

Part I. (50 points) Do all calculations in SAS. Use a word processor of your choice to write a report. Insert computer text output and graphics to support what you are saying, but you need to write something that looks like an academic paper — not a pile of computer output. Turn in a hard copy of your HW in class (i.e., don't email me your HW). Also:

1. Clearly specify parameters and hypotheses when appropriate.
2. Write a coherent conclusion.

(50^{pts}) **1. Amazon forest clearing and butterflies:**

As an introduction to their study on the effect of Amazon forest clearing¹, the researchers stated “Fragmentation of once continuous wild areas is a major way in which people are altering the landscape and biology of the planet.” Their study takes advantage of a Brazilian requirement that 50% of the land in any development project remain in forest and tree cover. As a consequence of this requirement, “islands” of forest of various sizes remain in otherwise cleared areas. The data below are the number of butterfly species found in 16 such islands. The area of the island is in hectares.

Summarize the role of area in the distribution of number of butterfly species. Write a brief report including a summary of statistical findings, graphical displays, and a section dealing with the methods used to answer the question of interest.

Reserve	Area (hectares)	Species (types)
1	1	14
2	1	50
3	1	55
4	1	34
5	1	40
6	1	57
7	10	43
8	10	103
9	10	33
10	10	53
11	10	50
12	100	110
13	100	70
14	100	119
15	100	60
16	1000	145

Create a SAS program, and provide the program as part of your writeup, to read in the data, and carry out the subsequent parts of the HW. It would be preferable for you create a text file with the data in it and use an INFILE statement to read the data in from that file. Include relevant comments as part of the SAS program; see the class notes for how to do this.

Remark: This problem is somewhat open-ended, but is easily formulated as a regression problem, where you are interested in whether the number of species changes with area, and if so, then how. You might consider the following steps in an analysis: (1) plot the data, (2) consider transforming the data to a scale where the relationship is roughly linear, and (3) if transformation is suggested, do the statistical analysis on the transformed scale, else do the analysis on the given data. Finally, (4) check model assumptions and (5) summarize your findings.

(a) (10 pts) Plot the data, consider transforming the data and plot on transformed scale.

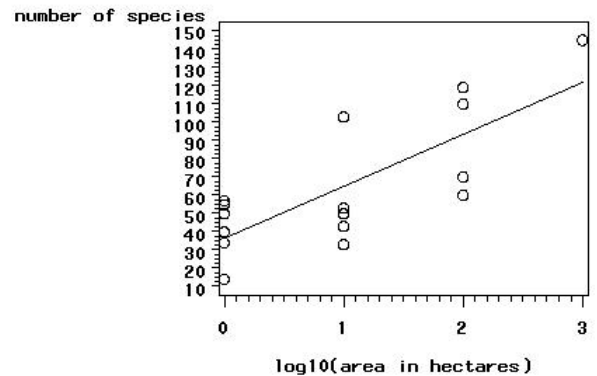
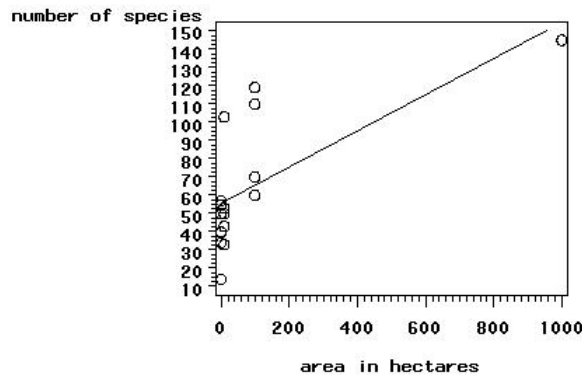
Solution: The plot with log(area) is roughly linear. This transformation will be used for analysis.

¹Lovejoy, T.E., Rankin, J.M., Bierregaard, R.O., Brown, K.S., Emmons, L.H., and Van der Woot, M.E. (1984). “Ecosystem Decay of Amazon Forest Remnants” in *Extinctions*; edited by M. H. Nitecki. Chicago: University of Chicago Press. Example from *The Statistical Sleuth* by Ramsey and Schafer.

50 pts

50 pts

Amazon butterfly species by area Amazon butterfly species by area



Program editor contents:

```
* options for sas session;
options ls=79 nodate nocenter;

* part (a) *****;
* read data, create variables, assign labels;
data amazon;
  * read data file;
  infile 'F:\Dropbox\UNM\teach\ADA2_stat528\assess\ADA2_HW_02_amazonbutterflies.dat';

  * assign variables to columns in infile;
  input reserve area species;

  * create area in log scale for linear relationship with species;
  arealog10 = log10(area);

  label reserve = 'reserve site number'
         area = 'area in hectares'
         arealog10 = 'log10(area in hectares)'
         species = 'number of species';

run;

* scatterplot;
symbol1 v=circle c=black i=r; * i=r makes a regression line;
title 'Amazon butterfly species by area';
proc gplot data=amazon;
  plot species*(area arealog10);
run;
```

(b) (20 pts) Do a statistical analysis on the scale (transformed or original) where the relationship is close to linear.

Solution: The linear regression slope is significant (t -stat=4.48 with p -value of 0.0005) and the regression explains about $R^2 = 59\%$ of the variability in number of species.

The slope is estimated at $\hat{\beta}_1 = 28.5$, thus we expect the number of butterfly species to increase by 28.5 for each unit increase in $\log_{10}(\text{area})$.

Variable	DF	Parameter Estimates				
		Parameter Estimate	Standard Error	t Value	DF	Pr > t
Intercept	1	36.25000	8.70185	4.17	1	0.0010
arealog10	1	28.50000	6.35494	4.48	1	0.0005
Root MSE		23.77799	R-Square	0.5896		
Dependent Mean		64.75000	Adj R-Sq	0.5603		
Coeff Var		36.72277				

Refitting the model without the influential large area (1000) is still significant, but explains less variability ($R^2 = 43\%$).

The slope is estimated at $\hat{\beta}_1 = 23.4$. Thus, the conclusions do not greatly change (28.5 vs 23.4, both with standard errors of about 7) with or without this observation.

Parameter Estimates

0 pts

Variable	DF	Parameter Estimate	Standard Error	t Value	DF	Pr > t
Intercept	1	39.11644	8.85455	4.42	1	0.0007
arealog10	1	23.40411	7.48346	3.13	1	0.0080
Root MSE		23.34712	R-Square	0.4293		
Dependent Mean		59.40000	Adj R-Sq	0.3855		
Coeff Var		39.30491				

Program editor contents:

```

* part (b) *****;
* regression analysis and checking model assumptions;
title 'Reg of butterfly species by log10(area)';
proc reg data=amazon;
  model species=arealog10/p r;
  plot student.*predicted. nqq.*student. cookd.*obs.; * diagnostic plots;
run;

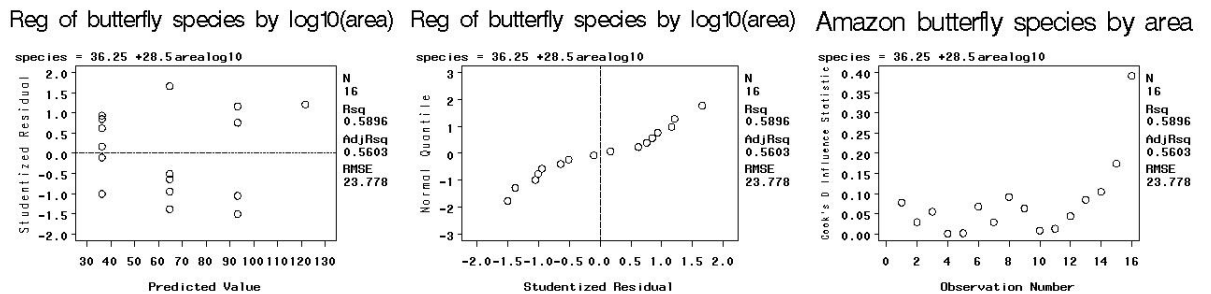
* check how much influence the single largest area observation has;
data amazon2;
  set amazon;
  if area ^= 1000; * delete single large area since large Cook's D;
run;

* scatterplot;
symbol1 v=circle c=black i=r; * i=r makes a regression line;
title 'Amazon butterfly species by area (del obs)';
proc gplot data=amazon2;
  plot species*(area arealog10);
run;

* part (b) *****;
* regression analysis and checking model assumptions;
title 'Del obs reg of butterfly species by log10(area)';
proc reg data=amazon2;
  model species=arealog10/p r;
  plot student.*predicted. nqq.*student. cookd.*obs.; * diagnostic plots;
run;
    
```

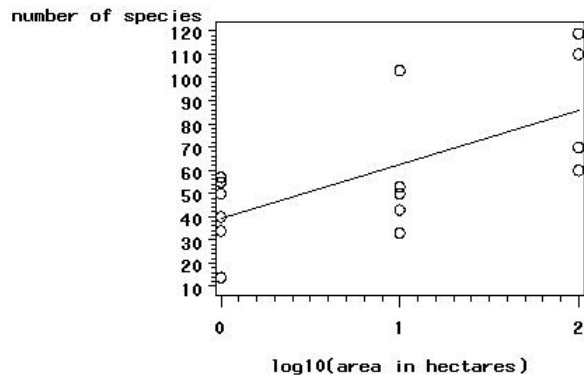
(c) (10 pts) Check model assumptions.

Solution: For the model with all data points, the model assumptions are reasonably met. There is no strong pattern for the residuals by predicted value (left), while the residuals seem short-tailed, they do not deviate strongly from normality (middle), though the Cook's distance shows the area=1000 observation as influential (right). This large Cook's D is why I refit the model without that observation.



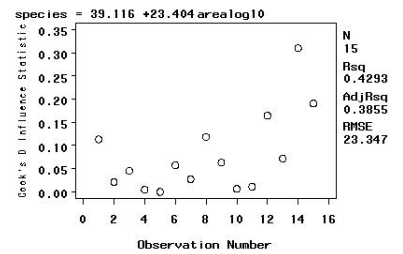
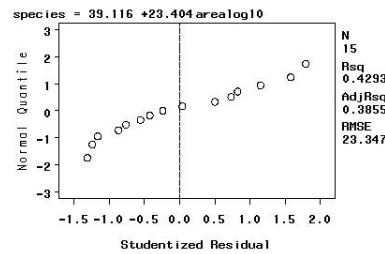
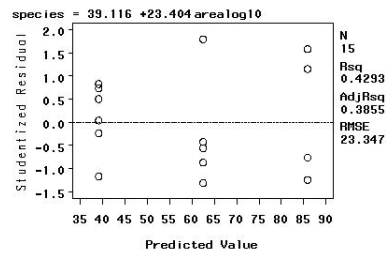
Without the influential observation, the data are still roughly linear.

Amazon butterfly species by area (del obs)



The model assumptions are similarly met as before.

Del obs reg of butterfly species by log10(area) Del obs reg of butterfly species by log10(area) Del obs reg of butterfly species by log10(area)



(d) (10 pts) Summarize your findings.

Solution: Using the full data model, there is a significant relationship between number of species and the logarithm of the size of their habitat. In particular, with each 10-fold increase in reserve area, we expect almost 30 more species to exist in that reserve with a 95% CI of give or take roughly 12 species.