Rank Regression with Normal Residuals using the Gibbs Sampler

Stephen P Smith
email: hucklebird@aol.com, 2018

Abstract Yu (2000) described the use of the Gibbs sampler to estimate regression parameters where the information available in the form of depended variables is limited to rank information, and where the linear model applies to the underlying variation beneath the ranks. The approach uses an imputation step, which constitute nested draws from truncated normal distributions where the underlying variation is simulated as part of a broader Bayesian simulation. The method is general enough to treat rank information that represents ties or partial orderings.

1. Introduction

Rank regression is a set of robust regression techniques based on rank information. The rank information contained in a set of dependent variables, y_i (i=1, ⋯, N), remains unchanged under a monotonic increasing transformation of y_i. The transformation is unspecified, and otherwise unknown, yet an idealized transformation is declared such that $f(y_i)=u_i+e_i$ where u_i is the i-th mean effect that is subjected to linear modeling and e_i is the i-th residual that conforms to a standardized statistical distribution. Robustness owes its existence to the observation that f() is unspecified.

When the residual, e_i, is log-exponential the model reduces to the Cox regression model and comes with very tractable calculations for maximum likelihood estimation (Kalbfleisch and Prentice 1973). Smith and Hammond (1988) treated the case when the residual is log-gamma, but this innovation came with a more complex set of calculations. Rank regression has been a historically very hard problem because of the challenging numerical integration implied with maximum likelihood estimation for the case when the residuals come from the normal distribution. Simple approximations have been used by replacing $f(y_i)$ with first moments of the order statistics (scores) calculated where $u_i=0$, and subjecting the scores to linear regression (Fisher and Yates 1938). A weighted regression is possible using an additional weight matrix derived from second derivatives of the log-likelihood, again evaluated at $u_i=0$ (Pettitt 1982, 1983).

Doksum (1987) describes the use of monte carol simulation, or importance sampling, as a means to calculate the log-likelihood for the proposed purpose of maximum likelihood estimation. Pettitt (1987) reviews and evaluates this method, and others, in the estimation of regression parameters from rank data. Cuzick (1988) presented an alternative estimation method, described as a hybrid of previously published methods. Yu (2000) reviews the use of rank data in the applied areas of psychology and econometrics, and develops a completely Bayesian solution to the challenges of rank regression.

The Bayesian approach offers a very powerful way to perform rank regression because of the theoretical ease of replacing $f(y_i)$ with imputed values as part of a broader Bayesian simulation. The theoretical ease is well demonstrated with the treatment of censored data, missing observations, and imbedded variation that is part of a hierarchical model (Gelfand and Smith 1990, Li 1988, Tanner and Wong 1987). In particular, what is needed is a simulation step that starts with the current estimates of u_i, and simulates the various e_i where u_i+e_i preserves the rank order. Therefore, the apparent theoretical ease can be prevented from becoming a practical tool by the perceived complexity coming with ranks and multivariate simulation[1] that thwarts the needed step. Yu (2000) gets beyond this concern using a remarkably simple remedy: the Gibbs sampler itself provides a very easy an attractive solution to the multivariate simulation problem, by trivially turning the multivariate simulation into an iterated collection of nested univariate simulations.

The rank regression model is described in Section 2, including the possibility of strata that allows for the possibility of different transformations, or f(), for different strata. Section 3.1 describes the simulation of residuals using the Gibbs sampler. Section 3.2 describes the broader simulation steps where u_i is allowed to vary. The methods are illustrated in Section 4 using data involving the placing of Herford heifers in the 13th Junior Polled Hereford National of 1986 in Tulsa, Oklahoma. The conclusion followings in Section 5.

2. Model Specifications

The rank regression model is describe as follows for K strata,

$$f_k(y_{ik})=\mathbf{x}_{ik}^T\mathbf{b} + e_{ik}$$

where $f_k()$, k=1, ⋯, K, is a collection of K strictly monotonic transformations that are unspecified, y_{ik} is the i-th observation that is nested in the k-th strata (i=1, ⋯, N_k), $\mathbf{x}_{ik}$ is a column vector of covariates that relate the regression parameters **b** to the ik-th observation, and e_{ik} is the ik-th residual that is assumed to be a standardized normal (0,1) deviate. The set of residuals are assumed to be statistically independent.

Define the vectors $\mathbf{f}_k^T=[\, f_k(y_{1k}), f_k(y_{2k}),\ \cdots, f_k(y_{Nk\,k})]^T$, $\mathbf{e}_k^T=(e_{1k}, e_{2k},\ \cdots, e_{Nk\,k})^T$ and the matrix $\mathbf{X}_k^T=(\mathbf{x}_{1k}, \mathbf{x}_{2k},\ \cdots, \mathbf{x}_{Nk\,k})^T$. Therefore, for the k-th strata alone the model is presented below.

$$\mathbf{f}_k=\mathbf{X}_k\mathbf{b}+\mathbf{e}_k$$

---

[1] This simulation is simple when $u_i=0$, i=1, ⋯ , N. Just simulate N independent normal (0,1) deviates, and turn them into the needed order statistics by sorting them.

A mean for the k-th strata is not needed as part of this model because it is absorbed into the unspecified transformation. When $\mathbf{X}_k$ contain continuous covariates a mean effect can be induced as the average effect given by $\mathbf{1}^T\mathbf{X}_k\mathbf{b}/N_k$, where $\mathbf{1}$ is a column vector of ones, and might be the source of rounding error and interfere[2] with the estimation of $\mathbf{b}$. This potential interference is removed by centering all the continuous covariates within strata, such that $\mathbf{X}_k^T\mathbf{1}=\mathbf{0}$, a column vector of zeros.

Define $\mathbf{f}^T=[\mathbf{f}_1^T, \mathbf{f}_2^T, \cdots, \mathbf{f}_K^T]$, $\mathbf{e}^T=(\mathbf{e}_1^T, \mathbf{e}_2^T, \cdots, \mathbf{e}_k^T)$ and $\mathbf{X}^T=(\mathbf{X}_1^T, \mathbf{X}_2^T, \cdots, \mathbf{X}_K^T)$. Then the complete model is given again below, but now in matrix notation.

$$\mathbf{f}=\mathbf{X}\mathbf{b}+\mathbf{e} \qquad (1)$$

3. Gibbs Sampler

3.1 Imputing Underlying Variation from Ranks

Without loss of generality assume that $y_{1k} < y_{2k} < \cdots < y_{Nk\,k}$, representing the rank information for the k-th strata. Restrict attention to the transformed scale. Then with $u_i=\mathbf{x}_{ik}^T\mathbf{b}$ the rank information becomes $u_1+e_1 < u_2+e_2 < \cdots < u_N+e_N.$, where the index k is dropped to simplify notation. The nested Gibbs simulation is outlined below, essentially Yu's (2000) method.

1. Simulate e_1 with u_1 and u_2+e_2 fixed,
2. Simulate e_2 with u_2, u_1+e_1 and u_3+e_3 fixed,
3. More generally, simulate e_i with u_i, $u_{i-1}+e_{i-1}$ and $u_{i+1}+e_{i+1}$ fixed, for i=3, 4, ⋯, N-1,
4. Finally, simulate e_N with u_N and $u_{N-1}+e_{N-1}$ fixed.

5. Repeating this nested simulation as many times as needed to exceed the burn-in.

Steps 1 through 4 involve simulating from a truncated normal (0,1) deviate. Define the cumulative distribution for the standard normal distribution as follows,

$$\Phi(T) = \int_{-\infty}^{T} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\,dx$$

and define the inverse function $\Phi^{-1}(P)=T$ such that $\Phi(T)=P$. The computer function "erfcc", for the complementary error function (Press et al., 1992), can be used in the

[2] Note that $\mathbf{1}^T\mathbf{e}_k/N_k$ tends to zero, $\mathbf{1}^T\mathbf{f}_k/N_k$ tends to the strata mean. Therefore, to avoid interference follow Section 3.1 exactly. When $\mathbf{e}_k$ is completely imputed, the imputation of $\mathbf{f}_k$ is given by $\mathbf{f}_k=\mathbf{X}_k\mathbf{b}+\mathbf{e}_k$.

approximation of Φ(T). Brophy (1985) shows that the Odeh and Evans approximation is a reliable calculation that leads to $\Phi^{-1}(P)$. Any of the first 4 steps in the nested simulation can be represented by the following calculations.

$p \leftarrow \Phi(u_{i-1}+e_{i-1} - u_i)$
Select c from a uniform [0,1] distribution
$q \leftarrow p + c \times \{\Phi(u_{i+1}+e_{i+1} - u_i)-p\}$
$e_i \leftarrow \Phi^{-1}(q)$

For step 1, note that p=0 by convention. Likewise, for step 4 use $\Phi(u_{i+1}+e_{i+1} - u_i)=1$. The results are otherwise immediate for steps 3 and 4.

With $e_1, e_2, \cdots, e_N$ imputed, they are added back to $u_1, u_2, \cdots, u_N$ to reconstitute $\mathbf{f}_k$, with the index k now reintroduced. The vectors $\mathbf{f}_k$ are collected into **f** as the imputation moves from strata to strata.

The nested simulation is very easy to adapt for tied rank information that is represented by groups: $g_1 < g_2 < \cdots < g_N$, where there can be several observations representing ties because they are indistinguishable in their membership in one of the groups. The maximum value (in the form u+e) in group g_{i-1} sets the lower limit for group g_i, whereas the minimum value (in the form u+e) in group g_{i+1} sets the upper limit for group g_i. Having found the lower and upper limits dynamically, the simulation follows the same calculations.

Likewise, the nested simulation is easy to adapt to partial orderings that contain less than the full rank information. Once a lower and upper limit is identified for any observation that is up for imputation, the calculations follow the same approach.

3.2 The Combined Simulation

The flat non-informative prior is used for **b**, given that these effect impact only on location. This completes the Bayesian specifications beyond (1).

Conditional on **f** in (1), and with **e** a vector of independent normal (0,1) deviates, the posterior distribution of **b** is a multivariate normal distribution (Tanner 1993, page 12):

$$\mathbf{b}\,|\,\mathbf{f} \sim MVN\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{f},\ \ (\mathbf{X}^T\mathbf{X})^{-1}\right)$$

There is no need to consider a residual variance because under the unspecified transformation the residuals are standard normal. Random effects can be part of the linear model (1), each set coming with its unknown variance component, and these would need to be included as part of the simulation. However, in the present case the

matrix **X** is restricted to continuous covariates.

Let the lower triangular matrix **L** be the Cholesky decomposition of $\mathbf{X}^T\mathbf{X}$, where $\mathbf{L}^T\mathbf{L}=\mathbf{X}^T\mathbf{X}$. The full Gibbs sampler is provided below.

1. Find suitable starting values for **b** and **e**. Setting **b**=**0** and **e** to the percentiles representing order statistics, provide good starting values.

2. Iterate the following steps beyond the burn-in to provide the sample of **b**.

A. Holding **b** fixed, impute the vector **f** using the nested simulation described in Section 3.1.

B. Holding **f** fixed, simulate the vector **b** using the following calculations.

- Solve the vector **w** in the linear system, $\mathbf{Lw}=\mathbf{X}^T\mathbf{f}$, using forward substitution.
- Simulate a vector **z**, where its i-th element is z_i being an independent normal (0,1) deviate, then set $\mathbf{w} \leftarrow \mathbf{w}+\mathbf{z}$.
- Solve **b** in the linear system, $\mathbf{L}^T\mathbf{b}=\mathbf{w}$, using backward substitution.

4. Example

The rank data presented by Smith and Hammond (1988) is used again to illustrate the new method. The data represents three strata, showing the show-ring placing (or ranks) of 53 Hereford heifers along with the height (inches) and weight (pounds) of each heifer. The same linear model will be used again, to allow comparisons:

$$f_{s(i)}(y_i)=b_1 h_i+b_2 h_i^2+b_3 w_i + e_i$$

where $f_s()$ is the unspecified transformation that is defined for the s-th strata, y_i is the underlying and unobserved variation representing the placing[3] of the i-th heifer, h_i and w_i is the height and weight of the i-th heifer, and e_i is the i-th residual that is treated as normal (0,1) rather than log-gamma as used in Smith and Hammond (1988). The parameters b_1, b_2 and b_3 are to be simulated using the Gibbs sampler.

[3] Coming in 1st corresponds to a small value of y in the present analysis, but it is equally valid to reverse this given that the normal distribution is symmetric.

The results for 1000 iterations of the Gibbs sampler are presented in Table 1, showing a side by side comparison with the results of Smith and Hammond (1988). The summary statistics were calculated separately for iterations 1 to 200 (case 1), and 201 to 1000 (case 2), to evaluate any effect of burn-in. The mean and median estimates for the three parameters, b_1, b_2 and b_3, were within the statistical precision when comparing case 1 and case 2. The burn-in effect was small. Nevertheless, the case 2 estimates were in very close agreement with the maximum likelihood estimates reported by Smith and Hammond for the log-gamma distribution that used the shape parameter η=100. This confirmed the prior work of Smith and Hammond and verified that the Gibbs sampler worked quite well in the present investigation.

Table 1. Summary results for 1000 iterations of the Gibbs sampler, and the corresponding estimates from Smith and Hammond (1988).

Notes	b_1	b_2	b_3
Gibbs Sampler [A]			
Iterates 1-200	-20.09(-20.52)±6.322	0.2037(0.2076)±0.0664	-0.0103(-0.0106)±0.0039
Iterates 201-1000	-21.97(-21.69)±7.737	0.2227(0.2206)±0.0805	-0.0107(-0.0107)±0.0043
Smith and Hammond's log-gamma estimates [B]			
η=100 [C]	-21.94±7.550	0.2225±0.0787	-0.0106±0.0042

A - Results in the form mean(median)±prediction error.
B - Results in the form m.l.e.±standard error.
C - The shape parameter η=100 was the largest considered by Smith and Hammond, where the log-gamma distribution is well approximated by the normal distribution.

One of the advantages of Bayesian simulation is that given the validity of the model, and given enough iterations, the results are exact. There is no need to consider alternative or approximate estimators that are to be judged on bias and statistical efficiency. It is quite reasonable to produce the statistical distributions for the three parameters, b_1, b_2 and b_3, as an alternative to point estimation. These three distributions that were derived from case 2 iterations are displayed in Figure 1, 2 and 3. They were formed using kernel density estimation (Simonoff 1996, Chapter 3) while choosing a smoothness parameter (or bandwidth) smaller than the innate variation to enable just seeing a little of the statistical noise.

Figure 1. Frequency plot of b_1 from Gibbs Sampler.

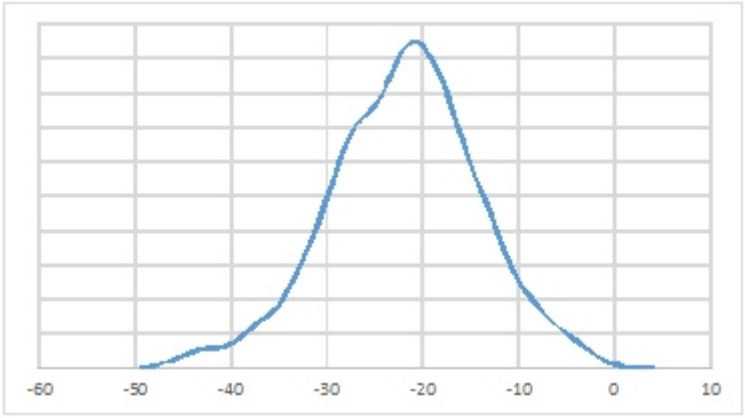


Figure 2. Frequency plot of b_2 from Gibbs Sampler.

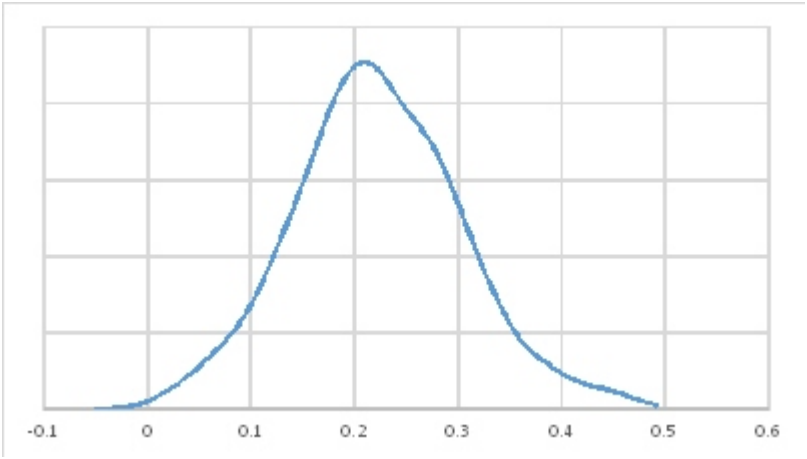
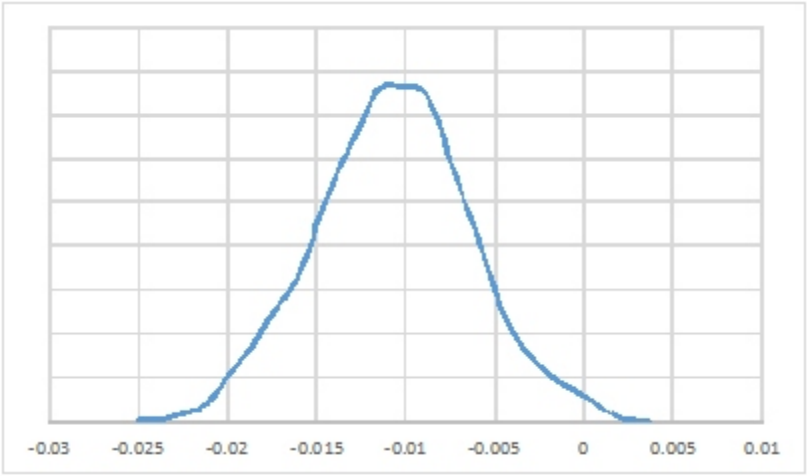


Figure 3. Frequency plot of b_3 from Gibbs Sampler.



5. Conclusion

The Gibbs sampler was successfully used to simulate the underlying variation of rank data, and this permitted a full Bayesian simulation to conduct a rank regression using continuous covariates to predict the underlying variation. The rank preserving simulation involves nested draws from a truncated normal distribution, and is easily appended to a full Gibbs sampler that involves the entire range of possibilities that may come with any nominated statistical model. The approach is general enough for rank data that shows ties, or when the rank information is replaced by a partial ordering.

Yu's (2000) approach proved to be powerful and fast. The amount of iterations required to permit burn-in did not seem to be excessive in the example studied.

References

Brophy, A.L., 1985, Approximation of the inverse normal distribution function, *Behavior Research Methods, Instruments & Computers*, 17 (3), 415-417.

Cuzick, J, 1988, Rank regression, *The Annals of Statistics*, 16 (4), 1369-1389.

Doksum, K.A., 1987, An extension of partial likelihood methods for proportional hazard models to general transformation models, *The Annals of Statistics*, 15 (1), 325-345.

Fisher, R.A., and F. Yates, 1938, *Statistical Tables for Biological, Agricultural and Medical Research*, Edinburgh and London, Oliver and Boyd.

Gelfand, A.E., and A.F.M. Smith, 1990, Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85 (410), 398-409.

Kalbfleisch, J.D., and R.L. Prentice, 1973, Marginal likelihood based on Cox's regression and life model, Biometrika, 60 (2), 267-278.

Li, K.H., 1988, Imputation using Markov chains, *Journal of Statistical Computation and Simulation*, 30 (1), 57-79.

Pettitt, A.N., 1982, Inferences for the linear model using a likelihood based on ranks, *The Journal of the Royal Statistical Society*, Series B, 44 (2), 234-243.

Pettitt, A.N., 1983, Approximate method using ranks for regression with censored data, *Biometrika*, 70 (1), 121-132.

Pettitt, A.N., 1987, Estimates for a regression parameter using ranks, *The Journal of the Royal Statistical Society*, Series B, 49 (1), 58-67.

Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, 1992, *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd Edition, Cambridge , Cambridge University Press.

Simonoff, J.S, 1996, *Smoothing Methods in Statistics*, New York, Springer.

Smith, S.P., and K. Hammond, 1988, Rank regression with log-gamma residuals, *Biometrika*, 75 (4), 741-751.

Tanner, M.A., 1993, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd Edition, New York, Springer-Verlag.

Tanner, MA, and W.H. Wong, 1987, The calculations of posterior distribution by data augmentation, *Journal of the American Statistical Association*, 82 (398), 582-540.

Yu, P.L.H, 2000, Bayesian Analysis of Order-statistics for Ranking Data, *Psychometrika*, 65 (3), 281-299.