BIS 621 - Regression Models

Poisson Regression Lecture & Lab

Josemari Feliciano

A brief warning

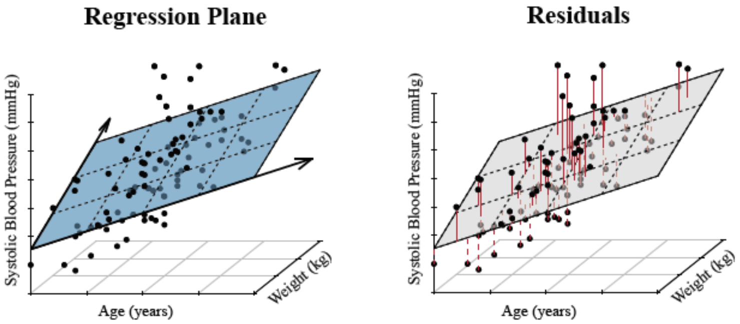
This is probably a bit more dense than usual.

But there are few resources on Poisson and count models in general.

Objective: Give you a general exposure of GLMs and count models. To establish when Poisson regression is appropriate.

Why study GLMs?

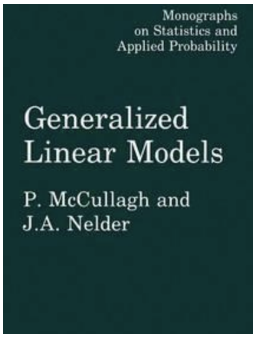
- Acronym stands for *generalized linear models.*
- Previously, we focused on models with normal distribution which are *general* linear models.
- *General* linear models include multiple regression, ANOVA.



A linear algebra perspective by showing what model fitting looks like with 2 independent variables.

Again, why study GLMs?

- The seminal text in this field:



Monographs
on Statistics and
Applied Probability

Generalized
Linear Models

P. McCullagh and
J.A. Nelder

- Why extend *general* models to *generalized*?

One reason is the use of *scale*: *'For example . . . count error structure is well approximated by Poisson and effects of some factors are multiplicative. [With GLM] normality . . . are no longer a requirement . . . for* *the error component while additivity of systematic effects can be specified to hold on a transformed scale if necessary.'*

Components of GLMs:

- Your typical notation:

$$g(\mu_i) = g(E(Y_i)) = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}$$

- Systematic Component: *the right handside of the equation*
- Link Function:
  - Simplest link function is $g(\mu_i) = \mu_i$. *e.g. linear reg.*
  - For Poisson loglinear model, $g(\mu_i) = log(\mu_i)$.
- Variance Function:
  - Describes how variance depends on the mean.
  - $Var(Y_i) = \phi V(E(Y_i)) = \phi V(\mu)$, for $\phi > 0$ and the function V where it is called variance function.

Poisson Regression

Key Assumptions:

- Outcome variable consists of count data {0,1,2,...}
- Independence of observations, this means each observation is independent of the other observations; that is, one observation cannot provide any information on another observation.
- The mean and variance of the model are identical.
- You have one or more independent variables, which can be measured on a continuous or nominal scale (categorical).

Notation:

$$log(\mu_i) = log(E(Y_i)) = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}$$

Let us keep it simple:

$$log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Our focus this lab: To measure the impact of pharmaceutical discount programs on patient abandonment counts of their medication.

Data Set

```
PROC IMPORT OUT= work.drug
    DATAFILE= "\\Client\H$\Documents\Yale_Teaching\BIS621\Poisson\drug_abandonment.xlsx"
    DBMS=EXCEL REPLACE;
    RANGE="drug_abandonment$";
    GETNAMES=YES;
    MIXED=NO;
    SCANTEXT=YES;
    USEDATE=YES;
    SCANTIME=YES;
RUN;
```

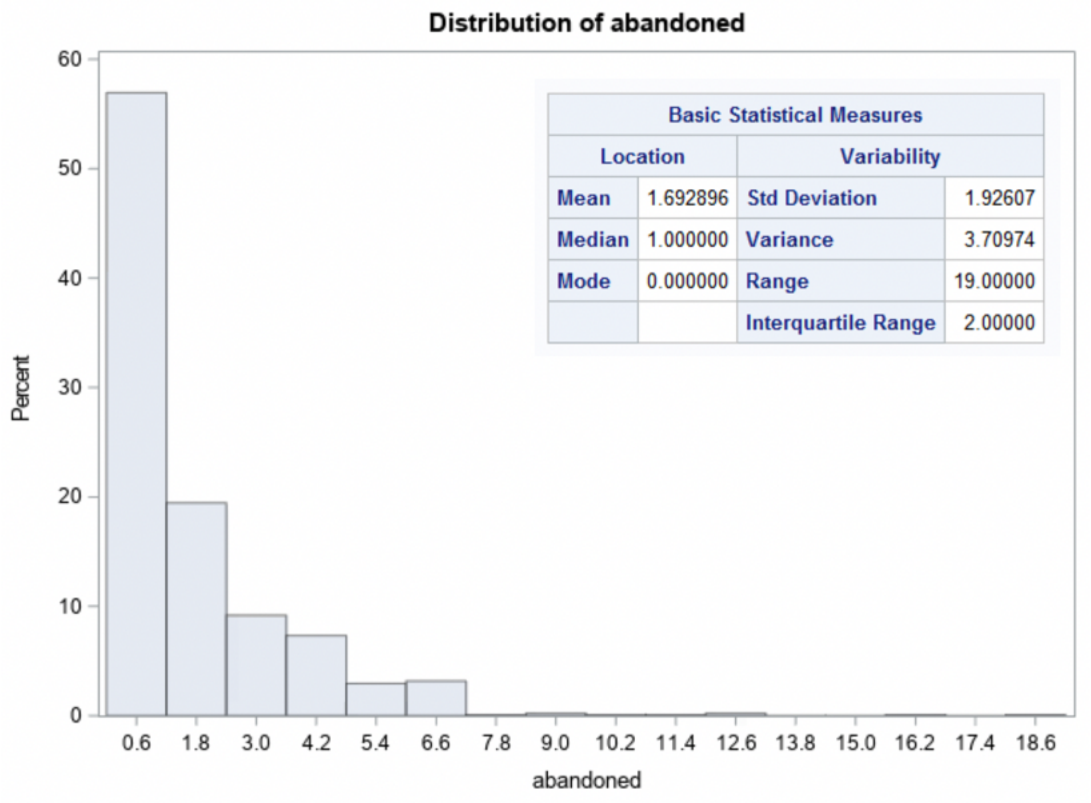
Data Descriptors:

abandoned: 1 if patient abandoned prescription at pharmacy, 0 if kept.
discount: 1 if patient offered e-voucher discount by pharmacy tech, 0 if not.
female: 1 if female, 0 otherwise.
num_other_meds: number of other medications the patients have to pick up.
final_copay: final price for the medication of interest.
pharmacy_distance: distance of pharmacy from patients.

Basic Data Processing

```
proc freq data=drug; run;

proc univariate data=drug;
    var  abandoned; histogram  abandoned;
run;
```



Basic Statistical Measures

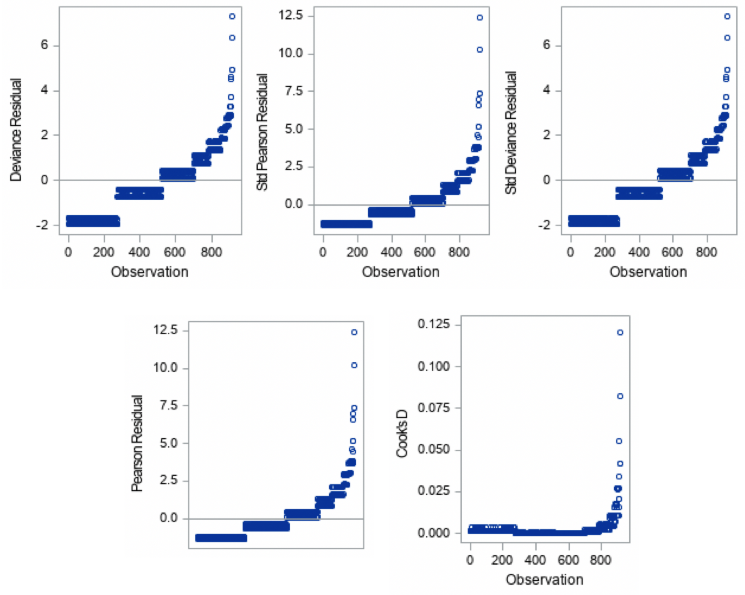
Location		Variability	
Mean	1.692896	Std Deviation	1.92607
Median	1.000000	Variance	3.70974
Mode	0.000000	Range	19.00000
		Interquartile Range	2.00000

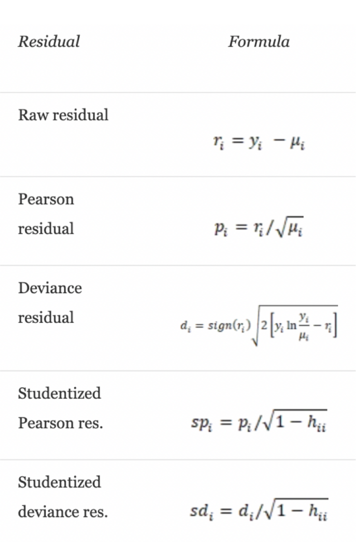
We want the mean and variance to be as close as possible. Ideally, we want the ratio of variance over mean to be less than 2. But it is not far off (≈ 2.2) plus we have other modifications to Poisson that will help correct for this.

The histogram also demonstrates Poisson-like behavior.

Poisson Fitting - Residuals

```
proc genmod data=drug plot=all;
    model abandoned = discount female / dist = poisson link = log type1 type3;
run;
```



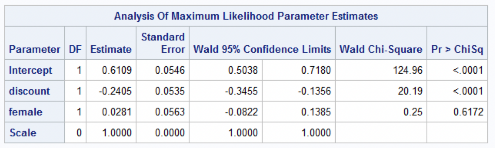


Residual	*Formula*
Raw residual	$r_i = y_i - \mu_i$
Pearson residual	$p_i = r_i / \sqrt{\mu_i}$
Deviance residual	$d_i = sign(r_i)\sqrt{2\left[y_i \ln\frac{y_i}{\mu_i} - r_i\right]}$
Studentized Pearson res.	$sp_i = p_i / \sqrt{1 - h_{ii}}$
Studentized deviance res.	$sd_i = d_i / \sqrt{1 - h_{ii}}$

Good to visually inspect to see any trends.

See any problematic trends?

Model Results

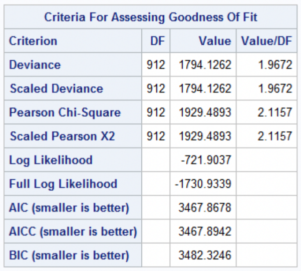


Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6109	0.0546	0.5038	0.7180	124.96	<.0001
discount	1	-0.2405	0.0535	-0.3455	-0.1356	20.19	<.0001
female	1	0.0281	0.0563	-0.0822	0.1385	0.25	0.6172
Scale	0	1.0000	0.0000	1.0000	1.0000		

Since we are working with a log link, we need to exponentiate the estimates first. Accordingly, $e^{-0.24} = 0.79$.

`Interpretation: Holding all other independent variable(s) constant, access to the discount program is associated with a 21% reduction in drug abandonment at the pharmacy.`



Criteria For Assessing Goodness Of Fit

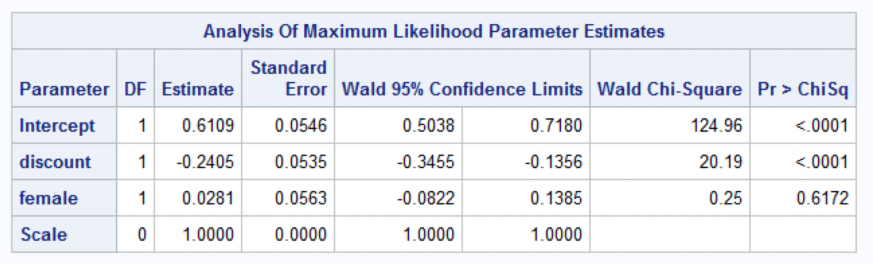
Criterion	DF	Value	Value/DF
Deviance	912	1794.1262	1.9672
Scaled Deviance	912	1794.1262	1.9672
Pearson Chi-Square	912	1929.4893	2.1157
Scaled Pearson X2	912	1929.4893	2.1157
Log Likelihood		-721.9037	
Full Log Likelihood		-1730.9339	
AIC (smaller is better)		3467.8678	
AICC (smaller is better)		3467.8942	
BIC (smaller is better)		3482.3246	

Note: Pay close attention to either deviance or pearson which are important to adjusting our models later on. Moreover, pay attention to AIC, short for Akaike Information Criterion. We will revisit them later.

Scaling Our Variance Assumption

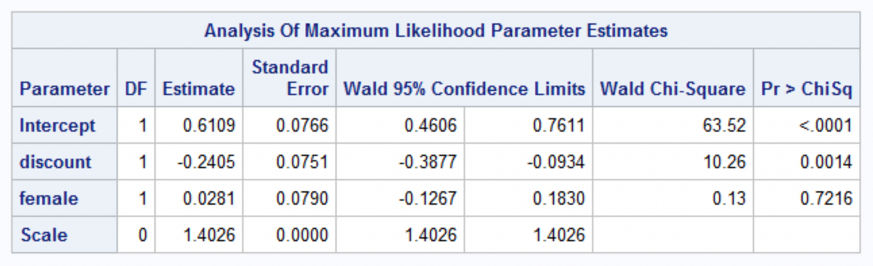
```
proc genmod data=drug;
  model abandoned = discount female / dist = poisson link = log scale = deviance;
run;
```

Notice that we added a `scale = deviance` argument. You may substitute `deviance` for `pearson` too.



Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6109	0.0546	0.5038	0.7180	124.96	<.0001
discount	1	-0.2405	0.0535	-0.3455	-0.1356	20.19	<.0001
female	1	0.0281	0.0563	-0.0822	0.1385	0.25	0.6172
Scale	0	1.0000	0.0000	1.0000	1.0000		



Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6109	0.0766	0.4606	0.7611	63.52	<.0001
discount	1	-0.2405	0.0751	-0.3877	-0.0934	10.26	0.0014
female	1	0.0281	0.0790	-0.1267	0.1830	0.13	0.7216
Scale	0	1.4026	0.0000	1.4026	1.4026		

Running an overdispersed count model underestimates the standard errors. To adjust for this, we can use the $Value/DF$ output earlier for Deviance (1.9672), then take its square-root (1.40). Using this value, we then scale the standard errors for our parameters by multiplying them with this scale. This "scale" is formally known as the *scale parameter*.

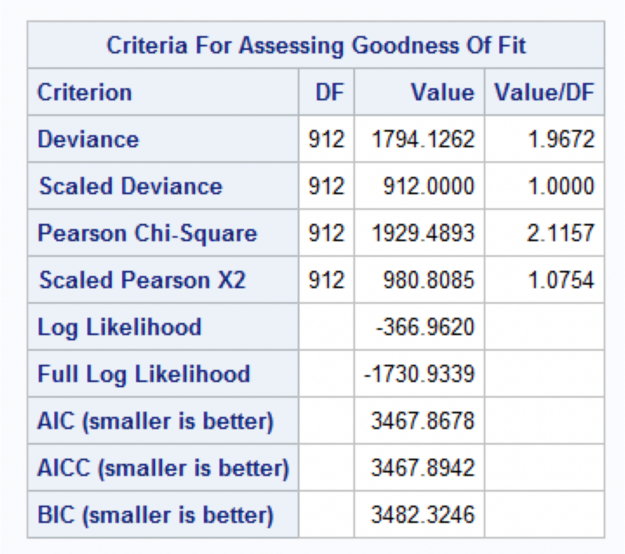
Negative Binomial (Poisson-Gamma Mixture)

```
proc genmod data=drug;
  model abandoned = discount female / dist = negbin link = log;
run;
```

Note: Negative Binomial (NB) is not a Poisson model since it is a different distribution but it is a common alternative for count data. Its "Poisson-Gamma" pseudonym highlights its variance component $(\mu + k\mu^2)$ and its similarities to both Poisson and Gamma. Remember the variance component for Poisson is μ whereas the component for Gamma is μ^2.

Both models include the same covariates thus comparison via AIC is appropriate.

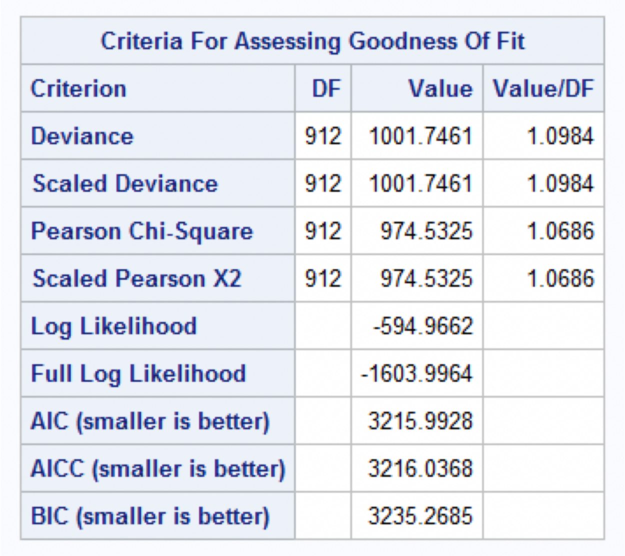
Which is the better model? (The SAS output tells you)



Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	912	1794.1262	1.9672
Scaled Deviance	912	912.0000	1.0000
Pearson Chi-Square	912	1929.4893	2.1157
Scaled Pearson X2	912	980.8085	1.0754
Log Likelihood		-366.9620	
Full Log Likelihood		-1730.9339	
AIC (smaller is better)		3467.8678	
AICC (smaller is better)		3467.8942	
BIC (smaller is better)		3482.3246	

Output for Poisson.



Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	912	1001.7461	1.0984
Scaled Deviance	912	1001.7461	1.0984
Pearson Chi-Square	912	974.5325	1.0686
Scaled Pearson X2	912	974.5325	1.0686
Log Likelihood		-594.9662	
Full Log Likelihood		-1603.9964	
AIC (smaller is better)		3215.9928	
AICC (smaller is better)		3216.0368	
BIC (smaller is better)		3235.2685	

Output for Negative Binomial.

Lab Write Up:

```
proc genmod data = drug;
  model abandoned = discount female yourothervariables / dist=zinb;
  zeromodel ?;
run;
```

Prompt 1:

In lecture, I fitted three drug abandonment models with the `discount` and `female` covariates. First, I did a regular Poisson model. Second, I scaled my Poisson model to adjust for overinflation by inflating the standard error of the estimates. Third, I performed a negative binomial model to see if it can give me a lower (better) AIC. Perform the same steps with additional covariates from the dataset.

Prompt 2 (Optional):

Somehow, all the Biostatisticians you know are gone for vacation and you must run a model called zero-inflated negative binomial. Here's the idea, some if not all of the zeros in your count dataset are generated by another data-generating process (binomial) that may help you adjust for overinflation. While you are not a Biostatistician, you know that models with lower AIC are generally better. Place one or two variables in the `?` that may contribute to any excess zeroes. Check the AIC. Perform the same for another set of variables. Check the AIC.