Data that are not Continuous

Matthew W Wheeler

Introduction:

All statistical tests are based upon the underlying assumed variance.

To this point, all we have really considered are normal models, which have constant variance.

For large, or really large, samples this is a pretty good approximation, but it doesn’t make sense for certain types of data.

Different Types of Data

There are many situations where what we measure is not a continuous variable:

- We could measure the number insurance claims of an individual (count)
- We can measure if an individual visits a website each day for a week (proportion/probability)
- We can measure if a person ends up in prison again after an intervention (proportion/probability)

For many situations, we can approximate continuous data, but there are other situations where it doesn’t make any sense.

The biggest issue is that when we have non-normal our typical data assumptions do not apply:

- For count data, the Poisson distribution is usually used to model such data. For this distribution the variance = mean (or rate).

- For success/failure data, the Binomial distribution is used frequently. For this distribution the variance is at a maximum when the probability of response p = 0.5 as variance = np(1-p), where n is the number of trials.

- We are going to use two popular models for such data:
  - Logistic – Binomial
  - Log-Linear – Poisson

Types of Data

Poisson

Experiments results in count y

- The rate of occurrence is λ
- Sometimes we have this observed over some unit time or area t
- We expect λ/t counts per unit.
- Variance $Y/t = \lambda/t$
- We model $\log(\lambda/t) = \beta_0 + \beta_1 x$

Binomial

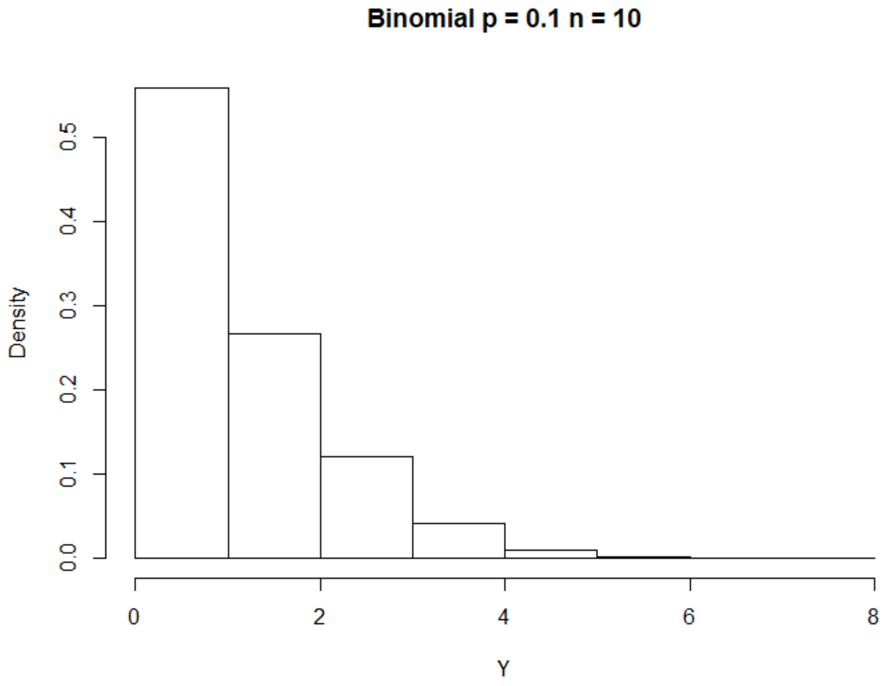
Experiment with n trials and y successes.

- Success occurs with probability *p*
- We expect *np* successes.
- The variance of y is $np(1-p)$
- We model

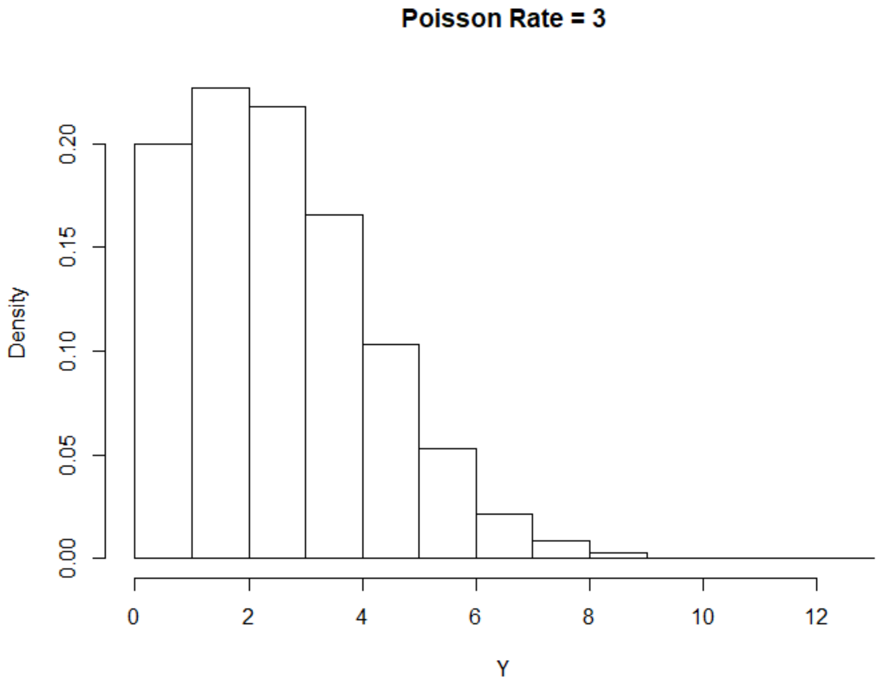
$$\log [p/(1-p)] = \beta_0 + \beta_1 x$$

How bad is non-constant variance?

Binomial



Poisson



Modeling is straightforward using the SAS proc genmod.

For us, the majority of modeling statements will be the same, which makes our life much easier.

The reason why we use these methods over glm is because our analyses will be more accurate because we are not relying on some normal approximation.

When I have huge sample sizes. Normality is a safe assumption.

Example 1

An insurer is interested in looking at the rate of claims by the age of the person insured and the size of the car. A random sample of the companies policies are taken and we are interested in seeing if the claim rate is different between the car size and age.

Response: Number of Claims

Offset: Number of Policies

Variables: Age/Size

Data:

```
/*insurance data*/
data insure;
      input n claims car$ age;
      ln = log(n);
      datalines;
   500   42  small  1
   1200  37  medium 1
   100    1  large  1
   400  101  small  2
   500   73  medium 2
   300   14  large  2
   ;
   run;
```

```
/*initial cut at the data*/
proc genmod data=insure;
              class car age/param=glm; *param=glm codes it like glm
                                   *so we can think of it in the same way;
              model claims = car age/dist   = poisson
                                  /*our data is count data we use
                                   the poisson distribution*/
                                  link   = log TYPE3
                                  /*TYPE3 IS LIKE TYPE III SUMS OF SQUARES*/
                              offset = ln; /*THIS IS OVER A NUMBER OF CUSTOMERS*/
              contrast 'Large vs. small' car 1 0 -1;
       run;
quit;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	2.8207	1.4103
Scaled Deviance	2	2.8207	1.4103
Pearson Chi-Square	2	2.8416	1.4208
Scaled Pearson X2	2	2.8416	1.4208
Log Likelihood		837.4533	
Full Log Likelihood		-16.4638	
AIC (smaller is better)		40.9276	
AICC (smaller is better)		80.9276	
BIC (smaller is better)		40.0946	

This should be close to 1 if it is not, we need to adjust so our tests statistics are more correct. The value 1.4 is good enough.

Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
Large vs. small	1	65.37	<.0001	LR

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
car	2	72.82	<.0001
age	1	104.64	<.0001

Tests are called Likelihood Ratio Tests

All tests are Likelihood ratio tests:

They are similar to our standard F-test, but are used in cases where we do not have normality.

They are essentially the fit of the model with and without the parameters. The -2 Log of this ratio is distributed approximately chi-squared with degrees of freedom the difference in parameters constrained.

There are two parameters constrained by car.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
car	2	72.82	<.0001
age	1	104.64	<.0001

There is one parameter constrained by age

Though there are a lot of difference theoretically between normal glm based ANOVA and poisson genmod ANALYSIS OF DEVIANCE, from a SAS perspective (or R) there are not that many differences when you are analyzing the data.

Further, you design the study in THE EXACT SAME MANNER as anything we have done so far. All you have to worry about is the analysis.

Analysis 2

An experiment is conducted to look at the probability an ingot is not ready for rolling after a several treatments. We are interested in finding the optimal conditions for rolling.

```
/*example 2 ingots data*/
data ingots;
   input Heat Soak r n @@;
   datalines;
7 1.0 0 10  14 1.0 0 31  27 1.0 1 56  51 1.0 3 13
7 1.7 0 17  14 1.7 0 43  27 1.7 4 44  51 1.7 0  1
7 2.2 0  7  14 2.2 2 33  27 2.2 0 21  51 2.2 0  1
7 2.8 0 12  14 2.8 0 31  27 2.8 1 22  51 4.0 0  1
7 4.0 0  9  14 4.0 0 19  27 4.0 1 16
;
```

```
/*binomial distribution
  logit analysis*/
proc genmod data=ingots;
       class heat soak;
       /*r observations total of n ingots*/
       model r/n = heat soak/ dist=bin   /*binomial distribution*/
                                            TYPE3      /*type 3 analysis of deviance*/
                                            link=logit; /*logit link*/;
       contrast 'Background vs 27 heat' heat 0 0 1 -1;
       contrast 'Backtroung vs 14 heat' heat 0 1 0 -1;
       /*These are not the Wald tests above,
         they will not have the same P-value*/
run;
```

The probability of not being rolled decreases as the heat ‘goes down’
Also the effect of 7 is what it is because there were no ‘failures’ for all 7 s’s

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3236	1.2450	-3.7637	1.1164	1.13	0.2877
Heat	7	1	-26.4235	127806.9	-250523	250470.4	0.00	0.9998
Heat	14	1	-3.2610	1.0737	-5.3653	-1.1566	9.22	0.0024
Heat	27	1	-1.9254	0.8507	-3.5927	-0.2581	5.12	0.0236
Heat	51	0	0.0000	0.0000	0.0000	0.0000	.	.
Soak	1	1	-0.2829	1.1962	-2.6273	2.0616	0.06	0.8131
Soak	1.7	1	0.5951	1.1639	-1.6860	2.8763	0.26	0.6091
Soak	2.2	1	0.4395	1.2739	-2.0573	2.9363	0.12	0.7301
Soak	2.8	1	-0.1305	1.4574	-2.9870	2.7261	0.01	0.9287
Soak	4	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Heat	3	13.08	0.0045
Soak	4	1.41	0.8422

Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
Background vs 27 heat	1	4.54	0.0332	LR
Backtroung vs 14 heat	1	9.05	0.0026	LR

What about a linear term?

```
/*Look at as a linear term*/
proc genmod data=ingots;
       *class heat soak;
       /*r observations total of n ingots*/
       model r/n = heat / dist=bin   /*binomial distribution*/
                              TYPE3      /*type 3 analysis of deviance*/
                       link=logit; /*logit link*/;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.4152	0.7275	-6.8411	-3.9892	55.40	<.0001
Heat	1	0.0807	0.0224	0.0369	0.1245	13.03	0.0003
Scale	0	1.0000	0.0000	1.0000	1.0000		

Which one is better?

The models are not nested so we have to compare the

AIC: 32.109 vs 41.6018

For prediction I would pick the linear logistic model, for the comparisons do the analysis of deviance for the first model.