Applied Logistic Regression

Week 8

1. Homework week 7: highlights
2. R^2-type measures / numerical problems
3. Illustrating numerical problems using statistical software packages
4. The ICU study
5. Using logistic regression model for individual patient decision making I
6. Using logistic regression model for individual patient decision making II
7. Homework
8. Homework week 8: highlights

Stanley Lemeshow, Professor of Biostatistics

College of Public Health, The Ohio State University

THE OHIO STATE UNIVERSITY

WEEK 8: R^2-TYPE MEASURES

- You may, on occasion, see a statistic which is called R^2.
- This statistic has been formulated to be analogous to the R^2 from linear regression
- It is formulated as a % decrease in the likelihood relative to a saturated model. By definition:

$$R_L^2 = 100 \times \frac{\left(L_0 - L_p\right)}{\left(L_0 - L_s\right)}$$

where L_k = log-likelihood for model k, $k = 0, p, s$.

In terms of the deviance, $D_k = 2\left(L_k - L_s\right)$, R_L^2 is

$$R_L^2 = 1 - \frac{L_p}{L_0} = 1 - \frac{D_p}{D_0}$$

This is pseudo R^2 in STATA

As can be seen from the first expression $\left(\text{when } J = n, L_s = 0\right)$, R_L^2 is a measure of the Type I likelihood ratio test $\left(\text{model } (0) \textit{ vs } \text{model}(\text{p})\right)$.

*R^2 measures are not measures of model adequacy, since they do not compare the fitted model to the saturated model.

Other measures have been suggested that correspond better to the R^2 in linear regression. e.g.,

$$r^2 = \frac{\left[\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(\hat{\pi}_i - \bar{\pi}\right)\right]^2}{\left[\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right]\left[\sum_{i=1}^{n}\left(\hat{\pi}_i - \bar{\pi}\right)^2\right]} \quad \text{or} \quad R_{SS}^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{\pi}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

where $\bar{y} = \bar{\pi} = \frac{n_1}{n}$

One problem that may occur when using (some) logistic regression software is that the program produces results when it should not. The source of these problems can be:

- A "zero" cell
- Complete separation
- Colinearity

In each instance the tip-off that something is wrong is the presence, in the output, of one or more large estimated coefficients accompanied by a very large estimated standard error.

This combination should serve as an automatic flag to the user that there are numerical problems in the data.

1. A Zero Cell

Consider fitting a logistic regression model to the following contingency table:

	X			
Y	1	2	3	TOTAL
1	7	12	20	39
0	13	8	0	21
Total	20	20	20	60

The results of the fit are as follows:

STATISTIX OUTPUT

```
PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR        COEF/SE          P
---------     -----------     ---------        -------      ------
CONSTANT        -0.61904        0.46881         -1.32       0.1867
X1               1.02450        0.65430          1.57       0.1174
X2               10.1850        16.2100          0.63       0.5298

DEVIANCE                  52.82
P-VALUE                  0.0000
DEGREES OF FREEDOM            2

CASES INCLUDED 5   MISSING CASES 1
```

* Note: No Warning!!!

JMP OUTPUT

Response: y

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept		0.61903921	0.4688072	1.74	0.1867
x1		-1.0245043	0.6543039	2.45	0.1174
x2	Unstable	-12.821926	99.842438	0.02	0.8978

```
. logit y i.x [w=freq]
(frequency weights assumed)

note: 3.x != 0 predicts success perfectly
      3.x dropped and 1 obs not used

Iteration 0:   log likelihood = -27.675866
Iteration 1:   log likelihood = -26.409925
Iteration 2:   log likelihood = -26.409166
Iteration 3:   log likelihood = -26.409166

Logistic regression                                     Number of obs   =         40
                                                        LR chi2(1)      =       2.53
                                                        Prob > chi2     =     0.1115
Log likelihood = -26.409166                             Pseudo R2       =     0.0458

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |
          2  |    1.024504   .6543039     1.57   0.117    -.2579077    2.306916
          3  |   (empty)
             |
       _cons |   -.6190392   .4688072    -1.32   0.187    -1.537884    .2998061
------------------------------------------------------------------------------

. logistic y i.x [w=freq]
Logistic regression                                     Number of obs   =         40
                                                        LR chi2(1)      =       2.53
                                                        Prob > chi2     =     0.1115
Log likelihood = -26.409166                             Pseudo R2       =     0.0458

------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |
          2  |    2.785714   1.822704     1.57   0.117     .7726665    10.04341
          3  |   (empty)
------------------------------------------------------------------------------
```

STATA OUTPUT

Dropped x = 3 data

SPSS OUTPUT

---------------------- Variables in the Equation ----------------------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X1	1.0245	.6543	2.4517	1	.1174	.0762	2.7857
X2	10.8219	36.7328	.0868	1	.7683	.0000	50104.96
Constant	-.6190	.4688	1.7436	1	.1867		

>Warning # 3211
>On at least one case, the value of the weight variable was zero, negative, or
>missing. Such cases are considered missing by statistical procedures which
>need positively weighted cases, but remain on the file and are processed by
>non-statistical facilities such as LIST and SAVE.

SYSTAT OUTPUT

```
L-L AT ITER      1 IS       -41.589
L-L AT ITER      2 IS       -28.949
L-L AT ITER      3 IS       -27.260
:
L-L AT ITER     14 IS       -26.409
L-L AT ITER     15 IS       -26.409

ITERATION LIMIT OF 15 EXCEEDED.
FAILED TO SATISFY CHANGE TOLERANCE.
MAXIMUM RELATIVE PARAMETER CHANGE ELEMENT:         0.059

RESULTS OF ESTIMATION
=====================

LOG LIKELIHOOD: -26.409

   PARAMETER                    ESTIMATE         S.E.      T-RATIO      P-VALUE
-------------------------------------------------------------------------------
 1 CONSTANT                 |      -0.619        0.469       -1.320        0.187
 2 X1                       |       1.025        0.654        1.566        0.117
 3 X2                       |      16.822      447.458        0.038        0.970
-------------------------------------------------------------------------------

                                                  95.0% BOUNDS
   PARAMETER                  ODDS RATIO        UPPER         LOWER
-----------------------------------------------------------------
 2 X1                       |       2.786       10.043        0.773
 3 X2                       |  .202150E+08         .          0.000
-----------------------------------------------------------------
LOG LIKELIHOOD OF CONSTANTS ONLY MODEL = LL(0) = -38.847
2*[LL(N)-LL(0)] = 24.875 WITH 2 DOF, CHI-SQ P-VALUE = 0.000
MCFADDEN'S RHO-SQUARED = 0.320
-------------------------------------------------------------------------------
```

EGRET OUTPUT

```
WARNING:  * means coeff increments halved

 --[not maximized]-----------------RESULTS-----------------------------[LR]--
 OUTCOME= y
 TERM                          COEFFICIENT     STD ERROR     P-VALUE     ODDS RATIO

%GM                              -.6190          ?             ?           .5385
x1                               1.025           ?             ?          2.786
x2                               33.89           ?             ?          .5228E+15

                  DEVIANCE ON    57 DF =          .000
  LIKELIHOOD RATIO STATISTIC ON   3 DF =        83.178, p < .001
```

NOTE:

LARGE COEFFICIENT + LARGE STANDARD ERROR = LARGE TROUBLE

The problem of a zero cell is most likely to occur when an Interaction term between two categorical variables is added to the model.

Consider the following example:

		Z					
		1		2		3	
	X:	1	0	1	0	1	0
Y	1	5	2	10	2	15	1
	0	5	8	2	6	0	4
Total		10	10	12	8	15	5

We now fit logistic regression models fit to these data:

Variable			SYSTAT		SPSS		STATISTIX		EGRET	
	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$
X	2.77	.72	1.39	1.01	1.39	1.01	1.39	1.01	1.39	?
Z_1	1.19	.81	0.29	1.14	0.29	1.14	0.29	1.14	0.29	?
Z_2	2.04	.89	0.00	1.37	0.00	1.37	0.00	1.37	0.00	?
$X \times Z_1$			1.32	1.51	1.32	1.51	1.32	1.51	1.32	?
$X \times Z_2$			16.20	516.69	10.20	42.44	8.57	11.45	27.61	?
constant	-2.32	.77	-1.39	0.79	-1.39	0.79	-1.39	0.79	-1.39	?

same for all programs

warnings provided

warnings provided

no warning

warnings provided

numerically stable main effects model

when interactions are added the program produces garbage due to the 0 cell

Solution to the zero cell problem:

(1) Find the zero cell by forming an $x \times y \times z$ table (as above)

(2) Collapse categories in a biologically meaningful way

OR (3) Throw the category out

2. Complete Separation

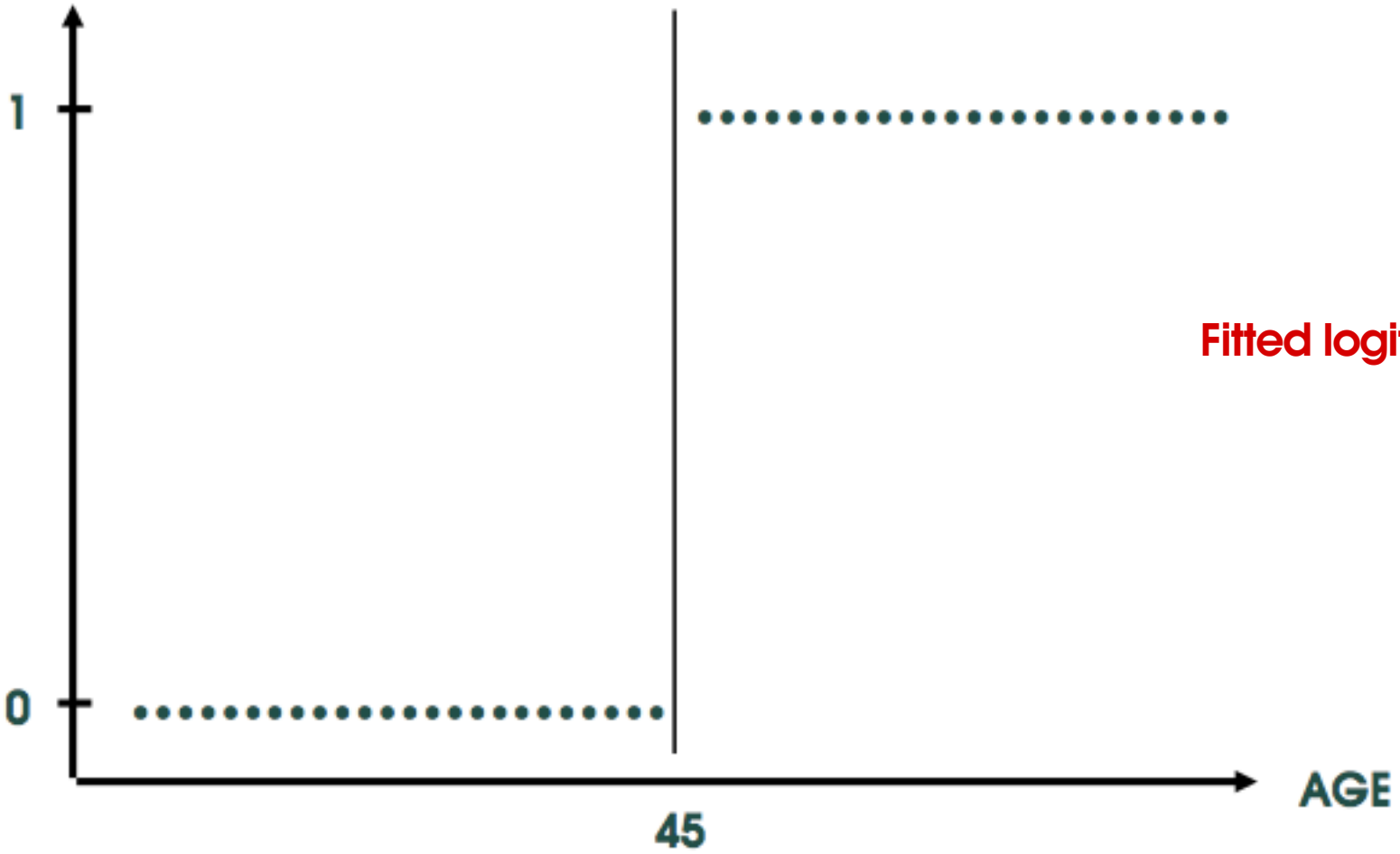
This problem also manifests itself with large estimates and standard errors and is most likely to occur in small samples.

- **In multivariable data it will be difficult to find the exact source.**
- **The essential idea is that all subjects with $y = 1$ may be separated from all subjects with $y = 0$ by their data.**

i.e., for subjects with $y = 1$, $\hat{\pi} = 1$

for subjects with $y = 0$, $\hat{\pi} = 0$

This can be shown pictorially as follows:



Fitted logit has $\hat{\beta} = \infty$

Here we see that, for $AGE < 45$, $y = 0$

for $AGE \geq 45$, $y = 1$

Obviously, the chance of getting a set of data where this would happen depends on n .

The above idea can be extended to more than 1 independent variable and we need not have the two regions separated by a plane.

Solution to Complete Separation:

- Plots
- Delete variables
- Collect more data

3. Colinearity in the x's

As in any regression model, the independent variables may be so highly correlated that $\underline{x}$, the design matrix, is nearly singular.

Unfortunately, some logistic regression software will run when this is the case.

The colinear variables will have large estimated coefficients, usually of opposite sign, and very large estimated standard errors.

When a variable is nearly constant the estimated coefficient may not be large enough to cause concern, but the estimated standard errors will be.

Example

$X_3 = C + \varepsilon$ ← error, constant

A sample of size n = 10 was selected from a population where x_1 and x_2 are highly related and x_3 and x_0 are highly related.

SEQUENCE OF MODELS

Variable	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$
x_1	1.4	1.0	502.1	444.3			500.2	446.5
x_2			−500.4	443.8			−498.6	446.0
x_3					1.59	19.9	−1.3	28.5
Constant	−1.0	0.83	1.28	1.98	−2.5	21.0	2.7	29.8

note very large $\hat{\beta}$ and $\widehat{SE}$ among colinear variables

Solution to colinearity

- Use the same methods suggested for use in linear regression.
- Get the usual diagnostics for colinearity among the x's by running a linear regression package
- Other possibilities include ridge regression type methods. These have been proposed but have not been studied that closely.

"While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to."

A. Conan Doyle

Sherlock Holmes: The Sign of Four

There were two data sets available.

- In Data Set I, consecutive admissions to the medical and surgical intensive care units (ICUs) in four hospitals were studied.
- Burn, coronary care, cardiac surgery, and patients < 18 years of age were excluded.

The four hospitals were:

Baystate Medical Center (Springfield, MA)

U. of Massachusetts Medical Center (Worcester, MA)

Ellis Hospital (Schenectady, NY)

Albany Medical Center (Albany, NY)

The following tables show the mortality rates in each hospital.

Data Set I - Developmental Part

(April 17, 1989 - July 31, 1990)

Hospital	N	N(Lived)	%	N(Died)	%
BMC	1641	1315	80.1	326	19.9
UMMC	1377	1128	81.9	249	18.1
AMC	802	682	85.0	120	15.0
EH	404	310	76.7	94	23.3
Total	4224	3435	81.3	789	18.7

Data Set I - Validation Part

(September 1, 1990 - May 10, 1991)

Hospital	N	N(Lived)	%	N(Died)	%
BMC	846	691	81.7	155	18.3
UMMC	670	542	80.9	128	19.1
AMC	550	450	81.8	100	18.2
Total	2066	1683	81.5	383	18.5

Detailed information was collected on each patient at ICU admission; at 24, 48, and 72 hours in the ICU; at ICU discharge; and at hospital discharge.

Variables collected dealt with each patient's condition, treatment, and vital status.

There were 4252 eligible patients in the developmental data set, of which 27 had to be eliminated because of missing medical records or missing values in key variables, and one was still in the hospital when data collection ended.

- This left 4224 patients for developing multiple logistic regression models.

In Data Set II, consecutive admissions to medical and surgical ICUs in 137 hospitals in 12 countries were studied.

- Patients were enrolled in the study between 30 Sept and 27 Dec, 1991 and followed in the hospital until 28 Feb, 1992.
- Any patients remaining in the hospital on 28 Feb, 1992 were dropped from the study.

As in Data Set I, burn, coronary care, cardiac surgery, and patients <18 years of age were excluded.

The number of participating ICUs, number of patients, and mortality rates for each country are given in the following table.

Data Set II

Country	Number of ICUs	N	N(Lived)	%	N(Died)	%
Belgium	11	1091	854	78.3	237	21.7
England	4	136	92	67.6	44	32.4
Finland	7	720	593	82.4	127	17.6
France	14	1393	990	71.1	403	28.9
Germany/Austria	15 *	1807	1523	84.3	284	15.7
Holland	11	939	751	80.0	188	20.0
Italy	20	1297	891	68.7	406	31.3
Spain	17	1270	926	72.9	344	27.1
Switzerland	11	756	652	86.2	104	13.8
USA/Canada	27 **	3732	2998	80.3	734	19.7
Total	137	13141	10270	78.2	2871	21.8

* 14 Germany, 1 Austria

** 25 USA, 2 Canada

65% of these patients were selected at random to be the Developmental part of the data set and the remaining 35% became the validation part of the data set.

Variables and Coefficients in the MPM_0 ($n = 12610$)

VARIABLE	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	$\widehat{OR}$	95% CI
Constant	-5.46836			
Physiology				
Coma or deep stupor at admission	1.48592	0.07877	4.42	3.79-5.16
Systolic blood pressure ≤ 90 mmHg	1.06127	0.07877	2.89	2.48-3.37
Mechanical ventilation at admission	0.79105	0.05625	2.21	1.98-2.46
CPR prior to admission	0.56995	0.11228	1.77	1.42-2.20
Heart rate ≥ 150 beats/minute	0.45603	0.14469	1.58	1.19-2.10
Chronic Diagnoses				
Metastatic neoplasm	1.19979	0.09843	3.32	2.74-4.03
Cirrhosis	1.13681	0.12648	3.12	2.43-3.99
Chronic renal failure, no acute exacerbation	0.91906	0.1047	2.51	2.04-3.08
Acute Diagnoses				
Acute renal failure, with or without chronic history	1.4821	0.08854	4.4	3.70-5.24
Intracranial mass effect	0.86533	0.08838	2.38	2.00-2.82
Gastrointestinal bleed	0.39653	0.09444	1.49	1.24-1.79
Cardiac dysrhythmia	0.28095	0.06758	1.32	1.16-1.51
Cerebrovascular incident	0.21338	0.08854	1.24	1.04-1.47
Other				
Admission not for elective surgery	1.19098	0.07362	3.29	2.85-3.80
Age	0.03057	0.00161		
(10-year OR)			1.36	1.32-1.40

Goodness-of-Fit Tests for the MPM_0 - Developmental Data Set

Pr(Dying)	Survived		Died	
	Observed	Expected	Observed	Expected
.000-.031	1196	1185.2	16	26.8
>.031-.045	1259	1259.2	49	48.8
>.045-.063	1180	1177.9	66	68.1
>.063-.086	1182	1180.2	93	94.8
>.086-.117	1140	1134.6	121	126.4
>.117-.161	1080	1087.9	182	174.1
>.161-.224	1011	1017.9	249	242.1
>.224-.338	916	918.1	348	345.9
>.338-.572	691	702.6	568	556.4
>.572-1.00	323	314.3	940	948.7
Total	9978		2632	

$$\hat{C} = 6.21 \qquad df = 8 \qquad p - value = 0.623$$

$$\hat{C} = \sum_{\text{all 20 cells}} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Pr(Dying)	Survived		Died	
	Observed	Expected	Observed	Expected
.00-.10	5430	5412.0	283	301.0
>.10-.20	2244	2256.3	392	379.7
>.20-.30	1076	1075.9	345	345.1
>.30-.40	493	502.0	276	267.0
>.40-.50	261	266.3	222	216.7
>.50-.60	188	194.8	242	235.2
>.60-.70	135	133.8	246	247.2
>.70-.80	93	87.1	257	262.9
>.80-.90	42	39.3	213	215.7
>.90-1.00	16	10.4	156	161.6
Total	9978		2632	

$$\hat{H} = 6.65 \qquad df = 8 \qquad p - value = 0.575$$

$$\hat{H} = \sum_{\text{all 20 cells}} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Area under ROC Curve: 0.837

Goodness-of-Fit Test for the MPM_0 - Validation Data Set (n = 6514)

Pr(Dying)	Survived		Died	
	Observed	Expected	Observed	Expected
.000-.031	644	641.0	12	15.0
>.031-.048	618	617.4	24	24.6
>.048-.065	611	613.7	39	36.3
>.065-.087	626	619.3	44	50.7
>.087-.117	563	575.0	76	64.0
>.117-.161	552	557.2	94	88.8
>.161-.233	535	533.3	126	127.7
>.233-.342	453	464.6	191	179.4
>.342-.558	372	367.0	282	287.0
>.558-1.00	197	170.6	455	481.4
Total	5171		1343	

$\hat{C} = 11.41 \quad df = 10 \quad p - value = 0.327$

Area under ROC Curve: 0.824

The following is an example of how you would use the MPM_0 to calculate, at the time of ICU admission, a patient's probability of dying in the hospital.

A *60-year-old* man with alcohol-induced *cirrhosis* is an *emergency admission* to an ICU. He has suffered a significant *GI bleed* resulting in a drop in his systolic blood pressure to *below 90 mmHg*. At the time of ICU admission his heart rate is 130 beats/minute.

A physician, nurse, or data collector would answer the following questions on a microcomputer:

Physiology:	
Coma or deep stupor at admission	no
Systolic blood pressure ≤ 90	yes
Mechanical ventilation	no
CPR prior to admission	no
Heart rate > 150	no
Chronic Diagnoses:	
Metastatic neoplasm	no
Cirrhosis	yes
Chronic renal failure, no acute exacerbation	no
Acute Diagnoses:	
Acute renal failure, with or without chronic	no
Intracranial mass effect	no
Gastrointestinal bleed	yes
Cardiac dysrhythmia	no
Cerebrovascular incident	no
Other:	
Admission not for elective surgery	yes
Age	60

A microcomputer program would then convert these answers to a probability by multiplying the responses and corresponding coefficients.

VARIABLE	$\hat{\beta}$	x	$\hat{\beta}x$
Constant	-5.46836		-5.46836
Coma or deep stupor at admission	1.48592	0	0
Systolic blood pressure ≤ 90	1.06127	1	1.06127
Mechanical ventilation	0.79105	0	0
CPR prior to admission	0.56995	0	0
Heart rate > 150	0.45603	0	0
Metastatic neoplasm	1.19979	0	0
Cirrhosis	1.13681	1	1.13681
Chronic renal failure, no acute exacerbation	0.91906	0	0
Acute renal failure, with or without chronic	1.48210	0	0
Intracranial mass effect	0.86533	0	0
Gastrointestinal bleed	0.39653	1	0.39653
Cardiac dysrhythmia	0.28095	0	0
Cerebrovascular incident	0.21338	0	0
Admission not for elective surgery	1.19098	1	1.19098
Age	0.03057	60	1.83420
	LOGIT:		0.15143

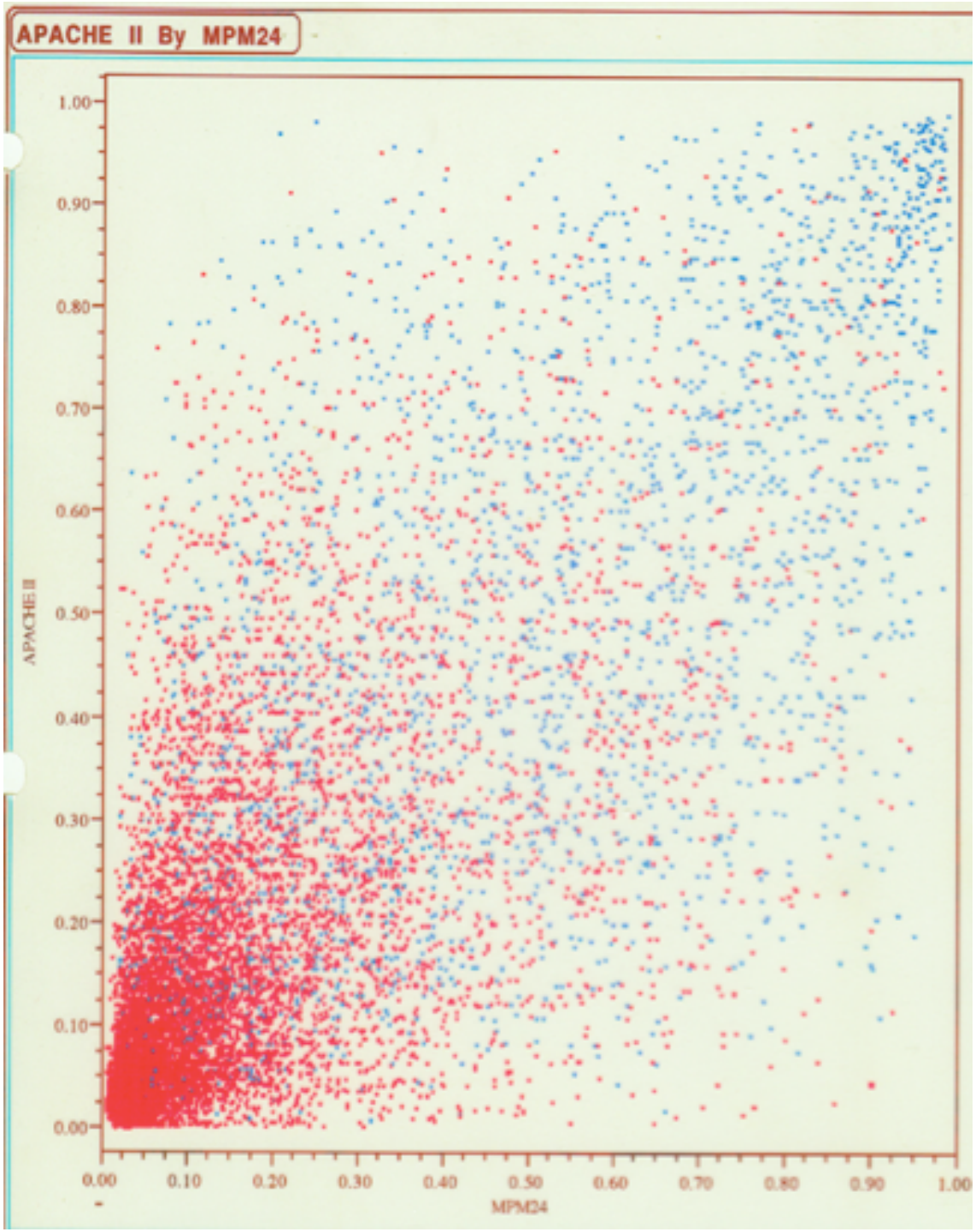
The probability of hospital mortality is computed by first calculating the logit, defined as:

$$\text{logit} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Then,

$$\text{Probability of hospital mortality} = \frac{e^{\text{logit}}}{1 + e^{\text{logit}}} = 0.54$$

i.e., this patient has a 54% probability of not surviving to hospital discharge.


APACHE II By MPM24
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
APACHE II
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
MPM24