TMA4300 Computer Intensive Statistical Methods

Exercise 3, Spring 2024

Problem 1: Bootstrapping a GLM

Suppose that $Y_1, Y_2, \dots, Y_n$ are independent random variables, and that each $Y_i \sim \text{bin}(m_i, p_i)$ where $\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i$ for $i = 1, 2, \dots, n$. This a generalized linear model that can be fitted by maximum likelihood in R using the code

```
mod <- glm(cbind(y, m - y) ~ x, family=binomial, data=data)
```

The ML estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, can be extracted from the fitted model object by using the `coef` function. Based on the result that $\text{Var}\,\hat{\boldsymbol{\beta}}$ in general is asymptotically equal to the inverse of the Fisher information matrix $F(\boldsymbol{\beta})$, an approximate estimate of $\text{Var}\,\hat{\boldsymbol{\beta}}$ is given $(F(\hat{\boldsymbol{\beta}}))^{-1}$. This estimate can be extracted from the fitted model object using the `vcov` function.

The data that we will use is available in the data frame `data` that you can load into R with the command

```
load(file=url("https://www.math.ntnu.no/emner/TMA4300/2024v/data.Rdata"))
```

Do this and then fit the model using the above code.

a) Write an R function that takes the fitted model object `mod` as input and that simulates $B = 10000$ bootstrap samples by resampling the observations triplets (m_i, y_i, x_i) with replacements. Then refit the above model to each bootstrap sample to obtain bootstrap replicates $\hat{\boldsymbol{\beta}}^{b*}$, $b = 1, 2, \dots, B$ of $\hat{\boldsymbol{\beta}}$. You function should then return these bootstrap replicates as a $(B \times 2)$-matrix.

b) Calculate an estimate of $\text{Var}\,\hat{\boldsymbol{\beta}}$ based on the bootstrap replicates $\hat{\boldsymbol{\beta}}^{b*}$. How to these compare to the approximate/asymptotic obtained using `vcov(mod)`?

c) Estimate the bias of the MLEs of the intercept and slope parameters. Do the estimators appear to be significantly biased? If so, compute bias-corrected estimates.

d) Using the percentile method, compute approximate 95%-confidence intervals for each model parameter. Compare these to the confidence intervals obtained based on the profile likelihood of each parameter. These can be computed using the `confint` function.

e) Redo points a) to d) using instead parametric bootstrapping. Briefly discuss differences you see.

Problem 2: Bootstrap confidence intervals

Suppose that $X_1, X_2, \dots, X_n$ is an iid sample from an exponential distribution with scale parameter β.

a) Show that pivotal quantity $2\sum_{i=1}^{n} X_i/\beta$ is chi-square with $2n$ degrees of freedom. Use this to derive an exact $(1-\alpha)$ confidence interval for β.

b) Suppose that we instead were to use parameteric bootstrapping and constructed a bootstrap confidence interval for β using the percentile method (Givens & Hotings, section 9.3.1), that is, using the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of bootstrap replicates $\hat{\beta}^*$ of $\hat{\beta}$ where $\hat{\beta}$ is the MLE of β. When $\hat{\beta}^*$ are based on bootstrap samples from $F(x;\hat{\beta})$, what is the exact distribution of $\hat{\beta}^*$? Find analytic formulas for the resulting confidence limits as functions of $\hat{\beta}$ for a given sample size n.

c) Find an expression for the exact coverage of the parametric bootstrap percentile interval in point b) in terms of the cdf and quantile function of the chi-square distribution. Compute the exact coverage for $n = 5, 10, 20, 50, 100$ for $\alpha = 0.05$.

d) Write an R function that takes the observed sample $x_1, x_2, \ldots, x_n$ and α as input and that computes a two-sided $(1-\alpha)$ BC_a-confidence interval for β based on parametric bootstrapping (see Givens & Hoeting, section 9.3.2.1 for details). Again, since the distribution of bootstrap replicates $\hat{\beta}^*$ is known analytically, there is no need to carry out simulations to find quantiles of this distribution as well as the constants a and b.

e) Assuming that the true value of $\beta = 1$, estimate the coverage of the BC_a interval in point d) for $\alpha = 0.05$ by simulating 10000 random samples, each of size $n = 10$ from the model and checking if the associated BC_a interval contains β. Briefly comment on how the performance of the interval compares to the intervals in point a) and b).

Problem 3: The EM-algorithm and bootstrapping

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be independent random variables, where the x_i's have an exponential distribution with intensity λ_0 and the y_i's have an exponential distribution with intensity λ_1. Assume we do not observe $x_1, \ldots, x_n, y_1, \ldots, y_n$ directly, but that we observe

$$z_i = \max(x_i, y_i) \quad \text{for } i = 1, \ldots, n \tag{1}$$

and

$$u_i = I(x_i \geq y_i) \quad \text{for } i = 1, \ldots, n, \tag{2}$$

where $I(A) = 1$ if A is true and 0 otherwise. Thus, for each $i = 1, \ldots, n$ we observe the largest value of x_i and y_i and we know whether the observed value is x_i or y_i. Based on the observed $(z_i, u_i), i = 1, \ldots, n$ we will use the EM algorithm to find the maximum likelihood estimates for (λ_0, λ_1)

a) Write down the log likelihood function for the complete data $(x_i, y_i), i = 1, \ldots, n$. Use this to show that

$$\begin{aligned} E\left[\ln f(\mathbf{x}, \mathbf{y}|\lambda_0, \lambda_1)|\mathbf{z}, \mathbf{u}, \lambda_0^{(t)}, \lambda_1^{(t)}\right] &= n(\ln\lambda_0 + \ln\lambda_1) \\ &- \lambda_0 \sum_{i=1}^{n}\left[u_i z_i + (1-u_i)\left(\frac{1}{\lambda_0^{(t)}} - \frac{z_i}{\exp\{\lambda_0^{(t)} z_i\} - 1}\right)\right] \\ &- \lambda_1 \sum_{i=1}^{n}\left[(1-u_i) z_i + u_i\left(\frac{1}{\lambda_1^{(t)}} - \frac{z_i}{\exp\{\lambda_1^{(t)} z_i\} - 1}\right)\right] \end{aligned}$$

b) Using the EM algorithm, use the result you found in point 1) to find a recursion in $(\lambda_0^{(t)}, \lambda_1^{(t)})$ for finding the maximum likelihood estimates for (λ_0, λ_1). Implement the recursion and find the maximum likelihood estimates applied to the data vectors `u` and `v` that you can load into R by doing

```
u <- scan(file="https://www.math.ntnu.no/emner/TMA4300/2024v/u.txt")
z <- scan(file="https://www.math.ntnu.no/emner/TMA4300/2024v/z.txt")
```

Visualise the convergence of the algorithm in a plot.

c) Use bootstrapping to estimate the standard deviations and the biases of each of $\widehat{\lambda}_0$ and $\widehat{\lambda}_1$ and to estimate $\mathrm{Corr}[\widehat{\lambda}_0, \widehat{\lambda}_1]$. Present pseudocode for your bootstrap algorithm. Discuss briefly whether you would prefer the maximum likelihood estimates or the bias corrected estimates for λ_0 and λ_1 in this case.

d) For the situation defined here, you find an analytical formula for $f_{Z_i,U_i}(z_i, u_i|\lambda_0, \lambda_1)$?

Is it possible to find analytical formulas for the maximum likelihood estimators $\widehat{\lambda}_0$ and $\widehat{\lambda}_1$? Find the mle for $\widehat{\lambda}_0$ and $\widehat{\lambda}_1$ analytically or numerically. What are the advantages of optimizing the likelihood directly compared to the EM algorithm?