UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Examination in	STK4900/9900 — Statistical methods and applications
Day of examination:	11. juni 2008
Examination hours:	09.00 – 12.00
This problem set consists of 2 pages.	
Appendices:	Table over the t-distribution, F-distribution og chisquare-distr
Permitted aids:	All printed and written plus an approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

As a part of a study of the profitability of dairy farms, the total work input in person years was registered for 50 farms, and these farms were categorized according to how many dairy cows they had. The categories that were used, were:

Category 1	Category 2	Category 3	Category 4
< 10 cows	$10 - 20$ cows	$21 - 31$ cows	> 31 cows

a) On the basis of this there was performed a one way analysis of variance with the work input as a dependent variable and the 4 categories as classes. The analysis of variance table was:

.	Df	SS	MS	F	P
Category	3	5.505	1.835	4.89	?
Residual	46	17.267	0.3754		
Total	49	22.772			

Explain how the degrees of freedom are found and how F is found. What can you say about the P-value from these numbers?

b) Assume that one wants to analyse the same data material by a simple linear regression, where the x-variable is 1 for category 1, 2 for category 2, 3 for category 3 and 4 for category 4. Formulate the model, and explain what the different terms mean. What will the number of degrees of freedom for the residual be for this analysis? Explain the principle for testing whether the slope of the regression line is zero.

c) What is the difference between the analysis in a) and the analysis in b)? Look upon a situation where the analysis in a) shows significance while the regression analysis in b) does not show significance. How will you interprete this result? If appropriate, draw a figure.

(Continued on page 2.)

Problem 2

A total of $n = 32$ students were registrered for STK4900/9900 this year. The number of students who met at the practical exercises were:

w = week			1					2		
d = day	1	2	3	4	5	6	7	8	9	10
dw = day within week	1	2	3	4	5	1	2	3	4	5
number of students	30	24	23	21	23	22	21	17	22	16

a) We are interested in stydying the time development of the number of students met at the practical exercises. Explain why it is reasonable to use a logistic regression model.

b) Such a model with d as a regression variable gave the following partial output:

Coefficients :	Estimate	Std.error	z − value
(Intercept)	1.65944	0.28845	5.753
d	−0.15447	0.04379	−3.527

Null deviance: 21.5932 on 9 degrees of freedom
Residual deviance: 8.5831 on 8 degrees of freedom.

Use these numbers to test in two different ways if the expected number of students met depends on time. Give a conclusion.

c) A multiple logistic regression with dw and w as regression variables gave:

Coefficients	Estimate	Std.error
(Intercept)	2.47268	0.50039
dw	−0.21029	0.08783
w	−0.68704	0.24789

Null deviance: 21.5932 on 9 degrees of freedom
Residual deviance: 8.0406 on 7 degrees of freedom.

Use the output table to test if there is any effect of week.

d) The model in b) can be looked upon as nested within the model in c). Why? Use this and the given deviance values to test if there is any effect of week? Why does this give a conclusion which is different from the conclusion that you found under c)? According to your opinion, is there a net effect of week in these data? In what sense? Give reasons for your answer.

THE END