

Should you always toss in a dispersion term ϕ ?

Here's some SAS code:

```
data example;
input x y n @@; x_sq=x*x;
datalines;
-2.0 86 100 -1.5 58 100 -1.0 25 100 -0.5 17 100 0.0 10 100
 0.5 17 100  1.0 25 100
;
proc genmod; * fit simple linear term in x & check for overdispersion;
  model y/n = x / link=logit dist=bin;
proc genmod; * adjust for apparent overdispersion;
  model y/n = x / link=logit dist=bin scale=pearson;
proc genmod; * what if instead we try a more flexible mean?;
  model y/n = x x_sq / link=logit dist=binom;
proc logistic; * residual plots from simpler model;
  model y/n = x; output out=diag1 reschi=p h=h xbeta=eta;
data diag2; set diag1; r=p/sqrt(1-h);
proc gplot; plot r*x; plot r*eta;
```

Output from fit of logistic model with logit link:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	74.6045	14.9209
Pearson Chi-Square	5	79.5309	15.9062

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.3365	0.1182	-1.5682	-1.1047	127.77	<.0001
x	1	-1.0258	0.0987	-1.2192	-0.8323	108.03	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

The coefficient for x is highly significant. Note that $P(\chi_5^2 > 74.6) < 0.0001$ and $P(\chi_5^2 > 79.5) < 0.0001$. Evidence of overdispersion? There's good replication here, so certainly *something* is not right with the model.

Let's include a dispersion parameter ϕ :

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	74.6045	14.9209
Scaled Deviance	5	4.6903	0.9381
Pearson Chi-Square	5	79.5309	15.9062
Scaled Pearson X2	5	5.0000	1.0000

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.3365	0.4715	-2.2607	-0.4123	8.03	0.0046
x	1	-1.0258	0.3936	-1.7972	-0.2543	6.79	0.0092
Scale	0	3.9883	0.0000	3.9883	3.9883		

We have $\hat{\phi} = 4.0$ and the standard errors are increased by this factor.
The coefficient for x is still significant.

Problem solved!!! Or is it?

Instead of adding ϕ to a model with a linear term, what happens if we allow the mean to be a bit more flexible?

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	4	1.7098	0.4274
Pearson Chi-Square	4	1.6931	0.4233

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.9607	0.1460	-2.2468	-1.6745	180.33	<.0001
x	1	-0.0436	0.1352	-0.3085	0.2214	0.10	0.7473
x_sq	1	0.9409	0.1154	0.7146	1.1671	66.44	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Here, we are *not* including a dispersion term ϕ . There is no evidence of overdispersion when the *mean is modeled correctly*. Adjusting SE's using the quasiliikelihood approach relies on *correctly modeling the mean*, otherwise ϕ becomes a measure of dispersion of data about *an incorrect mean*. That is, ϕ attempts to pick up the slop left over from specifying a mean that is too simple.

A correctly specified mean can obviate overdispersion. How to check if the mean is okay? Hint:

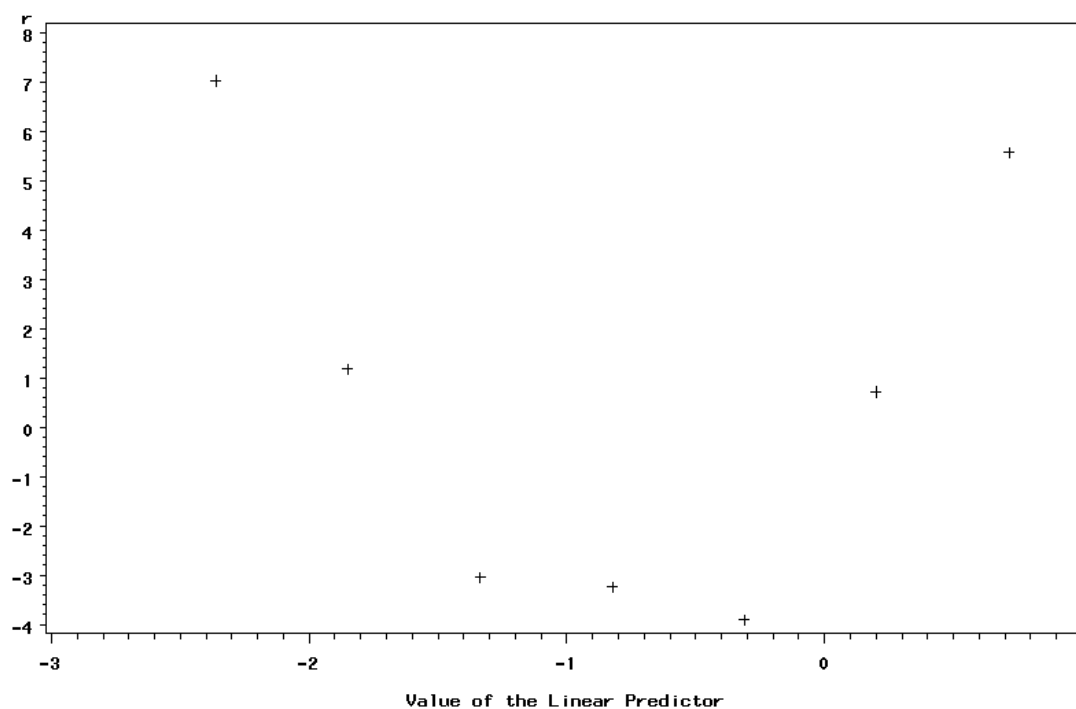


Figure 1: Residual plot r_i versus η_i for made-up data.

Chapter 6 – Logistic Regression III

6.1 Model selection

Two competing goals:

- Model should fit the data well.
- Model should be simple to interpret (smooth rather than overfit – principle of parsimony).

Often hypotheses on how the outcome is related to specific predictors will help guide the model building process.

Agresti points out a rule of thumb: at least 10 events and 10 non-events should occur for each predictor in the model (including dummies). So if $\sum_{i=1}^N y_i = 40$ and $\sum_{i=1}^N n_i = 830$, you should have no more than $40/10 = 4$ predictors in the model.

6.1.1 Horseshoe crab data

Recall that in all models fit we strongly rejected $H_0 : \text{logit } \pi(\mathbf{x}) = \beta_0$ in favor of $H_1 : \text{logit } \pi(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.5565	7	<.0001
Score	36.3068	7	<.0001
Wald	29.4763	7	0.0001

However, it was not until we carved superfluous predictors from the model that we showed significance for the included model effects.

This is an indication that several covariates may be highly related, or correlated. If one or more predictors are perfectly predicted as a linear combination of other predictors the model is overspecified and unidentifiable. Here's an example:

$$\text{logit } \pi(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 - 3x_2).$$

The MLE $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ is not unique and the model is said to be unidentifiable. The variable $x_1 - 3x_2$ is totally predicted and redundant given x_1 and x_2 .

Although a perfect linear relationship is usually not met in practice, often variables are *highly* correlated and therefore one or more are redundant. We need to get rid of some!

Although not ideal, automated model selection is necessary with large numbers of predictors. With $p - 1 = 10$ predictors, there are $2^{10} = 1024$ possible models; with $p - 1 = 20$ there are 1,048,576 to consider.

Backwards elimination starts with a large pool of potential predictors and step-by-step eliminates those with (Wald) p -values larger than a cutoff (the default is 0.05 in SAS PROC LOGISTIC).

```

proc logistic data=crabs1 descending;
  class color spine / param=ref;
  model y = color spine width weight color*spine color*width color*weight
  spine*width spine*weight width*weight / selection=backward;

```

When starting from all main effects and two-way interactions, the default p -value cutoff 0.05 yields only the model with width as a predictor

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	color*spine	6	9	0.0837	1.0000
2	width*color	3	8	0.8594	0.8352
3	width*spine	2	7	1.4906	0.4746
4	weight*spine	2	6	3.7334	0.1546
5	spine	2	5	2.0716	0.3549
6	width*weight	1	4	2.2391	0.1346
7	weight*color	3	3	5.3070	0.1507
8	weight	1	2	1.2263	0.2681
9	color	3	1	6.6246	0.0849

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.3508	2.6287	22.0749	<.0001
width	1	0.4972	0.1017	23.8872	<.0001

Let's change the criteria for removing a predictor to $p\text{-value} \geq 0.15$.

```
model y = color spine width weight color*spine color*width color*weight
spine*width spine*weight width*weight / selection=backward slstay=0.15;
```

Yielding a more complicated model:

Summary of Backward Elimination

Step	Effect	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	color*spine	6	9	0.0837	1.0000
2	width*color	3	8	0.8594	0.8352
3	width*spine	2	7	1.4906	0.4746
4	weight*spine	2	6	3.7334	0.1546
5	spine	2	5	2.0716	0.3549

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	13.8781	14.2883	0.9434	0.3314
color	1	1.3633	5.9645	0.0522	0.8192
color	2	-0.6736	2.6036	0.0669	0.7958
color	3	-7.4329	3.4968	4.5184	0.0335
width	1	-0.4942	0.5546	0.7941	0.3729
weight	1	-10.1908	6.4828	2.4711	0.1160
weight*color 1	1	0.1633	2.3813	0.0047	0.9453
weight*color 2	1	0.9425	1.1573	0.6632	0.4154
weight*color 3	1	3.9283	1.6151	5.9155	0.0150
width*weight	1	0.3597	0.2404	2.2391	0.1346

Let's test if we can simultaneously drop width and width*weight from this model. From the (voluminous) output we find:

Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	196.841
SC	230.912	228.374
-2 Log L	225.759	176.841

Fitting the simpler model with color, weight, and color*weight yields

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	197.656
SC	230.912	222.883
-2 Log L	225.759	181.656

There are 2 more parameters in the larger model (for width and width*weight) and we obtain $-2(L_0 - L_1) = 181.7 - 176.8 = 4.9$ and $P(\chi_2^2 > 4.9) = 0.07$. We barely accept that we can drop width and width*weight at the 5% level.

Forward selection starts by fitting each model with one predictor separately and including the model with the smallest p -value under a cutoff (default=0.05 in PROC LOGISTIC). When we instead have SELECTION=FORWARD in the MODEL statement we obtain the model with only width. Changing the cutoff to SLENTRY=0.15 gives the model with width and color.

Starting from main effects and working backwards by hand, we ended up with width and color in the model. We further simplified color to dark and non dark crabs. Using backwards elimination with a cutoff of 0.05 we ended up with just width. A cutoff of 0.15 and another “by hand” step (at the 0.05 level) yielded weight, color, and weight*color.

Your book considers backwards elimination starting with a three-way interaction model including color, spine condition, and width. The end model is color and width.

6.1.4 AIC & model selection

“No model is correct, but some are more useful than others.” – George Box.

It is often of interest to examine several competing models. In light of underlying biology or science, one or more models may have relevant interpretations within the context of why data were collected in the first place.

In the absence of scientific input, a widely-used model selection tool is the Akaike information criterion (AIC),

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

The $L(\hat{\beta}; \mathbf{y})$ represents model fit. If you add a parameter to a model, $L(\hat{\beta}; \mathbf{y})$ has to increase. If we only used $L(\hat{\beta}; \mathbf{y})$ as a criterion, we'd keep adding predictors until we ran out. The p penalizes for the number of the predictors in the model.

The AIC has very nice properties in large samples in terms of prediction. The smaller the AIC is, the better the model fit (asymptotically).

Model	AIC
W	198.8
$C+Wt+C * Wt$	197.7
$C+W$	197.5
$D + Wt + D * Wt$	194.7
$D + W$	194.0

If we pick one model, it's $W + D$, the additive model with width and the dark/nondark category.

6.2 Diagnostics

GOF tests are global checks for model adequacy.

The data are (\mathbf{x}_i, Y_i) for $i = 1, \dots, N$. The i^{th} fitted value is an estimate of $\mu_i = E(Y_i)$, namely $\widehat{E(Y_i)} = \hat{\mu}_i = n_i \hat{\pi}_i$ where $\pi_i = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}}$ and $\hat{\pi}_i = \frac{e^{\hat{\beta}' \mathbf{x}_i}}{1 + e^{\hat{\beta}' \mathbf{x}_i}}$. The raw residual is what we see Y_i minus what we predict $n_i \hat{\pi}_i$. The Pearson residual divides this by an estimate of $\sqrt{\text{var}(Y_i)}$:

$$e_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}.$$

The Pearson GOF statistic is

$$X^2 = \sum_{i=1}^N e_i^2.$$

The standardized Pearson residual is given by

$$r_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i) (1 - \hat{h}_i)}},$$

where \hat{h}_i is the i^{th} diagonal element of the *hat* matrix

$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2} \mathbf{X} (\mathbf{X}' \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{W}}^{1/2}$ where \mathbf{X} is the design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{N,p-1} \end{bmatrix},$$

and

$$\hat{\mathbf{W}} = \begin{bmatrix} n_1 \hat{\pi}_1 (1 - \hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & n_2 \hat{\pi}_2 (1 - \hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_N \hat{\pi}_N (1 - \hat{\pi}_N) \end{bmatrix}.$$

Alternatively, (6.2, p. 220) defines a deviance residual.

Comments:

- Plots of residuals r_j versus one of the $p - 1$ predictors x_{ij} , for $j = 1, \dots, N$ might show systematic lack of fit (i.e. a pattern). Adding nonlinear terms or interactions can improve fit.
- An overall plot is r_j versus the linear predictor $\hat{\eta}_j = \hat{\beta}' \mathbf{x}_j$. This plot will tell you if the model tends to over or underpredict the observed data for ranges of the linear predictor.

- The r_i are approximately $N(0, 1)$ when n_i is not small. With truly continuous predictors $n_i = 1$ and the residual plots *will* have a distinct pattern.
- I usually flag $|r_i| > 3$ as being ill-fit by the model.
- You can look at individual r_i to determine model fit. For the crab data, this might flag some individual crabs as ill-fit or unusual relative to the model.
- The *model* can't tell the difference between, e.g., two nondark crabs with carapace widths 23 *cm*. You can aggregate over like values of the predictors to slightly “improve” the residuals. This way the approximate $N(0, 1)$ may be a bit better. Ill fitting residuals then become *predictor values* where the aggregated number of events don't match what we'd expect under the model.

Let's look at $W + D$ for the crab data. We'll consider both width and width truncated to an integer cm . The DATA step is

```
data crabs1; input color spine width satell weight;
  weight=weight/1000; color=color-1;
  y=0; n=1; if satell>0 then y=1;
  dark=1; if color=4 then dark=2;
  w=int(width); * round down;
```

Two models fit & r_i plotted:

```
proc logistic data=crabs1 descending; * each crab has n_i=1;
  class dark; model y = dark width;
  output out=diag1 reschi=p h=h xbeta=eta;
data diag2; set diag1; r=p/sqrt(1-h);
proc gplot; plot r*width; plot r*dark; plot r*eta; * plot r_i vs width, dark, eta_i;
proc sort data=crabs1; by w dark; * aggregate over coarser widths;
proc means data=crabs1 noprint; by w dark; var y n; output out=crabs2 sum=sumy sumn;
proc logistic data=crabs2;
  class dark; model sumy/sumn = dark w;
  output out=diag3 reschi=p h=h xbeta=eta;
data diag4; set diag3; r=p/sqrt(1-h);
proc gplot;
  plot r*w; plot r*dark; plot r*eta;
```

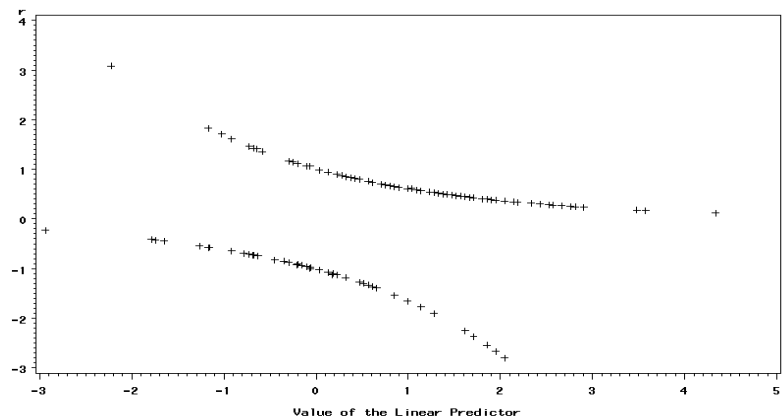
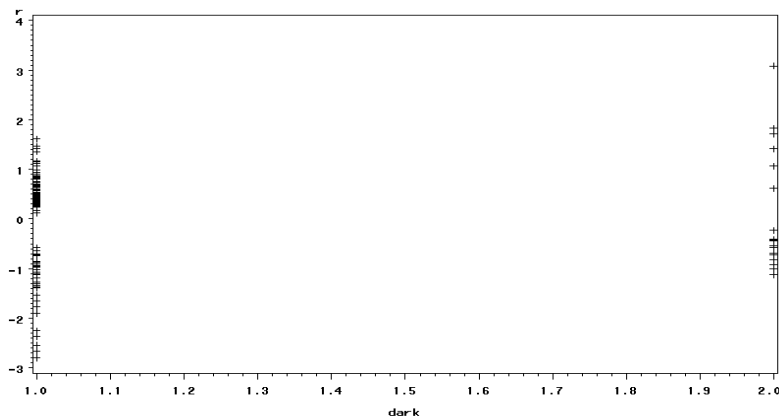
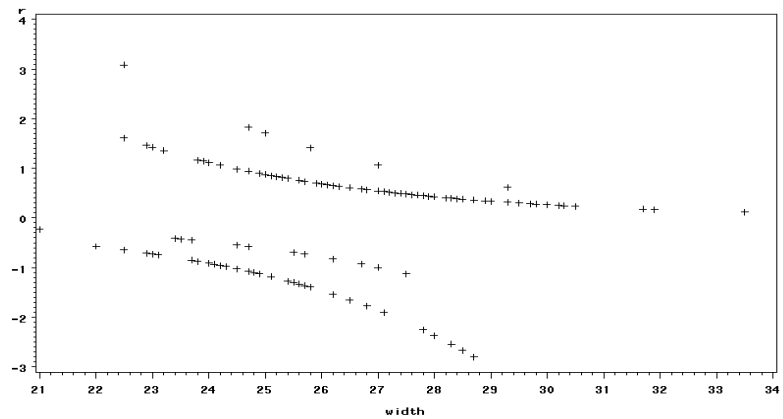
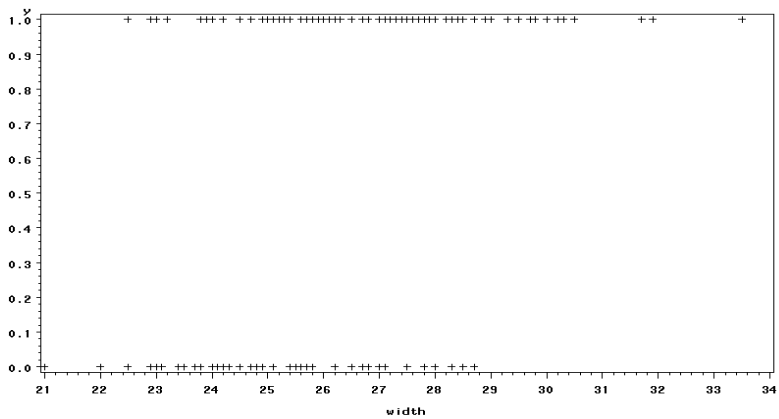


Figure 2: Raw data & residual plots, $n_i = 1$.

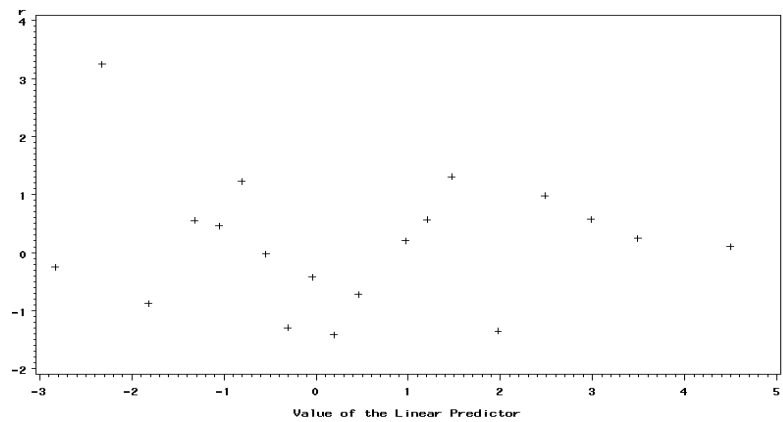
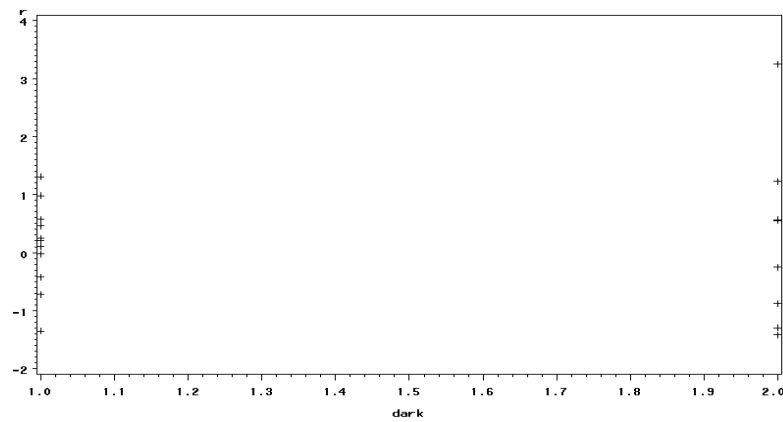
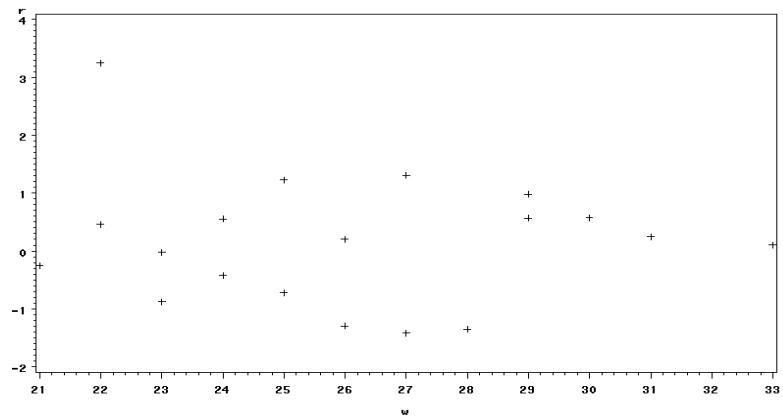
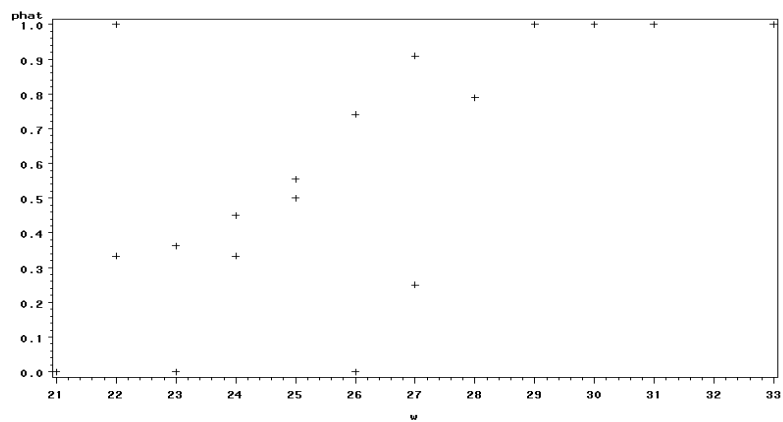


Figure 3: Raw data & residual plots, aggregated.

6.2.4 Influence

Unlike linear regression, the leverage \hat{h}_i in logistic regression depends on the model fit $\hat{\beta}$ as well as the covariates \mathbf{X} . Points that have extreme predictor values \mathbf{x}_i may not have high leverage \hat{h}_i if $\hat{\pi}_i$ is close to 0 or 1. Here are the influence diagnostics available in PROC LOGISTIC:

- Leverage \hat{h}_i . Still may be useful for detecting “extreme” predictor values \mathbf{x}_i .
- $c_i = e_i^2 \hat{h}_i / (1 - \hat{h}_i)^2$ measures the change in the joint confidence region for β when i is left out.
- DFBETA_{ij} is the standardized change in $\hat{\beta}_j$ when observation i is left out.
- The change in the X^2 GOF statistic when obs. i is left out is $\text{DIFCHISQ}_i = e_i^2 / (1 - \hat{h}_i)$.

I suggest looking at plots of c_i vs. i , and possibly the DFBETA's versus i . Let's look output from the aggregated crab data:

```
proc logistic data=crabs2;
class dark; model sumy/sumn = dark w / influence iplots;
```

Case Number	Covariates		Pearson Residual (1 unit = 0.4)					Deviance Residual (1 unit = 0.27)									
	dark1	w	Value	-8	-4	0	2	4	6	8	Value	-8	-4	0	2	4	6
1	-1.0000	21.0000	-0.2431			*				-0.3389			*				
2	1.0000	22.0000	0.4131				*			0.4021				*			
3	-1.0000	22.0000	3.1960						*	2.1987							*
4	1.0000	23.0000	-0.0239				*			-0.0240				*			
5	-1.0000	23.0000	-0.8053			*				-1.0964			*				
6	1.0000	24.0000	-0.3574			*				-0.3578				*			
7	-1.0000	24.0000	0.5160				*			0.4876					*		
8	1.0000	25.0000	-0.6239			*				-0.6189			*				
9	-1.0000	25.0000	1.0202				*			0.9797					*		
10	1.0000	26.0000	0.1850				*			0.1861				*			
11	-1.0000	26.0000	-1.2135			*				-1.4856			*				
12	1.0000	27.0000	1.1509				*			1.2527					*		
13	-1.0000	27.0000	-1.2035			*				-1.2174			*				
14	1.0000	28.0000	-1.1861			*				-1.0905			*				
15	1.0000	29.0000	0.9143				*			1.2671					*		
16	-1.0000	29.0000	0.5469				*			0.7234				*			
17	1.0000	30.0000	0.5503				*			0.7687				*			
18	1.0000	31.0000	0.2469				*			0.3466				*			
19	1.0000	33.0000	0.1054				*			0.1487				*			

Case Number	Hat Matrix Diagonal							Intercept								
	Value	(1 unit = 0.02)						Value	(1 unit = 0.07)							
		0	2	4	6	8	12	16		-8	-4	0	2	4	6	8
1	0.0237		*						-0.0283			*				
2	0.1839				*				0.1929				*			
3	0.0298		*						0.3672					*		
4	0.2486						*		-0.0128			*				
5	0.1467			*					-0.1877		*					
6	0.2844						*		-0.1639		*					
7	0.1331			*					0.0770				*			
8	0.2460						*		-0.0894			*				
9	0.3171							*		0.1331				*		
10	0.2255						*		-0.0340			*				
11	0.1232			*					0.0245			*				
12	0.2244						*		-0.4487		*					
13	0.2755						*		0.2139				*			
14	0.2323						*		0.5935						*	
15	0.1307			*					-0.3317		*					
16	0.0702		*						-0.0836			*				
17	0.0754		*						-0.1490		*					
18	0.0222		*						-0.0352			*				
19	0.00738		*						-0.00875			*				

Case Number	dark1 (1 unit = 0.1)					w (1 unit = 0.08)				
	DfBeta Value	-8	-4	0	2 4 6 8	DfBeta Value	-8	-4	0	2 4 6 8
1	0.0263			*		0.0258			*	
2	0.0390			*		-0.1901		*		
3	-0.4330		*			-0.3253		*		
4	-0.00363			*		0.0125			*	
5	0.3020				*				*	
6	-0.0788			*		0.1569				*
7	-0.1947		*			-0.0578		*		
8	-0.1464			*		0.0751				*
9	-0.7822	*				-0.0551			*	
10	0.0384			*		0.0381			*	
11	0.4486				*			*		
12	0.1861				*					*
13	0.7676				*		*			
14	-0.1497		*			-0.6119	*			
15	0.0598				*					*
16	-0.1153			*		0.0956				*
17	0.0208			*		0.1519				*
18	0.00400			*		0.0358			*	
19	0.000718			*		0.00887			*	

		Confidence Interval Displacement C						Confidence Interval Displacement CBar							
Case Number	Value	(1 unit = 0.05)						Value	(1 unit = 0.03)						
		0	2	4	6	8	12		16	0	2	4	6	8	12
1	0.00147	*						0.00144	*						
2	0.0471	*						0.0385	*						
3	0.3235			*				0.3139			*				
4	0.000252	*						0.000190	*						
5	0.1306		*					0.1115		*					
6	0.0710	*						0.0508	*						
7	0.0471	*						0.0409	*						
8	0.1685		*					0.1270		*					
9	0.7075					*		0.4832					*		
10	0.0129	*						0.00997	*						
11	0.2359			*				0.2069		*					
12	0.4940				*			0.3831			*				
13	0.7600					*		0.5507					*		
14	0.5544					*		0.4256			*				
15	0.1446		*					0.1257		*					
16	0.0243	*						0.0226	*						
17	0.0267	*						0.0247	*						
18	0.00142	*						0.00139	*						
19	0.000083	*						0.000083	*						

Case Number	Delta Deviance (1 unit = 0.32)						Delta Chi-Square (1 unit = 0.66)									
	Value	0	2	4	6	8	12	16	Value	0	2	4	6	8	12	16
1	0.1163	*							0.0606	*						
2	0.2001	*							0.2091	*						
3	5.1483						*		10.5284						*	
4	0.000764	*							0.000763	*						
5	1.3135			*					0.7600	*						
6	0.1788	*							0.1785	*						
7	0.2787	*							0.3071	*						
8	0.5101	*							0.5163	*						
9	1.4429			*					1.5239	*						
10	0.0446	*							0.0442	*						
11	2.4138				*				1.6794	*						
12	1.9524			*					1.7078	*						
13	2.0328			*					1.9990	*						
14	1.6148			*					1.8326	*						
15	1.7313			*					0.9616	*						
16	0.5459	*							0.3217	*						
17	0.6156	*							0.3276	*						
18	0.1215	*							0.0623	*						
19	0.0222	*							0.0112	*						

Obs. 3 has a large e_i (and larger r_i) and is flagged as ill-fit. Obs. 3 also has the largest DIFCHISQ. Obs. 3 is $n_3 = 1$ skinny ($22cm$) dark crab that had a satellite. Recall that the probability of having a satellite increases for light crabs and for wider crabs. This observation does not have much of an effect on $\hat{\beta}$ as measured by c_i and the DFBETAs, perhaps because it's only 1 crab.

Obs. 9 and 13 have the largest c_i . Refining their influence, both 9 and 13 have the largest (in magnitude) DFBETAs for the dark dummy variable. However, with relatively small $|e_i|$, these observations are not ill-fit. Obs. 9 represents $y_9 = 3$ dark crabs out of $n_9 = 6$ that have satellites at width $25cm$. Obs. 13 is $y_{13} = 1$ out of $n_{13} = 4$ dark crabs at $27cm$. These affect the estimate of dark's regression coefficient (adjusting for width) more than the other observations, *but are fit well by the model*. These two could be driving the significance of the color effect.

Obs	w	dark	sumy	sumn
1	21	2	0	1
2	22	1	2	6
3	22	2	1	1
4	23	1	4	11
5	23	2	0	4
6	24	1	9	20
7	24	2	1	3
8	25	1	15	27
9	25	2	3	6
10	26	1	20	27
11	26	2	0	2
12	27	1	20	22
13	27	2	1	4
14	28	1	15	19
15	29	1	10	10
16	29	2	1	1
17	30	1	6	6
18	31	1	2	2
19	33	1	1	1

6.3: $2 \times 2 \times K$ tables

Clinical trial w/ 8 centers, two creams compared to cure infection.

Center $Z = k$	Treatment X	Response Y		$\hat{\theta}_{XY(k)}$
		Success	Failure	
1	Drug	11	25	1.2
	Control	10	27	
2	Drug	16	4	1.8
	Control	22	10	
3	Drug	14	5	4.8
	Control	7	12	
4	Drug	2	14	2.3
	Control	1	16	
5	Drug	6	11	∞
	Control	0	12	
6	Drug	1	10	∞
	Control	0	10	
7	Drug	1	4	2.0
	Control	1	8	
8	Drug	4	2	0.3
	Control	6	1	

Have:

- Y binary outcome (e.g. success/failure of treatment).
- X binary predictor (e.g. treatment).
- Stratum Z (e.g. treatment center).

Want to test $X \perp Y|Z$ versus an alternative. Let

$\pi_{ik} = P(Y = 1|X = i, Z = k)$ and

$$\theta_{XY(k)} = \frac{P(Y = 1|X = 1, Z = k)/P(Y = 2|X = 1, Z = k)}{P(Y = 1|X = 2, Z = k)/P(Y = 2|X = 2, Z = k)}.$$

Recall $X \perp Y|Z$ when $\theta_{XY(k)} = 1$. This happens under the model

$$\text{logit } \pi_{ik} = \alpha + \beta_k^Z.$$

This is an ANOVA-type specification where instead of listing $K - 1$ dummy variables, we concisely include a subscript on Z 's effect β_k^Z . So there are K effects for Z , $\beta_1^Z, \beta_2^Z, \dots, \beta_K^Z$ and they sum to zero.

An additive alternative model specifies

$$\text{logit } \pi_{ik} = \alpha + \beta I\{i = 1\} + \beta_k^Z.$$

Under this model $\theta_{XY(k)} = e^\beta$ for all k . The odds *ratios* are the same across strata, but the strata-specific probabilities of success change with $Z = k$. $X \perp Y|Z$ if we accept $H_0 : \beta = 0$.

The most general alternative is

$$\text{logit } \pi_{ik} = \alpha + \beta I\{i = 1\} + \beta_k^Z + \beta_k^{XZ} I\{i = 1\}.$$

This is a saturated model and allows

$\theta_{XY(1)} \neq \theta_{XY(2)} \neq \dots \neq \theta_{XY(K)}$. $X \perp Y|Z$ if we accept $H_0 : \beta = 0, \beta_k^{XZ} = 0$ for $k = 1, \dots, K$.

Both of these alternatives allow testing $H_0 : X \perp Y|Z$ in PROC LOGISTIC with a Wald test.

Cochran-Mantel-Haenszel statistic

$$\text{CMH} = \frac{\left[\sum_{k=1}^K (n_{11k} - \hat{\mu}_{11k}) \right]^2}{\sum_{k=1}^K \text{var}(n_{11k})},$$

where $\hat{\mu}_{11k} = n_{1+k}n_{+1k}/n_{++k}$ and

$$\text{var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k}-1).$$

- Motivated by retrospective studies, e.g. case-control, so response (column) totals are assumed fixed. Then row (treatment) totals are sufficient and conditioned on. Leaves only one free parameter in each table, say n_{11k} which is hypergeometric under H_0 :
- Null hypothesis is $H_0 : X \perp Y|Z$.
- $\hat{\mu}_{11k} = E(n_{11k})$ and $\text{var}(n_{11k})$ are under H_0 .
- When H_0 true, $\text{CMH} \overset{\bullet}{\sim} \chi_1^2$.

A bit more detail why n_{11k} are hypergeometric ...

	$Y = 1$	$Y = 2$	
$X = 1$	n_{11k}	n_{12k}	n_{1+k}
$X = 2$	n_{21k}	n_{22k}	n_{2+k}
	n_{+1k}	n_{+2k}	n_{++k}

- There are n_{1+k} “red balls” $X = 1$ and n_{2+k} “green balls” $X = 2$.
- We choose n_{+1k} balls (controls $Y = 1$) from the urn. Under independence one cannot tell the difference between a case and a control. The number n_{11k} out of n_{+1k} that are “red,” i.e. exposures $X = 1$, is hypergeometric (under H_0).
- See page 91, (3.16) in Section 3.5.1.
- Back to logistic regression formulation...

The additive alternative looks in a certain direction for deviations from conditional independence $X \perp Y|Z$. It can be more powerful when the additive model truly holds.

The interaction, saturated model can be more powerful when the additive alternative does not hold.

The CMH test is equivalent to a score test for testing $H_0 : \beta = 0$ in the additive model; see your book (p. 232). This test can be carried out in PROC FREQ.

```
data cmh;
input center $ treat response count;
datalines;
a 1 1 11
a 1 2 25
a 2 1 10
a 2 2 27
b 1 1 16
b 1 2 4
...
h 1 1 4
h 1 2 2
h 2 1 6
h 2 2 1
;
proc freq; weight count; tables center*treat*response / cmh;
```

With annotated output:

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.3841	0.0115
2	Row Mean Scores Differ	1	6.3841	0.0115
3	General Association	1	6.3841	0.0115

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	2.1345	1.1776	3.8692
	Logit **	1.9497	1.0574	3.5949
Cohort (Col1 Risk)	Mantel-Haenszel	1.4245	1.0786	1.8812
	Logit **	1.2194	0.9572	1.5536
Cohort (Col2 Risk)	Mantel-Haenszel	0.8129	0.6914	0.9557
	Logit	0.8730	0.7783	0.9792

** These logit estimators use a correction of 0.5 in every cell of those tables that contain a zero.

We see CMH= 6.384 with $p = 0.0115$ and so we reject that $X \perp Y|Z$ in favor of a *common odds ratio* estimated as $\hat{\theta}_{XY} = 2.13 (1.18, 3.87)$.

Alternatively, we can fit the three logit models:

```
data cmh2;
input center $ treat y n; treat=abs(treat-2);
datalines;
a 1 11 36
a 2 10 37
b 1 16 20
b 2 22 32
...
h 1 4 6
h 2 6 7
;
proc logistic data=cmh2; class center; model y/n = center;
proc logistic data=cmh2; class center; model y/n = treat center;
proc logistic data=cmh2; class center; model y/n = treat center treat*center;
```

Label the models (1), (2), and (3) respectively. The fit of (2) corresponds to the alternative in the CMH test:

Type 3 Analysis of Effects

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
treat	1	6.4174	0.0113
center	7	58.4897	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.2554	0.2692	21.7413	<.0001
treat	1	0.7769	0.3067	6.4174	0.0113
center a	1	-0.0667	0.3133	0.0453	0.8315
center b	1	1.9888	0.3556	31.2789	<.0001
center c	1	1.0862	0.3596	9.1236	0.0025
center d	1	-1.4851	0.5707	6.7711	0.0093
center e	1	-0.5866	0.4582	1.6390	0.2005
center f	1	-2.2136	0.9171	5.8260	0.0158
center g	1	-0.8644	0.7016	1.5178	0.2180

We reject $H_0 : \beta = 0$ ($p = 0.0113$) and thus reject $X \perp Y|Z$. We estimate the common odds ratio to be $e^{-0.777} = 2.18$ (1.19, 3.97) (from excised output).

By adding `/ aggregate scale=none;` to the MODEL statement, we find the Pearson GOF $X^2 = 8.03$ on $df = 16 - (1 + 1 + 7) = 7$ with $p = 0.33$. The additive model does not show gross LOF.

Let's examine the full interaction (saturated) model anyway...

The $-2 \text{ Log } L$ from (1) is 283.689 (under Model Fit Statistics) and from (3) is 267.274. The number of parameters added to (1) to get (3) is 8. The p -value is $P(\chi_8^2 > 16.415) = 0.0368$.

We reject that $H_0 : \beta = 0, \beta_k^{XY} = 0$ in the saturated model (3) and hence also reject $X \perp Y|Z$. Notice the p -value is about 3 times larger though; we lost some power by considering a *very* general alternative. By accepting this more complex alternative we have lost interpretability as well, the estimated odds ratio $\hat{\theta}_{XY(k)}$ changes with center k . From (3)'s fit **Type 3 Analysis of Effects**:

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
treat	1	0.0064	0.9362
center	7	24.2036	0.0010
treat*center	7	4.0996	0.7682

The Type III effects table shows we can drop the treat*center from the model and so we go with the analysis and results from the CMH analysis and/or logit analysis on the previous slide.

6.4 Better living through models

Consider an $I \times 2$ table where X is categorical and Y is binary. When the probability of $Y = 2$ is the same for each level of $X = i$, $\pi(i) = P(Y = 2|X = i) = \pi$, we have $X \perp Y$. In terms of log-odds this is

$$\text{logit } \pi(i) = \alpha.$$

1. If X is nominal, allowing a separate probability for each level of X gives

$$\text{logit } \pi(i) = \alpha + \beta_i,$$

for $i = 1, \dots, I$; the saturated model.

2. When X is ordinal, we can use the above alternative model, or instead use scores $u_1 \leq u_2 \leq \dots \leq u_I$ in place of X and fit the model

$$\text{logit } \pi(i) = \alpha + \beta u_i.$$

In the first case a test of $H_0 : \beta_1 = \dots = \beta_I = 0$ is a test of $H_0 : X \perp Y$ versus the most general possible alternative. The test statistic (score, Wald, or LRT) has a χ^2_{I-1} distribution under H_0 .

In the second case a test of $H_0 : \beta = 0$ tests $X \perp Y$ versus a focused, *linear* alternative. The test statistic has a χ^2_1 distribution under H_0 .

The gist of pp. 236–240 is:

- If X is ordinal and the logistic regression model treating X as continuous fits okay, you can increase your power to reject $H_0 : X \perp Y$ by looking in one particular direction (linear log-odds of scores).
- If the model *does not* fit then you can *lose* power by looking in only one place to the exclusion of other alternatives.
- For nominal X we pretty much can only test the saturated model to the intercept model.