1 Methodik

In diesem Kapitel wird als erstes die Theorie hinter den Modellen erklärt. Danach gehen wir auf die Schwierigkeiten ein, denen wir gegenüberstanden. Im letzten Teil stellen wir, dann die Ergebnisse vor.

1.1 Theorie

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. Für das bessere Verständis bauen wir den Theorie-Teil von einem einfachen Modell bis hin zu dem von uns verwendeten Modell auf.

1.1.1 Lineares Regressionsmodell

Bei einem linearen Regressionsmodell wird der Erwartungswert einer Zielvariable durch die Linearkombination von Einflussgrößen, auch Kovariablen genannt, beschrieben. Das lineare Regressionsmodell nimmt dabei folgende Form an:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i \tag{1.1}$$

Die Zielgröße wird durch y_i abgebildet, welche im Fall des linearen Regressionsmodells normalverteilt ist. Die Kovariablen werden durch $x_{i1}, ..., x_{ik}$ gekennzeichnet und ϵ_i stellt den Störterm dar. In Anwendungen bei denen es sich nicht um eine normalverteilte Zielvariable handelt ist das lineare Regressionsmodell unzureichend. Um das Problem zu lösen, wird im nächsten Kapitel auf das generalisierte lineare Regressionsmodell eingegangen. (Quelle: Fahrmeir S.62)

1.1.2 Generalisierte lineare Regressionsmodell

Das generalisierte lineare Regressionsmodell eignet sich für die Modellierung von Zielvariablen die nicht unbedingt normalverteilt sind. Die abhängige Variable nimmt

typischerweise eine Verteilung aus einer Exponentialfamilie (z.B. Binomial, Poisson, Multinomial, Normal,...) an. In unserem Fall handelt es sich um eine durch 0 und 1 kodierte Zielvariable y_i. Die beobachtete Zielgröße, der Anteil an Personen mit LVS-Gerät, ist Binomial-verteilt mit $y_i|x_i \sim B(1, \pi_i)$:

$$y_i = \begin{cases} 1 & \text{,wenn Person mit LVS-Gerät identifiziert wird} \\ 0 & \text{,wenn Person ohne LVS-Gerät identifiziert wird.} \end{cases} \tag{1.2}$$

Bei der generalisierten Regression mit binärer Zielvariable wird der Effekt der Kovariablen durch die (bedingte) Wahrscheinlichkeit π_i beschrieben.

$$\pi_i = P(y_i = 1|x_{i1}, ..., x_{ik}) = E(y_i|x_{i1}, ..., x_{ik}) \tag{1.3}$$

Für den Erwartungswert der Zielvariable gilt:

$$E(y_i) = \pi_i = \frac{exp(\eta_i)}{1 + exp(\eta_i)} = h(\eta_i). \tag{1.4}$$

Der Erwartungswert der Zielvariable wird mit dem linearen Prädiktor η_i mittels der Responsefunktion $\pi_i = h(\eta_i)$ verknüpft. Die Logit-Linkfunktion g ist die Umkehrung der Responsefunktion, daher gilt:

$$g(\pi_i) = \log(\frac{\pi_i}{1 - \pi_i}). \tag{1.5}$$

In unserer Arbeit machen wir Gebrauch von dem Logit-Link. Diese Transformation ist zur mathematischen Berechnung unseres Modells notwendig. In Kapitel (...) werden wir den Logit-Link zurücktransformieren um die Interpretation der Ergebnisse zu vereinfachen. Das generalisierte lineare Regressionsmodell beschreibt lineare Einflüsse der Kovariablen auf die Zielvariable. In praktischen Anwendungen ist dies oft unzureichend, da neben den linearen Einflüssen auch nicht-lineare, flexible Einflüsse von Kovariablen auf die abhängige Variable wirken können (Quelle: Fahrmeir S.190). Daher wird im nächsten Kapitel die additive Regression näher erläutert.

1.1.3 Additives Regressionsmodell

Das additive Modell stellt in der Statistik ein semi-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 hat die Form:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}}_{\text{parametrische Effekte}} + \underbrace{f_1(z_{i1}) + ... + f_q(z_{iq})}_{\text{semi-parametrische Effekte}} + \epsilon_i \tag{1.6}$$

Der parametrische Teil des Standardmodells hat die gleiche Form, wie schon im linearen Modell aufgezeigt. Die Funktionen $f_1(z_1), ..., f_q(z_q)$ werden mit nichtparametrischen Effekten geschätzt. (Quelle: Fahrmeir S.47) Auf die Theorie hinter der Funktion f wird im nächsten Kapitel eingegangen.

1.1.4 Splines

Die Schätzfunktion f wird durch sogenannte Splines modelliert. Splines stellen bestimmte Funktionen dar, die stückweise aus Polynomen eines bestimmten Grades zusammengesetzt werden. Es gibt verschiedene Arten von Splines. In unserer Arbeit beschränken wir uns auf die Penalisierten Splines, die Zyklische Splines und die Thin-Plate Splines.

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:

$$f(z) = \gamma_0 + \gamma_1 z + ... + \gamma_l z^l, \tag{1.7}$$

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.

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:

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z) \tag{1.8}$$

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:

$$B_j^0(z) = \mathbb{1}_{[k_j, k_{j+1}]}(z) = \begin{cases} 1 & k_j \leq z < k_{j+1}; \;\; mit \;\; j = 1, ..., d-1; \\ 0 & \text{sonst.} \end{cases} \tag{1.9}$$

Für B-Splines höheren Grades gilt:

$$B_j^l(z) = \frac{z - k_j}{k_{j+l} - k_j} B_j^{l-1}(z) + \frac{k_{j+l+1} - z}{k_{j+l+1} - k_j + 1} B_{j+1}^{l-1}(z). \tag{1.10}$$

Die Abbildung 1.1 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 1.1: 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.

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 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:

$$\lambda \int (f''(z))^2 dz. \tag{1.11}$$

In der Anwendung wird für die Approximation der zweiten Ableitungen, die zweite Differenz des Paramters γ_j verwendet. Wenn der Glättungsparameter λ gegen unendlich konvergiert, erhält man eine nahezu lineare Schätzfunktion $f(z)$ und der Strafterm wird primär durch λ reguliert. Im Fall von $\lambda \to 0$ verfügt der Glättungsparameter (λ) nur über ein geringers Gewicht.

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**)