Projects for "Bayesian Computation with R"

Rainer Hirk

WS 2021/2022

1 S&P Rating Data

On the homepage of this course you can find a time series for Standard & Poors default data in the different rating categories. The data set is available as `spdata.df.rda` and consists of the yearly default data as well as the number of companies at risk for the five rating categories (A, BBB, BB, B, CCC) over the years 1981–2000.
We propose the following random-effects model.

```
> library("lme4")
> load(file.path("..", "../data/spdata.df.rda"))
> glmer(cbind(defaults, firms-defaults) ~ 0 + rating + (1 | year),
+       family=binomial(probit), data=spdata.df)

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( probit )
Formula: cbind(defaults, firms - defaults) ~ 0 + rating + (1 | year)
   Data: spdata.df
     AIC      BIC   logLik deviance df.resid
 404.338  419.969 -196.169  392.338       94
Random effects:
 Groups Name        Std.Dev.
 year   (Intercept) 0.2415
Number of obs: 100, groups:  year, 20
Fixed Effects:
  ratingA  ratingBBB   ratingBB    ratingB    ratingC
  -3.4318    -2.9185    -2.4039    -1.6895    -0.8378
```

Reproduce the analysis using Bayesian methods in `JAGS`.

References

[1] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques, and tools.* Princeton university press, 2010 (Chapter 8).

2 Quarterly growth rate of U.S. real GNP

In package **FinTS** the quarterly growth rate data of the U.S. real gross national product which is seasonally adjusted is available from the second quarter of 1947 to the first quarter of 1991. ML estimation is used to fit an autoregressive model (AR(p)).
An AR(p) process can be fitted using `arima` function in `R` by ML estimation. For $p = 1$ one has

```
> data("q.gnp4791", package = "FinTS")
> p <- 1
> (m <- arima(q.gnp4791, order = c(p, 0, 0),
+                  method = "ML"))

Call:
arima(x = q.gnp4791, order = c(p, 0, 0), method = "ML")

Coefficients:
         ar1  intercept
      0.3786     0.0077
s.e.  0.0698     0.0012

sigma^2 estimated as 9.801e-05:  log likelihood = 562.47,  aic = -1118.94
```

1. After inspecting the ACF plot, choose the optimal p of the AR(p) model by using a suitable information criterion (e.g., AIC).

2. After identifying the best model, use it to make 2-steps ahead predictions (`predict(m, n.ahead = 2)`).

3. Reproduce this analysis for the chosen model using Bayesian methods in `JAGS` including model fitting and prediction.

References

[1] Ruey S Tsay. *Analysis of financial time series*, Volume 543. John Wiley & Sons, 2005.

3 Home mortgage disclosure act data

The Boston HMDA data set in the **AER** was collected by researchers at the Federal Reserve Bank of Boston and combines information from mortgage applications and a follow-up survey of the banks and other lending institutions that received these mortgage applications. The data pertain to mortgage applications made in 1990 in the greater Boston metropolitan area. In the following a subset of the original data is used by restricting the observations only to single-family residences (thereby excluding data on multi-family homes) and to black and white applicants (thereby excluding data on applicants from other minority groups). This leaves 2380 observations. Documentation for the data can be found at `http://artax.karlin.mff.cuni.cz/r-help/library/AER/html/HMDA.html`. We fit two generalized linear models using the `glm` function:

```
> data("HMDA", package = "AER")
> hmda_probit <-  glm(deny ~ pirat + lvrat + chist + phist +
+                       selfemp + insurance + afam + single + hschool,
+                      data = HMDA, family = binomial("probit"))
> summary(hmda_probit)


Call:
glm(formula = deny ~ pirat + lvrat + chist + phist + selfemp +
    insurance + afam + single + hschool, family = binomial("probit"),
    data = HMDA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8048  -0.4357  -0.2970  -0.2089   3.2101

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.70630    0.33580  -8.059 7.67e-16 ***
pirat         2.34503    0.41529   5.647 1.64e-08 ***
lvrat         0.81771    0.25269   3.236 0.001212 **
chist2        0.32649    0.10568   3.089 0.002005 **
chist3        0.43286    0.16162   2.678 0.007400 **
chist4        0.77480    0.18140   4.271 1.94e-05 ***
chist5        0.62071    0.13015   4.769 1.85e-06 ***
chist6        0.83035    0.12331   6.734 1.66e-11 ***
phistyes      0.72652    0.11940   6.084 1.17e-09 ***
selfempyes    0.34538    0.11191   3.086 0.002028 **
insuranceyes  2.57596    0.28208   9.132  < 2e-16 ***
afamyes       0.37070    0.09710   3.818 0.000135 ***
singleyes     0.24705    0.07881   3.135 0.001721 **
hschoolyes   -0.61472    0.24005  -2.561 0.010443 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1744.2  on 2379  degrees of freedom
Residual deviance: 1270.2  on 2366  degrees of freedom
AIC: 1298.2

Number of Fisher Scoring iterations: 6
```

```
> data("HMDA", package = "AER")
> hmda_logit <- glm(deny ~ pirat + lvrat + chist + phist +
+                      selfemp + insurance + afam + single + hschool,
+                    data = HMDA, family = binomial("logit"))
> summary(hmda_logit)


Call:
glm(formula = deny ~ pirat + lvrat + chist + phist + selfemp +
    insurance + afam + single + hschool, family = binomial("logit"),
    data = HMDA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9179  -0.4296  -0.2986  -0.2166   3.0928

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.3607     0.6377  -8.407  < 2e-16 ***
pirat          4.6272     0.7800   5.932 2.99e-09 ***
lvrat          1.9166     0.5010   3.826 0.000130 ***
chist2         0.7157     0.2113   3.388 0.000705 ***
chist3         0.8446     0.3093   2.730 0.006328 **
chist4         1.5343     0.3298   4.653 3.28e-06 ***
chist5         1.2459     0.2437   5.112 3.19e-07 ***
chist6         1.5819     0.2274   6.957 3.48e-12 ***
phistyes       1.2743     0.2072   6.151 7.72e-10 ***
selfempyes     0.6602     0.2109   3.130 0.001749 **
insuranceyes   4.6072     0.5570   8.272  < 2e-16 ***
afamyes        0.6513     0.1771   3.678 0.000235 ***
singleyes      0.4402     0.1519   2.898 0.003753 **
hschoolyes    -1.1260     0.4257  -2.645 0.008166 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1744.2  on 2379  degrees of freedom
Residual deviance: 1265.6  on 2366  degrees of freedom
AIC: 1293.6

Number of Fisher Scoring iterations: 6
```

Choose the suitable link function. For this model, reproduce the analysis using Bayesian methods in `JAGS`.

References

[1] Alicia H Munnell, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney. Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review*, pages 25–53, 1996.

4 Compustat financial ratios and credit scores

The compRatios dataset contains financial ratios computed from annual balance sheet information from Compustat North America between the years 2009 and 2013, for 447 companies in the US and Canada. In addition, a credit score is constructed from rating data and is monotonically decreasing with creditworthiness (the higher the score the closer to default a firm is). The dataset compRatios.rda can be found on the webpage of this course. The documentation for the variables, including formula and category for each financial ratio can be found in Table 1.

Table 1: Documentation

Code	Description	Category
R1	interest expenses / assets	interest coverage
R2	interest expenses / debt	interest coverage
R3	cash / liabilities	liquidity
R4	debt / EBITDA	leverage
R5	retained earnings / assets	profitability
R6	EBIT / assets	profitability
R7	cash-flow / equity	cash-flow
R8	sales / assets	efficiency

We fit the following mixed effects model using the lmer function in the **lme4** package:

```
> load(file.path("..", "../data/compRatios.rda"))
> m <- lmer(score ~ year + R1 + R2 + R3 + R4 + R5 + R6 + R7 + R8 +
+           (1|company), data = compRat)
> print(summary(m), correlation = FALSE)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
score ~ year + R1 + R2 + R3 + R4 + R5 + R6 + R7 + R8 + (1 | company)
   Data: compRat

REML criterion at convergence: -597.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.8632 -0.3334 -0.0169  0.3112  7.4203

Random effects:
 Groups   Name        Variance Std.Dev.
 company  (Intercept) 0.09414  0.3068
 Residual             0.01709  0.1307
Number of obs: 1645, groups:  company, 447

Fixed effects:
             Estimate Std. Error  t value
(Intercept) -2.478009   0.016445 -150.680
year2010    -0.243512   0.009529  -25.554
year2011    -0.461407   0.009833  -46.924
year2012    -0.570582   0.010000  -57.057
year2013    -0.682234   0.025580  -26.671
R1           0.096592   0.013755    7.022
R2           0.038381   0.009059    4.237
R3           0.014425   0.008224    1.754
R4           0.104043   0.009948   10.459
R5          -0.104538   0.012281   -8.512
R6          -0.039846   0.010211   -3.902
R7           0.001224   0.007243    0.169
R8           0.087840   0.013548    6.484
```

Reproduce the analysis using using Bayesian methods in JAGS.

5 Volatility modelling of Microsoft log-returns

We consider the Microsoft (MSFT) daily adjusted prices and the corresponding log-returns for the period 1997-2000. The data can be downloaded using the **quantmod** package in `R`. Figure 1 shows the time series of raw log-returns. We estimate a GARCH(1,1) model with a leverage effect using the `R` package **fGarch**. A constant mean term is used for modelling the conditional mean, i.e., the equation for the conditional mean has the following form:

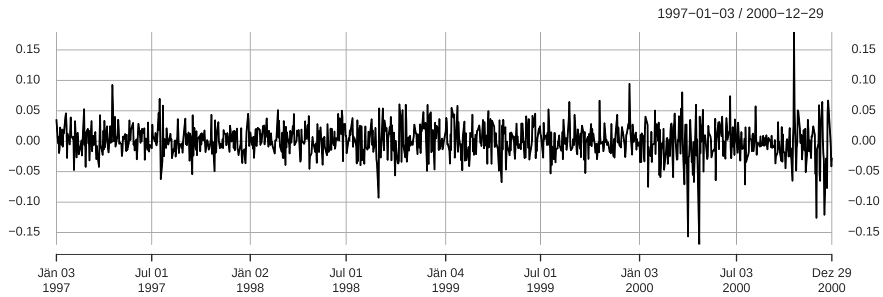
$$y_t = \mu + \epsilon_t.$$

The conditional variance is given by:

$$\sigma_t^2 = \omega + \alpha_1(y_{t-1} - \gamma_1|y_{t-1}|)^2 + \beta_1\sigma_{t-1}^2 \qquad \omega > 0, \alpha_1 \geq 0, \beta_1 \geq 0, \alpha_1 + \beta_1 < 1, \gamma_1 \in [-1, 1].$$

The conditional distribution of the innovations ϵ_t can be assumed to follow, e.g., a normal distribution $\epsilon_t \sim N(0, \sigma_t^2)$ or a Student t-distribution $\epsilon_t \sim t_\nu(0, \sigma_t^2)$.

Figure 1: Microsoft – raw log-returns



1. Fit the model in `R` and use a conditional normal distribution and a conditional Student t-distribution for the innovations:

```
> library(fGarch)
> fitNormal <- garchFit(~ garch(1,1), data = y,
+                       cond.dist = "norm", leverage = TRUE)
> fitStudentT <- garchFit(~ garch(1,1), data = y,
+                         cond.dist = "std", leverage = TRUE)
>
```

   Inspect the residuals of the two models (e.g., by using Q-Q plots). Explain which model you consider more appropriate and why.

2. Reproduce the analysis for the chosen model using Bayesian methods in `JAGS`.

References

[1] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools (Revised Edition).* Princeton university press, 2015 (Chapter 4, Example 4.24).

6 Airline data

The data consists of monthly airline passenger numbers (in thousands) from 1949–1960. This data set is also referred to as the classic Box & Jenkins airline data.
A seasonal ARIMA process with a periodicity of 12 months is fitted to the time series after taking $\log_{10}$. It is assumed that for the nonseasonal as well as the seasonal part the time series follows an MA(1) process after taking the first difference of the series and the first seasonal difference.

```
> data("AirPassengers", package = "datasets")
> AirPassengers <- log10(AirPassengers)
> (air_arima <- arima(AirPassengers, c(0, 1, 1),
+                      seasonal = list(order = c(0, 1, 1), period = 12),
+                      method = "ML"))

Call:
arima(x = AirPassengers, order = c(0, 1, 1), seasonal = list(order = c(0, 1,
    1), period = 12), method = "ML")

Coefficients:
          ma1     sma1
      -0.4018  -0.5569
s.e.   0.0896   0.0731

sigma^2 estimated as 0.0002543:  log likelihood = 353.96,  aic = -701.92

> pred <- predict(air_arima, n.ahead = 2)
> 10^pred$pred

          Jan      Feb
1961 450.4224 425.7172

> 10^cbind(LB = pred$pred - 1.96 * pred$se,
+          UB = pred$pred + 1.96 * pred$se)

               LB       UB
Jan 1961 419.1476 484.0307
Feb 1961 391.4747 462.9550
```

Reproduce this analysis using Bayesian methods in `JAGS` including model fitting and prediction. Please note that the proposed model implies

$$(1 - L^{12})(1 - L)\log_{10}(\text{AirPassengers}_t) = (1 + \alpha_1 L)(1 + \alpha_2 L^{12})\epsilon_t,$$

where L denotes the lag operator, i.e, shifts the time series one step such that $Lx_t = x_{t-1}$. This is equivalent to

$$\text{dlogAirPassengers}_t = (1 + \alpha_1 L + \alpha_2 L^{12} + \alpha_1\alpha_2 L^{13})\epsilon_t$$

where $\text{dlogAirPassengers}_t = (1 - L^{12})(1 - L)\log_{10}(\text{AirPassengers}_t)$.

References

[1] Ruey S Tsay. *Analysis of financial time series*, Volume 543. John Wiley & Sons, 2005.