Selecting the Best Distribution for Your Response

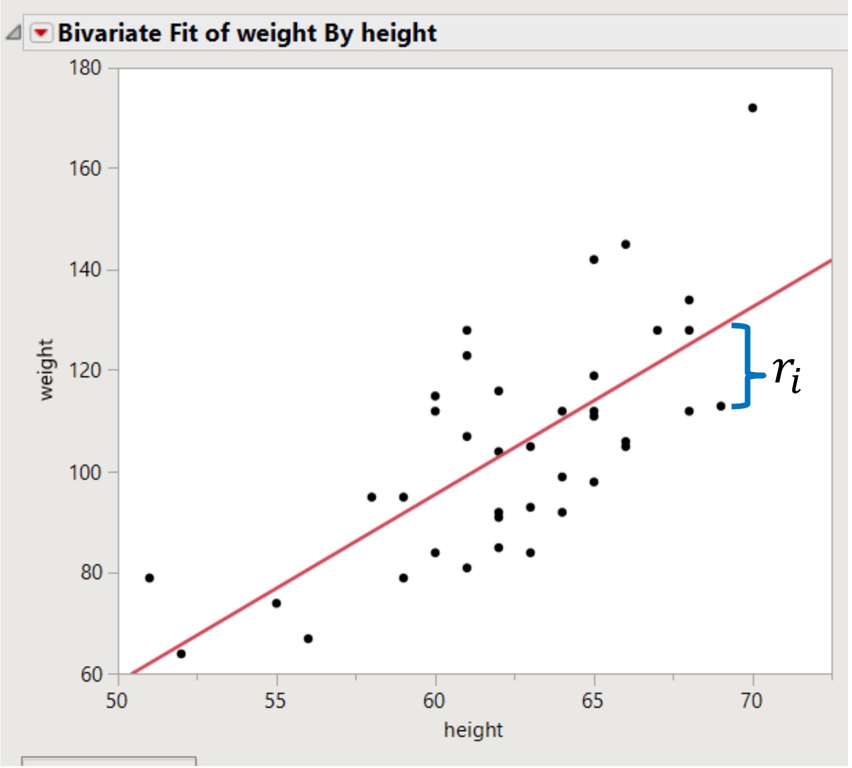
Developer Tutorial

Clay Barker, PhD

JMP Principal Research Statistician Developer

Copyright © SAS Institute Inc. All rights reserved.

Simple Linear Regression



What is simple linear regression?

$$Y_i = \beta_0 + \ \beta_1 x_i + \epsilon_i$$

Usually we assume

$$\epsilon_i \sim N(0, \sigma^2)$$

We don’t have to assume normality, but it makes inference easier.

Copyright © SAS Institute Inc. All rights reserved.

Simple Linear Regression

Assuming that the errors (and response) are normal makes life easier.

Why? Among other reasons, things like estimation and inference tend to have an explicit form.

For example:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$cov(\hat{\beta}) = (X^T X)^{-1} \sigma^2$$

We get a lot of mileage out of assuming normality, but it isn’t always valid.

Copyright © SAS Institute Inc. All rights reserved.

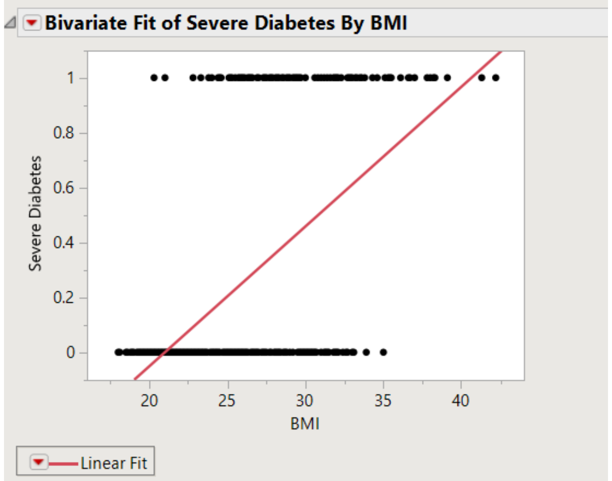
Simple Linear Model

Normality

What happens if you assume normality when you shouldn't?

Two main concerns:

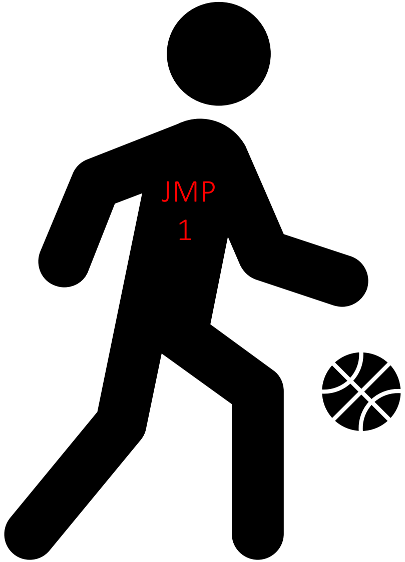
1. Predictions outside of meaningful range (maybe not a big deal)
2. Inference is not reliable (probably a bigger deal)



Copyright © SAS Institute Inc. All rights reserved.

What if normality is not appropriate?

Let’s say we want to model steals for a basketball player.

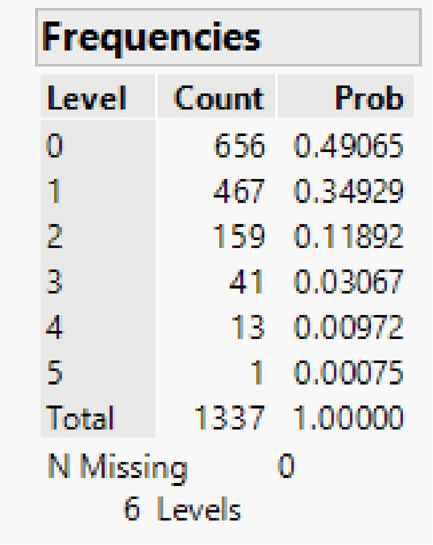


What might impact performance?

Experience?
Opponent?
Home/Away?
How much rest?
…

Copyright © SAS Institute Inc. All rights reserved.

Steals in Basketball



Frequencies

Level	Count	Prob
0	656	0.49065
1	467	0.34929
2	159	0.11892
3	41	0.03067
4	13	0.00972
5	1	0.00075
Total	1337	1.00000

N Missing 0
6 Levels

The response will only take integer values.

And even for the best players, the response will only take a couple values.

$Y = \{0,1,2,3,4,5\}$ for Steve Nash

Normality isn't appropriate at all here, but we still need to build a model.

What should we do?

Copyright © SAS Institute Inc. All rights reserved.

Overview

1. Overview of Generalized Linear Models (GLMs)
2. How to evaluate your models
    1. Know your data and your distribution
    2. R-square
    3. Information criteria
3. Examples

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Models

A quick overview

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Models

But first, back to the linear model

Our beloved linear model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

$$= x_i^T \beta + \epsilon_i \qquad \text{where } x_i \text{ is a } p+1 \text{ vector}$$

We assume that our errors are independent and normally distributed.

$$\epsilon_i \sim N(0, \sigma^2)$$

So given our predictor vector x_i, we know the distribution of the response

$$y_i | x_i \sim N(x_i^T \beta, \sigma^2)$$

$$\mathrm{E}(y_i|x_i) = x_i^T \beta \qquad \mathrm{Var}(y_i|x_i) = \sigma^2$$

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Model

Same idea as linear regression but instead of normality, we assume that $y_i|x_i$ has some other distribution.

Many cases where we need to do this

1. Count data (ex: number of defects on a product)
2. Skewed data (ex: salaries)
3. Proportions
4. Labels (ex: good/neutral/bad or yellow/blue/green)

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Model

Formal Statement

We assume a probability function for our response

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

This is called an *exponential family* distribution.

Do we need to spend any more time on this level of detail?

Nope.

Instead, let’s focus on **using** these models.

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Models

Three key ingredients to a GLM:

1. A distribution for the response given the predictors (the random piece)

2. A linear predictor $x_i^T \beta$ (the systematic piece)

3. A link function (the piece that connects 1 and 2)

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Model

The Distribution

When we talk about the distribution in our GLM, we're talking about the distribution of the **response given the predictors**.

This is a critical piece!

In general it **is not** the distribution of the residuals.

Of course there are exceptions (Normal, t-distribution, …)

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Model

The Distribution

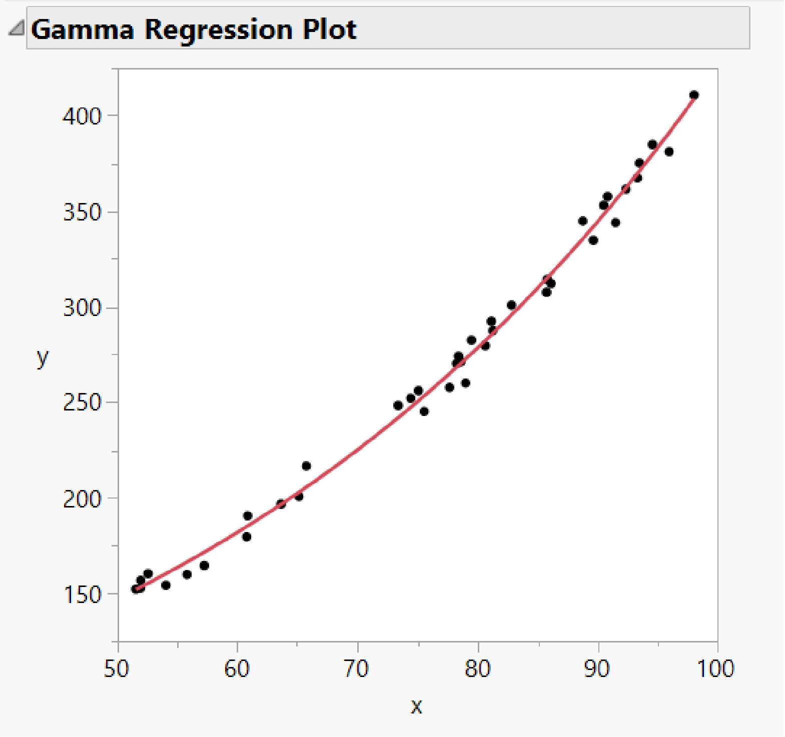
The Normal may trick us into thinking of the distribution of the residuals.

How can we avoid this mistake?

Helpful reminder:

The Gamma distribution is strictly positive.

The residuals for this Gamma regression are positive and negative.



Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Model

The Distribution

Another common gotcha.

The distribution **is not** the distribution of the response.

It’s the distribution of the response given the predictors.

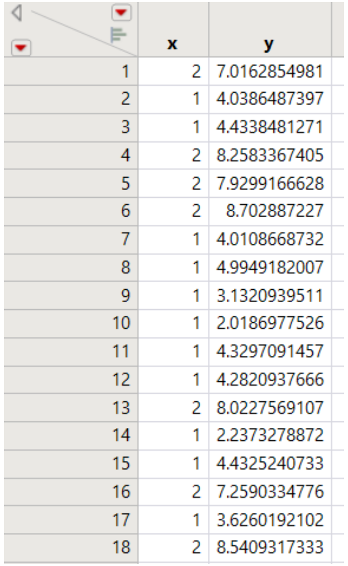
Why is this distinction important? Let’s look at an example.

Copyright © SAS Institute Inc. All rights reserved.

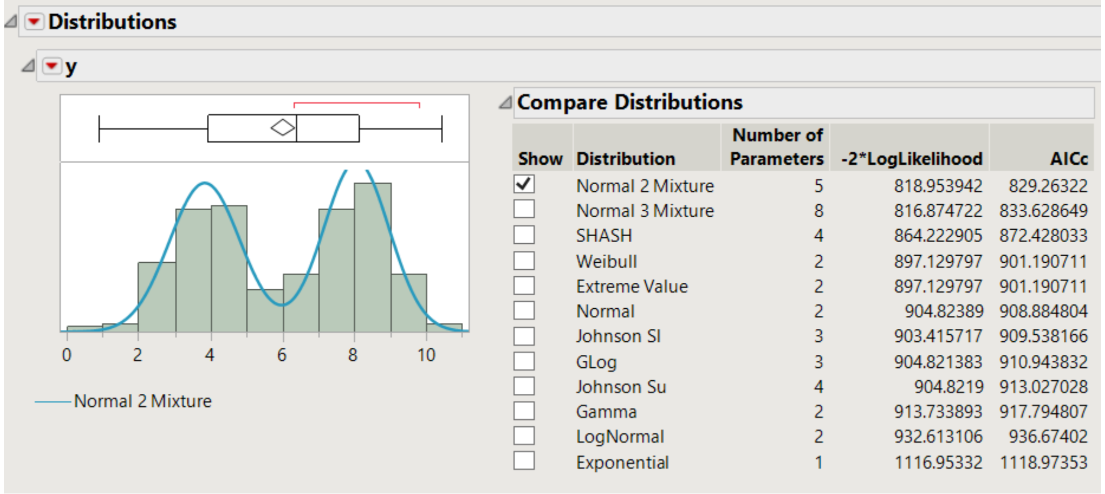
Generalized Linear Model

The Distribution

We have a single effect to model the response.
Based on this histogram, should we do a mixture model regression?



	x	y
1	2	7.0162854981
2	1	4.0386487397
3	1	4.4338481271
4	2	8.2583367405
5	2	7.9299166628
6	2	8.702887227
7	1	4.0108668732
8	1	4.9949182007
9	1	3.1320939511
10	1	2.0186977526
11	1	4.3297091457
12	1	4.2820937666
13	2	8.0227569107
14	1	2.2373278872
15	1	4.4325240733
16	2	7.2590334776
17	1	3.6260192102
18	2	8.5409317333



Distributions

y

Compare Distributions

Show	Distribution	Number of Parameters	-2*LogLikelihood	AICc
☑	Normal 2 Mixture	5	818.953942	829.26322
□	Normal 3 Mixture	8	816.874722	833.628649
□	SHASH	4	864.222905	872.428033
□	Weibull	2	897.129797	901.190711
□	Extreme Value	2	897.129797	901.190711
□	Normal	2	904.82389	908.884804
□	Johnson Sl	3	903.415717	909.538166
□	GLog	3	904.821383	910.943832
□	Johnson Su	4	904.8219	913.027028
□	Gamma	2	913.733893	917.794807
□	LogNormal	2	932.613106	936.67402
□	Exponential	1	1116.95332	1118.97353

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Model

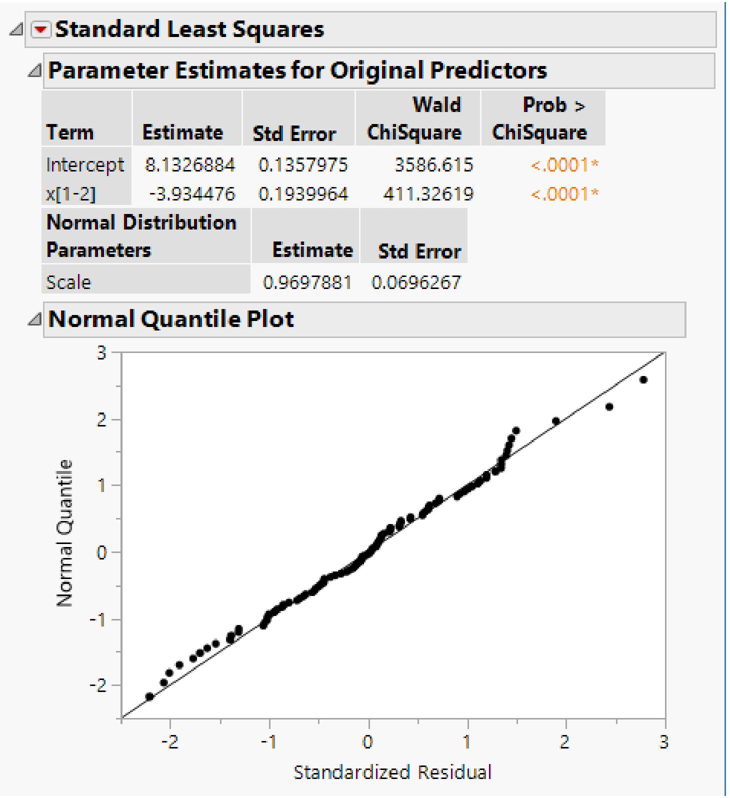
The Distribution

Absolutely not!

The truth is that this is just a simulated One-way ANOVA model

$$y_i = 4 + 4 * \mathrm{I}(x_i = 2) + z_i$$
$$z_i \sim N(0,1)$$
$$x_i = \{1,2\}$$

The histogram of the response ignores our predictor(s), so it provides limited information.



Standard Least Squares

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare
Intercept	8.1326884	0.1357975	3586.615	<.0001*
x[1-2]	-3.934476	0.1939964	411.32619	<.0001*

Normal Distribution Parameters	Estimate	Std Error
Scale	0.9697881	0.0696267

Normal Quantile Plot

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Models

The Linear Predictor

A linear function that ties our predictors to the mean of the distribution.

$$x_i^T \beta = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

Exactly what it sounds like: a linear combination of predictors we specify.

$x_i^T \beta$ can take any value, we may need to map it into a meaningful range…

Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Models

The Link Function

Converts linear predictor into the correct range for the distribution’s mean.

$$x_i^T \beta = g(\mu) \qquad g^{-1}\left(x_i^T \beta\right) = \mu$$

Some important link functions

1. Identity: $g^{-1}\left(x_i^T \beta\right) = x_i^T \beta$ maps into $(-\infty, \infty)$
2. Log: $g^{-1}\left(x_i^T \beta\right) = \exp(x_i^T \beta)$ maps into $(0, \infty)$
3. Logit: $g^{-1}\left(x_i^T \beta\right) = {}^{1}\!/_{1+\exp(-x_i^T \beta)}$ maps into $(0,1)$

There are plenty of others, but these are the big ones.

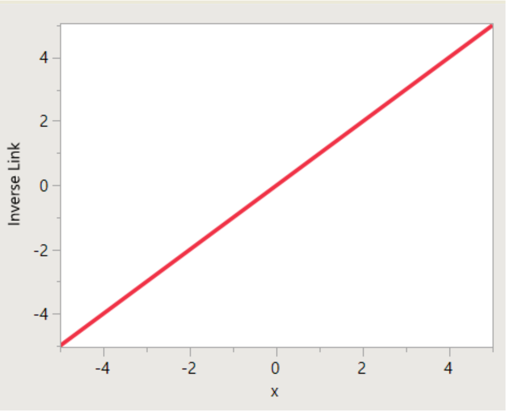
Note: Genreg picks the most appropriate link for you.

Copyright © SAS Institute Inc. All rights reserved.

Generalize Linear Models

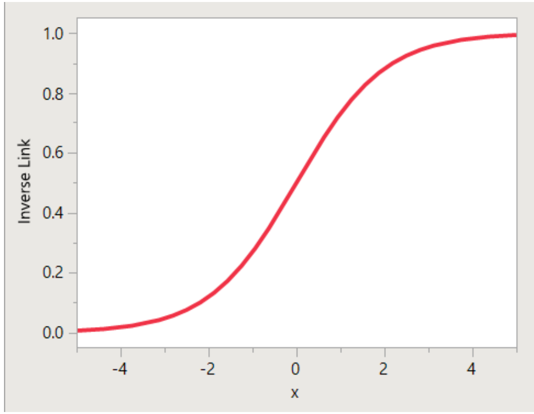
Inverse Link Functions

Identity



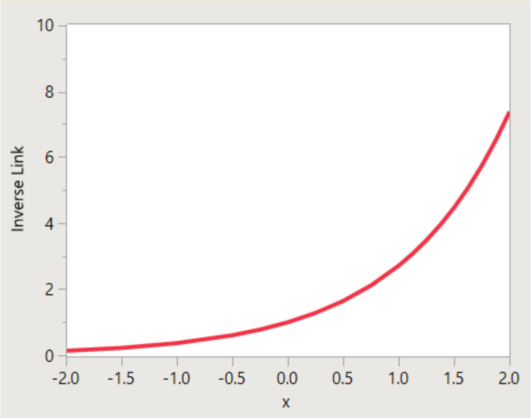
For when the response can take any value

Logit



The response should be in [0,1] (probably probabilities)

Log



The response needs to be positive

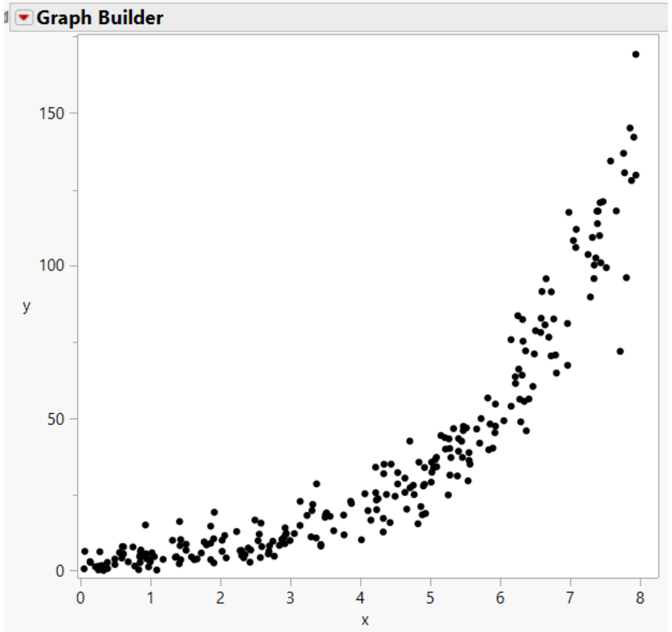
Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Model

An Example

Put the three pieces together and what do we have?

Let’s look at a simple example.



Y increases as a function of X

Y seems to become more variable with X (less obvious)

The Gamma is a natural choice.

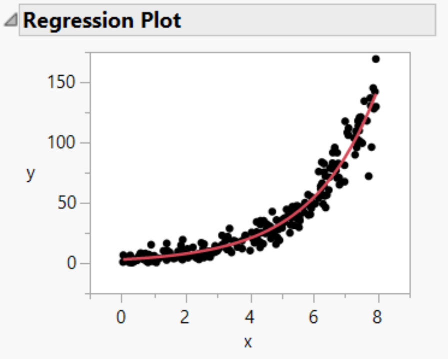
Gamma is defined for $y \in (0, \infty)$, so the log link makes sense.

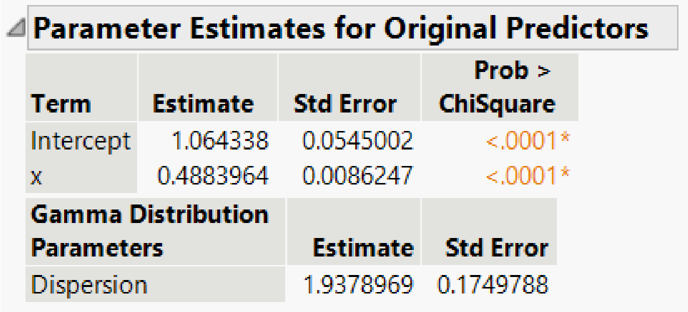
Copyright © SAS Institute Inc. All rights reserved.

Generalized Linear Model

An Example

Genreg output





Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Prob > ChiSquare
Intercept	1.064338	0.0545002	<.0001*
x	0.4883964	0.0086247	<.0001*

Gamma Distribution Parameters	Estimate	Std Error
Dispersion	1.9378969	0.1749788

This is the linear predictor.

Recall our model looks like

$$y|x \sim \text{Gamma}(\mu, \sigma) \qquad \mu = \exp(\beta_0 + \beta_1 x)$$

And $\hat{\beta}_0 = 1.064 \qquad \hat{\sigma} = 1.938$

$\hat{\beta}_1 = .488$

Copyright © SAS Institute Inc. All rights reserved.

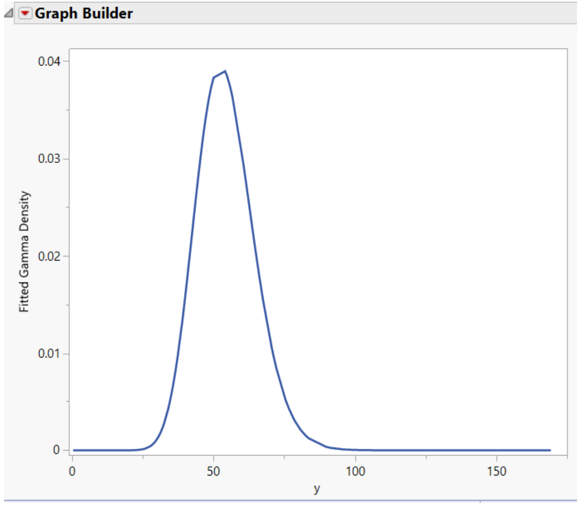
Generalized Linear Model

An Example

So what does that tell us about our response at say, x=6?

$$\exp(1.064 + .488 * 6) \approx 54.2$$

$$y|x_{=6} \sim \text{Gamma}(54.2, 1.938)$$



$$E(y|x = 6) = 54.2$$
$$\text{Var}(y|x = 6) = 54.2 * 1.938$$

Copyright © SAS Institute Inc. All rights reserved.

Choosing a Distribution

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

We probably know some things about our response.

1. Is it always positive?
2. Is it always integer valued?
3. Is the variance constant or is it proportional to the mean?
4. Is the response a proportion?
5. Is it even numeric?

Using what we know about the response, we can usually narrow it down to a couple of distributions.

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

Positive Responses

If your response is always positive, that narrows it down a little.

Ex: Most physical measurements, time, ...

Consider strictly positive distributions with a log link.

Some natural choices:

1. Gamma and Exponential
2. Lognormal
3. Weibull

And if we know we have count data...

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

Count Data

Is it a binomial?

Are we counting independent events for a given number of trials?

Ex: Number of heads out of 10 coin flips?

...Or is it Beta-Binomial?

Are we counting correlated events for a given number of trials?

The correlation causes the response to be more variable.

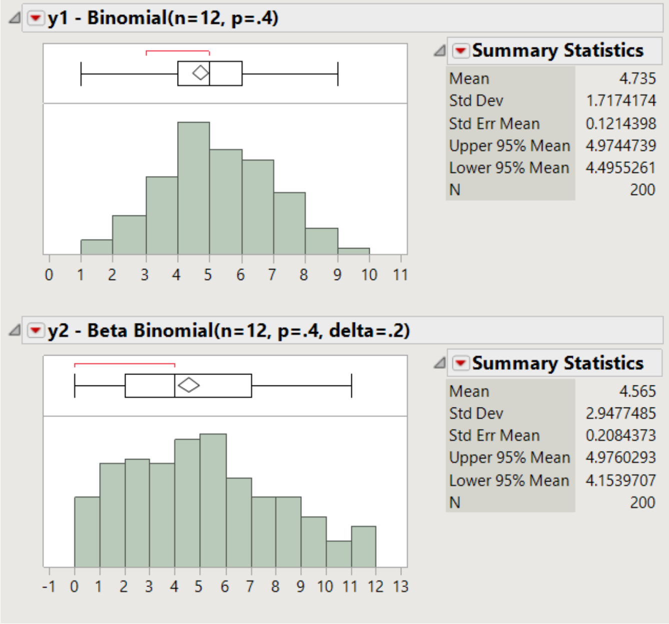
Ex: Number of shots made out of 10 attempts in a basketball game.

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

Binomial vs Beta-Binomial

200 simulated observations from each distribution



y1 - Binomial(n=12, p=.4)

Summary Statistics

Mean	4.735
Std Dev	1.7174174
Std Err Mean	0.1214398
Upper 95% Mean	4.9744739
Lower 95% Mean	4.4955261
N	200

y2 - Beta Binomial(n=12, p=.4, delta=.2)

Summary Statistics

Mean	4.565
Std Dev	2.9477485
Std Err Mean	0.2084373
Upper 95% Mean	4.9760293
Lower 95% Mean	4.1539707
N	200

Binomial

$$\mathrm{E}(y) = np$$
$$\mathrm{Var}(y) = np(1-p)$$

Beta-Binomial

$$\mathrm{E}(y) = np(1-p)$$
$$\mathrm{Var}(y) = np(1-p)[1+(n-1)\delta]$$

Copyright © SAS Institute Inc. All rights reserved.

Building Models

Count Data

What if we're not counting binary outcomes?

Ex: Number of defects on a product

Number of cars that pass through an intersection in a day

Then we probably need to use the Poisson distribution.

The Poisson is unique in that $\mathrm{E}(y) = \mathrm{var}(y) = \lambda$

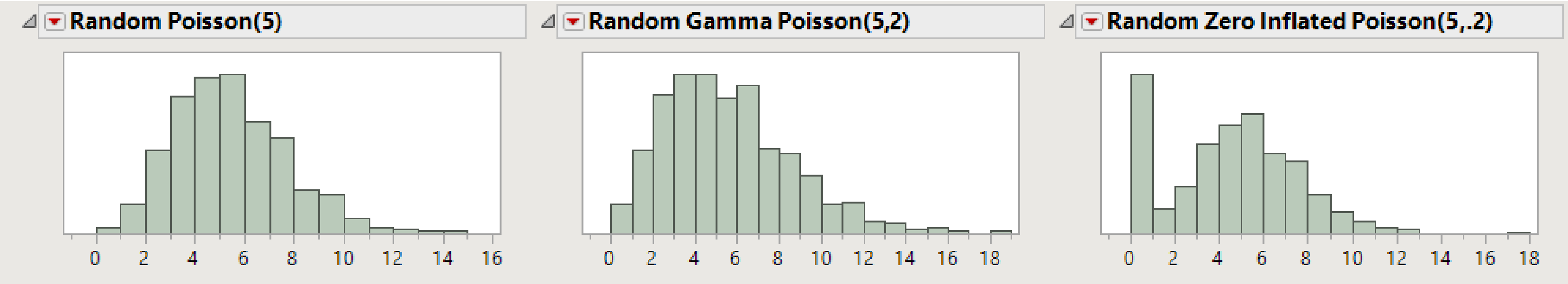
And if we need to accommodate overdispersion (extra variance)?

Choose the Negative-Binomial where $\mathrm{E}(y) = \lambda$ $\quad$ $\mathrm{var}(y) = \sigma\lambda$

Copyright © SAS Institute Inc. All rights reserved.

Building Models

Three flavors of the Poisson



Poisson(λ)
E(y) = λ
Var(y) = λ

Need extra variation?
Gamma Poisson(λ, σ)
E(y) = λ
Var(y) = $\lambda\sigma$
Also known as the
Negative Binomial.

Need extra zeros?
ZI-Poisson(λ, π)
E(y) = $(1 - \pi)\lambda$
Var(y) = $\lambda(1 - \pi)(1 + \lambda\pi)$

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

Coefficient of Determination

From working with least-squares models, we all know and love R^2

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

For GLMs, R^2 isn't useful since we're not working with square loss.

What is R^2 measuring? How well a model fits compared to the mean.

We can extend that idea to GLMs.

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

Generalized R-square

For generalized linear models,

$$R_g^2 = 1 - \left(\frac{L_0}{L_M}\right)^{2/n}$$

Where $L_0 =$ likelihood for an intercept only model

$L_M =$ likelihood for our fitted model.

If our model isn't very good, $L_M \approx L_0$ and R_g^2 will be close to zero.

If our model is great, $L_M \gg L_0$ and R_g^2 will be close to 1.

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

Be Careful!

Let’s say we narrow it down to the gamma or lognormal for our model.

Gamma: $R_g^2 = .85$.

Lognormal: $R_g^2 = .95$.

Is the lognormal model better? Maybe.

...but maybe the intercept-only lognormal model just fits very poorly.

We can use the R-square to compare models **within** a distribution.

To compare **between** distributions, use an information criteria.

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

Information Criteria

The AIC and BIC are information criteria that we use to compare models.

AIC = $2p - 2\log(L)$

AICc = $2p - 2\log(L) + \frac{2p(p+1)}{n-p-1}$ small sample correction

BIC = $\log(n) * p - 2\log(L)$

where *p* is the number of parms fit, *L* is the likelihood, and *n* is sample size.

These measures balance model fit with model complexity.

Smaller values are better.

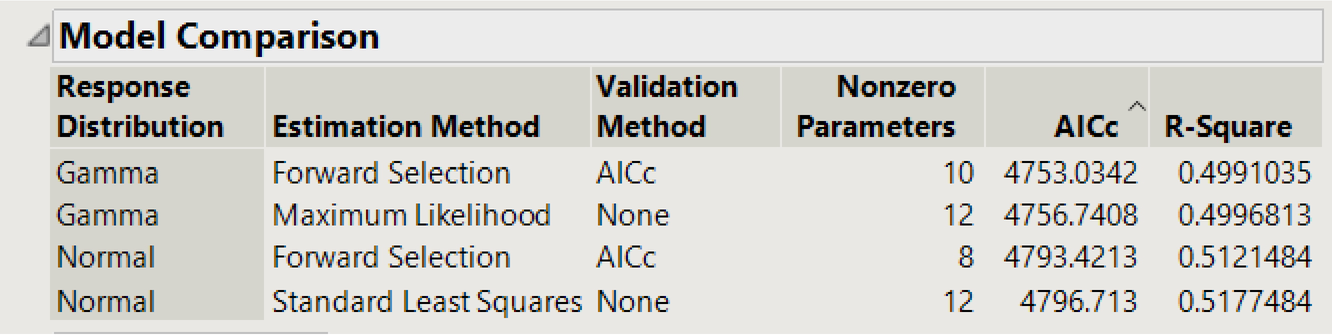
Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

Information Criteria

The AIC and BIC estimate the Kullback-Leibler divergence, which is the distance from the fitted model to the truth.

So we can use them to compare models **within** the same distribution and **across** different distributions.



Model Comparison

Response Distribution	Estimation Method	Validation Method	Nonzero Parameters	AICc	R-Square
Gamma	Forward Selection	AICc	10	4753.0342	0.4991035
Gamma	Maximum Likelihood	None	12	4756.7408	0.4996813
Normal	Forward Selection	AICc	8	4793.4213	0.5121484
Normal	Standard Least Squares	None	12	4796.713	0.5177484

Copyright © SAS Institute Inc. All rights reserved.

Evaluating Models

AICc and BIC

The AICc and BIC are great all-purpose tools for

- …comparing 2 or more models
- …that don’t have to be nested
- …that don’t even have to be from the same response distribution

We do need a likelihood and degree of freedom, which can be a limitation.

Ex: The degrees of freedom for a tree isn’t well defined.

Rule of thumb: AIC tends to overfit and BIC tends to underfit.

Copyright © SAS Institute Inc. All rights reserved.

Choosing the Response Distribution

Discrete Responses

Using our intuition, we can narrow it down to a few distributions and then use the AICc or BIC to guide us.

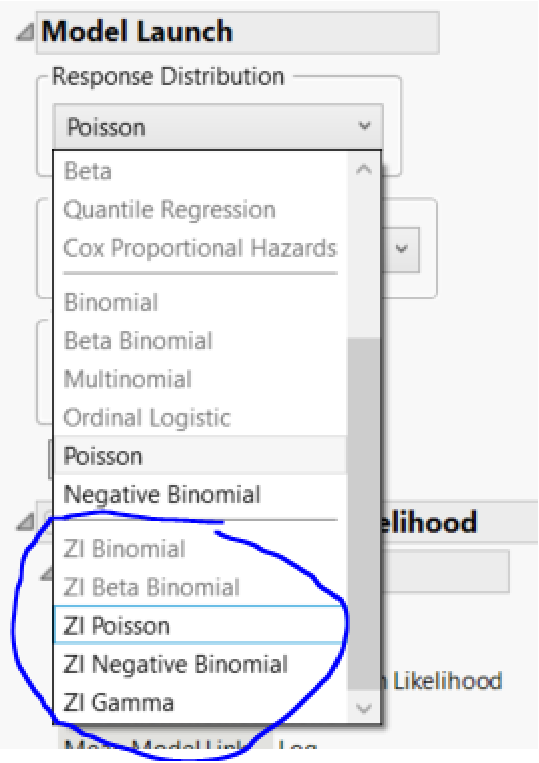
If we have count data…usually we think of the Poisson.

- Events out of trials? -> binomial or beta-binomial
- Do we need to account for overdispersion? -> negative binomial
- Do we have extra zeros? -> zero-inflated distribution
- Only observe a couple of distinct values? -> consider switching to logistic

Copyright © SAS Institute Inc. All rights reserved.

Choosing the Response Distribution

Discrete Responses with extra zeros



Response Distribution: Poisson

The ZI stands for “Zero Inflated”.
…which means there are more zeros than otherwise expected.

Suppose we ask folks leaving a park how many fish they caught while visiting.
There are two ways to observe a zero.

1. A visitor doesn’t fish
2. A visitor fishes, but doesn’t catch anything.

Copyright © SAS Institute Inc. All rights reserved.

Choosing the Response Distribution

Continuous Responses

And if we have a continuous response…

- Do we have negative values? -> normal
- Is it bound to (0,1)? -> beta
- Does variance increase with the mean? -> gamma, Weibull, lognormal
- Is it time-to-event/censored? -> probably Weibull or lognormal
- A pretty good catch-all? -> normal
- Do we suspect that we have outliers? -> Cauchy or t(5).

Copyright © SAS Institute Inc. All rights reserved.

Choosing the Response Distribution

What if our response isn’t even numeric???

- Is it two-level? Use the binomial.

Ex: Yes/No or A/B.

- 3+ levels and order matters? Use Ordinal logistic.

Ex: Low/Medium/High or Small/Medium/Large.

- 3+ levels and order doesn’t matter? Use the Multinomial.

Ex: Pizza/Hamburger/Burrito or Red/Blue/Green/Orange.

Copyright © SAS Institute Inc. All rights reserved.

Wrap-up

Copyright © SAS Institute Inc. All rights reserved.

Wrap Up

- GLMs are an important piece of your modeling toolbox.
- Genreg makes them easy to fit and use.
- How should I choose the response distribution?
  - Narrow it down to a handful of meaningful options
  - Compare AICc or BIC values to pick the “best”.
  - When all else fails, the Normal and Gamma are a good start.

Copyright © SAS Institute Inc. All rights reserved.

Thanks!

Clay.Barker@sas.com

sas.com

Copyright © SAS Institute Inc. All rights reserved.