1. Lösungen zu Kapitel 3

Übungsaufgabe 3.1

Systematische Antwortverweigerung kann erhebliche Probleme bei der Schätzung der Anhängerschaft von Fußballvereinen haben. Selektion in die Stichprobe wäre dann kein Problem, wenn man die Identifikationsannahme treffen würde, diese Selektion sei exogen, sprich die Informationen der fehlenden Antworten würden denen der vorliegenden Stichprobe entsprechen. In diesem Fall wäre die Stichprobe repräsentativ für die Grundgesamtheit.

Dies ist bei einem emotionsgeladenen Thema wie Fußball eine sehr fragwürdige Annahme. Die Richtung der Verzerrung ist nicht vorherzusagen. So kann es auf zum einen sein, dass es einem Anhänger eines erfolgsverwöhnten Vereins wie dem FC Bayern München leichter fällt, sich zu „outen “bzw. stolz zu seinem Verein zu stehen. Einem Anhänger des Regionalliga-Vereins Rot-Weiß Essen könnte es hingegen schwerer fallen, sich öffentlich zu seinem Verein zu bekennen (obwohl natürlich auch dieser Verein Erfolge aufzuweisen hat, u.a. DFB-Pokal und deutscher Meister in den 50er Jahren). Darüber hinaus kann das Ergebnis der Umfrage regional variieren. Lokalpatriotismus und Traditionsbewusstsein können regional verschieden stark ausgeprägt sein und zu einer leidenschaftlichen Zuneigungsbekundung führen. Ohne weitere Informationen ist es daher nicht möglich, Annahmen über den Selektionsprozess zu treffen.

Übungsaufgabe 3.2

Aus **Tabelle 3.1.** lassen sich drei bedingte Wahrscheinlichkeiten bzgl. des Schulabschluss „Hauptschule“ ($Y = 3$) ablesen. Es gibt drei Altersklassen mit den Jüngeren zwischen 20 und 30 Jahren ($X = 1$), den 30- bis unter 50-Jährigen ($X = 2$) in der mittleren Altersklasse und den Älteren mit 50 Jahren oder älter ($X = 3$). So ist die bedingte Wahrscheinlichkeit einen Hauptschluss zu haben, wenn man zu der unteren Altersklasse gehört 24,72% ($P(Y = 3|X = 1) = 0,2472$). Es ist allerdings zu beachten, dass dies nicht die bedingte Wahrscheinlichkeit der jüngeren Bevölkerung ist, einen Hauptschulabschluss zu besitzen. Es ist die bedingte Wahrscheinlichkeit der jüngeren Bevölkerung der *Erwerbstätigen*. Betrachtet man nun zusätzlich die Zufallsvariabel Z, mit $Z = 0$ für Nicht-Erwerbstätige und $Z = 1$ für Erwerbstätige, dann lassen sich die 24,72% als $P(Y = 3|X = 1, Z = 1)$ schreiben.

Übungsaufgabe 3.3

Nehmen wir nun wie in Beispiel 3.5 beschrieben an, dass die Ausfallhäufigkeiten über die Altersgruppen hinweg schwanken, innerhalb der Altersgruppen jedoch exogene Selektion vorliegt.

Die Zufallsvariable Z beschreibt nun den Beobachtungsstatus und nimmt den Wert 1 an, wenn eine Beobachtung vorliegt, und den Wert 0, wenn keine Beobachtung vorliegt. Die exogene Selektion lässt sich folgendermaßen beschreiben:

$$P(Y = 3|X = x, Z = 1) = P(Y = 3|X = x, Z = 0) = P(Y = 3|X = x).$$

Dies bedeutet, dass die in **Tabelle 3.1.** gegebenen bedingten Wahrscheinlichkeiten immer noch korrekt sind, die marginalen Wahrscheinlichkeiten jedoch nicht stimmen.

In den drei beschriebenen Altersgruppen ist der Anteil mit einem Hauptschulabschluss am geringsten unter den Jungen. Ist die Ausfallwahrscheinlichkeit bei den Jungen jedoch höher als bei den anderen Altersgruppen, so sind sie unterrepräsentiert. Dies bedeutet, dass der geringe Anteil der jüngeren Erwerbstätigen zu gering ins Gewicht fällt und insgesamt der Anteil der Erwerbstätigen mit Hauptschulabschluss überschätzt würde.

Übungsaufgabe 3.4

Die bedingte Prognose der Wachstumsrate des Bruttoinlandsprodukts Y gegeben den Ölpreis X $(P(Y = y|X = x, t = 1))$ hängt stark von Annahmen über die Ölpreisentwicklung ab. Eine Förderentscheidung der OPEC bestimmt diese Ölpreisentwicklung und lässt genauere Prognosen über die Ölpreisentwicklung zu. Diese zusätzliche Information sollte selbstverständlich für eine Korrektur der ursprünglichen Prognose verwendet werden. Revisionen der Prognose können auch dann getroffen werden, wenn die OPEC-Entscheidung im Verlauf des Prognosejahrs getroffen wird.

Häufig wird erwartet, dass die Wahrscheinlichkeitsverteilung $P(Y|X = x, t = 1)$ der möglichen Wachstumsraten auf eine Zahl, die *Punktprognose* $m(x)$, reduziert wird. Es gibt verschiedene mögliche Ausprägungen von $m(x)$, die theoretisch denkbar wären. Hierunter fallen z.B. die Ausprägung von Y, die die höchsten Wahrscheinlichkeit aufweist, der Mittelwert oder auch der Median. Diese Entscheidung lässt sich systematisch fällen, indem eine Verlustfunktion aufgestellt wird, die den möglichen Prognosefehler minimiert. Eine beliebte Verlustfunktion ist die gewichtete quadratische Abweichung der möglichen Ausprägung von Y und des Punktschätzers. Dieses Optimierungsproblem lässt sich über die erste Ableitung lösen und ergibt den Mittelwert als optimale Wahl für den Punktschätzer. Diese Wahl beruht allerdings auf der Wahl der funktionalen Form der Verlustfunktion. Eine quadratische Funktion bewertet große Abweichungen als besonders schwerwiegend. Alternativ könnte auch der absolute Prognosefehler minimiert werden, was wiederum zum Median als Punktschätzer führen würde.

Übungsaufgabe 3.5

Die Verlustfunktion, die den absoluten Fehler minimiert, sieht folgendermaßen aus:

$$Q = p_1|(1-m)| + \ldots + p_6|(6-m)|.$$

Der Punktschätzer m, der diese Funktion minimiert, ist der Median, der die Bedingung $P(x \leq m) = \frac{1}{2}$ erfüllt. Dies soll an einem Beispiel verdeutlicht werden.

Angenommen die Zustände 1 bis 6 haben die Verteilung $p_1 = p_2 = p_6 = 1/3$ und $p_3 = p_4 = p_5 = 0$. Der Mittelwert μ dieser Verteilung ist 3 und der Median 2. Diese Werte kann man in die Verlustfunktion einsetzen und erhält:

$$\begin{aligned} Q_\mu &= \frac{1}{3}|(1-3)| + \frac{1}{3}|(2-3)| + \frac{1}{3}|(6-3)| = 2 \\ Q_{Median} &= \frac{1}{3}|(1-2)| + \frac{1}{3}|(2-2)| + \frac{1}{3}|(6-2)| = 1\frac{2}{3} \end{aligned}$$

Der erwartete Fehler Q fällt bei der absoluten Verlustfunktion unter Verwendung des Medians kleiner aus als unter Verwendung des Mittelwerts.

Übungsaufgabe 3.6

a) Die gesuchte bedingte Wahrscheinlichkeit setzt sich folgendermaßen zusammen:

$$\begin{aligned} P(A = a|X = x) = {} & P(A = a|X = x, D = 1) \cdot P(D = 1|X = x) \\ & + P(A = a|X = x, D = 0) \cdot P(D = 0|X = x). \end{aligned}$$

Das Identifikationsproblem besteht in der Tatsache, dass beobachtet werden kann, ob ein Haushalt die Frage nach den Alkoholausgaben beantwortet ($P(D = 1)$), jedoch die Alkoholausgaben der zensierten Haushalte nicht beobachtet werden können ($P(A = a|X = x, D = 0)$). Aus diesem Grunde kann die bedinge Wahrscheinlichkeit von Interesse, die monatlichen Haushaltsausgaben für Alkohol gegeben bestimmter Charakteristika wie z.B. das Einkommen oder die Anzahl an Kindern ($P(A = a|X = x)$), nicht erfolgreich geschätzt werden.

b) Auf Grundlage der vorliegenden Stichprobe kann die gesuchte, aber nicht identifizierte bedingte Wahrscheinlichkeit zumindest durch obere und untere Schranken eingegrenzt werden. Zu diesem Zweck werden Annahmen über die zensierten Beobachtungen getroffen. Die untere Grenze ergibt sich, wenn man annimmt, dass die Ausgaben für Alkohol der Haushalte, die keine Angaben zu ihren monatlichen Ausgaben für Alkohol gemacht haben, mit einer Wahrscheinlichkeit von 0 in dem Intervall B liegen ($P(A \in B|X = x, D = 0) = 0$).
Die untere Schranke ergibt sich folglich als

$$\begin{gathered} P(A \in B|X = x, D = 1) \cdot P(D = 1|X = x) \\ +P(A \in B|X = x, D = 0) \cdot P(D = 0|X = x) \\ = P(A \in B|X = x, D = 1) \cdot P(D = 1|X = x) \\ \leq P(A \in B|X = x). \end{gathered}$$

Die obere Schranke ergibt sich durch die Annahme, dass für die zensierten Haushalte die Alkoholausgaben mit hundertprozentiger Wahrscheinlichkeit in dem Intervall B lieben.

$$\begin{gathered} P(A \in B|X = x) \leq \\ P(A \in B|X = x, D = 1) \cdot P(D = 1|X = x) \\ +P(A \in B|X = x, D = 0) \cdot P(D = 0|X = x) \\ = P(A \in B|X = x, D = 1) \cdot P(D = 1|X = x) \\ +P(D = 0|X = x) \end{gathered}$$

c) Es kann versucht werden durch beobachtbare Eigenschaften X als Kontrollvariablem mehr Informationen zu erhalten. Anstatt exogene Selektion anzunehmen, kann von systematischer Variation im Antwortverhalten in Abhängigkeit von X ausgegangen werden. Mit Hilfe von X können Teilausschnitte aus der Stichprobe ausgewählt werden, für die exogene Selektion angenommen wird. Das Selektionsproblem kann dann wiederum durch eine geeignete Umgewichtung der Stichprobeninformation reduziert werden.

c) Mit Hilfe einer Instrumentvariablen kann für unbeobachtete Heterogenität in der Stichprobe kontrolliert werden. Wie zuvor muss die Instrumentvariable X mit dem Antwortverhalten verbunden sein

$$P(D = 1|X = j) \neq P(D = 1|X = k) \text{ für } (j \neq k),$$

darf jedoch nicht mit den Alkoholausgaben der Haushalten korreliert sein

$$P(Y = y) = P(Y = y|X = j).$$

Die Instrumentvariable ermöglicht es, die obere und die untere Schranke der bedingten Wahrscheinlichkeiten weiter einzugrenzen. Aus der Schar der unteren Schranke kann die höchste und aus der Schar der oberen

Schranke kann die niedrigste Schranke gewählt werden. Im günstigsten Fall fallen die beiden Schranken auf einen Wert zusammen, was eine exakte Identifikation der bedingten Wahrscheinlichkeit zulässt.