第三章　研究設計

本章將依文獻探討，採用最小偏差法及類神經網路做為本文的研究方法。比較最小化分類誤差或最小化個體誤差二個準則下，何者對整體誤差的降低具有較小變異。使用 Microsoft SQL Server 2000 建立實證資料庫，利用 S-plus 6.2 統計軟體撰寫程式，並以 STATISCA 6.0 類神經網路模組確認程式結果。以下第一節描述資料現況，第二、三節說明本文對研究模型的設計條件。最後以單年度的預測誤差及續保年度的收支均衡原則，做為衡量模型的標準。

3.1 資料描述與分析

3.1.1 資料描述

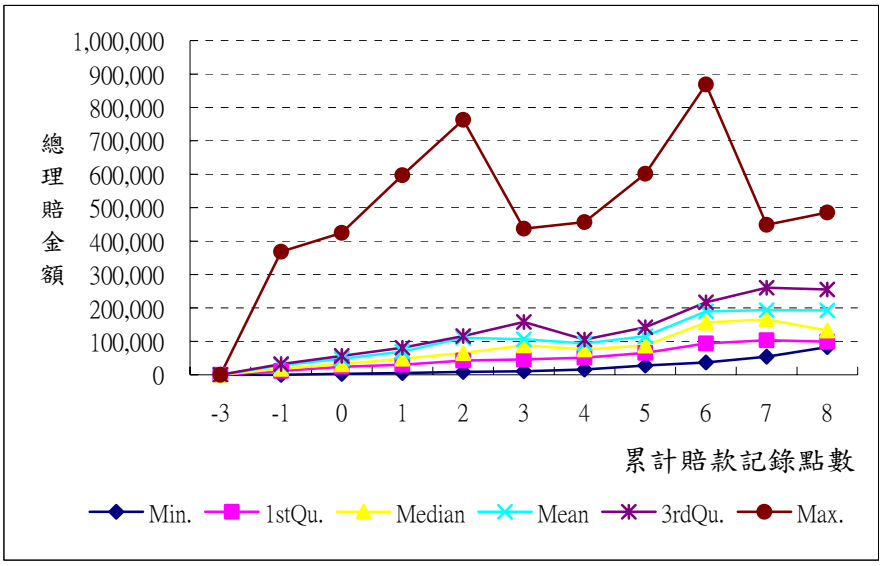
本文以 1999 年新投保汽車車體損失險甲、乙、丙三式合計，且續保至 2002 年的被保險人(共 9104 筆)為研究對象，故在此三年內加入及退出的被保險人不在研究對象之內。保費收入與賠款支出均以保單年度做為業績歸屬的基礎，限定為己車己開(駕駛人與被保險人同一人)，以理賠金額計次，整合後的資料記錄如下：

表 3-1-1: 資料量彙整表

年度	1999	2000		2001		2002	
承保資料量	30.8MB	33.2MB		34.2MB		36.2MB	
理賠資料量	9.15MB	10.3MB		10.7MB		10.8MB	
當年度個人戶(人數)	39,824	41,814		40,676		50,915	
當年度新投保個人戶(人數)	20,973	21,176		23,785		29,195	
1 年續保戶 (人數)		16,252	40.80%	17,934	42.89%	19,769	48.60%
2 年續保戶(人數)	1999 ~ 2001	11,010	27.64%	2000 ~2002		12,715	30.41%
3 年續保戶(人數)	1999 年新保戶三年續約至 2002 年					9,104	22.86%
9,104(100%)	三年皆投保甲式					1,263	13.87%
	三年皆投保乙式					2,251	24.72%
	三年皆投保丙式					3,614	39.69%
	三年中混合投保					1,976	21.70%

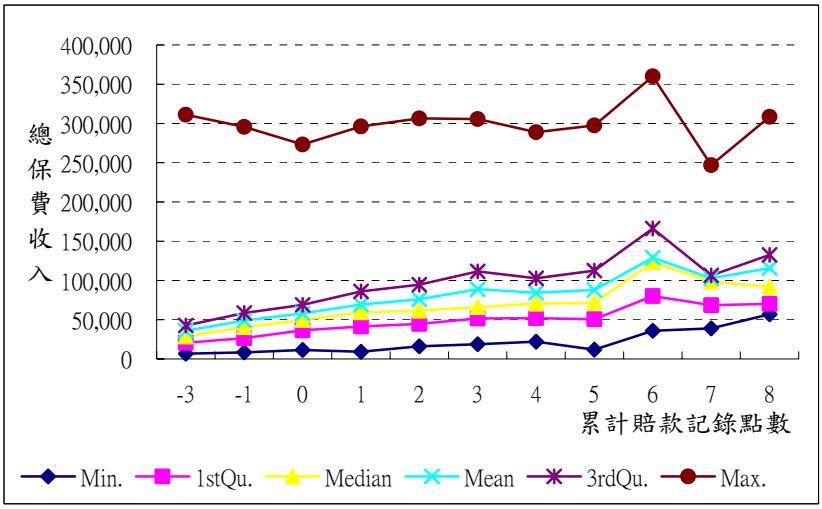
3.1.2 資料分析

圖 3-1-1: 累計賠款記錄點數理賠金額基礎統計量



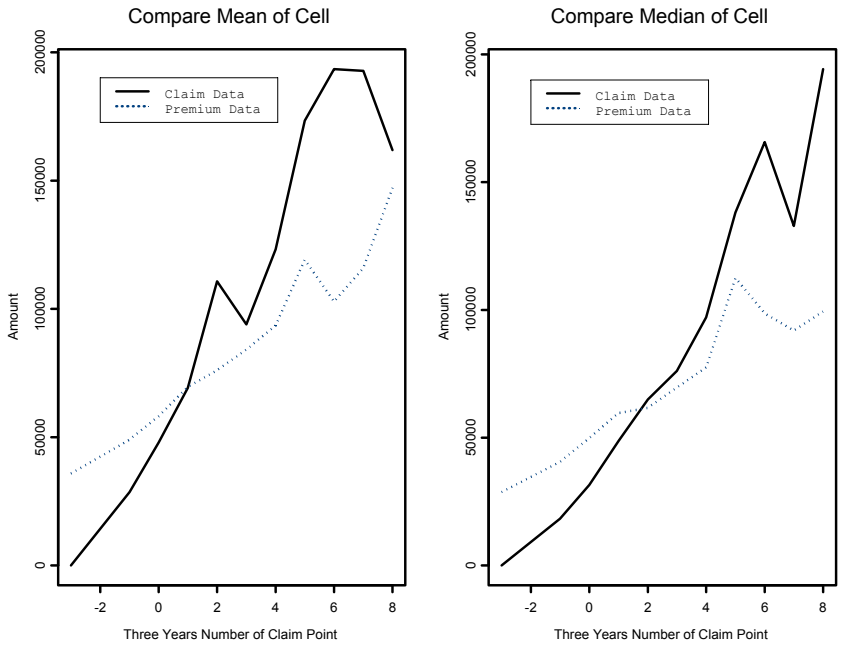
為了解資料初象，對資料進行基礎資料分析是必要的步驟。現行制度下，當年度無理賠記錄計為-1 點，發生 1 次理賠計為 0(不加費)，2 次得 1 點並以此類推，每點加減費幅度 20%。採用投保前三年的損失經驗累計賠款記錄點數，各組內賠款平均數大於中位數(圖 3-1-1)，表示理賠金額呈右偏分配，若以平均數做為衡量風險的指標易受離群值(Outlier)影響。依資料分賠款記錄點數在 1 點以下的群組，過去三年所收取的保費收入足以支付賠款成本外(例:點 1 的平均保費收入 59,631 大於平均理賠支出 48,478)，賠款記錄點數 2 點以上者，過去三年所收的保費則不足以支付賠款 (例:點 3 的平均保費收入 66,007 小於平均理賠支出 88,304)。

圖 3-1-2: 累計賠款記錄點數總保費收入基礎統計量



由於各群組中的賠款金額呈右偏分配(圖 3-1-1)，採用中位數取代平均數可減輕異常值的影響(圖 3-1-3)，各群組中的平均賠款與平均保費的差距，隨累計賠款記錄的增加而增加，由於高賠款記錄群組的保費低估，而造成損失率上升的現象。

圖 3-1-3: 2002 年投保前三年各賠款記錄點數賠款支出與保費收入比較圖



3.2 最小偏差估計模型設計

3.2.1 關於賠款記錄係數

因分類增加造成資料過少，實證上將增加計算上的困難，魏長賢(民83)觀察到純保費與賠款記錄點數呈線性關係時，即以純保費的線性估計值代替分類中被保險人的純保費。本文的實證研究時亦遇到相同的問題，本文試行另一種替代的方法：以過去三年累計賠款記錄點數所繳付的總保費為自變數，應變數為過去三年所發生的總理賠金額進行迴歸分析，迴歸係數即代表各累計賠款記錄點數群所需調整的比例。

3.2.2 關於費率代號係數

由基礎資料分析(附圖 A-5)，新車與發照年份四年以上者發生理賠的機率較高，並非為直線關係。目前費率代號係數表依貝里氏乘法模型，按發照年份與費係代號計算各組的最適純保費，並以費率代號第七等級(車價 55 萬至 60 萬)為基準，依線性關係求得最後從車因素費率係數表。為符合資料現況，費率代號係數的損失頻率與損失幅度計算平均損失成本，再依線性調整各代號群係數，再上貝里氏加型模式估計發照年份與性別的分類係數(以民國 90 年為基準)，依此計得當年度發照的各級費率代號係數表。

3.3 類神經網路模型

3.3.1 網路設計[1]

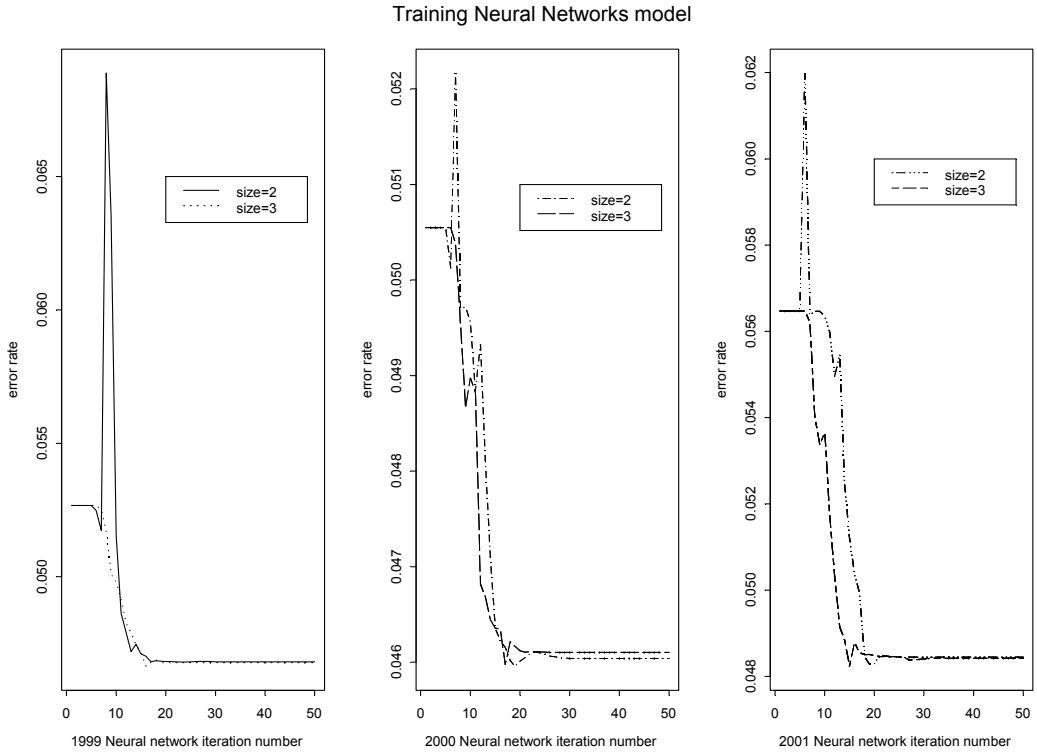
設計網路首先需決定網路架構，可自文獻中歸納出廣泛適用的規則，一般問題用一層，複雜的問題選用兩層，根據影響的重要性，使用類神經網路前，網路設計的步驟為：首先、以單一隱藏層試行，其次、

[1] 網路分析的一般性介紹，詳參中文參考文獻[16],CH18

固定網路學習率為 0.5 選定隱藏層單位元。第三、決定疊代次數。第四、固定隱藏層與疊代次數決定網路學習率。第五、條件不變下衡量是否需再加入隱藏層。最後、依試行結果參考經驗法則以決定網路變數設定。依嘗試錯誤法設定本文類神經網路為單層隱藏層、三個隱藏層單位元、學習率為 0.01、疊代次數 50 次，試行過程如下所述。

(1) 選擇疊代次數

圖 3-3-1: 隱藏層與疊代次數比較圖

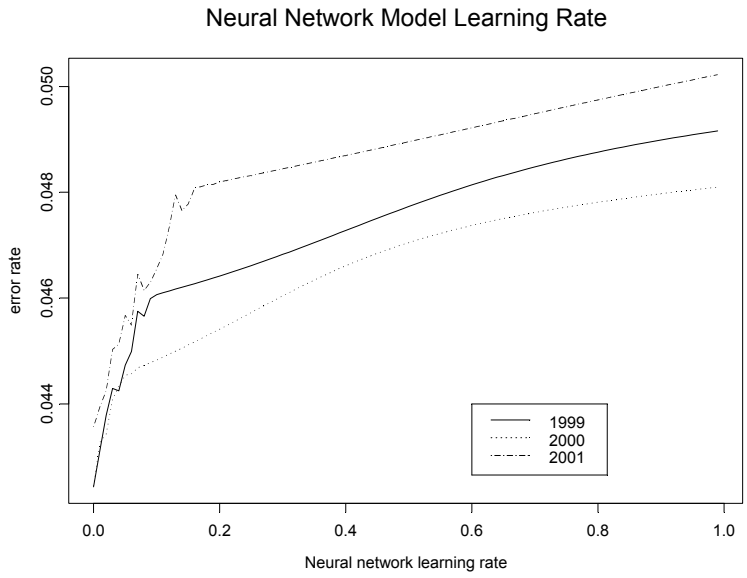


若網路訓練過早停止，則無法達到收斂的最佳狀態。輸入值採現行加減費因素，選用一層隱藏層、固定學習率為 0.5 的條件下，以疊代次數 500 次試行，比較隱藏層的單位元(圖 3-3-1)，疊代次數 32 次趨於收斂，隱藏層為二個單位元，誤差於疊代次數 10 至 13 時有明顯的不穩定跳動，選擇三個單位元則穩定地趨於收斂，依試行結果隱藏層中選擇三個單位元，並設定疊代次數為 50 次。

(2) 選擇網路學習率

學習率在類神經網路中扮演著調整權重值大小的角色，功能類似於轉彎時的方向盤，當轉角大時需多打幾圈，轉角小時只需要稍微轉動一下即可，以 0 至 1 間隔 0.001 測試學習率，誤差隨學習率上升而上升(圖 3-3-2)，為避免矯枉過正，本文將學習率設定為 0.01。

圖 3-3-2: 學習率誤差比較圖



(3) 選擇隱藏層數

本文採經驗法則(輸入與輸出層單位元數總和的一半)設定處理單位元。關於隱藏層的設定，簡單的模型由於參數少，對於目標問題可能產生較大的近似誤差，相對的，估計誤差則較小；複雜的模型由於參數較多，近似目標問題的誤差較小，但估計誤差較大，以 1999 年續保資料訓練類神經網路權重值，並以 2000、2001 年續保資料為測試資料集。使用第二層隱藏層時，訓練誤差微幅增加，測試誤差則互有增減，比較誤差變異數後，顯示複雜的模型不一定表現就比較好，應該針對問題選擇適合的模型。

表 3-3-1 :類神經網路隱藏層誤差比較表

Neural Networks Model			1999	2000	2001
Inputs	Hidden (1)	Hidden (2)	Train Error	Test Error	Test Error
7	3	0	0.054474	0.047212	0.057097
7	3	3	0.058341	0.054085	0.041936
Statistics	Data Mean	Data S.D.	Error Mean	Error S.D.	Correlation
Hidden 1	0.0206	0.0694	0.0008	0.0605	0.4912
Hidden 2	0.0206	0.0694	0.0008	0.0648	0.4795

3.3.2 類神經網路驗證

為建立網路權重，一般以隨機的方式，抽取資料的三分之二做為訓練資料集，並以剩餘的三分之一留做測試資料集，以防止網路產生過度配適(Over Fitting)的現象，意即網路過份擬合訓練資料，以至於對其他資料產生很大的預測誤差。關於網路的測試以均方誤差〔Mean Square Error〕定義如下：

$$MSE = \sqrt{\frac{\sum_{P}^{M}\sum_{j}^{N}\left(T_j^P - Y_j^P\right)^2}{M \cdot N}} \qquad (3\text{-}3\text{-}1)$$

T_j^P :第 p 個範例的第 j 個輸出單位元的目標輸出值

Y_j^P :第 p 個範例的第 j 個輸出單位元的推論輸出值

M :範例數目　N :輸出層處理單位元數

各輸入單位元對於輸出值的解釋標準，則以相關係數定義如下：

$$\rho = \frac{\left(\sum_{P}^{M}\sum_{j}^{N} T_j^P \cdot Y_j^P\right) - n \cdot \mu_t \cdot \mu_y}{(n-1) \cdot \delta_t \cdot \delta_y} \qquad (3\text{-}3\text{-}2)$$

$\mu_t\ \delta_t$:目標輸出值的平均數與標準差;y 表推論輸出值

3.3.3 類神經網路輸出值在加減費上的應用

若不對網路輸出值加以修正，以類神經網路估計費率將形同自己保

險(Self Insurance)。由於訓練網路之初，按資料逐筆修正權重值，輸出是個別被保險人經驗損失率，而不是在承保團體下的經驗費率，故對於類神經網路在加減費的應用，本文修正定義如下：

$$保險費 = BP \times \{1 + (Y - \mu(Y) \cdot Max(Y)) \times f\} = BP \times \{1 + \Delta \text{Loss Ratio} \times f\} \quad (3\text{-}3\text{-}3)$$

BP：基本保費 Y：網路輸出值 ；Max(Y)：網路輸出最大值
u(Y)：網路輸出值期望值 ；f：預期損失率（依資料分析，本文選擇 f=0.6）

意即個人經驗與平均損失率的差異需考慮分攤比例，藉由式 3-3-3 使類神經網路得將風險程度(網路輸出值)轉換為純保費基礎。例如：被保險人今年的損失經驗高於平均損失率 60%，但由於有加入風險集合(pooling)，故只需加費 0.6f，而 0.6(1-f)的比例則需由自己承擔。另外，由於調整費率的因素需與計算保險費所考量的因素一致，故本文在類神經網路的費率調整因素加入性別、年齡、發照年份、費率代號、賠款記錄等分類因素，及保險種類、自負額等純保費因素。

3.4 模型比較基礎

表 3-4-1 :模型比較表

模型 / 項目	貝里氏最小偏差法 布朗氏最小平方法	類神經網路模型
模型經驗資料期間	三年（1999~2002）	一年
最小化誤差準則	以降低分類群組的估計誤差，達到整體估計誤差的不偏性	以降低個體的估計誤差，達到整體估計誤差的不偏性
單一年度 (衡量估計誤差)	由 1999 年投保並續保至 2002 年的經驗資料，計算分類費率，據此估計 1999(2000、2001)各年度的平均理賠金額	以 1999(2000、2001)年的承保資料，預測滿期後的損失率
續保三年度 (衡量收支平衡)	以上述經驗資料計算出的分類費率，配適 1999(2000、2001)	以 1999 年資料訓練模型，預測滿期後(續保一年、二年、三年)的損失率