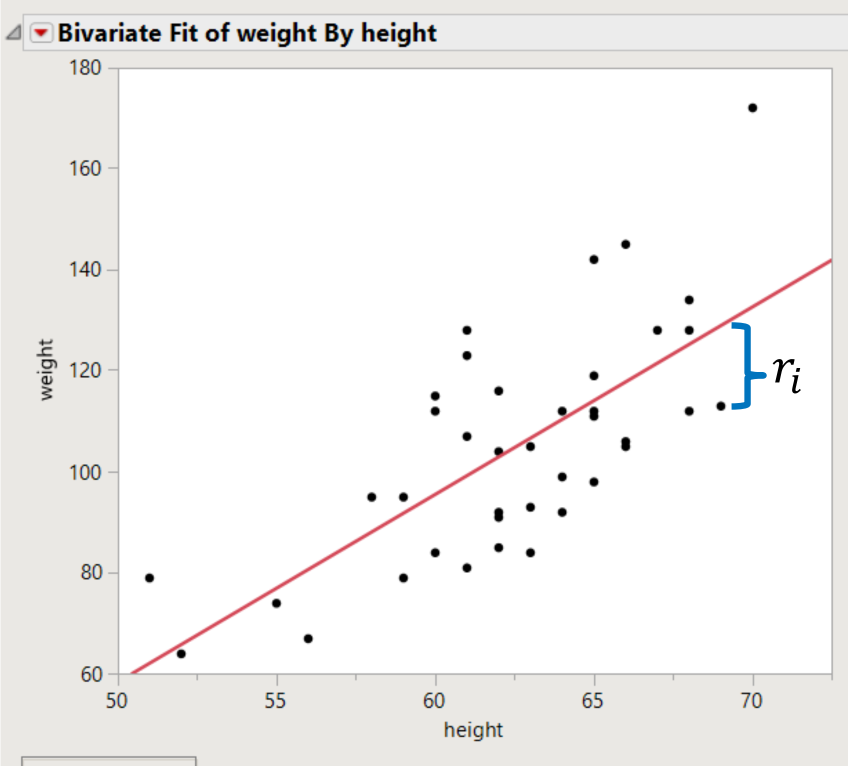
Not Quite Normal

Choosing the Best Distribution to Model your Response

Clay Barker, PhD

JMP Principal Research Statistician Developer

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 a lot easier.

Simple Linear Regression

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

Why?

Things like estimation and inference become much easier.

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 linear regression models, but there are times when it just is not appropriate.

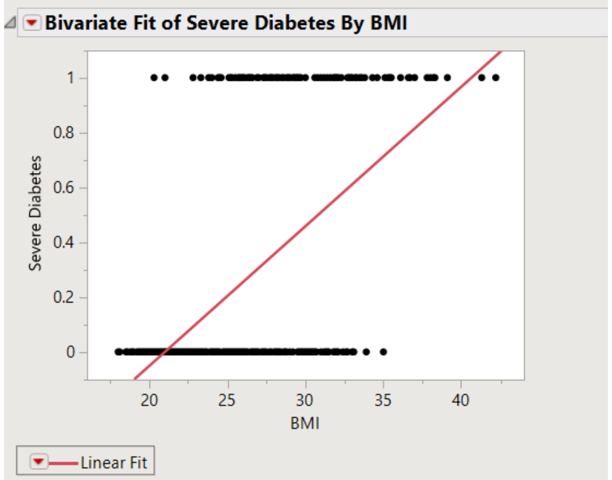
Simple Linear Model

Normality

So what happens when 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)



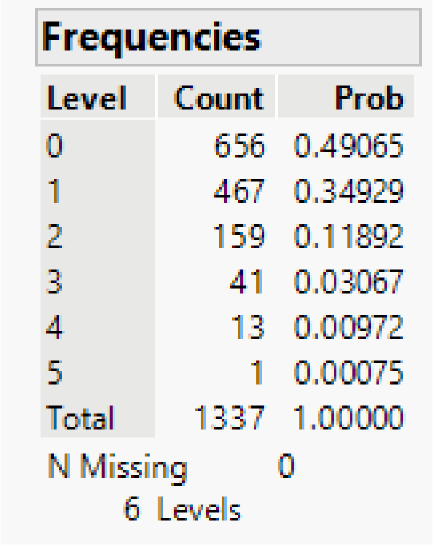
What if normality is not appropriate?

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

What might impact performance?

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

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 takes integer values.

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

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

The normality assumption doesn't seem appropriate at all here, but we still need to build a model.

What should we do?

Overview

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

Generalized Linear Models

A quick overview

Generalized Linear Models

But first, back to the linear model

Our beloved linear model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i \\ &= x_i^T \beta + \epsilon_i \end{aligned}$$

where x_i is a $p + 1$ 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

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

Generalized Linear Model

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

And there are lots of situations where $y_i|x_i$ may not be normal

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)

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.

Generalized Linear Model

Example

Normal

$$f(y;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$=\exp\left(\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{y^2}{2\sigma^2}-\log(\sigma)-\log(\sqrt{2\pi})\right)$$

So…

$\theta=\mu$ $\quad b(\theta)=\frac{\theta^2}{2}$

$\phi=\sigma$ $\quad a(\phi)=\phi^2$ $\quad c(y,\phi)={}^{y^2}\!/_{2\phi^2}+\log(\phi)+\log\left(\sqrt{2\pi}\right)$

Generalized Linear Model

Example

Poisson

$$f(y,\lambda) = \frac{\exp(-\lambda)\lambda^{y}}{y!}$$

$$= \exp(-\lambda + y\log(\lambda) - \log(y!))$$

And so…

$\theta = \log(\lambda)$ $b(\theta) = \exp(\theta)$

$a(\phi) = 1$ $c(y,\phi) = \log(y!)$

Generalized Linear Models

Do we really need to know about exponential families and all this?

…Not really, it’s more of a homework problem.

Instead we’ll focus on the most important parts of using GLMs.

Generalized Linear Models

There are 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)

Generalized Linear Model

The Distribution

When we specify the distribution of a GLM, we're specifying the distribution of the **response given the predictors**.

This is a critical piece!

In general it **is not** the distribution of the residuals. (Exception: Normal, ...)

Generalized Linear Model

The Distribution

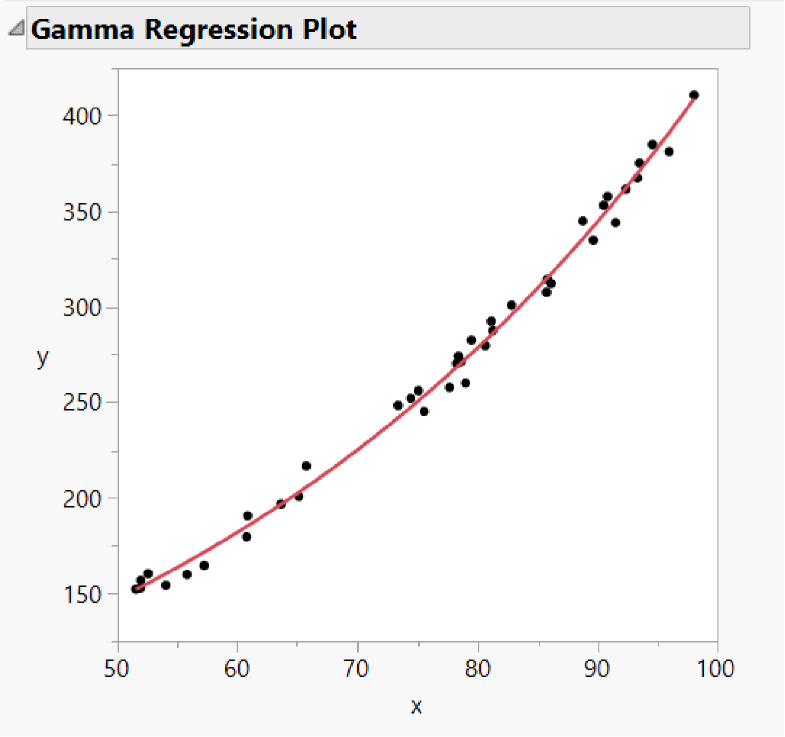
Because of the Normal distribution, it's easy to think that we're talking about the distribution of the residuals.

How can we avoid this mistake?

Here is a helpful reminder:

The Gamma distribution is strictly positive.

The residuals for this Gamma regression are positive and negative.



Generalized Linear Model

The Distribution

Another common mistake to avoid:

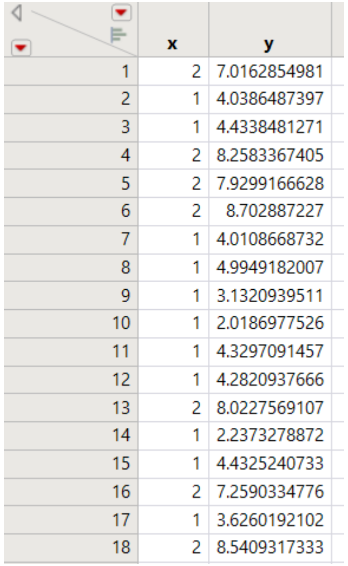
The distribution **is not** the distribution of the response, it’s the distribution of the response given the predictors.

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

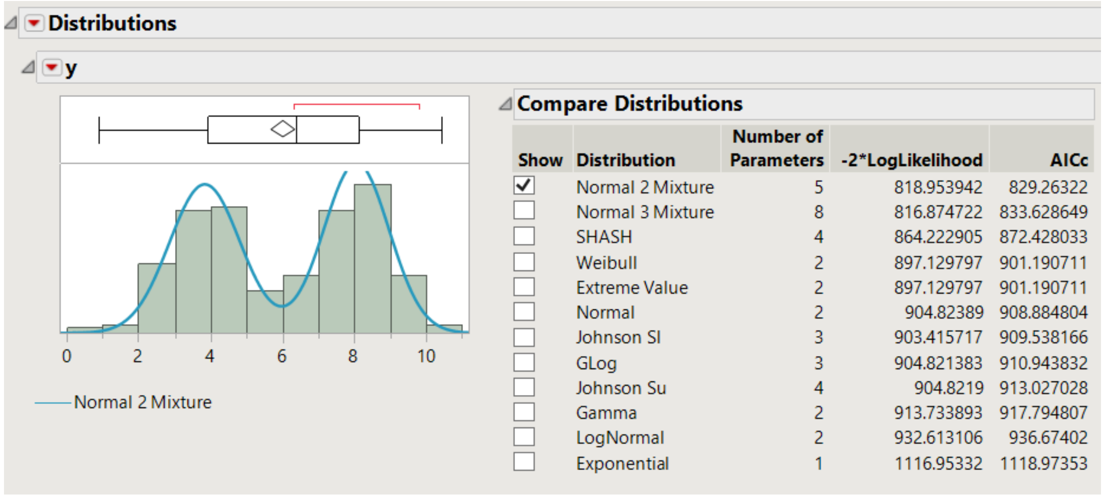
Generalized Linear Model

The Distribution

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

Generalized Linear Model

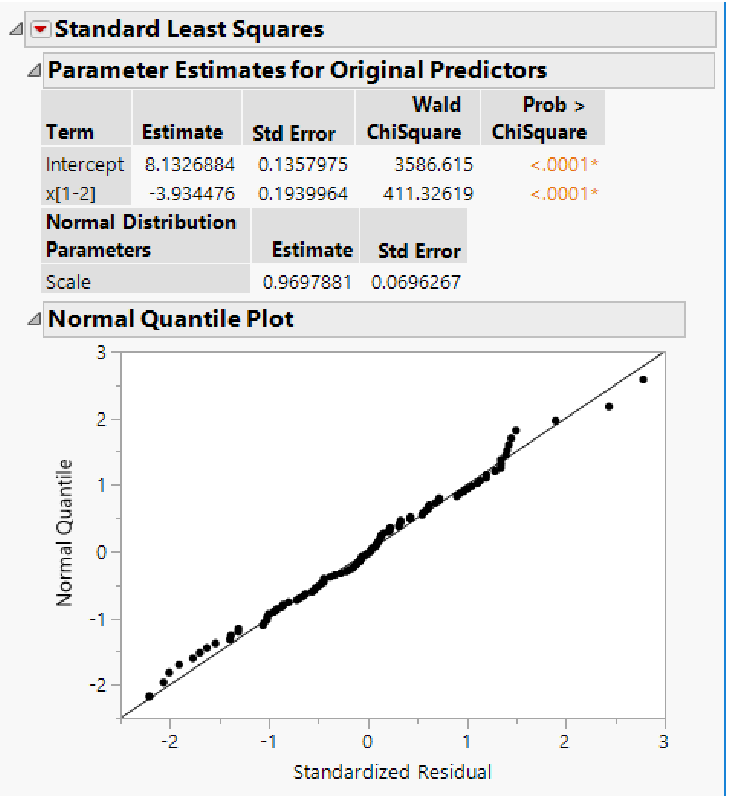
The Distribution

100% No!

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 be careful using it to choose a distribution.



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

Generalized Linear Models

The Linear Predictor

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

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

Pretty much exactly what it sounds like.

In ordinary least squares, our model is just the linear predictor.

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

Generalized Linear Models

The Link Function

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

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

Some important link functions

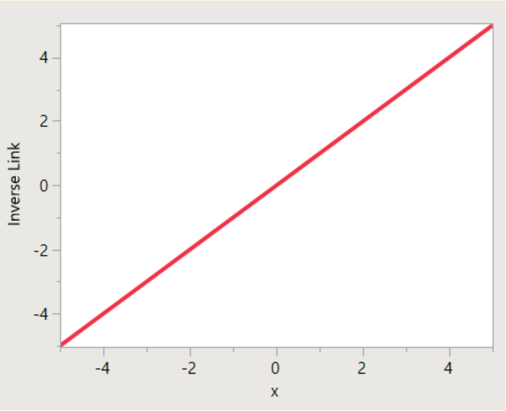
1. Identity: $g^{-1}(x_i^T\beta) = x_i^T\beta$ maps into $(-\infty, \infty)$
2. Log: $g^{-1}(x_i^T\beta) = \exp(x_i^T\beta)$ maps into $(0, \infty)$
3. Logit: $g^{-1}(x_i^T\beta) = {}^{1}/_{1+\exp(-x_i^T\beta)}$ maps into $(0,1)$

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

Generalize Linear Models

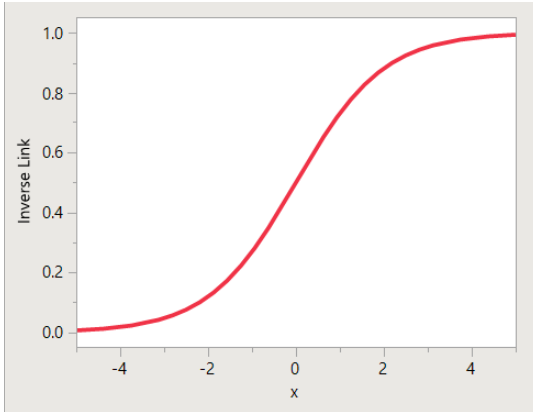
Inverse Link Functions

Identity



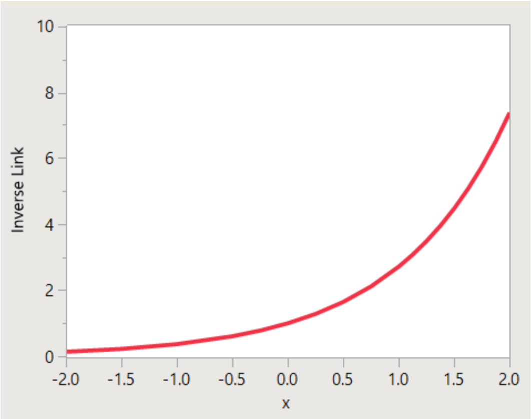
For when the response can take any value

Logit



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

Log



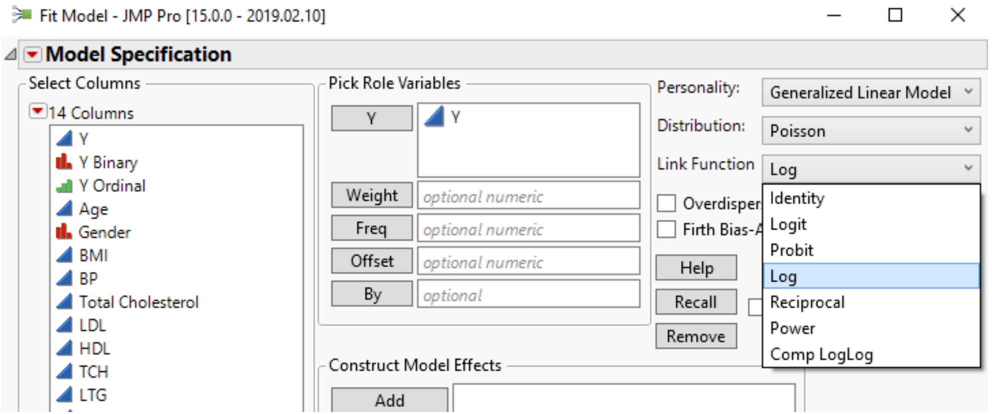
The response needs to be positive

Generalized Linear Models

The Link Function

We choose the link to convert an unbounded number to an appropriate range for the response…that’s the key.

Fit Model’s Generalized Linear Model personality lets you choose the link function, otherwise JMP uses the one that makes the most sense.



Fit Model - JMP Pro [15.0.0 - 2019.02.10]

Model Specification

Personality: Generalized Linear Model

Distribution: Poisson

Link Function: Log

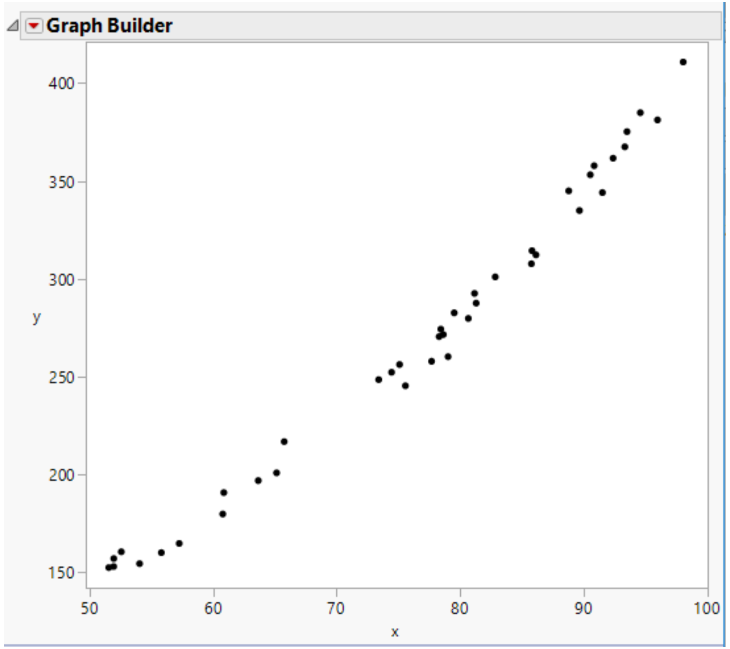
Y: Y

Generalized Linear Model

An Example

Put the three pieces together and what do we have?

Let’s look at a simple gamma example.



Y increases as a function of X

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

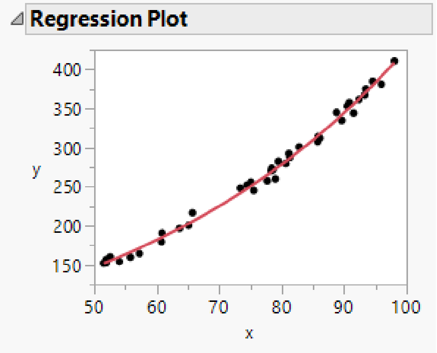
These make the Gamma a natural choice.

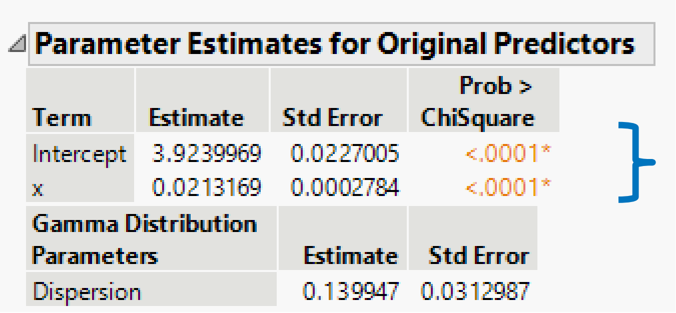
Gamma is defined for $y \in (0, \infty)$, so the log link is a natural choice.

Generalized Linear Model

An Example

Output from the Generalized Regression Platform in JMP Pro.





Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Prob > ChiSquare
Intercept	3.9239969	0.0227005	<.0001*
x	0.0213169	0.0002784	<.0001*

Gamma Distribution Parameters	Estimate	Std Error
Dispersion	0.139947	0.0312987

$\}$ 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 $\hat{\beta}_0 = 3.924 \qquad \hat{\sigma} = .140$

$\hat{\beta}_1 = .0213$

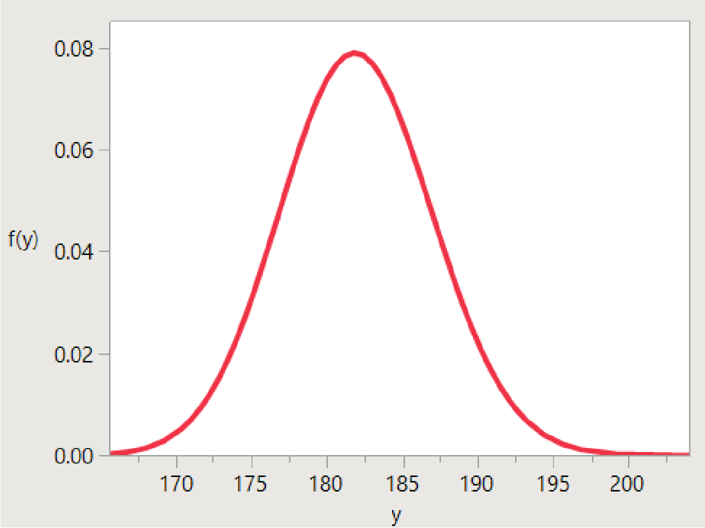
Generalized Linear Model

An Example

So what does that tell us about our response at say, x=60?

$$\exp(3.924 + .0213 * 60) \approx 181.6$$

$$y|x_{=60} \sim \text{Gamma}(181.6, .14)$$



$$E(y|x) = 181.6$$

$$\text{Var}(y|x) = 181.6 * .14$$

Fitting GLMs in JMP

Fitting GLMs in JMP

JMP fits GLMs in a variety of places…

1. Generalized Linear Model
2. Nominal and Ordinal Logistic
3. Generalized Regression in JMP Pro

And in some places that we won't cover…

1. Fit Y by X (simple logistic regression)
2. Parametric Survival (for when you have a censored response)
3. Choice/MaxDiff (very specific)

Fitting GLMs in JMP

Generalized Linear Model Personality

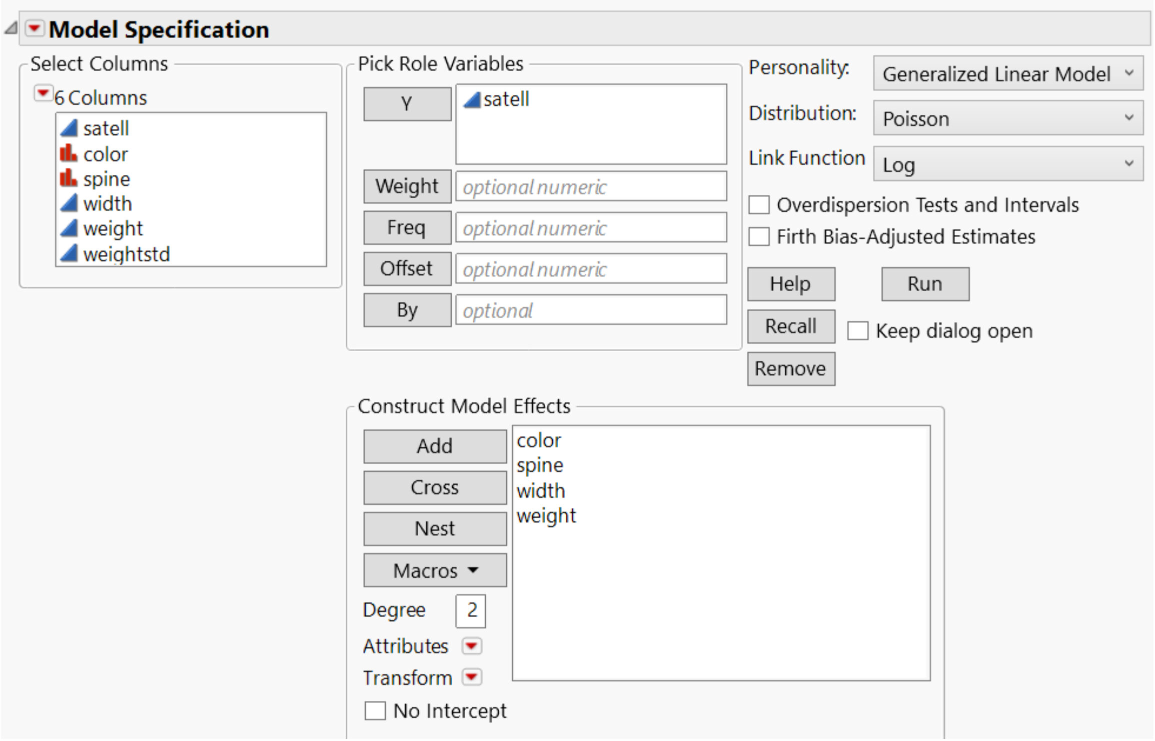
A personality within Fit Model. This is a great place to go when…

1. You want to pick your link function (logit, probit, log, …)
2. One of the available distributions is appropriate (Normal, Poisson, Binomial, and exponential)

Go to Analyze>Fit Model and choose the GLM Personality

Fitting GLMs in JMP

The GLM Platform



Model Specification

Select Columns

6 Columns

- satell
- color
- spine
- width
- weight
- weightstd

Pick Role Variables

Y: satell

Weight: *optional numeric*

Freq: *optional numeric*

Offset: *optional numeric*

By: *optional*

Personality: Generalized Linear Model

Distribution: Poisson

Link Function: Log

□ Overdispersion Tests and Intervals

□ Firth Bias-Adjusted Estimates

□ Keep dialog open

Construct Model Effects

- color
- spine
- width
- weight

Degree: 2

□ No Intercept

Fitting GLMS in JMP

Nominal and Ordinal Logistic Regression

If our response is not continuous, then we should use the Nominal Logistic or Ordinal Logistic personalities in Fit Model.

If our response takes ordered values, us ordinal logistic.

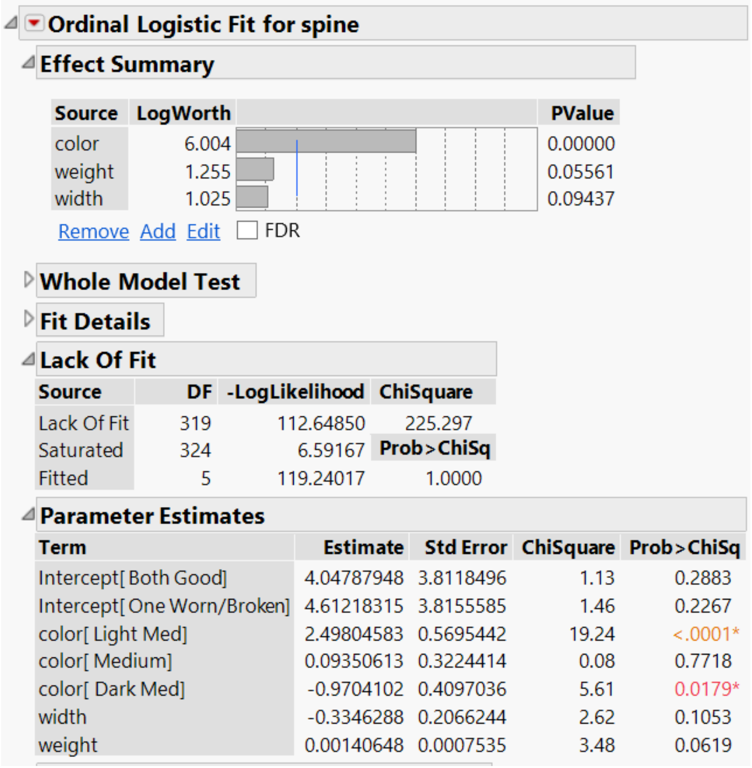
Ex: Y={low, medium, high, alert}

Otherwise, go with Nominal Logistic. But keep in mind, Nominal Logistic models use up a lot of parameters.

Ex: Y={red, blue, green, yellow}   or   Y={yes, no}

Fitting GLMs in JMP

Ordinal and Nominal Logistic



Ordinal Logistic Fit for spine

Effect Summary

Source	LogWorth	PValue
color	6.004	0.00000
weight	1.255	0.05561
width	1.025	0.09437

Remove Add Edit □ FDR

Whole Model Test

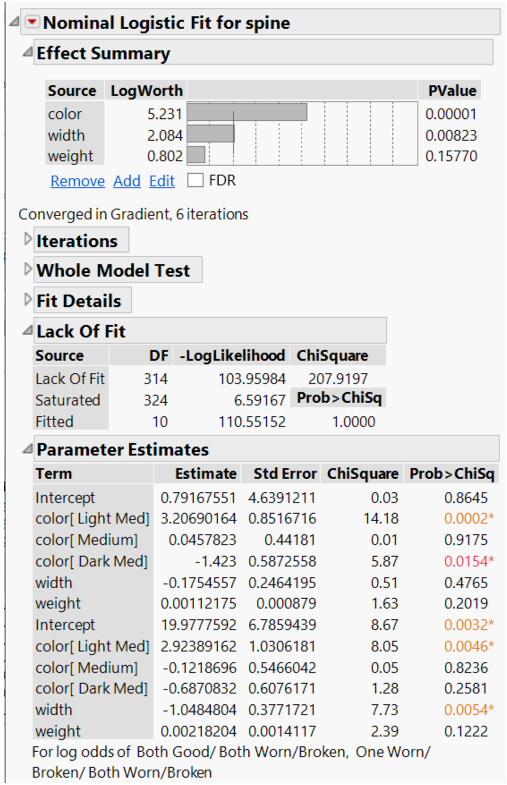
Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	319	112.64850	225.297
Saturated	324	6.59167	Prob>ChiSq
Fitted	5	119.24017	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[Both Good]	4.04787948	3.8118496	1.13	0.2883
Intercept[One Worn/Broken]	4.61218315	3.8155585	1.46	0.2267
color[Light Med]	2.49804583	0.5695442	19.24	<.0001*
color[Medium]	0.09350613	0.3224414	0.08	0.7718
color[Dark Med]	-0.9704102	0.4097036	5.61	0.0179*
width	-0.3346288	0.2066244	2.62	0.1053
weight	0.00140648	0.0007535	3.48	0.0619



Nominal Logistic Fit for spine

Effect Summary

Source	LogWorth	PValue
color	5.231	0.00001
width	2.084	0.00823
weight	0.802	0.15770

Remove Add Edit □ FDR

Converged in Gradient, 6 iterations

Iterations

Whole Model Test

Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	314	103.95984	207.9197
Saturated	324	6.59167	Prob>ChiSq
Fitted	10	110.55152	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	0.79167551	4.6391211	0.03	0.8645
color[Light Med]	3.20690164	0.8516716	14.18	0.0002*
color[Medium]	0.0457823	0.44181	0.01	0.9175
color[Dark Med]	-1.423	0.5872558	5.87	0.0154*
width	-0.1754557	0.2464195	0.51	0.4765
weight	0.00112175	0.000879	1.63	0.2019
Intercept	19.9777592	6.7859439	8.67	0.0032*
color[Light Med]	2.92389162	1.0306181	8.05	0.0046*
color[Medium]	-0.1218696	0.5466042	0.05	0.8236
color[Dark Med]	-0.6870832	0.6076171	1.28	0.2581
width	-1.0484804	0.3771721	7.73	0.0054*
weight	0.00218204	0.0014117	2.39	0.1222

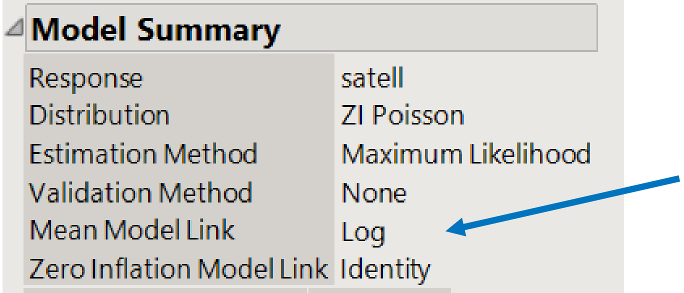
For log odds of Both Good/ Both Worn/Broken, One Worn/ Broken/ Both Worn/Broken

Fitting GLMs in JMP Pro

The Generalized Regression Platform

JMP Pro personality of Fit Model

1. Supports 20 distributions, including Nominal and Ordinal logistic.
2. Uses default link functions.
3. Includes 10 different variable selection methods including Lasso and step based methods.
4. Supports censoring for some distributions.



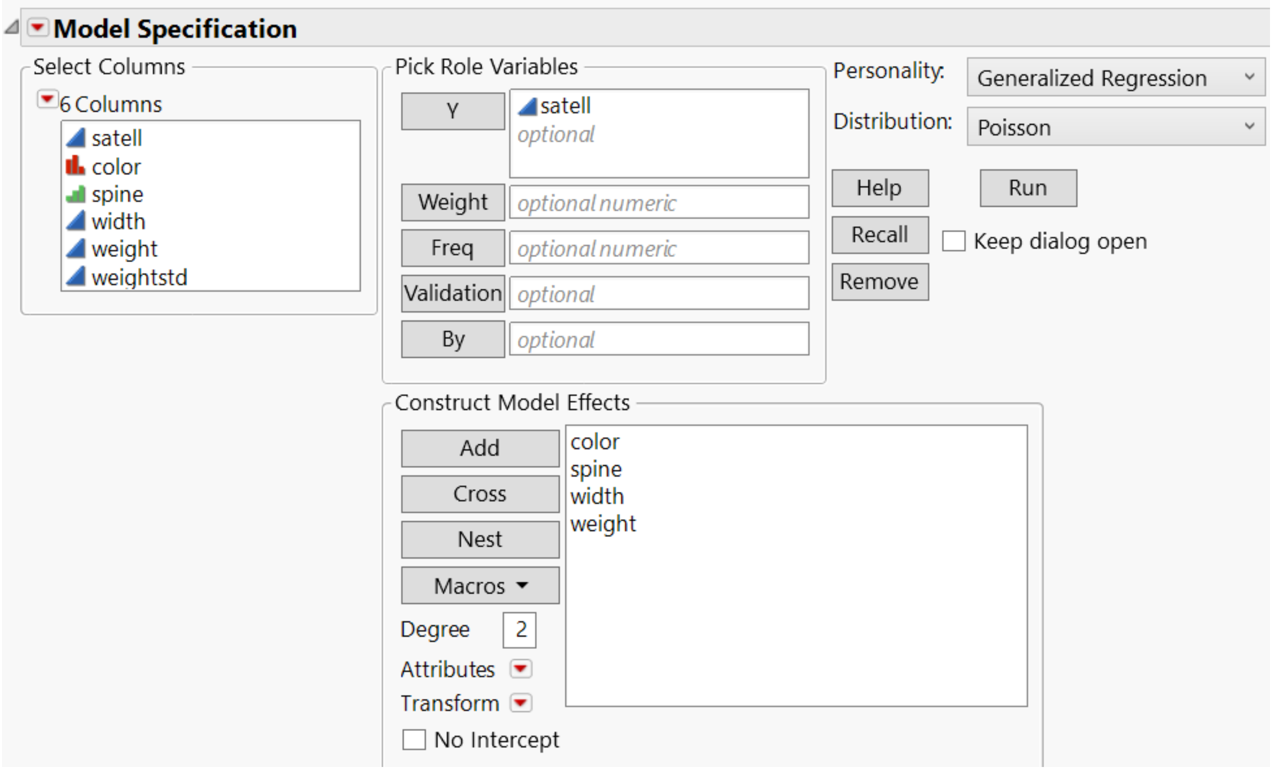
Model Summary

Response	satell
Distribution	ZI Poisson
Estimation Method	Maximum Likelihood
Validation Method	None
Mean Model Link	Log
Zero Inflation Model Link	Identity

Genreg tells you which links it used.

Fitting GLMs in JMP Pro

Genreg



Model Specification

Select Columns

6 Columns

- satell
- color
- spine
- width
- weight
- weightstd

Pick Role Variables

Y: satell

Weight: *optional numeric*

Freq: *optional numeric*

Validation: *optional*

By: *optional*

Personality: Generalized Regression

Distribution: Poisson

□ Keep dialog open

Construct Model Effects

color
spine
width
weight

Degree: 2

□ No Intercept

Choosing a Distribution

Evaluating Models

You probably know a few things about your 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.

Evaluating Models

Positive Responses

If your response is always positive and you want to insure positive predictions, that narrows it down a little….

Ex: Most physical measurements, time, …

Consider strictly positive distributions, probably with a log link.

Some natural choices:

1. Gamma and Exponential
2. Lognormal
3. Weibull

And if we know we have count data…

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.

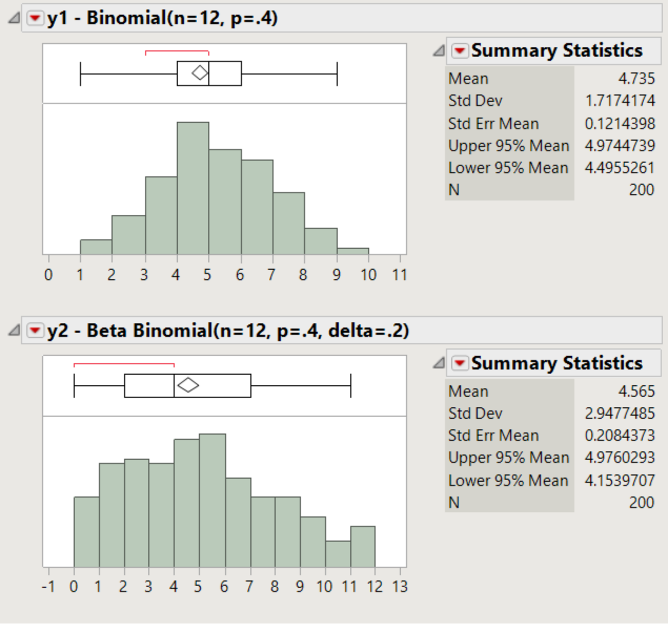
Ex: Number of shots made out of 10 attempts in a basketball game.

*The beta-binomial is available in Distribution and Genreg

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

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

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.

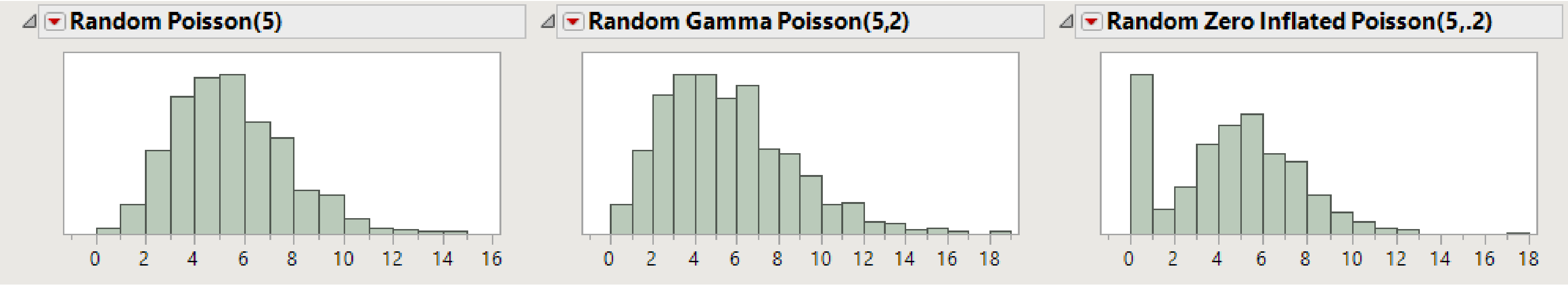
The Poisson is unique in that $\mathrm{E}(y) = \mathrm{var}(y) = \lambda$

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

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

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) = $\lambda(1 - \pi)(1 + \lambda\pi)$

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, this quantity isn’t terribly useful since we’re no longer working with square loss.

What is R^2 measuring?

How well a model fits compared to the mean, which we can extend to GLMs.

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.

Evaluating Models

Be Careful!

For our predictors, we narrowed it down to the gamma or lognormal.

For our gamma model, $R_g^2 = .85$.

For the 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, we should use an information criteria.

Evaluating Models

Information Criteria

The AIC and BIC are both popular 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.

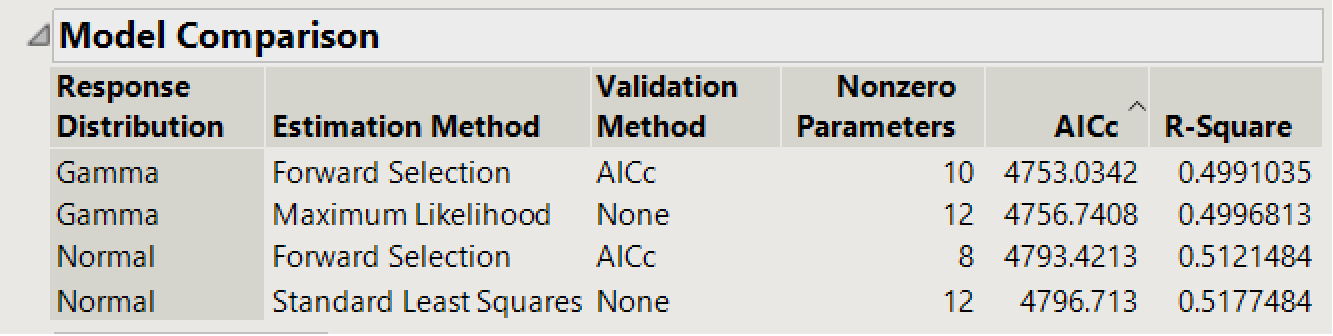
Smaller values are better.

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

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

But we have to have a likelihood and degrees 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.

Choosing the Response Distribution

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

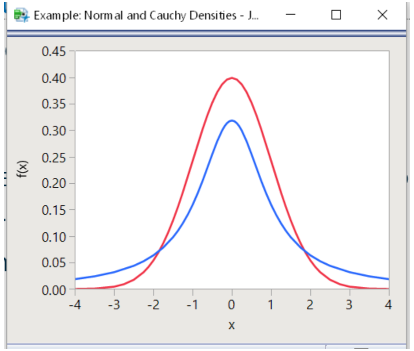
If we have count data…usually we think of the Poisson.

- 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

Choosing the Response Distribution

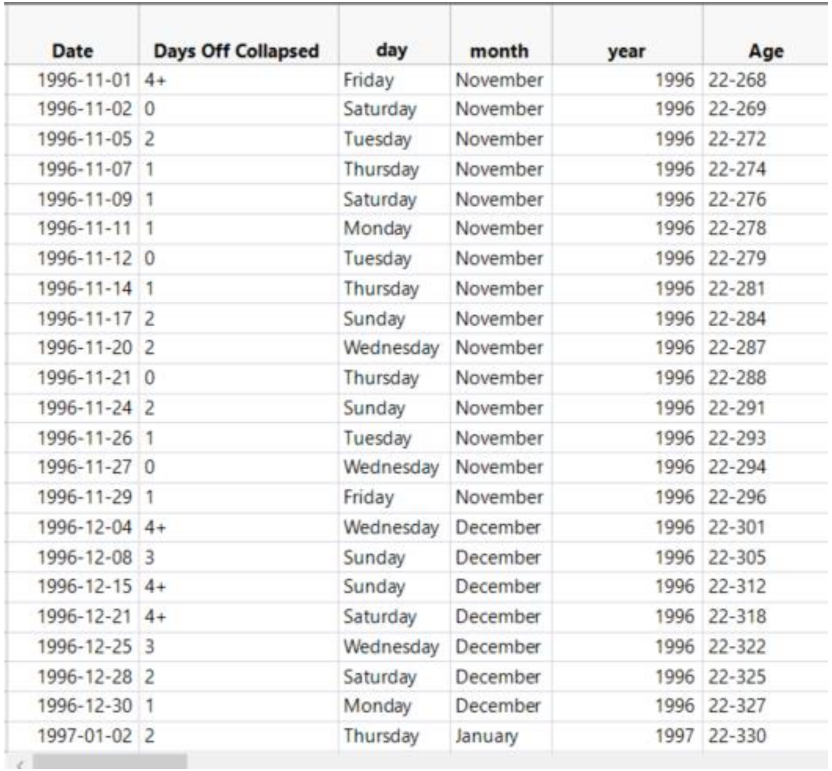
And if we have a continuous response…

- 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 expect to have outliers? -> Cauchy



Back to Our Basketball Data

Now let’s try building some models



Date	Days Off Collapsed	day	month	year	Age
1996-11-01	4+	Friday	November	1996	22-268
1996-11-02	0	Saturday	November	1996	22-269
1996-11-05	2	Tuesday	November	1996	22-272
1996-11-07	1	Thursday	November	1996	22-274
1996-11-09	1	Saturday	November	1996	22-276
1996-11-11	1	Monday	November	1996	22-278
1996-11-12	0	Tuesday	November	1996	22-279
1996-11-14	1	Thursday	November	1996	22-281
1996-11-17	2	Sunday	November	1996	22-284
1996-11-20	2	Wednesday	November	1996	22-287
1996-11-21	0	Thursday	November	1996	22-288
1996-11-24	2	Sunday	November	1996	22-291
1996-11-26	1	Tuesday	November	1996	22-293
1996-11-27	0	Wednesday	November	1996	22-294
1996-11-29	1	Friday	November	1996	22-296
1996-12-04	4+	Wednesday	December	1996	22-301
1996-12-08	3	Sunday	December	1996	22-305
1996-12-15	4+	Sunday	December	1996	22-312
1996-12-21	4+	Saturday	December	1996	22-318
1996-12-25	3	Wednesday	December	1996	22-322
1996-12-28	2	Saturday	December	1996	22-325
1996-12-30	1	Monday	December	1996	22-327
1997-01-02	2	Thursday	January	1997	22-330

Wrap-up

Wrap Up

- GLMs are an important piece of our modeling toolbox.
- The platforms in JMP make 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.

- And remember: it's the distribution of the response given the predictors.

Wrap-up

More Resources

Search the JMP Community for past blog posts and talks

A good intro to GLMs

- "Foundations of Linear and Generalized Linear Models" by Agresti (Wiley, 2015)

Lots of information about the AIC and BIC.

- “Model Selection and Multimodel Inference” by Burnham and Anderson (Springer, 2003)
- As well as Ken Burnham’s webpage

Thanks!

Clay.Barker@sas.com

sas.com