Lecture 23: Poisson Regression
Stat 704: Data Analysis I, Fall 2010

Tim Hanson, Ph.D.

University of South Carolina

Poisson regression

* Regular regression data $\{(\mathbf{x}_i, Y_i)\}_{i=1}^n$, but now Y_i is a positive integer, often a count: new cancer cases in a year, number of monkeys killed, etc.
* For Poisson data, $\text{var}(Y_i) = E(Y_i)$; variability increases with predicted values. In regular OLS regression, this manifests itself in the "megaphone shape" for r_i versus $\hat{Y}_i$.
* If you see this shape, consider whether the data could be Poisson (e.g. blood pressure data, p. 428).
* Any count, or positive integer could potentially be approximately Poisson. In fact, binomial data where n_i is really large, is approximately Poisson.

Log and identity links

Let $Y_i \sim \text{Pois}(\mu_i)$.
The **log-link** relating μ_i to $\mathbf{x}_i'\boldsymbol{\beta}$ is standard:

$$Y_i \sim \text{Pois}(\mu_i), \quad \log \mu_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1},$$

the log-linear **Poisson regression** model.

The **identity** link can also be used

$$Y_i \sim \text{Pois}(\mu_i), \quad \mu_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1}.$$

Both can be fit in PROC GENMOD.

Interpretation for log-link

We have

$$Y_i \sim \text{Pois}(\mu_i).$$

The log link $\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$ is most common:

$$Y_i \sim \text{Pois}(\mu_i), \quad \mu_i = e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}},$$

or simply $Y_i \sim \text{Pois}\left(e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}\right)$.
Say we have $k = 3$ predictors. The mean satisfies

$$\mu(x_1, x_2, x_3) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}.$$

Then increasing x_2 to $x_2 + 1$ gives

$$\mu(x_1, x_2 + 1, x_3) = e^{\beta_0+\beta_1 x_1+\beta_2(x_2+1)+\beta_3 x_3} = \mu(x_1, x_2, x_3)e^{\beta_2}.$$

In general, increasing x_j by one, but holding the other predictors the constant, increases the mean by a factor of e^{β_j}.

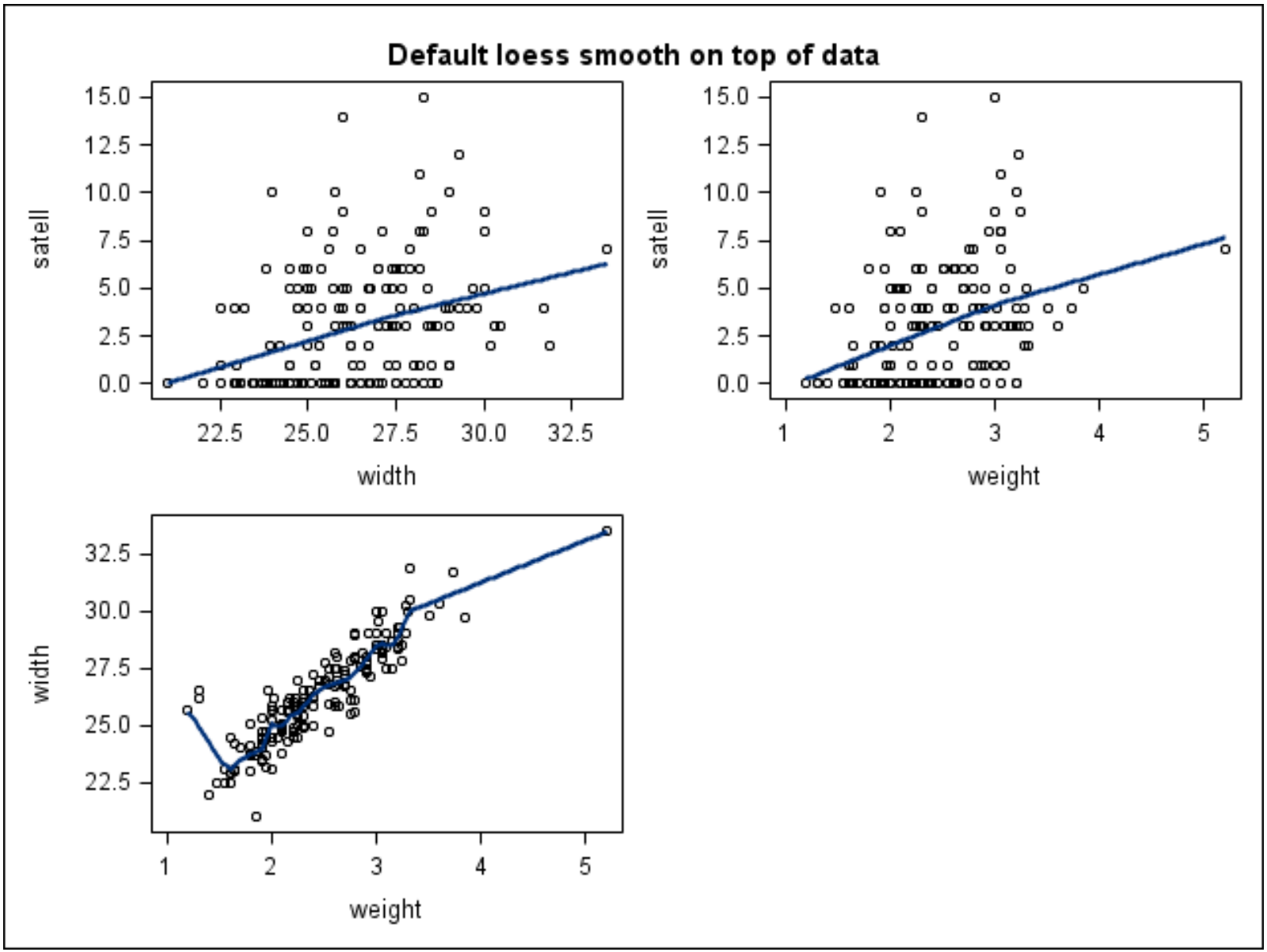
Example: Crab mating
Data on female horseshoe crabs.

- C = color (1,2,3,4=light medium, medium, dark medium, dark).
- S = spine condition (1,2,3=both good, one worn or broken, both worn or broken).
- W = carapace width (cm).
- Wt = weight (kg).
- Sa = number of satellites (additional male crabs besides her nest-mate husband) nearby.

Looking at the data...

We initially examine width as a predictor for the number of satellites. A raw scatterplot of the numbers of satellites versus the predictors does not tell us much. Superimposing a smoothed fit helps & shows an approximately linear trend in weight.

```
ods png; ods graphics on;
options nodate;
proc sgscatter data=crabs;
  title "Default loess smooth on top of data";
  plot satell*(width weight) width*weight / loess;
run;
ods graphics off; ods png close;
```


Default loess smooth on top of data
satell
width
weight
width
weight

We'll fit three models using `proc genmod`.

$$Sa_i \sim \text{Pois}(e^{\alpha+\beta W_i}),$$

$$Sa_i \sim \text{Pois}(\alpha + \beta W_i),$$

and

$$Sa_i \sim \text{Pois}(e^{\alpha+\beta_1 W_i+\beta_2 W_i^2}).$$

SAS code

SAS code:

```
data crab; input color spine width satell
weight;
   weight=weight/1000; color=color-1;
   width_sq=width*width;
datalines;
3  3  28.3  8  3050
4  3  22.5  0  1550
...et cetera...
5  3  27.0  0  2625
3  2  24.5  0  2000
;
proc genmod;
   model satell = width / dist=poi link=log ;
proc genmod;
   model satell = width / dist=poi link=identity ;
proc genmod;
   model satell = width width_sq / dist=poi link=log ;
run;
```

Output from fitting the three Poisson regression models:

SAS output

The GENMOD Procedure

Model Information

Data Set	WORK.CRAB
Distribution	Poisson
Link Function	Log
Dependent Variable	satell

Number of Observations Read	173
Number of Observations Used	173

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	171	567.8786	3.3209
Scaled Deviance	171	567.8786	3.3209
Log Likelihood		68.4463	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-3.3048	0.5422	-4.3675	-2.2420	37.14	<.0001
width	1	0.1640	0.0200	0.1249	0.2032	67.51	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

SAS output

The GENMOD Procedure

Model Information

Data Set	WORK.CRAB
Distribution	Poisson
Link Function	Identity
Dependent Variable	satell

Number of Observations Read	173
Number of Observations Used	173

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	171	557.7083	3.2615
Scaled Deviance	171	557.7083	3.2615
Log Likelihood		73.5314	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-11.5321	1.5104	-14.4924	-8.5717	58.29	<.0001
width	1	0.5495	0.0593	0.4333	0.6657	85.89	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

SAS output

The GENMOD Procedure

Model Information

Data Set	WORK.CRAB
Distribution	Poisson
Link Function	Log
Dependent Variable	satell

Number of Observations Read	173
Number of Observations Used	173

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	170	558.2359	3.2837
Scaled Deviance	170	558.2359	3.2837
Log Likelihood		73.2676	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-19.6525	5.6374	-30.7017	-8.6034	12.15	0.0005
width	1	1.3660	0.4134	0.5557	2.1763	10.92	0.0010
width_sq	1	-0.0220	0.0076	-0.0368	-0.0071	8.44	0.0037
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

- Write down the fitted equation for the Poisson mean from each model.
- How are the regression effects interpreted in each case?
- How would you pick among models? Recall

$$\text{AIC} = -2[L(\hat{\boldsymbol{\beta}}; \mathbf{y}) - p].$$

$$\begin{aligned} -2(73.27 - 3) &= -140.54, \\ -2(73.53 - 2) &= -143.06, \\ -2(68.44 - 2) &= -132.88. \end{aligned}$$

- Are there any potential problems with any of the models? How about prediction?

Offsets

* Sometimes counts are collected over different amounts of time, space...
* For example, we may have numbers of new cancer cases per *month* from some counties, and per *year* from others.
* If time periods are the same from for all data, then μ_i is the mean count per time period.
* Otherwise we specify μ_i as a rate per unit time period and have data in the form $\{(\mathbf{x}_i, Y_i, t_i)\}_{i=1}^n$ where t_i is the amount of time that the Y_i accumulates over.
* Model: $Y_i \sim \text{Pois}(t_i \mu_i)$.
* For the log-link we have

$$Y_i \sim \text{Pois}\left(e^{\mathbf{x}_i'\boldsymbol{\beta}+\log(t_i)}\right).$$

$\log(t_i)$ is called an *offset*.

Ache monkey hunting

* Data on the number of capuchin monkeys killed by $n = 47$ Ache hunters over several hunting trips were recorded; there were 363 total records.
* I'll describe the hunting process in class; it involves splitting into groups, chasing monkeys through the trees, and shooting arrows straight up.

Let Y_i be the number of monkey's killed by hunter i ($i = 1, \ldots, 47$) over several hunting trips lasting different amounts of days, totaling t_i. Let μ_i be the hunter i's kill rate (per day).

$$Y_i \sim \text{Pois}(\mu_i t_i),$$

where

$$\log \mu_i = \beta_0 + \beta_1 a_i + \beta_2 a_i^2.$$

What goes up...

* Monkey hunting is dangerous. What goes up...
* We include a quadratic effect because we expect a "leveling off" effect or possible decline in ability with age.
* Of interest is when hunting ability is greatest. Hunting prowess contributes to a man's status within the group.
* a_i is hunter i's age in years.
* Looking ahead, the fitted *monkey kill rate* is

$$\mu(a) = \exp(-5.4842 + 0.1246a - 0.0012a^2).$$

* Monkey kill rate? Now who says statistics isn't exciting!

Dinner

SAS code

```
data ache; input id age kills days; logdays=log(days); rawrate=kills/days;
datalines;
 1      67        0         3
 2      66        0        89
 3      63       29       106
 4      60        2         4
 5      61        0        28
 6      59        2        73
 7      58        3         7
 8      57        0        13
 9      56        0         4
10      56        3       104
11      55       27       126
12      54        0        63
13      51        7        88
14      50        0         7
15      48        3         3
16      49        0        56
17      47        6        70
18      42        1        18
19      39        0         4
20      40        7        83
21      40        4        15
22      39        1        19
23      37        2        29
24      35        2        48
25      35        0        35
26      33        0        10
27      33       19        75
28      32        9        63
29      32        0        16
30      31        0        13
```

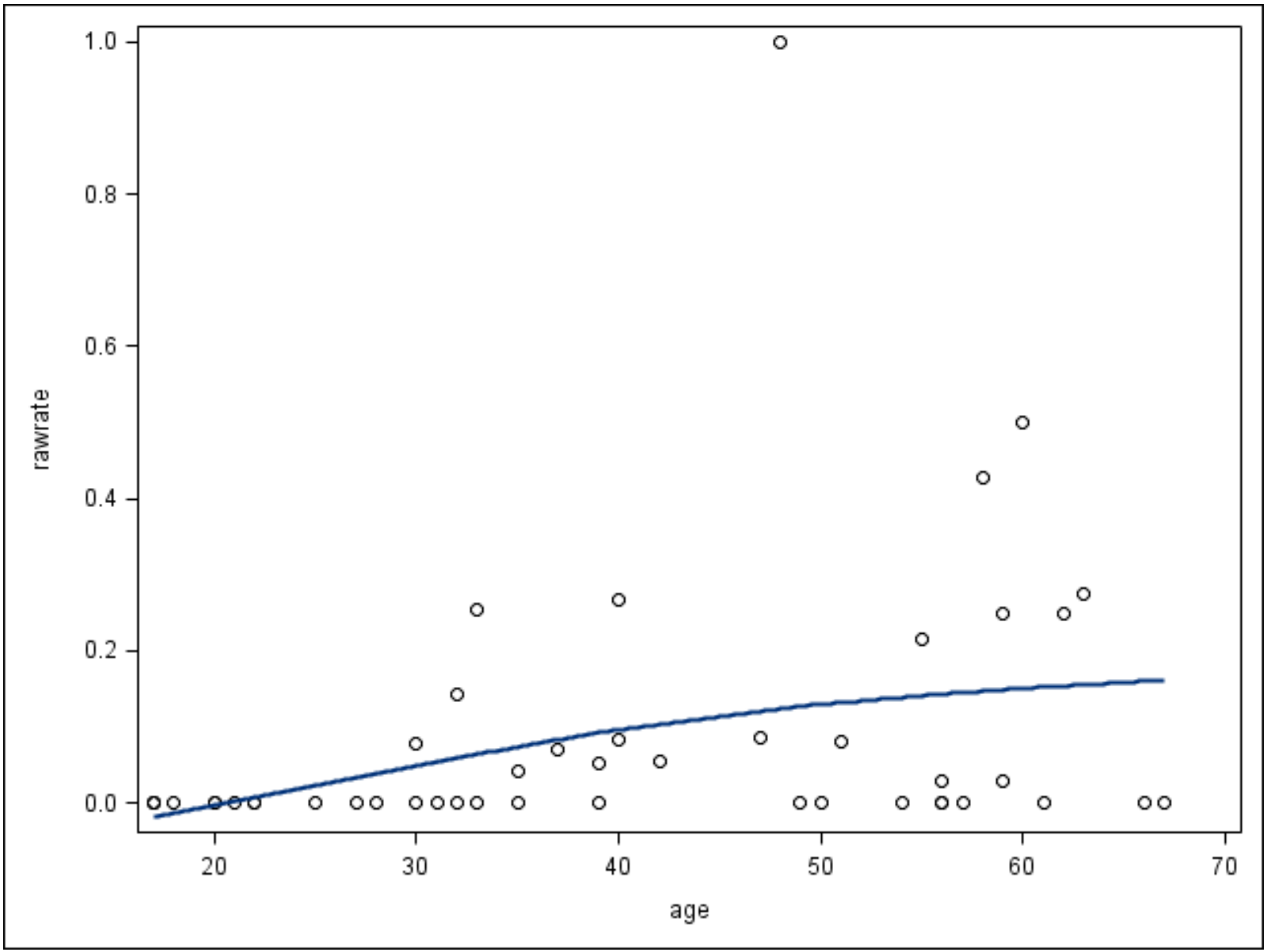
SAS code

```
31      30      0       20
32      30      2       26
33      28      0        4
34      27      0       13
35      25      0       10
36      22      0       16
37      22      0       33
38      21      0        7
39      20      0       33
40      18      0        8
41      17      0        3
42      17      0       13
43      17      0        3
44      56      0       62
45      62      1        4
46      59      1        4
47      20      0       11
;
ods jpeg; ods graphics on; * not weighted by how many days...;
proc sgscatter data=ache;
 plot rawrate*age / loess;
run;
ods graphics off; ods jpeg close;

proc genmod data=ache;
 model kills=age age*age / dist=poisson link=log offset=logdays;
 output out=out p=p reschi=r;

ods png; ods graphics on;
proc sgscatter data=out;
 plot r*(p age) / loess; run;
ods graphics off; ods png close;
```

Raw rates.



SAS output

Model Information

Data Set	WORK.ACHE
Distribution	Poisson
Link Function	Log
Dependent Variable	kills
Offset Variable	logdays

Number of Observations Read	47
Number of Observations Used	47

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	44	186.0062	4.2274
Scaled Deviance	44	186.0062	4.2274
Pearson Chi-Square	44	197.7941	4.4953
Scaled Pearson X2	44	197.7941	4.4953
Log Likelihood		98.7129	
Full Log Likelihood		-124.8921	
AIC (smaller is better)		255.7841	
AICC (smaller is better)		256.3423	
BIC (smaller is better)		261.3346	

Algorithm converged.

SAS output

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.4842	1.2448	-7.9240	-3.0445	19.41	<.0001
age	1	0.1246	0.0568	0.0134	0.2359	4.82	0.0281
age*age	1	-0.0012	0.0006	-0.0024	0.0000	3.78	0.0520
Scale	0	1.0000	0.0000	1.0000	1.0000		

Goodness of fit

The Pearson residual is

$$r_{P_i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

As in logistic regression, the sum of these gives the Pearson GOF statistic

$$X^2 = \sum_{i=1}^{n} r_{P_i}^2.$$

$X^2 \sim \chi^2_{n-p}$ when the regression model fits. Alternative is "saturated model."

Deviance statistic is

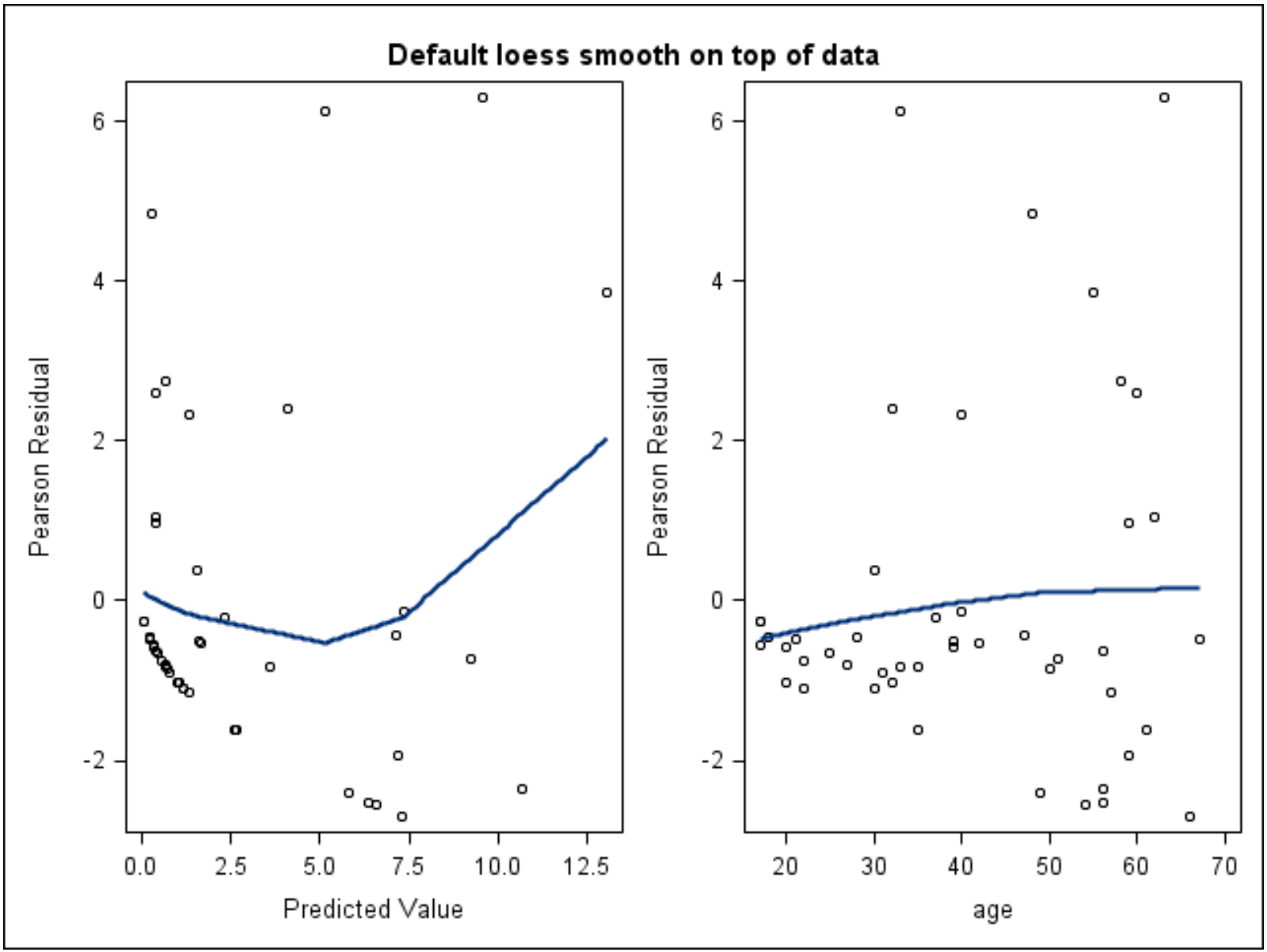
$$D^2 = -2 \sum_{i=1}^{n} \left[Y_i \log(\hat{\mu}_i / Y_i) + (Y_i - \hat{\mu}_i) \right].$$

Replace $\hat{\mu}_i$ by $\hat{\mu}_i t_i$ when offsets are present. $D^2 \sim \chi^2_{n-p}$ when the regression model fits. Page 621 defines "deviance residual" dev_i.

Diagnostics

* From SAS we can get Cook's distance c_i (cookd), leverage h_i (h), predicted $\hat{Y}_i = e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}$ (p) Pearson residual r_{P_i} (reschi; have variance < 1), studentized Pearson residual r_{SP_i} (stdreschi; have variance $= 1$).
* Residual plots have same problems as logistic regression for counts Y_i close to zero. Think of when the normal approximation to the Poisson works okay...same idea here.
* Can do smoothed versions; Ache hunting data...

Model doesn't fit very well; $\text{var}(r_{P_i}) < 1$...



If there's time, we'll consider an analysis of the full crab mating data in SAS...