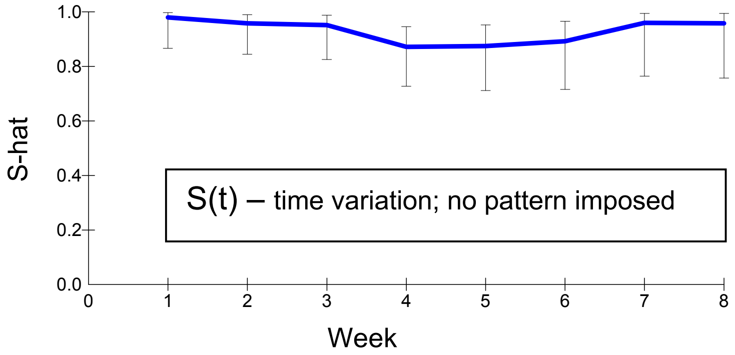
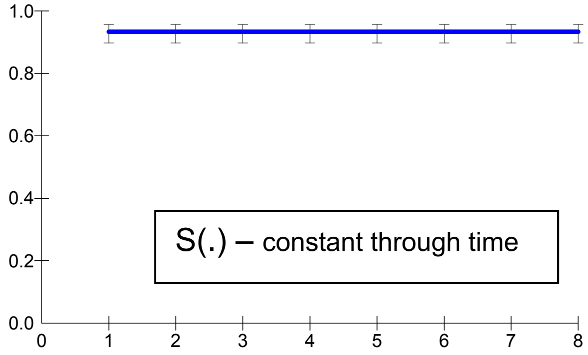
Modeling known-fate data on survival – binomials & logistic regression

Consider the following dataset:

Week	# Females	# Died	# Survived	Ppn Survived	(p*(1-p))/n	se-hat
1	48	1	47	0.979	0.00042	0.021
2	47	2	45	0.957	0.00087	0.029
3	41	2	39	0.951	0.00113	0.034
4	39	5	34	0.872	0.00287	0.054
5	32	4	28	0.875	0.00342	0.058
6	28	3	25	0.893	0.00342	0.058
7	25	1	24	0.960	0.00154	0.039
8	24	1	23	0.958	0.00166	0.041
TOTAL	**284**	**19**	**265**	**0.9331**	**0.00022**	**0.015**





Model = S(.) – Constant survival – all data are pooled

MLE's: S-hat = 0.933, se-hat = 0.015

-2lnL = $-2 \times \ln(0.933^{265} \times 0.067^{19}) = 139.47$

AIC = 139.47 + 2(1) = 141.47

AICc = 139.47 + 2(1) + 2(1)(1+1)/(284-1-1) = 141.47+ ~0.014 = ~141.484

Model = S(t) – Survival varies freely among weeks

-2lnL = $-2 \times \ln \{[(0.979)^{47}(1-0.979)^{1}]$
$\times [(0.957)^{45}(1-0.957)^{2}]$
$\times [(0.951)^{39}(1-0.951)^{2}]$
$\times [(0.871)^{34}(1-0.871)^{5}]$
$\times [(0.875)^{28}(1-0.875)^{4}]$
$\times [(0.893)^{25}(1-0.893)^{3}]$
$\times [(0.960)^{24}(1-0.960)^{1}]$
$\times [(0.958)^{23}(1-0.958)^{1}]\}$

-2lnL = 132.01

AIC = 132.01 + 2(8) = 148.01

AICc = 132.01 + 2(8) + 2(8)(8+1)/(284-8-1) = 132.01 + 16 + 0.52 = 148.53

Likelihood Ratio Test

H_0: S(.) fits the data as well as S(t)

$X^2 = -2\ln[L(S.)/L(S_t)]$
$= -2\ln[(5.17686E-31)/(2.16094E-29)]$
$= -2\ln[0.023956514]$
$= 7.463$

Alternatively, [-2lnL(S(.)] – [-2lnL(S(t)] = 139.47 - 132.01 = 7.46

df = 7 … S(t) has 8 parameters and S(.) has 1, and 8-1=7.
P = 0.3823 … cannot reject the null hypothesis – considering weekly estimates does not improve the fit of the model to the data – don't have strong evidence of weekly variation in survival rate.

Information-theoretic - AIC comparison

ΔAICc for S(.) = 0.00
ΔAICc for S(t) = 7.05

S(.) is a more parsimonious model for this dataset than is S(t)

BUT WHAT OF OTHER MODELS?

Imagine that the average temperature (in degrees C) varied by week

1	2	3	4	5	6	7	8
11	9	8	4	6	5	10	9

To find out if there's evidence of a relationship between weekly survival rate and temperature, we need a more flexible method: LOGISTIC REGRESSION is one such tool and one we'll use often.

Building and Evaluating Competing Models of Binomial Processes

So far, we've simply worried about estimating *p* based on an observed number of heads out of *N* trials. This is useful (9/37 walruses survive; p-hat = 0.243, se-hat = 0.07). But … this is also inadequate for most problems. That is, we'll want to know if *p* (probability of surviving 1 year) varies among the walruses based on factors such as age, year, individual characteristics, etc. We need to evaluate a variety of models to check for **heterogeneity** in *p* (or *S*). A useful method for evaluating competing models of binomial responses is logistic regression.

This method assumes that errors in Y are **binomially distributed** and uses a **logit link** between Y and the regression string. Let's work through what these features entail.

$$\Pr(y_i = 1 \mid \beta_0, \beta_1, x_i) = \frac{\exp(\beta_0 + \beta_1 \cdot X_1)}{1 + \exp(\beta_0 + \beta_1 \cdot X_1)} = \pi(x) = E(Y \mid x)$$

So, $E(y_i \mid x_i) = \pi(x)$ and $y_i = \pi(x) + \varepsilon_i$,

where the quantity can assume 1 of 2 possible values:

(1) if y_i = 1, then $\varepsilon_i = 1 - \pi(x)$ with probability $\pi(x)$; and

(2) if y_i = 0, then $\varepsilon_i = -\pi(x)$ with probability $1 - \pi(x)$.

(3) Thus, ε has a distribution with mean 0 and variance = $\pi(x) \cdot [1 - (\pi(x)]$.

The use of $\pi(x)$ provides us with a simpler notation for working with the regression problem and nicely shows how the errors are explicitly assumed to be **binomially distributed**.

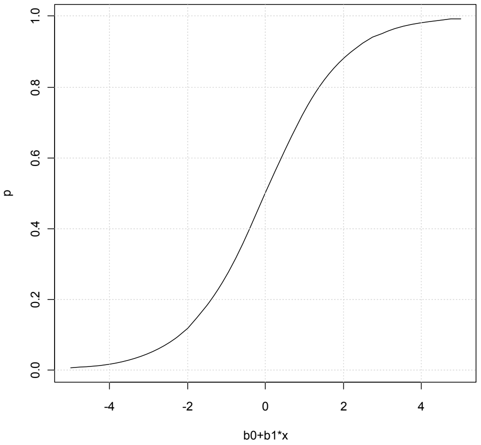
We can use the logit transformation on both sides of the equation and simplify the right side considerably! Aha, so that's what the **logit link** refers to – let's look at that a bit more closely.

$$\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 \cdot X_1$$

This is nice but leads to a transformed response variable: ln(S/(1-S)) = log odds or the natural logarithm of the ratio of S & 1-S. Here, the relationship between ln(S/(1-S)) and $\beta_0 + \beta_1 \cdot X_1$ is linear. The relationship between (Pr(y=1)) and

$\frac{\exp(\beta_0 + \beta_1 \cdot X_1)}{1 + \exp(\beta_0 + \beta_1 \cdot X_1)}$ is S-shaped and constrains the resulting values so that they range between 0 & 1, which is appropriate for probabilities.

To get some experience with how the function behaves, try using a range of values from -5 to +5 (e.g., by step sizes of 0.25) in place of $(\beta_0 + \beta_1 \cdot X_1)$ in the equation. When you do, you'll see that the resulting values range from ~0 to ~1 and are at exactly 0.5 when $(\beta_0 + \beta_1 \cdot X_1) = 0$, which makes sense given that exp(0)=1.
In **R**, you can obtain these probabilities from the logistic distribution with: `plogis(seq(-5,5,.25))`



The error term and link function lead to a pdf and likelihood function for the logistic equation. These functions should look quite familiar to you.

$$L(\hat{\beta}_0, \hat{\beta}_1 \mid X) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

As we discussed before, actual estimation is done on lnL.
But … the basic concepts here are that:

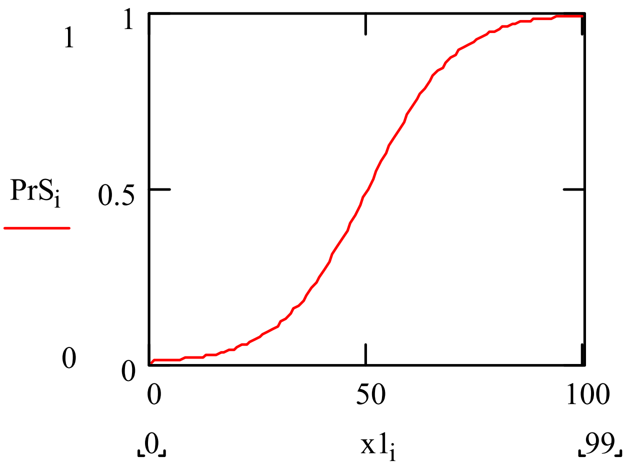
1. $\pi(x)$ gives the conditional probability that y=1 given *x* for any estimates of β_0 and β_1.
2. $1-\pi(x)$ gives the conditional probability that y=0 given *x* for any estimates of β_0 and β_1.
3. For those pairs (x_i,y_i) where y_i = 1 the contribution to the likelihood function is $\pi(x)$ [1].
4. For those pairs (x_i,y_i) where y_i = 0 the contribution to the likelihood function is $1-\pi(x)$ [1].
5. Thus, the most likely estimates of β_0 and β_1 are those that lead to high values of $\pi(x)$ when y_i = 1 and low values of $\pi(x)$ when y_i = 0 (OR high values of $1-\pi(x)$ when y_i = 0).

Beyond this set of difficulties, everything else about model building and model selection is the same as what you learned in regression problems with normally distributed errors in **y**. The principles that we're using to guide analysis of linear regression also apply in logistic regression (model lists, model selection, etc.).

Useful information about interpreting beta's in simple logistic regression

The parameter β_1 determines the rate of increase or decrease of the S-shaped curve. The sign of β_1 indicates whether the curve ascends or descends, and the rate of change increases as $|\beta_1|$ increases. When the model holds with β_1=0, the

Pr(S=1) simplifies to a constant value = $\frac{e^{\beta_0}}{1+e^{\beta_0}}$. In this case, $\pi(x)$ is identical for all *x*, i.e., *y* is independent of *x*. Examine the curve below and note that the rate of change in $\pi(x)$ per unit change in *x* varies. Straight lines drawn tangent to the curve at any particular *x* value describe the rate of change at that point. The logistic regression's line has a slope equal to $\beta_1 \cdot \pi(x) \cdot [1-\pi(x)]$. Thus, the line tangent to the curve at *x* for which $\pi(x)$ = 0.5 has slope β_1(0.5)(0.5), and when $\pi(x)$ = 0.9 or 0.1, the line has slope β_1(0.9)(0.1). Thus, the line has its steepest slope when $\pi(x)$ = 0.5, and the slope approaches 0 as $\pi(x)$ approaches 0 or 1. It turns out for a simple model that's only $(\beta_0 + \beta_1 \cdot X_1)$, the *x* value for which $\pi(x)$ = 0.5 can be determined by *x* = - β_0 / β_1 and is sometimes called the median effective level (EL$_{50}$).



B0=-5, B1 =0.1

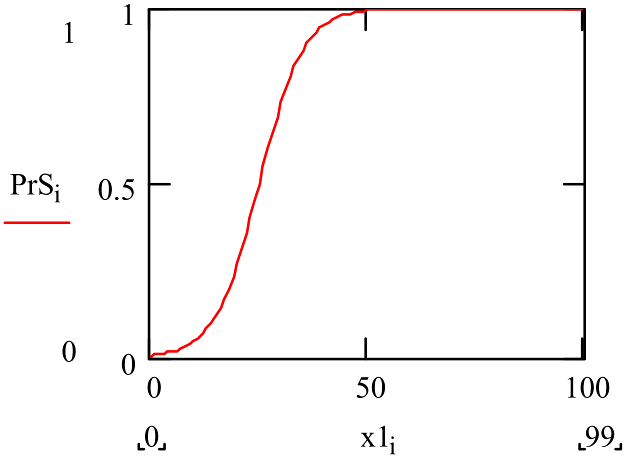
-(-5/0.1)= EL$_{50}$ = 50

Pr(S=1|x = 50) = 0.5

$$\frac{e^{-5+.1\cdot 50}}{1+e^{-5+.1\cdot 50}} = 0.5$$

slope @ x=50 = .1(.25) = .025

slope @ x=5 is .1(.011)(.989)=0.001



B0=-5, B1 =0.2

-(-5/0.2)= EL$_{50}$ = 25

Pr(S=1|x = 25) = 0.5

$$\frac{e^{-5+.2\cdot 25}}{1+e^{-5+.2\cdot 25}} = 0.5$$

slope @ x= 25 = .2(.25) = .05

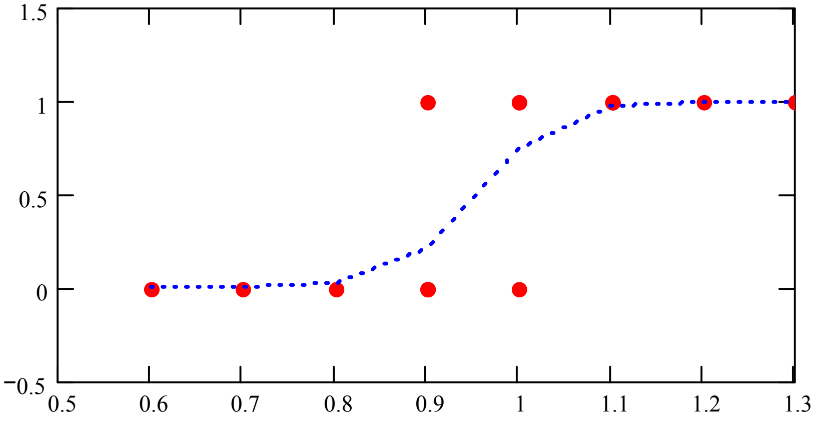
slope @ x = 5 is .2(.018)(.982) = 0.004

Logistic Regression – MLE

DATA

x=c(0.6, 0.7, 0.8, 0.9, 0.9, 0.9, 0.9, 1.0, 1.0, 1.0, 1.1, 1.1, 1.1, 1.2, 1.3, 1.3, 1.3)
y=c(0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1)

Graphical Representation of data and fitted line



Calculating the Likelihood of different combinations of parameter estimates

$$\pi_i = \frac{\exp(\beta_0+\beta_0\cdot x_i)}{1+\exp(\beta_0+\beta_0\cdot x_i)} \qquad \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_0 \cdot x_i$$

$$L(\beta_0,\beta_1 \mid N) = \prod_{i=1}^{N} \frac{n_i!}{y_i!\cdot(n_i-y_i)!}\pi_i^{y_i}\cdot(1-\pi_i)^{(n_i-y_i)}$$

$$\ln L(\beta_0,\beta_1) = \sum_{i=1}^{N}\{y_i\cdot\ln[\pi(x_i)]+(1-y_i)\cdot\ln[1-\pi(x_i)]\}$$

lnL(-15.000, 21.000) = -20.383

lnL(-30.000, 11.000) = -175.700

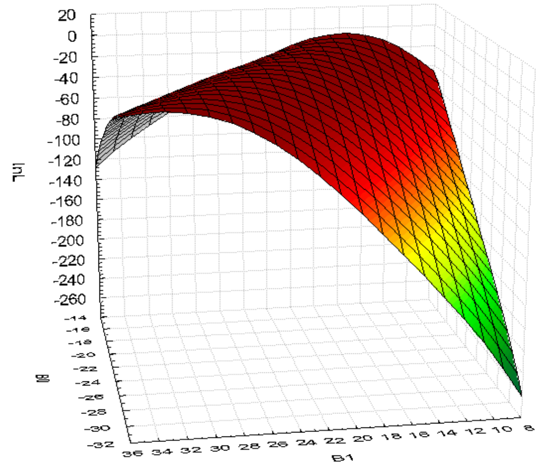
lnL(-22.000, 23.000) = -4.347

MLE’s are:

Parameter	Estimate	Std. Error	95% Confidence Intervals	
β_0	-22.7465	12.26336	-46.7823	1.28922
β_1	23.8061	12.81938	-1.3194	48.93159
lnL	-4.34485			

Calculating the Likelihood and log-Likelihood of β$_0$ = -22.7465 and β$_1$ = 23.8061 for the dataset under consideration (obtain π_i using $\pi_i = \frac{\exp(\beta_0+\beta_0 \cdot x_i)}{1+\exp(\beta_0+\beta_0 \cdot x_i)}$)

π_i	y_i	$\pi_i^{y_i}$	$(1-\pi_i)$	$(1-y_i)$	$(1-\pi_i)^{(1-y_i)}$	$\pi_i^{y_i} \cdot (1-\pi_i)^{(1-y_i)}$	$\ln(\pi_i^{y_i} \cdot (1-\pi_i)^{(1-y_i)})$
0.000	0	1.000	1.000	1	1.000	1.000	-0.000
0.002	0	1.000	0.998	1	0.998	0.998	-0.002
0.024	0	1.000	0.976	1	0.976	0.976	-0.024
0.211	0	1.000	0.789	1	0.789	0.789	-0.237
0.211	0	1.000	0.789	1	0.789	0.789	-0.237
0.211	0	1.000	0.789	1	0.789	0.789	-0.237
0.211	1	0.211	0.789	0	1.000	0.211	-1.558
0.743	0	1.000	0.257	1	0.257	0.257	-1.357
0.743	1	0.743	0.257	0	1.000	0.743	-0.298
0.743	1	0.743	0.257	0	1.000	0.743	-0.298
0.969	1	0.969	0.031	0	1.000	0.969	-0.032
0.969	1	0.969	0.031	0	1.000	0.969	-0.032
0.969	1	0.969	0.031	0	1.000	0.969	-0.032
0.997	1	0.997	0.003	0	1.000	0.997	-0.003
1.000	1	1.000	0.000	0	1.000	1.000	-0.000
1.000	1	1.000	0.000	0	1.000	1.000	-0.000
1.000	1	1.000	0.000	0	1.000	1.000	-0.000
Properties of last 2 columns are used to get the likelihood & lnL of the pair of values for β$_0$ and β$_1$ that were used to calculate the values of *pi$_i$* in the table.						**Product of column =** **0.013** = likelihood	**Sum of column =** **-4.35** = log-likelihood



Do you think that this is a desirable shape? Now, take a look back at the estimates, the associated SE's, and the dataset: we've got a small sample size and poor precision.

Estimated Variance-Covariance Matrix

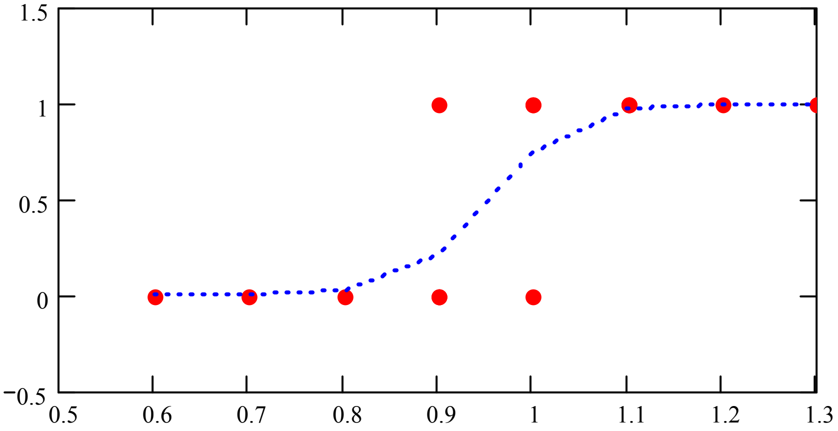
Parameter	β_0	β_1
β_0	150.390	-156.823
β_1	-156.823	164.336

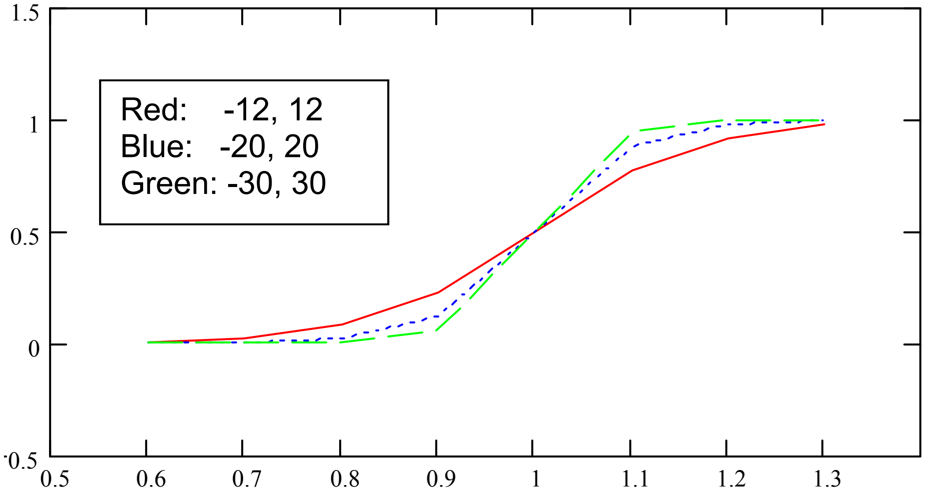
Do these values make sense given the estimates & graph you saw above?

Estimated Correlation Matrix

Correlation or $\left(\frac{cov_{1,2}}{se_1 \cdot se_2}\right)$ for the 2 estimates = -156.823/(12.26336 x 12.81938) = ?

Parameter	β_0	β_1
β_0	1.000000	-0.997550
β_1	-0.997550	1.000000





Goodness of Fit

Deviance Residuals (formal testing available [chi-square distribution] but problematic)

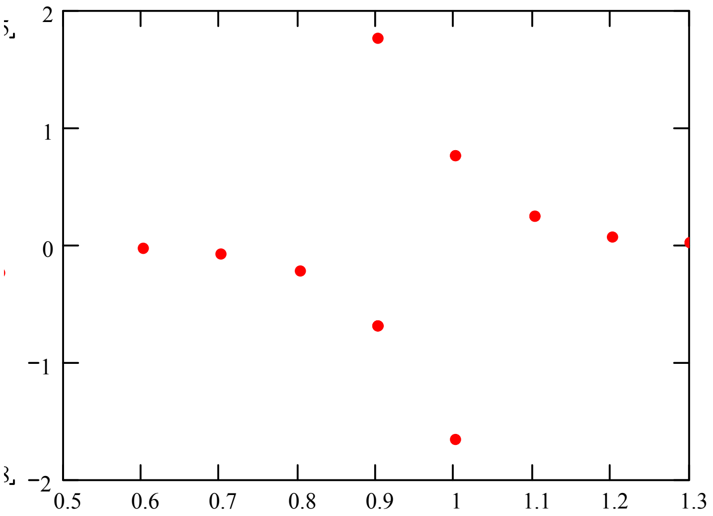
$$dev_i = \pm\{-2[Y_i \cdot \log_e(\hat{\pi}_i) + (1 - Y_i) \cdot \log_e(1 - \hat{\pi}_i)]\}^{1/2}$$

The sign of a *deviance residual* is positive if $y_i \geq \hat{\pi}_i$ and negative if $y_i < \hat{\pi}_i$

π_i	y_i	Deviance$_i$	Dev$_i^2$
0.000	0	-0.021	0.0004
0.002	0	-0.068	0.0046
0.024	0	-0.221	0.0488
0.211	0	-0.688	0.4733
0.211	0	-0.688	0.4733
0.211	0	-0.688	0.4733
0.211	1	1.765	3.1152
0.743	0	-1.648	2.7159
0.743	1	0.771	0.5944
0.743	1	0.771	0.5944
0.969	1	0.251	0.0630
0.969	1	0.251	0.0630
0.969	1	0.251	0.0630
0.997	1	0.077	0.0059
1.000	1	0.023	0.0005
1.000	1	0.023	0.0005
1.000	1	0.023	0.0005

$$\sum_{i=1}^{n} (dev_i)^2 = ModelDeviance = -2\ln L = \text{-2 (-4.3449)} = \ 8.6898$$

Deviance Residuals (y-axis) against values of *x* (x-axis)

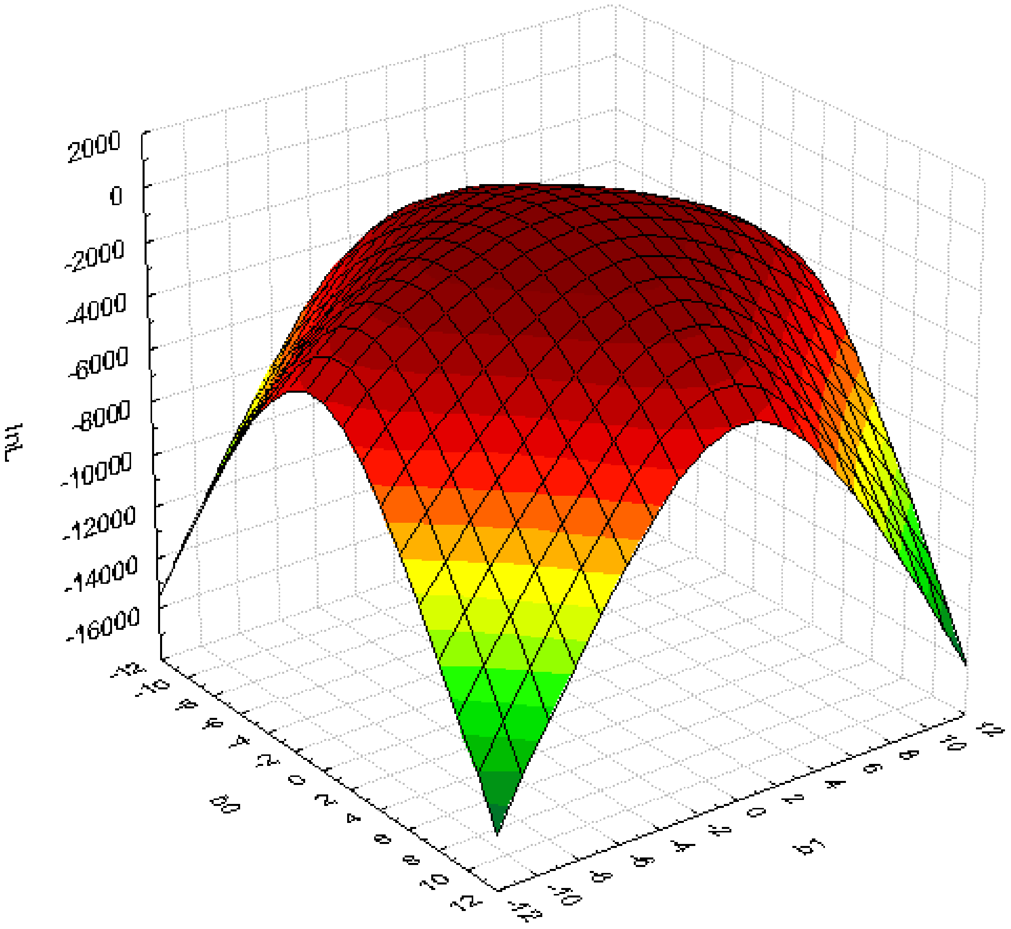


More interesting problems will have:

- Larger datasets
- Competing models

Under these circumstances, we'll:

- Find the MLEs for each model,
- Use the lnL values to compare models (more on how to do that later)
- Evaluate Goodness of fit, if possible, for our most complex model (more on that later too!).

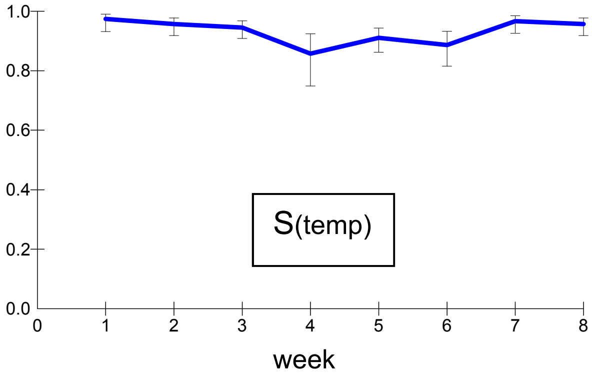


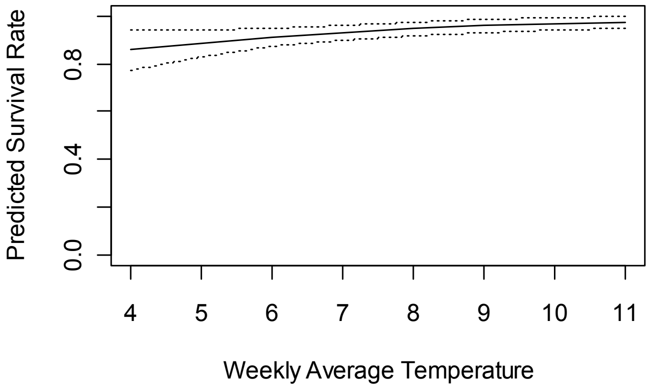
OK – so now that you have some background in logistic regression. Let's go back to our dataset on weekly survival and see if we can work with some more interesting models.

Week	# Females	# Died	# Survived	Ppn Survived	(p*(1-p))/n	se-hat
1	48	1	47	0.979	0.00042	0.021
2	47	2	45	0.957	0.00087	0.029
3	41	2	39	0.951	0.00113	0.034
4	39	5	34	0.872	0.00287	0.054
5	32	4	28	0.875	0.00342	0.058
6	28	3	25	0.893	0.00342	0.058
7	25	1	24	0.960	0.00154	0.039
8	24	1	23	0.958	0.00166	0.041
TOTAL	**284**	**19**	**265**	**0.9331**	**0.00022**	**0.015**

Average temperature (in degrees C) varied by week. Does this relate to variation in survival?

Week	1	2	3	4	5	6	7	8
Temp	11	9	8	4	6	5	10	9





Model S(temp)

The MLEs for this model are $\hat{\beta}_0$ = 0.75; $\hat{\beta}_1$ = 0.26 – consider their role in the likelihood equation and how they would have been found.

$$\pi_i = \frac{\exp(\beta_0+\beta_0 \cdot x_i)}{1+\exp(\beta_0+\beta_0 \cdot x_i)} \qquad \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_0 \cdot x_i$$

$$lnL(\beta_0, \beta_1) = \sum_{i=1}^{N} [y_i \cdot (\beta_0 + \beta_0 \cdot x_i)] - \sum_{i=1}^{N} \ln[1 + \exp(\beta_0 + \beta_0 \cdot x_i)]$$

$\hat{\pi}_i$ = exp(0.75 + (0.26 x temp$_{\text{week } i}$)) / [1 + exp(0.75 + (0.26 x temp$_{\text{week } i}$))]

$$\textbf{-2lnL} = -2 \times \ln \{[(0.974)^{47}(1-0.974)^{1}] \times [(0.957)^{45}(1-0.957)^{2}] \times [(0.945)^{39}(1-0.945)^{2}]$$
$$\times \quad [(0.858)^{34}(1-0.858)^{5}] \times [(0.910)^{28}(1-0.910)^{4}] \times [(0.887)^{25}(1-0.887)^{3}]$$
$$\times \quad [(0.967)^{24}(1-0.967)^{1}] \times [(0.957)^{23}(1-0.957)^{1}]\}$$

-2lnL = 132.64

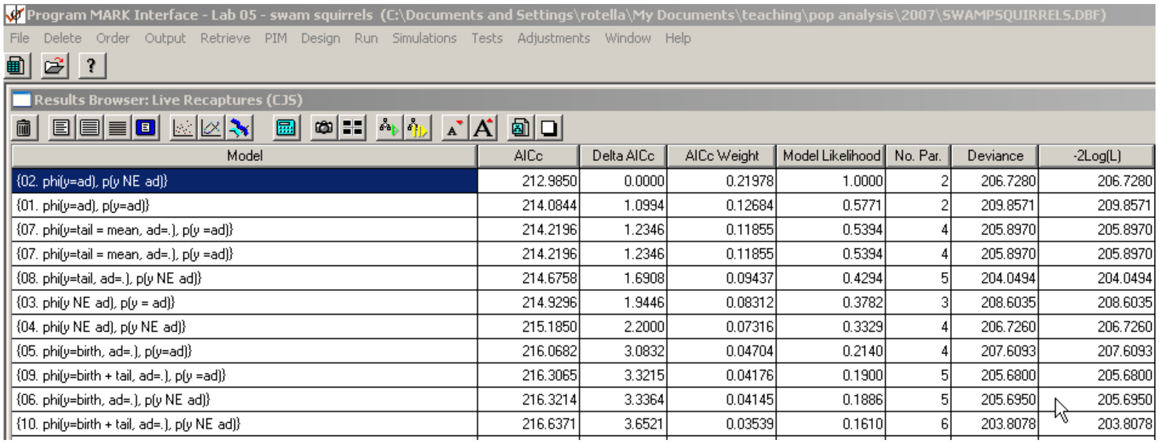
AIC = 132.64 + 2(2) = 136.64

AICc= 132.64 + 2(2) + 2(2)(2+1)/(284-2-1) = 132.64 + 4 + 0.04 = 136.68

Model	AICc	Delta AICc	AICc Weight	#Par	~-2lnL	~Deviance
{S(temp)}	136.690	0.00	0.87171	2.0000	132.64	0.63
{S(.)}	141.486	4.80	0.07924	1.0000	139.47	7.46
{S(t)}	148.532	11.84	0.00234	8.0000	132.01	0.00

Aha – so, it looks bringing in covariates through logistic regression can be useful.

Now what you need is some software for analyzing data. Luckily, Program MARK does a nice job at this. One might also use other packages such as R, SAS, etc. We'll use MARK because it is also very useful for so many other related analyses that we'll do.


Program MARK Interface - Lab 05 - swam squirrels (C:\Documents and Settings\rotella\My Documents\teaching\pop analysis\2007\SWAMPSQUIRRELS.DBF)
File
Delete
Order
Output
Retrieve
PIM
Design
Run
Simulations
Tests
Adjustments
Window
Help
?
Results Browser: Live Recaptures (CJS)

Model	AICc	Delta AICc	AICc Weight	Model Likelihood	No. Par.	Deviance	-2Log(L)
{02. phi(y=ad), p(y NE ad)}	212.9850	0.0000	0.21978	1.0000	2	206.7280	206.7280
{01. phi(y=ad), p(y=ad)}	214.0844	1.0994	0.12684	0.5771	2	209.8571	209.8571
{07. phi(y=tail = mean, ad=.), p(y =ad)}	214.2196	1.2346	0.11855	0.5394	4	205.8970	205.8970
{07. phi(y=tail = mean, ad=.), p(y =ad)}	214.2196	1.2346	0.11855	0.5394	4	205.8970	205.8970
{08. phi(y=tail, ad=.), p(y NE ad)}	214.6758	1.6908	0.09437	0.4294	5	204.0494	204.0494
{03. phi(y NE ad), p(y = ad)}	214.9296	1.9446	0.08312	0.3782	3	208.6035	208.6035
{04. phi(y NE ad), p(y NE ad)}	215.1850	2.2000	0.07316	0.3329	4	206.7260	206.7260
{05. phi(y=birth, ad=.), p(y=ad)}	216.0682	3.0832	0.04704	0.2140	4	207.6093	207.6093
{09. phi(y=birth + tail, ad=.), p(y =ad)}	216.3065	3.3215	0.04176	0.1900	5	205.6800	205.6800
{06. phi(y=birth, ad=.), p(y NE ad)}	216.3214	3.3364	0.04145	0.1886	5	205.6950	205.6950
{10. phi(y=birth + tail, ad=.), p(y NE ad)}	216.6371	3.6521	0.03539	0.1610	6	203.8078	203.8078