Modellierung der Salmonella–Prävalenz bei Hähnchen

ABSCHLUSSARBEIT

zur Erlangung des akademischen Grades

Bachelor of Science (B.Sc.)

im Bachelorstudiengang Statistik

an der Wirtschaftswissenschaftlichen Fakultät

Humboldt-Universität zu Berlin

von

Alena Myšičková

geboren am 19.03.1981 in Čáslav

Matrikelnr.: 184568

Gutachter:

Prof. Dr. Wolfgang Härdle
Prof. Dr. Bernd Rönz

Berlin, den 17. März 2005

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel und Quellen angefertigt habe.
Die Arbeit hat keiner anderen Prüfungsbehörde vorgelegen.

Berlin, den 17. März 2005 Alena Myšičková

Unterschrift

Inhaltsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die hier vorliegende Arbeit hat zum Ziel eine Untersuchung zum Vorkommen von Salmonellen bei deutschem Geflügelfleisch, insbesondere eine Analyse des Einflusses von Eigenschaften des Geflügelmastbetriebes auf die Salmonella-Prävalenz.
Sowohl das Thema als auch die Daten stammen von dem Bundesinstitut für Risikobewertung in Berlin, aus einer dort im Jahr 1999 ausgeführten Studie.
Das Vorkommen von Salmonellen bei Nutzgeflügel wurde gewählt, weil salmonella–enthaltende Lebensmittel ein erhebliches gesundheitliches und ökonomisches Problem darstellen. In Praxis wird Geflügelfleisch nicht nur erhitzt verbraucht, da in zunehmendem Umfang aus Geflügelfleisch auch Rohfleischerzeugnisse hergestellt werden und Salmonellen durch Kreuzkontamination im Küchenbereich in Speisen gelangen können, die vor dem Verzehr nicht mehr erhitzt werden.
Vor weiteren Einleitung werden auf dieser Stelle wichtige Begriffe definiert.
Unter **Prävalenz** oder auch Grundanteil versteht man in der medizinischen Statistik und der Epidemiologie die Häufigkeit, in der eine bestimmte Krankheit (oder ein bestimmtes Merkmal) in einer bestimmten Bevölkerung (Population) vorkommt[1].
Mit **Kreuzkontamination** bezeichnet man die direkte oder indirekte Übertragung von pathogenen (krankheitserregenden) Mikroorganismen von verdorbenen Lebensmitteln auf andere Lebensmittel. Sie ist eine Hauptursache von Lebensmittelvergiftungen. [2]
Um die Salmonella–Prävalenz bei den Hähnchen in dem Zusammenhang mit den Eigenschaften des Betriebes modellieren zu können, wird ein Regression Modell verwendet. Da die Prävalenz nur Werte aus dem Intervall $[0, 1]$ annehmen kann, wird man aus der Klasse der generalisierten linearen Modellen auswählen, konkret wird in diesem Fall das Logit Modell benutzt. Das Salmonella–Vorkommen in entnommenen Proben wird als abhängige Variable betrachtet, als erklärende Variablen werden die Informationen über die Betriebe aufgenommen.
Im anschließenden Kapitel wird der Datensatz, sowohl auch die Daten– und Materialerhebung beschrieben.
Das dritte Kapitel stellt die verwendete statistische Methode, ihre Schätzung und die Hypothesenprüfung vor.
Im Kapitel 4 werden die mit dem statistischen Programm XploRe gewonnenen Ergebnisse präsentiert und die Interpretation der geschätzten Koeffizi-

[1] Quelle: de.wikipedia.org
[2] Quelle: www.eufic.org

enten erklärt.
Das fünfte Kapitel faßt die wichtigsten Punkte der Arbeit zusammen.
Die Quantlets des statistischen Programms XploRe sind dem siebten Kapitel zu entnehmen.

2 Daten und Material

2.1 Datensatz

Der für diese Arbeit zu Grunde liegende Datensatz stammt aus dem Bundesinstitut für Risikobewertung in Berlin (ehemaliger Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin). Der Datensatz wurde bereits studiert, im Jahr 1999 wurde eine Studie über Sallmonellen–Infektionen bei Hähnchen durchgeführt. Die Studie hatte als Ziel die Daten zum Infektions– und Kontaminationsgrad von Hähnchen aus verschiedenen Betrieben in Deutschland zu erheben und weiterhin ein wiederkehrendes, eventuell permanentes Auftreten von bestimmten Salmonella–Stämmen zu überprüfen. Die Ergebnisse damaliger Analyse wurden im Februar 2001 in einem Bericht "Untersuchungen zum Vorkommen von Salmonellen bei deutschem Nutzgeflügel und Geflügelfleisch" von L. Ellerbroek, H. Wichmann-Schauer und M. Haarmann veröffentlicht.
Es wurden insgesamt 189 Herden beprobt. Unter einer Geflügel–Herde versteht man eine ganze Population allen Tieren, die sich in dem Zeitpunkt der Entnahme im Stall befinden. Alle Herden kamen aus insgesamt 66 Betrieben und aus 5 verschiedenen Regionen in Deutschland, die aus Datenschutzgrunden nachfolgend als Region A, B, C, D und E kodiert werden. Die Betriebe wurden in der Regel mehrmals (zwei– bis dreimal) untersucht. Die Hähnchen werden so gezüchtet, dass der Stall vor dem Anlegen einer neuen Herde entleert und gereinigt wird. Nach einer bestimmten Zeit werden dann alle überlebenden Tiere ins Schlachthof transportiert und geschlachtet. Eine neue Herde wird erst nach einer Neureinigung des Stalles gebracht. Bei der Probenentnahmen waren die Zeitabstände groß genug, damit immer verschiedene Herden untersucht werden könnten. Deswegen können wir die Beobachtungen als unabhängig betrachten.
Die Geflügelmastbestände wurden in 2 Gruppen nach der Jahresproduktion verteilt; Betriebe mit einer jährlichen Produktion von mehr als 20 000 Hähnchen und eigenem Management werden weiterhin als *Großbetriebe*, die Betriebe mit weniger als 20 000 Stück Schlachtgeflügel pro Jahr als *Kleinbetriebe* bezeichnet.
Für die obengenannte Studie wurden 3 verschiedene Probenarten entnommen. Das Sammeln der sogenannten *Gazekotproben* erfolgte durch Mullgazeschläuche, welche beim Durchlaufen des Stalles von Mitarbeitern über den Schutzstiefeln getragen wurden. Dabei kann man davon ausgehen, dass die gesamte Stallfläche möglichst representativ beprobt wurde. Etwa 3 Wochen später wurden während der Schlachtung bei denselben Herden *Halshaut–*

und *Kloakentupferproben* entnommen. Die Kloakentupferproben wurden aufgrund der geringen Tagesschlachtzahlen bei Kleinbetrieben nicht untersucht. Dieser Arbeit beschäftigt sich nur mit der Analyse der Halshautproben, da die Kloakentupferproben nur bei einem Teil der Herden entnommen wurden. Bei den Gazekotporben könnte man nicht die einzelnen Tiere, sondern nur ganze Herden untersuchen, was zu einer mehr komplexen Arbeit führen würde.

2.2 Poolproben

Das entnommene Material wurde zu sogenannten *Poolproben* vereinigt, so dass die Halshautproben von 5 verschiedenen Tieren einer Herde eine Poolprobe gebildet haben. Die Poolproben wurden dann in 5 Laboren [3] kulturell aufgearbeitet und jede Poolprobe auf das Vorkommen von Salmonella untersucht.
Die Bestimmung des positiven (bzw. negativen) Befundes wurde folgenderweise definiert:
"Die Poolprobe wurde als positiv beurteilt, wenn nach Kultivierung auf mindestens einer Selektivplatte einer Selektivanreichung verdächtige Kolonien gewachsen waren, die Lactose–negativ waren und mit dem omnivalenten Antiserum agglutinierten, und wenn das Referenzlabor die Zugehörigkeit von mindestens einem Isolat dieser Probe zur Gattung Salmonella bestätigen konnte. Dementsprechend wurden Proben als negativ eingestuft, wenn sich aus ihnen keine derartigen Kolonien anzüchten ließen.[4] *"*
Dieser Arbeit wird sich mit weiteren biologischen Details über den kulturellen Nachweis von Salmonellen nicht beschäftigen, die Beurteilung der Positivität (bzw. Negativität) wurde als eine eindeutige Aussage angenommen.
Ein kleiner Überblick über die untersuchten Poolproben gibt die Tabelle (1).

[3] Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin(BgVV) Berlin und Jena, die Außenstelle für Epidemiologie der Tierärztlichen Hochschule Hannover in Bakum, das Institut für Anatomie, Physiologie und Hygiene der Haustiere der Universität Bonn, das Institut für Hygiene und Technologie der Lebensmittel der Universität München.

[4] Siehe bgvv (2001).

	Halshautproben	Gazekotproben	Kloakentupferproben
Großbetriebe	840	620	840
Kleinbetriebe	136	267	–
Gesamt	976	887	840

Tabelle 1: Anzahl der untersuchten Poolproben.

2.3 Betrieb

Für die Bewertung der Untersuchungsergebnisse wurden in Mastbetrieben Daten über Management, baulichen Bedingungen und zur Bestimmung des Hygienestatus erhoben. Die ausgewählten Faktoren, bei denen ein eventueller Einfluss auf die Salmonella–Prävalenz zu erwarten ist, sind:

- die Größe des Betriebes,
- die Region in Deutschland,
- das Vorkommen von anderem Geflügel auf dem Hof,
- die Durchführung von einer Schädlingsbekämpfung,
- die Entfernung von einem anderem Betrieb,
- die Woche der Entnahme und
- die Klassifizierung der Hygienekonditionen.

Es wird ein positiver Effekt auf die Salmonella–Prävalenz [5] bei einem schlechten Hygienestand des Betriebes und bei nicht Durchführung der Schlädlingsbekämpfung erwartet. Einen positiven Zusammenhang zwischen das Vorhanden des anderen Geflügels und eine größere Häufigkeit des Salmonella-Vorkommens könnte auch auftreten, da die Kreuzkontamination von anderen Tieren zu erwarten ist.
Andererseits sollte sich kein Einfluss der Entnahmewoche auf die Salmonella–Prävalenz zeigen, da die Salmonella Bakterien allgemein unabhängig von der Jahreszeiten auftreten.

[5] Hier in dem Sinne des positiven Einfluss auf das Salmonella–Vorkommen, obwohl allgemein für Menschen und die Betriebe das Auftreten von Salmonellen eine negative Auswirkung hat.

3 Statistische Analyse

Wie bereits erwähnt, ist das Ziel dieser Arbeit eine statistische Modellierung der Salmonella–Prävalenz in der Abhängigkeit von den Betriebseigenschaften.
Diese Idee und auch der Datenzustand weist darauf hin, dass man ein Regression–Modell verwenden soll. Da die abhängige Variable *Salmonella–Prävalenz* nur Werte aus einem beschränkten Intervall ([0,1]) annehmen kann, wird ein Modell aus der Klasse der Generalisierten linearen Modellen (GLM) ausgewählt.

3.1 GLM – Generalisierte lineare Modelle

Generalisierte (verallgemeinerte) lineare Modelle stellen einen Spezialfall der Regression–Modellen dar. Regression–Modellen ist eine Klasse von Modellen, die die Beziehungen – insbesondere die Stärke und Form – zwischen zwei oder mehreren Variablen, denen ein beobachteter Datensatz zugrunde liegt, untersuchen.
Ein generalisiertes lineares Modell ist ein Regression–Modell in der Form:

$$\mathrm{E}Y = G(\mathbf{x}^\top \boldsymbol{\beta}) \,, \tag{1}$$

wobei $\mathrm{E}Y$ für den Mittelwert der response Variable Y steht, $\mathbf{x}$ ist der Vektor der erklärenden Variablen und $\boldsymbol{\beta}$ der Vektor der unbekannten, zu schätzenden Parameter, darstellt.
Die Funktion $G(\bullet)$ wird als die *Linkfunktion* [6] bezeichnet, weil sie den Mittelwert $\mathrm{E}Y$ auf die Linearkombination $\eta \stackrel{\text{def}}{=} \mathbf{x}^\top \boldsymbol{\beta}$ bezieht.

GLM beinhalten eine ganze Reihe von weitverwendeten Modellen wie z.B.: lineare Regression, Varianzanalyse, Logit– oder Probit–Modell. Verallgemeinerte lineare Modelle finden ein breites Spektrum von Anwendungen in medizinischen, biometrischen und auch sozio–ökonomischen Studien.
In den GLM (ebenso wie im Regression–Modell) unterscheidet man zwischen 2 Typen von Variablen: den unabhängigen oder auch erklärenden Variablen und den abhängigen, erklärten oder auch response Variablen.

[6] Bei McCullagh and Nelder (1989) wurde die Inversefunktion G^{-1} als die Linkfunktion bezeichnet.

3.2 Variablen im Prävalenz–Modell

Als die abhängige Variable wird die Zufallsvariable $Y = (Y_1, \ldots, Y_n)$ betrachtet. n bezeichnet in diesem Fall die Anzahl der untersuchten Poolproben, in unserer Studie $n = 976$. Die Zufallsvariable Y_i kann folgende Werte annehmen:

$$Y_i = \begin{cases} 1 & i-\text{te Poolprobe ist Salmonella–positiv} \\ 0 & i-\text{te Poolprobe ist Salmonella–negativ} \end{cases} \tag{2}$$

Die Wahrscheinlichkeit, dass die Poolprobe i Salmonella–positiv ist, wird als π_i bezeichnet:

$$\pi_i \stackrel{\text{def}}{=} \mathrm{P}(Y_i = 1) \; , \tag{3}$$

und so gilt:

$$1 - \pi_i = \mathrm{P}(Y_i = 0) \; .$$

Die response Variable Y ist Bernoulli–verteilt mit dem Parameter π_i.

Bevor die erklärenden Variablen ins Modell angenommen werden können, mussten einige zur Verfügung stehende Variablen aus dem Datensatz umkodiert werden. Die mehrkategoriale Variable *Woche der Entnahme* wurde in eine zweikategoriale Variable, die nur zwischen den Sommmerwochen und anderen Jahreszeiten unterscheidet, transponiert. Die metrisch skalierte Variable *Entfernung von nächsten Hühnerbestand* wurde auch in einer zweikategorialen Variable (Entfernung größer oder kleiner als 1 km) umkodiert. Dies erleichtert die Analyse, weil man im Falle der Entfernung als eine stetige Variable große Anzahl von Kovariten–Gruppen erhalten würde, was zu einer komplizierten Interpretation des Modells führen könnte. Bei der Ermittlung von der Hygieneklassifizierung wurden von den Mitarbeitern Fragebögen ausgefüllt und nach der Auswertung bei den Großbetrieben 3, bei den Kleinbetrieben 2 verschiedene Hygieneklassen gebildet. Dieser Klassifizierung würde aber für eine weitere Analyse allen Betrieben unpassend. Es wurde deshalb eine neue Variable *Hygienegruppe* gebildet, die für die Kleinbetriebe unverändert bleibt und bei den Großbetrieben die ehemaligen Klassen I und II in die Hygienegruppe 1 zusammenfügt. Die Hygieneklasse 3 bei den Großbetrieben wurde dann in die Hygienegruppe 2 umgewandelt. Alle genannten Umkodierungen wurden nach einer Beratung mit den Spezialisten des Bundesinstituts für Risikobewertung durchgeführt.

Insgesamt wurden 7 erklärende Variablen betrachtet:

X_1 : die Region mit den Kategorien A, B, C, D und E

X_2 : die Betriebsgröße aufgeteilt in Kleinbetrieb und Großbetrieb (0 = Kleinbetrieb, 1 = Großbetrieb)

X_3 : ein Indikator für anderes Geflügel auf dem Hof (0 = ja, 1 = nein)

X_4 : ein Indikator für die Durchführung der Schädlingsbekämpfung (0 = ja, 1 = nein)

X_5 : die Entfernung vom nächsten Hühnerbestand aufgeteilt in kleiner als 1km und größer als 1 km ($0 = < 1000$m, $1 = \geq 1000$ m)

X_6 : ein Indikator für das Sommer (0 = Frühjahr/Herbst, 1 = Sommer (Juni – September))

X_7 : die Hygienegruppe aufgeteilt in 2 Kategorien

Daraus lassen sich folgende konkrete Fragen an die statistische Untersuchng bilden:

In wieweit wird die Salmonella–Prävalenz von der erklärenden Variablen beeinflusst?

Welche Faktoren haben einen signifikanten Einfluss auf die Prävalenz?

Für die Antwort wurde ein Spezialfall der Generalisierten linearen Modellen - das Logit Modell - verwendet.

3.3 Modellierung binären Daten und Logit Modell

Aufgrund der binären Verteilung der response Variable Y soll die Linkfunktion das Intervall $[-\infty, \infty]$ auf das Intervall [0,1] abbilden:

$$\begin{aligned} \lim_{\eta \to \infty} G(\eta) &= 1 \\ \lim_{\eta \to -\infty} G(\eta) &= 0\,. \end{aligned}$$

Diese Transformation ist in diesem Modell notwendig, weil die Linearkombination $\eta = \mathbf{x}^\top \boldsymbol{\beta}$ theoretisch die Werte zwischen $-\infty$ und ∞ annehmen kann, wobei die Wahrscheinlichkeit π nur Werte aus dem $[0, 1]$–Interval annehmen

kann.
In dem Logit Modell wird als Linkfunktion die logistische Verteilungsfunktion verwendet:

$$G(\mathbf{x_i}^\top \boldsymbol{\beta}) = \frac{e^{\mathbf{x_i}^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x_i}^\top \boldsymbol{\beta}}} = \pi_i \ , \tag{4}$$

mit der Inversefunktion:

$$\mathbf{x_i}^\top \boldsymbol{\beta} = \log\left(\frac{\pi_i}{1-\pi_i}\right) = g(\pi_i) \ . \tag{5}$$

Der Bruch $\frac{\pi_i}{1-\pi_i}$ wird als Odds (Chancen) eines Erfolges bezeichnet, da er den Verhältnis der Erfolgswahrscheinlichkeit zur Misserfolgswahrscheinlichkeit darstellt. Die logarithmierten Odds werden in der Literatur als log Odds bezeichnet.

Weil ein nichtlinearer Zusammenhang zwischen den Wahrscheinlichkeiten π_i und den erklärenden Variablen X_i besteht, ist die Interpretation der geschätzten Koeffizienten $\hat{\beta}_i$ nicht so einfach wie im Falle der linearen Regression.
Die mehrkategoriale erklärende Variable "Region" ist nicht direkt im Logit–Modell verwendbar, sie musst in 4 neuen Variablen (als Kontrast–Variablen bezeichnet) mit einer Bezugskategorie umkodiert werden. Für die Erstellung der Kontrast–Variablen wird in diesem Modell eine Indikator–Kodierung [7] in Dummy–Variablen (0;1-Variablen) verwendet. Mit den neuen Kontrast–Variablen erhöht sich die Anzahl der erklärenden Variablen und damit auch die Anzahl der zu schätzenden Parametern $\hat{\beta}_j$, die dann immer bezüglich der Referenzkategorie zu interpretieren sind.

Hinsichtlich der allgemeinen Interpretation von Koeffizienten entspricht der Parameter $\hat{\beta}_j$ der Veränderung in den log Odds (siehe Formel(5)), falls sich die zugehörige Variable X_j um eine Einheit erhöht (oder eine andere Kategorie als die Referenzkategorie annimmt), bei gleichzeitiger Konstanz der Werten aller anderen Einflussfaktoren. Hinsichtlich der Erfolgswahrscheinlichkeit π kann der Regressionparameter nur über die Richtung der Veränderung aussagen (Ist der Parameter positiv, wird sich die Erfolgswahrscheinlichkeit mit Erhöhung der zugehörigen Variable X_j um eine Einheit und Fixierung allen anderen Variablen vergrößern; wird der Parameter negativ, wird sich die Wahrscheinlichkeit bei der gleichen Veränderung der erklärenden Variablen verkleinern.). Erhöht sich die Variable X_j um eine Einheit und werden die

[7] Siehe Rönz (2001).

Werte der Einflussfaktoren konstant gehalten, verändern sich die Odds Ratios[8] multiplikativ mit dem Faktor $\exp(\hat{\beta}_j)$.

3.4 Schätzung des Logit–Modells mit ML–Methode

Zur Schätzung der Modellkoeffizienten β_i wird meistens in GLM die Maximum–Likelihood Methode verwendet. Für die log–Likelihood–Funktion gilt:

$$l(\boldsymbol{\pi}, \mathbf{Y}) = \log f(\mathbf{Y}, \boldsymbol{\beta}) = \sum_{k=1}^{K} \log f_k(Y_k, \beta_k) \ , \tag{6}$$

wobei $f(\mathbf{Y}, \boldsymbol{\beta})$ die Dichtefunktion von $\mathbf{Y}$ für festen Parametern $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_J)$ darstellt. Nach Verwendung der Linkfunktion $G(\mathbf{x_i}^\top \boldsymbol{\beta})$ in die Dichtefunktion werden die unbekannten Wahrscheinlichkeiten π_i durch unbekannten Parametern β_j ersetzt. Index K bezeichnet die Anzahl der Kovariaten–Gruppen, n_k die Anzahl der Beobachtungen in der k–ten Kovariaten–Gruppe und für den gesamten Stichprobenumfang gilt: $n_1 + n_2 + \cdots + n_K = n$.
Im Logit–Modell ist die Anzahl der Erfolge in jeder Kovariaten–Gruppe binomialverteilt (Summe von Bernoulli–verteilten Variablen) mit Parametern π_k (Erfolgswahrscheinlichkeit in der k–ten Gruppe) und n_k. Die Dichtefunktion führt dann zu:

$$f_k(y_k, \pi_k) = \begin{pmatrix} n_k \\ y_k \end{pmatrix} \pi_k^{y_k} (1 - \pi_k)^{n_k - y_k} \ . \tag{7}$$

Die log–likelihood–Funktion ist dann:

$$l(\pi_1, \ldots, \pi_K; y_1, \ldots, y_K) = \sum_{k=1}^{K} \left[y_k \log \left(\frac{\pi_k}{1 - \pi_k} \right) + n_k \log(1 - \pi_k) + \log \begin{pmatrix} n_k \\ y_k \end{pmatrix} \right] \ . \tag{8}$$

Nach Vernachlässigung des letzten Terms (spielt keine Rolle für die Schätzung der π_k) ergibt sich:

$$l(\pi_1, \ldots, \pi_K; y_1, \ldots, y_K) = \sum_{k=1}^{K} \left[y_k \log \left(\frac{\pi_k}{1 - \pi_k} \right) + n_k \log(1 - \pi_k) \right] \ .$$

[8]Das Verhältnis zweier Chancen - Chance des Erfolges unter den Fällen mit Annehmen von X_j einer Kategorie (bzw.Referenzkategorie) im Gegensatz zu den Fällen mit Annehmen von X_j einer anderen Kategorie.

Die log Odds wurden durch (5) ersetzt und die log–Likelihood $l(\boldsymbol{\beta}, \mathbf{y})$ wird eine Funktion von Beobachtungsdaten und den unbekannten Parametern β_j. Die partielle Ableitung nach β_j wird gleich Null gesetzt und damit die log–Likelihood Funktion maximiert. Dies führt aber zu nichtlinearen Gleichungen, die iterativ z.B. mit der Fisher Scoring Methode [9] gelöst werden.

3.5 Hypothesenprüfung und Güte der Anpassung

Die Likelihood Funktion wird auch für den Vergleich des geschätzten Modells mit dem maximalen Modell[10] in der Deviance verwendet:

$$D = -2[l(\hat{\boldsymbol{\pi}}; \mathbf{y}) - l(\hat{\boldsymbol{\pi}}_{max}; \mathbf{y})] \; , \tag{9}$$

wobei $l(\hat{\boldsymbol{\pi}}; \mathbf{y})$ ist der maximale Wert der log–Likelihood–Funktion an der Stelle der geschätzten Wahrscheinlichkeit $\hat{\boldsymbol{\pi}}$ und $l(\hat{\boldsymbol{\pi}}_{max}; \mathbf{y})$ ist der maximale Wert der log–Likelihood–Funktion bei dem maximalen Modell an der Stelle $\hat{\pi}_k = y_k; \quad k = 1, \ldots, n$. Unter der Voraussetzung, dass die Anzahl der Beobachtungen n festgelegt ist, $K < n$ und $n_k \to \infty$ für jede k, folgt D approximativ einer χ^2–Verteilung mit $n - J$ Freiheitsgraden. Ist das Modell mit J Parameter korrekt, so dass $D \approx \chi^2_{n-J}$, kann man erwarten[11]

$$D \cong n - J \; . \tag{10}$$

Die Approximation an die χ^2–Verteilung ist relativ schlecht, wenn einige n_k klein sind. Deshalb wird die Deviance vielmehr für einen Vergleich zweier Modellen mit unterschiedlicher Parameterzahl verwendet. Unter der Nullhypothese wird ein Modell M_0 mit J erklärenden Variablen (damit J zu schätzenden Parametern) gegen ein Modell M_1 mit $J + t$ erklärenden Variablen (und somit $J + t$ zu schätzenden Parametern) getestet. Dann ergibt sich für die Deviance:

$$\Delta D = D_0 - D_1 = 2[l_1(\hat{\boldsymbol{\pi}}; \mathbf{y}) - l_0(\hat{\boldsymbol{\pi}}; \mathbf{y})] \; ,$$

wobei D_0 die zum Modell M_0 gehörende Deviance und D_1 die zum Modell M_1 gehörende Deviance bezeichnet. ΔD ist approximativ χ^2–verteilt mit

[9] Für eine detaillierte Beschreibung der Methode siehe McCullagh and Nelder (1989) oder Rönz (2001).

[10] Das maximale oder auch saturierte Modell enthält keine Konstante und keine erklärenden Variablen, siehe Rönz (2001).

[11] Folgt aus der groben Approximation der χ^2–Verteilung, dass der Erwartungswert gleich die Anzahl der Freiheitsgrade ist, siehe Dobson (1990).

$f = (K - J) - [K - (J + t)] = t$ Freiheitsgraden.
Ein anderes Maß für die Diskrepanz des Modells ist die Pearson Statistik:

$$X^2 \stackrel{\text{def}}{=} \sum_k \frac{(y_k - \hat{\pi}_k)^2}{s(\hat{\beta}_k)} \ ,$$

wobei $s(\hat{\beta}_k)$ für das k–te Diagonalelement der geschätzten Varianz–Kovarianz–Matrix für $\hat{\beta}_k$ steht. Die Pearson Statistik folgt approximativ einer χ^2–Verteilung mit $f = n - J$ Freiheitsgraden (J ist die Anzahl der zu schätzenden Parametern β_j), wenn das Modell korrekt ist.
Ein Test zur Prüfung der einzelnen Parameter $\beta_j; j = 1, \ldots, J$ auf Signifikanz ist der t–Test[12] mit der Teststatistik:

$$t = \frac{\hat{\beta}_j}{\sqrt{s(\hat{\beta}_j)}} \ ,$$

worin $s(\hat{\beta}_j)$ das j–te Diagonalelement der geschätzten Varianz–Kovarianz–Matrix für $\hat{\beta}_j$ bezeichnet. Die Teststatistik t folgt approximativ einer t–Verteilung mit einem Freiheitsgrad.

[12]Verwendung einer Modifikation dieses Tests - der Wald–Test - ist auch möglich, siehe z.B. Rönz (2001).

4 Ergebnisse

Alle in diesem Kapitel beschriebenen Ergebnisse wurden mit Hilfe der statistischen Umgebung XploRe berechnet. Dazu verwendete Quantlets sind dem Anhang zu entnehmen.

4.1 Bivariate Analyse

Im folgenden Abschnitt werden einfache zweidimensionale Analysen durchgeführt, es wurden die Zusammenhänge zwischen der abhängigen Variable Y und der einzelnen erklärenden Variablen X_j geprüft. Dafür wird das Quantlet `crosstable` verwendet, welches die paarweisen Kontingenztabellen berechnet und die Ergebnisse der χ^2–Unabhängigkeitstests zusammen mit Überschreitungswahrscheinlichkeit und den Kontingenzkoeffizienten gibt. Die Ergebnisse sind in Tabellen (2) bis (8) zusammengefasst. Da die Wahrscheinlichkeiten des Nichtablehnen der Unabhängigkeitsnullhypothese bei allen Variablen außer "Entfernung von nächstem Hühnerstand" $(33,34\%)$ und "Schädlingsbekämpfung" (7.43%) Werte kleiner als 5% annimmt, scheint die Auswahl der exogenen Variablen für das Modell sinnvoll.
Aus der Analyse der Variable "Region" ist deutlich zu sehen, dass eine höhere Salmonella–Prävalenz (59,1 und 48,7%) in den Regionen A und E vorkommt, wobei in den anderen 3 Regionen der Anteil der positiven Poolproben bei maximal 20%(siehe Tabelle (2)) liegt. Der größte Einfluss auf die Salmonella–Prävalenz weist die Variable "Betriebsgröße" auf, in den großen Mastbeständen kommt die Salmonella bei mehr als 40% der Poolproben vor, in den Kleinbetrieben dann nur bei weniger als 10% der entnommenen Proben (Tabelle (3)). Keine große Unterschiede wurden andererseits bei den Variablen "Zeitraum" (was auch zu erwarten wurde) und "Entfernung von dem nächsten Hühnerstand" gefunden (Tabellen (7) und (6)). Überraschend ist das Ergebniss der Analyse der Variable "Hygienegruppe", weil die Betriebe mit einem besseren Hygienemanagement eine deutlich größere (fast doppelte) Salmonella–Prävalenz aufweisen (Tabelle (8)). Ein Gegeneffekt wurde auch bei den Indikatoren für anderes Geflügel auf dem Hof (Tabelle (4)) und für die Schädlingsbekämpfung (Tabelle (5)) erwartet, da von anderen Tieren auf dem Hof zu der Kreuzkontamination ankommen kann und damit sich der Anteil der positiven Poolproben erhöhen kann. Bei den Betrieben, wo die Schädlingsbekämpfung durchgeführt wurde, ist auch eine niedrigere Prävalenz zu erwarten. Bei beiden Variablen hat sich dieser Einfluss in der bivariaten Analyse nicht gezeigt. Ob sich die Effekte der einzelnen zweidi-

mensionalen Analysen auch in dem gesamten Logit Modell beweisen, zeigt der nächste Abschnitt.

	Poolproben		
	negativ	positiv	Gesamt
A	108	156	264
	40,9%	**59, 1**%	100,0%
B	148	38	186
	79,6%	**20, 4**%	100,0%
C	48	6	54
	88,9%	**11, 1**%	100,0%
D	202	34	236
	85,6%	**14, 4**%	100,0%
E	121	115	236
	51,3%	**48, 7**%	100%
χ^2–Statistik			159.99
Signifikanzniveau			0.000

Tabelle 2: Die Kontingenztabelle für die Variablen Region und Salmonella–Vorkommen.

`bianalyse.xpl`

	Poolproben		
	negativ	positiv	Gesamt
Kleinbetriebe	126	10	136
	92.6%	**7.4**%	100%
Großbetriebe	501	339	840
	59.6%	**40.4**%	100%
χ^2–Statistik			55.50
Signifikanzniveau			0.000

Tabelle 3: Die Kontingenztabelle für die Variablen Betriebsgröße und Salmonella–Vorkommen.

bianalyse.xpl

	Poolproben		
	negativ	positiv	Gesamt
And.Geflügel ja	92	26	118
	78.0%	**22.0**%	100%
And.Geflügel nein	535	323	858
	62.4%	**37.6**%	100%
χ^2–Statistik			11.01
Signifikanzniveau			0.001

Tabelle 4: Die Kontingenztabelle für die Variablen Anderes Geflügel und Salmonella–Vorkommen.

bianalyse.xpl

	Poolproben		
	negativ	positiv	Gesamt
Schädlingsbekämpfung ja	226	146	372
	60.8%	**39.2**%	100%
Schädlingsbekämpfung nein	401	203	604
	66.4%	**33.6**%	100%
χ^2–Statistik			3.19
Signifikanzniveau			0.074

Tabelle 5: Die Kontingenztabelle für die Variablen Schädlingsbekämpfung und Salmonella–Vorkommen.

bianalyse.xpl

	Poolproben		
	negativ	positiv	Gesamt
Entfernung $<$ 1000 m	340	178	518
	65.6%	**34.4**%	100%
Entfernung $\geq$ 1000 m	287	171	458
	62.7%	**37.3**%	100%
χ^2–Statistik			0.94
Signifikanzniveau			0.333

Tabelle 6: Die Kontingenztabelle für die Variablen Entfernung und Salmonella–Vorkommen.

bianalyse.xpl

	Poolproben		
	negativ	positiv	Gesamt
Frühjahr/Herbst	265	173	438
	60,5%	**39, 5**%	100,0%
Sommer	362	176	538
	67,3%	**32, 7**%	100,0%
χ^2–Statistik			4.84
Signifikanzniveau			0.028

Tabelle 7: Die Kontingenztabelle für die Variablen Zeitraum der Entnahme und Salmonella–Vorkommen.

bianalyse.xpl

	Poolproben		
	negativ	positiv	Gesamt
Hygienegruppe 1	359	273	632
	56,8%	**43, 2**%	100,0%
Hygienegruppe 2	268	76	344
	77,9%	**22, 1**%	100,0%
χ^2–Statistik			43.18
Signifikanzniveau			0.000

Tabelle 8: Die Kontingenztabelle für die Variablen Hygienegruppe und Salmonella–Vorkommen.

bianalyse.xpl

4.2 Ergebnisse der Logistischen Regression

Für die Schätzung der Regressionskoeffizienten $\hat{\beta}_j$ wird das Quantlet `glmest` verwendet. Als Input-Variablen gibt man die erklärenden Variablen X_j und die response Variable Y. Es wurde die Option "bilo" für die binäre logistische Regression (Logit Modell) benutzt. Als Output bekommt man die geschätzten Koeffizienten $\hat{\beta}_j$, die geschätzte Varianz–Kovarianz–Matrix der Koeffizienten und Statistiken, die über der Signifikanz der Koeffizienten und über der Güte der Anpassung des Modells aussagen. Für einen schönen Überblick der geschätzten Parametern und eine informative Grafik der Linearkombination $\eta = \mathbf{x}^\top \boldsymbol{\beta}$ vs. der Linkfunktion (bzw. Y) wurden die Quantlets `glmout` und `glmtable` gebraucht. Die Ergebnisse sind folgender Tabelle (9) und Abbildung (1) zu entnehmen.

Parameter	$\hat{\beta}_i$	S.E.	t-value	p-value	$\exp \hat{\beta}_i$
Konstante	-3.1729	0.7993	-3.9698	0.000	0.0419
Betriebsgröße	2.4389	0.3723	6.5506	0.000	11.4600
And. Geflügel	-0.0840	0.3744	-0.2243	0.823	0.9194
Schädlingsbekämpfung	0.2910	0.1966	1.4798	0.139	1.3378
Entfernung	0.2319	0.1810	1.2812	0.200	1.2609
Zeitraum	-0.4864	0.1654	-2.9405	0.003	0.6149
Hygienegruppe	-0.8078	0.2513	-3.2146	0.001	0.4459
Region A	-0.3977	0.2342	-1.6980	0.089	0.6719
Region B	-1.6198	0.2937	-5.5146	0.000	0.1979
Region C	-2.3699	0.4987	-4.7522	0.000	0.0935
Region D	-2.2078	0.2689	-8.2119	0.000	0.1099

Tabelle 9: Die geschätzten Parameter und Odds Ratios für das Logit Modell

`LogitModell.xpl`

Wie aufgrund der bivariaten Analyse zu erwarten war, unterscheiden sich die Groß– und Kleinbetriebe hinsichtlich der Salmonella–Prävalenz hochsignifikant. Das Odds Ratio (oder auch $\exp(\hat{\beta})$) sagt nämlich, dass es mehr als elfmal wahrscheinlicher ist, dass die Halshautprobe aus einem Großbetrieb Salmonella–positiv wird als die Halshautprobe aus einem Kleinbetrieb. Die Tatsache, dass sich auf dem Hof auch ein anderes Geflügel befindet, hat kein signifikanter Einfluss ($p = 0,823$) auf das Salmonella–Vorkommen. Die

möglich erwartende Kreuzkontamination durch andere Tiere hat sich nicht gezeigt.
Auch zwischen Proben, die aus einem desinfizierten oder nicht desinfiziertem Betrieb stammen, scheint es keinen signifikanten Unterschied ($p = 0,139$) zu geben.
Die Entfernung des Betriebes von einem anderem Hühnerstand zeigt kein Effekt auf die Salmonella–Prävalenz, was auch bei der bivariaten Analyse zu sehen war.
Gegen die Erwartung wurde ein signifikanter Unterschied zwischen den Proben, die im Sommer und in anderen Jahreszeiten entnommen wurden, gefunden. Bei den Proben vom Sommer sinkt die "Chance" Salmonella–positiv zu werden im Vergleich zu den Proben von dem Rest des Jahres um den Faktor 0,61. Dieses Ergebnis bestätigt die Theorie, dass sich die Salmonella Bakterien bei höhere Temperaturen nicht besser vermehren können.
Erstaunlich ist das Ergebnis hinsichtlich der Hygienegruppe. Der Hygienezustand des Betriebes hat zwar ein signifikanter Effekt auf das Salmonella-Vorkommen, aber genau im Gegenteil zur Erwartung. Die Chance Salmonella–positiv zu sein bei einer Poolprobe, die aus dem Betrieb, der in die 2. (befriedigende) Hygienegruppe klassifiziert wurde, stammt, ist mehr als zweimal kleiner als bei den Poolproben, die in dem Betrieb aus der 1. (guten) Hygienegruppe entnommen wurden. Das kann möglicherweise an der falschen Klassifizierung durch die Umkodierung der Variable "Hygieneklasse" liegen. Dabei wurden große Betriebe mit der Bewertung "gut" (Hygieneklasse I) und "zufriedenstellend" (Hygieneklasse II) in die Hygienegruppe 1 zusammengefügt. In der zweiten Hygieneklasse konnten sich viele Proben mit einem positivem Salmonella–Ergebnis befinden und damit das Regressionkoeffizient beeinflussen.
Die Variable "Region" zeigt auch einen signifikanten Effekt auf die Salmonella–Prävalenz. Zwischen den Regionen A und E (Referenzkategorie) scheint es einen signifikanten ($p < 0,1$) wenn auch geringen Effekt zu geben. Die Wahrscheinlichkeit, dass eine Poolprobe aus der Region A Salmonella–positiv wird, ist 0,67 mal geringer als bei einer Poolprobe aus der Region E. Poolproben aus anderen Regionen (B, C und D) haben eine noch geringere Chance Salmonella nachzuweisen als die Poolproben aus der Region E ($p < 0,0001$). Die Erklärung dafür könnte sein, dass in den Regionen A und E mehrere Großbetriebe untersucht wurden oder auch, dass in die Regionen A und E Küken aus einer Zuchtstation mit höherer Salmonella–Prävalenz geliefert wurden.

Grafik, die in der Abbildung (1) erscheint, zeigt auf der Horizontalachse die Linearkombination $\eta = \mathbf{x_i}^\top \boldsymbol{\beta}$ und auf der Vertikalachse die beobachteten $\mathbf{y}$–

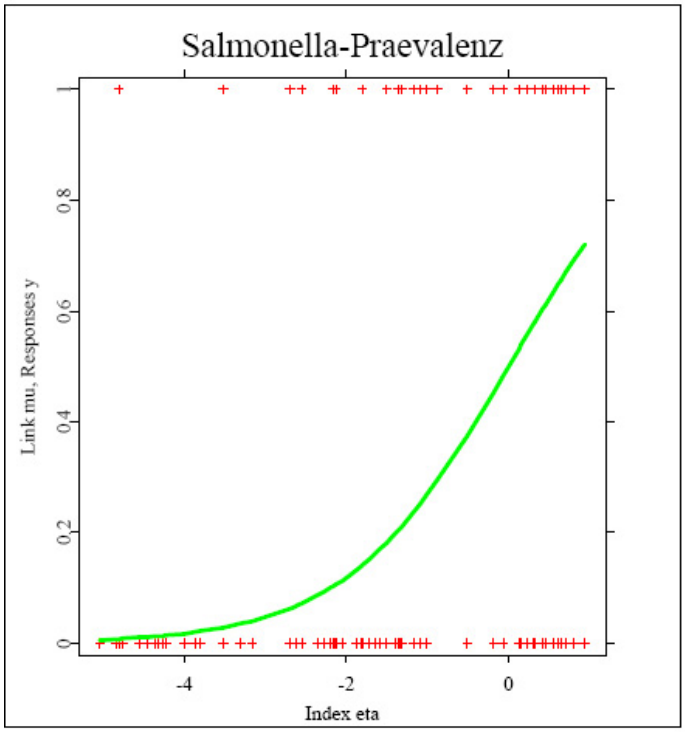

Abbildung 1: Schätzungsergebnisse des Logit–Modells und die Linkfunktion mit beobachteten Werten.

LogitModell.xpl

Werte als roten Kreuze. Die Linkfunktion $G(\mathbf{x_i}^\top \boldsymbol{\beta})$ wird als die grüne Linie dargestellt.

df	965
Deviance	1015.4669
Log-likelihood	-507.7335
Pearson's χ^2	1125.6191
R^2	0.2021
Adj. R^2	0.1939
AIC	1037.4669
BIC	1091.1850
Kovar. Gruppen	63

Tabelle 10: Statistiken für das Logit–Modell

LogitModell.xpl

In der Tabelle (10) sind die berechneten Statistiken für Beurteilung der Modellanpassung enthalten. Die Anzahl der Freihatsgrade bekommt man als die Anzahl der beobachteten Responses ohne die Anzahl der geschätzten Parameter: $f = n - J = 976 - 11 = 965$.
Der maximale Wert der log–Likelihood–Funktion an der Stelle $\hat{\pi}$ ist gleich $-507,73$ und damit die Deviance $D = 1015,47$. Wie bereits erwähnt im Abschnitt (3.5), folgt die Deviance approximativ einer χ^2–Verteilung mit $n - J$ Freiheitsgraden und folglich kann $D \cong n - J = 965$ grob erwartet werden.
Eine feinere Beurteilung der Güte der Anpassung liefert die Pearson' Statistik $X^2 = 1125,62$. Der kritische Wert der χ^2–Verteilung mit 965 Freiheitsgraden und mit dem Signifikanzniveau $\alpha = 0,05$ ist $\chi^2_{0,95;965} = 1038,4$. Die Nullhypothese, dass das angegebene Logit Modell das angemessene Modell ist, wird jedoch verworfen, weil $X^2 = 1125,6191 > 1038,4$. Dies deutet auf eine mögliche Misspezifikation des Modells bzw. fehlenden erklärenden Variablen im Modell. Da der Datensatz vorgegeben wurde, ist leider die Einnahme von weiteren Variablen nicht möglich.
Vergleicht man jedoch das angegebene Modell (M_1) mit $J = 11$ zu schätzenden Parametern gegen das Modell (M_0) mit einer Konstante, bekommt man $\Delta D = 257,28$. Der maximale Wert der log–likelihood–Funktion für das Modell M_0 wird folgendermaßen berechnet.

Das Modell ohne erklärenden Variablen lautet gemäß (5):

$$g(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \beta_0 \ ,$$

davon ergibt sich für den Parameter:

$$\hat{\beta}_0 = \log\left(\frac{\sum\limits_{k=1}^{n} y_k}{n - \sum\limits_{k=1}^{n} y_k}\right) = \log\left(\frac{349}{976-349}\right) = -0,5859 \ .$$

Die geschätzte Wahrscheinlichkeit π_k ist somit konstant für alle n und entspricht der relativen Häufigkeit des Salmonella–Vorkommens:

$$\hat{\pi}_k = \frac{\sum\limits_{k=1}^{n} y_k}{n} = \frac{349}{976} = 0,3576 \ .$$

Der Wert der log–likelihood–Funktion an der Stelle $\hat{\beta}_0 = -0,5859$ berechnet man aus (8):

$$l_0(\hat{\beta}_0, \mathbf{y}) = \hat{\beta}_0 \sum_{k=1}^{n} y_k - n \cdot \log\left(\frac{n}{n - \sum\limits_{k=1}^{n} y_k}\right) = -636,37 \ .$$

Für die Deviance ergibt sich:

$$\Delta D = 2[l_1(\hat{\boldsymbol{\pi}}; \mathbf{y}) - l_0(\hat{\boldsymbol{\pi}}; \mathbf{y})] = 2(-507,7335 + 636,37) = 257,28 \ .$$

Aus der Tabelle der χ^2–Verteilung findet man für $f = 11 - 1 = 10$ Freiheitsgraden und das vorgegebene Signifikanzniveau $\alpha = 0,05$ den kritischen Wert $\chi^2_{0,05;11} = 4,57$. Die Nullhypothese mit dem Modell M_0 wird auf dem 5%–Niveau abgelehnt, da $\Delta D = 257,28 > \chi^2_{0,05;11} = 4,57$ ist. Das bedeutet, dass das angegebene Logit Modell mit den 7 erklärenden Variablen besser ist als das Modell mit einer Konstante. Die Betriebseigenschaften haben einen systematischen Einfluss auf die Salmonella–Prävalenz.

Um erklärende Variable mit einem bedeutenden Einfluss auf die response Variable herauszufinden, wird eine Modell–Selektion–Methode mit AIC (Akaike Information Criterion) als Kriterium für die Aufnahme verwendet. Das AIC

Kriterium ist nichts anderes als die log–likelihood–Funktion mit einer Korrektur bezüglich des Anzahl der geschätzten Parameter:

$$\text{AIC} = -2l(\hat{\boldsymbol{\pi}}; \mathbf{y}) + 2J \ , \tag{11}$$

wobei $l(\hat{\pi}; \mathbf{y})$ der maximale Wert der log–likelihood–Funktion an der Stelle der geschätzten $\hat{\pi}$ ist und J der Anzahl der geschätzten Parameter bezeichnet. AIC ist ein relatives Maß für den Auswahl des besten Submodells. Für die Modell–Selektion wird das XploRe Quantlet `glmselect` verwendet, es wird die Option "aic" für das AIC Selektion–Kriterium gewählt. Die Ergebnisse sind in der Tabelle (11) zusammengefasst.

Bestes Modell	$\hat{\beta}_i$
Konstante	-3.0016
Betriebsgröße	2.3844
Schädlingsbekämpfung	0.2755
Zeitraum	-0.49735
Hygienegruppe	-0.7942
Region A	-0.4452
Region B	-1.7699
Region C	-2.3030
Region D	-2.2189
df	967
Deviance	1017.40
Log-likelihood	-508.71
Pearson's χ^2	1154.80
R^2	0.20
Adj. R^2	0.19
AIC	1035.40
BIC	1079.40
Kovar. Gruppen	63

Tabelle 11: Statistiken für das beste Modell

`LogitModell.xpl`

Es wird ein Modell mit Konstante und 5 erklärenden Variablen (Betriebsgröße, Schädlingsbekämpfung, Zeitraum der Entnahme, Hygienegruppe und

Region), die einen wesentlichen Einfluss auf die Salmonella–Prävalenz bei Hähnchen ausüben, als das beste Submodell gewählt.
Jedoch ist auch diese Modellanpassung nicht gut, da die Pearson' Statistik $X^2 = 1154,8 > \chi^2_{0,95;967} = 1040,5$. Die Nullhypothese auf eine gute Anpassung des Modells wird also abgelehnt.

5 Zusammenfassung und Ausblick

Das Vorkommen von Salmonellen bei Nutzgeflügel stellt bei der Gewinnung salmonellenfreier Lebensmittel ein erhebliches gesundheitliches und ökonomisches Problem dar. Fleisch von Hähnchen ist nach Angaben in der Literatur zu einem hohen Prozentsatz mit Salmonellen kontaminiert.
Diese Arbeit beschäftigt sich mit dem Einfluss verschiedener technischen und hygienischen Eigenschaften der Geflügelmastbetrieben in Deutschland auf die Salmonella–Prävalenz bei Hähnchen.
Dazu wurde eine logistische Regression mit sieben erklärenden Variablen (Eigenschaften des Betriebes) und mit einer abhängigen Variable Salmonella–Vorkommen verwendet. Für die Schätzung der unbekannten Parameter und Berechnung der Modell–Statistiken wurde das statistische Programm XploRe benutzt.
Innerhalb dieser Untersuchung wurde gezeigt, dass der größte Einfluss auf die Salmonella–Prävalenz die Größe des Betriebes hat. Es wurden noch Zusammenhänge zwischen weiteren Faktoren wie Region in Deutschland, Hygienebedingungen des Betriebes, Zeitraum der Probenentnahme und Durchführung einer Schädlingsbekämpfung und das Vorkommen von Salmonellen gefunden. Jedoch haben die Modell–Statistiken gezeigt, dass die Modelanpassung nicht adequat ist. Dies kann an einer Misspezifikation des Modells oder fehlenden weiteren erklärenden Variablen liegen.
Für die nächste Datenerhebung wäre nützlich noch weitere Informationen über die Betriebe und Hähnchen (wie z.B. Ort der Herkunft von Küken, Art bzw. Herkunft vom Futter) zu sammeln. Bei der Klassifizierung des Hygienestandes sollten die Betriebe nicht geteilt in Groß– und Kleinbetriebe betrachtet werden, sondern eine allgemeine Hygieneklassifizierung für alle Betriebe durchgeführt werden.
Dieser Arbeit sollte ein Teil von einem größeren Projekt, der sich mit einer dynamischen Analyse der Salmonella–Prävalenz beschäftigt, werden. Die Studie sollte eine Entwicklung der Samonella–Prävalenz vom Stall, über Schlachthof, Transportwege bis zur Haushalt und eventueller Infektion bei Menschen modellieren. Eine detaillierte Beschreibung der Methode ist bei Anderson, Kelly and Snary (2000) beschrieben.

6 Anhang

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;
;;; Quantlet bianalyse ;;
;;;;;;;;;;;;;;;;;;;;;;;;;;;
; gebrauchte libraries für die Analyse
library("stats")
library("xplore")
; Daten einlesen
dat = read("C:\Dokumenty\Salmonellen\Daten_xplore.dat")
;
;dim(dat)
;min(dat)
;max(dat)
; nur Halshautproben
dat = paf(dat, dat[,2]==2)
;
;dim(dat)
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; die response Variable y
y = dat[,3]
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; Dummies für die kategoriale Variable REGION
{r,reg}=categorize(dat[,1],"value",5)
;r
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; die erklärenden Variablen (mit Konstante in der ersten
  Spalte)
x = matrix(rows(dat))~dat[,4]~dat[,6:7]~dat[,9:11]~r
xvar ="konst"|"betrgr"|"and_gef"|"schädl"|"entfkat"|"zeitr2"
  |"hyggr"|"region"
yvar ="poolproben"
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; Kontingenztabellen
crosstable(x[,2]~y,xvar[2,]~yvar)
crosstable(x[,3]~y,xvar[3,]~yvar)
crosstable(x[,4]~y,xvar[4,]~yvar)
crosstable(x[,5]~y,xvar[5,]~yvar)
crosstable(x[,6]~y,xvar[6,]~yvar)
crosstable(x[,7]~y,xvar[7,]~yvar)
crosstable(dat[,1]~y,xvar[8,]~yvar)
```

```
;;;;;;;;;;;;;;;;;;;;;;;;;
;;; Quantlet LogitModell ;;
;;;;;;;;;;;;;;;;;;;;;;;;;
; gebrauchte libraries
library("stats")
library("xplore")
library("glm")
; Datensatz
dat = read("C:\Dokumenty\Salmonellen\Daten_xplore.dat")
; dim(dat)
;
dat = paf(dat, dat[,2]==2)     ; nur Halshautproben
; dim(dat)
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; die response Variable y
y = dat[,3]
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; Dummies für die kategoriale Variable Region,
; Region E als Referenzkategorie
{r,reg}=categorize(dat[,1],"value",5)
;r
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; die erklärenden Variablen (mit Konstante
; in der ersten Spalte)
x = matrix(rows(dat))~dat[,4]~dat[,6:7]~dat[,9:11]~r
;
xvar ="konst"|"betrgr"|"and_gef"|"schädl"|"entfkat"|"zeitr2"
  |"hyggr"|"regionA"|"regionB"|"regionC"|"regionD"
;
yvar ="poolproben"
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; logit Modell ;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; Schätzung des Modells
{b,bv,stat}=glmest("bilo",x,y)
; Variablen-Beschreibung
opt=glmopt("title","Salmonella-Prävalenz", "xvars", xvar)
; Wald Test
wald = (stat.tvalue)^2
; Tabellen-Output
glmtable("bilo",x,y,b,bv,stat,opt)
; Display-Output mit Grafik
glmout("bilo",x,y,b,bv,stat,opt)
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; Model selection (Suche nach dem besten Modell)

opt=glmopt("crit","aic")          ; AIC Kriterium
;opt=glmopt("fix",8:11,opt)       ; immer Variablen 8-11
```

```
modsel=glmselect("bilo",x,y)     ; selektiert das beste
  Submodell
modsel
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; Modell mit 5 erklärenden Variablen
;
x2 = x[,1:2]~x[,4]~x[,6:11]
;
```

```
proc(out)=glmtable(code,x,y,b,bv,stat,opt)
; ------------------------------------------------------------
; Library        glm
; ------------------------------------------------------------
; See_also      glmest glmstat glmopt
; ------------------------------------------------------------
; Macro        glmtable
; ------------------------------------------------------------
; Description  glmout creates a nice table for GLM.
; ------------------------------------------------------------
; Link    ../tutorials/glmstart.html Tutorial: GLM in XploRe
; ------------------------------------------------------------
; Keywords      GLM, Generalized Linear Model
; ------------------------------------------------------------
; Usage        glmtable(code,x,y,b,bv,stat{,opt})
; Input
;   Parameter    code
;   Definition   text string, the short code for the model
;                (e.g."bilo" for logit or "noid" for
;                ordinary PLM).
;   Parameter    x
;   Definition       n x p  matrix, the predictor variables.
;   Parameter    y
;   Definition       n x 1  vector, the response variables.
;   Parameter    b
;   Definition       p x 1, estimated coefficients.
;   Parameter    bv
;   Definition       p x p, estimated covariance of b.
;   Parameter    stat
;   Definition       list, containing statistics produced
;                    by glm functions
;   Parameter    opt
;   Definition       optional, a list with optional input.
;                    "glmopt" can be used to set up this
;                    parameter. The order of the list
;                    elements is not important.Parameters
;                    which are not given are replaced by
;                    defaults (see below).
;   Parameter    opt.pow
;   Definition       scalar, power for power link. If not
;                    given, set to 0 (logarithm).
;   Parameter    opt.nbk
;   Definition       scalar, extra parameter k for negative
;                    binomial distribution. If not given,
;                    set to 1 (geometric distribution).
;   Parameter    opt.xvars
;   Definition       p x 1 string vector, variable names
;                    for the output. Note, that only up
```

```
49 ;                    to 15 characters are used.
50 ;   Parameter    opt.title
51 ;   Definition       string, title for the output.
52 ;                    If not given, a default is set.
53 ; Output
54 ;   Parameter    table
55 ;   Definition       string output, containing estimation
56 ;                    results (estimated coefficients, S.E.,
57 ;                    p-values, t-values and exp(betas))
58 ;                    and modell statistics.
59 ; -------------------------------------------------------------
60 ;   Example    library("glm")
61 ;            ; ==========================
62 ;            ;    simulate data
63 ;            ; ==========================
64 ;               n=100
65 ;               b=1|2
66 ;               p=rows(b)
67 ;               x=2.*uniform(n,p)-1
68 ;               y=( 1./(1+exp(-x*b)).>uniform(n) )
69 ;             ; ==========================
70 ;             ;   GLM fit and table
71 ;             ; ==========================
72 ;               {b,bv,stat}=glmest("bilo",x,y)
73 ;               glmtable("bilo",x,y,b,bv,stat)
74 ; -------------------------------------------------------------
75 ;   Result
76 ;
77 ;[ 1,] "=================================================="
78 ;[ 2,] " GLM Estimation, 'bilo', n=100"
79 ;[ 3,] "=================================================="
80 ;[ 4,] "PARAMETERS    b      SE      t-value  p-value  exp(b) "
81 ;[ 5,] "--------------------------------------------------"
82 ;[ 6,] " b[1]      0.9831   0.4301   2.2856    0.022    2.673
       "
83 ;[ 7,] " b[2]      1.8685   0.4197   4.4518    0.000    6.479
       "
84 ;[ 8,] "=================================================="
85 ;[ 9,] "                  Modell Statistics                "
86 ;[10,] " "
87 ;[11,] " df                                  98            "
88 ;[12,] " Deviance                          107.4288        "
89 ;[13,] " Log-likelihood                    -53.7144        "
90 ;[14,] " Pearson's chi^2                   104.0121        "
91 ;[15,] " R^2                                 0.2169        "
92 ;[16,] " Adj. R^2                            0.2089        "
93 ;[17,] " AIC                               111.4288        "
94 ;[18,] " BIC                               116.6391        "
95 ;[19,] "=================================================="
```

```
;[20,] " iterations                3                          "
;[21,] " distinct obs.           100                          "
;
; ------------------------------------------------------------
;   Author    A. Mysickova, 20050210
; ------------------------------------------------------------
;
; classify our algo
;
  glmmodels=getglobal("glmmodels")
;
  binomial  = sum(code==(glmmodels.binomial)) >0
  nbinomial = sum(code==(glmmodels.nbinomial)) >0
  power     = sum(code==(glmmodels.power)) >0
  twoparfam = sum(code==(glmmodels.twoparfam)) >0
;
; check initial errors
;
  havearray=(rows(dim(x))>2)
  error(havearray,"x must be vector or matrix")
  error(rows(x)!=rows(y),"x and y have different number of
    rows")
  error((rows(dim(y))>1),"y must be a vector")
;
  n=rows(x)
  p=cols(x)
;
  error(rows(dim(b))>1,"b must be vector")
  error(rows(dim(bv))>2,"bv must be matrix")
  error(rows(b)!=p,"b and x have incompatible dimensions")
  error(dim(bv)!=(p|p),"b and bv have incompatible
    dimensions")
;
;  wx=1
;  off=0
  pow=0
  nbk=1
;  nopic=0
;  name="glm"
  xvars=" "
  havexvars=0
;  weights="prior"
;
; now check which optional values have been given
;
  if (exist(opt)>0)

    if (comp(opt,"pow")>0)
      notgood=(exist(opt.pow)!=1)
```

```
143       notgood=notgood || (dim(dim(opt.pow))!=1)
144       warning (notgood>0, "opt.pow was unusable, used
            default =0 instead")
145       if (notgood==0)
146         pow=opt.pow
147       endif
148     endif
149     if (comp(opt,"nbk")>0)
150       notgood=(dim(dim(opt.nbk))!=1)
151       warning (notgood>0, "opt.nbk was unusable, used
            default =1 instead")
152       if (notgood==0)
153         nbk=opt.nbk
154       endif
155     endif
156     if (comp(opt,"xvars")>0)
157       notgood=(exist(opt.xvars)!=2) || (rows(dim(opt.xvars))
            !=1)
158       notgood=notgood || (rows(opt.xvars)!=p)
159       warning (notgood>0, "opt.xvars not consistent with x")
160       if (notgood==0)
161         xvars=opt.xvars
162         havexvars=1
163       endif
164     endif
165     if (comp(opt,"title")>0)
166       notgood=(exist(opt.title)!=2) || (dim(dim(opt.title))!
            =1)
167       warning (notgood>0, "opt.title not a single string,
            used default instead")
168       if (notgood==0)
169         title=opt.title
170       endif
171     endif
172   endif
173 ;
174   titl="GLM Estimation, '"+code+"', n="+string("%1.0f",n)
175   s0="-------------------------------------------------------"
176   s1="======================================================="
177   if (exist(title)>0)
178     txt=    s1
179     txt=txt|(" "+title)
180     txt=txt|s0
181     txt=txt|(" ")
182     txt=txt|(" "+titl)
183     txt=txt|s1
184   else
185     txt=    s1
186     txt=txt|(" "+titl)
```

```
187       txt=txt|s1
188     endif
189   ;
190   ;
191     havenote=0
192     iconst=minind(abs(max(x)-min(x)),2)
193     if (max(x[,iconst])-min(x[,iconst]) ==0)
194       if (havexvars)
195         note = " * constant variable: "+paf(xvars,(1:p)==
            iconst)
196       else
197         note = " * constant variable: "+string("b[%-1.0f]",
            iconst)
198       endif
199       havenote=1
200     endif
201     ;
202   ; txt=txt|(" ")|("Estimates")
203   ; txt=txt|"-------------------------------------------------"
204     txt=txt|" PARAMETERS      b         SE       t-value     p-
        value    exp(b)   "
205     txt=txt|"-------------------------------------------------"
206
207     if (havexvars)
208       labl=xvars+"                    "
209       labl=substr(labl,1,16)
210       blank=string(" ",1:rows(labl))
211       lablnew=substr(labl,1,2)
212       j=1
213       while (j<15)
214         j=j+1
215         sub=substr(labl,j,j+1)
216         if (1-prod(sub==blank))
217         lablnew=lablnew+sub
218         endif
219       endo
220       labl=lablnew
221     else
222       if (rows(b)<10)
223         labl=string(" b[%-1.0f]   ",1:rows(b))
224       else
225         labl=string(" b[%-1.0f]    ",1:9)
226         labl=labl|string(" b[%-1.0f]   ",10:rows(b))
227       endif
228     endif
229     bstr=string(" %10.4f",b)
230     sstr=" "
231     tstr=" "
232     pstr=" "
```

```
233   ;expstr=" "
234   if (exist(stat.serror))
235     sstr=string(" %10.4f",stat.serror)
236   endif
237   if (exist(stat.tvalue))
238     tstr=string(" %10.4f",stat.tvalue)
239   endif
240   if (exist(stat.pvalue))
241     pstr=string(" %10.3f",stat.pvalue)
242   endif
243   expstr=string(" %10.4g    ",exp(b))
244   txt=txt|labl+bstr+sstr+tstr+pstr+expstr
245   ;
246
247   if (havenote)
248     txt=txt|s0|note
249   endif
250 ;
251   if (power)
252   txt=txt|" power = "+string("%8.4g",pow)
253   txt=txt|s1
254   endif
255   if (nbinomial)
256     txt=txt|" k = "+string("%8.4g",nbk)
257   txt=txt|s1
258   endif
259
260   havenote=0
261
262
263   txt=txt|s1|"                  Modell Statistics
                                        "
264   namestat=" "
265     if (comp(stat,"df")>0)
266       if (exist(stat.df)==1)
267         namestat=namestat|(" df                    "+string("
              %12.0f                                 ",stat.df))
268       endif
269     endif
270     if (comp(stat,"deviance")>0)
271       if (exist(stat.deviance)==1)
272         namestat=namestat|(" Deviance              "+string(
            "  %16.4f                            ",stat.deviance
            ))
273       endif
274     endif
275     if (comp(stat,"loglik")>0)
276       if (exist(stat.loglik)==1)
```

```
277         namestat=namestat|(" Log-likelihood         "+string(
              "  %16.4f                          ",stat.loglik))
278       endif
279     endif
280     if (twoparfam)
281     if (comp(stat,"dispersion")>0)
282       if (exist(stat.dispersion)==1)
283         namestat=namestat|(" Dispersion             "+string(
              "  %16.4f                          ",stat.
              dispersion))
284       endif
285     endif
286     endif
287     if (comp(stat,"pearson")>0)
288       if (exist(stat.pearson)==1)
289         namestat=namestat|(" Pearson's chi^2        "+string(
              "  %16.4f                          ",stat.pearson)
              )
290       endif
291     endif
292     if (comp(stat,"r2")>0)
293       if (exist(stat.r2)==1)
294         namestat=namestat|(" R^2                    "+string(
              "  %16.4f                          ",stat.r2))
295       endif
296     endif
297     if (comp(stat,"adr2")>0)
298       if (exist(stat.adr2)==1)
299         namestat=namestat|(" Adj. R^2               "+string(
              "  %16.4f                          ",stat.adr2))
300       endif
301     endif
302     if (comp(stat,"aic")>0)
303       if (exist(stat.aic)==1)
304         namestat=namestat|(" AIC                    "+string(
              "  %16.4f                          ",stat.aic))
305       endif
306     endif
307     if (comp(stat,"bic")>0)
308       if (exist(stat.aic)==1)
309         namestat=namestat|(" BIC                    "+string(
              "  %16.4f                          ",stat.bic))
310       endif
311     endif
312     namestat=namestat|s1
313     if (comp(stat,"it")>0)
314       if (exist(stat.it)==1)
315         if (stat.it>0)
```

```
316             namestat=namestat|(" iterations      "+string("
                  %12.0f                                  ",stat.it
                  ))
317           endif
318         endif
319       endif
320       if (comp(stat,"nr")>0)
321         if (exist(stat.nr)==1)
322           namestat=namestat|(" distinct obs.   "+string("%12.0f
                                                  ",stat.nr))
323         endif
324       endif
325       if (comp(stat,"ret")>0)
326         if (exist(stat.ret)==1)
327           if (stat.ret==-1)
328             note = " ! missing values occurred"
329             havenote=1
330           endif
331           if (stat.ret==1)
332             note = " ! max. number of iterations reached"
333             havenote=1
334           endif
335         endif
336       endif
337       txt=txt|namestat
338       if (havenote)
339         txt=txt|(" ")|note
340       endif
341      out = txt
342    ;exec(name+"out=txt")
343    ;putglobal(name+"out")
344  endp
```

Literatur

Anderson, W., Kelly, L. and Snary, E. (05/2000). *Exposure Assessment of Salmonella spp. in broilers*, preliminary report.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*, second edition, Chapman and Hall, London.

Dobson, A.R. (1990). *An Introduction to Generalized Linear Models*, Chapman and Hall, London.

Ellerbroek, L. and Wichmann-Schauer, H. and Haarmann, M. (2001). *Untersuchungen zum Vorkommen von Salmonellen bei deutschem Nutzgeflügel und Geflügelfleisch*, bgvv Heft, Berlin.

Härdle, W., Klinke, S., and Turlach, B.A. (1995). *XploRe: An Interactive Statistical Computing Environment*, Springer, New York.

Härdle, W., Müller, M., and Klinke, S. (2001). *XploRe Learning Guide*, Springer, Berlin.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, second edition, Chapman and Hall, London.

Rönz, B. (2001). *Generalisierte lineare Modelle*, Vorlesung Skript.