Statistics for Analytics (BAN 100)

Assignment 6

by : Aaron Gonsalves

(161288196)

PROBLEM 1

CODES

```
proc import datafile  =  '/home/u58712040/BAN100/files/Customer.xlsx'
 out  =  Customer
 dbms  =  xlsx
 replace;
 getnames=yes;
run;

data Customer;
    set Customer;
    if lowcase(Rating) in ('very good', 'excellent') then Y=1;
    if lowcase(Rating) in ('good', 'fair') then Y=0;
run;

title"Customer data";
proc print data=Customer;
run;
```

Customer data

Obs	Manufacturer	Price	Rating	Y
1	Bernard Callebaut	3.17	Very Good	1
2	Candinas	3.58	Excellent	1
3	Fannie May	1.49	Good	0
4	Godiva	2.91	Very Good	1
5	Hershey,Äôs	0.76	Good	0
6	L.A. Burdick	3.7	Very Good	1
7	La Maison du Chocolate	5.08	Excellent	1
8	Leonidas	2.11	Very Good	1
9	Lindt	2.2	Good	0
10	Martine,Äôs	4.76	Excellent	1
11	Michael Recchiuti	7.05	Very Good	1
12	Neuchatel	3.36	Good	0
13	Neuchatel Sugar Free	3.22	Good	0
14	Richard Donnelly	6.55	Very Good	1
15	Russell Stover	0.7	Good	0
16	See,Äôs	1.06	Very Good	1
17	Teuscher Lake of Zurich	4.66	Very Good	1
18	Whitman,Äôs	0.7	Fair	0
19	Whitman,Äôs Sugar Free	1.21	Fair	0

a) Write the logistic regression equation relating x = price per serving to y.

$$P(Y|X) = \frac{e^{\beta 0+\beta 1\, X1}}{1+e^{\beta 0+\beta 1\, X1}}$$

β0 is the intercept and β1 is the coefficient of X1.

b) Use SAS to compute the estimated logit.

CODES

```
proc logistic data=Customer;
            model y = Price;
run;
```

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CUSTOMER
Response Variable	Y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	19
Number of Observations Used	19

Response Profile

Ordered Value	Y	Total Frequency
1	0	8
2	1	11

Probability modeled is Y=0.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	27.864	20.399
SC	28.808	22.288
-2 Log L	25.864	16.399

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.4648	1	0.0021
Score	7.3311	1	0.0068
Wald	4.9924	1	0.0255

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.8050	1.4316	3.8387	0.0501
Price	1	-1.1492	0.5143	4.9924	0.0255

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Price	0.317	0.116	0.868

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	86.4	Somers' D	0.727
Percent Discordant	13.6	Gamma	0.727
Percent Tied	0.0	Tau-a	0.374
Pairs	88	c	0.864

- The coefficient of Price, β1 is -1.1492
- The intercept, β0 is 2.8050
- From the above procedure, it is clear that our model is built for level 0.
- So, the value that we get from the equation represents the probability of the quality rating ‘good’ or ‘fair’.

c) Use the estimated logit computed in part (b) to compute an estimate of the probability
a chocolate that has a price per serving of $4.00 will have a quality rating of very good or excellent.

$$\hat{y} = p(y=0 \mid x) = \frac{e^{\beta 0 + \beta 1\, x1}}{1 + e^{\beta 0 + \beta 1\, x1}} = \frac{e^{2.8050 - 1.1492x}}{1 + e^{2.8050 - 1.1492\, x}} = \frac{e^{2.8050 - 1.1492\ (4)}}{1 + e^{2.8050 - 1.1492\ (4)}}$$

$$= \frac{0.1667}{1.1667} = 0.1429 \quad = 14.29\ \%$$

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.8050	1.4316	3.8387	0.0501
Price	1	-1.1492	0.5143	4.9924	0.0255

- 14.29% shows the probability of a chocolate that has a price per serving of $4.00 will have a quality rating ‘good’ or ‘fair’.

- The probability of a chocolate that has a price per serving of $4.00 will have a quality rating of ‘Very good’ or ‘Excellent’

= 1-0.1429 = 0.8571 =85.71%

d) What is the estimate of the odds ratio? What is its interpretation?

- The probability of a chocolate that has a price per serving of $4.00 will have a quality rating ‘good’ or ‘fair’ is 14.29%

- The probability of a chocolate that has a price per serving of $4.00 will have a quality rating of ‘Very good’ or ‘Excellent’ is 85.71%

$$\text{odds}_0 = \frac{0.1429}{0.8571} = 0.1666$$

- Increasing Price to $5.00

- The probability of a chocolate that has a price per serving of $4.00 will have a quality rating ‘good’ or ‘fair’ is 5.02%

- The probability of a chocolate that has a price per serving of $4.00 will have a quality rating of ‘Very good’ or ‘Excellent’ is 94.98% (1-0.0502)

$$\hat{y} = \mathrm{p}(\mathrm{y}=0 \mid \mathrm{x}) = \frac{e^{\beta 0+ \beta 1\, x1}}{1+ e^{\beta 0+ \beta 1\, x1}} = \frac{e^{2.8050\, -1.1492 x}}{1+ e^{2.8050\, -1.1492\, x}} = \frac{e^{2.8050\, -1.1492\, (5)}}{1+ e^{2.8050\, -1.1492\, (5)}}$$

$$= \frac{0.0529}{1.0529} = 0.0502$$

$$\text{odds}_1 = \frac{0.0502}{0.9498} = 0.05281$$

$$\text{odds}_0 = \frac{0.1429}{0.8571} = 0.1666$$

$$\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0}$$

$$\text{Odds Ratio} = \frac{0.05281}{0.16666} = 0.317$$

Interpretation

- The odds of Chocolate quality rating ‘good’ or ‘fair’ over quality rating of ‘Very good’ or ‘Excellent’ is 5% / 94% that is 0.0581
- The Odds Ratio for Price is 0.317, 95% and the Confidence interval is 0.116 to 0.868
- The odds for both events are the same, if the value of the odds ratio is 1.
- A ratio of odds greater than 1 infers that there are greater odds of the event happening versus the non-happening.
- A ratio of odds less than 1 infers that there are lesser odds of the event happening versus the non-happening.
- The odd ratio is 0.317 here, which is less than 1 and thus, infers that the odds for Chocolate quality rating being ‘Good’ or ‘Fair’ is lesser than the odds of quality rating being ‘Very Good’ or ‘Excellent’.

PROBLEM 2

CODES

```
proc import datafile  =  '/home/u58712040/BAN100/files/titanic.csv'
 out  =  Titanic
 dbms  =  CSV
 replace;
 getnames=yes;
run;

title "Listing of Titanic dataset";
proc print data=Titanic(Obs=10);
run;
```

Listing of Titanic dataset

Obs	PassengerId	Survived	Class	Name	Sex	Age	SiblingSpouse	ParentChild	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	6	0	3	Moran, Mr. James	male	.	0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

a) Write the logistic regression equation relating Class and Survived.

$$\text{Probability of Y given X, } P(Y|X) = \frac{e^{\beta 0+\beta 1\, X1+\beta 2\, X2}}{1+e^{\beta 0+\beta 1\, X1+\beta 2\, X2}}$$

- β0 is the intercept and β1 is the coefficient of X1 and β2 is the coefficient of X2

Dependent Variable

y=1 i.e. passenger was survived

y=0 i.e. passenger was not survived

Independent Variable

Class - First, Second & Third class passengers (1, 2 & 3)

b) For the Titanic data, use SAS to compute the estimated logistic regression equation.

CODES

```
proc logistic data=Titanic;
class Class param=ref;
model Survived= Class;
run;
```

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TITANIC
Response Variable	Survived
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	891
Number of Observations Used	891

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	549
2	1	342

Probability modeled is Survived='1'.

Class Level Information			
Class	Value	Design Variables	
Class	1	1	0
	2	0	1
	3	0	0

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1188.655	1089.108
SC	1193.447	1103.485
-2 Log L	1186.655	1083.108

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	103.5471	2	<.0001
Score	102.8890	2	<.0001
Wald	96.6294	2	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Class	2	96.6294	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.1398	0.1053	117.1232	<.0001
Class	1	1	1.6704	0.1759	90.1689	<.0001
Class	2	1	1.0310	0.1814	32.3116	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Class 1 vs 3	5.314	3.765	7.502
Class 2 vs 3	2.804	1.965	4.001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	51.2	Somers' D	0.363
Percent Discordant	14.9	Gamma	0.549
Percent Tied	33.9	Tau-a	0.172
Pairs	187758	c	0.681

- The coefficient of Class 1 i.e., β1 is 1.6704
- The coefficient of Class 2 i.e., β1 is 1.0310
- The value of intercept, β0 is -1.1398
- From the above procedure, it is clear that our model is built for level 0.
- So, the value we get from the equation shows the probability of the passenger who did not survive.

c) What is the interpretation of $E(y)$ when $x_2 = 2$?

- Estimated probability that a passenger in second class will survive.

d) Estimate the probability of surviving the 2nd class passengers and the 3rd class passengers.

CODES

```
proc logistic data=Titanic;
class Class (ref='2' ref='3') / param=ref;
model Survived(event='1') = Class;
run;
```

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TITANIC
Response Variable	Survived
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	891
Number of Observations Used	891

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	549
2	1	342

Probability modeled is Survived='1'.

Class Level Information			
Class	Value	Design Variables	
Class	1	1	0
	2	0	1
	3	0	0

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1188.655	1089.108
SC	1193.447	1103.485
-2 Log L	1186.655	1083.108

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	103.5471	2	<.0001
Score	102.8890	2	<.0001
Wald	96.6294	2	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Class	2	96.6294	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.1398	0.1053	117.1232	<.0001
Class	1	1	1.6704	0.1759	90.1689	<.0001
Class	2	1	1.0310	0.1814	32.3116	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Class 1 vs 3	5.314	3.765	7.502
Class 2 vs 3	2.804	1.965	4.001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	51.2	Somers' D	0.363
Percent Discordant	14.9	Gamma	0.549
Percent Tied	33.9	Tau-a	0.172
Pairs	187758	c	0.681

e) What is the estimated odds ratio? What is the interpretation?

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Class 1 vs 3	5.314	3.765	7.502
Class 2 vs 3	2.804	1.965	4.001

Interpretation

- The passenger in 2nd Class has 2.804 times odds of surviving than that of passenger in 3rd Class.

- The passenger in 1st Class have 5.314 times odds of surviving than that of passenger in 3rd Class.

- The Logistic Regression Model Accuracy is 51.2%

THANK YOU