UNIVERSITEIT VAN DIE VRYSTAAT
UNIVERSITY OF THE FREE STATE

STSB 6816

WISKUNDIGE STATISTIEK & AKTUARIËLE WETENSKAP/
MATHEMATICAL STATISTICS & ACTUARIAL SCIENCE

Test 2 — 23 July 2020

TYD/TIME: 1800 Minutes PUNTE/MARKS: 100

INSTRUCTIONS:

- Answer all questions in a single R Markdown document. Knit to Word at the end and submit both for assessment in a single submission. Alternatively, you may work directly in Word at a penalty of 2% of the total test mark allocation.
- Label questions clearly using headings, as it is done on this question paper. Verify that you have done this correctly by noting the structure in the Word Navigation Pane.
- All results accurate to 2 decimal places.
- Show all derivations, formulas, code, sources and reasoning. Sources must be cited in text and referenced at the end, with links to websites where information or code was found.
- Intervals should cover 95% probability unless stated otherwise, and hypothesis tests should assume $\alpha = 0.05$.
- No communication software, no communication devices, and no communication capable websites may be accessed prior to submission, except for the purpose of posing questions to the lecturer. You may not (nor even appear to) attempt to communicate or pass information to another student. Thus, your final submission must be truly unique to you in all respects.

Question 1

The Generalised Pareto Distribution (GPD) is used to model extreme observations above a threshold or cut-off point. One situation where it is commonly used is reinsurance claims. The density is

$$f_X(x) = \sigma^{-1}\left[1 + \frac{\xi(x-\mu)}{\sigma}\right]^{-\xi^{-1}-1}$$

The key parameter of the GPD is the extreme value index (EVI), which we will denote ξ (xi), that measures the extremeness of the extreme values. This parameter is generally assumed constant but unknown for a specific data generating process.

Another parameter is the threshold (μ) which we will assume to be 0 for this problem. In any situation where the threshold value is known it can be subtracted from all larger observations and then assumed to be 0. In reinsurance contracts the threshold is usually specified in writing and is thus known.

The last parameter is the scale parameter σ, this parameter stretches out the distribution to fit the scale of the data, and must be estimated carefully as it is negatively correlated with ξ. The problem is that the scale can easily change over time due to many possible factors. One possible factor is inflation.

You are provided a set of extreme claims with the threshold already subtracted and set to 0. You are also provided with an inflation index matching the time of each claim. The claim times are given as days since the introduction of this type of policy on 1 July 2014.

It is your task to fit two GPD distribution models to the claims:

Model 1: Assume that the scale parameter grows with inflation ($\sigma_i = \sigma * Index_i$), so that the observations can be divided by the index to arrive at an i.i.d. GPD sample with a fixed scale and fixed EVI. Assume a positive EVI to avoid boundary issues. Use the MDI prior for the GPD $\propto \sigma^{-1}\exp(-\xi)$ or $c_1 - \log\sigma - \xi$ on the log scale (c_1 being an unknown constant), or $c_2 + \log\alpha - \log\beta - \alpha$ if you express the GPD as a gamma mixture of exponential random variables.

Model 2: Ignore the inflation index entirely and instead assume that the scale parameter increases gradually over time in a gradual exponential curve, *i.e.* σ_i (*or* β_i) = $\exp(\sigma_0 + \delta * t_i)$ where t_i is the time since the start of the contract (also given in your data). This model should take into account that δ must be positive. Again assume a positive EVI to avoid boundary issues, perhaps with an Exp(1) prior or similar.

(a) Load the data corresponding to your student number and draw a plot of claims (as points) and inflation (as a line) versus time. [4]

(b) For Model 1, implement the model in any way you choose, as best you can. Illustrate the joint posterior distribution of the two parameters on a graph of your choosing. [10]

[Remember to divide the claims by the inflation index first then fit the GPD as if you had an i.i.d. sample.]

(c) Redraw your data plot but now include the 95th upper percentile suggested by Model 1 as an additional line. Explain how parameter uncertainty was handled in deciding where to draw the line, or draw fuzzy/additional lines that illustrate the parameter uncertainty.
Note that the quantile function of the GPD is $[(1-p)^{-\xi}-1]\sigma/\xi$. [6]

(d) Now revert back to the original data set and fit Model 2 as best you can. Illustrate the posterior distribution of the slope parameter (δ). [11]

(e) Redraw your original data plot but now replace inflation with the growth curve suggested by Model 2. Be sure to also illustrate the uncertainty in your estimation of the growth curve. [6]

(f) Compare Model 1 to Model 2 considering generalisability or parsimony. [4]

Total for Question 1: 41

Question 2

In the file provided you will find a set of data on the sheet corresponding to your student number. The data has many variables of different types. You are tasked with building appropriate regression models for the first two of those variables, dependent on all the others, and interpreting the results.

The first variable is an objective measure of subject illness on an arbitrary real scale. The second variable is a subjective measure of subject illness on a scale of 0 to [maximum in your data]. You also have available the subject age, various measures that might explain subject illness, and the subject ID.

(a) Begin by drawing small histograms of all numeric *Xvar* variables to see whether any have excessive skewness. If so, take the natural log of those variables and draw new histograms to see whether the impact was good. Explain why this step is useful. [8]

(b) Regress the first variable on all the others except the second variable (as linear main effects only), assuming a t density for the dependent variable conditional on the explanatory model. Discuss which explanatory variables, other than Subject(s), can be considered statistically significant. [Tip: use as prior the assumption that the degrees of freedom parameter ν (nu) follows an exponential(0.1) density.] [12]

(c) Explain when and why one might want to use a t density instead of a Gaussian density for a dependent variable. [3]

(d) Calculate 90% prediction intervals for the values of your dependent variable and calculate the empirical coverage of your model. Give only the empirical coverage and attempt to explain any deviation from 90% coverage. [5]

(e) Adapt your model to make the Subject variable (last column) a random effect grouping variable instead of a fixed effect as in the previous model. Report the estimated standard deviation of the random effect as a proportion of the estimated standard deviation of the residuals (point estimates are fine here). [8]

(f) When is it appropriate to drop insignificant explanatory variables and refit the model, and when should a variable not be dropped even if it appears insignificant? [3]

(g) Adapt your model to use the prior density given below. Illustrate the posterior density of ν that results. [5]

$$p(\nu) \propto \left[\psi'(0.5\nu) - \psi'(0.5(\nu+1)) - 2(\nu+3)\nu^{-1}(\nu+1)^{-2}\right]^{0.5}$$

where ψ' is the trigamma function.

(h) Regress the second variable on all the others except the first variable (as linear main effects only), assuming an appropriate density of your choosing for the dependent variable conditional on the explanatory model. Estimate the coefficients of the explanatory variables (except Subject) and interpret the values of those coefficients which appear significant. Also give the probability that the Age coefficient is larger than the others. [15]

Total for Question 2: 59