WORKING PAPER

9/2015

Bayesian Inference in Regression Models with Ordinal Explanatory Variables

Sune Karlsson and Asrat Temesgen
Statistics

ISSN 1403-0586

http://www.oru.se/Institutioner/Handelshogskolan-vid-Orebro-universitet/Forskning/Publikationer/Working-papers/
Örebro University School of Business
701 82 Örebro
SWEDEN

Bayesian Inference in Regression Models with Ordinal Explanatory Variables[1]

September 11, 2015

Sune Karlsson[2]
Örebro University School of Business

Asrat Temesgen
Örebro University School of Business

Abstract

This paper considers Bayesian inference procedures for regression models with ordinally observed explanatory variables. Taking advantage of a latent variable interpretation of the ordinally observed variable we develop an efficient Bayesian inference procedure that estimates the regression model of interest jointly with an auxiliary ordered probit model for the unobserved latent variable. The properties of the inference procedure and associated MCMC algorithm are assessed using simulated data. We illustrate our approach in an investigation of gender based wage discrimination in the Swedish labor market and find evidence of wage discrimination.

Keywords: Markov Chain Monte Carlo; latent variables; ordered probit; wage discrimination

JEL-codes: C11, C25, C35, J31

[1] We are grateful for useful comments and suggestions from Neby Bekele.
[2] Sune.Karlsson@oru.se

1. Introduction

The use of ordinal explanatory variables in regression models is a complicated but somewhat neglected issue. The basic problem stems from the fact that the relevant information in an ordinal variable is the ordering or rank of different observations, the values or labels assigned to different categories is irrelevant as long as the order is preserved. Directly including an ordinal explanatory variable in the model will consequently lead to results that to a large extent is determined by the particular values assigned to the categories.

The simplest and probably most common solution is to replace the ordinal variable with dummy variables for the categories. This has two disadvantages; it increases the number of parameters to estimate and is likely to introduce a measurement error. The measurement error comes about when the categories of the ordinal variable do not correspond to well defined and homogenous states, for example the order customers arrive in a queue. In our experience ordinal data is, for the most part, not of this nature. Instead the data represents loosely defined categories such as “disagree”, “neutral” and “agree”, well – but not uniquely – defined categories such as educational attainment categorized as “completed primary education”, “completed secondary education”, “university degree” etc. or it is only recorded which interval, e.g. income bracket, a variable falls into. In these cases there is an underlying latent variable, the effect of which is the real object of interest and it is this situation that we focus on.

That is, we assume that there is a well-defined latent variable z_i^* that one wishes to control for or estimate the effect of in the analysis but that there is only an ordinal indicator z_i,

$$z_i = k \text{ if } \mu_{k-1} < z_i^* \leq \mu_k, k = 1, \dots, m,$$

available and that the model of interest includes the latent variable as a linear regressor, i.e. $y_i = \mathbf{x}_i'\boldsymbol{\beta} + z_i^*\gamma + \varepsilon_i$. Building on this assumption we propose an efficient Bayesian procedure for inference in linear regression models with ordinally observed explanatory variables.

Our model setup is closely related to Breslaw and McIntosh (1998) who develop a simulation based procedure based on the assumption that the latent variable z_i^* is normal conditional on the variables $\boldsymbol{w_i}$ (which may contain $\boldsymbol{x}_i$). This assumption leads to an ordered probit model for z_i and the distribution of z_i^* conditional on $\boldsymbol{w}_i$ and z_i is obtained as a truncated normal. Breslaw and McIntosh use a draw from this conditional distribution as a proxy for z_i^* and a second draw as an instrument to overcome the measurement error problem. We depart from this in that our approach is Bayesian and, crucially, that we estimate the equations for y_i and z_i^* jointly.

Earlier related literature includes Hsiao and Mountain (1985) who consider the problem of estimating the equation for y_i in a context where the cut points, μ_k, for discretizing the latent variable are known. They suggested using an estimate of the expectation of, $E(z_i^*|z_i, \boldsymbol{x}_i)$, conditional on z_i and the fully observed explanatory variables, $\boldsymbol{x}_i$, as a proxy for z_i^*. This procedure will yield consistent estimates if the conditional expectation is estimated consistently.

Estimating the expectation conditional on both z_i and $\boldsymbol{x_i}$ is complicated and Hsiao and Mountain also considered using the expectation conditional on z_i only, $E(z_i^*|z_i)$. This is easier to estimate and will yield consistent estimates if $z_i^* - E(z_i^*|z_i)$ is uncorrelated with $\boldsymbol{x_i}$. This is unlikely to be the case but it turns out that $E(z_i^*|z_i)$ is a valid instrument for z_i^* and they suggested an instrumental variable estimator based on this. Terza (1987) generalized Hsiao and Mountain by allowing for unknown cut points and proposed an estimator of $E(z_i^*|z_i)$ based on the assumption that the marginal distribution of z_i^* is normal. Using this as a proxy yields consistent estimates when $z_i^* - E(z_i^*|z_i)$ is uncorrelated with $\boldsymbol{x}_i$.

The outline of the paper is as follows. Section 2 introduces the model and the prior distribution and gives details of the posterior distribution of the parameters. Section 3 suggests and evaluates a Markov chain Monte Carlo algorithm for exploring the posterior distribution and in section 4 we estimate a wage equation where the true educational attainment is unobserved. Finally, section 5 concludes.

2. Regression with ordinally observed explanatory variables

The basic model we consider is a linear regression model where one of the explanatory variables is only observed in the form of an ordinal indicator representing an underlying trait of fundamental interest. This could, for example, be "credit worthiness", "job performance", "health status", "amount of schooling" – as in our application – or any difficult to measure characteristic that has been rated on an ordinal scale. Ideally we would like to estimate the model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + z_i^*\gamma + \varepsilon_i \qquad (1)$$

where $\mathbf{x}_i$ denotes fully observed explanatory variables and z_i^* is a latent variable representing the true value of the unobserved trait. While z_i^* is unobserved we have at our disposal an ordinal indicator with *m* levels given by

$$z_i = k \text{ if } \mu_{k-1} < z_i^* \leq \mu_k, k = 1, \dots, m$$

which provides partial information about z_i^* for some unknown cut points μ_k. To complete the model we follow Breslaw and McIntosh (1998) and make the additional assumption that z_i^* is well described by a second linear regression

$$z_i^* = \mathbf{w}_i'\boldsymbol{\delta} + \eta_i. \qquad (2)$$

Assuming that η_i is normal Breslaw and McIntosh proposes a two stage estimation strategy where (2) is estimated as an ordered probit model and use the estimated model to simulate values of z_i^* that are then used to estimate (1) in the second stage using an IV-type estimator. We depart from Breslaw and McIntosh by estimating (1) and (2) jointly and thus making efficient use of all the information in the data. For convenience we retain the assumption that $\eta_i \sim N(0,1)$ while also assuming that $\varepsilon_i \sim N(0,1/\tau)$. These assumptions are far from necessary but simplify the analysis and the proposed MCMC scheme for simulating from the posterior distribution.

Augmenting the data by explicitly introducing the latent variable z_i^* in the analysis simplifies the analysis considerably and is key to our approach as it leads to a straightforward MCMC procedure for exploring the posterior distribution.

2.1 Prior distributions

We specify a simple prior structure based on natural conjugate priors. That is, we use the priors

$$\boldsymbol{\beta} \sim \mathrm{MVN}(\underline{\boldsymbol{\beta}}, \underline{\boldsymbol{\Sigma}}_{\beta})$$

$$\gamma \sim \mathrm{N}(\underline{\gamma}, \underline{\sigma}_{\gamma}^{2})$$

$$\boldsymbol{\delta} \sim \mathrm{MVN}(\underline{\boldsymbol{\delta}}, \underline{\boldsymbol{\Sigma}}_{\delta})$$

$$\tau \sim \mathrm{Gamma}\ (a, b)$$

for the unknown parameters in (1) and (2) together with a uniform prior on the cut points $\mu_k$subject to the restriction

$$\mu_1 = 0 < \mu_2 < \cdots < \mu_{m-1} < \mu_m = \infty$$

where μ_1 is set to zero for identification purposes and the lower limit of the first class is $\mu_0 = -\infty$.

2.2 Posterior distribution

The joint posterior distribution is intractable but the full conditional posteriors follow from standard results for the linear regression model if we also condition on the latent variable z^*. The full conditional posteriors for the regression parameters are normal and the conditional posterior for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\gamma, \tau, \boldsymbol{\delta}, \boldsymbol{z}^*, \boldsymbol{y} \sim \mathrm{MVN}(\,\overline{\boldsymbol{\beta}}, \overline{\boldsymbol{\Sigma}}_{\beta})$$

with $\overline{\boldsymbol{\Sigma}}_{\beta} = \left(\underline{\boldsymbol{\Sigma}}_{\beta}^{-1}\underline{\boldsymbol{\beta}} + \tau \boldsymbol{X}'\tilde{\boldsymbol{y}}\right)^{-1}$, $\overline{\boldsymbol{\beta}} = \overline{\boldsymbol{\Sigma}}_{\beta}\left(\underline{\boldsymbol{\Sigma}}_{\beta}^{-1}\underline{\boldsymbol{\beta}} + \tau \boldsymbol{X}'\tilde{\boldsymbol{y}}\right)$, and $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \gamma \boldsymbol{z}^*$. For γ we have

$$\gamma|\boldsymbol{\beta}, \tau, \boldsymbol{\delta}, \boldsymbol{z}^*, \boldsymbol{y} \sim N\left(\overline{\gamma}, \overline{\sigma}_{\gamma}^{2}\right)$$

with $\overline{\sigma}_{\gamma}^{2} = \left(\underline{\sigma}_{\gamma}^{-2} + \tau \boldsymbol{z}^{*\prime}\boldsymbol{z}^*\right)^{-1}$, $\overline{\gamma} = \overline{\sigma}_{\gamma}^{2}\left(\underline{\sigma}_{\gamma}^{-2}\underline{\gamma} + \tau \boldsymbol{z}^{*\prime}\breve{\boldsymbol{y}}\right)$ and $\breve{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. We will also make use of the joint conditional posterior for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma)'$. Let $\underline{\boldsymbol{\theta}} = \left(\underline{\boldsymbol{\beta}}', \underline{\gamma}\right)'$, $\underline{\boldsymbol{\Sigma}}_{\theta} = \begin{pmatrix} \underline{\boldsymbol{\Sigma}}_{\beta} & \mathbf{0} \\ \mathbf{0} & \underline{\sigma}_{\gamma}^{2} \end{pmatrix}$ and $\boldsymbol{H} = (\boldsymbol{X}, \boldsymbol{z}^*)$ we then have

$$\boldsymbol{\theta}|\tau, \boldsymbol{\delta}, \boldsymbol{z}^*, \boldsymbol{y} \sim N\left(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\Sigma}}_{\theta}\right) \tag{3}$$

for $\overline{\boldsymbol{\Sigma}}_{\theta} = \left(\underline{\boldsymbol{\Sigma}}_{\theta}^{-1} + \tau \boldsymbol{H}'\boldsymbol{H}\right)^{-1}$ and $\overline{\boldsymbol{\theta}} = \overline{\boldsymbol{\Sigma}}_{\theta}\left(\underline{\boldsymbol{\Sigma}}_{\theta}^{-1}\underline{\boldsymbol{\theta}} + \tau \boldsymbol{H}'\boldsymbol{y}\right)$.

The full conditional posterior for τ is Gamma,

$$\tau|\boldsymbol{\beta}, \gamma, \boldsymbol{z}^*, \boldsymbol{y} \sim \mathrm{Gamma}\left(\frac{n}{2} + a, \frac{\sum_1^n \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} - z_i^*\gamma\right)^2}{2} + b\right). \tag{4}$$

The full conditional posterior for $\boldsymbol{\delta}$ is given by

$$\boldsymbol{\delta}|z^* \sim \text{MVN}(\overline{\boldsymbol{\delta}}, \overline{\boldsymbol{\Sigma}}_\delta) \tag{5}$$

for $\overline{\boldsymbol{\delta}} = \overline{\boldsymbol{\Sigma}}_\delta \left(\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}^{-1}\underline{\boldsymbol{\delta}} + \mathbf{W}'\boldsymbol{z}^*\right)$ and $\overline{\boldsymbol{\Sigma}}_\delta = \left(\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}^{-1} + \boldsymbol{W}'\boldsymbol{W}\right)^{-1}$.

There is obviously a high degree of dependence between the distribution of the cut points, μ_k, and the latent variable z_i^* and we will consider these distributions in some detail. The full conditional posterior for the cut points, μ_k, follows from Albert and Chib (1993) as the uniform distributions

$$\mu_k|\mu_{k-1}, \mu_{k+1}, \boldsymbol{z}^* \sim U\left(\max\left(\max_{z_i=k} z_i^*, \mu_{k-1}\right), \min\left(\min_{z_i=k+1} z_i^*, \mu_{k+1}\right)\right). \tag{6}$$

Turning to the distribution of z_i^* conditional on the parameters in the model, the joint likelihood for (1) and (2) together with the restrictions imposed by the cut points gives the kernel of the conditional distribution as

$$\begin{aligned} &p(z_i^*|\boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \tau, \boldsymbol{\mu}, y_i, z_i) \\ &\quad \propto \exp\left[-\frac{\tau}{2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} - z_i^*\gamma)^2\right] \exp\left[-\frac{1}{2}\left(z_i^* - \boldsymbol{w}_i'\boldsymbol{\delta}\right)^2\right] \mathrm{I}\left(\mu_{z_i-1} \le z_i^* \le \mu_{z_i}\right). \end{aligned} \tag{7}$$

Completing the square for z_i^* yields

$$p(z_i^*|\boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \tau, \boldsymbol{\mu}, y_{\boldsymbol{i}}, z_{\boldsymbol{i}}) \propto \exp\left[-\frac{\tau_z}{2}(z_i^* - m_i)^2\right] \mathrm{I}\left(\mu_{z_i-1} \le z_i^* \le \mu_{z_i}\right)$$

for $m_i = \tau_z^{-1}\left[\tau\gamma\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) + \boldsymbol{w}_i'\boldsymbol{\delta}\right]$ and $\tau_z = \tau\gamma^2 + 1$. That is, a truncated normal distribution

$$z_i^*|\boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \tau, \boldsymbol{\mu}, y_i, z_i \sim N(m_i, \tau_z^{-1}) \mathrm{I}\left(\mu_{z_i-1} \le z_i^* \le \mu_{z_i}\right). \tag{8}$$

With a uniform prior on the cut points, the form of the joint posterior of $\boldsymbol{z}^*$ and $\mu_2, \dots, \mu_{m-1}$ follows from (7) and we have

$$p(\boldsymbol{z}^*, \mu_2, \dots, \mu_{m-1}|\boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \tau, \boldsymbol{y}, \boldsymbol{z}) \propto \prod_{i=1}^n \exp\left[-\frac{\tau_z}{2}(z_i^* - m_i)^2\right] \mathrm{I}\left(\mu_{z_i-1} \le z_i^* \le \mu_{z_i}\right).$$

Integrating out the latent variable $\boldsymbol{z}^*$ yields the joint conditional posterior of the cut points as

$$p(\mu_2, \dots, \mu_{m-1}|\boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \tau, \boldsymbol{y}, \boldsymbol{z}) \propto \prod_{i=1}^n \left\{\Phi\left[\sqrt{\tau_z}\left(\mu_{z_i} - m_i\right)\right] - \Phi\left[\sqrt{\tau_z}\left(\mu_{z_{i-1}} - m_i\right)\right]\right\} \tag{9}$$

where Φ is the standard normal CDF.

3. Markov chain Monte Carlo

Our strategy for constructing a posterior sampler is to combine a standard Gibbs sampler for the linear regression (1) with a sampler for the ordinal probit model (2). For the linear regression part we can sample from the full conditionals (3) and (4). Albert and Chib (1993) suggested a straightforward Gibbs sampler for the ordinal probit model sampling in turn from the full conditional posteriors of the latent variable, the regression parameters and the cut points. In our case this is the distributions (8), (5) and (6) where (8) differs from Albert and Chib and reflects the fact that z_i^* appears in both the linear regression and the ordinal probit.

As pointed out by Cowles (1996) the sampler of Albert and Chib tends to converge slowly when the sample size is large and there are many observations in each category of the ordinal variable. Cowles attributed this to the high degree of dependence between the cut points and the latent variable and proposed the use of a Metropolis-Hastings step that updates μ_k and z_i^* jointly. A draw from the joint conditional posterior of $\mu_2, \dots, \mu_{m-1}$ and $\boldsymbol{z}^*$ can be obtained by first drawing $\mu_2, \dots, \mu_{m-1}$ from (9) and then generating $z_i^*, i = 1, \dots, n$ from (8). Cowles suggests proposing new values for the cut points from truncated normal distributions centered at the current cut points, generate proposals for z_i^* from (8) conditional on the proposed cut points and accept or reject the proposals based on the standard M-H acceptance ratio. Since (8) amounts to a Gibbs step the acceptance ratio does not depend on the current or proposed values of $\boldsymbol{z}^*$ which simplifies the calculations.

Algorithm 1: Metropolis-Hastings within Gibbs sampler

Choose starting values $\boldsymbol{\beta}^{(0)}$, $\gamma^{(0)}$, $\tau^{(0)}$, $\boldsymbol{\delta}^{(0)}$, $\boldsymbol{\mu}^{(0)}$ with $\mu_0^{(0)} = -\infty$, $\mu_1^{(0)} = 0$ and $\mu_m^{(0)} = \infty$ and a value for the tuning parameter σ_{MH}^2. σ_{MH}^2 determines the acceptance rate of the Metropolis-Hastings step and should be set to achieve a reasonable acceptance rate, Johnson and Albert (1999) suggests $0.05/m$ as a default.

For $j = 1$ to $B + R$

1) Set $f_0 = -\infty$, $f_1 = 0$ and $f_m = \infty$.
   a. For $k = 2, \dots, m-1$ generate proposals f_k from a $N\left(\mu_k^{(j-1)}, \sigma_{MH}^2\right)$ distribution truncated to the interval $\left(f_{k-1}, \mu_{k+1}^{(j-1)}\right)$.
   b. Set $\mu_k^{(j)} = f_k$ with probability
   $$\alpha = \prod_{i=1}^{n} \left\{ \frac{\Phi\left[\sqrt{\tau_z}\left(f_{z_i} - m_i\right)\right] - \Phi\left[\sqrt{\tau_z}\left(f_{z_i-1} - m_i\right)\right]}{\Phi\left[\sqrt{\tau_z}\left(\mu_{z_i}^{(j-1)} - m_i\right)\right] - \Phi\left[\sqrt{\tau_z}\left(\mu_{z_i-1}^{(j-1)} - m_i\right)\right]} \right\} \times \prod_{k=2}^{m-1} \left\{ \frac{\Phi\left[\left(\mu_{k+1}^{(j-1)} - \mu_k^{(j-1)}\right)/\sigma_{MH}\right] - \Phi\left[\left(f_{k-1} - \mu_k^{(j-1)}\right)/\sigma_{MH}\right]}{\Phi\left[\left(f_{k+1} - f_k\right)/\sigma_{MH}\right] - \Phi\left[\left(\mu_{k-1}^{(j-1)} - f_k\right)/\sigma_{MH}\right]} \right\}$$
   otherwise set $\boldsymbol{\mu}^{(j)} = \boldsymbol{\mu}^{(j-1)}$.
2) Generate $z_i^{*(j)}$ from the full conditional posterior $z_i^*|\boldsymbol{\beta}^{(j-1)}, \gamma^{(j-1)}, \boldsymbol{\delta}^{(j-1)}, \tau^{(j-1)}, \boldsymbol{\mu}^{(j)}, y_i, z_i$ in (8).
3) Generate $\boldsymbol{\delta}^{(j)}$ from the full conditional posterior $\boldsymbol{\delta}|\boldsymbol{z}^{*(j)}$ in (5).
4) Generate $\tau^{(j)}$ from the full conditional posterior $\tau\left|\boldsymbol{\beta}^{(j-1)}, \gamma^{(j-1)}, \boldsymbol{z}^{*(j)}, \mathrm{y}\right.$ in (4).
5) Generate $\boldsymbol{\theta}^{(j)} = \left(\beta^{(j)\prime}, \gamma^{(j)}\right)'$ from the full conditional posterior $\boldsymbol{\theta}|\tau^{(j)}, \boldsymbol{\delta}^{(j)}, \boldsymbol{z}^{*(j)}, \boldsymbol{y}$ in (3).

Discard the first B draws as burn-in.

We have experimented with both approaches and found that Cowles' approach leads to a Markov chain that mixes well with lower autocorrelation in the chain than the pure Gibbs sampler. Our preferred sampler is summarized in Algorithm 1. Note that we take advantage of the fact that the acceptance ratio does not depend on z_i^* and generate new values for the latent variable irrespective of if the proposed cut points are accepted or not in order to improve the mixing of the chain.

3.1 Performance of the MCMC algorithm and inference procedure

To check the properties of the sampler in Algorithm 1 we generated a synthetic data set with $n = 500$ observations from the data generating process

$$z_i^* = 1 + 2.5x_{1i} + 1.5x_{2i} - 3x_{3i} + \eta_i$$
$$y_i = 2 + 1.4x_{3i} + 0.5x_{4i} + 3x_{5i} + 4z_i^* + \varepsilon_i$$

with the explanatory variables $x_{ji} \sim N(1,1)$, $\eta_i \sim N(0,1)$ and $\varepsilon_i \sim N(0,4)$. Note that the explanatory variable x_{3i} is common to the two equations. For the vector of cut points we set $\mu_2 = 2$ and $\mu_3 = 4$ leaving 159, 84, 86 and 171 observations in the four categories of the ordinal variable z_i.

Using an uninformative prior with

$$\boldsymbol{\beta}_0 = (1.5, 1.6, 0.7, 2.8)'$$
$$\boldsymbol{\Sigma}_{0\beta} = 10000\boldsymbol{I}$$
$$\gamma_0 = 0.5$$
$$\tau_\gamma = 0.00001$$
$$\boldsymbol{\delta}_0 = (0.8, 2.5, 1.6, -3.4)'$$
$$\boldsymbol{\Sigma}_{0\delta} = 10000\boldsymbol{I}$$
$$a = b = 0.1$$

we run a number of experiments to assess the convergence properties of the sampler. Throughout we set $\sigma_{MH}^2 = 0.0225$ which yields an acceptance rate of about 0.35 for the Metropolis-Hastings step.

Based on the output from one chain with 100 000 draws the diagnostic of Geweke (1992) indicates that convergence is relatively quick but also that there are some outliers produced by the chain which will affect the results if a too short run is used. To confirm we run 5 chains with random overdispersed starting values, each with 20 000 draws after discarding 5 000 draws as burn-in. Applying the Gelman and Rubin (1992) diagnostic yields potential scale reduction factors of 1.01 or less for the parameters. Running means for the first 1 500 draws from the five chains and the trace plot from one of the chains is displayed in Figure 1 for two of the parameters, γ and μ_3, which appear to be among the ones that are slowest to converge.

The autocorrelations in the chain are small for parameters on the observed explanatory variables in the equation for y_i but quite persistent for other parameters. The autocorrelation functions for the draws of selected parameters are displayed in Figure 2.

Figure 1: Running means and trace plots for simulated data

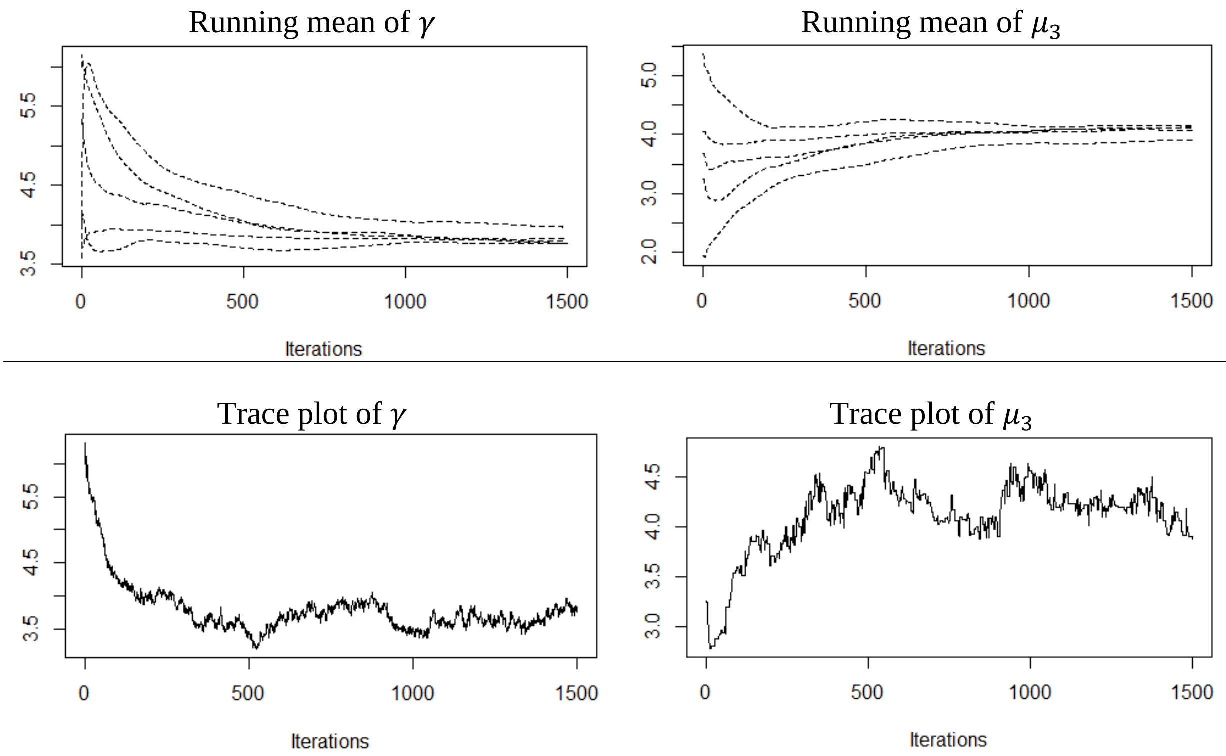
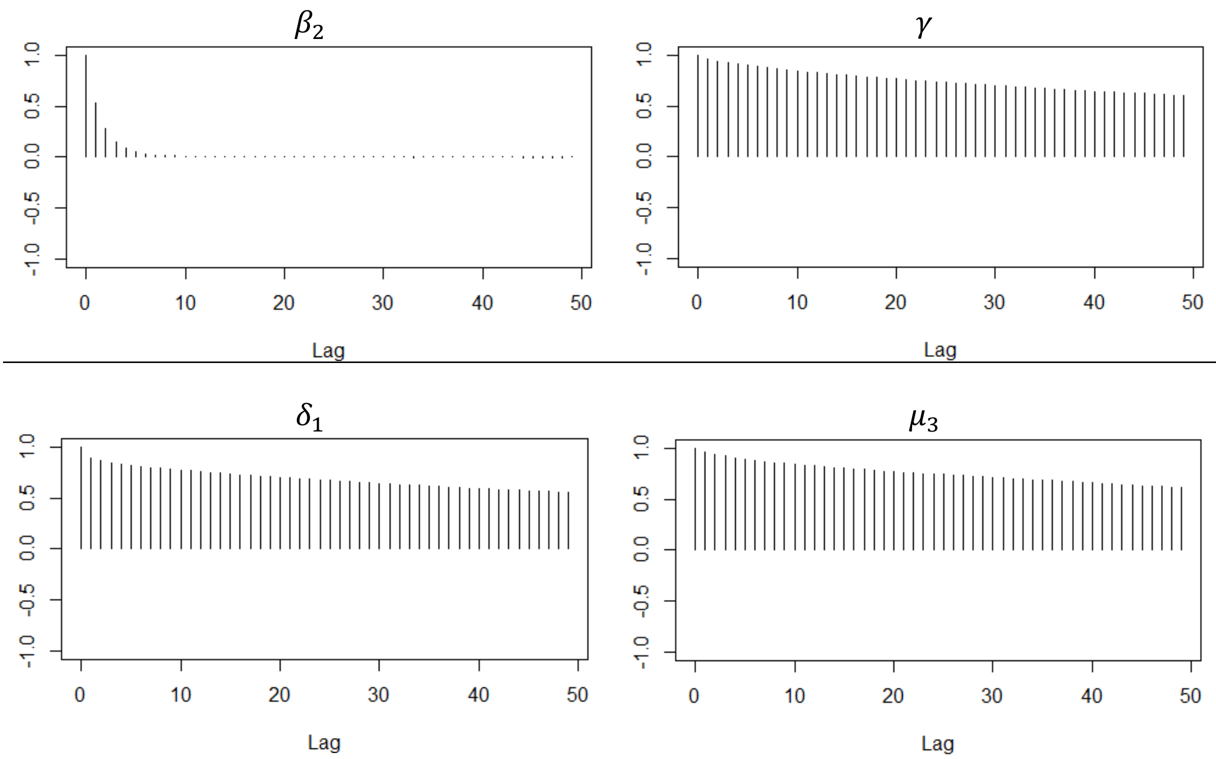


Figure 2. Autocorrelation functions for simulated data.



Summary statistics for the output of the long chain after discarding 5 000 draws as burn-in are displayed in Table 1 together with the in practice infeasible OLS estimates of the two equations using the actual values of z_i^* as comparison. The posterior distributions are well centered on the true parameter values with the possible exception of γ and τ and even for these parameters the 95% highest posterior density regions includes the true parameter values. Compared to the OLS estimates the Bayes estimates does equally well although there is a loss in precision due to the inability to observe the latent variable.

The coefficient γ on the latent variable z_i^* can be difficult to interpret due to the lack of a fixed scale for the variable. The cut points in the ordinal probit equation do, however, carry information about the scale of z_i^*. Recall that the cut point between the first and second category is fixed at 0. This together with, for example, the posterior means of the other cut points – in this case 2.09 and 4.15 – allows us to interpret γ. A move from the borderline between the first and second category to the borderline between the second and third categories would, for example, correspond to a change of the dependent variable y by about 8 units.

Table 1: Posterior distribution for simulated data

Parameter	True value	OLS[a]	Posterior mean	Posterior std. dev.	Percentiles 2.5%	Percentiles 50%	Percentiles 97.5%	Num. std. err.
β_0	2.0	2.12	2.01	0.53	0.98	2.01	3.05	0.0068
β_1	1.4	1.22	1.36	0.27	0.84	1.36	1.88	0.0020
β_2	0.5	0.58	0.53	0.16	0.21	0.53	0.84	0.0009
β_3	3.0	2.88	3.13	0.16	2.82	3.12	3.43	0.0009
γ	4.0	3.98	3.75	0.18	3.41	3.75	4.11	0.0081
τ	0.25	0.25[b]	0.17	0.02	0.13	0.17	0.23	0.0004
δ_0	1.0	1.04	1.09	0.15	0.80	1.09	1.39	0.0034
δ_1	2.5	2.48	2.62	0.14	2.36	2.61	2.89	0.0058
δ_2	1.5	1.52	1.65	0.09	1.47	1.65	1.83	0.0035
δ_3	–3.0	–2.98	–3.21	0.17	–3.55	–3.21	–2.90	0.0077
μ_2	2.0		2.09	0.16	1.78	2.08	2.41	0.0055
μ_3	4.0		4.15	0.23	3.71	4.14	4.63	0.0112

[a] Infeasible estimate using actual values of the latent variable z_i^* .
[b] Inverse of residual variance.

4. Gender based wage discrimination

Palme and Wright (1992) studied gender discrimination in the Swedish labor market. Here we estimate a version of their wage equation with job characteristics and individual specific characteristics one of which is an ordinal variable representing the level of education. This is a rather coarse classification of the true, latent, education level and simply including this as a set of dummy variables would introduce a measurement error in the model. Consequently we use the latent variable estimation procedure introduced in section 2 to estimate the wage equation.

We model the natural logarithm of the wage as a function of sex, age, marital status, work experience, integration in Swedish society (as measured by the home language when growing up), education level, region of residence and job characteristics. The education level is in turn modeled as a function of the education level of the parents. The data set is taken from the year 2000 wave of the Swedish Level of Living Survey (Levnadsnivåundersökningen) (Johansson and Eriksson, 2000) and consists of 2 767 observations after restricting the sample to those who were employed at the time of the survey and removing observations with missing data. The variables are summarized in Table 2.

Following Palme and Wright we capture the job characteristics by taking the first few principal components of a large set of indicator variables describing mainly the physical aspects of the job. Based on a Scree plot we choose to use the first four principal components which account for 43% of variation in the data. See Table 3 for details.

In order to study the existence of gender discrimination we decompose the wage differential into four different components along the lines of Oaxaca (1973) and Blinder (1973). That is we estimate an extended version of the model discussed in section 2,

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{x}_{Mi}'\boldsymbol{\beta}_M + z_i^*\gamma + z_{Mi}^*\gamma_M + \boldsymbol{p}_i'\boldsymbol{\theta} + \boldsymbol{p}_{Mi}'\boldsymbol{\theta}_M + \varepsilon_i$$

$$z_i^* = \boldsymbol{w}_i'\boldsymbol{\delta} + \boldsymbol{w}_{Mi}'\boldsymbol{\delta}_M + \eta_i.$$

where $\boldsymbol{x}_i$ represents individual characteristics, z_i^* the unobserved level of schooling, $\boldsymbol{p}_i$ the principal components representing job characteristics and $\boldsymbol{w}_i$ is the education level of the parents. The subscript *M* on the variables indicate that the variables have been interacted with the dummy variable for males and the corresponding subscripted coefficient measures the difference between males and females.[3]

The wage differential between men and women can then be decomposed into

$$\bar{y}_M - \bar{y}_F = (\bar{\boldsymbol{x}}_M' - \bar{\boldsymbol{x}}_F')\boldsymbol{\beta} + (\bar{z}_M - \bar{z}_F)\gamma + (\bar{\boldsymbol{p}}_M' - \bar{\boldsymbol{p}}_F')\boldsymbol{\theta} + \bar{\boldsymbol{x}}_M'\boldsymbol{\beta}_M + \bar{z}_M\gamma_M + \bar{\boldsymbol{p}}_M'\boldsymbol{\theta}_M$$

the differential due to different individual characteristics, different levels of schooling, different job characteristics and the differential (or wage discrimination) due to differences in the compensation for individual characteristics, levels of schooling and job characteristics.

The model is estimated with the variables in Table 2 (including the squares of Age, ExperWork, ExperCompany and ExperJob), the four principal components in Table 3, the unobserved education level and interaction terms with the dummy for males. The education level is modeled as a function of the mother and fathers education levels, again including interaction terms with the dummy for males.

Specifying an uninformative prior we run a first chain with 50 000 replicates. Inspection of the output indicates that the convergence properties are quite good. The lag 10 autocorrelation is negligible for all parameters except the constant term and the cut points in the equation for the latent education level which are 0.19, 0.52 and 0.53. Running 5 shorter chains with

[3] This requires minor modifications to the sampler to account for the interaction term between males and the unobserved level of schooling. The mean and variance of the truncated normal distribution for z_i^* in (8) is now given by $m_i = \tau_{z_i}^{-1}[\tau(\gamma + M_i\gamma_M)(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) + \boldsymbol{w}_i'\boldsymbol{\delta}]$ and $\tau_{z_i} = \tau(\gamma + M_i\gamma_M)^2 + 1$ where M_i is the dummy variable for males.

overdispersed starting values and 25 000 draws and applying the Gelman and Rubin (1992) diagnostic indicates that 5 000 draws as burn-in is more than sufficient.

Table 2: Data description

Variable	Description	Characteristics
Wage equation		
lnWage	Log of hourly wage	Mean = 4.71
Male	1 = Male	50.4%
Age	Age in 2000	Mean = 43.1
msMarried	1 = Married or cohabiting	74.8%
ExperWork	Years of work experience	Mean = 20.5
ExperCompany	Years at current employer	Mean = 8.98
ExperJob	Years in current position	Mean = 6.67
hlSwedish	1 = Swedish as home language	91%
hlNordic	1 = Other Nordic language	3%
hlEuro	1 = Other European language	3%
hlOther	1 = Other home language	3%
Education level	1 = Less than 7 years of primary education	7.4%
	2 = 9 years of primary education, vocational education	42.5%
	3 = High school degree, other tertiary education	35.0%
	4 = University degree, graduate education	15.1%
rStockholm	1 = residing in Stockholm	17.4%
rGothenburg	1 = residing in Gothenburg	7.6%
rMalmo	1 = residing in Malmö	4.1%
rLargeTown	1 = residing in town with more than 30 000 inhabitants	20.3%
rSmallTown	1 = residing in small town	19.5%
rRural	1 = rural resident	30.1%
Education level		
elFather1	Fathers education level 1 = Primary education or less	58.7%
elFather2	1 = Vocational education, high school	30.1%
elFather3	1 = University education	11.3%
elMother1	Mothers education level 1 = Primary education or less	61.6%
elMother2	1 = Vocational education, high school	29.2%
elMother3	1 = University education	9.1%

Table 3: Principal components capturing job characteristics

	Loadings			
Variable	PC 1	PC 2	PC 3	PC 4
PunctualYes	0.274	–0.011	0.154	–0.516
ClockYes	0.019	0.037	–0.332	0.043
InflexYes	–0.319	0.006	–0.159	0.647
Lift11	0.027	0.021	0.036	0.020
Lift12	0.066	0.043	0.025	0.045
Lift13	0.118	0.021	0.028	0.057
PhysicalDemYes	0.445	0.103	0.108	0.192
SweatYes	0.313	0.087	0.008	0.175
MentalYes	0.005	–0.469	0.502	0.199
RepetYes	0.186	–0.126	–0.339	–0.102
Stress1	–0.185	0.631	0.005	–0.030
Stress2	0.088	–0.541	–0.311	0.033
PhysicalExh1	0.023	–0.003	0.005	–0.008
PhysicalExh2	0.082	–0.018	0.003	–0.029
PhysicalExh3	0.209	0.037	0.141	0.300
Noise111	0.074	0.061	0.031	0.062
Noise112	0.028	–0.002	0.066	–0.012
Noise121	0.089	0.031	–0.058	–0.035
Noise122	0.064	–0.004	–0.023	–0.039
MonoBodyYes	0.232	0.001	–0.559	0.046
UnpleasBodyYes	0.441	0.104	0.080	0.212
Gas11	0.104	0.024	–0.087	–0.024
Gas12	0.083	0.027	0.001	0.030
Gas13	0.067	0.087	0.032	0.120
Dirty11	0.282	0.132	0.076	0.069
Dirty12	0.079	0.048	–0.019	0.082
Poison11	0.029	0.003	–0.005	0.000
Poison12	0.028	0.019	0.015	0.018
Poison13	0.066	0.062	–0.006	0.124
Variance proportion	0.176	0.114	0.077	0.053

Table 4: Posterior distribution of parameters in wage equation					
Parameter	Posterior mean	Posterior std. dev.	Percentiles 2.5%	97.5%	Numerical std. error $\times 10^4$
Constant	4.22878	0.11957	3.99473	4.46384	5.4621
age	0.00870	0.00678	–0.00464	0.02200	0.3111
ageSq	–0.00013	0.00007	–0.00027	0.00002	0.0034
hlScand	–0.02916	0.03675	–0.10053	0.04299	1.6407
hlEuro	–0.09613	0.04104	–0.17616	–0.01562	1.8930
hlOther	–0.10083	0.04061	–0.18006	–0.02041	1.8634
msMarried	0.03907	0.01546	0.00883	0.06917	0.7044
ExperWork	0.01180	0.00378	0.00439	0.01919	0.1757
ExperWorkSq	–0.00014	0.00007	–0.00028	0.00001	0.0034
ExperCompanyYr	0.00715	0.00343	0.00043	0.01388	0.1573
ExperCompanyYrSq	–0.00013	0.00010	–0.00033	0.00008	0.0047
ExperJobYr	–0.00410	0.00377	–0.01150	0.00329	0.1734
ExperJobYrSq	0.00006	0.00013	–0.00019	0.00030	0.0059
rGothenburg	–0.04510	0.02706	–0.09840	0.00786	1.2451
rMalmo	–0.08833	0.03413	–0.15473	–0.02137	1.5644
rLargeTown	–0.10904	0.02025	–0.14868	–0.06919	0.9268
rSmallTown	–0.14359	0.02159	–0.18616	–0.10144	0.9865
rRural	–0.11940	0.01922	–0.15702	–0.08163	0.8862
PC1	–0.07036	0.00883	–0.08763	–0.05309	0.4141
PC2	–0.01409	0.01083	–0.03531	0.00715	0.4988
PC3	0.01202	0.01133	–0.01006	0.03409	0.5248
PC4	0.05502	0.01334	0.02893	0.08103	0.6096
male	0.42968	0.17777	0.07822	0.77853	8.1922
m_age	–0.01666	0.01026	–0.03673	0.00353	0.4751
m_ageSq	0.00023	0.00011	0.00000	0.00045	0.0053
m_hlScand	0.09359	0.05880	–0.02199	0.20861	2.7238
m_hlEuro	0.04557	0.05547	–0.06302	0.15307	2.5651
m_hlOther	–0.14377	0.05604	–0.25347	–0.03445	2.5759
m_msMarried	0.02252	0.02177	–0.02036	0.06512	0.9975
m_ExperWork	0.00476	0.00551	–0.00599	0.01556	0.2572
m_ExperWorkSq	–0.00010	0.00010	–0.00031	0.00010	0.0047
m_ExperCompanyYr	0.00659	0.00463	–0.00254	0.01559	0.2126
m_ExperCompanyYrSq	–0.00015	0.00013	–0.00041	0.00011	0.0061
m_ExperJobYr	–0.00624	0.00518	–0.01646	0.00395	0.2384
m_ExperJobYrSq	0.00008	0.00016	–0.00024	0.00040	0.0077
m_rGothenburg	–0.03497	0.03932	–0.11223	0.04165	1.7993
m_rMalmo	–0.03581	0.05038	–0.13491	0.06317	2.3066
m_rLargeTown	–0.03339	0.02999	–0.09221	0.02544	1.3800
m_rSmallTown	–0.03457	0.03082	–0.09505	0.02616	1.4128
m_rRural	–0.04372	0.02831	–0.09910	0.01140	1.3066
m_PC1	–0.00903	0.01220	–0.03292	0.01479	0.5760
m_PC2	–0.02214	0.01419	–0.04985	0.00568	0.6585
m_PC3	–0.02023	0.01693	–0.05319	0.01289	0.7883
m_PC4	–0.04520	0.01848	–0.08177	–0.00906	0.8472
Z*	0.09367	0.00784	0.07834	0.10900	0.5345
m_Z*	0.01077	0.01183	–0.01246	0.03400	0.7012
tau	18.48276	0.52565	17.47386	19.53953	28.4691

Tables 4 and 5 summarizes the posterior distribution of the parameters obtained by running a second long chain with 50 000 draws after discarding 10 000 draws as burn-in. As expected, the education level has a positive effect on the wage. There is also a positive effect of experience, in general and with the same employer, although there appears to be a small negative effect from staying in the same position. It is also clear that the wage level is higher in Stockholm (the base category) than in other areas of the country. We can also note that the marriage premium found in many other studies is confirmed for both males and females. There is a distinct disadvantage to having been brought up with a non-Scandinavian home language, in particular for males with a non-European home language. Finally we note that the wage is substantially higher for males than for females while, for the most part, the interaction terms with the dummy for males have posterior means close to zero and there is considerable uncertainty about the sign.

Turning to the equation for the education level in Table 5 we find that the mothers and fathers education level has a positive effect on the education level with a tendency for males to have a lower education level than females. In addition, the effect of the mothers' education is somewhat weaker for males than females.

Table 6 and Figure 3 display the posterior distribution of the decomposition of the wage differential between males and females. In the sample the mean log wage is 4.80 for males and 4.62 for females, corresponding to geometric means of the hourly wage of 121.51 SEK and 101.63 SEK. Overall females have more favorable characteristics but males are better compensated, i.e. there is evidence of wage discrimination. The difference in compensation between males and females is largely driven by the compensation for the individual characteristics and the dummy for males, the property of being male. Similarly, females tend to have a higher education level but males are better rewarded for their educational achievements. Interestingly, there is a slight tendency for males to have better paying jobs while at the same time females tend to receive a higher reward for the job characteristics.

Table 5: Posterior distribution of parameters in equation for education level

Parameter	Posterior mean	Posterior std. dev.	Percentiles 2.5%	Percentiles 97.5%	Numerical std. error $\times 10^4$
Const2	1.17569	0.04911	1.08018	1.27268	7.5641
elFather2	0.37130	0.06919	0.23570	0.50636	3.7632
elFather3	0.75303	0.11830	0.51941	0.98345	6.6027
elMother2	0.52244	0.06994	0.38522	0.65991	3.6637
elMother3	0.89225	0.12526	0.64692	1.13879	6.8815
male	–0.02871	0.05778	–0.14183	0.08441	2.9968
m_elFather2	0.04690	0.09813	–0.14435	0.23805	5.1683
m_elFather3	0.07791	0.16076	–0.23772	0.39390	8.5800
m_elMother2	–0.12314	0.09935	–0.31705	0.07220	5.1358
m_elMother3	–0.35220	0.17460	–0.69207	–0.01061	9.3177
c2	1.54805	0.03951	1.47255	1.62673	10.1250
c3	2.69977	0.04777	2.60767	2.79368	12.2313

Table 6: Posterior distribution of decomposition of the wage differential $\bar{y}_M - \bar{y}_F$					
Parameter	Posterior mean	Posterior std. dev.	Percentiles 2.5%	97.5%	Numerical std. error $\times 10^4$
Difference due to individual characteristics	–0.00674	0.00336	–0.01337	–0.00016	0.1563
Difference in compensation for individual characteristics	0.17730	0.02049	0.13711	0.21766	1.1654
Difference due to education level	–0.00766	0.00170	–0.01110	–0.00445	0.0921
Difference in compensation for education level	0.01650	0.01808	–0.01885	0.05184	1.0519
Difference due to job characteristics	0.00198	0.00216	–0.00224	0.00620	0.0992
Difference in compensation for job characteristics	–0.00263	0.00150	–0.00560	0.00029	0.0695
Total difference due to characteristics	–0.01242	0.00412	–0.02052	–0.00432	0.1933
Total difference due to compensation	0.19117	0.00976	0.17197	0.21013	0.4549

5. Concluding remarks

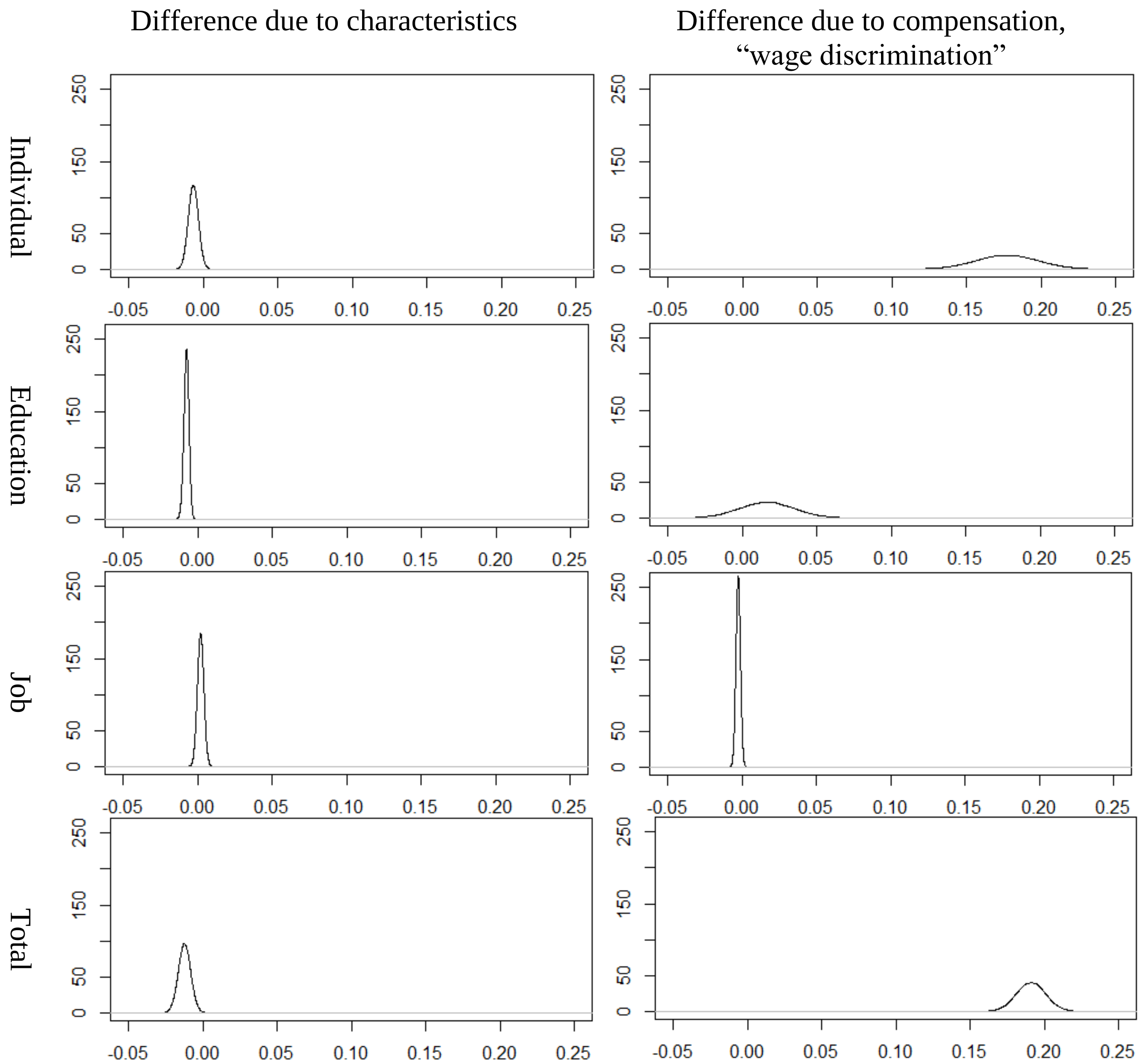
In this paper we propose a general solution to the problem of estimating linear regression models where one of the explanatory variables is ordinally observed. We extend the earlier literature in two directions. Firstly we consider a Bayesian formulation of the problem, secondly we explicitly consider both the process generating the underlying latent variable and the regression model of primary interest and estimate them jointly for improved efficiency.

We propose a straightforward Metropolis-Hastings within Gibbs sampler for exploring the posterior distribution and demonstrate that it performs well on both real and simulated data.

While we have not discussed this explicitly it should be clear that our inference procedure is also applicable in the simpler situation when the cut points for converting the unobserved latent variable into an ordinal indicator are known. E.g. when survey respondents are asked to report in which interval their income falls.

The usefulness of the modeling approach and inferential procedure is illustrated in an investigation of gender based wage discrimination in the Swedish labor market where the education level is only ordinally observed. We find substantial evidence of wage discrimation, part of which is due to females being less rewarded than males for their educational achievements.

Figure 3: Posterior distribution of decomposition of the wage differential $\bar{y}_M - \bar{y}_F$



References

Albert, J.H. and Chib, S., (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association*, **88,** 669-79.

Blinder, A.S., (1973), 'Wage Discrimination: Reduced Form and Structural Estimates', *The Journal of Human Resources*, **8**, 436-455.

Breslaw, J. A. and McIntosh, J., (1998), 'Simulated latent variable estimation of models with ordered categorical data', *Journal of Econometrics* **87**, 25-47.

Cowles, M. K. (1996), ' Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models,' *Statistics and Computing*, **6**, 101-111.

Gelman, A and Rubin, D. B., (1992), Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457-511.

Geweke, J., (1992), 'Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments' in Bernardo, J. M.; Berger, J. O.; David, A. P. & Smith, A. F. M. (ed.), *Bayesian Statistics 4,* Clarendon Press. p 169-193.

Hsiao, C. and Mountain, D. (1985) 'Estimating the short-run income elasticity of demand for electricity by using cross-sectional categorized data', *Journal of the American Statistical Association* **80**, 259-265.

Johansson, S and R Eriksson, (2000), *SND 0178 Levnadsnivåundersökningen,* http://snd.gu.se/sv/catalogue/study/389.

Johnson, V.E. and Albert, J.H. (1999), *Ordinal Data Modeling*, Springer-Verlag.

Oaxaca, R., (1973), 'Male-Female Wage Differentials in Urban Labor Markets', *International Economic Review*, **14**, 693-709.

Palme, M. and Wright, R., (1992) 'Gender discrimination and compensating differentials in Sweden', *Applied Economics*, **24**, 751-759.

Terza, J. V. (1987), 'Estimating linear models with ordinal qualitative regressors', *Journal of Econometrics* **34**, 275-291.