Ludwig-Maximilians-Universität
Institut für Statistik
Wintersemester 2019/2020

Abschlussbericht zum Projekt LVS-IR-Taubenstein

Projektpartner: Sascha Filimon, Roman Ossner
Gruppenbetreuer: Dr. André Klima
Projektgruppe: Alexander Fogus, Lea Vanheyden, Zorana Spasojević

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Als beliebtes Ziel für Touristen und Wintersportler besteht im Alpengebiet eine besondere Konfliktsituation zwischen Mensch- und Tierreich. Routen für Spaziergänger, Skifahrer und Skitourengänger grenzen oft direkt an Lebensräume von Wildtieren an und führen so zu Stress für das Wildtierreich. Die vom Deutschen Alpenverein (DAV) in Kooperation mit dem Freistaat Bayern auf den Weg gebrachte Kampagne „Natürlich auf Tour“ soll für eine Sensibilisierung und Informationsgebung rund um das Thema Naturschutz dienen.

Neben der Aufklärung ist ein weiteres Ziel dabei, das Verhalten der menschlichen Besucher zu analysieren um daraus abzuleiten inwiefern man es womöglich steuern kann. In diesem Sinne untersuchte der DAV in Zusammenarbeit mit dem Departement für Geographie der LMU am Berg Taubenstein im Mangfallgebirge rund um den Spitzingsee in der Saison 18/19 und 19/20 den Anteil der Skitourengänger mit sogenannten LVS-Geräten. LVS-Gerät ist die Abkürzung für Lawinenverschüttetensuchgerät, mit Hilfe dieser Geräte können von Lawinen verschüttete Personen schnell gefunden werden. Personen, die ein LVS-Gerät dabeihaben, können mit diesem andere LVS-Geräte suchen und auch selbst gefunden werden.

Anhand der zur Verfügung gestellten Daten zur Saison 18/19 wird durch ein Modell der Anteil der Skitourengänger mit LVS-Gerät in Abhängigkeit von anderen Faktoren (wie z.B. Uhrzeit, Temperatur, Schneehöhe) analysiert .

Zudem wird untersucht, von welchen EinflussgröSSen die Messfehler abhängen, welcher Art (Über-/Unterschätzung) sie sind und ob eine Struktur (mögl. Verteilung) vorliegt.

Unter Berücksichtigung der Erkenntnisse über die Messfehler werden Hypothesen aufgestellt, in welcher Form die Messfehler die geschätzten Abhängigkeiten beeinflussen.

Noch schreiben, was wir rausgefunden haben.

Quelle:

`https://de.wikipedia.org/wiki/Lawinenverschüttetensuchgerät`

2 Datenbasis

2.1 Datengrundlage

Um auf den Anteil der Skitourengängen, die ein LVS-Gerät bei sich hatten, schlieSSen zu können, hat man Checkpoints aufgestellt. Für den Aufstieg am Taubenstein gibt es zwei Routen, eine Nord- und eine Südroute. Die genauen Lagen kann man Abbildung 2.1 entnehmen. An beiden Routen wurde jeweils ein Checkpoint aufgestellt, an dem vorbeigehende Besucher gemessen werden. Durch Infrarotmessung wird erfasst, ob ein Mensch am Checkpoint vorbeigeht. AuSSerdem werden LVS-Geräte, die auf Sendebetrieb geschaltet sind, erfasst. Für jede einzelne Checkpointmessung liegt das jeweilige Datum mit Uhrzeit vor und an welcher der zwei Routen (N oder S) sie erfasst wurde.

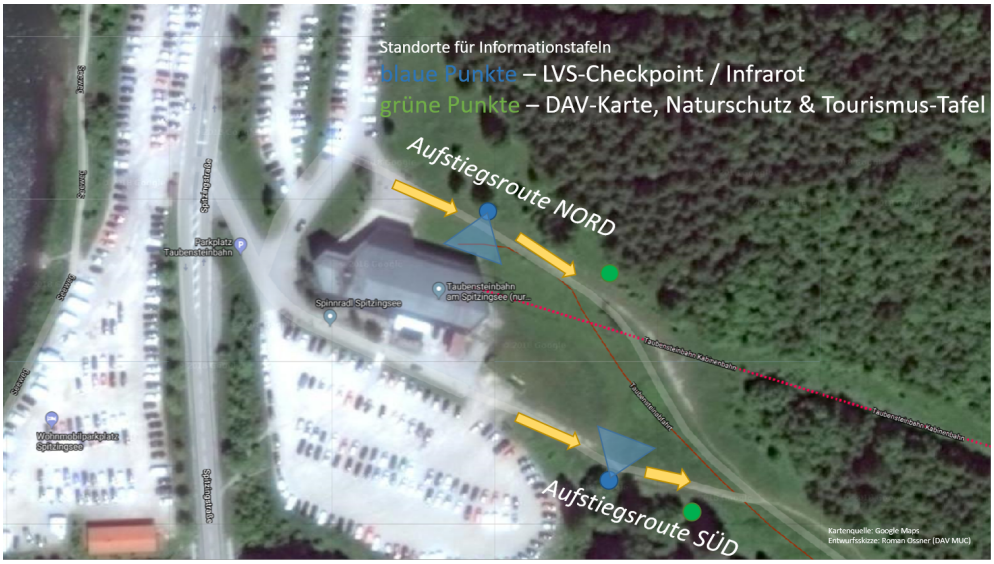

Abbildung 2.1: Satellitenbild, dass die Lage der zwei Routen und der Checkpoints verdeutlich

Für die erste Untersuchung benutzen wir vorerst nur diese automatisch erfassten Daten zur Saison 18/19. Der umfasste Zeitraum läuft dabei vom 21.12.2018 bis

zum 13.04.2019. Anzumerken ist dabei, dass am 23.12. und 24.12. keine Messungen vorliegen, zudem werden Messungen vom 07.01. bis zum 15.01. auSSer Acht gelassen, da aufgrund von starkem Schneefall die Checkpoints bedeckt waren.

Zusätzlich zu diesen automatischen Messungen wurden manuell Gruppen von (?)Studenten/Mitarbeitern des Departments für Geographie(?) an bestimmten Tagen vor Ort eingesetzt, um durch Befragungen manuelle Daten zu gewinnen. Dabei wurde festgestellt, dass bei den durch die Checkpoints erhobenen Daten Messfehler vorliegen.

Quelle:

Folien vom Erstgespräch

Neben der Erfassung von Personen und LVS-Geräten liegen verschiedene weitere Daten vor.Für jeden Tag an dem gemessen wurde gibt es Information zu den Wetterbedingungen bzw. anderen möglichen Einflussvariablen. „snowhight“ bemisst die Schneehöhe in cm. „temperature“ ist die Temperatur des Tages um 12:00 mittags. „solar_radiation“ zeigt die Sonneneinstrahlung in W/cm^2. AuSSerdem sind die Lawinenwarnstufen des jeweiligen Tages angegeben. Es kann vorkommen, dass die Lawinenwarnstufe auf der Spitze des Berges eine andere ist als im Tal, deshalb gibt es zwei Variablen: „avalanche_report_top“ und „avalanche_report_down“. Diese geben die Lawinenwarnstufe an der Spitze und am FuSS des Berges an. Obwohl es Stufen von 1 (niedrig) bis 5 (sehr hoch) gibt, war die in dem beobachteten Zeitraum höchste Stufe eine 4. An Tagen an denen die Stufen unterschiedlich waren wurde auSSerdem in „avalanche_report_border“ der Höhenmeter angegeben, ab dem sich die Lawinengefahr unterscheidet. In „avalanche_report_comment“ ist vermerkt, ob es sich dabei um eine Waldgrenze handelt. „day“ gibt an, um welchen Tag der Woche es sich handelt und „day_weekday“, „day_weekend“ und „holiday“ geben jeweils an, ob der Tag ein Tag unter der Woche oder am Wochenende war und ob er sich innerhalb einer Ferienzeit befunden hat.

2.2 Datenaufbereitung

Der erste Schritt besteht darin, die gegebenen Daten um weitere

Bearbeitung der Daten durch uns:

(noch mehr schreiben)

sunnrise und sunset

„count_beacon“ und „count_infrared“ enthalten die Anzahl der gemessenen

Tabelle 2.1: Übersicht aller benutzten Variablen mit kurzer Beschreibung

Name	**Beschreibung**	**Werte**
date	Datum	Datum vom 25.12.2018 bis zum 15.01.2019
time	Uhrzeit(minutengenau)	Uhrzeit von 03:59 bis 03:59
position	Route an der gemessen wurde	diskret (2 Ausprägungen): S,N
lvs	Hat die gemessene Person ein LVS-Gerät mitgeführt?	diskret (2 Auspr.): ja, nein
day	Wochentag	diskret (7 Auspr.): Montag, Dienstag,...
snowhight	Schneehöhe in cm (am Tag der Messung)	stetig: 16-212
temperature	Temperatur in řC um 12:00	stetig: (-7.9)-(9.7)
solar_radiation	Sonneneinstrahlung in W/cm^2	stetig: 14-792
avalanche_report	berechnete Lawinenwarnstufe	diskret (7 Auspr.): 1, 1.5, 2,...
holiday	Handelte es sich um einen Ferientag?	diskret (2 Auspr.): ja, nein
sunrise	Uhrzeit des Sonnenaufgangs	Uhrzeit von 05:29 bis 08:02
sunset	Uhrzeit des Sonnenuntergangs	Uhrzeit von 16:24 bis 18:58
day_length	berechnete Länge des Tages (von Sonnenauf- bis Sonnenuntergang)	von 08:24 bis 13:28 h

LVS-Geräte bzw. Infrarotmessungen pro Tag.

Umwandlung der Messungen zu Personendaten. Umstellung des Tages von 04:00 bis 04:00.

Eine Übersicht über alle Variablen die verwendet wurden ist in Tabelle 2.1 zu finden.

3 Deskriptive Analyse

Insgesamt 37593 Messungen an 114 Tagen

8468 Beacons, 29125 Infrareds (vor Umkodierung)

nach Umkodierung: 31574 Personen

8468 mit LVS-Gerät, 23106 ohne LVS-Gerät

pr

Die meisten Leute zwischen 09:00 und 18:00 unterwegs

–

Schneehöhe nimmt bis Mitte Januar stark zu und fällt ab Mitte Februar ab

Temperatur nimmt im Trend bus Mitte Januar ab und steigt danach

Sonnenstrahlung steigt bis März leicht und danach stark

—

Anteil schwankt in den ersten Wochen deutlich

generell viele AusreiSSer, aber kein groSSe Veränderung bei Schneehöhe, Temperatur und Sonneneinstrahlung

Anteile zur Mittagszeit geringer

mit steigender Lawinengefahr steigt die Anzahl

4 Methodik

In folgendem Kapitel soll in einem Modell der Zusammenhang einer beobachteten abhängigen Variable, in unserem Fall, der Anteil der Skitourengänger mit LVS-Gerät, durch mehrere unabhängige Variablen erklärt werden. Mithilfe der generalisierten additiven Regression lässt sich diese Fragestellung auf eine flexible und strukturierte Weise lösen. Die Modellierung wird mittels nichparametrischer Regression durchgeführt. Dabei wird im ersten Schritt eine einfache nichtparametrische Regression mit normalverteilter Zielvariable und einer einzigen Kovariable angenommen. Die Modellierung dient zu Veranschaulichung der Polynom- und Penalisierungs-Splines, sowie deren Basisfunktion. Im letzten Schritt wird die Modellierung der Splines mit dem generalisierten additiven Modell verknüpft und durch ein Modell mit mehreren Kovariablen erweitert. Die mathematischen Formeln und Konzepte aus diesem Kapitel basieren, wenn nicht anders vermerkt, aus dem Buch „Regression-Modelle, Methoden und Anwendungen“ von Ludwig Fahrmeir.

4.1 Nichtparametrische Regression

Bei einem linearen Regressionsmodell wird der Erwartungswert einer Zielvariable durch die Linearkombination von Einflussgrößen, auch Kovariablen genannt, beschrieben. In praktischen Anwendungen ist dies oft unzureichend, da neben den linearen Einflüssen auch nicht-lineare, flexible Einflüsse von stetigen Kovariablen auf die abhängige Variable wirken können. Die univariate Glättung stellt ein Verfahren dar, welche eine fexible Modellierung genau einer metrischen unabhängigen Variable auf eine metrische Zielvariable ermöglicht. Es wird davon ausgegangen, dass die Messungen (y_i, z_i) mit $i = 1, ..., n$ gegeben sind, wobei y_i die Beobachtungen der Zielvariable und z_i die Werte eine metrischen Kovariable abbilden. Die Zielvariable soll nun durch eine Funktion der Kovariable und einen additiven Störterm (ϵ_i) erklärt werden:

$$y_i = f(z_i) + \epsilon_i; \;\; E(\epsilon_i) = 0 \;\; und \;\; Var(\epsilon_i) = \sigma^2 \;\; mit \;\; i = 1, ..., n. \tag{4.1}$$

Daraus ergibt sich:

$$E(y_i) = f(z_i) \;\; Var(y_i) = \sigma^2 \;\; mit \;\; i = 1, ..., n, \tag{4.2}$$

was bedeutet, dass die Schätzfunktion f den Erwartungswert der abhängigen Variable darstellt.

4.1.1 Polynom-Splines

Eine Möglichkeit, die Funktion f abzubilden, sind Splines, die über Polynome modelliert werden und Polynom-Splines, oder auch Regressions-Splines, genannt werden. Das polynomiale Modell hat die Form:

$$f(z) = \gamma_0 + \gamma_1 z + ... + \gamma_l z^l, \tag{4.3}$$

wobei das Polynom vom Grad l durch den Einfluss von z auf y modelliert wird und γ_j die Regressionskoeffizienten darstellen. Um eine flexible Schätzfunktion gewährleisten zu können, wird der Definitionsbereich von z in Intervalle geteilt und Polynome werden für das jeweilige Intervall geschätzt. Die Zerlegung von z erfolgt stückweise auf Basis von sogenannten Knotenmengen k_1 bis k_m. Ein Problem, welches dabei entstehen könnte ist, dass an den Intervallgrenzen keine *glatte* Funktion entstehen. Der Graph einer *glatten* Funktion hat an den Intervallgrenzen keine „Ecken“. Deshalb muss für die „mathematische“ Glattheit die zusätzliche Bedingung, dass die Funktion $f(z)$ an den Intervallgrenzen $(l-1)$-mal stetig differenzierbar ist, eingeführt werden. Der Grad des Splines und die Anzahl der Knoten können somit im hohen Maße die Funktion beeinflussen, je höher die Anzahl an Knoten ist, desto *rauer* ist die Funktion $f(z)$. Um Polynom-Splines in der Praxis anwenden zu können, bedarf es einer Darstellung der Menge der Polynom-Splines. Eine mögliche Darstellung eines Polynom-Splines stellt die B-Spline-Basisfunktion dar.

4.1.2 Basis Splines

Der B-Spline dient als Basisfunktion für Polynom-Splines. Jedoch wird sich auch der Penalisierungs-Spline, welcher im darauffolgenen Kapitel vorgestellt wird, die

B-Spline-Basisfunktion zu Nutzen machen. Das Ziel bei der Konstruktion der B-Spline-Basisfunktion ist es die Polynomstücke des gewünschten Grades an den Knoten ausreichend *glatt* zusammenzusetzen. Die Basisfunktion besteht dabei aus $(l+1)$ Polynomstücken vom Grad l, welche stetig an den Knoten zusammengesetzt werden. Die Funktion $f(z)$ lässt sich durch die Linearkombination der Basisfunktionen:

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z) \tag{4.4}$$

darstellen. Die Koeffizienten γ_j werden mit Hilfe der KQ-Methode geschätzt und dienen zur Skalierung der Basisfunktionen. Die Anzahl der erforderlichen Basisfunktionen $d = m + l - 1$, setzt sich aus der Zahl der Knoten m und den Polynomstücken des gewünschten Grades l, welche $(l-1)$-mal stetig differenzierbar an den Knoten aneinandergefügt werden, zusammen.
Für B-Splines vom Grad 0 folgt:

$$B_j^0(z) = \mathbb{1}_{[k_j, k_{j+1}]}(z) = \begin{cases} 1 & k_j \leq z < k_{j+1}; \;\; mit \;\; j = 1, ..., d-1; \\ 0 & \text{sonst.} \end{cases} \tag{4.5}$$

Für B-Splines höheren Grades gilt:

$$B_j^l(z) = \frac{z - k_j}{k_{j+l} - k_j} B_j^{l-1}(z) + \frac{k_{j+l+1} - z}{k_{j+l+1} - k_j + 1} B_{j+1}^{l-1}(z). \tag{4.6}$$

Die Abbildung 4.1 visualisiert die Schätzung eines B-Splines anhand eines fiktiven Datenbeispiels schematisch. Abbildung (a) demonstriert eine B-Spline Basis vom dritten Grad. In Abbildung (b) wird die Basisfunktion mit dem Kleinste-Quadrate-Schätzer $\hat{\gamma}_j$ skaliert. Die Abbildung (c), zeigt die erhaltene Schätzung, wenn die skalierten Basisfunktionen addiert werden.

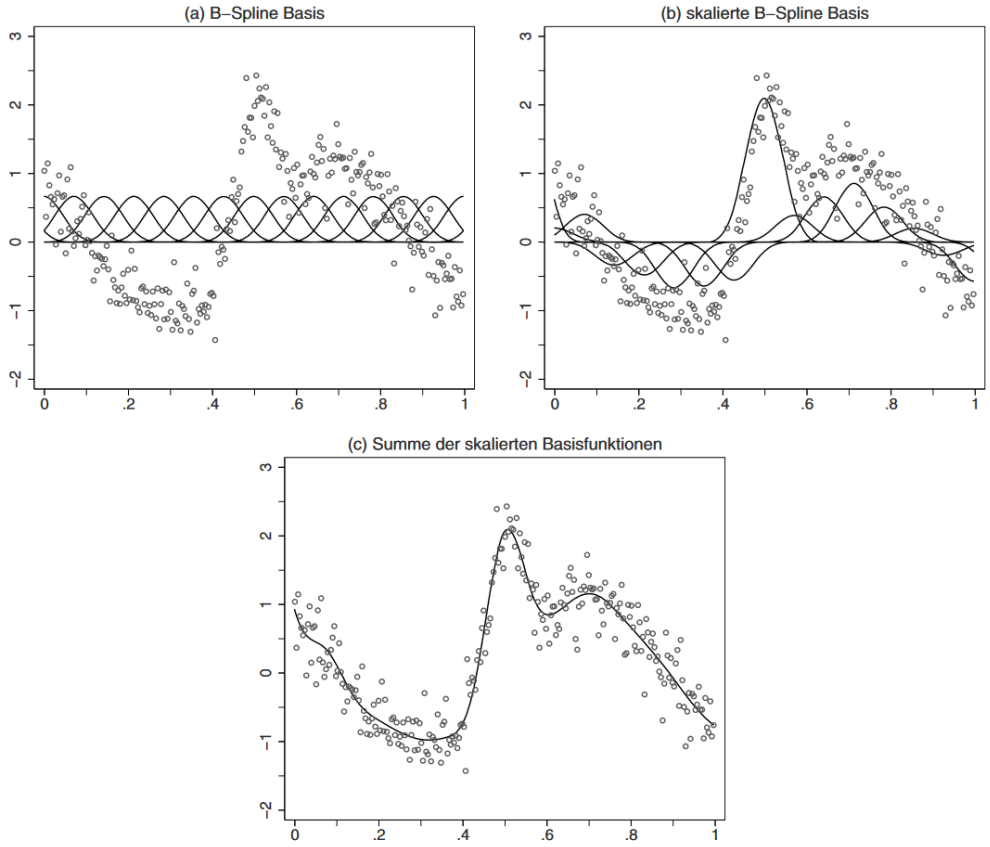

Abbildung 4.1: Schematische Darstellung der Schätzung eines nichtparametrischen Effekts mit B-Splines

Das nächste Kapitel behandelt P-Splines, welche auf B-Splines basieren und durch einen Strafterm erweitert worden sind. In unserer Arbeit beschränken wir uns auf P-Splines.

4.1.3 Penalisierte Splines basierend auf B-Splines

Im vorherigen Kapitel, haben wir die Wichtigkeit der Knotenmenge kennengelernt. Die Glattheit und Flexibilität der Schätzfunktion $f(z)$ hängt stark von der Anzahl der Knoten ab. Das Ziel ist es die Funktion $f(z)$ auf Basis einer B-Spline Funktion, mit einer groSSen Zahl an Knoten, zu approximieren um so die Flexibilität der Schätzfunktion sicherzustellen und andererseits zu hohe Variabilität in Form eines Strafterms zu sanktionieren. Der Penalisierte-Spline, gekürzt P-Spline, kombiniert somit eine B-Spline-Basis verbunden mit einem Strafterm. Der Strafterm für B-Splines wird häufig durch die quadrierte zweite Ableitung der Schätzfunktion $f(z)$ modelliert, da diese Form im R-Paket *mgcv* von Simon Woods implementiert wird.

Die zweite Ableitung stellt ein MaSS für die Krümmung der Funktion dar und ist daher in der Lage die Glattheit einer Funktion einzuschätzen. Der Strafterm:

$$\lambda \int (f''(z))^2 dz \tag{4.7}$$

beruht auf der quadrierten zweiten Ableitung der Schätz-Funktion, welche nach z integriert wird und durch den Glättungsparameter λ dargestellt wird. Die Glattheit der Schätzung des Modells im penalisierten Fall, wird primär durch den Glättungsparameter (λ) reguliert, d.h. je größer λ, desto glatter ist die Schätzung.

Zyklische P-Splines

Eine Glättungsfunktion heißt *zyklisch*, wenn die Funktion dieselben Werte und ersten Ableitungen an ihrer oberen und unteren Grenze aufweist. Beispielsweise kann man die Funktion für die Variable Woche mit Hilfe von zyklische P-Splines modellieren. So könnte man sicher stellen, dass die Werte am Ende einer Woche zusammenhängend zu den Werten am Anfang der Woche sind. Beispielsweise könnte der Fall eintreten, dass Werte von Sonntag den Wert von Montag beeinflussen und andersherum. Der Penalisierungsansatz für zyklische P-Splines ist simultan zu der Penalisierung von P-Splines basierend auf B-Splines (siehe 4.6)

(**Simon Woods Buch in Bib leihen und Zyklische Splines weiter ausführen**)

4.2 Additives Modell

In diesem Kapitel, werden wir vom einfachen Modell mit normalverteilter Zielvariable und genau einer Kovariable, auf ein additives Modell mit mehreren Kovariablen übergehen. Das additive Modell stellt in der Statistik ein nicht-parametrisches Regressionsmodell dar. Neben mehreren nicht-linearen flexiblen Einüssen können hier auch lineare Einflüsse von Kovariablen auf die abhängige Variable modelliert werden.
Das Standardmodell der additiven Regression (ohne Interaktionen) hat die Form:

$$y_i = \underbrace{f_1(z_{i1}) + ... + f_q(z_{iq})}_{\text{nicht-parametrische Effekte}} + \underbrace{\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}}_{\text{parametrische Effekte}} + \epsilon_i \tag{4.8}$$

Die Funktionen $f_1(z_1), ..., f_q(z_q)$ werden mit nichtparametrischen Effekten, die wir bereits im vorherigen Kapitel (4.1) erläutert haben, geschätzt. Der Störterm be-

sitzt die gleichen Annahmen, wie das einfache Modell mit nur einer Kovariable. Das Modell (4.8) kann durch Interaktionseffekte erweitert werden um Wechselwirkungen zwischen den Kovariablen zu integrieren. Seien z_1 und z_2 metrische Variablen, dann ergibt sich durch das additive Einbeziehen der Prädiktor:

$$\eta_i = f_1(z_{i1}) + f_2(z_{i2}) + f_{1,2}(z_{i1}, z_{i2}) + ..., \tag{4.9}$$

welcher durch die zweidimensionale Funktion $f_{1,2}(z_1, z_2)$ erweitert wurde. Die Schätzung der Funktion $f_{1,2}(z_1, z_2)$ erfolgt über eine nichtparametrische bivariate Methode.

4.2.1 Generalisierte additive Modell

Nichtlineare Effekte von metrischen Kovariablen treten auch bei Regressionsmodellen auf, die keine normalverteilte Zielvariable haben. Die abhängige Variable nimmt aber typischerweise eine Verteilung aus einer Exponentialfamilie (z.B. Binomial, Poisson, Multinomial, Normal,...) an. In unserem Fall, handelt es sich um eine binär verteilte Zielvariable, die sich folgendermaßen darstellen lässt:

$$y_i = \begin{cases} 1 & \text{,wenn Person mit LVS-Gerät identifiziert wird} \\ 0 & \text{,wenn Person ohne Lvs-Gerät identifiziert wird.} \end{cases} \tag{4.10}$$

Allgemein sieht das generalisierte additive Modell mit binär verteilter Zielvariable, $y_i|x_i \sim B(1, \pi_i)$, folgendermaßen aus. Der Erwartungswert μ der Zielvariable y wird mit dem additiven Prädiktor:

$$\eta_i = f_1(z_{i1}) + ... + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, \tag{4.11}$$

durch eine Responsefunktion h bzw. eine Linkfunktion $g = h^{-1}$ zu:

$$\mu = h(\eta) \; bzw. \; \eta = g(\mu) \tag{4.12}$$

verknüpft. Dabei ist der Prädiktor ein wichtiges Hilfsmittel bei der effizienten Codierung von additiven Modellen. Wie schon im vorherigen Kapitel (4.1) beschrieben, sind $f_1(z_{i1}), ..., f_q(z_{iq})$ *glatte* Funktionen, die grundsätzlich nicht durch eine starre parametrische Form festgelegt werden, sondern durch nicht-parametrische Methoden geschätzt werden. Die Parameter $\beta_0, ..., \beta_k$ werden mittels der Maximum-Likelihood-Methode:

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{4.13}$$

bestimmt.
Für den Erwartungswert der Zielvariable gilt:

$$E(y_i) = \pi_i = \frac{exp(\eta_i)}{1 + exp(\eta_i)}. \tag{4.14}$$

Dies lässt sich mit dem Logit-Link zu

$$\log(\pi_i) = \log(\frac{\pi_i}{1 - \pi_i}) \tag{4.15}$$

trasformieren.
Des weiteren muss die Varianzhomogenität, auch Homoskedastizität genannt, nicht unbedingt befriedigt sein. Das heißt, die Varianz ist in einigen Fällen nicht konstant und hat die Form:

$$Var(y_i) = \frac{\pi_i(1 - \pi_i)}{w_i}. \tag{4.16}$$

Der Gewichtungsfaktor w_i wurde eingeführt um den Fall von Individual- und gruppierten-Daten gleich zu behandeln. Für Individualdaten gilt $w = 1$, für gruppierte Daten nimmt w_i den jeweiligen Umfang der Gruppe an. Auf die Bedeutung von gruppierten Daten, wird im nächsten Kapitel näher eingegangen. Beim Störterm ϵ_i geht man davon aus, dass dieser unabhängig sein muss, jedoch ist ϵ_i im binären Fall nicht normalverteilt.
(Quelle von Störterm: https:/online.stat.psu.edustat504node220)

4.3 Gruppierte Daten

Im bisherigen Szenario sind wir davon ausgegangen, dass Individualdaten vorliegen. Für jedes Individuum i aus der Stichprobe n liegt eine Beobachtung (y_i, x_i) vor. Das heißt jeder binäre Wert der Zielvariable gehört zu genau einer Beobachtung $i = 1, ..., n$. Eine weitere Möglichkeit, welche Speicherplatz und Rechenzeit spart, wäre die Daten nach identischen Zeilen in der Datematrix zu gruppieren. Man erhält, dann eine Matrix mit unterschiedlichen Kovariablenvektoren x_i. In unserer Arbeit werden die Daten nach der Kovariable *Datum* gruppiert. Die Datenmatrix nimmt, dann folgende Gestalt an:

$$\begin{matrix} Gruppe\ 1 \\ \vdots \\ Gruppe\ i \\ \vdots \\ Gruppe\ G \end{matrix} \begin{bmatrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_G \end{bmatrix} \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_i \\ \vdots \\ \bar{y}_G \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \vdots \\ x_{G1} & \cdots & x_{Gk}. \end{bmatrix} \tag{4.17}$$

Dabei ist $\bar{y}_i$ die relative Häufigkeit, kann jedoch auch als absolute Häufigkeit durch $n_i\bar{y}_i$ dargestellt werden. Die Zahl an verschiedenen Kovariablenvektoren G ist im binären Modell deutlich kleiner als der Stichprobenumfang n.

4.4 Überdispersion

Die Varianz ist ein wichtiges Instrument für die Charakterisierung einer Verteilung. Sie gilt als Streuungsmaß und beschreibt die Konzentration der Verteilung um den Erwartungswert. Für gruppierte Daten lässt sich die Varianz innerhalb der Gruppe abschätzen. In der Praxis ist diese *empirische* Varianz oft bedeutend höher als durch ein Binäres Modell geschätzt wurde. Diese Unterschätzung der Varianz wird als Überdispersion bezeichnet. Gründe dafür können die unbeobachtete Heterogenität und postive Korrelation zwischen binären Beobachtungen, die zum gleichen *cluster* gehören, sein. Eine mögliche Lösung dafür, wäre die Einführung eines multiplikativen Überdispersionsparameters $\phi > 1$ in die Varianzformel, d.h.

$$Var(y_i|\mathbf{x}_i) = \phi\frac{\pi_i(1-\pi_i)}{n_i}. \tag{4.18}$$

(Quelle von Modell-Fit: https:/online.stat.psu.edustat504node220)
(Quelle von Überdispersion: Fahrmeir Regression)