Statistics 516- Spring 2002 - Exam 1 ***(modified for Spring 2003 practice)***

Part I: Answer the following three questions. First two are eight points each, the third is seven points.

1) Define what is meant by the p-value (or empirical significance level of a test).

2) In peforming a linear regression to predict y from x, what four assumptions must be satisfied?

3) $\sum_{i=1}^{n}(x_i-\bar{x})^2 = 22.0$ $\sum_{i=1}^{n}(y_i-\bar{y})^2 = 21752.0$ $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = 236.5$ $\bar{x} = 22.5$ $\bar{y} = 332.0$ $n = 7$

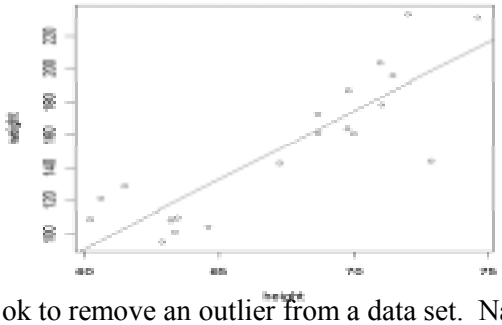
Complete the Table

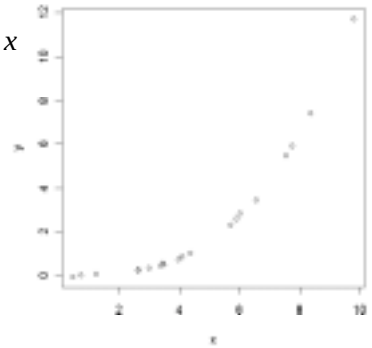
Source	SS	DF	MS	F	Prob>F
Regression	_____	___	_____	_____	0.4529
Error	_____	5	3841.9250		
Total	_____	___			

Part II: Answer eleven of the following twelve questions. If you answer more than eleven I will grade only the first eleven. Seven points each.

1) SAS produces a t-statistic of -1.87 and a p-value of 0.075 for testing the null hypothesis $\beta_1=0$ vs. the alternate hypothesis of $\beta_1\neq 0$. What would the p-value be for testing the alternate hypothesis $\beta_1<0$?

2) Consider the scatter plot of x vs. y shown at the right. Suggest a transformation of either *x* or *y* that may allow you to perform linear regression on the data.





3) The regression line shown to the right is: height = -407.771 + 8.320 weight. Even if all of the assumptions are met, why couldn't we use this regression equation to predict the weight of someone who is 55" tall?

4) In class we gave two situations where it was ok to remove an outlier from a data set. Name one of those two situations.

Questions 5-8 use the attached data set *Minnows* is from a study by Ryan, Hubert, Sprague, and Parrot. It concerned the effect of Zinc and Copper in the water on the amount of protein in minnow larvae. One unit of `Metals` equals a total Copper parts per million + 0.1 Zinc parts per million of one. One unit of `Protein` is a micro-gram. In each case the researchers controlled the level of the metals and randomly assigned the groups of minnow larvae to one of the contaminated tanks. Use the above description and the accompanying SAS output to answer the following questions.

5) For each 100 unit increase in the amount of metal, what is the predicted change in the amount of protein?

6) Is there a statistically significant relationship (α=0.05) between the amount of metal and the amount of protein? How did you check this?

7) Give a 95% confidence interval for the range of protein values a new set of minnow larvae would be expected to fall in, if the metal amount was 0.

8) Comment on the regression assumptions for this problem. Which are met or not met, and how could you tell?

Questions 9-12 use the attached data set *Galapagos* from an article in the journal *Science* by Johnson and Raven. It concerns the number of native species on the various Galapagos Islands based on the number of non-native species (`NonNative`), the area of the island in km^2 (`Area`) , the elevation of the island in m (`Elev`), the distance from the nearest other island in km (`DistNear`), and the distance from Santa Cruz in km (`DistSC`). Use the accompanying SAS output to answer the following questions. You may assume the assumptions for performing regression are met.

9) What percentage of the variation in the number of native species is explained by the five predictor variables?

10) Carefully state what null and alternate hypotheses go with the p-value of 0.0008 in the *Type I Tests* box. Do we accept or reject that null hypothesis?

11) Assuming the the regression with five variables fit, which simpler models are also acceptable? How could you tell this?

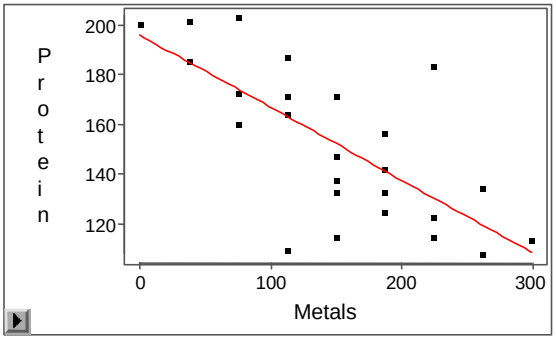
12) State which of the islands number of native species is predicted worst by the model and how you found it.

```
DATA Minnows;
INPUT Protein Metals;
CARDS;
201      0
186      37.5
173      75
110      112.5
115      150
202      37.5
161      75
172      112.5
138      150
133      187.5
204      75
165      112.5
148      150
143      187.5
123      225
188      112.5
172      150
157      187.5
115      225
108      262.5
133      150
125      187.5
184      225
135      262.5
114      300
;

PROC INSIGHT;
OPEN Minnows;
FIT Protein = Metals;
RUN;

PROC GLM DATA=Minnows;
MODEL Protein = Metals / ALPHA=0.05 CLI;
RUN;

PROC GLM DATA=Minnows;
MODEL Protein = Metals / ALPHA=0.05 CLM;
RUN;
```

Protein = Metals
Response Distribution: Normal
Link Function: Identity

Model Equation

Protein = 195.880 - 0.2912 Metals



Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Pr > F
	1	1	11924.6400	23	496.8417	0.5106	24.00	<.0001

Summary of Fit

Mean of Response	152.2000	R-Square	0.5106
Root MSE	22.2899	Adj R-Sq	0.4894

Analysis of Variance

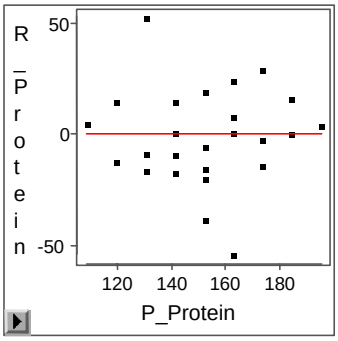
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	1	11924.6400	11924.6400	24.00	<.0001
Error	23	11427.3600	496.8417		
C Total	24	23352.0000			

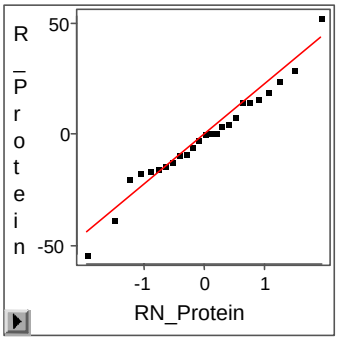
Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Metals	1	11924.6400	11924.6400	24.00	<.0001

Parameter Estimates

Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept	1	195.8800	9.9684	19.65	<.0001	.	0
Metals	1	-0.2912	0.0594	-4.90	<.0001	1.0000	1.0000





The GLM Procedure

Observation	Observed	Predicted	Residual	95% Confidence Limits for Individual Predicted Value	
1	201.00000000	195.88000000	5.12000000	145.36873153	246.39126847
2	186.00000000	184.96000000	1.04000000	135.94411215	233.97588785
3	173.00000000	174.04000000	-1.04000000	126.12080324	221.95919676
4	110.00000000	163.12000000	-53.12000000	115.87103480	210.36896520
5	115.00000000	152.20000000	-37.20000000	105.17656814	199.22343186
6	202.00000000	184.96000000	17.04000000	135.94411215	233.97588785
7	161.00000000	174.04000000	-13.04000000	126.12080324	221.95919676
8	172.00000000	163.12000000	8.88000000	115.87103480	210.36896520
9	138.00000000	152.20000000	-14.20000000	105.17656814	199.22343186
10	133.00000000	141.28000000	-8.28000000	94.03103480	188.52896520
11	204.00000000	174.04000000	29.96000000	126.12080324	221.95919676
12	165.00000000	163.12000000	1.88000000	115.87103480	210.36896520
13	148.00000000	152.20000000	-4.20000000	105.17656814	199.22343186
14	143.00000000	141.28000000	1.72000000	94.03103480	188.52896520
15	123.00000000	130.36000000	-7.36000000	82.44080324	178.27919676
16	188.00000000	163.12000000	24.88000000	115.87103480	210.36896520
17	172.00000000	152.20000000	19.80000000	105.17656814	199.22343186
18	157.00000000	141.28000000	15.72000000	94.03103480	188.52896520
19	115.00000000	130.36000000	-15.36000000	82.44080324	178.27919676
20	108.00000000	119.44000000	-11.44000000	70.42411215	168.45588785
21	133.00000000	152.20000000	-19.20000000	105.17656814	199.22343186
22	125.00000000	141.28000000	-16.28000000	94.03103480	188.52896520
23	184.00000000	130.36000000	53.64000000	82.44080324	178.27919676
24	135.00000000	119.44000000	15.56000000	70.42411215	168.45588785
25	114.00000000	108.52000000	5.48000000	58.00873153	159.03126847

Observation	Observed	Predicted	Residual	95% Confidence Limits for Mean Predicted Value	
1	201.00000000	195.88000000	5.12000000	175.25886100	216.50113900
2	186.00000000	184.96000000	1.04000000	168.33470623	201.58529377
3	173.00000000	174.04000000	-1.04000000	160.99804656	187.08195344
4	110.00000000	163.12000000	-53.12000000	152.80943050	173.43056950
5	115.00000000	152.20000000	-37.20000000	142.97794628	161.42205372
6	202.00000000	184.96000000	17.04000000	168.33470623	201.58529377
7	161.00000000	174.04000000	-13.04000000	160.99804656	187.08195344
8	172.00000000	163.12000000	8.88000000	152.80943050	173.43056950
9	138.00000000	152.20000000	-14.20000000	142.97794628	161.42205372
10	133.00000000	141.28000000	-8.28000000	130.96943050	151.59056950
11	204.00000000	174.04000000	29.96000000	160.99804656	187.08195344
12	165.00000000	163.12000000	1.88000000	152.80943050	173.43056950
13	148.00000000	152.20000000	-4.20000000	142.97794628	161.42205372
14	143.00000000	141.28000000	1.72000000	130.96943050	151.59056950
15	123.00000000	130.36000000	-7.36000000	117.31804656	143.40195344
16	188.00000000	163.12000000	24.88000000	152.80943050	173.43056950
17	172.00000000	152.20000000	19.80000000	142.97794628	161.42205372
18	157.00000000	141.28000000	15.72000000	130.96943050	151.59056950
19	115.00000000	130.36000000	-15.36000000	117.31804656	143.40195344
20	108.00000000	119.44000000	-11.44000000	102.81470623	136.06529377
21	133.00000000	152.20000000	-19.20000000	142.97794628	161.42205372
22	125.00000000	141.28000000	-16.28000000	130.96943050	151.59056950
23	184.00000000	130.36000000	53.64000000	117.31804656	143.40195344
24	135.00000000	119.44000000	15.56000000	102.81470623	136.06529377
25	114.00000000	108.52000000	5.48000000	87.89886100	129.14113900

```
DATA Galapagos;
INPUT Native  NonNative Are Elev  DistNear DistSC Island $;
CARDS;
23     35     25.09  332    0.60   0.60   Baltra
21     10     1.24   109    0.60   26.30  Bartolome
3      0      0.21   114    2.80   58.70  Caldwell
9      16     0.10   46     1.90   47.40  Champion
1      1      1.05   130    1.90   1.90   Coamano
11     7      0.34   119    8.00   8.00   Daphne_Major
12     12     0.08   93     6.00   12.00  Daphne_Minor
7      3      2.33   168    34.10  290.20 Darwin
4      4      0.03   46     0.40   0.40   Eden
2      0      0.18   112    2.60   50.20  Enderby
26     71     58.27  198    1.10   88.30  Wspanola
35     58     634.49 1494   4.30   95.30  Fernandina
17     41     0.57   49     1.10   93.10  Gardner_(Esp)
4      1      0.78   227    4.60   62.20  Gardner_(SM)
19     21     17.35  76     47.40  92.20  Genovesa
89     258    4669.32       1707   0.70   28.10  Isabela
23     28     129.49 343    29.10  85.90  Marchena
2      0      0.01   25     3.30   45.90  Onslow
37     67     59.56  777    29.10  119.60 Pinta
33     75     17.95  458    10.70  10.70  Pinzon
9      3      0.23   84     0.50   0.60   Las_Plazas
30     40     4.89   367    4.40   24.40  Rabida
65     215    551.62 716    45.20  66.60  San_Cristobal
81     156    572.33 906    0.20   19.80  San_Salvador
95     349    903.82 864    0.60   0.00   Santa_Cruz
28     34     24.08  259    16.50  16.50  Santa_Fe
73     212    170.92 640    2.60   49.20  Santa_Maria
16     28     1.84   154    0.60   9.60   Seymour
8      8      1.24   186    6.80   50.90  Tortuga
12     9      2.85   253    34.10  254.70 Wolf
;

PROC INSIGHT;
OPEN Galapagos;
RUN;

PROC REG DATA=Galapagos;
MODEL Native = Area DistNear DistSC Elev NonNative /
      SELECTION = RSQUARE ADJRSQ CP;
RUN;
```

13	Int	Int	Int	Int	Int	Int	Nom	Int	Int	Int	Int	Int	Int
30	Native	NonNative	Are	Elev	DistNear	DistSC	Island	R_Native	P_Native	RN_Native	H_Native	RT_Native	D_Native
1	23	35	25.09	332	0.6	0.6	Baltra	1.7283	21.2717	0.4728	0.0715	0.2474	8.1772426E-04
2	21	10	1.24	109	0.6	26.3	Bartolom	10.4803	10.5197	1.6098	0.0643	1.5704	2.6628346E-02
3	3	0	0.21	114	2.8	58.7	Caldwell	-4.5632	7.5632	-0.6680	0.0646	-0.6560	5.0747641E-03
4	9	16	0.10	46	1.9	47.4	Champion	-1.3756	10.3756	-0.3820	0.0720	-0.1969	5.2173892E-04
5	1	1	1.05	130	1.9	1.9	Coamano	-8.5136	9.5136	-1.3609	0.0790	-1.2636	2.2261489E-02
6	11	7	0.34	119	8.0	8.0	Daphne_M	-0.1954	11.1954	-0.2087	0.0783	-0.0280	1.1623109E-05
7	12	12	0.08	93	6.0	12.0	Daphne_M	0.3811	11.6189	0.1246	0.0694	0.0544	3.8392075E-05
8	7	3	2.33	168	34.1	290.2	Darwin	0.7848	6.2152	0.2087	0.4691	0.1484	3.3834780E-03
9	4	4	0.03	46	0.4	0.4	Eden	-4.5834	8.5834	-0.7755	0.0860	-0.6667	7.1342761E-03
10	2	0	0.18	112	2.6	50.2	Enderby	-5.7320	7.7320	-0.8929	0.0625	-0.8275	7.7030500E-03
11	26	71	58.27	198	1.1	88.3	Wspanola	0.8842	25.1158	0.2942	0.1225	0.1301	4.1038974E-04
12	35	58	634.49	1494	4.3	95.3	Fernandi	-9.5459	44.5459	-2.0403	0.6986	-2.7643	2.3117521E+00
13	17	41	0.57	49	1.1	93.1	Gardner_	1.8381	15.1619	0.6680	0.1391	0.2733	2.0919621E-03
14	4	1	0.78	227	4.6	62.2	Gardner_	-5.9688	9.9688	-1.0241	0.0566	-0.8600	7.4713147E-03
15	19	21	17.35	76	47.4	92.2	Genovesa	3.6286	15.3714	0.8929	0.3737	0.6371	4.1401421E-02
16	89	258	4669.32	1707	0.7	28.1	Isabela	1.0605	87.9395	0.3820	0.9824	1.1315	1.1778903E+01
17	23	28	129.49	343	29.1	85.9	Marchena	2.9350	20.0650	0.7755	0.1153	0.4316	4.1869452E-03
18	2	0	0.01	25	3.3	45.9	Onslow	-4.3121	6.3121	-0.5677	0.0717	-0.6216	5.1067392E-03
19	37	67	59.56	777	29.1	119.6	Pinta	0.2616	36.7384	0.0414	0.1989	0.0403	7.0002927E-05
20	33	75	17.95	458	10.7	10.7	Pinzon	-0.9706	33.9706	-0.2942	0.0809	-0.1395	2.9769217E-04
21	9	3	0.23	84	0.5	0.6	Las_Plaz	-0.0467	9.0467	-0.0414	0.0820	-0.0067	7.0110843E-07
22	30	40	4.89	367	4.4	24.4	Rabida	7.0858	22.9142	1.1758	0.0553	1.0269	1.0266771E-02
23	65	215	551.62	716	45.2	66.6	San_Cris	-8.0139	73.0139	-1.1758	0.4320	-1.5381	2.8373850E-01
24	81	156	572.33	906	0.2	19.8	San_Salv	21.9287	59.0713	2.0403	0.1488	4.4802	3.2578008E-01
25	95	349	903.82	864	0.6	0.0	Santa_Cr	-9.5176	104.5176	-1.6098	0.5453	-2.1268	7.8836373E-01
26	28	34	24.08	259	16.5	16.5	Santa_Fe	7.1713	20.8287	1.3609	0.0922	1.0619	1.8985133E-02
27	73	212	170.92	640	2.6	49.2	Santa_Ma	4.8762	68.1238	1.0241	0.2489	0.7853	3.4610055E-02
28	16	28	1.84	154	0.6	9.6	Seymour	-0.1257	16.1257	-0.1246	0.0628	-0.0179	3.7234155E-06
29	8	8	1.24	186	6.8	50.9	Tortuga	-3.4139	11.4139	-0.4728	0.0469	-0.4841	1.9843901E-03
30	12	9	2.85	253	34.1	254.7	Wolf	1.8340	10.1660	0.5677	0.3296	0.3092	8.1388914E-03

Native = Area DistNear DistSC Elev NonNative

Response Distribution: Normal

Link Function: Identity

Model Equation

Native = 6.7437 - 0.0025 Area + 0.0982 DistNear - 0.0265 DistSC + 0.0184 Elev + 0.2409 NonNative

Summary of Fit

Mean of Response	26.5000	R-Square	0.9428
Root MSE	7.1070	Adj R-Sq	0.9308

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	5	19963.2724	3992.6545	79.05	<.0001
Error	24	1212.2276	50.5095		
C Total	29	21175.5000			

Type I Tests

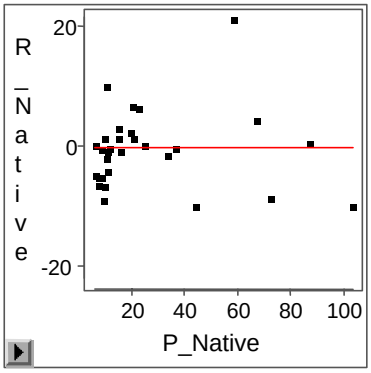
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Area	1	8123.3224	8123.3224	160.83	<.0001
DistNear	1	106.6735	106.6735	2.11	0.1591
DistSC	1	736.3862	736.3862	14.58	0.0008
Elev	1	5165.2902	5165.2902	102.26	<.0001
NonNative	1	5831.6001	5831.6001	115.46	<.0001

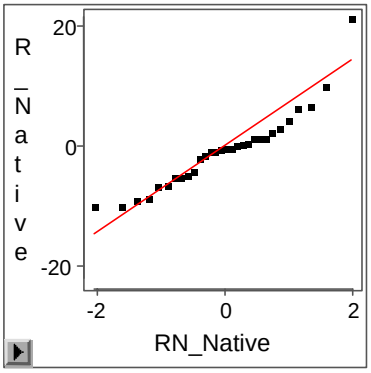
Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Area	1	56.0507	56.0507	1.11	0.3026
DistNear	1	34.0055	34.0055	0.67	0.4200
DistSC	1	54.0551	54.0551	1.07	0.3112
Elev	1	554.4144	554.4144	10.98	0.0029
NonNative	1	5831.6001	5831.6001	115.46	<.0001

Parameter Estimates

Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept	1	6.7437	2.1945	3.07	0.0052	.	0
Area	1	-0.0025	0.0024	-1.05	0.3026	0.4104	2.4365
DistNear	1	0.0982	0.1197	0.82	0.4200	0.5965	1.6765
DistSC	1	-0.0265	0.0256	-1.03	0.3112	0.5735	1.7437
Elev	1	0.0184	0.0056	3.31	0.0029	0.3174	3.1510
NonNative	1	0.2409	0.0224	10.75	<.0001	0.4422	2.2615





The REG Procedure
Model: MODEL1
Dependent Variable: Native

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	0.9133	0.9102	10.3452	NonNative
1	0.6253	0.6119	131.0884	Elev
1	0.3836	0.3616	232.4104	Area
1	0.0276	-.0071	381.6643	DistSC
1	0.0000	-.0357	393.2369	DistNear
2	0.9372	0.9326	2.3191	Elev NonNative
2	0.9153	0.9090	11.5007	Area NonNative
2	0.9137	0.9073	12.1833	DistNear NonNative
2	0.9133	0.9069	12.3417	DistSC NonNative
2	0.6489	0.6229	123.2110	DistSC Elev
2	0.6265	0.5989	132.5713	Area Elev
2	0.6254	0.5977	133.0412	DistNear Elev
2	0.3945	0.3496	229.8550	Area DistSC
2	0.3887	0.3434	232.2985	Area DistNear
2	0.0450	-.0257	376.3713	DistNear DistSC
3	0.9401	0.9332	3.1207	Area Elev NonNative
3	0.9379	0.9308	4.0141	DistSC Elev NonNative
3	0.9376	0.9304	4.1650	DistNear Elev NonNative
3	0.9160	0.9063	13.2280	Area DistNear NonNative
3	0.9153	0.9056	13.4967	Area DistSC NonNative
3	0.9141	0.9041	14.0277	DistNear DistSC NonNative
3	0.6666	0.6282	117.7638	DistNear DistSC Elev
3	0.6491	0.6086	125.1227	Area DistSC Elev
3	0.6268	0.5837	134.4600	Area DistNear Elev
3	0.4234	0.3569	219.7193	Area DistNear DistSC
4	0.9411	0.9317	4.6733	Area DistSC Elev NonNative
4	0.9402	0.9306	5.0702	Area DistNear Elev NonNative
4	0.9401	0.9305	5.1097	DistNear DistSC Elev NonNative
4	0.9166	0.9032	14.9764	Area DistNear DistSC NonNative
4	0.6674	0.6141	119.4555	Area DistNear DistSC Elev
5	0.9428	0.9308	6.0000	Area DistNear DistSC Elev NonNative