Predictive Inference Tools for Researchers

by

Voyze G. Harris III

A Thesis Submitted to the Faculty of the

Statistics and Data Science Graduate Interdisciplinary Program

In Partial Fulfillment of the Requirements
For the Degree of

Master of Science

In the Graduate College

The University of Arizona

2021

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Master's Committee, we certify that we have read the thesis prepared by Voyze Gabriel Harris III, titled *[Enter Thesis Title]* and recommend that it be accepted as fulfilling the dissertation requirement for the Master's Degree.

_____ Date: _____
Dr. Dean Billheimer

_____ Date: _____
Dr. Edward Bedrick

_____ Date: _____
Dr. Walter Piegorsch

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to the Graduate College.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the Master's requirement.

_____ Date: _____
Dr. Dean Billheimer
Master's Thesis Committee Chair
Biostatistics

Contents

1 Thesis Abstract

An obstacle to widespread employment of Bayesian predictive inference in scientific research is the lack of suitable computing tools. In this thesis I document several established useful models, and provide an applicable set of tools for statisticians. For each of the included models, some basic notes on mathematical derivation are presented, and predictive inference is illustrated with examples. For the details of the models and some of the examples I relied primarily on Seymour Geisser's Predictive Inference: An Introduction (1993) and Peter D. Hoff's A First Course in Bayesian Statistical Methods (2009).

An R package has been developed, the main purpose of which is to provide the researcher with a means of producing random samples from predictive distributions. For all the models, the package includes random sample generators. For those models with analytical solutions, density and distribution functions are also provided. The standard R naming convention for these function classes has been adopted: density functions are prefixed with the letter "d," distribution functions with the letter "p," and random generation functions with the letter "r." Also included in all function names is the abbreviation "pred" (for predictive) and an initialism or abbreviation identifying the model itself. For example, the density function for the Beta-Binomial model is named "`dpredBB()`." The R code for each function is included in Appendix X-insert link to appendix here.

2 Introduction: Predictive Inference

My understanding of Bayesian Predictive Inference began with the University of Arizona course "Bayesian Statistical Theory and Applications" under [is it appropriate to put "Dr." here and elsewhere?] Edward Bedrick. Since then it has been shaped by exposure to various sources, including articles by Bedrick and Dean Billheimer, and to a greater extent the aforementioned texts by Hoff and Geisser, and others, including Bayesian Data Analysis by Andrew Gelman et. al., and Statistical Prediction Analysis by J. Aitchison and I.R. Dunsmore. In the next section and what follows, the ideas expressed are an amalgam of what was learned from this body of scholarship.

2.1 Why Predictive Inference?

The main purpose of statistics is to predict future events based on observed data. Prediction about meaningful quantities that are relevant to the object of study facilitates scientific progress in multiple ways. Advantages include enhancing scientific reproducibility, enabling corroboration or refutation of current hypotheses through future experimentation, informing decision-making by summarizing quantities of direct interest to the researcher, and shifting the focus of statistical analysis from estimation of hypothetical parameters to statements about concrete observables.

It is not the intent of this thesis to suggest that parametric inference should be abandoned in statistical analyses. Conventional statistical inference techniques are useful for summarizing information about large quantities of data in a handful of usable values, and leveraging such summaries to determine whether a particular problem merits continued

attention. Indeed, the scientific discipline of statistics developed along frequentist lines, and the evolution of Bayesian methods has occurred atop that foundation.

Prediction is a means of discriminating between scientific hypotheses. Generally, a model may be judged by the quality of its predictions. Given competing models, the better predictor will be given more weight, and a useful model increases in utility as it its predictive capability improves.

To illustrate the potential difference between results from Bayesian prediction and using plug-in estimators, consider the game Pass the Pigs®, a push-your-luck dice game in which the "dice" are actually rubber pig figures. Two pig dice are thrown, and points are scored according to the combination of positions in which they come to rest. Details about the game can be found on Wikipedia here: https://en.wikipedia.org/wiki/Pass_the_Pigs

Figure 1: Pass The Pigs®

For the purpose of this example, consider the probability of a single pig landing in the "Razorback" position, which occurs when the pig is lying on its back with its legs extended upward. The irregular shape of the pig makes it difficult to assign probabilities to results other than by means of experimentation. Such an experiment was conducted at Duquesne University, and an article describing the experiment as well as Bayesian predictive inference performed on the results appeared in the *Journal of Statistics Education* Volume 14, Number 3, in 2006. The article can be accessed here: http://jse.amstat.org/v14n3/datasets.kern.html. Of the 11,954 recorded results for individual pigs, approximately 22.4% were Razorbacks.

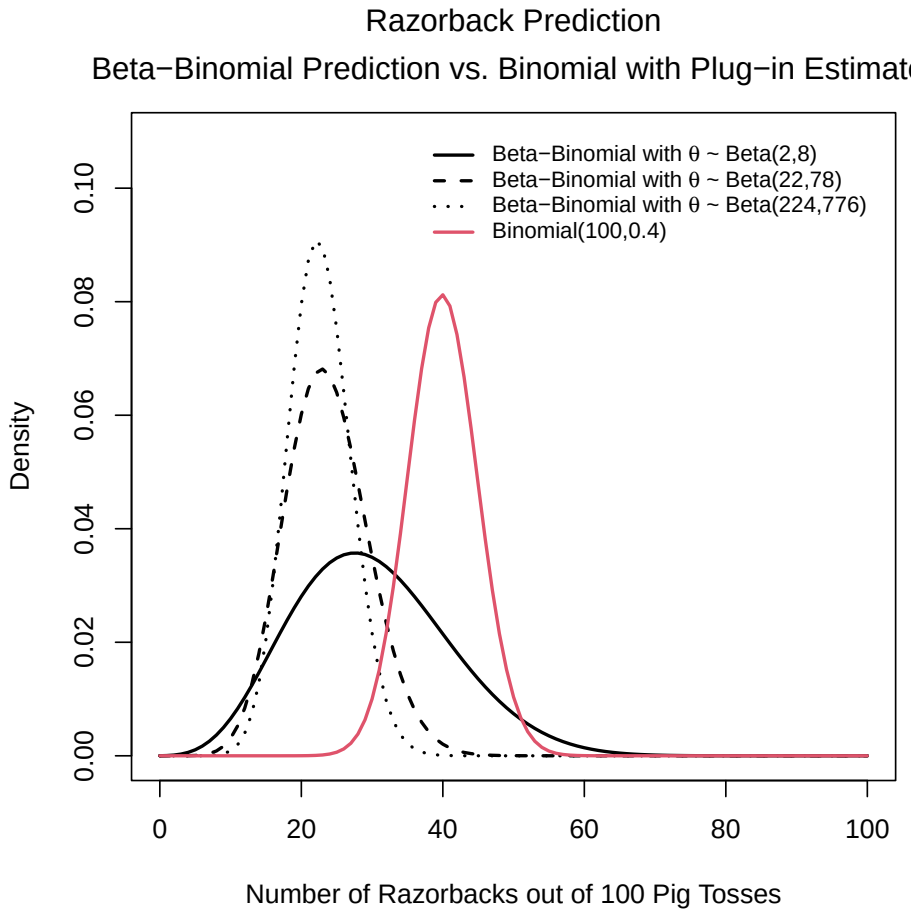
Suppose $t = 4$ Razorbacks have been observed out of $N = 10$ tosses of a single pig die, suggesting a straightforward binomial distribution with $\theta = \Pr(\text{Razorback}) = t/N = 0.4$. Taking the Duquesne experiment into consideration, we'll perform Bayesian prediction using three prior distributions for θ: $\theta \sim \text{Beta}(2, 8)$, $\theta \sim \text{Beta}(22, 78)$, and $\theta \sim \text{Beta}(224, 776)$, and compare these results to predictions obtained from the plug-in estimator $\theta = 0.4$. Any number of prior distributions on θ would satisfy the condition that $E(\theta) \approx 0.224$, suggested by the prior information. The specific choice of a Beta prior is made largely for computational convenience.

In this example the question asked by the researcher is, "For $M = 100$ future observations, how many Razorbacks are predicted?" The density curves in the plot below show the influence of the details of the choice of prior on the location and variance of the predictive distribution. Essentially, each pair of shape parameters (α, β) in the Beta prior reflects the researcher's level of reliance on the results of the Duquesne experiment, with the "weight"

given to that knowledge increasing with the shape parameters by orders of magnitude. The choice of shape parameters might be influenced by such things as pig tossing method (perhaps the researcher is throwing them by hand rather than dropping them by the carefully controlled method used in the Duquesne experiment), or by a need to account for pig-to-pig variation, or anything else the researcher believes introduces a deviation from the events upon which the prior information is based. Use of the scaled-up Big Pigs™variant, for example might give reason to use a very low-weighted prior such as $\theta \sim \text{Beta}(2, 8)$.

Figure 2: Big Pigs™to Original Approximate Scale

The plot and table below illustrate the effects of the Bayesian prediction method. Perhaps most notable is the location disparity between the plug-in (Binomial) prediction and the family of Bayesian (Beta-binomial) predictive distributions. The consideration of prior knowledge has a significant effect. In this example, the strong influence of the choice of shape factors for the Beta prior on the mean and variance of the predictive distribution provides options for the future prognosticator. If closely duplicating the Duquesne experimental conditions, for example, predictions might be based on the result from the Beta(224, 776) prior. A lower-weighted prior might be used for any experimental element the researcher believes introduces a deviation from the knowledge upon which the prior information is based.



#Razorbacks Predicted out of 100 Tosses				
Prediction Method	(α, β)	$\mathbf{E}(\theta)$	**mean**	**SD**
Beta-Binomial	(2,8)	0.2	30.19	11.01
Beta-Binomial	(22,78)	0.22	23.42	5.82
Beta-Binomial	(224,776)	0.224	22.68	4.53
Binomial$(100, 0.4)$	–	–	40	4.9

2.2 The Bayesian Parametric Prediction Format

We want to predict future outcomes based on current knowledge. Specifically we're asking: for observed values $Y_1 = y_1, ..., Y_n = y_n$, what is likely to be the value of the next observation, $\tilde{Y} = \tilde{y}$? We want to compute $Pr(\tilde{Y}|y_1, ..., y_n)$, where $y_1, ..., y_n$ are conditionally independent and identically distributed (i.i.d.) with respect to a population parameter (or parameters) θ. We assign prior distribution $\pi(\theta)$ based on some existing knowledge or beliefs. Here we are careful to satisfy ourselves that $Y_1, ..., Y_n$ are *exchangeable*, which enables us to rely on de Finetti's representation theorem for the i.i.d. assumption. Thus we can write

$$
\begin{aligned}
p(\tilde{Y} = \tilde{y}|Y_1 = y_1, ..., Y_n = y_n) &= \frac{p(\tilde{y}, y_1, ..., y_n)}{p(y_1, ..., y_n)} \\
&= \frac{\int p(\tilde{y}, y_1, ..., y_n|\theta)\pi(\theta)d\theta}{p(y_1, ..., y_n)} \\
&= \frac{\int p(\tilde{y}|\theta)p(y_1, ..., y_n|\theta)\pi(\theta)d\theta}{p(y_1, ..., y_n)} \\
&= \frac{\int p(\tilde{y}|\theta)p(\theta|y_1, ..., y_n)p(y_1, ..., y_n)d\theta}{p(y_1, ..., y_n)} \\
&= \int p(\tilde{y}|\theta)p(\theta|y_1, ..., y_n)d\theta \qquad (1)
\end{aligned}
$$

For prediction, we need only to characterize the observed data $(y_1, ..., y_n)$ conditional θ and supply a suitable prior distribution $\pi(\theta)$. From there we compute posterior

$$p(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{\int_\theta p(\theta)p(y|\theta)d\theta}$$

and make our prediction using (1).

3 Predictive Problems with Conjugate Priors

When data can be modeled with a distribution that suggests a conjugate prior for the parameter(s) of interest, and those parameters can reasonably be represented by that conjugate prior, the posterior calculations are greatly simplified. The four classes of models addressed in this section all exhibit this feature. The first three are single-parameter exponential families and have closed-form solutions for prediction. The fourth is a two-parameter exponential family with unkown mean and variance. While this one does not admit an analytical solution, prediction is easily accomplished by means of simple monte carlo (MC) sampling. The four classes of predictive models presented in this section are:

- Beta-Binomial ($T = \sum Y \sim \text{Binom}(N, \theta)$ with $\theta \sim$ Beta)
- Exponential-Gamma ($Y \sim \text{Exp}(\theta)$ data with $\theta \sim$ Gamma)
- Poisson-Gamma ($Y \sim \text{Poi}(\theta)$ data $\theta \sim$Gamma)
- Normal-Inverse Gamma ($Y \sim \text{Normal}(\theta, \sigma)$ data with $\theta \sim$ Normal and $\sigma \sim$ Inverse Gamma)

Throughout this section, N is used for the observed data sample size, and S is the number of predictions desired. The predicted future result is indicated by surmounting the data variable with a tilde. For example, when the observed data is represented by y, the prediction is designated $\tilde{y}$ Parameter variable names are clearly identified as they appear.

3.1 Prediction of Future Successes: Beta-Binomial (Geisser p. 73)

3.1.1 Derivation

Let $Y_1, ..., Y_N$ be independent binary variables with $\Pr(Y_i = 1) = \theta$, with $Y_i = 1$ indicating success and $Y_i = 0$ indicating failure. The number of observed successes can be represented by $T = \sum Y_i$, which is sufficient for θ and has a binomial(N, θ) distribution. That is,

$$Pr(T = t|\theta) = \binom{N}{t} \theta^t (1-\theta)^{N-t}.$$

Assuming $\theta \sim \text{Beta}(\alpha, \boldsymbol{\beta})$, we have prior distribution

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}.$$

The posterior distribution of θ given $Y_1, ..., Y_N$, then, is

$$p(\theta|Y_1,...,Y_N) = p(\theta|t) = \frac{p(t|\theta)\pi(\theta)}{\int p(t|\theta)\pi(\theta)d\theta}$$

$$= \frac{\binom{N}{t}\theta^t(1-\theta)^{N-t}\frac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}}{\int \binom{N}{t}\theta^t(1-\theta)^{N-t}\frac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}d\theta}$$

$$= \frac{\theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}}{\int \theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}d\theta}$$

$$= \frac{\Gamma(N+\alpha+\beta)}{\Gamma(t+\alpha)\Gamma(N-t+\beta)}\theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}$$

$$= \text{Beta}(t+\alpha, N-t+\beta)$$

Note that the ratio of Gamma functions in the final step appears as a scaling constant that enables the Beta$(t+\alpha, N-t+\beta)$ density function under the integrand in the denominator of the previous step to resolve to 1.

We want to predict the number $\tilde{T}$ of successes out of M future observations. That is, $\tilde{T} = \sum_{i=1}^{M} Y_{N+i}$, and we have Beta-Binomial predictive distribution

$$\Pr[\tilde{T} = \tilde{t}|T = t] = \int p(\tilde{T} = \tilde{t}|\theta)p(\theta|t)d\theta$$

$$= \int \binom{M}{\tilde{t}}\theta^{\tilde{t}}(1-\theta)^{M-\tilde{t}}\frac{\Gamma(N+\alpha+\beta)}{\Gamma(t+\alpha)\Gamma(N-t+\beta)}\theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}d\theta$$

$$= \frac{M!}{\tilde{t}!(M-\tilde{t})!}\frac{\Gamma(N+\alpha+\beta)}{\Gamma(t+\alpha)\Gamma(N-t+\beta)}\int \theta^{\tilde{t}+t+\alpha-1}(1-\theta)^{M+N-\tilde{t}-t+\beta-1}d\theta$$

$$= \frac{\Gamma(M+1)\Gamma(N+\alpha+\beta)\Gamma(\tilde{t}+t+\alpha)\Gamma(M+N-\tilde{t}-t+\beta)}{\Gamma(\tilde{t}+1)\Gamma(M-\tilde{t}+1)\Gamma(t+\alpha)\Gamma(N-t+\beta)\Gamma(M+N+\alpha+\beta)}, \quad (2)$$

an impressive combination of Gamma functions. Note that the last two factors in the numerator together with the final factor in the denominator comprise the reciprocal of the scale factor corresponding with the Beta$(\tilde{t}+t+\alpha, M+N-\tilde{t}-t+\beta)$ kernel in the integrand, enabling the integral to resolve to 1.

3.1.2 R Implementation (Beta-Binomial)

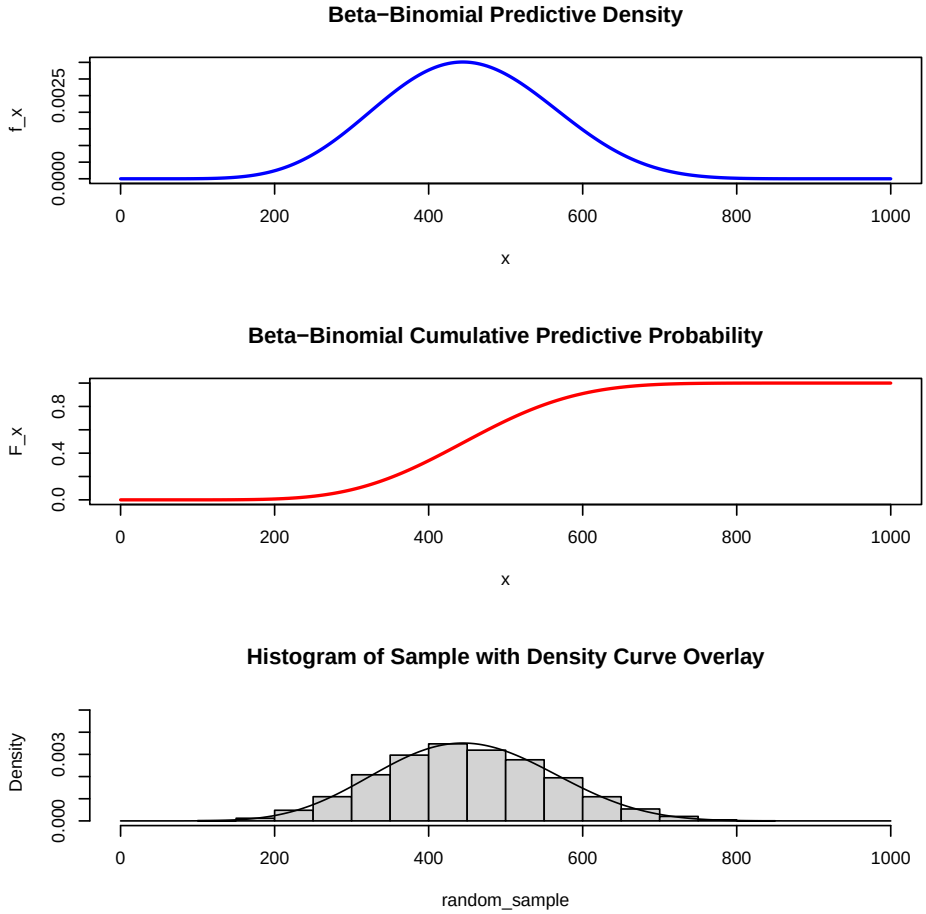
This result has been used to create R functions `dpredBB()`, `ppredBB()`, and `rpredBB()` for the Beta-Binomial predictive distribtuion for density, cumulative probability, and ran-

dom sampling, respectively (see appendix for the R code). `dpredBB()` and `rpredBB()` were used in the Pass the Pigs example in the introduction. The following generic example exercises all three functions.

The density function `dpredBB()` relies on the R function `lgamma()` to evaluate the numerator and denominator factor by factor logarithmically, and then exponentiates for the final result, evaluated at each integer value from 1 to the total number of future trials. The cdf `ppredBB()` simply calls `dpredBB()` and returns the cumulative sum of that discrete set of results. The random sampler `rpredBB()` makes use of the inverse transform method and the output from the cdf `ppredBB()`.

3.1.3 Example

Recapitulating the Pass the Pigs® example, suppose $t = 7$ Razorbacks have been observed out of $n = 10$ tosses of Big Pigs™, and the researcher has settled on prior Pr(Razorback) $= \theta \sim \text{Beta}(2, 8)$. For $N = 1000$ future observations, how many successes are predicted? The figures below show the predictive distribution from `dpredBB()`, the cumulative distribution from `ppredBB()`, and a histogram of random draws from `rpredBB()`.



3.2 Survival Time: Exponential-Gamma (Geisser p. 74)

3.2.1 Derivation

Suppose $Y_1, ..., Y_d$ represent fully observed copies from an exponential survival time density

$$p(y|\theta) = \theta e^{-\theta y}$$

and $Y_{d+1}, ..., Y_N$ represent censored copies surviving beyond the experimental time limit.

The usual exponential likelihood is used for the fully observed copies, whereas for the censored copies we need $\Pr(Y > y|\theta) = 1 - \Pr(Y \leq y|\theta) = 1 - F(y|\theta) = 1 - (1 - e^{-\theta y}) = e^{-\theta y}$. Here F denotes the cumulative distribution function.

Thus, with $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, the overall likelihood is

$$L(\theta|y) = \prod_{i=1}^{d} \theta e^{-\theta y_i} \prod_{i=d+1}^{N} e^{-\theta y_i} = \theta^d e^{-\theta N \bar{y}}$$

Assuming a Gamma(δ, γ) prior for θ,

$$\pi(\theta) = \frac{\gamma^\delta \theta^{\delta-1} e^{-\gamma\theta}}{\Gamma(\delta)}$$

we obtain the posterior

$$\begin{aligned}
p(\theta|Y_1, ..., Y_N) &= \frac{p(Y_1, ..., Y_N|\theta)\pi(\theta)}{\int p(Y_1, ..., Y_N|\theta)\pi(\theta)d\theta} \\
&= \frac{\theta^d e^{-\theta N\bar{y}} \cdot \frac{\gamma^\delta \theta^{\delta-1} e^{-\gamma\theta}}{\Gamma(\delta)}}{\int \left(\theta^d e^{-\theta N\bar{y}} \cdot \frac{\gamma^\delta \theta^{\delta-1} e^{-\gamma\theta}}{\Gamma(\delta)}\right) d\theta} \\
&= \frac{\left(\theta^{d+\delta-1} e^{-\theta(\gamma+N\bar{y})}\right)}{\int \left(\theta^{d+\delta-1} e^{-\theta(\gamma+N\bar{y})}\right) d\theta} \\
&= \frac{(\gamma + N\bar{y})^{d+\delta}}{\Gamma(d+\delta)} \theta^{d+\delta-1} e^{-\theta(\gamma+N\bar{y})} \\
&= \text{Gamma}(d+\delta, \gamma + N\bar{y}),
\end{aligned}$$

with the Gamma$(d+\delta, \gamma+N\bar{y})$ density in the denominator of next to last step integrating to 1.

The survival time predictive probability density then is

$$p(\tilde{Y} = \tilde{y}|Y_1, ..., Y_N) = \int p(\tilde{y}|\theta)p(\theta|y_1, ..., y_N)d\theta$$

$$= \int \theta e^{-\theta\tilde{y}} \cdot \frac{(\gamma + N\bar{y})^{d+\delta}\theta^{d+\delta-1}e^{-\theta(\gamma+N\bar{y})}}{\Gamma(d+\delta)}d\theta$$

$$= (d+\delta)(\gamma + N\bar{y})^{d+\delta} \int \frac{\theta^{(d+\delta+1)-1}e^{-\theta(\gamma+N\bar{y}+\tilde{y})}}{(d+\delta)\Gamma(d+\delta)}d\theta$$

$$= \frac{(d+\delta)(\gamma + N\bar{y})^{d+\delta}}{(\gamma + N\bar{y} + \tilde{y})^{d+\delta+1}} \int \frac{(\gamma + N\bar{y} + \tilde{y})^{d+\delta+1}\,\theta^{(d+\delta+1)-1}e^{-\theta(\gamma+N\bar{y}+\tilde{y})}}{\Gamma(d+\delta+1)}d\theta$$

$$= \frac{(d+\delta)(\gamma + N\bar{y})^{d+\delta}}{(\gamma + N\bar{y} + \tilde{y})^{d+\delta+1}}, \tag{3}$$

simplifying here by constructing a Gamma$(d + \delta + 1, \gamma + N\bar{y} + \tilde{y})$ density in the final integrand.

3.2.2 R Implementation (Exponential-Gamma)

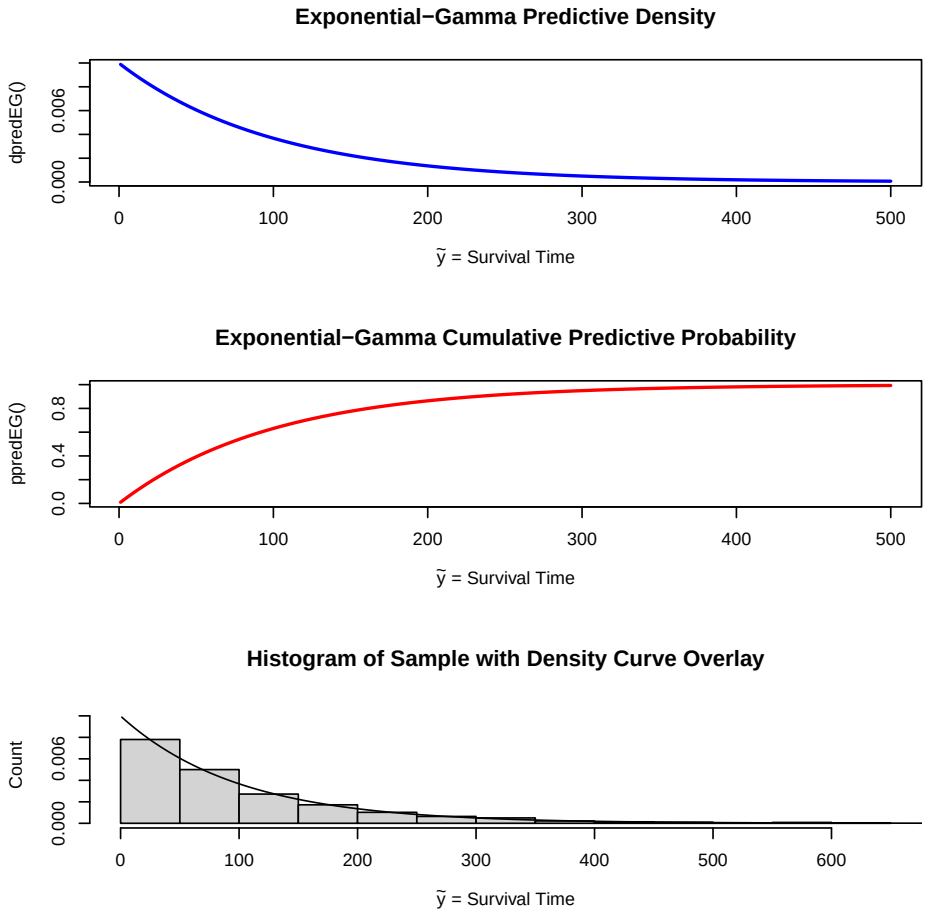
This result has been used to create R functions `dpredEG()`, `ppredEG()`, and `rpredEG()` for the Gamma-Exponential distribtuion for density, cumulative probability, and random sampling, respectively (see appendix for R code). These functions are exercised in the following example.

The density function `dpredEG()` evaluates the numerator and denominator of the predictive density logarithmically (using the R function `log()`) and then exponentiates to produce the result. The cdf `ppredEG()` integrates the pdf at each discrete value using the R function `integrate()`. The random sampler `rpredEG()` draws posterior $\theta|y_1, ..., y_i \sim$ Gamma$(d + \delta, \gamma + \sum y_i)$ and then draws predictions from Exp(θ).

3.2.3 Example

Suppose $d = 797$ out of $N = 1000$ copies have been observed, and the remaining 203 censored. We are interested in the number of survivors out of $M = 1000$ future observations. For the sake of this example, survival times were generated randomly as $y_i \sim$ Exp(0.01), with all times exceeding an arbitrary predetermined time cutoff of 160 being set to that time. These reset times are the censored copies. We know the value of the exponential parameter ($\theta = 0.01$) used to generate the data, because of the artificial way we have constructed this example. If we did not have that information, we could consider the mean of the generated survival times (observed and censored), which was 79.9. This indicates that the expected value $1/\theta$ of survival times must be greater than 79.9, and thus

we can assume $\theta < 1/79.9 \approx 0.0125$ with some margin. This is all to say that aiming for $\theta = 0.01$ when choosing parameters for the prior distribution of θ is reasonable. Since we are assuming prior $\theta \sim \text{Gamma}(\delta, \gamma)$ we want $\delta/\gamma \approx 0.01$. For this example, say $\delta = 0.5$ and $\gamma = 50$. The figures below illustrate the predictive probability using `dpredEG()` and `rpredEG()`, along with a histogram of a random sample taken using `rpredEG()`.



3.3 Poisson-Gamma Model (Hoff p. 43ff)

3.3.1 Derivation

Suppose we have count data observations $Y_1, ..., Y_N|\theta \overset{i.i.d.}{\sim} \text{Poisson}(\theta)$, with θ assumed to have prior Gamma(α, β) distribution. That is,

$$\begin{aligned} P\left(Y_1 = y_1, ..., Y_N = y_N|\theta\right) &= \prod_{i=1}^{N} p\left(y_i|\theta\right) \\ &= \prod_{i=1}^{N} \frac{1}{y!}\theta^{y_i}e^{-\theta} \\ &= \left(\prod_{i=1}^{N} \frac{1}{y!}\right)\theta^{\sum y_i}e^{-N\theta} \\ &= c\left(y_1, ..., y_N\right)\theta^{\sum y_i}e^{-N\theta} \end{aligned}$$

and

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \text{ with } \theta, \alpha, \beta > 0.$$

Then we have posterior

$$\begin{aligned} p\left(\theta|y_1, ..., y_N\right) &= \frac{p\left(y_1, ..., y_N|\theta\right)\pi(\theta)}{\int_\theta p\left(y_1, ..., y_N|\theta\right)p(\theta)} \\ &= \frac{p\left(y_1, ..., y_N|\theta\right)\pi(\theta)}{p\left(y_1, ..., y_N\right)} \\ &= \frac{1}{p\left(y_1, ..., y_N\right)}\theta^{\sum y_i}e^{-N\theta}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\ &= C\left(y_1, ..., y_N, \alpha, \beta\right)\theta^{\alpha+\sum y_i-1}e^{-(\beta+N)\theta} \\ &\propto \text{Gamma}\left(\alpha + \sum y_i, \beta + N\right). \end{aligned}$$

Here

$$
\begin{aligned}
C\left(y_1, ..., y_N, \alpha, \beta\right) &= \frac{1}{p\left(y_1, ..., y_N\right)} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \frac{1}{\int_\theta p\left(y_1, ..., y_N | \theta\right) \pi(\theta)} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \frac{1}{\int_\theta \left(\prod \frac{1}{y_i!}\right) \theta^{\sum y_i} e^{-N\theta} \cancel{\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)} \theta^{\alpha-1} e^{-\beta\theta}} \cdot \cancel{\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)} \\
&= \frac{1}{\left(\prod \frac{1}{y_i!}\right) \frac{\Gamma(\alpha+\sum y_i)}{(\beta+N)^{\alpha+\sum y_i}} \int_\theta \frac{(\beta+N)^{\alpha+\sum y_i}}{\Gamma(\alpha+\sum y_i)} \theta^{\sum y_i+\alpha-1} e^{-(\beta+N)\theta}} \\
&= \frac{\prod_{i=1}^N y_i! (\beta+N)^{\alpha+\sum y_i}}{\Gamma(\alpha+\sum y_i)}
\end{aligned}
$$

Call this constant C_N (for N observations).

Note that with an additional observation $\tilde{y} = y_{N+1}$ the constant becomes

$$C_{N+1} = \frac{\prod_{i=1}^{N+1} y_i! (\beta+N+1)^{\alpha+\sum_{i=1}^{N+1} y_i}}{\Gamma(\alpha+\sum_{i=1}^{N+1} y_i)}.$$

Also note that the marginal joint distribution of k observations is

$$p(y_1, ..., y_k) = \frac{1}{C_k} \frac{\beta^\alpha}{\Gamma(\alpha)}.$$

For future observation $\tilde{y}$, then, we compute predictive distribution

$$p\left(\tilde{y}|y_1,...,y_N\right) = \frac{p\left(y_1,...,y_N,\tilde{y}\right)}{p\left(y_1,...,y_N\right)} = \frac{p\left(y_1,...,y_{N+1}\right)}{p\left(y_1,...,y_N\right)} = \frac{\frac{1}{C_{N+1}}\cancel{\frac{\beta^\alpha}{\Gamma(\alpha)}}}{\frac{1}{C_N}\cancel{\frac{\beta^\alpha}{\Gamma(\alpha)}}} = \frac{C_N}{C_{N+1}}$$

$$= \frac{\frac{\prod_{i=1}^{N} y_i!(\beta+N)^{\alpha+\sum_{i=1}^{N} y_i}}{\Gamma(\alpha+\sum_{i=1}^{N} y_i)}}{\frac{\prod_{i=1}^{N+1} y_i!(\beta+N+1)^{\alpha+\sum_{i=1}^{N+1} y_i}}{\Gamma(\alpha+\sum_{i=1}^{N+1} y_i)}}$$

$$= \frac{\Gamma\left(\alpha+\sum_{i=1}^{N+1} y_i\right)(\beta+N)^{\alpha+\sum_{i=1}^{N} y_i}}{(y_{N+1}!)\,\Gamma\left(\alpha+\sum_{i=1}^{N} y_i\right)(\beta+N+1)^{\alpha+\sum_{i=1}^{N+1} y_i}}$$

$$= \frac{\Gamma\left(\alpha+\sum_{i=1}^{N} y_i+\tilde{y}\right)(\beta+N)^{\alpha+\sum_{i=1}^{N} y_i}}{(\tilde{y}!)\,\Gamma\left(\alpha+\sum_{i=1}^{N} y_i\right)(\beta+N+1)^{\alpha+\sum_{i=1}^{N} y_i+\tilde{y}}}$$

$$= \frac{\Gamma\left(\alpha+\sum y_i+\tilde{y}\right)}{\Gamma(\tilde{y}+1)\Gamma(\alpha+\sum y_i)} \cdot \left(\frac{\beta+N}{\beta+N+1}\right)^{\alpha+\sum y_i} \cdot \left(\frac{1}{\beta+N+1}\right)^{\tilde{y}} \quad (4)$$

This is a negative binomial distribution: $\tilde{y} \sim \mathrm{NB}\left(\alpha+\sum y_i, \beta+N\right)$

3.3.2 R Implementation (Poisson-Gamma)

This result has been used to create R functions `dpredPG()`, `ppredPG()`, and `rpredPG()` for the Poisson-Gamma distribution for density, cumulative probability, and random sampling, respectively (see appendix for R code). These functions are exercised in the example below.

The density function `dpredPG()` simply makes use of the R function `dnbinom()`. The cdf `ppredPG()` returns a cumulative sum of the results of `dpredPG()`. The random sampler `rpredPG()` is a bit more complicated. The difficulty is that the upper bound of the support of the predictive distribution $p(\tilde{y}|y_1,...,y_n)$ is not known. Since $p(\cdot)$ is negative binomial, we can count on it eventually decreasing toward 0 asymptotically. To establish the support, then, a method was employed to find where p comes "sufficiently close" to 0. To accomplish this a modified bisection method was devised as follows:

1. Set a desired tolerance ϵ for the distance of the predictive distribution above zero at the upper end of its support. Currently the function leverages the local device's numerical precision with $\epsilon =$ `sqrt(.Machine$double.eps)`.

2. Set lower endpoint L equal to the expected value of $\tilde{Y}$. That is, $L = E(\tilde{Y}|y_1, ..., y_N) = \frac{\alpha + \sum y_i}{\beta + N}$ (negative binomial).

3. Step to the right of L by increments in the sequence $\{L + 2^k : k = 1, 2, ...\}$, setting upper endpoint $U_{test} = L+2^k$ using the first value of k for which $\texttt{dpredPG}\left(L + 2^k\right) < 0$.

4. Bisect $[L, U_{test}]$ letting B be the integer nearest to the middle of the interval.

5. Test whether $0 \leq \texttt{dpredPG}(B) \leq \epsilon$. If so, accept $U = B$ as the upper end of the support. If not:

6. Establish a new interval $[L', U'_{test}]$ as follows:

   (a) if $[\texttt{dpredPG}(L), \texttt{dpredPG}(B)]$ straddles 0, then $[L', U'_{test}] = [L, B]$.

   (b) if $[\texttt{dpredPG}(B), \texttt{dpredPG}(U_{test})]$ straddles 0, then $[L', U'_{test}] = [B, U_{test}]$.

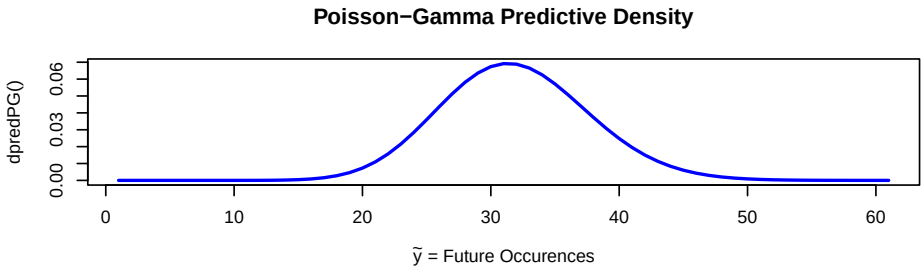
7. Repeat steps 3 - 5 above with the updated interval $[L, U_{test}] = [L', U'_{test}]$ until the condition in step 5 is reached, establishing upper bound U.

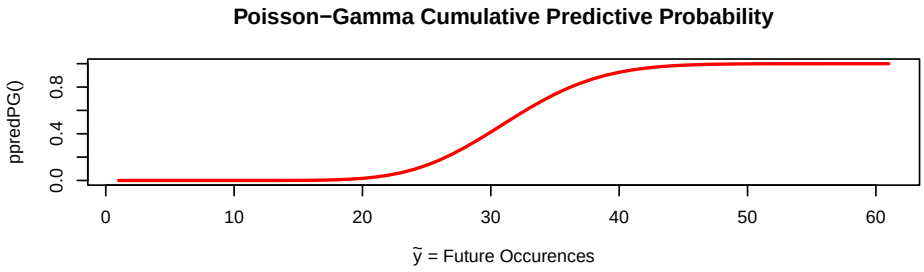
8. Use $[0, U]$ as the support upon which to draw a random sample.

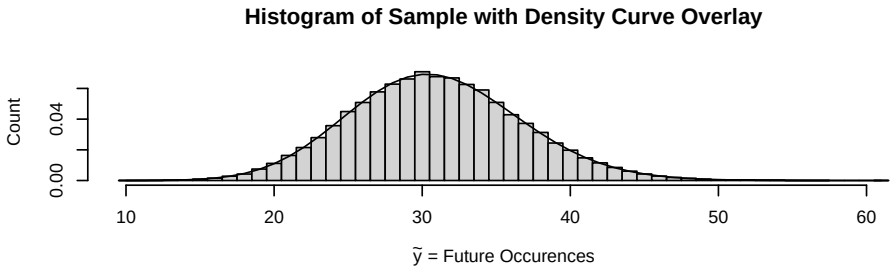
3.3.3 Example

I figured out the source of this example. I originally picked 10 random integers for the counts and liked the way the curves came out, so I kept them. My prior parameter choices were equally arbitrary. I could reproduce Hoff's birthrate example (p. 49) instead, which would be nicer than this completely abstracted example, but I hesitate to do that just because of the extra work. Thoughts?

Suppose we have 10 prior observations with counts 27, 79, 21, 100, 8, 4, 37, 15, 3, 97. Let $\alpha = 11$ and $\beta = 3$. For $\tilde{y} = 1 : 100$ possible future occurrences, the figures below show the predictive distribution from `dpredPG()`, the cumulative distribution from `ppredPG()`, and a histogram of random draws from `rpredPG()`.


Poisson−Gamma Predictive Density
dpredPG()
0.00
0.03
0.06
0
10
20
30
40
50
60
ỹ = Future Occurences


Poisson−Gamma Cumulative Predictive Probability
ppredPG()
0.0
0.4
0.8
0
10
20
30
40
50
60
ỹ = Future Occurences


Histogram of Sample with Density Curve Overlay
Count
0.00
0.04
10
20
30
40
50
60
ỹ = Future Occurences

3.4 Normal Observation with Normal-Inverse Gamma Prior

3.4.1 One sample Normal-Inverse Gamma

3.4.1.1 Derivation [Hoff p. 69ff]

Let $\{Y_1, ..., Y_N|\theta, \sigma^2\} \overset{i.i.d.}{\sim} \text{Normal}(\theta, \sigma^2)$. Then the joint sampling density is

$$\begin{aligned} p\left(y_1, ..., y_N|\theta, \sigma^2\right) &= \prod_{i=1}^{N} p\left(y_i|\theta, \sigma^2\right) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i-\theta}{\sigma}\right)^2} \\ &= \left(2\pi\sigma^2\right)^{-N/2} e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i-\theta}{\sigma}\right)^2}. \end{aligned}$$

Following Hoff (p. 74ff), for joint inference on both θ and σ, assume priors

$$\frac{1}{\sigma^2} \sim \text{Gamma}\left(\nu_0/2, \nu_0\sigma_0^2/2\right)$$

$$\theta|\sigma^2 \sim \text{Normal}\left(\mu_0, \sigma^2/\kappa_0\right)$$

where (σ_0^2, ν_0) are the sample variance and sample size of prior observations, and (μ_o, κ_0) are the sample mean and sample size of prior observations.

From this we derive joint posterior distribution

$$\left\{\theta|y_1, ..., y_N, \sigma^2\right\} \sim \text{Normal}\left(\mu_N, \sigma^2/\kappa_N\right)$$

$$\left\{\sigma^2|y_1, ..., y_N\right\} \sim \text{Inverse Gamma}\left(\nu_N/2, \sigma_N^2\nu_N/2\right).$$

where

$$\kappa_N = \kappa_0 + N$$

$$\mu_N = \frac{\kappa_0\mu_0 + N\bar{y}}{\kappa_N}$$

$$\nu_n = \nu_0 + N$$

$$\sigma_N^2 = \frac{1}{\nu_N}\left[\nu_0\sigma_0^2 + (N-1)s^2 + \frac{\kappa_0 N}{\kappa_N}\left(\bar{y} - \mu_0\right)^2\right].$$

Here $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the sample mean and $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$ is the sample variance.

From the joint posterior distribution we generate marginal samples by means of the Monte Carlo method (Hoff, p. 77):

$$\begin{array}{cc} \sigma^{2(1)} \sim \text{Inverse Gamma}\left(\nu_N/2, \sigma_N^2\nu_N/2\right), & \theta^{(1)} \sim \text{Normal}\left(\mu_N, \sigma^{2(1)}/\kappa_N\right) \\ \vdots & \vdots \\ \sigma^{2(S)} \sim \text{Inverse Gamma}\left(\nu_N/2, \sigma_N^2\nu_N/2\right), & \theta^{(S)} \sim \text{Normal}\left(\mu_N, \sigma^{2(S)}/\kappa_N\right) \end{array}$$

For prediction of future $\tilde{y}|y_1, ..., y_N, \theta, \sigma^2$, generate $\tilde{y}_i \sim \text{Normal}\left(\theta^{(i)}, \sigma^{2(i)}\right)$.

For prediction without the influence of any previous knowledge (Hoff p. 79), we can employ Jeffreys prior $\tilde{p}(\theta, \sigma^2) = 1/\sigma^2$. This leads to the same conditional distribution for θ but a gamma$\left(\frac{N-1}{2}, \frac{1}{2}\sum(y_i - \bar{y})^2\right)$ distribution for $1/\sigma^2$. This joint posterior distribution can be used to predict future $\tilde{y}$ by first drawing θ, σ^2 and then simulating $\tilde{y} \sim \text{Normal}(\theta, \sigma^2)$. Alternatively, the joint posterior can be integrated to show that

$$\frac{\theta - \bar{y}}{s/\sqrt{N}}|y_1, ..., y_N \sim t_{N-1}.$$

The resulting predictive distribution for $\tilde{y}$ is a t-distribution with location $\bar{y}$ and scale $s\sqrt{1 + 1/N}$ and $N - 1$ degrees of freedom (Gelman et. al. p. 66).

3.4.1.2 R Implementation (Normal-Inverse Gamma, 1-sample) R functions `dpredNormIG1()`, `ppredNormIG1()`, and `rpredNormIG1()` have been created for the Normal-Inverse Gamma distribution for density, cumulative probability, and random sampling, respectively (see appendix). These functions all include options for implementation with or without previous knowledge as desired. If Jeffreys prior is used, the functions simply implement R's Student's t-distribution functions `rt()`, `dt()`, and `pt()`, applying the location and scale parameters as described above. For predictions using previous knowledge, the functions work as follows: For the random sampler `rpredNormIG1()`, the Monte-Carlo method described above is directly employed. The predictive density and cumulative predictive density functions (`dpredNormIG1()` and `ppredNormID1()`, respectively) depend on the random sample. `ppredNormIG1()` utilizes the empirical cumulative density function `ecdf()` from R's stats package. `dpredNormIG1()` utilizes a Kernel Density Estimation (KDE) method and R's built-in `density()` function. The KDE is computed by definition, using a Normal kernel:

$$\hat{f}_K(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x - X_i}{h}\right),$$

where

X_i is the random sample generated using `rpredNormIG1()`

K is Normal$(0, 1)$

h is the bandwidth from R's `density()` function (that is, `h = density(X_i)$bw`)

These functions are exercised in the following example.

3.4.1.3 Example *Example (Hoff p. 72ff, using data from Grogan and Wirth (1981)): Midge wing length*

Grogan and Wirth (1981) provide 9 measurements of midge wing length, in millimeters: $y = \{1.64, 1.7, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08\}$. Previous studies suggest values $\mu_0 = 1.9$ and $\sigma_0^2 = 0.01$. We choose $\kappa_0 = \nu_0 = 1$ "...so that our prior distributions are only weakly centered around these estimates from other populations" (Hoff p. 76). We compute

$$\bar{y} = 1.804$$

$$\text{var}(y) = 0.0169$$

$$\kappa_N = 1 + 9 = 10$$

$$\mu_N = \frac{1 \cdot 1.9 + 9 \cdot 1.804}{10} = 1.814$$

$$\nu_N = 1 + 9 = 10$$

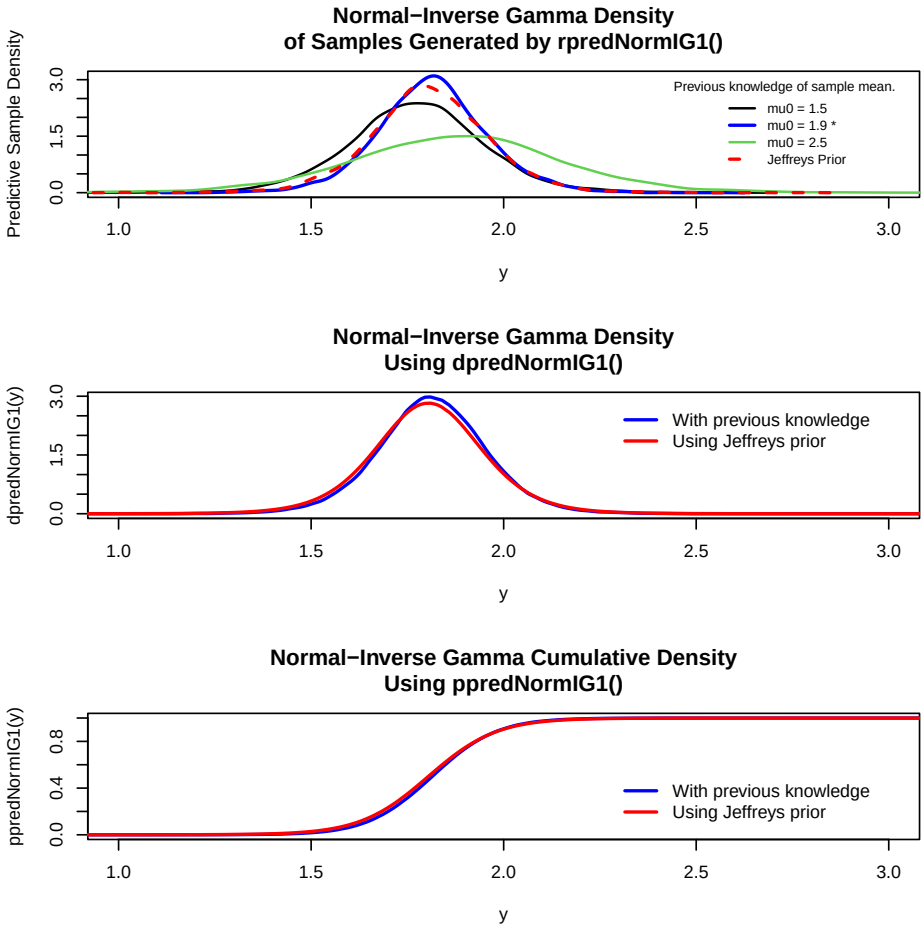
$$\sigma_N^2 = \frac{1}{10}\left[1 \cdot 0.01 + (9-1) \cdot 0.0169 + \frac{1 \cdot 9}{10}(1.804 - 1)^2\right] = 0.0153$$

Thus $\nu_N/2 = 5$ and $\nu_N\sigma_N^2/2 = 0.7662$ and we have posteriors

$$\left\{\theta | y_1, ..., y_N, \sigma^2\right\} \sim \text{Normal}\left(1.814, \sigma^2/10\right)$$

$$\left\{\sigma^2 | y_1, ..., y_N\right\} \sim \text{Inverse Gamma}(5, 0.7662)$$

The plot below illustrates the influence of previous knowledge of the population mean, and compares to the predictions resulting from Jeffreys prior.



3.4.2 Two-sample Normal-Inverse Gamma

3.4.2.1 Derivation For a Bayesian analysis comparing two groups $Y_{1,1}, ..., Y_{N_1,1}$ and $Y_{1,2}, ..., Y_{N_2,2}$ we use the following sampling model (Hoff p. 127):

$$\begin{aligned} Y_{i,1} &= \mu + \delta + \epsilon_{i,1} \\ Y_{i,2} &= \mu - \delta + \epsilon_{i,2} \\ \{\epsilon_{i,j}\} &\sim \text{i.i.d. Normal}\left(0, \sigma^2\right). \end{aligned}$$

Letting $\theta_1 = \mu + \delta$ and $\theta_2 = \mu - \delta$ we see that $\delta = (\theta_1 - \theta_2)/2$ is half the population difference in means, and $\mu = (\theta_1 + \theta_2)/2$ is the pooled average. We'll assume conjugate prior distributions

$$\begin{aligned} p\left(\mu, \delta, \sigma^2\right) &= p(\mu) \times p(\delta) \times p\left(\sigma^2\right) \\ \mu &\sim \text{Normal}\left(\mu_0, \gamma_0^2\right) \\ \delta &\sim \text{Normal}\left(\delta_0, \tau_0^2\right) \\ \sigma^2 &\sim \text{Inverse Gamma}\left(\nu_0/2, \nu_0\sigma_0^2/2\right), \end{aligned}$$

where ν_0 as before is the assumed prior sample size. The full conditional distributions follow:

$\{\mu|\mathbf{y}_1, \mathbf{y}_2, \delta, \sigma^2\} \sim \text{Normal}\left(\mu_N, \gamma_N^2\right)$, where

$$\mu_N = \gamma_N^2 \times \left[\frac{\mu_0}{\gamma_0^2} + \frac{\sum_{i=1}^{N_1}(y_{i,1} - \delta) + \sum_{i=1}^{N_2}(y_{i,2} + \delta)}{\sigma^2}\right]$$

$$\gamma_N^2 = \left[\frac{1}{\gamma_0^2} + \frac{(N_1 + N_2)}{\sigma^2}\right]^{-1}$$

$\{\delta|\mathbf{y}_1, \mathbf{y}_2, \mu, \sigma^2\} \sim \text{Normal}\left(\delta_N, \tau_N^2\right)$, where

$$\delta_N = \tau_N^2 \times \left[\frac{\delta_0}{\tau_0^2} + \frac{\sum_{i=1}^{N_1}(y_{i,1} - \mu) - \sum_{i=1}^{N_2}(y_{i,2} - \mu)}{\sigma^2}\right]$$

$$\tau_N^2 = \left[\frac{1}{\tau_0^2} + \frac{(N_1 + N_2)}{\sigma^2}\right]^{-1}$$

$\{\sigma^2|\mathbf{y}_1, \mathbf{y}_2, \mu, \delta\} \sim \text{Inverse Gamma}\left(\frac{\nu_N}{2}, \frac{\nu_N\sigma_N^2}{2}\right)$, where

$$\nu_N = \nu_0 + N_1 + N_2$$

$$\nu_N \sigma_N^2 = \nu_0 \sigma_0^2 + \sum_{i=1}^{N_1} \left(y_{i,1} - [\mu + \delta]\right)^2 + \sum_{i=1}^{N_2} \left(y_{i,2} - [\mu - \delta]\right)^2$$

3.4.2.2 R Implementation (Normal-Inverse Gamma, 2-sample) The R function `rpredNormIG2()` implements a Gibbs sampler to approximate the posterior distribution $p\left(\mu, \delta, \sigma^2 | \mathbf{y}_1, \mathbf{y}_2\right)$, from which to generate predictions for the two populations as follows:

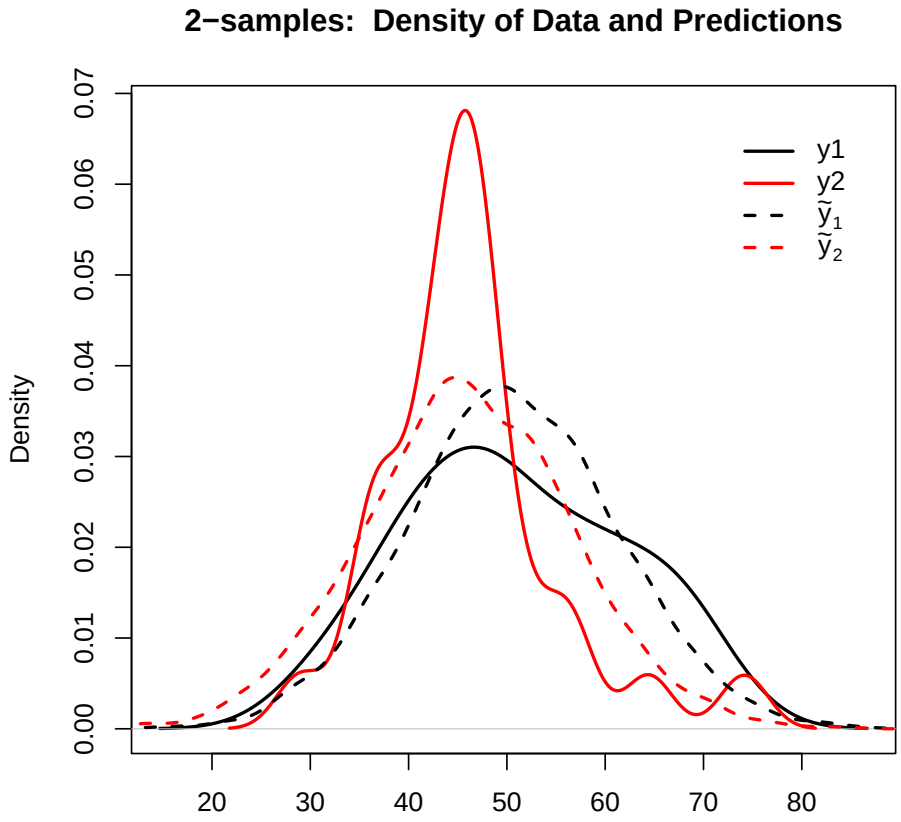
1. Set initial values $\mu = \frac{\theta_1 + \theta_2}{2}$ and $\delta = \frac{\theta_1 - \theta_2}{2}$
2. Generate a single $\sigma^2 | \mathbf{y_1}, \mathbf{y_2}, \mu, \delta$
3. Generate a single $\mu | \mathbf{y_1}, \mathbf{y_2}, \delta, \sigma^2$
4. Generate a single $\delta | \mathbf{y_1}, \mathbf{y_2}, \mu, \sigma^2$
5. Predict $\tilde{y}_1 \sim \text{Normal}\left(\mu + \delta, \sigma^2\right)$ and $\tilde{y}_2 \sim \text{Normal}\left(\mu - \delta, \sigma^2\right)$

The user provides the two samples $\mathbf{y_1}$ and $\mathbf{y_2}$ along with values for $\mu_0, \sigma_0^2, \delta_0, \tau_0^2, \nu_0$, and desired prediction sample size M. The function returns M predictions for each population and the vectors of generated values for μ, δ, and σ^2.

3.4.2.3 Example (Hoff p. 128-129 *Analysis of math score data*)
Hoff provides the following example, which we reproduce here.

Math score data for two schools were based on results of a national exam in the United States, standardized to produce a nationwide mean of 50 and a standard deviation of 10. Unless the two schools were known in advance to be extremely exceptional, reasonable prior parameters can be based on this information. For the prior distributions of μ and σ^2, we'll take $\mu_0 = 50$ and $\sigma_0^2 = 10^2 = 100$, although this latter value is likely to be an overestimate of the within-school sampling variability. We'll make these prior distributions somewhat diffuse, with $\gamma_0^2 = 25^2 = 625$ and $\nu_0 = 1$. For the prior distribution on δ, choosing $\delta_0 = 0$ represents the prior opinion that $\theta_1 > \theta_2$ and $\theta_2 > \theta_1$ are equally probable. Finally, since the scores are bounded between 0 and 100, half the difference between θ_1 and θ_2 must be less than 50 in absolute value, so a value of $\tau_0^2 = 25^2 = 625$ seems reasonably diffuse.

The results of a call to `rpredNormIG2()` with the above input values for $\mathbf{y}_1, \mathbf{y}_2, \mu_0, \sigma_0^2, \delta_0, \tau_0^2,$ and N are summarized in the following plot.



3.4.3 k-sample Normal-Inverse Gamma: Comparing multiple groups

For two-level data consisting of groups and units within groups, denote data $\mathbf{Y}_1, ..., \mathbf{Y_k}$ where $\mathbf{Y}_j = \{Y_{1,j}, ..., Y_{N_j,j}\}$. We have the hierarchical Normal model (Hoff p. 132ff):

$$\phi_j = \left\{\theta_j, \sigma^2\right\}, p\left(y|\phi_j\right) = \text{Normal}\left(\theta_j, \sigma^2\right) \text{ (within-group model)}$$

$$\psi_j = \left\{\mu, \tau^2\right\}, p\left(\theta_j|\psi\right) = \text{Normal}\left(\mu, \tau^2\right) \text{ (between-groups model)}$$

We use standard semiconjugate Normal and Inverse Gamma prior distributions for the fixed but unknown parameters in the model:

$$\sigma^2 \sim \text{Inverse Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

$$\tau^2 \sim \text{Inverse Gamma}\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right)$$

$$\mu \sim \text{Normal}\left(\mu_0, \gamma_0^2\right)$$

3.4.3.1 Derivation As with the two-sample problem, joint posterior inferences for the unknown parameters can be made by constructing a Gibbs sampler to approximate the posterior distribution $p\left(\theta_1, ..., \theta_k, \mu, \tau^2, \sigma^2 | \mathbf{y_1}, ..., \mathbf{y_k}\right)$. For this we need the full conditional distribution of each parameter (Hoff pp. 134-135):

$$\left\{\mu | \theta_1, ..., \theta_k, \tau^2\right\} \sim \text{Normal}\left(\frac{\frac{k\bar{\theta}}{\tau^2} + \frac{\mu_0}{\gamma_0^2}}{\frac{k}{\tau^2} + \frac{1}{\gamma_0^2}}, \frac{1}{\frac{k}{\tau^2} + \frac{1}{\gamma_0^2}}\right)$$

$$\left\{\tau^2 | \theta_1, ..., \theta_k, \mu\right\} \sim \text{Inverse Gamma}\left(\frac{\eta_0 + k}{2}, \frac{\eta_0 \tau_0^2 + \sum \left(\theta_j - \mu\right)^2}{2}\right)$$

$$\left\{\theta_j | y_{1,j}, ..., y_{n,j}, \sigma^2\right\} \sim \text{Normal}\left(\frac{\frac{n_j \bar{y}_j}{\sigma^2} + \frac{1}{\tau^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

$$\left\{\sigma^2 | \theta, \mathbf{y_1}, ..., \mathbf{y_k}\right\} \sim \text{Inverse Gamma}\left(\frac{1}{2}\left[\nu_0 + \sum_{j=1}^{k} n_j\right], \frac{1}{2}\left[\nu_0 \sigma_0^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{i,j} - \theta_j\right)^2\right]\right).$$

Note that $\sum \sum \left(y_{i,j} - \theta_j\right)^2$ is the sum of squared residuals across all groups, conditional on the within-group means, and so the conditional distribution concentrates probability around a pooled-sample estimate of the variance.

3.4.3.2 R Implementation (Normal-Inverse Gamma, k-samples) The R function `rpredNormIGk()` implements a Gibbs sampler for posterior approximation of each unknown quantity by sampling from its full conditional distribution. From these posteriors, predictions are generated, as follows:

1. Set prior parameter values:

$$\begin{aligned} &\nu_0, \sigma_0^2 \text{ for } p\left(\sigma^2\right) \\ &\eta_0, \tau_0^2 \text{ for } p\left(\tau^2\right) \\ &\mu_0, \gamma_0^2 \text{ for } p\left(\mu\right). \end{aligned}$$

2. Set initial states for the unknown parameters:

$$\begin{aligned} \theta_1^{(1)} = \bar{\mathbf{y}}_\mathbf{1}, ..., \theta_k^{(1)} &= \bar{\mathbf{y}}_\mathbf{k} \\ \mu^{(1)} &= \text{mean}\left(\theta_1^{(1)}, ..., \theta_k^{(1)}\right) \\ \tau^{2(1)} &= \text{var}\left(\theta_1^{(1)}, ..., \theta_k^{(1)}\right) \\ \sigma^{2(1)} &= \text{mean}\left(\text{var}\left(\mathbf{y}_1\right), ..., \text{var}\left(\mathbf{y}_k\right)\right)) \end{aligned}$$

3. For $s \in \{1, ..., S\}$, sample

   (a) $\mu^{(s+1)} \sim p\left(\mu | \theta_1^{(s)}, ..., \theta_k^{(s)}, \tau^{2(s)}\right)$

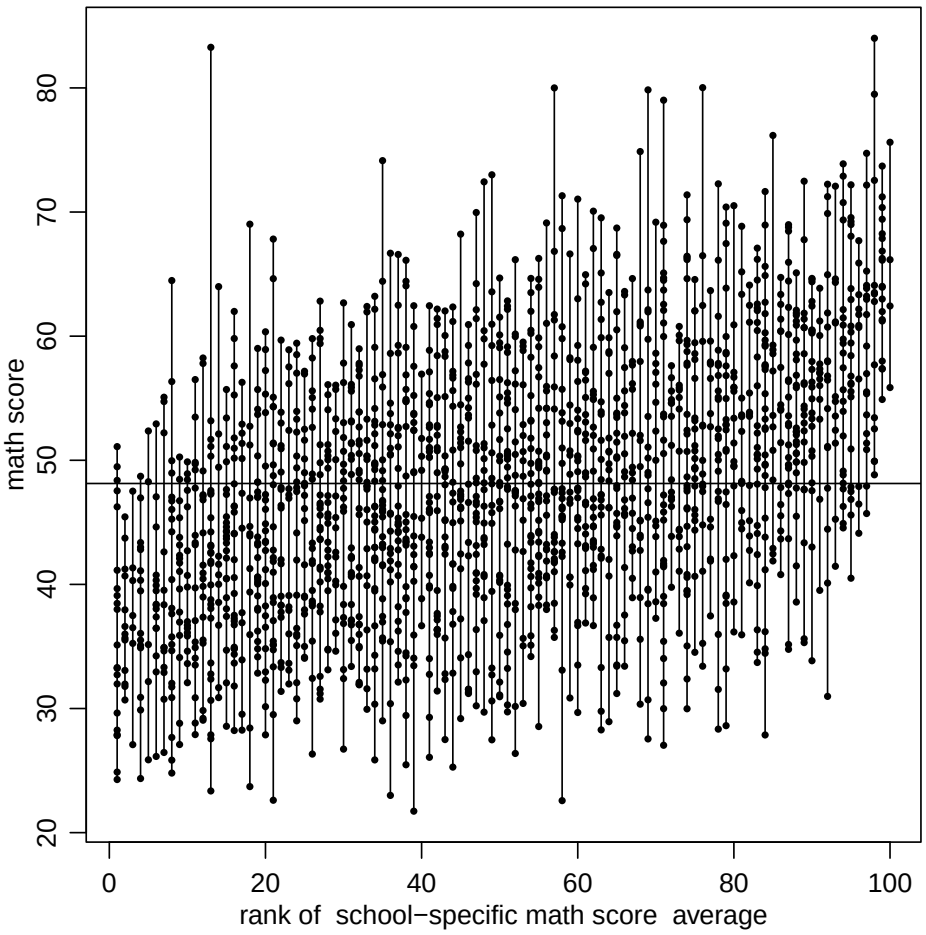
(b) $\tau^{2(s+1)} \sim p\left(\tau^2 | \theta_1^{(s)}, ..., \theta_k^{(s)}, \mu^{(s+1)}\right)$

(c) $\sigma^{2(s+1)} \sim p\left(\sigma^2 | \theta_1^{(s)}, ..., \theta_k^{(s)}, \mathbf{y}_1, ..., \mathbf{y}_k\right)$

(d) $\theta_j^{(s+1)} \sim p\left(\theta_j | \mu^{(s+1)}, \tau^{2(s+1)}, \sigma^{2(s+1)}, \mathbf{y}_j\right)$ for $j \in \{1, ..., k\}$

4. For $s \in \{1, ..., S\}$, generate prediction $\tilde{y}_j^{(s)} \sim \text{Normal}\left(\theta_j^{(s)}, \sigma^{2(s)}\right)$

3.4.3.3 Example Returning to the math scores example, data for 10th-grade students from 100 large urban schools (each having 10th-grade enrollment of at least 400) is summarized in the following plot. Each school's results are represented by a vertical segment, with points plotted at the individual students' math scores. The horizontal line is drawn at the grand mean of the individual school means.



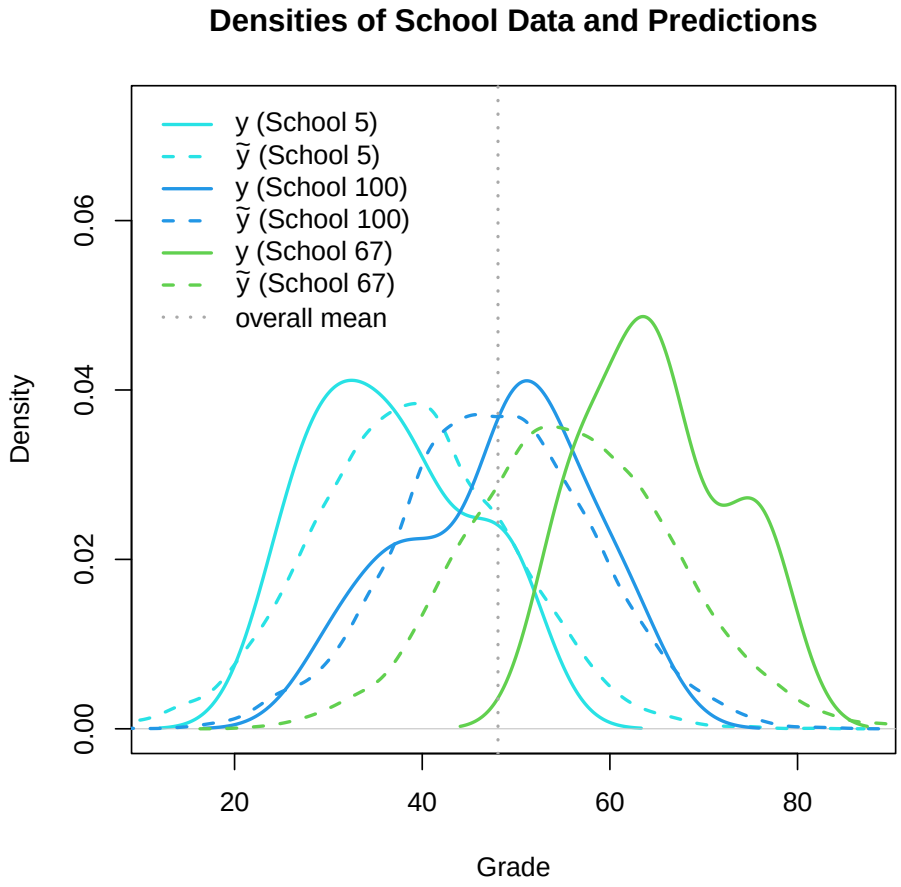
For prediction, we'll use the following prior values (Hoff p. 137):

σ_0^2 : 100 (within-school variance)
ν_0 : 1 (prior sample size)
τ_0^2 : 100 (between-school variance)
η_0 : 1 (prior sample size)
μ_0 : 50 (prior mean of school means)
γ_0^2 : 25 (prior variance of school means)

In the example below the observed test score data from three individual schools are compared with their predictions. The schools chosen were numbers 5, 67, and 100 from the study, which had the minimum average math score, maximum average math score, and closest to the overall average math score, respectively.

	school	average
max average	67	65.02
min average	5	36.58
grand mean	–	48.13
closest to overall	92	48.18

The plot below shows the density curves of the math scores for these three schools and their predictions. The observed data is shown using solid lines, and the predictive densities are displayed with dashed lines. The predictions are "pulled" toward the overall mean, which is indicated on the plot with the dotted gray vertical line.


Densities of School Data and Predictions
y (School 5)
ỹ (School 5)
y (School 100)
ỹ (School 100)
y (School 67)
ỹ (School 67)
overall mean
Density
0.00
0.02
0.04
0.06
20
40
60
80
Grade

4 Normal Regression

Starting with observations $Y_1, ..., Y_N$ and explanatory variables $\mathbf{x}_1, ..., \mathbf{x}_n$, with $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, ..., x_{i,p}\}$ where $Y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$ and $\epsilon_1, ..., \epsilon_n \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2)$, we have joint probability density

$$\begin{aligned} p\left(y_1, ...y_n | \mathbf{x}_1, ..., \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2\right) &= \prod_{i=1}^{n} p\left(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^T \mathbf{x}_i\right)^2\right\}. \end{aligned} \tag{5}$$

If we let $\mathbf{y} = (y_1, ..., y_n)^T$ and let $\mathbf{X}$ be the $n \times p$ matrix whose ith row is $\mathbf{x}_i$, then we can express this joint probability in terms of the Multivariate Normal distribution. The Normal regression model is

$$\{\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2\} \sim \text{Multivariate Normal}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}\right),$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} E\left[Y_1 | \boldsymbol{\beta}, \mathbf{x}_1\right] \\ \vdots \\ E\left[Y_n | \boldsymbol{\beta}, \mathbf{x}_n\right] \end{pmatrix}$$

The density (5) depends on $\boldsymbol{\beta}$ through the residuals $\left(y_i - \boldsymbol{\beta}^T \mathbf{x}_i\right)$. We compute the ordinary least squares estimates

$$\hat{\boldsymbol{\beta}}_{ols} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

and

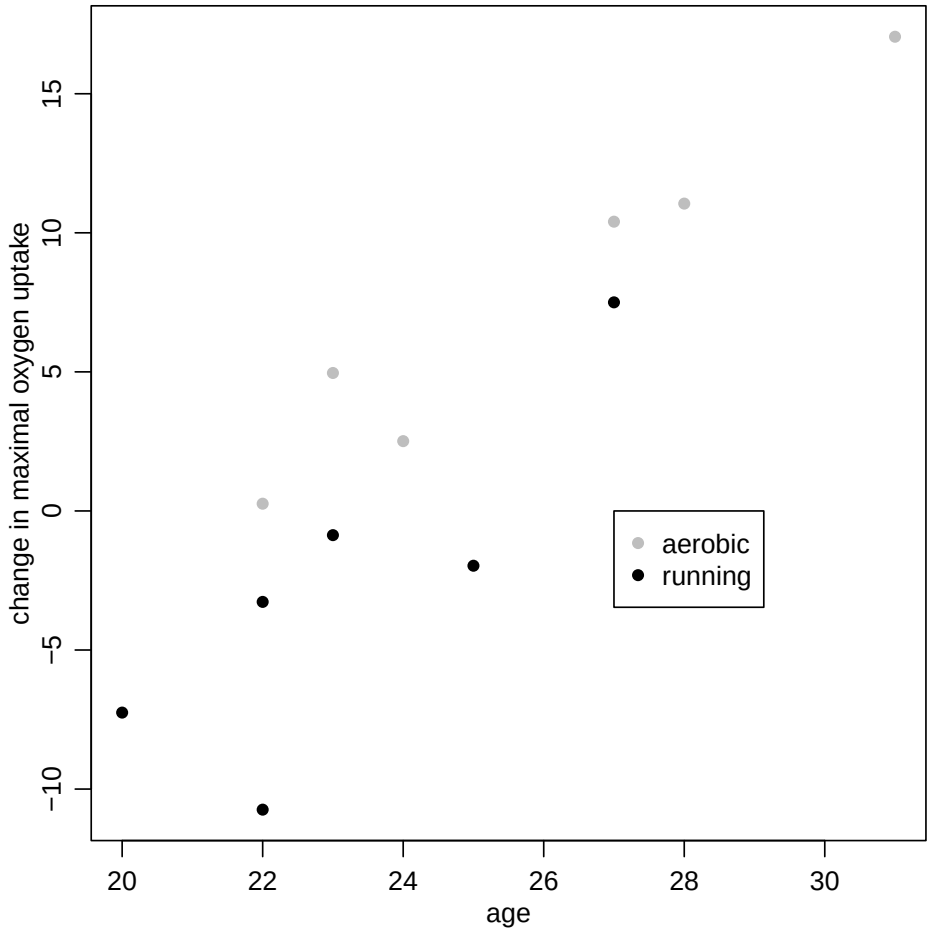
$$\hat{\sigma}^2_{ols} = \frac{SSR\left(\hat{\boldsymbol{\beta}}_{ols}\right)}{(n-p)} = \frac{\sum \left(y_i - \hat{\boldsymbol{\beta}}^T_{ols} x_i\right)^2}{(n-p)}.$$

4.1 Least Squares Estimation Example (Hoff p. 149ff.)

Here we reproduce the example provided by Hoff, which will be used to illustrate Bayesian prediction for a regression model.

Example: Oxygen uptake (from Kuehl (2000), Hoff p. 149ff)
Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimens on oxygen uptake. Six of the twelve men were randomly assigned to a 12-week flat-terrain running program, and the remaining six were assigned to a 12-week step aerobics program. The maximum oxygen uptake of each subject was measured (in liters per minute) while running on an inclined treadmill, both before and after the 12-week program. Of interest is how a subject's change in maximal

oxygen uptake may depend on which program they were assigned to. However, other factors, such as age, are expected to affect the change in maximal uptake as well. The results are shown here:



Hoff's regression model:

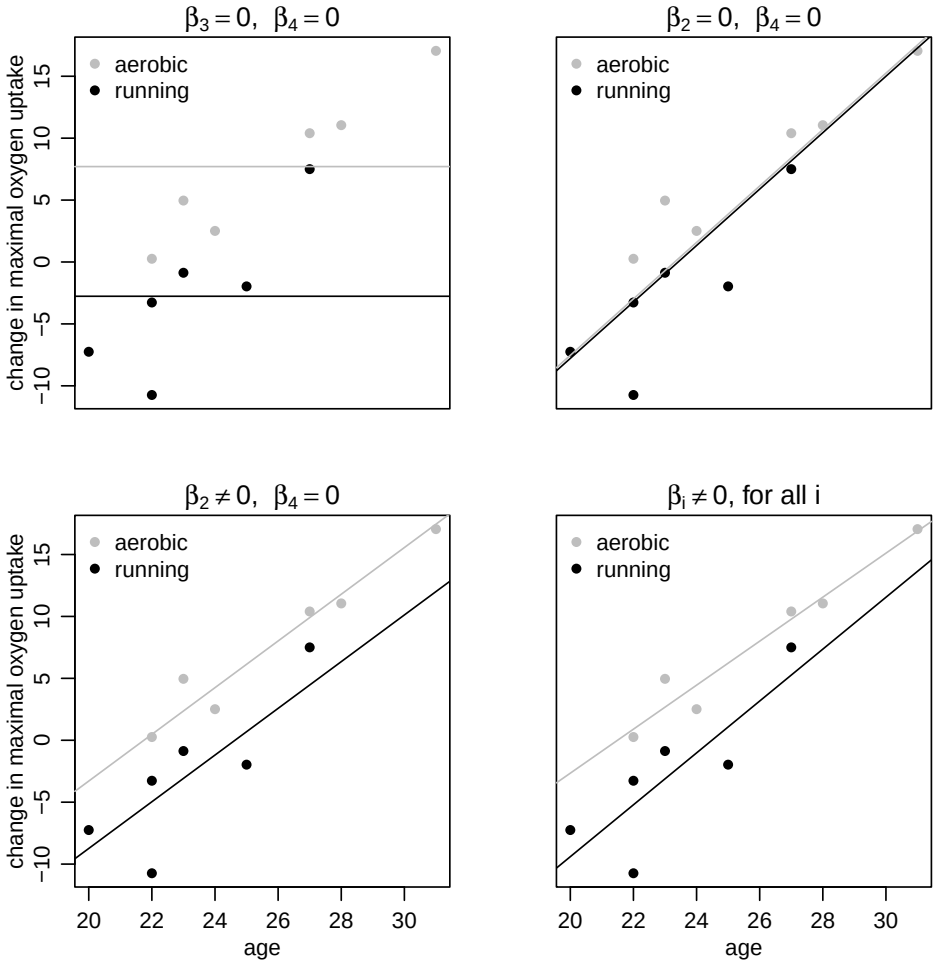
$$
\begin{aligned}
Y_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \text{ where} \\
x_{i,1} &= 1 \text{ for each subject } i \\
x_{i,2} &= 0 \text{ if subject } i \text{ is on the running program, 1 if on aerobic} \\
x_{i,3} &= \text{ age of subject } i \\
x_{i,4} &= x_{i,2} \times x_{i,3}
\end{aligned}
\tag{6}
$$

Under this model the conditional expectations of Y for the two different levels of $x_{i,2}$ are

$$
\begin{aligned}
E[Y|\mathbf{x}] &= \beta_1 + \beta_3 \times (age) \text{ if } x_{i,2} = 0 \text{ (running program), and} \\
E[Y|\mathbf{x}] &= (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times (age) \text{ if } x_{i,2} = 1 \text{ (aerobic program)}
\end{aligned}
$$

In other words, the model assumes that the relationship is linear in age for both exercise groups, with the difference in intercepts given by β_2 and the difference in slopes given by β_4. If we assumed that $\beta_2 = \beta_4 = 0$, then we would have identical lines for both groups.

If we assumed $\beta_2 \neq 0$ and $\beta_4 = 0$ then we would have a different line for each group but they would be parallel. Allowing all coefficients to be non-zero gives us two unrelated lines. Some different possibilities are depicted graphically below:



Let's find the least squares regression estimates for the model (6), and use the results to evaluate the differences between the two exercise groups. The ages of the 12 subjects, along with their observed changes in maximal oxygen uptake, are

$$\mathbf{x}_3 = (23, 22, 22, 25, 27, 20, 31, 23, 27, 28, 22, 24)$$
$$\mathbf{y} = (-0.87, -10.74, -3.27, -1.97, 7.50, -7.25, 17.05, 4.96, 10.40, 11.05, 0.26, 2.51),$$

with the first six elements of each vector corresponding to the subjects in the running group and the latter six corresponding to subjects in the aerobics group. After constructing the 12×4 matrix $\mathbf{X} = (\mathbf{x}_1\, \mathbf{x}_2\, \mathbf{x}_3\, \mathbf{x}_4)$, the matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ can be computed, from which we get $\boldsymbol{\beta}_{ols} = (-51.29, 13.11, 2.09, -0.32)^T$:

This means that the estimated linear relationship between uptake and age has an intercept and slope of -51.29 and 2.09 for the running group, and -51.29 + 13.11 = -38.18 and 2.09 - 0.32 = 1.77 for the aerobics group. These two lines are plotted in the fourth panel of Figure XX. We obtain unbiased estimate $\sigma^2 = SSR(\hat{\boldsymbol{\beta}}_{ols})/(n-p) = 8.54$, and use this to

compute the standard error of the components of $\hat{\boldsymbol{\beta}}_{ols}$, which are 12.25, 15.76, 0.53, and 0.65, respectively. Comparing the values of $\hat{\boldsymbol{\beta}}_{ols}$ to their standard errors suggests that the evidence for differences between the two exercise regimens is not very strong.

4.2 Bayesian Estimation for a Regression Model (Hoff p. 154ff)

4.2.1 Derivation

4.2.1.1 A semiconjugate prior distribution Hoff proposes a semiconjugate prior distribution for $\boldsymbol{\beta}$ and σ^2 to be used when there is information available about the parameters. The sampling density of the data is

$$p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2) \propto \exp\{-\frac{1}{2\sigma^2}\text{SSR}(\boldsymbol{\beta})\} = \exp\{-\frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}]\},$$

and for priors we choose $\boldsymbol{\beta} \sim$ Multivariate Normal$(\boldsymbol{\beta}_0, \Sigma_0)$ and $1/\sigma^2 = \gamma \sim$ Gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$.

Thus we have

$$\begin{aligned}
&p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) \\
&\quad \propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2) \times p(\boldsymbol{\beta}) \\
&\quad \propto \exp\{-\frac{1}{2}(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}/\sigma^2 + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}/\sigma^2) - \frac{1}{2}(-2\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta})\} \\
&\quad = \exp\{\boldsymbol{\beta}^T(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^T\mathbf{y}/\sigma^2) - \frac{1}{2}\boldsymbol{\beta}^T(\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)\boldsymbol{\beta}\}
\end{aligned}$$

and

$$\begin{aligned}
p(\gamma|\mathbf{y},\mathbf{X},\boldsymbol{\beta}) &\propto p(\gamma)p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\gamma) \\
&\propto \left[\gamma^{\nu_0/2-1}\exp(-\gamma \times \nu_0\sigma_0^2/2)\right] \times \left[\gamma^{n/2}\exp(-\gamma \times \text{SSR}(\boldsymbol{\beta})/2)\right] \\
&= \gamma^{(\nu_0+n)/2-1}\exp(-\gamma[\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta})]/2),
\end{aligned}$$

which we recognize as a gamma density, so that

$$\{\sigma^2|\mathbf{y},\mathbf{X},\boldsymbol{\beta}\} \sim \text{Inverse Gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta})]/2).$$

Constructing a Gibbs sampler to approximate the joint posterior distribution $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y},\mathbf{X})$ is then straightforward: given current values $\{\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}\}$, new values can be generated by

1. updating $\boldsymbol{\beta}$:
   (a) compute $\mathbf{V} = \text{Var}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^{2(s)}]$ and $\mathbf{m} = \text{E}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^{2(s)}]$
   (b) sample $\boldsymbol{\beta}^{(s+1)} \sim$ Multivariate Normal$(\mathbf{m},\mathbf{V})$
2. updating σ^2:
   (a) compute SSR$(\boldsymbol{\beta}^{(s+1)})$

(b) sample $\sigma^{2(s+1)} \sim$ Inverse Gamma$([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta}^{(s+1)})]/2)$.

To create a sample from the predictive distribution of responses: for each $s \in \{1, ..., S\}$, draw $\epsilon^{(s)} \sim \text{Normal}(0, \sigma^{2(s)})$. Then compute

$$y^{(s)} = \boldsymbol{\beta}^{(s)T}\mathbf{X} + \epsilon.$$

4.2.1.2 Default and weakly informative prior distributions In situations where prior information is unavailable or difficult to quantify, an alternative "default" class of prior distributions is given. Specification of the prior parameters $(\boldsymbol{\beta}_0, \Sigma_0)$ and (ν_0, σ_0^2) that represent actual prior information for a Bayesian analysis can be difficult. For a prior distribution that is not going to represent real prior information about the parameters, we choose one that is as minimally informative as possible. The resulting posterior distribution, then, will represent the posterior information of someone who began with little knowledge of the population being studied. Here we will employ Zellner's "g-prior" (Zellner, 1986). We choose $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\Sigma_0 = k(\mathbf{X}^T\mathbf{X})^{-1}, k = g\sigma^2, g > 0$, which satisfies a desired condition that the regression parameter estimation be invariant to changes in the scale of the regressors. With this, equations **??** and **??** reduce to

$$\text{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] = [\mathbf{X^T X}/(g\sigma^2) + \mathbf{X^T X}/\sigma^2]^{-1} = \frac{g}{g+1}\sigma^2(\mathbf{X^T X})^{-1} \tag{7}$$

$$\text{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] = [\mathbf{X^T X}/(g\sigma^2) + \mathbf{X^T X}/\sigma^2]^{-1}\mathbf{X^T y}/\sigma^2 = \frac{g}{g+1}\sigma^2(\mathbf{X^T X})^{-1}\mathbf{X^T y}. \tag{8}$$

Letting

$$\mathbf{V} = \frac{g}{g+1}\sigma^2(\mathbf{X^T X})^{-1} \text{ and } \mathbf{m} = \frac{g}{g+1}\sigma^2(\mathbf{X^T X})^{-1}\mathbf{X^T y}$$

we arrive at posteriors

$$\{\sigma^2|\mathbf{y}, \mathbf{X}\} \sim \text{Inverse Gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}_g]/2) \tag{9}$$

$$\{\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2\} \sim \text{Multivariate Normal}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{ols}, \frac{g}{g+1}\sigma^2[\mathbf{X^T X}]^{-1}\right). \tag{10}$$

Here $\text{SSR}_g = \mathbf{y^T y} - \mathbf{m^T V^{-1} m} = \mathbf{y^T}(\mathbf{I} - \frac{g}{g+1}\mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T})\mathbf{y}$.

Simple Monte Carlo approximation can be used to sample from the joint posterior density $p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ as follows. Here g is typically set to the number of prior observations. Then:

1. sample $\sigma^2 \sim$ Inverse Gamma$([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}_g]/2)$
2. sample $\boldsymbol{\beta} \sim$ Multivariate Normal $\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{ols}, \frac{g}{g+1}\sigma^2[\mathbf{X^T X}]^{-1}\right)$.

To create a sample from the predictive distribution of responses, draw $\epsilon \sim \text{Normal}(0, \sigma^2)$. Then for each triplet $(\beta, \sigma^2, \epsilon)$ we have

$$y = \boldsymbol{\beta}^T\mathbf{X} + \epsilon.$$

4.2.2 R Implementation (Normal Regression)

The R function

```
rpredNormReg(S=1,Xpred,X,y,beta0,Sigma0,nu0=1,s20=1,gprior = TRUE)
```

approximates the joint posterior density $p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ using one of the two methods described above, generates S triplets $(\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}, \epsilon^{(s)} \sim \text{Normal}(0, \sigma^{2(s)})$, and returns S predictions $y = X_{pred}\boldsymbol{\beta}^{(s)} + \epsilon^{(s)}$.

The function defaults to Zellner's location-invariant g-prior, in which case input values for `beta0, Sigma0, nu0`, and `s20` are ignored. If the user wants to employ Hoff's semi-conjugate prior as defined in section 4.2.1.1 above, all input variables must be specified, with gprior = FALSE.

4.2.3 Example

In the example below (Hoff data and code found here) to employ Hoff's semi-conjugate prior we use

$\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_{ols} = (-51.29, -51.29, -51.29, -51.29)$ (ordinary least squares estimator of $\boldsymbol{\beta}$)

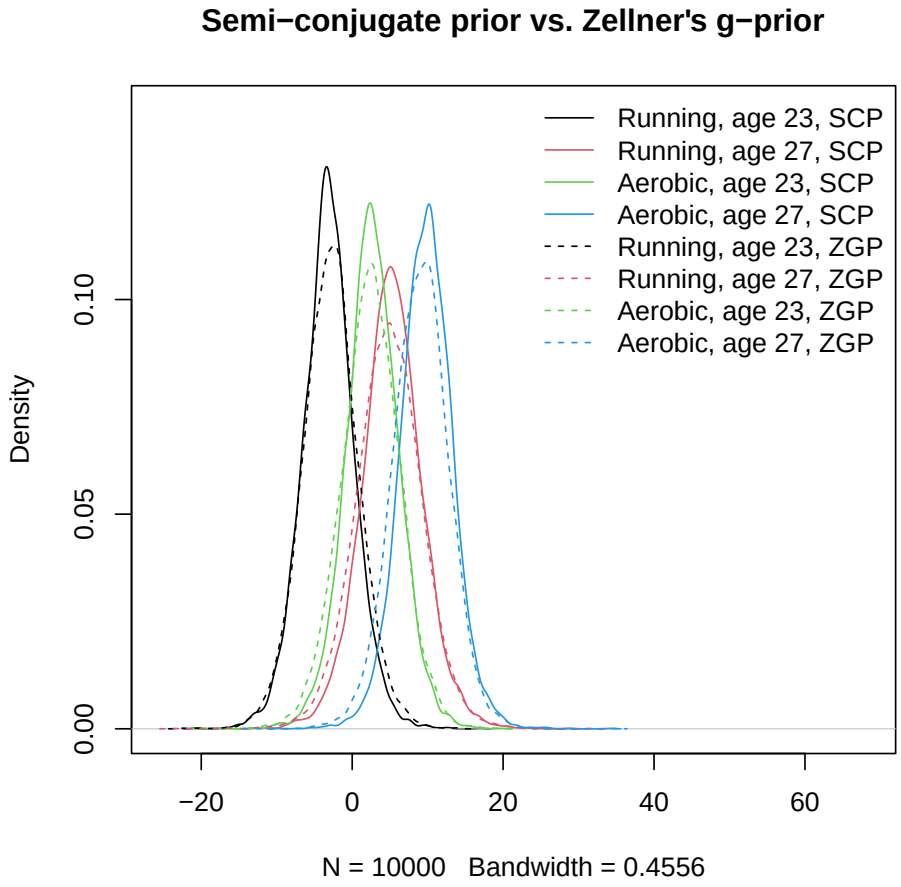
$\Sigma_0 = (X^T X)^{-1}\sigma^2 n = \begin{pmatrix} 1801.4 & -1801.4 & -77.02 & 77.02 \\ -1801.4 & 2981.28 & 77.02 & -122.03 \\ -77.02 & 77.02 & 3.32 & -3.32 \\ 77.02 & -122.03 & -3.32 & 5.07 \end{pmatrix}$ (sampling variance of $\hat{\boldsymbol{\beta}}_{ols}$)WHY TIM

$\nu_0 = 1$ (prior sample size)

$\sigma_0^2 = \dfrac{\sum e_i}{n-1} = 6.21$ (variance of the residuals)

$S = 5000$ (sample size for predictive distribution random draw)

Inspection of the plot below shows the predicted distributions using Zellner's g-prior shrink toward 0, and have greater variance than those predicted using Hoff's semi-conjugate prior.


Semi−conjugate prior vs. Zellner's g−prior
Running, age 23, SCP
Running, age 27, SCP
Aerobic, age 23, SCP
Aerobic, age 27, SCP
Running, age 23, ZGP
Running, age 27, ZGP
Aerobic, age 23, ZGP
Aerobic, age 27, ZGP
Density
0.00
0.05
0.10
−20
0
20
40
60
N = 10000 Bandwidth = 0.4556

5 Conclusion

6 Appendix

7 References