Part A: Linear regression model

Consider the following linear regression model:

$$\begin{aligned} \texttt{DIV}_i &= \beta_1 + \beta_2\,\texttt{LDIV}_i + \beta_3\,\texttt{EPS}_i + \beta_4 \ln(\texttt{MCAP}_i) + \beta_5 \ln(\texttt{OWN}_i) + \varepsilon_i, \\ \varepsilon_i &\overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad \sigma > 0, \end{aligned} \tag{1}$$

where $i = 1, \dots, 517$ represents the companies.

Question A.1: Explain *briefly* the economic content of this model (e.g., justify the choice of explanatory variables).

Suggested answer

This model assumes a relationship between dividends and a few explanatory variables: previous dividends, which give a signal of how likely a company is to pay out a dividend in the current period, earnings per share, which are expected to explain dividends (by definition, a dividend is a share of companies' earnings paid out to shareholders). Market capitalization, as a measure of corporate size, and the number of unique shareholders, are likely to explain dividends as well (see, e.g., the agency cost hypothesis of dividend theory), and are therefore included in the analysis. Some variables are used in log, to allow a nonlinear effect on dividends, despite the linear structure of the model.

Question A.2: Fit the linear regression model for the data described above using the Gibbs sampler. You can use the code provided in class (`gibbs_linreg.R`, available on Absalon under `Files/code`), your own code or a package of your choice. Justify your choice of prior parameters.

[In this question, you are only asked to make inference, not to derive the sampler or explain how the code is constructed.]

Suggested answer

The Gibbs sampler coded in class is run on the data to fit the linear regression model. For the prior, we specify $\beta \sim N_5(0, 100 \times I_5)$ and $\sigma^2 \sim IG(2, 1)$, which are rather noninformative prior distributions. A more informed analyst could decide to incorporate some prior information, for example based on previous studies that can give an idea of the sign and of the magnitude of the effect of the different explanatory variables used. The normal prior would then be centered around this 'prior guess,' and the strength of the belief in this value would be adjusted through the variance of the prior.

We run the Gibbs sampler for 11,000 iterations and discard the first 1,000 as burn-in. The corresponding code is provided in `MAIN.R`.

Question A.3: Summarize and explain the results. What can you conclude about the determinants of dividends, both from a statistical and from an economic point of view?

Suggested answer

Before summarizing the posterior distribution, convergence and mixing should be assessed. A visual check reveals that after the burn-in of 1,000 iterations the posterior distribution looks stationary, and mixing appears to be very good. This is confirmed by a plot of the autocorrelations, which disappear after one lag only. The inefficiency factors are close to 1 for all parameters, also indicating that the Gibbs sampler is almost as good as iid sampling in this case. The sampled values of the parameters can therefore be used for posterior inference.

The following table shows the posterior means, standard deviations and 95% highest posterior density intervals of the parameters:

	Mean	SD	[95%	HPD]
CONS	-0.857	0.749	-2.299	0.575
LDIV	0.832	0.020	0.794	0.872
EPS	0.034	0.006	0.024	0.046
log(MCAP)	-0.008	0.056	-0.122	0.099
log(OWN)	0.196	0.124	-0.050	0.439
SIGMA2	0.372	0.023	0.328	0.418

The results show evidence that lagged dividends and earnings per share have an impact on dividends, while market capitalization and the number of unique shareholders do not seem to play an important role. This can be seen from the 95% highest posterior density intervals in the last two columns of the table, which provide credible intervals for the corresponding parameters: the lower bounds of these intervals is far from zero for the first two variables (this is especially true for lagged dividends), but zero is included in the intervals of the last two parameters, indicating little evidence of an impact of these two variables. Lagged dividends appear to be the best predictor of current dividends, which makes sense from an economic point of view: some companies are used to rewarding their shareholders on a regular basis, others are known for not doing so.

Part B: Extending the linear regression model

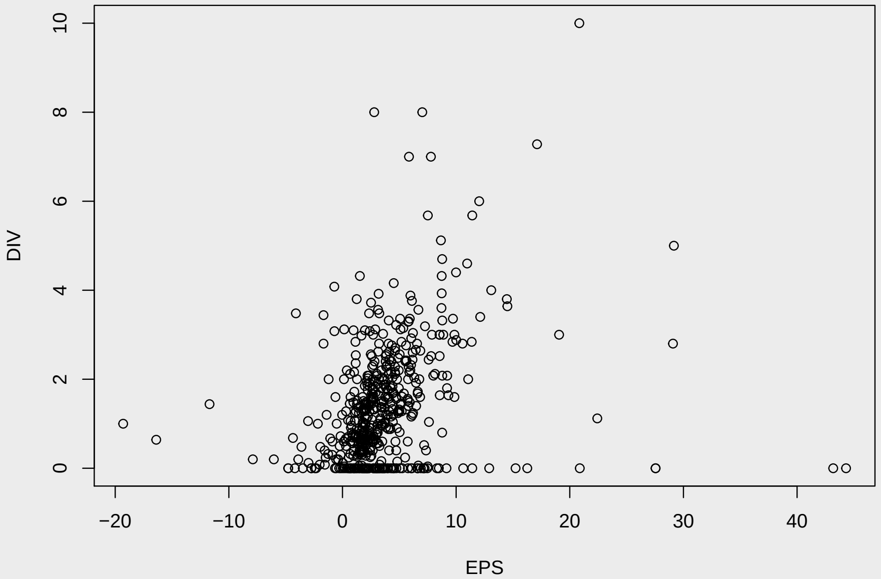
Some companies do not pay any dividends to their shareholders. This results in a particular structure of the dependent variable `DIV` that may be problematic for the use of the linear regression model.

The goal of Part B is to investigate a potential misspecification of the model used in Part A. That is, derive a model that is better specified for the problem at hand, and compare the results from both approaches.

Question B.1: To what extent does the particular structure of the dependent variable `DIV` create a problem for the use of the linear regression model? Show a pertinent figure that illustrates the problem.

Suggested answer

The dependent variable `DIV` represents the dividends paid out. Since not all companies pay out dividends, this variable contains zero values and is censored below zero. This can be seen in the following figure showing the joint values of `DIV` and `EPS`, where `DIV` clearly appears censored below zero, with a bunch of zero observations scattered on the X-axis. There are 96 companies not paying out dividends in this data set (i.e., 18.6% of the sample).



The linear regression model does not take into account this censoring of the dependent variable and treats the zero observations as regular data points, ignoring the underlying mechanism that determines if companies pay out dividends or not. Ignoring this particular feature of the data may therefore result in distorted results (biased estimator).

Question B.2: One of your colleagues suggests you work with the following likelihood function as an alternative to the linear regression model:

$$\ell(\beta, \sigma^2; Y, X) = \prod_{i=1}^{N} \left[1 - \Phi\left(\frac{X_i'\beta}{\sigma}\right)\right]^{\mathbb{1}\{Y_i=0\}} \left[\frac{1}{\sigma}\phi\left(\frac{Y_i - X_i'\beta}{\sigma}\right)\right]^{\mathbb{1}\{Y_i>0\}}, \qquad (2)$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote, respectively, the cumulative distribution function (CDF) and the probability density function (PDF) of the standard normal distribution $N(0,1)$, and $\mathbb{1}\{\cdot\}$ is the indicator function that is equal to 1 if the corresponding condition is fulfilled, to 0 otherwise. The vector of regression coefficients is $\beta = (\beta_1, \ldots, \beta_5)'$. The variable Y_i corresponds to the dependent variable $\texttt{DIV}_i$, and the vector X_i contains the same explanatory variables as in eq. (1) for company i, where the first entry of this vector is equal to 1 for the intercept term.

Specify the model that corresponds to this likelihood function. Explain how this model directly accounts for companies not paying dividends, and at the same time how it explains the amount of dividends paid out by companies who do.

Suggested answer

The likelihood function provided by your colleague clearly treats the censored observations (the zeros) differently from the observed dividends. The underlying model can be obtained from the linear regression model specified in eq. (1), but treating the dependent variable as a latent variable $Y_i^\star$, such that

$$Y_i^\star = X_i'\beta + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad (3)$$

where $X_i = \begin{pmatrix} 1 & \texttt{LDIV}_i & \texttt{EPS}_i & \ln(\texttt{MCAP}_i) & \ln(\texttt{OWN}_i) \end{pmatrix}'$. The variable $Y_i \equiv \texttt{DIV}$ is then obtained from the following observational rule:

$$Y_i = \begin{cases} Y_i^\star & \text{if } Y_i^\star > 0 \text{ (dividend paid out)}, \\ 0 & \text{if } Y_i^\star \leq 0 \text{ (\textit{no} dividend paid out)}, \end{cases} \qquad (4)$$

which is equivalent to $Y_i = \max\{0, Y_i^\star\}$.

The normality assumption on the error term provides the following conditional distribution for the latent variable:

$$Y_i^\star \mid X_i, \beta, \sigma^2 \overset{\text{ind}}{\sim} N(X_i'\beta, \sigma^2). \tag{5}$$

The likelihood is constructed according to the observational rule specified in eq. (4), which provides the following densities/probabilities for each of the two possible outcomes:

$$Y_i = \begin{cases} Y_i^\star & \text{with density } p(Y_i^\star \mid X_i, \beta, \sigma^2), \\ 0 & \text{with probability } Pr(Y_i^\star \leq 0 \mid X_i, \beta, \sigma^2). \end{cases}$$

The first density is obtained from eq. (5):

$$p(Y_i^\star \mid X, \beta, \sigma^2) = \frac{1}{\sigma}\phi\bigg(\frac{Y_i^\star - X_i'\beta}{\sigma}\bigg) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i^\star - X_i'\beta)^2\right\},$$

where $\phi(\cdot)$ denotes the probability density function (PDF) of the standard normal distribution $N(0,1)$.

The second probability is equal to:

$$\begin{aligned} Pr(Y_i^\star \leq 0 \mid X_i, \beta, \sigma^2) &= Pr(X_i'\beta + \varepsilon_i \leq 0 \mid X_i, \beta, \sigma^2), \\ &= Pr\left(\frac{\varepsilon_i}{\sigma} \leq -\frac{X_i'\beta}{\sigma} \mid X_i, \beta, \sigma^2\right), \\ &= \Phi\left(-\frac{X_i'\beta}{\sigma}\right), \\ &= 1 - \Phi\left(\frac{X_i'\beta}{\sigma}\right), \end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution, and the last line is obtained thanks to the symmetry of the normal distribution.

Given the observational rule specified in eq. (4), these two cases can be combined to produce the density function, which, expressed for the whole sample as the product of the individual densities thanks to the independence assumption (for $i = 1, \ldots, N$), corresponds to the likelihood function specified in eq. (2) (a function of the model parameters β and σ^2).

This model is called a *tobit model* and takes into account the censoring of the dependent variable `DIV` by distinguising its zero values from its positive values.

Question B.3: Explain why the likelihood function specified in eq. (2) is difficult to use directly for Bayesian inference. Propose an alternative solution that makes sampling feasible.

Suggested answer

The likelihood specified in eq. (2) is a complicated function of the model parameters β and σ^2, as it relies on the CDF of the standard normal disitribution, which has no closed-form solution. Therefore, it is impossible to find a conjugate prior or even a non-conjugate prior that would provide a posterior distribution of a known family that could easily be simulated.

As an alternative, it is possible to *augment* the likelihood with the latent variable $Y^\star$ for the censored values of the dividends. By doing so, the well-known results about the linear regression model can be applied to sample the regression coefficients β and the variance σ^2 based on eq. (3).

The corresponding *augmented* density is:

$$p(Y_i, Y_i^\star \mid X_i, \beta, \sigma^2) = p(Y_i \mid Y_i^\star, X_i, \beta, \sigma^2)p(Y_i^\star \mid X_i, \beta, \sigma^2),$$

where $p(Y_i \mid Y_i^\star, X_i, \beta, \sigma^2)$ is equal to 1 if dividends are positive and $Y_i = Y_i^\star$ at the same time (the latent variable is observed in that case), or if Y_i is censored and the corresponding latent variable $Y_i^\star$ is negative. It is equal to zero in all the other cases (for instance, $Y_i^\star$ cannot be negative is Y_i if nonzero).

A sampling scheme can then be implemented by simulating the latent variable for the censored observations, and then sampling the remaining parameters β and σ^2 conditional on the sampled $Y^\star$.

Question B.4: Specify a prior distribution for the model parameters. Explain the concept of conjugacy and use natural conjugate priors for your analysis, if such conjugate priors exist.

Suggested answer

A prior distribution needs to be assumed for the regression coefficients β and for the variance of the error term σ^2. A prior is said to be a natural conjugate if, multiplied by the likelihood, it leads to a posterior distribution that belongs to the same family of distribution.

In the latent variable model, the normal distribution is a conjugate prior for β and the inverse-gamma is a conjugate prior for σ^2 (similarly to the linear regression model):

$$\beta \sim N(b_0, B_0),$$
$$\sigma^2 \sim IG(c_0, d_0),$$

where $b_0 \in \mathbb{R}^K$ and B_0 is a $K \times K$ covariance matrix (K is the number of explanatory variables), and $c_0 > 0$, $d_0 > 0$. In our application we will use $b_0 = 0$ and $B_0 = 100 \times I_K$, $c_0 = 2$ and $d_0 = 1$, which are rather noninformative priors.

Question B.5: Propose a Markov chain Monte Carlo (MCMC) sampler for this model. Derive the (conditional) posterior distribution(s) used to update the model, and provide details about the different steps of your MCMC sampler.

Suggested answer
The sampling scheme relies on data augmentation and simulates the latent variable for the censored values of Y. The latent variable model is then a simple linear regression model. The usual updating process can be applied to the regression coefficients and to the variance of the error term, using Bayes' theorem.

The conditional posterior distributions of β and σ^2 are identical to those derived in class for the linear regression model. They are reproduced here for the sake of completeness.

Conditional posterior distribution of the regression coefficients. Using compact form notation, such that $Y^\star$ is the vector of length N containing the latent variable and X is the $N \times K$ matrix containing the explanatory variables, the application of Bayes' theorem provides:

$$\begin{aligned}
p(\beta \mid Y^\star, X, \sigma^2) &\propto p(Y \mid X, \beta, \sigma^2)\, p(\beta) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\beta' X' X \beta - 2\beta' X' Y^\star\right)\right\} \exp\left\{-\frac{1}{2}\left(\beta' B_0^{-1} \beta - 2\beta' B_0^{-1} b_0\right)\right\}, \\
&\propto \exp\left\{ -\frac{1}{2}\left[\beta' \underbrace{\left(\frac{X'X}{\sigma^2} + B_0^{-1}\right)}_{B_p} \beta - 2\beta' \underbrace{\left(\frac{X'Y^\star}{\sigma^2} + B_0^{-1} b_0\right)}_{b_p}\right]\right\},
\end{aligned}$$

which corresponds to the kernel of the following normal distribution:

$$\beta \mid Y, X, \sigma^2 \sim N(B_{\mathrm{p}}^{-1} b_{\mathrm{p}}, B_{\mathrm{p}}^{-1}).$$

Conditional posterior distribution of the variance of the error term.

$$\begin{aligned}p(\sigma^2 \mid Y, X, \beta) &\propto p(Y \mid X; \beta, \sigma^2)\, p(\sigma^2),\\ &\propto (\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2\right\}(\sigma^2)^{-c_0-1}\exp\left\{-\frac{d_0}{\sigma^2}\right\},\\ &\propto (\sigma^2)^{-\frac{N}{2}-c_0-1}\exp\left\{-\frac{1}{\sigma^2}\left(d_0 + \frac{1}{2}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2\right)\right\},\end{aligned}$$

providing the kernel of an inverse-gamma distribution:

$$\sigma^2 \mid Y, X, \beta \sim IG\left(c_0 + \frac{N}{2}, d_0 + \frac{1}{2}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2\right).$$

Conditional distribution of the latent variables. Since the latent variable is equal to the observed dependent variable when this one is positive, it is only necessary to derive the conditional distribution of $Y_i^\star$ when Y_i is equal to zero. For each $i = 1, \ldots, N$, we have:

$$\begin{aligned}p(Y_i^\star \mid Y_i = 0, X_i, \beta, \sigma^2) &\propto p(Y_i = 0 \mid Y_i^\star, X_i, \beta, \sigma^2) p(Y_i^\star \mid X_i, \beta, \sigma^2),\\ &\propto \mathbb{1}\{Y_i = 0\}\, \mathbb{1}\{Y_i^\star \leq 0\}\, \frac{1}{\sigma}\phi\left\{\frac{Y_i^\star - X_i'\beta}{\sigma}\right\},\end{aligned}$$

where $p(Y_i = 0 \mid Y_i^\star, X_i, \beta, \sigma^2)$ is obtained using the observational rule specified in eq. (4): Since the sign of $Y_i^\star$ completely determines Y_i when the latter one is equal to 0, this probability is equal to 1 if both conditions are fulfilled simultaneously.

The last expression shows that $Y_i^\star$ is sampled from the following truncated normal distribution when the corresponding Y_i is equal to 0:

$$Y_i^\star \mid Y_i = 0, X_i, \beta, \sigma^2 \overset{\text{ind}}{\sim} TN_{(-\infty,0]}(X_i'\beta, \sigma^2) \qquad \text{if } Y_i = 0. \tag{6}$$

Gibbs sampler for the tobit model. Initialize model parameters with starting values $\beta^{(0)}$, $\sigma^{2(0)}$. Repeat the following steps until practical convergence, for $t = 1, \ldots, T$:

(a) For each $i = 1, \ldots, N$:

  i. set $Y_i^\star = Y_i$ if $Y_i > 0$,

  ii. otherwise if $Y_i = 0$, sample $Y_i^\star$ from the truncated normal specified in eq. (6).

(b) Sample $\beta^{(t)}$ from $p(\beta \mid Y^\star, X, \sigma^{2(t-1)})$.

(c) Sample $\sigma^{2(t)}$ from $p(\sigma^2 \mid Y^\star, X, \beta^{(t)})$.

Question B.6: Write a computer program that implements your sampler. *[You can use the function `gibbs_linreg()` provided in class, available on Absalon under `Files/code`, and extend it for your needs.]*

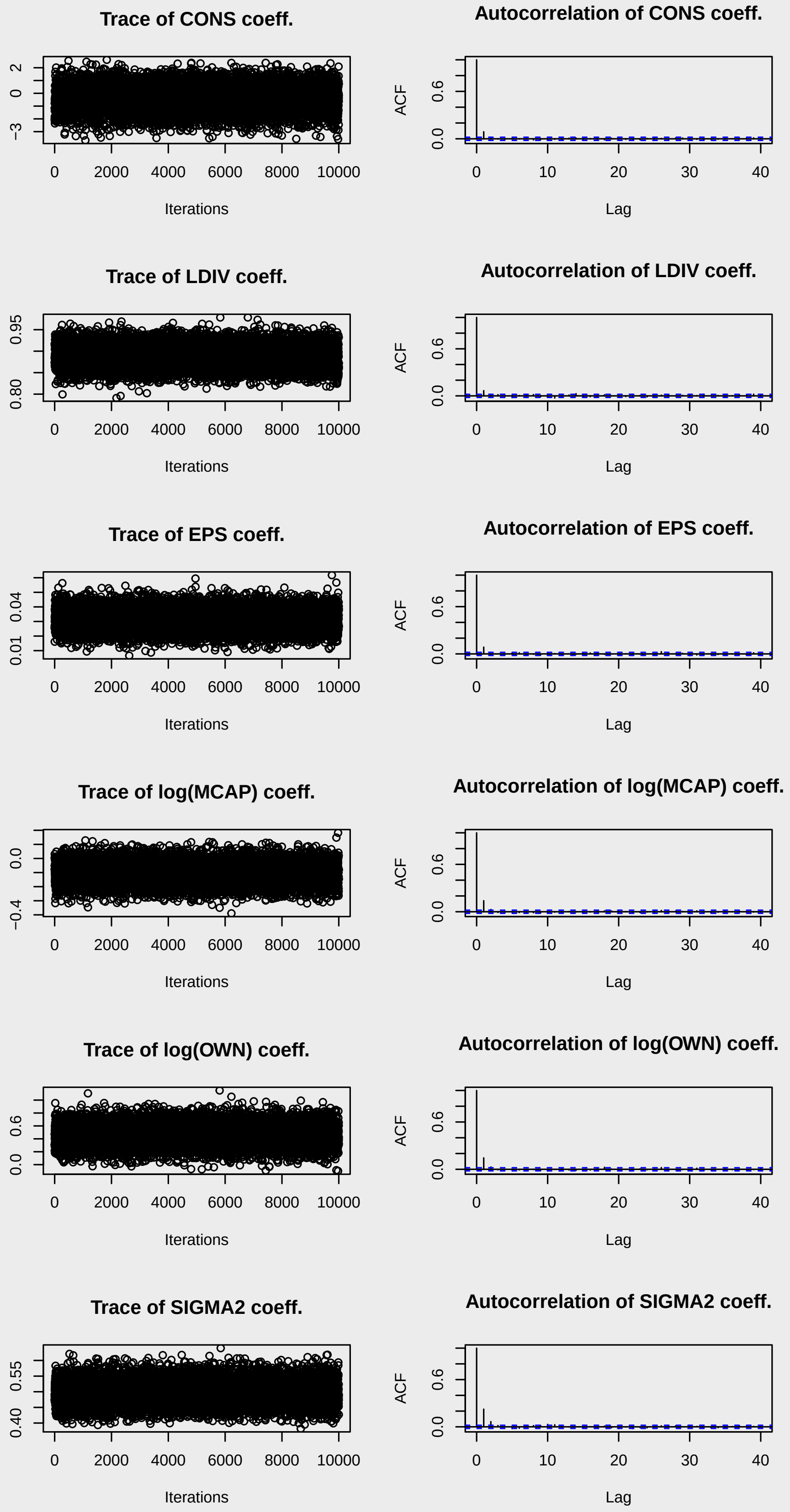
Suggested answer

See function `gibbs_tobit()` in file `gibbs_tobit.R`.

Question B.7: Run your sampler on the data. Summarize the posterior results and compare them to those obtained from the linear regression model. What are the main differences? You may illustrate your answer with a figure and/or a table.

Suggested answer

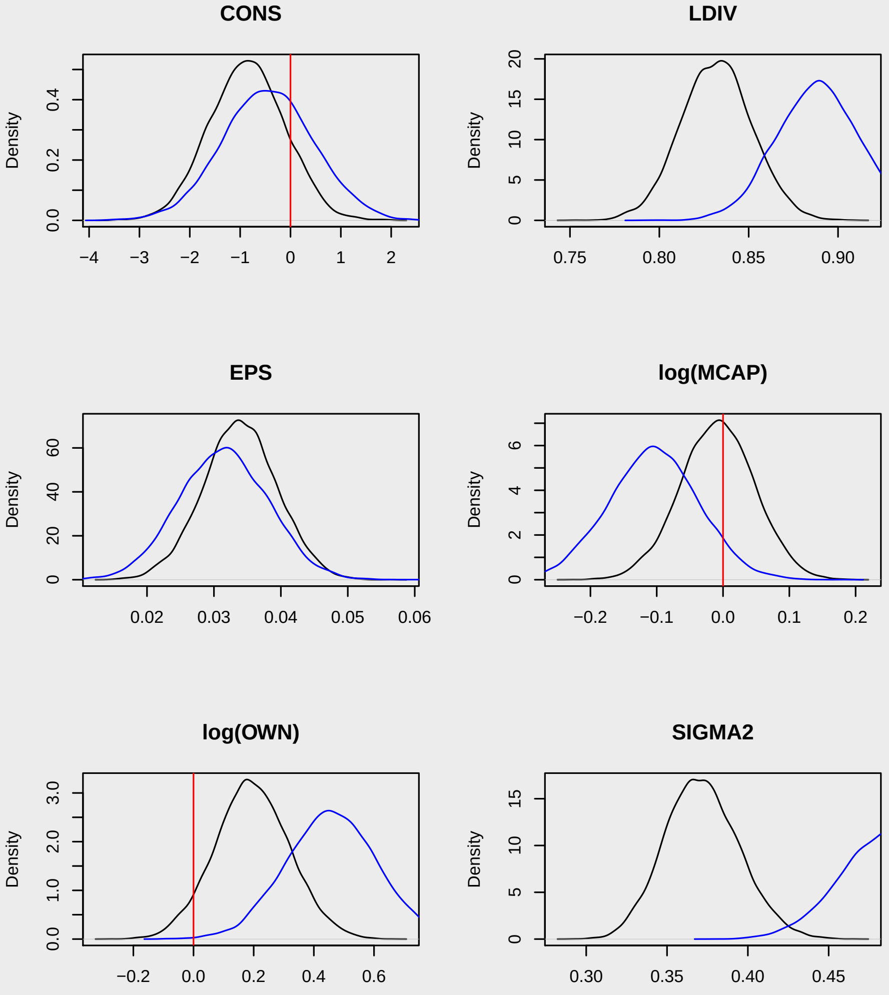
Before summarizing the posterior distribution, it is necessary to check convergence and mixing. This can be done visually using trace plots and autocorrelograms, which show that the posterior distribution looks stationary and autocorrelations fade out very quickly (indicating good mixing). These posterior draws can therefore be used for posterior inference.



The posterior is summarized in the following table, where the posterior means, standard deviations (SD) and 95% HPD intervals are displayed for each parameter:

	Mean	SD	[95%	HPD]
CONS	-0.445	0.907	-2.167	1.368
LDIV	0.889	0.023	0.845	0.937
EPS	0.031	0.007	0.018	0.044
log(MCAP)	-0.105	0.068	-0.240	0.026
log(OWN)	0.460	0.153	0.164	0.752
SIGMA2	0.495	0.035	0.428	0.565

The main differences between these posterior results and those obtained from the linear regression model used previously appear more obvious when plotting the estimated posterior densities against each other, for each parameter (linear model in black, censored model in blue, zero in vertical red line):



It appears that when using the linear regression model and ignoring the censoring of the dependent variable, the impact of the previous dividend (`LDIV`) and of the number of unique shareholders (OWN) are underestimated, while the impact of market capitalization (MCAP) is overestimated. There is, however, no real difference for the earnings per share ratio (EPS).

Interestingly, this improved specification of the model reveals more evidence from the data that market capitalization (MCAP) plays a role in determining dividends: with the linear regression model, the posterior probability that the corresponding coefficient (β_4) is negative was $Pr(\beta_4 < 0 \mid Y, X) = 55.87\%$, whereas with the censored model it is equal to 94.02%.

Therefore, ignoring the censoring in the dependent variable distorts the results. It is important to take into account the zero dividends to measure correctly the impact of the different variables of interest on dividend payout.

Question B.8: According to the theory, the number of shareholders should be related to the level of dividend payout *(agency cost hypothesis of dividend theory).*

Do you find evidence in the data supporting this theory? Justify your answer using test statistics.

Suggested answer

This hypothesis can be tested by computing a credible interval for β_5. The 95% highest posterior density interval is provided in the previous table. The lower bound of this interval for β_5 is clearly larger than zero, indicating evidence in the data in favor of this theory. This is confirmed by the computation of the posterior probability $Pr(\beta_5 > 0 \mid Y, X)$, which is equal to 99.85%.

Part C: Accounting for sector heterogeneity

There might be large differences between sectors that explain dividend payout. Assume you now have access to a variable SEC indicating to which sector each company belongs to. This variable could, for instance, be the 11 categories of the Global Industry Classification Standard (GICS) sector.

Question C.1. Propose an extension to the model used in Part B that allows to capture heterogeneity across sectors. Your model should be able to account for differences in levels, structural differences with respect to the explanatory variables, and differences in unobserved heterogeneity across sectors.

Provide details about the model specification, including prior specification. Justify the relevance of your specification.

Suggested answer

Companies' heterogeneity across sectors can be accounted for by specifying a *hierarchical model* that adds more layers to the original model, in order to introduce heterogeneity.

For example, differences in levels can be captured by random intercepts, structural differences with respect to the explanatory variables by random regression coefficients, and remaining differences in unobsered heterogeneity by adding a mixing parameter to the variance of the error term. The resulting latent variable model would be expressed as:

$$Y_i^\star \overset{\text{ind}}{\sim} N(X_i'\beta_{\texttt{SEC}_i}, \lambda_{\texttt{SEC}_i}\sigma_\varepsilon^2),$$

with a first hierarchical level:

$$\begin{aligned} \sigma_\varepsilon^2 &\sim IG(c_0, d_0), \\ \lambda_j &\overset{\text{iid}}{\sim} IG(g_0, h_0), \qquad j = 1, \ldots, \text{card}(\texttt{SEC}), \\ \beta_j \mid b, B &\overset{\text{iid}}{\sim} N_5(b, B), \end{aligned}$$

and a second hierarchical level:

$$\begin{aligned} b &\sim N_5(b_0, V_0), \\ B &\sim IW_5(\nu_0, S_0), \end{aligned}$$

for $i = 1, \ldots, N$, where card(SEC) is the number of single industry sectors in the variable SEC, $N_K(\cdot, \cdot)$ denotes the multivariate normal distribution of dimension K, and $IW_K(\cdot, \cdot)$ denotes the inverse-Wishart distribution of dimension K.

The second level of hierarchy would allow to learn more about the structural differences, compared to the standard tobit model used in Part B, as we would obtain different posterior distributions for the regression coefficients across sectors.

Question C.2. How would you modify your MCMC sampler derived in Part B to accommodate this/these new feature/s of the model? Describe *briefly* the modifications. *[Only sketch the resulting sampler, do not derive any posterior distributions].*

Suggested answer

The MCMC sampler would have additional steps to sample the parameters of the additional layers: The parameters that are sector-specific, β_j and λ_j, would be sampled sequentially within each sector (i.e., card(`SEC`) substeps for each parameter, which would be done in a loop, or vectorized). The hyperprior parameters b and B would be sampled in an additional step, conditional on the sampled values of β_j. All of these substeps would be performed using simple Gibbs updates, as the corresponding conditional distributions can all be derived in closed-form solutions and correspond to known families.