北美SAS (香港考場) SBA 質素機經 02.01.2016(ddmmyyyy) by mikeleung110

請留意，這並不是最強的機經，我想說這個世界永遠沒有最強的，只有更好更高質素的機經，我在此希望所有享用及讀過這機經的朋友，希望你們參考之時能再把我這個機經不斷不斷的改善加強，我更加想將這些機經和LEGENDS發揚光大，把分享機經的精神宣揚出去，使得日後使用的朋友在學習上更加事半功倍！

內容主要有三大部分：

Contents

Legends:

FIB=FILL IN THE BLANKS=填空題
CBSC=CHANGED BUT SAME CONCEPT=題目有變但概念大致相同
CBSA=CHANGED BUT SAME ANSWER=題目有變但相同的答案
CH=CHANGED=題目有變
MDI=MIND THE DISTRUBED ITEMS=小心干擾的項目
ANS=正確答案

Small Legends for SBA

DS=Data Set
TR=Training Data Set
TE=Testing Data Set
V=Validation Data Set
MV=Missing Value

Main Notes Legends

1. Prob=Probability
2. CO=Cut Off
3. ROC=Receiver Operating Curve
4. DS=Data Set
5. TR=Training Data Set
6. TE=Testing Data Set
7. V=Validation Data Set
8. MV=Missing Value
9. SEN=Sensitivity
10. Spec=Specificity
11. Rep=Replacement
12. RD=Random Draw
13. PV+/- =Positive/Negative Predicted Value
14. T+/- = True Positive/Negative
15. TOA+/- = Total Actual Positive/Negative
16. TOP+/- = Total Predicted Positive/Negative
17. PC/AC= Predicted Class/Actual Class
18. HA=Honest Assessment
19. QCS=Quasi-Complete Separation
20. IV=input variables
21. L=Lift
22. d=depth
23. NSD=Not significant Different
24. S=Spearman; P=Pearson

1. 變題機經

<table>
<tr><td></td><td>**變題機經(02.01.2016,ddmmyyyy) (prepared by mikeleung110)**</td></tr>
<tr><td>**Q**</td><td>**Details (updated on 02.01.2016, ddmmyyyy<<如要參考使用表格內容或作更改的話，請你標註日期的月份/日子排序，因為國內常用 mmddyyyy 跟香港的 ddmmyyyy 不一樣，很混亂，日期的標註真的很重要)**</td></tr>
<tr><td></td><td>**以下是在 65 題出到的內容，後面沒有說明的就表示一樣的內容沒有變，注意答案的選項位置可能有變化，以下我都盡量精簡說明得非常非常清楚。(讓你們見識一下何謂質素機經，沒有最強的機經，只有更好更高質素的機經！(香港是說質素，反之國內是說素質，真的是給你們玩了))**</td></tr>
<tr><td>3</td><td></td></tr>
<tr><td>4</td><td></td></tr>
<tr><td>5</td><td></td></tr>
<tr><td>6</td><td></td></tr>
<tr><td>7</td><td></td></tr>
<tr><td>8</td><td>雖沒有考但仍說 SOL:
因為在第二個程序中輸入的 SUB-DATASET 僅僅包含了 TG=1 的情況(即 EVENT 實際發生的部分)。求出的均值相當於 Sensitivity.
ANS: Sensitivity</td></tr>
<tr><td>9</td><td>CH: 舊題「A」變「B」
OLD:「A 圖是曲線；TR=90.5%；V=75.5%，B 圖是直線；TR=83%；V=78.3%」
NEW:「A 圖是直線；TR=83%；V=78.3%，B 圖是曲線；TR=90.5%；V=75.5%」
SOL: 比較模型時，主要看 V 的 ACCURACY。單看 TR 不夠，會出現 Overfitting.
ANS(NEW): Model **A**. It is simpler with higher accuracy than model **B** on validation data.</td></tr>
<tr><td>10</td><td>UNCH: 注意 200 為 PROFIT，而不是 REVENUE
SOL:
<table>
<tr><td></td><td></td><td colspan="2">Solicit</td></tr>
<tr><td></td><td></td><td>0</td><td>1</td></tr>
<tr><td rowspan="2">Purch</td><td></td><td>0</td><td>-10</td></tr>
<tr><td>1</td><td>0</td><td>200</td></tr>
</table>
ANS: Profit=(P_R>0.05)*Purch*200+(-10)*(1-purch)*(P_R>0.05)</td></tr>
<tr><td>11</td><td>雖沒有考但仍說 SOL:
HA(Honest Assessment)一定有 TR 和 V，TE 可以沒有
Summary:
V= To compare models and select and fine-tune the final model
TE=To provide an unbiased measure of assessment for the final model
TR= To build the predictive model</td></tr>
<tr><td>12</td><td>雖沒有考但仍說 SOL:
ROC 下的面積愈大，模型愈好。</td></tr>
<tr><td>13</td><td>CH: 舊題「500 PROFIT」變「500 REVENUE」</td></tr>
</table>

<table>
<tr><td></td><td>Profit=Revenue-Cost=500-50=450</td></tr>
<tr><td>14</td><td>UNCH &amp; ANS: Sensitivity & specificity are not affected by oversampling
SOL from books: Sensitivity and specificity, however, are not affected by separate sampling because they don’t depend on the proportion of each class in the sample.</td></tr>
<tr><td>15</td><td>CH:舊題「deployed has 5% event；nine times 」變「deployed has 10% event；<b>nineteen</b> times 」
SOL:
Priorevent=the probability of event in population=0.10
P_1 範圍從此求出：
<table>
<tr><td></td><td></td><td colspan="2">Solicit</td></tr>
<tr><td></td><td></td><td>0</td><td>1</td></tr>
<tr><td rowspan="2">Respond</td><td>0</td><td>0</td><td>-1</td></tr>
<tr><td>1</td><td>0</td><td><b>19</b></td></tr>
</table>
19p+(1-p)(-1)>0
20p>1
p>0.05

ANS:X=0.10 Y=0.05</td></tr>
<tr><td>16</td><td>UNCH
另外加了一條問 TR 和 V 的用處(注意 TR 和 TE 並不一樣，我考的時候背答案太快混淆了)
CAUTION: TE 是個別問的；而 TR 和 V 是兩者一起問的。
ANS: V is used to compare models and select and fine-tune the final model while TR is used to build the predictive model
MDI: V is used to compare models and select and fine-tune the final model while TR is to provide an unbiased measure of assessment for the final model.
Summary:
V= To compare models and select and fine-tune the final model
TE=To provide an unbiased measure of assessment for the final model
TR= To build the predictive model</td></tr>
<tr><td>17</td><td>UNCH:
SOL: SEN=T+/TOA+ =25/(23+25)=25/28
另外加了一題同時同 Accuracy & Error Rate
ANS: Accuracy=83/150; Error Rate=67/150
Accuracy=((T-)+(T+))/(Total Cases)
Error Rate=((F-)+(F+))/(Total Cases)
Legends & formula 的詳情在這裡不詳細說了，看我附件的簡單記法。</td></tr>
<tr><td>20</td><td></td></tr>
<tr><td>21</td><td>CH:變 FIB(填空)
FIB:hovtest</td></tr>
</table>

22	CH:多了個 MDI:XL and 2XL are only the groups are significantly different from all groups CBSA: Only XL and 2XL are not significantly different from each other.(真的要仔細地看清楚答案的每一個字眼，因為試題不僅捉 GRAMMAR 還會將句子重組，雖則這題沒有句子重組，畢竟真的要看清楚字眼，不然只懂背一個答案的話看了 MDI 還是會選錯的，有時候真的大意錯在這裡。)
23	CH:舊題「MODEL 15716」變「MODEL 7266」 ANS: SSR/SST=7266/20761=35%
24	CH: Given GLM 產生的圖，找明顯的 ASSUMPTION VIOLATION SOL: Histogram 不是正態分佈；QQ PLOT 不是斜對角線 ANS: Normality violates
25	CH:改變圖片：右邊的豎綫在陰影外面 ANS: Medium wrist size is significantly different than small wrist size. SOL: 1) Control=S 就是說用 Small size 和其他組(Medium, Large) 比較，Small 為 REF 組 2) 陰影部分 若豎綫在陰影部分之內，說明 NSD (not significantly different) 若豎綫在陰影部分之外，說明 SD
26	UNCH: 注意 TTEST 跟 VAR CLASS specifies classification variables for analysis. MODEL specifies dependent & independent variables for analysis.
28	CH: 舊「MV(Missing value)」變「Redundant」 ANS: Varclus
29	
30	
31	
32	
33	CH:舊「Including」變「Excluding/Eliminating」 ANS: Stabilize parameter estimate and decrease the risk of overfitting.
35	CH: Given plots, ask for remedial solution(改善方法) ANS: add a log transformed variable x to the existing model

36	CH:舊「Pearson」變「Spearman」
	ANS: OUTS=
	OUTH=Specifies the output DS with Hoeffding’s Statistics
	OUTK=Specifies the output DS with Kendall correlation statistics
	OUTP=Specifies the output DS with Pearson correlation statistics
	OUTS=Specifies the output DS with Spearman correlation statistics

39	CH:舊「is added」變「is excluded」 ANS: Decrease in R-sq
40	CH: Given eqt: Logit(p)=0.005income+0.004 age+…

<table>
<tr><td></td><td>For which age=30, income=unknown
ANS: unpredictable/cannot calculated 的字眼因為是 MV</td></tr>
<tr><td>41</td><td>CH:變 FIB
FIB:0.4115(4 d.p.)
R-sq=SSR/SST=33033/80265=0.4115</td></tr>
<tr><td>42</td><td></td></tr>
<tr><td>43</td><td>CH: Ask for total observation for DS (n)
SOL:
<table>
<tr><td></td><td>DF</td><td>SS</td><td>MS</td></tr>
<tr><td>Regression/Model</td><td>K</td><td>SSR</td><td>MSR=SSR/k</td></tr>
<tr><td>Error</td><td>n-k-1</td><td>SSE</td><td>MSE=SSE/(n-k-1)</td></tr>
<tr><td>Total</td><td>n-1</td><td>SST</td><td></td></tr>
</table>
ANS: n-1=99>>>n=100</td></tr>
<tr><td>44</td><td></td></tr>
<tr><td>45</td><td></td></tr>
<tr><td>46</td><td></td></tr>
<tr><td>47</td><td>CH:舊「AIC」變「SBC」
SOL:跟 AIC 一樣都是愈細愈好: smaller SBC value are preferable;Lower AIC values indicates more desirable model.
ANS: 選 SBC VALUE=63 裡的 VAR</td></tr>
<tr><td>48</td><td></td></tr>
<tr><td>49</td><td>UNCH:注意 REG 是沒有/SOLUTION; GLM 是有/SOLUTION 跟的和 CLASS 後只有一個 VAR</td></tr>
<tr><td>50</td><td>CH:舊「Collinearity」變「influential factor」
ANS: cooksd</td></tr>
<tr><td>51</td><td>UNCH: 只多了「Salary is in 1000 units.」</td></tr>
<tr><td>52</td><td>UNCH: 注意 Return 則會得到一個新模型</td></tr>
<tr><td>53</td><td></td></tr>
<tr><td>54</td><td></td></tr>
<tr><td>55</td><td></td></tr>
<tr><td>56</td><td></td></tr>
<tr><td>57</td><td>CH:舊「Concordant」變「Discordant」
ANS: An observation with the event has <b>lower</b> predicted probability than the observation <b>without</b> the event.</td></tr>
<tr><td>58</td><td>UNCH
SOL:
1-Spec= F+ =did not & incorrectly classified=25%
SEN= T+ =did & correctly classified=85%(從 25%的 1-Spec 看 85%的 SEN)</td></tr>
<tr><td>59</td><td></td></tr>
<tr><td>60</td><td>UNCH</td></tr>
</table>

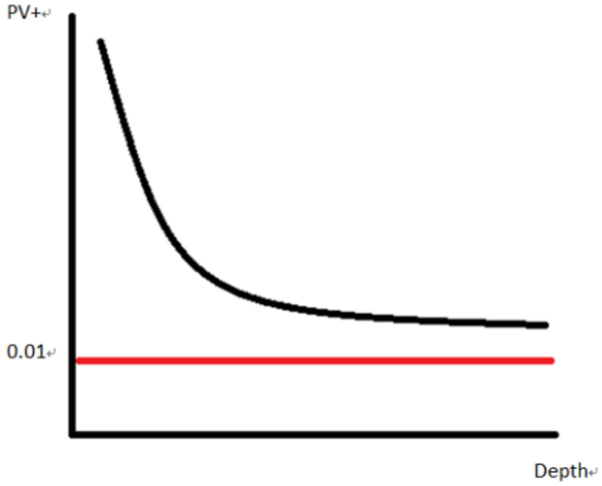
	SOL:
	A: p= 1/(1+e^-D)
	B: As O=e^D, therefore
	p=o/(1+o)
	p=e^D/(1+e^D)
	p=1/(e^-D+1)
	p= 1/(1+e^-D)

61	CH: Which statistics is better model if larger? ANS: Adj-R-sq
64	雖則沒考，但是都說一下 SOL As P value for HOME is significant so being deleted. Greatest importance=Greatest ABSOLUTE value of the estimate(Last Column) Least importance=Smallest ABSOLUTE value of the estimate(Last Column) ANS: Greatest DOWN_AMT; Least CASH
65	

2. 新題庫

		新題庫 **updated on 02.01.2016(ddmmyyyy) (prepared by mikeleung110)**
Q	**Real Q**	**Details (updated by 02.01.2016, ddmmyyyy<<如要參考使用表格內容或作更改的話，請你標註日期的月份/日子排序，因為國內常用 mmddyyyy 跟香港的 ddmmyyyy 不一樣，很混亂，日期的標註真的很重要); REAL Q 是真正考試的排序次序**
		(讓你們見識一下何謂質素機經，沒有最強的機經，只有更好更高質素的機經！(香港是說質素，反之國內是說素質，真的是給你們玩了))
1	4	Given a eqt Logit(p)=0.005income+0.004 age+… For which age=30, income=unknown ANS: 類似 cannot calculated 的字眼，記不清楚了
2	7	There is N observation data, k parameter variables, and a categorical variable with 20 levels. How many additional parameters variables are added to the model? A 20 B 19 C N-1 D k+20 ANS:A(Not Sure)
3	9	Given a logits plots, ask the remedial solution ANS: add a log transformed variable x to the existing model
4	11	What is the relationship with the correlation of the coefficient of pearson between variables? (類似問這些，主是問 PEARSON CORREALTION) A linear & monotonic correlation between variables B non-linear & monotonic correlation between variables C linear & non-monotonic correlation between variables D non-linear & non-monotonic correlation between variables ANS:A(should be)
5	12	Which is the improper use of LOGISTIC proc? A ranked of likelihood of the default of the loan B predict WHEN customers to buy a house within one-six month C predict WHICH customer to use the internet to buy a house within six month D predict WHICH customer to refinance the mortgage within one month 因為印象中 LOGISTIC 老是在 RANK 什麼和排序東西的,所以 RANK, WHICH*2 相同的我都排除了，剩下的就是 WHEN 是最另類的了。選 B(不確定)

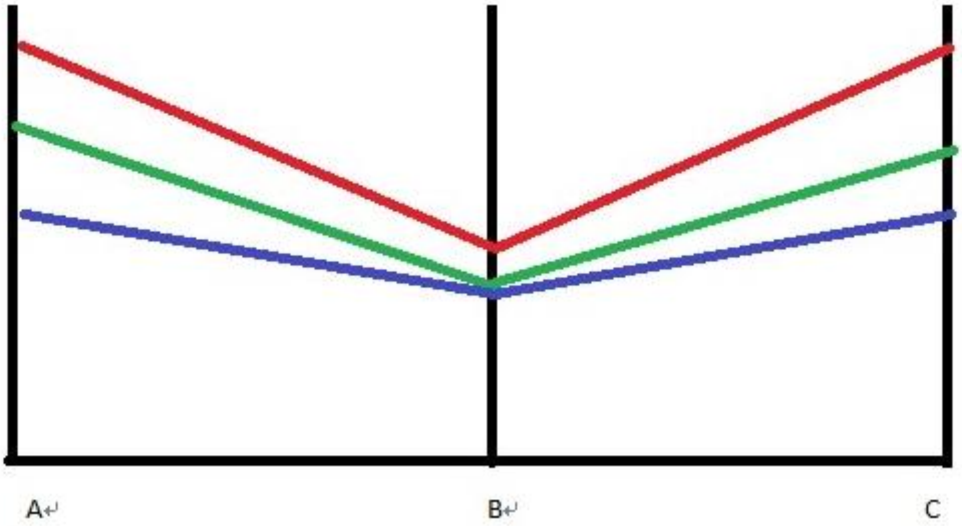
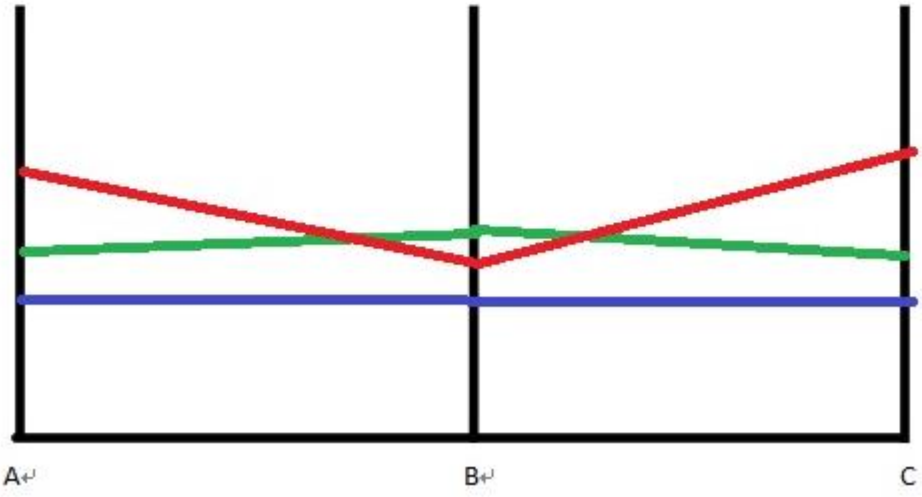
6	17	Given variables
		Gender(F,M)
		INCOME(Low, Medium, High)
		Age

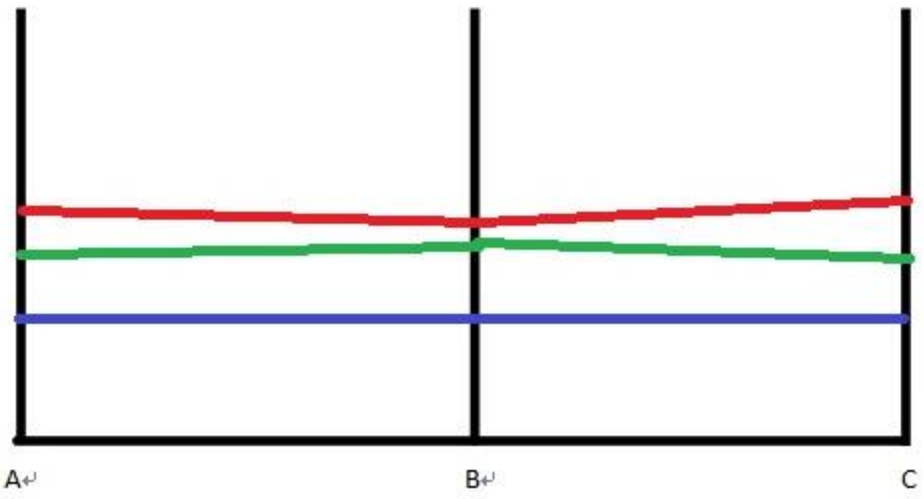
		It is required to use High income to compare with the two income level, and also take account with age and gender Which code is used? A Class Gender(“F”) Income(“high”)/param=ref Model=Gender income age B Class Gender Income Model=Gender income age C Model=Gender income age D Class Gender Income/param=ref Model=Gender income age 前人的機經只是提到了選有 param=, ref= ，這裡有兩個選，前人就他媽的沒有提到其他了，我想因為題目問用 HIGH COMPARED WITH TWO INCOME，所以選了 A (不確定) NOTES: PARAM=option in CLASS statement specifies the parameterization method for classification variables REF=option specifies the ref lv
7	37	Given a gain chart, What is the use of the reference line(red)?  A prior event rate B False Positive of probability C False Negative of probability

		D proportional cases which cannot be classified ANS: A(Not Sure)
8	41	Given one final output, and one ANOVA with 3 variables for 3 p-values, which is correct? TABLE 1: Final output 的 variable “size” p-value=0.192 (這個數是一定的) TABLE2: 3-p-values tables: P-value Intercept 0.004(不確定這數是否在這一列) Size S 0.380(不確定這數是否在這一列) Size M 0.192(不確定這數是否在這一列) Size L xxxxx(不確定這數是否在這一列) A Significant difference between variable “M” & “S” as P=0.004 B Effect is not significant for “size” due to the effect as P=0.380 C Effect is significant for “size” due to the effect as P=0.192 D NOT significant difference between variable “M” & “S” as P=0.380 我選 C (不確定)
9	47	How to generate roc curve Proc 後面都有一段 CODE 的，抱歉記不清楚了 A proc reg data B proc roc C proc genmod D proc logistic data=xxxx; ANS:D

10	52	忘記了問什麼啦，只記了答案
		A proc surveyselect data=frame out=sample sampsize=800 outall;
		B proc surveyselect data=frame out=sample sampsize=(800) outall;
		C proc surveyselect data=frame out=sample sampsize=800;
		D proc surveyselect data=frame out=sample sampsize=(800);
		這一條他媽的原來前人的機經有提到但我忘記了掙扎了該有 OUTALL 還是沒有，選了 C 是錯的，正確是選 A SAMP 是沒有括號的和要有 OUTALL
		ANS:A

11	53	~Q22 變題 A Only XL and 2XL are not significantly different from each other. B XL and 2XL are only the groups are significantly different from all groups ANS:A
12	60	Interaction between webpage(A,B,C) and income(L,M,H) An analyst claimed that there is a great interaction between variable webpage with HIGH income, so which plot of interaction graphs indicates the results?

		Low income 🟦 Medium income 🟩 High income 🟥 A  B  C

		 D 忘記了 ANS: B(Not Sure)
13		Large different between performance on TR & TE usually indicates overfitting
14		Accuracy=((T-)+(T+))/(Total Cases) Error Rate=((F-)+(F+))/(Total Cases) Two items are asked in the same question.
15		VIF>10 presence of strong collinearity in the model VIF<10 not a problem of collinearity in the model 題中 VIF 顯示 FULL NAME=Variance Inflation Factor
16		Which two follow Hierarchy principle=single? A Model= Region Campaign B Model= Region*Campaign C Model= Region Region*Campaign D Model= Region Campaign Region*Campaign ANS:A,C Notes: Hierarchy=single indicates only 1 effect can enter or leave the model at one time.

3. 65題主要詳解 by mikeleung110 (請先看Main Notes Legends)

<table>
<tr><td></td><td></td><td>**65 題主要詳解 updated on 02.01.2016(ddmmyyyy) (prepared by mikeleung110)**</td></tr>
<tr><td>**Q**</td><td>**Bk Pages**</td><td>**Book Pages LP=Logistic Book; AP=ANOVA Book; Details:**</td></tr>
<tr><td>1</td><td>LP172(4-34)</td><td>As more along the curve, prob CO changes,
CO increase, increases Cases allowed to Class 1(Class 0 if CO decreases), SEN increase, spec decrease. (ALL vice versa, 即 decrease 所有東西的符號都調轉)</td></tr>
<tr><td>2</td><td>LP144(4-6)</td><td>HA: Split Data into TR+V, V will be used for assessment, results of the analysis on TR need to be applied to V, not recalculated.</td></tr>
<tr><td>3</td><td>LP176(4-38)</td><td>OUTROC=option creates an output DS with SEN(_SENSIT_) and one minus spec(_1MSPEC_) calculated for full range of CO prob</td></tr>
<tr><td>4</td><td></td><td></td></tr>
<tr><td>5</td><td>LP145(4-7);
LP264(A-36)</td><td>SMPRATE=option specifies what portion of develop DS should be selected.
OUTALL=used to return the initial DS augmented by a flag to indicated selection in the sample.</td></tr>
<tr><td>6</td><td>LP183-184
(4-45-46)</td><td></td></tr>
<tr><td>7</td><td>LP183(4-45)</td><td>Improvement: Add L=1 (Base Line)
Restrict to focus the region by 0.005<d&lt;0.5</td></tr>
<tr><td>8</td><td></td><td>TG=1 ：EVENT 發生，求均值，即 SEN</td></tr>
<tr><td>9</td><td></td><td>SOL: 比較模型時，主要看 V 的 ACCURACY。單看 TR 不夠，會出現 Overfitting.</td></tr>
<tr><td>10</td><td>LP194(4-56)</td><td>SOL:
<table>
<tr><td></td><td></td><td colspan="2">PC</td></tr>
<tr><td></td><td></td><td>0</td><td>1</td></tr>
<tr><td rowspan="2">AC</td><td>0</td><td>0</td><td>-1</td></tr>
<tr><td>1</td><td>0</td><td>99</td></tr>
</table>
Profit=Revenue-Cost=100-1=99
E(Profit | p_i, solicit) > E(Profit | p_i, do not solicit)
p_i*99 + (1-p_i)*(-1) &gt; p_i*0 + (1-p_i)*(0)
99*p_i -1 + p_i &gt; 0
100*p_i -1 &gt; 0
p_i &gt; 0.01.</td></tr>
<tr><td>11</td><td></td><td>HA 一定有 TR 和 V，TE 可以沒有
Summary:
V= To compare models and select and fine-tune the final model
TE=To provide an unbiased measure of assessment for the final model
TR= To build the predictive model</td></tr>
<tr><td>12</td><td></td><td>ROC 下的面積愈大，模型愈好。</td></tr>
<tr><td>13</td><td></td><td>Profit=Revenue-Cost=500-50=450</td></tr>
</table>

<table>
<tr><td>14</td><td>LP174(4-36)</td><td>ANS: Sensitivity & specificity are not affected by oversampling
SOL from books: Sensitivity and specificity, however, are not affected by separate sampling because they don’t depend on the proportion of each class in the sample.</td></tr>
<tr><td>15</td><td>LP194(4-56)</td><td>Priorevent=the probability of event in population=0.05
<table>
<tr><td></td><td></td><td colspan="2">Solicit</td></tr>
<tr><td></td><td></td><td>0</td><td>1</td></tr>
<tr><td rowspan="2">Purch</td><td></td><td>0</td><td>-1</td></tr>
<tr><td>1</td><td>0</td><td>9</td></tr>
</table>

9p+(1-p)(-1)>p*0+(1-p)*(0)
9p+(1-p)(-1)>0
p>0.1</td></tr>
<tr><td>16</td><td></td><td>Summary:
V= To compare models and select and fine-tune the final model
TE=To provide an unbiased measure of assessment for the final model
TR= To build the predictive model</td></tr>
<tr><td>17</td><td>LP170-171
(4-32-33)</td><td>
<table>
<tr><td rowspan="2" colspan="2"></td><td colspan="2">PC</td><td rowspan="2"></td></tr>
<tr><td>0</td><td>1</td></tr>
<tr><td rowspan="2">AC</td><td>0</td><td>T-</td><td>F+</td><td>TOA-</td></tr>
<tr><td>1</td><td>F-</td><td>T+</td><td>TOA+</td></tr>
<tr><td colspan="2"></td><td>TOP-</td><td>TOP+</td><td></td></tr>
</table>

1) Accuracy=((T-)+(T+))/(Total Cases)
2) Error Rate=((F-)+(F+))/(Total Cases)
3) PV+ = T+/TOP+
4) PV- = T+/TOP-
5) SEN=T+/TOA+
6) Spec=T-/TOA-
7) TOA+ = (F-)+(T+)
8) TOA- = (F+)+(T-)
9) TOP+ = (F+)+(T+)
10) TOP- = (F-)+(T-)
11) Total Cases= Overall Sum= T+F= (F+)+(F-)+(T-)+(T+)
這樣寫法很容易明了吧？比書本上的一堆很冗贅的句子好得非常多吧？
希望這樣的簡稱能夠流芳百世……</td></tr>
<tr><td>18</td><td></td><td>V 最重要(記法:V=VICTORY)</td></tr>
<tr><td>19</td><td></td><td></td></tr>
</table>

20		An interaction occurs when change lv of one factor result in change different between lvs of other factors.
21		
22		
23		ANS: SSR/SST=15716/20761=76%
24		
25		1) Control=S 就是說用 Small size 和其他組(Medium, Large) 比較，Small 為 REF 組 2) 陰影部分 若豎綫在陰影部分之內，說明 NSD (not significantly different) 若豎綫在陰影部分之外，說明 SD
26		注意 TTEST 跟 VAR CLASS specifies classification variables for analysis. MODEL specifies dependent & independent variables for analysis.
27		看答案已解
28		P84(3-28) VARCLUS: eliminate redundant dimensions which related to principal components analysis. P166(4-28) STDIZE: 1) Output a DS than contains the relevant info about the imputed values for every input 2) Impute TR in V P146(4-8)~Q32 CLUSTER 1) Perform Greenacre's Corr. Analysis 2) Group those lvs together P245 (A-17)
29	LP107(3-51)	A useful plot to detect non-linear relationship is plot of empirical logits Scatter Plot: In regression analysis, standard practice to examine scatter plots of target verus each input variable)
30		
31		QCS affects the convergence of estimation algorithm.
32	LP134(3-78)	If there are nominal input variables with numerous lv, Lvs should be collapsed to reduce likelihood of QCS & reduce redundancy among lv (use clustering)
33		
34	LP70(3-14)	QCS occurs when lv of categorical input has a target event rate of 0% or 100%. If QCS occurs, 1) One of logits-→infinite 2) LV estimate of that coefficient-→infinite

		3) Affect convergence of estimate algorithm
35		Logistic reg 是和 Log-odds 有關的 and continuous

36		Found from Google & SAS(書是找不到的)
		OUTH=Specifies the output DS with Hoeffding’s Statistics
		OUTK=Specifies the output DS with Kendall correlation statistics
		OUTP=Specifies the output DS with Pearson correlation statistics
		OUTS=Specifies the output DS with Spearman correlation statistics

37		

38		Spearman(S) vs Pearson(P)
		S=use ranks of data (記法：Spear=矛，用一枝矛刺穿所有已排好的東西吧)
		S=computed on ranks→depicts monotonic relationships
		P=use observed values when variables is numeric (一個 RANK 一個 NUM，同時記了兩個，記了 S=RANK，剩下的推理都知道是 P=NUM 吧)
		P=on true values→depicts linear relationships
		S can be interpreted as P correlation between ranks on Variable X and ranks on variable Y.

39		Addition of predicted variables= increase R-sq
40		Error~i.i.d. N(0, constant variance) (i.e. Error is independent and identically disturbed with normal distribution of 0 mean & constant variance, 夠清楚了吧？)
41		R-sq=1-(SSE/SST)=SSR/SST=33033/80265=0.4115
42		
43		The adjusted R^2 is the R^2 that is adjusted for the number of parameters in the model or it takes account the number of terms in the model
44 45		SLENTRY=specifies the significant lv for **entry** in model used in **FORWARD** and STEPWISE SLSTAY= specifies the significant lv for **staying** in model used in **BACKWARD** and STEPWISE. SLS should be in the range of (0,1) Default values: FORWARD=0.5 BACKWARD=0.1 STEPWISE=0.15
46		If Y=b0+b1X1+…+bnXn X1=X2=…Xn=0-→Y=b0 for which b0=intercept

47		P129(3-73)
		Smaller SBC values are preferable.
		P312(5-35)
		Lower values of AIC indicates more desirable model.

48		Exclude y X1, X12→2 3 lvs→3 Interaction terms→1 Intercept b0→1 No. of parameters= 2+3+1+1=7
49		注意 REG 是沒有/SOLUTION; GLM 是有/SOLUTION 跟的和 CLASS 後只有一個 VAR
50	LP268(4-48)	PROC REG use VIF, COLLIN, COLLINOINT to assess the magnitude of collinearity problem Other notes: Collinearity Problems 1) Variance of coefficient increase, results in decrease precise estimation of parameters and predicted values 2) But no a violation of assumption 3) R-sq, F→Significantly Large 4) P-value of more than 2 variables are statistically significant (Large) Collinearity Diagnostics VIF: measure of magnitude of collinearity COLLIN: include intercept vector when analyzing X'X matrix COLLINOINT: exclude intercept vector when analyzing X'X matrix VIF>10 presence of strong collinearity in the model VIF<10 not a problem of collinearity in the model
51		
52		注意 Return 則會得到一個新模型
53		
54		
55		
56		Reference lv =intercept
57		
58		1-Spec= F+ =did not & incorrectly classified=25% SEN= T+ =did & correctly classified=85%(從 25%的 1-Spec 看 85%的 SEN)
59		Correction for oversampling is simply an adjustment to intercept.

60		A: p= 1/(1+e^-D)→這是鐵定的
		B: As O=e^D, therefore
		p=o/(1+o)
		p=e^D/(1+e^D)
		p=1/(e^-D+1)
		p= 1/(1+e^-D)

61		

62		
63		
64		As P value for HOME is significant so being deleted. Greatest importance=Greatest ABSOLUTE value of the estimate(Last Column) Least importance=Smallest ABSOLUTE value of the estimate(Last Column)
65		