Example 2 – Subset Selection

This section presents an example of how to conduct a subset selection. The data used are stored in the Leukemia dataset. This analysis will search for the best model from among a pool of the six numeric variables.

Setup

To run this example, complete the following steps:

1 Open the Leukemia example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Leukemia** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Remiss**
Numeric X's	**Cell, Smear, Infil, LI, Blast, Temp**
Terms	**1-Way**
Subset Selection Tab	
Search for the Best Subset from the X's	**Checked**
Search Method	**Hierarchical Forward Selection**
Stop search when number of terms reaches	**6**
Reports Tab	
Run Summary	**Checked**
Subset Summary	**Checked**
Subset Detail	**Checked**
Coefficient Significance Tests	**Checked**
All Other Reports	**Unchecked**
Plots Tab	
All Plots	**Unchecked**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Item	Value
Y Variable	Remiss	Rows Processed	29
Reference Value	0	Rows Used	27
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	2
Numeric X Variables	6	Rows Freq Miss. or 0	0
Categorical X Variables	0	Rows Prediction Only	0
Final Log Likelihood	-10.87752	Unique Rows (Y and X's)	27
Model R^2	0.36707	Sum of Frequencies	27
Actual Convergence	2.081623E-06	Likelihood Iterations	9
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	6	Completion Status	Quasi-Separation
Priors	Equal		
Subset Selection Method	Hierarchical Forward Selection		

******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********
Your dataset had QUASI-COMPLETE SEPARATION which means that the maximum likelihood routine did NOT converge so the statistical tests are not valid. Although the prediction equations correctly classified much of your data, they may not do so for other observations. Quasi-Complete Separation often occurs because your sample size is too small.
******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********

The first thing we notice is the warning message about quasi-separation. If quasi-separation occurs, the maximum likelihood estimates do not exist and all results are suspect. We note that 9 likelihood iterations occurred and the Actual Convergence is near the Target Convergence. We decide to rerun the analysis after resetting the Max Terms in Subset box from 6 to 5. Note that this error message often occurs when a small set of data is fit with a model with too many terms.

At this point, reset the value for **Stop search when number of terms reaches** (on the Subset Selection tab) to **5** manually or load the template **Example2b**. Now, rerun the analysis.

Run Summary

Run Summary

Item	Value	Item	Value
Y Variable	Remiss	Rows Processed	29
Reference Value	0	Rows Used	27
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	2
Numeric X Variables	6	Rows Freq Miss. or 0	0
Categorical X Variables	0	Rows Prediction Only	0
Final Log Likelihood	-10.92900	Unique Rows (Y and X's)	27
Model R^2	0.36407	Sum of Frequencies	27
Actual Convergence	7.136538E-07	Likelihood Iterations	7
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	5	Completion Status	Normal Completion
Priors	Equal		
Subset Selection Method	Hierarchical Forward Selection		

The warning message has disappeared and the algorithm finished normally.

Subset Selection Summary

Subset Selection Summary
Subset Selection Method = Hierarchical Forward Selection

No. Terms	No. X's	Log Likelihood	R² Value	R² Change
1	1	-17.18588	0.00000	0.00000
2	2	-13.03648	0.24144	0.24144
3	3	-12.17036	0.29184	0.05040
4	4	-10.97669	0.36130	0.06946
5	5	-10.92900	0.36407	0.00277

This report shows the best log-likelihood value for each subset size. In this example, it appears that four terms (the intercept and three variables) provides the best model. Note that adding the fifth variable does not increase the R-squared value very much.

No. Terms

The number of terms. Note that this includes the intercept.

No. X's

The number of *X*'s that were included in the model. Note that in this case, the number of terms matches the number of *X*'s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function evaluated at the maximum likelihood estimates. Our goal is to find a subset size above which little is gained by adding more variables.

R^2 Value

This is the value of R^2 calculated using the formula

$$R_L^2 = \frac{L_p - L_0}{L_0 - L_S}$$

as discussed in the introduction. We are looking for the subset size at which this value does not increase by a meaningful amount.

R^2

This is the increase in R^2 that occurs when each new subset size is reached. Search for the subset size below which the R^2 value does not increase by more than 0.02 for small samples or 0.01 for large samples.

In this example, the optimum subset size appears to be four terms.

Subset Selection Detail

Subset Selection Detail
Subset Selection Method = Hierarchical Forward Selection

Step	Action	No. of Terms	No. of X's	Log Likelihood	Term Entered	Term Removed
1	Add	1	1	-17.18588	Intercept	
2	Add	2	2	-13.03648	LI	
3	Add	3	3	-12.17036	Cell	
4	Add	4	4	-10.97669	Temp	
5	Add	5	5	-10.92900	Smear	

This report shows the highest log likelihood for each subset size. In this example, it appears that four terms (the intercept and three variables) provide the best model. Note that adding the fifth variable does not increase the *R*-squared value very much.

Action

This item identifies the action that was taken at this step. A term was added, removed, or two were switched.

No. Terms

The number of terms. Note that this includes the intercept.

No. X's

The number of *X*'s that were included in the model. Note that in this case, the number of terms matches the number of *X*'s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function after the completion of this step. Our goal is to find a subset size above which little is gained by adding more variables.

Terms Entered and Removed

These columns identify the terms added, removed, or switched.

Discussion of Example 2

After considering these reports, it was decided to include Cell, LI, and Temp in the final logistic regression model. Another run should now take place using only these independent variables. A complete residual analysis is necessary before the equation is finally adopted.

Example 3 – One Categorical X Variable

The independent variables in logistic regression may be categorical as well as numerical. This example is of the simplest categorical case of a binary response and a binary independent variable. More complicated examples will be shown below.

In this example, a simple yes-no question is asked of each member of two groups. The following two-by-two table presents the results. The analyst wants to understand the relationship between group membership and response to the question.

	Response		
Group	**Yes**	**No**	**Total**
A	91	9	100
B	93	27	120
Total	184	36	220

These data would normally be analyzed using the methods for comparing two proportions such as Fisher's exact test or the chi-square test for independence in a contingency table. The following table presents the results of this analysis.

Two Proportions Output

Counts and Proportions

Group	Response: No Count	Response: Yes Count	Total Count	Proportion*
A	9	91	100	p1 = 0.0900
B	27	93	120	p2 = 0.2250

* Proportion = No / Total

Proportions Analysis

Statistic	Value
Group 1 Event Rate (p1)	0.0900
Group 2 Event Rate (p2)	0.2250
Absolute Risk Difference |p1 - p2|	0.1350
Number Needed to Treat 1/|p1 - p2|	7.41
Relative Risk Reduction |p1 - p2|/p2	0.60
Relative Risk p1/p2	0.40
Odds Ratio o1/o2	0.34

Two-Sided Tests of the Difference (P1 - P2)
H0: P1 = P2 vs. Ha: P1 ≠ P2

Test Statistic Name	p1	p2	Difference p1 - p2	Test Statistic Value	Prob Level	Reject H0 at α = 0.05?
Wald Z	0.0900	0.2250	-0.1350	-2.695	0.0070	Yes
Fisher's Exact	0.0900	0.2250	-0.1350	0.010	0.0097	Yes

The conclusion of this analysis is to reject the null hypothesis that the two proportions are equal. The significance levels are 0.0097 using Fisher's exact test and 0.0070 using the normal approximation which is equivalent to the chi-square test for independence. Note that the odds ratio is 0.34.

We will now see how to analyze these data using logistic regression. The data must be entered into a database so that they can be processed. The following table shows how these data are rearranged and entered. These data have been entered into a database named 2BY2.

2By2 dataset (subset)

Group	Response	Count
A	No	9
A	Yes	91
B	No	27
B	Yes	93

Setup

To run this example, complete the following steps:

1 Open the 2By2 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **2By2** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Response**
Categorical X's	**Group**
Default Recoding Scheme	**Binary**
Frequencies	**Count**
Priors	**Equal across Y Values**
Reports Tab	
Run Summary	**Checked**
Y Variable Summary	**Checked**
Coefficient Significance Tests	**Checked**
Odds Ratios	**Checked**
Analysis of Deviance	**Checked**
Log-Likelihood and R^2	**Checked**
All Other Reports	**Unchecked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Logistic Regression Output

Run Summary

Item	Value	Item	Value
Y Variable	Response	Rows Processed	4
Reference Value	No	Rows Used	4
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	Count	Rows X's Missing	0
Numeric X Variables	0	Rows Freq Miss. or 0	0
Categorical X Variables	1	Rows Prediction Only	0
Final Log Likelihood	-94.23344	Unique Rows (Y and X's)	4
Model R²	0.06908	Sum of Frequencies	220
Actual Convergence	2.559022E-11	Likelihood Iterations	6
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	2	Completion Status	Normal Completion
Priors	Equal		

Y Variable Summary

Y Response	Count	Unique Rows (Y and X's)	Y Proportion	Y Prior	R² (Y vs Pred. Probability)	Percent Correctly Classified
No	36	2	0.16364	0.50000	0.03302	75.000
Yes	184	2	0.83636	0.50000	0.03302	49.457
Total	220	4				53.636

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept	0.68222	0.29814	2.288	0.02212	1.97826
(Group="B")	-1.07687	0.41218	-2.613	0.00898	0.34066

Odds Ratios

Independent Variable X	Regression Coefficient b(i)	Odds Ratio Exp(b(i))	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Intercept	0.68222	1.97826	1.10282	3.54863
(Group="B")	-1.07687	0.34066	0.15187	0.76413

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
All	1	196.08640	7.61951	0.00577
Group	1	196.08640	7.61951	0.00577
None(Model)	1	188.46689		

Log Likelihood & R²

Term(s) Omitted	DF	Log Likelihood	R² of Remaining Term(s)	Reduction From Model R²	Reduction From Saturated R²
All	1	-98.04320	0.00000		
Group	1	-98.04320	0.00000	0.06908	1.00000
None(Model)	1	-94.23344	0.06908	0.00000	0.93092
None(Saturated)	4	-42.89226	1.00000		0.00000

Although a casual comparison between this report and that of the Two Proportion procedure shows little in common, a more detailed report shows many similarities. First of all, notice that the significance level of the test of GROUP in the Analysis of Deviance Section of 0.00577 compares very closely with the 0.007037 from the chi-square test. Also notice that the odds ratios from both reports round to 0.34066. The confidence limits of these two reports are not exactly the same, but they are close.

To summarize the logistic regression analysis, we can conclude that there is a significant relationship between response and group.

This example has shown the similarities between these two approaches to the analysis of two proportions. Usually, you would analyze these data using the two proportions approach. However, that approach is not as easily extended to the case of several independent variables including a mixture of categorical and numeric.

Example 4 – Logit Model Validation with BMDP PR

This example will serve three purposes. First of all, it will be the first example of a dataset whose Y variable has more than two outcomes. Second, it will be an example of what the output looks like when all of the independent variables are categorical. And finally, it will validate the procedure by allowing the comparison of the **NCSS** output with that of the **BMDP PR** program which also performs multiple-group logistic regression. This example comes from the **BMDP** manual. The database containing the data used in this example is named NC Criminal

The NC Criminal dataset contains data that will be used to study the relationship between a cases verdict and three factors: race, county, and type of offense. The variables that are on the database are as follows.

Count contains the number of individuals with the characteristics specified on that row.

Verdict is the response variable. Three outcomes are given in the database: *G* for guilty, *NG* for not guilty, and *NP* for not prosecuted.

Race gives the race of the individual. It has two values: *A* and *B*.

County refers to county in North Carolina in which the offense was considered. The possible values are: *Durham* and *Orange*.

Offense contains the particular offense that the individual was accused of. These are *Drunk*, *Violence*, *Property*, *Major Traffic*, and *Speeding.*

You can view the data by loading the NC Criminal dataset, so they will not be displayed here.

Setup

To run this example, complete the following steps:

1 Open the NC Criminal example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **NC Criminal** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 4** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Verdict**
Reference Value	**NP**
Categorical X's	**Race(B;A) County(B;Durham) Offense(B;Drunk)**
Frequencies	**Count**
Priors	**Ni/N (Y-Value Proportions)**
Reports Tab	
Run Summary	**Checked**
Y Variable Summary	**Checked**
Coefficient Significance Tests	**Checked**
Analysis of Deviance	**Checked**
Log-Likelihood and R^2	**Checked**
All Other Reports	**Unchecked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Logistic Regression Output

Run Summary

Item	Value	Item	Value
Y Variable	Verdict	Rows Processed	60
Reference Value	NP	Rows Used	57
Number of Y-Values	3	Rows for Validation	0
Frequency Variable	Count	Rows X's Missing	0
Numeric X Variables	0	Rows Freq Miss. or 0	3
Categorical X Variables	3	Rows Prediction Only	0
Final Log Likelihood	-408.29185	Unique Rows (Y and X's)	60
Model R^2	0.69779	Sum of Frequencies	615
Actual Convergence	4.751915E-11	Likelihood Iterations	6
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	14	Completion Status	Normal Completion
Priors	Ni/N		

Y Variable Summary

Y Verdict	Count	Unique Rows (Y and X's)	Y Proportion	Y Prior	R^2 (Y vs Pred. Probability)	Percent Correctly Classified
G	445	20	0.72358	0.72358	0.17107	93.933
NG	123	20	0.20000	0.20000	0.10397	20.325
NP	47	20	0.07642	0.07642	0.06628	0.000
Total	615	60				72.033

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept					
G	2.82983	0.44457	6.365	0.00000	16.94253
NG	1.24012	0.48781	2.542	0.01102	3.45604
(Race="B")					
G	0.26083	0.33984	0.767	0.44279	1.29800
NG	-0.10324	0.36248	-0.285	0.77579	0.90191
(County="Orange")					
G	-0.89593	0.33719	-2.657	0.00788	0.40823
NG	-0.12175	0.36036	-0.338	0.73547	0.88537
(Offense="MjTraffic")					
G	-0.21380	0.62893	-0.340	0.73390	0.80751
NG	0.48012	0.67038	0.716	0.47387	1.61627
(Offense="Property")					
G	-0.91853	0.57784	-1.590	0.11193	0.39911
NG	0.00928	0.61911	0.015	0.98804	1.00932
(Offense="Speed")					
G	0.49546	0.51245	0.967	0.33361	1.64126
NG	-0.26697	0.57599	-0.463	0.64301	0.76570
(Offense="Violence")					
G	-2.23014	0.51372	-4.341	0.00001	0.10751
NG	-0.57863	0.53748	-1.077	0.28168	0.56067

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi^2)	P-Value
All	12	925.59805	109.01434	0.00000
Race	2	819.21845	2.63475	0.26784
County	2	832.03780	15.45409	0.00044
Offense	8	898.18115	81.59744	0.00000
None(Model)	12	816.58371		

Log Likelihood & R^2

Term(s) Omitted	DF	Log Likelihood	R^2 of Remaining Term(s)	Reduction From Model R^2	Reduction From Saturated R^2
All	2	-462.79903	0.00000		
Race	2	-409.60923	0.68093	0.01686	0.31907
County	2	-416.01890	0.59887	0.09892	0.40113
Offense	8	-449.09057	0.17549	0.52230	0.82451
None(Model)	12	-408.29185	0.69779	0.00000	0.30221
None(Saturated)	120	-384.68551	1.00000		0.00000

The output format is similar to previous examples. Notice in the analysis of deviance section that the variable *Race* is not significant. That is, in these data, the race of the defendant is not related to the verdict.

The *Coefficient Significance Tests* report combines the two logistic regression equations on one report. This makes it a bit more complicated to read, but it allows a quick comparison to be made of the corresponding regression coefficients. For each independent variable, the regression coefficient from each equation is shown. Thus, 2.82983 is the intercept for the *G* equation and 1.24012 is the intercept for the *NG* equation. No coefficient is shown for *NP* because it is the reference value.

Also note that the definition of the binary variables is as before. Thus the independent variable *County="Orange"* refers to a binary variable that was generated from the *County* variable. This binary variable is one when the county value is *Orange* and zero otherwise.

Validation

In order to validate this module, the estimated regression coefficients and the log likelihood generated by the ***BMDP*** (refer to page 1165 of version 7.0 of the ***BMDP*** manual) are displayed below.

Outcome: G	Coefficient	Std Error
1 RACE	0.2608	0.340
2 COUNTY	-0.8959	0.337
3 OFFENSE(1)	-2.230	0.514
4 OFFENSE(2)	-0.9185	0.578
5 OFFENSE(3)	-0.2138	0.629
6 OFFENSE(4)	0.4955	0.512
7 CONST1	2.830	0.445

Outcome: NG	Coefficient	Std Error
8 RACE	-0.1032	0.362
9 COUNTY	-0.1218	0.360
10 OFFENSE(1)	-0.5786	0.537
11 OFFENSE(2)	0.9281E-02	0.619
12 OFFENSE(3)	0.4801	0.670
13 OFFENSE(4)	-0.2670	0.576
14 CONST1	1.240	0.488

As you can see, these results match those displayed by **NCSS** exactly.

Example 5 – Logit Model with Interaction

This example continues with the analysis of the data given in Example 4. In that example, no interactions were included in the model. This example will include the two-way interactions in the model.

Setup

To run this example, complete the following steps:

1 Open the NC Criminal example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **NC Criminal** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 5** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**Verdict**
Reference Value	**NP**
Categorical X's	**Race(B;A) County(B;Durham) Offense(B;Drunk)**
Frequencies	**Count**
Terms	**Up to 2-Way**
Priors	**Ni/N (Y-Value Proportions)**
Reports Tab	
Run Summary	**Checked**
Y Variable Summary	**Checked**
Coefficient Significance Tests	**Checked**
Analysis of Deviance	**Checked**
Log-Likelihood and R^2	**Checked**
All Other Reports	**Unchecked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Logistic Regression Output

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept					
G	2.00583	0.50400	3.980	0.00007	7.43225
NG	0.72258	0.57465	1.257	0.20860	2.05975
(Race="B")					
G	1.44835	0.86924	1.666	0.09567	4.25608
NG	-1.10628	1.08369	-1.021	0.30733	0.33079
(County="Orange")					
G	0.14731	1.15368	0.128	0.89840	1.15871
NG	1.83395	1.18755	1.544	0.12251	6.25854
(Offense="MjTraffic")					
G	-0.30745	1.10221	-0.279	0.78029	0.73532
NG	-0.25450	1.23436	-0.206	0.83665	0.77531
(Offense="Property")					
G	-0.72178	0.83542	-0.864	0.38760	0.48589
NG	0.35757	0.89267	0.401	0.68874	1.42985
(Offense="Speed")					
G	1.93682	1.08041	1.793	0.07303	6.93666
NG	0.87254	1.19650	0.729	0.46586	2.39297
(Offense="Violence")					
G	-0.15836	0.87409	-0.181	0.85624	0.85354
NG	1.07460	0.91294	1.177	0.23916	2.92882
(Race="B")*(County="Orange")					
G	0.19528	0.81517	0.240	0.81067	1.21566
NG	0.83286	0.85899	0.970	0.33225	2.29990
(Race="B")*(Offense="MjTraffic")					
G	-1.17876	1.35078	-0.873	0.38285	0.30766
NG	1.16592	1.50638	0.774	0.43894	3.20886
(Race="B")*(Offense="Property")					
G	-0.83367	1.27452	-0.654	0.51305	0.43445
NG	1.35214	1.42888	0.946	0.34400	3.86569
(Race="B")*(Offense="Speed")					
G	-1.78987	1.25551	-1.426	0.15398	0.16698
NG	0.24862	1.45010	0.171	0.86387	1.28225
(Race="B")*(Offense="Violence")					
G	-2.31322	1.19041	-1.943	0.05199	0.09894
NG	0.51640	1.30133	0.397	0.69150	1.67598
(County="Orange")*(Offense="MjTraffic")					
G	0.45137	1.52019	0.297	0.76653	1.57046
NG	-0.53668	1.61710	-0.332	0.73998	0.58469
(County="Orange")*(Offense="Property")					
G	0.04871	1.41697	0.034	0.97258	1.04992
NG	-2.10279	1.47544	-1.425	0.15410	0.12212
(County="Orange")*(Offense="Speed")					
G	-1.39431	1.37573	-1.014	0.31082	0.24800
NG	-2.66093	1.48387	-1.793	0.07294	0.06988
(County="Orange")*(Offense="Violence")					
G	-2.42314	1.36627	-1.774	0.07614	0.08864
NG	-3.93664	1.38198	-2.849	0.00439	0.01951

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
All	30	925.59805	146.82239	0.00000
Race	2	797.83870	19.06304	0.00007
County	2	788.31126	9.53560	0.00850
Offense	8	802.98614	24.21048	0.00211
Race*County	2	780.53878	1.76312	0.41414
Race*Offense	8	795.98619	17.21053	0.02799
County*Offense	8	798.81172	20.03607	0.01020
None(Model)	30	778.77566		

Log Likelihood & R^2

Term(s) Omitted	DF	Log Likelihood	R^2 of Remaining Term(s)	Reduction From Model R^2	Reduction From Saturated R^2
All	2	-462.79903	0.00000		
Race	2	-398.91935	0.81778	0.12202	0.18222
County	2	-394.15563	0.87877	0.06104	0.12123
Offense	8	-401.49307	0.78483	0.15497	0.21517
Race*County	2	-390.26939	0.92852	0.01129	0.07148
Race*Offense	8	-397.99309	0.82964	0.11016	0.17036
County*Offense	8	-399.40586	0.81155	0.12825	0.18845
None(Model)	30	-389.38783	0.93980	0.00000	0.06020
None(Saturated)	120	-384.68554	1.00000		0.00000

Notice how the interactions are labeled. For example, the variable labeled *(Race="B")*(Offense="Violence")* is the interaction variable is generated by multiplying the binary variable defined by (Race="B") with the binary variable defined by (Offense="Violence"). The resulting variable is one if both of these conditions are true and zero otherwise.

Note that the R^2 is now 0.93980, so this model is almost as good as the saturated model.

Looking at the analysis of deviance table, we note that all terms are significant except for the Race*County interaction.

Example 6 – Odds Ratios for Categorical X's

Lachin (2000) pages 90, 91, and 257 presents an analysis of hypothetical data from an ulcer healing clinical trial conducted to study the effectiveness of a drug over a placebo. There were 100 patients assigned to the group receiving the drug and another 100 patients assigned to the group receiving the placebo. The ulcers were stratified into one of three types: 1. Acid-dependent, 2. Drug dependent, and 3. Intermediate. Each ulcer was followed for a period of time after which it was considered healed or not. The data for this experiment are given below. These data have been entered into a database named **Lachin91**.

Lachin91 dataset (subset)

Count	Ulcer	Drug	Healed
16	1	1	1
26	1	1	0
20	1	0	1
27	1	0	0
9	2	1	1
3	2	1	0
4	2	0	1
5	2	0	0
28	3	1	1
18	3	1	0
16	3	0	1
28	3	0	0

Setup

To run this example, complete the following steps:

1 Open the Lachin91 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Lachin91** and click **OK**.

2 Specify the Logistic Regression procedure options

- Find and open the **Logistic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 6** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option **Value**

Variables, Model Tab
Y .... **Healed**
Categorical X's....................................... **Ulcer Drug**
Frequencies .... **Count**
Priors .... **Equal across Y Values**

Reports Tab
Run Summary.. **Checked**
Coefficient Significance Tests .... **Checked**
Odds Ratios .... **Checked**
Analysis of Deviance .... **Checked**
All Other Reports .... **Unchecked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Logistic Regression Output

Run Summary

Item	Value	Item	Value
Y Variable	Healed	Rows Processed	12
Reference Value	0	Rows Used	12
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	Count	Rows X's Missing	0
Numeric X Variables	0	Rows Freq Miss. or 0	0
Categorical X Variables	2	Rows Prediction Only	0
Final Log Likelihood	-134.84531	Unique Rows (Y and X's)	12
Model R^2	0.54106	Sum of Frequencies	200
Actual Convergence	1.10275E-10	Likelihood Iterations	4
Target Convergence	1E-06	Maximum Iterations	20
Model D.F.	4	Completion Status	Normal Completion
Priors	Equal		

Coefficient Significance Tests

Independent Variable X	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald P-Value	Odds Ratio Exp(b(i))
Intercept	-0.48951	0.21833	-2.242	0.02496	0.61293
(Ulcer=2)	0.83527	0.50247	1.662	0.09645	2.30543
(Ulcer=3)	0.32777	0.30424	1.077	0.28132	1.38787
(Drug=1)	0.50234	0.28845	1.742	0.08159	1.65259

Odds Ratios

Independent Variable X	Regression Coefficient b(i)	Odds Ratio Exp(b(i))	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Intercept	-0.48951	0.61293	0.39955	0.94027
(Ulcer=2)	0.83527	2.30543	0.86109	6.17243
(Ulcer=3)	0.32777	1.38787	0.76451	2.51949
(Drug=1)	0.50234	1.65259	0.93894	2.90864

Analysis of Deviance

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
All	3	276.27807	6.58746	0.08628
Ulcer	2	272.87155	3.18094	0.20383
Drug	1	272.74521	3.05460	0.08051
None(Model)	3	269.69061		

Note that neither Drug nor Ulcer is statistically significant at the 0.05 level using either the deviance tests in the *Analysis of Deviance* table or the Wald tests in the *Coefficient Significance Tests* section. From the *Odds Ratios* section, we see that the odds of healing are increased 1.65259 when the drug is administered.