

Aufgabe 1:

Der `cd4` Datensatz umfasst 2376 Beobachtungen der Zahl der CD4-Zellen im Blut von 369 HIV-infizierten/AIDS-kranken Männern vor und nach dem Zeitpunkt, zu dem zum ersten Mal HIV-Antikörper in ihrem Blut nachgewiesen wurden (Seroconversion). Die Zahl der CD4 -zellen dient dabei als Biomarker für den Zustand des Immunsystems. Das Hauptkenntnisinteresse liegt darin, die Form des Verlaufes des CD4-Gehalts über die Zeit zu bestimmen.

Lesen Sie die Daten ein und wandeln Sie `drug` und `ID` in Faktorvariablen um:

```
R>url <- "http://www.statistik.lmu.de/institut/lehrstuhl/semwiso/longitudinal_ss08/"
R>cd4 <- read.table(paste(url, "download/cd4_Diggle.txt", sep = ""),
+   header = T)
R>cd4$drug <- factor(cd4$drug, labels = c("yes", "no"))
R>cd4$ID <- factor(cd4$ID)
```

(a) Benutzen Sie den folgenden Code um die Verläufe der einzelnen Patienten zu plotten

```
R>library("lattice")
R>xyplot(CD4 ~ Time, data = cd4, group = ID, panel = function(x,
+   y, ...) {
+   panel.superpose(x, y, ..., col = "grey", type = "l", lwd = 0.1)
+   panel.xyplot(x, y, col = "black", type = "p", pch = 19, cex = 0.3)
+   panel.loess(x, y, lwd = 2.5, col = "blue")
+   panel.loess(x, y, lwd = 2.5, span = 1, degree = 2, col = "green")
+   panel.loess(x, y, lwd = 2.5, span = 0.1, evaluation = 100,
+     col = "red")
+ })
```

- (i) Was kontrollieren die `panel.loess`-Parameter `span` und `degree`? Was stellt also die grüne Linie dar? Überlegen sie sich wie die Schätzwerte zustande kommen. Wie ändert sich das Schätzverfahren wenn `span` nicht 1 ist?
- (ii) Untersuchen Sie wie sensibel die geglättete Kurve bezüglich Änderungen dieser beiden Parameter ist. Für sehr kleine Werte von `span` sollten sie `evaluation` entsprechend erhöhen - warum?

(b) Wir vergrößern die Zeitmessung um uns einen besseren Überblick über die Struktur im Datensatz zu verschaffen. Mit dem folgenden Code legen wir eine neue Variable `Time2` an, die `Time` in bestimmte vordefinierten Zeitintervalle kategorisiert:

```
R>plot(table(round(4 * cd4$Time)/4))
R>timepoints <- c(floor(min(cd4$Time)), -1.25, -0.75, -0.25, 0,
+   0.25, 0.75, 1.25, 1.75, 2.25, 2.75, 3.25, ceiling(max(cd4$Time)))
R>cd4$Time2 <- cut(cd4$Time, breaks = timepoints)
R>(table(cd4$Time2))
```

Benutzen Sie die R-Hilfe um den obigen Code nachzuvollziehen. Überlegen Sie warum die Intervallgrenzen anders als in der Vorlesung gewählt wurden.

- (c) Wir wollen die Korrelationen der CD4-Werte ähnlich wie auf Folie 2.23 darstellen. Dazu muss der Datensatz umstrukturiert werden:

```
R>m <- lm(CD4 ~ Time, data = cd4)
R>cd4$eps <- resid(m)
R>o <- order(cd4$Time)
R>cd4.kat <- reshape(cd4[o, ], v.names = c("CD4", "eps"), timevar = "Time2",
+   idvar = "ID", direction = "wide", drop = c("Age", "Time",
+   "cigpacks", "drug", "sexpartners", "cesd"))
R>colnames(cd4.kat)[seq(2, 24, by = 2)] <- levels(cd4$Time2)
R>colnames(cd4.kat)[seq(3, 25, by = 2)] <- paste("e.", levels(cd4$Time2),
+   sep = "")
```

- (i) Vergleichen Sie die Einträge für ID “10005” in den beiden Datensätzen. Was ist problematisch?
- (ii) Benutzen sie `spiom()` um `cd4.kat` grafisch darzustellen.
- (iii) Berechnen Sie die Korrelationen zwischen den Messungen in den verschiedenen Intervallen. Berechnen Sie die Korrelationen zwischen den Residuen einer Regression mit linearem Zeittrend in den verschiedenen Intervallen.
- (d) Stellen sie auf Basis der vergrößerten Zeitpunkte den Verlauf der Messungen und der Varianz der Messungen über die Zeit dar. Benutzen sie `tapply()` um die Varianzen in den Intervallen zu bestimmen. Warum ist es hier besser den Datensatz im “long”-Format zu benutzen?