類別資料分析期末考

姓名:葉冠宏　學號:108753208

Problem1

1. The dataset "loan.txt" contains 1000 applicants' 11 variables. The loan department reviewed each applicant's profile and decided whether the applicant is creditable.

Y=0 if not creditability; 1 if creditiability.
AcctBal l: 0 if no account or $0 in account; 1 if account balance > $0.
DuraCred: in months

PrevPay: 0 if delayed payment; 1 if no delayed payment.
CredAmnt: in $100
Savings: 0 if none or low in savings; 1 if some savings.
CurrEmpl: 0 if length of current employment < 4 years; 1 if at least 4 years.
Gender: 0 if male; 1 if female.
Age: in years.
House: 0 if not owned; 1 if owned.

Logistic regression model is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.

(a) Report descriptive analysis for each variable.

DuraCre, CredAmnt, Age 都是 numerical data

AcctBal, PrevPay, Savings, CurrEmpl, Gender, House 都是 binary data

DuraCard:

可以看到 DuraCard 的平均是 20.9, 中位數是 18, 眾數是 24, 標準差是 12, range 是 68

The SAS System

The UNIVARIATE Procedure
Variable: DuraCred

Moments			
N	1000	Sum Weights	1000
Mean	20.903	Sum Observations	20903
Std Deviation	12.0588145	Variance	145.415006
Skewness	1.09418417	Kurtosis	0.91978136
Uncorrected SS	582205	Corrected SS	145269.591
Coeff Variation	57.689396	Std Error Mean	0.3813332

Basic Statistical Measures			
Location		Variability	
Mean	20.90300	Std Deviation	12.05881
Median	18.00000	Variance	145.41501
Mode	24.00000	Range	68.00000
		Interquartile Range	12.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	54.81558	Pr > |t|	<.0001
Sign	M	500	Pr >= |M|	<.0001
Signed Rank	S	250250	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	72
99%	60
95%	48
90%	36
75% Q3	24
50% Median	18
25% Q1	12
10%	9
5%	6
1%	6
0% Min	4

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
4	545	60	808
4	538	60	812
4	537	60	848
4	531	60	891
4	239	72	720

CredAmnt:

可以看到 CredAmnt 的平均是 3271, 中位數是 2319, 眾數是 1258, 標準差是 2823, range 是 18174

The SAS System

The UNIVARIATE Procedure
Variable: CredAmnt

Moments			
N	1000	Sum Weights	1000
Mean	3271.248	Sum Observations	3271248
Std Deviation	2822.75176	Variance	7967927.5
Skewness	1.94959429	Kurtosis	4.29248061
Uncorrected SS	1.8661E10	Corrected SS	7959959570
Coeff Variation	86.2897512	Std Error Mean	89.2632483

Basic Statistical Measures			
Location		Variability	
Mean	3271.248	Std Deviation	2823
Median	2319.500	Variance	7967927
Mode	1258.000	Range	18174
		Interquartile Range	2608

Note: The mode displayed is the smallest of 5 modes with a count of 3.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	36.6472	Pr > |t|	<.0001
Sign	M	500	Pr >= |M|	<.0001
Signed Rank	S	250250	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	18424.0
99%	14248.5
95%	9214.0
90%	7201.0
75% Q3	3972.5
50% Median	2319.5
25% Q1	1365.0
10%	933.5
5%	708.5
1%	417.5
0% Min	250.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
250	313	15653	696
276	384	15672	811
338	287	15857	646
339	301	15945	963
343	374	18424	977

Age:

可以看到 Age 的平均是 35.54, 中位數是 33, 眾數是 27, 標準差是 11.35, range 是 56

The SAS System

The UNIVARIATE Procedure
Variable: Age

Moments			
N	1000	Sum Weights	1000
Mean	35.542	Sum Observations	35542
Std Deviation	11.3526701	Variance	128.883119
Skewness	1.02471202	Kurtosis	0.62052948
Uncorrected SS	1391988	Corrected SS	128754.236
Coeff Variation	31.9415625	Std Error Mean	0.35900295

Basic Statistical Measures			
Location		Variability	
Mean	35.54200	Std Deviation	11.35267
Median	33.00000	Variance	128.88312
Mode	27.00000	Range	56.00000
		Interquartile Range	15.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	99.00197	Pr > |t|	<.0001
Sign	M	500	Pr >= |M|	<.0001
Signed Rank	S	250250	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	75.0
99%	67.5
95%	60.0
90%	52.0
75% Q3	42.0
50% Median	33.0
25% Q1	27.0
10%	23.0
5%	22.0
1%	20.0
0% Min	19.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
19	820	74	214
19	299	74	260
20	990	74	815
20	988	75	603
20	926	75	625

AcctBal:

可以看到 AcctBal 有 606 個 0, 394 個 1。

在 AcctBal 為 0 時，DuraCred 的平均是 21.519，最高數值是 72。在 AcctBal 為 1 時，DuraCred 的平均是 19.95，最高數值是 60。所以在 account balance>$0 時 DuraCred 的平均比較低。

在 AcctBal 為 0 時，CredAmnt 的平均是 3361，最高數值是 18424。在 AcctBal 為 1 時，CredAmnt 的平均是 3133，最高數值是 15653。所以在 account balance>$0 時 CredAmnt 的平均比較低。

在 AcctBal 為 0 時，Age 的平均是 34.98，最高數值是 75。在 AcctBal 為 1 時，Age 的平均是 36.40，最高數值是 74。所以在 account balance>$0 時 Age 的平均比較高。

The SAS System

The MEANS Procedure

AcctBal	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	606	DuraCred	606	21.5198020	12.4916675	4.0000000	72.0000000
		CredAmnt	606	3361.06	2983.07	276.0000000	18424.00
		Age	606	34.9834983	11.6811521	19.0000000	75.0000000
1	394	DuraCred	394	19.9543147	11.3109168	4.0000000	60.0000000
		CredAmnt	394	3133.11	2554.16	250.0000000	15653.00
		Age	394	36.4010152	10.7862224	19.0000000	74.0000000

PrevPay:

可以看到 PrevPay 有 89 個 0, 911 個 1。

在 PrevPay 為 0 時，DuraCred 的平均是 25，最高數值是 60。在 PrevPay 為 1 時，DuraCred 的平均是 20.5，最高數值是 72。所以在 no delayed payment 時 DuraCred 的平均比較低。

在 PrevPay 為 0 時，CredAmnt 的平均是 4226.13，最高數值是 18424。在 PrevPay 為 1 時，CredAmnt 的平均是 3177.96，最高數值是 15857。所以在 no delayed payment 時 CredAmnt 的平均比較低。

在 PrevPay 為 0 時，Age 的平均是 35.26，最高數值是 74。在 PrevPay 為 1 時，Age 的平均是 35.56，最高數值是 75。所以在 no delayed payment 時 Age 略高一些。

The SAS System

The MEANS Procedure

PrevPay	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	89	DuraCred	89	25.0224719	14.2158503	6.0000000	60.0000000
		CredAmnt	89	4226.13	3789.23	339.0000000	18424.00
		Age	89	35.2696629	10.5202878	22.0000000	74.0000000
1	911	DuraCred	911	20.5005488	11.7589222	4.0000000	72.0000000
		CredAmnt	911	3177.96	2694.59	250.0000000	15857.00
		Age	911	35.5686059	11.4357960	19.0000000	75.0000000

CurrEmpl:

可以看到 CurrEmpl 有 573 個 0, 427 個 1。

在 CurrEmpl 為 0 時，DuraCred 的平均是 20.1，最高數值是 72。在 CurrEmpl 為 1 時，DuraCred 的平均是 21.96，最高數值是 60。所以在 length of current employment >=4 years 時 DuraCred 的平均比較高。

在 CurrEmpl 為 0 時，CredAmnt 的平均是 3191.52，最高數值是 18424。在 CurrEmpl 為 1 時，CredAmnt 的平均是 3378.24，最高數值是 15653。所以在 length of current employment >=4 years 時 CredAmnt 的平均比較高。

在 CurrEmpl 為 0 時，Age 的平均是 33.022，最高數值是 75。在 CurrEmpl 為 1 時，Age 的平均是 38.92，最高數值是 74。所以在 length of current employment >=4 years 時 Age 比較高。

The SAS System

The MEANS Procedure

CurrEmpl	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	573	DuraCred	573	20.1082024	11.6513521	4.0000000	72.0000000
		CredAmnt	573	3191.52	2762.22	250.0000000	18424.00
		Age	573	33.0226876	10.6746796	19.0000000	75.0000000
1	427	DuraCred	427	21.9695550	12.5195251	4.0000000	60.0000000
		CredAmnt	427	3378.24	2901.82	338.0000000	15653.00
		Age	427	38.9227166	11.3686053	20.0000000	74.0000000

Savings:

可以看到 Savings 有 603 個 0, 397 個 1。

在 Savings 為 0 時，DuraCred 的平均是 20.44，最高數值是 60。在 Savings 為 1 時，DuraCred 的平均是 21.60，最高數值是 72。所以如果有 some savings 時 DuraCred 的平均比較高。

在 Savings 為 0 時，CredAmnt 的平均是 3187.82，最高數值是 18424。在 Savings 為 1 時，CredAmnt 的平均是 3397.97，最高數值是 14782。所以如果有 some savings 時 CredAmnt 的平均比較高。

在 Savings 為 0 時，Age 的平均是 35.13，最高數值是 75。在 Savings 為 1 時，Age 的平均是 36.15，最高數值是 75。所以如果有 some savings 時 Age 比較高。

The SAS System

The MEANS Procedure

Savings	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	603	DuraCred	603	20.4411277	11.7411634	4.0000000	60.0000000
		CredAmnt	603	3187.82	2787.71	276.0000000	18424.00
		Age	603	35.1376451	11.4931422	19.0000000	75.0000000
1	397	DuraCred	397	21.6045340	12.5082712	4.0000000	72.0000000
		CredAmnt	397	3397.97	2874.06	250.0000000	14782.00
		Age	397	36.1561713	11.1221770	19.0000000	75.0000000

Gender:

可以看到 Gender 有 908 個 0, 92 個 1。

在 Gender 為 0 時，DuraCred 的平均是 21.22，最高數值是 60。在 Gender 為 1 時，DuraCred 的平均是 17.7，最高數值是 72。所以女性的 DuraCred 平均比較低。

在 Gender 為 0 時，CredAmnt 的平均是 3395.44，最高數值是 18424。在 Gender 為 1 時，CredAmnt 的平均是 2045.54，最高數值是 9398。所以女性的 CredAmnt 平均比較低。

在 Gender 為 0 時，Age 的平均是 36.07，最高數值是 75。在 Gender 為 1 時，Age 的平均是 30.34，最高數值是 61。所以女性的 Age 平均比較低。

The SAS System

The MEANS Procedure

Gender	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	908	DuraCred	908	21.2268722	12.1891227	4.0000000	60.0000000
		CredAmnt	908	3395.44	2882.85	250.0000000	18424.00
		Age	908	36.0682819	11.4850829	19.0000000	75.0000000
1	92	DuraCred	92	17.7065217	10.2050817	6.0000000	72.0000000
		CredAmnt	92	2045.54	1721.92	276.0000000	9398.00
		Age	92	30.3478261	8.3737355	20.0000000	61.0000000

House:

可以看到 House 有 893 個 0, 107 個 1。

在 House 為 0 時，DuraCred 的平均是 20.14，最高數值是 72。在 House 為 1 時，DuraCred 的平均是 27.26，最高數值是 60。所以如果有房子 DuraCred 的平均比較高。

在 House 為 0 時，CredAmnt 的平均是 3078.34，最高數值是 18424。在 House 為 1 時，CredAmnt 的平均是 4881.21，最高數值是 14782。所以如果有房子 CredAmnt 的平均比較高。

在 House 為 0 時，Age 的平均是 34.53，最高數值是 75。在 House 為 1 時，Age 的平均是 43.93，最高數值是 75。所以如果有房子 Age 的平均比較高。

The SAS System

The MEANS Procedure

House	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	893	DuraCred	893	20.1410974	11.4277915	4.0000000	72.0000000
		CredAmnt	893	3078.34	2640.19	250.0000000	18424.00
		Age	893	34.5363942	10.8284662	19.0000000	75.0000000
1	107	DuraCred	107	27.2616822	15.0266031	6.0000000	60.0000000
		CredAmnt	107	4881.21	3675.00	700.0000000	14782.00
		Age	107	43.9345794	12.1789108	22.0000000	75.0000000

Y:

可以看到 Y 有 300 個 0(信用不好), 700 個 1(信用好)。

在 Y 為 0 時，DuraCred 的平均是 24.86，最高數值是 72。在 Y 為 1 時，DuraCred 的平均是 19.2，最高數值是 60。所以信用較好的人 DuraCred 的平均比較低。

在 Y 為 0 時，CredAmnt 的平均是 3938，最高數值是 18424。在 Y 為 1 時，CredAmnt 的平均是 2985.44，最高數值是 15857。所以信用較好的人 CredAmnt 的平均比較低。

在 Y 為 0 時，Age 的平均是 33.96，最高數值是 74。在 Y 為 1 時，Age 的平均是 36.22，最高數值是 75。所以信用較好的人 Age 的平均比較高。

The SAS System

The MEANS Procedure

Y	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	300	DuraCred	300	24.8600000	13.2826389	6.0000000	72.0000000
		CredAmnt	300	3938.13	3535.82	433.0000000	18424.00
		Age	300	33.9600000	11.2251986	19.0000000	74.0000000
1	700	DuraCred	700	19.2071429	11.0795643	4.0000000	60.0000000
		CredAmnt	700	2985.44	2401.50	250.0000000	15857.00
		Age	700	36.2200000	11.3474279	19.0000000	75.0000000

(b) Let “Creditability” be the response variable, *Y*, fit a logistic regression using **EVENT='0’**, with all other variables in the model. Use **α=0.05**.

(b1) Remove nonsignificant variables and verify that models with and without these variables do not differ significantly. What is the final model at this stage?

ANS:

Final model at this stage:

Log(PI(Y=0|X) / (1- PI(Y=0|X)))= 0.1077 + 0.0395 * DuraCred -1.5073 * AcctBal -1.0533 * PrevPay -0.6120 * Savings -0.4705 * CurrEmpl

以下為 with nonsignificant variables 到 without nonsignificant variables，其他變數 beta 值的前後變化:

DuraCred 0.0338->0.0395 (16.8%)

AcctBal -1.4932->-1.5073 (-0.9%)

PrevPay -1.0431->-1.0533 (-0.97%)

Savings -0.6126->-0.6120 (0.09%)

CurrEmpl -0.4071->-0.4705 (-15.57%)

=>可以發現 beta 的前後變化都不超過 20% => do not differ significantly.

作法:

首先，fit the univariate logistic model for each covariate。最後我們發現 Gender is insignificant

```
PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=DuraCred;

PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=CredAmnt;

PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=Age;

PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=AcctBal;

PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=PrevPay;

PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=Savings;

PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=CurrEmpl;

PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=House;

PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=Gender;
```

接著我們用上一步得到的那些 significant 的 covariate 一起在 fit 一次 model. 然後我們用 alpha=0.05 去篩選，得到最後的 covariate 就是 DuraCred , AcctBal, PrevPay , Savings, CurrEmpl。

然後我們去一個一個看之前被篩掉的 covariate 如果加回去 fit 是不是還是 insignificant, 結果 CredAmnt, Age, Gender, House 都還是 insignificant.

以下為此階段的最終模型 fit 的結果:

The SAS System

The LOGISTIC Procedure

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

Number of Observations Read	1000
Number of Observations Used	1000

Response Profile		
Ordered Value	Y	Total Frequency
1	0	300
2	1	700

Probability modeled is Y='0'.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1223.729	1037.252
SC	1228.636	1066.699
-2 Log L	1221.729	1025.252

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	196.4763	5	<.0001
Score	181.6117	5	<.0001
Wald	146.8989	5	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.1077	0.2855	0.1423	0.7060
DuraCred	1	0.0395	0.00632	39.0797	<.0001
AcctBal	1	-1.5073	0.1848	66.5381	<.0001
PrevPay	1	-1.0533	0.2508	17.6370	<.0001
Savings	1	-0.6120	0.1666	13.5003	0.0002
CurrEmpl	1	-0.4705	0.1596	8.6941	0.0032

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
DuraCred	1.040	1.027	1.053
AcctBal	0.222	0.154	0.318
PrevPay	0.349	0.213	0.570
Savings	0.542	0.391	0.752
CurrEmpl	0.625	0.457	0.854

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	76.3	Somers' D	0.538
Percent Discordant	22.6	Gamma	0.544
Percent Tied	1.1	Tau-a	0.226
Pairs	210000	c	0.769

(b2) Add interactions between 2 variables in (b1), and fit the model. Then, remove nonsignificant variables or interactions, and verify that models with and without these variables do not differ significantly. Be cautious that if an interaction between A and B is significant, then both A and B should be kept in the model no matter being significant or not.

Ans:

Final model:

Log(PI(Y=0|X) / (1- PI(Y=0|X))) = 1.0005 + 0.00162*DuraCred -1.1853* AcctBal -2.4353*PrevPay + 0.0635*Savings -0.2848*CurrEmpl + 0.0575*DuraCred*PrevPay -0.0292*DuraCred*Savings - 1.0170 * AcctBal * CurrEmpl

Beta 值加上 interaction 前後變化:

DuraCred 0.0395->0.00162 (-95.89%)

AcctBal -1.5073->-1.1853 (-21.36%)

PrevPay -1.0533->-2.4353 (131.206%)

Savings -0.6120->0.0635 (-110.37%)

CurrEmpl -0.4705->-0.2848 (-39.46%)

=>可以發現 beta 值變化還蠻大的，因為 interaction 的加入，有 Confounding effect

作法:

我們一個個加(b1)階段各種 interaction 的排列組合，去看看是否 significant。結果最終得出了 DuraCred*PrevPay, DuraCred*Savings , AcctBal * CurrEmpl 這些排列組合都是 significant. 即便把這三個 interaction 同時加入 model 也都還是 significant.

```
/*STEP6: Explore possible interaction, add one pair one by one to see if it is significant.*/
/*
PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=DuraCred AcctBal PrevPay Savings CurrEmpl;


DuraCred*AcctBal (x)
DuraCred*CurrEmpl (x)
AcctBal*PrevPay (x)
AcctBal*Savings (x)
PrevPay*Savings (x)
PrevPay*CurrEmpl (x)
Savings*CurrEmpl (x)

DuraCred*PrevPay (o)
 DuraCred*Savings (o)
 AcctBal*CurrEmpl (o)
```

最終模型 fit 後的統計數據如下:

The SAS System

The LOGISTIC Procedure

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

Number of Observations Read	1000
Number of Observations Used	1000

Response Profile		
Ordered Value	Y	Total Frequency
1	0	300
2	1	700

Probability modeled is Y='0'.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1223.729	1020.711
SC	1228.636	1064.881
-2 Log L	1221.729	1002.711

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	219.0177	8	<.0001
Score	200.1792	8	<.0001
Wald	153.2273	8	<.0001

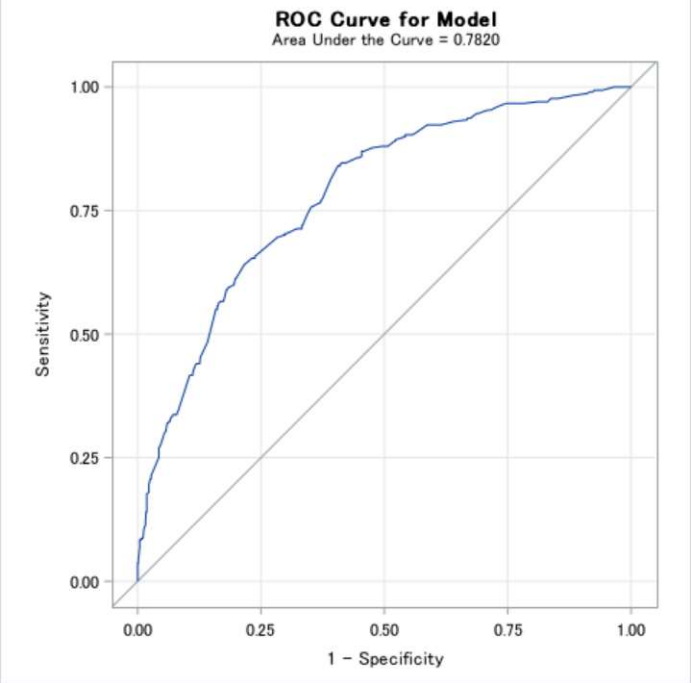
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.0005	0.5026	3.9630	0.0465
DuraCred	1	0.00162	0.0176	0.0085	0.9264
AcctBal	1	-1.1853	0.2232	28.2111	<.0001
PrevPay	1	-2.4353	0.5080	22.9814	<.0001
Savings	1	0.0635	0.3398	0.0349	0.8518
CurrEmpl	1	-0.2848	0.1824	2.4388	0.1184
DuraCred*PrevPay	1	0.0575	0.0180	10.2325	0.0014
DuraCred*Savings	1	-0.0292	0.0131	4.9883	0.0255
AcctBal*CurrEmpl	1	-1.0170	0.4145	6.0198	0.0141

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	77.7	Somers' D	0.564
Percent Discordant	21.3	Gamma	0.570
Percent Tied	1.1	Tau-a	0.237
Pairs	210000	c	0.782

(b3)Provide classification table, goodness-of-fit test, ROC curve, and AUC for the model in (b2).

ANS:

Roc curve 如下，AUC 值是 0.7820，是一個 acceptable discrimination。



Classification table:

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.000	300	0	700	0	30.0	100.0	0.0	30.0	.
0.050	291	114	586	9	40.5	97.0	16.3	33.2	92.7
0.100	286	190	510	14	47.6	95.3	27.1	35.9	93.1
0.150	277	278	422	23	55.5	92.3	39.7	39.6	92.4
0.200	267	321	379	33	58.8	89.0	45.9	41.3	90.7
0.250	253	386	314	47	63.9	84.3	55.1	44.6	89.1
0.300	211	467	233	89	67.8	70.3	66.7	47.5	84.0
0.350	183	560	140	117	74.3	61.0	80.0	56.7	82.7
0.400	165	586	114	135	75.1	55.0	83.7	59.1	81.3
0.450	125	613	87	175	73.8	41.7	87.6	59.0	77.8
0.500	104	640	60	196	74.4	34.7	91.4	63.4	76.6
0.550	97	653	47	203	75.0	32.3	93.3	67.4	76.3
0.600	81	660	40	219	74.1	27.0	94.3	66.9	75.1
0.650	75	670	30	225	74.5	25.0	95.7	71.4	74.9

0.700	44	687	13	256	73.1	14.7	98.1	77.2	72.9
0.750	21	691	9	279	71.2	7.0	98.7	70.0	71.2
0.800	11	699	1	289	71.0	3.7	99.9	91.7	70.7
0.850	4	700	0	296	70.4	1.3	100.0	100.0	70.3
0.900	0	700	0	300	70.0	0.0	100.0	.	70.0
0.950	0	700	0	300	70.0	0.0	100.0	.	70.0
1.000	0	700	0	300	70.0	0.0	100.0	.	70.0

goodness-of-fit test: 可以發現 pr=0.0897, 略高於 alpha=0.05

Partition for the Hosmer and Lemeshow Test					
		Y = 0		Y = 1	
Group	Total	Observed	Expected	Observed	Expected
1	110	7	3.74	103	106.26
2	103	7	7.91	96	95.09
3	99	9	12.71	90	86.29
4	99	15	20.13	84	78.87
5	99	32	25.62	67	73.38
6	82	21	24.26	61	57.74
7	102	31	34.54	71	67.46
8	102	52	41.85	50	60.15
9	99	51	53.46	48	45.54
10	105	75	75.77	30	29.23

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
13.7080	8	0.0897

(c)Use stepwise selection by specifying the option "**SELECTION= Forward slentry=0.15 slstay=0.20**".

(c1)Based on the original 9 predictors, report only the model in the last step. Remove nonsignificant variables and verify that models with and without these variables do not differ significantly. What is the model at this stage?

ANS:

Final model at this stage:

Log(PI(Y=0|X) / (1- PI(Y=0|X)))= 0.4949 + 0.0391* DuraCred -1.5029 * AcctBal -1.0580* PrevPay -0.6156* Savings -0.4031* CurrEmpl -0.0115*Age

Remove nonsignificant variables 的 Beta 前後變化，可以發現只有 age 的變化超過 20%:

DuraCred 0.0337->0.0391 (16%)

Age -0.0152->-0.0115 (-24%)

AcctBal -1.4949->-1.5029 (0.53%)

PrevPay -1.0384->-1.0580 (1.88%)

Savings -0.6120->-0.6156 (0.58%)

CurrEmpl -0.4096->-0.4031 (-1.58%)

作法:

```
/*PROC LOGISTIC DATA=loan;
 MODEL Y(EVENT='0')=DuraCred CredAmnt Age AcctBal PrevPay Savings CurrEmpl Gender House;
*/
/*############################################################################*/

PROC LOGISTIC DATA=loan;
 MODEL Y(EVENT='0')=DuraCred CredAmnt Age AcctBal PrevPay Savings CurrEmpl Gender House/SELECTION= Forward slentry=0.15 s
```

以下為 model at last step 的統計數據:

Step 6. Effect Age entered:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1223.729	1036.648
SC	1228.636	1071.003
-2 Log L	1221.729	1022.648

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	199.0803	6	<.0001
Score	183.3552	6	<.0001
Wald	148.6329	6	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4949	0.3745	1.7463	0.1863
DuraCred	1	0.0391	0.00632	38.2315	<.0001
Age	1	-0.0115	0.00720	2.5555	0.1099
AcctBal	1	-1.5029	0.1849	66.0440	<.0001
PrevPay	1	-1.0580	0.2508	17.7949	<.0001
Savings	1	-0.6156	0.1669	13.6115	0.0002
CurrEmpl	1	-0.4031	0.1651	5.9616	0.0146

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
DuraCred	1.040	1.027	1.053
Age	0.989	0.975	1.003
AcctBal	0.222	0.155	0.320
PrevPay	0.347	0.212	0.568
Savings	0.540	0.390	0.749
CurrEmpl	0.668	0.484	0.924

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	77.0	Somers' D	0.541
Percent Discordant	22.9	Gamma	0.541
Percent Tied	0.0	Tau-a	0.228
Pairs	210000	c	0.771

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
2.6105	3	0.4556

Analysis of Effects Eligible for Entry			
Effect	DF	Score Chi-Square	Pr > ChiSq
CredAmnt	1	0.7772	0.3780
Gender	1	0.2646	0.6070
House	1	1.8948	0.1687

Note: No (additional) effects met the 0.15 significance level for entry into the model.

Summary of Forward Selection					
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	AcctBal	1	1	103.9648	<.0001
2	DuraCred	1	2	40.9520	<.0001
3	PrevPay	1	3	18.8406	<.0001
4	Savings	1	4	15.1634	<.0001
5	CurrEmpl	1	5	8.7626	0.0031
6	Age	1	6	2.5664	0.1092

Partition for the Hosmer and Lemeshow Test					
		Y = 0		Y = 1	
Group	Total	Observed	Expected	Observed	Expected
1	101	3	5.28	98	95.72
2	100	14	8.40	86	91.60
3	100	7	12.12	93	87.88
4	100	14	17.05	86	82.95
5	100	29	24.66	71	75.34
6	100	29	30.81	71	69.19
7	101	37	36.99	64	64.01
8	100	47	43.07	53	56.93
9	100	53	53.12	47	46.88
10	98	67	68.49	31	29.51

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.1324	8	0.2559

(c2) Add interactions between 2 variables in (c1), and fit the model. Then, remove nonsignificant variables or interactions, and verify that models with and without these variables do not differ significantly. Be cautious that if an interaction between A and B is significant, then both A and B should be kept in the model no matter being significant or not.

ANS:

Final model:

Log(PI(Y=0|X) / (1- PI(Y=0|X)))= 0.9458 +0.00313*DuraCred -1.2147* AcctBal -2.0333*PrevPay + 0.2352*Savings + 0.8370*CurrEmpl -0.0106*Age + 0.0556* DuraCred*PrevPay - 0.0295*DuraCred*Savings -0.9592*AcctBal*CurrEmpl -1.0402* PrevPay*CurrEmpl - 0.4384*Savings*CurrEmpl;

加入 interaction 前後 beta 的變化如下，可以發現變化都很大:

DuraCred 0.0391->0.00313 (-91.99%)

AcctBal -1.5029->-1.2147 (-19.17%)

PrevPay -1.0580->-2.0333 (92.18%)

Savings -0.6156->0.2352 (-138.20%)

CurrEmpl -0.4031->0.8370 (-307.64%)

Age -0.0115->-0.0106 (-7.8260%)

作法:

我們去把上一步的 variables 做各種排列組合的 interaction, 用 slentry=0.15 當作篩選標準，看看各種排列組合是否 significant。得出有 DuraCred*PrevPay ,DuraCred*Savings ,AcctBal*CurrEmpl , PrevPay*CurrEmpl , Savings*CurrEmpl, PrevPay* Savings 這些 interaction 是 significant。

篩選進來之後，我們同時加入那些 significant 的 variables 去 fit 那個 model，然後用 slstay=0.2 去做篩選，結果 PrevPay* Savings 被篩掉了。

```
PROC LOGISTIC DATA=loan;
 MODEL Y(EVENT='0')=DuraCred Age AcctBal PrevPay Savings CurrEmpl;

slentry=0.15

DuraCred*Age (x)0.84
DuraCred*AcctBal (x) 0.4
DuraCred*CurrEmpl (x) 0.72
Age*AcctBal (x) 0.45
Age*PrevPay (x) 0.1536
Age*Savings (x) 0.56
Age*CurrEmpl (x) 0.3225
AcctBal*PrevPay (x) 0.807
AcctBal*Savings (x) 0.47

AcctBal*CurrEmpl (o) 0.01
PrevPay*Savings (o) 0.1239
PrevPay*CurrEmpl (o) 0.0624
Savings*CurrEmpl (o) 0.0757
DuraCred*PrevPay (o) 0.0018
DuraCred*Savings (o) 0.0078
*/

/*
PROC LOGISTIC DATA=loan;
MODEL Y(EVENT='0')=DuraCred Age AcctBal PrevPay Savings CurrEmpl AcctBal*CurrEmpl PrevPay*Savings PrevPay*CurrEmpl Saving
*/
```

以下為最終模型的統計數據:

The SAS System

The LOGISTIC Procedure

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

Number of Observations Read	1000
Number of Observations Used	1000

Response Profile		
Ordered Value	Y	Total Frequency
1	0	300
2	1	700

Probability modeled is Y='0'.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1223.729	1019.256
SC	1228.636	1078.149
-2 Log L	1221.729	995.256

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	226.4731	11	<.0001
Score	204.6882	11	<.0001
Wald	153.8296	11	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9458	0.5634	2.8185	0.0932
DuraCred	1	0.00313	0.0176	0.0315	0.8591
AcctBal	1	-1.2147	0.2228	29.7114	<.0001
PrevPay	1	-2.0333	0.5294	14.7523	0.0001
Savings	1	0.2352	0.3547	0.4398	0.5072
CurrEmpl	1	0.8370	0.5259	2.5335	0.1115
Age	1	-0.0106	0.00734	2.0835	0.1489
DuraCred*PrevPay	1	0.0556	0.0181	9.4423	0.0021
DuraCred*Savings	1	-0.0295	0.0133	4.9316	0.0264
AcctBal*CurrEmpl	1	-0.9592	0.4214	5.1809	0.0228
PrevPay*CurrEmpl	1	-1.0402	0.5327	3.8131	0.0509
Savings*CurrEmpl	1	-0.4384	0.3530	1.5423	0.2143

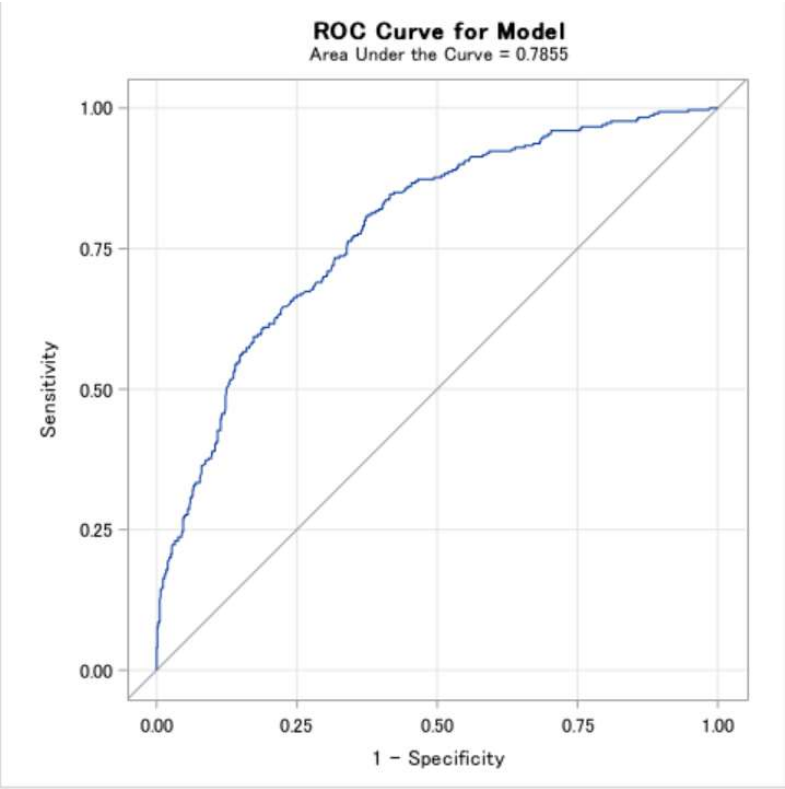
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.989	0.975	1.004

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.5	Somers' D	0.571
Percent Discordant	21.4	Gamma	0.571
Percent Tied	0.0	Tau-a	0.240
Pairs	210000	c	0.785

(c3) Provide classification table, goodness-of-fit test, ROC curve, and AUC for the model in (c2).

Ans:

ROC curve 如下，AUC 為 0.7855，是 acceptable discrimination。



goodness-of-fit test 結果如下，可以看到 pr=0.0579，略高於 0.05:

Partition for the Hosmer and Lemeshow Test					
		Y = 0		Y = 1	
Group	Total	Observed	Expected	Observed	Expected
1	100	5	2.77	95	97.23
2	100	7	6.87	93	93.13
3	100	11	12.59	89	87.41
4	100	15	18.18	85	81.82
5	100	22	24.55	78	75.45
6	100	36	30.40	64	69.60
7	100	26	35.23	74	64.77
8	100	54	41.52	46	58.48
9	100	53	55.28	47	44.72
10	100	71	72.61	29	27.39

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
15.0657	8	0.0579

Classification Table:

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.000	300	0	700	0	30.0	100.0	0.0	30.0	.
0.050	293	122	578	7	41.5	97.7	17.4	33.6	94.6
0.100	288	180	520	12	46.8	96.0	25.7	35.6	93.8
0.150	274	270	430	26	54.4	91.3	38.6	38.9	91.2
0.200	263	336	364	37	59.9	87.7	48.0	41.9	90.1
0.250	246	412	288	54	65.8	82.0	58.9	46.1	88.4
0.300	217	461	239	83	67.8	72.3	65.9	47.6	84.7
0.350	187	532	168	113	71.9	62.3	76.0	52.7	82.5
0.400	161	593	107	139	75.4	53.7	84.7	60.1	81.0
0.450	127	620	80	173	74.7	42.3	88.6	61.4	78.2
0.500	115	630	70	185	74.5	38.3	90.0	62.2	77.3
0.550	94	649	51	206	74.3	31.3	92.7	64.8	75.9
0.600	83	658	42	217	74.1	27.7	94.0	66.4	75.2
0.650	66	669	31	234	73.5	22.0	95.6	68.0	74.1
0.700	42	692	8	258	73.4	14.0	98.9	84.0	72.8
0.750	30	696	4	270	72.6	10.0	99.4	88.2	72.0
0.850	3	700	0	297	70.3	1.0	100.0	100.0	70.2
0.900	0	700	0	300	70.0	0.0	100.0	.	70.0
0.950	0	700	0	300	70.0	0.0	100.0	.	70.0
1.000	0	700	0	300	70.0	0.0	100.0	.	70.0

Problem2:The following table displays primary food choice for a sample of alligators, classified by length (≤2.3 meters, >2.3 meters) and by the lake in Florida in which they were caught.

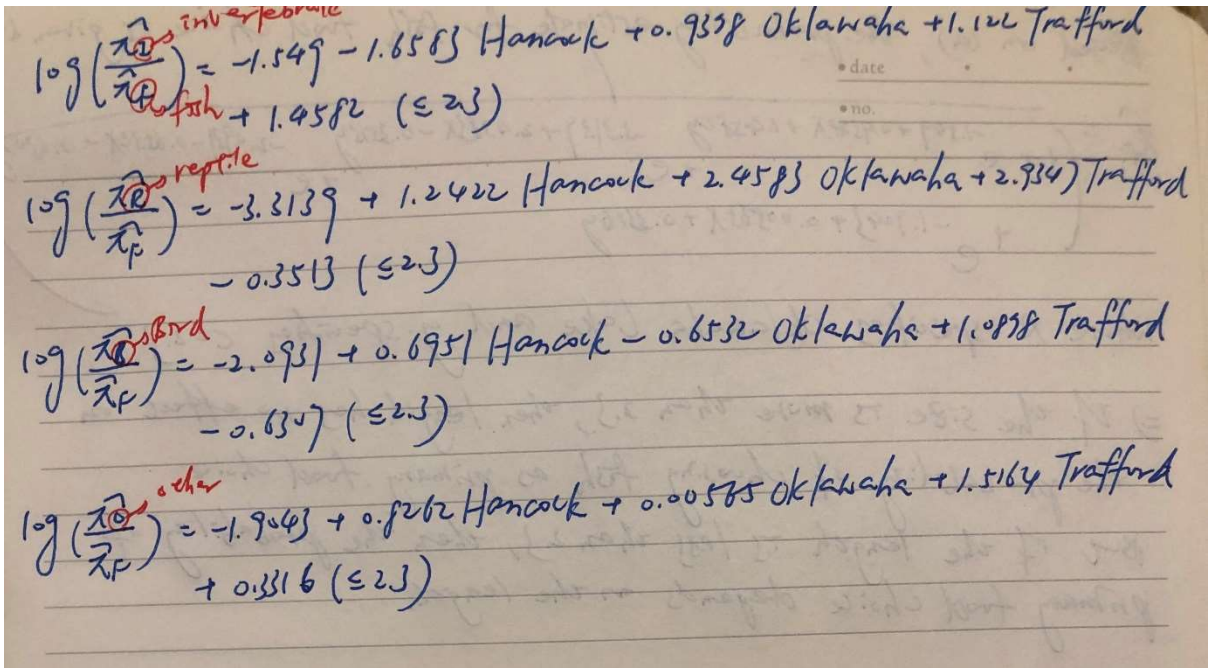
		Primary Food Choice				
Lake	Size	Fish	Invertebrate	Reptile	Bird	Other
Hancock	≤2.3	23	4	2	2	8
	>2.3	7	0	1	3	5
Oklawaha	≤2.3	5	11	1	0	3
	>2.3	13	8	6	1	0
Trafford	≤2.3	5	11	2	1	5
	>2.3	8	7	6	3	5
George	≤2.3	16	19	1	2	3
	>2.3	17	1	0	1	3

Source: Wildlife Research Laboratory, Florida Game and Fresh Water Fish Commission.

(a)Fit a model to describe effects of length and lake on primary food choice.Report the prediction equations.

Ans:

Fit 的 model equation 如下:



The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.GATOR
Response Variable	food
Number of Response Levels	5
Frequency Variable	count
Model	generalized logit
Optimization Technique	Newton-Raphson

Number of Observations Read	40
Number of Observations Used	36
Sum of Frequencies Read	219
Sum of Frequencies Used	219

Response Profile		
Ordered Value	food	Total Frequency
1	bird	13
2	fish	94
3	inverteb	61
4	other	32
5	reptile	19

Logits modeled use food='fish' as the reference category.

Note: 4 observations having nonpositive frequencies or weights were excluded since they do not contribute to the analysis.

Class Level Information				
Class	Value	Design Variables		
lake	Hancock	1	0	0
	Oklawaha	0	1	0
	Trafford	0	0	1
	George	0	0	0
size	<=2.3	1		
	>2.3	0		

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

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	17.0798	12	1.4233	0.1466
Pearson	15.0429	12	1.2536	0.2391

Number of unique profiles: 8

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	612.363	580.080
SC	625.919	647.862
-2 Log L	604.363	540.080

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	64.2826	16	<.0001
Score	57.2475	16	<.0001
Wald	49.7584	16	<.0001

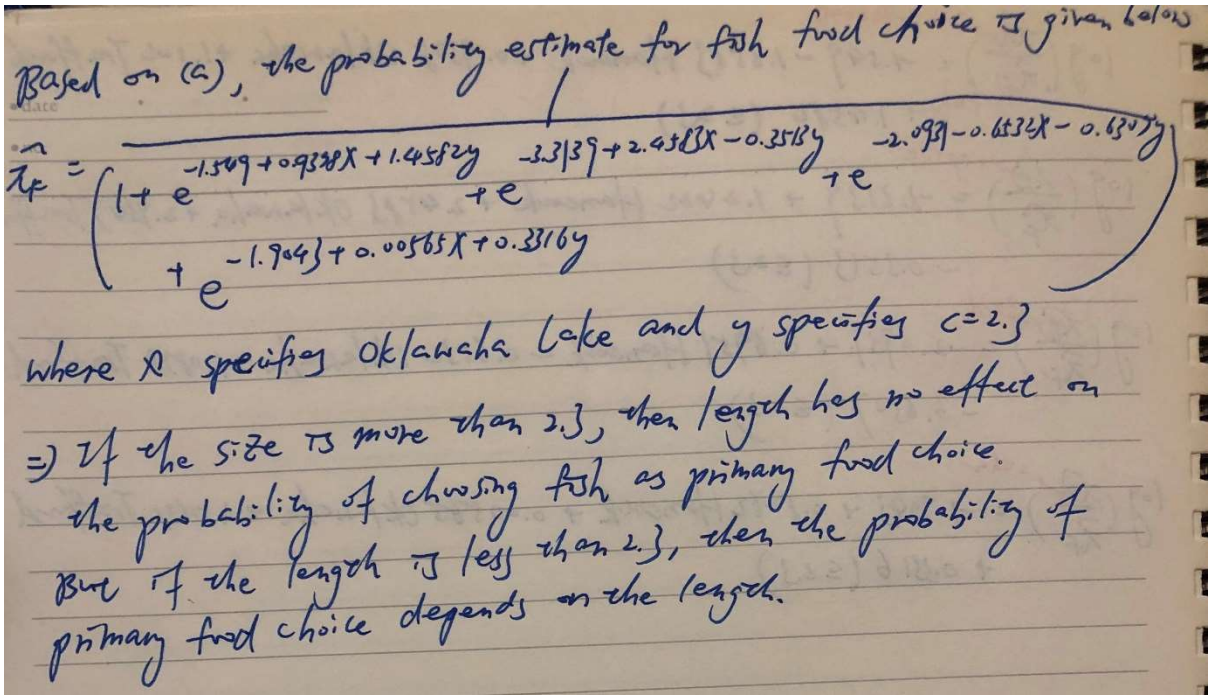
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
lake	12	35.4890	0.0004
size	4	18.7593	0.0009

Analysis of Maximum Likelihood Estimates							
Parameter		food	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		bird	1	-2.0931	0.6622	9.9894	0.0016
Intercept		inverteb	1	-1.5490	0.4249	13.2890	0.0003
Intercept		other	1	-1.9043	0.5258	13.1150	0.0003
Intercept		reptile	1	-3.3139	1.0528	9.9081	0.0016
lake	Hancock	bird	1	0.6951	0.7813	0.7916	0.3736
lake	Hancock	inverteb	1	-1.6583	0.6129	7.3216	0.0068
lake	Hancock	other	1	0.8262	0.5575	2.1959	0.1384
lake	Hancock	reptile	1	1.2422	1.1852	1.0985	0.2946
lake	Oklawaha	bird	1	-0.6532	1.2021	0.2953	0.5869
lake	Oklawaha	inverteb	1	0.9372	0.4719	3.9443	0.0470
lake	Oklawaha	other	1	0.00565	0.7766	0.0001	0.9942
lake	Oklawaha	reptile	1	2.4583	1.1179	4.8360	0.0279
lake	Trafford	bird	1	1.0878	0.8417	1.6703	0.1962
lake	Trafford	inverteb	1	1.1220	0.4905	5.2321	0.0222
lake	Trafford	other	1	1.5164	0.6214	5.9541	0.0147
lake	Trafford	reptile	1	2.9347	1.1161	6.9131	0.0086

size	<=2.3	bird	1	-0.6307	0.6425	0.9635	0.3263
size	<=2.3	inverteb	1	1.4582	0.3959	13.5634	0.0002
size	<=2.3	other	1	0.3316	0.4483	0.5471	0.4595
size	<=2.3	reptile	1	-0.3513	0.5800	0.3668	0.5448

Odds Ratio Estimates				
Effect	food	Point Estimate	95% Wald Confidence Limits	
lake Hancock vs George	bird	2.004	0.433	9.266
lake Hancock vs George	inverteb	0.190	0.057	0.633
lake Hancock vs George	other	2.285	0.766	6.814
lake Hancock vs George	reptile	3.463	0.339	35.343
lake Oklawaha vs George	bird	0.520	0.049	5.490
lake Oklawaha vs George	inverteb	2.553	1.012	6.437
lake Oklawaha vs George	other	1.006	0.219	4.608
lake Oklawaha vs George	reptile	11.685	1.306	104.508
lake Trafford vs George	bird	2.968	0.570	15.447
lake Trafford vs George	inverteb	3.071	1.174	8.032
lake Trafford vs George	other	4.556	1.348	15.400
lake Trafford vs George	reptile	18.815	2.111	167.717
size <=2.3 vs >2.3	bird	0.532	0.151	1.875
size <=2.3 vs >2.3	inverteb	4.298	1.978	9.339
size <=2.3 vs >2.3	other	1.393	0.579	3.354
size <=2.3 vs >2.3	reptile	0.704	0.226	2.194

(b) Using the fit of your model, estimate the probability that the primary food choice is “fish,” for each length in Lake Oklawaha. Interpret the effect of length.



Problem3: The following table results from a clinical trial for the treatment of small-cell lung cancer. Patients were randomly assigned to two treatment groups. The sequential therapy\administered the same combination of chemotherapeutic agents in each treatment cycle. The alternating therapy used three different combinations, alternating from cycle to cycle.

		Response to Chemotherapy			
Therapy	Gender	Progressive Disease	No Change	Partial Remission	Complete Remission
Sequential	Male	28	45	29	26
	Female	4	12	5	2
Alternating	Male	41	44	20	20
	Female	12	7	3	1

Source: Holtbrugge, W. and Schumacher, M., *Appl. Statist.*, **40**: 249–259, 1991.

(a)Fit a cumulative logit model with main effects for treatment and gender. Interpret the estimated treatment effect

ANS:

解釋:

Based on the table of analysis of maximum likelihood estimates, the estimated effect of therapy is -0.5807. The estimated odds that a sequential therapy's response is in progressive disease direction rather than the complete remission direction equal exp(-0.5807)=0.5595 times the estimated odds for alternating therapy.

The estimated effect of gender is -0.5414. The estimated odds that a male's response is in progressive disease direction rather than the complete remission direction equal

exp(-0.5414)=0.5819 times the estimated odds for females.

所 fit 的 model 統計數據如下:

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.LUNGCANCER
Response Variable	response
Number of Response Levels	4
Frequency Variable	value
Model	cumulative logit
Optimization Technique	Fisher's scoring

Number of Observations Read	16
Number of Observations Used	16
Sum of Frequencies Read	299
Sum of Frequencies Used	299

Response Profile		
Ordered Value	response	Total Frequency
1	prog	85
2	no	108
3	partial	57
4	complete	49

Probabilities modeled are cumulated over the lower Ordered Values.

Class Level Information		
Class	Value	Design Variables
therapy	sequenti	1
	alternat	0
gender	male	1
	female	0

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

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
2.9280	4	0.5699

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	5.5677	7	0.7954	0.5910
Pearson	5.3527	7	0.7647	0.6170

Number of unique profiles: 4

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	805.968	799.057
SC	817.069	817.559
-2 Log L	799.968	789.057

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10.9113	2	0.0043
Score	10.6693	2	0.0048
Wald	10.7442	2	0.0046

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
therapy	1	7.5131	0.0061
gender	1	3.3619	0.0667

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	prog	1	-0.1960	0.2947	0.4424	0.5060
Intercept	no	1	1.3713	0.3059	20.0903	<.0001
Intercept	partial	1	2.4221	0.3276	54.6609	<.0001
therapy	sequenti	1	-0.5807	0.2119	7.5131	0.0061
gender	male	1	-0.5414	0.2953	3.3619	0.0667

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
therapy sequenti vs alternat	0.560	0.369	0.848
gender male vs female	0.582	0.326	1.038

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	39.0	Somers' D	0.149
Percent Discordant	24.1	Gamma	0.236
Percent Tied	36.9	Tau-a	0.109
Pairs	32431	c	0.575

(b) Fit the model that also contains an interaction term between treatment and gender. Interpret the interaction term by showing how the estimated treatment effect varies by gender.

ANS:

解釋如下:

Based on the table of analysis of maximum likelihood estimates, the estimated effect of interaction between therapy and gender is 0.5906. The estimated odds that a sequential therapy's response by a male is in progressive disease direction rather than the complete remission direction equal exp(0.5906)=1.8 times the estimated odds for alternating therapy's response by female.

所 fit 的 model 統計數據如下:

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.LUNGCANCER
Response Variable	response
Number of Response Levels	4
Frequency Variable	value
Model	cumulative logit
Optimization Technique	Fisher's scoring

Number of Observations Read	16
Number of Observations Used	16
Sum of Frequencies Read	299
Sum of Frequencies Used	299

Response Profile		
Ordered Value	response	Total Frequency
1	prog	85
2	no	108
3	partial	57
4	complete	49

Probabilities modeled are cumulated over the lower Ordered Values.

Class Level Information		
Class	Value	Design Variables
therapy	sequenti	1
	alternat	0
gender	male	1
	female	0

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

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
3.8245	6	0.7004

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	4.5209	6	0.7535	0.6066
Pearson	4.4151	6	0.7359	0.6207

Number of unique profiles: 4

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	805.968	800.010
SC	817.069	822.213
-2 Log L	799.968	788.010

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.9581	3	0.0075
Score	11.5405	3	0.0091
Wald	11.5767	3	0.0090

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
therapy	1	3.8490	0.0498
gender	1	4.0268	0.0448
therapy*gender	1	0.9901	0.3197

Note: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all the parameters associated with that effect are zero. Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Analysis of Maximum Likelihood Estimates							
Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	prog		1	0.0770	0.3986	0.0373	0.8468
Intercept	no		1	1.6484	0.4102	16.1462	<.0001
Intercept	partial		1	2.6978	0.4260	40.1002	<.0001
therapy	sequenti		1	-1.0786	0.5498	3.8490	0.0498
gender	male		1	-0.8646	0.4309	4.0268	0.0448
therapy*gender	sequenti	male	1	0.5906	0.5935	0.9901	0.3197

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	39.0	Somers' D	0.149
Percent Discordant	24.1	Gamma	0.236
Percent Tied	36.9	Tau-a	0.109
Pairs	32431	c	0.575

(c) Does the interaction model give a significantly better fit?

ANS:

Based on the tables in (a), (b), we compare the AIC and SC values between the model with and without interaction.

The model with the smallest AIC is considered the best. SC penalizes the number of predictors in the model and smallest SC is more desirable.

The value of AIC and SC is smallest for the model without interaction. So this model is a better fit. Thus the model with interaction doesn't give a significantly better fit than the model without interaction.