Modeling of salmonella-prevalence in the case of chicken

Alena Myšičková
Wolfgang Härdle

Institute for Statistics and Econometrics
Humboldt-University at Berlin

Outline

Prevalence

”In medical statistics or epidemiology prevalence or basic component describes the frequency of the appearence of a certain disease (or a certain symptom) in a given population.”

This work studies:

prevalence of salmonella infection in the population of german chicken
aim: analysis of the prevalence depending on the properties of the farms

Data

Data from the Bundesinstitut für Risikobewertung, Berlin

L. Ellerbroek, H. Wichmann-Schauer, M. Haarmann:
Analysis of the prevalence of salmonella in the case of German poultry (02/2001)

Time and place of measuring

- 5 regions in Germany (A, B, C, D, E)
- Year 1999
- 66 farms
- 189 flocks

1 flock = whole population in a barn

Data Collection

Farms (populations of mast-poultry) divided into two groups:

large farms: yearly production $\geq$ 20 000 chicken

small farms: yearly production $<$ than 20 0000 chicken

Are there significant differences between large and small farms?

Samples have been taken at three different places:

dirt-samples – taken from the employees´ protection boots after they walked through the barn – whole area of the barn is represented

neck-skin-samples – taken from single chicken during slaughtering

cloaca-pad-samples – taken from the cloaca during slaughtering, only in the case of flocks on large farms

Today´s topic: intermediate results of the analysis of the neck-skin-samples

Pooled Samples

Withdrawn material has been pooled and analysed in five laboratories.
One pooled sample = 5 chicken
Alltogether 976 pooled samples:

	neck-skin-samples
large farms	840
small farms	136
Sum	976

Statistical Analysis

Aim:

- Modeling of Salmonella-Prevalence
- Dependency from other aspects (properties of the farms, regions, . . .)

⇒ Method: Generalized Linear Models

Properties of the Farms

Possible influencing factors:

- size of the farm
- category of hygiene of the farm
- other animals/poultry on the farm
- pest control
- distance to other farms
- week of withdrawal

Model

considers a random variable $Y_i, i = 1, \ldots 976$:

$$Y_i = \begin{cases} 1, & i - \text{th pooled sample salmonella–positive} \\ 0, & i - \text{th pooled sample salmonella–negative} \end{cases}$$

Probability for sample i to be salmonella–positive:

$$\begin{aligned} \pi_i &= \mathrm{P}(Y_i = 1) \\ 1 - \pi_i &= \mathrm{P}(Y_i = 0)\,. \end{aligned}$$

Code conversion of the variables

- period of withdrawal: variable with numerous categories (weeks) $\rightarrow$ variable with two categories (summer; other seasons)
- distance: metric variable $\rightarrow$ variable with two categories (more or less than 1km)
- category of hygiene: large farms – I, II, III; small farms – I, II $\rightarrow$ category of hygiene: large farms: category I + II $\rightarrow$ group 1, category III $\rightarrow$ group 2; small farms: same classification

explanatory variables:

- X_1 : region (5 categories)
- X_2 : size of the farm (1 = small farms, 2 = large farms)
- X_3 : other poultry on the farm (1 = yes, 2 = no)
- X_4 : active pest control (1 = yes, 2 = no)
- X_5 : distance to the next chicken farm ($1 = < 1000$m, $2 = \geq 1000$ m)
- X_6 : period of withdrawal (1 = spring/fall, 2 = summer (june – september))
- X_7 : category of hygiene (2 categories)

Question:

To which extent is salmonella-prevalence influenced by the explanatory variables?

Which of the factors do have significant influence on the prevalence?

Answer by means of the logit-model

Logit-Model

Regression Model of the form:

$$\mathsf{E}\,Y = P(Y=1) = \pi = G(X^\top \beta)$$

- $\mathsf{E}\,Y$ – average of the response-variable Y
- distribution of the response-variable Y out of the family of exponential distributions (Bernoulli-distribution)
- X – vector of explanatory variables

- β – vector of unknown parameters
- $G(\bullet)$ – known **link-funktion**

using the logistic distribution function as **link-funtion**:

$$G(X_i^\top \beta) = \frac{e^{X_i^\top \beta}}{1 + e^{X_i^\top \beta}} = \pi_i$$

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

The inverse of the link-funktion:

$$X_i^\top \beta = log\left(\frac{\pi_i}{1-\pi_i}\right) = g(\pi_i)$$

$\frac{\pi_i}{1-\pi_i}$ = odds (chance) of success
$log\left(\frac{\pi_i}{1-\pi_i}\right)$ = log odds

Interpretation of the Parameters

non-linear correlation between π_i and $X_i^\top \beta$

- $\hat{\beta}_i$ – change in log odds, only in the direction of the change of the probability of success π_i
- $\exp(\hat{\beta}_i)$ – change of the odds-ratio, if X_i increases by one unit given all other X-variables being fixed then they change by a multiplicator $\exp(\hat{\beta}_i)$

Estimation of the GLM model by ML method

log–likelihood–function:

$$l(\pi, \mathbf{y}) = \log f(\mathbf{y}, \theta) = \sum_k \log f_k(y_k, \theta_k) \ ,$$

$f(\mathbf{y}, \theta)$ – density function of $\mathbf{y}$ for a fixed parameter θ.

Use the link function $G(X_i^\top \beta)$, which replaces π by β.
maximize $l(\beta, y)$:

- $\frac{\partial l}{\partial \beta_i} \stackrel{!}{=} 0$
- solve the non-linear equation iterative with the Fisher Scoring Method

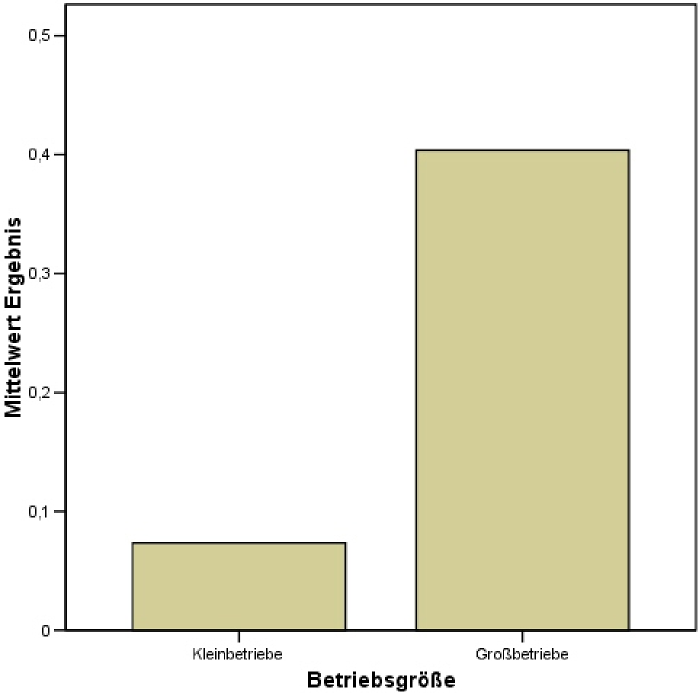
in the logit–model:

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

Results

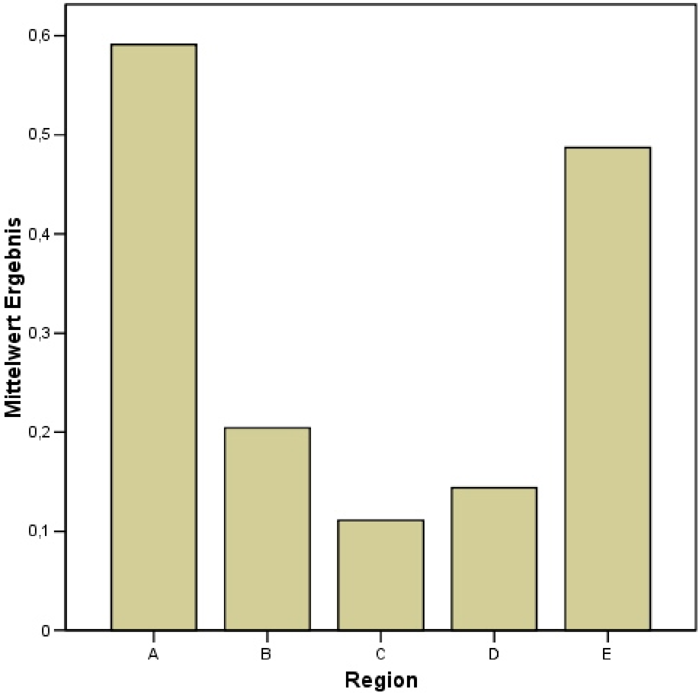
bivariate analysis of explanatory variables – tables and graphs
X_2 = size of farm:

	pooled samples		
	negative	positive	sum
small farms	126	10	136
	92.6%	**7.4%**	100%
large farms	501	339	840
	59.6%	**40.4%**	100%


Mittelwert Ergebnis
0,5
0,4
0,3
0,2
0,1
0
Kleinbetriebe
Großbetriebe
Betriebsgröße

$X_1 =$ region:

	pooled samples		
	negative	positive	sum
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%



0,6
0,5
0,4
0,3
0,2
0,1
0
Mittelwert Ergebnis
A
B
C
D
E
Region

X_6 = period of withdrawal:

	pooled samples		
	negative	positive	sum
spring/fall	265	173	438
	60,5%	$\mathbf{39,5}\%$	100,0%
summer	362	176	538
	67,3%	$\mathbf{32,7}\%$	100,0%

X_7 = category of hygiene:

	pooled samples		
	negative	positive	sum
category 1	359	273	632
	56,8%	**43,2%**	100,0%
category 2	268	76	344
	77,9%	**22,1%**	100,0%

logit-model with 7 exogeneous variables

variable	$\hat{\beta}_i$	s.e.	p–value	$exp(\hat{\beta}_i)$
constant	-2.89	1.05	0.006	0.06
region E (Ref.)			0.000	1
region(A)	-0.40	0.24	0.090	0.67
region(B)	-1.62	0.29	0.000	0.20
region(C)	-2.37	0.50	0.000	0.01
region(D)	-2.21	0.27	0.000	0.11
size of farm	2.44	0.38	0.000	11.46
other poultry	-0.08	0.38	0.823	0.92
pest control	0.29	0.20	0.140	1.34
distance	0.23	0.18	0.201	1.26
period of w.	-0.49	0.17	0.003	0.62
cat. of hyg.	-0.81	0.25	0.001	0.45

logit–model with 4 exogeneous variables (using LR-backwards-selective method)

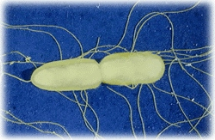
variable	$\hat{\beta}_i$	s.e.	p–value	$exp(\hat{\beta}_i)$
constant	-2.50	0.80	0.002	1,26
region E (Ref.)			0.000	1
region(A)	-0.37	0.22	0.104	0,67
region(B)	-1.67	0.25	0.000	0,20
region(C)	-2.45	0.48	0.000	0,09
region(D)	-2.09	0.25	0.000	0,11
size of farm	2.36	0.36	0.000	11,46
period of w.	-0.51	0.16	0.002	0,92
cat. of hyg.	-0.66	0.22	0.003	1,34

Conclusion

4 factors with significant influence on prevalence:

- size of farm, region, category of hygiene, period of withdrawal

salmonella prevalence only influenced by some of the factors (about 30% of the variance explained by the model) → salmonella bacteria to a certain extent always appear



Outlook

This work in the context of a bigger project:
Dynamic analysis of salmonella prevalence in

- barn
- slaughterhouse
- during transport
- household (kitchen)
- human being