

Parametric and nonparametric mixed normal regression models

Murray Aitkin

Education Statistics Services Institute

Washington DC

`MAitkin@air.org`

and Department of Statistics

University of Newcastle-upon-Tyne, UK

`Murray.Aitkin@ncl.ac.uk`

Summary

Normal mixture regression models are fitted to Brownlee's stackloss data. The t -distribution is a parametric (gamma) mixture on the variance; the overdispersed normal is a finite (nonparametric) mixture on the mean.

The two analyses lead to very different conclusions.

Models for overdispersion and outliers

We model the presence of *outliers* in regression models by incorporating an unobserved *random effect* in the model. For the random effect model for the normal mean, the conjugate distribution is also

For eg
normal model is observed:
 $i267.2(T53)Tj/T41Tf0-0.12288260$

Finite mixtures have been used to represent outliers by adding a second component distribution which contains the outliers:

$$Y | Z \sim (1 - Z) * N(\mu, \phi^2) + Z * N(\beta' \mathbf{x}, \sigma^2)$$
$$Z \sim b(1, \pi)$$

marginally, where π is close to 1.

This model can be extended to K components for multiple outliers, with one “good” component and $K - 1$ “outlier” components (Aitkin and Tunnicliffe Wilson 1980).

An attractive parametric alternative to finite mixtures is to introduce the random effect into the *variance*.

Let $Y|Z \sim N(\mu, \sigma^2/Z)$ with Z having a conjugate gamma $(r/2, 2/r)$ distribution with mean 1. The marginal distribution of Y is then

$$\begin{aligned}
 m(y) &= \frac{(r/2)^{r/2}}{\sqrt{2\pi}\sigma\Gamma(r/2)} \int z^{\frac{r-1}{2}} \exp \left\{ -z \left[\frac{r}{2} + \frac{(y-\mu)^2}{2\sigma^2} \right] \right\} dz \\
 &= \frac{(r/2)^{r/2}\Gamma(\frac{r+1}{2})}{\sqrt{2\pi}\sigma\Gamma(r/2)} \left[\frac{r}{2} + \frac{(y-\mu)^2}{2\sigma^2} \right]^{-\frac{(r+1)}{2}} \\
 &= \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi r}\sigma\Gamma(r/2)} \left[1 + \frac{(y-\mu)^2}{r\sigma^2} \right]^{-\frac{(r+1)}{2}}
 \end{aligned}$$

which is a t -distribution of $(y - \mu)/\sigma$ with r degrees of freedom. Thus Y has mean μ (provided $r > 1$) and variance $r\sigma^2/(r - 2)$ (provided $r > 2$).

The model can be fitted conveniently by iterative weighted least squares using an EM algorithm; the weights w_i are related to the residuals from the normal model:

$$w_i = \left[1 + \frac{(y_i - \mu_i)^2}{r\sigma^2} \right]^{-1} .$$

The weighting achieves a smooth accommodation of large outliers from the model - their influence on the parameter estimates is reduced, giving a form of *robust* estimation.

The *df* parameter r can be estimated, or it can be varied over a grid to examine its effect on the analysis.

Example - the Brownlee stack-loss data

We reproduce here the analysis of Brownlee's (1965) stack-loss data from Lange, Little and Taylor (1989).

The data consist of 21 observations on stack-loss y (the loss of acid through the stack) in a chemical plant for the conversion of ammonia to nitric acid, with three explanatory variables: air flow x_1 , cooling water inlet temperature x_2 and acid concentration x_3 . The stack-loss data have been analysed many times for outliers: the observations numbered 1,3,4 and 21 have repeatedly been identified as outliers.

Table 1: Stackloss data

y	x_1	x_2	x_3	y	x_1	x_2	x_3	y	x_1	x_2	x_3
42	80	27	89	20	62	24	93	8	50	18	89
37	80	27	88	15	58	23	87	7	50	18	86
37	75	25	90	14	58	18	80	8	50	19	72
28	62	24	87	14	58	18	89	8	50	19	79
18	62	22	87	13	58	17	88	9	50	20	80
18	62	23	87	11	58	18	82	15	56	20	82
19	62	24	93	12	58	19	93	15	70	20	91

Brownlee used the data to illustrate the desk-calculator computations for least squares fitting of the three-variable regression model.

For the three-explanatory-variable normal model the value of the “disparity” $-2 \log L_{max}$ is

$$n(1 + \log(2\pi) + \log(RSS/n)) = 104.58$$

. Parameter estimates are given below and in the table for $r = \infty$.

	estimate	s.e.	parameter
1	-39.92	11.90	1
2	0.716	0.135	X1
3	1.295	0.368	X2
4	-0.152	0.156	X3
scale parameter			2.918

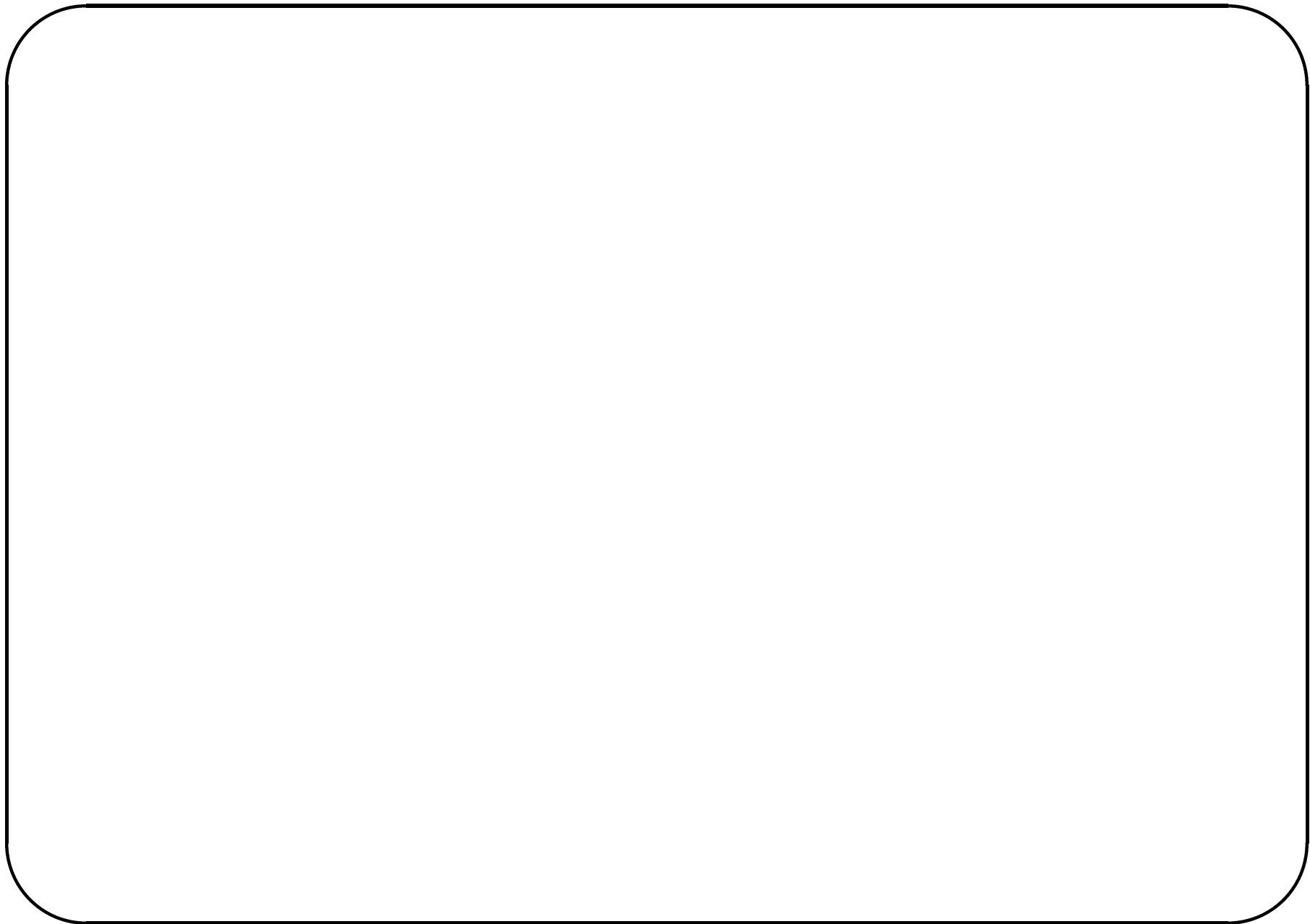


Table 2: Stackloss t-regression estimates

r	1	x_1	x_2	x_3	σ	disparity
∞	-39.9	0.716	1.30	-0.152	2.92	104.58
10	-40.7	0.793	1.03	-0.133	2.56	104.12
6	-40.7	0.835	0.876	-0.124	2.30	103.56
5	-40.5	0.848	0.817	-0.120	2.19	103.27
4	-40.1	0.857	0.746	-0.115	2.03	102.85
3	-39.1	0.854	0.657	-0.104	1.76	102.14
2	-38.1	0.848	0.557	-0.089	1.34	100.63
1.1	-38.4	0.852	0.491	-0.071	0.93	99.14
1	-38.6	0.852	0.489	-0.068	0.88	99.16
0.5	-40.8	0.840	0.536	-0.044	0.38	101.09

The likelihood is fairly flat for large r but more peaked near \hat{r} – the change in disparity between the normal model at $r = \infty$ and that at $\hat{r} = 1.1$ is 5.44. The disparity is not monotone for small r – it increases as r decreases, from 99.14 at $r = 1.1$ to 101.91 at $r = 0.40$, and then decreases again with further decrease in r . With decreasing r , x_1 becomes more important and x_2 and x_3 less important. We now examine the weights for $r = 1.1$.

0.032	0.914	0.028	0.014	0.491	0.301	0.762
0.817	0.451	0.987	0.771	0.997	0.096	0.250
0.342	0.982	0.882	0.979	0.643	0.234	0.010

The weights on observations 1,2,4 and 21 are very low, below 0.033, and that on observation 13 is 0.096. An issue of concern is that the total weight may be substantially less than n - here it is 10.98.

Nearly half the data have been “lost” in the downweighting, and the parameter estimates become more strongly dependent on the remaining observations with high weights, particularly 2,7,8,10,12 and 17.

For smaller values of r these effects become extreme. The weights for $r = 0.5$ are:

0.003	0.954	0.002	0.001	0.047	0.023	0.066
1.000	0.036	1.000	0.310	0.846	0.008	0.018
0.053	0.995	0.703	0.791	0.145	0.024	0.001

The regression is now being determined only by observations 2, 8, 10, 12, 16, 17 and 18 – the total weight is only 7.03. The very small degrees of freedom given by \hat{r} raises questions about the validity of the t -distribution.

Finite mixture analysis of the stack-loss data

Aitkin and Tunnicliffe Wilson (1980) gave references to earlier work, and a detailed investigation of normal mixture models for outliers, with suspected outliers assigned to specific mixture components. These components had their own means, unrelated to the regression structure modelled for the “good” observations.

The overdispersion approach here (following Aitkin 1996) is different since the “outliers” still contribute to the estimation of the regression model parameters.

For all the mixture models, parameter estimates and the disparity are sensitive to starting values. For the two-component model the best estimates give a disparity of 100.39; parameter estimates are shown for this model, and for three- and four-component models, below. The mixture models are fitted without intercepts.

Table 3: Disparities for stackloss data

K	disparity
1	104.58
2	100.39
3	85.72
4	72.96

Table 4: Mixture fits for stackloss data

K	x_1	x_2	x_3	σ	K_(1)	K_(2)	K_(3)	K_(4)
1	0.716	1.30	-0.152	2.92	-39.9			
2	0.605	1.83	-0.280	1.38	-30.5	-36.2		
3	0.766	0.62	-0.061	1.03	-30.0	-37.2	-45.5	
4	0.874	0.91	0.033	0.51	-50.9	-55.5	-59.5	-67.4

Posterior probabilities of component membership are given for the same three models below.

Table 5: Posterior probabilities of component membership

i	K	2		3			4			
		1	2	1	2	3	1	2	3	4
1		0.999	0.001	1	0	0	0	1	0	0
2		0.000	1.000	0	1	0	0	0	1	0
3		1.000	0.000	1	0	0	0	1	0	0
4		1.000	0.000	1	0	0	1	0	0	0
5		0.002	0.998	0	1	0	0	0	1	0
6		0.000	1.000	0	1	0	0	0	1	0
7		0.000	1.000	0	1	0	0	0	1	0

Table 6: Posterior probabilities of component membership

i	K k	2		3			4			
		1	2	1	2	3	1	2	3	4
8		0.002	0.998	0	1	0	0	0	1	0
9		0.000	1.000	0	1	0	0	0	1	0
10		0.993	0.007	0	1	0	0	1	0	0
11		1.000	0.000	0	1	0	0	1	0	0
12		1.000	0.000	0	1	0	0	1	0	0
13		0.093	0.907	0	1	0	0	0	1	0
14		0.987	0.013	0	1	0	0	0	1	0

Table 7: Posterior probabilities of component membership

i	K k	2		3			4			
		1	2	1	2	3	1	2	3	4
15		1.000	0.000	0	1	0	0	1	0	0
16		0.971	0.029	0	1	0	0	1	0	0
17		0.000	1.000	0	1	0	0	1	0	0
18		0.009	0.991	0	1	0	0	1	0	0
19		0.002	0.998	0	1	0	0	1	0	0
20		0.910	0.090	0	1	0	0	1	0	0
21		0.000	1.000	0	0	1	0	0	0	1

The two-component model splits the observations into two groups, with each observation having posterior probability of at least 0.907 of falling in one component.

The three-component model isolates observations 1, 3 and 4 in the first component, and observation 21 in the third. (Observations 1,3,4 and 21 are those repeatedly identified as outliers in previous analyses.)

For the four-component model the MLE of the residual SD is 0.5, half a measurement unit. Aitkin and Tunnicliffe Wilson (1980) noted a similar phenomenon in their mixture modelling: with four components the residual standard deviation was so small that the normal density representation of the likelihood breaks down unless measurements can be recorded to another decimal place. This result seems unreasonable.

The four-component model splits the observations into four groups, each observation having posterior probability at least 0.999 of falling in one component. Observation 4 falls in the first component, and 21 in the fourth. The other observations are split in several long sequences between the second and third components. This result strongly suggests a real latent class model, with a “missing factor” from the regression model. The appearance of the successive observations 5-9 in the third component and observations 15-20 in the second component strongly suggests a five-day week effect.

We define a `week` factor with five levels: week 1 contains observations 1-4, week 2 observations 5-9, week 3 observations 10-14, week 4 observations 15-19 and week 5 observations 20 and 21. Adding `week` to the three-variable normal model with no mixture structure gives a disparity of 76.28 which is close to that for the four-component mixture model.

	estimate	s.e.	parameter
	0.371	0.128	X1
	0.884	0.592	X2
	-0.026	0.102	X3
	-10.88	2.01	WEEK(2)
	-10.38	4.10	WEEK(3)
	-13.05	3.81	WEEK(4)
	-11.80	3.34	WEEK(5)
	-12.01	15.34	K_(1)
sigma	1.488		
disparity	76.28		

The disparity change between the normal regression model and that including week is 28.30 on 4 *df*. Equivalently, the *F*-statistic for the contribution of week is 9.25 for $F_{4,13}$; these are both significant beyond the 0.1% level. This factor is important.

The importance of x_2 has decreased in this model and it appears that airflow x_1 is the only important variable when the weekly time sequence of the data is taken into account. The weeks after week 1 show a dramatically lower stack-loss, with week 4 lower than weeks 2, 3 and 5.

“Week” however is not an explanatory variable in any real sense: the need for this factor in the model is a strong indication that there are other very important omitted process variables which apparently changed their values in successive operating weeks, as did the variables x_1 and x_2 . (This analysis was repeated with six- and seven-day weeks, and for five-day weeks starting with different observations in the sequence, with totally negative results: there is no evidence at all for a week effect longer than five days, or of the sequence starting at a different day.)

Does the “week” model fully account for the outliers? We refit the t regression model with week included as an explanatory variable. Now the disparity *increases* as r decreases from ∞ to 6, showing that the normal model is the best-supported over this range. As r decreases further from 6 to 3, the disparity *decreases*, and goes to $-\infty$ as $r \rightarrow 3$. At $r = 3$ only eight observations have full weight 1, and all the others have weight zero. Since the model has eight parameters, it fits the data exactly with zero variance. This shows again the unexpected effects of the “downweighting” induced by the ML fitting of the model.

We conclude that the persistent identification of outliers in these data is a result of model mis-specification: the data were apparently consecutive daily observations of a five-day working week process, in which changes in the process were made at the ends of some working weeks. These changes were apparently in addition to the substantial changes in the level of airflow and water temperature between weeks. When the week is included as a factor in the model, the evidence for a mixture, or for outliers, disappears.

It is of interest that some listings of these data report the data as being daily observations, though this is not mentioned in Brownlee's original listing of the data. We are able to reconstruct the time sequence as a latent class because of the substantial changes in stack loss in successive weeks.