Modelling Longitudinal Data using the Stat-JR Package

William Browne and George Leckie

Centre for Multilevel Modelling
University of Bristol

7 July 2014

1. Models for Repeated Measures
Growth Curve Models

Introduction

Longitudinal studies

- A longitudinal study tracks the same group of subjects over time
    - Subjects may be individuals, organisations etc.
- Classic longitudinal designs are prospective
    - A cohort study samples a group who experience an event at the same time (usually born in a given time period)
    - A panel study samples a cross-section of the population
- But longitudinal data may also be collected retrospectively, e.g. asking people to recall timing of events (births, marriages)

Causal inference using longitudinal data

- Cross-sectional data cannot be used for causal inference (unless from a randomised experiment). Longitudinal data can greatly reduce bias in estimating causal effects
  - E.g. how can we interpret cross-sectional associations between exercise (E), diet (D) and weight (W)?
  - E and D influence W, but high W may influence (encourage or discourage) changes in E and D

- Repeated measurements taken on the same individuals can help to disentangle complex interrelationships between all three variables

Age vs cohort effects

- How should we interpret a positive cross-sectional association between age and voting in elections?
  - Perhaps people become more socially responsible as they get older (an age effect)
  - Or perhaps people who grew up during periods of national instability take a greater interest in politics (a cohort effect)
- Longitudinal data can be used to distinguish age and cohort effects
  - Does a person's chance of voting change with age? Is a 20-year old born in 1960 more likely to vote than a 20-year old born in 1970?

Types of longitudinal data

- Repeated measures
  - The same variable is measured on several occasions, e.g. height, test scores, attitudes
  - Collected prospectively
- Event history (duration) data
  - The duration until some event occurs (measured from time of becoming 'at risk' of the event), e.g. duration from marriage to divorce
  - Collected retrospectively or prospectively

This course focuses on the analysis of repeated measured data.

Example research questions and methods of analysis

- How does cognitive development vary between children (GCM)?
- Is there a gender difference in the rate of cognitive development (GCM)?
- What is the effect of income on subsequent mental health (adjusting for prior mental health) (DM)?
- What is the effect of a *change* in income on mental health (adjusting for mental health before income change) (DM)?

GCM - Growth curve model
DM - Dynamic model

Repeated Measures Data;
Introduction to Growth Curve Models

Example: Reading development

- Data from children of female respondents to the U.S. National Longitudinal Survey of Youth [*]
- Reading scores for 221 children on four occasions (only complete cases considered)
- Occasions spaced two years apart (1986, 1988, 1990 and 1992); children aged 6-8 in 1986
- Other variables: antisocial behaviour (also repeated measures), gender, amount of cognitive support at home

Interested in variation between children in their reading development (or reading trajectories)

[*] See `http://www.unc.edu/~curran/srcd-docs/srcdmeth.pdf`

Reading data in wide form

Repeated measures data often come in the form of 1 record per individual, with different measures stored as separate variables.

child	male	homecog	read1	read2	read3	read4
1	1	9	2.1	2.9	4.5	4.5
2	0	9	2.3	4.5	4.2	4.6

Reading data in long form

Most methods of longitudinal data analysis require data to be restructured so there is 1 record per measurement occasion.

child	year	male	homecog	read
1	1	1	9	2.1
1	2	1	9	2.9
1	3	1	9	4.5
1	4	1	9	4.5
2	1	0	9	2.3
2	2	0	9	4.5
2	3	0	9	4.2
2	4	0	9	4.6

It is straightforward to convert data from wide to long format.

Questions about reading development

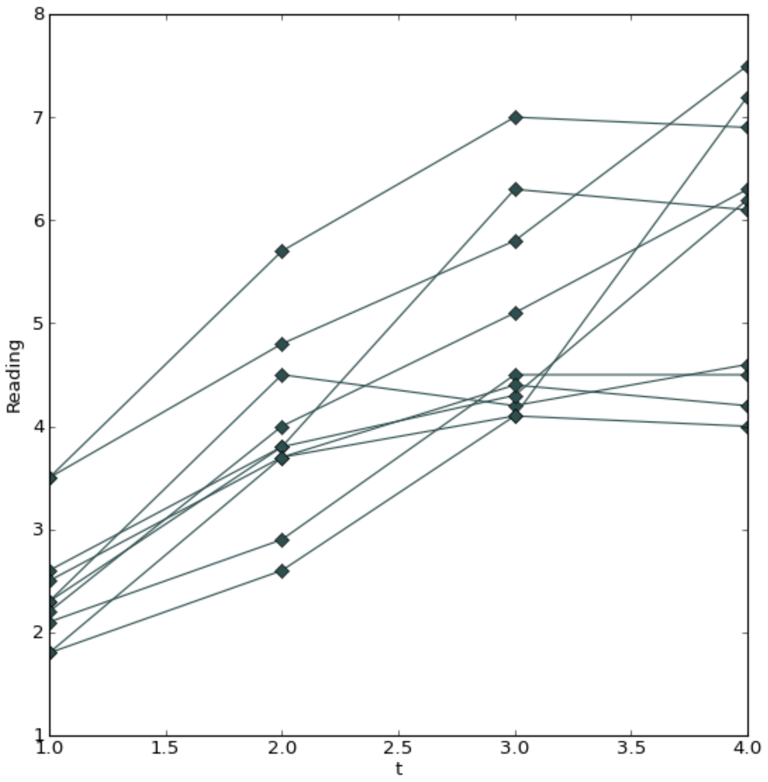
- What is the nature of reading development with age? Linear or nonlinear?
- How much do children vary in their initial reading score and in the rate of development?
- Does the initial score and rate of change depend on child/family characteristics, e.g. amount of cognitive support at home?

Summary statistics for reading by year

	Year			
	1	2	3	4
Mean	2.52	4.04	5.02	5.80
SD	0.88	1.00	1.10	1.22

- Mean reading score increases by year
- But likely to be a large amount of variation between children in reading ability and speed of progress

Observed reading trajectories for 10 children



- Trajectories fairly monotonic (increasing with age), but nonlinear
- Individual variation in level (intercept) and rate of change (slope)

Repeated measures data: Notation

y_{ti} is the response at occasion t ($t = 1, \ldots T$) for individual i ($i = 1, \ldots, n$).

Suppose that z_{ti} is the timing of occasion t for individual i. For simplicity, we assume $z_{ti} = t$ but in many applications time $\equiv$ age, and individuals may vary in age at occasion t.

- Occasions need not be equally spaced
- Individuals may have missing data because of attrition or by design

Can view data as having a 2-level hierarchical structure: responses (level 1) within individuals (level 2).

Basic linear growth model

We begin with a model that allows for individual variation in the level of y for individuals.

Ignoring covariates for now, a simple linear growth curve model is:

$$\begin{aligned} y_{ti} &= \beta_{0i} + \beta_1 t + e_{ti} \\ \beta_{0i} &= \beta_0 + u_{0i} \qquad \text{(individual variation in level of } y\text{)} \end{aligned}$$

where u_{0i} is an individual-specific residual (or random effect) representing unmeasured individual characteristics that are fixed over time and e_{ti} are occasion-level residuals.

Basic linear growth model: Assumptions and estimation

- Usually assume $u_{0i} \sim N(0, \sigma^2_{u0})$ and $e_{ti} \sim N(0, \sigma^2_e)$ are normally distributed.

- Assume $\text{cov}(e_{ti}, e_{si}) = 0$, i.e. correlation between an individual's y-values over time is explained by u_{0i}.

- The model can be viewed as a 'random intercept' multilevel model, where u_{0i} allows the level of y (or intercept) to vary across individuals.

- The growth rate (coefficient of t, β_1) is fixed across individuals, which will usually be an unrealistic assumption.

Application of random intercept model: Reading

Results from MLwiN (Iterative Generalised Least Squares)

Parameter	Est	(SE)
Intercept (β_0)	2.719	(0.068)
Slope of t (β_1)	1.084	(0.020)
Level 2 (between-child) variance (σ^2_{u0})	0.729	(0.080)
Level 1 (within-child) variance (σ^2_e)	0.422	(0.023)
Deviance	2202.68	

After accounting for a common time effect, $0.73/(0.73 + 0.42) = 63\%$ of the variation in y_{ti} is due to differences between individuals

Linear growth model with random slopes

A random intercept model will rarely be realistic for repeated measures data, so we now move to a random slope model.

$$
\begin{aligned}
y_{ti} &= \beta_{0i} + \beta_{1i} t + e_{ti} \\
\beta_{0i} &= \beta_0 + u_{0i} && \text{(individual variation in level of } y\text{)} \\
\beta_{1i} &= \beta_1 + u_{1i} && \text{(individual variation in growth rate)}
\end{aligned}
$$

$$
\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}\right] \quad e_{ti} \sim N(0, \sigma^2_e)
$$

Interpretation

- $y_{ti} = \beta_0 + \beta_1 t$ is the average trajectory (but may not represent trajectory of any individual)
- u_{0i} is the individual departure about the intercept of this line
- u_{1i} is the individual departure about the slope of this line
- σ^2_{u0} is the between-individual variance in the mean of y at $t = 0$ (Code t so that 0 is in observed range, e.g. 1st occasion or mid-point.)
- σ^2_{u1} is the between-individual variance in the growth rate
- σ_{u01} is the covariance between the intercepts and slopes of the individual linear trajectories

Application of random slope model: Reading

Results from MLwiN (IGLS)

Parameter	Est	(SE)
Intercept (β_0)	2.719	(0.057)
Slope of t (β_1)	1.084	(0.024)
Level 2 (between-child)		
Variance of intercepts (σ^2_{u0})	0.516	(0.071)
Intercept-slope covariance (σ_{u01})	0.029	(0.022)
Variance in slopes of t (σ^2_{u1})	0.069	(0.013)
Level 1 (within-child)		
Variance (σ^2_e)	0.306	(0.021)
Deviance	2119.05	

t is recoded $0, 1, 2, 3$ (i.e. centred at $t = 1$) so $\hat{\sigma}^2_{u0} = 0.52$ is the between-child variance at the first occasion

Testing for between-individual variation in growth rate

Compare random intercept and random slope model using a likelihood ratio test.

Null hypothesis is $\text{H}_0 : \sigma^2_{u1} = \sigma_{u01} = 0$

Compare change in deviance (twice the difference in the log-likelihoods between the 2 models) with a chi-squared distribution on 2 d.f.

In reading example, $\text{LR} = 2202.68 - 2119.05 = 83.6$ so strong evidence against the null $\implies$ the random slope model is preferred.

Interpretation of random slope model: Reading

- $\text{read}_{ti} = 2.72 + 1.08t$ is the equation of the fitted average line, but children vary in their intercepts and slopes about this line

- Between-child variance in the mean reading score at 1st occasion (age 6-8) is 0.52

- Between-child variance in slope of time (growth rate) is 0.069

- Covariance between child intercepts and slopes is 0.029 (translates to correlation of $0.029/(\sqrt{0.52 \times 0.069}) = 0.15$

- Within-child variance (between occasions) is 0.31

Interpretation of intercept-slope covariance

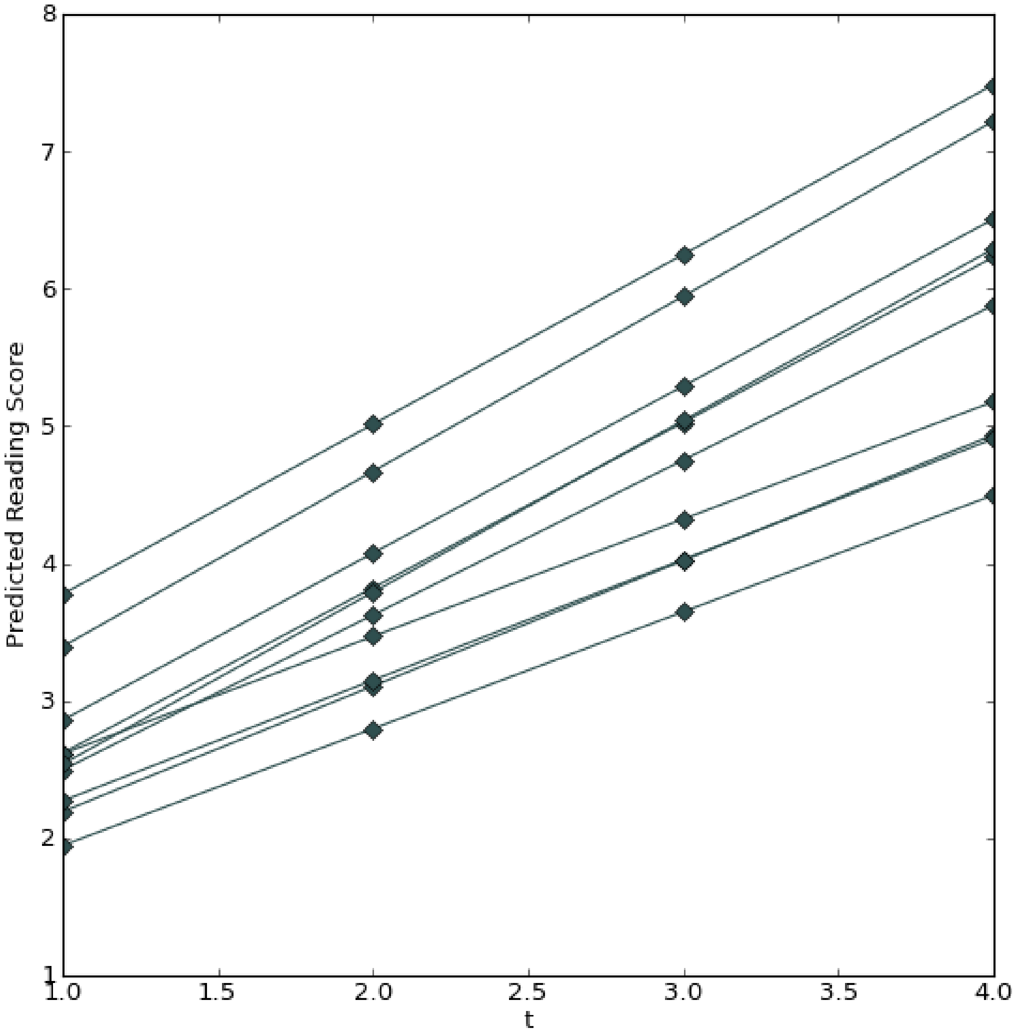
σ_{u01} is the covariance between the intercepts and slopes of the individual linear trajectories

E.g. $\sigma_{u01} > 0$ implies individuals with $u_{0i} > 0$ (above-average y at $t = 0$) tend to have $u_{1i} > 0$

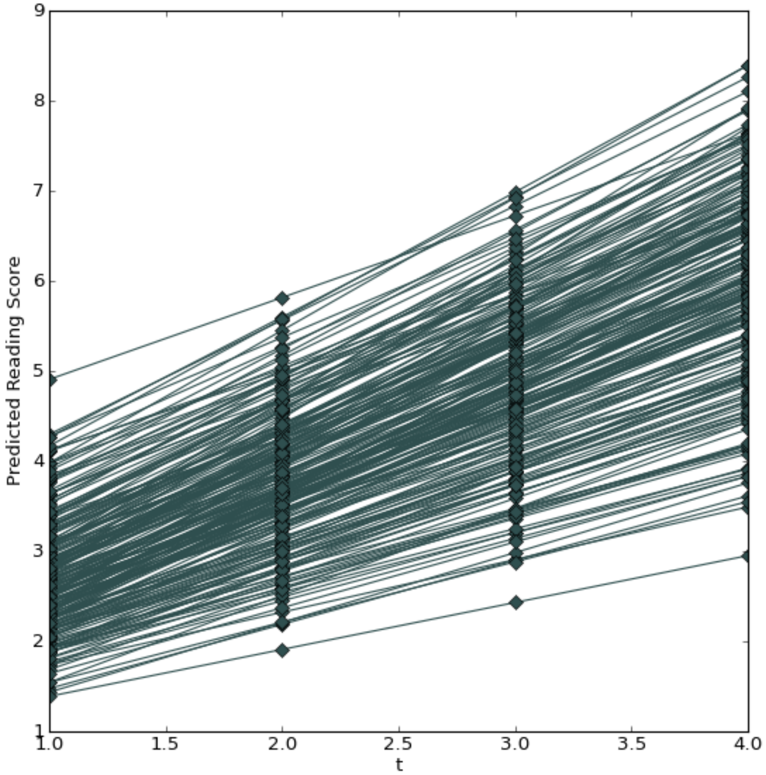
If 'average' slope $\beta_1 > 0$ then $u_{1i} > 0$ suggests a steeper-than-average growth rate

If $\beta_1 < 0$ then $u_{1i} > 0$ suggests a flatter-than-average growth rate

Predicted linear trajectories for first 10 children

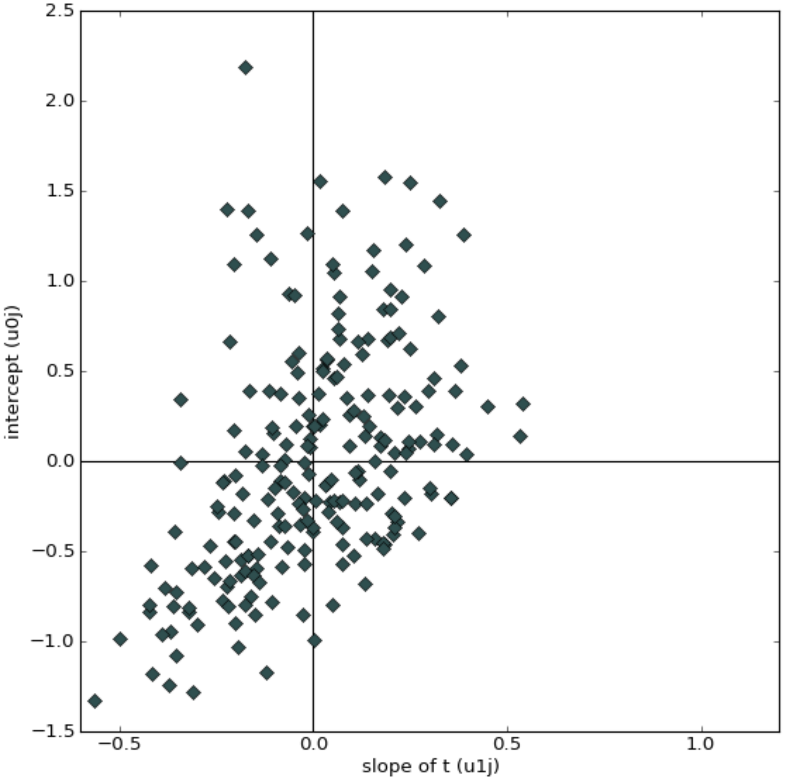


Predicted linear trajectories for all 221 children



Some suggestion that the variability in reading scores increases with t (age).

Reading: Intercept vs slope residuals ($\hat{u}_{0i}$ vs $\hat{u}_{1i}$)



Weak positive correlation: children with above-average scores at $t = 0$ ($u_{0i} > 0$) tend also to progress more quickly ($u_{1i} > 0$).

Between-individual variance

Random slope for t implies that the between-individual variance depends on t:

$$\begin{aligned} \text{var}(u_{0i} + u_{1i}t) &= \text{var}(u_{0i}) + 2\text{cov}(u_{0i}, u_{1i})t + \text{var}(u_{1i})t^2 \\ &= \sigma^2_{u0} + 2\sigma_{u01}t + \sigma^2_{u1}t^2 \end{aligned}$$

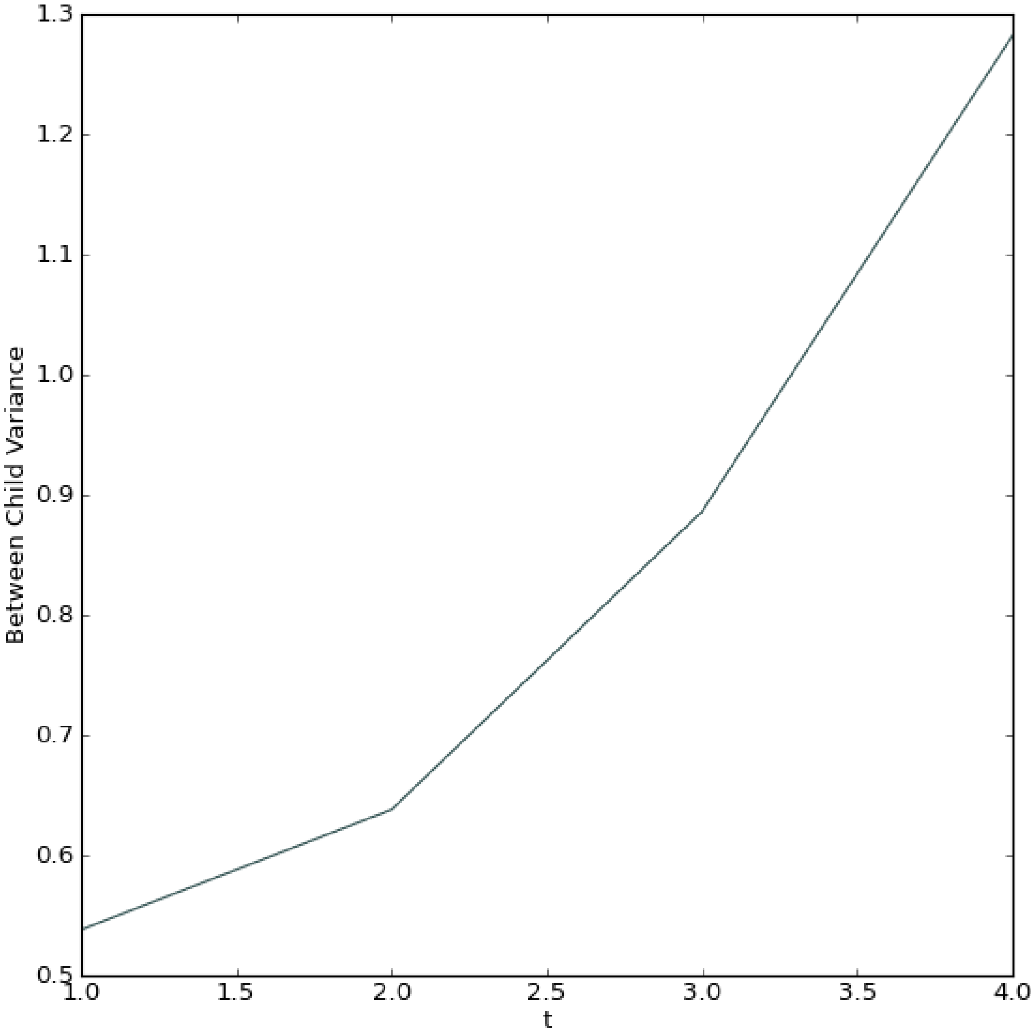
i.e. a quadratic function of time with 'coefficients' σ^2_{u0}, σ_{u01} and σ^2_{u1}.

For children's reading, between-child variance estimated as

$$0.52 + 0.058t + 0.069t^2$$

i.e. strictly increasing with age (as implied by 'fanning out' pattern in prediction lines)

Between-child variance in reading by time (age)



2. An Introduction to MCMC methods and Stat-JR

What will we cover in this first session?

- What is Bayesian Statistics? (as opposed to classical or frequentist statistics)
- What is MCMC estimation?
- MCMC algorithms and Gibbs Sampling
- MCMC diagnostics
- MCMC Model comparisons
- Stat-JR

WHAT IS BAYESIAN STATISTICS?

Why do we need to know about Bayesian statistics?

- In the practicals in this workshop we will be using Stat-JR and primarily MCMC methods which are a family of estimation methods used for fitting realistically complex models.
- MCMC methods are generally used on Bayesian models which have subtle differences to more standard (frequentist) models.
- As most statistical courses are still taught using classical or frequentist methods we need to describe the differences before going on to consider MCMC methods.

Bayesian Inference

In Bayesian inference there is a fundamental distinction between

- Observable quantities *x,* i.e. the data
- Unknown quantities θ

θ can be statistical parameters, missing data, latent variables…

- Parameters are treated as random variables

In the Bayesian framework we make probability statements about model parameters

In the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data.

Prior distributions

As with all statistical analyses we start by positing a model which specifies p(x| θ)

This is the **likelihood** which relates all variables into a ‘**full probability model**’

However from a Bayesian point of view :

- θ is unknown so should have a probability distribution reflecting our uncertainty about it before seeing the data
- Therefore we specify a **prior distribution** p(θ)

Non-informative priors

We often do not have any prior information, although true Bayesian’s would argue we always have some prior information!

We would hope to have good agreement between the frequentist approach and the Bayesian approach with a non-informative prior.

Diffuse or flat priors are often better terms to use as no prior is strictly non-informative!

For an example with an unknown mean, candidate priors are a Uniform distribution over a large range or a Normal distribution with a huge variance.

Posterior Distributions

Also *x* is known so should be conditioned on and here we use Bayes theorem to obtain the conditional distribution for unobserved quantities given the data which is known as the **posterior distribution**.

$$p(\theta \mid x) = \frac{p(\theta)\,p(x \mid \theta)}{\int p(\theta)\,p(x \mid \theta)\,d\theta = (p(x))} \propto p(\theta)\,p(x \mid \theta)$$

← Bayes Thereom

The prior distribution expresses our uncertainty about θ **before** seeing the data.

The posterior distribution expresses our uncertainty about θ **after** seeing the data.

Point and Interval Estimation

In Bayesian inference the outcome of interest for a parameter is its full posterior distribution however we may be interested in summaries of this distribution.

A simple point estimate would be the mean of the posterior. (although the median and mode are alternatives.)

Interval estimates are also easy to obtain from the posterior distribution and are given several names, for example credible intervals, Bayesian confidence intervals and Highest density regions (HDR). All of these refer to the same quantity.

MCMC METHODS

How does one fit models in a Bayesian framework?

Let us now consider a simple linear regression:

$$weight_i = \beta_0 + \beta_1 height_i + e_i$$

$$e_i \sim N(0, \sigma^2)$$

With conjugate priors:

$$\beta_0 \sim N(0, m_0), \beta_1 \sim N(0, m_1),$$

$$\sigma^2 \sim \Gamma^{-1}(\varepsilon, \varepsilon)$$

$$\text{where } m_0 = m_1 = 10^6, \varepsilon = 10^{-3}$$

So our goal now is to make inferences on the joint posterior distribution:

$$p(\beta_0, \beta_1, \sigma^2 \mid y)$$

MCMC Methods

Goal: To sample from joint posterior distribution:

$$p(\beta_0, \beta_1, \sigma^2 \mid y)$$

Problem: For complex models this involves multidimensional integration

Solution: It may be possible to sample from conditional posterior distributions,

$$p(\beta_0 \mid y, \beta_1, \sigma^2), p(\beta_1 \mid y, \beta_0, \sigma^2), p(\sigma^2 \mid y, \beta_0, \beta_1)$$

It can be shown that after *convergence* such a sampling approach generates dependent samples from the joint posterior distribution.

Gibbs Sampling

When we can sample directly from the conditional posterior distributions then such an algorithm is known as Gibbs Sampling.

This proceeds as follows for the linear regression example:

Firstly give all unknown parameters starting values,

$$\beta_0(0), \beta_1(0), \sigma^2(0).$$

Next loop through the following steps:

Gibbs Sampling ctd.

Sample from

$p(\beta_0 \mid y, \beta_1(0), \sigma^2(0))$ to generate $\beta_0(1)$ and then from

$p(\beta_1 \mid y, \beta_0(1), \sigma^2(0))$ to generate $\beta_1(1)$ and then from

$p(\sigma^2 \mid y, \beta_0(1), \beta_1(1))$ to generate $\sigma^2(1)$.

These steps are then repeated with the generated values from this loop replacing the starting values. The chain of values produced by this procedure is known as a Markov chain, and it is hoped that this chain converges to its equilibrium distribution which is the joint posterior distribution.

Calculating the conditional distributions

In order for the algorithm to work we need to sample from the conditional posterior distributions.

If these distributions have standard forms then it is easy to draw random samples from them.

Mathematically we write down the full posterior and assume all parameters are constants apart from the parameter of interest.

We then try to match the resulting formulae to a standard distribution.

The Stat-JR software gives the derivations!

Algorithm Summary

Repeat the following three steps

1. Generate β_0 from its Normal conditional distribution.
2. Generate β_1 from its Normal conditional distribution.
3. Generate $1/\sigma^2$ from its Gamma conditional distribution

Convergence and burn-in

Two questions that immediately spring to mind are:

1. We start from arbitrary starting values so when can we safely say that our samples are from the correct distribution?
2. After this point how long should we run the chain for and store values?

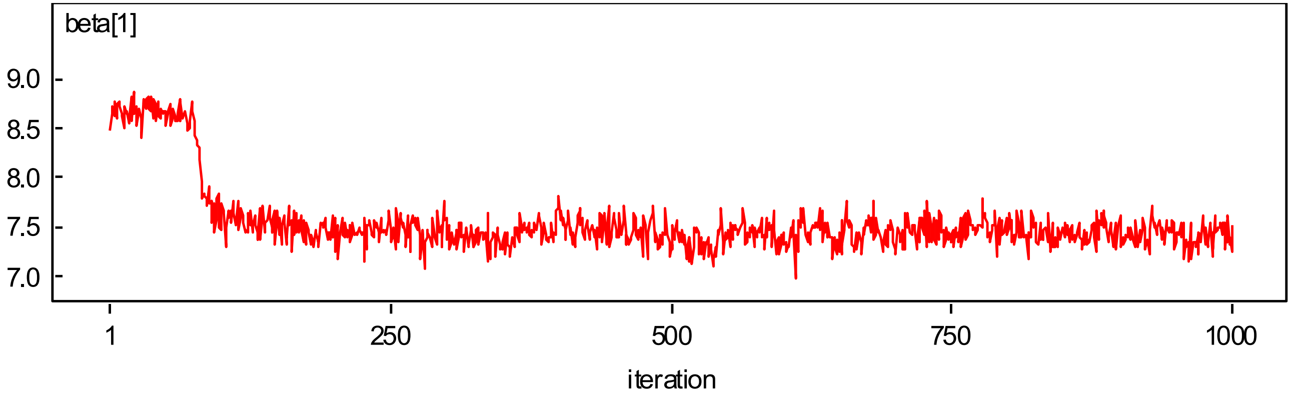
MCMC DIAGNOSTICS

Checking Convergence

This is the researchers responsibility!

Convergence is to a target **distribution** (the required posterior), not to a single value as in ML methods.

Once convergence has been reached, samples should look like a random scatter about a stable mean value.



Convergence occurs here at around 100 iterations.

How many iterations after convergence?

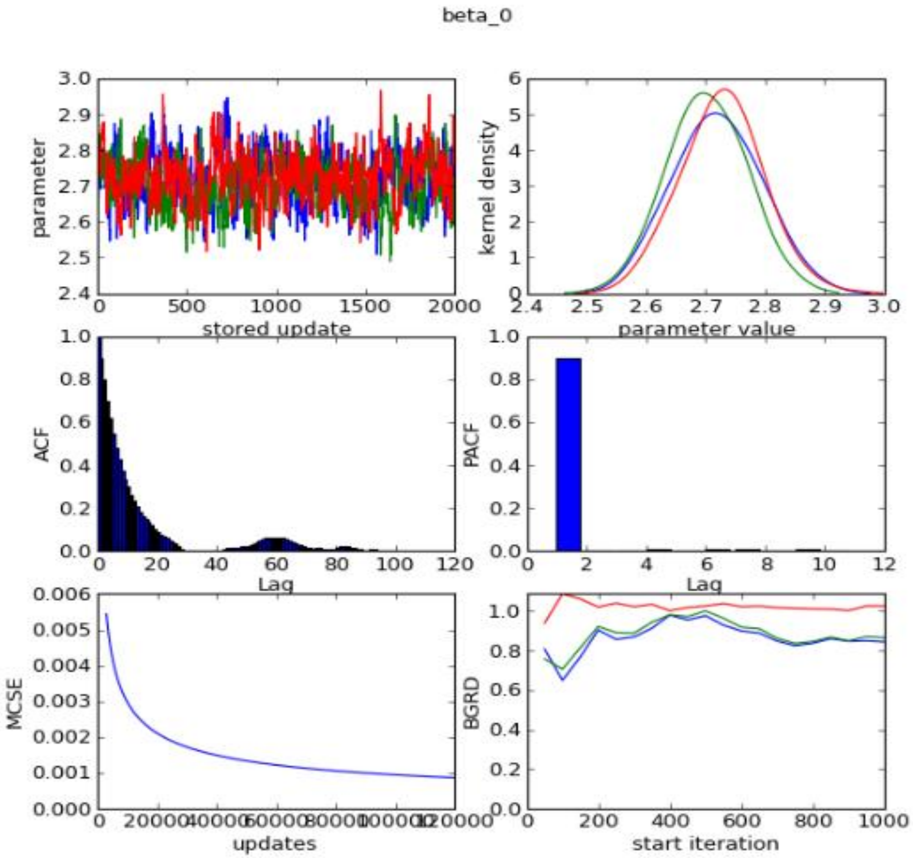
After convergence, further iterations are needed to obtain samples for posterior inference.

More iterations = more accurate posterior estimates.

MCMC chains are dependent samples and so the dependence or autocorrelation in the chain will influence how many iterations we need.

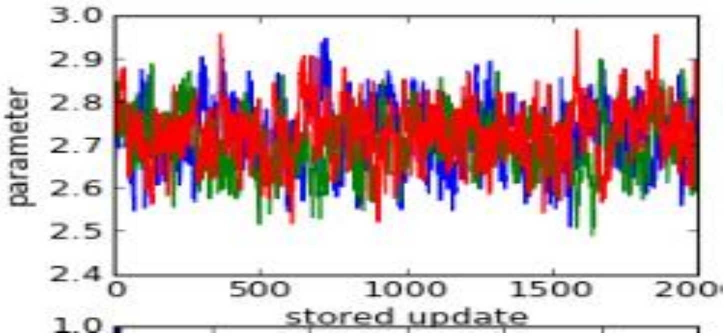
Accuracy of the posterior estimates can be assessed by the Monte Carlo standard error (MCSE) for each parameter.

MCMC diagnostics (Example from Stat-JR)



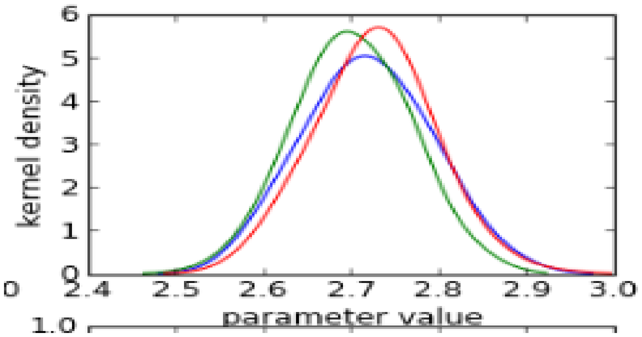
We will describe each pane separately – Note MLwiN has similar six way plots!

Trace plot



This graph plots the generated values of the parameter against the iteration number.
When multiple chains are run as here each chain is a different colour.
A crude test of mixing with 1 chain is the ‘blue finger’ test.
Here the red chain was plotted last so as mixing improves the plot becomes redder.
These chains don’t mix that well but could be worse!

Kernel Density plot



This plot is like a smoothed histogram.

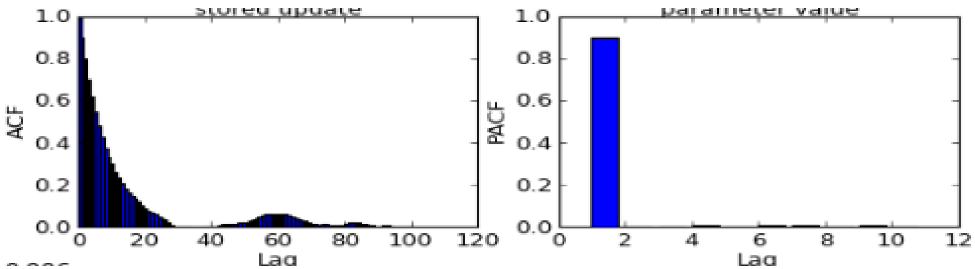
Instead of counting the estimates into bins of particular widths like a histogram, the effect of each iteration is spread around the estimate via a Kernel function e.g. a normal distribution.

This means that at each point we get the sum of the Kernel function parts for each iteration.

The Kernel density plot has a smoothness parameter that can be modified.

With multiple chains we hope each kernel plot is the same – here we see some variability due to short chain lengths.

Time series diagnostics



Here we have the Auto correlation function (ACF) and partial autocorrelation function (PACF) plots.

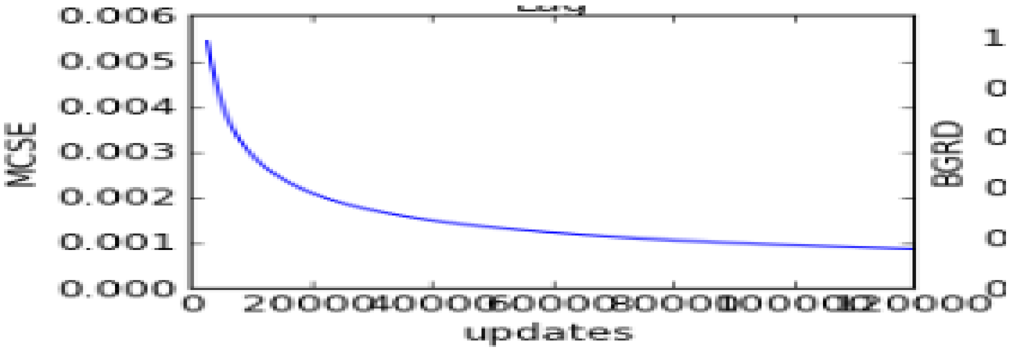
The ACF measures how correlated the values in the chain are with their close neighbours. The lag is the distance between the two chains to be compared.

An independent chain will have approximately zero autocorrelation at each lag.

A Markov chain should have a power relationship in the lags i.e. if ACF(1) = ρ then ACF(2) = ρ^2 etc. This is known as an AR(1) process.

The PACF measures discrepancies from such a process and so should normally have values 0 after lag 1.

Monte Carlo Standard Error



The Monte Carlo Standard Error (MCSE) is an indication of how much error is in the estimate due to the fact that MCMC is used.

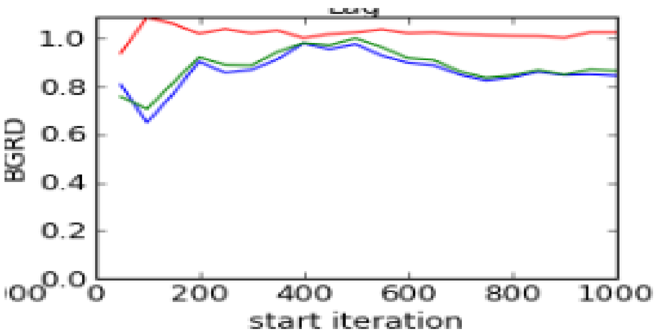
As the number of iterations increases the MCSE→0.

For an independent sampler it equals the SD/√n.

However it is adjusted due to the autocorrelation in the chain.

The graph above gives estimates for the MCSE for longer runs.

Brooks Gelman Rubin diagnostic



A multiple chain diagnostic that looks at how well the chains converge to the same distribution.

The green and blue lines show the between and within chain variability and the ratio shown in red is the diagnostic which should converge to 1.0.

Summary Statistics and Accuracy Diagnostics (available from Summary Statistics template)

Statistics given are:

- Mean and SD – from the chain.
- Median – by sorting the chain and finding the middle value.
- Other quantiles for 90% and 95% CIs as well as minimum, maximum and IQR. Thus one can obtain a non-symmetric interval.

Accuracy diagnostics:

- Brooks-Draper works on quoting the mean to *n* significant figures. It’s formulae uses the estimate, it’s s.d. and the ACF and it can often give very small or very large values!
- The ESS will be discussed next.

Effective Sample Size

This quantity gives an estimate of the equivalent number of independent iterations that the chain represents.

This is related to the ACF and the MCSE.

Its formula is: $n/\kappa \text{ where } \kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k).$

In Stat-JR it is given in the ModelResults object as well as summary statistics and for the parameter above is 420 despite running for 3 chains of 2,000 iterations each!

Inference using posterior samples from MCMC runs

A powerful feature of MCMC and the Bayesian approach is that all inference is based on the joint posterior distribution.

We can therefore address a wide range of substantive questions by appropriate summaries of the posterior.

Typically report either the mean or median of the posterior samples for each parameter of interest as a point estimate

2.5% and 97.5% percentiles of the posterior sample for each parameter give a 95% posterior credible interval (interval within which the parameter lies with probability 0.95)

Derived Quantities

Once we have a sample from the posterior we can answer lots of questions simply by investigating this sample.

Examples:

What is the probability that $\theta > 0$?

What is the probability that $\theta_1 > \theta_2$?

What is a 95% interval for $\theta_1/(\theta_1 + \theta_2)$?

MODEL COMPARISON

Model Comparison in MCMC

In frequentist statistics there are many options including:

- Likelihood ratio (deviance) tests
- Wald Tests
- Information Criterion – e.g. AIC/BIC

Here we look at a criterion that can be used with MCMC and which for a linear regression model is equivalent to the AIC – the Deviance information criterion (DIC).

DIC

A natural way to compare models is to use a criterion based on a trade-off between the fit of the data to the model and the corresponding complexity of the model.

DIC does this in a Bayesian way.

DIC = ‘goodness of fit’ + ‘complexity’.

Fit is measured by deviance $D(\theta) = -2\log L(data \mid \theta)$

Complexity is measured by an estimate of the ‘effective number of parameters’ defined as

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta])$$
$$= \overline{D} - D(\overline{\theta})$$

i.e. Posterior mean deviance minus the deviance evaluated at the posterior mean of the parameters.

DIC (continued)

The DIC is then defined analagously to AIC as

$$DIC = D(\overline{\theta}) + 2p_D$$
$$= \overline{D} + p_D$$

Models with smaller DIC are better supported by the data.

- DIC is available in Stat-JR in the ModelResults object.
- DIC can be monitored in other packages such as MLwiN under the Model/MCMC menu and WinBUGS from (Inference/DIC menu).

Some guidance on DIC

- any decrease in DIC suggests a better model
- But stochastic nature of MCMC; so, with small difference in DIC you should confirm if this is a real difference by checking the results with different seeds and/or starting values.
- More experience with AIC, and common rules of thumb………
- A model with a Δ value within 1-2 of the best model has substantial support in the data, and should be considered along with the best model.
- A Δ value within 4-7 of the best model has considerably less support.
- A Δ value > 10 indicates that the worse model has virtually no support and can be omitted from further consideration.

SUMMARY & COMPARISON WITH FREQUENTIST APPROACH

Markov chain Monte Carlo (MCMC)

- MCMC methods are Bayesian estimation techniques which can be used to estimate multilevel models
- MCMC works by drawing a random sample of values for each parameter from its probability distribution
- The mean and standard deviation of each random sample gives the point estimate and standard error for that parameter

Estimating a model using MCMC estimation

- We start by specifying the **model** and our **prior** knowledge for each parameter (nearly always no knowledge!)
- Next we specify **initial values** for the model parameters
- We then run the MCMC algorithm and obtain the **parameter chains**
- We discard the initial **burn-in** iterations when the chains are settling down (converging to their **posterior** distributions)
- Summary statistics for the remaining **monitoring iterations** are then calculated:
  - Point estimates and standard errors are given by the means and standard deviations of the chains

Frequentist (IGLS) vs. MCMC (1)

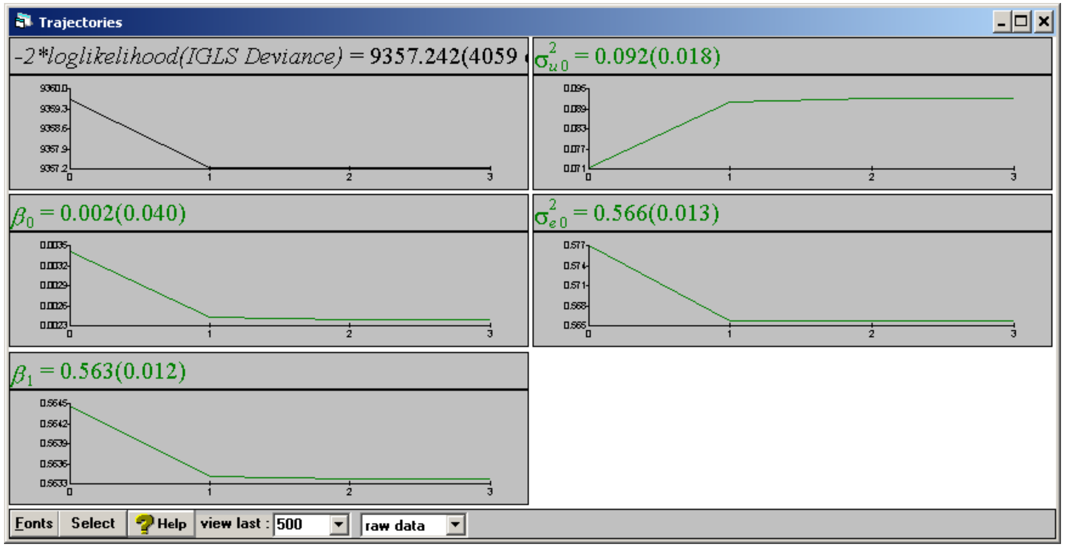
IGLS	MCMC
Fast	Slow
Uses MQL/PQL approximations to fit discrete response models, which can sometimes produce biased estimates	Produces unbiased estimates
Cannot incorporate prior information	Can incorporate prior information

- Note that in practice we often do not incorporate prior information
- We want to protect our inferences from being influenced by our prior beliefs
  - True Bayesians have a very different take

IGLS vs. MCMC (2)

IGLS	MCMC
Confidence intervals based on normality are unreasonable for variance parameters	Normality not assumed
Hard to calculate confidence intervals for functions of parameters	Easy to calculate confidence intervals for arbitrarily complex functions of parameters
Difficult to extend to new models	Easy to extend
Model convergence is judged for you	You have to judge model convergence for yourself

IGLS vs. MCMC (3)



Trajectories

-2*loglikelihood(IGLS Deviance) = 9357.242(4059

σ^2_{u0} = 0.092(0.018)

β_0 = 0.002(0.040)

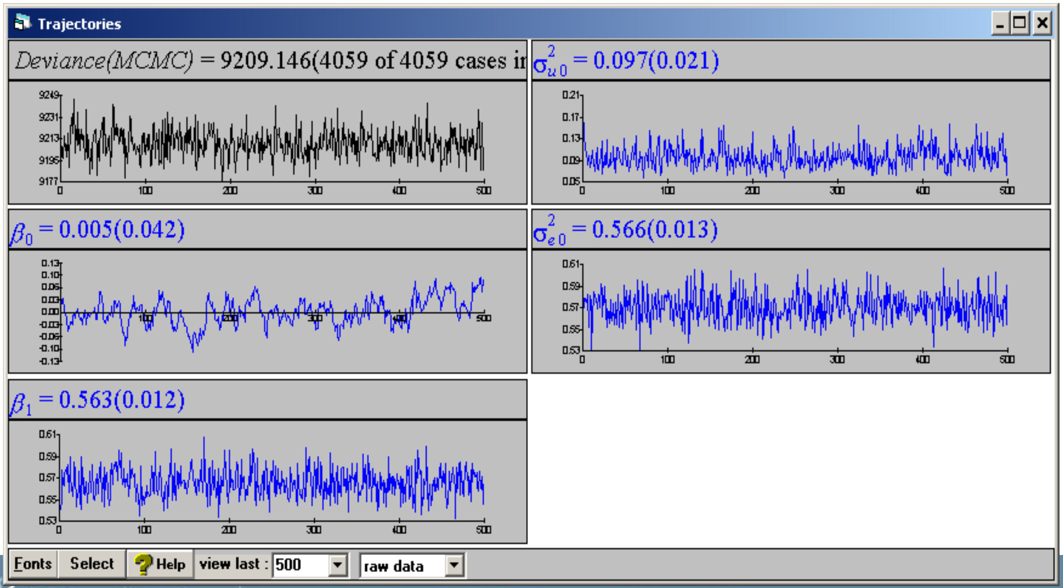
σ^2_{e0} = 0.566(0.013)

β_1 = 0.563(0.012)

Fonts Select Help view last : 500 raw data

IGLS algorithm converges deterministically to a point

Convergence is therefore judge for you



Trajectories

Deviance(MCMC) = 9209.146(4059 of 4059 cases in

σ^2_{u0} = 0.097(0.021)

β_0 = 0.005(0.042)

σ^2_{e0} = 0.566(0.013)

β_1 = 0.563(0.012)

Fonts Select Help view last : 500 raw data

MCMC algorithm converges stochastically to the equilibrium probability distribution

You have to judge convergence for yourself

Priors

- Our prior knowledge for each parameter is summarised by a probability distribution referred to as the **prior distribution**
  - Typically, we specify that we have no prior knowledge as we like the ‘data to speak for it self’
  - We therefore specify **vague**, **diffuse** or **uninformative priors**

$$\beta_1 \sim N(0{,}10000) \approx U(-\infty, \infty)$$

MCMC samplers

- At the t^{th} iteration we want to sample from the posterior distribution of each parameter in turn
  - If we can write down an analytical expression for the posterior distribution then we can use **Gibbs sampling**
    - Computationally efficient algorithm
    - Continuous response models
  - If we can't write down an analytical expression for the posterior then we use **Metropolis-Hastings** sampling

Deviance information criterion (DIC) for model comparison

- **DIC** can be viewed as an **AIC** or **BIC** statistic for MCMC
- DIC balances goodness of fit and model complexity (i.e. deviance and number of parameters)
- Want to maximise fit and minimise complexity
  - – Lower deviance and fewer parameters
- So "better" models have smaller DIC
- Note that the DIC does not have universal approval!

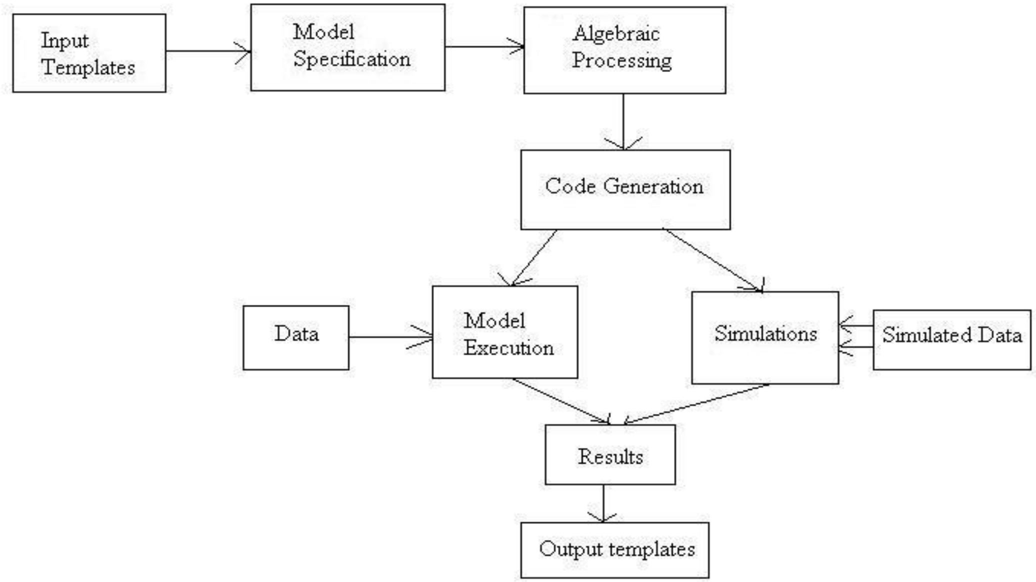
STAT-JR

Stat-JR

- Stat-JR is a new statistical software package named after our former colleague Jon Rasbash
- It is based around the concept of templates that perform a specific statistical task and can be slotted together to form a statistical software package.
- The big vision was an all-singing all-dancing system where expert users could add functionality easily and which interoperates with other software. Stat-JR has an underpinning algebra system which can be used to create model fitting templates.

STAT-JR component based approach

Below is an early diagram of how we envisioned the system. Here you will see boxes representing components some of which are built into the STAT-JR system. The system is written in Python with currently a VB.net algebra processing system. A team of coders work together on the system.



Templates

Backbone of Stat-JR.

Consist of a set of code sections for advanced users to write. A bit like R packages.

For a model template it consists of at least:

- an *inputs* method which specifies inputs and types
- A *model* method that creates (BUGS like) model code for the algebra system
- An (optional) *latex* method can be used for outputting LaTeX code for the model.

Other optional functions required for more complex templates

Regression 1 Example

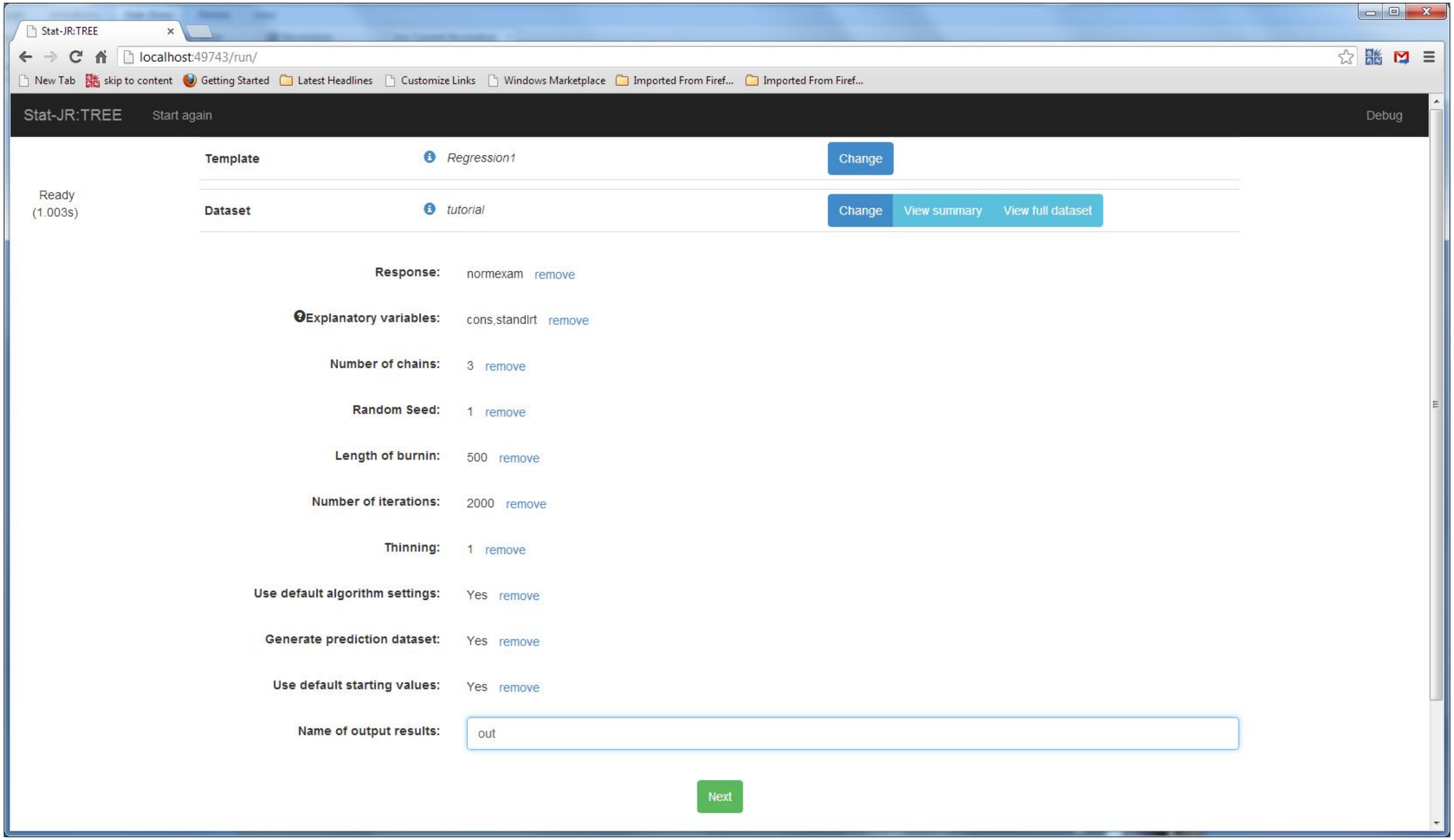
```
from EStat.Templating import *

class Regression1(Template):
    'A model template for fitting 1 level Normal multiple
         regression model in eStat only.‘
 tags = [ 'Model', '1-Level', 'eStat', 'Normal' ]
 engines = ['eStat']
 inputs = '''
y = DataVector('Response: ')
x = DataMatrix('Explanatory variables: ', allow_cat=True,
         help= 'predictor variables')
beta = ParamVector(parents=[x], as_scalar=True)
tau = ParamScalar()
sigma = ParamScalar(modelled = False)
sigma2 = ParamScalar(modelled = False)
deviance = ParamScalar(modelled = False)
'''
```

```
 model = '''
model{
    for (i in 1:length(${y})) {
        ${y}[i] ~ dnorm(mu[i], tau)
        mu[i] <- ${mmult(x, 'beta', 'i')}
    }

    # Priors
    % for i in range(0, x.ncols()):
    beta${i} ~ dflat()
    % endfor
    tau ~ dgamma(0.001000, 0.001000)
    sigma2 <- 1 / tau
    sigma <- 1 / sqrt(tau)
}
'''
       latex = r'''
\begin{aligned}
 \mbox{${y}}_i & \sim \mbox{N}(\mu_i, \sigma^2) \\
\mu_i & =
  ${mmulttex(x, r'\beta', 'i')} \\
%for i in range(0, len(x)):
\beta_${i} & \propto 1 \\
%endfor
\tau & \sim \Gamma (0.001,0.001) \\
\sigma^2 & = 1 / \tau
\end{aligned}
'''
```

An example of STAT-JR – setting up a model



Template Regression1 Change

Dataset tutorial Change View summary View full dataset

Response: normexam remove

Explanatory variables: cons,standlrt remove

Number of chains: 3 remove

Random Seed: 1 remove

Length of burnin: 500 remove

Number of iterations: 2000 remove

Thinning: 1 remove

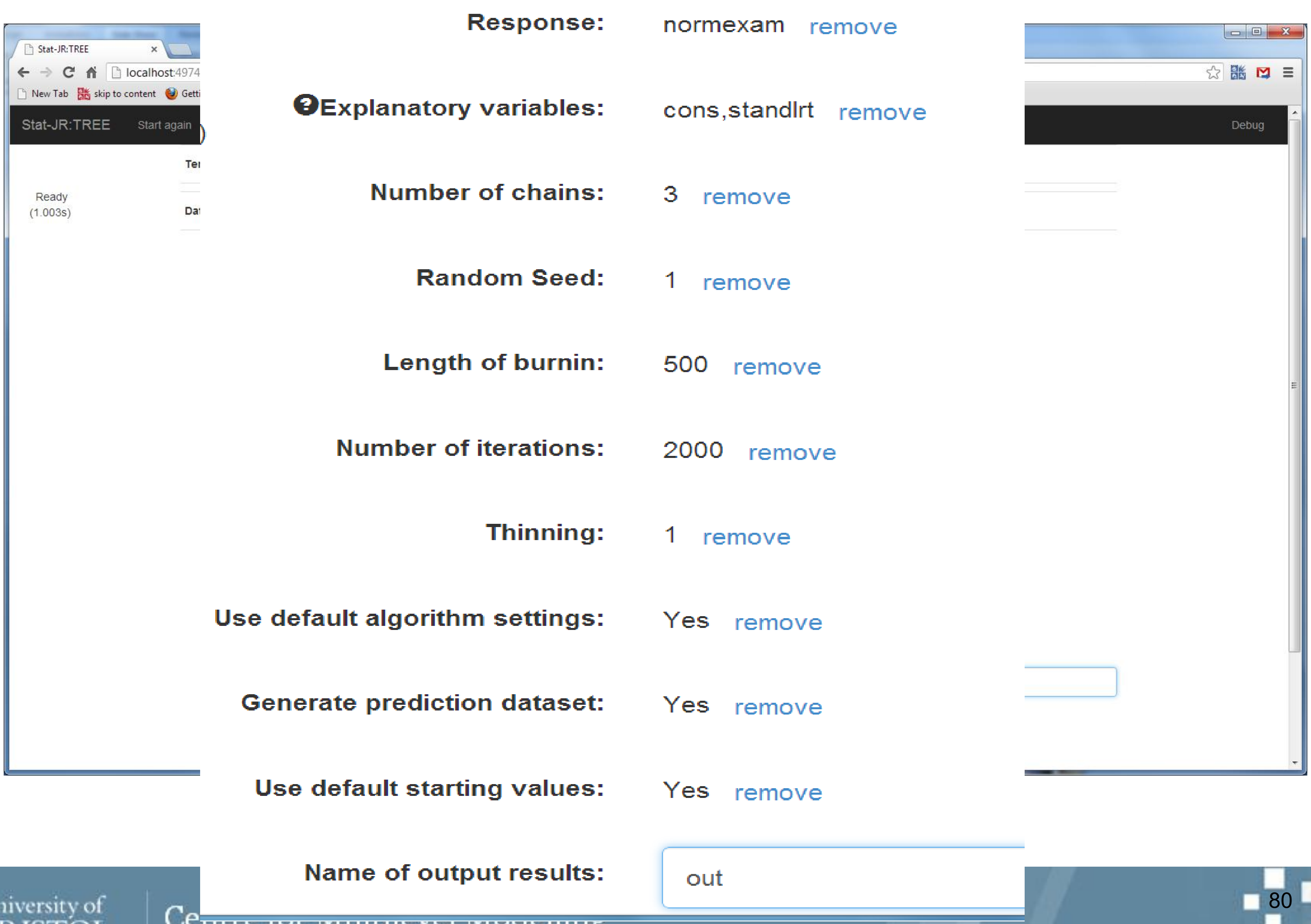
Use default algorithm settings: Yes remove

Generate prediction dataset: Yes remove

Use default starting values: Yes remove

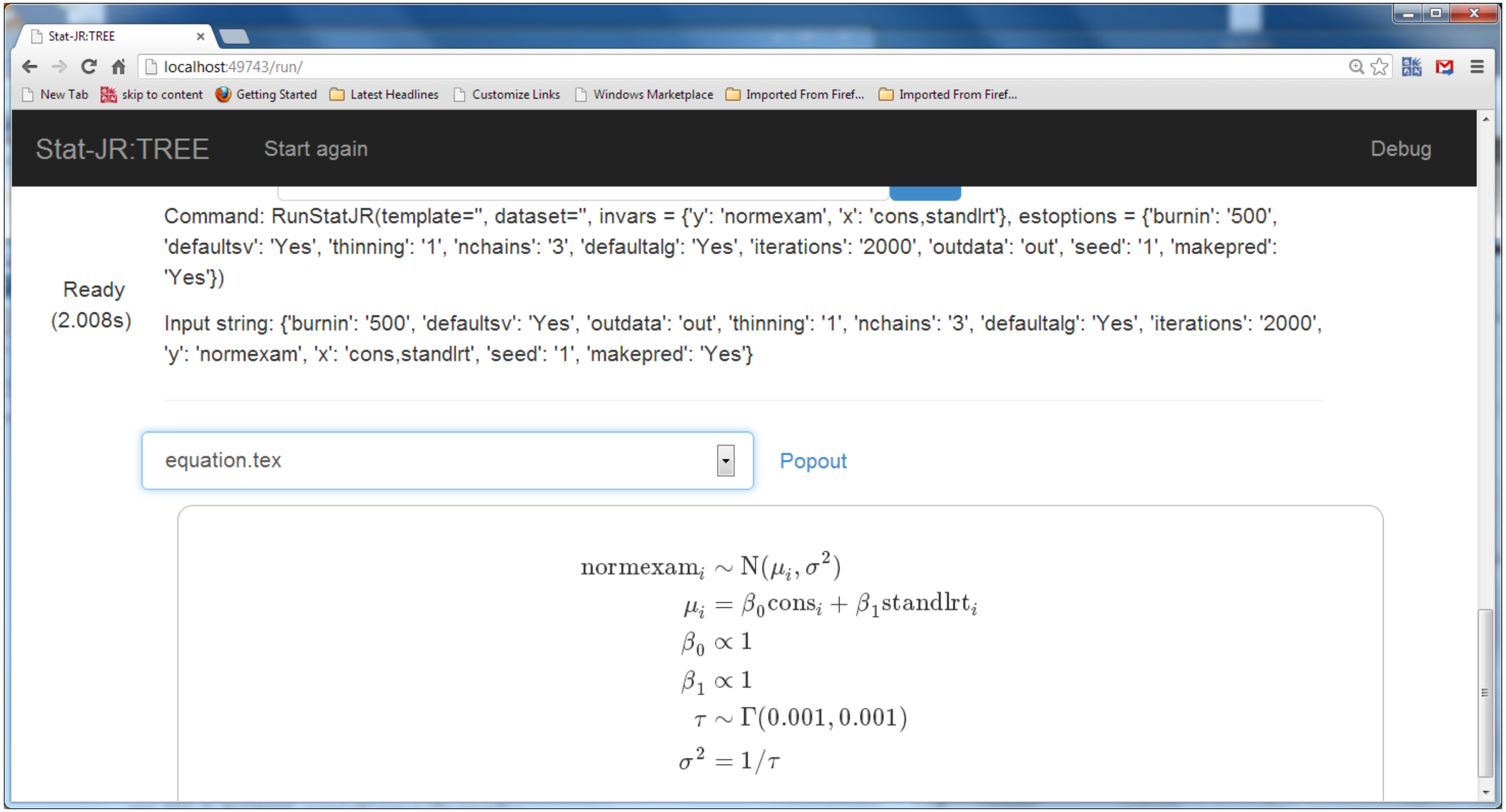
Name of output results: out

An example of STAT-JR – setting up a model



Response:	normexam remove
Explanatory variables:	cons,standlrt remove
Number of chains:	3 remove
Random Seed:	1 remove
Length of burnin:	500 remove
Number of iterations:	2000 remove
Thinning:	1 remove
Use default algorithm settings:	Yes remove
Generate prediction dataset:	Yes remove
Use default starting values:	Yes remove
Name of output results:	out

Equations for model and model code



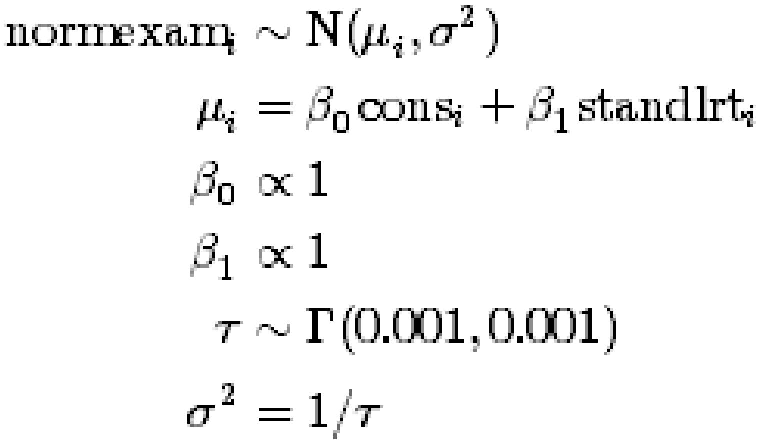
Command: RunStatJR(template='', dataset='', invars = {'y': 'normexam', 'x': 'cons,standlrt'}, estoptions = {'burnin': '500', 'defaultsv': 'Yes', 'thinning': '1', 'nchains': '3', 'defaultalg': 'Yes', 'iterations': '2000', 'outdata': 'out', 'seed': '1', 'makepred': 'Yes'})

Input string: {'burnin': '500', 'defaultsv': 'Yes', 'outdata': 'out', 'thinning': '1', 'nchains': '3', 'defaultalg': 'Yes', 'iterations': '2000', 'y': 'normexam', 'x': 'cons,standlrt', 'seed': '1', 'makepred': 'Yes'}

$$
\begin{aligned}
\mathrm{normexam}_i &\sim \mathrm{N}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 \mathrm{cons}_i + \beta_1 \mathrm{standlrt}_i \\
\beta_0 &\propto 1 \\
\beta_1 &\propto 1 \\
\tau &\sim \Gamma(0.001, 0.001) \\
\sigma^2 &= 1/\tau
\end{aligned}
$$

All objects created available from one pull down and can be popped out to separate tabs in browser.

Equations for model and model code



$$
\begin{aligned}
\mathrm{normexam}_i &\sim \mathrm{N}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 \mathrm{cons}_i + \beta_1 \mathrm{standlrt}_i \\
\beta_0 &\propto 1 \\
\beta_1 &\propto 1 \\
\tau &\sim \Gamma(0.001, 0.001) \\
\sigma^2 &= 1/\tau
\end{aligned}
$$

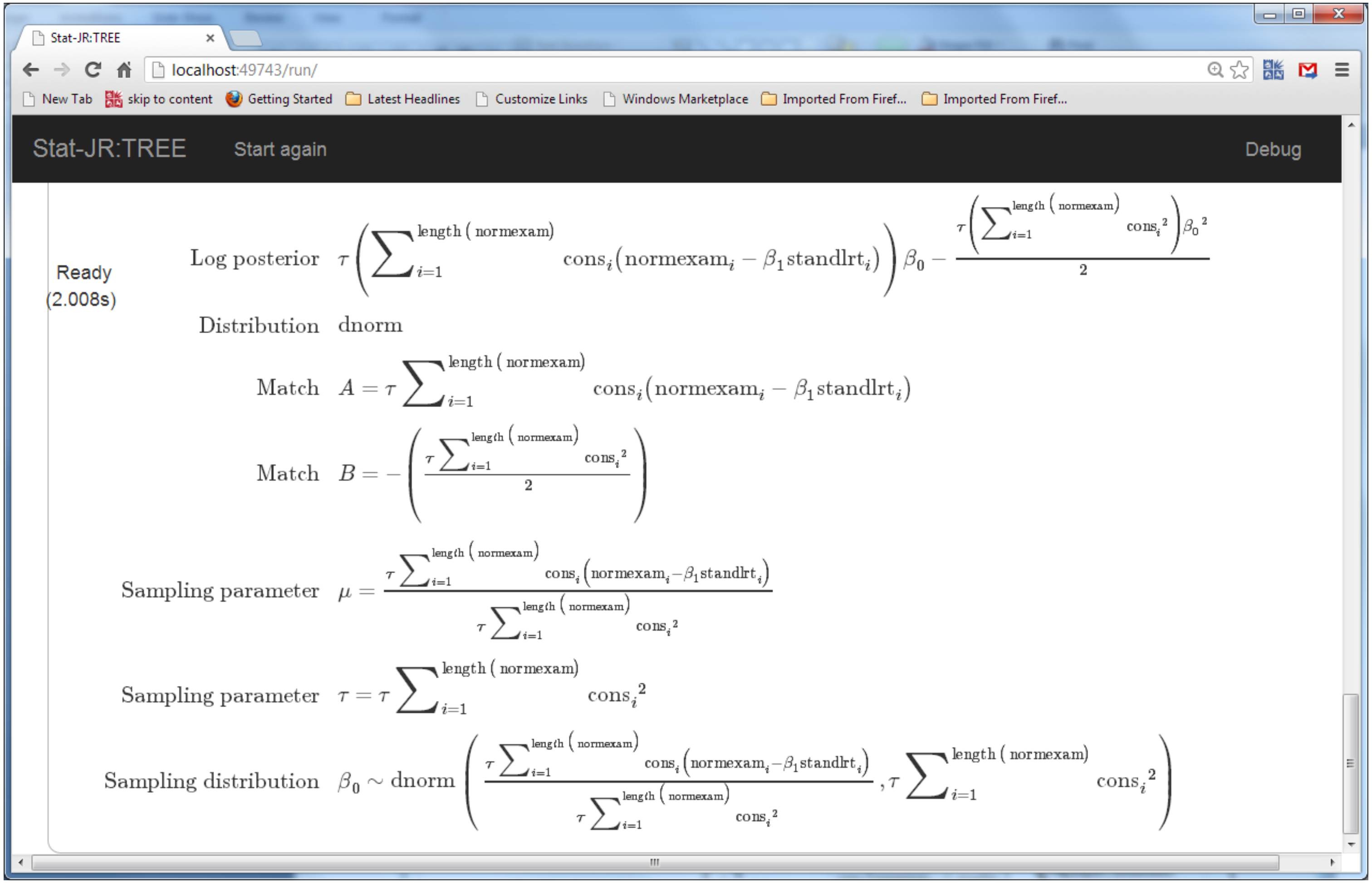
- Note: Equations use MATHJAX and so underlying LaTeX can be copied and paste. The model code is based around the WinBUGS language with some variation.

Model code in detail

```
model{
   for (i in 1:length(normexam)) {
      normexam[i] ~ dnorm(mu[i], tau)
      mu[i] <- cons[i] * beta0 + standlrt[i] * beta1
 }
# Priors
   beta0 ~ dflat()
   beta1 ~ dflat()
   tau ~ dgamma(0.001000, 0.001000)
   sigma2 <- 1 / tau
   sigma <- 1/sqrt(tau)
}
```

For this template the code is, aside from the length function, standard WinBUGS model code.

Algebra system steps



Ready (2.008s)

Log posterior $\tau\left(\sum_{i=1}^{\text{length}(\text{normexam})} \text{cons}_i\left(\text{normexam}_i - \beta_1\text{standlrt}_i\right)\right)\beta_0 - \frac{\tau\left(\sum_{i=1}^{\text{length}(\text{normexam})} {\text{cons}_i}^2\right){\beta_0}^2}{2}$

Distribution dnorm

Match $A = \tau\sum_{i=1}^{\text{length}(\text{normexam})} \text{cons}_i\left(\text{normexam}_i - \beta_1\text{standlrt}_i\right)$

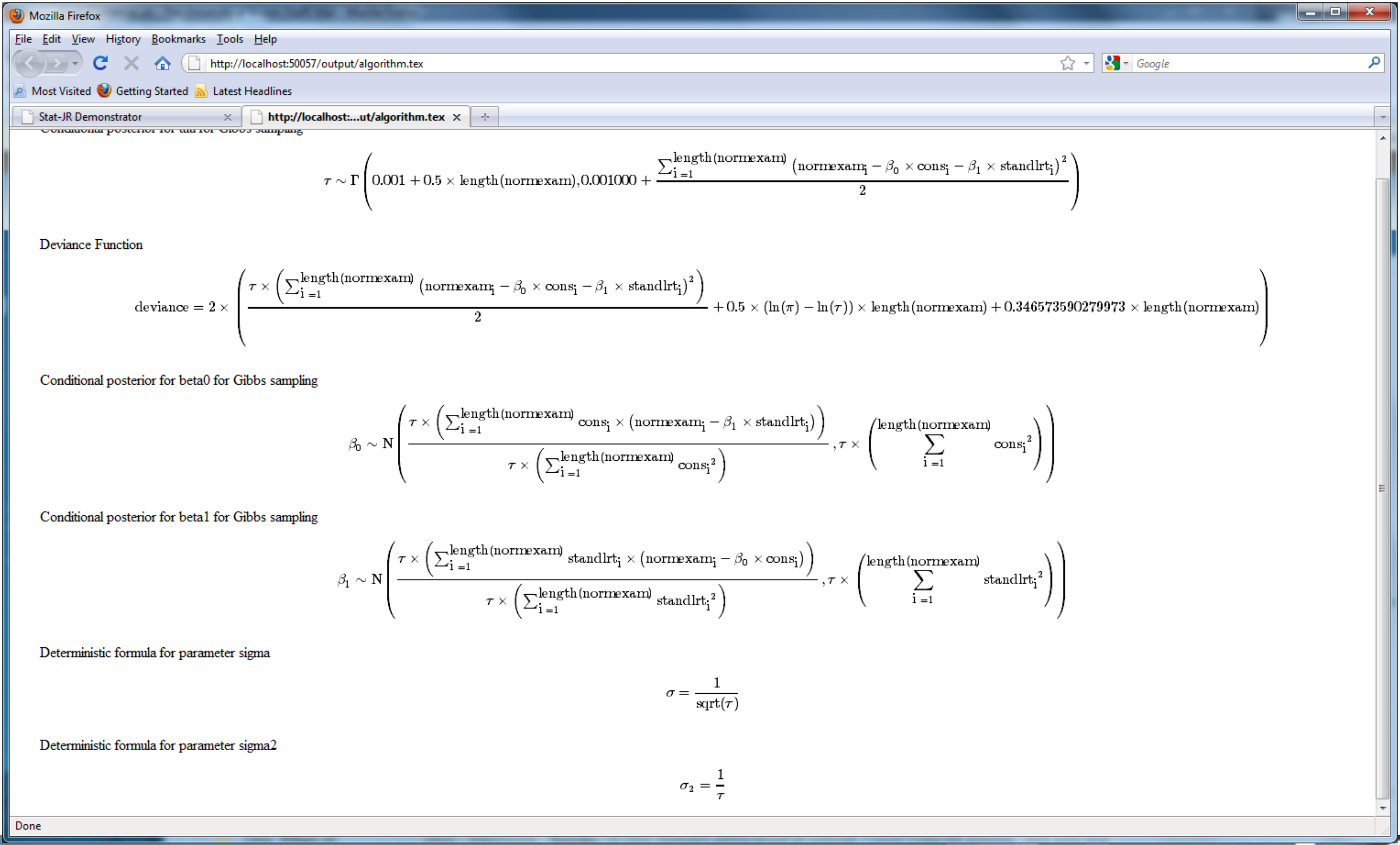
Match $B = -\left(\frac{\tau\sum_{i=1}^{\text{length}(\text{normexam})} {\text{cons}_i}^2}{2}\right)$

Sampling parameter $\mu = \frac{\tau\sum_{i=1}^{\text{length}(\text{normexam})} \text{cons}_i\left(\text{normexam}_i - \beta_1\text{standlrt}_i\right)}{\tau\sum_{i=1}^{\text{length}(\text{normexam})} {\text{cons}_i}^2}$

Sampling parameter $\tau = \tau\sum_{i=1}^{\text{length}(\text{normexam})} {\text{cons}_i}^2$

Sampling distribution $\beta_0 \sim \text{dnorm}\left(\frac{\tau\sum_{i=1}^{\text{length}(\text{normexam})} \text{cons}_i\left(\text{normexam}_i - \beta_1\text{standlrt}_i\right)}{\tau\sum_{i=1}^{\text{length}(\text{normexam})} {\text{cons}_i}^2}, \tau\sum_{i=1}^{\text{length}(\text{normexam})} {\text{cons}_i}^2\right)$

Algebra system steps


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:50057/output/algorithm.tex
Google
Most Visited
Getting Started
Latest Headlines
Stat-JR Demonstrator
http://localhost:...ut/algorithm.tex
Done

Conditional posterior for tau for Gibbs sampling

$$\tau \sim \Gamma\left(0.001 + 0.5 \times \text{length}(\text{normexam}), 0.001000 + \frac{\sum_{i=1}^{\text{length}(\text{normexam})} (\text{normexam}_i - \beta_0 \times \text{cons}_i - \beta_1 \times \text{standlrt}_i)^2}{2}\right)$$

Deviance Function

$$\text{deviance} = 2 \times \left(\frac{\tau \times \left(\sum_{i=1}^{\text{length}(\text{normexam})} (\text{normexam}_i - \beta_0 \times \text{cons}_i - \beta_1 \times \text{standlrt}_i)^2\right)}{2} + 0.5 \times (\ln(\pi) - \ln(\tau)) \times \text{length}(\text{normexam}) + 0.346573590279973 \times \text{length}(\text{normexam})\right)$$

Conditional posterior for beta0 for Gibbs sampling

$$\beta_0 \sim \text{N}\left(\frac{\tau \times \left(\sum_{i=1}^{\text{length}(\text{normexam})} \text{cons}_i \times (\text{normexam}_i - \beta_1 \times \text{standlrt}_i)\right)}{\tau \times \left(\sum_{i=1}^{\text{length}(\text{normexam})} \text{cons}_i^2\right)}, \tau \times \left(\sum_{i=1}^{\text{length}(\text{normexam})} \text{cons}_i^2\right)\right)$$

Conditional posterior for beta1 for Gibbs sampling

$$\beta_1 \sim \text{N}\left(\frac{\tau \times \left(\sum_{i=1}^{\text{length}(\text{normexam})} \text{standlrt}_i \times (\text{normexam}_i - \beta_0 \times \text{cons}_i)\right)}{\tau \times \left(\sum_{i=1}^{\text{length}(\text{normexam})} \text{standlrt}_i^2\right)}, \tau \times \left(\sum_{i=1}^{\text{length}(\text{normexam})} \text{standlrt}_i^2\right)\right)$$

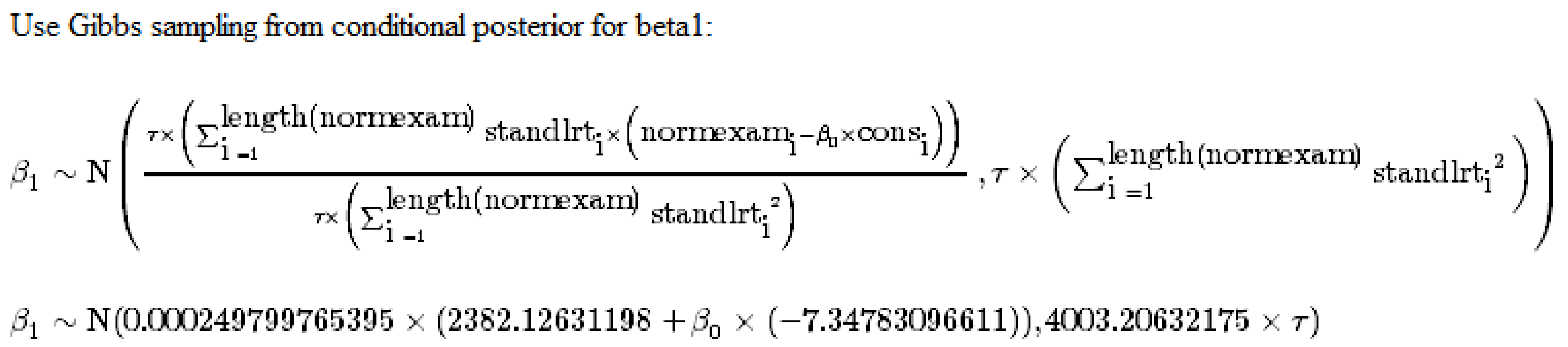
Deterministic formula for parameter sigma

$$\sigma = \frac{1}{\text{sqrt}(\tau)}$$

Deterministic formula for parameter sigma2

$$\sigma_2 = \frac{1}{\tau}$$

Algebra system steps



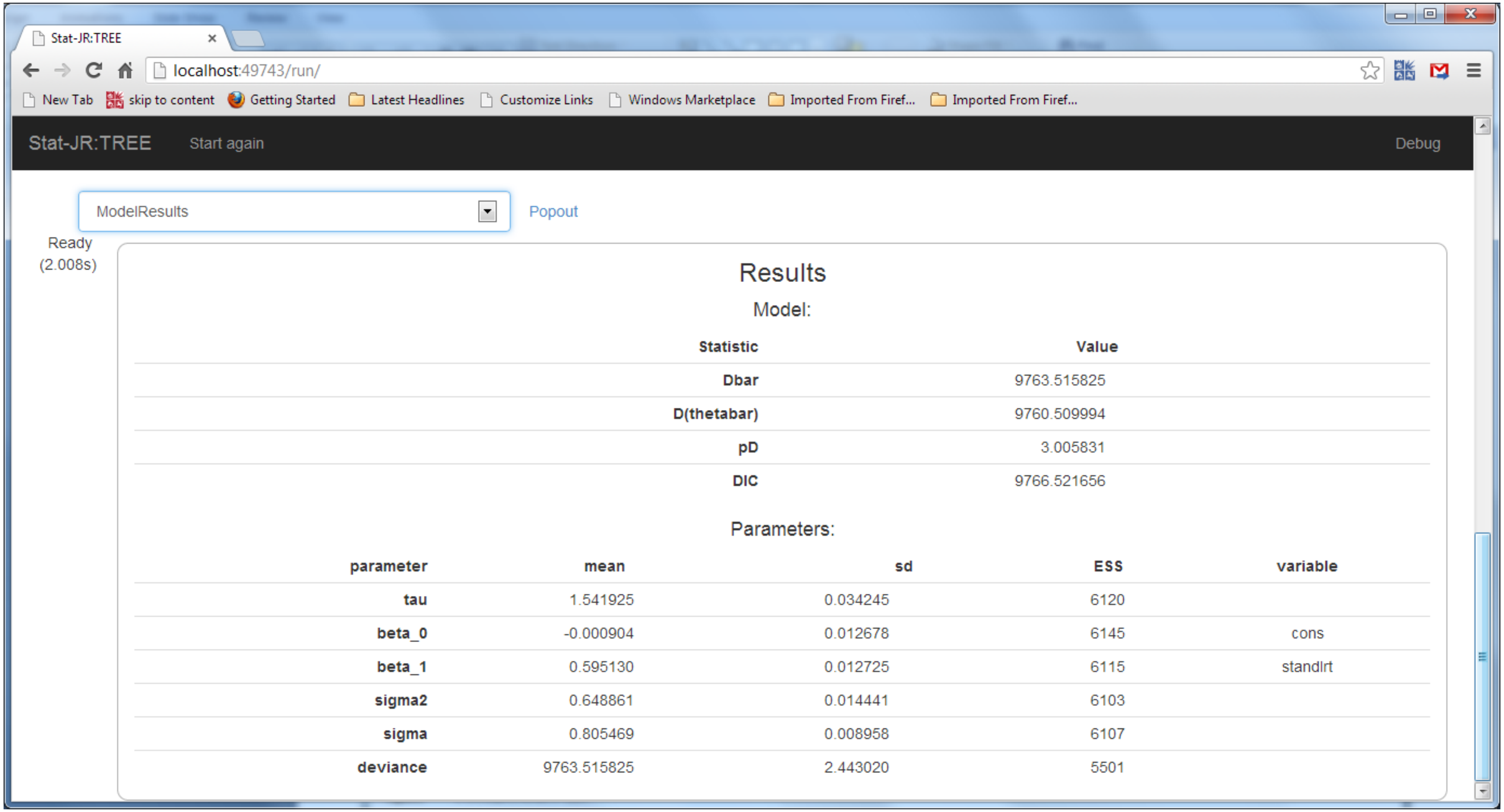
Use Gibbs sampling from conditional posterior for beta1:

$$\beta_1 \sim \mathrm{N}\left(\frac{\tau \times \left(\sum_{i=1}^{\mathrm{length(normexam)}} \mathrm{standlrt}_i \times \left(\mathrm{normexam}_i - \beta_0 \times \mathrm{cons}_i \right) \right)}{\tau \times \left(\sum_{i=1}^{\mathrm{length(normexam)}} \mathrm{standlrt}_i^{\,2} \right)}, \tau \times \left(\sum_{i=1}^{\mathrm{length(normexam)}} \mathrm{standlrt}_i^{\,2} \right) \right)$$

$$\beta_1 \sim \mathrm{N}(0.000249799765395 \times (2382.12631198 + \beta_0 \times (-7.34783096611)), 4003.20632175 \times \tau)$$

Here the first line is what is returned by the algebra system – which works solely on the model code.
The second line is what can be calculated when values are added for constants and data etc.
System then constructs C code and fits model

Output from the E-STAT engine



Ready
(2.008s)

Results

Model:

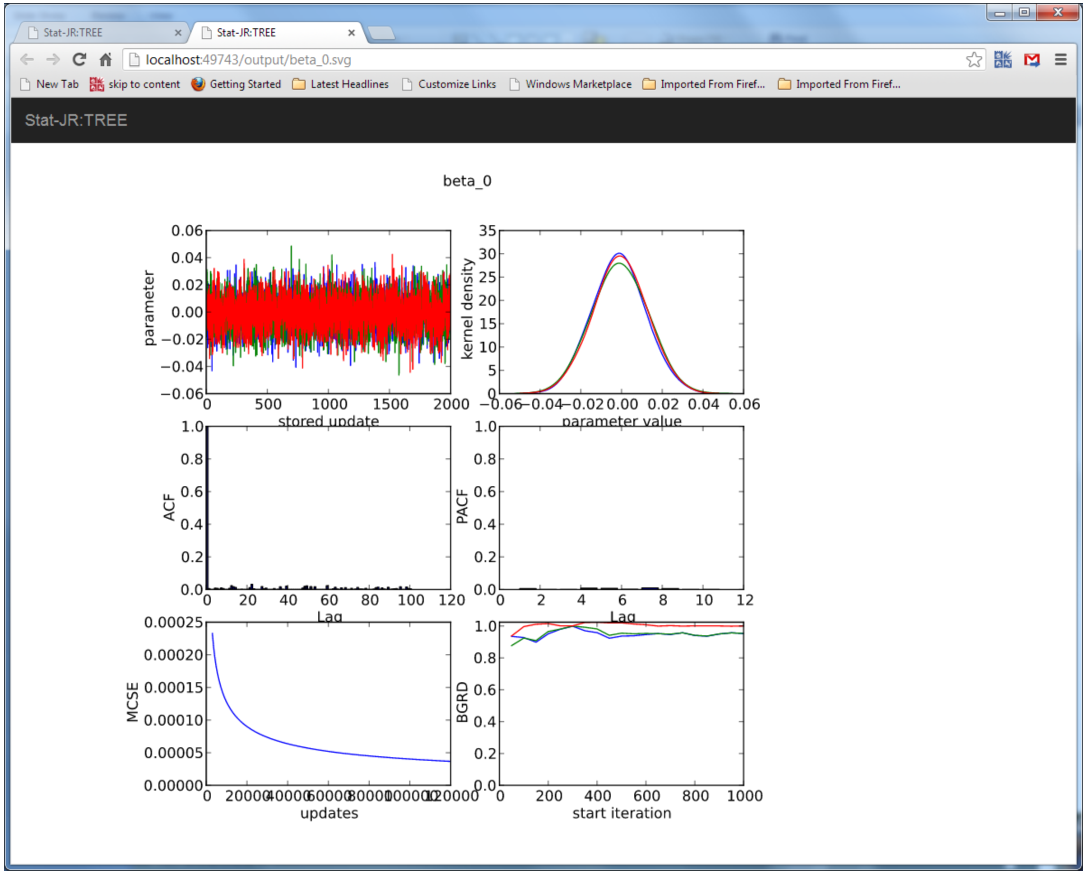
Statistic	Value
Dbar	9763.515825
D(thetabar)	9760.509994
pD	3.005831
DIC	9766.521656

Parameters:

parameter	mean	sd	ESS	variable
tau	1.541925	0.034245	6120	
beta_0	-0.000904	0.012678	6145	cons
beta_1	0.595130	0.012725	6115	standlrt
sigma2	0.648861	0.014441	6103	
sigma	0.805469	0.008958	6107	
deviance	9763.515825	2.443020	5501	

Estimates and the DIC diagnostic can be viewed for the model fitted.

Output from the E-STAT engine



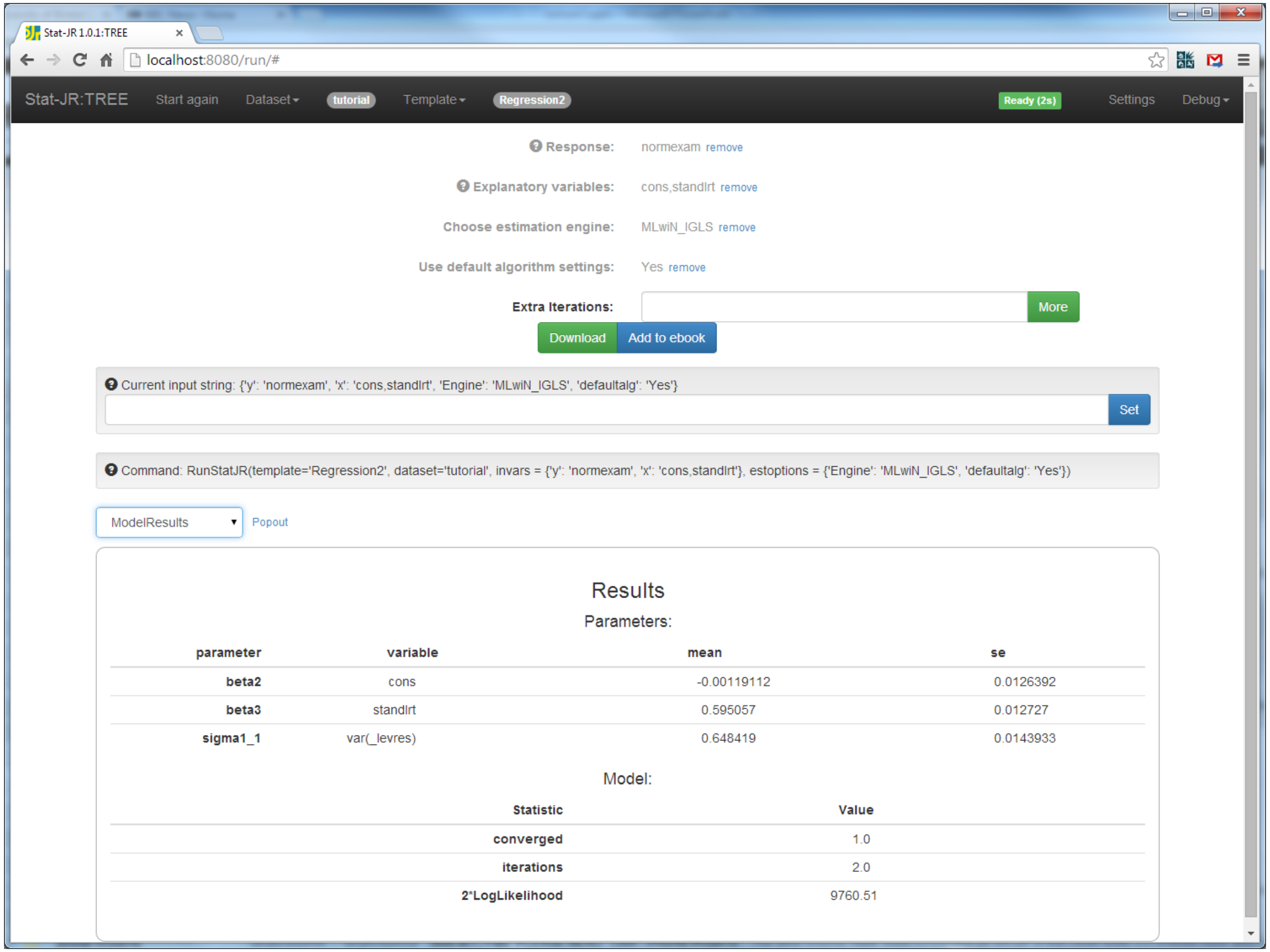
E-STAT offers multiple chains so that we can use multiple chain diagnostics to aid convergence checking.

Otherwise the graphs are borrowed from the MLwiN 6-way plotting.

Graphics are in svg format so scale nicely.

INTEROPERABILITY

Interoperability with MLwiN



Response: normexam remove

Explanatory variables: cons,standlrt remove

Choose estimation engine: MLwiN_IGLS remove

Use default algorithm settings: Yes remove

Extra Iterations:

Current input string: {'y': 'normexam', 'x': 'cons,standlrt', 'Engine': 'MLwiN_IGLS', 'defaultalg': 'Yes'}

Command: RunStatJR(template='Regression2', dataset='tutorial', invars = {'y': 'normexam', 'x': 'cons,standlrt'}, estoptions = {'Engine': 'MLwiN_IGLS', 'defaultalg': 'Yes'})

ModelResults

Results

Parameters:

parameter	variable	mean	se
beta2	cons	-0.00119112	0.0126392
beta3	standlrt	0.595057	0.012727
sigma1_1	var(_levres)	0.648419	0.0143933

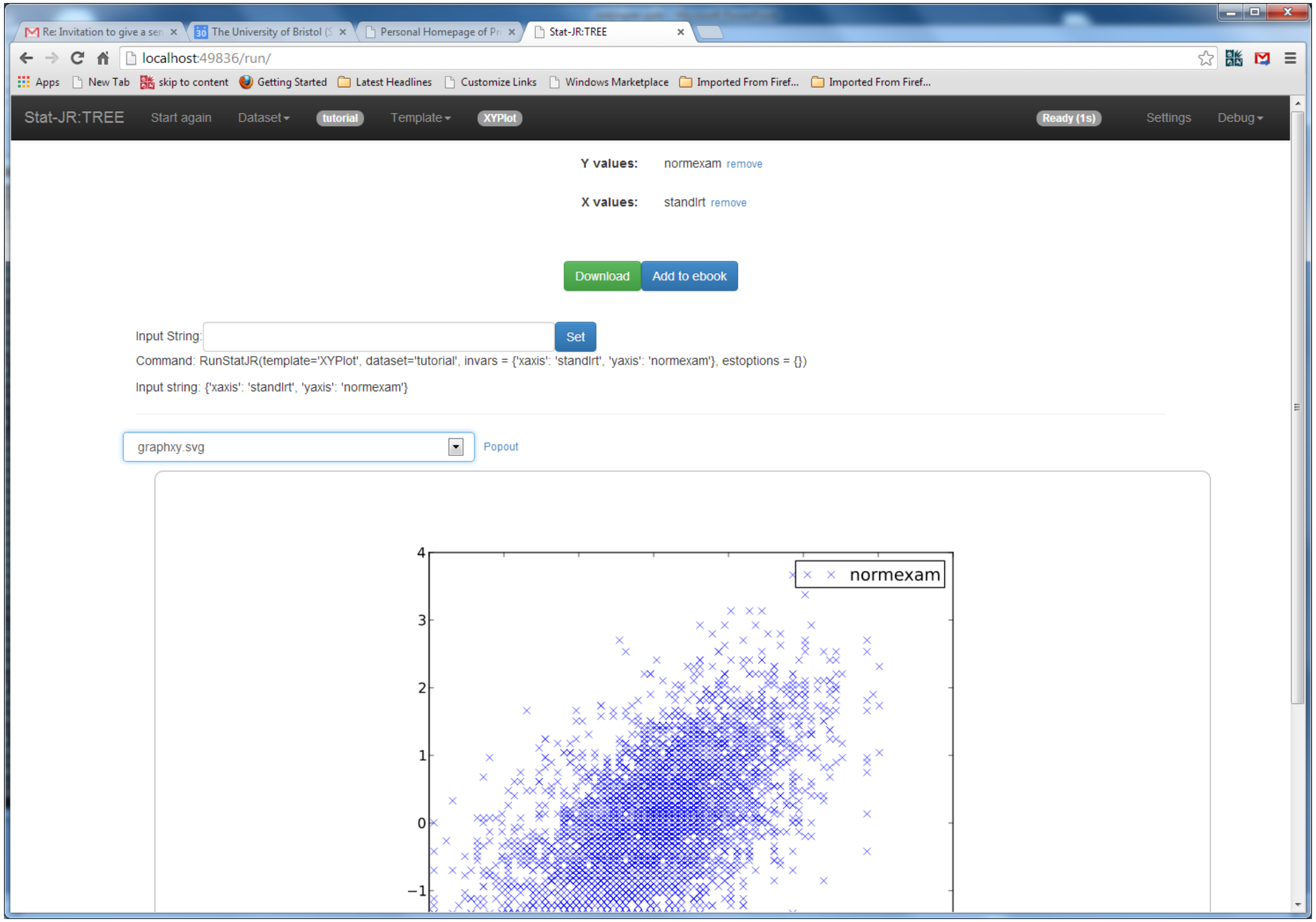
Model:

Statistic	Value
converged	1.0
iterations	2.0
2*LogLikelihood	9760.51

MLwiN can be chosen as an alternative estimation engine. Here macro files to be run in MLwiN are constructed and the output from MLwiN is translated into a ModelResults object.

Currently we are unable to get windows back from MLwiN into Stat-JR

Other templates - XYplot



Y values: normexam remove

X values: standlrt remove

Input String:

Command: RunStatJR(template='XYPlot', dataset='tutorial', invars = {'xaxis': 'standlrt', 'yaxis': 'normexam'}, estoptions = {})

Input string: {'xaxis': 'standlrt', 'yaxis': 'normexam'}

graphxy.svg

There are also templates for plotting.

For example here is a plot using the Xyplot template.

Shown is the plot whilst the Python command script is also available.

Different forms of STAT-JR and E-books

- TREE (Template Reading and Execution Environment) - the format we have demonstrated up to now. Allows user to investigate 1 template and 1 dataset. A dataset can be output from 1 template and then used by the next.
- Cmdtest – this format involves the use of a Python script and allows the template to be called from within a script. Helpful for our test suite and potentially for tasks like simulations. We have code to run Stat-JR from Stata and R using this interface.
- DEEP (Ebooktest) – mixing up templates with textboxes to make executable books – this is covered later.
- Note Stat-JR currently runs on Windows but should soon be platform independent.

(If applicable) results outputted as dataset to be fed back in…

Stat-JR prompts user for input

Template

+

Dataset

Stat-JR writes commands, etc., to perform requested function

Function performed

Results of function produced

(If applicable) external software opened, run, then closed, with results returned to Stat-JR. E.g…

```
myModel<- glm(normexam~
Summary(myModel)
plot(myModel,1)
```

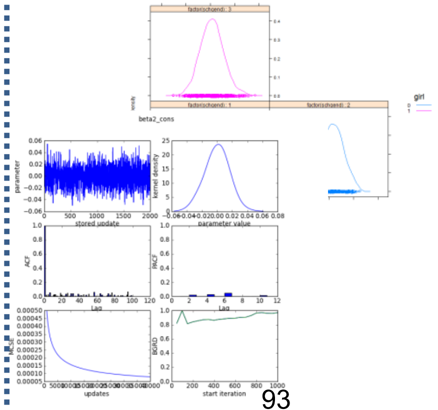
Select **Open Worksheet**
Select **datafile.dta**
Select **Equations** from **Fi**

Scripts *Macros* *Equations* *Point & click instructions*

$$
\begin{aligned}
\mathrm{normexam}_i &\sim \mathrm{N}(\mu_i, \sigma^2)\\
\mu_i &= \beta_0 \mathrm{cons}_i\\
\beta_0 &\propto 1\\
\tau &\sim \Gamma(0.001, 0.001)\\
\sigma^2 &= 1/\tau
\end{aligned}
$$

Results
Model:
DIC: 9766.506
Parameters:
Beta1: 0.594

Results tables *Charts*



3. Further Growth Curve Models

Allowing for nonlinear growth

We can allow for nonlinear growth, e.g. a quadratic polynomial

$$\begin{aligned} y_{ti} &= \beta_{0i} + \beta_{1i} t + \beta_{2i} t^2 + e_{ti} \\ \beta_{0i} &= \beta_0 + u_{0i} \\ \beta_{1i} &= \beta_1 + u_{1i} \\ \beta_{2i} &= \beta_2 + u_{2i} \end{aligned}$$

where the three individual-level random effects (u_{0i}, u_{1i}, u_{2i}) follow a trivariate normal distribution.

t and t^2 treated as explanatory variables in multilevel model for y.

Maximum order of polynomial

The order of polynomial that can be fitted depends on the number of occasions T.

For T observations can fit up to a $T - 1$ order polynomial, e.g. cubic for $T = 4$.

But for $T = 4$ the covariance matrix for y has 10 parameters: 4 variances and 6 covariances. Fitting cubic leads to a 4×4 covariance matrix at individual level, i.e. 10 parameters. So cannot also estimate an occasion-level variance.

Reading: Quadratic growth with fixed t^2 effect

We begin by adding t^2 to the 'fixed' part of the model only (i.e. as an explanatory variable with fixed coefficient).

	IGLS		MCMC	
Parameter	Est	(SE)	Mean	(SD)
Intercept (β_0)	2.534	(0.060)	2.529	(0.060)
t (β_1)	1.641	(0.055)	1.646	(0.058)
t^2 (β_2)	-0.186	(0.016)	-0.187	(0.017)
Level 2 (between-child)				
Intercept variance (σ^2_{u0})	0.564	(0.070)	0.586	(0.075)
Intercept-t covariance (σ_{u01})	0.008	(0.022)	0.007	(0.023)
t variance (σ^2_{u1})	0.083	(0.013)	0.087	(0.013)
Level 1 (within-child)				
Variance (σ^2_e)	0.238	(0.016)	0.239	(0.016)
	Deviance = 2006.4		DIC = 1576.3	

Does the addition of t^2 improve the model?

- The coefficient of t^2 is -0.187 with a standard error of 0.017 which is highly significant
- The change in deviance is $2119 - 2006 = 113$ which far exceeds the critical value of 3.84 for $\chi^2_{1;0.05}$
- The DIC decreases from 1766 to 1576, a difference of 190

So we conclude that the addition of the quadratic in t improves model fit. We now allow the coefficient of t^2 to vary randomly between children.

For all following analyses we present only MCMC results (see Practicals for details of number of chains, burn-in, and number of iterations).

Reading: Quadratic growth with random t^2 effect

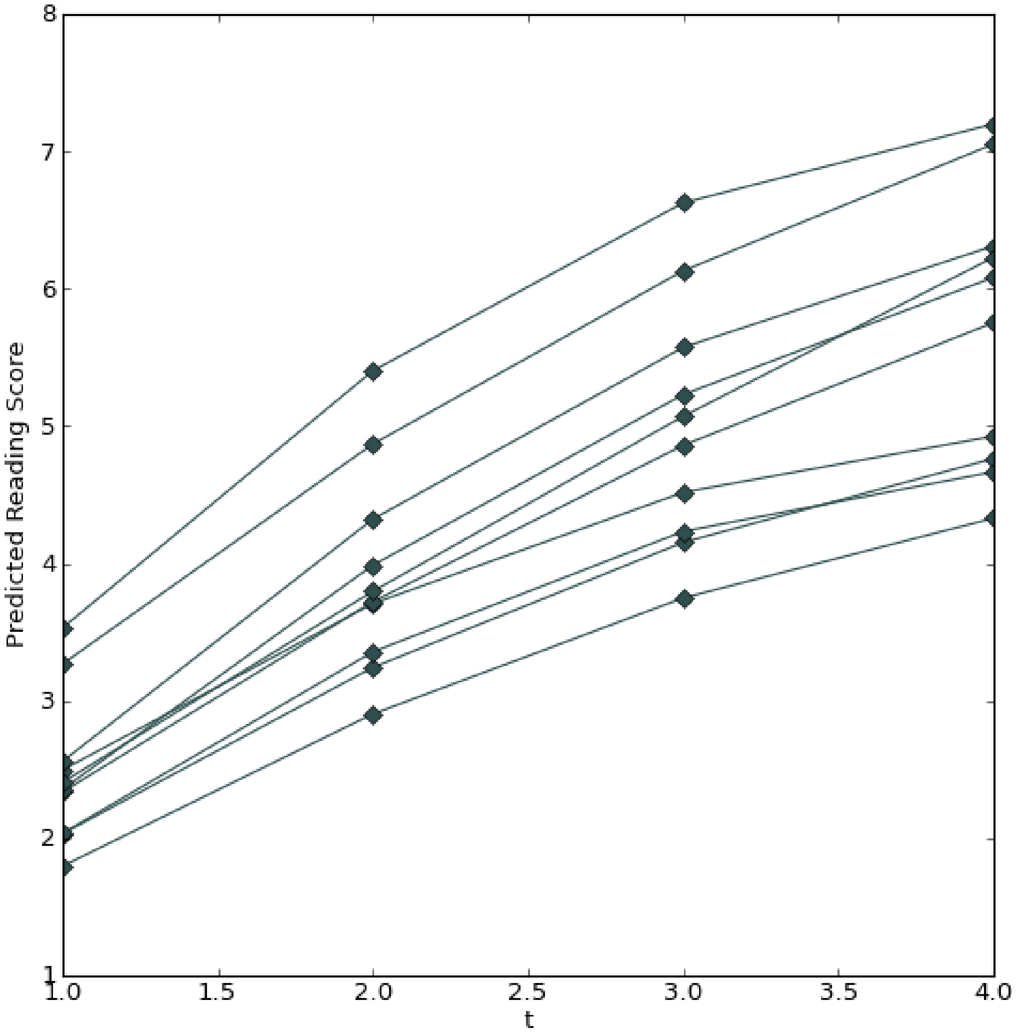
We next allow the coefficient of t^2 to vary randomly between children. The coefficient of t^2 becomes $\beta_{2i} = \beta_2 + u_{2i}$.

3 new terms are added to the level 2 random part of the model:

- Covariance between intercepts and coefficients of t^2, $\text{cov}(u_{0i}, u_{2i}) = \sigma_{u02}$
- Covariance between coefficients of t and coefficients of t^2, $\text{cov}(u_{1i}, u_{2i}) = \sigma_{u12}$
- Variance of coefficients of t^2, $\text{var}(u_{2i}) = \sigma^2_{u2}$

The DIC decreases from 1576 to 1492, so conclude random coefficient for t^2 is necessary to better capture between-individual variation in trajectories.

Predicted quadratic trajectories for first 10 children



Adding covariates

Covariates can be added to a growth model, and these can be individual-level (fixed over time) or time-varying.

Often interested in how trajectories differ between groups, e.g.:

- Does level of reading score depend on amount of cognitive support at home (x_i)?
    - Test by adding x_i as a covariate to growth model
    - Does x_i explain variation in intercept (σ^2_{u0})?

- Does reading progress depend of cognitive support?
    - Test by adding x_i plus its interaction with time, $x_i t$, to model
    - Does $x_i t$ explain variation in progress (σ^2_{u1})?

Reading: Effects of cognitive support on reading level

Results from MCMC estimation (fixed part coefficients only):

Parameter	Mean	(SD)
Intercept (β_0)	2.049	(0.204)
t (β_1)	1.645	(0.060)
t^2 (β_2)	-0.186	(0.017)
homecog (β_3)	0.053	(0.022)

- A 1-unit increase in homecog is associated with a 0.053 point increase in reading score
- But little change in between-individual intercept variances (see Practical)

Reading: Effects of cognitive support on reading progress

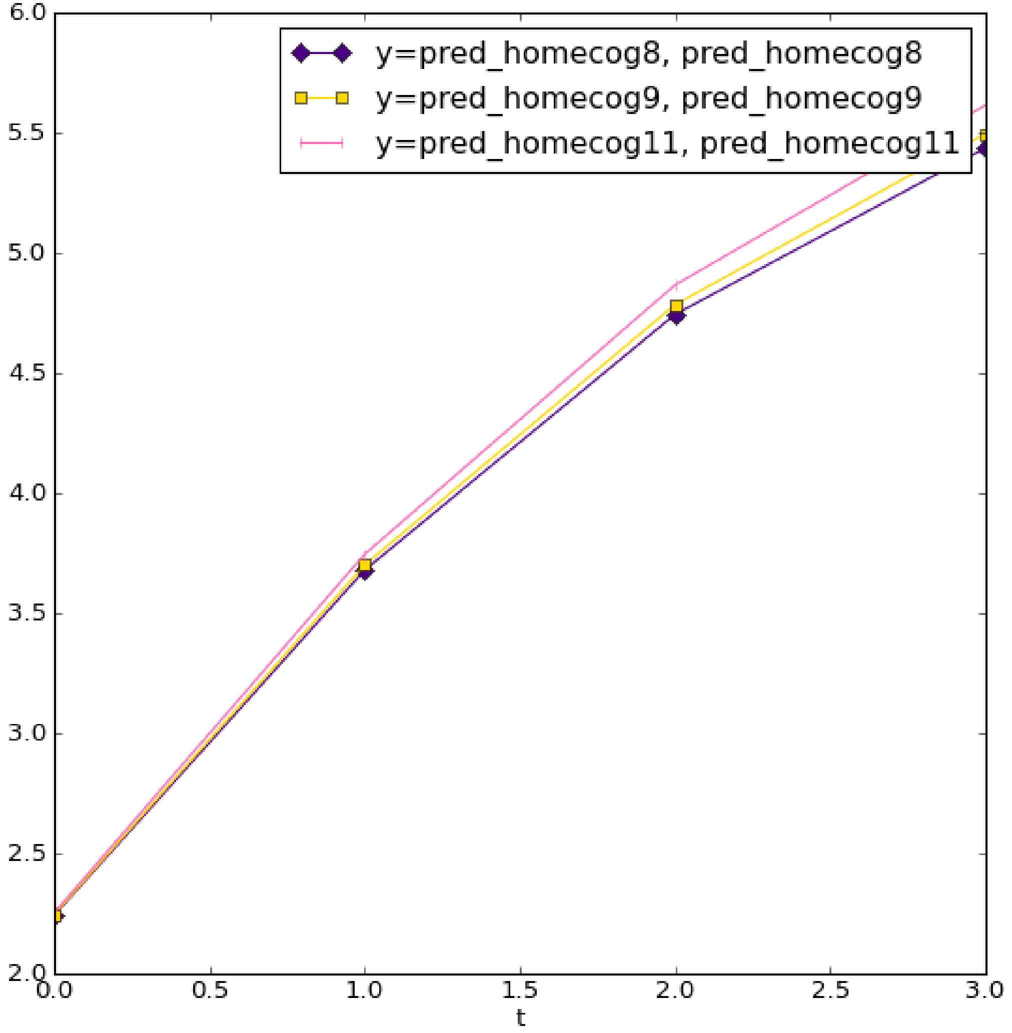
Does the effect of `homecog` depend on the child's age? Add an interaction between t and `homecog`.

Parameter	Mean	(SD)
Intercept (β_0)	2.166	(0.243)
t (β_1)	1.471	(0.115)
t^2 (β_2)	-0.186	(0.017)
`homecog` (β_3)	0.041	(0.025)
$t\times$ `homecog` (β_4)	0.019	(0.010)

- At baseline ($t = 0$), a 1-unit increase in `homecog` is associated with a 0.041 point increase in reading score
- Effect of `homecog` increases with t
- But little change in between-individual variances/covariances (see Practical)

Reading progress by cognitive support at home

Predicted reading scores for quartiles of `homecog`



Allowing for residual autocorrelation

So far occasion-level (time-varying) residuals e_{ti} have been assumed independent. But we might expect residuals to be correlated across occasions, especially if measurements are close together in time.

A common autocorrelation structure is the 1st-order autoregressive, AR(1), structure:

$$\text{corr}(e_{ti}, e_{si}) = \alpha^{|t-s|}$$

which, for $|\alpha| < 1$, implies the correlation decays with time difference $|t - s|$.

AR(1) structure for reading data

For 4 measurements, the correlation matrix for $(e_{1i}, e_{2i}, e_{3i}, e_{4i})$ is:

$$\begin{pmatrix} 1 & & & \\ \alpha & 1 & & \\ \alpha^2 & \alpha & 1 & \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}$$

Allowing for AR(1) autocorrelation therefore involves one additional parameter α.

We can test for autocorrelation using a (frequentist) t-test of $H_0 : \alpha = 0$.

If Bayesian estimation is used, can examine credible intervals.

Reading model with AR(1) residuals

- Using Stata `xtmixed` the autocorrelation parameter α is estimated as -0.291 with a standard error of 0.213 $\implies$ we cannot reject $\mathrm{H}_0 : \alpha = 0$

- Using MCMC in Stat-JR $\hat{\alpha} = -0.327$ (SE $= 0.222$). However chains show poor mixing, which suggests model may not be well identified.

Random intercept model with AR(1) residuals

A flexible model (with random coefficients on time variables) may sufficiently capture individual variation in trajectories, removing need to allow for autocorrelation at occasion level.

Suppose we fit a quadratic for t but allow only the intercepts to vary across children.

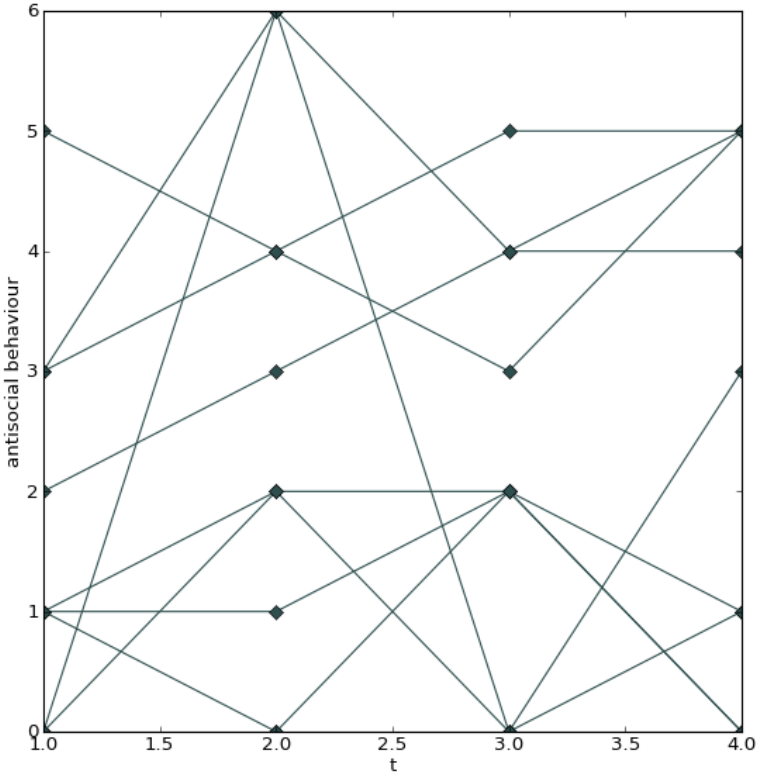
- Using `xtmixed` α is now estimated as 0.67 with a SE of 0.08, so strongly significant
- Using MCMC in Stat-JR, α estimated as 0.693 (SD $=$ 0.065)
- But we know that this simple model insufficiently captures variation between children's trajectories

Handling highly variable trajectories

Growth curve models are most useful for developmental processes (e.g. cognitive measures, height, weight).

Although they can be applied in any situation where the concept of a 'trajectory' is useful, the model may need to be extremely complex to adequately represent change that is highly nonlinear with large between-individual variation in the shape of trajectories.

Observed trajectories in antisocial behaviour for 10 children



A low-order polynomial growth curve will not capture this variation, even with random coefficients.

Treating time as categorical: multivariate model

Include dummies for t in fixed part of model (t_1, t_2, t_3, t_4) and estimate separate residual variance for each t.

$$\text{read}_{ti} = \beta_0 t_1 + \beta_1 t_2 + \beta_2 t_3 + \beta_3 t_4 + e_{ti}$$

where

$$\begin{pmatrix} e_{0i} \\ e_{1i} \\ e_{2i} \\ e_{3i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{e0} & & & \\ \sigma_{e01} & \sigma^2_{e1} & & \\ \sigma_{e02} & \sigma_{e12} & \sigma^2_{e2} & \\ \sigma_{e03} & \sigma_{e13} & \sigma_{e23} & \sigma^2_{e3} \end{pmatrix} \right]$$

Means and covariances for antisocial behaviour

When modelling change, we aim to capture the means and variances of y_{ti} at each t, and their covariances across t, as simply as possible.

Variable	Mean	n
anti1	1.50	221
anti2	1.84	221
anti3	1.88	221
anti4	2.07	221

Sample covariance matrix (correlations in brackets)

	anti1	anti2	anti3	anti4
anti1	2.37			
anti2	1.16 (0.42)	3.21		
anti3	1.22 (0.44)	1.63 (0.51)	3.24	
anti4	1.35 (0.42)	2.00 (0.54)	2.24 (0.60)	4.35

MCMC results from fitting multivariate model to antisocial behaviour

Fixed part estimates

Parameter	Mean	(SD)
t1 (β_0)	1.50	(0.11)
t2 (β_1)	1.84	(0.12)
t3 (β_2)	1.88	(0.12)
t4 (β_3)	2.07	(0.14)

Random part estimates

	Uniform		Wishart	
Param	Mean	(SD)	Mean	(SD)
σ^2_{e0}	2.48	(0.25)	2.40	(0.23)
σ_{e01}	1.21	(0.21)	1.16	(0.20)
σ^2_{e1}	3.36	(0.33)	3.25	(0.31)
σ_{e02}	1.28	(0.22)	1.23	(0.21)
σ_{e12}	1.70	(0.26)	1.64	(0.25)
σ^2_{e2}	3.39	(0.34)	3.28	(0.32)
σ_{e03}	1.41	(0.25)	1.35	(0.24)
σ_{e13}	2.10	(0.31)	2.01	(0.29)
σ_{e23}	2.35	(0.32)	2.25	(0.30)
σ^2_{e3}	4.55	(0.45)	4.38	(0.42)

The model perfectly reproduces the observed means. Var/cov estimates sensitive to choice of prior; Wishart estimates closest to observed values.

Notes on multivariate model

- Very flexible and allows for autocorrelation
- But flexibility means that we cannot discern any general patterns in individual trajectories
- Equivalent to $T - 1$ polynomial (with no occasion-level residual) - same number of parameters and likelihood value, but multivariate model easier to interpret
- May be useful if trajectories are of limited interest and few measurement occasions, e.g. if effects of time-varying x of major interest

4. Dynamic (Autoregressive) Models

Research questions about 'dynamic' relationships

The growth curve approach is appropriate when interest centres on trajectories in y over time, and how they differ across groups.

For some types of process, we might expect a direct dependency of y_t on previous y, e.g. earnings, depression.

We might also be interested in more dynamic questions, e.g.:

- How does health at t depend on a change in employment status between $t-1$ and t (adjusting for health prior to t)?
- How does starting a new treatment at $t-1$ affect depression at t (adjusting for earlier depression)?

1st-order AR(1) dynamic model

The most commonly applied dynamic model assumes that y_t depends on past y through y_{t-1}:

$$y_{ti} = \delta y_{t-1,i} + \beta x_{ti} + u_i + e_{ti}, \qquad t = 2, 3, \dots, T$$

where $u_i \sim N(0, \sigma_u^2)$ and $e_{ti} \sim N(0, \sigma_e^2)$

- In practice we also include an intercept term in the model. This is set to zero here for simplicity.
- δ commonly assumed to be equal across individuals
- u_i can be treated as fixed rather than random (see brief discussion later)

State dependence vs unobserved heterogeneity

Residual correlation between y_t and y_{t-1} is $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$.

Is correlation between y_t and y_{t-1} due to:

- Causal effect of y_{t-1} on y_t?

  $\Rightarrow |\delta|$ close to 1 and ρ close to 0 (state dependence)

- Mutual dependence on time-invariant omitted variables?

  $\Rightarrow |\delta|$ close to 0 and ρ close to 1 (unobserved heterogeneity)

Example of state dependence vs unobserved heterogeneity

E.g. Explanations for pattern of high income over time:

- Current income determined by past income

  $\Rightarrow |\delta|$ close to 1 and ρ close to 0 (state dependence)

- Dependence of income at all t on unmeasured characteristics (qualifications, skills, ambition etc)

  $\Rightarrow |\delta|$ close to 0 and ρ close to 1 (unobserved heterogeneity)

Endogeneity of lagged outcomes

$$y_{ti} = \delta y_{t-1,i} + \beta x_{ti} + u_i + e_{ti}, \qquad t = 2, 3, \ldots, T$$

A standard assumption of regression models is that residuals are uncorrelated with explanatory variables (y_{t-1} and x_t here):

$$\begin{aligned} &\text{cov}(u_i, y_{t-1,i}) = 0, \quad &&\text{cov}(e_{ti}, y_{t-1,i}) = 0 \\ &\text{cov}(u_i, x_{ti}) = 0, &&\text{cov}(e_{ti}, x_{ti}) = 0 \end{aligned}$$

Assumption that $\text{cov}(u_i, y_{t-1,i}) = 0$ is especially problematic as:

$$y_{t-1,i} = \delta y_{t-2,i} + \beta x_{t-1,i} + u_i + e_{t-1,i}$$

y_{t-1} is said to be endogenous with respect to y_t.

Dependence of y_t on earlier y for AR(1) model

Can show y_t $(t = 2, \ldots, T)$ depends on earlier t entirely through y_1. Omitting covariates for simplicity:

$$\begin{aligned} y_{2i} &= \delta y_{1i} + u_i + e_{2i} \\ \\ y_{3i} &= \delta y_{2i} + u_i + e_{3i} \\ &= \delta^2 y_{1i} + (1 + \delta) u_i + \delta e_{2i} + e_{3i} \\ \\ y_{4i} &= \delta y_{3i} + u_i + e_{4i} \\ &= \delta^3 y_{1i} + (1 + \delta + \delta^2) u_i + \delta^2 e_{2i} + \delta e_{3i} + e_{4i} \end{aligned}$$

By repeated substitution, we find the effect of y_1 on subsequent y_t is δ^{t-1} which diminishes with t for $|\delta| < 1$

The 'initial conditions' problem

As y_t depends on previous y through y_1, our assumptions about y_1 are important.

y_1 may not be measured at the start of the process

Can view as a missing data problem:

$$\begin{array}{lr} \text{Observed} & (y_1, \ldots, y_T) \\ \text{Actual} & (y_{-k}, \ldots, y_0, y_1, \ldots, y_T) \end{array}$$

where first $k + 1$ measures are missing.

We can allow for endogeneity of y_{t-1} by specifying a model for y_1.

Modelling y_1

Our model for y_t $(t = 2, 3, \ldots, T)$ is:

$$y_{ti} = \delta y_{t-1,i} + \beta x_{ti} + u_i + e_{ti}$$

A general model for y_1 is:

$$y_{1i} = \gamma_0 + \gamma_1 x_{1i} + v_i + e_{1i}$$

Estimating the joint model for $y_1, y_2, \ldots, y_T$

We can combine the models for $t = 1$ and $t \geq 2$ in one model.

Define t_1 and t_{2T} as dummies for $t = 1$ and $t \geq 2$.

Form interaction variables $t_1 x_{1i}$, $t_{2T} y_{t-1,i}$ and $t_{2T} x_{ti}$.

$$\begin{aligned} y_{ti} &= t_1\{\gamma_0 + \gamma_1 x_{1i} + v_i + e_{1i}\} \\ &+ t_{2T}\{\delta y_{t-1,i} + \beta x_{ti} + u_i + e_{ti}\} \end{aligned}$$

This can be viewed as a multilevel model with level 2 random coefficients for t_1 and t_{2T} and complex level 1 variance.

To recover the 2 equations, set (i) $t_1 = 1, t_{2T} = 0$ and (ii) $t_1 = 0, t_{2T} = 1$.

Identification of model with initial conditions

Some restrictions must be placed on the residuals v_i and e_{1i} in the model for y_1.

This is because we cannot separate between-individual and within-individual variances for $t = 1$.

Possible approaches include:

1. Allow separate variances for e_{1i} and e_{ti} $(t > 1)$ but set $v_i = u_i$
2. Estimate variance of v_i and its covariance with u_i but set $\text{var}(e_{1i}) = \text{var}(e_{ti})$ for $t > 1$

What if we ignore the initial conditions?

Suppose we do not model y_1, and fit only the model for $y_2, \dots y_T$:

$$y_{ti} = \delta y_{t-1,i} + \beta x_{ti} + u_i + e_{ti}$$

- The estimate of δ will be biased upwards because we have failed to allow for the shared dependency of y_t and y_{t-1} on u_i.
    - The association between y_t and y_{t-1} is partly due to unmeasured time-invariant factors affecting y across time
- Estimates of covariate effects will also be biased if x correlated with y at t and $t - 1$

Application of AR(1) Model to Antisocial Behaviour

Results ignoring initial condition

$$\text{anti}_{ti} = \beta_0 + \delta \text{anti}_{t-1,i} + \beta_1 \text{male}_i + u_i + e_{ti}, \qquad t = 2, 3, 4$$

Parameter	Mean	(SD)
anti_{t-1}	0.48	(0.09)
male	0.43	(0.15)
cons	0.87	(0.16)
Child-level variance (σ_u^2)	0.22	(0.27)
Occasion-level variance (σ_e^2)	2.44	(0.24)

$\hat{\delta} = 0.48$ (SD $= 0.09$) and $\hat{\sigma}_u^2 = 0.22$ (SD $= 0.27$) $\Longrightarrow$ state dependence rather than unobserved heterogeneity.

Models with initial condition

$$
\begin{aligned}
\text{anti}_{1i} &= \gamma_0 + \gamma_1 \text{male}_i + v_i + e_{1i} \\
\text{anti}_{ti} &= \beta_0 + \delta \text{anti}_{t-1,i} + \beta_1 \text{male}_i + u_i + e_{ti}, \qquad t = 2, 3, 4
\end{aligned}
$$

Estimate two versions of model for y_1:

1. Allow separate variances for e_{1i} and e_{ti} $(t > 1)$ but estimate a common child-level random effect u_i
2. Estimate variance of v_i and its covariance with u_i but assume variance of occasion-level residual is same for all t

Results for alternative AR(1) models

	No IC		IC (1)		IC (2)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
anti$_{t-1}$	0.48	(0.09)	0.20	(0.05)	0.06	(0.06)
male	0.43	(0.17)	0.62	(0.17)	0.71	(0.20)
cons	0.87	(0.16)	1.26	(0.14)	1.45	(0.17)
Child level						
var at $t = 1$	-	-	0.96	(0.18)	0.83	(0.22)
var at $t > 1$	0.22	(0.27)	0.96	(0.18)	1.70	(0.32)
covariance	-	-	-	-	1.05	(0.20)
Occasion level						
var at $t = 1$	-	-	1.55	(0.19)	1.65	(0.11)
var at $t > 1$	2.44	(0.24)	1.97	(0.14)	1.65	(0.11)

Comments on AR(1) results

- Effect of y_{t-1} overestimated and between-child variance (σ_u^2) underestimated if initial condition ignored
  - This impacts on the estimate for `male`

- Effect of y_{t-1} depends on how y_1 is modelled (0.06 vs 0.20)
  - But $\text{corr}(u_i, v_i) = 1.05/\sqrt{0.83 \times 1.70} = 0.88$ which suggests model IC(2) is over-parameterised (2 extra parameters versus 1 extra for IC(1))

- In general, estimates can be sensitive to assumptions about initial condition for short panels

- In a real application, model for y_1 should include covariates x_{1i} that capture history of y up to $t = 1$ for better identification

Application of AR(1) Model to Mental Health and Employment

Another example: Mental health and employment

Question: What is the effect of change in employment status on subsequent mental health (adjusting for previous mental health)?

- Data from British Household Panel Study (25% sub-sample)
- 2808 men of working age observed annually for 2-18 years
- Mental health measured by General Health Questionnaire (scores 0-36, mean = 10)
- Employment status at each wave: employed (E) or non-employed (N)
  - Focus on *change* in employment status between $t-1$ and t: EE, NN, EN and NE

Model specification

$$\begin{aligned}
\text{GHQ}_{1i} &= \gamma_0 + \gamma_1 \text{age}_{1i} + \gamma_2 \text{age}^2_{1i} + \gamma_3 \text{employ}_{1i} + v_i + e_{1i} \\
\text{GHQ}_{ti} &= \beta_0 + \delta \text{GHQ}_{t-1,i} \\
&+ \beta_1 \text{empNN}_{t-1,i} + \beta_2 \text{empEN}_{t-1,i} + \beta_3 \text{empNE}_{t-1,i} \\
&+ \beta_4 \text{age}_{ti} + \beta_5 \text{age}^2_{ti} + u_i + e_{ti}, \qquad t = 2, \ldots 18
\end{aligned}$$

- $t = 1$ corresponds to 1st wave of observation for an individual
- employ is a dummy for employment status at $t = 1$ (1=employed, 0=not employed)
- empNN, empEN and empNE are dummies for change in employment status between $t - 1$ and t: non-employed at both (NN), become non-employed (EN) and become employed (NE). Reference category is employed at both (EE)

Effects of GHQ and employment transitions for alternative AR(1) models

	No IC		IC (1)		IC (2)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
GHQ_{t-1}	0.26	(0.01)	0.23	(0.01)	0.23	(0.02)
empNN	0.85	(0.10)	0.83	(0.10)	0.84	(0.10)
empEN	1.41	(0.13)	1.42	(0.13)	1.41	(0.13)
empNE	−0.83	(0.13)	−0.75	(0.15)	−0.85	(0.13)

- Assumption about y_1 has little impact in this long panel. Extended model for y_1 also included experience of unemployment up to $t = 1$ but conclusions the same

- Compared to men who remain employed, being non-employed at both waves or becoming non-employed is associated with higher depression scores, while becoming employed is associated with lower depression