Predicting Regression Probability Distributions

with Imperfect Data

through Optimal Transformations

Jerome H. Friedman

Stanford University

MACHINE LEARNING / REGRESSION

$y = F(\mathbf{x}, \mathbf{z})$

$y =$ outcome variable

$\mathbf{x} = (x_1 \cdot \cdot\cdot, x_p)$ observed predictor variables

$\mathbf{z} = (z_1, z_2, \cdots)$ other variables

Goal: estimate $E\,[y \,|\, \mathbf{x}]$ given data $\{y_i, \mathbf{x}_i\}_{i=1}^{N}$

STATISTICAL MODEL

$$y = f(\mathbf{x}) + s(\mathbf{x}) \cdot \epsilon$$

$f(\mathbf{x}) = E\,[y \,|\, \mathbf{x}]$ location function

$s(\mathbf{x}) > 0$ scale function

$\varepsilon =$ random variable, $E\,[\varepsilon \,|\, \mathbf{x}] = 0$

Prediction: $\hat{y} = f(\mathbf{x})$

$s(\mathbf{x}) \cdot \epsilon =$ "irreducible error" (unavoidable)

REDUCIBLE ERROR

$$r(\mathbf{x}) = E \,|\, f(\mathbf{x}) - \hat{f}(\mathbf{x}) \,|$$

$f(\mathbf{x})$ = optimal location (target) function

$\hat{f}(\mathbf{x})$ = estimate based on training data & ML method

ML goal: methods to reduce $r(\mathbf{x})$

Statistics goal: methods to estimate $r(\mathbf{x})$

Prediction error (y) = Reducible + Irreducible

Usually: Irreducible $s(\mathbf{x}) >>$ Reducible $r(\mathbf{x})$

USUAL ASSUMPTIONS

$$s(\mathbf{x}) = s = \text{constant (homoscedasticity)}$$

$$\epsilon \sim N(0,1) \quad \text{(normality)}$$

Neither very likely

Tukey:

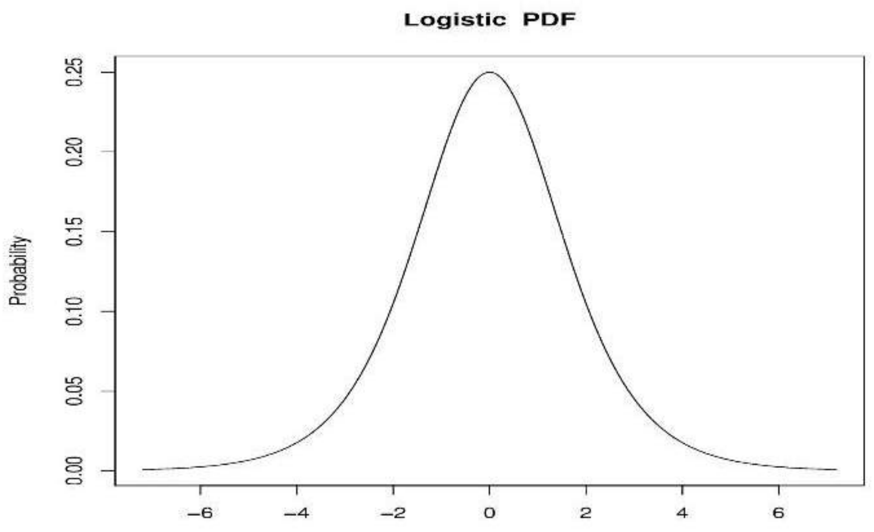
"small residuals $\simeq$ normal, larger have heavier tails."

LOGISTIC DISTRIBUTION

$$\epsilon \,|\, \mathbf{x} = (y - f(\mathbf{x}))/s(\mathbf{x})$$

$$\tilde{p}(\epsilon) = \frac{e^{-\epsilon}}{s(1+e^{-\epsilon})^2}$$

small $|\,\varepsilon\,| \sim$ normal, large $|\,\varepsilon\,| \sim$ exponential



Prediction: $\hat{y} = \hat{f}(\mathbf{x})$

$$\hat{f}(\mathbf{x}) = \arg\min_{f \in F} \sum_{i=1}^{N} [\varepsilon_i + 2\log(1 + e^{-\varepsilon_i})]$$

$$\varepsilon_i = (y_i - f(\mathbf{x}_i))/s(\mathbf{x}_i)$$

minimized at $f(\mathbf{x}_i) = y_i$ indep $s(\mathbf{x}_i)$

$1/s(\mathbf{x}_i) \sim$ "weight" for obs i

controls relative influence of i to fit

Using incorrect $s(\mathbf{x})$ to estimate $f(\mathbf{x})$

increases variance, not bias

assume $s(\mathbf{x}) =$ constant usually not too bad

ESTIMATE $\hat{s}(\mathbf{x})$

(1) Improve $\hat{f}(\mathbf{x})$ in high variance settings.

(2) Important inferential statistic:

(a) prediction interval $\sim$ accuracy of $\hat{y}$-prediction:

logistic: $IQR\,[\,y\,|\,\mathbf{x}] = 2\,s(\mathbf{x})/\log(3)$

(b) can affect decision

CENSORING

Data: $\{y_i, \mathbf{x}_i\}_1^N \rightarrow \{a_i, b_i, \mathbf{x}_i\}_1^N$

$$a_i \leq y_i \leq b_i$$

$a_i = b_i = y_i \Rightarrow y$-value known

$a_i = -\infty \Rightarrow$ censored below b_i

$b_i = \infty \Rightarrow$ censored above a_i

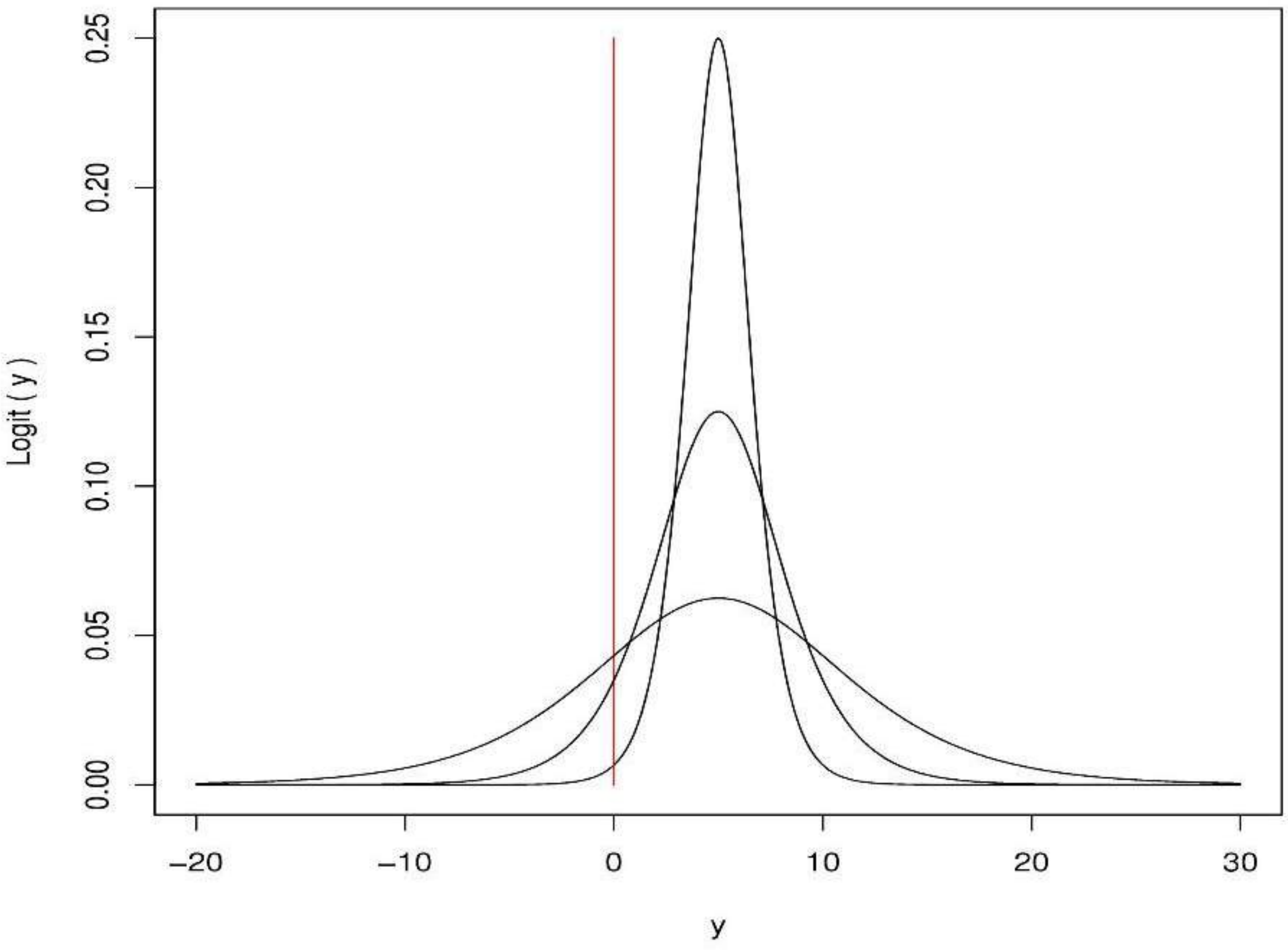
Otherwise: interval censored $[a_i, b_i]$

LIKELIHOOD

$$\mathsf{Pr}(a \leq y \leq b) = \frac{1}{1+e^{-(b-f)/s}} - \frac{1}{1+e^{-(a-f)/s}}$$

Depends strongly on *both* f *and* s

Need to estimate *both* $f(\mathbf{x})$ *and* $s(\mathbf{x})$


0.00
0.05
0.10
0.15
0.20
0.25
Logit (y)
−20
−10
0
10
20
30
y

EXERCISE

$$(\hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})) = \arg\min_{f,s \in F} \sum_{i=1}^{N} L(a_i, b_i, f(\mathbf{x}_i), s(\mathbf{x}_i))$$

$$L(a, b, f(\mathbf{x}), s(\mathbf{x})) = \log\left[\frac{1}{1 + \exp((f(\mathbf{x}) - a)/s(\mathbf{x}))} - \frac{1}{1 + \exp((f(x) - b)/s(\mathbf{x}))}\right]$$

GRADIENT BOOSTED TREE ENSEMBLES

Ann. Statist, **29**. 1189 – 1232 (2001)

$\hat{f}(\mathbf{x}) = \sum_{k=1}^{K_f} T_k^{(f)}(\mathbf{x})$

$\log(\hat{s}(\mathbf{x})) = \sum_{k=1}^{K_s} T_k^{(s)}(\mathbf{x})$

$T_k(\mathbf{x}) =$CART–tree $(\mathbf{x})$

ITERATIVE GRADIENT BOOSTING

Start: $\hat{s}(\mathbf{x}) = \text{constant}$

Loop {

$$\hat{f}(\mathbf{x}) = \text{tree–boost}\ [\ f(\mathbf{x}) \mid \hat{s}(\mathbf{x})\ ]$$

$$\log(\hat{s}(\mathbf{x})) = \text{tree–boost}\ \ [\log(s(\mathbf{x})) \mid \hat{f}(\mathbf{x})\ ]$$

}

Until no change

OPTIMAL TRANSFORMATIONS

$$g(y) = f(\mathbf{x}) + s(\mathbf{x}) \cdot \varepsilon$$

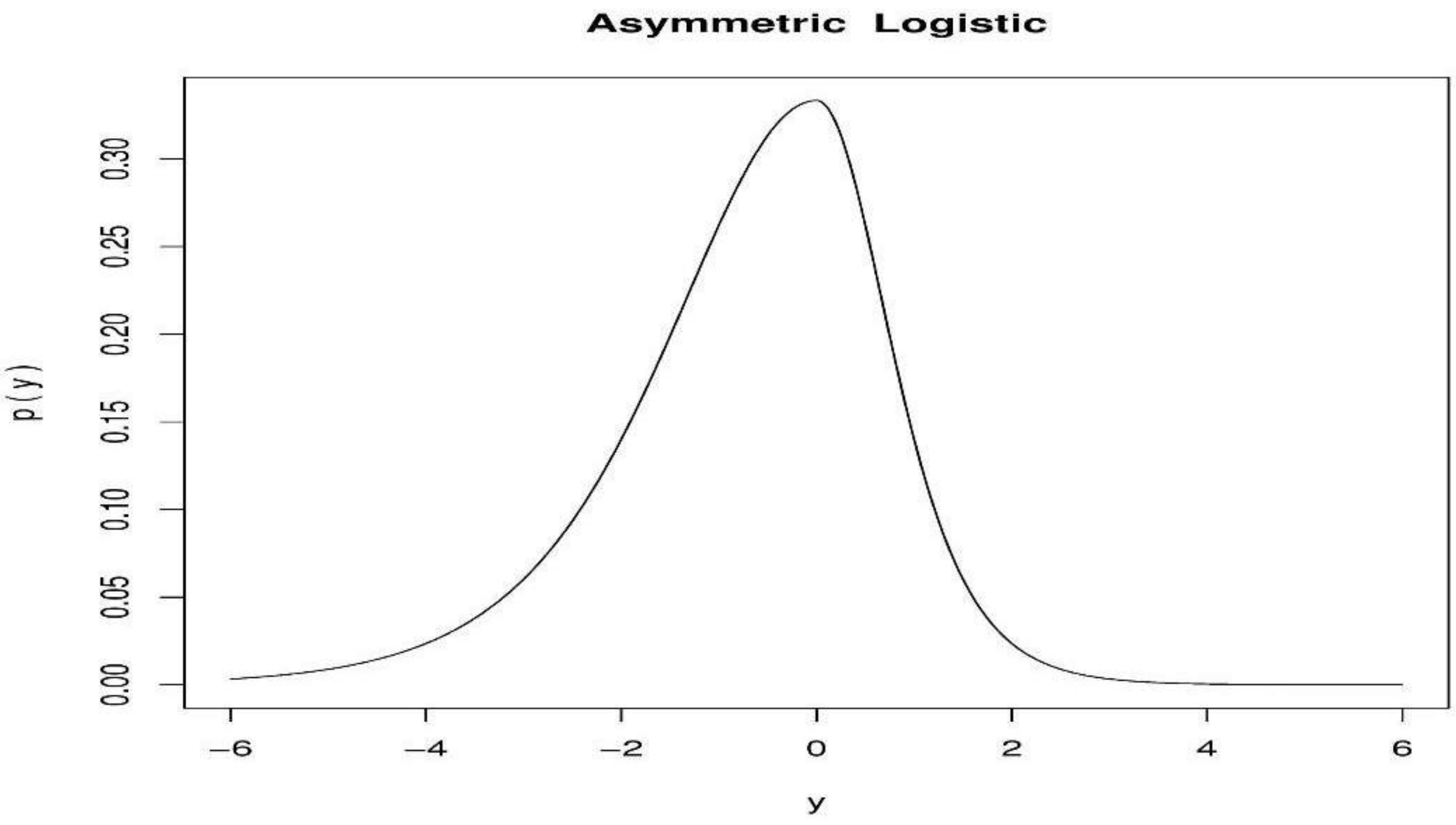
$g(y) =$ unknown monotonic function

$$[\hat{g}(y), \hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})] = \text{arg min}_{g,(f,s) \in F} \sum_{i=1}^{N} L[g(y_i), f(\mathbf{x}_i), s(\mathbf{x}_i)]$$

$$\frac{1}{N} \sum_{i=1}^{N} CDF(\hat{g}(y), \hat{f}(\mathbf{x}_i), \hat{s}(\mathbf{x}_i)) = CDF_y(y)$$

$$CDF(g(y), f(\mathbf{x}), s(\mathbf{x})) = \frac{1}{1+\text{exp}((f(\mathbf{x})-g(y))/s(\mathbf{x}))}$$

ASYMMETRY



ASYMMETRIC LOGISTIC DISTRIBUTION

$$p(z \mid f, s_l, s_u) = \frac{2}{s_l + s_u}\left[\frac{I(z \leq f)\ \mathsf{exp}((f-z)/s_l)}{(1+\mathsf{exp}((f-z)/s_l))^2} + \frac{I(z > f)\ \mathsf{exp}((f-z)/s_u)}{(1+\mathsf{exp}((f-z)/s_u))^2}\right]$$

$$g(y) = f(\mathbf{x}) + \eta, \quad \begin{cases} \eta = -s_l(\mathbf{x}) \cdot |\,\varepsilon\,|, & prob = s_l(\mathbf{x})/(s_l(\mathbf{x}) + s_u(\mathbf{x})) \\ \eta = +s_u(\mathbf{x}) \cdot |\,\varepsilon\,|, & prob = s_u(\mathbf{x})/(s_l(\mathbf{x}) + s_u(\mathbf{x})) \end{cases}$$

$f =$ mode, $s_l =$ lower scale, $s_u =$ upper scale

$$\varepsilon \sim L(0, 1)$$

ASYMMETRIC GRADIENT BOOSTING

Start: $\hat{s}_l(\mathbf{x}) = \hat{s}_u(\mathbf{x}) =$ constant

Loop {

$\hat{f}(\mathbf{x}) =$ tree–boost $f(\mathbf{x})$ given $\hat{s}_l(\mathbf{x})$ & $\hat{s}_u(\mathbf{x})$

$\widehat{\log(s_l(\mathbf{x}))} =$ tree–boost $s_l(\mathbf{x})$ given $\hat{f}(\mathbf{x})$ & $\hat{s}_u(\mathbf{x})$

$\widehat{\log(s_u(\mathbf{x}))} =$ tree–boost $s_u(\mathbf{x})$ given $\hat{f}(\mathbf{x})$ & $\hat{s}_l(\mathbf{x})$

} Until change $<$ threshold.

OPTIMAL TRANSFORMATIONS

$$g(y)\,|\,\mathbf{x} \sim \text{Logistic}\,(\hat{f}(\mathbf{x}), \hat{s}_l(\mathbf{x}), \hat{s}_u(\mathbf{x}))$$

$$[\hat{g}(y), \hat{f}(\mathbf{x}), \hat{s}_l(\mathbf{x}), \hat{s}_u(\mathbf{x})]$$

$$= \text{arg min}_{g,(f,s_l,s_u)\in F} \textstyle\sum_{i=1}^{N} L[g(y_i), f(\mathbf{x}_i), s_l(\mathbf{x}_i), s_u(\mathbf{x}_i)]$$

$$\tfrac{1}{N} \textstyle\sum_{i=1}^{N} CDF(\hat{g}(y), \hat{f}(\mathbf{x}_i), \hat{s}_l(\mathbf{x}_i), s_u(\mathbf{x}_i)) = CDF_y(y)$$

$$CDF(g(y), f(\mathbf{x}), s(\mathbf{x})) = \text{messy closed form}$$

DIAGNOSTICS

For $\mathbf{x}_i \in S$:

Model: $\mathsf{Pr}(\hat{g}(y) < z) = \frac{1}{|S|}\sum_{i\in S} CDF(z, \hat{f}(\mathbf{x}_i), \hat{s}_l(\mathbf{x}_i), \hat{s}_u(\mathbf{x}_i))$

Empirical: $\widehat{\mathsf{Pr}}(\hat{g}(y) < z) = \frac{1}{|S|}\sum_{i\in S} I(\hat{g}(y_i) \le z)$

Symmetric: $(g(y_i) - f(x_i))/s(x_i) \sim L(0,1)$

Asymmetric : $I(-)\ (\hat{g}(y_i) - \hat{f}(\mathbf{x}_i))/\hat{s}_l(\mathbf{x}_i)$

$$+ I(+)\ (\hat{g}(y_i) - \hat{f}(\mathbf{x}_i))\,/\hat{s}_u(\mathbf{x}_i) \sim L(0,1)$$

DATA SUBSETS S

$$S = \{i \,|\, r < \hat{f}(\mathbf{x}_i) \leq t \;\&\; u < \hat{s}(\mathbf{x}_i) \leq v\}.$$

$$S = \{i \,|\, r < \hat{f}(\mathbf{x}_i) \leq t \;\&\; u < \sqrt{\hat{s}_l(\mathbf{x}_i)\hat{s}_u(\mathbf{x}_i)} \leq v\}$$

$$S = \{i \,|\, r < \hat{f}(\mathbf{x}_i) \leq t \;\&\; u < \hat{s}_l(\mathbf{x}_i) \leq v\}.$$

$$S = \{i \,|\, r < \hat{f}(\mathbf{x}_i) \leq t \;\&\; u < \hat{s}_{ul}(\mathbf{x}_i) \leq v\}.$$

QUESTIONNAIRE DATA ($N = 8856$)

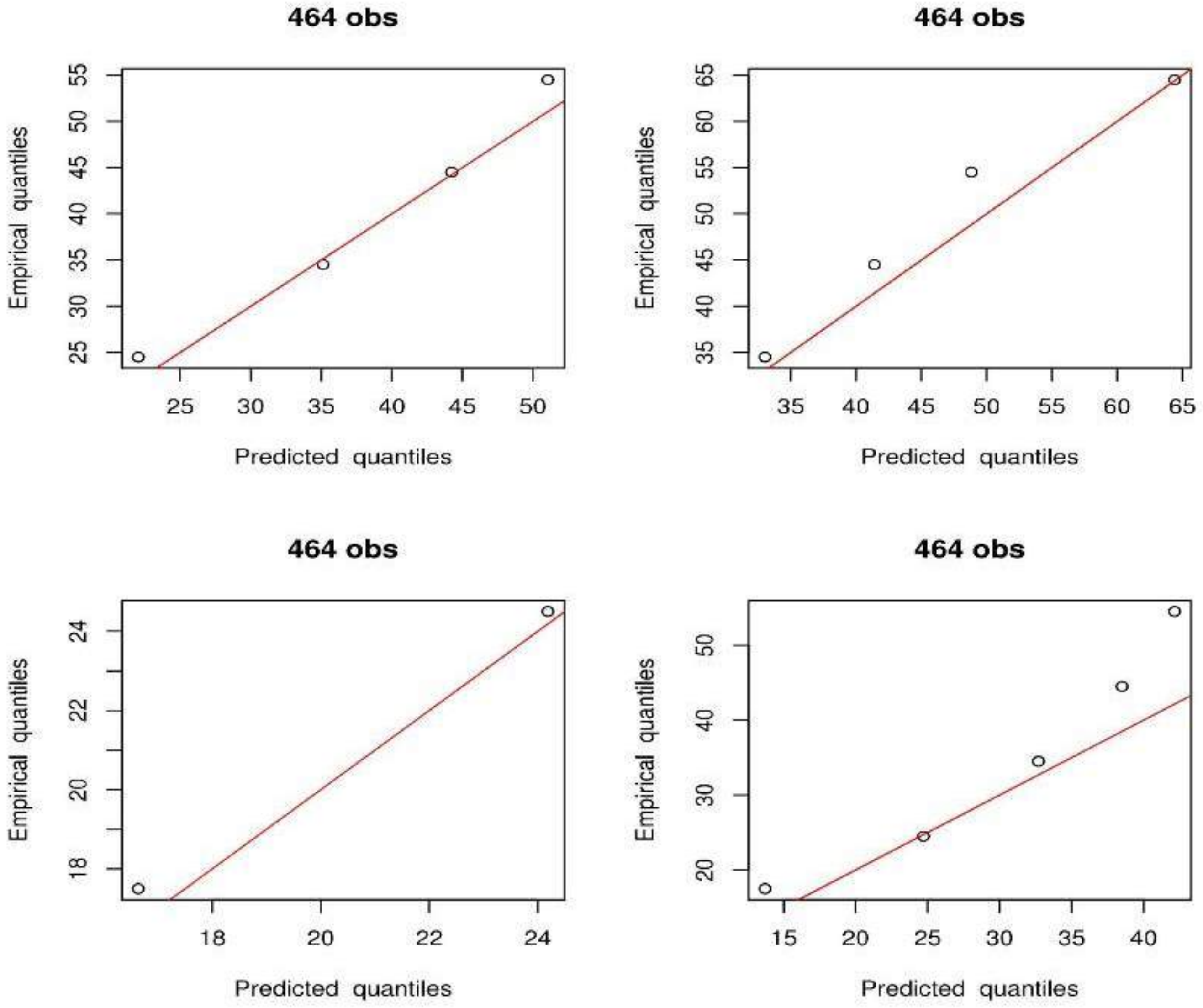
Bay Area Shopping Mall Customers

AGE $\in$

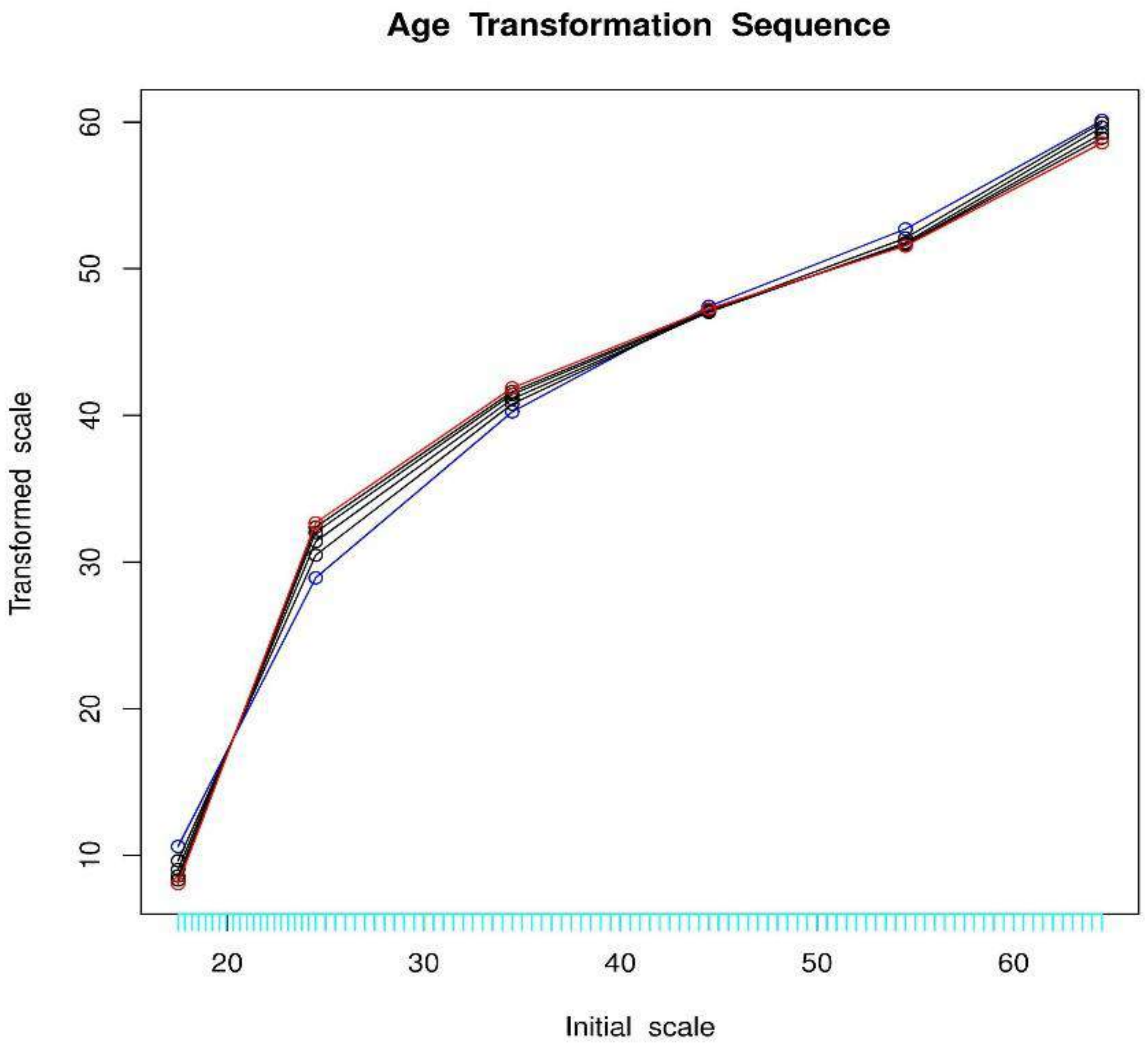
1	2	3	4	5	6	7
17 & under	18–24	25–34	35–44	45–54	55–64	65 & older

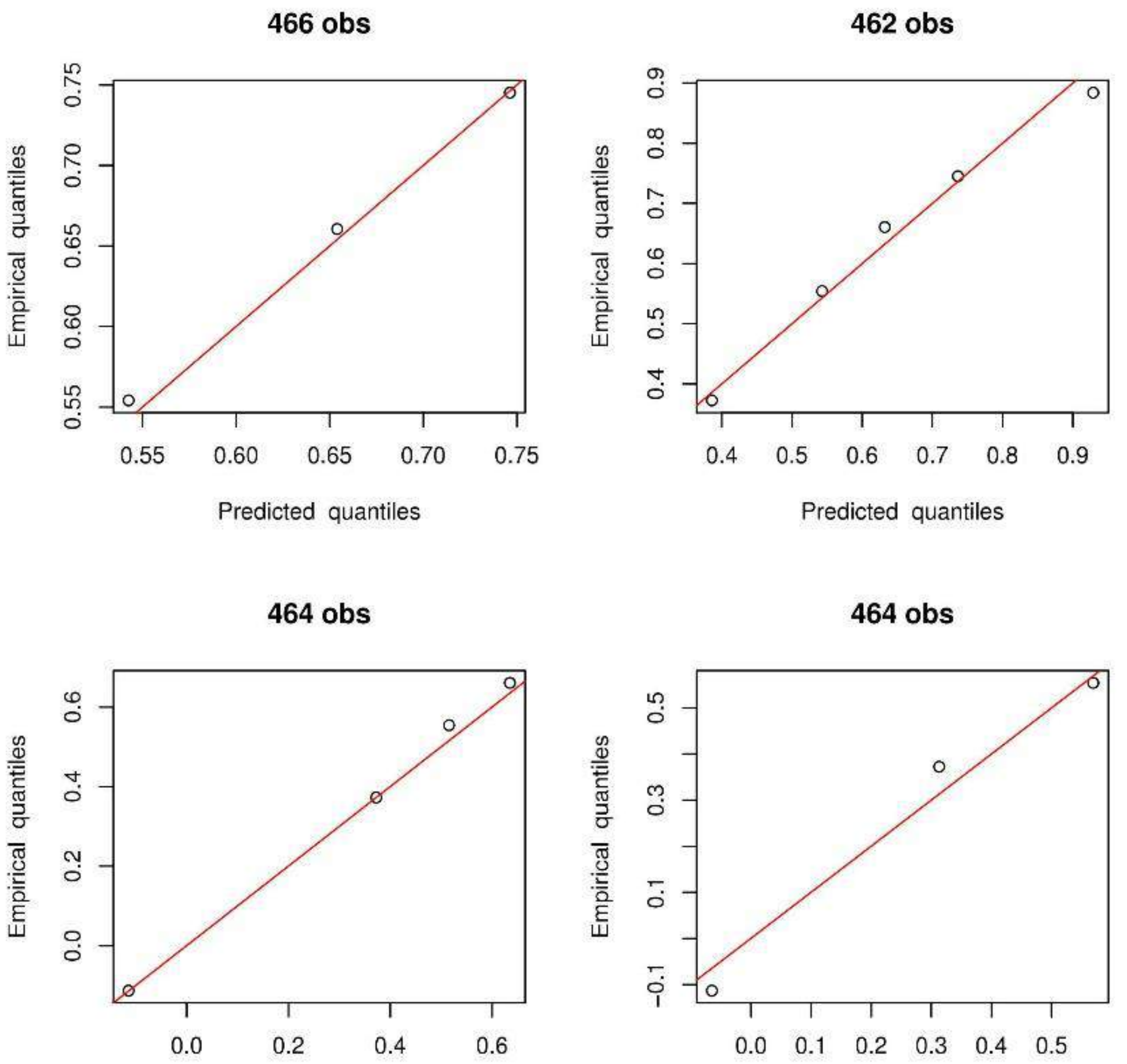
$\mathbf{x} =$

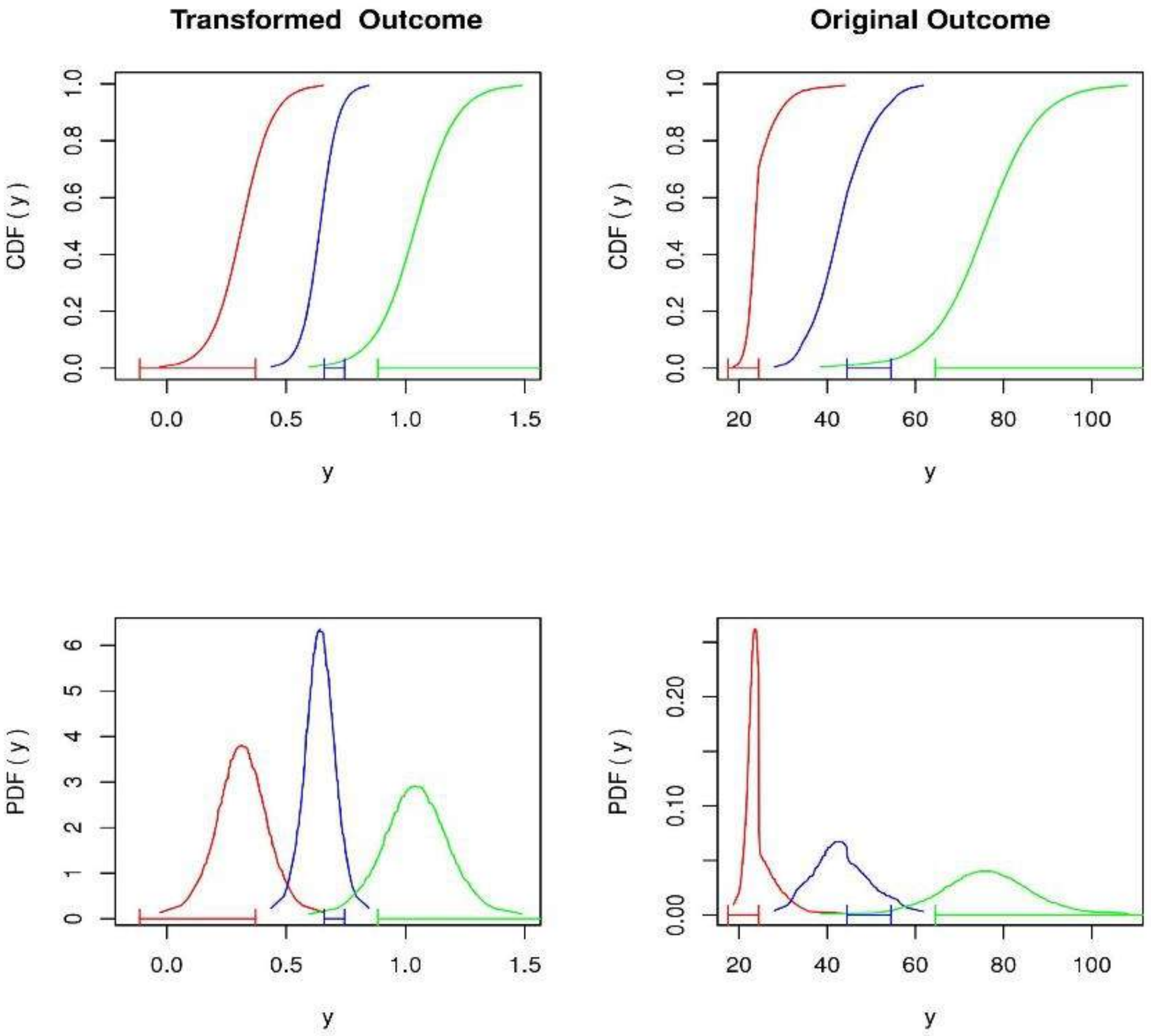
1	Occupation
2	Type of home
3	Gender
4	Martial status
5	Education
6	Annual income
7	Lived in Bay Area
8	Dual Incomes
9	Persons in household
10	Persons in household under 18
11	Householder status
12	Ethnic classification
13	Language



Q–Q plots of empirical versus predicted quantiles for untransformed Omnireg solution.


Age Transformation Sequence
Transformed scale
Initial scale
10
20
30
40
50
60
20
30
40
50
60


466 obs
462 obs
464 obs
464 obs
Empirical quantiles
Predicted quantiles
0.55
0.60
0.65
0.70
0.75
0.4
0.5
0.6
0.7
0.8
0.9
0.0
0.2
0.4
0.6
−0.1
0.1
0.3
0.5
0.0
0.1
0.2
0.3
0.4
0.5


Transformed Outcome
Original Outcome
CDF (y)
PDF (y)
y
0.0
0.5
1.0
1.5
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
20
40
60
80
100
0.00
0.10
0.20

MASHABLE ONLINE NEWS POPULARITY

(Irvine Repository)

$N = 39644$ news articles

25000 train, 7322 mod. sel., 7322 test

$y =$ popularity (number of shares on social media)

$\mathbf{x} = 59$ numerical attributes

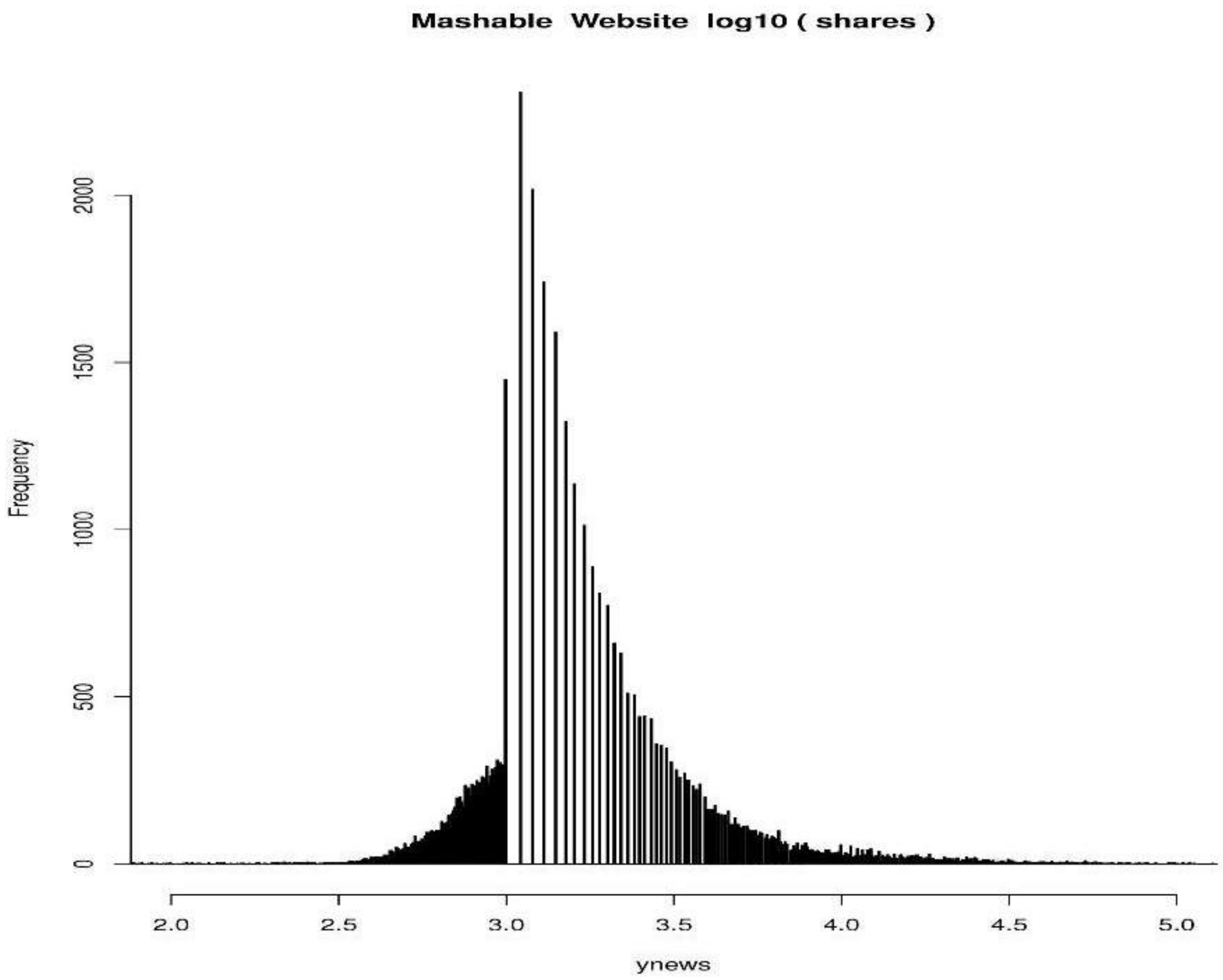
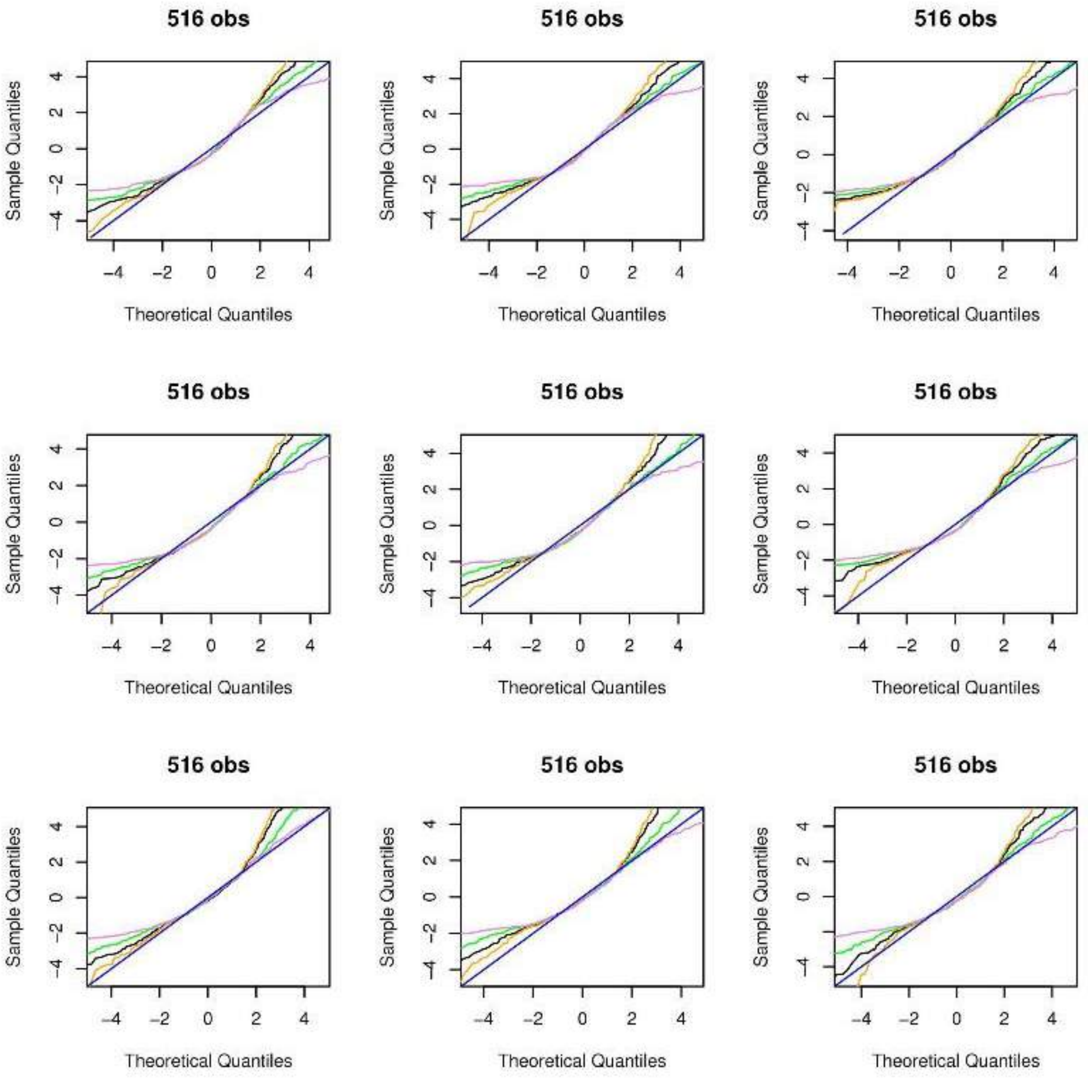
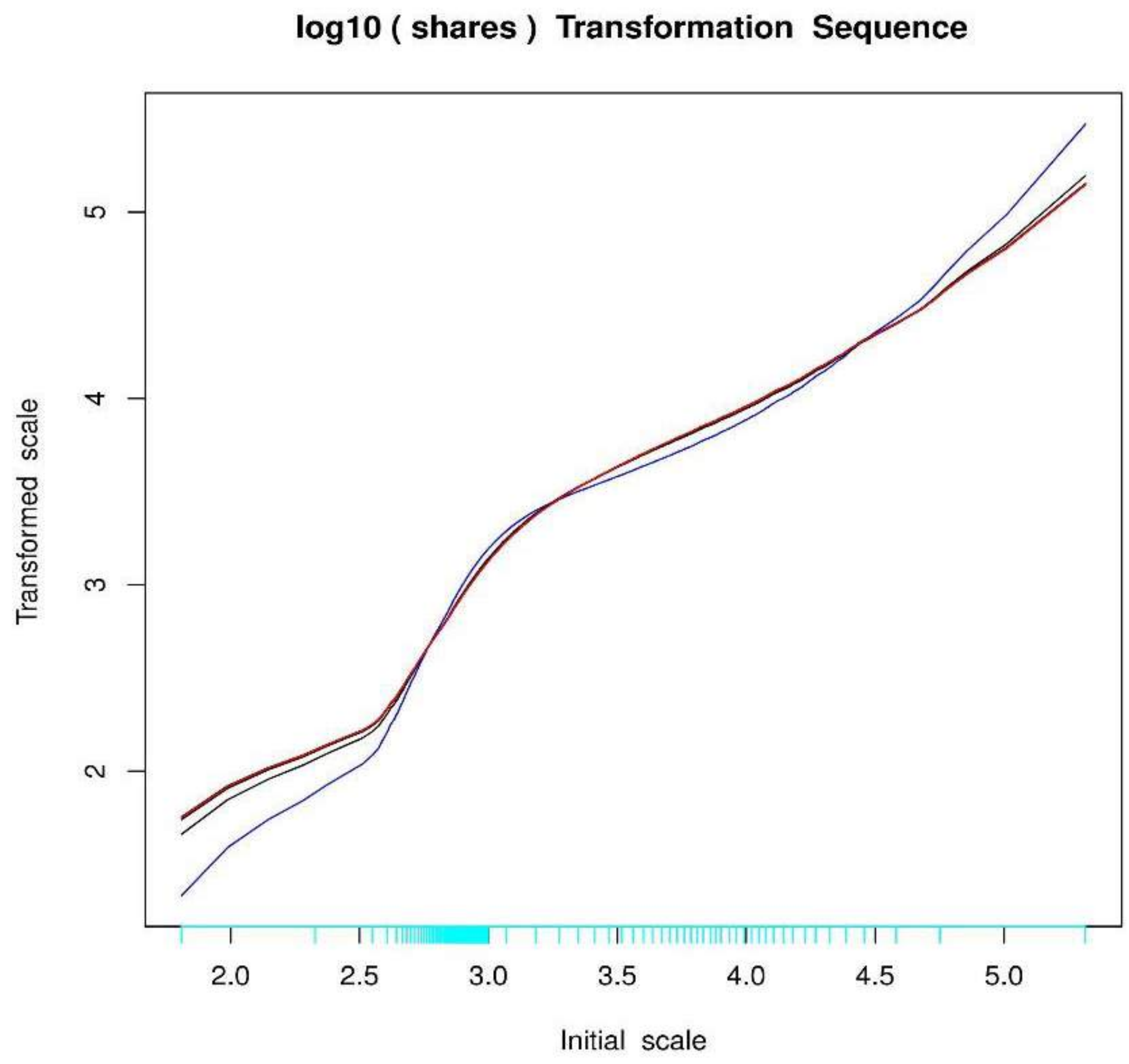
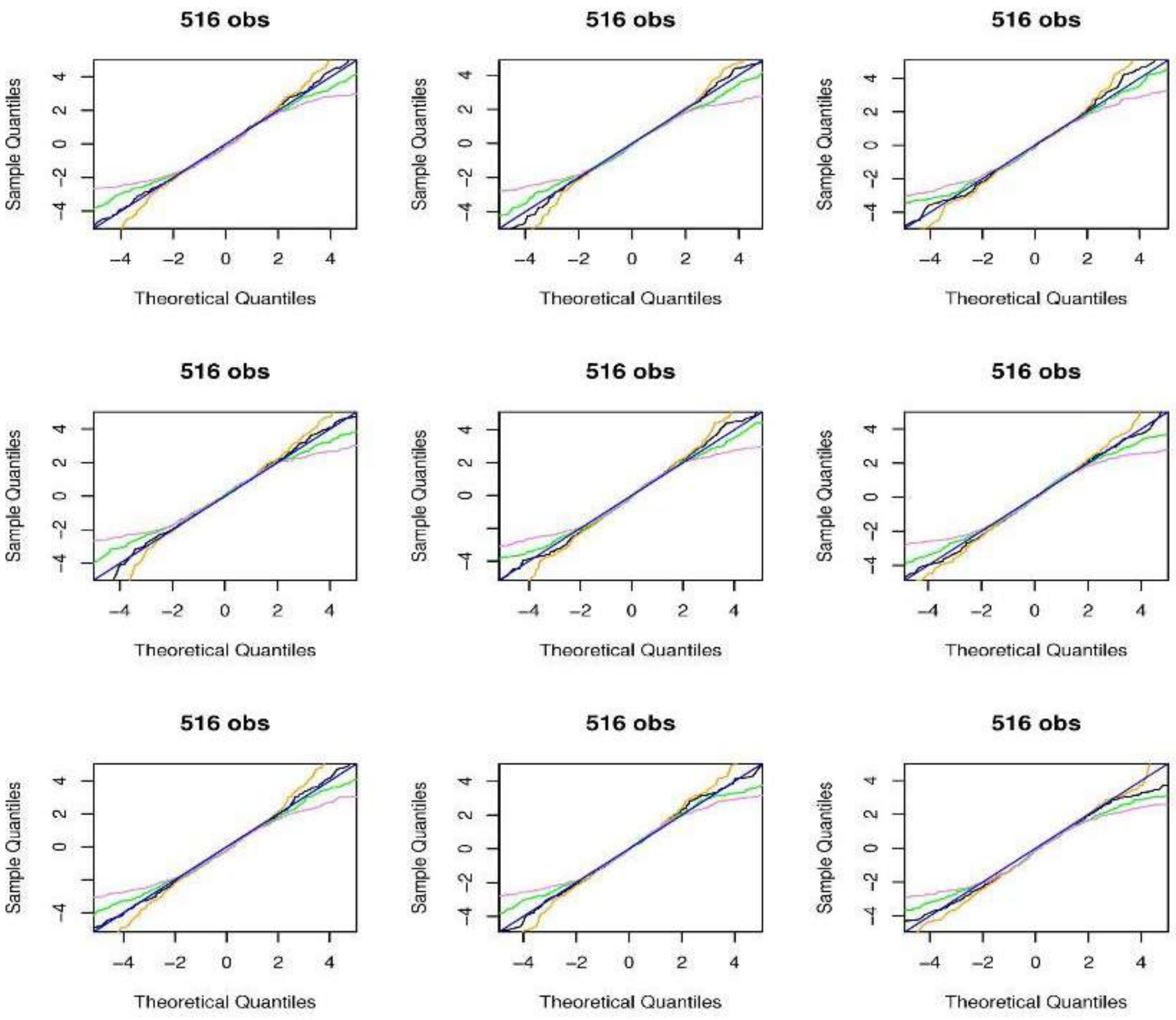

Mashable Website log10 (shares)
Frequency
0
500
1000
1500
2000
2.0
2.5
3.0
3.5
4.0
4.5
5.0
ynews

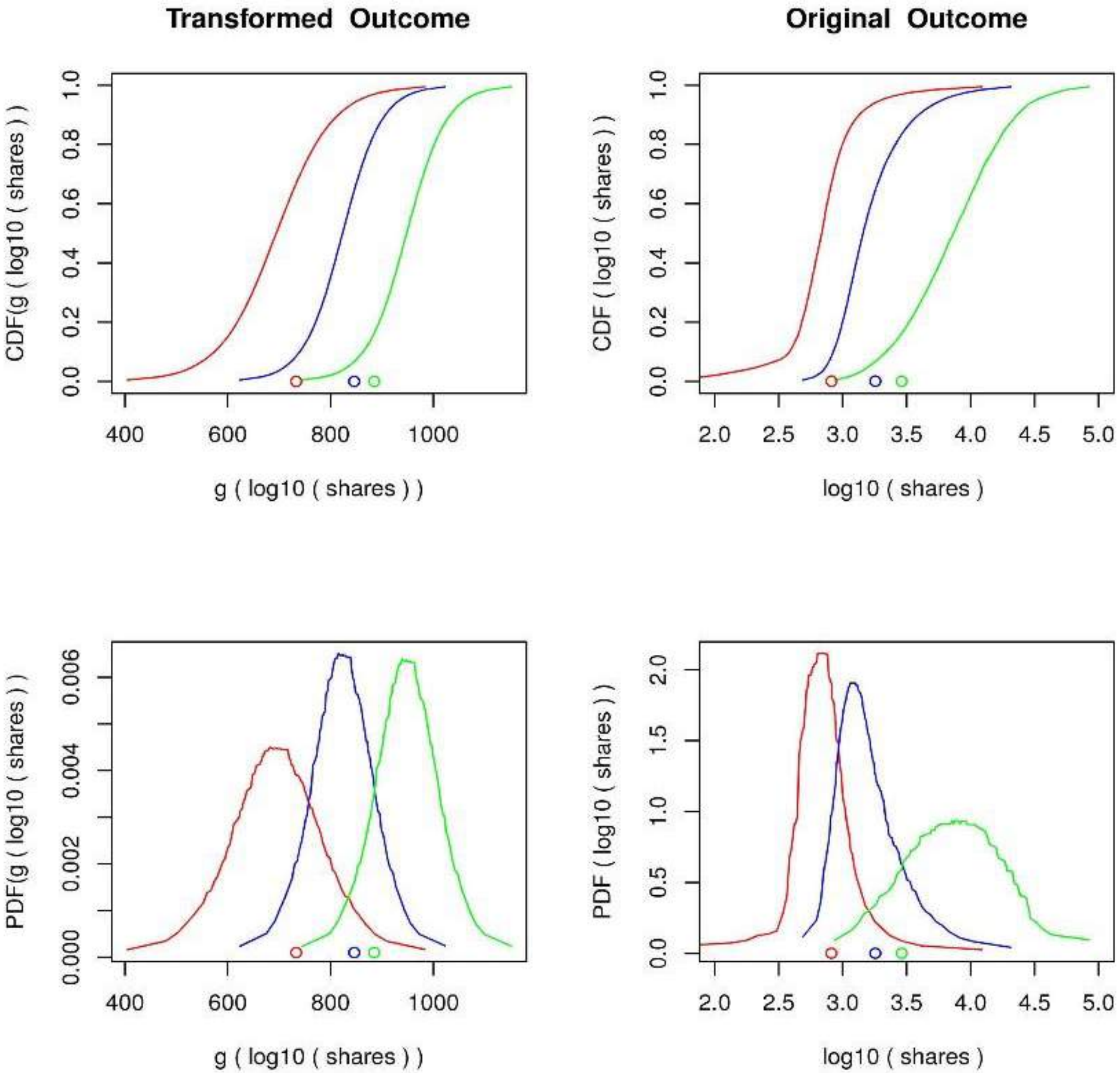
Fig 10


516 obs
Sample Quantiles
Theoretical Quantiles
-4
-2
0
2
4


log10 (shares) Transformation Sequence
Transformed scale
Initial scale
2
3
4
5
2.0
2.5
3.0
3.5
4.0
4.5
5.0



516 obs
Sample Quantiles
Theoretical Quantiles
516 obs
Sample Quantiles
Theoretical Quantiles
516 obs
Sample Quantiles
Theoretical Quantiles
516 obs
Sample Quantiles
Theoretical Quantiles
516 obs
Sample Quantiles
Theoretical Quantiles
516 obs
Sample Quantiles
Theoretical Quantiles
516 obs
Sample Quantiles
Theoretical Quantiles
516 obs
Sample Quantiles
Theoretical Quantiles
516 obs
Sample Quantiles
Theoretical Quantiles


Transformed Outcome
Original Outcome
CDF(g (log10 (shares))
CDF (log10 (shares))
g (log10 (shares))
log10 (shares)
PDF(g (log10 (shares))
PDF (log10 (shares))
g (log10 (shares))
log10 (shares)

(1/2) – Million Song Dataset

Song Recordings

(Irvine Repository)

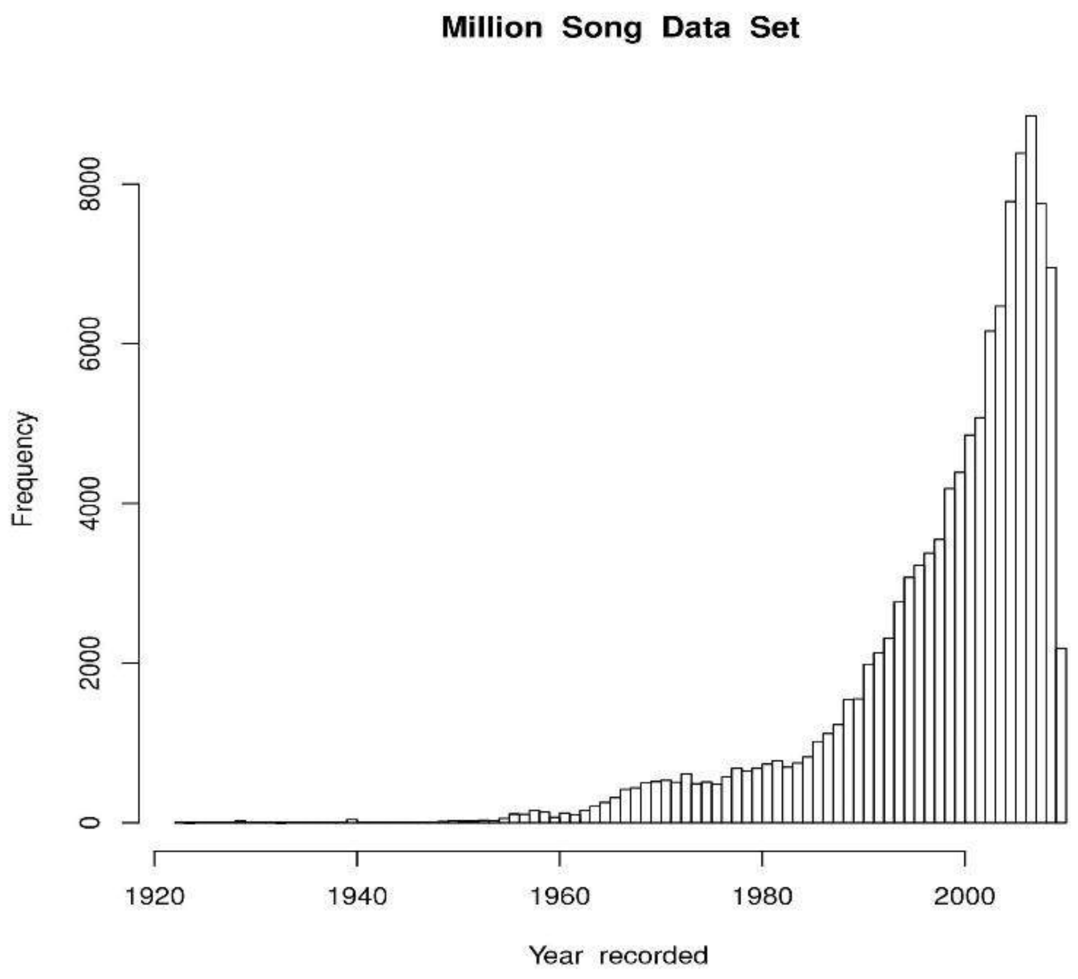
$N(train) = 463715$ song recordings

50000 train, 10000 mod. selection

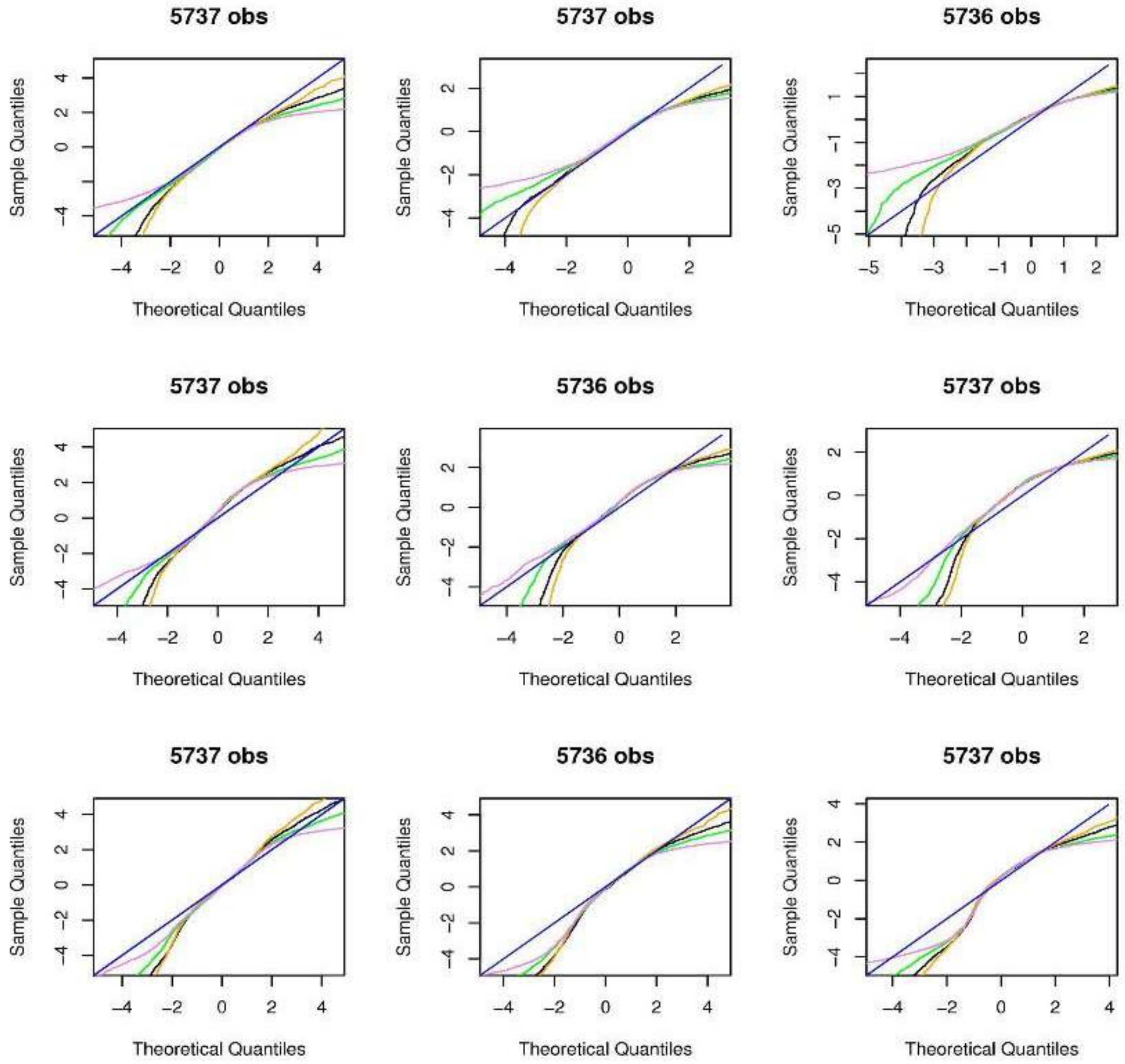
$N(test) = 51630$

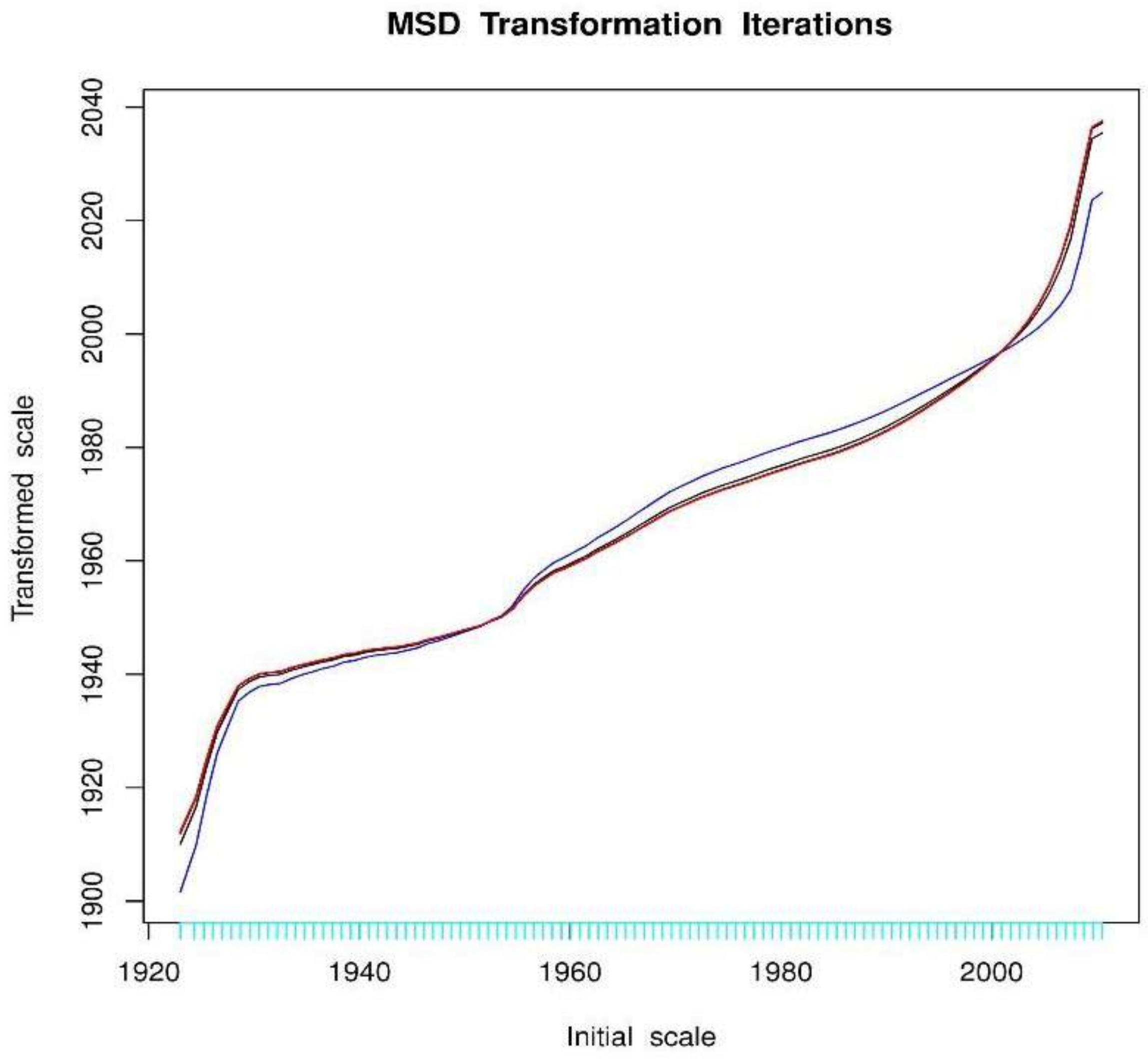
$y =$ year released (1922 – 2011)

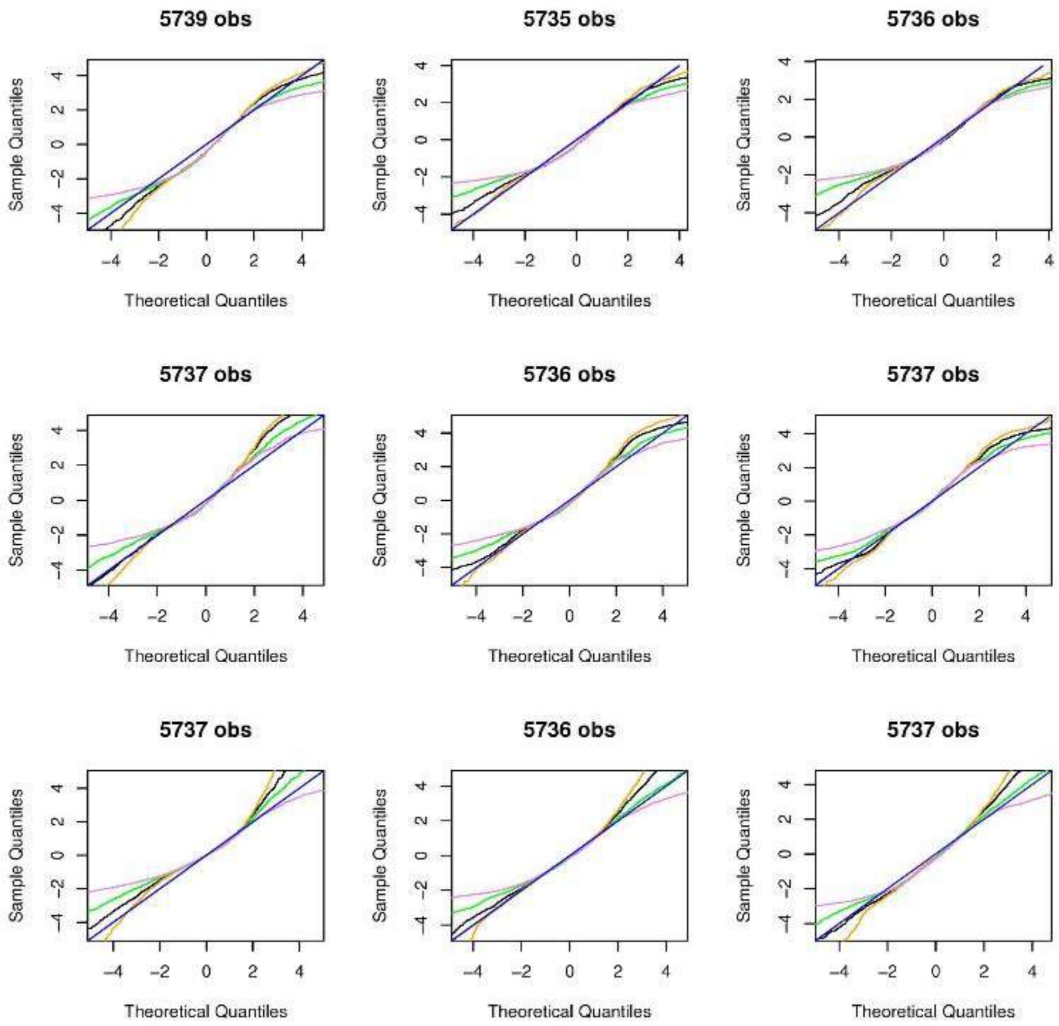
$\mathbf{x} = 89$ acustic measurements

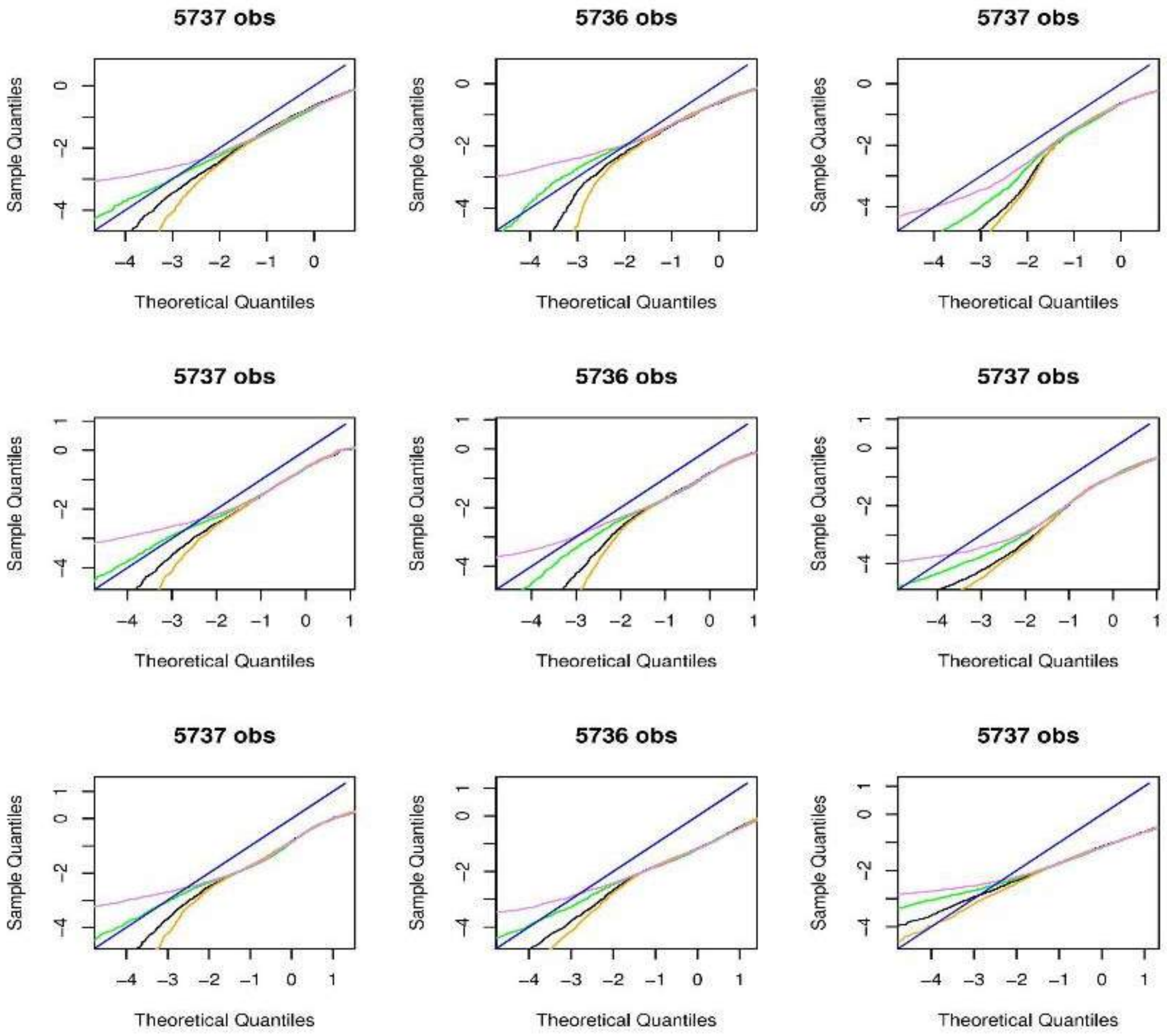


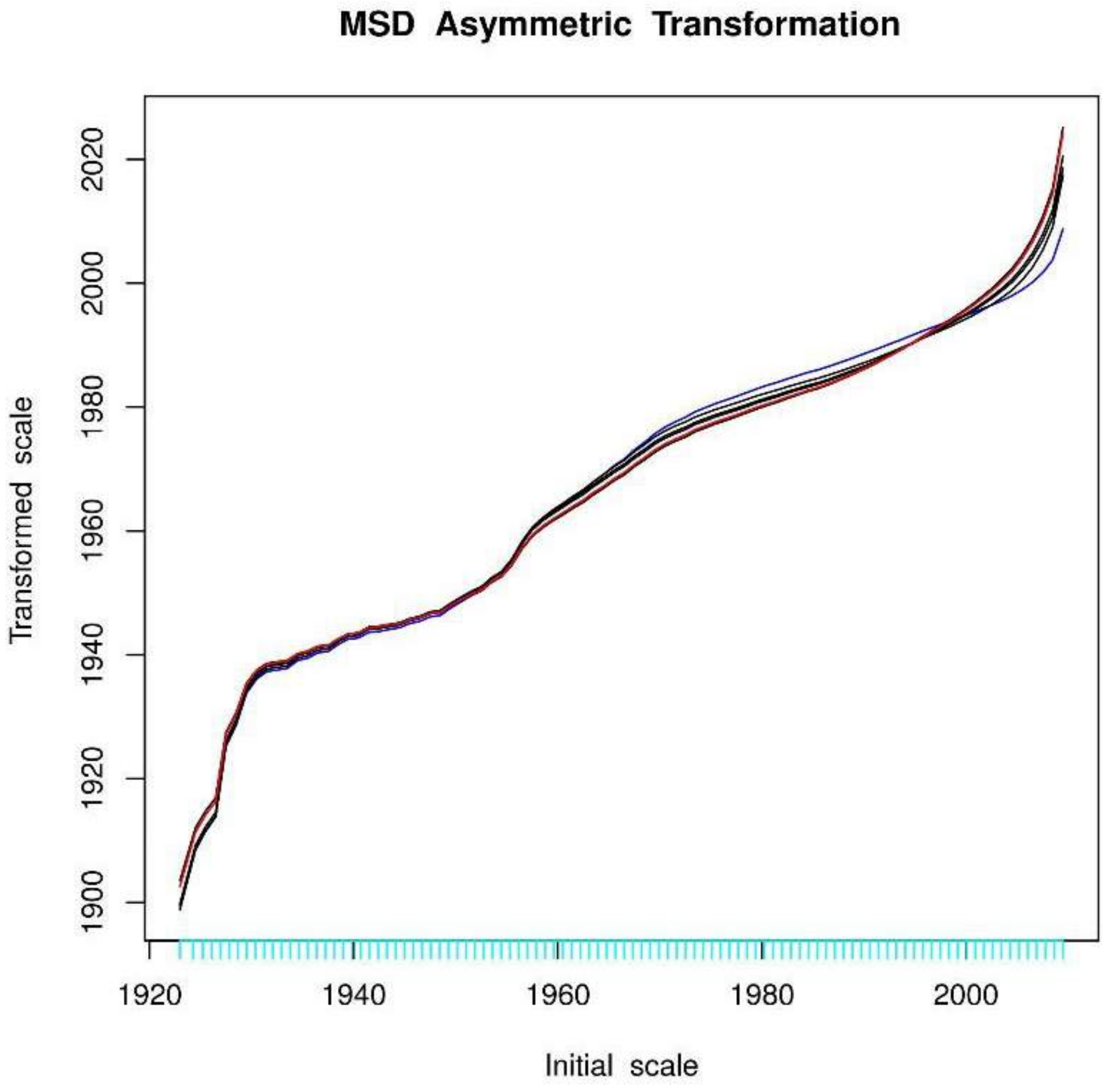
Million Song Data Set
Frequency
0
2000
4000
6000
8000
1920
1940
1960
1980
2000
Year recorded


5737 obs
5737 obs
5736 obs
5737 obs
5736 obs
5737 obs
5737 obs
5736 obs
5737 obs
Sample Quantiles
Theoretical Quantiles

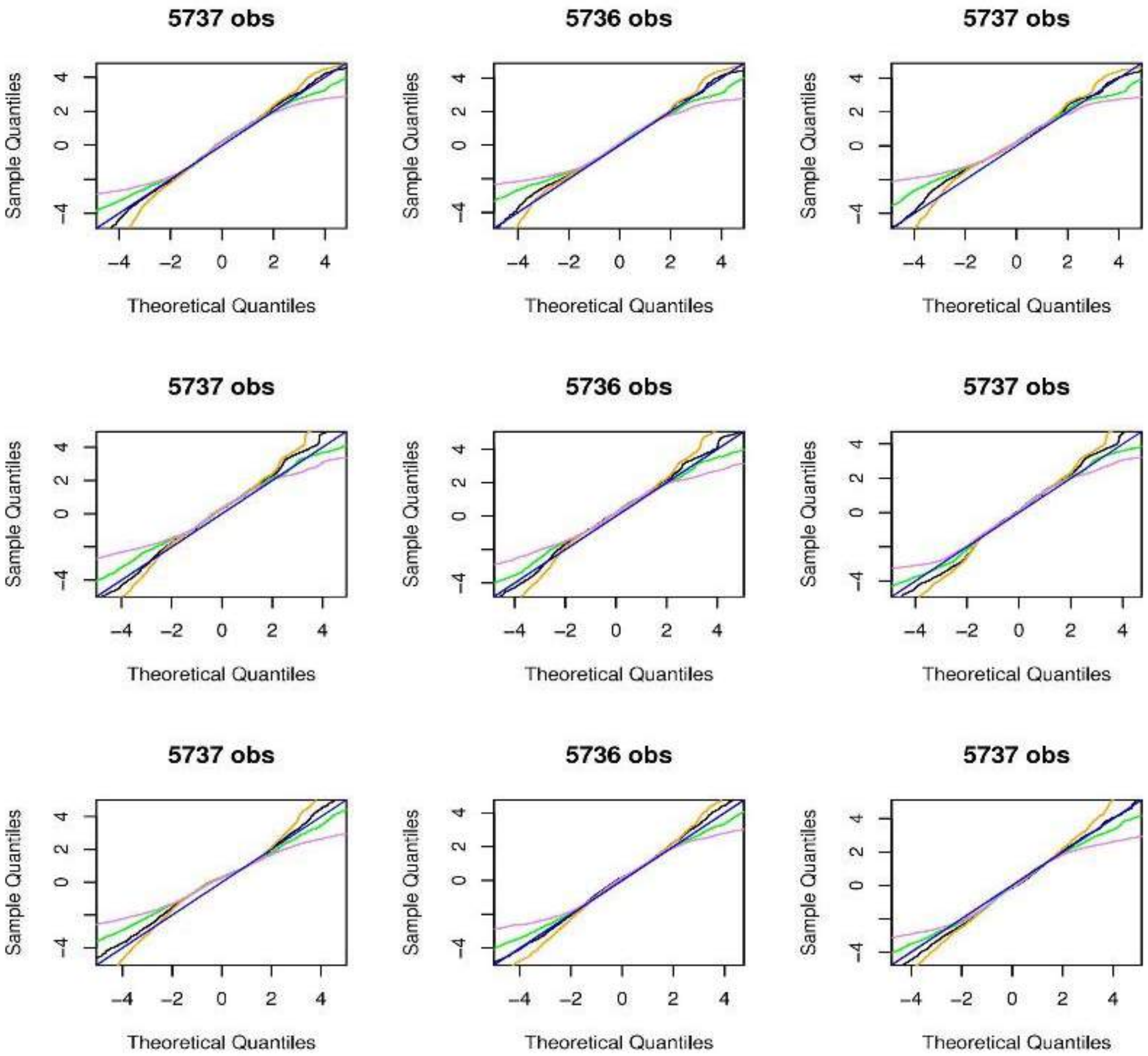

MSD Transformation Iterations
Transformed scale
Initial scale
1900
1920
1940
1960
1980
2000
2020
2040
1920
1940
1960
1980
2000


5739 obs
5735 obs
5736 obs
5737 obs
5736 obs
5737 obs
5737 obs
5736 obs
5737 obs
Sample Quantiles
Theoretical Quantiles
-4
-2
0
2
4

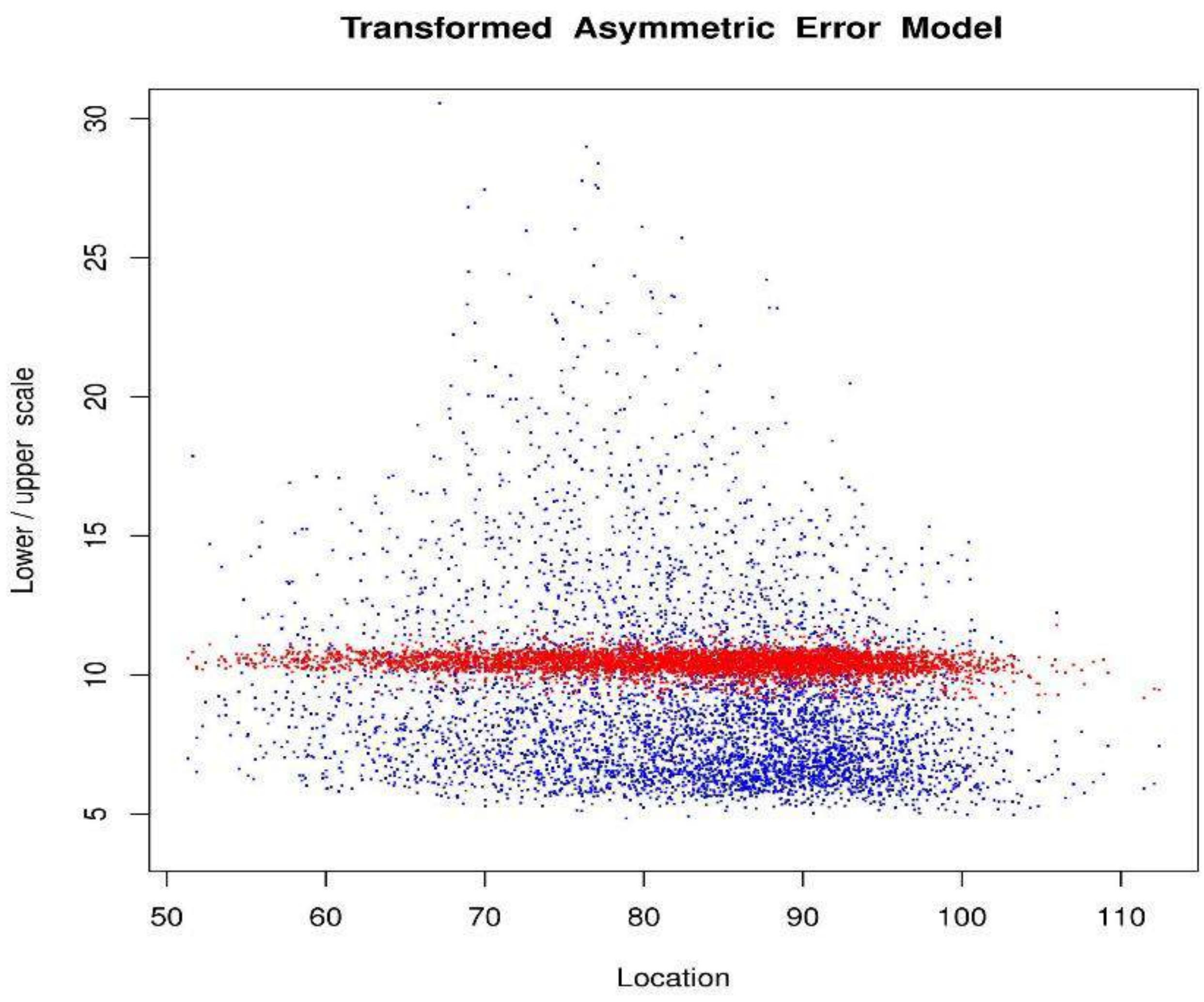

5737 obs
5736 obs
5737 obs
5737 obs
5736 obs
5737 obs
5737 obs
5736 obs
5737 obs
Sample Quantiles
Theoretical Quantiles

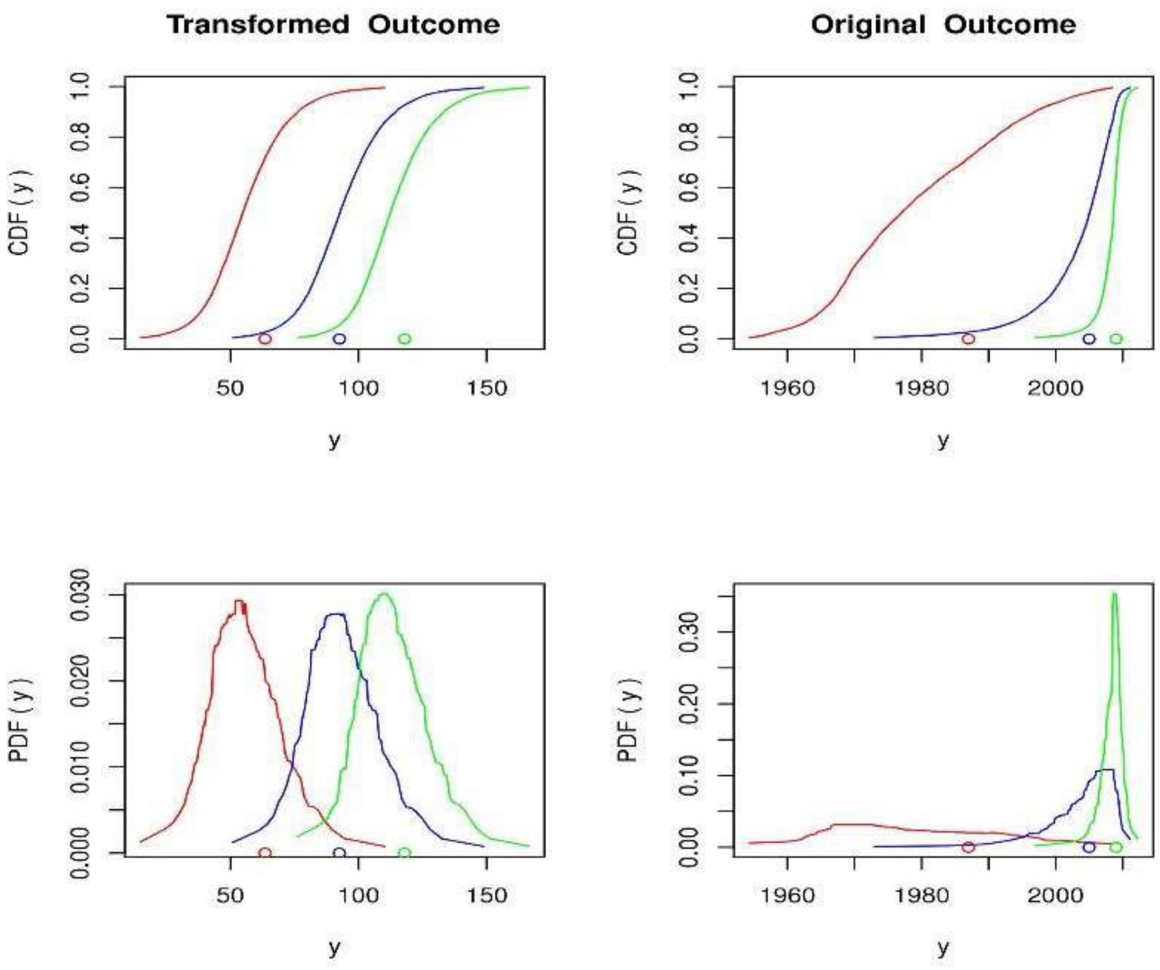


MSD Asymmetric Transformation
Transformed scale
1900
1920
1940
1960
1980
2000
2020
Initial scale



5737 obs
Sample Quantiles
Theoretical Quantiles
5736 obs
Sample Quantiles
Theoretical Quantiles
5737 obs
Sample Quantiles
Theoretical Quantiles
5737 obs
Sample Quantiles
Theoretical Quantiles
5736 obs
Sample Quantiles
Theoretical Quantiles
5737 obs
Sample Quantiles
Theoretical Quantiles
5737 obs
Sample Quantiles
Theoretical Quantiles
5736 obs
Sample Quantiles
Theoretical Quantiles
5737 obs
Sample Quantiles
Theoretical Quantiles


Transformed Asymmetric Error Model
Lower / upper scale
5
10
15
20
25
30
50
60
70
80
90
100
110
Location


Transformed Outcome
Original Outcome
CDF (y)
PDF (y)
y
0.0
0.2
0.4
0.6
0.8
1.0
50
100
150
1960
1980
2000
0.000
0.010
0.020
0.030
0.00
0.10
0.20
0.30

LESSONS

1. Don't confuse reducible and irreducible error.

2. Tukey was right (normality is rare)

3. Even approximate homoscedasticity is rare

4. Even approximate symmetry is rare

5. Optimal (nonobvious) transformations can help

Slides:

http://statweb.stanford.edu/~jhf/talks/baidu.pdf