First mandatory assignment in STK3100/STK4100-f13

Posted: Monday September 30th.

Deadline: Thursday October 17th, 14:30 pm.

The answers shall be delivered in the box un the hallway on the 7'th floor in Niels Henrik Abels house. Use the standard front page which can be downloaded from www.mn.uio.no/math/studier/admin/obligatorisk-innlevering/obligforside.pdf and indicate whether you attend STK3100 or STK4100.

This is the first out of two mandatory assignments in STK31000/STK4100-f13. It consists of two problems. Both handwritten reports and answers using a word processing system are acceptable. Where you use R (or some other statistical package) the relevant part of the output must be enclosed or pasted into the report. It is OK if you cooperate, but each student must deliver a separate and individually formulated report. Also in case of cooperation it should be indicated in the report whom the others are. There is more information in "Regelverk for obligatoriske oppgaver" which can be obtained from the course web page.

Problem 1

This problem is about testing of two sided fully specified hypotheses for a scalar parameter. There are several approaches: likelihood ratio test, Wald test and the score test. In the exercise you shall compare them in a model where the variable is binomially distributed.

Suppose $Y \sim \mathrm{Bin}(n, p)$, i.e. binomially distributed with n trials and probability of success p.

a) Find the log likelihood $l(p)$, score function $U(p)$ and expected information $\mathcal{J}(p)$ based on Y.

b) Describe the approximate distribution for the maximum likelihood estimater (MLE) $\hat{p} = Y/n$ when n is large and explain why the test statistic for the Wald test for $\mathrm{H}_0 : p = p_0$ is given by

$$Z_W(p_0) = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n}$$

so that at the test statistic is approximately standard normally distributed under the null hypotheses.

c) Show that the score test for the same null hypothesis has test statistic

$$Z_U(p_0) = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}.$$

Also explain why the distribution of $Z_U(p_0)$ is approximately standard normally distributed under H_0..

d) Show that the test statistic for the likelihood ratio test for $\mathrm{H}_0 : p = p_0$ can be written

$$Z_{LR}^2(p_0) = 2[Y \log(\frac{\hat{p}}{p_0}) + (n - Y)\log(\frac{1-\hat{p}}{1-p_0})]$$

What is the approximate distribution of $Z_{LR}^2(p_0)$ when n is large?

e) With $n = 100$ and $y = 30$ implement and compare the three tests for $\mathrm{H}_0 : p = 0.5$.

f) With $n = 100$ and $y = 5$ implement and compare the three tests for $\mathrm{H}_0 : p = 0.15$.

g) For each of the test statistics from parts b)-d) the approximate 95% confidence interval for p have the form $\{p_0 : Z_{\cdot}^2(p_0) < 3.84\}$. Explain why.

h) Show that when $\mathrm{se}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ the confidence interval based on the Wald statistic $Z_W(p_0)$ is $\hat{p} \pm 1.96\mathrm{se}(\hat{p})$.

i) Show that the confidence interval based on the score statistic $Z_U(p_0)$ is the solution of a quadratic equation.
(Solution:$< \hat{p}_L, \hat{p}_U >= \frac{y+1.92}{n+3.84} \pm \frac{1.96\sqrt{n\hat{p}(1-\hat{p})+3.84/4}}{n+3.84}$.)

j) Compute the confidence intervals when $n = 100$ and $y = 30$. (For the interval based on the likelihood ratio statistic $Z_{LR}^2(p_0)$ there is no explicit formula for the end points of the intervals, but you can read them off a plot of $(p_0, Z_{LR}^2(p_0)))$.

k) Compute the confidence intervals when $n = 100$ and $y = 5$.

An alternative parameterization for the binomial distribution use the canonical parameter $\theta = \log(p/(1-p))$, i.e. log-odds.

l) Find the approximate distribution for the MLE $\hat{\theta}$ and show that the estimator of the standard error for $\hat{\theta}$ can be expressed as $\sqrt{1/Y + 1/(n-Y)}$.

m) Formulate the approximate 95% Wald confidence interval $< \hat{\theta}_L, \hat{\theta}_U >$ for θ. Explain why

$$< \frac{\exp(\hat{\theta}_L)}{1+\exp(\hat{\theta}_L)}, \frac{\exp(\hat{\theta}_U)}{1+\exp(\hat{\theta}_U)} >$$

is an approximate 95% confidence interval for p.

n) Compute this interval with $n = 100$, $Y = 30$ and $y = 5$. Compare with the other confidence intervals.

Problem 2

The data set "Third-party-claims" is described on page 17 in deJong and Heller. In the examples in chapter 6 the number of claims is treated as response and the number of accidents as co-variate. We will in this exercise also consider including the other covariates.

a) To check that the data are properly entered reproduce figure 4.1 on page 50 in deJong and Heller.

b) To get an impression of how the continuous covariates are related, use command `pairs(third[,3:7])`. What is your impression?

c) Compute the mean and empirical variance of the number of claims in each of the 13 locations. What is the best way of checking for over dispersion: categorize using covariates as above or categorizing using groups defined by the response as done in deJ&H? Give your reasons.

d) Start by fitting a general negative binomial regression model of the same type as in deJong and Heller but with the covariates `log(ki)`, `sd` and `log(pop-density)` used in addition. Comment on the result. How do you explain the sign of the coefficients of `log(ki)` and `log(accidents)`?

e) Check whether the covariates `sd` and `log(pop-density)` can be dropped.

f) Compute the covariance matrix/ correlation matrix of the estimated coefficients. Comment on the correlations corresponding to the estimate of the coefficients of `log(ki)` and `log(accidents)`.

g) Discuss whether it is reasonable to retain both covariates `log(ki)` and `log(accidents)`. If one has to be deleted, which one is the most reasonable to keep?

h) As an alternative to the fitted negative binomial model you chose in part g) use the same covariates to fit a model using quasi-likelihood. Compare this to your chosen negative binomial model.

i) Compare the residuals for the two models from part h) with the residuals from the corresponding Poisson model. Comment on what you find.