Estimation Techniques in Mixture of Regression Model

Subhrajyoty Roy (BS - 1613)

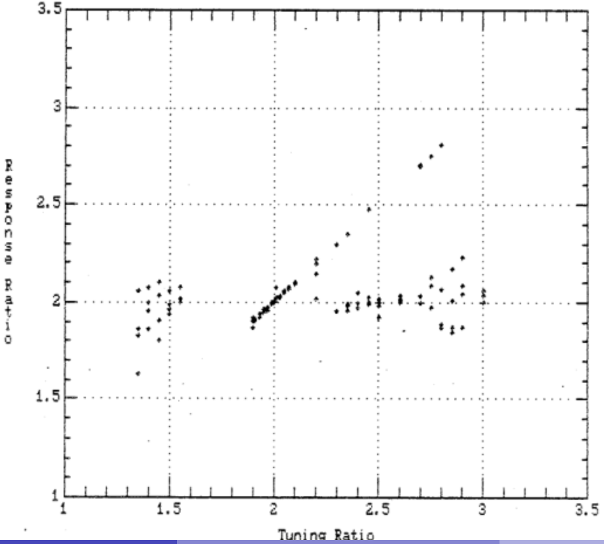
March 4, 2019

The Historical Example

Elizabeth Cohen, in 1980, wanted to study music perception of eastern and western musicians by an experiment conducted at Center for Research in Music and Acoustics (CRMA), Stanford University.

1. **Tuning Ratio:** The ratio of the harmonics to the fundamental tone (A major i.e. 440 cps). The harmonics is played to the musicians. It is chosen to be varied from 1.35 to 3.0
2. **Response Ratio:** The ratio of harmonics to the fundamental tone that the musician tunes in his/her instrument.

Figure: Scatterplot of Cohen's experiment



The Mathematical Model

Cohen's experiment was not the first. The model existed long before in **Econometrics**, introduced by Richard Quandt and James Ramsey, under the name of switching regression.

Let us consider the usual setup of regression where y_i's are the responses and x_i's are the predictor vectors with errors denoted by ϵ_i's. The regression coefficients are β_j vector. Consider new parameters π_j's for $j = 1, 2 \ldots k$, with $\sum_j \pi_j = 1$.

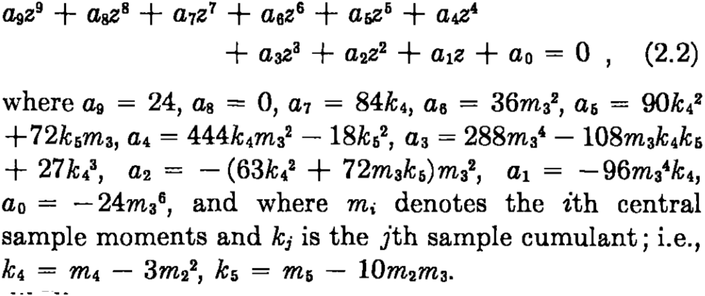
$$y_i = \begin{cases} {x'}_i \beta_1 + \epsilon_{i1} & \text{w.p. } \pi_1 \\ {x'}_i \beta_2 + \epsilon_{i2} & \text{w.p. } \pi_2 \\ \ldots & \ldots \\ {x'}_i \beta_k + \epsilon_{ik} & \text{w.p. } \pi_k \end{cases}$$

where $\epsilon_{ij} \sim N(0, \sigma_j^2)$ and these are independent.

The free parameters are $\theta = \left(\beta_1, \ldots \beta_k, \sigma_1^2, \ldots, \sigma_k^2, \pi_1, \ldots \pi_{k-1}\right)$

Why it has not been studied before?

MOM: Karl Pearson tried to estimate parameters of 2 sized mixture of Normal distributions $f(x) = \lambda N(\mu_1, \sigma_1^2) + (1-\lambda)N(\mu_2, \sigma_2^2)$ using method of moments technique in 1894. Equate first five population and sample central moments, and after some **messy** algebra, he showed that it requires the solution for the equation;



$$a_9z^9 + a_8z^8 + a_7z^7 + a_6z^6 + a_5z^5 + a_4z^4 + a_3z^3 + a_2z^2 + a_1z + a_0 = 0 \,, \quad (2.2)$$

where $a_9 = 24$, $a_8 = 0$, $a_7 = 84k_4$, $a_6 = 36m_3{}^2$, $a_5 = 90k_4{}^2 + 72k_5m_3$, $a_4 = 444k_4m_3{}^2 - 18k_5{}^2$, $a_3 = 288m_3{}^4 - 108m_3k_4k_5 + 27k_4{}^3$, $a_2 = -(63k_4{}^2 + 72m_3k_5)m_3{}^2$, $a_1 = -96m_3{}^4k_4$, $a_0 = -24m_3{}^6$, and where m_i denotes the ith central sample moments and k_j is the jth sample cumulant; i.e., $k_4 = m_4 - 3m_2{}^2$, $k_5 = m_5 - 10m_2m_3$.

Nobody ever dared to do the same for Regression Mixture!

Why it has not been studied before? (Contd.)

MLE: Consider the case with only $k = 2$.

$$
\begin{aligned}
\mathcal{L}(\lambda, \beta, \sigma) &= \prod_{i=1}^{n} \left[\sum_{j=1}^{2} \lambda_j (2\pi\sigma_j^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma_j^2} (y_i - x_i'\beta_j)^2 \right\} \right] \\
&\geq \prod_{i=1}^{n} \left[\sum_{j=1}^{2} (8\pi\sigma_j^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma_j^2} (y_i - x_i'\beta_j)^2 \right\} \right] \\
&= \prod_{i=2}^{n} \left[\sum_{j=1}^{2} (8\pi\sigma_j^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma_j^2} (y_i - x_i'\beta_j)^2 \right\} \right] \\
&\times \frac{1}{2\sqrt{2\pi}} \sum_{j=1}^{2} \frac{1}{\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2} (y_1 - x_1'\beta_j)^2 \right\} \\
&\geq \prod_{i=2}^{n} \frac{1}{2\sqrt{2\pi}\sigma_2} \exp\left\{ -\frac{1}{2\sigma_2^2} (y_i - x_i'\beta_2)^2 \right\} \times \frac{1}{2\sqrt{2\pi}\sigma_1}
\end{aligned}
$$

Day (1969) stated that *"Maximum likelihood clearly breaks down!"*

What is the way out then?

Even if both MOM and MLE approach does not show promises;

1. W.Y. Tan and W.C.Chang (1972) in economics.
2. Fryer J.G. and Robertson C.A (1972) in biological sciences.
3. Hosmer D. W. (1973) in biology.
4. Hosmer D.W. (1974) in communication science.

reported "good" estimates using iterative techniques based on maximum likelihood approach. However, everyone concluded that the choice of initial values is crucial.

Kiefer (1978), showed that MLE techniques have some hope.

De Veaux (1986 & 1989) thinks of using EM Algorithm in the problem.

Consistency of solution of Likelihood Equation

Theorem

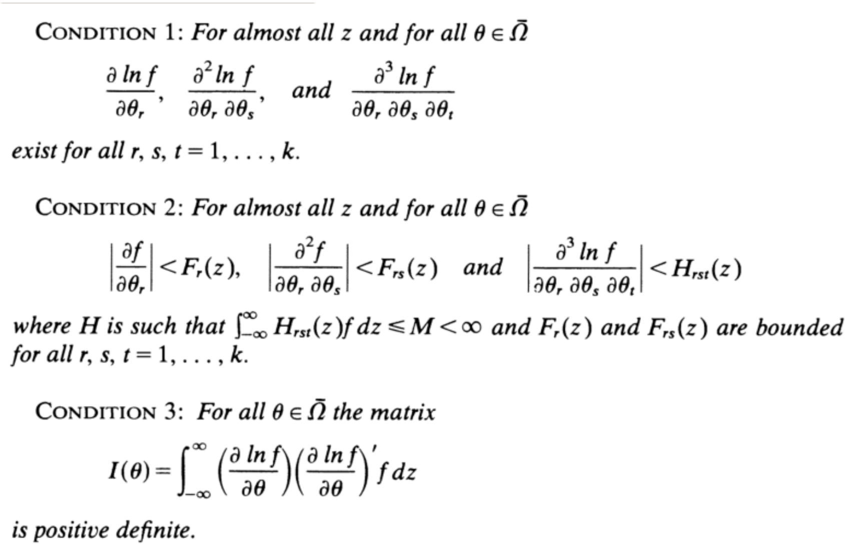
Let $f(z;\theta)$ be density, with unknown parameter vector θ residing in a closed parameter space Ω. $z_1, z_2, \ldots z_n$ are i.i.d. observations. Let θ_0 be true parameter value. Under some **regularity conditions**, *there exists θ_n, solution of the log-likelihood equation;*

$$\frac{\partial \log \mathcal{L}}{\partial \theta} = 0$$

such that $\sqrt{n}(\theta_n - \theta_0) \to N(0, I(\theta_0)^{-1})$.

The univariate version is proved by Harald Cramer, and the multivariate version is done by Tarone and Gruenhage.

The regularity conditions



Condition 1: *For almost all z and for all $\theta \in \bar{\Omega}$*

$$\frac{\partial \ln f}{\partial \theta_r}, \quad \frac{\partial^2 \ln f}{\partial \theta_r \, \partial \theta_s}, \quad \text{and} \quad \frac{\partial^3 \ln f}{\partial \theta_r \, \partial \theta_s \, \partial \theta_t}$$

exist for all $r, s, t = 1, \ldots, k$.

Condition 2: *For almost all z and for all $\theta \in \bar{\Omega}$*

$$\left|\frac{\partial f}{\partial \theta_r}\right| < F_r(z), \quad \left|\frac{\partial^2 f}{\partial \theta_r \, \partial \theta_s}\right| < F_{rs}(z) \quad \text{and} \quad \left|\frac{\partial^3 \ln f}{\partial \theta_r \, \partial \theta_s \, \partial \theta_t}\right| < H_{rst}(z)$$

where H is such that $\int_{-\infty}^{\infty} H_{rst}(z) f \, dz \leq M < \infty$ and $F_r(z)$ and $F_{rs}(z)$ are bounded for all $r, s, t = 1, \ldots, k$.

Condition 3: *For all $\theta \in \bar{\Omega}$ the matrix*

$$I(\theta) = \int_{-\infty}^{\infty} \left(\frac{\partial \ln f}{\partial \theta}\right)\left(\frac{\partial \ln f}{\partial \theta}\right)' f \, dz$$

is positive definite.

Kiefer's Conclusion

- Consider a closed region Ω containing the true parameter value θ_0 for mixture of regression setup, which does not contain the boundary cases $\lambda = 0, \lambda = 1, \sigma_1 = 0$ and $\sigma_2 = 0$, then the mixture density satisfies the regularity conditions.
- Kiefer verified this by simply differentiating the log likelihood.
- He noted that $I(\theta)$ is not positive definite if and only if y_i/x_i is a fixed constant for any $i = 1, 2, \ldots n$.
- Kiefer also noted that the above theorem does not show whether there are multiple consistent estimators, also how to find them. He exclaimed we can simply use Newton-Raphson method to find the solution of log likelihood equation.
- Quandt and Ramsey performed some simulation exercises using Newton Raphson method, and found the root to heavily depend on the initial value. They suspected the likelihood has numerous local maxima, and only a few of them are actually consistent.

A basic EM Algorithm

As soon as EM Algorithm emerges, people found that Mixture of Regression can be iteratively solved using EM Algorithm.

- $\ell(\theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j \phi_j(y_i|x_i) \right)$ where $\phi_j(\cdot|x_i)$ is normal density with mean $x_i'\beta_j$ and variance σ_j^2.
- Think of the complete data by introducing the latent variable w_{ij}, denoting the indicator whether y_i comes from component j.
- Given $\theta^{(t)}$, expected "complete" log-likelihood is;

$$Q(\theta; \theta^{(t)}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} -w_{ij}^{(t)} \left(\frac{y_i - x_i'\beta_j}{\sigma_j} \right)^2 \right) + \dots$$

where $w_{ij}^{(t)} = \frac{\pi_j^{(t)} \phi_j(y_i|x_i)}{\sum_{j=1}^{k} \pi_j^{(t)} \phi_j(y_i|x_i)}$.

- Maximizing $Q(\theta;\theta^{(t)})$ with respect to β_j is same as performing a weighted regression exercise, of $y_1, y_2, \dots y_n$ on $x_1, x_2, \dots x_n$ with weights $w_{1j}, w_{2j}, \dots w_{nj}$.
- The updation rule hence is;

$$\beta_j^{(t+1)} = \left(X' W_j X\right)^{-1} X' W_j y$$

$$\sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(t)} (y_i - x_i' \beta_j^{(t+1)})^2}{\sum_{i=1}^{n} w_{ij}^{(t)}}$$

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(t)}}{n}$$

where W_j is the diagonal matrix with elements $w_{1j}, w_{2j}, \dots w_{nj}$.

Other types of EM: ECM

- At the end of E step, we have $w_{ij}^{(t)}$.
- C-step classifies the sample points to the components. Create a partition $P^{(t+1)} = \left\{ P_1^{(t+1)}, P_2^{(t+1)}, \dots P_k^{(t+1)} \right\}$, where

$$P_j^{(t+1)} = \left\{ (x_i, y_i) : j = \arg_h \max w_{ih}^{(t)} \right\}$$

- The M-step is exactly same as before, with the same $w_{ij}^{(t)}$'s.
- **Q: Why bother doing C-step?**
  Allows dynamic number of components. If any of $P_j^{(t+1)}$ is empty or *has only one observation*, restart with $(k-1)$ components.

Other types of EM: ESM

- Same E-step as basic EM.
- Rather than classifying the observation to components based on highest posteriors, simulate the classification according to multinomial distribution. Therefore,

$$P_j^{(t+1)} = \{(x_i, y_i) : e_j \sim Multinomial(1, w_{i1}, w_{i2}, \dots w_{ik})\}$$

  where e_j is the vector of size k with 1 at position j and 0 at other places.
- The M-step is same as in basic EM.
- Allows dynamic component numbers, but is not too rigid about classifying.
- Emperically performs better when the initial estimates are random.

Is EM a good thing to do?

Theorem (De Veaux, 1986)

Let $x_i, y_i, i = 1, 2, \ldots n$ be independent samples from mixture of regression problem. Let $\hat{\theta}_n$ be a $\sqrt{n}$-consistent estimator of $\theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$. Let $\left\{\theta^{(t)}\right\}_{t \geq 0}$ be the EM algorithm sequence starting at $\hat{\theta}_n$. Then, $\theta_n^* = \lim_{t\to 0} \theta^{(t)}$ is asymptotically efficient for estimating θ.

Some important results of EM Algorithm

Theorem (Dempster, Laird, Rubin 1977)

EM Algorithm never decreases the log-likelihood.

$$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$$

.

Theorem (Wu, 1983)

If both "complete" and "incomplete" likelihoods are continuously differentiable w.r.t. θ and the parameter space θ is a compact subset of $\mathbb{R}^*$, then any sequence of $\theta^{(t)}$ produced by EM Algorithm converges to some stationary point θ^* of the log-likelihood.

De Veaux's Result on Convergence of EM Algorithm

Theorem (De Veaux, 1986)

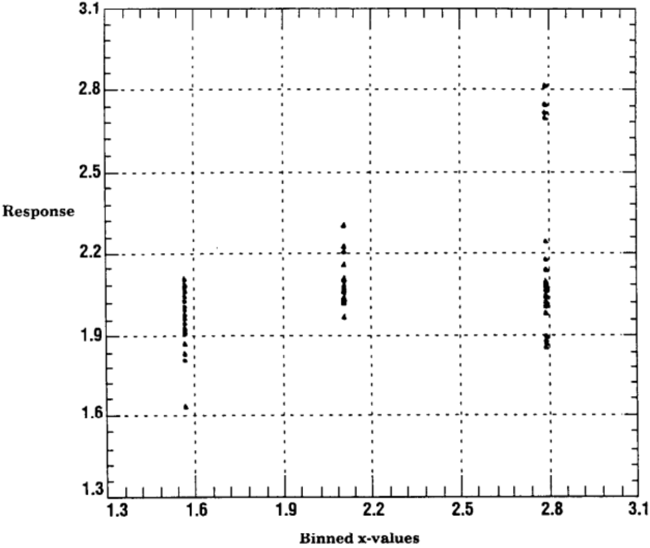
Let $x_i, y_i, i = 1, 2, \ldots n$ be independent samples from mixture of regression problem. Let $\hat{\theta}_n$ be a $\sqrt{n}$-consistent estimator of $\theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$. Let $\left\{\theta^{(t)}\right\}_{t \geq 0}$ be the EM algorithm sequence starting at $\hat{\theta}_n$. Then, $\theta_n^* = \lim_{t \to 0} \theta^{(t)}$ is asymptotically efficient for estimating θ.

Note: The proof do not require the initial estimates to be actually $\sqrt{n}$-consistent. $\hat{\theta}_n$ should be "close" so that we have a convex log-likelihood there.

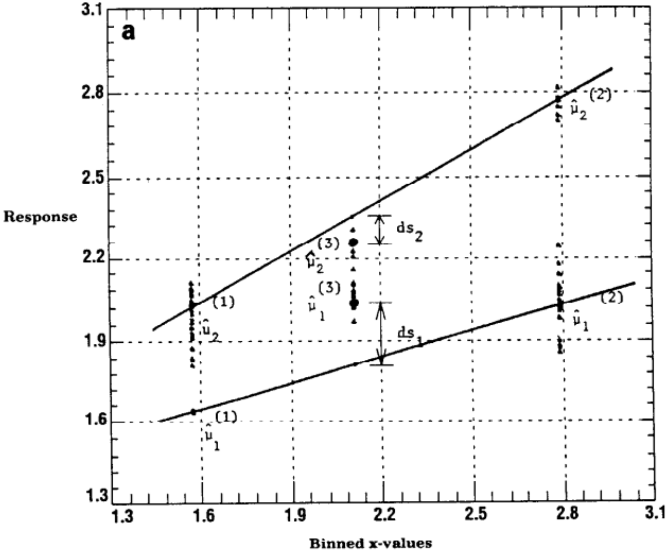
Good MM estimates to start with

1. Choose three non-overlapping bins I_1, I_2, I_3 on x-axis.
2. Assume each bin contains same number of points. (!important)
3. Obtain estimates for each bin $(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ as $(0.5, \bar{y} - s/2, \bar{y} + s/2, s/2, s/2)$. Assume, $\hat{\mu}_1 < \hat{\mu}_2$.
4. Obtain the pair of straight lines, the straight pair $(\hat{\alpha}_{1s}, \hat{\alpha}_{2s}, \hat{\beta}_{1s}, \hat{\beta}_{2s})$ and the crossed pair $(\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c})$.
5. Calculate total distances from pair of regression lines to the points $(\bar{x}_3, \hat{\mu}_1^{(3)})$ and $(\bar{x}_3, \hat{\mu}_2^{(3)})$.
6. Choose the minimizing one as initial estimates for beta coefficients.

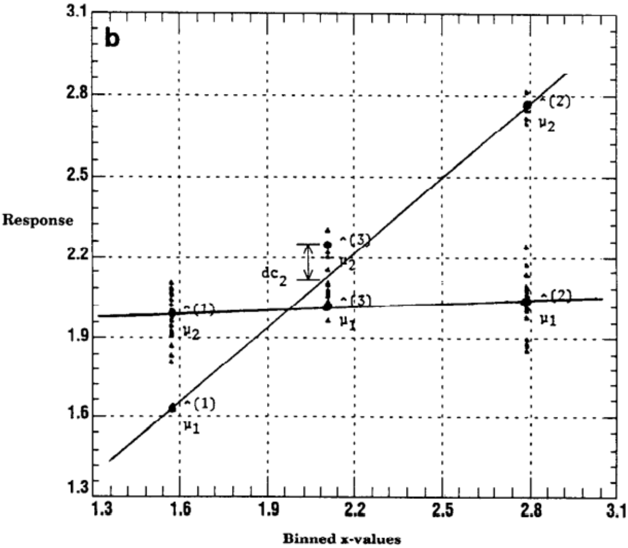
Binning Algorithm in Action (Step 1)



Binning Algorithm in Action (Step 2)



Binning Algorithm in Action (Step 3)



Does this binning Algorithm gives a "good" starting point?

No, in the way we described it.
Yes, if the estimates for each interval is $\sqrt{m}$-consistent, where m is the number of datapoints in each bin.

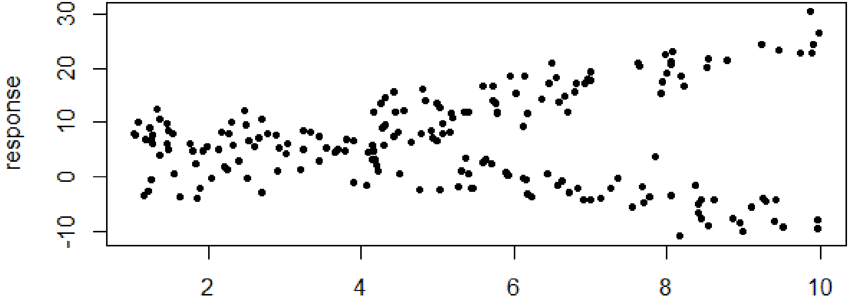
- Later, it has been proved that if $\lambda_0 = 0.5$ (Jones & MchLachlan, 1992 and Turner, 2000), then the binning procedure gives $\sqrt{n}$-consistent starting point for EM.
- Current Algorithms implemented in different software packages, tries with random values of λ as starting point, and performs the estimation several times.
- Standard errors of the estimates are obtained using Bootstrapping.

Some Simulations

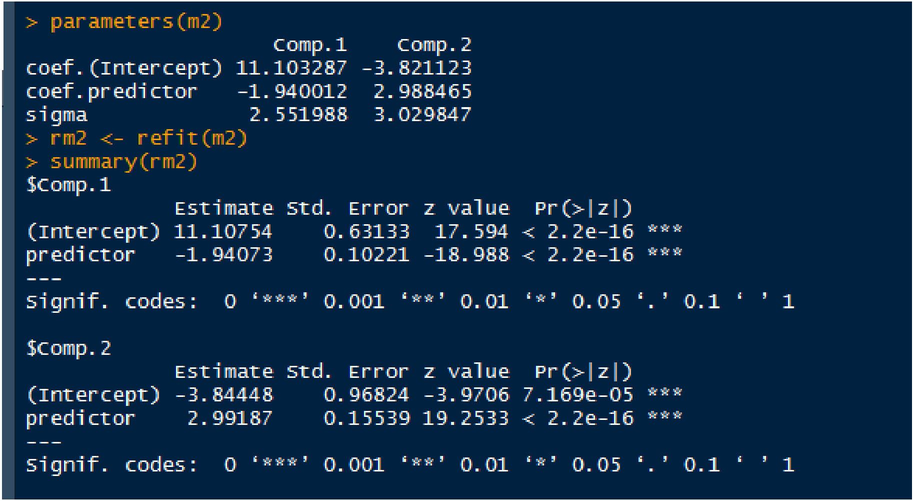
A dataframe was generated with a single predictor x and the response y modeled as;

$$y = \begin{cases} -4 + 3x + \epsilon & \text{for class 1} \\ 12 - 2x + \epsilon & \text{for class 2} \end{cases}$$

where $\epsilon \sim N(0, 9)$.

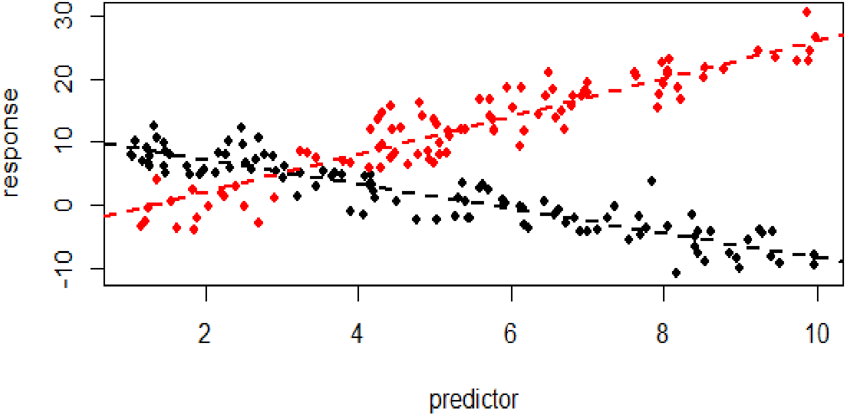


Some Simulations (Contd.)



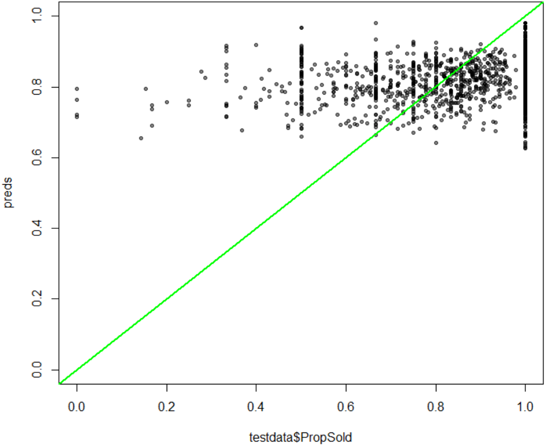
```
> parameters(m2)
                    Comp.1    Comp.2
coef.(Intercept) 11.103287 -3.821123
coef.predictor   -1.940012  2.988465
sigma             2.551988  3.029847
> rm2 <- refit(m2)
> summary(rm2)
$Comp.1
            Estimate Std. Error z value  Pr(>|z|)    
(Intercept) 11.10754    0.63133  17.594 < 2.2e-16 ***
predictor   -1.94073    0.10221 -18.988 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$Comp.2
            Estimate Std. Error z value  Pr(>|z|)    
(Intercept) -3.84448    0.96824 -3.9706 7.169e-05 ***
predictor    2.99187    0.15539 19.2533 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Some Simulations (Contd.)



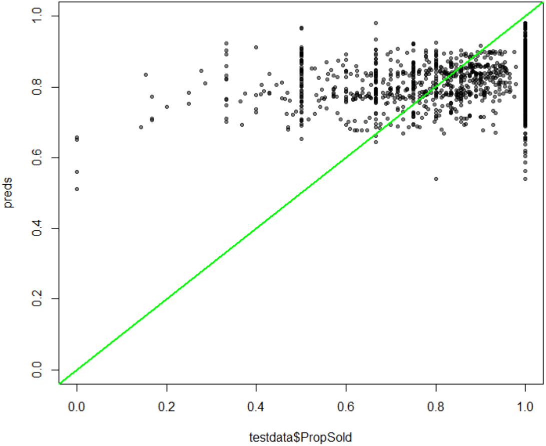
A Practical Example: Predicting Sold Status of Tea Lots

Figure: Predicted Probabilities vs Actual Probabilities for Logistic Regression



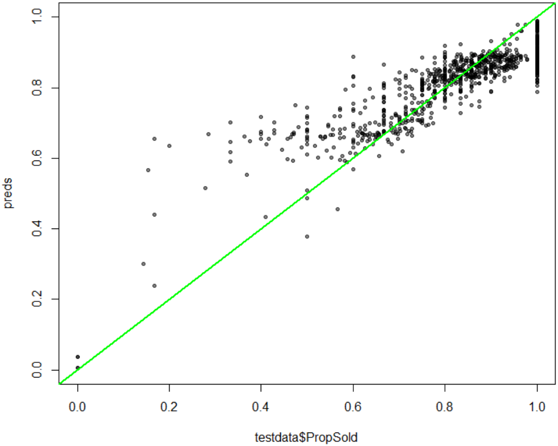
A Practical Example: Predicting Sold Status of Tea Lots

Figure: Predicted Probabilities vs Actual Probabilities for GAM



A Practical Example: Predicting Sold Status of Tea Lots

Figure: Predicted Probabilities vs Actual Probabilities for 3 component Mixture of Logistic Regression



References

1. *Mixtures of Linear Regression*, Richard De Veaux (1989) Computational Statistics & Data Analysis Vol - 8, pp 227-245.
2. *Parameter Estimation for a mixture of Linear Regressions*, Richard De Veaux, Technical Report 247, April 1986, Dept. of Statistics, Stanford University.
3. *Fitting mixtures of linear regressions*, Susana Faria & Gilda Soromenho, Journal of Statistical Computation and Simulation Vol.80, no. 2, February 2010, pp 201-225.
4. *Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model*, Nicholas M. Kiefer, Econometrica, Vol. 46, No. 2, March 1978, pp 427-434.
5. *Estimating Mixtures of Normal Distributions and Switching Regressions*, Richard E. Quandt & James B. Ramsey (1978), Journal of American Statistical Association, 73:364, pp 730-738.

THANK YOU