# A NON-SYMMETRIC LOGIT MODEL AND GROUPED PREDICTAND CATEGORY DEVELOPMENT

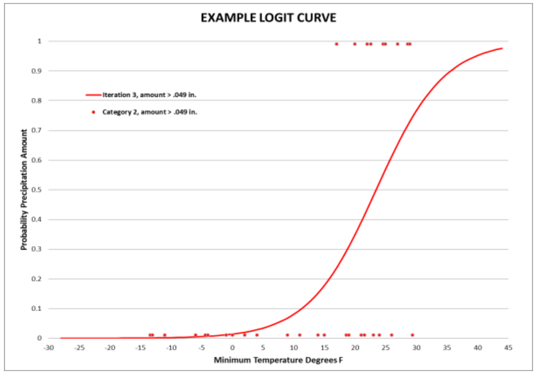
Bob Glahn

February 6, 2014

# LOGISTIC REGRESSION

Fits a binary predictand to one or more predictors

$$P = \frac{e^{f(x)}}{1+e^{f(x)}}$$




f(x) = a + bx

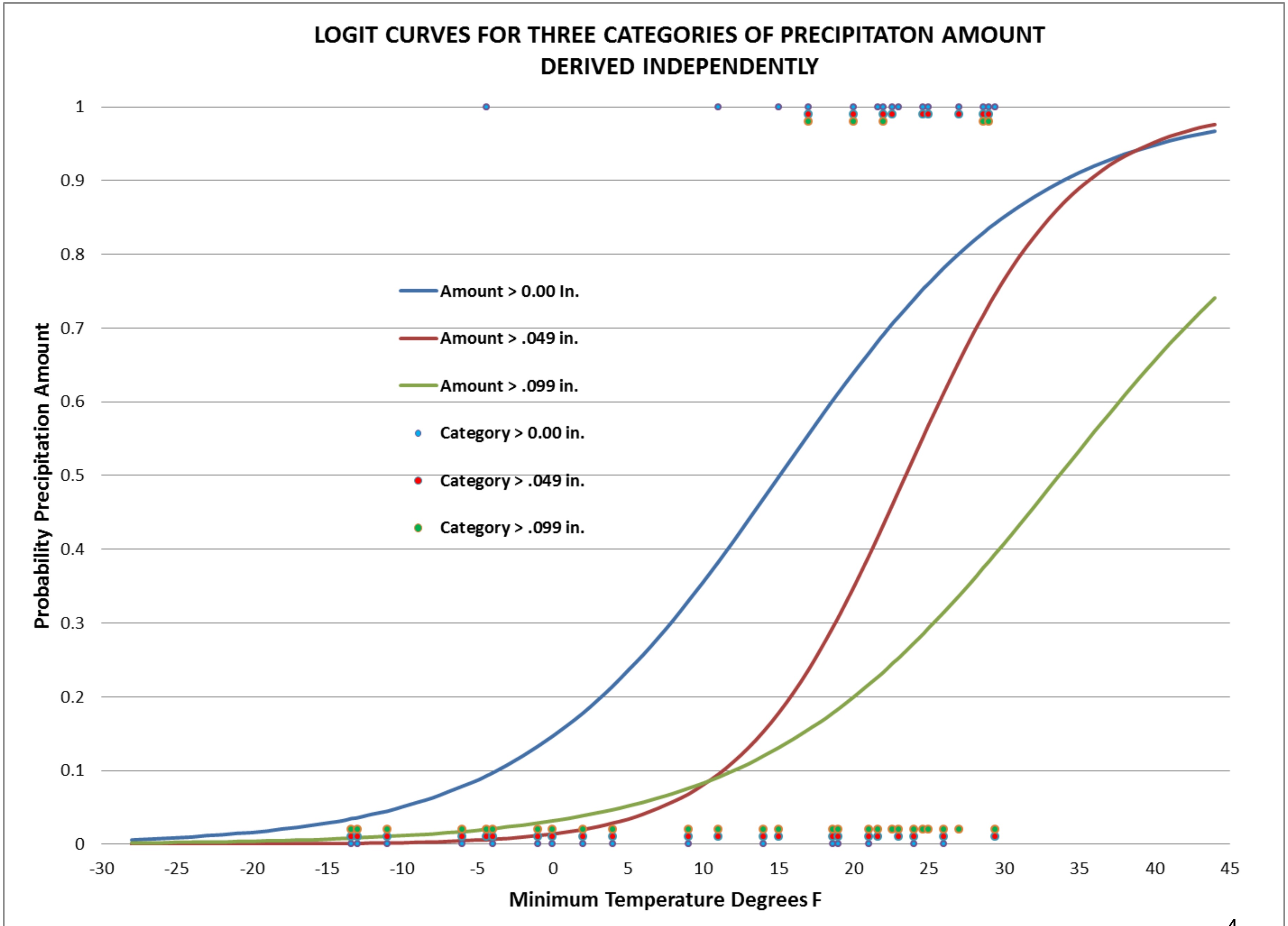
Produces a symmetric curve

# Questions

- **Why should the curve be symmetric**
  – A non-symmetric curve may fit the data better

- **What about multiple predictand categories?**
  – A quasi-continuous variable can be categorized into several binary variables
  – Each binary can be fit with logistic regression
  – However, the curves may overlap, and the estimates will be inconsistent, e.g.,

    P (1 inch snow) > P (2 inches snow)


LOGIT CURVES FOR THREE CATEGORIES OF PRECIPITATON AMOUNT
DERIVED INDEPENDENTLY
Probability Precipitation Amount
Minimum Temperature Degrees F
Amount > 0.00 In.
Amount > .049 in.
Amount > .099 in.
Category > 0.00 in.
Category > .049 in.
Category > .099 in.
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
35
40
45

# Example and Data

Example

Relate probability of precip in categories to cube root of model qpf

Data used

GFS output, 24-h projection, 12-h accumulated qpf in mm

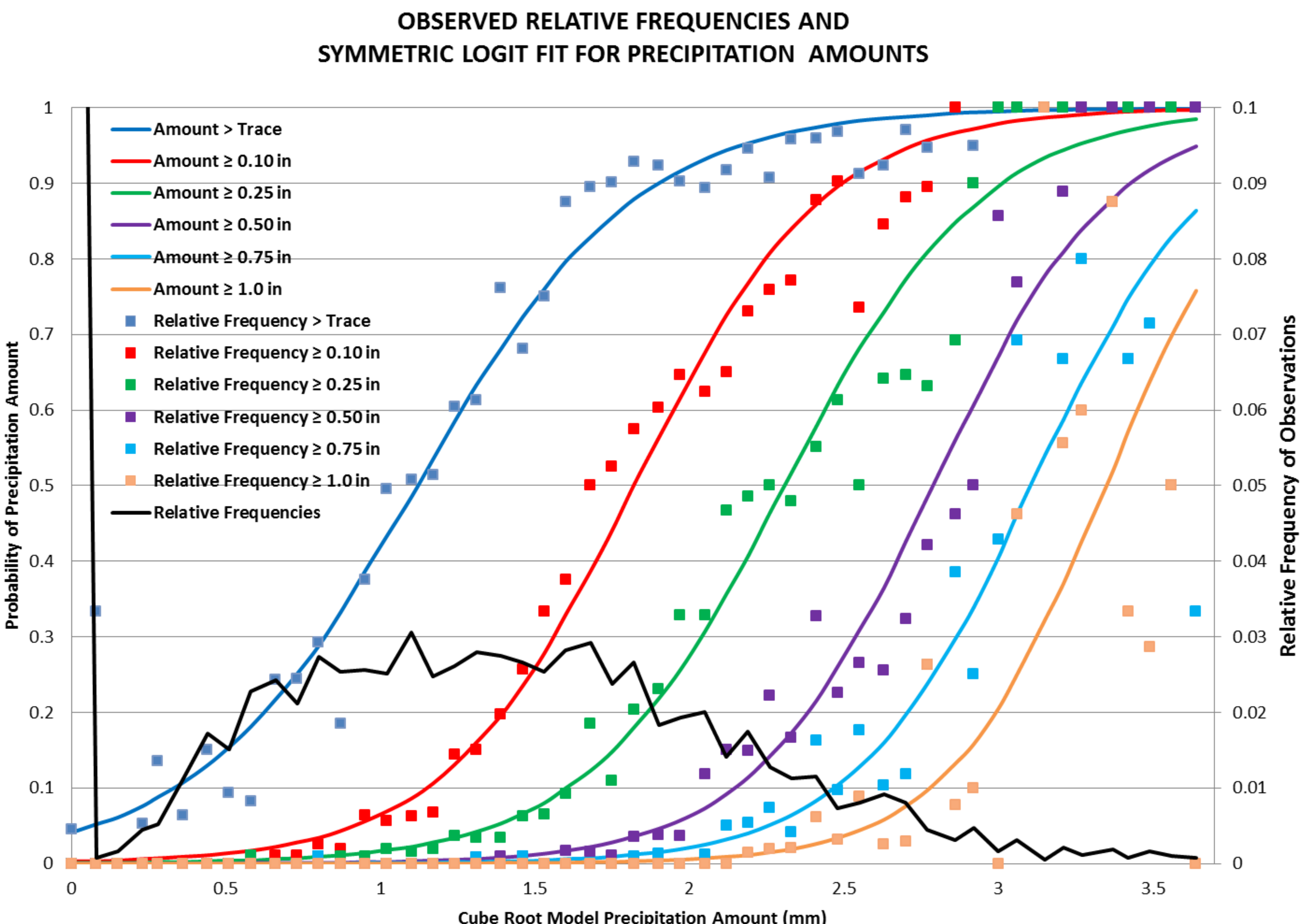
Observations of 12-h precip in inches 24 hours after model run time

24 Stations in Pacific Northwest

Two cool seasons, October-March

2011-12 (development); 2012-13 (independent)

Categories ≥ .01, ≥ 0.1, ≥ 0.25, ≥ 0.5, ≥ 0.75, ≥ 1.0



OBSERVED RELATIVE FREQUENCIES AND
SYMMETRIC LOGIT FIT FOR PRECIPITATION AMOUNTS
Amount > Trace
Amount ≥ 0.10 in
Amount ≥ 0.25 in
Amount ≥ 0.50 in
Amount ≥ 0.75 in
Amount ≥ 1.0 in
Relative Frequency > Trace
Relative Frequency ≥ 0.10 in
Relative Frequency ≥ 0.25 in
Relative Frequency ≥ 0.50 in
Relative Frequency ≥ 0.75 in
Relative Frequency ≥ 1.0 in
Relative Frequencies
Probability of Precipitation Amount
Relative Frequency of Observations
Cube Root Model Precipitation Amount (mm)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
0
0.5
1
1.5
2
2.5
3
3.5

# Non-Symmetric Logistic Regression

$f(q) = a + bq$

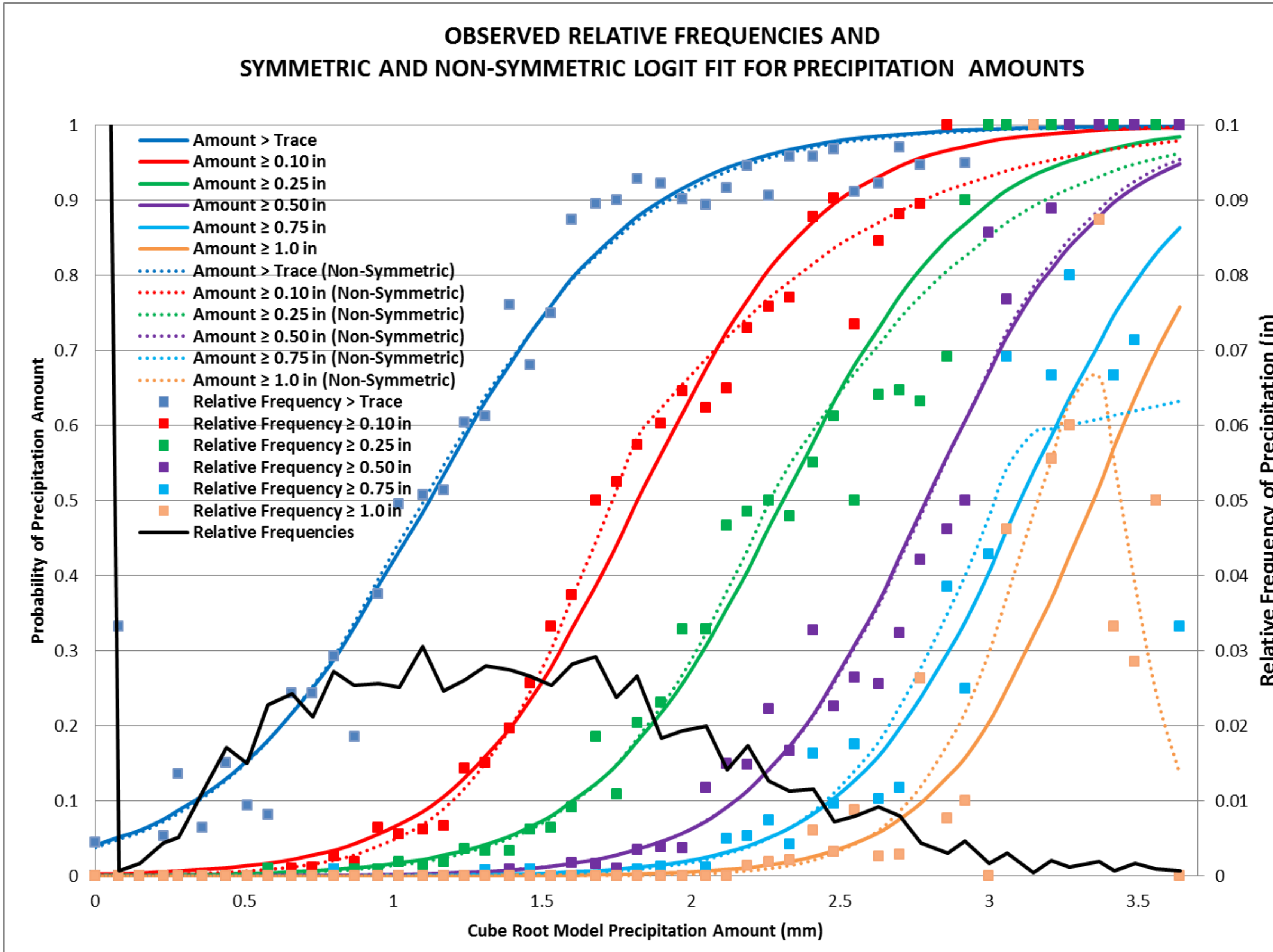
Include another predictor $q'$

$q' = q - q^*$ when $q > q^*$

$q' = 0$ otherwise

Parallels piecewise linear in linear regression (changes slope at $q^*$)

$f(q) = a_{ns} + b_{ns}q + c_{ns}q'$


OBSERVED RELATIVE FREQUENCIES AND
SYMMETRIC AND NON-SYMMETRIC LOGIT FIT FOR PRECIPITATION AMOUNTS
Amount > Trace
Amount ≥ 0.10 in
Amount ≥ 0.25 in
Amount ≥ 0.50 in
Amount ≥ 0.75 in
Amount ≥ 1.0 in
Amount > Trace (Non-Symmetric)
Amount ≥ 0.10 in (Non-Symmetric)
Amount ≥ 0.25 in (Non-Symmetric)
Amount ≥ 0.50 in (Non-Symmetric)
Amount ≥ 0.75 in (Non-Symmetric)
Amount ≥ 1.0 in (Non-Symmetric)
Relative Frequency > Trace
Relative Frequency ≥ 0.10 in
Relative Frequency ≥ 0.25 in
Relative Frequency ≥ 0.50 in
Relative Frequency ≥ 0.75 in
Relative Frequency ≥ 1.0 in
Relative Frequencies
Probability of Precipitation Amount
Relative Frequency of Precipitation (in)
Cube Root Model Precipitation Amount (mm)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
0
0.5
1
1.5
2
2.5
3
3.5

# Summary—Non symmetric Development

- Demonstrated concept but was of limited practical use on this data set.
- Total sample size is not the only consideration, but sample on both sides of q*
- Potential for non-symmetric in situations where relationship is non-monotonic

# Grouped Category Development—Method 1 Separation Constant Option

Independently-derived curves for different thresholds may give inconsistent results.

Possible Inconsistencies can be avoided by grouping the data for all categories and including a 2nd predictor which denotes group membership, as Wilks has suggested (*Meteorological Applications*, **16** , 2009).

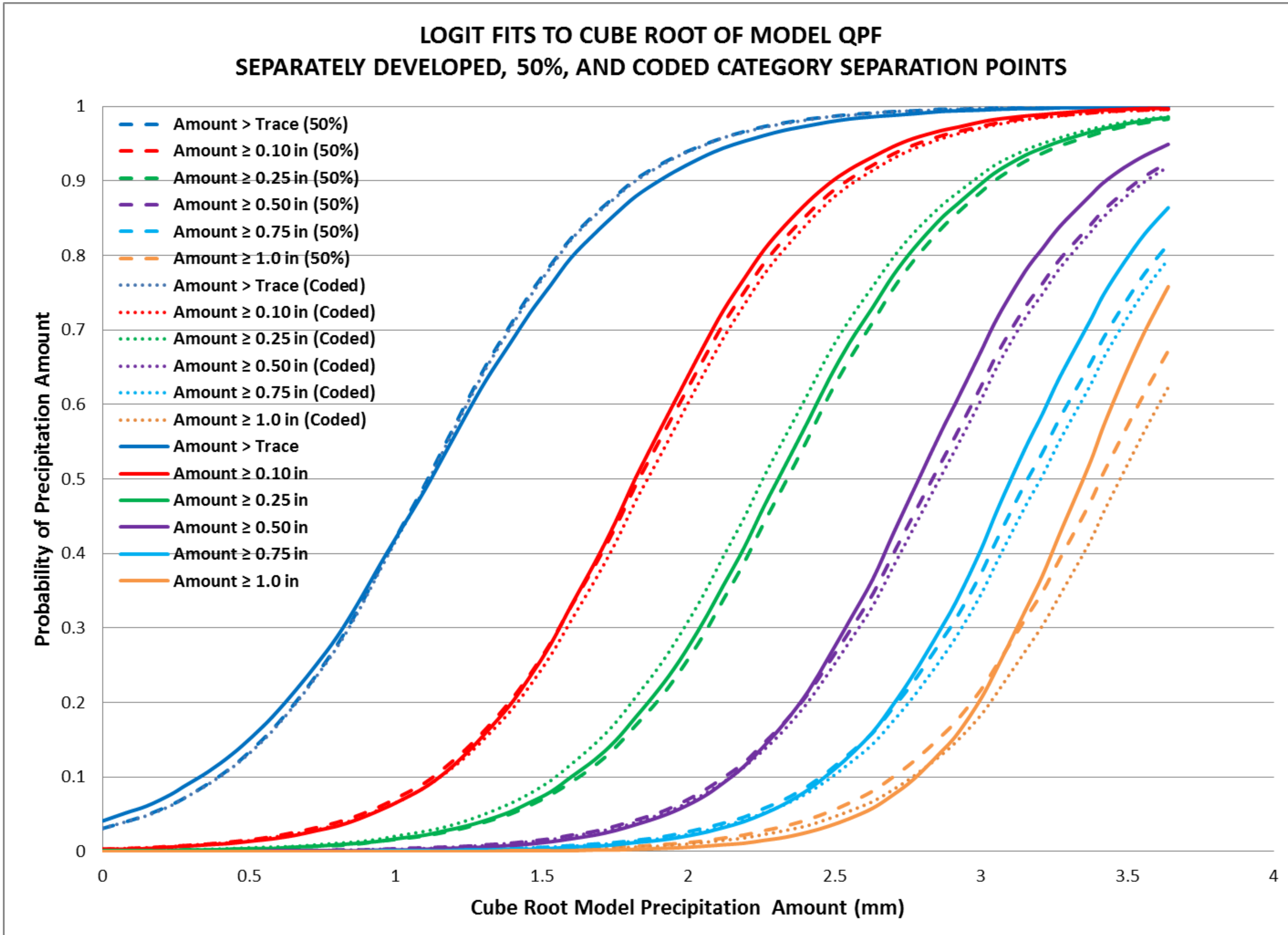
$f(q) = a_1 + b_1 q + c_1 g$

g for each category can be given the value where the individual curves cross 50% in the units of the predictor, cube root mm

# Grouped Category Development—Method 2
# Binary Predictor Option

Possible Inconsistencies can be avoided by grouping the data for all categories and including binary predictors which denote group membership

$$f(q) = a_2 + b_2 q + \sum c_i B_i$$



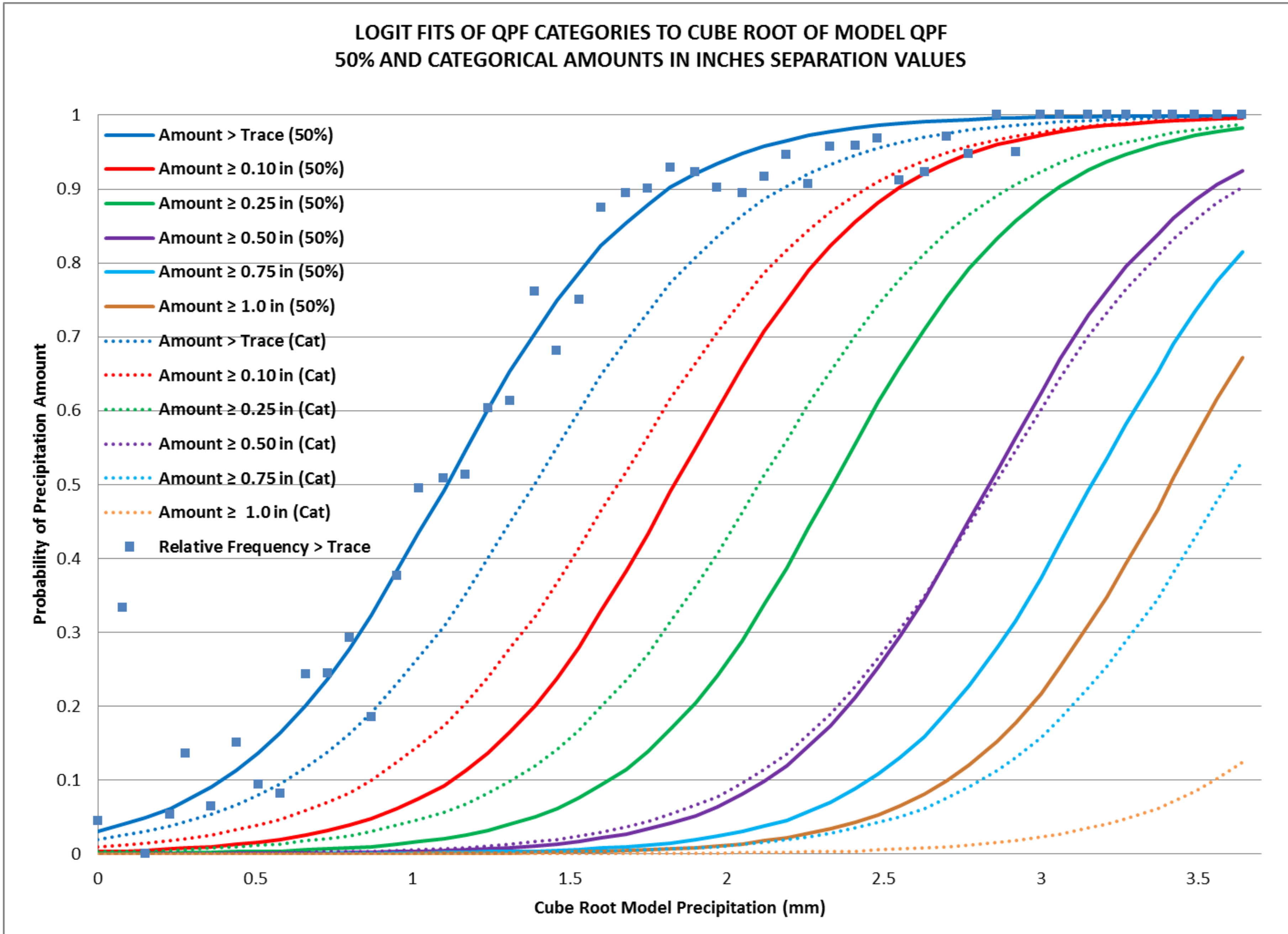
LOGIT FITS TO CUBE ROOT OF MODEL QPF
SEPARATELY DEVELOPED, 50%, AND CODED CATEGORY SEPARATION POINTS
Amount > Trace (50%)
Amount ≥ 0.10 in (50%)
Amount ≥ 0.25 in (50%)
Amount ≥ 0.50 in (50%)
Amount ≥ 0.75 in (50%)
Amount ≥ 1.0 in (50%)
Amount > Trace (Coded)
Amount ≥ 0.10 in (Coded)
Amount ≥ 0.25 in (Coded)
Amount ≥ 0.50 in (Coded)
Amount ≥ 0.75 in (Coded)
Amount ≥ 1.0 in (Coded)
Amount > Trace
Amount ≥ 0.10 in
Amount ≥ 0.25 in
Amount ≥ 0.50 in
Amount ≥ 0.75 in
Amount ≥ 1.0 in
Probability of Precipitation Amount
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3
3.5
4
Cube Root Model Precipitation Amount (mm)

# Grouped Category Development—Method 1

$f(q) = a_1 + b_1 q + c_1 g$

The choice of g is critical.

Suppose g was given the value of the individual category limits in inches of precipitation (e.g., 0.01, 0.25).


LOGIT FITS OF QPF CATEGORIES TO CUBE ROOT OF MODEL QPF
50% AND CATEGORICAL AMOUNTS IN INCHES SEPARATION VALUES
Amount > Trace (50%)
Amount ≥ 0.10 in (50%)
Amount ≥ 0.25 in (50%)
Amount ≥ 0.50 in (50%)
Amount ≥ 0.75 in (50%)
Amount ≥ 1.0 in (50%)
Amount > Trace (Cat)
Amount ≥ 0.10 in (Cat)
Amount ≥ 0.25 in (Cat)
Amount ≥ 0.50 in (Cat)
Amount ≥ 0.75 in (Cat)
Amount ≥ 1.0 in (Cat)
Relative Frequency > Trace
Probability of Precipitation Amount
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3
3.5
Cube Root Model Precipitation (mm)

# Summary
# Grouped Category Development—Method 1

Probabilities are consistent (lines do not cross).

Defining a value for group membership is important when > 2 groups

When only 2 groups, values do not matter

Definition of group membership requires knowledge of predictor predictand relationship. A good definition for one predictor may not be good for another predictor, either individually or in combination with the first.

If a functional form for group membership is known, then the probability of <u>any</u> group definition can be calculated from the single equation.

For a single predictor, only three constants are fit.

# Summary
# Grouped Category Development—Method 2

Probabilities are consistent (lines do not cross).

Defining a value for group membership is by definition = 1

When only 2 groups, grouping not necessary—one equation is sufficient

Definition of group membership does not require knowledge of predictor predictand relationship; extension to multiple predictors is not a problem

There is no functional form for group membership, and the probability of any group definition cannot be calculated.

For a single predictor, the number of constants fit = the number of categories plus 1