

## Generalisierte Lineare Modelle – Übungen: Blatt 2

1. (Titanic Daten) Einer der wohl bekanntesten Datensätze mit kategoriellen Inhalten basiert auf Aufzeichnungen über die 2201 Passagiere an Bord der Titanic. Die Daten selbst stehen auf unserer WebPage zum download bereit und beinhalten Informationen über die Reiseklasse (**Class**) mit den vier Stufen **First**, **Second**, **Third** und **Crew**, das Alter (**Age**) mit den beiden Ausprägungen **Adult** und **Child**, das Geschlecht (**Gender**) mit **Female** oder **Male**, sowie über den binären Überlebensstatus (**Survived**) mit den Stufen **Yes** oder **No**. Sehr interessant sind auch die Ergebnisse einer *google search* auf *titanic data*.
  - (a) Finde ein passendes Modell für den Überlebensstatus und interpretiere, wie der Umstand zu überleben von den anderen Attributen abhängt.
  - (b) Berechne alle relevanten *odds*, sowie die für sie (wahrscheinlich doch für **Age** = **Adult** und die entsprechende Faktorstufe für **Gender**) als Passagier relevanten *odds ratios* und interpretiere diese kurz verbal.
  - (c) Bestimme das odds ratio für ihre favorisierte Reiseklasse, das einen Altersvergleich zulässt (wäre es besser, wenn sie doch noch ein Kind gewesen wären?).
  - (d) Welcher Passagiertyp hatte generell die höchste und welcher die geringste Chance unter dem gefundenen Modell zu überleben?
  
2. (Agresti, Problem 5.2) Während 23 Space Shuttle Flüge vor der Challenger Katastrophe 1986 (**Ft** Flugnummer) wurden Daten über die vorherrschende Temperatur (in °F) aufgezeichnet, sowie die Information erhoben, ob zumindest einer der primären O-Ringe einen Hitzeschaden hatte (**Thermal Distress TD**, 1 ja, 0 nein).

Ft	°F	TD	Ft	°F	TD	Ft	°F	TD	Ft	°F	TD	Ft	°F	TD
1	66	0	2	70	1	3	69	0	4	68	0	5	67	0
6	72	0	7	73	0	8	70	0	9	57	1	10	63	1
11	70	1	12	78	0	13	67	0	14	53	1	15	67	1
16	75	0	17	70	0	18	81	0	19	76	0	20	79	0
21	75	1	22	76	0	23	58	1						

- (a) Verwende ein logistisches Regressionsmodell um den Effekt der Temperatur auf die Wahrscheinlichkeit eines Schadens zu modellieren. Zeichne ein Bild des geschätzten Modells und interpretiere es.
- (b) Schätze die Wahrscheinlichkeit eines Schadens für eine Temperatur von 31°F, wie sie zum Zeitpunkt der Katastrophe in 1986 vorherrschte.
- (c) Konstruiere ein Konfidenzintervall für den Temperatureffekt auf die Chance (odds) eines Hitzeschadens ( $P(Y = 1|t)/P(Y = 0|t) = \exp(\beta_0) \exp(\beta_t)t$ ), d.h. für  $\exp(\beta_t)$ , und teste die statistische Signifikanz dieses Effekts.
- (d) Prüfe die Modellanpassung mittels eines Vergleichs mit einem komplexeren Modell, z.B. mit einem Modell in dem auch  $t^2$  inkludiert ist.

3. (Agresti, Problem 6.19) In einer Studie über nicht-metastasenbildenden Knochenkrebs wurde untersucht, unter welchen Umständen ein Subjekt drei Jahre krankheitsfrei war.

Lymphocytische Infiltration	Geschlecht	Osteoplastische Pathologie	krankheitsfrei	
			ja	nein
stark	weiblich	nein	3	0
		ja	2	0
	männlich	nein	4	0
		ja	1	0
schwach	weiblich	nein	5	0
		ja	3	2
	männlich	nein	5	4
		ja	6	11

- (a) Zeige, dass manch einzelner Prädiktor einen signifikanten Effekt hat, wenn dieser ohne die beiden andere Prädiktoren im Modell verwendet wird.
- (b) Versuche ein logistisches Regressionsmodell mit allen drei Haupteffekten anzupassen. Erkläre, warum der MLE für den Effekt der Infiltration nicht existiert. Um dies zu sehen, setze die Genauigkeit auf `epsilon = 1e-16` und erhöhe die Anzahl an Iterationen, z.B. `maxit = 100`.