Model Based Statistics in Biology.
Part IV. The General Linear Model. Multiple Explanatory Variables
Chapter 12.2 Multiple Regression. Several Explanatory Variables.

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning based on models combined with statistics.
ReCap Part II (Chapters 5,6,7)
Hypothesis testing uses the logic of the null hypothesis to declare a decision.
Estimation is concerned with the specific value of an unknown population parameter.
ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.
ReCap (Ch 12) Multiple Regression with Two Explanatory Variables.

Today: Multiple Regression with more than two explanatory variables.

[Add example with stepwise regression or other selection procedures.]

Wrap-up.

Introduction

Analysis of species number on the Galapagos Islands.

Data from Johnson, M.P. and P.H. Raven (1973) Species number and endemism: The Galapagos revisited. *Science* 179: 893-895.

Does the number of endemic plants species depend on factors other than island area?

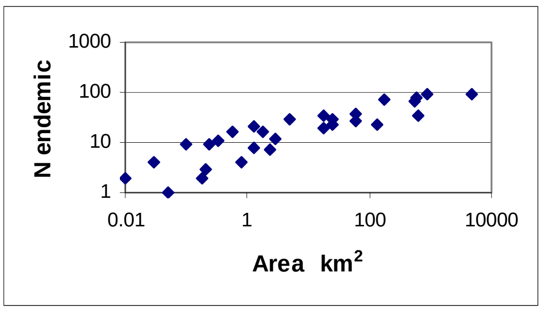
Plant species diversity on islands. Galapagos data from Johnson and Raven (1973)								
Island	All species	Endemic species	Area km^2	Elev m	Dist from nearest island km	Dist from Santa Cruz km	Area adjacent island km2	Non endemic species
Baltra	58	23	25.09	na	0.6	0.6	1.84	35
Bartolome	31	21	1.24	109	0.6	26.3	572.33	10
Caldwell	3	3	0.21	114	2.8	58.7	0.78	0
Champion	25	9	0.1	46	1.9	47.4	0.18	16
Coamano	2	1	0.05	na	1.9	1.9	903.82	1
Daphne Major	18	11	0.34	119	8	8	1.84	7
Daphne Minor	24	na	0.08	93	6	12	0.34	na
Darwin	10	7	2.33	168	34.1	290.2	2.85	3
Eden	8	4	0.03	na	0.4	0.4	17.95	4
Enderby	2	2	0.18	112	2.6	50.2	0.1	0
Espanola	97	26	58.27	198	1.1	88.3	0.57	71
Fernandina	93	35	634.49	1494	4.3	95.3	4669.32	58
Gardner	58	17	0.57	49	1.1	93.1	58.27	41
Gardner	5	4	0.78	227	4.6	62.2	0.21	1
Genovesa	40	19	17.35	76	47.4	92.2	129.49	21
Isabela	347	89	4669.32	1707	0.7	28.1	634.49	258
Marchena	51	23	129.49	343	29.1	85.9	59.56	28
Onslow	2	2	0.01	25	3.3	45.9	0.1	0
Pinta	104	37	59.56	777	29.1	119.6	129.49	67
Pinzon	108	33	17.95	458	10.7	10.7	0.03	75
Las Plazas	12	9	0.23	na	0.5	0.6	25.09	3
Rabida	70	30	4.89	367	4.4	24.4	572.33	40
San Cristobal	280	65	551.62	716	45.2	66.6	0.57	215
San Salvador	237	81	572.33	906	0.2	19.8	4.89	156
Santa Cruz	444	95	903.82	864	0.6	0	0.52	349
Santa Fe	62	28	24.08	259	16.5	16.5	0.52	34
Santa Maria	285	73	170.92	640	2.6	49.2	0.1	212
Seymour	44	16	1.84	na	0.6	9.6	25.09	28
Tortuga	16	8	1.24	186	6.8	50.9	17.95	8
Wolf	21	12	2.85	253	34.1	254.7	2.33	9

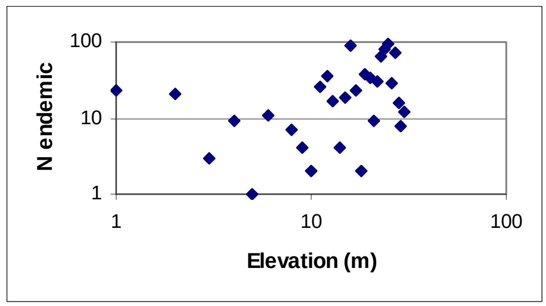
1. Construct model

Verbal model. The number of endemic species depends on factors other than island area, such as elevation and geographical factors likely to affect dispersal, including distance to nearest island, area of nearest island, distance from largest island, and distance from Santa Cruz, the island with the most species.

1. Construct model - Verbal model.
Response variable is number of endemic species. *N*
Explanatory variable is island area. *A* = km^2
Explanatory variable is maximum elevation. *Elev* = m
Explanatory variable is distance from nearest island. *Dni* = km
Explanatory variable is distance from Santa Cruz Island. *DSC* = km
Explanatory variable is area of nearest island. *Ani* = km.
All variables are on a ratio type of scale.

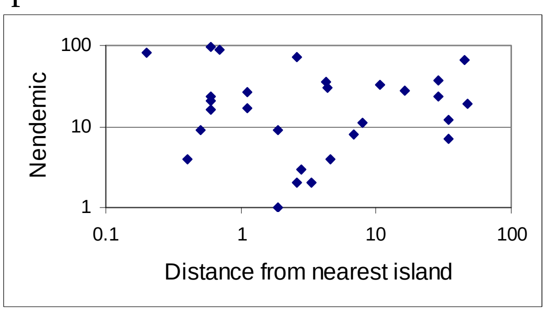
Graphical model.
Plot of response variable against each explanatory variable, keeping in mind that relation of response to any particular explanatory variable may change if the effects of another explanatory variable are removed by regression analysis.

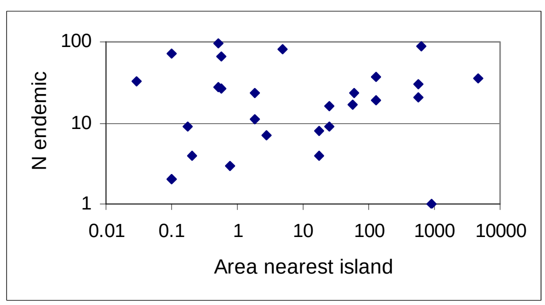


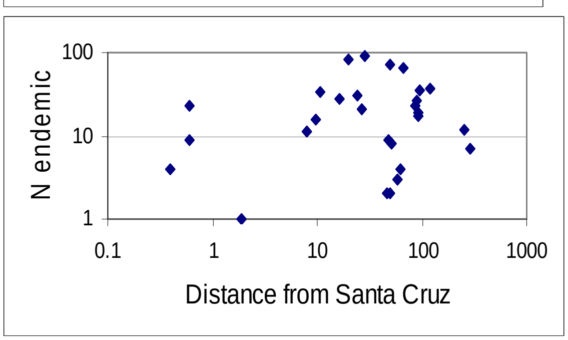


The graphs show a clear relation of species number to island area. There is some indication of a relation as well to island elevation, although this may be an effect of island area, as elevation and area are associated.

Endemic species number appears to be poorly related to area of nearest island, distance from nearest island, or distance from the central island (Santa Cruz), which has the most species.







Are these impressions borne out by partial regression analysis ? Such an analysis examines the relation of the response to each explanatory variable, taking into account the relation of response variable to other explanatory variables.

1. Construct the model

First, a model with just area, a relation substantiated by many previous studies of species number in relation to island area.

The power law is: $N = c\, A^{\beta A}$

Hence: $\ln(N) = \ln(c) + \beta_A \cdot \ln(A)$

The statistical model is: $\ln(N) = \beta_o + \beta_A \cdot \ln(A) + \text{residual}$ $\beta_o = \ln(c)$

Hence: $\ln N = \mu + \text{residual}$ normal residual

$\mu = \beta_o + \beta_A \cdot \ln(A)$

The parameter β_A is the exponent of the power law relation of species number to area. It is a simple regression coefficient.

Next, a model with all five explanatory variables.

$\ln N = \mu + \text{residual}$

$\mu = \beta_o + \beta_A \cdot \ln(A) + \beta_{Elev} \cdot \ln(Elev) + \beta_{Dni} \cdot \ln(Dni) + \beta_{Ani} \cdot \ln(Ani) + \beta_{DSC} \cdot \ln(DSC)$

In this model the parameter β_A stands for rate of change species number with area, controlled for the other four explanatory variables. β_A is the partial regression coefficient $\beta_{A:(Elev,Dni,Ani,DSC)}$, symbol that is read as 'the partial regression of species number on area, given elevation, distance to the nearest island, area of the nearest island, and distance from Santa Cruz.

2. Execute analysis.

Place data in model format. Create and label a column for:

- -response variable.
- -each explanatory variable.
- -logarithm of response variable.
- -logarithm of each explanatory variable.

Code the model statement in statistical package according to the GLM

The Minitab code is:

```
MTB > glm 'lnN' =  'lnA'  'lnElev'  'lnDnr'  'lnAnr'  'lnDSC' ;
SUBC> covariate   'lnA'  'lnElev'  'lnDnr'  'lnAnr'  'lnDSC' ;
SUBC> fits c4;
SUBC> residuals c5.
```

The SAS code is:

```
Proc glm;
   Model  'lnN' =  'lnA'  'lnElev'  'lnDnr'  'lnAnr'  'lnDSC' ;
```

The R code is:

```
DiversityModel <- lm(lnN ~ lnA + lnElev + lnDnr + lnAnr + lnDSC')
```

2. Execute analysis.

The parameter estimates for the species-area curve (regression on *A* only) are

$\ln(N) = 2.195 + 0.312 \ln A$ the exponent is close to typical value of 0.3

The intercept on the log scale (2.195) is calculated for a line fit though the mean on a log scale. mean(ln(*N*)) = $\hat{\beta}_o$ = 2.72 (n = 29)

Back transformation to an arithmetic scale yields the geometric mean $e^{2.72}$ = 15.23

Compare this to the arithmetic mean number of endemic species $\overline{N}$ = 27

The parameter estimates for the multiple regression equation are based on all five explanatory variables.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.632766063	1.33308164	2.73	0.0144
lnArea	0.306859555	0.07238696	4.24	0.0006
lnElev	-0.077426139	0.23398989	-0.33	0.7448
lnDnear	-0.011885158	0.08207882	-0.14	0.8866
lnDSCruz	-0.263359721	0.14262671	-1.85	0.0823
lnAnear	0.025496286	0.03732542	0.68	0.5038

SAS estimates (above) differ somewhat from Minitab estimates (below)

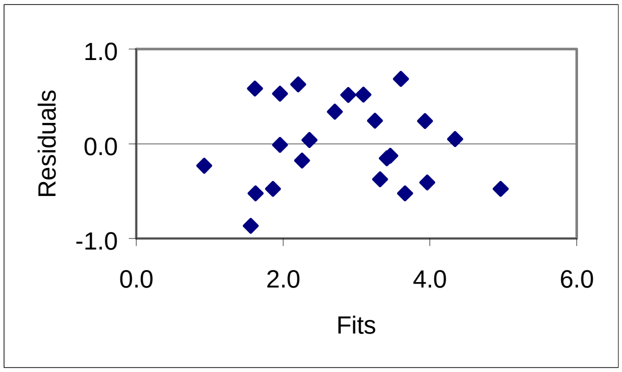
Term	Coef	SE Coef	T	P
Constant	4.593	1.234	3.72	0.002
lnA	0.36391	0.07319	4.97	0.000
lnElev	-0.2618	0.2158	-1.21	0.242
lnDni	-0.01094	0.07798	-0.14	0.890
lnDSC	-0.2805	0.1365	-2.05	0.056
lnAni	0.02752	0.03508	0.78	0.444

These are the estimates of the partial regression coefficients. Because the explanatory variables are themselves correlated, the partial regression estimates will not be the same as the estimates of the simple regression coefficients. These coefficients are used to compute the fitted values, which in turn are used to compute the residuals.

Plot residuals versus fitted values.

3. Evaluate model.

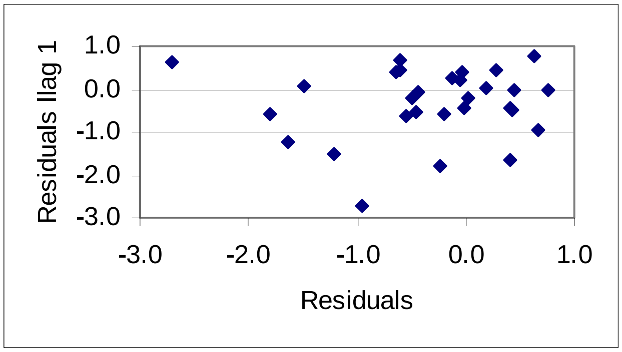
a. Some indication of arch, and so straight line assumption may not be correct for at least one of the explanatory variables. The plot of species number vs elevation suggests a non-linear relation to elevation.



b. Residuals homogeneous ?
Yes, from residual vs fit plot.

3. Evaluate model.

c. Other distributional assumptions



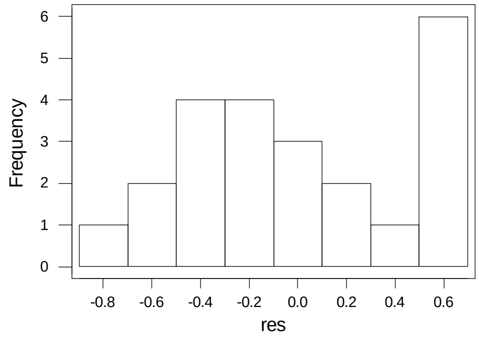
Independent ? Yes.

The plot of residuals were ordered by distance to nearest island to evaluate independence. The spatially ordered residuals versus themselves (at lag 1) shows no positive or negative trends.

Normal residuals ?

No. Histogram shows strong skew.

Confidence limits and p-values based on t-distribution may be incorrect.



4. State population and whether sample is representative.

The population is number of endemic species on all islands above 0.01 km^2 in the Galapagos archipelago.

The population is all values of species number per island, assuming a fixed error on a logarithmic scale. The population represented by the model is considered applicable to similar archipelagos: relatively young in geological age and lacking in wet habitat at higher elevations.

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

The goal of the study was to decide whether species number depends on factors other than island area. Thus we are interested in hypothesis testing with respect to each of the explanatory variables, other than area.

6. State H_A / H_o with tolerance for Type I error

Here are the hypothesis pairs listed in the order in which they appear in the model:

The first term concerns the effect of area, controlled for the other four explanatory variables. H_A: $\beta_A \neq 0$ Equivalently H_A: $var(\beta_A \cdot \ln(A)) > 0$

H_o: $\beta_A = 0$ H_o: $var(\beta_A \cdot \ln(A)) = 0$

The remaining H_A/H_o pairs are

H_A: $\beta_{Elev} \neq 0$	H_A: $\beta_{Dni} \neq 0$	H_A: $\beta_{Ani} \neq 0$	H_A: $\beta_{DSC} \neq 0$
H_o: $\beta_{Elev} = 0$	H_o: $\beta_{Dni} = 0$	H_o: $\beta_{Ani} = 0$	H_o: $\beta_{DSC} = 0$

Test statistic will be the F-ratio. Distribution will be F-distribution.

Tolerance for Type I error. $\alpha = 5\%$

7. ANOVA - Calculate df and variance, partition according to model.

Compute total df, partition according to model.

```
GLM model at top of board, on left
ANOVA table at top, on right.
```

$$\ln N = \beta_o + \beta_A \cdot \ln(A) + \beta_{Elev} \cdot \ln(Elev) + \beta_{Dni} \cdot \ln(Dni) + \beta_{Ani} \cdot \ln(Ani) + \beta_{DSC} \cdot \ln(DSC) + \text{error}$$

Source	Total	ln*A*	ln*Elev*	ln*Dni*	ln*Ani*	ln*DSC*	res
df	23 – 1 =	+ 1	+ 1	+ 1	+ 1	+1	+17

Source	df	Seq SS	Adj SS	MS	F	----> p
lnA	1	21.7915	6.4626	6.4626	24.73	0.000116
lnElev	1	0.2211	0.3848	0.3848	1.47	0.242
lnDnr	1	0.3031	0.0051	0.0051	0.02	0.890
lnDSC	1	0.9777	1.1033	1.1033	4.22	0.056
lnAnr	1	0.1608	0.1608	0.1608	0.62	0.444
Error	17	4.4434	4.4434	0.2614		
Total	22	27.8977				

The explanatory variables are correlated and so the adjusted SS will differ from the sequential SS. The adjusted SS is for each explanatory variable when it is entered <u>last</u> into the GLM. The Minitab estimates (above) differ somewhat from the SAS estimates (below)

Source	df	Adj SS	MS	F	----> p
ln*A*	1	5.0705	5.0705	17.97	0.0006
ln*Elev*	1	0.03089	0.03089	0.11	0.7448
ln*Dni*	1	0.005916	0.005916	0.02	0.8866
ln*DSC*	1	0.9620	0.9620	3.41	0.0823
ln*Ani*	1	0.1317	0.1317	0.47	0.5038
residual	17	4.7967			
Total	22				

The sequential SS add up to SS_{tot} = 27.8977 in both analyses.

The adjusted SS will not sum to the sum of the deviations from the grand mean (SS_{tot})

8. Recompute p-value if necessary.

The violation of the assumption of normal residuals was judged to be substantial so p-values were recomputed by randomization. To do this, the response variable was randomized, the regression was run, and the coefficients for each term were collected. The proportion of randomized coefficients that exceeded the observed estimate was the randomized p-value. The results for 5000 randomizations were as follows.

Source	df	F ---> p	n/5000 = p	
ln*A*	1	0.000116	102/5000	0.0204
ln*Elev*	1	0.242	3021/5000	0.604
ln*Dni*	1	0.890	4753/5000	0.951
ln*DSC*	1	0.056	1798/5000	0.360
ln*Ani*	1	0.444	3622/5000	0.724

Note that the p-values changed substantially in two cases.
The p-value for area changed by a factor of 0.0204/0.000116 = 176
The p-value for distance from Santa Cruz changed by a factor of 0.36/0.056 = 6
Despite the substantial change, none of the decisions changed.

These unusually large changes in p-value arise from several sources.
Large outliers were present -- these have a strongly distorting effect.
There were multiple explanatory variables that were correlated.

9. Declare decision about model terms, with evidence

Reject H_o: $\beta_A = 0$ $\quad 0.02 = p_{rand} < \alpha = 0.05$

Cannot reject H_o: $\beta_{Elev} = 0$ $\quad$ H_o: $\beta_{Dni} = 0$ $\quad$ H_o: $\beta_{Ani} = H_o$ $\quad$ H_o: $\beta_{DSC} = 0$

$p_{rand} = 0.60$ $\quad$ $p_{rand} = 0.95$ $\quad$ $p_{rand} = 0.72$ $\quad$ $p_{rand} = 0.36$

Conclusion: On the Galapagos islands, number of endemic species increases along with island area. Number does not depend on proximity to other islands, area of adjacent island, or distance from centre of the archipelago. It may, however, depend on island elevation in non-linear fashion.

10. Analysis of parameters of biological interest.

The parameter estimates are of no interest for those variables where the p-values were far from significant. The parameter estimate is of interest for island area, which was statistically significant. The model estimates (all 29 islands) is

$$N = e^{2.195} A^{0.312}$$

Johnson and Raven (1973) concluded that number of endemic plant species depended only on island area (according to a power law). They concluded, contrary to an earlier study based on a less complete lists of plant species, that other geographic factors (elevation, distance to nearest island, distance from centre of archipelago, area of nearest island) have no effect on plant species number. They provide a biological explanation for the lack of effect of elevation. They note that the Galapagos are a relatively young archipelago, with few endemic species inhabiting cooler and moister habitats at upper elevations, in contrast to other archipelagos such as Hawai'i.