STATISTICAL ANALYSIS OF CAMPYLOBACTER RISK FACTORS AT BROILER FARMS

Weyermair, K.[1], Pless, P.[2], Matt, M.[3]

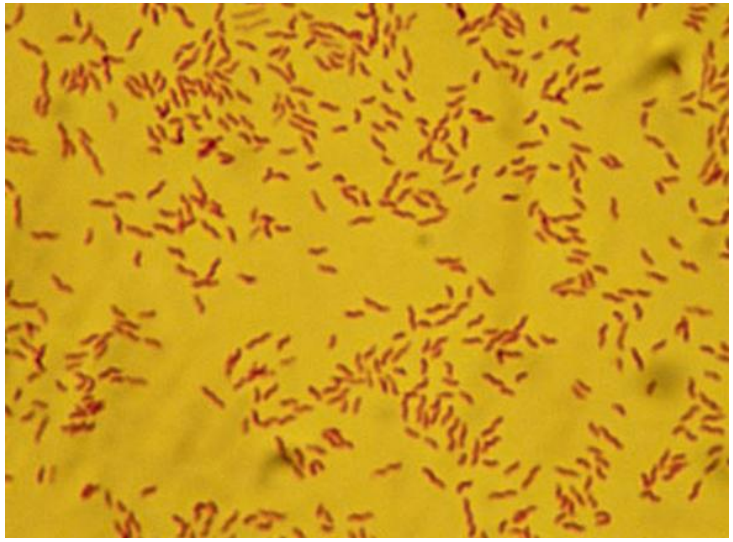
[1] Department Data, Statistics, Risk Assessment, Austrian Agency for Health and Food Safety, Graz, Austria
[2] Department of Veterinary Administration, Styrian Government, Graz, Austria
[3] Department Data, Statistics, Risk Assessment, Austrian Agency for Health and Food Safety, Innsbruck, Austria

Karin Weyermair
Department Data, Statistics, Risk Assessment

Österreichische Statistiktage
Graz, 8.9.2011

Campylobacter (C.)

- Bacterium
- Most common cause of bacterial foodborne disease in Austria
- Mainly related to poultry



Microscopic picture of Campylobacter bacteria, magnification 1:1000

Investigation of C. risk factors on broiler farms

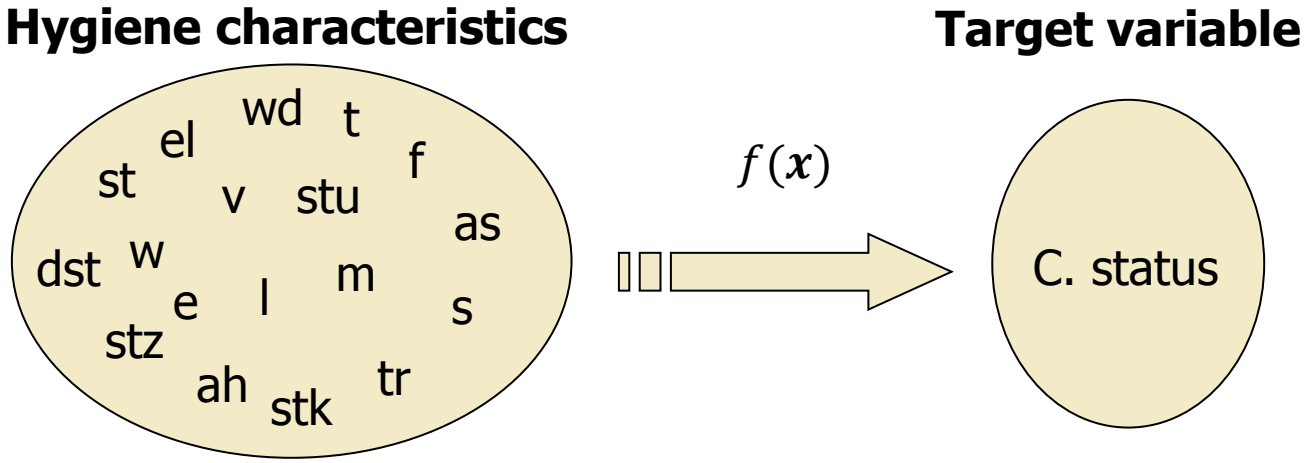
- Audits by official veterinarians
- 53 Austrian flocks
- 18 hygiene characteristics, rated with values 1 to 5
- Determination of C. status of the flock by faecal samples at slaughter over several periods, rated with values 1 to 5

Data description

Rating frequencies		"low hygiene" to "high hygiene"				
Variable description	**Abb.**	**1**	**2**	**3**	**4**	**5**
Number of stables	st	1	9	11	14	18
Water supply	w	1	7	13	5	27
Manure storage	m	4	7	18	15	9
Other animals at the farm	t	7	15	22	6	3
Stable surroundings	stu	1	9	22	19	2
Structural stable conditions	stz	0	11	24	17	1
Ventilation system	l	0	6	34	11	2
Pest security	s	1	15	16	21	0
Chicken watering systems	tr	0	2	23	26	2
Feeding	f	3	7	14	25	4
Stable cloth	stk	4	7	29	12	1
Hygiene sluice	v	2	12	26	13	0
Cleaning and disinfection equipment	wd	2	12	28	11	0
Litter type	e	0	2	38	10	3
Litter storage	el	0	4	30	18	1
Stable cleaning	dst	0	4	29	19	1
Collection system	as	0	16	24	9	4
Thinning frequency	ah	17	6	19	1	10
		"always C. positive" to "never C. positive"				
		1	**2**	**3**	**4**	**5**
Campylobacter status of farm	faeces	15	12	12	10	4

Analysis idea

- Model the influence of the 18 hygiene characteristics on the *C.* status



- Ordinal logistic regression
- Correlation structure among hygiene characteristics?

Correlation structure

Spearman Correlation Coefficients, N = 53																		
	st	**w**	**m**	**t**	**stu**	**stz**	**l**	**s**	**tr**	**f**	**stk**	**v**	**wd**	**e**	**el**	**dst**	**as**	**ah**
st	1	0.00	0.12	0.03	0.03	-0.02	0.02	-0.10	-0.13	-0.12	-0.02	0.12	0.12	-0.21	0.03	-0.27	0.33	**0.60**
w		1	-0.06	0.01	0.03	0.19	0.14	0.04	-0.20	0.06	0.27	0.08	0.07	0.08	0.13	-0.04	0.01	0.13
m			1	0.15	-0.05	0.07	0.01	0.12	-0.05	-0.06	-0.05	-0.10	0.03	0.35	0.28	0.12	0.11	0.11
t				1	**0.56**	0.36	0.29	0.37	0.04	0.09	0.21	0.16	0.21	0.01	0.13	0.24	-0.04	0.05
stu					1	**0.51**	**0.51**	**0.50**	0.13	0.44	0.49	0.43	0.47	0.16	0.26	0.34	0.21	0.15
stz						1	**0.74**	**0.73**	0.22	**0.62**	0.49	**0.60**	0.45	0.15	0.48	**0.58**	-0.01	-0.24
l							1	**0.65**	0.18	**0.62**	0.37	0.49	0.30	0.17	0.36	0.44	-0.08	-0.20
s								1	0.33	0.46	0.32	**0.50**	0.32	0.11	**0.50**	**0.60**	0.03	-0.23
tr									1	0.35	0.16	0.31	0.17	0.00	0.30	0.22	-0.07	-0.13
f										1	0.46	**0.50**	0.37	0.13	0.38	0.36	-0.06	-0.30
stk											1	**0.67**	**0.66**	0.23	0.36	0.35	0.11	-0.06
v												1	**0.53**	-0.02	0.41	0.48	0.13	0.03
wd													1	0.10	0.30	0.28	0.23	0.02
e														1	0.29	0.27	-0.01	-0.11
el															1	0.39	-0.04	-0.14
dst																1	0.08	-0.24
as																	1	0.36
ah																		1

Yellow background: p-value <0.05 **Bold**: correlation > 0.5

Correlated variables

- Hygiene characteristics are correlated
- E.g. stu (stable surroundings) correlates ($\rho > 0.5$) with
  - t (other animals at the farm)
  - stz (structural stable condition)
  - l (ventilation system)
  - s (pest security)
- Effective dimension of the data set?
- Variable bundles?

Principal component analysis (PCA)

- Orthogonal transformation of the original set of variables in a new set of uncorrelated variables
- Principal components: $\mathbf{F} = (\boldsymbol{f_1, \ldots, f_p})$

$$\mathbf{F} = \mathbf{ZT\Lambda}^{-1/2} = \mathbf{ZR^{-1}L}$$

  - Standardized data matrix: $\mathbf{Z} = (z_{ij}), i = 1, \ldots, 53; j = 1, \ldots, 18$

$$z_{ij} = \frac{y_{ij} - \overline{y}_j}{\sqrt{n-1}s_j}$$

  - Empirical correlation matrix: $\mathbf{R} = \mathbf{Z'Z} = \mathbf{LL'}$
  - Eigenvalues of $\mathbf{R}$: $\mathbf{\Lambda} = diag\{\lambda_1, \ldots, \lambda_p\}$ $\quad \lambda_1 \geq \lambda_2 \ldots \geq \lambda_p$
  - Eigenvectors of $\mathbf{R}$: $\mathbf{T} = \mathbf{Eigenvec(R)}$
  - Matrix of loadings: $\mathbf{L} = \mathbf{T\Lambda}^{1/2}$

PCA: Dimension reduction

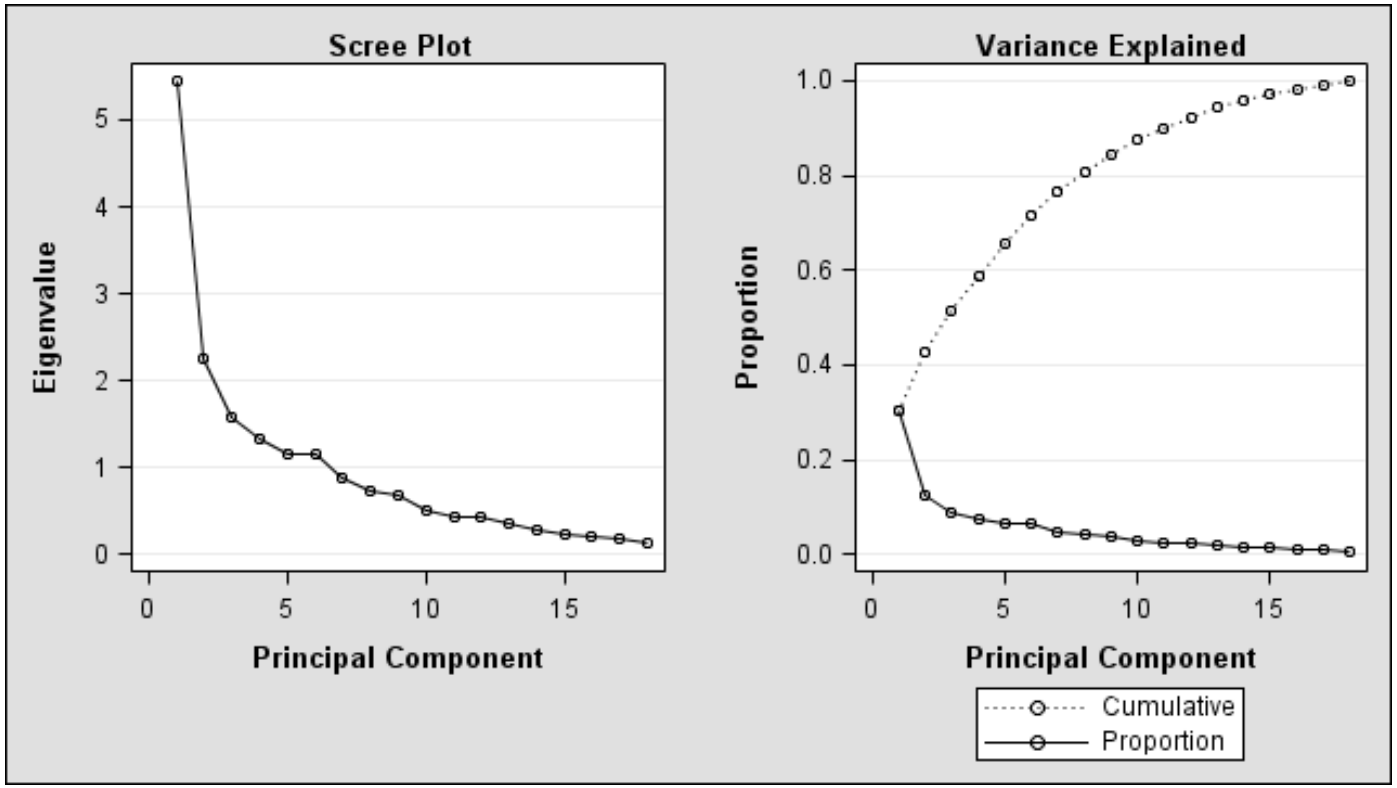
- Dimension reduction while retaining most of the variability in the data
- Full decomposition of $\mathbf{Z}$: $\mathbf{Z} = \mathbf{F}\mathbf{L}'$
- Extraction of first k components: $\mathbf{F}^{(\mathrm{k})} = \mathbf{Z}\,\mathbf{T}^{(\mathrm{k})}\boldsymbol{\Lambda}^{-1/2(\mathrm{k})}$

$$\Rightarrow \mathbf{Z} = \mathbf{F}^{(\mathrm{k})}\mathbf{L}'^{(\mathrm{k})} + \mathbf{E}^{(\mathrm{k})}$$

- Total variability: $v = \mathrm{tr}(\mathbf{R})$
- Explained variability by $\mathbf{F}^{(\mathrm{k})}$: $(\lambda_1 + \cdots + \lambda_k)/v$

PCA: Dimension reduction

- 6 Eigenvalues of correlation matrix > 1 (Eigenvalue criterium)
- 72% variance explained with 6 components
- Choice: $k = 6$



PCA: Interpretation

- Meaning of 6 „new" variables?
- Loadings describe relation between items and factors
- Interpretation of loadings in $\mathbf{L}^{(k)}$ difficult
- Varimax rotation
  - Orthogonal rotation matrix: $\mathbf{M} = \left(m_{ij}\right), i, j = 1, \ldots, k;\ \mathbf{M}\mathbf{M}' = \mathbf{I}$

$$\tilde{\mathbf{L}} = \mathbf{L}^{(k)}\mathbf{M},\ \tilde{\mathbf{F}} = \mathbf{F}^{(k)}\mathbf{M},$$

$$\mathbf{Z} = \tilde{\mathbf{F}}\tilde{\mathbf{L}}' + \mathbf{E}^{(k)} = \mathbf{F}^{(k)}\mathbf{M}\mathbf{M}'\mathbf{L}'^{(k)} + \mathbf{E}^{(k)} = \mathbf{F}^{(k)}\mathbf{L}'^{(k)} + \mathbf{E}^{(k)}$$

  - Each original variable tends to be associated with one factor and each factor represents only a small number of variables
  - Leads to a plausible interpretation of components

PCA: Interpretation

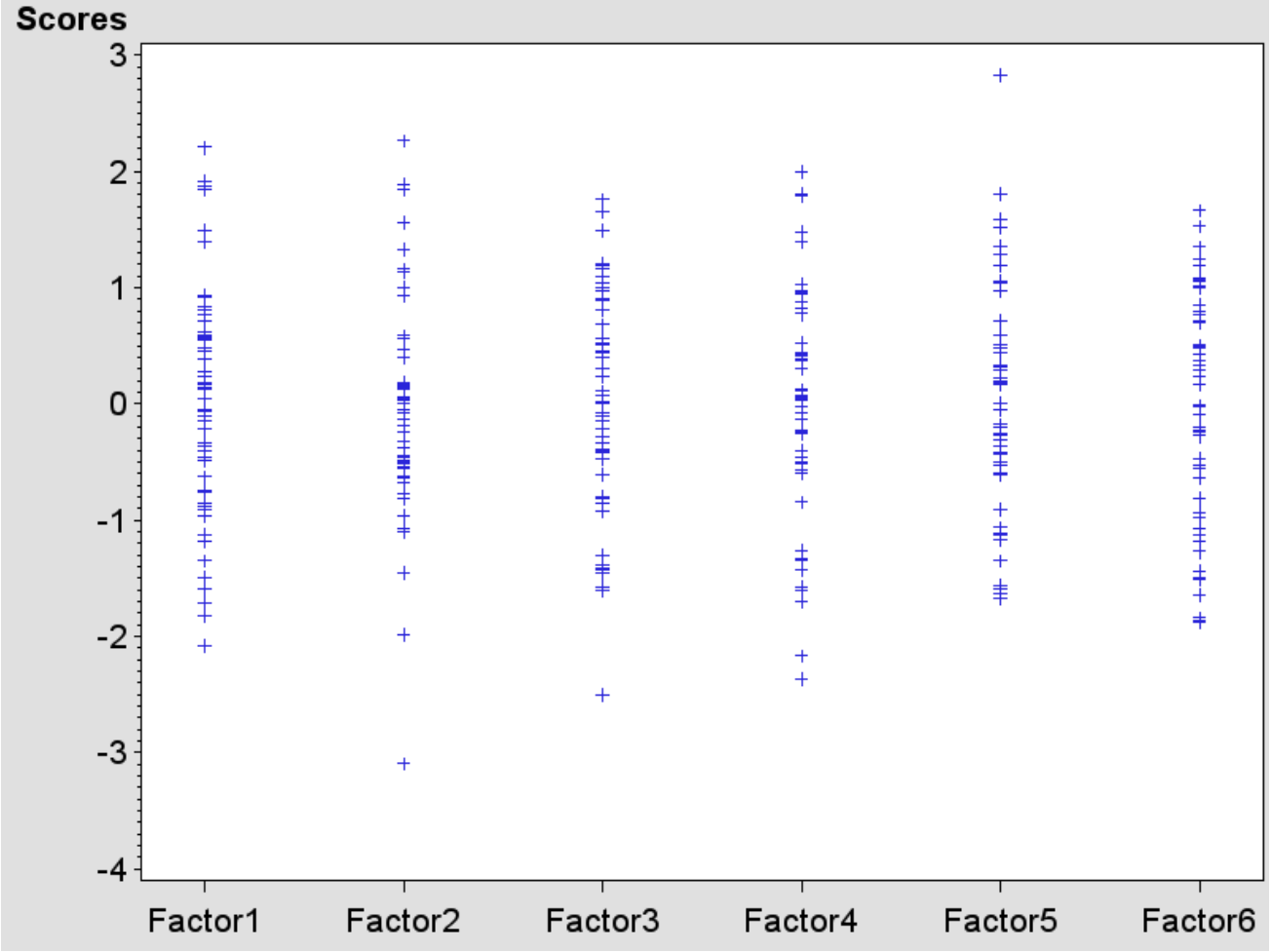
Rotated Factor Pattern						
	L1	**L2**	**L3**	**L4**	**L5**	**L6**
st	0.06	0.05	0.91	-0.01	0.00	0.04
w	0.20	0.04	0.10	-0.11	0.00	0.87
m	0.00	-0.07	0.19	0.14	0.83	-0.08
t	0.16	0.00	0.06	0.85	0.05	-0.03
stu	0.34	0.46	0.04	0.63	-0.08	0.01
stz	0.80	0.16	-0.10	0.32	0.11	0.15
l	0.78	0.04	-0.03	0.31	0.02	0.17
s	0.72	0.02	-0.09	0.39	0.23	-0.06
tr	0.52	-0.01	-0.07	-0.18	-0.06	-0.60
f	0.73	0.17	-0.20	-0.01	-0.06	-0.07
stk	0.41	0.74	-0.09	-0.02	0.05	0.19
v	0.70	0.46	0.06	0.05	-0.13	-0.03
wd	0.33	0.75	0.06	0.04	0.05	-0.04
e	0.03	0.19	-0.32	-0.06	0.71	0.17
el	0.61	0.13	0.05	-0.16	0.53	-0.03
dst	0.43	0.28	-0.40	0.32	0.30	-0.10
as	-0.19	0.65	0.26	0.15	0.10	-0.01
ah	-0.27	0.22	0.78	0.11	0.01	0.11

Coloured background: |Loading| >0.5

- L1: stz, l, s, tr, f, v, el➔ „Building structure and stable management "
- L2: stk, wd, as ➔ „Hygiene on the farm"
- L3: st, ah➔ „Size"
- L4: t, stu ➔ „Surrounding"
- L5: m, e, el ➔ „Litter"
- L6: w, tr ➔ „Watering"

PCA: „New" variables

- Scores: $\tilde{\mathbf{F}} = \mathbf{F}^{(k)}\mathbf{M} = \mathbf{Z}\underbrace{\mathbf{T}^{(k)}\boldsymbol{\Lambda}^{-1/2(k)}\mathbf{M}}_{\textbf{std.scoring coeff.}}$



Ordinal logistic regression

- M1: „Score model"
- Target variable $\mathbf{Y}$: C.status
- Predictor variables: $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_6$
- Cumulative model: $p_i = Pr(Y \leq i|\boldsymbol{f}), i = 1, \ldots, 5$

$$g(p_i) = \alpha_i + \beta_1 \boldsymbol{f}_1 + \cdots + \beta_6 \boldsymbol{f}_6$$

- Link function $g()$: $logit(p_i) \equiv log\left(\frac{p_i}{1-p_i}\right)$

Results M1 „Score model“

- Stepwise variable selection: f_1 (management), f_2 (hygiene), f_3 (size), f_4 (surrounding) remain in the model
- Assumption of parallel lines not violated (p=0.38)
- Interpretation: hygiene f_2 increased by 1 results in reduction of odds $\frac{p_i}{1-p_i}$ by $exp(-1.12) = 0.32$
- Most important hygiene component followed by size, surrounding, management

Analysis of ML-Estimates			
Parameter		**Estimate**	**Pr > ChiSq**
Intercept	**1**	-1.4237	0.0001
Intercept	**2**	-0.0705	0.8306
Intercept	**3**	1.3561	0.0004
Intercept	**4**	3.4194	<.0001
f1		-0.5700	0.0333
f2		-1.1212	0.0002
f3		-0.7621	0.0058
f4		-0.6263	0.0230

Interest of farmers

- What is the effect on the C. status,
  if one hygiene characteristic
  e.g. stu (stable surroundings) is improved to the next level?

  - Simpler model?
    Idea: simple sum scales of identified variable bundles

  - Recalculation of coefficients for the original variables from model equation

Transformation model equation

- Reversing linear transformations of original data in PCA (standardization, rotation, scoring)

$$g(p_i) = \alpha_i + \beta_1 \boldsymbol{f}_1 + \cdots + \beta_4 \boldsymbol{f}_4$$

$$= \widetilde{\alpha_i} + \widetilde{\beta_1} \boldsymbol{st} + \cdots + \widetilde{\beta_{18}} \boldsymbol{ah}$$

- Coefficients are linear combinations of scoring coefficients and regression parameters scaled by standard deviation of the original variable

$$\widetilde{\beta_j} = \frac{\beta_1 s_{j1} - \beta_2 s_{j2} - \beta_3 s_{j3} - \beta_4 s_{j4}}{sd(\boldsymbol{y_j})}$$

$$\widetilde{\alpha_i} = \alpha_i - \left(\widetilde{\beta_1}\overline{st} + \cdots + \widetilde{\beta_{18}}\overline{ah}\right)$$

Results of transformation M1

Parameter	Description	Estimate
Intercept 1		7.314
Intercept 2		8.667
Intercept 3		10.094
Intercept 4		12.157
wd	Cleaning and disinfection equipment	-0.478
as	Collection system	-0.459
stu	Stable surroundings	-0.426
v	Hygiene sluice	-0.312
stk	Stable cloth	-0.304
st	Number of stables	-0.297
ah	Thinning frequency	-0.270
t	Other animals at the farm	-0.208
stz	Structural stable conditions	-0.110
l	Ventilation system	-0.083
dst	Stable cleaning	-0.067
s	Pest security	-0.064
m	Manure storage	-0.021
f	Feeding	-0.015
el	Litter storage	-0.011
tr	Chicken watering systems	-0.011
w	Water supply	0.058
e	Litter type	0.187

Simpler model

- Creation of **scales** from items with loadings >0.5 in PCA:
  - „Building structure and stable management“:

    $$v_1 = stz + l + s + tr + f + v + el;$$

  - „Hygiene“: $v_2 = stk + wd + as$;
  - „Size“: $v_3 = st + ah$;
  - „Surrounding“: $v_4 = t + stu$;
  - „Litter“: $v_5 = m + e + el$;
  - „Watering“: $v_6 = w + tr$;

Scales: „New“ variables

- Frequencies of new variables

v1	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Freq									1	1	7	1	1	5	6	6	3	5	5	3	3		3	3					

v2	3	4	5	6	7	8	9	10	11	12	13	14	15
Freq		1	1	2	4	17	13	4	6	1	4		

v3	2	3	4	5	6	7	8	9	10
Freq	1	7	6	5	6	12	6	2	8

v4	2	3	4	5	6	7	8	9	10
Freq		6	5	9	12	12	7	2	

v5	3	4	5	6	7	8	9	10	11	12	13	14	15
Freq				1	5	4	14	8	11	7	1	2	

v6	2	3	4	5	6	7	8	9	10
Freq				4	10	8	19	12	

Reliability analysis

- Cronbach's alpha:

$$\alpha = \frac{k}{k-1} \frac{\sum_{i \neq j} \mathrm{Cov}(Y_i Y_j)}{\mathrm{Var}(Y_1 + \cdots + Y_k)}$$

Scale	Variables	k	Cronbach's alpha
v1	stz l s tr f v el	7	0.85
v2	stk wd as	3	0.65
v3	st ah	2	0.75
v4	t stu	2	0.69
v5	m e el	3	0.59
v6	w tr	2	-0.53

- Categorization of variable w (water supply) conceptually wrong?

Ordinal logistic regression

- M2: „Scale model"
- Target variable $\mathbf{Y}$: C.status
- Predictor variables: $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_6$
- Cumulative model: $p_i = Pr(Y \leq i|\boldsymbol{v}), i = 1, \ldots, 5$

$$g(p_i) = \alpha_i + \beta_1 \boldsymbol{v}_1 + \cdots + \beta_6 \boldsymbol{v}_6$$

- Link function $g(): logit(p_i) \equiv log\left(\frac{p_i}{1-p_i}\right)$

Results M2 „Scale model“

- Stepwise variable selection: variables v_1 (management), v_2 (hygiene), v_3 (size) remain in the model
- Assumption of parallel lines (sharply) not violated (p=0.09)
- Interpretation: hygiene v_2 increased by 1 results in reduction of odds $\frac{p_i}{1-p_i}$ by $\exp(-0.36) = 0.70$
- Most important size scale followed by hygiene and management

Analysis of ML-Estimates			
Parameter		**Estimate**	**Pr > ChiSq**
Intercept	**1**	8.7148	<.0001
Intercept	**2**	10.048	<.0001
Intercept	**3**	11.459	<.0001
Intercept	**4**	13.4739	<.0001
v1		-0.1922	0.0147
v2		-0.3626	0.0249
v3		-0.3941	0.0021

Model comparison: effects

Parameter	Description	M1	M2
Intercept 1		7.314	8.715
Intercept 2		8.667	10.048
Intercept 3		10.094	11.459
Intercept 4		12.157	13.474
wd	Cleaning and disinfection equipment	-0.478	-0.363
as	Collection system	-0.459	-0.363
stu	Stable surroundings	-0.426	-
v	Hygiene sluice	-0.312	-0.192
stk	Stable cloth	-0.304	-0.363
st	Number of stables	-0.297	-0.394
ah	Thinning frequency	-0.270	-0.394
t	Other animals at the farm	-0.208	-
stz	Structural stable conditions	-0.110	-0.192
l	Ventilation system	-0.083	-0.192
dst	Stable cleaning	-0.067	-
s	Pest security	-0.064	-0.192
m	Manure storage	-0.021	-
f	Feeding	-0.015	-0.192
el	Litter storage	-0.011	-0.192
tr	Chicken watering systems	-0.011	-0.192
w	Water supply	0.058	-
e	Litter type	0.187	-

Model comparison: fit and prediction

Model Fit Statistics			
Criterion	**Intercept Only**	**M1**	**M2**
AIC	171.192	151.500	150.309
SC	179.073	167.263	164.101
-2 Log L	163.192	135.500	136.309

M1 Score model

	Prediction					
Result	**1**	**2**	**3**	**4**	**5**	
1	11	1	2	1		9%
2	6	2	2	2		
3	3	1	5	3		
4	2		2	6		
5				2	2	11%
	9%				21%	49%

M2 Scale model

	Prediction					
Result	**1**	**2**	**3**	**4**	**5**	
1	11		3	1		11%
2	5	2	3	2		
3	4		5	3		
4	1	2	1	6		
5				3	1	11%
	13%				17%	47%

Scenario farms

- Prediction of scenarios
- Some differences in predicted categories by M1 and M2

Scenario	Abb.	st	w	m	t	stu	stz	l	s	tr	f	stk	v	wd	e	el	dst	as	ah	M1	M2
Reference (mode per variable)	A	5	5	3	3	3	3	3	4	4	4	3	3	3	3	3	3	3	3	**3**	**3**
Reference (average category per variable)	B	4	4	3	2	3	3	3	3	4	3	3	3	3	3	3	3	3	3	**3**	**3**
Average stable, no hygiene understanding of farmer, no additional costs	C	4	5	5	3	5	5	5	5	5	5	1	1	2	5	5	1	3	1	**1**	**3**
Average stable, high hygiene understanding of farmer, no additional costs	D	4	4	5	5	4	3	3	3	4	3	5	4	4	3	3	3	3	3	**4**	**4**
Best stable, high hygiene understanding of farmer, low additional costs	E	5	5	5	5	5	3	3	4	4	4	5	5	5	3	5	4	5	3	**5**	**5**
Minimum requirements of poultry hygiene directive / mode	F	4	3	4	2	4	4	3	4	4	4	4	5	5	3	3	4	4	4	**5**	**4**

Grey background: items are not used in M2

Model comparison

M1 Score model

- Uncorrelated predictor variables
- Values of predictor variables „continuous"
- Most important hygiene
- Transformation of regression coefficients necessary for interpretation
- Includes all 18 original variables
- Different „weight" of each original variable

M2 Scale model

- Minor correlations (<0.5) between predictor variables
- Values of predictor variables „discrete"
- Most important size
- Regression coefficients directly interpretable
- Includes only 12 original variables
- Variables within a scale have the same „weight"

Next steps…

- Further observations of new farms
- Observation of changes in C. status on farms which improve one or more hygiene characteristics
- Evaluation of models M1, M2 (agreement of prediction and observed C. status)
- Establishing a „Campylobacter-herd-status prediction tool"
  - Assessment of intervention strategies (effect vs. costs)
  - Informatory application for farmers, vets, interested stakeholders