```
1          ********************************************************;
2          *** Logistic Regression - Disease outbreak example     ***;
3          *** NKNW table 14.3  (Appendix C3)                     ***;
4          *** Study of a disease outbreak from a mosquito born   ***;
5          *** disease within two sectors of a city.              ***;
6          ********************************************************;
7
8          dm'log;clear;output;clear';
9          options nodate nocenter nonumber ps=512 ls=132 nolabel;
10         ODS HTML style=minimal rs=none
body='C:\Geaghan\Current\EXST7034\Fall2005\SAS\DiseaseOutbreak01.html' ;
NOTE: Writing HTML Body file: C:\Geaghan\Current\EXST7034\Fall2005\SAS\DiseaseOutbreak01.html
11
12         TITLE1 'Logistic Regression - NKNW Example 14.3';
13         data Disease; infile cards missover;
14            input case Age Status1 Status2 sector Disease;
15            label case = 'case number'
16                  age  = 'Patients age'
17                  status  = 'Socioeconomic status upper, middle and lower'
18                   disease = 'Disease present = 1';
19         * Status classes are upper (0, 0), Middle (1, 0) and Lower (0, 1);
20         Cards;

NOTE: Variable status is uninitialized.
NOTE: The data set WORK.DISEASE has 98 observations and 6 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.03 seconds
119        ;
120           ods html;
121           ods graphics on;
NOTE: ODS Statistical Graphics will require a SAS/GRAPH license when it is declared production.
122
123        proc logistic data=Disease DESCENDING alpha=0.05;
124           TITLE2 'Logistic regression on Disease data';
125           model Disease = Age Status1 Status2 Sector / lackfit RSQ iplots;
126           output out=next1 PREDICTED=yhat Lower=lcl95 Upper=ucl95 dfbetas=_ALL_
127                  resdev=resdev difdev=difdev;
128        run;
NOTE: PROC LOGISTIC is modeling the probability that Disease=1.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
WARNING: Statistical graphics displays created with ODS are experimental in this release.
NOTE: There were 98 observations read from the data set WORK.DISEASE.
NOTE: The data set WORK.NEXT1 has 98 observations and 17 variables.
NOTE: At least one W.D format was too small for the number to be printed. The decimal may be
shifted by the "BEST" format.
NOTE: The PROCEDURE LOGISTIC printed page 1.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           4.77 seconds
      cpu time            3.12 seconds
```

Logistic Regression - NKNW Example 14.3
Logistic regression on Disease data

The LOGISTIC Procedure

Model Information

Data Set	**WORK.DISEASE**
Response Variable	**Disease**
Number of Response Levels	**2**
Model	**binary logit**
Optimization Technique	**Fisher's scoring**

Number of Observations Read	**98**
Number of Observations Used	**98**

```
          Response Profile
  Ordered                      Total
   Value      Disease      Frequency
       1            1             31
       2            0             67
Probability modeled is Disease=1.

            Model Convergence Status
                     Convergence criterion (GCONV=1E-8) satisfied.
                     Model Fit Statistics
                                         Intercept
                         Intercept             and
            Criterion         Only      Covariates
            AIC            124.318         111.054
            SC             126.903         123.979
            -2 Log L       122.318         101.054

            R-Square    0.1950    Max-rescaled R-Square    0.2736
```

Model fit statistics

1) Akaike Information Criterion $\text{AIC} = -2\log(L) + 2p$

where Log(L) is the log likelihood and p is the number of parameters

2) Schwarz Criterion $\text{SC} = -2\log(L) + p\log(\sum_j f_j)$

3) -2log L $-2\sum_{i-1}^{n}\left[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i)\right]$

This is analogous to the SSE in regression and is given in SAS as the “-2 Log L”.

Two models (full and reduced) can be compared by calculating the difference in “-2 Log L” for both models. This difference follows a chi square distribution with a d.f. equal to the difference in d.f. for the two models.

4) Generalized R^2 $1-\left(\frac{L(0)}{L(\theta)}\right)^{\frac{2}{n}}$, where L(0) is the intercept only model.

Since this value reaches its maximum of less than 1 for discrete models an adjustment has been proposed. This is called the Max-rescaled Rsquare in SAS. $\frac{R^2}{R^2_{\max}}$

```
        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq
Likelihood Ratio        21.2635        4         0.0003
Score                   20.4067        4         0.0004
Wald                    16.6437        4         0.0023
```

Wald : used to test individual parameter estimates and to place confidence intervals. It is based on a large sample assumption of asymptotic normality.

Chi-square Test $\beta_i^2 / Var(\beta_i) = [\beta_i / Stderr(\beta_i)]^2$

Confidence interval $P\left(e^{(\hat{\beta}_i - 1.96 S_{\hat{\beta}_i})} \le \beta i \le e^{(\hat{\beta}_i - 1.96 S_{\hat{\beta}_i})}\right) = 0.95$

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3127	0.6426	12.9545	0.0003
Age	1	0.0297	0.0135	4.8535	0.0276
Status1	1	0.4088	0.5990	0.4657	0.4950
Status2	1	-0.3051	0.6041	0.2551	0.6135
sector	1	1.5746	0.5016	9.8543	0.0017

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.030	1.003	1.058
Status1	1.505	0.465	4.868
Status2	0.737	0.226	2.408
sector	4.829	1.807	12.907

Association of Predicted Probabilities and Observed Responses

Percent Concordant	77.5	Somers' D	0.554
Percent Discordant	22.1	Gamma	0.556
Percent Tied	0.3	Tau-a	0.242
Pairs	2077	c	0.777

Association of Predicted Probabilities and Observed Responses

Observations with different responses are paired and compared. If the one with the lower observed response has a lower predicted value it is said to be concordant. Otherwise they are discordant or tied. SAS reports concordant, discordant, ties and the number of pairs examined.

A number of other statistics are based on the same information of the number concordant (n_c) and the number discordant (n_d). Where "t" is the total number of pairs with different responses and N is the sum of observation frequencies in the data then the following statistics can be derived. Note that ties are given by $t-n_c-n_d$. The statistic "c" is equal to $(n_c + 0.5(t-n_c-n_d))/t$. Somers' D is equal to $(n_c-n_d)/t$. The Goodman-Kruskal Gamma is $(n_c-n_d)/(n_c+n_d)$ and Kendall's Tau-a is $(n_c-n_d)/(0.5N(N-1)/$

Partition for the Hosmer and Lemeshow Test

Group	Total	Disease = 1 Observed	Disease = 1 Expected	Disease = 0 Observed	Disease = 0 Expected
1	10	0	0.79	10	9.21
2	10	1	1.02	9	8.98
3	11	2	1.51	9	9.49
4	10	1	1.78	9	8.22
5	10	3	2.34	7	7.66
6	10	4	3.09	6	6.91
7	10	7	3.91	3	6.09
8	11	3	5.51	8	5.49
9	10	5	6.32	5	3.68
10	6	5	4.75	1	1.25

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
9.1871	8	0.3268

Goodness-of-fit

Most Goodness-of-fit tests (Pearson and deviance) require replication in subpopulations. This is often a problem with continuous variables in the model. The Hosmer-Lemeshow Goodness-of-Fit Test can be used for sparser data. This test is only available for binary models.

In this approach the data are sorted on the basis of their response probability (default), and divided into approximately 10 groups (minimum = 3). See SAS help for details on the grouping.

Once in groups a Chi square statistic is calculated. For each group we have S_i as the observed number of "successes" in the group, n_i as the total number of observations in the group and $\bar{\pi}_i$ as the mean predicted probability in each group (from the model). For the "g" groups the Chi square statistic is then calculated as:

$$\chi^2 = \sum_{i=1}^{g} \frac{(S_i - n_i\bar{\pi}_i)^2}{n_i\bar{\pi}_i(1-\bar{\pi}_i)}$$

The usual interpretations apply to this lack of fit test. Small values of P would indicate an inadequate model.

Deviance

The deviance in logistic regression can be calculated as

$$DEV(X_0, X_1, ..., X_{p-1}) = -2\sum_{i-1}^{n}\left[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i)\right]$$

Partial deviances can be calculated for reduced models and the difference in the two models (with n−p and n−p+*diff* d.f.) should follow a Chi square distribution with *diff* d.f. This test of partial deviances is also called the likelihood ratio test.

Deviance residuals

Residual analysis is not as simple with Logistic Regression as it was with Linear Regression. Since the dependent variable is 0 or 1 the not normally distributed and in fact the distribution is not known. As a result residual analysis in Logistic Regression is done on "Deviance" residuals. These are calculated as:

$$dev_i = \pm\sqrt{\{-2[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i)]\}}$$

where the sign is + if $Y_i \geq \hat{\pi}_i$ and the sign is - if $Y_i < \hat{\pi}_i$. Note that the SS of these residuals will sum to the model deviance.

Pearson residuals

Given the residual $e_i = Y_i - \hat{\pi}_i$, the predicted probabilities $\hat{\pi}_i = Y_i - e_i$ and r_i = the number of events in a given observation with n_i trials, then the Pearson residual is given as:

$$\frac{r_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Confidence interval displacement diagnostics

SAS provides two measures of confidence interval displacement as influence diagnostics, C for individual observations and CBAR for overall change in the parameter estimates when individual observations are removed.

Logistic Regression - NKNW Example 14.3
Listing of one kept value for each by group from Logistic Reg

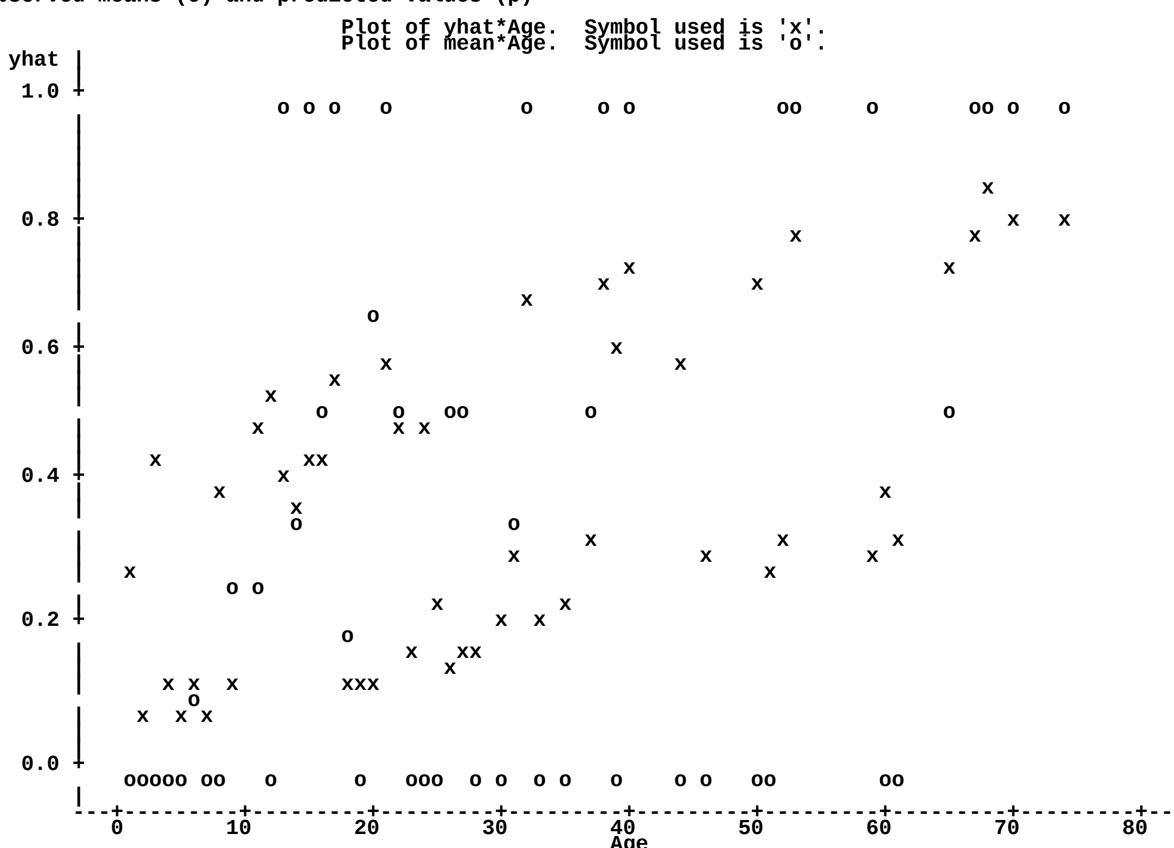
Obs	case	Age	Status1	Status2	sector	Disease	_LEVEL_	yhat	lcl95	ucl95	resdev	DFBETA_Intercept	DFBETA_Age	DFBETA_Status1	DFBETA_Status2	DFBETA_sector	difdev
1	40	1	0	1	1	0	1	0.26630	0.09293	0.56251	-0.78696	-0.01939	0.09579	0.01031	-0.10891	-0.08726	0.65146
2	28	2	0	1	0	0	1	0.07187	0.02346	0.19976	-0.38621	-0.02297	0.02517	0.00406	-0.02016	0.01472	0.15105
3	37	3	1	0	1	0	1	0.44026	0.20277	0.70864	-1.07729	-0.03653	0.13065	-0.15334	-0.00929	-0.09777	1.23059
4	34	4	0	0	0	0	1	0.10031	0.03273	0.26867	-0.45980	-0.06299	0.03831	0.03101	0.03402	0.03349	0.21527
5	62	5	0	1	0	0	1	0.07805	0.02670	0.20716	-0.40315	-0.02281	0.02410	0.00395	-0.02219	0.01582	0.16460
6	3	6	0	0	0	0	1	0.10581	0.03573	0.27423	-0.47295	-0.06449	0.03745	0.03228	0.03561	0.03517	0.22774
7	91	7	0	1	0	0	1	0.08244	0.02904	0.21258	-0.41482	-0.02260	0.02318	0.00385	-0.02364	0.01660	0.17427
8	18	8	0	0	1	0	1	0.37751	0.18144	0.62395	-0.97368	-0.12774	0.11676	0.10849	0.08850	-0.07308	0.98811
9	14	9	0	0	0	1	1	0.11456	0.04062	0.28333	2.08167	0.51506	-0.27619	-0.26478	-0.29462	-0.29239	4.59669
10	24	11	1	0	1	0	1	0.49946	0.25967	0.73950	-1.17650	-0.00272	0.08844	-0.18022	-0.01563	-0.11255	1.46058
11	46	12	1	0	1	0	1	0.50690	0.26696	0.74370	-1.18915	0.00209	0.08229	-0.18374	-0.01651	-0.11446	1.49154
12	42	13	0	0	1	1	1	0.41304	0.21594	0.64260	1.32982	0.16963	-0.13835	-0.16158	-0.13291	0.11495	1.85198
13	25	14	0	1	1	0	1	0.34824	0.14743	0.62277	-0.92529	0.01835	0.05833	0.00400	-0.14741	-0.11664	0.89961
14	20	15	0	0	1	1	1	0.42754	0.23031	0.65085	1.30362	0.15429	-0.11829	-0.15489	-0.12785	0.11266	1.77482
15	23	16	0	0	1	1	1	0.43483	0.23757	0.65515	1.29058	0.14686	-0.10860	-0.15161	-0.12536	0.11150	1.73728
16	39	17	1	0	1	1	1	0.54397	0.30330	0.76572	1.10351	-0.02357	-0.04073	0.16920	0.01774	0.10412	1.27659
17	5	18	0	1	0	1	1	0.11082	0.04422	0.25134	2.09755	0.15482	-0.11485	-0.02231	0.26748	-0.17253	4.60704
18	96	19	0	1	0	0	1	0.11378	0.04577	0.25578	-0.49152	-0.01878	0.01312	0.00263	-0.03438	0.02201	0.24494
19	30	20	0	1	0	0	1	0.11682	0.04734	0.26041	-0.49845	-0.01822	0.01186	0.00247	-0.03545	0.02252	0.25194
20	21	21	1	0	1	1	1	0.57330	0.33160	0.78442	1.05485	-0.03820	-0.01341	0.16141	0.01876	0.09840	1.16292
21	16	22	0	0	1	1	1	0.47909	0.28142	0.68354	1.21315	0.10549	-0.05528	-0.13272	-0.11097	0.10445	1.52577
22	11	23	0	0	0	0	1	0.16403	0.06903	0.34176	-0.59860	-0.07437	0.01912	0.04453	0.05175	0.05261	0.36518
23	43	24	0	0	1	0	1	0.49395	0.29583	0.69399	-1.16715	-0.09070	0.03837	0.12371	0.10385	-0.09963	1.40958
24	81	25	1	0	0	0	1	0.23861	0.09829	0.47395	-0.73839	-0.03390	0.00011	-0.08578	0.01437	0.07126	0.56267
25	6	26	0	1	0	0	1	0.13653	0.05707	0.29233	-0.54184	-0.01385	0.00251	0.00128	-0.04256	0.02582	0.29819
26	13	27	0	0	0	0	1	0.18100	0.07840	0.36473	-0.63194	-0.07550	0.01099	0.04777	0.05627	0.05763	0.40740
27	70	28	0	0	0	0	1	0.18545	0.08079	0.37099	-0.64051	-0.07569	0.00867	0.04860	0.05745	0.05895	0.41866
28	68	30	0	0	0	0	1	0.19461	0.08559	0.38416	-0.65792	-0.07592	0.00366	0.05028	0.05987	0.06166	0.44206
29	8	31	1	0	0	1	1	0.27253	0.11409	0.52147	1.61246	0.06279	0.06227	0.27211	-0.03847	-0.21553	2.76692
30	22	32	1	0	1	1	1	0.65079	0.40153	0.83810	0.92689	-0.06762	0.04432	0.13975	0.02029	0.08316	0.89443

31	1	33	0	0	0	0	1	0.20898	0.09281	0.40556	-0.68473	-0.07590	-0.00487	0.05285	0.06363	0.06591	0.47951
32	2	35	0	0	0	0	1	0.21898	0.09759	0.42094	-0.70308	-0.07561	-0.01128	0.05460	0.06624	0.06888	0.50612
33	10	37	1	0	0	0	1	0.30931	0.12945	0.57423	-0.86030	-0.00925	-0.05360	-0.12060	0.01410	0.09116	0.77259
34	50	38	1	0	1	1	1	0.69019	0.43337	0.86648	0.86115	-0.07776	0.06620	0.12804	0.02044	0.07527	0.77207
35	57	39	0	0	1	0	1	0.60396	0.39037	0.78411	-1.36107	-0.02646	-0.08139	0.13312	0.11555	-0.12830	1.92752
36	49	40	1	0	1	1	1	0.70277	0.44296	0.87546	0.83992	-0.08033	0.07216	0.12419	0.02040	0.07271	0.73465
37	35	44	0	1	1	0	1	0.56602	0.28366	0.81117	-1.29210	0.19641	-0.15575	-0.02924	-0.26635	-0.20543	1.79985
38	69	46	0	0	0	0	1	0.28001	0.12255	0.51991	-0.81059	-0.06920	-0.05866	0.06465	0.08206	0.08725	0.67970
39	58	50	0	0	1	0	1	0.67901	0.43957	0.85086	-1.50754	0.04032	-0.19919	0.13590	0.12156	-0.15094	2.39781
40	84	51	0	1	0	0	1	0.24960	0.09562	0.51133	-0.75782	0.03299	-0.08525	-0.01037	-0.08886	0.04451	0.59705
41	74	52	0	0	0	1	1	0.31736	0.13470	0.58133	1.51506	0.13190	0.20421	-0.15128	-0.19742	-0.21258	2.45114
42	41	53	1	0	1	1	1	0.77682	0.49439	0.92531	0.71070	-0.08867	0.09691	0.10008	0.01925	0.05719	0.52764
43	65	59	0	1	0	1	1	0.29676	0.10555	0.60143	1.55875	-0.15028	0.32987	0.04183	0.26448	-0.12459	2.65895
44	4	60	0	0	0	0	1	0.37100	0.14908	0.66506	-0.96293	-0.04443	-0.15768	0.07794	0.10593	0.11613	0.98523
45	29	61	0	1	0	0	1	0.30932	0.10784	0.62395	-0.86032	0.07270	-0.15548	-0.01986	-0.11808	0.05478	0.78738
46	48	65	0	1	1	0	1	0.70895	0.35026	0.91671	-1.57116	0.40560	-0.43105	-0.07048	-0.36607	-0.27791	2.80759
47	17	67	0	0	1	1	1	0.77815	0.49016	0.92752	0.70829	-0.04972	0.12152	-0.03751	-0.03561	0.05310	0.52488
48	51	68	1	0	1	1	1	0.84467	0.53631	0.96236	0.58104	-0.08489	0.10238	0.07521	0.01660	0.04190	0.35400
49	44	70	0	0	1	1	1	0.79317	0.49700	0.93704	0.68076	-0.05250	0.12278	-0.03380	-0.03250	0.05013	0.48565
50	52	74	0	0	1	1	1	0.81201	0.50546	0.94806	0.64535	-0.05521	0.12315	-0.02928	-0.02867	0.04631	0.43731

```
122       proc sort data=next1 nodupkey; by Age; run;
NOTE: There were 98 observations read from the data set WORK.NEXT1.
NOTE: 48 observations with duplicate key values were deleted.
NOTE: The data set WORK.NEXT1 has 50 observations and 17 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
123        proc print data=next1;
124           TITLE2 'Listing of one kept value for each by group from Logistic Reg';
125        run;
NOTE: There were 50 observations read from the data set WORK.NEXT1.
NOTE: The PROCEDURE PRINT printed page 2.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.21 seconds
      cpu time            0.07 seconds
126
127        options ps=56 ls=111;
128        proc sort data=Disease; by Age; run;
NOTE: There were 98 observations read from the data set WORK.DISEASE.
NOTE: The data set WORK.DISEASE has 98 observations and 6 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

```
129        proc sort data=next1; by Age; run;
NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
130        proc means data=Disease noprint; by Age; var Disease;
131             output out=next3 n=n mean=mean var=var; run;
NOTE: There were 98 observations read from the data set WORK.DISEASE.
NOTE: The data set WORK.NEXT3 has 50 observations and 6 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds
132
133        data three; set next1 next3; run;
NOTE: There were 50 observations read from the data set WORK.NEXT1.
NOTE: There were 50 observations read from the data set WORK.NEXT3.
NOTE: The data set WORK.THREE has 100 observations and 22 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
134        proc plot data=three;  plot yhat*Age='x' mean*Age='o' / overlay;
135             TITLE2 'Plot of observed means (o) and predicted values (p)';
136        run;
138           ods graphics off;
139           ods html close;
NOTE: There were 100 observations read from the data set WORK.THREE.
NOTE: The PROCEDURE PLOT printed page 3.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.24 seconds
      cpu time            0.07 seconds
```

Logistic Regression - NKNW Example 14.3
Plot of observed means (o) and predicted values (p)

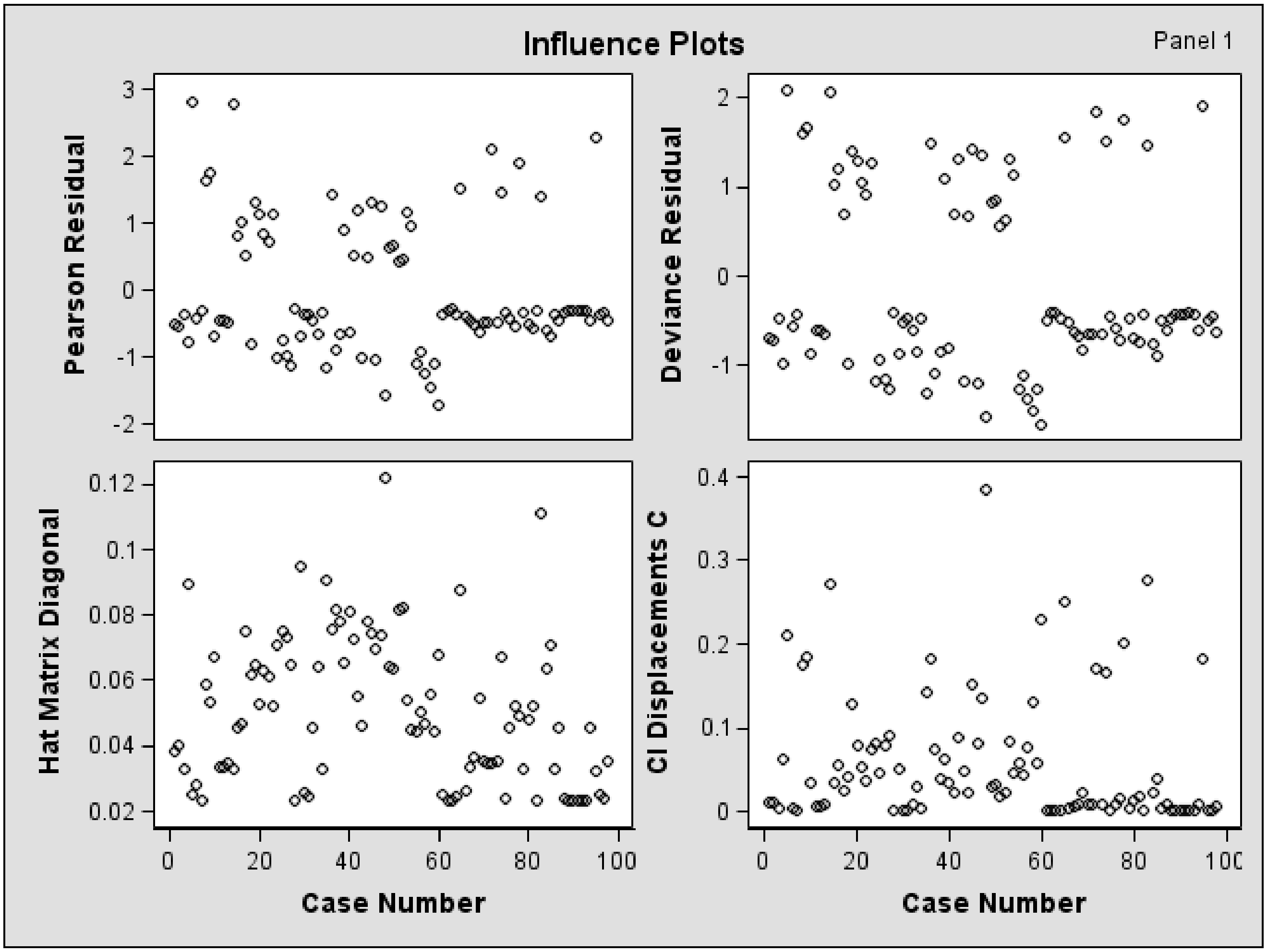


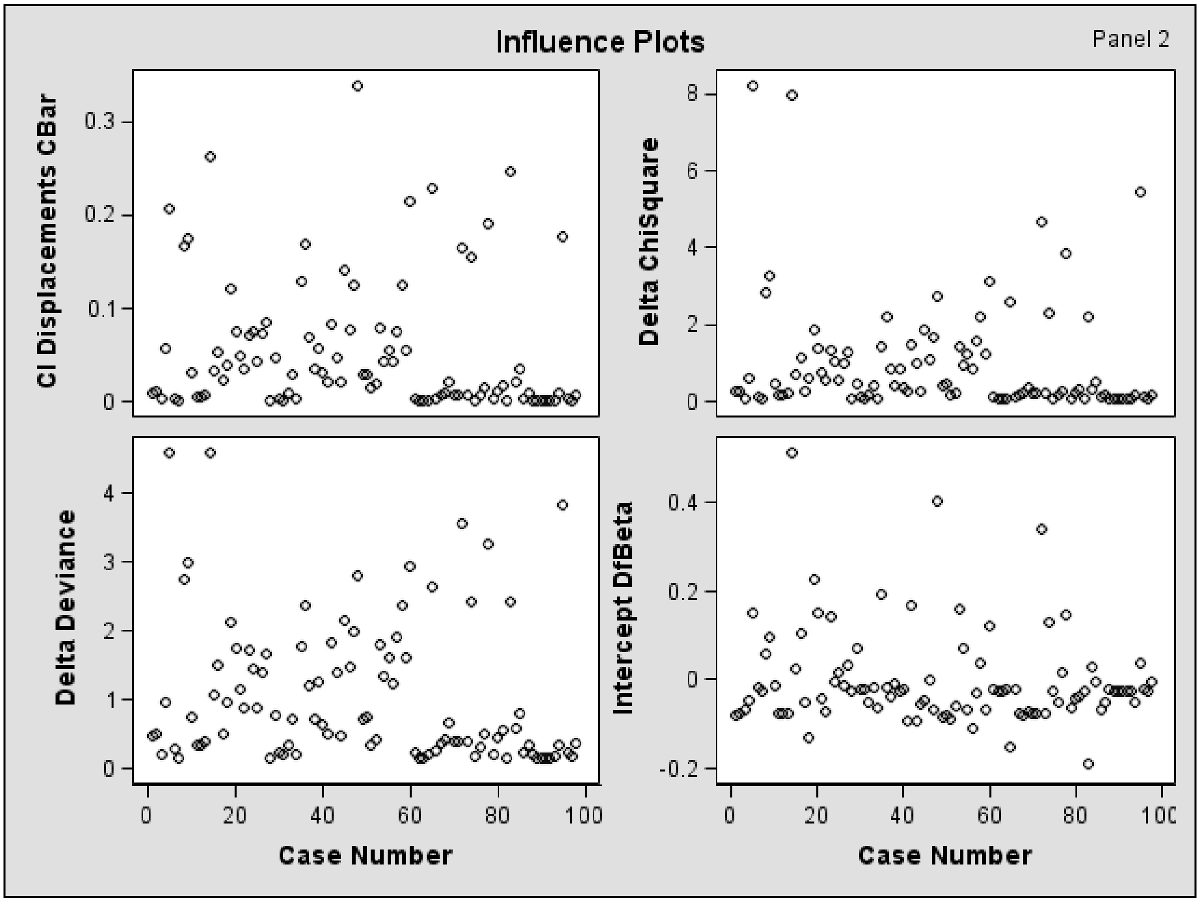
NOTE: 100 obs had missing values.

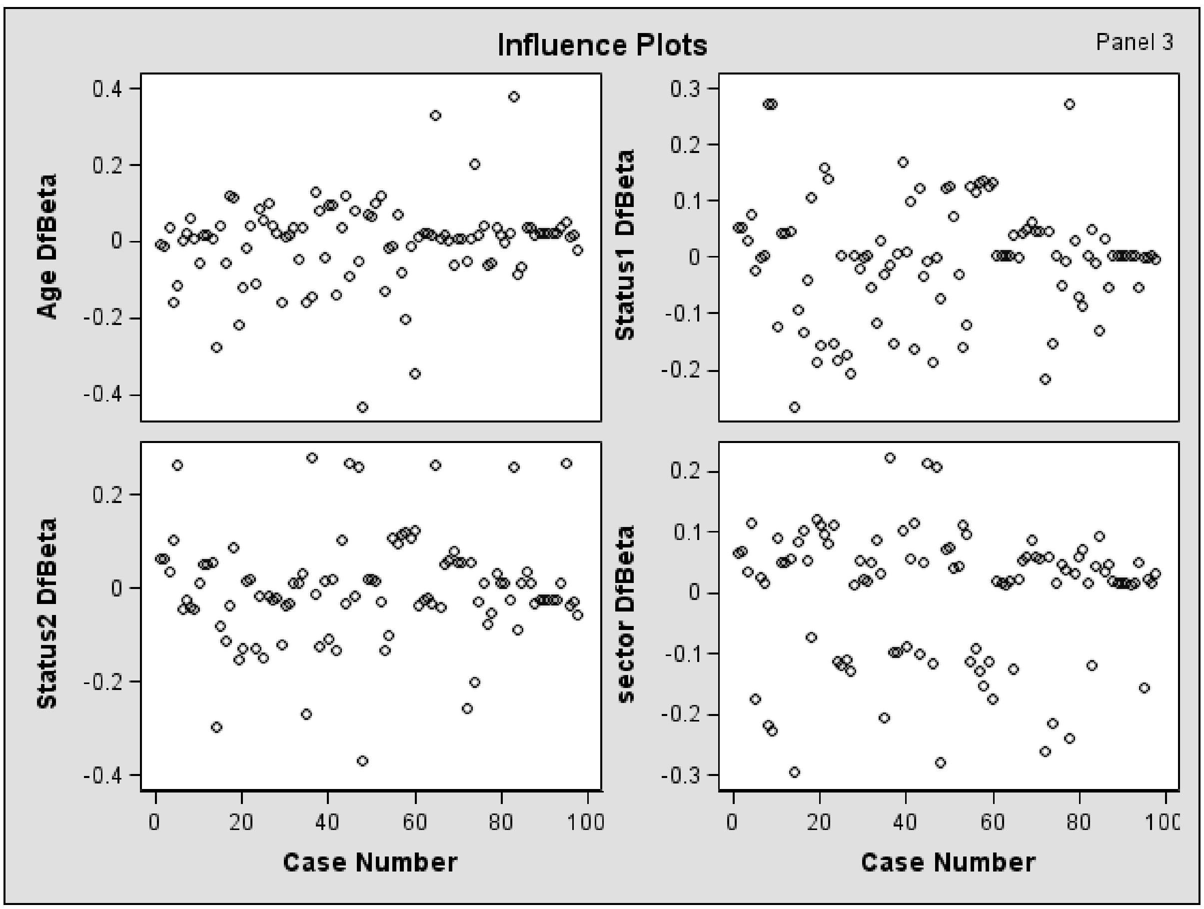
A number of new graphics are available, but are still reported as "experimental". They are activated with ODS statements as follows.

```
120            ods html;
121            ods graphics on;
```

The graphics are then places in the HTML output.




Influence Plots
Panel 2
CI Displacements CBar
Delta ChiSquare
Delta Deviance
Intercept DfBeta
Case Number
Case Number


Influence Plots
Panel 3
Age DfBeta
Status1 DfBeta
Status2 DfBeta
sector DfBeta
Case Number
Case Number