Logistic Regression Analysis of the Occurrence of Diabetes in Pima Indian Women

Instructor: Dr.Javier Cabrera

Venkata Sasikiran Goteti, Michael Pannucci,
and Juan Zhang

March 24, 2004

PRG:	Number of times pregnant
PLASMA:	Plasma glucose concentration in saliva
BP:	Diastolic blood pressure
THICK:	Triceps skin fold thickness
INSULIN:	Two hours serum insulin
BODY:	Body mass index (Weight/Height)
PEDIGREE:	Diabetes pedigree function
AGE:	In years
RESPONSE:	1:Diabetes, 0:Not

Goal: Multiple logistic regression of the response on the eight co-variates along with any two-way interactions

> For reasons that have plagued scientists for years, many Pima Indian women have diabetes.

> NIH researchers have been trying to investigate this through genetic research.

> This report will focus on a multiple logistic regression model fit to a data set containing 768 observations.

> In order to model this data set, we must employ a method of variable selection.

> With eight first-order variables and 28 possible second-order variables, step-wise selection is a good preliminary selection doctrine.

> The forward algorithm did not yield useful results, and the backward and both algorithms failed to converge.

> A more traditional approach involving tests of significance and the analysis of deviance were conducted to determine which variables should be included in the model.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.890e+01	3.913e+00	-4.830	1.37e-06	***
PRG	1.022e+00	2.682e-01	3.809	0.000139	***
PLASMA	1.098e-01	3.034e-02	3.619	0.000296	***
BP	-1.688e-02	3.870e-02	-0.436	0.662680	
THICK	-8.304e-03	5.886e-02	-0.141	0.887817	
INSULIN	-1.139e-02	9.465e-03	-1.203	0.229000	
BODY	2.375e-01	9.913e-02	2.396	0.016589	*
PEDIGREE	5.295e+00	2.196e+00	2.411	0.015893	*
AGE	5.056e-02	8.446e-02	0.599	0.549373	
PRG:PLASMA	-1.460e-03	1.264e-03	-1.155	0.248197	
PRG:BP	-1.545e-03	1.763e-03	-0.876	0.380773	
PRG:THICK	1.304e-03	2.236e-03	0.583	0.559800	
PRG:INSULIN	2.772e-05	4.291e-04	0.065	0.948500	
PRG:BODY	-9.253e-03	5.062e-03	-1.828	0.067543	.
PRG:PEDIGREE	1.746e-01	1.065e-01	1.639	0.101132	
PRG:AGE	-1.057e-02	3.115e-03	-3.394	0.000690	***
PLASMA:BP	-2.810e-04	2.501e-04	-1.124	0.261098	
PLASMA:THICK	-2.285e-04	2.549e-04	-0.896	0.370064	

...

AIC: 735.44

	DF	Deviance		Resid. Df	Resid. Dev
NULL				767	993.48
PRG	1	37.27	*	766	956.21
PLASMA	1	171.26	*	765	784.95
BP	1	0.89		764	784.06
THICK	1	4.00	*	763	780.06
INSULIN	1	1.97		762	778.09
BODY	1	41.24	*	761	736.85
PEDIGREE	1	10.88	*	760	725.97
AGE	1	2.52		759	723.45
PRG:PLASMA	1	2.63		758	720.82
PRG:BP	1	0.27		757	720.55
PRG:THICK	1	0.75		756	719.79
PRG:INSULIN	1	3.59		755	716.21
PRG:BODY	1	1.93		754	714.28
PRG:PEDIGREE	1	0.85		753	713.43
PRG:AGE	1	8.74	*	752	704.69
PLASMA:BP	1	2.51		751	702.17
PLASMA:THICK	1	0.88		750	701.30
...					

These statistics are known to follow a chi-square(1) distribution. Thus, if we were to select a variable based on, say, a 5% significance level, the relevant critical value is 3.841. Variables with a * next to their deviance value and/or are statistically significant at the 5% level will thus be used in our model.

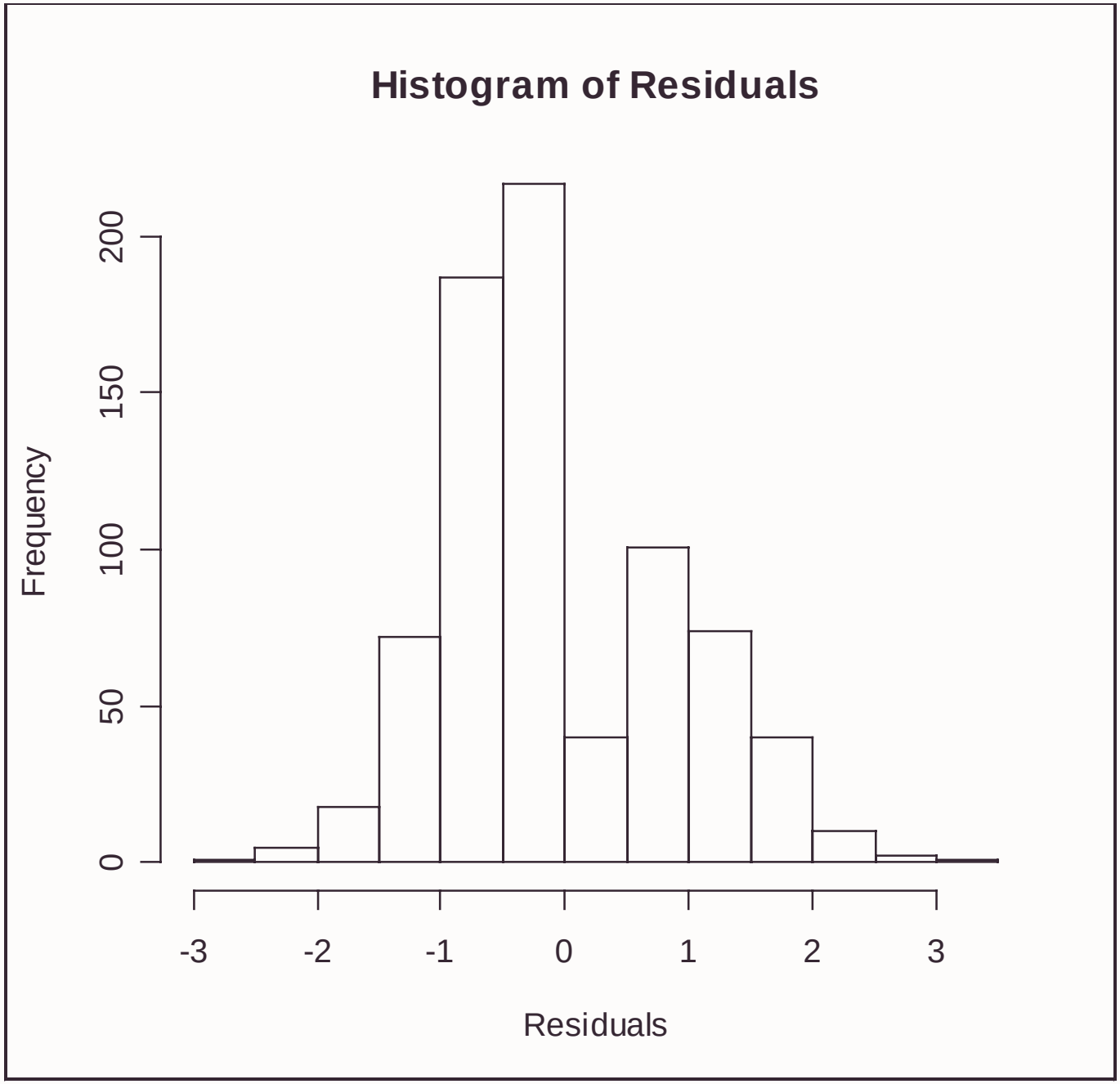
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.348e+01	2.175e+00	-6.197	5.74e-10	***
PRG	7.327e-01	1.839e-01	3.984	6.76e-05	***
PLASMA	8.314e-02	1.400e-02	5.939	2.86e-09	***
BP	-4.923e-02	1.851e-02	-2.659	0.007828	**
THICK	-2.375e-02	1.275e-02	-1.863	0.062398	.
INSULIN	-8.102e-03	3.022e-03	-2.681	0.007349	**
BODY	1.215e-01	2.316e-02	5.247	1.55e-07	***
PEDIGREE	2.629e+00	1.545e+00	1.702	0.088742	.
AGE	8.835e-02	5.189e-02	1.703	0.088621	.
PRG:BODY	-7.101e-03	3.820e-03	-1.859	0.063017	.
PRG:AGE	-9.950e-03	2.915e-03	-3.413	0.000642	***
PLASMA:PEDIGREE	-1.893e-02	1.140e-02	-1.661	0.096715	.
PLASMA:AGE	-1.003e-03	2.991e-04	-3.353	0.000801	***
BP:AGE	1.105e-03	5.336e-04	2.071	0.038340	*
THICK:PEDIGREE	5.340e-02	2.221e-02	2.404	0.016197	*
INSULIN:PEDIGREE	-3.429e-03	2.331e-03	-1.471	0.141392	
INSULIN:AGE	2.502e-04	7.418e-05	3.373	0.000743	***

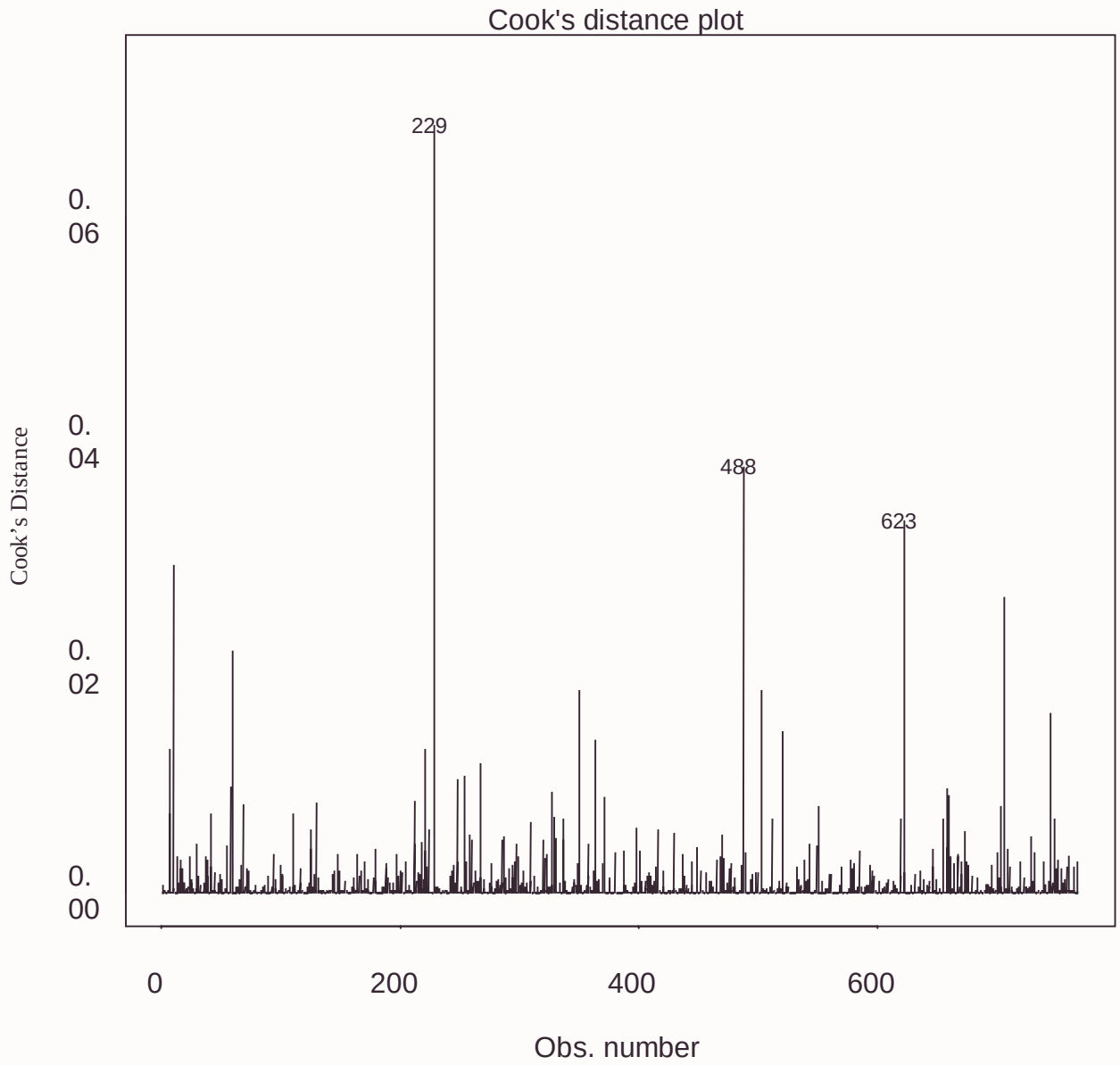
Null deviance: 993.48 on 767 degrees of freedom
Residual deviance: 675.57 on 751 degrees of freedom
AIC: 709.57

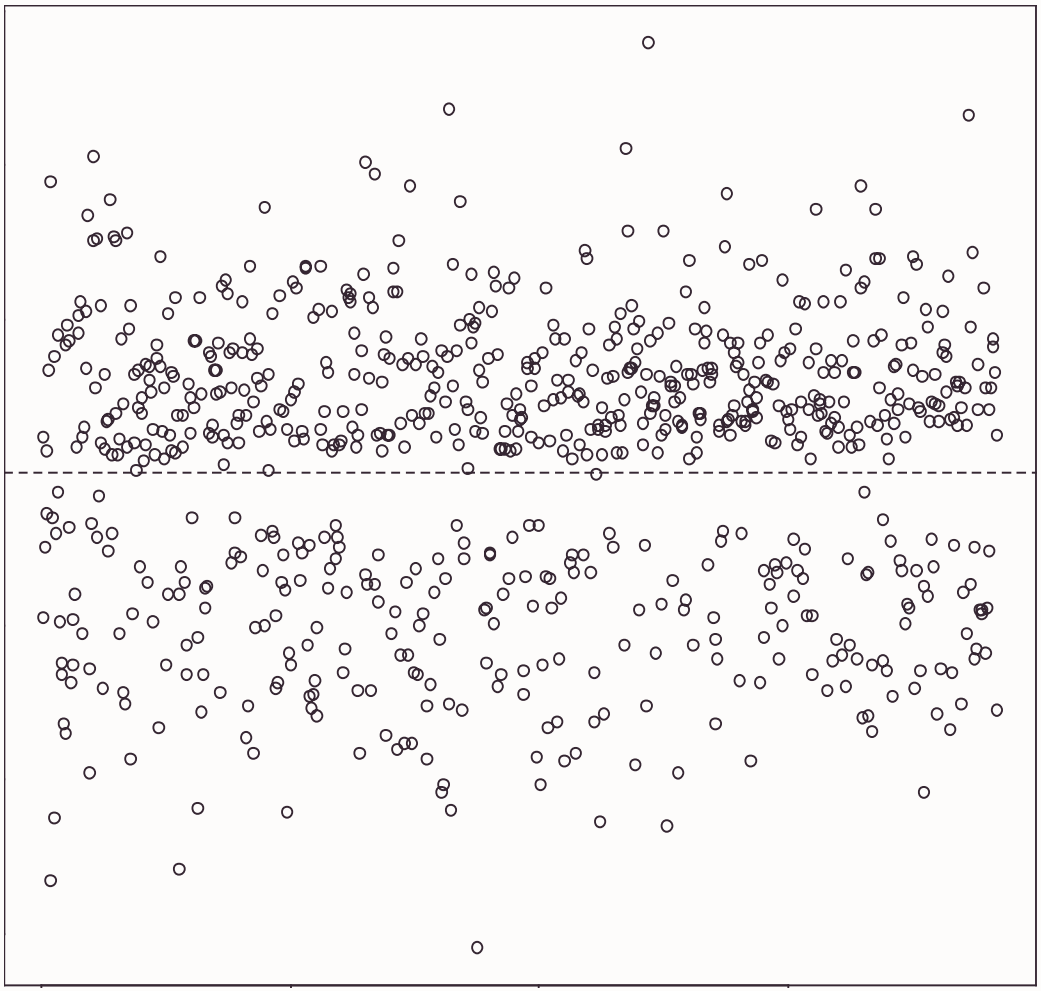
The following illustrates, in descending order of log-odds, the strongest factors in terms of increased likelihood of diabetes and decreased likelihood of diabetes. The order is determined by the relative magnitude of the variable's *z*-score.

Increased Log-Odds: PLASMA, BODY, PRG, INSULIN*AGE, THICK*PEDIGREE, BP*AGE, AGE, PEDIGREE

Decreased Log-Odds: PRG*AGE, PLASMA*AGE, INSULIN, BP, THICK, PRG*BODY, PLASMA*PEDIGREE, INSULIN*PEDIGREE


Histogram of Residuals
Frequency
0
50
100
150
200
-3
-2
-1
0
1
2
3
Residuals


Cook's distance plot
229
488
623
0.
06
0.
04
0.
02
0.
00
Cook's Distance
0
200
400
600
Obs. number



Deviance Residual Plot

> In terms of prediction, this model predicts the log-odds that someone has diabetes.

> If we invert the logistic predictions to probabilities and use the cut-off 44%, where observations with a predicted response greater than 44% are classified as having diabetes and observations with a predicted response below 44% are healthy.

> The cut-off point was chosen through trial-and-error and resulted in 87 false-positives and 82 false-negatives. Thus, there were 181 correct predictions of diabetes (out of 268) and 418 correct predictions of non-diabetes (out of 500).

> So, there is a 77.99% success rate for detecting diabetes correctly.

References:

http://diabetes.niddk.nih.gov/dm/pubs/pima/pathfind/pathfind.htm

(Title: The Pima Indians : Pathfinders for health)

http://diabetes.niddk.nih.gov/dm/pubs/pima/genetic/genetic.htm

(Title: The Pima Indians and genetic research)

Thank You