Practice Questions: Final

1. Explain the difference between a logistic linear regression model and loglinear model for three categorical variables X, Y, and Z.

2. In which case in particular is it more meaningful to analyze the relative risk versus the difference between two proportions.

3. The likelihood ratio test is based on what idea?

4. In categorical data analysis Spearman's correlation coefficient is used for what purpose?

5. Give two examples for a GLM, specify link function, random and systematic component for the models.

6. Describe the process of finding a good model in the presence of 5 predictor variables. Which considerations are necessary for deciding to include interaction terms, higher order terms?

7. In 2003 there was a report in the news that an AIDS vaccine tested in Thailand didn't show any effect.

   The data quoted in the news is presented in the two–way table below. Every participant in the study was either vaccinated or received a placebo, and was 5 years later tested for HIV and was either found HIV positive or HIV negative:

	Placebo	Vaccine
HIV+	105	106
HIV-	1168	1167

   Let $X = HIV$ (yes or no) and Y (placebo or vaccine).

   (a) Estimate $P(X = yes \text{ and } Y = vaccine)$

   (b) Estimate the conditional distribution(s) of X.

   (c) Estimate the odds for $X = yes$ for people who received the vaccine

   (d) Estimate the odds ratio for X and Y, interpret the result.

   (e) Find a 95% confidence interval for the odds ratio for X and Y.

   (f) Does the confidence interval indicate a difference in the odds of getting HIV for the two treatment groups?

8. Use the data from the previous problem and test if the variables X and Y are independent.

9. Use the data from the previous problem to find a 99% confidence interval for the difference in the proportion of people who were infected with HIV between those who received the vaccine and those who received placebo.

   Does the confidence interval indicate a significant difference in the proportion of people who get infected with HIV for the two groups?

10. We use a logistic regression model for the data given above.

    The results are in the table below

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	.004	1	.952
	Block	.004	1	.952
	Model	.004	1	.952

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1454.963[a]	.000	.000

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	vacc(1)	.009	.144	.004	1	.952	1.009
	Constant	-2.408	.102	558.688	1	.000	.090

a. Variable(s) entered on step 1: vacc.

	HIV	vacc	freq	SRE_1	ZRE_1	DEV_1
1	1.00	1.00	106.00	2.23089	3.31947	2.23001
2	1.00	.00	105.00	2.23443	3.33381	2.23356
3	.00	1.00	1168.00	-.41698	-.30125	-.41682
4	.00	.00	1167.00	-.41526	-.29996	-.41510
5						

SRE_1=Standardized Residuals
ZRE_1=Studentized Residuals
DEV_1=Deviance Residuals

(a) Give the model equation describing the population.

(b) Give the estimated model equation and interpret the estimates.

(c) Assess the model fit given the information provided.

(d) Find a Wald confidence interval for the slope of the model. What are the conclusions you can draw from the confidence interval.

(e) If you would have to guess which loglinear model fits this data well, which one would it be? Why?

11. Assume we add Z=gender to the data.

    (a) Give a loglinear model for the three variables assuming conditional independence of X and Y.

    (b) Which loglinear model do you think will give a good fit according to what you have found before. Justify your answer.