# Computational Statistics, Assessed Practical

Week 1, HT 2021

- This practical sheet contains two exercises. Write a single report with answers to both exercises.
- The report has soft word limit at 2000 words and a hard limit at 2500 words. This word limit is on the main body of the report. Equations, tables, figures, captions, appendices to your report and computer code do not contribute to the word count.
- You should use your anonymous practical ID (and not your real name) for the cover page of the report, and you should name the PDF file you upload using that same ID (e.g. "P042.pdf").
- The submission deadline is 12 noon Monday 3 May.

## Exercise 1

The data in `dvis.csv` relate to the number of visits to a family doctor, or GP, and is the same data set as used in the Applied Statistics assessed practical. The dependent count variable $Y$ is `docvis`, and the linear predictor is given by $\eta = \mathbf{x}^T\beta$ where $\mathbf{x}$ contains the constant, `age`, `hhninc`, `female`, `hhkids`,`educyrs` and `addins`.

1. Estimate the above model specification using the Poisson GLM with canonical link function. An estimator for the variance of the parameter estimates that is robust to general forms of heteroskedasticity is the so-called sandwich estimator. This estimator can be obtained using the `vcovHC` command from the 'sandwich' package in R. Use this estimator to compute the standard errors of the estimated coefficients and compare to the standard Poisson GLM standard errors. Comment on the findings.

2. Compute bootstrap standard errors of the estimated coefficients using the nonparametric paired bootstrap. How do these compare to the ones obtained in 1.?

3. Compute the Wald test statistics and p-values for testing $H_0 : \beta_{hhkids} = \beta_{educyrs} = 0$, using the standard Poisson GLM, the sandwich and the bootstrap estimates of the variance matrix of the estimated parameters. What do you conclude?
   *Note:* Let $\hat{\beta}$ denote the Poisson MLE of $\beta$, then the formula for the Wald test statistic is

$$W = \begin{pmatrix} \hat{\beta}_{hhkids} \\ \hat{\beta}_{educyrs} \end{pmatrix}^T \left( v\hat{a}r \begin{pmatrix} \hat{\beta}_{hhkids} \\ \hat{\beta}_{educyrs} \end{pmatrix} \right)^{-1} \begin{pmatrix} \hat{\beta}_{hhkids} \\ \hat{\beta}_{educyrs} \end{pmatrix}.$$

4. Compute a bootstrap p-value for the Wald test statistic based on the sandwich variance estimator. Does the result change compared to that found in 3.?

5. What do the above analysis and results imply when used for model selection?

## Exercise 2

The variables inflFR and year in the data set infl.csv are time series observations on the real consumer price annual inflation rate for France for the years 1956-2020.

Investigate and write a report on the time series features of the inflation data. The main goal here is to fit an AR(1) model with time varying coefficients using the Kalman filter.

1. Perform an exploratory analysis of the inflation data and summarise your findings.

2. Let $Y$ denote the inflation rate. Estimate the AR(1) specification

   $$Y_t = \alpha + Y_{t-1}\beta + W_t$$

   for $t = 1, \ldots, T$, with $t = 1$ for the year 1957, and so $Y_0$ is the inflation rate in 1956 and $T = 64$. Describe and interpret your findings and discuss the adequacy of this specification to capture the time series features/serial correlation of the inflation data.

3. Specify an AR(1) model with time varying coefficients as a linear Gaussian state-space model for $t = 1, \ldots, T$,

   $$\begin{aligned} \beta_t &= \beta_{t-1} + V_t \\ Y_t &= Y_{t-1}\beta_t + W_t \end{aligned}$$

   where, given $Y_0 = y_0$, the random variables $\beta_0, V_1, V_2, \ldots, V_T, W_1, W_2, \ldots, W_T$ are independent, with $\beta_0 \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ and for $t = 1, \ldots, T$, $V_t \sim \mathcal{N}\left(0, \sigma_v^2\right)$ and $W_t \sim \mathcal{N}\left(0, \sigma_w^2\right)$.

   (a) Adapt the R function kalman provided in the lecture notes or use your own function to compute the filtering mean $\mu_{t|t} = \mathbb{E}[\beta_t | y_0, \ldots, y_t]$ and variance $\sigma_{t|t}^2 = \mathbb{V}(\beta_t | y_0, \ldots, y_t)$. Assume that $\beta_0 \sim \mathcal{N}(0, 1)$ and set $\sigma_v^2 = 0.01$ and $\sigma_w^2 = 4$, and present your results for the filtering mean in a graph, including the 95% credible intervals as detailed in the lecture notes.

   *Note:* As we have omitted the constant for ease of calculation, take the variables in deviation from their means:[1]

```
data <- read.csv("infl.csv")
inflFR<-data$inflFR
year<-data$year

T1 <- length(inflFR)
infl <- as.matrix(inflFR[2:T1])
laginfl <- as.matrix(inflFR[1:T1-1])
year <- as.matrix(year[2:T1])
T <- length(year)

infl <- infl-mean(infl)
laginfl <- laginfl-mean(laginfl)
```

[1] Also note that input “y” in the kalman function is a row vector, so “t(infl)”.

(b) Keeping $\beta_0 \sim \mathcal{N}(0, 1)$, we now would like to choose the optimal values of $\sigma_v^2$ and $\sigma_w^2$ by computing the loglikelihood $\log p(y_1, \dots, y_T | y_0)$ for different values of $\sigma_v^2$ and $\sigma_w^2$. Note that

$$\log p(y_1, \dots, y_T | y_0) = \log p(y_1 | y_0) + \sum_{t=2}^{T} \log p(y_t | y_0, \dots, y_{t-1})$$

where $p(y_1|y_0) = \mathcal{N}(y_1; \hat{y}_{1|0}, s_1^2)$ and $p(y_t|y_0, \dots, y_{t-1}) = \mathcal{N}(y_t; \hat{y}_{t|t-1}, s_t^2)$, where $\hat{y}_{t|t-1} := \mathbb{E}[Y_t|y_0, \dots, y_{t-1}]$ and $s_t^2 := \mathbb{E}[\left(Y_t - \hat{y}_{t|t-1}\right)^2 |y_0, \dots, y_{t-1}]$

i. Modify the `R` function `kalman` so that it also returns the loglikelihood $\log p(y_1, \dots, y_T | y_0)$.

ii. By computing the loglikelihood over a grid of values, numerically find the maximum likelihood estimates of $\sigma_v^2$ and $\sigma_w^2$.

(c) For the estimated values $\sigma_v^2$ and $\sigma_w^2$, repeat question (a). Summarise and interpret your findings.