Model Based Statistics in Biology.

Part IV. The General Linear Model. Multiple Explanatory Variables

Chapter 12.2 Multiple Regression. Three Explanatory Variables.

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11)
12 Multiple Regression. Introduction
12.1 Two Explanatory Variables
12.2 Three Explanatory Variables
13 GLM multiway ANOVA
14 GLM ANCOVA
15 Review - GLM with multiple explanatory variables.

on chalk board

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.

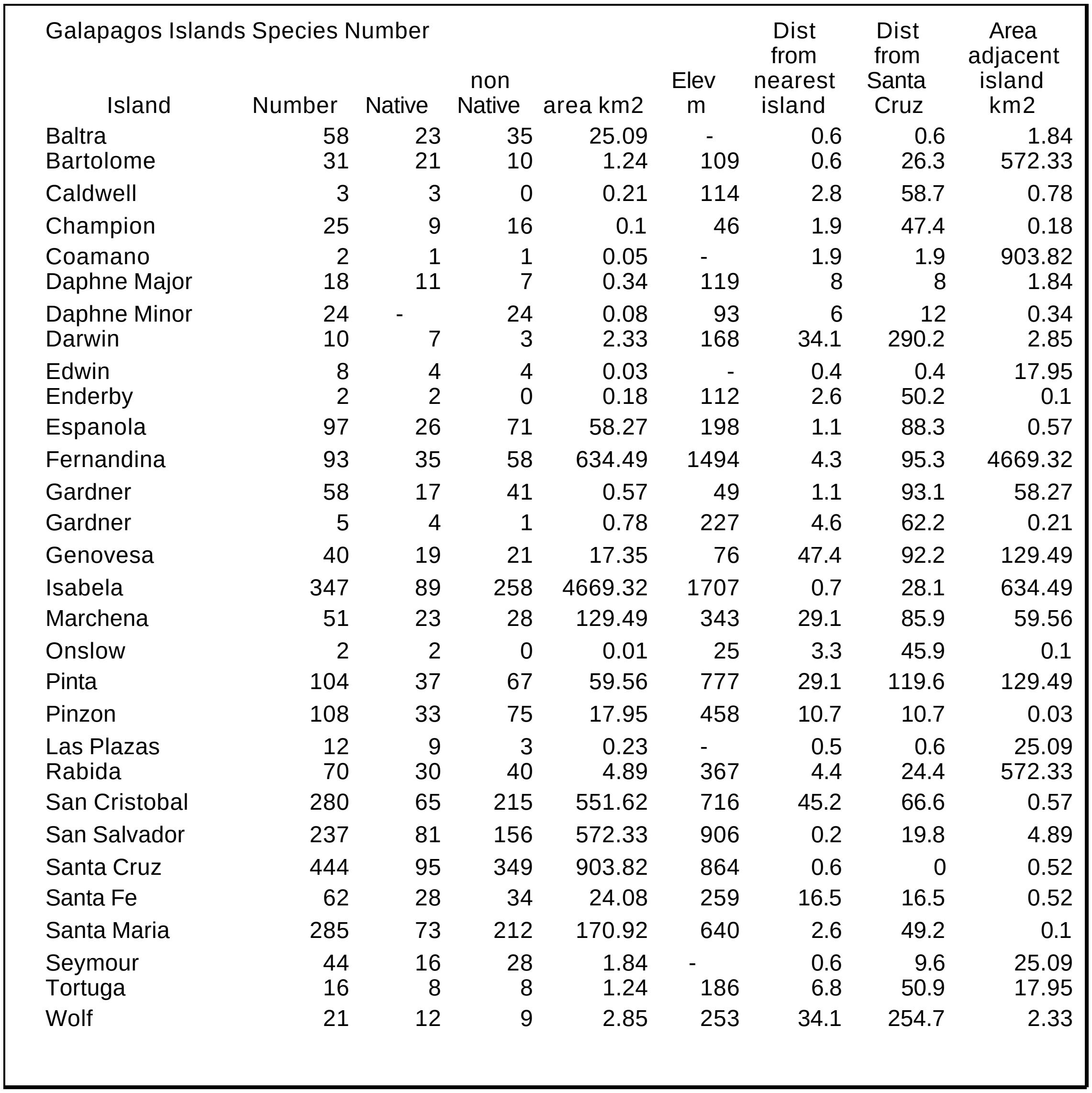
Today: Multiple Regression with more than two explanatory variables.

[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?



Galapagos Islands Species Number

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

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.

Response variable is number of endemic species. N

Explanatory variable is island area. A = km^2

Explanatory variable is maximum elevation. H = m

Explanatory variable is distance from nearest island. Dnr = km

Explanatory variable is distance from Santa Cruz Island. DSC = km

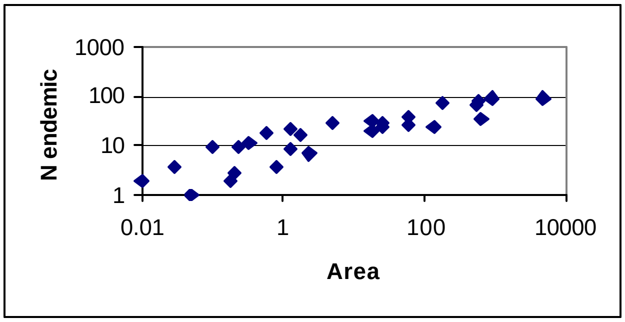
Explanatory variable is area of nearest island. Anr = km.

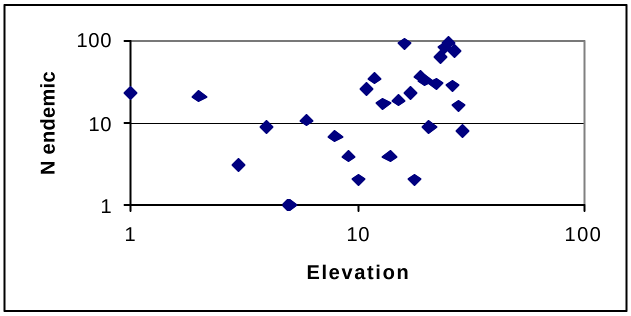
All variables are on a ratio type of scale.

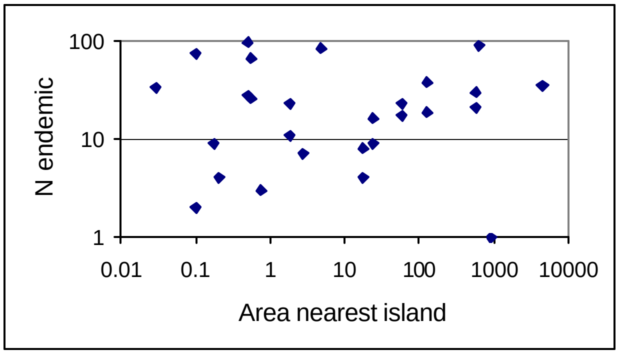
1. Construct model

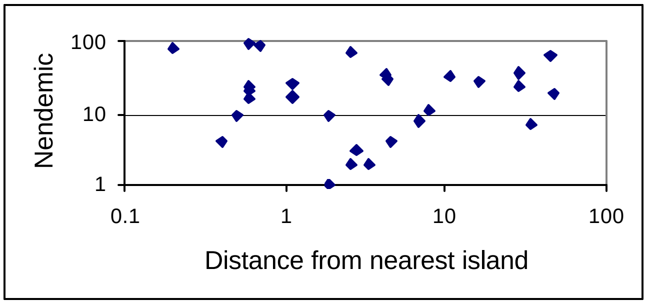
Graphical model.

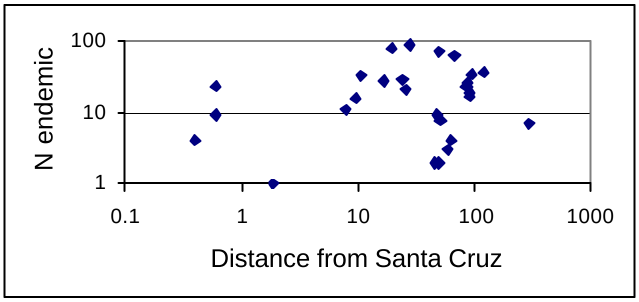
Plot of response variable against each explanatory variable, keeping in mind that relation of response to any particular explanatory variable may change if the effects of another explanatory variable are removed by regression analysis.











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. There is some suggestion of a relation to area of nearest island. Species number appears to be completely unrelated to 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: $N = c\ A^{\$A}$

Hence: $\ln(N) = \ln(c) + \$_A @ \ln(A)$

The statistical model is: $\ln(N) = \$_o + \$_A @ \ln(A) + \text{residual}$ $\quad \$_o = \ln(c)$

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

$: = \$_o + \$_A @ \ln(A)$

The parameter $\$_A$ 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.

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

$$: = \$_o + \$_A @ \ln(A) + \$_{Elev} @ \ln(Elev) + \$_{Dnr} @ \ln(Dnr) + \$_{Anr} @ \ln(Anr) + \$_{DSC} @ \ln(DSC)$$

In this model the parameter $\$_A$ stands for rate of change species number with area, controlled for the other four explanatory variables. $\$_A$ represents the partial regression coefficient $\$_{A:(Elev,Dnr,Anr,DSC)}$, a 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' ;
```

Fits and residuals calculated from:

model statement output of fitted values and residuals (as in Minitab code).
parameters reported by GLM routine

2. Execute analysis.

The overall mean is

mean(ln(N)) = $\hat{\$}_o$ = 2.72 (n = 29)

Logarithms have no units, so this average has no units.

The regression equation for A only is

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

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

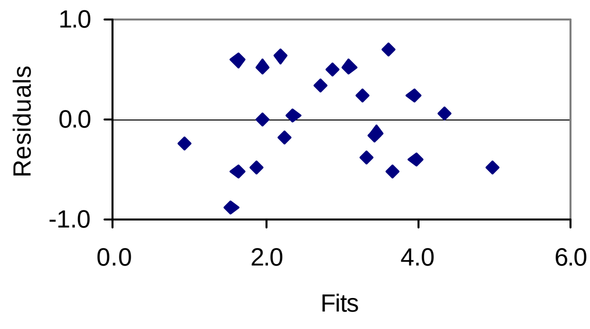
```
                                      Standard
Parameter          Estimate              Error    t Value    Pr > |t|

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
lnAnr           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
lnDnr      -0.01094                    0.07798                 -0.14       0.890
lnDSC      -0.2805                     0.1365                  -2.05       0.056
lnAnr       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.



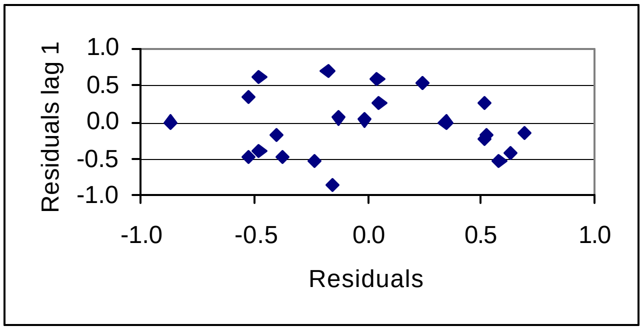
3. Evaluate model.

a. No bowls or arches are evident in plot of residuals against fitted values, so straight line assumption valid for regression of log transformed variables.

b. Residuals homogeneous ?
Yes.

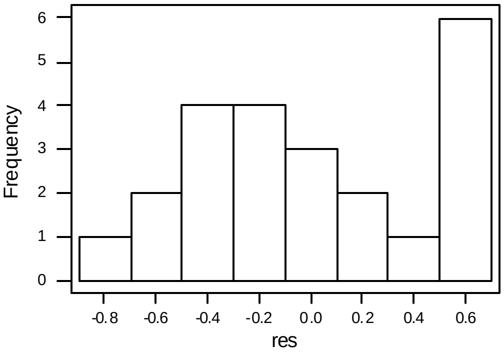
3. Evaluate model.

c. Other distributional assumptions



Independent ? Yes.

The plot of 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 not enumerable.
The population is all values of species number per island, if the observational study were run repeatedly. The population is represented by the model

$$\ln N = \$_o + \$_A @ \ln(A) + \$_{Elev} @ \ln(Elev) + \$_{Dnr} @ \ln(Dnr)$$
$$+ \$_{DSC} @ \ln(DSC) + \$_{Anr} @ \ln(Anr)$$

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: $\$_A \ldots 0$ Equivalently H_A: $\mathrm{var}(\$_A @ A) > 0$
H_o: $\$_A = 0$ H_o: $\mathrm{var}(\$_A @ A) = 0$

The remaining H_A/H_o pairs are

H_A: $\$_{Elev} \ldots 0$ H_A: $\$_{Dnr} \ldots 0$ H_A: $\$_{DSC} \ldots 0$ H_A: $\$_{Anr} \ldots 0$
H_o: $\$_{Elev} = 0$ H_o: $\$_{Dnr} = 0$ H_o: $\$_{DSC} = 0$ H_o: $\$_{Anr} = 0$

Test statistic will be F-ratio. Distribution will be F-distribution.
Tolerance for Type I error. " = 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.
```

GLM: $\ln N = \$_o + \$_A @ \ln A + \$_{Elev} @ \ln Elev + \$_{Dnr} @ \ln Dnr + \$_{Anr} @ \ln Anr + \$_{DSC} @ \ln DSC + \text{res}$

Source	Total	ln*A*	ln*Elev*	ln*Dnr*	ln*Anr*	ln*DSC*	res
df	23 ! 1 =	1	+ 1	+ 1	+ 1	+ 1	+17

Source	df	SS	MS	F	---->	p
ln*A*	1					
ln*Elev*	1					
ln*Dnr*	1					
ln*Anr*	1					
ln*DSC*	1					
residual	?					? = 22 ! 1 ! 1 ! 1 ! 1
Total	23 ! 1					

Fill in df.

Obtain Type III (adjusted) SS for each term in model.

That is, we use the SS for each explanatory variable when it is entered <u>last</u> into the GLM.

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

```
Source          df          Adj SS      MS          F      ---->  p
lnA              1          5.0705      5.0705      17.97   0.0006
lnElev           1          0.03089     0.03089     0.11    0.7448
lnDnr            1          0.005916    0.005916    0.02    0.8866
lnDSC            1          0.9620      0.9620      3.41    0.0823
lnAnr            1          0.1317      0.1317      0.47    0.5038
residual        17          4.7967
   Total        22
```

The sequential SS added up to SS_{tot} = 27.8977 in both analyses.
The adjusted SS cannot be summed.

8. Recompute p-value if necessary.

The effect of the violation of assumption of normal residuals was judged to be substantial so p-values recomputed by randomization. The response variable was randomized, the regression was run, 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*Dnr*	1	0.890	4753/5000	0.951
ln*DSC*	1	0.056	1798/5000	0.360
ln*Anr*	1	0.444	3622/5000	0.724

None of the decisions changed.
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

Several factors contributed to these unusually large changes in p-value.
Large outliers were present and these have a strongly distorting effect.
There were multiple explanatory variables, that were highly correlated.

9. Declare decision about model terms, with evidence

Reject H_o: $\$_A = 0$ 0.02 = p < " = 0.05
Accept H_A: $\$_A > 0$

Accept H_o: $\$_{Elev} = 0$ H_o: $\$_{Dnr} = 0$ H_o: $\$_{Anr} = 0$ H_o: $\$_{DSC} = 0$
Reject H_A: $\$_{Elev}$ ... 0 H_A: $\$_{Dnr}$ ... 0 H_A: $\$_{Anr}$ ... 0 H_A: $\$_{DSC}$ ... 0

Conclusion: Number of endemic species on the Galapagos islands depends on island area. Number does not depend on island elevation, proximity to other islands, area of adjacent island, or distance from centre of the archipelago.

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 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 plants, 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.