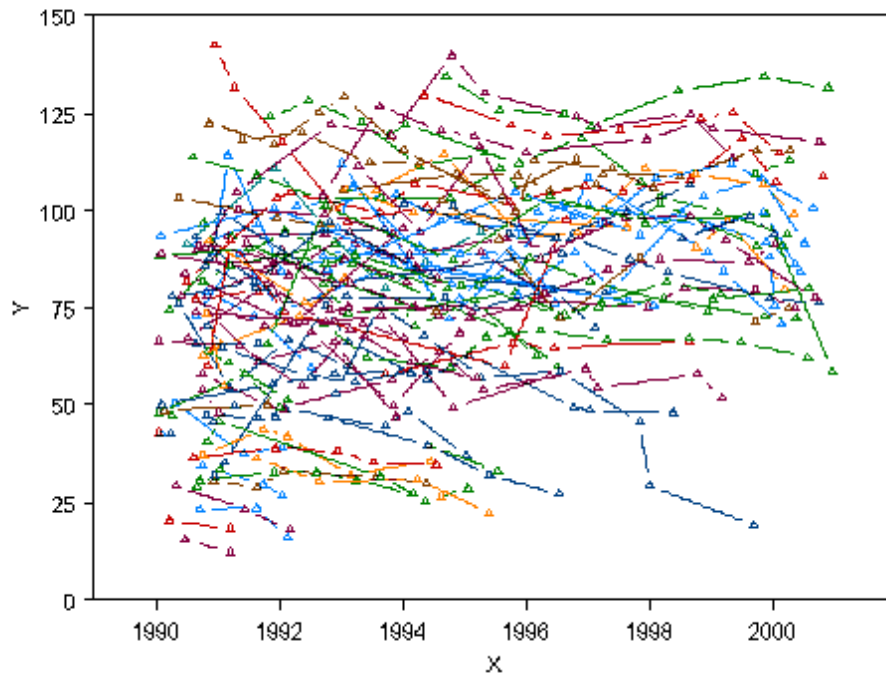


Abb. 1: Lungenfunktionswerte im Verlauf

Für die Auswertung gibt es im Rahmen von GEE mehrere Möglichkeiten für die Modellwahl:

1. Alle Koeffizienten fest (für alle Patienten identisch)

$$\begin{aligned}
 Y_{ij} & && \text{(Messung zum Zeitpunkt Nr. } j \text{ bei Patient Nr. } i \text{ (} X_{ij} \text{))} \\
 = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ij}^2 & && \text{(feste Effekte)} \\
 + E_{ij} & && \text{(Zufallsfehler)}
 \end{aligned}$$

1. Alle Zufallsfehler unabhängig, mit identischen Varianzen:

$$E_i \sim N(0, R) \quad \text{mit} \quad R = \sigma^2 \mathbf{I} \quad (2.2)$$

Dann ist man im üblichen linearen Modell. Dies ist in der Regel nicht geeignet, da dann keine Korrelationen zwischen Messwerten, auch nicht innerhalb der Patienten, zugelassen werden.

2. "Compound Symmetry": Alle Varianzen identisch; Kovarianzen möglich, aber alle identisch:

$$R = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_2^2 & \dots & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \dots & \sigma_2^2 \\ \vdots & \vdots & \dots & \vdots \\ \sigma_2^2 & \sigma_2^2 & \dots & \sigma_1^2 + \sigma_2^2 \end{pmatrix} \quad (2.3)$$

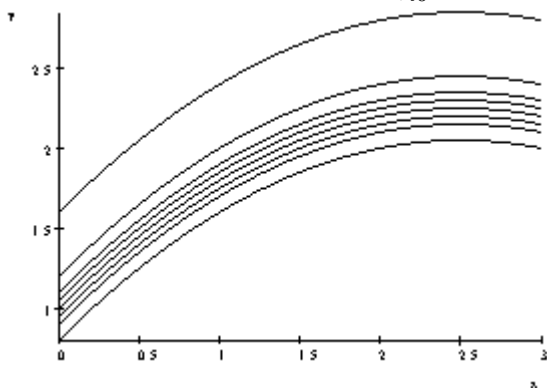
Aus den nächsten Beispielen wird deutlich, woher eine solche Modellannahme stammt:

2. Feste Koeffizienten für den parametrischen Kurvenverlauf (für alle Patienten identisch), aber zusätzlich:

Zufälliger "intercept" in der linearen Beziehung:

$$\begin{aligned} Y_{ij} & \quad \text{(Messung zum Zeitpunkt Nr. } j \text{ bei Patient Nr. } i \text{ (} X_{ij} \text{))} \\ = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ij}^2 & \quad \text{(feste Effekte)} \\ + \gamma_{i0} & \quad \text{(zufällige Effekte: Intercept)} \\ + E_{ij} & \quad \text{(Zufallsfehler)} \end{aligned} \quad (2.4)$$

Das heißt: Gegenüber dem für alle Patienten gemeinsamen (erwarteten) Kurvenverlauf darf jeder Patient (zusätzlich zu seinen spezifischen Messfehlern) noch seine eigene, "zufällige" Niveau-Verschiebung γ_{i0} darüberlegen (Beachte: $E(\gamma_{i0}) = 0$):



Weitere Annahme: Unabhängige, identisch verteilte Fehler:

$$E_i \sim N(0, R) \quad \text{mit} \quad R = \sigma^2 \mathbf{I} \quad (2.5)$$

Unter diesen Modellannahmen entstehen Korrelationen innerhalb der Messwerte eines Patienten dadurch, dass jeder Patient sein eigenes Niveau im Kurvenverlauf hat und die Werte dadurch größere Gemeinsamkeiten innerhalb als zwischen den Patienten erhalten:

$$V_i : = cov(Y_i) = Z_i G Z_i' + R_i \quad (2.6)$$

$$= var(\gamma_{i0}) + \sigma^2 \mathbf{I} \quad (2.7)$$

$$= \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_2^2 & \dots & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \dots & \sigma_2^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_2^2 & \sigma_2^2 & \dots & \sigma_1^2 + \sigma_2^2 \end{pmatrix} \quad \text{mit } \sigma_1^2 = \sigma^2 \text{ und } \sigma_2^2 = var(\gamma_{i0}) \quad (2.8)$$

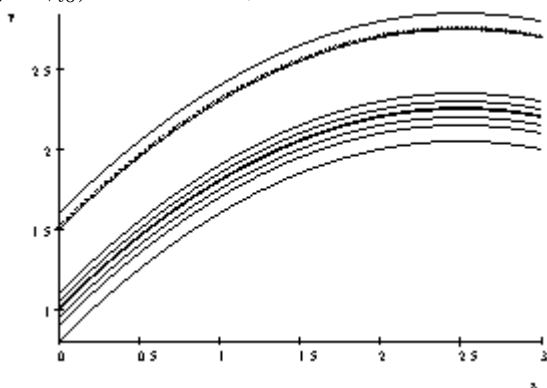
Folgerung:

Die CS (Compound Symmetry)-Annahme für die Fehler-Kovarianzmatrix ergibt sich aus der Voraussetzung unabhängiger Fehler E_{ij} und der Überlagerung der Messungen durch unabhängige Patienteneffekte.

Beide Modellspezifikationen sollten also zum selben Modell führen. Dennoch besteht ein Unterschied: Bei der Wahl 1.1 (nur feste Effekte und Kovarianzmatrix CS) werden auch negative Kovarianzen zugelassen (trotz der irreführenden Schreibweise " $\sigma_1^2 + \sigma_2^2$ ").

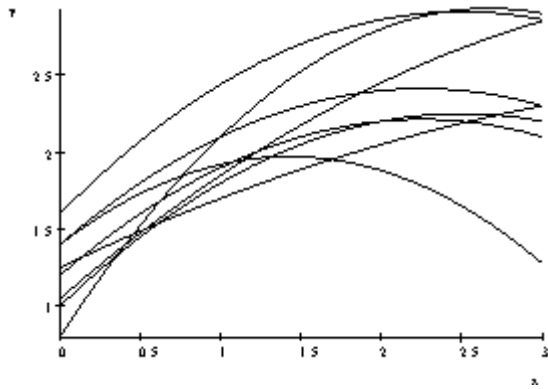
Frage zur Anwendung: Kann ein Patient zur Untersuchung des Kurvenverlaufs beitragen ,wenn er nur *einen* Messwert hat?

Antwort: Ja; der gemeinsame Kurvenverlauf (feste Effekte) wird durch die Gesamtheit aller Daten festgelegt; zum einzelnen Messwert wird das Niveau der Kurve (Intercept $\beta_0 + \gamma_{i0}$) so ermittelt, dass die Patientenkurve durch diesen Punkt geht:



3. Auch nach Abzug der Messfehler hat jeder Patient komplett seinen eigenen Kurvenverlauf: Die festen Koeffizienten (Erwartungswerte für die Gesamtpopulation) werden je Patient überlagert von zufälligen Abweichungen mit Erwartungswert 0

$$\begin{aligned}
 & Y_{ij} && \text{(Messung zum Zeitpunkt Nr. } j \text{ bei Patient Nr. } i \text{ (} X_{ij} \text{))} \\
 = & \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ij}^2 && \text{(feste Effekte)} \\
 & + \gamma_{i0} + \gamma_{i1} X_{ij} + \gamma_{i2} X_{ij}^2 && \text{(zufällige Effekte)} \\
 & + E_{ij} && \text{(Zufallsfehler)}
 \end{aligned} \quad (2.9)$$



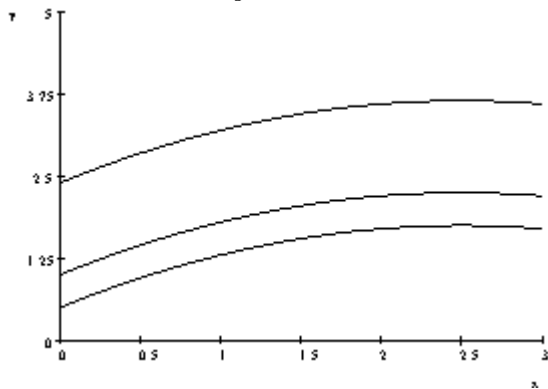
Zusätzliche Modellspezifikationen betreffen wieder die Kovarianzmatrix der Residuen (\mathbf{R}) und nun auch der zufälligen Effekte (\mathbf{G}). Die einfachste Version für \mathbf{G} ist die der Unabhängigkeit der Zufallseffekte γ_{ik} voneinander. Identische Varianzen sind aber nicht anzunehmen. In SAS wird dies durch die Spezifikation VC (Variance Components) realisiert. Ganz frei lässt man die Struktur durch die Spezifikation UN (UnStructured).

2.2. Mehrere Gruppen, additive Gruppeneffekte

Annahme: Parametrische Verlaufskurve wie oben, zusätzliche konstante (zeitunabhängige) Gruppeneffekte:

$$f(t) = b_0 + b_1 t + b_2 t^2 + g_g \quad (2.10)$$

mit Gruppeneffekt g_g . Dies führt zur Annahme *paralleler* Verläufe:



Allgemeiner Ansatz:

$$Y_{ij} \quad (\text{Messung zum Zeitpunkt Nr. } j \text{ bei Patient Nr. } i \text{ (} X_{ij} \text{)}) \\ = g_g \quad (\text{Effekt der Gruppe Nr. } g) \quad (2.11)$$

$$= \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ij}^2 \quad (\text{feste Effekte}) \quad (2.12)$$

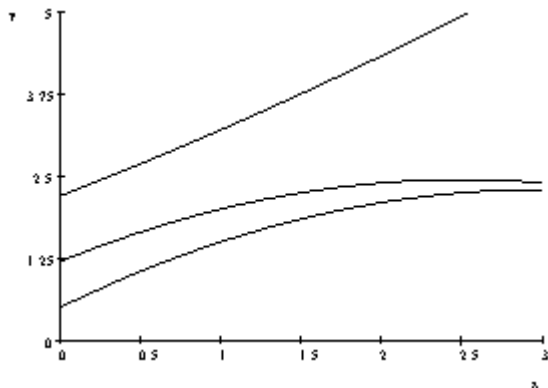
$$+ \gamma_{i0} + \gamma_{i1} X_{ij} + \gamma_{i2} X_{ij}^2 \quad (\text{zufällige Effekte}) \quad (2.13)$$

$$+ E_{ij} \quad (\text{Zufallsfehler})$$

Bei der Umsetzung in SAS wird dazu die Gruppenzugehörigkeit als fester Faktor in das Modell einbezogen. Die Spezifizierungen des Modells bzgl. Kovarianzmatrizen und zufälligen Effekten erfolgen wie oben.

2.3. Mehrere Gruppen, gruppenspezifische Verlaufparameter

Geht man davon aus, dass die Verlaufskurven zwischen den Gruppen nicht nur parallel verschoben sind, sondern auch darüber hinaus ihre gruppenspezifische Gestalt haben (s.Abb.)



lautet der allgemeine Ansatz:

$$Y_{ij} \quad (\text{Messung zum Zeitpunkt Nr. } j \text{ bei Patient Nr. } i \text{ (} X_{ij} \text{)})$$

$$= \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ij}^2 \quad (\text{generelle feste Effekte}) \quad (2.14)$$

$$+ g_g \quad (\text{genereller Effekt der Gruppe Nr. } g) \quad (2.15)$$

$$+ (g_g b)_1 X_{ij} + (g_g b)_2 X_{ij}^2 \quad (\text{gruppenspezifische Veränd.d.festen Effekte}) \quad (2.16)$$

$$+ \gamma_{i0} + \gamma_{i1} X_{ij} + \gamma_{i2} X_{ij}^2 \quad (\text{zufällige Effekte}) \quad (2.17)$$

$$+ E_{ij} \quad (\text{Zufallsfehler})$$

Die Umsetzung in SAS erfordert 2 Komponenten für die Gruppe: den Haupteffekt (Zeile 2.14) und die Wechselwirkung zum Zeitverlauf (Zeile 2.15).

2.4. Verläufe mit zeitabhängigen Kovariablen

Im Gegensatz zum Generellen Linearen Modell (GLM) können im GEE *zeitabhängige* Kovariablen untersucht werden. Hierzu gibt es viele Anwendungen, da in Verlaufsuntersuchungen häufig zu jedem Messzeitpunkt mehrere Variablen erhoben werden und deren Zusammenhang *im Zeitverlauf* interessiert. Dies kann dann realisiert werden, wenn man eine der interessierenden Variablen als Zielvariable Y deklariert und die andere(n) als Kovariablen. Wenn -wie bisher angenommen- weiterhin ein polynomialer Zeitverlauf unterlegt wird, lautet der Ansatz für Patient Nr. j :

$$Y_j \quad (\text{Messung zum Zeitpunkt Nr. } j \text{ (} X_{1j} \text{)})$$

$$= \beta_0 + \beta_1 X_{1j} + \beta_2 X_{1j}^2 \quad (\text{generelle feste Effekte: Zeitverlauf}) \quad (2.18)$$

$$+ c_2 X_{2j} + c_3 X_{3j} \quad (\text{fester Eff. d. Kovariabl. } X_2 \text{ und } X_3 \text{ zum Zeitp. } X_{1j}) \quad (2.19)$$

$$+ \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{1j}^2 \quad (\text{zufällige Effekte: Zeitverlauf}) \quad (2.20)$$

$$+ E_{ij} \quad (\text{Zufallsfehler})$$

Formal kann man dem Modell noch zufällige Effekte der Kovariablen hinzufügen ($\delta_2 X_{2j} + \delta_3 X_{3j}$). Das ist dann so zu interpretieren, dass der "Einfluss" einer Kovariablen auf die Zielvariable für jeden Patienten unterschiedlich intensiv sein kann.

2.5. Messwiederholungen ohne Struktur

In den bisherigen Beispielen waren die Messwiederholungen innerhalb eines Patienten durch die *Zeit* geordnet. Strukturierungen können auch durch einen qualitativen Faktor gegeben sein (z.B.: Messungen je Patient zu verschiedenen Proben aus den Geweben Tumor/ Fibrose/ Fett/ Haut). Dann können verschiedene Annahmen über die Kovarianzmatrix R plausibel sein, z.B. die, dass jede Faktorstufe ihre eigene Fehlervarianz generiert und die Fehler ansonsten unabhängig sind.

Sind die Messwiederholungen hingegen ohne Struktur (z.B. das Gewicht der Nachkommen eines Wurfes im Tierversuch), so erscheinen nur noch Annahmen über die Kovarianzmatrix R sinnvoll, die von der Reihenfolge der Beobachtungen unabhängig sind, also "Compound Symmetry" oder unabhängige identische Verteilung der Fehler: $R = \sigma^2 \mathbf{I}$.

Ansatz (mit einem Gruppeneffekt) für Beobachtungseinheit ("Wurf") i :

$$Y_j \quad (\text{Messung Nr. } j)$$

$$= \beta_0 \quad (\text{genereller Erwartungswert}) \quad (2.21)$$

$$+ g_g \quad (\text{genereller Effekt der Gruppe Nr. } g) \quad (2.22)$$

$$+ \gamma_0 \quad (\text{zufällige Effekte für Einheit (Wurf) } i) \quad (2.23)$$

$$+ E_j \quad (\text{Zufallsfehler})$$

3. SAS-Beispiele

3.1. Beispiel 1

```
/* -----*/
/* Beispiel 1:
   Verlaufsanalysen (Polynome 2. Grades)
   in e i n e r Gruppe
```

```

----- */

/* Nur feste Effekte; unabhängige Fehler
----- */

proc mixed    maxiter=200 maxfunc=200;
  class  patient ;
  model  y = alter0|alter0 /solution;
  repeated /subject = patient type = vc r ;
run;

/* Zum Vergleich: GLM */

proc glm;
  class  patient ;
  model  y = alter0|alter0 ;
run;

/* Nur feste Effekte; Compound Symmetrie */
/* -----*/

proc mixed    maxiter=200 maxfunc=200;
  class  patient ;
  model  y = alter0|alter0 /solution;
  repeated / subject = patient type =cs r ;
run;

/* Feste Effekte (Zeitverlauf)
   Zusätzlich: Zufälliger Patienteneffekt (intercept)
   unabhängige Fehler */
/* -----*/

proc mixed    maxiter=200 maxfunc=200 empirical covtest ic
              method = reml;
  class  patient ;
  model  y = alter0|alter0 /solution;
  random int /subject = patient type = un g gcorr S V ;
  repeated / subject = patient type = vc r ;
run;

/* Feste Effekte (Zeitverlauf)
   Zusätzlich: Zufällige Überlagerungen der ersten beiden

```

```

    Kurvenparameter (intercept , linear )
    und unabhängige Fehler */
/* -----*/

proc mixed    maxiter=200 maxfunc=200 empirical covtest ic
              method = reml;
  class patient ;
  model  y = alter0|alter0 /solution;
  random int alter0 /subject = patient type = un g gcorr S V ;
  repeated / subject = patient type = vc r ;
run;

/* Feste Effekte (Zeitverlauf)
   Zusätzlich: Zufällige Überlagerungen aller Kurvenparameter
   (intercept , linear, quadratisch )
   und unabhängige Fehler */
/* -----*/

proc mixed    maxiter=200 maxfunc=200 empirical covtest ic
              method = ml;
  class patient ;
  model  y = alter0|alter0 /solution;
  random int alter0|alter0 /subject = patient type = un g gcorr S V ;
  repeated / subject = patient type = vc r ;
run;

/* Ende Beispiel 1*/
/* -----*/

```

3.2. Beispiel 2

```

/* -----*/
/* Beispiel 2:
   Verlaufsanalysen (Polynome 2. Grades)
   in 2 oder mehr Gruppen (1 oder mehr Faktoren) */
/* -----*/

/* -----*/
/* Teil 1:
   Polynom (lin. quadrat.) unabhängig von Gruppe
   Gruppe hier: Geschlecht
   -----*/

```

```

/* Nur feste Effekte; unabhängige Fehler */
/* -----*/

```

```

proc mixed    maxiter=200 maxfunc=200;
  class patient sex;
  model y = alter0|alter0 sex /solution;
  repeated /subject = patient type = vc r ;
run;

```

```

/* Nur feste Effekte; Compound Symmetrie */
/* -----*/

```

```

proc mixed    maxiter=200 maxfunc=200;
  class patient ;
  model y = alter0|alter0 sex /solution;
  repeated / subject = patient type =cs r ;
run;

```

```

/* Feste Effekte (Zeitverlauf)
   Zusätzlich: Zufälliger Patienteneffekt (intercept)
   unabhängige Fehler */
/* -----*/

```

```

proc mixed    maxiter=200 maxfunc=200 empirical covtest ic
              method = reml;
  class patient sex ;
  model y = alter0|alter0 sex /solution;
  random int /subject = patient type = un g gcorr S V ;
  repeated / subject = patient type = vc r ;
run;

```

```

/* Feste Effekte (Zeitverlauf)
   Zusätzlich: Zufällige Überlagerungen der ersten beiden
   Kurvenparameter (intercept , linear )
   und unabhängige Fehler */
/* -----*/

```

```

proc mixed    maxiter=200 maxfunc=200 empirical covtest ic
              method = reml;
  class patient sex ;
  model y = alter0|alter0 sex /solution;
  random int alter0 /subject = patient type = un g gcorr S V ;
  repeated / subject = patient type = vc r ;

```

```

run;

/* Feste Effekte (Zeitverlauf)
   Zusätzlich: Zufällige Überlagerungen aller Kurvenparameter
   (intercept , linear, quadratisch )
   und z   unabhängige Fehler */
/* -----*/

proc mixed    maxiter=200 maxfunc=200 empirical covtest ic
             method = ml;
class patient sex ;
model y = alter0|alter0 sex/solution;
random int alter0|alter0 /subject = patient type = un g gcorr S V ;
repeated / subject = patient type = vc r ;
run;

/* -----*/
/* -----*/
/* Teil 2:
   Wechselwirkung zw. Steigung /quadr. Anteil
   und Gruppe: */
/* -----*/

/* Nur feste Effekte; unabhängige Fehler */
/* -----*/

proc mixed    maxiter=200 maxfunc=200;
class patient sex;
model y = alter0|alter0 sex sex*alter0|alter0 /solution;
repeated /subject = patient type = vc r ;
run;

/* Nur feste Effekte; Compound Symmetrie */
/* -----*/

proc mixed    maxiter=200 maxfunc=200;
class patient sex ;
model y = alter0|alter0 sex sex * alter0|alter0 /solution;
repeated / subject = patient type =cs r ;
run;

```

```

/* Feste Effekte (Zeitverlauf)
   Zusätzlich: Zufälliger Patienteneffekt (intercept)
   unabhängige Fehler */
/* -----*/

proc mixed    maxiter=200 maxfunc=200 empirical covtest ic
              method = reml;
  class patient sex ;
  model  y = alter0|alter0 sex  sex *alter0|alter0 /solution;
  random int /subject = patient type = un g gcorr S V ;
  repeated / subject = patient type = vc r ;
run;

/* Ende Beispiel 2*/
/* -----*/

```