Section 14.13 Poisson Regression

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

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 used most often:

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

yielding what is commonly called the **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

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.

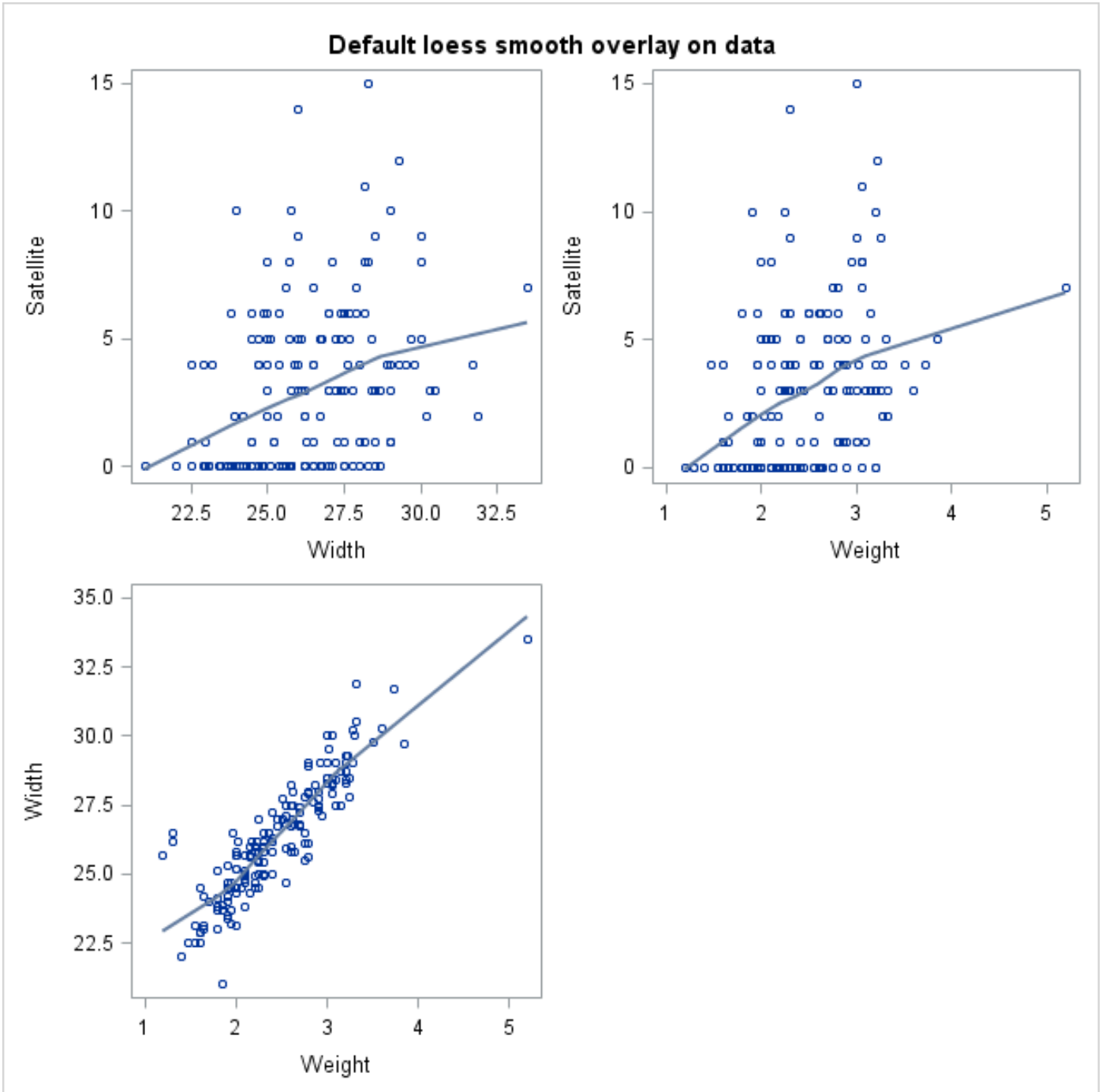
Using logistic regression we explored whether a female had *one or more* satellites. Using Poisson regression we can model the actual *number* of satellites directly.

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.

Note that variability increases with width and weight!

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


Default loess smooth overlay on data
Satellite
Width
Satellite
Weight
Width
Weight

Three competing models using width as predictor

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

$$Sa_i \sim \text{Pois}(e^{\beta_0+\beta_1 W_i}),$$

$$Sa_i \sim \text{Pois}(\beta_0 + \beta_1 W_i),$$

and

$$Sa_i \sim \text{Pois}(e^{\beta_0+\beta_1 W_i+\beta_2 W_i^2}).$$

SAS code

SAS code:

```
data crabs; input color spine width satellite
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
;

*Problems with residuals seen here apply to other models as well;
proc genmod data=crabs plots=all;
   model satellite = width / dist=poi link=log ;
proc genmod data=crabs;
   model satellite = width / dist=poi link=identity ;
proc genmod data=crabs;
   model satellite = 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.

Inference

- 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

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

For log-link quadratic, identity-link linear, and log-link linear we have

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

- Are there any potential problems with any of the models? How about prediction? Are there any potential problems common to *all* these models? We will return to this topic later.

Offsets

- Sometimes counts are collected over different intervals or areas 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 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*–a covariate with fixed coefficient 1.

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.

The hunting process involves splitting into groups, chasing monkeys through the trees, and shooting arrows straight up.

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

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

where

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

A quadratic effect is included to accommodate 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.

Aiming for...

...dinner!

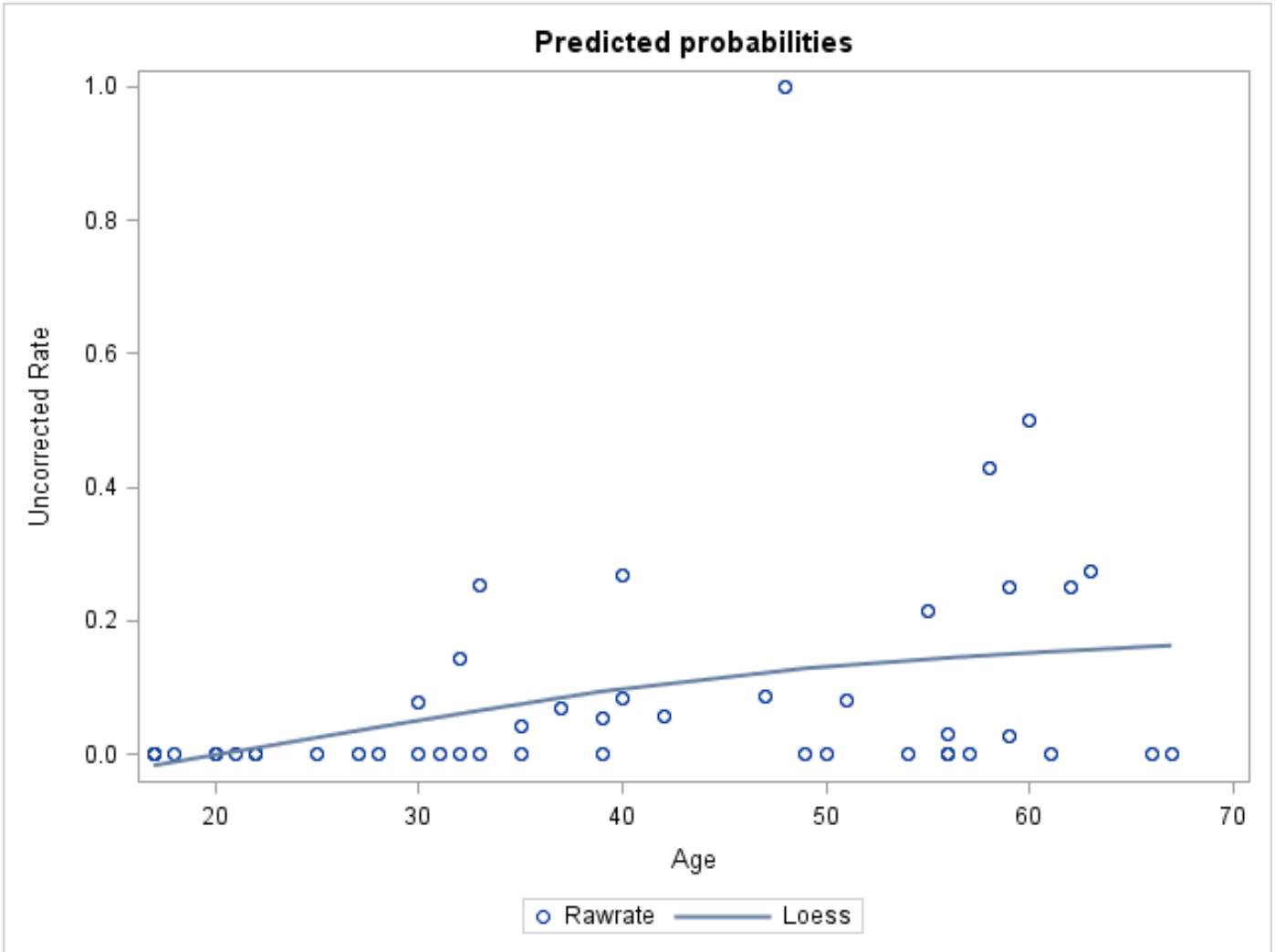
SAS code

```
data ache; input age kills days @@; logdays=log(days); rawrate=kills/days;
datalines;
67       0        3 66       0       89 63      29      106 60       2        4
61       0       28 59       2       73 58       3        7 57       0       13
56       0        4 56       3      104 55      27      126 54       0       63
51       7       88 50       0        7 48       3        3 49       0       56
47       6       70 42       1       18 39       0        4 40       7       83
40       4       15 39       1       19 37       2       29 35       2       48
35       0       35 33       0       10 33      19       75 32       9       63
32       0       16 31       0       13 30       0       20 30       2       26
28       0        4 27       0       13 25       0       10 22       0       16
22       0       33 21       0        7 20       0       33 18       0        8
17       0        3 17       0       13 17       0        3 56       0       62
62       1        4 59       1        4 20       0       11
;
proc sgscatter data=ache; * not weighted by how many days...;
 plot rawrate*age / loess;

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

proc sgscatter data=out;
 plot r*(p age) / loess; run;
```

Raw rates with loess smooth



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		

The fitted *monkey kill rate* is

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

At what age, typically, is monkey hunting ability maximized?

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. The alternative is the "saturated model."

The deviance statistic is

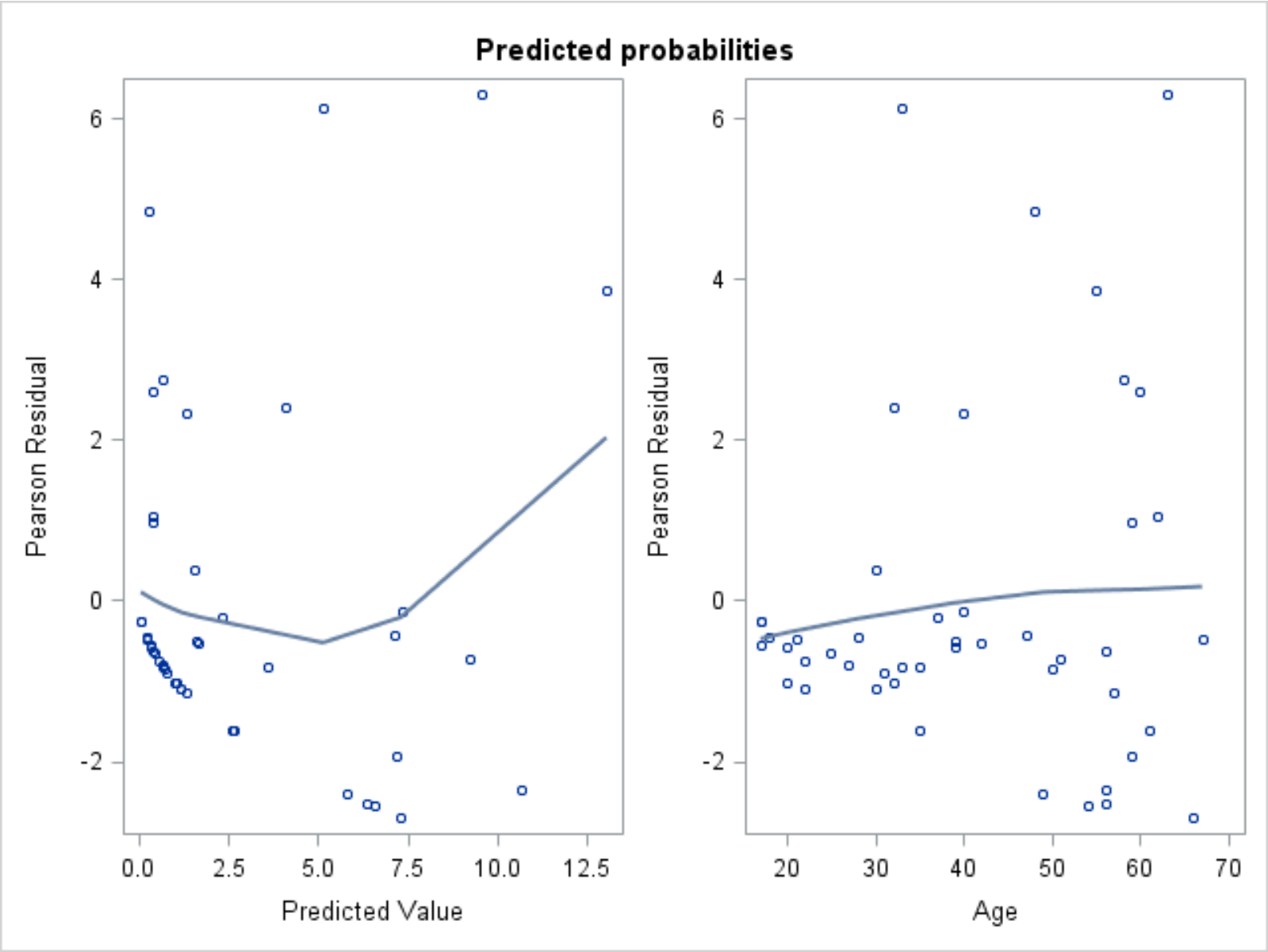
$$D^2 = -2 \sum_{i=1}^{n} \left[Y_i \log(\hat{\mu}_i / Y_i) + (Y_i - \hat{\mu}_i) \right].$$

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; variance < 1), studentized Pearson residual r_{SP_i} (stdreschi; 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.
- We can do smoothed versions; see SAS code for Ache hunting data.

The model doesn't fit very well; $\text{var}(r_{P_i}) \gg 1$...



Overdispersion–blocking

The variability in the Pearson residuals is much higher than what we should see; there are many poorly fit observations. This extra-Poisson variability is often referred to as "overdispersion."

Recall that in Chapters 21, 25, and 27 we discussed *blocking* on individuals to reduce variability. The Ache hunters actually took part in many hunting trips, i.e. there are repeated measures on each hunter. We can instead consider hunting trip j from hunter i of length L_{ij} days, and posit a mixed model

$$Y_{ij} \sim \text{Pois}(\lambda_{ij} L_{ij}), \quad \log(\lambda_{ij}) = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + u_i,$$

where

$$u_1, \ldots, u_{47} \overset{iid}{\sim} N(0, \sigma^2)$$

are random *hunter ability* effects.

This model, fit in `proc glimmix`, reduces variability by appropriately blocking the repeated measures on hunter. We'll fit this model in our lecture on GLMM's.

Overdispersion–Negative binomial regression

Overdispersion is also an issue with the three models in Slide 9, indicated cursorily by the goodness of fit criteria being consistently greater than 3. One simple solution is to assume a distribution for which the variance can be larger than the mean, e.g., the negative binomial. SAS uses the following parameterization for the dispersion parameter k, which yields $V(Y) = \mu + k\mu^2$:

$$f(y) = \frac{\Gamma(y + 1/k)}{\Gamma(y+1)\Gamma(1/k)} \frac{(k\mu)^y}{(1+k\mu)^{y+1/k}}, \; y = 0, 1, \ldots$$

$k = 1/r$ and $\mu = rp/q$ for the more familiar parameterization of the Negative Binomial distribution.

Negative binomial regression

We can model a negative binomial regression with log link in PROC GENMOD

```
proc genmod data=crabs;
   model satellite = width / dist=nb link=log ;
   run;
```

Model Information

Data Set	WORK.CRAB
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	Satellite

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	171	195.8112	1.1451
Pearson Chi-Square	171	144.7507	0.8465
Log Likelihood		154.3889	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-14.0525	1.2642	-6.5303	-1.5747	10.28	0.0013
width	1	0.1921	0.0476	0.0987	0.2854	16.27	<0.0001
Dispersion	1	1.1055	0.1971	0.7795	1.5679		

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

Overdispersion-Zero-Inflated Poisson

Excessive zeroes in a data set can be one source of overdispersion; this appears to be the case for both the Ache data and the horseshoe crab data (though not the only source of over-dispersion in either case).
We use a mixture model, including a distribution degenerate at 0:

$$Y \sim F,\ F = \gamma F_0 + (1-\gamma)F_\mu$$

For $F_0, P_{F_0}(Y = 0) = 1$, while F_μ is a Pois(μ) distribution.

Overdispersion-ZIP

PROC NLIN in SAS can actually optimize a likelihood, provided the contribution to the likelihood for each value of Y is provided.

$$\begin{aligned} P_F(Y=0) &= \gamma P(Y=0|F_0) + (1-\gamma)P(Y=0|F_\mu) \\ &= \gamma + (1-\gamma)exp(-\mu(\mathbf{x})) \end{aligned}$$

$$\begin{aligned} P_F(Y=y) &= \gamma P(Y=y|F_0) + (1-\gamma)P(Y=y|F_\mu) \\ &= (1-\gamma)exp(-\mu(\mathbf{x}))\mu(\mathbf{x})^y/y!,\ y=1,2,3,\ldots \end{aligned}$$

Horseshoe Crab ZIP model

ZIP (and ZINB) models can actually be fit in GENMOD now. Here is a likelihood-based approach instead.

```
proc nlmixed data=crabs;
*Starting values.  For p0, use the frequency table;
*For b0 and b1, use fit from Poisson regression model;
parms p0=.5 b0=-3.0 b1=0.2;
mu0=exp(b0+b1*width);
if satellite=0 then do;
prob=p0+(1-p0)*exp(-mu0);
loglike=log(prob);
end;
else loglike=log(1-p0)+satellite*log(mu0)-mu0-lgamma(satellite+1);
model satellite~general(loglike);
run;
```

Horseshoe Crab ZIP model

In PROC GENMOD, it is straightforward to model $\gamma(\mathbf{x})$, the probability that Y is sampled from distribution F_0, as a logistic regression model.

```
proc genmod data=crabs;
  model satellite = width /dist=zip obstats;
  zeromodel  /link = logit; *we need to transform the answer to recover gamma;
run;

*We can add covariates to a logit model for whether or not a zero is observed;
proc genmod data=crabs;
  model satellite = width /dist=zip obstats;
  zeromodel width /link = logit;
run;
```

Miller Lumber

Miller lumber is large retailer of lumber and other household supplies. During a two-week period customers were surveyed. The store wanted to model the numbers Y_i of individuals coming from $n = 110$ census tracts over the same two-week period as a function of

- x_1 number of housing units.
- x_2 average income in \$.
- x_3 average housing unit age in years.
- x_4 distance to nearest competitor in miles.
- x_5 distance to Miller Lumber in miles.

These data are analyzed on pp. 621–623 (Table 14.14). We will also analyze these data if time permits.