CATEGORICAL DATA ANALYSIS, 5 credits (intermediate studies), 3.9.–22.10.2015. Literature: Alan Agresti. An Introduction to Categorical Data Analysis, 2. edition. Lecturer: Pekka Pere.

4th exercise set (2.10.)

1. Let π_j ($j = 1, \ldots, c$) be the probability for random variate Y to take a value in category j and N be the number of such independent experiments. Frequencies N_j of Y falling into category j follow the Multinomial distribution

$$\mathsf{P}(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

($\sum_{i=1}^{c} N_i = n$ and $\sum_{i=1}^{c} \pi_i = 1$). Explain why $n\pi_j$ and $n\pi_j(1 - \pi_j)$ are the expected value and variance for the number of observations falling into class j.

2. Let a (multivariate) random variate follow the Multinomial distribution

$$\mathsf{P}(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

($\sum_{i=1}^{c} N_i = n$ ja $\sum_{i=1}^{c} \pi_i = 1$).

a) Derive the logarithm of the multinomial likelihood function

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{c} n_i \log \pi_i.$$

Above $\boldsymbol{\pi} = [\pi_1 \ldots \pi_c]'$.

b) Derive the likelihood equation

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0$$

and prove that the MLE $\hat{\pi}_j$ obeys

$$\frac{\hat{\pi}_j}{\hat{\pi}_c} = \frac{n_j}{n_c}, \quad j = 1, \ldots, c - 1.$$

c) Prove that

$$\sum_{i=1}^{c} \hat{\pi}_i = 1 = \frac{\hat{\pi}_c n}{n_c}$$

and that the MLE is

$$\hat{\pi}_j = \frac{n_j}{n}.$$

(Hint: The MLEs fulfill the equation $\sum_{i=1}^{c} \hat{\pi}_i = 1$.)

3. Let the frequencies N_{ij} in the 2×2 contingency table ($i, j = 1, 2$)

N_{11}	N_{12}	N_{1+}
N_{21}	N_{22}	N_{2+}
N_{+1}	N_{+2}	n

be multinomially distibuted ($\sum_{i=1}^{2} \sum_{j=1}^{2} N_{ij} = n$). Explain why the MLE for the odds ratio is

$$\hat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}}.$$

Above n_{ij}s are the observed frequencies. (Hint1: $\hat{\pi}_{ij} = n_{ij}/n$ (justify this as well) in an obvious notation. Hint2: The properties of MLEs in transformations: $\hat{g}(\boldsymbol{\pi}) = g(\hat{\boldsymbol{\pi}})$.)[1]

4.

a) Let a (multivariate) random variate follow the Multinomial distribution

$$\mathsf{P}(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

($\sum_{i=1}^{c} N_i = n$ ja $\sum_{i=1}^{c} \pi_i = 1$). Derive the likelihood ratio test statistic for the null hypothesis $\pi_1 = \pi_{10}, \ldots, \pi_c = \pi_{c0}$:

$$2 \sum_{i=1}^{c} n_i \log \frac{n_i}{\mu_i}.$$

Here $\mu_i = n\pi_{i0}$ is the expected frequency in class i if the null hypothesis is valid.

b) As above but let the classes above be arranged into a $I \times J$ -contingency table ($c = IJ$). The null hypothesis is that the class probabilities are π_{ij0}. Derive the likelihood ratio statistic (2.7):

$$2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \frac{n_{ij}}{\mu_{ij}}.$$

Above $\mu_{ij} = n\pi_{ij0}$ is the expected frequency in cell ij under the null hypothesis.

c) As b) but let us suppose that the null hypothesis is that the classifying variables are independent. The cell probabilities π_{ij} and marginal probabilities π_{i+} ja π_{+j} are connected as follows: $\pi_{ij} = \pi_{i+}\pi_{+j}$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Derive the likelihood ratio statistic (2.8):

$$2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}.$$

[1] E.g. P. Nieminen and P. Saikkonen (2013): Tilastollisen päättelyn kurssi. Helsingin yliopisto. P. 16. (`Http://www.helsinki.fi/~pjniemin/paattely.pdf`; read 27.9.2014.) B.W. Lindgren (1976): *Statistical Theory, 3rd edition*. MacMillan. New York. Pp. 244–245.

Above $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ (in obvious notation) is the MLE for the expected frequency in cell ij.

5. Exercise 3.6 continued. It will be proved that the statistics z_s^2 and X^2 equal:

$$z_s^2 \equiv \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}(1-\hat{\pi})/n_{1+} + \hat{\pi}(1-\hat{\pi})/n_{2+}} = \sum_{i=1}^{2}\sum_{i=j}^{2} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \equiv X^2.$$

Above $\hat{\pi}_i = n_{i1}/n_{i+}$, $\hat{\pi} = n_{+1}/n$, $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ and $i, j = 1, 2$.

a) Prove that

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{1+}n_{2+}}.$$

b) Prove that

$$\frac{\hat{\pi}(1-\hat{\pi})}{n_{1+}} + \frac{\hat{\pi}(1-\hat{\pi})}{n_{2+}} = \frac{1}{n}\frac{n_{+1}n_{n+2}}{n_{1+}n_{2+}}.$$

c) Prove now that $z_s^2 = X^2$. (Hint: Points a)–b) and the result of exercise 3.6.)

6. Harrell (2015, Section 9.3.4) ponders the pros and cons of the three test statistics Wald (W), likelihood ratio (LR) and score (S) in general.[2] He concludes (p. 193):

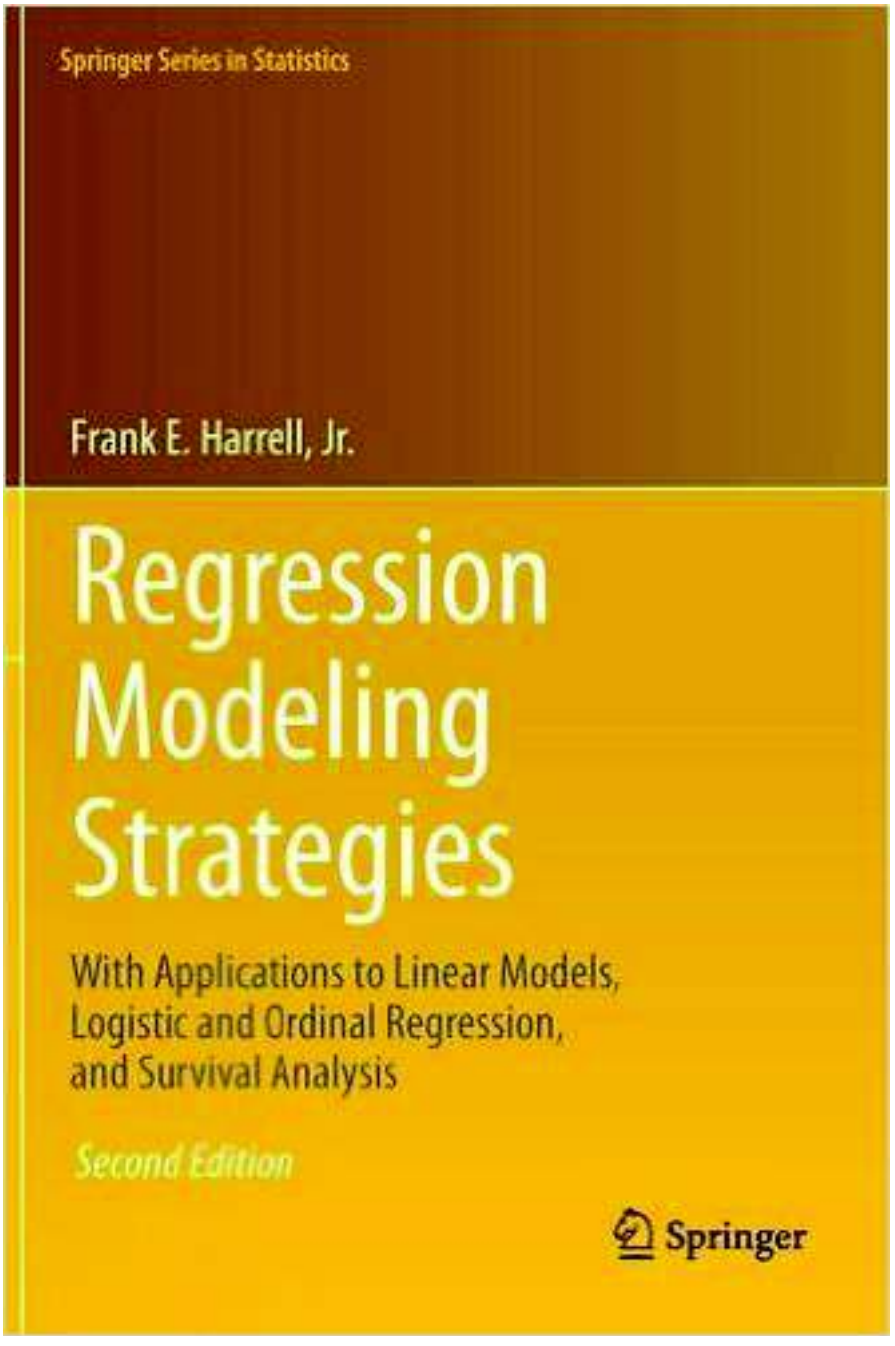
From the standpoint of statistical properties, LR is the best statistic, followed by S and W.

A case in point is testing the null hypothesis of equal event probabilities in two independent binomial samples (a 2×2 contingency table with fixed marginal row frequencies). Harrell considers the case in detail and reasons (p. 195; remarks of the composer of the exercise are in square brackets):

This [LR] *statistic* [for testing the null of equality of two proportions] *for large enough n_1 and n_2 has a χ^2 distribution with 1 d.f. – – It can be shown that the corresponding score statistic is equivalent to the Pearson χ^2 statistic.* [Exercise 4.5.] *The better LR statistic can be used routinely over the Pearson χ^2 for testing hypothesis in contingency tables.*

Consider these evaluations. Are they in harmony with the arguments in Agresti's (2007) text book?

[2] F.E. Harrell, Jr. (2015): *Regression Modeling Strategies. With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd edition.* Springer. Cham.


Springer Series in Statistics
Frank E. Harrell, Jr.
Regression Modeling Strategies
With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis
Second Edition
Springer