Additive Hazard Regression Model

Teil 1

Georg Vogt

26. Juni 2006

Seminar über Statistik: SS 06 ETHZ

Seminar für Statistik ETH Zürich
Prof. Dr. Künsch
Assistent: Marco Cattaneo

1 Einleitung

Im Folgenden wird ein einfaches Beispiel beschrieben und anschliessend erweitert um damit Eigenschaften des Additiven Hazard Modells aufzuzeigen.

Sie möchten an einem Fest ein Würfel-Gewinnspiel durchführen und überlegen sich welche Konsequenzen dies für Sie haben könnte.

Der Spieler wirft einen Farbwürfel mit den 6 Seiten weiss, gelb, orange, rot, blau und schwarz alle vier Sekunden während einer Minute (15 Würfe).

Für einen Wurf Weiss oder Gelb zählen Sie zwei Punkte, für Orange oder Rot vier Punkte und für Blau bzw. Schwarz sechs Punkte.

Die für einen Gewinn benötigte Summe der gewürfelten Punkte legen Sie auf 24 fest. Wie viele der Spieler werden gewinnen? Um diese Frage zu beantworten modellieren Sie das Würfelspiel durch die zwei hintereinander geschalteten Zufallsprozesse, den Würfelprozess und den Zählprozess. Der Pfad des Würfelprozesses liefert den Input für den Zählprozess und die berechnete Verteilung für das Ereignis: der Spieler gewinnt, ergibt, dass jeder Teilnehmer auch gewinnen wird.

Das Spiel wird nun um zwei Komponenten erweitert und vor allem entschliessen Sie sich die Verläufe aufzuzeichnen und später auszuwerten. Zum Spass möchten Sie einerseits herausfinden ob Ihre berufstätigen Freunde anders würfeln als die studierenden Kollegen und andererseits vermuten Sie, dass ein höherer Gewinn einen stärkeren Anreiz schneller zu würfeln auslöst.

Beim erweiterten Spiel können die Teilnehmer eine Minute lang frei würfeln. Die Auswertung nehmen Sie mit den Additiven Hazard Regressions Modell vor:

$$h\left[t \mid Z_j(t)\right] = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t)\, Z_{j,k}(t)$$

Für Sie bedeutet $h\,[t \mid Z_j(t)]$ das Risiko, dass Sie zur Zeit t einen Gewinn auszahlen müssen. Mit $Z_{j,k}(t)$ ist die zusätzlich zu den aufgezeichneten Spielen vorhandene Information pro Teilnehmer bezeichnet.

$$Z_{j,1}(t) \quad \begin{cases} 1 & \text{falls erwerbstätig} \\ 0 & \text{falls nicht erwerbstätig} \end{cases}$$

$$Z_{j,2}(t) \quad \begin{cases} 1 & \text{Spiel mit tiefem Gewinn} \\ 0 & \text{Spiel mit hohem Gewinn} \end{cases}$$

Damit lassen sich die Spieler in verschiedene Gruppen einteilen und es kann untersucht werden ob ein Zusammenhang zwischen den Merkmalen einer Gruppe und dem Eintreffen eines Gewinnes existiert. Die Art wie die Information kodiert wird ist wichtig, weil damit auch bestimmt wird welche der Teilnehmer als Referenzgruppe dienen.

Bei der Referenzgruppe sind die Kovariablen Z_k gerade Null und diese Beobachtungen werden verwendet um die sogenannte baseline hazard rate β_0 zu schätzen. Der Beitrag der anderen Gruppen β_k wird als positive oder negative Ergänzung zur baseline hazard rate berechnet.

Die oben aufgeführte Definition der Kovariablen $Z_{j,k}$ führt zu den vier Gruppen:

Gruppe 0	Gruppe 1	Gruppe 2	Gruppe 3
nicht erwerbstätig Spiel mit hohem Gewinn	nicht erwerbstätig Spiel mit tiefem Gewinn	erwerbstätig Spiel mit hohem Gewinn	erwerbstätig Spiel mit tiefem Gewinn
Kodierung			
$Z_{j,1}(t)=0$ und $Z_{j,2}(t)=0$	$Z_{j,1}(t)=0$ und $Z_{j,2}(t)=1$	$Z_{j,1}(t)=1$ und $Z_{j,2}(t)=0$	$Z_{j,1}(t)=1$ und $Z_{j,2}(t)=1$

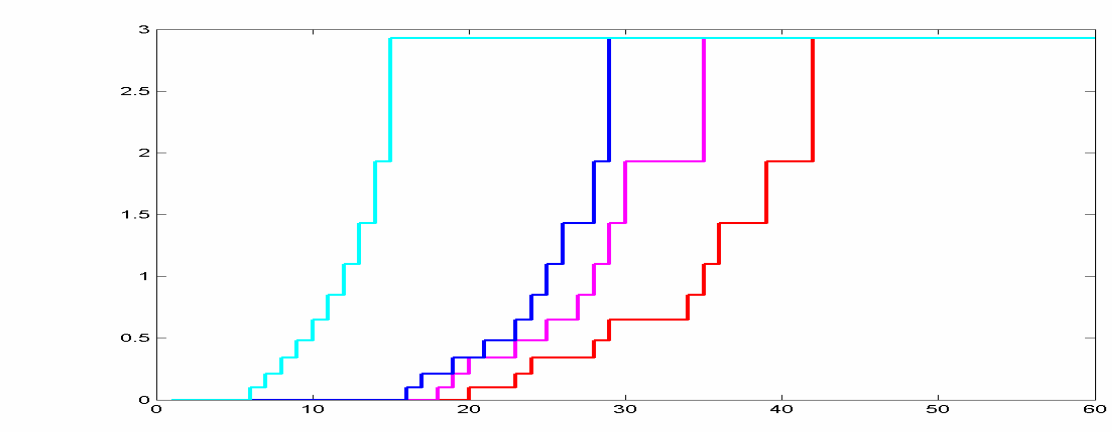
Diese Art der Kodierung sollte vermieden werden, weil sie bei der Auswertung zu Schwierigkeiten führt. Eine bessere Kodierung wird erreicht, indem man für p Gruppen (p-1) Kovariablen einführt:

Gruppe 0: falls $Z_{j,1}(t)=0$ und $Z_{j,2}(t)=0$ und $Z_{j,3}(t)=0$

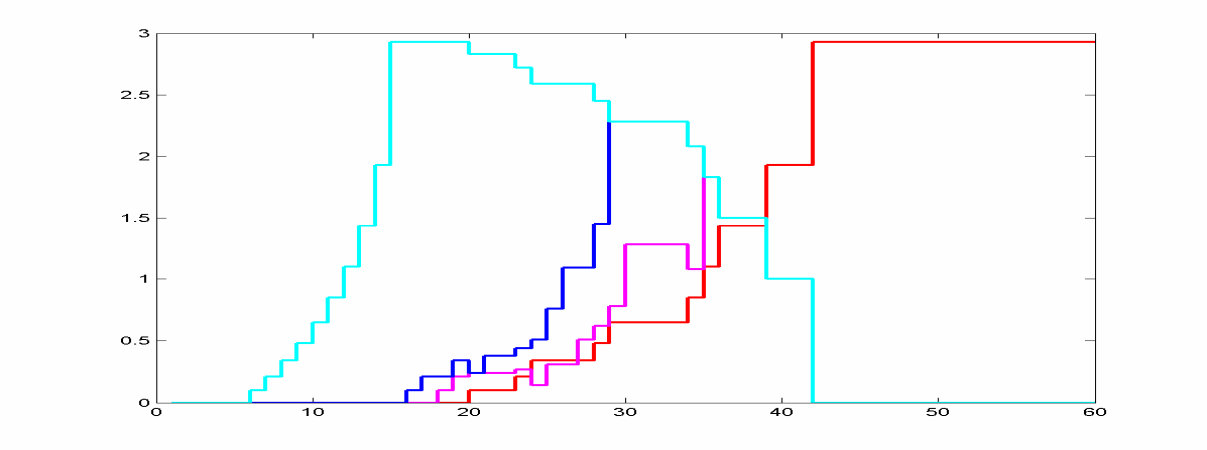
Gruppe 1: $Z_{j,1}(t)$ $\begin{cases} 1 & \text{falls nicht erwerbstätig und Spiel mit tiefem Gewinn} \\ 0 & \text{falls in Gruppe } 0,2,3 \end{cases}$

Gruppe 2: $Z_{j,2}(t)$ $\begin{cases} 1 & \text{falls erwerbstätig und Spiel mit hohem Gewinn} \\ 0 & \text{falls in Gruppe } 0,1,3 \end{cases}$

Gruppe 3: $Z_{j,3}(t)$ $\begin{cases} 1 & \text{falls nicht erwerbstätig und Spiel mit hohem Gewinn} \\ 0 & \text{falls in Gruppe } 0,1,2 \end{cases}$

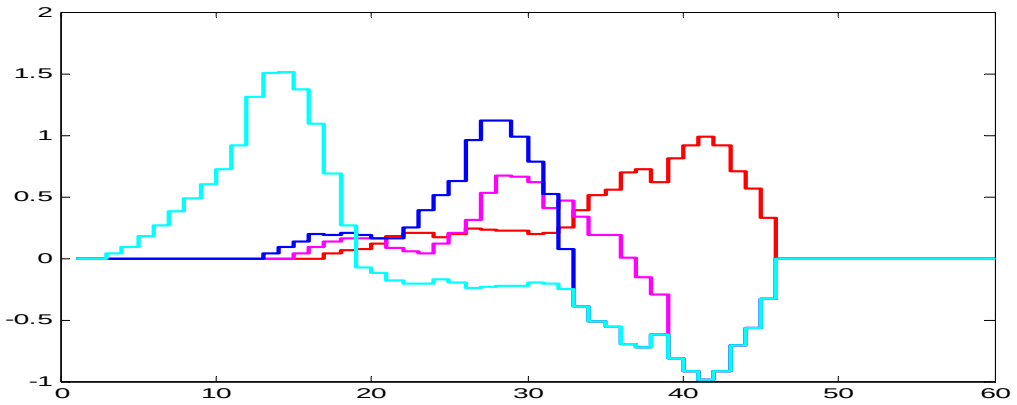


geschätzte kumulative Hazardfunktionen $\hat{B}_0$, $\hat{B}_1$, $\hat{B}_2$, $\hat{B}_3$ für die einzelnen Gruppen



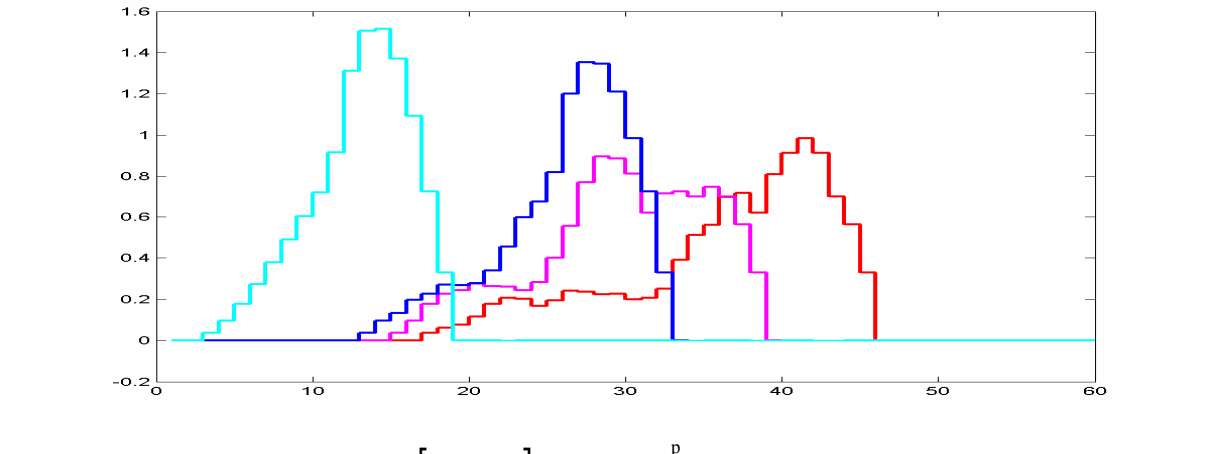
geschätzte kumulative Hazardfunktionen $\hat{B}_0$, $\hat{B}_1 - \hat{B}_0$, $\hat{B}_2 - \hat{B}_0$, $\hat{B}_3 - \hat{B}_0$

Als Gruppe 0 wurden die Teilnehmer definiert, die nicht zu einer anderen Gruppe gehören d.h. diejenigen, die erwerbstätig sind und bei einem Spiel mit tiefem Gewinn mitmachen. Es wird vermutet, dass diese Spieler am wenigsten motiviert sind und deshalb länger brauchen um die erforderliche Punktezahl zu würfeln.



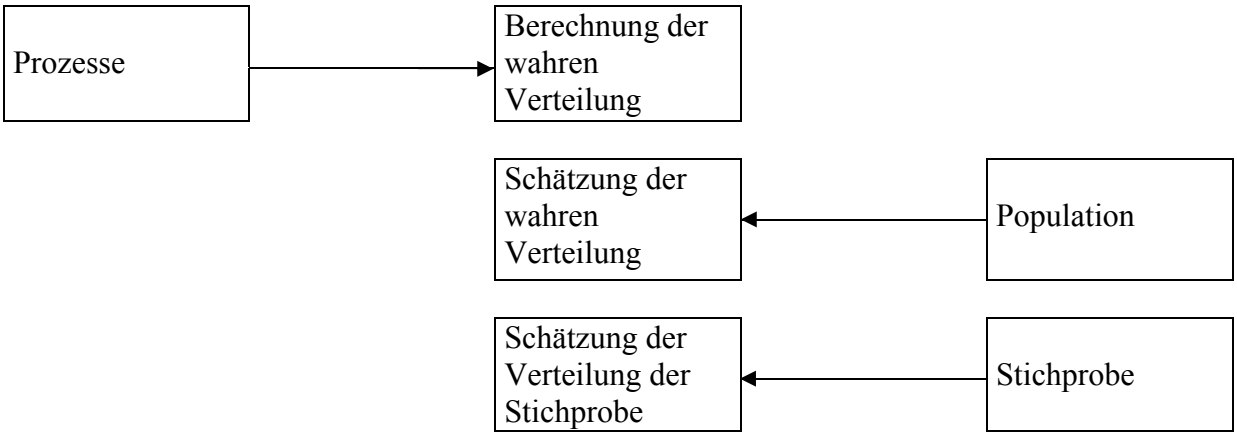
geschätzte Funktionen für $\hat{\beta}_0$, $(\hat{\beta}_1 - \hat{\beta}_0)$, $(\hat{\beta}_2 - \hat{\beta}_0)$, $(\hat{\beta}_3 - \hat{\beta}_0)$

Das Additive Hazard Regressions Modell ermöglicht die Auswertung von Beobachtungen aus nicht homogener Population und auch wenn die Umweltzustände nicht konstant sind.



Hazardfunktionen $\hat{h}\left[t \mid Z_j(t)\right] = \hat{\beta}_0(t) + \sum_{k=1}^{p} \hat{\beta}_k(t)\, Z_{j,k}(t)$ des additiven Modells

Während das einfache Spiel berechnet werden konnte wird das erweiterte Spiel geschätzt.



2 Grundlagen

2.1 Würfelprozess

Der betrachtete stochastische Prozess W ist eine Folge von Zufallsvariablen $W = \{W(t) : t \in \Gamma\}$ mit Indexmenge $\Gamma = \{0, 1, 2, \cdots\}$ für Prozesse in diskreter Zeit, die alle auf dem gleichen Wahrscheinlichkeitsraum (Ω, *F*, P) definiert sind.

Für einen derartigen Prozess W, werden die Zufallsfunktionen $W(t,\omega) : R^+ \to R$, $\omega \in \Omega$, Pfad oder Trajektorien des Prozesses genannt.

Eine stochastische Basis ist ein Wahrscheinlichkeitsraum (Ω, *F*, P) ausgestattet mit einer rechtsseitigen (d.h. $F_{t+} = F_t$) Filtration (Ω, *F*, $\{F_t : t \geq 0\}$,P). Die natürlichste Filtration ist der Verlauf eines stochastischen Prozesses.

Die Basismenge Ω enthält die Elementarereignisse ω_i, $i \in [1,\cdots,6]$: {ω_1 = weiss, ω_2 = gelb, ω_3 = orange, ω_4 = rot, ω_5 = blau, ω_6 = schwarz} also alle möglichen Resultate des Zufallsexperimentes.

Mit A_i $i \in [1,\cdots,3]$: A_1 := {ω_1, ω_2}, A_2 := {ω_3, ω_4}, A_3 := {ω_5, ω_6} werden die interessierenden Ereignisse bezeichnet. Die Menge *A* der Ereignisse A_i ist eine Teilmenge der Potenzmenge *P*(Ω) von Ω.

Durch die Menge A wird die σ – Algebra *F* definiert, für die gelten muss:

- $\Omega \in F$
- $A_i \in F \to A_i^c \in F$ mit $A_i^c := \Omega \setminus A_i$
- mit (A_n) (n = 1, 2, 3) einer Folge von Elementen aus *A* gehört auch die Vereinigung $\bigcup_{n=1}^{\infty} A_n$ zu *A*.

Man kann auch sagen die σ – Algebra F enthält Ω und ist unter Komplementierung und abzählbaren Vereinigungen abgeschlossen. Ist (Ω, F) ein messbarer Raum, was hier zutrifft, dann ist das Wahrscheinlichkeitsmass P auf F eine Abbildung, die jedem Ereignis $A_i \in F$ eine Zahl $P(A_i)$ derart zuordnet, dass gilt:

- $P(A_i) \geq 0; \quad \forall\, A_i \in F$
- $P(\Omega) = 1$
- $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$ d.h. für disjunkte Ereignisse ist das Wahrscheinlichkeitsmass additiv.

Die Wahrscheinlichkeit der interessierenden Ereignisse ist $P(A_i) = \frac{1}{3} \quad i \in [1, 2, 3]$
(Genau genommen sind es die Pfade des Würfelprozesses, die 24 Punkte erreicht haben und damit zum Ereignis des Zählprozesses führen, die man betrachtet.)

Als Borel-σ-Algebra B^1 der reellen Geraden R bezeichnet man die von der Menge der abgeschlossenen und beschränkten Intervalle $\{[a, b]: \quad a \leq b\}$ erzeugte σ-Algebra. Ihre Elemente heissen Borelmengen.

Der anfangs verwendete Begriff der Zufallsvariablen kann jetzt wie folgt definiert werden: Für den gegebenen Wahrscheinlichkeitsraum (Ω, F, P) bezeichnet man die messbare Funktion W von (Ω, F) in (R, B^1) als reelle Zufallsvariable. Messbar bedeutet hier, dass die Inverse Funktion $W^{-1}(B^1)$ auf (Ω, F) abbildet.

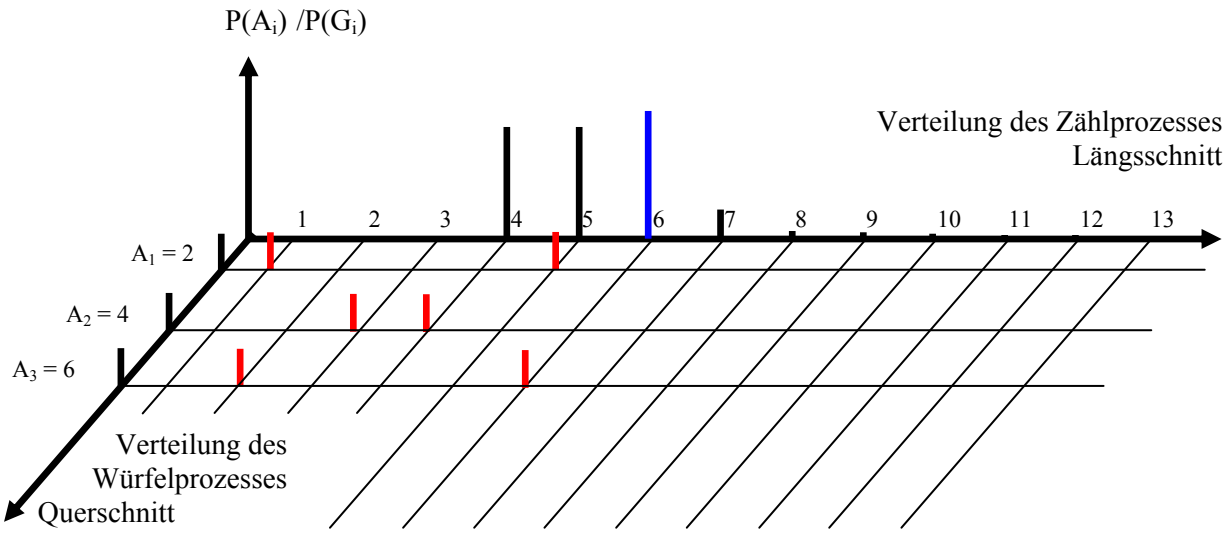
Als Verteilungsfunktion von W ergibt sich diejenige Funktion, die jeder reellen Zahl x die folgendermassen definierte Zahl F(x) zuordnet:

$$F(x) = P\{W \leq x\} = P_w(]-\infty, x])$$

Als Punktgewicht von W bezeichnet man die durch $\pi(x) = P\{X = x\}$ definierte Abbildung $\pi(\cdot): R \to R^+$. Das Punktgewicht $\pi(\cdot)$ ist vollständig durch die Verteilung (und damit auch durch die Verteilungsfunktion F) bestimmt. Die Zahl $\pi(x)$ ist die Sprunghöhe von F in x. Für die betrachtete Zufallsvariable W und allgemein für diskrete Zufallsvariablen gilt, dass die Summe $\sum_x \pi(x)$ gleich Eins ist.

2.2 Zählprozess

Der Zählprozess ist im wesentlichen eine Auswertung des zugrunde liegenden Würfelprozesses. Man könnte sagen, dass der Zählprozess die zeitliche Dimension erfasst.



Aus dem Seminarvortrag vom 2. Mai 06, Einführung in die Überlebensanalyse: Teil 1, seien einige Definitionen nochmals angegeben.

Die Überlebensfunktion ist definiert als: $S(t) = P[T_j > t]$

Ist die Ausfallszeit T diskret mit den Werten T_j, j = 1, 2, ... so dass $T_1 < T_2 < ...$, dann ist S(t) eine monoton fallende Treppenfunktion und es gilt:

$$S(t) = \sum_{Tj>t} P[T = T_j] =: \sum_{Tj>t} p(T_j)$$

Die Hazardfunktion ist definiert als: $$\lambda(t) := \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t \mid T \ge t]}{\Delta t} \ge 0$$

Für diskrete Ausfallzeiten T schreibt man mit $S(T_0) = 1$: $$\lambda(T_j) = P[T = T_j \mid T > T_j] = \frac{p(T_j)}{S(T_{j-1})}, \quad j = 1, 2, ...$$

Die kumulierte Hazardfunktion lautet im diskreten Fall:

$$\Lambda(t) = \sum_{Tj \le t} \lambda(T_i)$$

Der Kaplan-Meier Schätzer für die Überlebensfunktion S(t):

$$\hat{S}(t) = \prod_{s \le t} \left\{ 1 - \frac{\Delta \bar{N}(s)}{\bar{Y}(s)} \right\}$$

Der Nelson-Aalen Schätzer für die kumulierte Hazardfunktion Λ(t):

$$\hat{\Lambda}(t) = \sum_{s \le t} \frac{\Delta \bar{N}(s)}{\bar{Y}(s)}$$