Example of Generalized Linear Model (GzLM) using Splus.	Raw Fish	Abun	dance
Data violate assumptions for ANOVA (General Linear Model)	Transect Y	ear1	Year2
	1	222	171
Prepared for NPS Inventory and Monitoring Program	2	125	101
hy David Schweider March 2007	3	69	57
by David Schneider. March 2007.	4	92	161
	5	121	792
The presentation uses a structured, model-based approach to	6	97	121
statistical analysis of data in biology.	7	153	119
	8	609	360
1 Construct model	9	147	93
Passarch question Does number of fish depend on year at corel reafs	10	135	130
in Here View works?	11	32	73
in Hawai ian parks?	12	113	200
	13	51	48
Verbal model: Reef fish numbers depend on year, controlling for	14	62	92
systematic variation in abundance at 20 locations along a monitoring	15	78	105
transect at Hanalei on Kauai (data from Eric Brown).	16	26	35
	17	10	6
Define response variable: $N = Abundance (in columns 2 and 2)$	18	5	10
Define response variable. $N = Abundance (in columns 2 and 3)$	19	59	62
Define explanatory variable of interest:	20	87	48
Voor - Vr this is a categorical variable with two classes			

Year = Yr this is a categorical variable, with two classes. Define secondary explanatory variable for statistical control:

Transect = Tr This is also categorical, with 20 classes.

Write formal model $N = \mu + Normal \ error$ $\mu = \beta_o + \beta_{Tr} \cdot Tr + \beta_{Yr} \cdot Yr$

The notation differs from that for the general linear model. However, if we substitute the second expression into the first, we obtain the same model as for the general linear model. This notation will be needed when we move from normal errors to other error distributions.

This is the structural model for a two-way ANOVA with no interaction term.	Raw Fi	sh Abur	ndance
Because Vr only has two classes it is also the model for a paired	IN	Yearin	ansect
because 11 only has two classes, it is also the model for a parted	222	1	1
comparison design.	125	1	2
	69	1	3
	92	1	4
2. Execute model.			
Reorganize data in model format	5	1	18
Reorganize data in model format	59	1	19
	87	1	20
One column of data for N, for Tr, and for Yr	171	2	1
	101	2	2
This can be done in SPlus, but if the data is already in a	57	2	3
spreadsheet, it is easier to reorganize in the spreadsheet than in	161	2	4
SPlus.			
	10	2	18
	62	2	19
	48	2	20

Paste data into structured data file in the statistical package, then label each column. In most packages, categorical variables are declared in the analytic routine.

However, SPlus requires that variables be declared as categorical.

Here is the conversion routine.

🔠 S-PLUS - Fish			
<u>F</u> ile <u>E</u> dit <u>V</u> iew <u>I</u> nsert	F <u>o</u> rmat	Data Statistics Graph Option	ns <u>W</u> indow <u>H</u> elp
0 🚅 🖬 🎒 .	% 🖻 🕻	Select Data	🗅 🦉 🧧 🐜 🎎 🛍 Lin
	.00 ÷.00	Copy Move	
		Transpose Exchange	ands
1 2	3 _	Remove •	Copyright (c) 1988, 20 wight Insightful Corp.
Abundance Transect	Year -	Clear	se Developer Version 7.
222 1	1	Restructure	data will be in C:\PROG
125 2	1	Fill	
69 3	1	Recode	
92 4	1	Create Categories	
121 5	1	Random Numbers	
97 6	1	Distribution Functions	
153 7	1	Split	
609 8	1	Subset	
147 9	1	Merge Tabulate	
135 10	1	Expand Grid	
32 11	1	Random Sample	
113 12	1	Convert to Data Frame	
51 13	1	Change Data Type	
62 14	1		
Converts column type			li.

This converts from default data type to character data.



Code the model in statistical package according to the structural model μ

$N = \rho_o + \rho_{Tr} Tr + \rho_{vr} TR + \epsilon$	
Proc Genmod; SAS comman	ıds
Class Tr Yr;	
Model N = Tr Yr/	
link=identity	
dist=normal	
type1 type3;	
OUTPUT out=RESPRED p=pred stdresdev=stdresdev;	
PROC PLOT data=RESPRED; plot stdresdev*pred/vref=0;	

The structural model consists of all of the explanatory terms. The structural model is specified in a model statement, which has the same format as the model statement for the general linear model.

In addition to the structural model, we need to state the error structure and the function that links the response variable to the structural model. For the general linear model, the error structure is normal and the link is the identify link.

In SPlus we can implement this model with the ANOVA routine, the general linear model routine, or generalized linear model routine.

	ANUVA
Here is the dialogue box opened via the ANOVA routine in SPlus.	Model Options Results Plot Compare Data Data Set: Fish Image: Compare Weights: Image: Compare
The model is written by choosing the dependent and independent variables.	Subset Bows with: Save Model Object Omit Rows with Missing Values Save As:
The ANOVA routine automatically assumes normal error and identity link.	Dependent: Abundance Independent: Abundance Independent: Abundance Iransect Year
Note the Formula —> The ANOVA dialogue box generates this model statement. This is the same	Eormula: Abundance~Transect+Year
model statement that was used in SAS (above).	OK Cancel Apply k Current Help

The dialogue box generates the code for a routine (R language) that SPlus runs. Here is the routine, as it would be written in R:

aov(formula = Abundance ~ Transect + Year, data = Fish, na.action = na.exclude)

Note the model structure: Abundance ~ Transect + Year This codes the model:

 $N = \beta_o + \beta_{Tr} \cdot Tr + \beta_{yr} \cdot YR + \epsilon$

The aov() routine is applied to a data set called Fish data = Fish

If we type (or paste) the R routine into the SPlus command window, this executes the routine.



Returning to the ANOVA dialogue box, we define the output by clicking on Results box. The default is Type I SS, which will be explained below. Model Options Results Saved Results Save In: Image: Save In: Results Fitted Values Image: Strength of Squares Image: Saved Results Image: Saved Results	turning to the ANOVA						
output by clicking on Results box. The default is Type I SS, which will be explained below. Estimated C ANOVA Estimated K Model Options Results Saved Results Save In: Estimated K Model Options Results	logue box, we define the	Model Options Results Plot Compare					
Results box. Image: Short Output Save In: The default is Type I SS, which will be explained below. Image: Type I Sums of Squares Image: Fitted Values Image: Short Output Image: Short Output Image: Short Output Image: Short Output Image: Short Output Image: Short Output Image: Short Output Image: Short Output	put by clicking on	Printed Results					
The default is Type I SS, which will be explained below. Image: Type I Sums of Squares Image: Fitted Values Image: Type I Sums of Squares Image: Fitted Values Image: Fitted Values Image: Type I Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Sums of Squares Image: Fitted Values Image: Fitted Values Image: Fitted Values Image: Fitted Values Image: Fitted Values Image: Fitted Values Image: Fitted Values Image: Fitted Values Image: Fitted Values	sults box.	ut Save <u>I</u> n:	•				
Ine default is Type ISS, which will be explained below. Estimated C ANOVA Estimated K Model Options Results Plot Compare	a defeat to Trues I CC	ns of Squares 📃 <u>F</u> itted Values					
below. Estimated C ANOVA Compare	which will be explained	ums of Squares 🛛 🗖 <u>R</u> esiduals					
Estimated K Model Options Results Plot Compare	ow.						
Model Options Results Plot Compare		ANOVA					
		Model Options Resu	s Plot Compare				
Plots Options		- Plots	Jptions Z Include Smooth				
			r include s <u>m</u> ooln				
J Sgrt Abs Residuals vs Fit J Include Rugplot		J Sgrt Abs Residuals vs Fit	Include Rugplot				
Response vs Fit Number of Extreme Points To Identify:		Response vs Fit	lumber of Extreme Points To Identify:				
Residuals Normal QQ		I▼ Residuals Normal <u>Q</u> Q	3				
Residual-Fit <u>Spread</u>		🗖 Residual-Fit <u>S</u> pread	Partial Residual Plot Options				
Cook's Distance		– 🔽 <u>C</u> ook's Distance	Include Partial <u>F</u> it				
Neut un ane si fu un li clude Rugplot	ut une en eifu uchiek	n Partial Residuals	Include Rugplo <u>t</u>				
nlots we want to diagnose	ts we want to diagnose		Z Common <u>Y</u> -axis Scale				
the residuals.	residuals.						
Click on Plot box	ck on Plot box						
	1						
choose a plot to diagnose whether residuals are homogeneous	oose a plot to diagnose whi iduals are homogeneous						
Residuals vs Fit	Residuals vs Fit						
UK Cancel Apply K > current Help		UK Cancel Apply K > C	rrent Help				
Choose a plot to diagnose whether	bose a plot to diagnose wh						
Residuals Normal OO	Residuals Normal OC						

This will produce a quantile-quantile plot.

Once the model and output are defined, the OK button executes the routine in R.

3. Evaluate model

- a. Straight line assumption. No need to evaluate this, as we are not fitting any lines.
- b. Normality of error assumption.

The quantile-quantile plot shows the residuals are not normal.

They do not all fall on straight line.

There are two extremely negative residuals and two extremely positive residuals.



C. Homogeneity of variance.



Conclusion. Any p-values

we compute cannot be trusted because the assumptions have been grossly violated. At this point the next step is to revise the model.

However, before going back and choosing a better error structure, it will be useful to repeat the two way ANOVA within the framework of the GzLM (i.e. categorical explanatory variables, identity link, normal error).

2. Execute model (GzLM routine).

Ì	🖽 S-	PLUS	5 - Fish							
	<u>F</u> ile	<u>E</u> dit	<u>V</u> iew <u>I</u> n	nsert f	= <u>o</u> rmat <u>D</u> ata	<u>S</u> tatistics	<u>G</u> raph C	D <u>p</u> tions	<u>W</u> in	dow <u>H</u> elp
		3		3 %		🖌 🛛 Data Su	ummaries		►	📕 🦗 🎼 🖬 Linear
		=		= +.0	.00 0.	Compar	e Samples	C	ł	
			= =	≡ .00	+.0 - 0	Design	and Sample	e Size	5	
	ia∎ F	ish			_			and the second second		
			1		2 3	Regres:	sion	intridit	•	Linear Robust MM
			Abunda	nce T	ransect Yea	Mixed E	ffects		•	Robust LTS
	1			222	1	Genera	lized Least	Squares	•	Stepwise
	2			125	2	Surviva	1		•	Generalized Additive
	3			69	3	Tree	e Modele		•	Local (Loess)
	4			92	4	Compar	e Models			Nonlinear
	5			121	5	Cluster	Analysis		ł	Generalized Linear
	6			07	5	Multiva	late		-	Log-linear (Poisson)
	7	10000		150	7	Quality	Control Ch	narts	1	Logistic
	-			100	/	Resamp	ina.		1	Probit
	8	1000		609	8	Time Se	ries		•	
	9	10000		147	9					,
Seneraliz	ed Lir	iear i	Models				Ē			1
Model		Opti	ions	Re	sults	Plot	Pred	lict		
Data	anna Marta				Model				ть	a dialogua hoy in the
<u>D</u> ata Set:		Fish		-	Family:	gau	issian	-	Gz	I M routine is similar to
Weights:				-	Link:	ider	ntity	-	tha	t for any general linear
Subset <u>Ro</u>	ws with				⊻ariance Fu	unction: cor	istant	-	mo	del (including a two-way
🔽 Omit R	ows wit	h <u>M</u> iss	ing Value	es	- Save Mod	el Object			AN	IOVA) except that now
				<u>anna a</u> Taith	Save As:				we	are free to state the error
									sut	acture and the milk.
Dependent		Abe	ndanaa						As	before, we set up the
Debeugen			nuance	<u> </u>					mo	del by choosing the
Independe	nt:	Abu	.L> Indance	<u>^</u>					dep	bendent and independent
		Tran	nsect						var	iables.
		1 Ba		~					1£	
Formula:		(Abu	indance~	Transe	rt+Year	an Manana Ad	Second Co		ш V (nc	we use a gaussian
_onnula.		1 L	indunce	Tunser	ot i cui			199	lin	k, we expect to see the
Create Fo	ormula								san	ne results as with the
OF	Canon		Applu		current			Help	AN	IOVA command.

2. Execute model (GzLM routine).

	Generalized Linear Models	
The ANOVA table –> in the Results tab will produce the analysis of deviance (ANODEV) table.	Model Options Res Printed Results Short Output Image: Short Output Image: Short Output Image: Short Output	Item Plot Predict Saved Results Save In: Image: Constraint of the second
	OK Cancel Apply K >	current Help

Here is the R command generated by the dialogue box.

```
glm(formula = Abundance ~ Transect + Year, family = gaussian, data = Fish,
na.action = na.exclude, control = list(epsilon = 0.0001, maxit = 50,
trace = F))
```

Note the model structure has not change: Abundance ~ Transect + Year

This codes the model we wrote earlier:

 $N = \beta_o + \beta_{Tr} \cdot Tr + \beta_{yr} \cdot YR + \epsilon$

With the glm() routine we can specify the error structure (normal = Gaussian) and we can specify the link (Identity) between the response variable and the explanatory variables.

Pressing the OK button runs the glm() routine that we have generated in SPlus

2. Execute model (GzLM routine).

```
Here are the results of calling up the glm() routine that we generated using SPlus
```

```
*** Generalized Linear Model ***
Call: glm(formula = Abundance ~ Transect + Year, family = gaussian, data = Fish,
      na.action = na.exclude, control = list(epsilon = 0.0001, maxit = 50,
      trace = F))
(Dispersion Parameter for Gaussian family taken to be 13828.5 )
   Null Deviance: 884540.8 on 39 degrees of freedom
Residual Deviance: 262741.5 on 19 degrees of freedom
Number of Fisher Scoring Iterations: 1
Analysis of Deviance Table
Gaussian model
Response: Abundance
Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev
                      39884540.820268768.5
   NULT.
Transect 19 615772.3
  Year 1 6027.0
                          19 262741.5
```

Here is a comparison of the ANODEV table (above) to the ANOVA table produced by the ANOVA routine.

```
        Df Sum of Sq
        Mean Sq
        F Value
        Pr(F)

        Transect
        19
        615772.3
        32409.07
        2.343643
        0.0354791

        Year
        1
        6027.0
        6027.03
        0.435841
        0.5170610

        Residuals
        <u>19</u>
        <u>262741.5</u>
        13828.50
        13828.50

        Total
        39
        884540.8
        13828.50
        13828.50
```

The ANOVA table and the ANODEV present the same numbers in different ways.

In the ANOVA table, the total SS (884540.8) is partitioned into components; then the component due to year is compared to the error component by means of an F-ratio, from which we compute a p-value.

The ANODEV table starts with the same measure (but now called the NULL deviance), then computes the <u>improvement</u> in fit due to adding a term to the model.

The analysis above shows the improvement for each term, in the order in which they are stated in the model. This is called Type I (or sequential) Sum of Squares.

The ANODEV table rests on likelihood: How likely is the data, given the model? The larger the deviance, the less likely the data.

2. Execute model (GzLM routine). - - More about Analysis of Deviance.

Response: Abundance Terms added sequentially

sequential.	ly (tirs	t to last)
Deviance Re	esid. Df	Resid. Dev
	39	884540.8
615772.3	20	268768.5
6027.0	19	262741.5
	615772.3 6027.0	Sequentially (first Deviance Resid. Df 39 615772.3 20 6027.0 19

The total SS (null deviance) is 884540.8. This is the likelihood of the data, given that it fits a single value (the null model).

The deviance drops substantially (likelihood improves) from 884540.8 to 268768.5 if we include location along the transect in the model.

The deviance drops (likelihood improves) by a little bit more, from 268768.5 to 262741.5 = 6027.0 if we include year in the model.

The change in deviance can used to compute p-values and make statistical decisions if:

the residuals are homogeneous

the null model is reasonable (deviance close to expected value, given df).

The residuals in this case are the same as those produced by the ANOVA, they do not meet the assumptions, so a p-value based on the change in deviance of interest (6027.0) cannot be trusted.

After this explanatory excursion, we return to the analysis of the reef fish data by going back to Step 1, and setting up a revised model with a better error structure.

1. Construct model

Write formal model $N = \mu + Gamma \ error$ $\mu = \beta_o + \beta_{Tr} \cdot Tr + \beta_{Yr} \cdot Yr$

The gamma error structure is a good choice because the variance increases with the mean, which is what we know to be the case from looking at the residual versus fit plot (cones opening to the right mean the variance increases with the mean)

2. Execute model.

Note the error structure is specified as Gamma.

The link is still identity, which means we are looking for additive effects of Transect and Year on fish abundance.

If we are interested in multiplicative effects, we would specify a log link, which is available in the Link box.

Generalized	Linear Models				×
Model	Options	Results	Plot	Predict	
Data		Model	an a		
<u>D</u> ata Set:	Fish	▼ Family:		Gamma 💌	
<u>W</u> eights:		✓ Link:	[identity 💌	I
Subset <u>R</u> ows v	vith:	Varianc	e Function:	constant 📃 💌	I
🔽 Omit Rows	with <u>M</u> issing Values	Save	Model Object		
		<u>S</u> ave A	.s:		
_ Variables —					
D <u>e</u> pendent:	Abundance	•			
Independent:	<all> Abundance Transect Year</all>	~			
<u>F</u> ormula:	Abundance~Tra	insect+Year			
Create Formu	la				
OK Car	ncel Apply K	current		Help	



Moving to the Results, we specify the same output as before.



3. Evaluate model.

- A. Straight line assumption. No need to check, not line has been fit.
- C. Homogeneity of variance.



The residuals are homogeneous (close to a uniform band from left to right). The assumption of homogeneity is met.

> Note: When evaluating models with normal error structures, the raw residuals are plotted against the fitted values. When evaluating models with non-normal errors the residuals must be scaled. These are called either Pearson or deviance residuals. SPlus automatically uses the deviance residuals.

As a further diagnostic, we examine whether the null deviance (our starting point) is acceptable. A rule of thumb is that the null deviance should be no more than twice the value of the associated degrees of freedom. This analysis meets that condition. The SPlus estimate of the null deviance is 38.05 on 39 degrees of freedom.

Response: Abundance

Terms added sequentially (first to last) Df Deviance Resid. Df Resid. Dev NULL 39 38.05088

4. State population and whether sample is representative.

The data were taken at random points on a transect and hence the sample is from a finite, enumerable population of all points on the transect. The sample represents the population along the transect.

5. Decide on mode of inference. Is hypothesis testing appropriate?

Yes. We wish to know whether the difference in abundance between years is greater than expected by chance.

6. State H_A / H_o pair, tolerance for Type I error

Does abundance depend on year ?

Deviance $(\beta_{Y_r}) > 0$ Same as $H_A: \beta_{Y_{r1}} \neq \beta_{Y_{r2}}$ Deviance $(\beta_{Y_r}) = 0$ Same as $H_o: \beta_{Y_{r1}} = \beta_{Y_{r2}}$

Statistic - Non-Pearsonian chisquare (G-statistic) Tolerance for Type I error set at 5%.

7. Analysis of Deviance (instead of analysis of variance).

Response: Abundance

Terms	add	led	sequentia	ally (f	irst	t to last)
		Df	Deviance	Resid.	Df	Resid. Dev
NU	JLL				39	38.05088
Transe	ect	19	34.77155		20	3.27932
Ye	ear	1	0.04929		19	3.23003

The improvement in fit is G = 0.04929 on 1 degree of freedom. To be statistically significant, the improvement must by G = 3.84 on 1 degree of freedom, based on the chisquare distribution. Clearly, this improvement is nothing more than chance. [The exact p-value for G = 0.04949 is p = (1-0.18) = 0.82].

8. Assess robustness of p-value in ANOVA table, assuming normal homogeneous errors. - Not applicable.

9. Declare decision about terms of interest in model.

Accept H_o that means are equal between years. Reject H_A that means differ. $0.82 = p > \alpha = 0.05$.

10. Analysis of parameters of biological interest.

The parameter of interest was the difference in means between years. The difference in means appeared large ($\beta_{Yr1} - \beta_{Yr2} = 114.65 - 139.2 = -24.55$) but the power of the test was too low to detect a significant difference, even after controlling for variation among sites and using an appropriate error structure.

As a matter of interest we compare this decision (with an appropriate error structure) to that from a standard ANOVA (with a normal error structure, which is inappropriate). The decision is the same, but the p-value is too low (too sensitive) for the (inappropriate) F-test.

GzLM	G = 0.05,	p = 0.82	no significant year effect
GLM (ANOVA)	F = 0.44	p = 0.52	no significant year effect