Universität Tübingen
Geographisches Institut
Tübingen, im Mai 2002
Sommersemester 2002
Seminar „Verarbeitung Geographischer Daten“
Dozent: Dr. H.-J. Rosner
Referenten: Katrin Oehlkers, Helke Neuendorff, Tobias Schiller

# Analyse kategorialer Variablen

**Schriftliche Ausarbeitung des Referats vom 23. Mai 2002**

## Inhalt

## 1. Einleitung

Hinsichtlich ihres Skalierungsniveaus werden in der Statistik folgende Variablentypen unterschieden:

| | **Kategoriale Variablen (topologisch / qualitativ)** | | **Metrische Variablen (quantitativ)** | |
|---|---|---|---|---|
| | Nominal | Ordinal | Intervall | Rational |
| **Relationen** | = ≠ | = ≠<br>< > | = ≠<br>< ><br>+ - | = ≠<br>< ><br>+ -<br>x / |
| **Beispiele** | Familienstand<br>Geschlecht | Schulnoten<br>Rangfolgen | Temperatur<br>in °C | Einkommen<br>Größe |

Tab. 1: Variablentypen nach Skalenniveaus (nach ROSNER 2001, S. 10)

Die im Verlauf des bisherigen Kurses beschrieben statistischen Analysemethoden setzen meist ein metrisches Skalenniveau der Variablen voraus. Ein Beispiel dafür ist die Analyse des Zusammenhangs von Verdunstung (in Litern pro Zeiteinheit), Lufttemperatur (in Grad Celsius) und relativer Luftfeuchte (in Prozent) mittels Korrelation und Regression.

In der Geographie liegen Daten jedoch oft in Form kategorial skalierter Variablen (nominal oder ordinal) vor, etwa aus Befragungen. Beispiele sind etwa das Geschlecht der befragten Personen oder das verwendete Verkehrsmittel für eine bestimmte Strecke. Oftmals liegen auch ursprünglich metrisch skalierte Daten (zum Beispiel „Haushaltseinkommen“ in Euro) auf Grund von Vereinfachungen (in Form von Klassenbildung) oder aus Datenschutzgründen nur in einem kategorialen Skalenniveau vor („Haushaltseinkommen“ = bis 500 / 501-1500 / 1501-3000 / über 3001).

Bei den nominal skalierten Variablen kann nun eine weitere Unterscheidung getroffen werden:

- **dichotome (binäre) Variablen**, welche genau zwei Ausprägungen annehmen können („Geschlecht“ = weiblich/männlich, „Raucher“ = ja/nein), sowie
- **polytome Variablen** mit mehr als zwei Ausprägungen („Verkehrsmittel“ = Bus/Bahn/Auto/Fahrrad/Fußgänger)

Ein weiteres Merkmal kategorialer Variablen ist, dass sie nur endlich viele Werte annehmen können. Sie können zwar durch Zahlenwerte (weiblich=1; männlich=2) dargestellt werden; wie auch die Übersicht über die Relationen in Tab. 1 nahe legt, dür-

fen diese aber nicht unverändert für metrische Analysemethoden wie die multiple Regressionsanalyse herangezogen werden.

Bei statistischen Fragestellungen sind nun die in Tab. 2 dargestellten Kombinationen von Variablen verschiedener Niveaus möglich. Für die unterschiedlichen Kombinationen stehen jeweils spezifische Analysemodelle zur Verfügung.

| **Abhängige Variablen** | **Unabhängige Variablen** | | | |
|---|---|---|---|---|
| | alle metrisch | gemischt | kategorial | keine |
| metrisch | Multiple Regressions-analyse | Multiple Regressions-analyse | Varianz-analyse | Korrelations-/ Faktorenanalyse |
| **kategorial** | **Logit-Modell** | **Logit-Modell** | **Logit-Modell & Loglineares Modell** | **Loglineares Modell** |

Tab. 2: Typen statistischer Zusammenhänge (ROSNER 2001, S. 57)

Zur Analyse von Zusammenhängen, bei denen kategoriale Variablen beteiligt sind, kommen also im Wesentlichen zwei Methoden zum Einsatz: Das so genannte „lineare Logit-Modell“ und das „Loglineare Modell.“ Im Folgenden soll nun vor allem das Erstgenannte näher beschrieben werden, auf das wesentlich kompliziertere loglineare Modell soll nur kurz eingegangen werden.

## 2. Das lineare Logit-Modell für die bivariate Analyse mit einer dichotomen abhängigen Variablen

Das Logit-Modell soll an dem einfachsten Fall verdeutlicht werden, nämlich einer bivariaten Analyse mit einer metrisch skalierten unabhängigen und einer abhängigen Variablen, wobei die abhängige Variable dichotom sei, also nur zwei Ausprägungen haben soll.

Wir definieren daher für die abhängige Variable:

$$Y = \begin{cases} 0 \ falls \ A < 0 \\ 1 \ falls \ A \geq 0 \end{cases}$$

## 2.1 Die Lineare Regressionsgleichung

Mit Hilfe des linearen Regressionsmodells ($y_j = \alpha + \beta x_j$) lässt sich rein formal die Regression von Y nach X und damit der Zusammenhang zwischen der abhängigen und der unabhängigen Variablen berechnen.

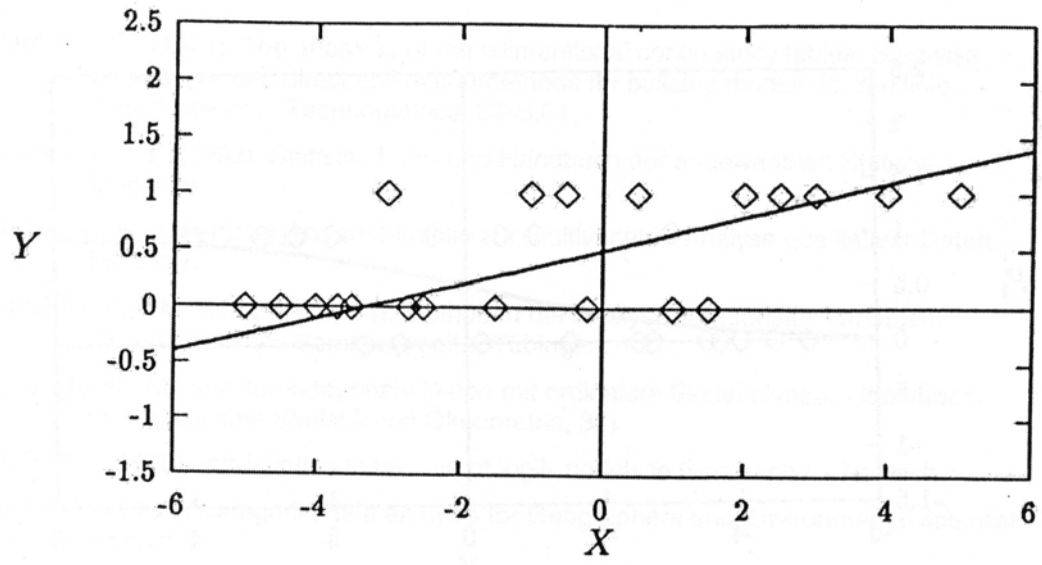


Abb. 1: Lineare Regression. (Rosner 2001: S. 59)

Die Regressionsgrade in Abb. 1 weist auf ein entscheidendes Manko dieser Regressionsanalyse hin. Die Gleichung gilt nämlich unbeschränkt, d. h. für beliebige $x_j$. Wenn $x_j$ allerdings einen bestimmten Wert über- bzw. unterschreitet, ergibt sich ein Wert größer 1 bzw. kleiner 0. Die Wahrscheinlichkeit bestimmter $x_j$ wäre damit also größer als 100% bzw. kleiner als 0%, was wenig sinnvoll ist.
Wir können also die Regressionsgleichung nur in einem bestimmten Intervall sinnvoll interpretieren, jedoch nicht mehr außerhalb dieses Intervalls.

## 2.2 Das lineare Wahrscheinlichkeitsmodell

Man kann das Problem formal dadurch beheben, indem man den Gültigkeitsbereich der entsprechenden Regressionsgleichung auf das Intervall beschränkt und Y oberhalb diese Intervalls gleich 1 und unterhalb dieses Intervalls gleich 0 setzt.

Allgemein würde man definieren:

$$p_{1j} = \begin{cases} 0 \;\; für\, a + bx_j < 0 \\ \hat{y}_j \;\; für\, 0 \le a + bx_j \le 1 \\ 1 \;\; für\, a + bx_j > 1 \end{cases}$$

$p_{1j}$ = Wahrscheinlichkeit mit der Y für X = $x_j$ den Wert 1 annimmt.

$p_{oj}$ = 1 - $p_{1j}$ = Wahrscheinlichkeit mit der Y für X = $x_j$ den Wert 0 annimmt.

(nach Bahrenberg et al 1992: S. 135)

Dieses Modell nennt man das *lineare Wahrscheinlichkeitsmodell*. Leider hat auch dieses Modell einige Nachteile, die seine Anwendbarkeit stark einschränken (vergleiche dazu auch Abb. 2). Es sind dies:

1. Schätzungen in der Nähe der Extremwerte 0 und 1 für $p_{1j}$ sind ungenau,
2. es ist nicht ohne weiteres auf den Fall erweiterbar, dass Y eine polytome Variable mit mehr als zwei Ausprägungen sei, und
3. erfahrungsgemäß entspricht der Kurvenverlauf bei einer wahrscheinlichkeitstheoretischen Interpretation eher einem S-förmigen Verlauf.

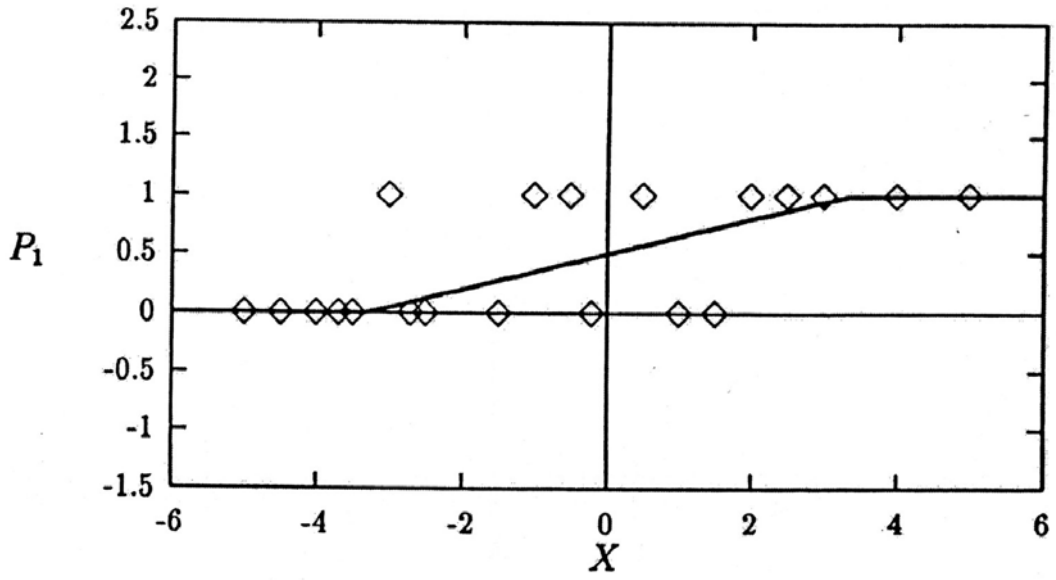


Abb. 2: Lineares Wahrscheinlichkeitsmodell (Rosner 2001: S. 59)

## 2.3 Das Logistische Modell

Die Schwächen des linearen Wahrscheinlichkeitsmodells lassen sich vermeiden, wenn man für die Regression von Y nach X nicht

$$y_j = p_{1j} = \alpha + \beta x_j$$

sondern folgende Regressionsgleichung wählt:

$$p_{1j} = \frac{e^{\alpha+\beta x_j}}{1+e^{-(\alpha+\beta x_j)}}$$

Bei dieser Gleichung kann $p_{1j}$ nur Wert zwischen 1 und 0 annehmen (s. Abb.3):

- Geht $x_j$ gegen $\infty$, geht $p_{1j}$ gegen 1.
- Geht $x_j$ gegen $-\infty$, geht $p_{1j}$ gegen 0.

Geht $x_j$ gegen $-\infty$, so geht der Zähler gegen 0, der Nenner geht gegen 1, und der Quotient geht insgesamt gegen 0.

Für die Regressionsanalyse ist diese Gleichung allerdings nicht direkt zu gebrauchen, weil $p_{1j}$ nicht linear von $x_j$ abhängig ist.

Wir formen die Gleichung deshalb wie folgt um:

$$p_{0j} = 1 - p_{1j} = 1 - \frac{e^{\alpha+\beta x_j}}{1+e^{-(\alpha+\beta x_j)}}$$

$$= \frac{1+e^{\alpha+\beta x_j} - e^{\alpha+\beta x_j}}{1+e^{-(\alpha+\beta x_j)}} = \frac{1}{1+e^{-(\alpha+\beta x_j)}}$$

$$\frac{p_{1j}}{p_{0j}} = \frac{p_{1j}}{1-p_{1j}} = e^{\alpha+\beta x_j}$$

$$l_i = \ln\frac{p_{1j}}{p_{0j}} = \ln\frac{p_{1j}}{1-p_{1j}} = \alpha + \beta x_j$$

(vgl. Bahrenberg 1992: S.136)

Der Ausdruck $\ln\frac{p_{1j}}{p_{0j}}$ heißt „Logit-Transformation“ oder kurz „Logit.“

Wir haben durch die „Logit-Transformation“ die abhängige Variable $P_1=p_{1j}$ in die abhängige Variable $L=l_j$ transformiert, wobei L jetzt linear von X abhängig ist. Wir können also auf L das einfache bivariate Regressionsmodell anwenden:

$$L = \alpha + \beta x + \varepsilon$$

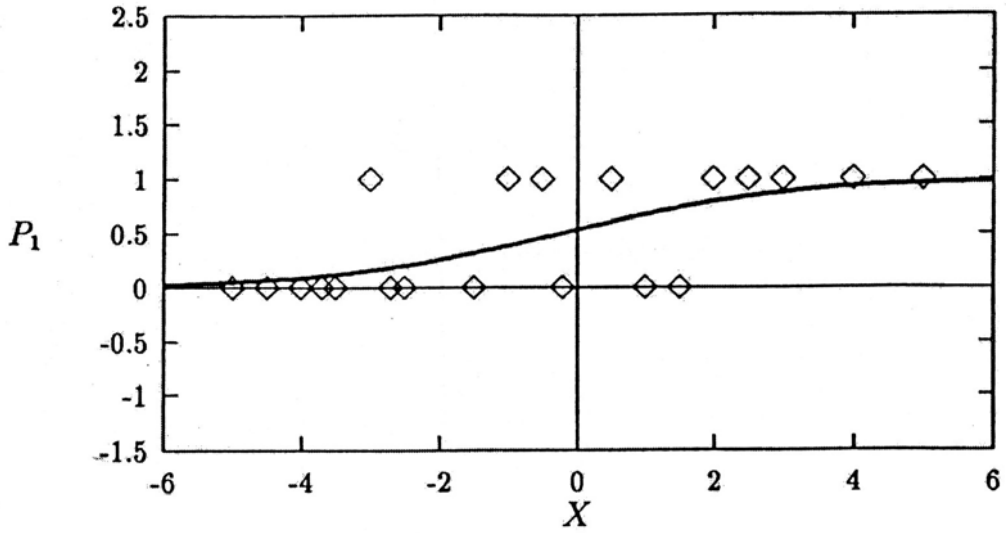


Abb. 3: Logistisches Modell (Rosner 2001: S. 60)

## 3. Anwendungsbeispiel: Pendlerverhalten von Angestellten

Dieses Beispiel soll die Anwendung des linearen Logit-Modells verdeutlichen. Hierzu wurde in einer Gemeinde eine stichprobenartige Befragung mit 12 Angestellten durchgeführt. Gefragt wurde, ob die Pendler motorisiert oder zu Fuß, bzw. mit dem Fahrrad zum Arbeitsplatz gelangen. Hierbei ist die dichotome Variable „Wahl des Verkehrsmittels“ (motorisiert / zu Fuß). Die unabhängige Variable ist „Entfernung des Arbeitsplatzes vom Wohnort in Kilometern“. Ziel ist es nun, die Wahrscheinlichkeit zu berechnen, für welche Strecke welches Fortbewegungsmittel gewählt wird.

In Tab. 3 sind die Ergebnisse der Befragung zusammengefasst, wobei der Ausprägung „motorisierte Fortbewegung“ der Wert 1 und Ausprägung „zu Fuß oder mit dem Fahrrad“ der Wert 0 zugewiesen wurde.

| | K | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Entfernung z. Arbeitsplatz in km: | $x'_k$ | 0,5 | 0,7 | 1,0 | 1,3 | 1,5 | 2,0 | 2,2 | 2,3 | 2,7 | 3,0 | 3,2 | 3,4 |
| Wahl des Verkehrsmittels: | $y'_k$ | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| „km – Gruppen“ | i | 1 | | | | 2 | | | | 3 | | | |
| Anteil der Ausprägungen i. d. jew. Km - Gruppe | $x_i$ | 1 | | | | 2 | | | | 3 | | | |
| | $p_i$ | 1/4 | | | | 2/4 | | | | 3/4 | | | |

Tab. 3: Ergebnisse der Erhebung (HARTUNG & ELPELT 1995: S. 129)

Nun soll zunächst mit Hilfe des linearen Regressionsmodells der Zusammenhang zwischen der unabhängigen Variablen (Entfernung zu Arbeitsplatz) und der abhängi-

gen Variablen (Art des Fortbewegungsmittels) ermittelt werden und erhalten somit folgende Regressionsgerade:

$$y = 0{,}0179 + 0{,}243x$$

Die berechnete Gerade des Regressionsmodells gibt einen Trend an, der allerdings in der Realität nicht möglich ist. Ab x > 4 wird der Wert für y nämlich größer als 1. Die Wahrscheinlichkeit müsste demnach ab diesem Wert über 100% liegen (vgl. Abb. 4). Ein lineares Wahrscheinlichkeitsmodell kann diesen Fehler verhindern, ist aber aufgrund der konstanten Wahrscheinlichkeiten unrealistisch.

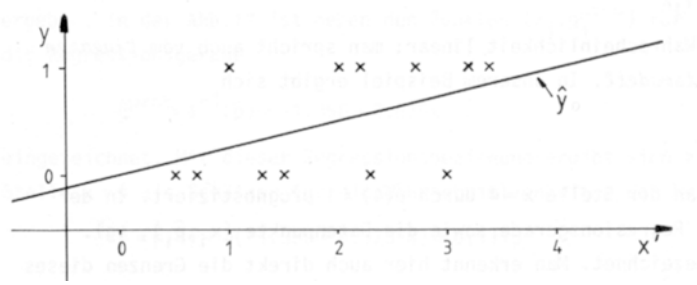


Abb. 4: Regressionsgerade für das Anwendungsbeispiel. (Hartung 1995: S. 129)

Der Sachverhalt wird am Besten mit Hilfe einer S-förmigen Kurve abgebildet, welcher folgende Gleichung zu Grunde liegt:

$$p = \frac{1}{1 + e^{\alpha + \beta x}}$$

Nach der Umformung der Gleichung (vgl. 2.3) ergibt sich:

$$\ln \frac{p}{1 - p} = \alpha + \beta x$$

Den Logarithmus des Quotienten der Wahrscheinlichkeit (der rechte Teil der Gleichung) bezeichnen wir dabei als Logit. Wenn nun die Werte 1 und 0 für p eingesetzt werden, erhält man keine Lösung, da der Logarithmus von 0 nicht existiert, oder eine Null im Nenner per Definition nicht erlaubt ist. Aus diesem Grund werden Gruppen von Wahrscheinlichkeiten gebildet (vgl. Tab. 3). Die gruppierten Logits liegen auf einer Geraden, so dass folgende Gleichung für die Regressionsgerade gegeben ist (vgl. Abb. 5):

$$g^{\lg t} = \ln \frac{p}{1 - p} = 2{,}1972 + 1{,}0986x$$

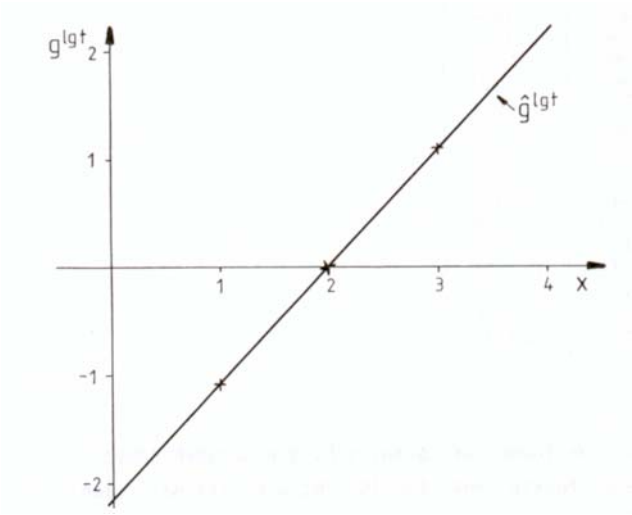


Abb. 5. (HARTUNG & ELPELT 1995: S. 133)

Nun kann also zum Beispiel an der Stelle x = 4 die Wahrscheinlichkeit p des Auftretens von Ausprägung „motorisiert“ prognostiziert werden.

$$P = p(4) = \frac{1}{1 + e^{\,2{,}1972 - 1{,}0986 \cdot 4}} = 0{,}8999$$

Es fahren also rd. 90% der Befragten bei einer Distanz von 4 km „motorisiert“ zum Arbeitsplatz.

## 4. Zum loglinearen Modell

Das loglineare Modell wird zur Lösung statistischer Probleme mit in der Regel mehr als zwei Variablen auf kategorialem Niveau (abhängige sowie ggf. unabhängige) verwendet. Es dient also der Analyse mehrdimensionaler Kontingenz- oder Kreuztabellen. Ein Beispiel hierfür wäre etwa die Untersuchung des statistischen Zusammenhangs zwischen den dichotomen Variablen „Verstädterungsgrad“ (mit den Ausprägungen niedrig / hoch) und „Arbeitsplatzentwicklung“ (mit den Ausprägungen negativ / positiv) sowie der polytomen Variable „Binnenwanderungssaldo“ (mit den Ausprägungen negativ / schwach positiv / stark). Eine Anwendung des loglinearen Modells ergibt im Wesentlichen eine Aussage darüber, *ob* überhaupt ein statistischer Zusammenhang zwischen diesen Variablen besteht und ggf. wie stark dieser ausgeprägt ist. Lässt sich das lineare Logit-Modell mit der Regressionsanalyse (metrisch skalierter) Variablen vergleichen (Aussage über die Form des Zusammenhangs),

entspricht das loglineare Modell also in etwa der Korrelationsanalyse (Stärke des Zusammenhangs).

## 5. Verwendete Literatur

HARTUNG, J. & B. ELPELT (1995): Multivariate Statistik. Lehr- und Handbuch der Angewandten Statistik. München.

BAHRENBERG, G., E. GIESE & J. NIPPER (1992): Statistische Methoden in der Geographie. Band 2: Multivariate Statistik. Stuttgart.

ROSNER, H.-J. (2001): Verarbeitung Geographischer Daten. Methodische Bausteine zu Statistik und GIS. Tübingen.