# *Mixture Models*

Econ 690

Purdue University

- In virtually all of the previous lectures, our models have made use of normality assumptions.
- From a computational point of view, the reason for this assumption is clear: combined with normal priors for regression parameters, this yields convenient posterior (or conditional) posteriors for regression parameters, whence standard simulation methods can be applied.
- However, such assumptions may not be supported at all by the data, and diagnostic checks could reveal evidence against normality.
- So, what should we do in these cases?
- Are there any more flexible alternatives which retain computational tractability?

- To this end, we first describe scale mixtures of normals models.
- The most popular of these involve generalizing our models to allow for Student-t errors, so that our model can accommodate fat tails in the data.
- Other distributions can also be obtained as a scale mixture of normals, including (among others): double exponential errors and logistic errors.
- Such models, though more flexible than the textbook Gaussian model, are symmetric and can not accommodate features such as skew and multimodality.
- We then discuss finite Gaussian mixtures as an alternative for these cases, and also talk (a little) about models with skew-Normal errors.

# Scale Mixtures of Normals Models

We first review how the Student-t density can be regarded as a scale mixture of the Gaussian density.

Suppose you specify:

$$y|\beta, \lambda, \sigma^2 \sim N(x\beta, \lambda\sigma^2)$$

and choose the following prior for $\lambda$ (treating $\nu$ as given):

$$\lambda|\nu \sim IG\left(\frac{\nu}{2}, \frac{2}{\nu}\right) \Rightarrow p(\lambda) = \left[\Gamma\left(\frac{\nu}{2}\right)\left(\frac{2}{\nu}\right)^{\nu/2}\right]^{-1} \lambda^{-[(\nu/2)+1]} \exp\left(-\frac{\nu}{2\lambda}\right)$$

so that the prior for $\lambda$ is independent of $\beta$ and $\sigma^2$.

For this model, we seek to:

(a) Derive the density

$$p(y|\beta,\sigma^2) = \int_0^\infty p(y|\beta,\lambda,\sigma^2)p(\lambda)\ d\lambda.$$

(b) Given the result in (a), comment on how the addition of $\lambda$ to the error variance can be a useful computational device for an applied Bayesian researcher.

It follows that

$$\begin{aligned} p(y|\beta,\sigma^2) &= \int_0^\infty \left( \left[\sqrt{2\pi}\sigma\right]^{-1} \lambda^{-(1/2)} \exp\left[-\frac{1}{2\lambda}\left(\frac{y-x\beta}{\sigma}\right)^2\right]\right) \times \\ &\quad \left( \left[\Gamma\left(\frac{\nu}{2}\right)\left(\frac{2}{\nu}\right)^{\nu/2}\right]^{-1} \lambda^{-[(\nu/2)+1]} \exp\left(-\frac{\nu}{2\lambda}\right)\right) d\lambda \\ &= \left[\sqrt{2\pi}\sigma\right]^{-1}\left[\Gamma\left(\frac{\nu}{2}\right)\left(\frac{2}{\nu}\right)^{\nu/2}\right]^{-1} \times \\ &\quad \int_0^\infty \lambda^{-[(\nu+3)/2]} \exp\left[-\frac{1}{\lambda}\left(\frac{1}{2}\left[\frac{y-x\beta}{\sigma}\right]^2 + \frac{\nu}{2}\right)\right] d\lambda. \end{aligned}$$

The integral above is the kernel of an

$$IG\left(\frac{\nu+1}{2}, \left(\frac{1}{2}\left[\left(\frac{y-x\beta}{\sigma}\right)^2 + \nu\right]\right)^{-1}\right)$$

density.

Thus,

$$p(y|\beta,\sigma^2) = \left[\sqrt{2\pi}\sigma\right]^{-1}\left[\Gamma\left(\frac{\nu}{2}\right)\left(\frac{2}{\nu}\right)^{\nu/2}\right]^{-1} \times \Gamma\left(\frac{\nu+1}{2}\right)\left(\frac{1}{2}\left[\left(\frac{y-x\beta}{\sigma}\right)^2+\nu\right]\right)^{-[(\nu+1)/2]}.$$

Rearranging and canceling terms, we obtain

$$p(y|\beta,\sigma^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\sigma\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{1}{\nu}\left(\frac{y-x\beta}{\sigma}\right)^2\right]^{-[(\nu+1)/2]},$$

which is in the form of a Student-t density, i.e.,

$$y|\beta,\sigma^2 \sim t(x\beta,\sigma,\nu).$$

- In a sense, we can think about this procedure as something like data augmentation. The parameter $\lambda$ is not necessarily an object of interest (though it could be), but is, instead, a useful device for allowing for Student-t errors.
- Specifically, conditioned on $\lambda$, all of our usual Gibbs sampling results will apply.
- Similarly, given all of the other parameters of the model, sampling from $\lambda$'s posterior conditional is also straight-forward.
- In other words, this result is useful to the applied Bayesian researcher as it, in conjunction with the Gibbs sampler, allows the estimation of models with Student-t errors, thus relaxing the Normality assumption.

To see this connection more formally, regard (a) as one observation's contribution to the likelihood function and note (adding $i$ subscripts to denote the individual observations)

$$p(\beta, \sigma^2, \{\lambda_i\}|y) \propto \left[\prod_{i=1}^{n} \phi(y_i; x_i\beta, \lambda_i\sigma^2)p(\lambda_i)\right] p(\beta, \sigma^2),$$

which implies that

Thus working with the seemingly more complicated joint posterior which contains the inverted-Gamma mixing variables $\lambda_i$ yields the same inference for $\beta$ and $\sigma^2$ that would be obtained by directly working with a regression model with Student-t errors.

We will show how this is done in the context of a linear regression model below.

Consider the regression model:

- 

We seek to show how the Gibbs sampler can be used to fit this model.

To implement the Gibbs sampler we need to obtain the complete posterior conditionals for the parameters $\beta$, $\sigma^2$ and $\{\lambda_i\}$.

The joint posterior distribution is given as

$$p(\beta, \{\lambda_i\}, \sigma^2|y) \propto \left[\prod_{i=1}^{n} \phi(y_i; x_i\beta, \lambda_i\sigma^2)p(\lambda_i)\right] p(\beta)p(\sigma^2).$$

Given this joint posterior, we need to obtain the posterior conditionals: $p(\beta|\lambda, \sigma^2, y)$, $p(\lambda|\beta, \sigma^2, y)$ and $p(\sigma^2|\beta, \lambda, y)$.

The following complete conditional posterior distributions are obtained:

$$\beta|\{\lambda_i\},\sigma^2|y \sim N\Big[\left(X'\Lambda^{-1}X/\sigma^2+V_\beta^{-1}\right)^{-1}\left(X'\Lambda^{-1}y+V_\beta^{-1}\mu_\beta\right), \left(X'\Lambda^{-1}X/\sigma^2+V_\beta^{-1}\right)^{-1}\Big],$$

where $\Lambda \equiv \text{diag}\{\lambda_i\}$ and thus $\Lambda^{-1} = \text{diag}\{\lambda_i^{-1}\}$, and $X$ and $y$ are stacked appropriately.

As for the posterior conditional for the variance parameter $\sigma^2$,

$$\sigma^2|\beta, \{\lambda_i\}, y \sim IG\left[\frac{n}{2} + a, \left(b^{-1} + \frac{1}{2}(y - X\beta)'\Lambda^{-1}(y - X\beta)\right)^{-1}\right].$$

Finally, we can apply our previous result to derive the posterior conditional for each $\lambda_i$:

- A Gibbs sampler involves cycling through these conditionals.
- Note that different choices of $\nu$ in the hierarchical prior for $\lambda_i$ yield models with different tail properties.
- Finally, note that this procedure can be extended to allow for double exponential errors under an *exponential* prior for $\lambda$ and logistic errors provided the mixing variables $\lambda$ have the *asymptotic distribution of the Kolmogorov distance statistic* [Andrews and Mallows (JRSS B, 1974)].

# *Finite Gaussian Mixtures*

Consider a two-component Normal mixture model

- 

Note that, to generate values $y$ from this model, one can first draw from a two-point distribution with probabilities $P$ and $1 - P$. Given a draw from this two-point distribution, one can then draw from the associated component of the mixture [either $N(\mu_1, \sigma_1^2)$ or $N(\mu_2, \sigma_2^2)$] to obtain a draw $y$.

Using the above intuition, we will augment the mixture model with a set of component indicator variables, say $\{\tau_i\}_{i=1}^n$, where $\tau_i$ is either zero or one, and $\tau_i = 1$ implies that the $i^{th}$ observation is drawn from the first component of the mixture. (When $\tau_i = 0$, the implication is that the $i^{th}$ observation is drawn from the second component).

We will also assign a hierarchical prior to $\tau_i$ so that the probability associated with the first component is $P$, and then place a Beta prior on $P$. Using this augmented structure, we will describe how a Gibbs sampler can be employed to fit the mixture model.

Before describing the augmented representation, let $\theta$ denote all the model's parameters and $\theta_{-x}$ denote all parameters other than $x$.

The model can be written as

$$\begin{aligned} p(y|\theta,\{\tau_i\}) &= \prod_{i=1}^{n} \left[\phi(y_i;\mu_1,\sigma_1^2)\right]^{\tau_i} \left[\phi(y_i\ \mu_2,\sigma_2^2)\right]^{1-\tau_i} \\ \tau_i &\overset{iid}{\sim} B(1,P), \quad i=1,2,\cdots n \\ P &\sim \beta(\underline{p}_1,\underline{p}_2), \\ \mu_i &\overset{ind}{\sim} N(\underline{\mu}_i,\underline{v}_i), \quad i=1,2 \\ \sigma_i^2 &\overset{ind}{\sim} IG(\underline{a}_i,\underline{b}_i), \quad i=1,2. \end{aligned}$$

In the above $B(1,P)$ denotes a Binomial density on one trial with "success" probability $P$, or equivalently, a Bernoulli density with success probability $P$. Similarly, $\beta(a,b)$ denotes the Beta density with parameters $a$ and $b$.

Note that when marginalizing the conditional likelihood $p(y|\theta, \{\tau_i\})$ over $\tau_i$, we are left with the two-component mixture model described at the outset of this section. To see this, note that the assumed conditional independence across observations, together with the fact that $\tau_i$ is binary, implies

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{n} \sum_{j=0}^{1} p(y_i|\theta, \tau_i = j)\Pr(\tau_i = j|\theta) \\ &= \prod_{i=1}^{n} \left[P\phi(y_i; \mu_1, \sigma_1^2) + (1-P)\phi(y_i; \mu_2, \sigma_2^2)\right]. \end{aligned}$$

Thus, the component indicators serve the practical purpose of facilitating computation, but their presence does not affect the joint posterior distribution of our parameters of interest.

The following complete posterior conditionals are obtained:

$$\mu_1|\theta_{-\mu_1}, y \sim N(D\mu_1 d\mu_1, D\mu_1)$$

where

$$D\mu_1 = \left(n_1/\sigma_1^2 + \underline{v}_1^{-1}\right)^{-1}, \quad d\mu_1 = \sum_i \tau_i y_i/\sigma_1^2 + \underline{v}_1^{-1}\underline{\mu}_1,$$

$n_1 \equiv \sum_i \tau_i$ denotes the number of observations "in" the first component of the mixture, and $n_2$ will be defined as $n_2 \equiv \sum_i (1-\tau_i) = n - n_1$. The complete conditional for $\mu_2$ follows similarly.

As for the conditional posterior distribution for the variance parameters,

$$\sigma_2^2|\theta_{-\sigma_2^2}, \{\tau_i\}, y \sim IG\left(n_2/2 + \underline{a}_2, \left[\underline{b}_2^{-1} + .5\sum_{i=1}^{n}(1-\tau_i)(y_i-\mu_2)^2\right]^{-1}\right),$$

and the complete conditional for $\sigma_1^2$ follows similarly.

Finally, for the component indicator variables, and component probability $P$,

- 

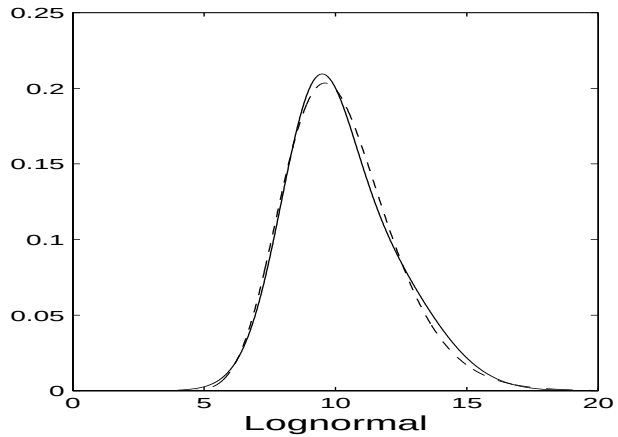
and

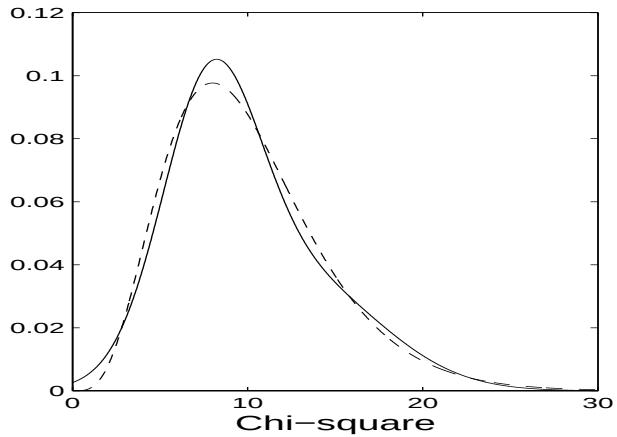
- 

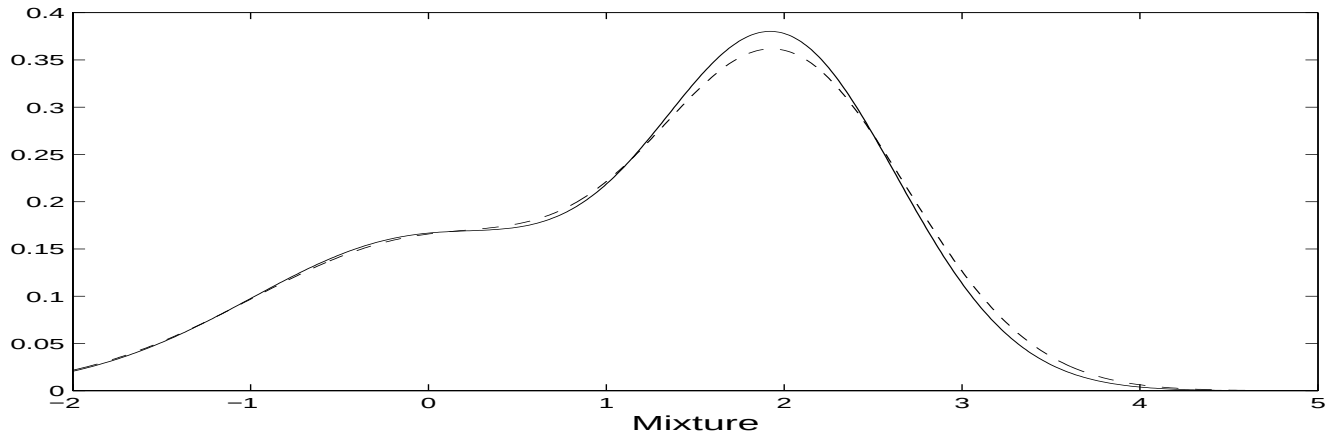
With these conditionals in hand, a Gibbs sampler can be implemented, noting, of course, that similar conditionals need to be obtained for $\mu_2$ and $\sigma_1^2$.

To illustrate the flexibility of the 2-component mixture model, we perform some generated data experiments. First, we generate:

- 2,000 observations from a lognormal distribution with parameters $\mu = \ln 10$ and $\sigma^2 = .04$.
- 5,000 observations from a Chi-square distribution with 10 degrees of freedom.
- (d) 3,000 observations from a two-component mixture model with $P = .4$, $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1^2 = 1$ and $\sigma_2^2 = .5$.


0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
Lognormal


0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
Chi−square


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−2
−1
0
1
2
3
4
5
Mixture

# A Regression Model with More than 2 Components

Consider the general set-up for a regression model using $G$ Normal mixture components:

- 

In this model we allow each mixture component to possess its own variance parameter, $\sigma_g$, and set of regression parameters, $\beta_g$.

This level of generality is not required - if desired, we could restrict some of these parameters to be constant across the mixture components.

For the purposes of computation, consider augmenting this model with a set of component label vectors, $\{z_i\}_{i=1}^n$ where

$$z_i = [z_{1i}\ z_{2i}\ \cdots\ z_{Gi}],$$

and $z_{gi} = 1$ implies that the $i^{th}$ individual is "drawn from" the $g^{th}$ component of the mixture.

To complete the augmentation step, we add a Multinomial prior (multivariate generalization of a Binomial) for the component label vector $z_i$ that depends on a vector of component probabilities $\pi$, and then specify a Dirichlet prior (multivariate generalization of the beta) for $\pi$.

The following priors are also employed:

$$\begin{aligned} \beta_g &\overset{ind}{\sim} N(\beta_{0g}, V_{\beta_g}), \quad g = 1, 2, \cdots, G \\ \sigma_g^2 &\overset{ind}{\sim} IG(a_g, b_g), \quad g = 1, 2, \cdots, G \end{aligned}$$

If we condition on the values of the component indicator variables, the conditional likelihood function can be expressed as

$$L(\theta) = \prod_{i=1}^{n} \left[\phi(y_i; x_i\beta_1, \sigma_1^2)\right]^{z_{1i}} \left[\phi(y_i; x_i\beta_2, \sigma_2^2)\right]^{z_{2i}} \cdots \left[\phi(y_i; x_i\beta_G, \sigma_G^2)\right]^{z_{Gi}}.$$

As stated, we add the following priors for the component indicators and component probabilities:

- 

Note that, taking the conditional likelihood and integrating out the component indicators gives an unconditional likelihood equivalent to our original model.

The augmented posterior density $p(\{\beta_g, \sigma_g^2, \pi_g\}_{g=1}^G, \{z_i\}_{i=1}^n | y)$ is proportional to the product of the augmented likelihood, the Multinomial and Beta priors, and the given priors for the regression and variance parameters.

It follows that the following complete posterior conditionals can be obtained:

$$\beta_g | \theta_{-\beta_g}, y \overset{ind}{\sim} N(D_{\beta_g} d_{\beta_g}, D_{\beta_g}), \quad g = 1, 2, \cdots G$$

where

$$D_{\beta_g} = \left[ (\sum_i z_{gi} x_i' x_i)/\sigma_g^2 + V_{\beta_g}^{-1} \right]^{-1}, d_{\beta_g} = (\sum_i z_{gi} x_i' y_i)/\sigma_g^2 + V_{\beta_g}^{-1} \beta_{0g}.$$

As for the variance parameters within each component,

$$\sigma_g^2|\theta_{-\sigma_g^2}, y \overset{ind}{\sim} IG\left(n_g/2 + a_g, \left[b_g^{-1} + (1/2)\sum_i z_{gi}(y_i - x_i\beta_g)^2\right]^{-1}\right) \quad g = 1, 2, \cdots G,$$

where $n_g \equiv \sum_{i=1}^N z_{gi}$ denotes the number of observations in the $g^{th}$ component of the mixture.

Finally,

$$z_i|\theta_{-z_i}, y \overset{ind}{\sim} M\left(1, \left[\frac{\pi_1\phi(y_i; x_i\beta_1, \sigma_1^2)}{\sum_{g=1}^G \pi_g\phi(y_i; x_i\beta_g, \sigma_g^2)} \; \frac{\pi_2\phi(y_i; x_i\beta_2, \sigma_2^2)}{\sum_{g=1}^G \pi_g\phi(y_i; x_i\beta_g, \sigma_g^2)} \right.\right.$$
$$\left.\left. \cdots \frac{\pi_G\phi(y_i; x_i\beta_G, \sigma_G^2)}{\sum_{g=1}^G \pi_g\phi(y_i; x_i\beta_g, \sigma_g^2)}\right]'\right),$$

and

- 

Fitting this model requires algorithms for drawing from the multinomial (a multivariate generalization of the binomial) and Dirichlet (a multivariate generalization of the beta) densities.

This is reasonably straight-forward: a Dirichlet draw can be obtained from a series of Beta draws, and likewise, a multinomial draw can be obtained from a series of binomial draws (I have code for doing this if you require it).

# *Skew-Normal Models*

Suppose that your error terms seem to be characterized by skew, but not multimodality, and you seek a more parsimonious alternative than the finite mixture approach. To this end, you consider a model of the form:

- 

Thus, $z$ has a half-Normal distribution.

We seek to answer the following questions related to this model:

(a) For $y$, a scalar generated from the above specification, derive the mixture density $p(y|x, \beta, \delta, \sigma^2)$. Comment on the role of $\delta$ in this conditional distribution.

(b) Let $\beta^* = [\beta' \ \delta]'$. Employing priors of the form $\sigma^2 \sim IG(a, b)$ and $\beta^* \sim N(0, V_\beta)$, derive a posterior simulator for fitting this regression model.

(a) For ease of exposition, let us drop the conditioning on $\beta$, $\sigma^2$ and $\delta$ in our notation and leave this implicit. We note

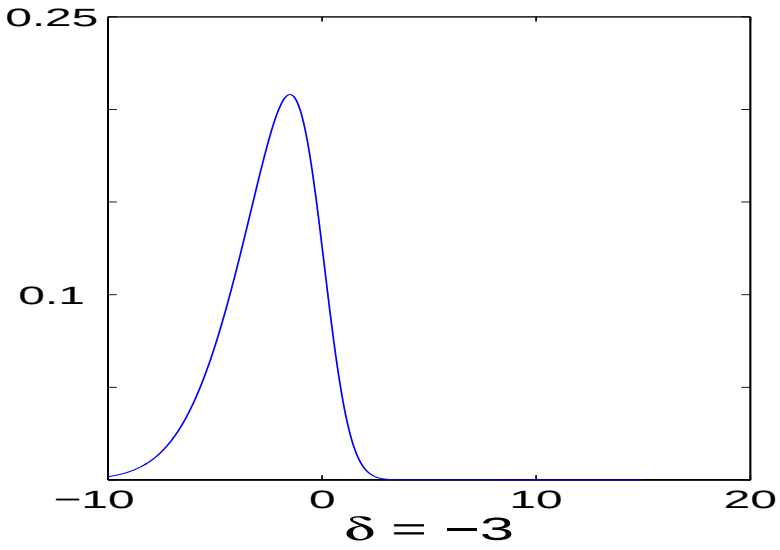
- 

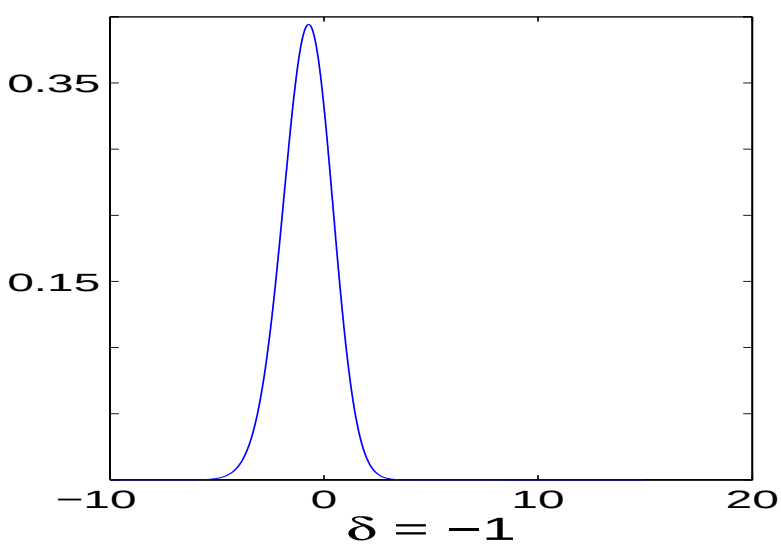
The density above is know n as a skew-Normal distribution and is sometimes written as $y \sim SN(x\beta, \sigma^2 + \delta^2, \delta/\sigma)$.

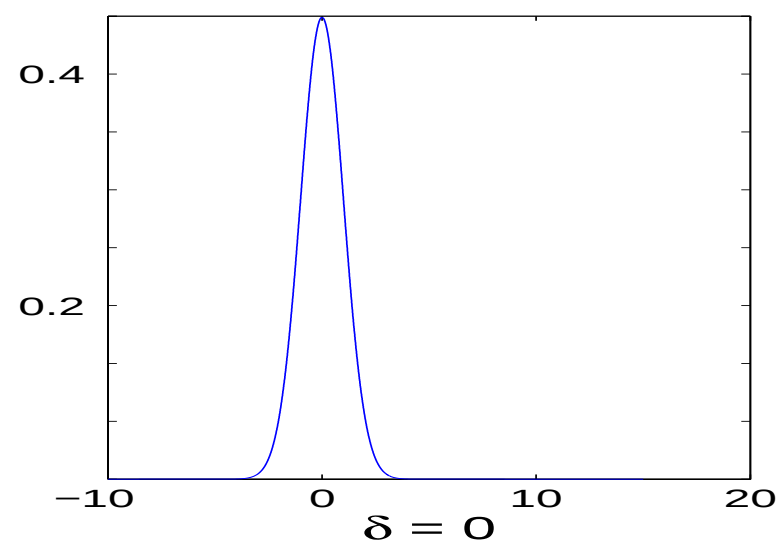
The parameter $\delta$ acts as a skewness parameter, and specifically,

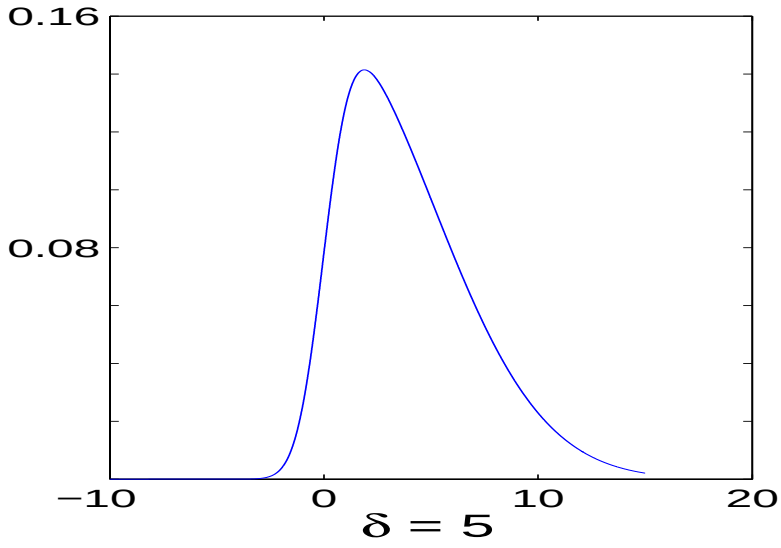
- 
- 
- When $\delta = 0$, the density is symmetric and we obtain $y \sim N(x\beta, \sigma^2)$.

On the following page, we provide plots of the skew-Normal density across different values of $\delta$ when $\sigma^2 = 1$ and $x\beta = 0$.


0.25
0.1
−10
0
10
20
δ = −3


0.35
0.15
−10
0
10
20
δ = −1


0.4
0.2
−10
0
10
20
δ = 0


0.16
0.08
−10
0
10
20
δ = 5

(b) For our posterior simulator, we make use of data augmentation and include $z = [z_1 \; z_2 \; \cdots \; z_n]'$ in the posterior distribution.

Before presenting these posterior conditionals, we first observe:

$$\begin{aligned} p(z, \beta^*, \sigma^2|y) &\propto p(\beta^*)p(\sigma^2)p(y, z|\beta^*, \sigma^2) \\ &\propto p(\beta^*)p(\sigma^2)\prod_{i=1}^{n}\phi(y_i; x_i\beta + z_i\delta, \sigma^2)\exp\left(-\frac{1}{2}z_i^2\right)I(z_i > 0). \end{aligned}$$

It follows that

$$z_i|\theta_{-z_i}, y \propto \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i\beta - z_i\delta)^2\right) \exp\left(-\frac{1}{2}z_i^2\right) I(z_i > 0).$$

Completing the square on $z_i$, and noting that $z_i$ is truncated at zero, we obtain

$$z_i|\theta_{-z_i}, y \overset{ind}{\sim} TN_{(0,\infty)}\left(\frac{\delta(y_i - x_i\beta)}{\sigma^2 + \delta^2}, \frac{\sigma^2}{\sigma^2 + \delta^2}\right), \ i = 1, 2, \cdots, n.$$

.

Let

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

and $W = [X \ z]$. With this notation, the posterior conditional for $\beta^*$ is of the form

$$\beta^* | \theta_{-\beta^*}, y \sim N(D_\beta d_\beta, D_\beta),$$

where

$$D_\beta = (W'W/\sigma^2 + V_\beta^{-1})^{-1}, \quad d_\beta = W'y/\sigma^2.$$

Finally,

$$\sigma^2 | \theta_{-\sigma^2}, y \sim IG\left(\frac{n}{2} + a, \left[b^{-1} + (1/2)(y - W\beta^*)'(y - W\beta^*)\right]^{-1}\right).$$

# Skew-Normal and Nonparametric Wage Estimates