Statistics 516 - Spring 2002 - Final Exam *(modified for Spring 2003 Practice)*

Part I: Answer the two following questions. Five points each.

1) In performing a regression, ANOVA, or ANCOVA, what four assumptions must be satisfied?

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

Part II: Answer fifteen of the following sixteen questions. Six points each.

Questions 1-4 refer to the attached results for performing a multiple regression using the data set `Speed`. Note that some parts of the output have been whited out.

1) Carefully state what null and alternate hypotheses the bold faced p-value is testing, identifying any model parameters you use. Do we accept or reject this null hypothesis at α=0.05?

2) Find the SSE, MSE, and their corresponding df.

3) Which of the cities would cause the model to be changed the most if it was removed? Which statistic did you use to tell this? Would the change in the model be very large?

4) What set of these variables forms the best regression model for predicting the death rate from heart attacks? What did you use to tell this?

Questions 5-8 refer to the attached results for performing an ANOVA on the data set `Rating`.

5) Complete the model equation for this ANOVA, identifying the parameters used. [NOT the estimated model equation!]

$$y_{ijk} = \mu_{base} +$$

y_{ijk} = the k^{th} observation in specialty i and city j
μ_{base} = the baseline mean

6) What hypothesis is being tested by `Contrast1`?

7) Say that the Holm procedure was used to see which specialties were significantly different from the other specialties. Why can we assume that these differences between the specialties are the same regardless of the city?

8) Suppose you could not tell whether the variances of the residuals were equal from the plots. What is the name of the test we used for this hypothesis?

Questions 9-12 refer to the attached results for performing an ANCOVA on the data set `Faculty`. Use α=0.05 if needed.

9) Assume the assumption of equal slopes is met. Is there a significant difference in the tolerance shown by professors of the same age but different ranks? If so, what is the average difference in the tolerance of an assistant and a full professor of the same age?

10) Assume the assumption of equal slopes is met. Is there a significant effect of `Age` on `Tolerance` for professors of a given rank? If so, what is the average change in `Tolerance` for each additional year of `Age`?

11) Assume the assumption of equal slopes is met. What percent of the variation in Tolerance is explained by this model using `Rank` and `Age`?

12) Is the assumption of equal slopes met for this model? (How could you tell?)

Questions 13-16 refer to the attached results for performing a logistic regression on the data set `Bumpus`.

13) Does a logistic regression seem to fit this data at α=0.05? (What test did you use to tell this?)

14) Assuming that the logistic regression does fit this data, is the length of the sparrow related to their chance of surviving at an α=0.05 level? (What test did you use to tell this?)

15) What is the predicted chance of surviving for a sparrow of length 200?

16) Why can't you trust the answer you gave in part c?

The data set `Speed` is from a 1990 article in *American Scientist*. The three independent variables are related to the pace of life in each of the 36 cities in the study:

`bank` the average time it takes a bank clerk to make change for two $20 bills
`walk` the average walking speed of pedestrians over 60ft on a clear summer business day downtown
`talk` the talking speed of postal clerks explaining the difference between regular, certified, and insured mail

The dependent variable is the age-adjusted death rate from ischemic heart disease. All four variables have been put on a standard scale so that units do not play a role in the analysis.

```
DATA Speed;
INPUT bank walk talk heart city $;
CARDS;
      31   28   24    24      Boston
      30   23   23    29      Buffalo
      29   24   18    31      NewYork
      28   28   23    26      SaltLake
      27   22   30    26      Columbus
      26   25   24    20      Worcester
      30   26   24    17      Providence
      28   30   21    19      Springfield
      33   22   18    26      Rochester
      33   22   22    24      KansasCity
      22   23   23    26      StLouis
      30   25   20    25      Houston
      32   23   23    14      Paterson
      29   18   25    11      Bakersfield
      25   27   27    19      Atlanta
      24   22   14    24      Detroit
      27   23   24    20      Youngstown
      26   22   24    13      Indianapolis
      24   23   25    20      Chicago
      31   12   19    18      Philadelphia
      27   23   17    16      Louisville
      28   20   18    19      Canton
      21   20   17    23      Knoxville
      19   22   18    11      SanFrancisco
      34   14   22    27      Chattanooga
      24   20   23    18      Dallas
      25   17   19    15      Oxnard
      25   26   19    20      Nashville
      20   19   22    18      SanDiego
      22   23   23    21      EastLansing
      26   13   22    11      Fresno
      29   16   21    14      Memphis
      25   17   18    19      SanJose
      22   17   15    15      Shreveport
      24   16   10    18      Sacramento
      13   20   12    16      LosAngeles
;

PROC INSIGHT;
OPEN Speed;
FIT heart = bank walk talk;
RUN;

PROC REG DATA=Speed;
MODEL  heart = bank walk talk /
      SELECTION = RSQUARE ADJRSQ CP;
RUN;
```

heart = bank walk talk

Response Distribution: Normal

Link Function: Identity

Model Equation

heart = 3.1787 + 0.4052 bank + 0.4516 walk - 0.1796 talk

Summary of Fit

Mean of Response	19.8056	R-Square	0.2236
Root MSE	4.8050	Adj R-Sq	0.1509

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	3	212.8264	70.9421	3.07	0.0416
Error					
C Total	35	951.6389			

Type I Tests

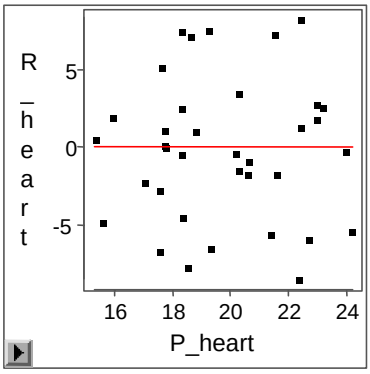
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
bank	1	95.9781	95.9781	4.16	0.0498
walk	1	101.7651	101.7651	4.41	**0.0438**
talk	1	15.0833	15.0833	0.65	0.4249

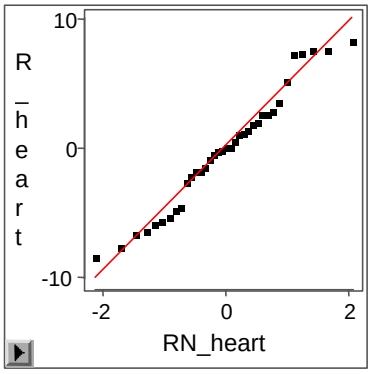
Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
bank	1	97.5837	97.5837	4.23	0.0480
walk	1	116.6941	116.6941	5.05	0.0316
talk	1	15.0833	15.0833	0.65	0.4249

Parameter Estimates

Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept	1	3.1787	6.3369	0.50	0.6194	.	0
bank	1	0.4052	0.1971	2.06	0.0480	0.8736	1.1447
walk	1	0.4516	0.2009	2.25	0.0316	0.8902	1.1233
talk	1	-0.1796	0.2222	-0.81	0.4249	0.7835	1.2763





	10	Int	Int	Int	Int	Nom	Int	Int	Int	Int	Int
36		bank	walk	talk	heart	city	R_heart	P_heart	H_heart	RT_heart	F_heart
■	1	31	28	24	24	Boston	-0.0746	24.0746	0.1211	-0.0163	-0.0061
■	2	30	23	23	29	Buffalo	7.4090	21.5910	0.0510	1.6227	0.3761
■	3	29	24	18	31	NewYork	8.4646	22.5354	0.0872	1.9196	0.5932
■	4	28	28	23	26	SaltLake	2.9614	23.0386	0.0974	0.6427	0.2111
■	5	27	22	30	26	Columbus	7.3335	18.6665	0.1975	1.7585	0.8724
■	6	26	25	24	20	Worceste	-0.6937	20.6937	0.0597	-0.1466	-0.0369
■	7	30	26	24	17	Providen	-5.7662	22.7662	0.0780	-1.2613	-0.3669
■	8	28	30	21	19	Springfi	-5.3010	24.3010	0.1606	-1.2130	-0.5305
■	9	33	22	18	26	Rocheste	2.7469	23.2531	0.1457	0.6124	0.2529
■	10	33	22	22	24	KansasCi	1.4653	22.5347	0.0947	0.3160	0.1022
■	11	22	23	23	26	StLouis	7.6507	18.3493	0.0823	1.7114	0.5124
■	12	30	25	20	25	Houston	1.9670	23.0330	0.0831	0.4220	0.1271
■	13	32	23	23	14	Paterson	-8.4014	22.4014	0.0768	-1.8917	-0.5457
■	14	29	18	25	11	Bakersfi	-7.5686	18.5686	0.1002	-1.7097	-0.5706
■	15	25	27	27	19	Atlanta	-1.6529	20.6529	0.1349	-0.3648	-0.1441
■	16	24	22	14	24	Detroit	3.6754	20.3246	0.1187	0.8104	0.2975
■	17	27	23	24	20	Youngsto	-0.1957	20.1957	0.0464	-0.0411	-0.0091
■	18	26	22	24	13	Indianap	-6.3389	19.3389	0.0503	-1.3723	-0.3159
■	19	24	23	25	20	Chicago	1.1995	18.8005	0.0843	0.2570	0.0780
■	20	31	12	19	18	Philadel	0.2529	17.7471	0.2075	0.0582	0.0298
■	21	27	23	17	16	Louisvil	-5.4530	21.4530	0.0738	-1.1867	-0.3349
■	22	28	20	18	19	Canton	-1.3238	20.3238	0.0526	-0.2790	-0.0658
■	23	21	20	17	23	Knoxvill	5.3331	17.6669	0.0778	1.1621	0.3374
■	24	19	22	18	11	SanFranc	-6.5801	17.5801	0.1101	-1.4783	-0.5199
■	25	34	14	22	27	Chattano	7.6729	19.3271	0.2139	1.8700	0.9755
■	26	24	20	23	18	Dallas	0.1951	17.8049	0.0632	0.0413	0.0107
■	27	25	17	19	15	Oxnard	-2.5738	17.5738	0.0598	-0.5463	-0.1377
■	28	25	26	19	20	Nashvill	-1.6382	21.6382	0.0799	-0.3505	-0.1033
■	29	20	19	22	18	SanDiego	2.0880	15.9120	0.1266	0.4592	0.1748
■	30	22	23	23	21	EastLans	2.6507	18.3493	0.0823	0.5697	0.1706
■	30	22	23	23	21	EastLans	2.6507	18.3493	0.0823	0.5697	0.1706
■	31	26	13	22	11	Fresno	-4.6337	15.6337	0.1695	-1.0603	-0.4791
■	32	29	16	21	14	Memphis	-4.3838	18.3838	0.0891	-0.9546	-0.2985
■	33	25	17	18	19	SanJose	1.2466	17.7534	0.0620	0.2640	0.0679
■	34	22	17	15	15	Shrevepo	-2.0765	17.0765	0.1029	-0.4506	-0.1526
■	35	24	16	10	18	Sacramen	-0.3334	18.3334	0.2311	-0.0779	-0.0427
■	36	13	20	12	16	LosAngel	0.6768	15.3232	0.3281	0.1692	0.1183

The REG Procedure

Dependent Variable: heart

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	0.1209	0.0950	4.2356	walk
1	0.1009	0.0744	5.0610	bank
1	0.0100	-.0191	8.8071	talk
2	0.2078	0.1598	2.6533	bank walk
2	0.1211	0.0678	6.2266	walk talk
2	0.1010	0.0465	7.0543	bank talk
3	0.2236	0.1509	4.0000	bank walk talk

The data set `Rating` is from Kleinbaum, Kupper, Muller, and Nizam (1998). It concerns the effectiveness of family nurse practioners (FNPs) with different specialties (`Spec`) from hospitals in three cities (`City`).

The three specialties are: Pediatrics (`PED`), Obstetrics and gynecology (`OBGYN`), Diabetes and hypertension (`DnH`). The three cities are simply numbered `1`, `2`, and `3`.

The dependent variable is a performance competency score.

```
DATA Rating;
INPUT Spec $ City $ Score @@;
CARDS;
  PED   1     91.7              OBGYN 1     80.1              DnH   1     71.5
  PED   1     74.9              OBGYN 1     76.2              DnH   1     49.8
  PED   1     88.2              OBGYN 1     70.3              DnH   1     55.1
  PED   1     79.5              OBGYN 1     89.5              DnH   1     75.4
  PED   2     86.3              OBGYN 2     71.3              DnH   2     80.2
  PED   2     88.1              OBGYN 2     73.4              DnH   2     76.1
  PED   2     92                OBGYN 2     76.9              DnH   2     44.2
  PED   2     69.5              OBGYN 2     87.2              DnH   2     50.5
  PED   3     82.3              OBGYN 3     90.1              DnH   3     48.7
  PED   3     78.7              OBGYN 3     65.6              DnH   3     54.4
  PED   3     89.8              OBGYN 3     74.6              DnH   3     60.1
  PED   3     84.5              OBGYN 3     79.1              DnH   3     70.8
  ;

  PROC INSIGHT;
  OPEN Rating;
  FIT Score = Spec City Spec*City;
  RUN;

  PROC GLM DATA=Rating ORDER=DATA;
  CLASS Spec City;
  MODEL Score = Spec City Spec*City;
  CONTRAST 'contrast 1'  Spec  1 0 -1;
  ESTIMATE 'contrast 1'  Spec  1 0 -1;
  RUN;
```

Nominal Variable Information

Level	Spec	City
1	DnH	1
2	OBGYN	2
3	PED	3

Parameter Information

Parameter	Variable	Spec	City
1	Intercept		
2	Spec	DnH	
3		OBGYN	
4		PED	
5	City		1
6			2
7			3
8	Spec*City	DnH	1
9		DnH	2
10		DnH	3
11		OBGYN	1
12		OBGYN	2
13		OBGYN	3
14		PED	1
15		PED	2
16		PED	3

Model Equation

Score = 83.8250 - 25.3250 P_2 - 6.4750 P_3 - 0.2500 P_5 + 0.1500 P_6 + 4.7000 P_8 + 4.1000 P_9 + 1.9250 P_11 - 0.3000 P_12

Summary of Fit

Mean of Response	74.3500	R-Square	0.5312
Root MSE	10.3676	Adj R-Sq	0.3923

Analysis of Variance

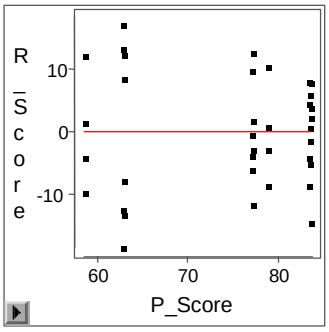
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	8	3288.9500	411.1188	3.82	0.0040
Error	27	2902.1800	107.4881		
C Total	35	6191.1300			

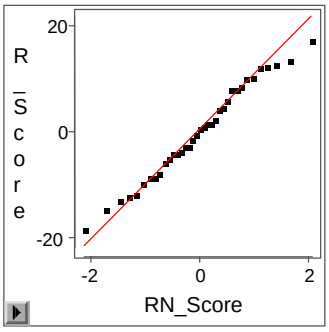
Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Spec	2	3229.8717	1614.9358	15.02	<.0001
City	2	24.5417	12.2708	0.11	0.8925
Spec*City	4	34.5367	8.6342	0.08	0.9877

Parameter Estimates

Variable	Spec	City	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept			1	83.8250	5.1838	16.17	<.0001	.	0
Spec	DnH		1	-25.3250	7.3310	-3.45	0.0018	0.2500	4.0000
	OBGYN		1	-6.4750	7.3310	-0.88	0.3849	0.2500	4.0000
	PED		0	0	.	.	.	.	.
City		1	1	-0.2500	7.3310	-0.03	0.9730	0.2500	4.0000
		2	1	0.1500	7.3310	0.02	0.9838	0.2500	4.0000
		3	0	0	.	.	.	.	.
Spec*City	DnH	1	1	4.7000	10.3676	0.45	0.6539	0.2813	3.5556
	DnH	2	1	4.1000	10.3676	0.40	0.6956	0.2813	3.5556
	DnH	3	0	0	.	.	.	.	.
	OBGYN	1	1	1.9250	10.3676	0.19	0.8541	0.2813	3.5556
	OBGYN	2	1	-0.3000	10.3676	-0.03	0.9771	0.2813	3.5556
	OBGYN	3	0	0	.	.	.	.	.
	PED	1	0	0	.	.	.	.	.
	PED	2	0	0	.	.	.	.	.
	PED	3	0	0	.	.	.	.	.





The GLM Procedure

Dependent Variable: Score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3288.950000	411.118750	3.82	0.0040
Error	27	2902.180000	107.488148		
Corrected Total	35	6191.130000			

R-Square	Coeff Var	Root MSE	Score Mean
0.531236	13.94438	10.36765	74.35000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Spec	2	3229.871667	1614.935833	15.02	<.0001
City	2	24.541667	12.270833	0.11	0.8925
Spec*City	4	34.536667	8.634167	0.08	0.9877

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Spec	2	3229.871667	1614.935833	15.02	<.0001
City	2	24.541667	12.270833	0.11	0.8925
Spec*City	4	34.536667	8.634167	0.08	0.9877

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
contrast 1	1	3008.320417	3008.320417	27.99	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
contrast 1	22.3916667	4.23257503	5.29	<.0001

The data set `Faculty` is from Kleinbaum, Kupper, Muller, and Nizam (1998). It concerns the political `Tolerance` of university faculty members (higher score equals more tolerance) based on the faculty members `Age` (in years) and `Rank` (Full, Associate, or Assistant Professor).

```
DATA Faculty;
INPUT Rank $ Age  Tolerance;
CARDS;
Full        65    3.03
Full        61    2.7
Full        47    4.31
Full        52    2.7
Full        49    5.09
Full        45    4.02
Full        41    3.71
Full        41    5.52
Full        40    5.29
Full        39    4.62
Assoc       34    4.62
Assoc       31    5.22
Assoc       30    4.85
Assoc       35    4.51
Assoc       49    5.12
Assoc       31    4.47
Assoc       42    4.5
Assoc       43    4.88
Assoc       39    5.17
Assoc       49    5.21
Assist      26    5.2
Assist      33    5.86
Assist      48    4.61
Assist      32    4.55
Assist      25    4.47
Assist      33    5.71
Assist      42    4.77
Assist      30    5.82
Assist      31    3.67
Assist      27    5.29
;
```

```
PROC INSIGHT;
OPEN Faculty;
FIT Tolerance = Age Rank;
RUN;
```

Parameter Information

Parameter	Variable	Rank
1	Intercept	
2	Age	
3	Rank	Assist
4		Assoc
5		Full

Model Equation

Tolerance = 5.8439 - 0.0364 Age + 0.3398 P_3 + 0.4034 P_4

Summary of Fit

Mean of Response	4.6497	R-Square	0.3429
Root MSE	0.7105	Adj R-Sq	0.2671

Analysis of Variance

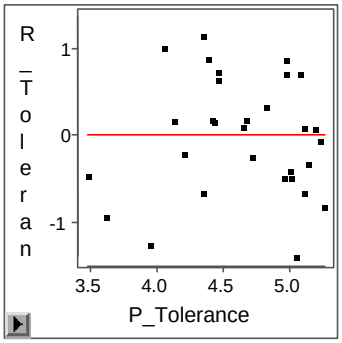
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	3	6.8484	2.2828	4.52	0.0111
Error	26	13.1237	0.5048		
C Total	29	19.9721			

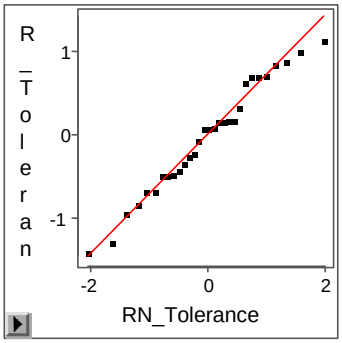
Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Age	1	2.2019	2.2019	4.36	0.0467
Rank	2	0.6434	0.3217	0.64	0.5368

Parameter Estimates

Variable	Rank	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept		1	5.8439	0.8651	6.75	<.0001	.	0
Age		1	-0.0364	0.0174	-2.09	0.0467	0.5816	1.7193
Rank	Assist	1	0.3398	0.4146	0.82	0.4199	0.4405	2.2700
	Assoc	1	0.4034	0.3598	1.12	0.2725	0.5849	1.7098
	Full	0	0	.	.	.	.	.






```
PROC INSIGHT;
OPEN Faculty;
FIT Tolerance = Age Rank Age*Rank;
RUN;
```

Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Age	1	1.1897	1.1897	2.92	0.1001
Rank	2	2.6243	1.3122	3.23	0.0574
Age*Rank	2	3.3610	1.6805	4.13	0.0287

The data set `bumpus` is an excerpt from the classic data set by Bumpus (1898). It concerns the mortality of house sparrows based on several measurements. In this case `survive=1` indicates the sparrow survived, `survive=0` indicates it did not. The predictor variable in this case is `length`.

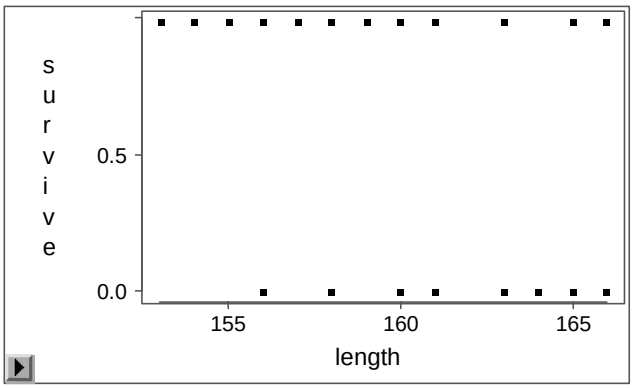
```
DATA bumpus;
INPUT survive length @@;
CARDS;
1     154                     0     160                     0     156
0     165                     1     160                     1     161
0     160                     0     161                     1     163
1     160                     1     160                     0     166
1     155                     1     161                     1     156
0     161                     0     160                     0     165
1     154                     1     160                     1     165
0     160                     0     165                     0     166
1     156                     1     159                     1     160
0     163                     0     161                     1     158
1     161                     1     158                     1     160
0     160                     0     161                     1     157
1     157                     1     159                     1     159
0     163                     0     160                     1     160
1     159                     0     164                     1     158
0     161                     1     166                     1     161
1     158                     0     158                     1     160
0     160                     1     159                     1     160
1     158                     1     160                     1     153
0     160                     0     160
;

PROC INSIGHT;
OPEN bumpus;
RUN;

PROC LOGISTIC DATA=bumpus DESCENDING;
MODEL survive = length / LACKFIT;
RUN;
```

Model Equation

Logit (survive) = 59.2919 - 0.3677 length



Summary of Fit

Mean of Response	0.5932	Deviance	68.0955	Pearson ChiSq	58.7201
SCALE	1.0000	Deviance / DF	1.1947	Pearson ChiSq / DF	1.0302
		Scaled Dev	68.0955	Scaled ChiSq	58.7201

Analysis of Deviance

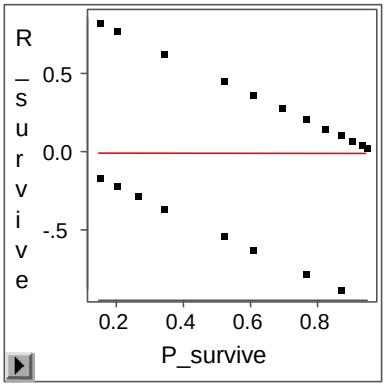
Source	DF	Deviance	Deviance / DF	Scaled Dev	Pr > Scaled Dev
Model	1	11.6330	11.6330	11.6330	0.0006
Error	57	68.0955	1.1947	68.0955	
C Total	58	79.7285			

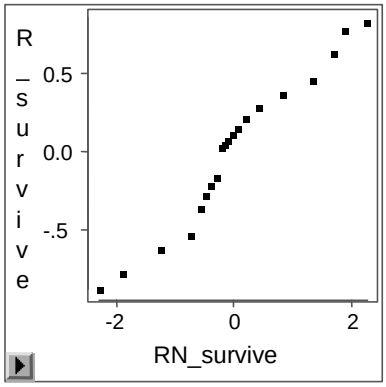
Type III (Wald) Tests

Source	DF	ChiSq	Pr > ChiSq
length	1	8.3611	0.0038

Parameter Estimates

Variable	DF	Estimate	Std Error	ChiSq	Pr > ChiSq
Intercept	1	59.2919	20.3888	8.4568	0.0036
length	1	-0.3677	0.1272	8.3611	0.0038





Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.6330	1	0.0006
Score	10.4794	1	0.0012
Wald	8.3611	1	0.0038

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
4.0389	6	0.6714