

Logistic Regression Example: Horseshoe Crab Data

- Study of nesting horseshoe crabs; taken from “An Introduction to Categorical Data Analysis”, by Alan Agresti, 1996, Wiley.
- Each female crab had a male attached to her in her nest; study investigated factors that affect whether the female had any other males (*satellites*), residing nearby her. Counts of number of satellites were recorded for each female.
- Explanatory variables thought to possibly affect this include the female's:
 - **color** (1=light med, 2=med, 3=dark med, 4=dark);
 - **spine** condition (1=both good, 2=one good, 3=both bad);
 - carapace **width** (cm);
 - **weight** (kg).
- We will focus on predicting presence or absence of **satellites (response)** using only **width (covariate)**.

Data and software code (SAS, SPSS, and R) available on Agresti's website:
<http://www.stat.ufl.edu/~aa/cda/software.html>

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	color	spine	width	satellites	weight														
2		3	28.3	8	3050														
3		4	3	22.5	0	1550													
4		2	1	26	9	2300													
5		4	3	24.8	0	2100													
6		4	3	26	4	2600													
7		3	3	23.8	0	2100													
8		2	1	26.5	0	2350													
9		4	2	24.7	0	1900													
10		3	1	23.7	0	1950													
11		4	3	25.6	0	2150													
12		4	3	24.3	0	2150													
13		3	3	25.8	0	2650													
14		3	3	28.2	11	3050													
15		5	2	21	0	1850													
16		3	1	26	14	2300													
17		2	1	27.1	8	2950													
18		3	3	25.2	1	2000													
19		3	3	29	1	3000													
20		5	3	24.7	0	2200													
21		3	3	27.4	5	2700													
22		3	2	23.2	4	1950													
23		2	2	25	3	2300													
24		3	1	22.5	1	1600													
25		4	3	26.7	2	2600													
26		5	3	25.8	3	2000													
27		5	3	26.2	0	1300													
28		3	3	28.7	3	3150													
29		3	1	26.8	5	2700													
30		5	3	27.5	0	2600													
31		3	3	24.9	0	2100													
32		2	1	29.3	4	3200													
33		2	3	25.8	0	2600													
34		3	2	25.7	0	2000													
35		3	1	25.7	8	2000													
36		3	1	26.7	5	2700													
37		3	3	23.7	0	1850													

Analysis using MTB: first create response variable (satell)

The screenshot shows the Minitab software interface. The main window displays a session log with the following text:

```
2/13/2008 11:03:21 AM

Welcome to Minitab, press F1 for help.
Executing from file: C:\Program Files\Minitab 15\English\Macros\Startup.mac

This Software was purchased for academic use only.
Commercial use of the Software is prohibited.

X
X
```

A "Calculator" dialog box is open, showing the following details:

- Store result in variable:
- Expression: `'satellites' > 0`
- Functions list: All Functions, Absolute value, Antilog, Any, Arcsine, Arccosine, Arctangent, And, Or, Not
- Buttons: Select, Assign as a formula (unchecked), Help, OK, Cancel

The "Worksheet 1" window shows a data table with the following columns and rows:

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
	color	spine	width	satellites	weight	satell					
1	3	3	28.3	8	3050	1					
2	4	3	22.5	0	1550	0					
3	2	1	26.0	9	2300	1					
4	4	3	24.8	0	2100	0					
5	4	3	26.0	4	2600	1					
6	3	3	23.8	0	2100	0					
7	2	1	26.5	0	2350	0					
8	4	2	24.7	0	1900	0					
9	3	1	23.7	0	1950	0					
10	4	3	25.6	0	2150	0					

The Windows taskbar at the bottom shows the Start button and several open applications: Minitab - Untitled, My Documents, Microsoft PowerPoint..., and Microsoft Excel - Hors... The system clock shows 12:49 PM on 2/13/2008.

Fit model, get influence diagnostic graphs, and goodness of fit measures

2/13/2008 11:03:21 AM

Welcome to Minitab, press F1 for help.
Executing from file: C:\Program Files\Minitab 15\Engl

This Software was purchased for academic use only.
Commercial use of the Software is prohibited.

X
X

	C1	C2	C3	C4	C5	C6
	color	spine	width	satellites	weight	satel
1	3	3	28.3	8	3050	
2	4	3	22.5	0	1550	
3	2	1	26.0	9	2300	
4	4	3	24.8	0	2100	
5	4	3	26.0	4	2600	1
6	3	3	23.8	0	2100	0
7	2	1	26.5	0	2350	0
8	4	2	24.7	0	1900	0
9	3	1	23.7	0	1950	0
10	4	3	25.6	0	2150	0

Binary Logistic Regression

Response in response/frequency format
Response: satell
Frequency (optional):
Response in event/trial format
Number of events:
Number of trials:
Model:
width
Factors (optional):
Select
Help
Graphs... Options...
Results... Storage...
OK Cancel

Binary Logistic Regression - Graphs

Diagnostic Plots

Involving Event Probability
 Delta chi-square vs probability
 Delta deviance vs probability
 Delta beta (Standardized) vs probability
 Delta beta vs probability

Involving Leverage (Hi)
 Delta chi-square vs leverage
 Delta deviance vs leverage
 Delta beta (Standardized) vs leverage
 Delta beta vs leverage
Help OK Cancel

Note: MTB calls categorical variables **factors**.

In **Graphs**, select these influence measures

In **Results**, select maximum number of items to display

Output: Fitted Model

Binary Logistic Regression: satell versus width

Link Function: Logit

Response Information

Variable	Value	Count	
satell	1	111	(Event)
	0	62	
Total		173	

The odds of a crab having a satellite are 1.64 times the odds for crabs that are 1 cm shorter in width (odds increase by 64% per unit increase in width).

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-12.3508	2.62873	-4.70	0.000			
width	0.497231	0.101736	4.89	0.000	1.64	1.35	2.01

Log-Likelihood = -97.226

Test that all slopes are zero: $G = 31.306$, $DF = 1$, $P\text{-Value} = 0.000$

Width is a significant predictor of incidence of satellites, as compared to just using the mean sample proportion, 111/173.

More on the Fitted Model

$$\hat{\pi}(x) = \frac{e^{-12.351+0.497x}}{1 + e^{-12.351+0.497x}}$$

At the mean width of $x=26.3$, the predicted prob of a satellite is 0.674, which corresponds to an odds of $0.674/(1-0.674)=2.07$.

At width of $x=26.3+1=27.3$, the predicted prob of a satellite is 0.773, which corresponds to an odds of $0.773/(1-0.773)=3.40$.

But this is an odds increase of 64%, i.e. $3.40=2.07(1.64)$.

Output: Goodness-Of-Fit

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	55.1779	64	0.776
Deviance	69.7260	64	0.291
Hosmer-Lemeshow	3.5615	8	0.894
Brown:			
General Alternative	1.1162	2	0.572
Symmetric Alternative	1.1160	1	0.291

Model passes all GOF tests

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Value	Group										Total	
	1	2	3	4	5	6	7	8	9	10		
1												
Obs	5	8	11	8	15	12	14	16	16	6	111	
Exp	5.4	7.6	8.6	9.9	15.4	12.9	13.3	16.8	15.3	5.7		
0												
Obs	14	10	6	9	9	6	3	4	1	0	62	
Exp	13.6	10.4	8.4	7.1	8.6	5.1	3.7	3.2	1.7	0.3		
Total	19	18	17	17	24	18	17	20	17	6	173	

Output: Predictive Ability

Measures of Association:

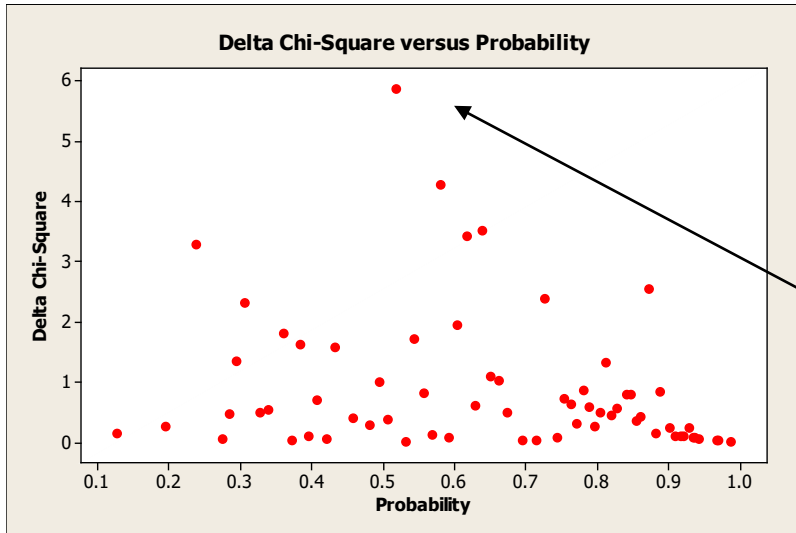
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	5059	73.5	Somers' D	0.48
Discordant	1722	25.0	Goodman-Kruskal Gamma	0.49
Ties	101	1.5	Kendall's Tau-a	0.22
Total	6882	100.0		

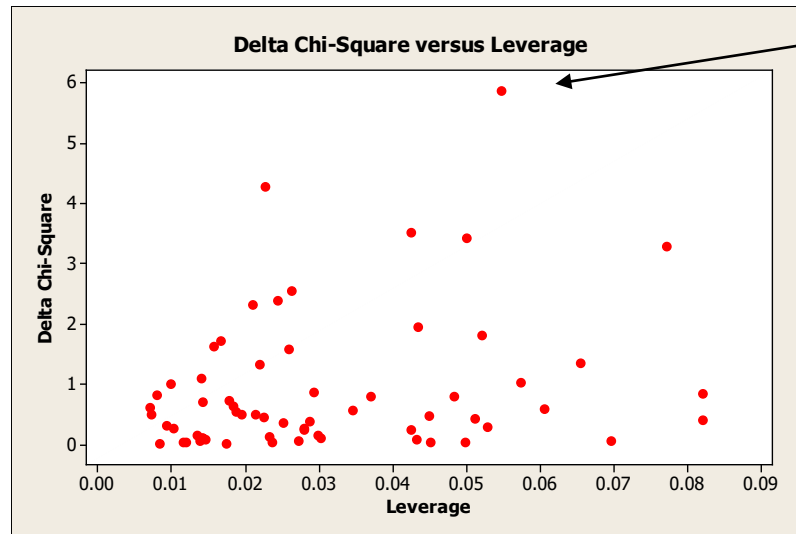
Use % concordant and discordant to compare the model to alternative models with different predictors and alternative link functions.

The **Summary Measures** attempt to summarize the concordant and discordant information. These measures vary between -1 and 1, with larger values denoting greater predictive/explanatory capability, and are the logistic regression equivalent of correlation between X and Y.

Output: Diagnostic Plots



A few obs are influential (leverage plot) and poorly fit (probability plot), esp. case #22 (Delta Chi-Square=5.86).



Delta values in excess of 3.8 are deemed too high.

Logistic Regression in SAS

```
proc logistic;  
model satell = width;
```

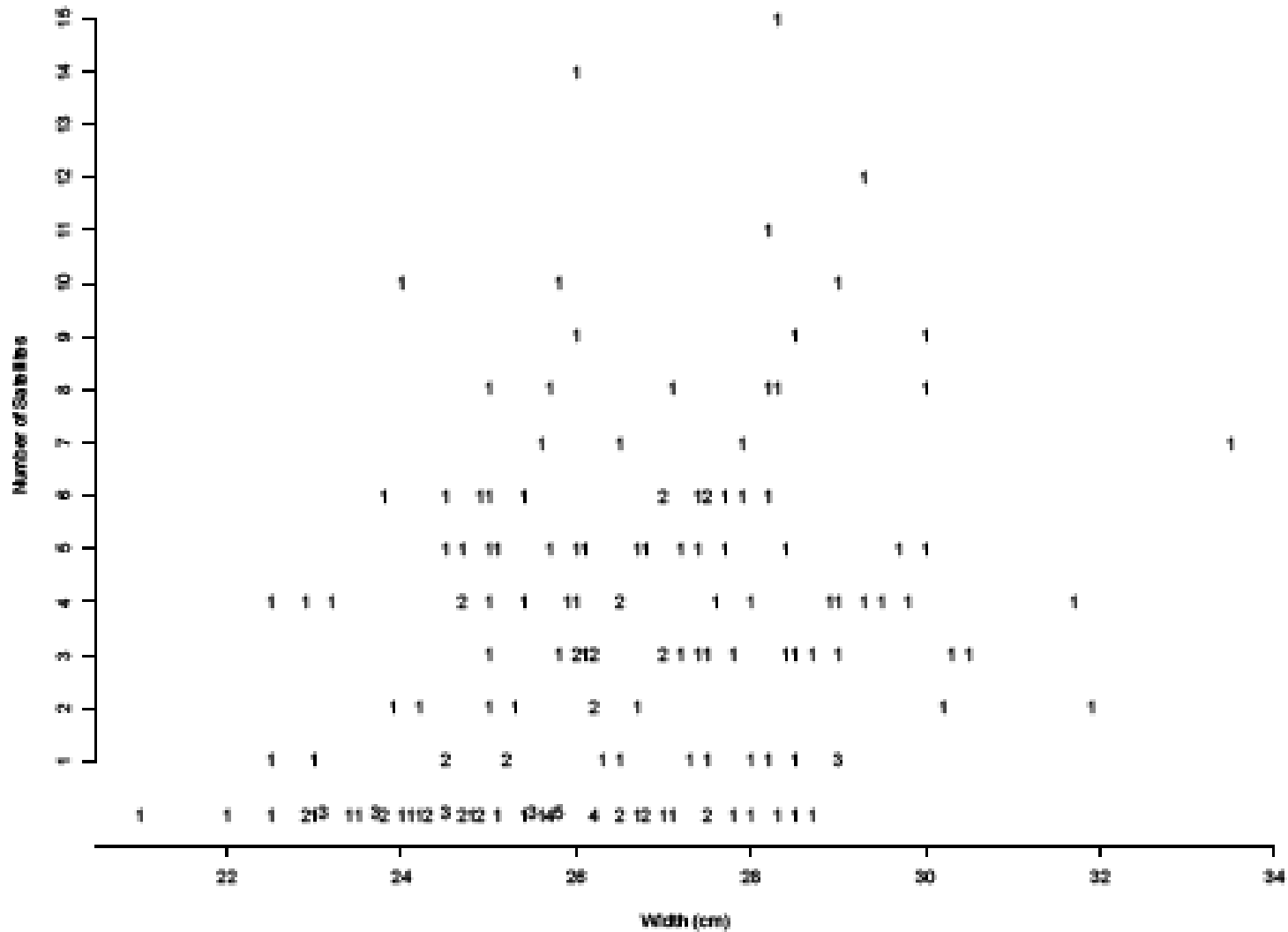
Logistic Regression in SPSS

ANALYZE > REGRESSION > BINARY LOGISTIC

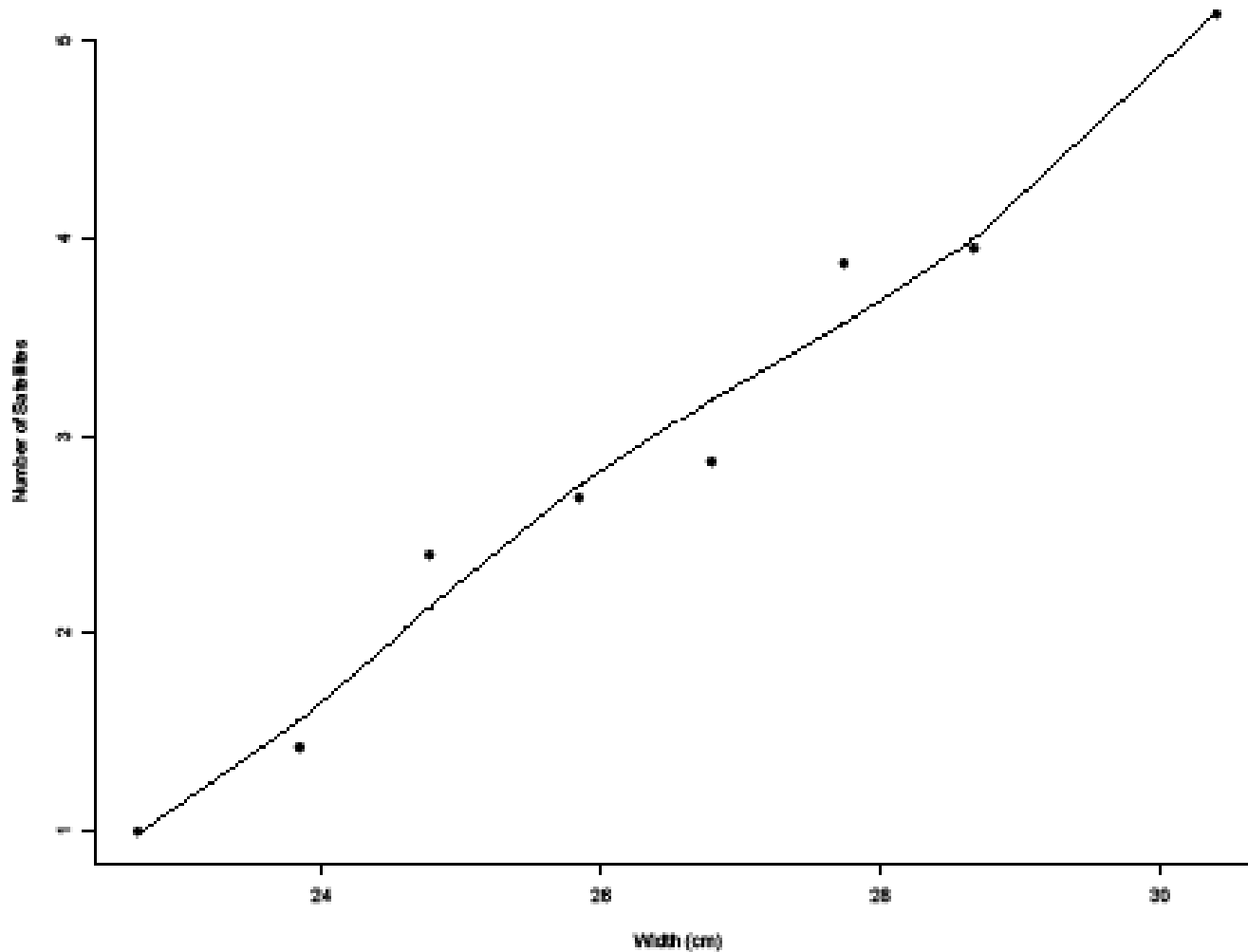
In LOGISTIC REGRESSION dialog box enter:

- response: **satell**
- covariate: **width**

Poisson Regression: Plot number of satellites vs. width



Smooth the plot (aggregate counts over width categories)



Poisson regression with log link (in R)

```
glm(formula = satellites ~ width, family = poisson(link = log),  
     data = crabs)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.8526	-1.9884	-0.4933	1.0970	4.9221

family=binomial
for logistic reg.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.30476	0.54224	-6.095	1.10e-09	***
width	0.16405	0.01997	8.216	< 2e-16	***

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 632.79 on 172 degrees of freedom
Residual deviance: 567.88 on 171 degrees of freedom
AIC: 927.18

Fitted model:

$$\log(\mu) = -3.305 + 0.164 \text{ Width}$$

LRT for comparing
model with and without
width is: 632.8-
567.9=64.9 on 1 df (sig.)

Poisson regression with identity link (in R)

```
glm(formula = satellites ~ width, family = poisson(link =  
  identity),  
  data = crabs, start = coef(log.fit))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.9113	-1.9598	-0.5405	1.0406	4.7988

Coefficients:

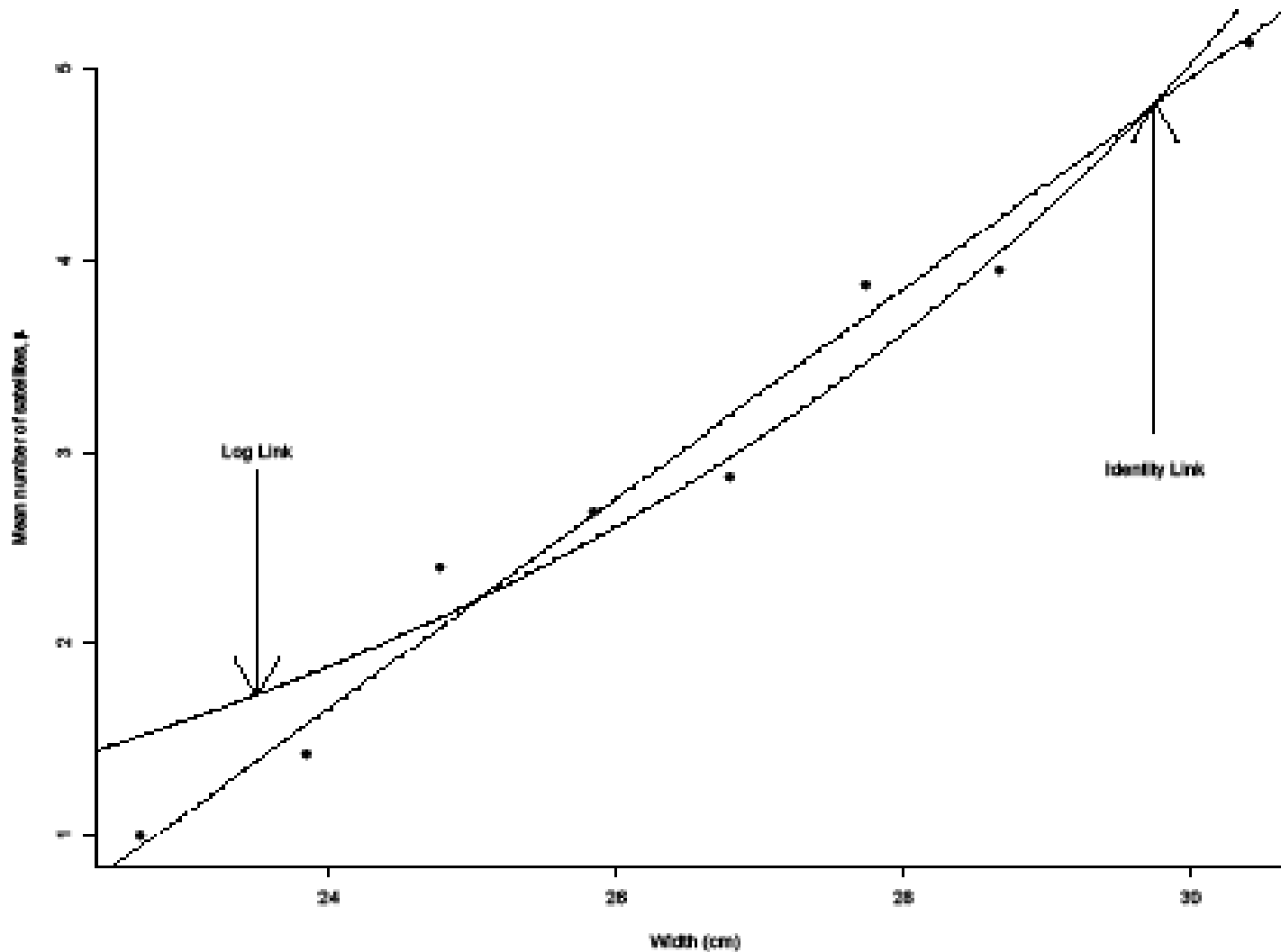
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-11.52547	0.67767	-17.01	<2e-16	***
width	0.54925	0.02968	18.50	<2e-16	***

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 632.79 on 172 degrees of freedom
Residual deviance: 557.71 on 171 degrees of freedom
AIC: 917.01

Fitted model: $\mu = -11.525 + 0.549 \text{ Width}$

Comparison of fitted lines for log vs. identity links



Identity link is a little better. (Verified by AIC.)

Note: cannot use LRT for this, must use AIC.