

جامعة القاهرة
معهد الدراسات والبحوث الإحصائية

المؤتمر السنوى الخمسين للإحصاء
وعلوم الحاسب وبحوث العمليات

إحصاء تطبيقى

٢٧-٣٠ ديسمبر ٢٠١٥

فهرس الإحصاء التتطبيقى

١٥-١	استخدام الارتباط القانونى لتحليل العلاقة بين ادارة الجودة الشاملة وادارة الازمات والكوارث (دراسة تحليلية على الشركة المصرية للاتصالات) عصام عطية عبد المنعم	١
٢٩-١٦	تقديم معلمات المجتمع بأستخدام اسلوب تقدير المساحات الصغيرة (دراسة حالة بحث القوى العاملة) راوية وجيه عبد المجيد السيد ياسمين ابراهيم	٢

**The 52nd Annual Conference on Statistics, Computer
Sciences and Operation Research**

Applied Statistics

25-27 Dec. 2017

جامعة القاهرة

Index

Applied STATISTICS

1	A General Delete-Group Jackknife Variance Estimator for Stratified Sampling Survey El Sayed, S. M. ¹ Abdel Latif, S. H. ²	1-9
2	A Generalized Linear Model for Female participation in Labor Force in Upper Egypt Reem Ismail Elsybaey* Yasmin Ibrahim	10-21
3	A Panel Data Analysis for Factors Effecting Employment on Egyptian Main Economic Sectors Naglaa A. Morad * Soad sayed Abd El rasoul	22-36
4	Some Properties of Double K- class Estimators in Linear Regression Under Multivariate t- Errors Abdel Raheim Mohamed Abdel Raheim	37-52
5	Robust Estimation of Parameters in logistic regression models Hazem Refaat Ahmed	53-71

استخدام الارتباط القانونى

تحليل العلاقة بين إدارة الجودة الشاملة وإدارة الأزمات والكوارث

(دراسة تحليلية على الشركة المصرية للاتصالات)

عصام عطية عبدالمنعم

الملخص

تعد إدارة الجودة الشاملة وفي ظل العديد من الأزمات والكوارث فى الآونة الأخيرة من المداخل العلمية الحديثة التي تمارسها المنظمات الإنتاجية والخدمية على حد سواء لتقديم مخرجات الجودة المناسبة المطلوبة وفقا لحاجات ورغبات الزبائن ، ويهدف البحث الى تحليل العلاقة بين إدارة الجودة الشاملة وأدارة الأزمات والكوارث ودور إدارة الجودة الشاملة في التقليل من اثارها وكلما استطاعت الشركة المصرية للاتصالات الساعية الى الاستمرار والنمو في ظل العديد من الأزمات والكوارث فى الآونة الأخيرة و في ظل المنافسة المتزايدة من تلبية وتوفير مستلزمات إدارة الجودة الشاملة المناسبة كلما أدى ذلك الى مواجهة الأزمات التي قد تجابهها بالاسلوب الذي يمكنها من معالجتها وأيجاد السبل الكفيلة لتفاديها مستقبلاً عند تكرارها خاصة وان الظروف الاقتصادية التي تمر بها الشركة المصرية للاتصالات تتطلب المزيد من التحسين والانتباه الى مصادر الأزمات والكوارث وادارتها بالاسلوب العلمي المناسب لتجاوزها دون اضرار ناهيك عن الخسائر المادية والمعنوية التي قد تتعرض لها الشركة جراء هذه الأزمات . وقد توصلت الدراسة الى عدة استنتاجات اعتمدت عليها التوصيات المناسبة .

طالب ماجستير جودة فى معهد الدراسات والبحوث الإحصائية

مقدمة البحث

شهدت العقود الأخيرة العديد من المتغيرات وثورة في مجال الاتصالات والمعلومات وفي ظل الازمة الاقتصادية العالمية ظهرت التحديات والمشكلات .وفي ظل مفهوم العولمة أصبح العالم اليوم ما يعرف بالقرية الصغيرة، إذ من خلال شبكة الانترنت أصبح بمقدور الفرد معرفة ما يدور في العالم على مدار الساعة وهو في غرفته أو مكتبه (الزعبي وآخرون ، ٢٠٠٩ م : ص ٤) ، فالعولمة ونظام التجارة الحرة كسرت الحواجز المركبة بين دول العالم وأصبحت الأسواق العالمية مفتوحة أمام جميع المنتجين ، هذه الأوضاع أفرزت منافسة شديدة وأصبحت ظاهرة العرض اكبر من الطلب في العديد من الأسواق ، وفي ظل هذه الظروف أصبحت الازمات التي تواجهها المنظمات اكبر من ذي قبل مما يتطلب مواجهتها وأدارتها للتقليل من اثارها (فرج وآخرون ، ١٩٩٧ م : ص ٤) ، و في ظل العديد من الازمات التي تمر بها البلاد فى تلك الالونة وبموجب ذلك اتجهت الأنظار إلى إدارة الجودة الشاملة التي تعدها الكثير من المنظمات وسيلة ناجحة لمواجهة الازمات والكوارث وادارتها والتقليل من اثارها واستخدامها كوسيلة فعالة من اجل إحداث تطورات جذرية في فلسفة إدارة الجودة ومشاركة كافة الأطراف لإجتياز الازمات والمخاطر البيئية باتجاه رضا الزبون وتلبية رغباته (العانى وآخرون ٢٠٠٢ م:ص ٦) ، إذ إن خسارة أي زبون يعني بحد ذاته مزيدا من المخاطر والتهديدات وبالتالي تعرض بقاء المنظمة للخطر والزوال .اما الجودة من وجهة نظر وفلسفة ادارة الجودة الشاملة (عقيلي ٢٠٠١ م: ص ٣١) فانها تعني تقديم سلعة او خدمة بمستوى عال من الجودة المتميزة تكون قادرة من خلالها على الوفاء باحتياجات ورغبات عملائها (منظمة ISO) بالصورة التي تتفق وتوقعاتهم ، وتحقيق الرضا والسعادة لديهم من خلال مقاييس موضوعة مسبقا لانتاج السلعة او تقديم الخدمة وإيجاد صفة التمييز فيها وفي ضوء ماتقدم تعني الجودة من وجهة نظر ادارة الجودة الشاملة مايلى "الاهتمام بكافة الجوانب وبالتفاصيل على حد سواء بهدف الوصول الى الكمال فلا مجال للصدفة والتخمين"

مشكلة البحث

حققت إدارة الجودة الشاملة وعلى مدار العقدين الماضيين نجاحات ملفته للنظر في العديد من المنظمات اليابانية والأمريكية والأوروبية معا ، (العانى وآخرون ٢٠٠٢م : ص ٢٣) ورغم أهميتها لهذه الدول والإنجازات التي حققتها في مختلف المجالات الصناعية والخدمية على حد سواء ، إلا أننا لم نجد لها مكانا متميزا في الشركة المصرية للاتصالات ، إذ تهتم الشركة المصرية للاتصالات بالايزو والحصول على شهادته (ISO 9001-2008) دون اكتراث بإدارة الجودة الشاملة التي اشتقت منه فلسفة المواصفة الدولية الايزو ، مما دفع الباحثين إلى تناول هذا الموضوع المهم من خلال دراسته وعلاقته بإدارة الازمات والكوارث كوسيلة ناجحة لضمان استمرار الشركة (شريف ، ١٩٩٨ م : ص ١٢٩) .

مما تقدم فان مشكلة البحث تبرز من خلال السؤال التالى :-

- هل هناك علاقة بين مستلزمات إدارة الجودة الشاملة وإدارة الازمات والكوارث ؟

أهمية البحث

إنبثقت أهمية الدراسة من تحديد العلاقة بين إدارة الجودة الشاملة و إدارة الأزمات و الكوارث من خلال تطبيق منظومة إدارة الجودة الشاملة بكفاءة و فاعلية مما تساعد فى مقاومة الأزمات و الكوارث و تعجل من سرعة إحتوائها و إعادة الأمن و الإستقرار و إزالة آثارها السلبية (قناوى وآخرون ، ١٩٩٧ م : ص ٦) و كذلك دراسة الأزمات و الكوارث السابقة بما يكفل فاعلية مواجهتها و إحباط آثارها مستقبلاً (الشعلان ، ٢٠٠٢ م : ص ٣٧) حيث أن التعرف على دور تطبيق إدارة الجودة الشاملة فى إدارة الأزمات و الكوارث يمكنها تلافى الآثار السلبية الناتجة عن عدم تطبيق إدارة الجودة الشاملة فى الأماكن التى سبق حدوث لها أزمات سابقة.

أهداف البحث

أن الهدف الرئيسى من هذا البحث هو ابراز أهمية دور تطبيق إدارة الجودة الشاملة بالشركة المصرية للاتصالات فى إدارة الأزمات و الكوارث والتقليل من آثارها وذلك من خلال قياس المتغيرات المستقلة المتمثلة فى مستلزمات إدارة الجودة الشاملة (صالح والسمان ، ٢٠٠٥م : ص ١١٠) ، والمتغيرات التابعة المتمثلة فى مراحل إدارة الازمة أو الكارثة ، (الشعلان ، ٢٠٠٢ م : ص ١٥٤) وذلك ببعض العبارات التى تناسب كل متغير .

منهج البحث وخطواته

سوف يتم استخدام المنهج الوصفي التحليلي عن طريق مدخل المسح باستخدام الاستبانة كأداة لجمع البيانات بالإضافة الى اساليب اخرى وفيما يلي توضيحها .

جمع البيانات

سوف يعتمد اسلوب العمل في أنجاز الجانب النظري من الدراسة الأسلوب الوصفي باستخدام المراجع العلمية والدوريات وأدبيات إدارة الإنتاج والعمليات فضلا عن شبكة الانترنت ومحتوياتها الحديثة. اما الجانب الميداني من الدراسة سوف يعتمد بشكل رئيسي على استمارات الاستبيان التي جاءت متوافقة ومستمدة من الجانب النظري من الدراسة بما يمكن تحديد مستلزمات ادارة الجودة الشاملة وادارة الازمات بالشركة المصرية للاتصالات اذ وزعت (٦٠) ستن استمارة على مسؤولي الوحدات ورؤساء الاقسام زيادة على ذلك المقابلات الشخصية والزيارات الميدانية للادارات ومواقع الشركة المختلفة بالمحافظات المختلفة . كما اعتمد اسلوب ليكرت الخاصي على تحليل الاستمارة وفضلا عن الارتباط بين متغيرات الدراسة الى جانب التكررات والنسب المئوية والمتوسطات والانحرافات المعيارية لوصف وتشخيص متغيرات الدراسة

متغيرات الدراسة

حيث تلاحظ ان متغيرات الدراسة ممثلة في الاستبيان والذي يوضح مجموعتين من المتغيرات وهما متغيرات المجموعة الأولى (X) الخاصة بمستلزمات إدارة الجودة الشاملة ومتغيرات المجموعة الثانية (Y) الخاصة بمرحلة إدارة الازمة وفضلا عن ايضاحها في استمارة الاستبيان وكيفية قياس كل متغير فيها .

الاسلوب البحثي

يستخدم هذا البحث اسلوب تحليل الارتباط القانوني Canonical Correlation والذي تقدم به Hotelling (1936)، حيث يهتم هذا الاسلوب بقياس التلازم أو الارتباط بين مجموعتين من المتغيرات. يفترض التحليل وجود مجموعتين من المتغيرات ، المجموعة الاولى ولترمز لها ب X وتشمل عدد P من المتغيرات بينما المجموعة الاخرى ورمز لها ب Y وتشمل عدد q من المتغيرات بحيث أن كلا من q, P اكبر من ٢ ، ولنفرض أن $p < q$. وعلى ذلك يمكن كتابة مصفوفة التغاير بين المتغيرات في صورتها المعيارية R بين الفئتين على النحو التالي :

$$R = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{yx} & R_{yy} \end{bmatrix}$$

حيث تمثل R_{xx} مصفوفة الارتباطات الداخلية بين متغيرات المجموعة X ، وبالمثل R_{yy} تعبر عن مصفوفة الارتباطات الداخلية بين المتغيرات المجموعة Y ، فى حين أن $R_{xy} = R_{yx}$ تمثل الارتباطات بين متغيرات المجموعة X ومتغيرات المجموعة Y والعكس على الترتيب .

المتغيرات المقننة Canonical Variables

تعرف التوليفات الخطية (u_k, v_k) من المتغيرات محل الدراسة والتي تحقق أعلى ارتباط بالمتغيرات المقننة (C.V.) Canonical Variables ويمكن صياغة الصورة العامة لهذه المتغيرات كما يلى :

$$U_k = a_k X \quad V_k = b_k Y$$

وتحقق تلك المتغيرات الخصائص التالية :-

$$\text{Cov}(U_k, U_l) = \text{Corr}(U_k, U_l) = 0 \quad k \neq l$$

$$\text{Cov}(V_k, V_l) = \text{Corr}(V_k, V_l) = 0 \quad k \neq l$$

$$\text{Cov}(U_k, V_l) = \text{Corr}(U_k, V_l) = 0 \quad k \neq l$$

حيث أن $k, l = 1, 2, 3, \dots, p \text{ or } q$.

الأوزان القانونية القياسية Standard Canonical Weights

تعرف معاملات التوليفات الخطية (a_k, b_k) والتي تدخل فى حساب المتغيرات المقننة (u_k, v_k) من المتغيرات الأساسية x, y وتحقق الخصائص السابقة بالأوزان المقننة (C.W.) وتحسب تلك الأوزان تحت قيد تساوى التباين للوحدة ، لذلك يطلق عليها أيضا الأوزان المقننة القياسية .

$$\text{Var}(U_k) = \text{Var}(V_k) = 1$$

وتحسب هذه الأوزان على النحو التالى :

$$a_k = e_k R_{11}^{-1/2} \quad b_k = f_k R_{22}^{-1/2}$$

حيث أن e_k يمثل المتجه الكامن eigenvector للمصفوفة التالية

$$R_{11}^{-\frac{1}{2}} R_{12} R_{22}^{-1} R_{21} R_{11}^{-\frac{1}{2}}$$

بينما f_k تمثل المتجه الكامن للمصفوفة التالية ،

$$R_{22}^{-\frac{1}{2}} R_{21} R_{11}^{-1} R_{12} R_{22}^{-\frac{1}{2}}$$

حيث أن $k = 1, 2, 3, \dots, p \text{ or } q$. وتدل تلك الأوزان القياسية على مقدار إسهام كل متغير فى التوليفة الخاصة بمجموعته .

معامل الارتباط القانونى Canonical Correlation Coefficients

يعرف الارتباط بين المتغيرات المقننة (U_k, V_k) بالارتباط المقنن Canonical Correlation (RC_k) ، والذي يحقق الشروط التالية ،

$$\max \text{Corr}(u_k, v_k) = RC_k$$

$$\text{Corr}(RC_k, RC_l) = 0 \quad k \neq l$$

وتكون قيم معاملات الارتباط المقنن RC_k تمثل الجذر التربيعى للقيم لمحدد المعادلة التالية

$$|R_{11}^{-1/2} R_{12} R_{22}^{-1} R_{21} R_{11}^{-1/2} - RC_k^2| = 0$$

كما يمثل RC_k^2 نسبة التباين المفسر explained variance بين أزواج المتغيرات المقننة المتناظرة (U_k, V_k) ويجب أن نذكر أن ،

$$RC_1 > RC_2 > RC_3 \dots > RC_p$$

ويستخدم متوسط مربعات معاملات الارتباط Canonical Correlation Square Mean (CCSM) كدليل على متوسط شدة العلاقة بين المجموعتين من المتغيرات محل الدراسة .

المعاملات الهيكلية Structure Coefficients

تعرف مصفوفة المعاملات الهيكلية S_y, S_x على انها معاملات الارتباط بين المتغيرات الأصلية فى x, y والمتغيرات المقننة (u, v) . لنفرض أن A تمثل مصفوفة الأوزان القياسية الخاصة بمتغيرات المجموعة x ، بينما B تمثل مصفوفة الأوزان القياسية الخاصة بمتغيرات المجموعة y على أن تأخذان الشكل التالى :

$$A = [a_1 \ a_2 \ \dots \ a_k]' , B = [b_1 \ b_2 \ \dots \ b_k]'$$

وعليه تتركب مصفوفة المعاملات الهيكلية كما يلى :

$$S_x = A * R_{xx} \quad , \quad S_y = B * R_{yy}$$

ويمثل مجموع مربعات المعاملات الهيكلية قيمة التباين المفسر بما يقابله من المتغير القانونى. وبقسمة هذا المقدار على عدد المتغيرات نحصل على نسبة التباين المفسر لتلك المتغير القانونى.

$$Pv_{x_i} = \frac{S_{x_i} S_{x_i}}{p} \quad , \quad Pv_{y_i} = \frac{S_{y_i} S_{y_i}}{q}$$

دليل التعددية Redundancy Index

يعرف دليل التعددية والذي تقدم به (Sterwart And Love (1968 لكلا من متغيرات المجموعة X ومتغيرات المجموعة Y على انه قياس لنسبة التباين فى Y المفسرة بواسطة التوليفة الخطية من X أو العكس . ويمكن تقديره من الصيغة التالية ،

$$Rd_{x_k} = Pv_{x_k} RC_k^2$$

$$Rd_{y_k} = Pv_{y_k} RC_k^2$$

وبجمع قيم أدلة التعددية نحصل على دليل الزيادة الكلى على النحو التالى

$$TR_{x_k} = \sum_j Rd_{x_k} \quad , \quad TR_{y_k} = \sum_j Rd_{y_k}$$

اختبار معنوية معاملات الارتباط القانونى

يعد اختبار Wilk's Lambda (1939) Bartlett من اشهر الاختبارات المستخدمة لاختبار معنوية معاملات الارتباط المقنن مجتمعة . هذا الاختبار خاص بفرض العدم

$$H_0 = RC_1 = RC_2 = RC_3 = \dots = RC_P = 0$$

ويعتمد هذا الاختبار على حساب ،

$$\Lambda = (1 - RC_1) (1 - RC_2) (1 - RC_3) \dots (1 - RC_P)$$

ومن ثم تحسب الإحصائية ،

$$\chi^2 = (N - 1 - 0.5) \text{Log}(\Lambda)$$

حيث N تمثل عدد المشاهدات . وتكون درجات الحرية المناظره (pq) .

عينة الدراسة

قام الباحث بسحب عينة من مختلف المستويات الوظيفية بالادارة العامة للجودة وجميع الاقسام التابعة لها فى قطاعات الشركة المختلفة بالمحافظات وكذلك الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى قطاعات الشركة المختلفة بالمحافظات وحيث أن مجتمع الدراسة المتمثل فى الادارة العامة للجودة وجميع الاقسام التابعة لها فى قطاعات الشركة المصرية للاتصالات المختلفة بالمحافظات والبالغ عدد مفردات الدراسة من واقع سجلات شئون العاملين وهم (٧٠) مفردة وكذلك الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى قطاعات الشركة المصرية للاتصالات المختلفة بالمحافظات والبالغ عدد مفردات الدراسة من واقع سجلات شئون العاملين وهم (٥٠) مفردة . وتم سحب عينة تمثل (٥٠ %) من مجتمع الدراسة اى تمثل (٦٠) مفردة منهم (٣٥) مفردة من الادارة العامة للجودة وجميع الاقسام التابعة لها ومنهم ايضا (٢٥) مفردة من الادارة العامة للإزمات وجميع الاقسام التابعة لها كما فى (جدول ١)

جدول (١) توزيع مفردات العينة على الإدارات

البيان	العدد	%
الإدارة العامة للجودة	20	33.3
اقسام الجودة فى الفروع	15	25.0
الادارة العامة للازمات	15	25.0
اقسام الازمات والطوارئ فى الفروع	10	16.7
المجموع	60	100.0

متغيرات الدراسة والأساليب الإحصائية

قام الباحث بإستخدام الحاسب الآلى على حزم البرامج الإحصائية SPSS و SAS لتحليل البيانات التى تم الحصول عليها من خلال إستمارة الاستبيان وذلك بإستخدام التحليل العاملى لتكوين ٦ مؤشرات للجودة الشاملة X وتكوين ٥ مؤشرات لإدارة الأزمة Y جدول (٢)

جدول (٢) الأوزان الناتجة من التحليل العاطلي الداخلة في تركيب المؤشرات

العوامل (Factors)	الأوزان القياسية	المتغيرات (Variables)
دعم الإدارة العليا ← X_1	.850 .800 .725	• تلزم الإدارة العليا شركتنا بتحقيق أعلى درجات الجودة • تعد إدارة الجودة الشاملة هدف جوهري لإدارة شركتنا • تعمل إدارة شركتنا على توفير منتجات بالجودة المناسبة .
مشاركة كافة العاملين في الشركة ← X_2	.841 .879 .844	• تدعم إدارة شركتنا مشاركة كافة العاملين في تحقيق الجودة المناسبة. • نمنح شركتنا لكافة العاملين الحوافز لمشاركتهم في تحقيق الجودة المناسبة • يشارك كافة عاملينا في تحقيق الجودة بدءا بالتصميم وانتهاء بتصميم
التحسين المستمر ← X_3	.809 .415 .812	• تعمل شركتنا باستمرار على تحسين جودة منتجاتها. • تطوير وتحسين جودة منتجاتنا يعد دافع لتحقيق الجودة المطلوبة. • التحسين المستمر لمنتجاتنا يعد هدفا مركزيا لكافة الأطراف
التركيز على الزبون ← X_4	.741 .893 .807	• تضع شركتنا أهدافها الإنتاجية بناءا على طلبات الزبون. • تقديم شركتنا نجاحها من خلال تلبية رغبات زبائننا المستمرة • تبحث شركتنا باستمرار على رغبات وتفضيلات الزبائن المتجددة .
التدريب والتعليم ← X_5	.767 .911 .674	• تشرك شركتنا عاملينا في دورات لتطوير منتجاتها • تعتمد شركتنا أسلوب التدريب المستمر لكافة اعاملينا. • يحمل عاملينا مؤهلات علمية وتقنية متخصصة لتحقيق الجودة المناسبة
التحفيز المستمر ← X_6	.841 .874 .819	• برنامجنا التحفيزية مبنية على أساس تحقيق أعلى جودة مطلوبة. • تعتمد شركتنا برنامج تحفيز الإبداع في تحقيق الجودة • تتابع شركتنا الجودة من خلال الحوافز الإيجابية والسلبية المناسبة

مرحلة اكتشاف إشارات الإنذار المبكر Y_1	.734 .885 .874 .831 .801	• يتوفر لدى الشركة قسم خاص، من مهامه رصد مؤشرات وقوع الأزمات. • تولي الإدارة العليا في الشركة اهتماما ودعمًا لرصد مؤشرات حدوث الأزمات. • يتم مسح بيئة عمل الشركة بصورة شاملة للتعرف على مؤشرات احتمال حدوث الأزمة. • تهتم إدارة الشركة بعمليات تصنيف وتبويب مؤشرات حدوث الأزمات. • تهتم إدارة الشركة بالتدريب المستمر لطواقم جمع وتحليل مؤشرات حدوث الأزمات.
مرحلة الاستعداد والوقاية Y_2	.828 .816 .871 .816 .715	• يتم تشكيل فرق مختلفة ومتعددة لحل العديد من الأزمات المحتملة في الشركة • يتوفر الدعم المناسب للفريق الذي يقوم بالتشخيص والتخطيط للأزمات المحتملة. • يسهل الحصول على الإمكانات المادية والبشرية والتقنية والمعلوماتية المطلوبة من الإدارات والأقسام الأخرى عند الحاجة لها من أجل التعامل مع الأزمات. • تتوفر برامج وخطط كافية وجاهزة لإدارة الأزمات في الشركة ويتم العمل على مراجعتها وتطويرها باستمرار. • توجد اتفاقيات ثنائية في إدارة الأزمات مع المؤسسات والجهات الحكومية ذات العلاقة.
مرحلة احتواء الأضرار Y_3	.735 .876 .857 .813 .763	• يؤخذ عامل الوقت عند التعامل مع الأزمات بعين الاعتبار وبدقة مناسبة. • تقوم إدارة الشركة بالاستجابة الفورية لاحتواء الأزمة عن طريق توزيع المهام وتحديد الصلاحيات بفترة قصيرة ومناسبة عند حدوث الأزمة. • يتم التأثير في الأحداث واستخدام إجراءات الطوارئ التي تقلل وتحد من الأضرار التي تسببها الأزمة بكفاءة. • تقوم إدارة الشركة بإعداد غرفة عمليات مناسبة ومجهزة بالتقنيات الحديثة لاحتواء أسباب وأضرار الأزمة. • هناك قدرة وسرعة مناسبة في تحريك الموارد المادية والبشرية الضرورية لاحتواء الأزمة.
مرحلة استعادة النشاط Y_4	.853	• تحافظ إدارة الشركة في ظروف الأزمات على اتخاذ الإجراءات اللازمة لمواصلة ممارسة النشاطات الاعتيادية في الشركة دون أي تأخير.

	.864 .842 .559	- تقوم إدارة الشركة بتحديد الاحتياجات اللازمة للمواقع المختلفة التي تأثرت بالأزمة لمعالجة تأثيرات الأزمة واستعادة النشاط الاعتيادي. - تعمل إدارة الشركة على اتخاذ كافة الإجراءات اللازمة للتخفيف من آثار الأزمة والحد من استمرار حدوثها. - تبادر إدارة الشركة بعمل حملات إعلامية مناسبة لجمهور المواطنين ووسائل الإعلام حول الأضرار التي سببتها الأزمة وكيف تم التعامل معها.
مرحلة التعلم ← y_5	.907 .896 .853 .855	- تستخلص إدارة الشركة بصورة فعالة الدروس والعبر من الأزمات التي واجهتها سابقا في محاولة للاستفادة منها مستقبلا. - تقوم إدارة الشركة بتقييم خطط وبرامج إدارة الأزمات السابقة بقصد تطويرها وتحسينها من أجل التعامل مع الأزمات المستقبلية. - تقوم إدارة الشركة بدمج الدروس المستفادة من النواقص والثغرات في الخطط السابقة بدقة عالية في خطط الأزمات المستقبلية. - تعمل إدارة الشركة على الاستفادة من أساليب معالجة الأزمات في الإدارات الأخرى في البلد أو في الدول الأخرى ذات التجارب المشابهة.

جدول (٢) يعرض الأوزان الناتجة من التحليل العاملي الداخلة في تركيب المؤشرات، حيث تشير النتائج الى: -

- أعلى الأوزان القياسية في مؤشر دعم الإدارة العليا (x_1) هو 850. والتي تمثل بعبارة تلتزم الإدارة العليا لشركتنا بتحقيق أعلى درجات الجودة .
- أعلى الأوزان القياسية في مؤشر مشاركة كافة العاملين في الشركة (x_2) هو 879. والتي تمثل بعبارة تمنح شركتنا لكافة العاملين الحوافز لمشاركتهم في تحقيق الجودة المناسبة .
- أعلى الأوزان القياسية في مؤشر التحسين المستمر (x_3) هو 812. والتي تمثل بعبارة التحسين المستمر لمنتجاتنا يعد هدفا مركزيا لكافة الأطراف.
- أعلى الأوزان القياسية في مؤشر التركيز على الزبون (x_4) هو 893. والتي تمثل بعبارة التحسين المستمر لمنتجاتنا يعد هدفا مركزيا لكافة الأطراف
- أعلى الأوزان القياسية في مؤشر التدريب والتعليم (x_5) هو 911. والتي تمثل بعبارة تعتمد شركتنا أسلوب التدريب المستمر لكافة العاملين.
- أعلى الأوزان القياسية في مؤشر التحفيز المستمر (x_6) هو 874. والتي تمثل بعبارة تعتمد شركتنا برنامج تحفيز الإبداع في تحقيق الجودة

- أعلى الأوزان القياسية فى مؤشر مرحلة اكتشاف إشارات الإنذار المبكر (Y_1) هو 885. والتي تمثل بعبارة تولي الإدارة العليا في الشركة اهتماما ودعمًا لرصد مؤشرات حدوث الأزمات.
- أعلى الأوزان القياسية فى مؤشر مرحلة الاستعداد والوقاية (Y_2) هو 871. والتي تمثل بعبارة يسهل الحصول على الإمكانيات المادية والبشرية والتقنية والمعلوماتية المطلوبة من الإدارات والأقسام الأخرى عند الحاجة لها من أجل التعامل مع الأزمات.
- أعلى الأوزان القياسية فى مؤشر مرحلة احتواء الأضرار (Y_3) هو 876. والتي تمثل بعبارة تقوم إدارة الشركة بالاستجابة الفورية لاحتواء الأزمة عن طريق توزيع المهام وتحديد الصلاحيات بفترة قصيرة ومناسبة عند حدوث الأزمة.
- أعلى الأوزان القياسية فى مؤشر مرحلة استعادة النشاط (Y_4) هو 864. والتي تمثل بعبارة تقوم إدارة الشركة بتحديد الاحتياجات اللازمة للمواقع المختلفة التي تأثرت بالأزمة لمعالجة تأثيرات الأزمة واستعادة النشاط الاعتيادي.
- أعلى الأوزان القياسية فى مؤشر مرحلة التعلم (Y_5) هو 907. والتي تمثل بعبارة تستخلص إدارة الشركة بصورة فعالة الدروس والعبر من الأزمات التي واجهتها سابقا في محاولة للاستفادة منها مستقبلا.

نتائج الدراسة

نستعرض فيما يلى نتائج تحليل الارتباط المقنن ، حيث يتم عرض نتائج قيم معاملات الارتباط المقنن ، الأوزان القياسية ، المعاملات الهيكلية واخيرا دليل الزيادة .

تمثل الجداول من (٣) الى (٤) نتائج تحليل الارتباط المقنن للشركة المصرية للاتصالات . فجدول (٣) يعرض الأوزان القياسية ومعاملات الارتباط المقننة بين إدارة الجودة الشاملة وإدارة الأزمات و الكوارث للشركة المصرية للاتصالات ، حيث تشير النتائج الى أن أعلى ارتباط بلغ ٠.٨٨ وهو ذات دلالة احصائية وقد بلغ (CCSM) والتي تمثل دليل لمتوسط العلاقة بين إدارة الجودة الشاملة وإدارة الأزمات والكوارث تساوى ٠.٢٣ وهو يوضح علاقة ضعيفة بين إدارة الجودة الشاملة وإدارة الأزمات والكوارث.

جدول (٣) الاوزان القياسية ومعاملات الارتباط المقننة بين إدارة الجودة الشاملة وإدارة

الأزمات و الكوارث بالشركة المصرية للإتصالات

U_k	X_1	X_2	X_3	X_4	X_5	X_6	RC_k	V_k	Y_1	Y_2	Y_3	Y_4	Y_5
u_1	.043	.312	-.011	.191	.255	.448	.883	v_1	.321	.246	.179	.241	.128
u_2	-.795	1.132	-.413	.610	.117	-.657	.436	v_2	-.997	-.963	.734	-.317	1.590
u_3	-.757	.268	1.203	-.379	-.984	.587	.330	v_3	-.709	1.425	.793	-.619	-.934
u_4	.431	.464	.778	-.194	-.146	-1.078	.233	v_4	1.242	1.098	-.977	.960	.204
u_5	-.628	-.602	.341	1.085	.159	-.164	.121	v_5	-.464	-.774	1.236	.965	-.875

* Lambda Statistic = 0.1484 with pr < 0.0001

هذا ويستكمل جدول (٤) نتائج التحليل حيث يعرض المعاملات الهيكلية ودليل التعددية بين مقاييس إدارة الأزمات والكوارث و إدارة الجودة الشاملة للشركة المصرية للإتصالات حيث توضح النتائج أن ١٥% من التباين في مقاييس إدارة الأزمات والكوارث راجع الى إدارة الجودة الشاملة وهذه النسبة تعني أن مساهمة ادارة الحودة مساهمة ضعيفة.

المراجع

العربية

- (١) الزعبي ، على والكساسبة ، حسن والسكرانه ، بلال (٢٠٠٩) " دور ادارة الجودة الشاملة فى تقليل المخاطر التسويقية فى ظل الازمة الاقتصادية العالمية فى شركات القطاع الخاص الاردنى " المؤتمر العلمى الدولى السنوى السابع ، جامعة الزرقاء ، الاردن .
- (٢) السمان، ثائر، صالح، ماجد محمد (٢٠٠٥) ادارة الجودة الشاملة، دراسة ميدانية فى المعمل الالبسة الولادية بالموصل ، المجلد ، ٢٧ العدد ٧٨.
- (٣) شريف، منى صلاح الدين(١٩٩٨) إدارة الازمات الوسيلة للبقاء . القاهرة : دار العلم للملايين .
- (٤) عقيلي ، عمر وصفي (٢٠٠١)، مدخل إلى المنهجية المتكاملة لإدارة الجودة الشاملة، الطبعة الأولى، دار وائل للنشر ، عمان، الأردن
- (٥) فرج ، ماجد إبراهيم سيد (١٩٩٧) " أهمية وجود اطر مؤسسية لمواجهة الازمات والكوارث المهددة للتنمية فى مصر "، المؤتمر الثانى لإدارة الازمات والكوارث المنعقد فى الفترة من ٢٥-٢٦ أكتوبر، القاهرة : جامعة عين شمس .
- (٦) قناوى ، فاروق وأخرون (١٩٩٧) " مواجهه لحالات الطوارئ بشركة جابكو "، المؤتمر الثانى لإدارة الازمات والكوارث المنعقد فى الفترة من ٢٥-٢٦ أكتوبر، القاهرة : جامعة عين شمس .

الأجنبية

- 7) Bartlett, M.S. (1939),"A Note On Test Of Significance In Multivariate Analysis", Proceedings Of Cambridge Philosophical Society, 35.
- 8) Hotelling. H. (1936), "Relations Between Two Sets Of Variates",Biometrika,Vol28,Pp.321-336
- 9) International Standard ISO 9001-2008: Quality Management Systems Requirements. 4rd Ed., Geneva: International Organization For Standardization, 2008-11-15.
- 10) Sterwart, D.K, And Love, W.A.(1968),"A General Canonical Correlation Index",Psychological Bulletin Vol,70,Pp.160-163.

تقدير معلمات المجتمع باستخدام أسلوب تقدير المساحات الصغيرة (دراسة حالة بحث القوى العاملة)

ياسمين إبراهيم**

راوية وجيه عبد المجيد السيد*
المستخلص:-

زادت الحاجة للإحصاءات المتعلقة بالمواقع الجغرافية الصغيرة ، حيث تكون البيانات ضعيفة جدا لدعم هذا النوع من طرق تقدير القياسية المستخدمة عادة على المستوى الوطني. وتستخدم الإحصاءات الرسمية "منطقة صغيرة" بشكل روئني لمجموعة متنوعة من الأغراض في مجموعة من الدول الأخرى حيث ان تقديرات المساحة الصغيرة تنتج بواسطة تقنيات الإحصاءات الجديدة التي تكون أكثر دقة من تقديرات المسح مباشرة وهناك تقنيات لحساب تقديرات مساحة صغيرة تجمع بين استخدام مصادر البيانات المساعدة من التعداد وبيانات المسح و السجلات الإدارية لإنتاج تقديرات لمستوى جغرافيا أكثر تفصيلا مما يمكن الحصول على تقديرات دقيقة بالاعتماد على مصادر البيانات الإدارية، حيث تستخدم تقديرات المساحات الصغيرة في تقييم مستوى الفقر للدول، ووضع السياسات العامة، وتخصيص الأموال في البرامج الحكومية المختلفة والسعى نحو تقديم خدمات أفضل من قبل الحكومة المحلية ومن الواضح أن الطلب على الإحصاءات مستوى منطقة صغيرة (الشيخة بدلا من المحافظة) يتزايد ومن هنا كانت الحاجة الى تقدير معدلات البطالة على مستوى الشياخة/ القرية بدلا من مستوى المحافظة او الجمهورية والهدف هو توافر بيانات على اقل مستوى جغرافى لمعدلات البطالة فى ظل سوء التصنيف الجيدللحالة العملية فى بيانات التعداد حيث التصنيف الجيد يتم من خلال اسئلة بحث القوى العاملة و هى تعالج القصور فى اسئلة الحالة العملية بالنسبة للتعداد ولكن لا يتم توافر تقديرات دقيقة لمعدلات البطالة على المستويات الجغرافية الصغرى من بحث القوى العاملة على الرغم من اسئلة البحث الأكثر تفصيلا للحالة العملية وذلك لصغر حجم العينة عند هذا المستوى الجغرافى مما ينتج تقديرات غير موثوق بها ولذلك تم تعويض ذلك بالاستعانة بخصائص التعداد لهذه المستويات الجغرافية الصغيرة لتعويض صغر حجم العينة و الحصول على تقديرات للمساحات الصغيرة (Small Area Estimation (SAE) موثوق بها كما سوف يتم التعرض الى اساليب تقدير SAE و التطبيق على بعضهم و المقارنة بينهم؛بالإضافة للمقارنة بين معدلات البطالة على مستوى المحافظة من بحث القوى العاملة Labor Force Survey (LFS) وتجميع التقديرات على مستوى الشياخة/ القرية باستخدام طرق SAE وذلك للتأكد من كفاءة هذه المقررات والتي تعتمد على البيانات المتاحة من كل من بحث القوى العاملة التى تعرض خصائص المستغلين و المتعطلين بالإضافة الى بيانات التعداد (من حيث الحالة العملية) ومدى الاعتماد عليها فى اتخاذ القرارات ورسم السياسات وللتعرف مدى الفجوة بين المشتغلين و المتعطلين فى التسلسل الهرمى الجغرافى وسوف تعرض هذه الورقة حالة هامة تؤثر على السياسات القومية ومدى عرض هذا البحث من تقديرات لمعدل البطالة على مستوى الشياخة او القرية وهو المستوى الذى يسبق اصغر مستوى جغرافى ومدى تأثير ذلك فى معالجة مشكلة البطالة بصورة أكثر دقة و تفصيلا للوصول الى الاسباب الجوهريه والتفصيلية للمشكلة من خلال الوصول الى اصغر مستوى جغرافى بالتطبيق على بيانات تعداد السكان لعام ٢٠٠٦ وبحث القوى العاملة لنفس العام والتطبيق على محافظة القليوبية لكونها محافظة ممثلة لخصائص الجمهورية الى حد كبير ومحاولة رسم السياسات لمواجهة هذه الظاهرة القومية . والتي تعكس الوضع الراهن لسوق العمل المصرى عند اصغر مستوى جغرافى و كيفية معالجة مشكلة البطالة والمساعدة فى رسم السياسات من الجهات المعنية بالوصول الى اقل مستوى جغرافى يعكس طبيعة المشكلة والمساهمة فى حلها.

*راوية وجيه عبد المجيد – باحث إحصائى ثالث بالجهاز المركزى للتعبئة العامة والإحصاء
**ياسمين إبراهيم _ مدرس احصاء بمعهد الدراسات والبحوث الإحصائية

تقدير معلمات المجتمع بأستخدام اسلوب تقدير المساحات الصغيرة (دراسة حالة بحث القوى العاملة)

راوية وجيه عبد المجيد السيد *

ياسمين ابراهيم **

المستخلص:-

زادت الحاجة للإحصاءات المتعلقة بالمواقع الجغرافية الصغيرة ، حيث تكون البيانات ضعيفة جدا لدعم هذا النوع من طرق تقدير القياسية المستخدمة عادة على المستوى الوطني. وتستخدم الإحصاءات الرسمية "منطقة صغيرة" بشكل روتيني لمجموعة متنوعة من الأغراض فى مجموعة من الدول الاخرى حيث ان تقديرات المساحة الصغيرة تنتج بواسطة تقنيات الإحصاءات الجديدة التي تكون اكثر دقة من تقديرات المسح مباشرة وهناك تقنيات لحساب تقديرات مساحة صغيرة تجمع بين استخدام مصادر البيانات المساعدة من التعداد و بيانات المسح و السجلات الادارية لإنتاج تقديرات لمستوى جغرافيا أكثر تفصيلا مما يمكن الحصول على تقديرات دقيقة بالاعتماد على مصادر البيانات الادارية، حيث تستخدم تقديرات المساحات الصغيرة فى تقييم مستوى الفقر للدول، ووضع السياسات العامة، وتخصيص الأموال فى البرامج الحكومية المختلفة والسعى نحو تقديم خدمات أفضل من قبل الحكومة المحلية ومن الواضح أن الطلب على الإحصاءات مستوى منطقة صغيرة (الشاخه بدلا من المحافظة) يتزايد ومن هنا كانت الحاجة الى تقدير معدلات البطالة على مستوى الشياخه / القرية بدلا من مستوى المحافظة او الجمهورية والهدف هو توافر بيانات على اقل مستوى جغرافى لمعدلات البطالة فى ظل سوء التصنيف الجيد للحالة العملية فى بيانات التعداد حيث التصنيف الجيد يتم من خلال اسئلة بحث القوى العاملة و هى تعالج القصور فى اسئلة الحالة العملية بالنسبة للتعداد ولكن لا يتم توافر تقديرات دقيقة لمعدلات البطالة على المستويات الجغرافية الصغرى من بحث القوى العاملة على الرغم من اسئلة البحث الاكثر تفصيلا للحالة العملية وذلك لصغر حجم العينة عند هذا المستوى الجغرافى مما ينتج تقديرات غير موثوق بها ولذلك تم تعويض ذلك بالاستعانة بخصائص التعداد لهذه المستويات الجغرافية الصغيرة لتعويض صغر حجم العينة و الحصول على تقديرات للمساحات الصغيرة (SAE) Small Area Estimation موثوق بها كما سوف يتم التعرض الى اساليب تقدير SAE و التطبيق على بعضهم و المقارنة بينهم ؛بالأضافة للمقارنة بين معدلات البطالة على مستوى المحافظة من بحث القوى العاملة Labor Force Survey (LFS) وتجميع التقديرات على مستوى الشياخه/ القرية بأستخدام طرق SAE وذلك للتأكد من كفاءة هذه المقدرات والتي تعتمد على البيانات المتاحة من كل من بحث القوى العاملة التى تعرض خصائص المستغلين و المتعطلين بالأضافة الى بيانات التعداد (من حيث الحالة العملية) ومدى الاعتماد عليها فى اتخاذ القرارات ورسم السياسات وللتعرف مدى الفجوة بين المشتغلين و المتعطلين فى التسلسل الهرمى الجغرافى وسوف تعرض هذه الورقة حالة هامة تؤثر على السياسات القومية ومدى عرض هذا البحث من تقديرات لمعدل البطالة على مستوى الشياخه او القرية وهو المستوى الذى يسبق اصغر مستوى جغرافى ومدى تأثير ذلك فى معالجة مشكلة البطالة بصورة اكثر دقة و تفصيلا للوصول الى الاسباب الجوهرية والتفصيلية للمشكلة من خلال الوصول الى اصغر مستوى جغرافى بالتطبيق على بيانات تعداد السكان لعام ٢٠٠٦ وبحث القوى العاملة لنفس العام والتطبيق على محافظة القليوبية لكونها محافظة ممثلة لخصائص الجمهورية الى حد كبير ومحاولة رسم السياسات لمواجهة هذه الظاهرة القومية . والتي تعكس الوضع الراهن لسوق العمل المصرى عند اصغر مستوى جغرافى و كيفية معالجة مشكلة البطالة والمساعدة فى رسم السياسات من الجهات المعنية للوصول الى اقل مستوى جغرافى يعكس طبيعة المشكلة والمساهمة فى حلها.

*راوية وجيه عبد المجيد – باحث إحصائى ثالث بالجهاز المركزى للتعبئة العامة والإحصاء

** ياسمين ابراهيم _ مدرس احصاء بمعهد الدراسات والبحوث الإحصائية

أولاً:مقدمة:

الحاجة إلى الإحصاءات الرسمية للمساعدة في تخطيط ومتابعة مشروعات التنمية أصبحت أكثر أهمية أو احتياجاً ، كما ان الدول اتجهت نحو تقديم خدمات أفضل من قبل الحكومة المحلية. وهو الواضح في زيادة الطلب على الإحصاءات على مستوى منطقة صغيرة وتقديرات المساحة الصغيرة تنتج بواسطة تقنيات الإحصاءات الجديدة التي تكون أكثر دقة من تقديرات المسح مباشرة وهناك تقنيات لحساب تقديرات مساحة صغيرة تجمع بين استخدام مصادر البيانات المساعدة من التعداد و بيانات المسح و السجلات الادارية لإنتاج تقديرات لمستوى جغرافيا أكثر تفصيلا مما يمكن الحصول على تقديرات دقيقة بالاعتماد على مصادر البيانات الادارية مما يساعد على دعم التخطيط في اتخاذ - أصبحت الحاجة للإحصاءات المتعلقة بالمواقع الجغرافية الصغيرة أكثر تزايداً، حيث تكون البيانات ضعيفة جداً لدعم هذا النوع من طرق التقدير القياسية المستخدمة عادة على المستوى الوطني. وتستخدم الإحصاءات الرسمية "منطقة صغيرة" بشكل روتيني لمجموعة متنوعة من الأغراض في مجموعة من الدول الأخرى ، بما في ذلك تقييم الرفاه الاقتصادية للدول، ووضع السياسات العامة وتخصيص الأموال في البرامج الحكومية المختلفة وذلك في ظل التطور السريع لمنهجية المسح والوكالات الحكومية المختلفة واستكشاف وسائل الجمع بين بيانات المسح الوطني مع السجلات الإدارية والتعداد لإنتاج إحصاءات مساحة صغيرة يمكن الاعتماد عليها

*من الأهمية المتزايدة لتقديرات المساحة الصغيرة سوف يتم عرض هذا الأسلوب من التقدير على حالة هامة تؤثر على السياسات القومية ومدى عرض هذا البحث لمشكلة البطالة من خلال تقديرات معدل البطالة على مستوى الشياخة او القرية وهو المستوى الذى يسبق اصغر مستوى جغرافى على مستوى الجمهورية وهى القطعة المساحية ومدى تأثير ذلك فى معالجة مشكلة البطالة بصورة أكثر دقة و تفصيلا للوصول الى الاسباب الجوهرية والتفصيلية للمشكلة من خلال الوصول الى أصغر مستوى جغرافى ومحاولة رسم السياسات لمواجهة هذه الظاهرة القومية وذلك يتم بتقدير معدل البطالة على مستوى الشياخة(حضر)/ القرية(ريف) لمحافظة القليوبية (فى كونها محافظة تعكس الى حد كبير خصائص المجتمع المصرى) وذلك من بحث القوى العاملة لعام ٢٠٠٦ و تعداد عام ٢٠٠٦ بالرغم من صغر حجم عينة مسح القوى العاملة على هذا المستوى الجغرافى حيث صغر حجم العينة لا يوفر تقديرات دقيقة ويتم معالجة هذه المشكلة من خلال البيانات المساعدة التى يتم الحصول عليها من بيانات التعداد (تعداد ٢٠٠٦) ويتم عمل مقارنة بين اساليب تقدير المساحة الصغيرة فى هذا البحث بين أسلوب التقدير المباشر للمساحات الصغيرة (طريقة المنطقة الواسعة بدون استخدام بيانات مساعدة) و التقدير غير المباشر (على مستوى الوحدة بأستخدام أسلوب الانحدار اللوجيستى المتعدد) ومقارنة النتائج بين اسلوبى التقدير وعمل تجميع لاساليب التقدير للوصول الى المستوى الجغرافى المتوافر له النتائج بالنسبة لمسح القوى العاملة لعكس مدى صحة نماذج التقدير(النموذج اللوجيستى) و التعداد ، وهذا يعكس الوضع الراهن لسوق العمل المصرى عند اصغر مستوى جغرافى و كيفية معالجة مشكلة البطالة والمساعدة فى رسم السياسات من الجهات المعنية بالوصول الى اقل مستوى جغرافى يعكس طبيعة المشكلة والمساهمة فى حلها.

ثانياً: أهمية المشكلة موضوع البحث والتعاريف المستخدمة: تعريف المساحات الصغيرة – القوى البشرية - قوة العمل – المستغلين – المتعطلين (هذه التعاريف صادرة عن الأمم المتحدة وخاصة منظمة العمل الدولية).

القوى البشرية:

تشمل القوى البشرية جميع السكان عدا:

- صغار السن أقل من ٦ سنوات.
 - الأفراد ٦٥ سنة فأكثر ولا يعملون ولا يرغبون في العمل ولا يبحثون عنه (مسنين).
 - المصاب بعجز كلي (غير قادر على العمل بصفة دائمة).
- وتمثل هذه الفئات مجتمعة ما يعرف بالسكان خارج القوى البشرية.

تنقسم القوى البشرية إلى:

أ- الأفراد داخل قوة العمل :

تشمل جميع الأفراد (١٥ سنة فأكثر) ويساهمون فعلاً بمجهوداتهم الجسمانية أو العقلية في أي نشاط إقتصادي يتصل بإنتاج السلع والخدمات (المشتغلون) وكذلك الذين يقدررون على أداء مثل هذا النشاط الإقتصادي ويرغبون فيه ويبحثون عنه ولكنهم لا يجدونه (المتعطلون).

(١)المشتغلون:

هم الأفراد (١٥ سنة فأكثر) الذين يزاولون أعمالاً بأي من الأنشطة الإقتصادية لبعض الوقت (على الأقل ساعة) خلال فترة الإسناد القصيرة للبحث (أسبوع) سواء داخل المنشآت أو خارجها.

و يعتبر في حكم المشتغلين الأفراد المرتبطون بعمل ولم يتمكنوا من ممارسته طوال فترة البحث بسبب المرض أو الإصابة أو الإجازة الاعتيادية أو منحة تدريبية أو دراسية أو لنزاع عمالي أو لعدم انتظام العمل في المنشأة لأسباب مؤقتة أو لأسباب ترجع لطبيعة العمل في بعض الأنشطة وإستثناءاً من ذلك لا يعتبر الفرد المرتبط بعمل والحاصل على إجازة بدون مرتب لمدة عام فأكثر مشغلاً طالما لم يزاول عملاً آخر خلال فترة المسح.

(٢) المتعطلون:

هم الأفراد في سن (١٥-٦٤ سنة) الذين يقدررون على العمل ويرغبون فيه ويبحثون عنه ولكنهم لا يجدونه ويعتبر الفرد متعطلاً خلال أسبوع البحث ولو كان قد تعاقد على عمل وحدد لتسلمه تاريخ لاحق لأسبوع البحث، ويدخل ضمن المتعطلين الموقوف عن العمل إيقافاً مؤقتاً أو لأجل غير مسمى ما دام يبحث عن عمل ولا يجده ولا يتقاضى أجراً أثناء أسبوع البحث وينقسم المتعطلون إلى:

- متعطّل قديم (سبق له العمل) لكنه لم يكن يعمل طوال فترة البحث طبقاً للتعريف السابق.
- متعطّل حديث (لم يسبق له العمل): مثل الخريج (دبلوم أو جامعة) الذي لم يلتحق بأي عمل بعد.

أهمية تقدير المساحات الصغيرة:

تقديرات مساحة صغيرة هامة لمساعدة الهيئات الحكومية المحلية في برامج التمويل / التقييم أو التخطيط الإقليمي الذي يتطلب تقديرات دقيقة محددة لمساعدتهم في اتخاذ قرارات مستنيرة بشأن كيفية تخصيص الموارد أو طلب للحصول على موارد إضافية. وغالباً ما تستخدم تقديرات مساحة صغيرة من قبل مسؤولي البرنامج لتحديد أو قياس المخصصات و كيفية تمويلها و دون معلومات عن منطقة صغيرة فإنهم يجدون صعوبة في تقييم الحاجة الفعلية للسلع والخدمات في كل منطقة وتقديرات المساحة

الصغيرة تقدم معلومات مفصلة عن المناطق الصغيرة مما يجعل الطلب الحكومة المحلية للبيانات منطقة صغيرة قد يزداد

تقديرات المساحة الصغيرة تنتج عندما يكون هناك طلب قوي ومبرر للمستخدم وكذلك أية بيانات بديلة على مستوى منطقة صغيرة من شأنها أن تخدم الغرض المطلوب. ويجب أن يكون هناك بيانات مسح كافية وبيانات إضافية للتأكد من أن المخرجات ستكون ذات نوعية جيدة لتناسب الغرض من التقدير. ولا بد دراسة طبيعة وخصائص المشكلة لضمان الحاجة إلى تقديرات موثوقة لمناطق صغيرة والنهج التقليدي للتقدير السكان داخل المجال الإحصائي من الدقة في عمليات المسح حيث أن العينة العادية لا يمكن أن توفر بيانات كافية عن مناطق صغيرة أو مجموعة من السكان لمتغير محل الاهتمام لاستدلال إحصائي جيد. ومن ناحية أخرى، لا يمكن إجراء عملية التعداد بسبب تكاليفه الباهظة في كثير من الأحيان للحصول على بيانات لرصد نمط التغيرات في الظواهر الديناميكية مثل التوظيف والبطالة. ومع ذلك، هناك عدة طرق أخرى يمكن استخدامها، على سبيل المثال، والجمع بين بعض بيانات المسح مع أحدث إحصاءة لنتائج التعداد، الإطارات أو السجلات الإدارية. وقد تم تطوير عدد من التقنيات لتقدير إحصاءات مساحة صغيرة من أجل الاستجابة لتزايد استخدام تقديرات مساحة صغيرة للتخطيط، كما أن التقنيات المتعلقة بتقدير المساحة الصغيرة مثل التعرف على أساليب التقدير المجال المباشر (أي تقدير مجال المسح)، وأساليب تقدير نطاق/ المجال غير المباشر (تقدير مركب)، أو نموذج الأساليب القائمة (نماذج مستوى المنطقة أو على مستوى وحدة) Rao (2003a) and Longford (2005a)

أهمية إحصاءات القوى العاملة:

يمثل العنصر البشري قوة العمل التي تمارس كافة الأنشطة الخدمية والإنتاجية والإستهلاكية في المجتمع، من هنا تأتي أهمية الإحصاءات والدراسات المتعلقة بنمو وتركيب وتوزيع القوى البشرية وقوة العمل بمختلف أنواعها وخصائصها. أهمية إحصاءات سوق العمل في قياس الفقر وعدم المساواة تستخدم مؤشرات نسبية مختلفة بين الناس كمقياس بديل لمستوى المعيشة. وذكر أن يرتبط الفقر بشكل وثيق مع البطالة. عدم وجود وظيفة يتوج في نقص الدخل من الأجور والحد من الموارد الأساسية. ويمكن أيضا أن تستخدم مؤشرات البطالة إلى تحديد مستوى الإكتفاء الاجتماعي في تنمية السكان. في استجابة للاحتياجات المتزايدة لرصد سوق العمل، والأولوية مسح القوى العاملة منذ عام ٢٠٠٦ كجزء من المسوح الأسرية لها، من أجل الحصول على تقديرات أكثر دقة لمستويات العمالة والبطالة. ومسح القوى العاملة (LFS) يقدم تقديرات لأعداد ونسب العاطلين عن العمل على المستوى الوطني والإقليمي. ولا يمكن للبيانات المسح أن توفر تقديرات موثوقة أو دقيقة للبطالة على مستوى (شياخة / قرية) لأنه لم يتم سحب مفردات العينة لتكون ممثلة على هذا المستوى وحجم العينة ليست كبيرة بما يكفي لإعطاء تقديرات دقيقة لمحاولة لتقدير البطالة على المستويات الوطنية وذلك باستخدام بيانات تعداد عام ٢٠٠٦ كما قدم تقديرات غير دقيقة بدرجة كافية من البطالة، بسبب أوجه القصور في تصميم لأسئلة التعداد في حالة القوى العاملة.

حدود دقة بحث القوى العاملة:

وقد أدرجت قياس البطالة كمسح اسرى منذ عام ٢٠٠٦. وقدمت اتجاهات ومستويات البطالة التي تقاس من أفراد الأسرة. ومن المهم أن نلاحظ أن بعض الخصائص الإضافية ضرورية عند مقارنة النتائج من الأسر المتعاقبة (الأسلوب التتبعي) في مسح القوى العاملة و لقد استمر استعراض الاستبيانات المستخدمة في هذا المسح، مع الحفاظ على المفاهيم والمعايير المعتمدة دوليا. وفي مقارنة تماما من البيانات المتاحة من تعداد السكان (تعداد ٢٠٠٦) وبحث القوى العاملة ٢٠٠٦ الذي يجري مرة سنويا، وقد بذلت تقديرات المسح لكي يمكن الاعتماد عليها على الأقل في مجال تقدير مستوى الجمهورية ثم

مستوى الأقليم ثم بعد إعادة النظر في تصميم العينة على أساس إطار التعداد عام ٢٠٠٦ ستمت التقديرات على مستوى المحافظة ؛ حيث أن البيانات لا يمكن أن تسفر عن تقديرات موثوقة من البطالة لأغراض التخطيط في وحدة الهدف الإدارية لتنفيذ السياسة المستهدفة مثل الشياخة/ القرية ، ولكي يتم الاعتماد على مصادر أخرى للبيانات عن إحصاءات سوق العمل كمتغيرات مساعدة مثل خصائص التعداد و البيانات الادارية . ولذلك هناك طلب متزايد على تمديد قياس البطالة من قبل بحث القوى العاملة على الأقل إلى أدنى مستوى صياغة السياسات المطلوبة . كما ان بحث القوى العاملة يهدف إلى :

- ١- قياس حجم قوة العمل المصرية المدنية داخل الدولة وخصائصها المختلفة.
- ٢- قياس مستوى التشغيل والتوظيف بالمجتمع وبالمناطق الجغرافية المختلفة بالجمهورية.
- ٣- رصد التوزيع الجغرافي للمشتغلين وفقاً لعدد من الخصائص أهمها :
(النوع – السن – الحالة التعليمية – الحالة العملية – المهنة – النشاط الإقتصادي – القطاع – الاستقرار في العمل – ساعات العمل) .
- ٤- قياس مستوى البطالة بالمناطق الجغرافية المختلفة بالجمهورية.
- ٥- رصد التوزيع الجغرافي للمتغلبين وفقاً للخصائص التالية :
(النوع – السن – الحالة التعليمية – مدة التعطل – نوع التعطل (سبق له العمل / لم يسبق له العمل) – المهنة والنشاط الإقتصادي بالنسبة للمتغلب الذي سبق له العمل) .

✚ فقر تقديرات التعداد ٢٠٠٦ لمعدلات البطالة:

على الرغم من أن تعداد عام ٢٠٠٦ أعطى تحسين الصورة العامة للخصائص الديموغرافية لسكان مصر، وكانت التقديرات النهائية تخضع للتعديل وفقاً لنطاقاً أقل من الحقيقي والمستمدة من تغييرات كبيرة جداً في الطريقة التي صيغت بها أسئلة فرص العمل والتي ينتج عنها تغييرات طفيفة في طريقة حساب معدل البطالة خلال هذه الفترة. وقد تم تصميم استبيان التعداد عام ٢٠٠٦ مع الأخذ بعين الاعتبار الصياغة الخاصة بالعمل؛ ومع ذلك كانت الأسئلة محدودة من حيث التحقيق من المستطلعين لتصنيف تلك الأنشطة اقتصادياً أم لا، وهناك عدد كبير من المستبعدين من المشاركة في سوق العمل لأنهم تخلوا عن البحث عن عمل. وكانت تقديرات البطالة من تعداد عام ٢٠٠٦ مختلفة إلى حد ما عن تقديرات بحث القوى العاملة ٢٠٠٦ . واقترح بعض المحاولات لتعديل الفرق حيث ان الاختلاف يرجع إلى نتائج مسح القوى العاملة كونها الاستبيان في القوى العاملة لديها أكثر تفصيلاً عن حالة العمل من بيانات التعداد . فإن اختلاف التقديرات من البطالة بين التعداد و LFS (بحث القوى العاملة) وهنا قد تكون الحاجة كبيرة لاستخدام المعلومات الإضافية من مصادر أخرى للتحقيق في ما إذا كان وضع تقديرات معدلات البطالة أكثر قبولاً أم لا.

• ثالثاً: الهدف من البحث :

تهدف الدراسة إلى التطلع إلى التحسينات الممكنة في قياس البطالة باستخدام معلومات التعداد وتدعيمها ببحث القوى العاملة للمنطقة الجغرافية المستهدفة الممثلة في المنطقة الجغرافية الصغيرة والهدف من هذه الدراسة هو تقدير معدل البطالة في منطقة جغرافية صغيرة مع عينة صغيرة الحجم في مسح القوى العاملة لهذه المنطقة مع استخدام المتغيرات المساعدة من تعداد عام ٢٠٠٦ مثل (وضع الجنسين) للحصول على تقدير موثوق بها مع صغر حجم العينة وتجميع هذه التقديرات لمستوى جغرافي صغير (الشياخة / القرية) ومقارنتها مع تقديرات التعداد السكاني ومسح القوى العاملة على المستوى الوطني حيث تعطي هذه الدراسة لمحة عامة عن الطرق المختلفة لتقدير

مساحة صغيرة المقترحة في هذا المجال، ويحقق استخدام هذه الأساليب لتوفير تقديرات أفضل على مستوى منطقة صغيرة (الشاخه / القرية).

وتهدف الدراسة أيضاً الى جمع المعلومات حول البطالة المقدرة من قبل LFS وغيرها من المعلومات من بيانات التعداد والبيانات الادارية، من أجل الحصول على تحسين للتقدير المكاني للبطالة في المنطقة المستهدفة (الشاخه / القرية). ولذلك فإن هذه الدراسة استكشاف القضايا المنهجية الناشئة من نماذج متكاملة مجتمعة لتقدير البطالة، وذلك باستخدام بيانات مسح القوى العاملة جنباً إلى جنب مع المتغيرات الاجتماعية والاقتصادية الأخرى في التعداد في نموذج الانحدار؛ حيث ان كل من المصادر الموثوق بها هي أقل نسبياً من التقدير الحقيقي المتوقع للمناطق المحلية بسبب قيود عينة التصميم في LFS على هذا المستوى وضعف القياس المباشر من البطالة من خلال تعداد عام ٢٠٠٦. وسيتم التعرض لطرق متنوعة من تقدير المساحات الصغيرة (SAE) على الرغم من أن الدراسة يمكن استخدام البيانات التي تم جمعها على مدى فترات مختلفة، والهدف من ذلك هو الحصول على تقدير دقيق وحقيقي من البطالة للمساحات الصغيرة في عام ٢٠٠٦، وذلك باستخدام LFS 2006 و التعداد العام للسكان لعام ٢٠٠٦ وذلك التعرف على الحدود الجغرافية المعلنة وبعض من القضايا التي يتم التحقيق فيها هي:

- التقييم ما إذا كان من الممكن لتقدير معدل البطالة ضمن مستويات مختلفة من التسلسل الهرمي الجغرافي.
- هل أساليب SAE قادرة على تسفر عن تقديرات مقبولة.
- إذا كان يتم الحفاظ على البيانات الشخصية لسوق العمل (على سبيل المثال البطالة بين الشباب مقابل نسبة البطالة بين الفئات العمرية الأخرى و أيضاً النوع سواء ذكر او انثى) على مستوى المجالات التقليدية للتقدير (الاقليم / المحافظة).

• رابعاً: منهجية البحث:

- كيفية تكوين نموذج لتقديرات المساحة الصغيرة لمعدل البطالة:
- بناء نموذج المساحة الصغيرة باستخدام بيانات المسح (مسح القوى العاملة ٢٠٠٦)
- حساب SAE من خلال تطبيق النموذج لبيانات التعداد (٢٠٠٦).
- حساب متوسط مربعات الخطأ للتقديرات .
- تطبيق نسب معدل البطالة المتوقع إلى مناطق تعداد السكان ٢٠٠٦
- مطابقة لضمان توافق نسب تقديرات SAE مع تقديرات المسح للمستوى المحلي.

• طرق تقدير المساحات الصغيرة:

اختيار طريقة لتقدير مساحة صغيرة يعتمد إلى حد كبير على الهدف المطلوب واحتياجات العملاء والبيانات المتاحة وتتوقف جودة البيانات والتقديرات على مختلف المستويات على توافر أحجام عينة كبيرة وتقنيات تقدير مساحة صغيرة تأتي عادة إلى الصدارة عندما تتوافر أحجام العينات صغيرة جداً بالنسبة للمنطقة الصغيرة مما أدى إلى تقديرات مباشرة مع ارتفاع الخطأ المعياري بشكل غير مقبول. تقوم النماذج في تقدير مساحة صغيرة على استخدام العلاقة بين السكان محل الاهتمام والمتغيرات التفسيرية للحصول على أفضل تقدير غير متحيز وفقاً لمدى توافر المتغيرات التفسيرية التي تساعد على تفسير المتغير التابع والمستوى المطلوب من تقدير مساحة صغيرة (الأسرة / المنطقة) وطبيعة المتغير التابع (مستمر أو ثنائي)، يتم استخدام الجودة المطلوبة SAE والهدف هو الحصول على إحصاءات وصفية لخصائص المناطق الصغيرة.

أولاً : طرق المساحات الصغيرة البسيطة (التقدير المباشر):

١-١ المقدر المباشر.

٢-١ مقدر المساحة الواسعة بدون بيانات مساعدة.

٣-١ مقدر المساحة الواسعة بيانات مساعدة.

ثانياً: طرق الانحدار (التقدير غير المباشر):

١-٢ التقدير على مستوى المنطقة.

٢-٢ التقدير على مستوى الوحدة .

أولاً: طرق المساحات الصغيرة البسيطة:

١-١ المقدر المباشر :

التقديرات المباشرة هي المقدرات القائم على التصميم التقليدي التي يتم الحصول عليها من خلال تطبيق أوزان المسح لوحدة العينة في كل منطقة صغيرة (Saei and Chambers, 2003). عند تصميم المسوح لتقديم تقديرات موثوقة فقط على النطاق المحلي.

٢-١ مقدر المساحة الواسعة بدون بيانات مساعدة:

هي واحدة من أبسط أنواع طرق التقدير البسيطة وهي تحسب من خلال تطبيق معدل ما(مثل معدل البطالة) للمنطقة الواسعة الذي تم الحصول عليه من بيانات المسح الى سكان المناطق الصغيرة (بيانات متاحة من تعداد السكان او التقديرات السكانية) ونجاح هذا المقدر يتوقف على اختيار المنطقة الواسعة في كونها كبيرة بالقدر الكافي للتقدير. ولحساب الوسط الحسابي لمنطقة ما كما يلي :

$$\hat{Y}_i = \hat{Y}_r = \frac{\hat{Y}_r}{\hat{N}_r}$$

حيث $\hat{Y}_r$ حيث هو المقدر المباشر للمنطقة r وهي تحسب من اجمالي حجم العينة للمجتمع المستهدف $\hat{Y}_r = \sum_{j \in r} w_j y_j$ واجمالي حجم السكان $\hat{N}_r = \sum_{j \in r} w_j$ نحصل عليه من البحث لكل فرد في العينة و y_j حيث مؤشر لكون الشخص في المجتمع المستهدف يحمل الصفة محل الدراسة.

٢-١ مقدر المساحة الواسعة بيانات مساعدة:

$$\hat{Y}_i^{(v)} = \sum_g \frac{N_{ig} \hat{Y}_{rg}^{(v)}}{N_{rg}}$$

وهذا التقدير يعتمد على تقدير وفقا لمجموعة فرعية في المجتمع وفقا لخصائص المنطقة الواسعة حيث ان :

$\hat{Y}_{rg}^{(v)}$ التقدير المباشر للطبقة الفرعية " v "

N_{ig} يمثل عدد السكان في الفئة g في المنطقة الصغيرة i

N_{rg} يمثل عدد السكان في الفئة g في المنطقة الواسعة r

ثانياً: طرق الانحدار:

هى تستخدم عندما يكون الغرض الحصول على تقديرات اكثر دقة وفى ظل توافر بيانات مساعدة جيدة وهى تستخدم عندما يكون هناك سهولة اكثر فى توضيح العلاقة بين المتغيرات واخذ ذلك فى الحسبان والاختيار بين اساليب الانحدار يعتمد على دقة التقديرات المتنبؤ بها و سهولة التنفيذ

٢-١ التقدير على مستوى المنطقة.

هى تستخدم عندما تكون المتغيرات المساعدة متاحة على مستوى المنطقة ولا يظهر تأثيرها على مستوى الفرد.

$$\hat{y}_i = x_i^t B + e_i$$

حيث ان :

$\hat{y}_i$ هى تمثيل التقدير المباشر للمجتمع محل الاهتمام.
 $x_i^t B$ هو يمثل متجة المتغيرات المساعدة .

e_i يمثل مقدار الخطأ (لتجاهل بعض المتغيرات اتى تؤثر على المتغير التابع)

٢-٢ التقدير على مستوى الوحدة.

هى تستخدم العلاقة بين حالة الشخص الفردية والمتغيرات على مستوى الشخص حيث ان المتغيرات المساعدة على مستوى الوحدة تكون متاحة لكل فرد فى المنطقة الصغيرة

$$\hat{y}_{ij} = x_{ij}^t B + e_{ij}$$

حيث ان :

e_{ij} يمثل مقدار الخطأ (لتجاهل بعض المتغيرات اتى تؤثر على المتغير التابع)

حيث ان المساحة الصغيرة يتم النظر اليها فى كونها مجموعة سكانية خاصة مثل عدد الافراد العاملين (المشتغلين) حيث ان المتغير التابع فى ذلك الوقت يكون متغير ثنائى حيث ان الفرد محل الاهتمام اما يكون يعمل او لا يعمل . ويكون نموذج الانحدار فى ذلك الوقت هو الانحدار اللوجيستى الثنائى حيث ان صور المتغير التابع اما = ١ يحمل الصفة محل الدراسة او ٠ لا يحمل الصفة محل الدراسة ويكون كما يلى :

$$\text{Ln} \left(\frac{p_{ij}^w}{1-p_{ij}^w} \right) = x_{ij}^t B + e_{ij}$$

حيث ان Ln تمثل اللوغارتم الطبيعى لاحتمال الحدث للفرد j الموجود فى المنطقة الصغيرة i ، x_{ij}^t هو متجة المتغيرات المستقلة و B هو معلمات او معاملات الانحدار والتى تقدر باستخدام دالة الامكان الاعظم.* وسوف يتم حساب طرق التقدير المباشر بالاعتماد على اوزان المسح و مقدار المساحة الواسعة بدون بيانات مساعدة اما بالنسبة لطرق التقدير غير المباشرة (التقدير على مستوى الوحدة) وسوف يتم التطبيق فى حساب معدل البطالة لشيخات وقرى محافظة القليوبية ويتم ذلك بواسطة الباحث.

(كل محافظة هى عبارة عن حضر وريف والحضر ينقسم لمجموعة من الاقسام و الريف ينقسم لمجموعة من المراكز و القسم ينقسم لمجموعة من الشياخات و على نفس المستوى الجغرافى المركز ينقسم لمجموعة من القرى وهذا المستوى الاقل جغرافيا الذى سوف يتم تقدير معدل البطالة له بطريقة المساحات الواسعة فى

ظل عدم وجود متغيرات مساعدة والمقدر المباشر كمقدرات مباشرة و التقدير على مستوى الوحدة بأستخدام الانحدار اللوجيستي كمقدر غير مباشر .

❖ التطبيق العملي للطرق المقدرات المختلفة و المقارنة بينهم:

بالتطبيق على اوزان بحث القوى العاملة تم الوصول الى هذه المعدلات للبطالة من خلال التقدير المباشر (طريقة المقدر المباشر ومقدر المساحة الواسعة) حيث عبارة عن حاصل قسمة المتعطلين على عدد افراد قوة العمل و استنتاج هذا المعدل على مستوى الشياخة/القرية ل ٤٠ شياخة و قرية داخل محافظة القليوبية بهذه الطرق:

المحافظة	حضر/ ريف	كود القسم / المركز	كود الشياخة / القرية	معدل البطالة للمقدر المباشر	معدل البطالة لمقدر للمنطقة الواسعة بدون بيانات مساعدة	معدل البطالة لمقدر الوحدة بأستخدام الانحدار اللوجيستي
القليوبية	1	1	5	2.34	8.3	5.57
	1	3	1	14.38	8.3	3.91
	2	3	7	6.39	8.3	1.18
	2	3	9	12.69	8.3	2.97

التقدير المباشر على مستوى المحافظة ٨.٨ % من بحث القوى العاملة حيث يتم مقارنة جميع معدلات البطالة لشياخات وقرى محافظة القليوبية للطرق السابقة للتقدير للتأكد من دقة التقديرات بمقارنتها بمعدل البطالة على مستوى المحافظة من بحث القوى العاملة .

• كيفية تطبيق التقدير على مستوى الوحدة :

يتم التطبيق في هذا النموذج بأستخدام نموذج الانحدار اللوجيستي الثنائي في حساب معدل البطالة من خلال التعرف على احتمال المتعطلين و المشتغلين حيث ان الحالة العملية هي المتغير التابع و تم اخذ كل من العمر و النوع و الحلة التعليمية و محل الإقامة (حضر/ ريف) كمتغيرات مستقلة .

$$Y = - 1.733 - 2.692 \text{ sex} + 3.066 \text{ age} + .510 \text{ edu_level} - .514 \text{ hadr}$$

$$Y = - 1.733 - 2.692 X_1 + 3.066 X_2 + .510 X_3 - .514 X_4.$$

ويقوم نموذج الانحدار اللوجستي على فرض أساسي هو أن المتغير التابع y الذي يتم دراسته هو

متغير ثنائي يأخذ القيمة (1) باحتمال (p) والقيمة 0 باحتمال (1-p) ، أي إلى حدوث الاستجابة

وعدم حدوثها ، وكما نعلم ان الانحدار الخطي يأخذ الشكل التالي:

$$Y = b_0 + b_1 X + e \quad \text{حيث ان}$$

ومن المعروف في الانحدار أن الطرف الأيمن لهذه النماذج يأخذ قيم من $-\infty$ (الى) $+\infty$ ، ولكن عندما

يكون المتغير التابع ثنائي ، فإن الانحدار الخطي البسيط لا يكون ملائم لان $E(y/x) = P$

$(y=1)=p$ ، وبذلك تكون قيمة الطرف الأيمن محصورة ما بين الرقمين (0,1) وبذلك يكون

النموذج غير قابل للتطبيق من وجهة نظر الانحدار، وإن إحدى طرق حل هذه المشكلة هو إدخال تحويل رياضية مناسبة على المتغير التابع Y .

$$\ln(p/1-p) = b_0 + b_1X.$$

ويسمى هذا النموذج بنموذج الانحدار اللوجستي، وتسمى التحويلة $\ln$ بتحويلة لوجت (Logit transformation)

وإن الدالة اللوجستية دالة مستمرة تأخذ القيم 0، 1، وتقترب y من الصفر كما تقترب الطرف الأيمن لدالة اللوجستية من ∞ وتقترب y من الواحد كما تقترب الطرف الأيمن لهذه الدالة من ∞ ، وهي دالة متماثلة عندما يكون الطرف الأيمن لهذه الدالة مساوي للصفر حيث أن

$$P = e^{a+bx} / 1 + e^{a+bx}$$

وبالتطبيق هذا النموذج نستنتج ما يلي :

حيث تم التطبيق حيث معدل او نسبة احتمال ان يكون الشخص يعمل الى كون الشخص لايعمل في المنطقة الصغيرة بشرط البيانات المساعدة المتوفرة.

a. Variable(s) entered on step 1: sex (النوع)

b. Variable(s) entered on step 2: Age (العمر).

c. Variable(s) entered on step 3: education (الحالة التعليمية) .

d. Variable(s) entered on step 4: hadr. (محل الإقامة).

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	315975.049 ^a	.176	.244
2	238877.206 ^a	.367	.509
3	237357.500 ^a	.370	.514
4	235096.259 ^a	.375	.521

Classification Table

Observed		Predicted		
		التابع المتغير		Percentage Correct
		0	1	
Step 4	0 التابع المتغير	177522	16888	91.3
	1	27483	70292	71.9
Overall Percentage				84.8

تم حساب النتائج السابقة بمعرفة الباحث.

خامساً: النتائج :

- ١- بالنسبة للنموذج اللوجيستي يعرض مدى معنوية المتغيرات المستقلة المدرجة فى النموذج من متغير التعليم و العمر والنوع ومحل الإقامة وعدم ظهور تأثير معنوى لتصنيف الحالة العملية (يعمل / لايعمل) بعد أضافة كلا من متغير الحالة التعليمية و محل الإقامة لهذا النموذج حيث لم ترتفع نسبة تصنيف النموذج ككل عن ٨٤.٨ % بعد اضافة هذان المتغيران على الرغم من التأثير المعنوى لهم.
- ٢- يلاحظ بنسبة مقبولة اتفاق النموذج اللوجيستي المقدر مع المنطق الاقتصادي فى منطقية المعلومات حيث ان زيادة متغير التعليم بمقدار واحد يؤدي الى زيادة لوغاريتم معامل الترجيح بمقدار ٥١٠. الذى يؤدي الى زيادة المتغير التابع بمقدار ١ مع ثبات تأثير باقى المتغيرات اما بالنسبة لمتغير النوع فأنها تدل على زيادة هذا المتغير بمقدار ١ سيؤدى الى خفض لوغاريتم معامل الترجيح بمقدار ٢.٦٩٢ للمتغير التابع يكون ١ فى ظل ثبات المتغيرات الاخرى وكذلك ينطبق التفسير على باقى المتغيرات المستقل التى تم ادراجها فى النموذج .
- ٣- اما بالنسبة للتصنيف الصحيح للنموذج من حيث يعمل ولايعمل بنسبة ٨٤.٨ % لمشاهدات العينة كما ان اختبار "WALD" لمعنوية المعلومات المقدره حيث ثبتت معنوية المعلومات المقدره جميعا.
- ٤- Nagelkerke R Square هو يعرض ان النموذج المقدر يمثل البيانات تمثيلا مقبولا الى حد ما بنسبة ٥٢.١% وان المتغيرات المستقلة التى تم ادراجها فى النموذج تساهم فى تفسير المتغير التابع بهذه النسبة على الرغم من ارتفاع نسبة تفسير المتغيرات المستقلة للمتغير التابع بأضافة كل من متغير الحالة التعليمية ومحل الإقامة وذلك فى ظل عدم لنموذج التصنيف.
- ٥- اما بالنسبة لتقديرات المساحة الصغيرة فى طريقة المقدر المباشر الذى يعتمد على اوزان المسح و طريقة مقدر المساحة الواسعة بدون بيانات مساعدة الذى يعتمد على تقدير المساحة الواسعة (المحافظة) وتطبيق نفس المعدل على المساحات الجغرافية الصغيرة داخل هذه المنطقة الواسعة فكانت قيمة معدل البطالة للمقدر المباشر بعد التجميع للمقارنة بمقدر البحث على مستوى المحافظة ٨.٥% اما بالنسبة لمقدر المساحة الواسعة ٨.٣% وهما يقتربان الى حد كبير من معدل البطالة على مستوى المحافظة من بيانات بحث القوى العاملة وهى ٨.٨% مما يدل على دقة هذه المقدرات اما بالنسبة للمقدر الغير مباشر باستخدام نموذج الانحدار اللوجيستي فان معدل البطالة قيمته ١٢.٢ % وهى قيمة مقبولة الى حد ما مقارنة بقيمة المعدل من نتائج المسح ولكن يظهر من ما سبق ان طرق التقدير المباشرة كانت درجة دقتها اعلى من طرق التقدير غير المباشرة بأستخدام الانحدار اللوجيستي مقارنة بمقدر المسح وربما هذا يرجع الى درجة التصنيف فى النموذج اللوجيستي بين يعمل / لايعمل او اهمال بعض المتغيرات المستقلة

الآخري التي يجب ادراجها فى النموذج والتي تؤثر تأثيراً معنوياً على المتغير التابع و بالتالى تزيد من دقة المقدرات الغير مباشرة

٦- بالنسبة للتطبيق العملى نستنتج وفقاً لهذه الدراسة فى المقارنة بين طرق تقدير المساحات الصغيرة ان المقدرات المباشرة كانت أكثر دقة و كفاءة من المقدرات الغير مباشرة مع امكانية التعديل فى المقدرات الغير مباشرة بمزيد من التجربة لزيادة دقتها وكفاءتها .

٧- الأهمية الفعالة لتقدير المساحات الصغيرة وما يصحبها من توفير فى الوقت والجهد والتكلفة والحصول على تقديرات دقيقة تساهم فى حل المشكلات مثل البطالة والفقر ... ومساعدة المسؤولين فى رسم السياسات واتخاذ القرارات فى ظل الأمكانيات المحدودة للدولة من خلال البيانات التفصيلية لأقل مستوى جغرافى.

سادساً: الخاتمة:

قامت الدراسة على محاولة عرض ومقارنة طرق تقدير المساحات الصغيرة المختلفة والتطبيق على أصغر مستوى جغرافى وهى الشياخة / القرية كمستوى أكثر تفصيلاً للمشكلة ، حيث يتم استخدام طرق التقدير المباشرة وذلك للوصول الى افضل معالجة لبيانات المشكلة موضوع الدراسة من المسح والتعداد السكانى. أما بالنسبة لطرق التقدير غير المباشرة فهى تعتمد على البيانات المتاحة وقوة العلاقة بين المتغير التابع والمتغيرات المستقلة المدرجة بالدراسة لكفاءة ودقة هذه المقدرات وعلى الرغم من المناهج النظرية المختلفة ؛ بالإضافة إلى التطبيق العملى لتقنيات SAE وجد انها تعتمد على البيانات المتاحة وكذلك على قوة العلاقة بين المتغيرات المستقلة بالإضافة الى مصادر المعلومات المختلفة ؛ حيث ان تقدير معدل البطالة فى الشياخات / القرى يعتمد على البيانات المستخدمة من بحث LFS وتعداد السكان ،وقد حاولت الدراسة ان تظهر نتائج المقارنة بين طرق التقدير المباشرة (طريقة المقدر المباشر و مقدر المساحة الواسعة بدون بيانات مساعدة) وقد أظهرت دقتها الى حد كبير أما بالنسبة للمقدر غير المباشر (طريقة التقدير على مستوى الوحدة بأستخدام نموذج الانحدار اللوجيستى) فقد اظهر النموذج درجة من الدقة المقبولة فى ظل البيانات المتوافرة والعلاقة بين المتغير التابع والمتغيرات المستقلة المدرجة فى النموذج والتى تعتمد زيادة كفاءتها على جودة المتغيرات المستقلة المدرجة فى النموذج ومدى تأثيرها على المتغير التابع وعرض ذلك من خلال حساب معدل البطالة بهذه الطرق لعرض مدى مساهمة ذلك فى معالجة مشكلة البطالة بالوصول إلى عرض المشكلة على أقل مستوى جغرافى وعرضها أكثر تفصيلاً والمساهمة فى حلها من خلال تقدير معدل البطالة على اصغر مستوى جغرافى حيث مدى أهمية طرق تقدير المساحات الصغيرة فى تقدير معالم المجتمع والتى لها تأثير كبير فى توفير الوقت والجهد والتكلفة والتى تساهم فى رسم السياسات القومية ودعم اتخاذ القرار .

المراجع :

مصادر عربية :

١- نشرة بحث القوى العاملة لعام ٢٠٠٦ الصادرة عن الجهاز المركزى للتعبئة العامة والإحصاء
من الموقع الإلكتروني لجهاز \ <http://www.capmas.org.eg>

مصادر اجنبية

- 1- Australian Bureau of Statistics (2005), Information Paper: Regional Research in Australia - the Statistical Dimension: an Information Development Plan for Rural and Regional Statistics, Catalogue No. 1362.0 (Due for release in early 2006).
- 2- Australian Bureau of Statistics (2005), A Guide to Small Area Estimation, available online
- 3- Rao, J. N. K. (2003). Small Area Estimation. Wiley Series in Survey Methodology.
- 4- Saei, A. and Chambers, R. (2003), Small Area Estimates: Methods Based on the Application of Mixed Models, Southampton Statistical Sciences Research Institute, Methodology Working Paper,

A General Delete-Group Jackknife Variance Estimator for Stratified Sampling Survey

El Sayed, S. M.¹

Abdel Latif, S. H.²

Abstract

In this article, the accuracy of point estimates in survey sampling is evaluated using variance estimates. The jackknife has been suggested as a useful resampling technique. The approximation of sampling relative error for the total in stratified sampling without replacement from finite populations will be reviewed, and a general delete-group jackknife method of Cao, R. et al (2013) is derived.

Keywords

Jackknife, Relative Error, Sampling Survey, Stratified Sampling, Variance Estimation.

1. Introduction

This section basically divided into two main parts of the review. The first part reviewed unequal probability sampling and the second part reviewed jackknife method as a resampling technique and its estimation of variance in survey sampling.

Firstly, Hansen and Hurwitz (1943) gave the first theoretical aspects of unequal probability sampling with replacement. Madow (1949) introduced the concept of unequal probability sampling without replacement but no theoretical aspects were given. Horvitz and Thompson (HT) (1952) were the first to develop the historical theoretical ideas of unequal probability sampling without replacement when they proposed an estimator of population total. Then they estimated a variance of the actual variance of their estimator.

1- Emeritus Professor of Statistics, Institute of Statistical Studies and Research, Cairo University.

2- Teaching Assistant, Institute of Statistical Studies and Research, Cairo University.

An alternative expression of HT estimator was developed by Yates and Grundy (1953). The previous variance estimators were very hard to compute with increase in the sample size. So, numerous survey statisticians have proposed approximate expression for variance of HT (1952) estimator to overcome these computing problems. The first attempt was made by Hartley and Rao (1962). Also, Rao_(1963), Hanif and Ahmad (2001), Shahbaz and Hanif (2003a, 2003b) , Basit and Shahbaz (2006), Shahbaz and Hanif, (2007), Alodat (2009), and Razvi and Khan (2013) have proposed approximate expressions for variance of HT estimator. Hassan et al. (2009) compared various approximate formulae for variance of HT estimator.

Secondly, in survey sampling, Durbin (1959) considered the first one who applied Quenouille's(1949, 1956) method in this area, then Rao (1965) and Rao and Webster (1966).

According to using jackknife method under unequal probability sampling, and specifically, in stratified sampling design. Lee (1973) attempted to develop an efficient scheme for variance estimation in stratified sampling whereby two primary selections per stratum are drawn, which form the parent sample. Lee's scheme is equivalent to the jackknife method of variance estimation used by Tukey (1958) for unstratified samples. However, the estimation procedure to be employed for his scheme would be different for the jackknife method. Campbell (1980) gave a general definition of a jackknife pseudo value that is applicable for unequal probability sampling and stratification, and generalized the jackknife variance estimate so that it applies to any sample design for which the variance of an estimated mean can be estimated. Shao and Tu (1995) presented the jackknife variance estimator for $\hat{\theta}$ when multistage sampling is used. Therefore, the resampling techniques, in the context of sampling survey, has been widely studied, Tukey's (1958) original idea of jackknife variance estimation has been developed to handle stratified multistage sampling by Jones (1974), Kish and Frankel (1974), Krewski and Rao (1981), Kovar et al. (1988), Rao et al. (1992), and Shao and Tu (1995).

Berger and Skinner (2005) established the consistency of Campbell's (1980) generalized jackknife variance estimator. They also compared the empirical performance of Campbell's jackknife (in a single stage context) with standard single stage jackknives such as in Tukey (1958), Kish and Frankel (1974), and Rao et al. (1992). Berger (2007) proposed a numerically simple

jackknife variance estimator. Berger's estimator is consistent under unistage stratified sampling without replacement. Berger proposed a modified Campbell's (1980) estimator. Escobar and Berger (2011) proposed two generalized jackknife variance estimators suitable for functions of HT (1952) point estimators. Regularity conditions under which the proposed estimators are design-consistent were also provided. The proposed estimators are compared with the Campbell (1980) jackknife variance estimator for a ratio.

Stratified multi-stage cluster sampling is commonly used in socio-economic surveys. Pioneering work on "delete-cluster" jackknife for variance estimation under stratified multi-stage cluster sampling is due to McCarthy (1969) and Kish and Frankel(1974).

Hall (2003) provided a scholarly historical account of Mahalanobis' seminal contributions to early development of survey sampling in India.

Escobar and Berger (2013) proposed a design-consistent jackknife variance estimator that takes account of all stages via deletion of clusters and observations within clusters. They also proposed a jackknife variance estimator for self-weighted two-stage (stratified) without replacement sampling. They made Monte-Carlo simulations which showed that the proposed jackknife could be more accurate than customary jackknife estimators for more than one stage such as the Rao et al. (1992) stratified multi-stage delete-cluster jackknife.

Cao, et al. (2013), handled the problem of approximating the sampling relative error of point estimates derived from large sample surveys on a finite population using stratified random sampling design without replacement. They proposed three jackknife methods and compared it with the plug-in and two bootstrap techniques. The first one (J1) considered by removing a sample value at each iteration, the second one (J2) constructed by removing a stratum at each iteration, and the third estimate (J3) constructed by considering the variance of $\hat{\theta}$ as a linear combination of variances of statistics constructed at stratum level. And these variances are previously estimated by jackknife in each stratum. They examined and compared the different procedures by extensive simulation study.

2. Jackknife Estimation

2.1 The Grouped Quenouille-Tukey Jackknife Method

For the sample of size n independent and identically distributed (*iid*) observations $x_1, x_2, \dots, x_n$. The sample is first divided at random into g groups $G_1, G_2, \dots, G_g$, each of size d , assuming that $n = dg$. The groups, G_i , are deleted in turn and the "delete-group" estimates $\hat{\theta}_i$, $i = 1, 2, \dots, g$, are computed, where $\hat{\theta}_i$ denotes the estimator of θ based on the sample of size $n - d = d(g - 1)$ after deleting G_i . Quenouille showed that the estimator

$$\hat{\theta}_J = \sum_{i=1}^g \tilde{\theta}_i / g,$$

where $\tilde{\theta}_i = g\hat{\theta} - (g - 1)\hat{\theta}_i$.

Tukey (1958) suggested regarding the $\tilde{\theta}_i$ as *iid* for general $\hat{\theta}$ and then using

$$\begin{aligned} v(\hat{\theta}_J) &= \sum_{i=1}^n (\tilde{\theta}_i - \hat{\theta}_J)^2 / n(n - 1). \\ &= \frac{(n-1)}{n} \sum_{i=1}^n (\hat{\theta}_i - \hat{\theta}_{(\cdot)})^2, \end{aligned}$$

where $\hat{\theta}_i$ is the jackknife pseudo estimate, that is the estimator calculated using the whole sample except for the i^{th} observation, $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$, $\hat{\theta}_{(\cdot)} = \sum_{i=1}^n \hat{\theta}_i / n$. As, $v(\hat{\theta}_J)$, the jackknife variance estimator of $\hat{\theta}_J$ or $\hat{\theta}$.

2.2 Cao et al Jackknife Method

The sampling error of $\hat{\theta}$ can be presented in relative terms, using the variation coefficient of the estimator given by:

$$E_{rel}(\hat{\theta}) = \sqrt{\frac{var(\hat{\theta})}{E(\hat{\theta})}}.$$

Cao, et al. (2013) handled the problem of approximating the sampling relative error of point estimates derived from large sample surveys on a finite population using stratified random sampling design without replacement as we mentioned in the previous section.

El Sayed and Abdel Latif (2015) introduced a general jackknife method of Cao, R. et al (2013) for their first method, jackknife leaving one sample value out, by removing a group of values out. And we will generalize the second method of Cao et al. (2013) in the following section.

3. Jackknife leaving group of stratum out

For simplicity and comparability of the results we will follow all symbols and notations and may be some scientific definitions as in Cao et al.(2013) and in generalizing his results we have to follow his approach so both results can be compared.

Let, L and N_i will denote the number of strata and the population size of the i^{th} stratum, $= 1, \dots, L$, respectively. And, let $\{x_{ij}, j = 1, \dots, n_i, i = 1, \dots, L\}$ be a stratified random sample without replacement of the variable of interest X , of size $n = \sum_{i=1}^L n_i$, n_i being the sample size within the i^{th} stratum. The unbiased estimator of the total, $\hat{\tau}$, and the sample variances, S_i^2 , are respectively given by;

$$\hat{\tau} = N\hat{\mu} = \sum_{i=1}^L N_i \bar{x}_{i.} = \sum_{i=1}^L F_i n_i \bar{x}_{i.} = \sum_{i=1}^L F_i x_{i.},$$

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2,$$

and, $F_i = N_i/n_i$ the elevation factors of each stratum. Being $\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$ and $x_{i.} = \sum_{j=1}^{n_i} x_{ij}$.

where, X_{ij} is the j^{th} element of the i^{th} population stratum, with $j = 1, \dots, N_i$, $i = 1, \dots, L$. Being $\bar{X}_{i.} = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}$ and $X_{i.} = \sum_{j=1}^{N_i} X_{ij}$.

It is a general jackknife of leaving one stratum out for Rao (2013) because we propose to calculate each pseudo value removing all the observations of one group of strata. Thus, the R^{th} jackknife pseudo value is based on the original sample without the observations of the eliminated strata R , i. e.,

$$\hat{\tau}_{(R)} = \frac{N}{N - N_R} \sum_{i \notin R}^L F_i x_{i.} \quad (3.1)$$

$$N_R = \sum_{r \in R} N_r = \frac{N}{N - N_R} (\hat{\tau} - \sum_{i \in R} F_i x_{i.}).$$

Now, two variants of the jackknife estimator are introduced by considering different ways of averaging the pseudo values $\hat{t}_{(R)}$. First, we use a weighted mean, where each pseudo value is weighted by the population size of some strata removed in calculation. Thus, we have

$$\begin{aligned}\hat{t}_{(\cdot)}^{1R} &= \sum_{R=1}^H \frac{N_R}{N} \hat{t}_{(R)} \\ &= \sum_{R=1}^H \frac{N_R}{N} \left(\frac{N}{N-N_R} (\hat{t} - \sum_{r \in R} F_r x_{r.}) \right) \\ \hat{t}_{(\cdot)}^{1R} &= \hat{t} \sum_{R=1}^H \frac{N_R}{N-N_R} - \sum_{R=1}^H \frac{N_R}{N-N_R} (\sum_{r \in R} F_r x_{r.})\end{aligned}\quad (3.2)$$

Then, the jackknife estimator of $var(\hat{t})$ takes the form;

$$\begin{aligned}\widehat{var}_{J,1R}(\hat{t}) &= \sum_{R=1}^H \frac{N_R}{N} \left(\frac{N-N_R}{N} \right) (\hat{t}_{(R)} - \hat{t}_{(\cdot)}^{1R})^2, \\ \widehat{var}_{J,1R}(\hat{t}) &= \sum_{R=1}^H \frac{N_R}{N} \left(\frac{N-N_R}{N} \right) \left[\frac{N(\hat{t} - \sum_{r \in R} F_r x_{r.})}{N-N_R} - \sum_{i=1}^H \frac{N_i(\hat{t} + \sum_{i \in R} F_i x_{i.})}{N-N_i} \right]^2.\end{aligned}\quad (3.3)$$

The jackknife estimator of the relative error is then calculated as:

$$\hat{\epsilon}_{J,1R}(\hat{t}) = \frac{1}{\hat{t}} \left(\widehat{var}_{J,1R}(\hat{t}) \right)^{1/2} \quad (3.4)$$

An alternative variant of the jackknife leaving some strata out is obtained if all the strata contribute with the same weight in the estimation, i. e. the pseudo values are directly averaged as follows

$$\begin{aligned}\hat{t}_{(\cdot)}^{2R} &= \frac{1}{H} \sum_{R=1}^H \hat{t}_{(R)} \\ \hat{t}_{(\cdot)}^{2R} &= \frac{1}{H} \sum_{R=1}^H \frac{N}{N-N_R} (\hat{t} - \sum_{r \in R} F_r x_{r.}) \\ &= \frac{N}{H} \left(\hat{t} \sum_{R=1}^H \frac{1}{N-N_R} - \sum_{R=1}^H \frac{\sum_{r \in R} F_r x_{r.}}{N-N_R} \right)\end{aligned}\quad (3.5)$$

Using $\hat{t}_{(\cdot)}^{2R}$, the jackknife estimator of $var(\hat{t})$ becomes

$$\begin{aligned}\widehat{var}_{J,2R}(\hat{t}) &= \frac{H-1}{H} \sum_{R=1}^H (\hat{t}_{(R)} - \hat{t}_{(\cdot)}^{2R})^2, \\ &= \frac{H-1}{H} \sum_{R=1}^H \left[\frac{N(\hat{t} - \sum_{r \in R} F_r x_{r.})}{N-N_R} - \frac{N}{H} \sum_{i=1}^H \frac{(\hat{t} + \sum_{i \in R} F_i x_{i.})}{N-N_i} \right]^2\end{aligned}\quad (3.6)$$

Not admitting a simpler explicit expression either.

And the jackknife estimator of the relative error with this criterion is:

$$\hat{\epsilon}_{J,2R}(\hat{t}) = \frac{1}{\hat{t}} \left(\widehat{var}_{J,2R}(\hat{t}) \right)^{1/2} \quad (3.7)$$

References

- [1]- **Alodat, N. A. (2009):** On unequal probability sampling without replacement sample size 2. *Int. J. Open Problems Comp. Math.*, Vol. 2, No. 1, pp. 108-112.
- [2]- **Basit, A. and Shahbaz, M. Q. (2006):** A class of selection procedure for unequal probability sampling without replacement and a sample of size 2. *J. Statist.*, Vol. 13, No.1, pp. 26-32.
- [3]- **Berger, Y. G. (2007):** A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, Vol. 94, No. 4, pp. 953-964.
- [4]- **Berger, Y. G. and Skinner, C. J. (2005):** A jackknife variance estimator for unequal probability sampling. *J. Roy. Statist. Soc. Ser. B*, Vol. 67, No. 1, pp. 79-89.
- [5]- **Cao, R., Vilar, J. A., Vilar, J. M., and López, A. K. (2013):** Sampling Error Estimation in Stratified Surveys. *Open Journal of Statistics*, Vol. 3, pp. 200-212.
- [6]- **Campbell, C. (1980):** A different view of finite population estimation . *Proc. Surv. Res. Meth. Sect. Am. Statist. Assoc.*, pp. 319-324.
- [7]- **Durbin, J. (1959):** A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, Vol. 46, pp. 477-480.
- [8]- **El Sayed, S. M. and Abdel Latif, S. H. (2015):** General jackknife variance estimator for stratified sampling survey. *International Mathematical Forum*, Vol. 10, No.8, pp. 377-383.
- [9]- **Escobar, E. L. and Berger, Y. G. (2011):** Jackknife variance estimation for functions of Horvitz & Thompson estimators under unequal probability sampling without replacement. *Int. Statistical Inst.: Proc. 58th World Statistical Congress*. Dublin (Session CPS046), pp. 5031-5034.
- [10]- **Escobar, E. L. and Berger, Y. G. (2013):** A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, Vol. 23, pp. 1-19.

- [11]- **Hall, P.** (2003): A short prehistory of the bootstrap. *Statistical Science*, Vol. 18, pp.158-167.
- [12]- **Hanif, M. and Ahmad, M.** (2001): Approximate variance formula for Horvitz-Thompson estimator using first order probabilities. Presented at ICCSS-VII held in Lahore.
- [13]- **Hansen, M. H. and Hurwitz, W. N.** (1943): On the theory of sampling from finite populations. *Ann. Math. Stat.*, Vol. 14, pp. 333-362.
- [14]- **Hartley, H. O. and Rao, J. N. K.** (1962): Sampling with unequal probabilities and without replacement, *Ann. Math. Stat.* Vol. 33, pp. 350-374.
- [15]- **Hassan, Y., Shahbaz, M. Q., and Hanif, M.** (2009): Empirical comparison of some approximate variance formulae of Horvitz-Thompson estimator. *World Applied Sciences Journal*, Vol. 5, pp. 597-599.
- [16]- **Horvitz, D. G., and Thompson, D. J.** (1952): A generalization of sampling without replacement from a finite universe. *J. Amer. Stat. Assoc.*, Vol. 47, pp. 663-685.
- [17]- **Jones, H. L.** (1974): Jackknife estimation of functions of stratum means. *Biometrika*, Vol. 61, pp. 343-348.
- [18]- **Kish, L. and Frankel, M. R.** (1974): Inference from complex samples (with discussion). *J. Roy. Statist. Soc. Ser. B.* Vol. 36, pp. 1-37.
- [19]- **Kovar, J. G., Rao, J. N. K. and Wu, C. F. J.** (1988): Bootstrap and other methods to measure errors in survey estimates. *Can. J. Statist.* Vol. 16, pp. 25-45.
- [20]- **Krewski, D. and Rao, J. N. K.** (1981): Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.* Vol. 9, pp. 1010-1019.
- [21]- **Lee, K.** (1973): Variance estimation in stratified sampling. *J. Am. Statist. Assoc.* Vol. 68, No. 342, pp. 336-342.
- [22]- **Madow, W. G.** (1949): On the theory of systematic sampling II. *Ann. Math. Stat.* Vol. 20, pp. 333-354.
- [23]- **Mahalanobis, P. C.** (1946): Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, Vol. 109, pp. 325-378.
- [24]- **McCarthy, P. J.** (1969): Pseudo-replication. *Review of the International Statistical Institute*, Vol. 37, pp. 239-264.
- [25]- **Miller, R. G.** (1974): The jackknife: a review. *Biometrika*, Vol. 61, pp. 1-15.
- [26]- **Quenouille, M. H.** (1949): Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B.* Vol. 11, pp. 68-84.

- [27]- **Quenouille, M. H.** (1956): Notes on bias in estimation. *Biometrika*, Vol. 34, pp.353-360.
- [28]- **Rao, J. N. K.** (1965): A note on the estimation of ratios by Quenouille's method. *Biometrika*, Vol. 52, pp. 647-649.
- [29]- **Rao, J. N. K.** (1963): On two systems of unequal probability sampling without replacement. *Ann. Inst. Stat. Math.*, Vol. 15, pp. 67-72.
- [30]- **Rao, J. N. K. and Webster, J. T.** (1966): On two methods for bias reduction. *Biometrika*, Vol. 53, pp. 571-577.
- [31]- **Rao, J. N. K., Wu, C. F. J. and Yue, K.** (1992): Some recent work on resampling methods for complex surveys. *Surv. Methodol.* Vol. 18, pp. 209-217.
- [32]- **Razvi, N. A. and Khan, M.A.** (2013): Sampling procedure with unequal probabilities of selection an generalization for fixed sample size. *International Journal of Statistika and Matematika*, Vol (7), Issue (1), pp. 12-13.
- [33]- **Shahbaz, M. Q., and Hanif, M.** (2003a): Variance formulas for Horvitz-Thompson estimator using first order inclusion probabilities. *J. Applied Statistical Sciences*, Vol. 12, No. 3, pp. 201-208.
- [34]- **Shahbas, M. Q., and Hanif, M.** (2003b): A sample procedure for unequal probability sampling without replacement sample size 2. *Pak. J. Statist.* Vol. 19, No. 1, pp. 151-160.
- [35]- **Shahbaz, M. Q. and Hanif, M.** (2007): A new approximate formula for variance of Horvitz-Thompson estimator using first order inclusion probabilities. *Pak .J .Stat. Oper. Res.*, Vol. III, No. 1, pp. 59-62.
- [36]- **Shao, J. and Tu, D.** (1995): *The Jackknife and Bootstrap*. Springer-Verlag New York, Ink.
- [37] - **Tukey, J. W.** (1958): Bias and confidence in not-quite large samples (Abstract). *Ann. Math. Statist.*, 29, 614.
- [38]- **Yates, F. and Grundy, P.M.** (1953): Selection without replacement from within strata with probability proportional to size. *J. Roy. Stat. Soc. Ser. B.* Vol. 15, pp. 253-261.

A Generalized Linear Model for Female participation in Labor Force in Upper Egypt

Reem Ismail Elsybaey*

Yasmin Ibrahim **

Abstract:

In the last two decades, the world has an obvious interest for the role of women in the family and society as a participant in development as well as men. According to many studies there exists a relationship between the number of participated women in labor force and some other determinants like the number of educated, married women and place of residence. This paper will apply a Poisson Regression Model for the number of working women in Upper Egypt. It will identify the determinants that have an impact on women's participation in the labor force. The test of model effects showed that all explanatory variables have a significant effect on the response variable. Also the results showed that the data is over dispersed, as the variance of the response variable is not equal to the mean in both years 2011 and 2013 of labor force survey.

Keywords: Working Woman, Labor Force Survey, Deviance.

1. Introduction

In the last two decades, the world has showed an obvious interest in the role of women in the family and society as a participant in development as well as men. There are many conferences referred to the need to promote the social status of women and strengthen their role in the political, economic, and social development. The most important conference is the International Conference on Population and Development which is concerned with Women's issues, held in Cairo in 1994. It called for women's equal right with men in all areas and the elimination of discrimination against them. Women's participation in the labor force is faced by a lot of obstacles; perhaps the most important are conditions and social beliefs, gender discrimination between men and women, which represents one of the structural problems in the labor market in Egypt. Discrimination appears in partitioning of jobs and in wages levels which reduce the motivation for woman's participation in the labor force. The research program which carried out by the (Social Research Center, 2009) showed that, in spite of all the development efforts to wide the educational base and increasing in health care as well as institutional reforms of laws and regulations that have been made in Egypt in recent decades, which in turn guarantees the rights of women, the participation of women in the labor force is still low, with less than 22% in 2006.

*Reem Elsybaey: The Author is a Master degree Student in ISSR and Statistician at (CAPMAS) National Statistics Office in Egypt.

** Yasmin Ibrahim: Assistant Professor in ISSR (Institute of Statistical Studies and Research) – Egypt, Author's Supervisor.

Table (1): The evolution of Egyptian female's labor force participation in age group (15 – 64)

Year	Females Labor force participation rate (%)
1990	27.60
1991	23.30
1992	23.90
1993	23.30
1994	24.10
1995	22.70
1996	21.80
1997	21.00
1998	20.30
1999	20.90
2000	21.10
2001	21.40
2002	20.00
2003	20.70
2004	21.40
2005	22.10
2006	22.40
2007	24.40
2008	24.70
2009	25.00
2010	25.30
2011	25.50
2012	25.60
2013	25.80

Source: World Bank Data, 2013

Table (1) shows that Egyptian female's Labor force participation rate in 1990 was 27.6% then it decreased in 2002 to be 20%, while in 2013 it increased again to be 25.8%, however this percentage doesn't exceed the quarter of the total of labor force participation.

This paper will apply a Poisson Regression Model for the number of working females in labor force in Upper Egypt. It will identify the determinants that have an impact on women's participation in the labor force.

2. Review of Literature

According to the review of literature, (Anker, 1983) discussed female labour force participation in developing countries and introduced a critique of current definitions and data collection methods for measuring, women's work. (Anker and Anker, 1989) examined the female labour force participation using official labour force data in Egypt. They suggested how recent changes in the procedures and questionnaires used by Central Agency for Public Mobilization and Statistics (CAPMAS) to collect these data have affected responses and

consequently official estimates of the female labor force. This is obvious in the great efforts made by Egyptian Labor Force Sample Survey (ELFSS) to improve reporting of women in the labor force, by specifically training interviewers to be aware of unpaid family labor, particularly labor on family farms. As a result of these efforts, the percentage of women reported to be working doubled from 5.8 percent in 1982 to 12.7 percent in 1983. Furthermore, the results of the methodological household survey jointly conducted by CAPMAS and ILO (International Labor Organization) were described and discussed several important ways of improving data on the female labour force. Also (Assad, 2002) stated that Egypt has gone through a structural adjustment program since 1970 like many other developing countries. Nevertheless these reforms have been accompanied by significant In-formalization of employment relations where the Egyptian labor market in particular paid employment outside the government sector is becoming increasingly de-feminized over this structural adjustment period. (Langsten and Salem, 2006) introduced two approaches to measure Women's Work in Developing Countries applied to the same population of ever-married women. The data of ever-married women were compared from two surveys which share the same basic definition of work, the first one is Egypt DHS (Demographic health survey) in 2003 and the second is Slow Fertility Transition (SFT) survey conducted in 2004. (Amer , 2007) studied the insertion of the youth in age group (15-29) years old into the Egyptian labor market. This study examined whether the patterns of school-to-work transition changed over time (1988-2006). Also (Spierings and Smits, 2007) discussed Women's labor market participation in MENA- region. This study stated that Women's participation in the economy, as integral part of their empowerment, is a major challenge facing the MENA-region, since women's labour market participation (LMP) is among the lowest throughout the world in these countries. (Sieverding, 2010) reviewed the Life Course Transitions and Female Labor Force Outcomes in Egypt. It used a longitudinal approach to the question of how young Egyptian women's labor force participation in different sectors of the economy changes after transition into marriage. It was limited to unmarried young women in 1998 and examined their labor force outcomes in 2006, then compared with those who had married in the intervening eight years and those who had not. The (population council, 2012) made a Survey of young people in Egypt. This survey stated that female youth face challenges in finding decent work both as women and as young people. Consequently, labor force participation among female youth remains low; also education and marriage are strongly associated with labor force participation. It's obvious that higher educational attainment is associated with greater labor force participation for female youth. According to (MENA Development Report, 2013) the worldwide average of the participation of women in the labor force approximately 50 percent while this percentage in MENA is half that at 25 percent. It is a fact that young women especially have high rates of

unemployment. This report discussed the improvements made to secure equal opportunities for women alongside men.

3. Methodology

This study will use data of (labor force survey for females in six of Upper Egypt governorates in years 2011 and 2013) due to available data on workers and the unemployed, whether male or female. The upper Egyptian females in the governorates (Benisuef - Fayoum- Menia- Asyuot- Suhag- Qena) were selected from the data of labor force survey. Thus, the final observations consisted of 30535 females in 2011 while in 2013 it was 27917. The study used a poisson regression model to test the significance of some factors and it will use two measures of goodness-of-fit like deviance and AIC. In statistics, poisson regression is a form of regression analysis used to model count data and contingency tables. It assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modelled by a linear combination of unknown parameters. In the case of rare events the Poisson distribution is more appropriate since the Poisson mean > 0 . So the logarithm of the response variable is linked to a linear function of explanatory variables such that $\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots$ etc. and so $Y = (e^{\beta_0}) (e^{\beta_1 X_1}) (e^{\beta_2 X_2}) \dots$ etc. This study used the Poisson regression (PR) model to identify the relationship between the endogenous variable "number of working women" and some other covariates such as number of educated women, number of married women, number of women in age group 15-64 and place of residence which is a dummy variable and consists of (urban =0; rural =1).

Assumptions in Poisson Regression:

The assumptions include:

1. Logarithm of the dependent variable changes linearly with equal increment increases in the exposure variable.
2. Changes in the rate from combined effects of different exposures or risk factors are multiplicative.
3. The number of cases has variance equal to the mean.
4. Observations are independent.

4. Results

Table (2): The descriptive statistics for the data in 2011 and 2013

Years Variables	2011			2013		
	Number of Shiakha	Mean	Std. Deviation	Number of Shiakha	Mean	Std. Deviation
Number of working women (Work)	146	34.53	33.55	147	25.48	22.591
Number of Educated women (Edu)	146	74.65	51.338	147	73.63	45.19
Number of Married women (Mar)	146	114.06	60.941	147	100.59	52.866
Number of women in age group (15-64)	146	176.18	94.864	147	161.94	85.837

- The upper Egyptian females in the governorates (Benisuef - Fayoum- Menia- Asyut- Suhag- Qena) were selected from the data of labor force survey. Thus, the final observations consisted of 30535 females in 146 Shiakha in year 2011 while in 2013 it was 27917 females in 147 Shiakha.
- The unconditional mean and variance of the response variable are very different in both years. The results show that the variance is greater than the mean which means that there is over dispersion in the data. Furthermore, the estimates in Poisson Regression model are still consistent but are inefficient, which leads to the invalidation of inference based on the estimated standard errors. In this case the Negative binomial regression can be used for over-dispersed count data. It can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the over-dispersion.

Multicollinearity Diagnostics:

The condition Index K can be used to test if there is a correlation between Explanatory variables or not and K is calculated as follows:

$$K = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \quad \text{where } \lambda_{\max} \text{ is the maximum Eigen Value and } \lambda_{\min} \text{ is the minimum Eigen Value.}$$

If the value of K is greater than 15 ($K \geq 15$) then the data used in the analysis has a correlated explanatory variables (Multicollinearity). Also if ($K > 30$) then there is a high correlation between the explanatory variables.

Table (3): Multicollinearity Diagnostics Results

Exp. Variables	2011		2013	
	Eigen value	Condition Index	Eigen value	Condition Index
Intercept	4.056	1.000	4.070	1.000
Ur_Ru	.730	2.357	.724	2.371
Edu	.173	4.839	.164	4.982
Married	.038	10.386	.040	10.131
Age	.003	36.514	.003	38.894

According to table (3) the explanatory variable (Number of women in age group (15-64)) is highly correlated with the other variables. After omitting this variable and checking the Multicollinearity again, the results were as follows:

Table (4): Multicollinearity Diagnostics (Cont.)

Exp. Variables	2011		2013	
	Eigen value	Condition Index	Eigen value	Condition Index
Intercept	3.106	1.000	3.117	1.000
Ur_Ru	.704	2.101	.702	2.108
Edu	.156	4.462	.146	4.615
Mar	.034	9.541 < 15	.035	9.440 < 15

The results in table (4) showed that the data used in the analysis has no correlated explanatory variables. Thus all the variables can be included in the analysis with Poisson regression Model.

Table (5): Goodness of fit

Goodness of Fit	2011			2013		
	Value	df	Value/df	Value	df	Value/df
Deviance	2971.306	142	20.925	1777.360	143	12.429
Pearson Chi-Square	3076.850	142	21.668	1790.216	143	12.519
Log Likelihood	-1833.021			-1231.674		
Akaike's Information Criterion (AIC)	3674.042			2471.348		
Bayesian Information Criterion (BIC)	3685.976			2483.310		

To assess the fit of the model, the goodness-of-fit chi-squared test and deviance are used, where the deviance statistic for Poisson Regression Model can be thought of as a measure of residual variation about the fitted model. Also the deviance $D(\hat{\beta})$ is assumed to have an approximate chi-square distribution with $n-k-1$ degrees of freedom. Furthermore the deviance is usually defined as the log likelihood of the final model, multiplied by (-2) . The deviance in 2011 was (2971.306) as Chi-square distributed with the model degrees of freedom (142), while in 2013 the deviance was less than 2013 and equal to (1777.360) with the model degrees of freedom (143).

Table (6): Poisson Regression results for the data of 2011 and 2013:

Exp. Variables	2011				2013			
	B	Std. Error	Sig.	Exp(B)	B	Std. Error	Sig.	Exp(B)
Intercept	3.103	.0572	0.000	22.263	2.648	.0647	0.000	14.119
Ur_Ru = 0 or 1	-.441	.0457	0.000	0.643	-.197	.0489	0.000	0.821
Number of educated women (Edu)	0.004	.0005	0.000	1.004	0.004	.0006	0.000	1.004
Number of married (women (Mar)	0.009	.0003	0.000	1.009	0.009	.0005	0.000	1.009

- Recall the form of the model equation in years 2011 and 2013:

In 2011:

$$\text{Log (Work)} = 3.103 - 0.441 (\text{Ur_Ru}) + 0.004 (\text{Edu}) + 0.009 (\text{Mar}).$$

This implies that:

$$\begin{aligned} \text{Work} &= \text{Exp} [3.103 - 0.441 (\text{Ur_Ru}) + 0.004 (\text{Edu}) + 0.009 (\text{Mar})]. \\ &= \text{Exp} (3.103) * \text{Exp} (-0.441 \text{Ur_Ru}) * \text{Exp} (+0.004 \text{Edu}) \\ &\quad * \text{Exp} (0.009 \text{Mar}). \end{aligned}$$

In 2013:

$$\text{Log (Work)} = 2.648 - 0.197 (\text{Ur_Ru}) + 0.004 (\text{Edu}) + 0.009 (\text{Mar}).$$

This implies that:

$$\begin{aligned} \text{Work} &= \text{Exp} [2.648 - 0.197 (\text{Ur_Ru}) + 0.004 (\text{Edu}) + 0.009 (\text{Mar})]. \\ &= \text{Exp} (2.648) * \text{Exp} (-0.197 \text{Ur_Ru}) * \text{Exp} (+0.004 \text{Edu}) \\ &\quad * \text{Exp} (0.009 \text{Mar}). \end{aligned}$$

- From table (6), the intercept is the Poisson regression estimate when all variables in the model are evaluated at zero. The log of the expected count for the Number of working women is 3.103 units in 2011 and in 2013 it was 2.648 units.
- The Educational level is considered one of the most important variables that have a significant effect on the number of working women in Upper

Egypt. The coefficient for the Number of educated women in years 2011 and 2013 was (0.004) units. This means that, for a one unit increase in the number of educated females the expected log count for the number of working females is expected to change by 0.004.

- Table (6) also refers to the important significant effect for the marital status, where the coefficient for the Number of married females in 2011 and 2013 was 0.009 units. This means that, for a one unit increase in the number of Married females the expected log count for the number of working females is expected to change by 0.009.
- The indicator variable Ur_Ru is the estimated Poisson regression coefficient comparing Rural to Urban. The coefficient for this dummy variable residence in the two years 2011 and 2013 is negative and statistically significant indicating that rural regions have fewer working females than urban regions.
- From table (6) we can notice that female's place of residence or female's region (Ur_Ru) has a significant effect on the number of working females. Also in both years 2011 and 2013 the expected log count of working females who are in rural areas decreased by about 36% and 18% respectively compared to the expected log counts of working females who are in urban areas.
- According to over dispersion the data is over dispersed, where the variance of the response variable is too large and not equal to the mean in both years 2011 and 2013.
- We tested over dispersion in the Poisson regression model using the ratio of the sum of Pearson chi square over the number of degrees of freedom (χ^2/df) or the (deviance / df). The results show that in 2011 the (χ^2/df) = 21.668 which is greater than 1 and in 2013 (χ^2/df) = 12.519 which confirmed over dispersion, however the data of 2013 has a lower value of dispersion than the data of 2011.
- For over-dispersed count data, Negative binomial regression Model is usually used for modeling this type of count variables. In this case of over dispersion, the Negative Binomial Regression Model can be used to handle this problem. Also it has the same form as the Poisson regression model and its parameter estimation performed by using MLE method. Bouche, (2009) discussed in a paper different methods that exist to solve the problem of underestimating variance in the Poisson regression model when over dispersion is present. This paper stated that the negative binomial regression model seems to be particularly accurate because of its theoretical distribution.

Table (7): Comparison between the results of Goodness of fit for Poisson Regression and Negative Binomial in years 2011 and 2013:

Goodness of Fit (2011)	Poisson Regression			Negative Binomial		
	Value	df	Value/df	Value	df	Value/ df
Deviance	2971.306	142	20.925	168.784	141	1.197
Pearson Chi-Square	3076.850	142	21.668	125.296	141	.889
Log Likelihood	-1833.021			-644.754		
Akaike's Information Criterion (AIC)	3674.042			1299.508		
Bayesian Information Criterion (BIC)	3685.976			1314.426		

Goodness of Fit (2013)	Poisson Regression			Negative Binomial		
	Value	df	Value/df	Value	df	Value/ df
Deviance	1777.360	143	12.429	160.670	142	1.131
Pearson Chi-Square	1790.216	143	12.519	134.594	142	.948
Log Likelihood	-1231.674			-599.834		
Akaike's Information Criterion (AIC)	2471.348			1209.669		
Bayesian Information Criterion (BIC)	2483.310			1224.621		

- Over-dispersion was tested in the Negative Binomial regression model using the ratio of the sum of Pearson chi square over the number of degrees of freedom (χ^2/df) or the (deviance / df). The results according to Poisson Regression Model in 2011 and 2013 was (χ^2/df) = (21.668, 12.519) respectively. However, the results of Negative Binomial Regression Model decreased to be (1.2, 1.1).

- According to Negative Binomial Regression Model the over-dispersion decreased more than the Poisson Regression Model.

Table (8): Comparison between the results of Negative Binomial regression Model in years 2011 and 2013:

Exp. Variables	2011				2013			
	B	Std. Error	Sig.	Exp(B)	B	Std. Error	Sig.	Exp(B)
Intercept	3.066	.3213	0.000	21.446	2.633	.2478	0.000	13.909
Ur_Ru = 0 or 1	-.524	.0750	0.000	.592	-.229	.0961	0.000	.796
(Edu)	.007	.0033	.046	1.007	.003	.0027	.238	1.003
(Mar)	.011	.0024	0.000	1.011	.009	.0021	0.000	1.009
Negative Binomial	0.755	.0908	—	—	0.516	.0640	—	—

- The Parameter Estimates includes the regression coefficients for each of the variables along with robust standard errors, p-values and 95% confidence intervals for the coefficients.
- From table (8), the intercept is the Negative Binomial Regression estimate when all variables in the model are evaluated at zero. The log of the expected count for the Number of working women is 3.066 units in 2011 and in 2013 it was 2.633 units.
- According to the results of Negative Binomial Regression Model the number of educated females has an insignificant effect on the number of working women in Upper Egypt. The coefficient for the Number of educated women in years 2011 and 2013 was (0.007, 0.003) units respectively. This means that, for a one unit increase in the number of educated females the expected log count for the number of working females is expected to change by (0.007 , 0.003) units respectively.
- Table (8) also refers to the important significant effect for the marital status, where the coefficient for the Number of married females in 2011 and 2013 was (0.011, 0.009) units. This means that, for a one unit increase in the number of Married females the expected log count for the number of working females is expected to change by 0.011 in year 2011 and will change by 0.009 in 2013.
- The indicator variable Ur_Ru is the estimated Negative Binomial regression coefficient comparing Rural to Urban. The coefficient for this dummy variable residence in the two years 2011 and 2013 is negative and statistically significant indicating that rural regions have fewer working females than urban regions.
- From table (8) we can notice that female's place of residence or female's region (Ur_Ru) has a significant effect on the number of working females. Also in both years 2011 and 2013 the expected log count of

working females who are in rural areas decreased by about 41% and 21% respectively compared to the expected log counts of working females who are in urban areas.

- Thus, we would expect the urban areas to have more number of working females than their rural counterparts.
- The dispersion parameter in years 2011 and 2013 was equal to (0.755, 0.516) respectively. The Negative Binomial Regression Model was able to handle and decrease over dispersion in the count data for both years 2011 and 2013. However, the over-dispersion still exists because the value of the dispersion parameter is greater than zero.
- Also the Generalized Poisson Regression Model (GPR) can be used to handle over-dispersion where the GPR is a generalization of the standard Poisson regression (PR) model.

Table (9): Choosing the best Model

Model	Significant Variables	AIC (2011)	AIC (2013)
Poisson Regression	Ur_Ru Edu Mar	3674.042	2471.348
Negative Binomial	Ur_Ru Mar	1299.508	1209.669

- For Akaike's Information Criterion (AIC) in table (10) the smaller the AIC, the better the model is.
- The smallest AIC value is a negative binomial regression model. Then the best model for the number of working females is obtained from the negative binomial.

5. Conclusions

- The test of model effects showed that the number of educated and married females have a significant effect on the number of working women in Upper Egypt in both years 2011 and 2013.
- Female's place of residence or female's region (Ur_Ru) has a significant effect on the number of working females. Also in both years 2011 and 2013 the expected log count of working females who are in rural areas decreased by about 36% and 18% respectively compared to the expected log counts of working females who are in urban areas.
- According to over dispersion the data is over dispersed, as the variance of the response variable is too large and not equal to the mean in both years 2011 and 2013.
- The Negative Binomial Regression Model was able to handle and decrease over dispersion in the count data for both years 2011 and 2013.

- The Negative Binomial is more adequate and appropriate in the case of over-dispersion than Poisson Regression.

Acknowledgment:

Authors would like to thank Dr. Amany Moussa professor of Institute of Statistical Studies and Research for supervising the research.

6. References

- Anker, R., and Anker, M., (1989), *Measuring the Female Labour Force in Egypt*, International Labour Review.
- Anker, R., 1983, *Female Labour Force Participation in Developing Countries: A Critique of Current Definitions and Data Collection Methods*, International Labour Review.
- Amer, M., (2007), *The Egyptian Youth Labor Market School-to-Work Transition 1998-2006*, ERF.
- Assaad, R., (2002), *Informalization and De-feminization: Explaining the Unusual Pattern in Egypt*, Cornell University.
- Bouche, G., (2009), *Application of Detecting and Taking Over-dispersion into Account in Poisson Regression Model*, University of Poitiers.
- Cameron, A.C, and Trivedi, P.K, (1998), *Regression Analysis of Count data*, Cambridge University press, New York.
- Elhadad, A., (2009), *Low Participation of Women in the Formal Labor Market and Wage Discrimination*, Social Research Center, AUC.
- Kleinbaum, D.G., Kupper, L.L. Muller, K.E., and Nizam, A., (2007), *Applied Regression Analysis and Other Multivariate Methods*, Thomson learning.Inc, 687-709.
- "Labor Force survey", (2011), Central Agency for Public Mobilization and Statistics.
- "Labor Force survey", (2013), Central Agency for Public Mobilization and Statistics.
- MENA Development Report, (2013), "Gender Equality and Development in the Middle East and North Africa", World Bank, Washington, DC.
- Sieverding, M., (2010), *Life Course Transitions and Female Labor Force Outcomes in Egypt*, California University.
- Spierings, N., and Smits, J., (2007), *Women's Labour Market Participation in Egypt, Jordan, Morocco, Syria & Tunisia: A three-level Analysis*, Radboud University.
- "Survey of Young People in Egypt, Policy Brief Series", (2012), Population Council.

A Panel Data Analysis for Factors Effecting Employment on Egyptian Main Economic Sectors

Naglaa A. Morad *

Soad sayed Abd El rasoul**

Abstract: In this paper, factors affecting employment is conducted in three important economic sectors (government, private and public sectors). A random sample from workers within enterprises in three sectors from Egyptian Labor Market Panel Survey (ELMPS) research, that was conducted at (1998, 2006, 2012), to study the effect of main factors (age, labor hours, educational status) on employees' wages. We chose 20 employees randomly from row data of ELMPS for government, private and public sectors, and follow the wage, hours, age and education status to these employees in the three periods (1998, 2006 and 2012).

Several models will be tested to determine the one that achieve the goal of this paper, and to study the effect of some factors like (age, labor hours, educational status) on employees' wages.

Keywords: Panel Data, Fixed Effect Model, Random Effect Model, Dummy variables, OLS estimator, log linear model, Average wage.

* Associated Professor, Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University

**Master student, Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University.

Introduction

The labor elements considered the most important elements of production and the essential foundation to improve the standard of living, and human resources development that depend on its states in the strategic planning for achieve sustainable development .wages and various compensations that relates to them the most important is the fundamental rights of workers, and is one of the most important obligations to the worker and the study of employment and wages under the current circumstances to determine the extent of its impact of current events on the level of wages and is it positive or negative impact, as well as employment .

Using Panel Data Model's, such as pooled model, fixed effect model and random model to determine which model is suitable for the Egyptian data,

Review of Literature Introduction

In recent years, there has been a growing interest in the specification and estimation of econometric relationships based on panel data. This interest can be explained by the fact that panel data offer researchers extended modeling possibilities as compared to purely cross-sectional data or time-series data. Panel data are generally more informative, and they contain more variation and less collinearity among the variables. The use of panel data results in a greater availability of degrees of freedom and hence increases efficiency in the estimation. Panel data also allow for the specification of more complicated behavioral hypotheses, including effects that cannot be addressed using pure cross-sectional or time-series data.

Elhorst (2003), showed a survey of the specification and estimation of spatial panel data models. This paper displayed four panel data models commonly used in applied research:

the fixed effects model, the random effects model, the fixed coefficients model, and the random coefficients model. The survey discusses the asymptotic properties of the estimators and provides guidance with respect to the estimation procedures, which should be useful for practitioners.

Baltagi (2005), provided an coverage of panel data techniques, especially for serial correlation, spatial correlation, heteroskedasticity, seemingly unrelated regressions, simultaneous equations, dynamic models, incomplete panels, limited dependent variables and nonstationary panels. Many of the estimation and testing techniques are illustrated with data sets which are available.

Reyna (2007), Explained two techniques used to analyze panel data (fixed and random effects), with fixed effects model showed (covariance model, within estimator, individual dummy variable model, least squares dummy variable model), while in random model showed random intercept, partial pooling model, then showed some panel data tests used STATA program.

Clark and Linzer (2012), showed two widely-used methods are the use of either “fixed” or “random” affects models, chosen between these approaches remains unclear in the applied literature. They employed a series of simulation experiments to evaluate the relative performance of fixed and random effects estimators for varying types of datasets.

Summery about the Egypt Labor Market Panel Survey (ELMPS)

The Egypt Labor Market Panel Survey, carried out by the Economic Research Forum (ERF) in cooperation with Egypt’s Central Agency for Public Mobilization and Statistics (CAPMAS) since 1998, has become the mainstay of labor market and human resource development research in Egypt, being the first and most comprehensive source of publicly available micro data on the subject. The 2012 round of the survey provides a unique opportunity to ascertain the impact of the momentous events accompanying the January 25th

revolution on the Egyptian economy and labor market and on the lives of Egyptian workers and their families. The Egypt Labor Market Panel Survey of 2012 (ELMPS 2012) is the third round of this longitudinal survey, which was also carried out in 2006.

Although the sample was designed to be nationally representative, this round and all subsequent rounds do not cover the five Border Governorates of Matruh, New Valley, Red Sea, North and South Sinai. These remote governorates are sparsely populated, containing no more than 2% of the population in 2006.(Assaad and Krafft , 2013).

Table(1.1): Estimate the Employees Numbers by Sectors

Government	Public	Private Within establishments	Investment
5,325,500	932,700	6,219,400	397,300

Source: Annual bulletin of labour force (2014).

Table(1.2) : Average Weekly Hours Worked and Average Daily Wage (L.E) of Workers Paid by Important Economic Activities

economic activities	average weekly hours	average daily wage (L.E)
Agriculture, forestry and fishing	44.4	48.3
Mining and quarrying	47.6	79.3
Manufacturing	50	55.1
Electricity, gaz, steam and air conditioning supply	46.7	77.7
Water supply, sewerage, waste management and remediation activities	47.1	54.5
Education	43	56.8
Human health and social work activities	45.3	50.2

Source: Annual bulletin of labour force (2014)

Models that will be applied in the study

1) Factors Affecting Wages in the Government Sector

1.1) fixed effect or LSDV model:

The equation for the fixed effects model becomes:

$$\text{Wage}_{it} = \alpha_i + \beta_1 \text{Age}_{it} + \beta_2 \text{Hours}_{it} + u_{it} \quad (1.1)$$

where

i is number of employees (i=20)

t is time (t=3 years)

α_i (i=1...n) is the unknown intercept for each entity (n entity-specific intercepts).

u_{it} is the error term. Allows for heterogeneity or individuality among employees by allowing to have its own intercept value. After testing fixed effect model, from table (1.1) the significant level of the full model was founded =0.01< 0.05, it means that some the coefficient of this model are not equal zero and model is good. By studying the effect of every independent variable on the dependent variable, it was found that all variables are significant.

Table (1.3) Fixed Effect or LSDV Test for Governmental Sector:

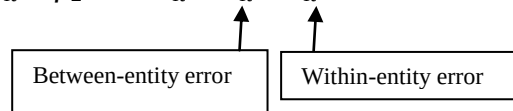
Variables	Estimated Parameters	S.E	P-value	Model Performance
Age	422.5	78.7	0.0001	Prob(F-statistic)=0.0001
Hours	-1116.7	450.3	0.02	R-squared=0.59
Constant	-5803.5			

Computed from ELMPS row data in (1998, 2006 and 2012).

Using E-Views program.

1.2) Random Effect model:

$$\text{Wage}_{it} = \alpha_i + \beta_1 \text{Age}_{it} + \beta_2 \text{Hours}_{it} + u_{it} + \varepsilon_{it} \quad (1.2)$$



The test for Random Effect model, from table (1.2), indicates that the significant level of the full model was $= 0.0001 < 0.05$, it means that some of the coefficient of this model are not equal zero and model is good. By studying the effect of every independent variable on the dependent variable, it was found that all variables are significant.

Table (1.4) Random Effect Test for Governmental Sector:

Variables	Estimated parameters	S.E	P-value	Model Performance
Age	292.1	62.5	0.0001	Prob(F-statistic)=0.0002
Hours	-1003	389.6	0.01	R-squared=0.26
Constant	-1545.1			

Computed from ELMPS row data in (1998, 2006 and 2012).

Using E-Views program.

To which model is appropriate? Fixed or Random by using Hausman test.

Hausman test

$$H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})' [\text{var}(\hat{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE})]^{-1} (\hat{\beta}_{RE} - \hat{\beta}_{FE})$$

Null hypothesis: Random-effects model appropriate.

Alternative hypothesis: Fixed Effect model is appropriate.

The Hausman test probability value =0.001 < α that means rejects null hypotheses and accept Alternative hypothesis fixed effect is appropriate model to explain wage, there is a positive relationship between age, work hours and wage.

The fixed effect model is

$$\hat{Wage}_{it} = -5803.5 + 422.5 Age_{it} - 1116.7 Hours_{it} \quad (1.3)$$

The effect of dummy variables on dependent variable and model is tested as follows:

$$Wage = C(1) + C(2)*age + C(3)*hours + C(4)*intermediate + C(5)*above_intermediate + C(6)*university + C(7)*age*intermediate + c(8)*age*above_intermediate + c(9)*age*university + c(10)*hours*intermediate + c(11)*hours*above_intermediate + c(9)*hours*university. \quad (1.4)$$

The dummy variables used to estimate fixed effect model

Null Hypothesis: All dummy variables are equal to zero=pooled regression model.

Alternative hypothesis: fixed effect model.

By using Wald test to check whether all dummy variables are zero or not from E-views result the F-statistic=0.9 > 0.05 that means accept the null Hypothesis: All dummy variables are equal to zero, pooled regression model is a good model.

In this pooled regression model, all observations in OLS regression are pooled meaning that we assumed that the coefficients (including the intercepts) are the same for all the individuals.

The appropriate model for governmental sector is model (1.3):

$$\hat{Wage}_{it} = -5803.5 + 422.5 Age_{it} - 1116.7 Hours_{it}$$

2) Factors affecting wages in Public sector

First we will estimate fixed effect model, then estimate random effect model and test which model is appropriate to data.

2.1) fixed effect or LSDV model:

The equation for the fixed effects model becomes:

$$\text{Wage}_{it} = \alpha_i + \beta_1 \text{Age}_{it} + \beta_2 \text{Hours}_{it} + u_{it} \quad (1.4)$$

where

i is number of employees (i=20)

t is time (t=3)

α_i (i=1...n) is the unknown intercept for each entity (n entity-specific intercepts).

u_{it} is the error term

Allows for heterogeneity or individuality among employees by allowing to have its own intercept value. The test for fixed effect model, from table (1.3) the significant level of the full model was found $=0.04 < 0.05$, it means that some of the coefficients of this model are not equal zero and model is good. By studying the effect of every independent variable on the dependent variable, it was found that age variable is significant but, work hours are not significant.

$$\hat{\text{Wage}}_{it} = -4287.3 + 322.8 \text{Age}_{it} \quad (1.5)$$

Table (1.5) Fixed Effect or LSDV Test for Public Sector:

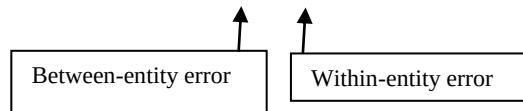
Variables	Estimated parameters	S.E	P-value	Model performance
Age	322.8	85.5	0.0005	Prob(F-statistic)=0.04 R-squared=0.52
Hours	-657.6	421.7	0.13	
Constant	-4287.3			

Computed from ELMPS row data in (1998, 2006 and 2012).

Using E-Views program.

2.2) Random Effect model:

$$\text{Wage}_{it} = \alpha_i + \beta_1 \text{Age}_{it} + \beta_2 \text{Hours}_{it} + u_{it} + \varepsilon_{it} \quad (1.6)$$



The test for Random Effect model, from table (1.4) significant level of the model was found $=0.002 < 0.05$, it means that some of the coefficients of this model are not equal zero and our model is good. By studying the effect of every independent variable on the dependent variable, it was found that age variable is significant but, work hours variable is not significant.

$$\text{Wage}_{it} = -1503.9 + 1881.1 \text{Age}_{it} \quad (1.7)$$

Table (1.6) Random Effect Test for Public Sector:

Variables	Estimated parameters	S.E	P-value	Model performance
Age	188.1	56.7	0.002	Prob(F-statistic)=0.002 R-squared=0.19
Hours	-316	364.9	0.4	
Constant	-1503.9			

Computed from ELMPS row data in (1998, 2006 and 2012).

Using E-Views program.

To test which model is appropriate? Fixed or Random by using Hausman test.

Null hypothesis: Random-effects model appropriate.

Alternative hypothesis: Fixed Effect model is appropriate.

For Hausman test the probability value =0.03 < 0.05 that means rejects null hypotheses and accept Alternative hypothesis fixed effect is appropriate model to explain wage, there is a positive relationship between age and wage.

The fixed effect model is

$$\hat{Wage}_{it} = -4287.3 + 322.8 Age_{it}$$

The effect of dummy variables on dependent variable and model is tested as follows:

$$\begin{aligned} Wage = & C(1) + C(2)*age + C(3)*hours + C(4)*less_intermediate + C(5)*intermediate + c(6)*above \\ & _intermediate + C(7)*university + C(8)*age*less_intermediate + C(9)*age*intermediate + c(10)* \\ & age*above_intermediate + C(11)*age*university + c(12)*hours*less_intermediate + C(13)*hour \\ & s*intermediate + c(14)*hours*above_intermediate + C(15)*hours*university. \end{aligned} \quad (1.8)$$

The dummy variables is used to estimate fixed effect model

Null Hypothesis: All dummy variables are equal to zero=pooled regression model .

Alternative hypothesis: fixed effect model.

By using Wald test to check whether all dummy variables are zero or not from the result the P-value for F-statistic=0.001 < 0.05 that means reject the null Hypothesis: All dummy variables are equal to zero, and the appropriate model for public sector is:

$$\hat{Wage} = 209926.8 + 322.8 *age - 220186.5*less_intermediate - 219603.3*intermediate - 216671.8*above_intermediate -$$

$$213326.7*university+2527.8*age*less_intermediate+2528.6*age*intermediate+2685.5*age*above_intermediate+2509.8*age*university+11764.7*hours*less_intermediate+12009*hours*intermediate+10346.3*hours*above_intermediate+11345*hours*university. \quad (1.9)$$

3) Factors affecting wages in private sector

First estimate the fixed effect model then estimate the random effect model and test which model is appropriate to data.

3.1) fixed effect or LSDV model :

The equation for the fixed effects model becomes:

$$Wage_{it} = \alpha_i + \beta_1 Age_{it} + \beta_2 Hours_{it} + u_{it} \quad (1.10)$$

From table (1.5) the significant level of the full model was founded $=0.005 < 0.05$, it means that some the coefficients of this model are not equal zero and model is good. By studying the effect of every independent variable on the dependent variable, it was found that age variable is significant but, work hours are not significant.

$$Wage_{it} = -6960 + 176 Age_{it} \quad (1.11)$$

Table (1.7) Fixed Effect or LSDV Test for Private Sector:

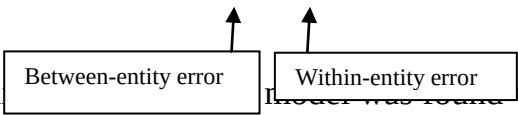
Variables	Estimated parameters	S.E	P-value	Model performance
Age	176	70.8	0.02	Prob(F-statistic)=0.0001
Hours	328	345	0.35	R-squared=0.73
Constant	-6960			

Computed from ELMPS row data in (1998, 2006 and 2012).

Using E-Views program.

3.2) Random Effect model:

$$\text{Wage}_{it} = \alpha_i + \beta_1 \text{Age}_{it} + \beta_2 \text{Hours}_{it} + u_{it} + \varepsilon_{it} \quad (1.12)$$

From table (1.6) sign  = 0.01 < 0.05, it means that some of the coefficients of this model are not equal zero and our model is good. By studying the effect of every independent variable on the dependent variable, it was founded that age variable is significant but, work hours variable is not significant.

$$\text{Wage}_{it} = -4393.4 + 95.7 \text{Age}_{it} \quad (1.13)$$

Table (1.8) Random Effect Test for Private Sector:

Variables	Estimated parameters	S.E	P-value	Model performance
Age	95.7	48.3	0.04	Prob(F-statistic)=0.01
Hours	364.7	249.8	0.15	R-squared=0.15
Constant	-4393.4			

Computed from ELMPS row data in (1998, 2006 and 2012).

Using E-Views program.

To test which model is appropriate? Fixed or Random by using Hausman test.

Null hypothesis: Random-effects model appropriate.

Alternative hypothesis: Fixed Effect model is appropriate.

For Hausman test the probability value = 0.001 < 0.05 that means rejects null hypotheses and accept Alternative hypothesis fixed effect is appropriate model to explain wage, there is a positive relationship between age and wage.

The fixed effect model is

$$\widehat{Wage}_{it} = -6960 + 176 Age_{it}$$

To check the effect of dummy variables on dependent variable and model that is to test:

$$salary = c(1) + c(2)*age + c(3)*hours + c(4)*illiterate + c(5)*Intermediate + c(6)*University + c(7)*age*illiterate + c(8)*age*Intermediate + c(9)*age*University + c(10)*hours*illiterate + c(11)*hours*Intermediate + c(12)*hours*University. \quad (1.14)$$

Null Hypothesis: All dummy variables are equal to zero=pooled regression model.

Alternative hypothesis: fixed effect model.

By using Wald test to check whether all dummy variables are zero or not from the result the P-value for the F-statistic = 0.001 < 0.05 that means reject the null Hypothesis: All dummy variables are equal to zero, and the appropriate model for private sector is:

$$\widehat{Wage}_{it} = -656.3 + 176 Age_{it} + 1054*illiterate - 2107.7*intermediate - 5149.9*university + 202.4*age*illiterate + 290*age*intermediate + 389.8*age*university - 1079*hours*illiterate - 1037.8*hours*intermediate - 973*hours*university. \quad (1.15)$$

4) Conclusions and Recommendation

This paper claims to determine the factors that affect average wages of the governmental, public and private sectors of Egypt.

Using panel data model, the results are

- 1- Fixed Effect Model is a good model for panel data model on governmental, public and private sectors.
- 2- Concerning the governmental sector, age has positive an effect on the wages, while the working hours have negative impact on the wages, if age increased by one unit, the wages would increase by 422.5 units. And if working hours increased by one unit, the wages would decrease by 1116.7 units, the educational status does not affect the

wages for the government sector, this is logic, when employee's age increases, wage will increase which called career progression in the government sector.

- 3- Concerning the public sector, age and the educational status have an effect on the wages, age has a positive effect on the wage, while the educational status have a negative effect but the educational status interaction with age and working hours have positive effect, while the working hours haven't an effect on the wages, the interaction between working hours and the educational status have an effect on the wages, if age increased by one unit, the wages would increase by 322 units.
- 4- Concerning the private sector, age and the educational status have an effect on the wages, age has a positive effect on the wages, while the educational status (illiterate and intermediate) have a negative effect but the educational status university has positive effect, the educational status interaction with age have positive effect, but the educational status interaction with working hours have negative effect, while the working hours haven't effect on the wages, if age increased by one unit, the wages would increase by 176 units.
- 5- Fixed effect model is a good model for main Egyptian sectors and for Egyptian labor market.

Table (1.10) Illustrate the Above Results Significant Factors for the Selected Model

Sector	Selected Model	Sign. Variables
Government	OLS Model	Age & Hours
Public	Fixed Effect Model	Age & Education status& Interaction between Age and Education status& Interaction between Hours and Education status
Private	Fixed Effect Model	Age & Education status& Interaction between Age and Education status& Interaction between Hours and Education status

References

1. Assaad, R. and Krafft, C. (2013), "*The Egypt Labor Market Panel Survey: Introducing The 2012 Round*", IZA Journal of Labor & Development.
2. Baltagi, H. B. (2005), "*Econometric Analysis of Panel Data*", 3rd ed. British Library Cataloguing in Publication Data.
3. Clark, S.T. Linzer, A.D. (2012), " *Should I Use Fixed or Random Effects?*", Department of Political, Science Emory University.
4. Elhorst, P.J. (2003). "*Specification and Estimation of Spatial Panel Data Models*", International Regional Science Review 26, 3: 244–268.
5. Reyna, T.O. (2007), "*Panel Data Analysis Fixed and Random Effects Using Stata*", Princeton University.
6. Sited From The Internet
 1. Andreica, E.M. Cristescu, A. and Pirciog, S. (2011), "*Simulation Scenarios of Employment on The Romanian Labor Market*",
<http://www.wseas.us/elibrary/conferences/2011/Prague/AICT/AICT-44.pdf>.
 2. Yaffee, A.R. (2003), "*A Primer for Panel Data Analysis* ",
http://www.nyu.edu/its/pubs/connect/fall03/yaffee_primer.html, Updated April 2005.
7. النشره السنوية لإحصاءات القوى العاملة - الجهاز المركزي للتعبئه العامه والاحصاء

Some Properties of Double k -class Estimators in Linear Regression Under Multivariate t -errors

Abdel Raheim Mohamed Abdel Raheim*

Abstract

This paper introduce a new analysis for the exact and asymptotic bias, moment matrix, and risk function for double k -class estimators in general linear regression model under multivariate t -errors. This paper study the performance of the double k -class estimators in linear regression model under multivariate t -errors. compare between double k -class estimators, and ordinary least squares estimators in the same regression model.

Keywords: bias estimators; double k -class estimators; multivariate t -errors; and risk function.

1- Introduction

It is well known that the non-linear and biased estimators of regression coefficients in a linear regression model can have smaller risk than linear and unbiased estimators under mild constraints. One such family is described by the double k -class estimators proposed by Ullah and Ullah (1978). The family of double k -class estimators is characterized by two characterizing scalars and encompasses many estimators as its particular cases, including the Stein rule family of estimators proposed by James and Stein (1961). In the context of linear regression models, the family of double k -class estimators was proposed under the assumption of spherical (homoskedastic) disturbances and assumption in this model under multivariate normal disturbances. Carter et al. (1993) found estimators having better performance than the least squares estimator in a linear regression model have led to the determinant of two characterizing

* * Department of Applied Statistics and econometrics, Institute of Statistical Studies and Research (ISSR), Cairo University, Egypte-mails: a_mohamedam@yahoo.com

scalars for several families of estimators and to the conditions for their dominance over least squares.

Later, Wan and Chaturvedi (2001) extended to the case when disturbances are non-spherical (hetroskedastic) and their variance covariance matrix is also unknown. They proposed the family of feasible generalized double k -class estimators and analyzed the quadratic risk performance of several estimators arising from as a particular cases under the large sample asymptotic approximation theory. The performance of the feasible generalized double k -class estimators under balanced loss function and general Pitman closeness criterion was studied by Chaturvedi and Shalabh (2004). The performance of the feasible generalized double k -class estimators to estimate the regression coefficients in a linear model under LINEX loss function under non-spherical disturbances with unknown variance covariance matrix by Shalabh et al (2012). The review of literature considered performance of the double k -class estimators in linear regression model under multivariate normal-errors.

Using the multivariate t distribution; First, when the data have longer tails than the normal distribution, the multiply imputed data sets using the t distribution allow more valid statistical inferences than those using the normal distribution with some "influential" observations deleted. Second, it is well known that the t distribution is widely used in applied statistics for robust statistical inferences; see Kotz and Nadarajah (2004).

The plan of the paper is as follows. The model and the estimators are described in Section 2. The properties of the double k -class estimators for the coefficients in a linear regression model with spherical disturbances are derived and analyze under multivariate t -errore in Section 3. And special function and some of the expectations required for the proofs in section (3) are presented in the Appendix.

2- The Model and Estimators

Consider the regression model

$$y = X\beta + u \tag{1}$$

Where y is a $T \times 1$ vector of observations on the dependent variable , X is a $T \times P$ matrix of observations on explanatory variables with rank p , β is a $P \times 1$ parameter vector ,and u is a $T \times 1$ disturbance vector .

With the following assumptions

Assumption 1 : $X_{T \times p}$ is nonstochastic matrix and of rank p

Assumption 2 : $u_{T \times 1} \sim T_v[\mu = 0, \sigma^2 I]$

Assumption 3 : $T > P$

Double k-class estimators are given by

$$\tilde{b}_{k_1, k_2} = \left[1 - \frac{k_1 \hat{u}\hat{u}}{\hat{y}y - k_2 \hat{u}\hat{u}} \right] b \quad (2)$$

Where k_1, k_2 are arbitrary scalars which may be stochastic or nonstochastic; $\tilde{b}_{k_1, k_2}$, represents a family of Double k-class estimators and $b = (X'X)^{-1}X'y$ is the OLS estimator of β .

The stein-rule estimator is given by

$$\tilde{b}_{k_1, 1} = \left[1 - \frac{k_1 \hat{u}\hat{u}}{\hat{y}y - \hat{u}\hat{u}} \right] b \quad (3)$$

3- The Main Results

The following section illustrates the main results

(3-1) The Exact Results

The formulas for bias, moment matrix and the risk function of the double k-class estimators under multivariate t-errors for $0 \leq k_2 \leq 1$ are derived.

1) The sampling error of the estimator in eq (2) is given by

$$(\tilde{b}_{k_1, k_2} - \beta) = (b - \beta) - k_1 cb \quad (4)$$

Where $c = \frac{\hat{y}M\hat{y}}{\hat{y}N\hat{y}}$; $M = I - x(x'x)^{-1}x'$, $N = I - k_2 M$

Further, M is an idempotent matrix with rank n and N is a nonnegative definite matrix provided $0 \leq k_2 \leq 1$

2) From assumption two we observe that

$$y \sim T_v[\bar{y}, \sigma^2 I] \quad , \bar{y} = xb \quad (5)$$

Now looking at eq (4), and eq (5) we note that to derive the bias and moment matrix and the risk function of $\tilde{b}_{k_1, k_2}$ the expectations of cb , cbb' and c^2bb' is needed. To obtain the expectations of $\tilde{b}_{k_1, k_2}$ we need the expectations of cb , cbb' and c^2bb' we first derive $E(c)$ and $E(c^2)$ [see Appendix B] by using characteristic function of the multivariate t distribution of Sutradhar (1986), Kotz and Nadarajah (2004) and then use the Ullah and Nagar (1974), Ullah and Ullah (1978), Gradshteyn and Ryzhik (2007), and Sawa (1972) techniques (see section 3) for $E(cb)$, $E(cbb')$ and $E(c^2bb')$, using these expectations, the moments of $\tilde{b}_{k_1, k_2}$ are expressed in terms of $R(\cdot)$ functions [see Appendix A].

For simplicity, we introduce the following notations and functions:

$$A_{g,e,o} = R(k_2^*, r^g, \theta, -2\text{Tr}+e, 2\text{Tr}-o); \quad (g, o = 0, 1, 2, \dots; e = \dots, -1, 0, 1, 2, \dots) \quad (6)$$

Where

$\theta = \frac{\beta' x' x \beta}{\sigma^2}$ is a noncentrality parameter, $0 \leq k_2^* \leq 1$, and the function $R(\cdot)$ is as defined in Appendix A.

Theorem 1: the exact bias of the double k-class estimators of β for $0 \leq k_2 \leq 1$ is given by

$$E(\tilde{b}_{k_1, k_2} - \beta) = -2nk_1\sqrt{\pi}\beta \left[\frac{v}{(v-2)} \frac{4T}{k_2^*} A_{2,-1,0} + \frac{1}{k_2^*} A_{1,0,0} \right] \quad (7)$$

Where $A_{2,-1,0}$ is as given in eq (6) for $g=2$, $e=-1$ and $o=0$; $A_{1,0,0}$ for $g=2$, $e=-1$ and $o=0$

Proof

The expectation on both sides of eq (4) is

$$E(\tilde{b}_{k_1, k_2} - \beta) = -k_1 E(cb) \quad (8)$$

Where $c = \frac{\dot{y}My}{\dot{y}Ny}$

M is idempotent matrix of rank n we can always obtain an orthogonal matrix $\omega_{T \times T}$ such that

$$\omega M \omega = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} = D_1, \quad \omega N \omega = \begin{bmatrix} (1 - k_2)I_n & 0 \\ 0 & I_{T-n} \end{bmatrix} = I - k_2 D_1 = D_2 \quad (9)$$

Where I_n represents the identity matrix of order $n \times n$ and 0 is the zero matrix of proper order. Further, consider D_1 a $T \times T$ diagonal matrix with first n elements unity and $T-n$ elements zero, and D_2 a $T \times T$ diagonal matrix with first n elements $(1 - k_2)$ and $T-n$ elements unity. D_2 will be nonnegative definite for $0 \leq k_2 \leq 1$. Thus, to obtain the expectation of cb on the right of eq (9) we write

$$E(cb) = (\dot{x}x)^{-1} \dot{x} \sigma \omega E(zc) \quad (10)$$

where

$$z = \frac{\dot{\omega} y}{\sigma} \sim T_v[\bar{z}, I], \quad \bar{z} = \frac{\dot{\omega} x \beta}{\sigma}, \quad ; c = \frac{\dot{y}My}{\dot{y}Ny} = \frac{\dot{z}D_1 z}{\dot{z}D_2 z} \quad (11)$$

note that

$$\begin{aligned} E(zc) &= E[(z - \bar{z})c] + \bar{z}E(c) \\ &= \frac{-v}{2\left(\frac{v}{2}-1\right)} \frac{\partial}{\partial \bar{z}} E(c) + \bar{z}E(c) \end{aligned} \quad (12)$$

Where

$$E(c) = \frac{2n\sqrt{\pi}}{k_2^*} R(k_2^*, r, \theta, -2Tr, 2Tr) \quad (13)$$

Next, the fact that

$$\frac{d}{d\bar{z}} E(c) = \frac{-8nT\sqrt{\pi}}{k_2^* \bar{z}} R(k_2^*, r^2, \theta, -2Tr, 2Tr) \quad (14)$$

And using of eq (A.4) see appendix A and eq (12) can be obtained as

$$E(zc) = \frac{-v}{(v-2)} \frac{-8nT\sqrt{\pi}}{k_2^* \bar{z}} R(k_2^*, r^2, \theta, -2Tr, 2Tr) + \bar{z} \frac{2n\sqrt{\pi}}{k_2^*} R(k_2^*, r, \theta, -2Tr, 2Tr) \quad (15)$$

Further, substuting from eq (15) in eq (10) we get.

$$E(cb) = 2n\sqrt{\pi}\beta \left[\frac{v}{v-2} \frac{4T}{k_2^*} R(k_2^*, r^2, \theta, -2Tr-1, 2Tr) + \frac{1}{k_2^*} R(k_2^*, r, \theta, -2Tr, 2Tr) \right] \quad (16)$$

Finally, substitute eq (6), and eq (16) in eq (8) we obtain the result stated in the theorem 1.

Theorem 2: the exact moment matrix of the double k-class estimators of β for $0 \leq k_2 \leq 1$ is given by

$$\begin{aligned}
 E[(\tilde{b}_{k_1, k_2} - \beta)(\tilde{b}_{k_1, k_2} - \beta)'] &= \frac{v}{v-2} \sigma^2 (\dot{X}X)^{-1} - 2k_1 \frac{v}{(v-2)} \sigma^2 (\dot{X}X)^{-1} \left[\frac{v}{(v-4)} \cdot \frac{8nT\sqrt{\pi}}{k_2^*} \right. \\
 &\quad \left. [-4TA_{(3,-1,0)} - A_{(2,-1,0)}] + \frac{16nT\sqrt{\pi}}{k_2^*} A_{(2,0,0)} + \frac{(v+T)}{((v+T)-2)} \frac{2n\sqrt{\pi}}{k_2^*} A_{(1,0,0)} \right] \\
 &\quad - k_1^2 \frac{v}{(v-2)} \sigma^2 (\dot{X}X)^{-1} \left[\frac{v}{(v-4)} \frac{4n\sqrt{\pi}}{k_2^{*2}} [-16nT^2 A_{(4,0,1)} + 4T(3n+2T)A_{(3,0,1)} \right. \\
 &\quad \left. - 2(2n+3T)A_{(2,0,1)} - A_{(1,0,1)}] + \frac{8n\sqrt{\pi}}{k_2^{*2}} [4nTA_{(3,1,1)} - (2n+2T)A_{(2,1,1)} + A_{(1,1,1)}] \right. \\
 &\quad \left. + \left(\frac{(v+T)}{((v+T)-2)} \right) \frac{2n\sqrt{\pi}}{k_2^{*2}} [2nA_{(2,1,1)} - A_{(1,1,1)}] \right] - \frac{2nk_1\beta\beta'\sqrt{\pi}}{k_2^*} \\
 &\quad \left\{ \frac{k_1}{k_2^*} [-2nA_{(2,1,1)} + A_{(1,1,1)}] - \frac{v8T}{v-2} A_{(2,1,0)} \right\} \quad (17)
 \end{aligned}$$

Proof

Using eq (4) we write the moment matrix of the double k-class estimators as

$$\begin{aligned}
 E[(\tilde{b}_{k_1, k_2} - \beta)(\tilde{b}_{k_1, k_2} - \beta)'] &= E(b - \beta)(b - \beta)' + 2k_1 \cdot E[cb]\beta' - 2k_1 \cdot E[cbb'] + k_1^2 \cdot E[c^2bb'] \quad (18)
 \end{aligned}$$

The first term on the right hand of eq (18) is $\frac{v}{v-2} \sigma^2 (\dot{X}X)^{-1}$

Next, considering the second term we note from eq (16) that

$$E(cb)\beta = 2n\sqrt{\pi}\beta\beta' \left[\frac{v}{v-2} \frac{4T}{k_2^*} R_{(k_2^*, r^2, \theta, -2Tr-1, 2Tr)} + \frac{1}{k_2^*} R_{(k_2^*, r, \theta, -2Tr, 2Tr)} \right] \quad (19)$$

Now, considering the third term on the right hand of eq (18), we write

$$E(cbb') = \sigma^2(\dot{x}x)^{-1}\dot{x}p. E[z\dot{z}c]. \dot{p}x(\dot{x}x)^{-1} \quad (20)$$

Where z and c are defined in eq (11), respectively. Using the procedure in eq (12), we can see

$$\begin{aligned} E[z\dot{z}c] &= E \left[[(z - \bar{z})(z - \bar{z})' + (z - \bar{z})\dot{z} + \bar{z}(z - \bar{z})' + \bar{z}\dot{z}]c \right] \\ &= \frac{-v^2}{(v-2)(v-4)} \cdot \frac{\partial^2}{\partial \bar{z} \partial \bar{z}'} E(c) - \frac{2v}{(v-2)} \bar{z} \frac{\partial}{\partial \bar{z}'} E(c) \\ &\quad + \left(\bar{z}\dot{z} + \frac{v(v+T)}{((v+T)-2)(v-2)} \right) E[c] \end{aligned} \quad (21)$$

Where $E(c)$ is as given in eq (13). Further, using of the appendix A, eq (21) can be rewritten as

$$\begin{aligned} E[z\dot{z}c] &= \frac{v^2}{(v-2)(v-4)} \cdot \frac{8nT\sqrt{\pi}}{k_2^*} \left[-4TR(k_2^*, r^3, -2Tr-1, 2Tr) - R(k_2^*, r^2, -2Tr-1, 2Tr) \right] + \\ &\quad \frac{2v}{(v-2)} \cdot \frac{8nT\sqrt{\pi}}{k_2^*} R(k_2^*, r^2, -2Tr, 2Tr) + \bar{z}\dot{z} \frac{2n\sqrt{\pi}}{k_2^*} R(k_2^*, r, -2Tr, 2Tr) + \\ &\quad \frac{v(v+T)}{((v+T)-2)(v-2)} \frac{2n\sqrt{\pi}}{k_2^*} R(k_2^*, r, -2Tr, 2Tr) \end{aligned} \quad (22)$$

and eq (20) can be written as

$$\begin{aligned} E(cbb') &= \sigma^2(\dot{x}x)^{-1} \left[\frac{v^2}{(v-2)(v-4)} \cdot \frac{8nT\sqrt{\pi}}{k_2^*} \left[-4TR(k_2^*, r^3, \theta, -2Tr-1, 2Tr) - R(k_2^*, r^2, \theta, -2Tr-1, 2Tr) \right] \right. \\ &\quad \left. + \frac{2v}{(v-2)} \frac{8nT\sqrt{\pi}}{k_2^*} R(k_2^*, r^2, \theta, -2Tr, 2Tr) + \frac{v(v+T)}{((v+T)-2)(v-2)} \frac{2n\sqrt{\pi}}{k_2^*} R(k_2^*, r, \theta, -2Tr, 2Tr) \right] \\ &\quad + \beta\beta' \frac{2n\sqrt{\pi}}{k_2^*} R(k_2^*, r, \theta, -2Tr, 2Tr) \end{aligned} \quad (23)$$

Similarly, considering the fourth term on the right hand side of eq (18) we have

$$E[z\dot{z}c^2] = \frac{-v^2}{(v-2)(v-4)} \cdot \frac{\partial^2}{\partial \bar{z} \partial \bar{z}'} E(c^2) - \frac{2v}{(v-2)} \bar{z} \frac{\partial}{\partial \bar{z}'} E(c^2) + \left(\bar{z}\dot{z} + \frac{v(v+T)}{((v+T)-2)(v-2)} \right) E[c^2] \quad (24)$$

From appendix B.

$$E(c^2) = \frac{2n\sqrt{\pi}}{k_2^{*2}} \left[-2n R(k_2^*, r^2, -2Tr, 2Tr-1) + R(k_2^*, r, -2Tr, 2Tr-1) \right] \quad (25)$$

The partial derivatives of $R(\cdot)$ given in the appendix A we have eq (24) and hence $E(c^2bb')$ as

$$E(c^2bb') = \sigma^2(\dot{x}x)^{-1} \left\{ \frac{v^2}{(v-2)(v-4)} \frac{4n\sqrt{\pi}}{k_2^{*2}} \left[16nT^2 R(k_2^*, r^4, \theta, -2Tr, 2Tr-1) \right] \right.$$

$$\begin{aligned}
& -4T(3n + 2T) R_{(k_2^*, r^3, \theta, -2Tr, 2Tr-1)} + 2(2n + 3T) R_{(k_2^*, r^2, \theta, -2Tr, 2Tr-1)} \\
& + R_{(k_2^*, r, \theta, -2Tr, 2Tr-1)} \Big] + \frac{2v}{(v-2)} \frac{4n\sqrt{\pi}}{k_2^{*2}} \Big[-4nT R_{(k_2^*, r^3, \theta, -2Tr+1, 2Tr-1)} \\
& + (2n + 2T) R_{(k_2^*, r^2, \theta, -2Tr+1, 2Tr-1)} - R_{(k_2^*, r, \theta, -2Tr+1, 2Tr-1)} \Big] - \left(\frac{v(v+T)}{((v+T)-2)(v-2)} \right) \\
& \frac{2n\sqrt{\pi}}{k_2^{*2}} \Big[2nR_{(k_2^*, r^2, \theta, -2Tr+1, 2Tr-1)} - R_{(k_2^*, r, \theta, -2Tr+1, 2Tr-1)} \Big] \Big\} \\
& - \beta\beta' \frac{2n\sqrt{\pi}}{k_2^{*2}} \Big[2nR_{(k_2^*, r^2, \theta, -2Tr+1, 2Tr-1)} - R_{(k_2^*, r, \theta, -2Tr+1, 2Tr-1)} \Big] \quad (26)
\end{aligned}$$

Finally, substituting from eq (6), eq (19), eq (23), and eq (26) in eq (18), the result stated in theorem 2 follows.

Theorem 3: the exact risk function of the double k-class estimators of β for $0 \leq k_2 \leq 1$ is given by

$$\begin{aligned}
E[(\tilde{b}_{k_1, k_2} - \beta)(\tilde{b}_{k_1, k_2} - \beta)'] &= \frac{v}{v-2} \sigma^2 (\hat{x}x)^{-1} - 2k_1 \frac{v}{(v-2)} \sigma^2 \text{tr}(\hat{x}x)^{-1} \left[\frac{v}{(v-4)} \cdot \frac{8nT\sqrt{\pi}}{k_2^*} \right. \\
& \left. \left[-4TA_{(3,-1,0)} - A_{(2,-1,0)} \right] + \frac{16nT\sqrt{\pi}}{k_2^*} A_{(2,0,0)} + \frac{(v+T)}{((v+T)-2)} \frac{2n\sqrt{\pi}}{k_2^*} A_{(1,0,0)} \right] \\
& - k_1^2 \frac{v}{(v-2)} \sigma^2 \text{tr}(\hat{x}x)^{-1} \left[\frac{v}{(v-4)} \frac{4n\sqrt{\pi}}{k_2^{*2}} \left[-16nT^2 A_{(4,0,1)} + 4T(3n + 2T) A_{(3,0,1)} \right. \right. \\
& \left. \left. - 2(2n + 3T) A_{(2,0,1)} - A_{(1,0,1)} \right] + \frac{8n\sqrt{\pi}}{k_2^{*2}} \left[4nTA_{(3,1,1)} - (2n + 2T) A_{(2,1,1)} \right. \right. \\
& \left. \left. + A_{(1,1,1)} \right] + \left(\frac{v+T}{v+T-2} \right) \frac{2n\sqrt{\pi}}{k_2^{*2}} \left[2nA_{(2,1,1)} - A_{(1,1,1)} \right] \right] - \frac{2nk_1\beta'\beta\sqrt{\pi}}{k_2^*} \\
& \left. \left\{ \frac{k_1}{k_2^*} \left[-2nA_{(2,1,1)} + A_{(1,1,1)} \right] - \frac{v}{v-2} \frac{8T}{k_2^*} A_{(2,1,0)} \right\} \quad (27)
\end{aligned}$$

Where: $\text{tr}(x'x)^{-1} = \sum_{j=1}^p \lambda_j$

λ_j is the jth characteristic root of $(x'x)^{-1}$

Note: all the theorems are based on assumption 1-3. Also, k_1 is considered as an arbitrary constant.

Proof

Theorem 3 follows by applying the trace on both sides of eq (17).

(3-2) large- θ asymptotic expansion

the following section is an asymptotic expansion of the bias, moment matrix and the risk function of the double k-class estimators for $0 \leq k_2 \leq 1$ under multivariate t-errors in terms of the inverse of θ . These results help in analyzing the complicated expressions of the exact bias, exact moment matrix and exact risk function given in eq (7), eq (17) and eq (27), respectively.

We note that sufficiently large θ which according to $\theta = \frac{\beta' X' X \beta}{\sigma^2}$ means relatively small σ . Thus, we consider the terms of order $\frac{1}{\theta^{2T-e}}$, $\frac{1}{\theta^{4T-e}}$ as the terms of order $\sigma^{2(2T-e)}$ and $\sigma^{2(4T-e)}$, respectively.

Theorem (4): the asymptotic expansion of the bias of double k-class estimators of β under multivariate t-errors in eq (7) up to order $\frac{1}{\theta^{4T+1}}$ is given by

$$E(\tilde{b}_{k_1, k_2} - \beta) = -2nk_1\sqrt{\pi} \left[\frac{(2T-1)}{(\theta)^{(2T)+1}} \left(\frac{v}{(v-2)} \frac{4T}{k_2^*} + \frac{\theta}{k_2^*} \right) \zeta_1 + \frac{(4T-1)}{(\theta)^{(4T)+1}} \left(\frac{v}{(v-2)} \frac{T}{k_2^*} + \frac{\theta}{8k_2^*} \right) \zeta_2 \right] \beta \quad (28)$$

Where:

$$\zeta_1 = (-1)^{T+1} \frac{3v^2 k_2^{*2n}}{8} \cdot \frac{\Gamma(2T-1)}{\left(\frac{v-4}{2}\right)_2}, \quad \zeta_2 = \frac{v^4 k_2^{*4n} \left(\frac{5}{2}\right)_2 \Gamma(4T-1)}{\left(\frac{v-8}{2}\right)_4}$$

Proof

Using eq (6) and substituting (A.7) in eq (7), the results in eq (28) can easily be established.

Theorem (5): The asymptotic expansion of the moment matrix of the double k-class estimators of β under multivariate t-errors in eq (17) up to order $\frac{1}{\theta^{4T+1}}$ is given by

$$\begin{aligned}
E[(\tilde{b}_{k_1, k_2} - \beta)(\tilde{b}_{k_1, k_2} - \beta)'] &= \frac{v}{v-2} \sigma^2 (\dot{X}X)^{-1} + \frac{2k_1}{(\theta)^{(2T)+1}} \sigma^2 (\dot{X}X)^{-1} \eta_1 (2T-1) \zeta_1 + \\
&\quad \frac{2k_1}{(\theta)^{(4T)+1}} \sigma^2 (\dot{X}X)^{-1} \eta_2 (4T-1) \zeta_2 - \frac{k_1}{(\theta)^{(2T)+1}} \beta' X' X \beta (\dot{X}X)^{-1} \eta_3 \zeta_1 - \\
&\quad \frac{k_1}{(\theta)^{(4T)}} \sigma^2 (\dot{X}X)^{-1} \eta_4 \zeta_2 - \frac{k_1^2}{(\theta)^{2T-1}} \sigma^2 (\dot{X}X)^{-1} \eta_5 \zeta_1 - \\
&\quad \frac{k_1^2}{(\theta)^{4T-1}} \sigma^2 (\dot{X}X)^{-1} \eta_6 \zeta_2 + \frac{2nk_1 \sigma^2 (\dot{X}X)^{-1} \sqrt{\pi}}{(\theta)^{(2T)-2} k_2^*} \left(2n \frac{k_1}{k_2^*} - \frac{k_1}{k_2^*} + \frac{8vT}{(v-2)} \right) \zeta_1 + \\
&\quad \frac{2nk_1 \sigma^2 (\dot{X}X)^{-1} \sqrt{\pi}}{(\theta)^{(4T)-2} k_2^*} \left(n \frac{k_1}{2k_2^*} - \frac{k_1}{8k_2^*} + \frac{2vT}{(v-2)} \right) \zeta_2
\end{aligned} \tag{29}$$

Where:

$$\begin{aligned}
\eta_1 &= \frac{8v^2 nT \sqrt{\pi}}{(v-2)(v-4)k_2^*} (4T+1), \quad \eta_2 = \frac{2v^2 nT \sqrt{\pi}}{(v-2)(v-4)k_2^*} (8T+1) \\
\eta_3 &= \frac{4vn \sqrt{\pi}}{(v-2)k_2^*} \left\{ \left[8T + \frac{(v+T)}{((v+T)-2)} \right] (2T-1) + k_1 \frac{v}{(v-4)} [-16nT^2 + 4T(3n+2T) - 2(2n+3T) - 1] \right\} \\
\eta_4 &= \frac{vn \sqrt{\pi}}{2(v-2)k_2^*} \left\{ \left[16T + \frac{(v+T)}{((v+T)-2)} \right] (4T-1) + k_1 \frac{v}{(v-4)} [-128nT^2 + 16T(3n+2T) - 4(2n+3T) - 1] \right\} \\
\eta_5 &= \frac{8vn \sqrt{\pi}}{(v-2)k_2^{*2}} \left\{ 4nT - (2n+2T) + 1 + \left(\frac{(v+T)}{((v+T)-2)} \right) \frac{2n \sqrt{\pi}}{k_2^{*2}} (2n-1) \right\} \\
\eta_6 &= \frac{8vn \sqrt{\pi}}{(v-2)k_2^{*2}} \left\{ 2nT - \frac{(n+T)}{2} + \frac{1}{8} + \left(\frac{(v+T)}{((v+T)-2)} \right) \frac{2n \sqrt{\pi}}{k_2^{*2}} \left(\frac{n}{2} - \frac{1}{8} \right) \right\}
\end{aligned}$$

Proof

Using eq (6) and substituting (A.7) in eq (17), the results in eq (29) can easily be established.

Theorem (6): The asymptotic expansion of risk function of the double k-class estimators of β under multivariate t-errors in eq (27) up to order $\frac{1}{\theta^{4T+1}}$ is given by

$$\begin{aligned}
E[(\tilde{b}_{k_1, k_2} - \beta)'(\tilde{b}_{k_1, k_2} - \beta)] &= \frac{v}{v-2} \sigma^2 \text{tr}(\dot{x}x)^{-1} + \frac{2k_1}{(\theta)^{(2T)+1}} \sigma^2 \text{tr}(\dot{x}x)^{-1} \eta_1(2T-1)\zeta_1 + \\
&\quad \frac{2k_1}{(\theta)^{(4T)+1}} \sigma^2 \text{tr}(\dot{x}x)^{-1} \eta_2(4T-1)\zeta_2 - \frac{k_1}{(\theta)^{(2T)}} \sigma^2 \text{tr}(\dot{x}x)^{-1} \eta_3\zeta_1 - \\
&\quad \frac{k_1}{(\theta)^{(4T)}} \sigma^2 \text{tr}(\dot{x}x)^{-1} \eta_4\zeta_2 - \frac{k_1^2}{(\theta)^{(2T-1)}} \sigma^2 \text{tr}(\dot{x}x)^{-1} \eta_5\zeta_1 - \\
&\quad \frac{k_1^2}{(\theta)^{(4T-1)}} \sigma^2 \text{tr}(\dot{x}x)^{-1} \eta_6\zeta_2 + \frac{2nk_1\sigma^2 \text{tr}(\dot{x}x)^{-1} \sqrt{\pi}}{(\theta)^{(2T)-2} k_2^*} \left(2n \frac{k_1}{k_2^*} - \frac{k_1}{k_2^*} + \frac{8vT}{(v-2)} \right) \zeta_1 + \\
&\quad \frac{2nk_1\sigma^2 \text{tr}(\dot{x}x)^{-1} \sqrt{\pi}}{(\theta)^{(4T)-2} k_2^*} \left(n \frac{k_1}{2k_2^*} - \frac{k_1}{8k_2^*} + \frac{2vT}{(v-2)} \right) \zeta_2
\end{aligned} \quad (30)$$

Proof

Using eq (6) and substituting (A.7) in eq (27), the results in eq (30) can easily be established

Corollary: the double k-class estimator of β in eq (2) dominates ordinary least squares estimator of β is $b = (\dot{x}x)^{-1}\dot{x}y$ in large θ asymptotics up to the order $\frac{1}{(\theta)^{(2T)-2}}$ when number of observations is even, in the sense that

$$\lim_{\theta \rightarrow \infty} (\theta)^{(2T)-2} \left\{ E[(\tilde{b}_{k_1, k_2} - \beta)'(\tilde{b}_{k_1, k_2} - \beta)] - E[(b - \beta)'(b - \beta)] \right\} < 0 \quad (31)$$

For $0 < k_1 \leq \frac{2}{n+2}(d-2)$, $d = \sum_{j=1}^p \frac{\lambda_j}{\lambda_L} > 2$; λ_j eigen values and λ_L largest of eigen values

And for any k_2 in $0 \leq k_2 \leq 1$

(3-3) Conclusion

Now looking the results of sections (3-1) and (3-2) we observe the following.

1- The following results regarding the exact and asymptotics bias of the double k-class estimator in eq (7) and eq (28) are true for $0 \leq k_2 \leq 1$ and $0 < k_1 \leq \frac{2}{n+2}(d-2)$

(a) $\tilde{b}_{k_1, k_2}$ is unbiased only for $k_1 = 0$ and in that case it is the ordinary least squares estimator

(b) The exact bias of an element of $\tilde{b}_{k_1, k_2}$ is a decreasing function of both k_1 and k_2 .

2- The exact risk function of ordinary least squares estimator in the nonorthogonal regression context, exists for $p \geq 3$; it is given by $E[(b - \beta)'(b - \beta)] = \frac{v}{v-2} \sigma^2 \text{tr}(\hat{X}X)^{-1}$.

$\tilde{b}_{k_1, k_2}$ estimator dominates the ordinary least squares estimator in eq (31) explains the difference between risk under quadratic loss function under multivariate t-error in large θ asymptotics up to the order $\frac{1}{(\theta)^{(2T)-2}}$ when number of observations is even.

in the eighth term on the right hand side of eq (30) will be negative up to the order $\frac{1}{(\theta)^{(2T)-2}}$ when number of observations is even.

Appendix

A. special function

(1) The Gamma functions Γ have the following power series representation (see Andrews)

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \quad x > 0$$

(2) The Beta function,

$$B(x, y) = \int_0^\infty \frac{t^{x-1}}{(1+t)^{x+y}} dt$$

(3) The pochhammer symbol,

$$(r)_0 = 1$$

$$(r)_n = r(r+1) \dots (r+n-1) \quad n = 1, 2, \dots$$

$$(r)_n = \frac{\Gamma(r+n)}{\Gamma(r)} \quad n = 0, 1, 2, \dots$$

(4) Similarly, the power series presentation of the function R,

$$R(k_2^*, r^g, \theta, -2Tr+e, 2Tr-o) = \sum_{r=0}^{\infty} (-1)^{Tr+r} \frac{v^{2r} r^g k_2^{*2nr} (\theta)^{-(2Tr)+e}}{2^{2r} r!} \cdot \frac{\left(r+\frac{1}{2}\right)_r \Gamma(2Tr-o)}{\left(\frac{v}{2}-2r\right)_r \left(\frac{v}{2}-r\right)_r} \quad (A.1)$$

Where $\theta = \hat{Z}\bar{Z} > 0$, $0 \leq k_2^* \leq 1$, $g, o = 0, 1, 2, \dots$; $e = \dots, -1, 0, 1, 2, \dots$

For $k_2^* = 1$, (A.1) reduce to

$$R_{(1,r^g, \theta, -2Tr+e, 2Tr-o)} = \sum_{r=0}^{\infty} (-1)^{Tr+r} \frac{v^{2r} r^g (\theta)^{-(2Tr)+e}}{2^{2r} r!} \cdot \frac{\left(\frac{r+1}{2}\right)_r \Gamma(2Tr-o)}{\left(\frac{v-2r}{2}\right)_r \left(\frac{v-r}{2}\right)_r} \quad (A.2)$$

The partial derivatives of R with respect to z can be written as

$$\begin{aligned} \frac{d}{dz} R_{(k_2^*, r^g, \theta, -2Tr+e, 2Tr-o)} &= 2\bar{z} \left[-2T R_{(k_2^*, r^{g+1}, \theta, -2Tr+e-1, 2Tr-o)} \right. \\ &\quad \left. + e R_{(k_2^*, r^g, \theta, -2Tr+e-1, 2Tr-o)} \right] \end{aligned} \quad (A.3)$$

$$\frac{\partial^2}{\partial \bar{z} \partial z} R_{(k_2^*, r^g, \theta, -2Tr+e, 2Tr-o)} = 16T^2 \bar{z} R_{(k_2^*, r^{g+2}, \theta, -2Tr+e-2, 2Tr-o)} - (16Te - 4T)$$

$$R_{(k_2^*, r^{g+1}, \theta, -2Tr+e-1, 2Tr-o)} + e(4e - 2) R_{(k_2^*, r^g, \theta, -2Tr+e-1, 2Tr-o)} \quad (A.4)$$

and partial derivatives of R with respect to k_2^* can be written as

$$\frac{\partial}{\partial k_2^*} R_{(k_2^*, r^g, \theta, -2Tr+e, 2Tr-o)} = \frac{2n}{k_2^*} R_{(k_2^*, r^{g+1}, \theta, -2Tr+e, 2Tr-o)} \quad (A.5)$$

$$\begin{aligned} \frac{\partial^2}{\partial k_2^{*2}} R_{(k_2^*, r^g, \theta, -2Tr+e, 2Tr-o)} &= \frac{4n^2}{k_2^{*2}} R_{(k_2^*, r^{g+2}, \theta, -2Tr+e, 2Tr-o)} \\ &\quad - \frac{2n}{k_2^{*2}} R_{(k_2^*, r^{g+1}, \theta, -2Tr+e, 2Tr-o)} \end{aligned} \quad (A.6)$$

Finally, for large θ , the asymptotic expansion of the function R, in (A.1) up to order $\frac{1}{\theta^{8T-e}}$ is given by

$$\begin{aligned} R_{(k_2^*, r^g, \theta, -2Tr+e, 2Tr-o)} &= (-1)^{T+1} \frac{v^2 k_2^{*2n}}{4 (\theta)^{(2T)-e}} \cdot \frac{\frac{3}{2} \Gamma(2T-o)}{\left(\frac{v-4}{2}\right)_2} + \frac{v^4 2^g k_2^{*4n}}{16 (\theta)^{(4T)-e}} \cdot \frac{\left(\frac{5}{2}\right)_2 \Gamma(4T-o)}{\left(\frac{v-8}{2}\right)_4} + \\ &\quad (-1)^{T+1} \frac{v^6 3^g k_2^{*6n}}{384 (\theta)^{(6T)-e}} \cdot \frac{\left(\frac{7}{2}\right)_3 \Gamma(6T-o)}{\left(\frac{v-12}{2}\right)_6} + o(\theta^{-8T+e}) \end{aligned} \quad (A.7)$$

Where : $2T \geq e$, T is number of observations, v is degrees of freedom, $k_2^* = 1 - k_2$, $0 \leq k_2 \leq 1$,
 $(g, o = 0, 1, 2, \dots; e = \dots, -1, 0, 1, 2, \dots)$, $\theta = \frac{\beta' x' x \beta}{\sigma^2}$

B. Evaluation of expectations required in subsection (3-1)

Let w be a $T \times 1$ Multivariate t distribution random vector such that

$$E(w) = 0, \quad E(w - \bar{w})(w - \bar{w})' = \frac{v}{v-2} I \quad (B.1)$$

$$; c = \frac{\dot{y}My}{\dot{y}Ny} = \frac{\dot{z}D_1z}{\dot{z}D_2z} = \frac{\dot{w}D_1w}{\dot{w}D_2w} \quad (B.2)$$

the joint characteriestic function (ch) of the quadratic forms $\dot{w}D_2w$ and $\dot{w}D_1w$ can be written as

$$\Phi_{c_w}(t_1, t_2) = E[e^{it_1 \dot{w}D_2w + it_2 \dot{w}D_1w}] \quad (B.3)$$

Note that

$$Q = t_1 - (k_2 t_1 - t_2) D_1, \quad Q\bar{z} = t_1 \bar{z} \quad (B.4)$$

We can obtain eq (B.3) as

$$\Phi_{c_w}(t_1, t_2) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \cdot \sum_{r=0}^{\infty} (-1)^r \frac{v^{2r} (k_2^* t_1 + t_2)^{2nr} t_1^{(T-n)2r}}{2r!} \cdot B\left(2r + \frac{1}{2}, \frac{v+1}{2} - 2r - \frac{1}{2}\right) \quad (B.5)$$

Where $k_2^* = 1 - k_2$.

The characteriestic function (ch) of the quadratic forms $\dot{z}D_2z$ and $\dot{z}D_1z$ can be written as

$$\Phi_{c_z}(t_1, t_2) = e^{iQ\dot{z}\bar{z}} \Phi_{c_w}(t_1, t_2) = e^{it_1 \dot{z}\bar{z}} \Phi_{c_w}(t_1, t_2) \quad (B.6)$$

The the following derivatives of (B.5) can then be obtained as follows.

$$\frac{-i \partial \Phi_{c_z}(t_1, t_2)}{\partial t_2} \Big|_{t_2=0} = \frac{-i \Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{\partial}{\partial k_2^*} \sum_{r=1}^{\infty} (-1)^r \frac{v^{2r} k_2^{*2nr} e^{it_1 \dot{z}\bar{z}} t_1^{2Tr-1}}{2r!} B\left(2r + \frac{1}{2}, \frac{v+1}{2} - 2r - \frac{1}{2}\right) \quad (B.7)$$

$$\frac{-\partial^2 \Phi_{c_z}(t_1, t_2)}{\partial t_2^2} \Big|_{t_2=0} = \frac{-\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{\partial^2}{\partial k_2^2} \sum_{r=1}^{\infty} (-1)^r \frac{v^{2r} k_2^{*2nr} e^{it_1 \bar{z} z} t_1^{2Tr-2}}{2r!} B\left(2r + \frac{1}{2}, \frac{v+1}{2} - 2r - \frac{1}{2}\right) \quad (B.8)$$

Finally using (B.7), (B.8), and (A.1) we obtain the following expectations for $0 \leq k_2^* \leq 1$.

$$\begin{aligned} E\left(\frac{\dot{z}D_1 z}{\dot{z}D_2 z}\right) &= i \int_0^\infty \frac{-i \partial \Phi_{c_z}(t_1, t_2)}{\partial t_2} \Big|_{t_2=0} dt_1 \\ &= \frac{2n\sqrt{\pi}}{k_2^*} R_{(k_2^*, r, \theta, -2Tr, 2Tr)} \end{aligned} \quad (B.9)$$

$$\begin{aligned} E\left(\frac{\dot{z}D_1 z}{\dot{z}D_2 z}\right)^2 &= i \int_0^\infty \frac{\partial^2 \Phi_{c_z}(t_1, t_2)}{\partial t_2^2} \Big|_{t_2=0} dt_1 \\ &= \frac{2n\sqrt{\pi}}{k_2^{*2}} \bar{z} z \left[-2nR_{(k_2^*, r^2, \theta, -2Tr, 2Tr-1)} + R_{(k_2^*, r, \theta, -2Tr, 2Tr-1)} \right] \end{aligned} \quad (B.10)$$

References

- 1- Andrews, L, C. (1992). *Special functions of mathematics for engineers. Second edition*, McGraw-Hill, Inc. New York.
- 2- Carter, R.A.L., Srivastava, V.K., Chaturvedi, A. (1993). Selecting a double k-class estimator for regression coefficients. *Statistics & Probability Letters* 18:363-371.
- 3- Chaturvedi, A., Shalabh. (2004). Risk and Pitman closeness properties of feasible generalized double k-class estimators in linear regression models with non-spherical disturbances under balanced loss function. *Journal of Multivariate Analysis* 90:229-256.
- 4- Gradshteyn, I, S., Ryzhik, I, M. (2007). *Table of Integrals, Series, and Products. Seventh Edition*, Elsevier, USA.
- 5- James, W., Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley University of California Press:361-379.
- 6- Kotz, S., Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge University Press, UK.

- 7- Shalabh., Garg, G., Heumann, C. (2012). Performance of double k-class estimators for coefficients in linear regression models with non-spherical disturbances under asymmetric losses. *Journal of Multivariate Analysis* 112:35-47.
- 8- Sawa, T. (1972). Finite-sample properties of the k-class estimators. *Econometrica* 40:715-721.
- 9- Sutradhar, B, C. (1986). On the characteristic function of multivariate Student t-distribution. *The Canadian Journal of Statistics* 14:329-337.
- 10- Ullah, A., Nagar, A, L. (1974). The exact mean of the two stage least squares estimator of the structural parameters in an equation having three endogenous variables. *Econometrica* 42:749-758.
- 11- Ullah, A., Ullah, S. (1978). Double k-class estimators of coefficients in linear regression. *Econometrica* 46:705-722.
- 12- Wan, A.T.K., Chaturvedi, A. (2001). Double k-class estimator in regression models with nonspherical disturbance. *Journal of Multivariate Analysis* 79:226-250.

Robust Estimation of Parameters in logistic regression models

Hazem Refaat Ahmed*

Abstract:

Classical Maximum likelihood method is the most commonly method for estimation of parameters in logistic regression model due to its efficiency. However, performance of maximum likelihood estimator(MLE) is undermined by existence of outliers and unusual observations Also, these outliers may have bad effects on the estimation of parameters of logistic regression models and affect predictions and classification error rate of these models. The current research paper tackle the problem of outliers and robust estimation in logistic regression through deriving the loss function of four existing robust logistic estimators.

Key words: Leverage points- Likelihood- Logistic Regression- Loss Function- Outliers-Robust estimates .

Introduction:

Logistic regression models are widely used in real life applications as classification models. These models describes the relation between a set of explanatory variables, which may be quantitative or categorical and a categorical response variable which may be dichotomous or polytomous . The usual model is the binary logistic model for which the response variable takes on only two values(for example: 1 for success and zero for failure)

Given an attribute x of an explanatory variable X and a classification variable Y . Then the simple binary logistic model is represented as the conditional probability for Y given x takes the following form:

* Teaching Assistant, Department of Applied Statistic and Econometrics, Institute of Statistical Studies and Research, Cairo University, Egypt.
E-Mail:hazemrefaatissr103@gmail.com

$$pr(Y_i = 1 | X_i = x) = \pi(x_i) = \frac{1}{e^{-(\alpha + \beta^T x_i)}} \quad , \quad i = 1, 2, \dots, n$$

Where : α and $B^T = (B_1, B_2, \dots, B_p)^T$ are the parameters of the model

Another representation of the above model is:

$pr(Y_i = 1 | X_i = x) = \pi(x_i) = F(x_i \beta)$, where F is the logistic distribution function

Let $\gamma = (\alpha, \beta^T)^T$ and $Z_i = (1, x_i^T)^T$

And the parameter estimates are those values that maximize this likelihood and they found by iterative methods as proposed by Cox(1966) .

obtaining the maximum likelihood estimates of logistic regression model is equivalent to finding the parameter estimates that maximize the log -likelihood function by solving the following problem:

$$\arg \max_{\gamma} \sum_{i=1}^n L(z_i^T \gamma, y_i)$$

where: L is the log-likelihood function given by:

$$L(z_i^T \gamma, y_i) = y_i \log F(Z_i^T \gamma) - (1 - y_i) \log(1 - F(Z_i^T \gamma)) \quad , i = 1, 2, \dots, n$$

The previous problem can be formulated $\sum_{i=1}^n |y_i - F(Z_i^T \gamma)| z_i = 0$ as:

$$\arg \min_{\gamma} \sum_{i=1}^n d(z_i^T \gamma, y_i)$$

Where

$d(.)$ is the negative log-likelihood function and is called deviance function. More formally , in literature , it is referred to as loss function.

Differentiating the loss function with respect to γ yields the likelihood score equation:

$$\sum_{i=1}^n |y_i - F(Z_i^T \gamma)| z_i = 0$$

solving this equation iteratively by either Newton -Raphson or scoring method results in maximum likelihood estimates.

The likelihood scoring equation is affected by outliers or extreme values and hence classical MLE is said to be not robust. Therefore, several robust alternatives of MLE have been created to remedy this problem.

This paper is organized as follows:

First, illustration of different types of outliers and their effect are presented in logistic regression

Second, we derive the formula of loss function for some existing robust logistic regression estimators, along with advantages and disadvantages for each estimator

1-Outliers in Logistic Regression: (Sanizah Ahmad et al.2010)

It is important to distinguish between the different cases of outlying observations in logistic regression. In a binary logistic model, outliers can occur in the Y-space, the X-space or in both spaces and 3 types of outliers are distinguished as follows:

1-Y- Outlier(Vertical Outlier):

Since in logistic regression the binary data of all y 's are either 0 or 1 For binary data, all the y 's are 0 or 1, hence an error in the y direction can only occur as a transposition $0 \rightarrow 1$ or $1 \rightarrow 0$. This type of outliers is referred to y -outlier or vertical outlier outlier.

2- Leverage Outlier:

An observation which is extreme in the design space X is called a leverage outlier or leverage point: a leverage point can be considered good or bad.

- good leverage point:

good leverage point occurs when $Y=1$ with a large value of $P(Y=1|x_i)$ or when $Y=0$ with a large value of $P(Y=0|x_i)$. good leverage points are not so harmful to logistic regression because they usually corresponds to small value of $|Y - \hat{\pi}|$.

- bad leverage point:

bad leverage point occurs when $Y=1$ with a small value of $P(Y=1|x_i)$ or when $Y=0$ with a large value of $P(Y=1|x_i)$. bad leverage points corresponds to a large value of $|Y - \hat{\pi}|$ and they are very important in logistic regression because they represent the misclassified observations.

A vertical outlier is an observation whose residual with respect to the fitted value, given by $y - F(z^T \gamma_0)$ is large. Hence, the influence of vertical outliers on the ML estimator is bounded since we always have $|y - F(z^T \gamma_0)| \leq 1$.

A good leverage point corresponds to an outlying z value, but with a small value for $y - F(z^T \gamma_0)$. When z increases, the value of the residual will typically tend much faster to zero than z to infinity, resulting in a small influence. Hence, it can be said that that good leverage points are not so harm for the ML-estimator.

A bad leverage point will always induce a large value of the difference between observed and fitted value. so, in general leverage points can be good or bad, depending on whether they have a small or large residual with respect to the imposed fit.

An important measure which we propose in our simulation study to compare between different logistic estimators is the influence function

Influence function (Hampel,1974) on a statistic T at a point x drawn from distribution F is simply defined as:

$$IF(T, x, F) = \lim_{\varepsilon \rightarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}, \text{ Where } X \sim F_\varepsilon \text{ if:}$$

$$X \sim \begin{cases} F & \text{with probability } 1 - \varepsilon \\ x & \text{with probability } \varepsilon \end{cases}$$

Where F_ε is a mixture of F and a point x .

Properties of the influence function

The influence function behaves like the score function (the first derivative of the log-likelihood function) in parameter estimation.. if we denote $IF(T, x, F)$ as a functional by $L_{F(x)}$, then, it satisfy the following nice properties:

$$\int L_{F(x)} dF(x) = 0$$

$$2-Var(T(\hat{F}_n)) \approx \int L_{F(x)}^2 dF(x) / n$$

So the second property states that there is a heuristic relation between the influence function and the asymptotic variance of an estimate. This relation can be used to quickly obtain an expression for the asymptotic variance and hence standard error of

the estimate .The maximum value of the influence function is called gross error sensitivity .Formally, the gross error sensitivity of an estimator T as a functional on F is defined as:

$$\xi^* = \xi^*(T, F) = \sup_{\{X : IF(x, T, F) \text{ exists}\}} |IF(x, T, F)|$$

2-Robust Estimators in Logistic Regression:

The deviance function(loss function) d of logistic regression is always positive and it is continuous and differentiable.

So it can, generally, be replaced by another general function ρ and the estimator of interest can therefore be defined by :

$$\hat{\gamma} = \arg \min_{\gamma} \sum_{i=1}^n \rho(d(z_i^T \gamma, y_i)) = \hat{\gamma}_{ML} = \arg \min_{\gamma} \sum_{i=1}^n \rho(d(s_i, y_i)) \quad (2.1)$$

where: the score value $s_i = z_i^T \gamma$ is a linear combination of the vector parameter γ and where : $\rho(.)$ is a positive, continuous and almost everywhere differentiable function and should satisfy $\rho(s;0) = \rho(-s;1)$.

let $\varphi(s) = \rho(s;0)$ gives for an observation for null y , the impact of a particular score s on the value of the objective function in (2.1).So, when $s \rightarrow \infty$, then, implying that a large negative score is not contributing to the objective function..

In the following , different robust estimators in logistic regression are presented:

(2.1) Pregibon estimator:

Pregibon(1982) suggested an estimator given by:

$$\hat{\gamma}_n = \arg \min h(x_i) \rho\left(\frac{d_i}{h(x_i)}\right), \text{ where } \rho \text{ is a loss function and } h(x_i) \text{ is a factor}$$

Which control leverage of each observation and d_i is the negative log-likelihood or deviance function of logistic regression.

The particular loss function Pregibon(1982) suggested to use is the Huber loss function(Huber,1964) which is given by:

$$\rho(u) = \begin{cases} u & u \leq k \\ 2(uk)^{1/2} - k & \text{o.w} \end{cases} \quad (2.2)$$

Where k is a pre-specified tuning constant that controls robustness of the function.

We can define this estimator, in terms of log-likelihood as follows:

$$\text{Let } f(x_i) = \alpha + B^T x_i, \quad i = 1, 2, \dots, n$$

Then:

$$\sum_i d_i = -\sum_i l(y_i, f(x_i)) \quad , \quad y_i = 0, 1 \quad , \quad i = 1, 2, \dots, n$$

$$= -\ell(0, f(x_i)) - \ell(1, f(x_i))$$

$$= -\log\left(\frac{1}{1+e^{f(x_i)}}\right) - \log\left(\frac{1}{1+e^{-f(x_i)}}\right)$$

Now, We can write $l(y_i, f(x_i))$ as follows:

$$l(y_i, f(x_i)) \quad , \quad y_i = 0, 1 \quad i = 1, 2, \dots, n$$

$$= l(1, f(x_i)) + l(0, f(x_i)) = l(1, f(x_i)) = -\log\left(\frac{1}{1+e^{-f(x_i)}}\right) = \log(1+e^{-f(x_i)})$$

So, instead of minimizing $\rho(d_i) = \sum_i d_i$, we want to:

$$\text{Min} - \sum_{i=1}^n \ell(y_i, f_i)$$

where:

$$\ell(y_i, f(x_i)) = \log(1+e^{-f(x_i)}) \quad i = 1, 2, \dots, n$$

Now, the Pregibon function is defined as:

$$\ell^{\text{Pregibon}}(u) = \rho(\ell(u)) = \begin{cases} \log(1+e^{-u}) & \log(1+e^{-u}) \leq H \\ -2(H \log(1+e^{-u}))^{1/2} - H & \text{otherwise} \end{cases}$$

Note:

$$\log(1+e^{-u}) \leq H \rightarrow e^{-u} \leq e^H - 1 \rightarrow u \geq -\log(e^H - 1)$$

Pregibon function can finally be written as:

$$\ell^{\text{Pregibon}}(u) = \begin{cases} \log(1+e^{-u}) & u \geq -\log(e^H - 1) \\ 0 & \text{otherwise} \end{cases} \quad (2.3)$$

The estimates obtained by (3.3) have approximately 95% asymptotic relative efficiency when $H = (1.345)^2$.

So, in view of loss function, we get the following problem:

$$\text{MIN} \sum_{i=1}^n \ell^{\text{Pregibon}}(y_i f(x_i)) \quad , \text{ where } \ell^{\text{Pregibon}}(y_i f(x_i)) \text{ is as defined in (2.3)}$$

Advantages of Pregibon function :

Pregibon function depends in its computation on Huber loss function which is a robust function .Using a tuning constant equal $(1.345)^2$, we obtain a high efficiency that reaches 95%.

Disadvantages of Pregibon function :

Pregibon function grows as $u \rightarrow \infty$ but less rapidly than the original function $l(u)$ of the logistic regression. The resulting coefficient estimates become less sensitive to extreme observations. However, the value of $\ell^{\text{pregibon}}(u)$ remains to be unbounded and hence the impact of outliers can still be large.

(2.2) Copas estimator:

Copas(1988) modeled contamination of class labels in the training data. Specifically, he introduced a misclassification model in which each response is misclassified with probability Π . As a result, the response y could be 1 with probability $P^*(x)$, where:

$$\begin{aligned} P^*(x) &= (1-\Pi)\pi(x) + \Pi(1-\pi(x)) \\ &= \pi(x) - \Pi\pi(x) + \Pi - \Pi\pi(x) \\ &= F(x^T \beta_{MC}) + \Pi(1-2F(x^T \beta_{MC})) = G(x^T \beta_{MC}, \Pi) \end{aligned}$$

Where B_{MC} denote the parameter under the misclassification model. Copas compared the two models:

$$1- \pi(x) = F(x^T B) \quad 2- P^*(x) = G(x^T B_{MC}, \Pi)$$

and concluded that the latter model is better because it has the same robustness properties of the former but it had less small to moderate sample size bias.

In terms of the log-likelihood function of logistic regression with $P^*(x)$ instead of $\pi(x)$, we have:

$$LOGL = \ell = \sum_{i=1}^n \{y_i \log[\pi(x_i) + \Pi(1 - 2\pi(x_i))] + (1 - y_i) \log[1 - \pi(x_i) - \Pi(1 - 2\pi(x_i))]\} \quad (2.4)$$

Let $S_1 = \log(\pi(x_i) + \Pi(1 - 2\pi(x_i)))$ denote the first term of the sum in in (2.4), then :

$$\begin{aligned} S_1 &= y_i \log(\pi(x_i) + \Pi(1 - 2\pi(x_i))) \\ &= y_i \log\left(\frac{1}{1 + e^{-f(x_i)}} + \Pi - \frac{2\Pi}{1 + e^{-f(x_i)}}\right) = y_i \log\left(\frac{1 + \Pi(e^{-f(x_i)} - 1)}{1 + e^{-f(x_i)}}\right) \end{aligned} \quad (2.5)$$

Let $S_2 = (1 - y_i) \log(1 - \pi(x_i) - \Pi(1 - 2\pi(x_i)))$ denote the second term in (2.4), then :

$$\begin{aligned} S_2 &= (1 - y_i) \log(1 - \pi(x_i) - \Pi(1 - 2\pi(x_i))) \\ &= (1 - y_i) \log(1 - \pi(x_i) - \Pi + 2\Pi\pi(x_i)) \\ &= (1 - y_i) \log\left(1 - \frac{1}{1 + e^{-f(x_i)}} - \Pi + 2\Pi \frac{1}{1 + e^{-f(x_i)}}\right) \\ &= (1 - y_i) \log\left[\frac{1 + e^{-f(x_i)} - 1 - \Pi(1 + e^{-f(x_i)}) + 2\Pi}{1 + e^{-f(x_i)}}\right] \\ &= (1 - y_i) \log\left[\frac{1 + e^{-f(x_i)} - 1 - \Pi(e^{-f(x_i)} - 1)}{1 + e^{-f(x_i)}}\right] \\ &= (1 - y_i) \log\left[\frac{e^{-f(x_i)} + \Pi(1 - e^{-f(x_i)})}{1 + e^{-f(x_i)}}\right] = (1 - y_i) \log\left[\frac{e^{-f(x_i)} + \Pi(e^{f(x_i)} e^{-f(x_i)} - e^{-f(x_i)})}{1 + e^{-f(x_i)}}\right] \end{aligned}$$

$$\begin{aligned}
 &= (1-y_i) \log \left[\frac{e^{-f(x_i)}(1+\Pi e^{-f(x_i)}-\Pi)}{1+e^{-f(x_i)}} \right] = (1-y_i) \log \left[\frac{e^{-f(x_i)}(1+\Pi e^{-f(x_i)}-\Pi)}{e^{-f(x_i)}e^{-f(x_i)}+e^{-f(x_i)}} \right] = \\
 &(1-y_i) \log \left[\frac{e^{-f(x_i)}(1+\Pi e^{-f(x_i)}-\Pi)}{e^{-f(x_i)}(e^{f(x_i)}+1)} \right] = (1-y_i) \log \left[\frac{1+\Pi e^{-f(x_i)}-\Pi}{e^{f(x_i)}+1} \right] \\
 &= (1-y_i) \log \left[\frac{1+\Pi(e^{-f(x_i)}-1)}{e^{f(x_i)}+1} \right] \quad (2.6)
 \end{aligned}$$

Substituting from (2. 5) and (2. 6) into (2.4),then we get:

$$LogL = l = \sum_{i=1}^n [y_i \log \left(\frac{1+\Pi(e^{-f(x_i)}-1)}{1+e^{-f(x_i)}} \right) + (1-y_i) \log \left(\frac{1+\Pi(e^{f(x_i)}-1)}{1+e^{f(x_i)}} \right)]$$

$LogL$ can be written as:

$$LogL = l = \sum_{i=1}^n [I_{(Y_i=1)} \log \left(\frac{1+\Pi(e^{-y_i f(x_i)}-1)}{1+e^{-y_i f(x_i)}} \right) + I_{(Y_i=0)} (1-y_i) \log \left(\frac{1+\Pi(e^{-y_i f(x_i)}-1)}{1+e^{-y_i f(x_i)}} \right)],$$

where I denote the indicator function. Then:

$$LogL = l = \sum_{i=1}^n \log \left[\frac{1+\Pi(e^{-y_i f(x_i)}-1)}{1+e^{-y_i f(x_i)}} \right] \quad (2.7)$$

And hence Copas function can be defined as:

$$l^{Cepas}(u) = \frac{1+e^{-u}}{1+\Pi(e^{-u}-1)} \quad (2.8)$$

And hence Copas function can be defined as:

$$l^{Cepas}(u) = \frac{1+e^{-u}}{1+\Pi(e^{-u}-1)}$$

So, In view of loss function, we get the equivalent problem of log-likelihood maximization in (2.7) as follows:

$$MIN \sum_{i=1}^n \ell^{Cepas}(y_i f(x_i)) \quad , \quad \text{where } \ell^{Cepas}(y_i f(x_i)) \text{ is as defined in (2.8)}$$

Advantages of Copas function :

Cairo University-Institute of Statistical Studies and Research

Assuming the worst cases of classification, that is, classification error rate $(\Pi)=0.5$. Then, for any values of Π smaller than 0.5, we find that $l^{Cepas}(u)$ is a decreasing function in u and bounded by $-\log \Pi$, hence, reducing the impact of outliers.

Disadvantages of Copas function :

Copas function depends in its computation on the unknown parameter Π which must be estimated.

(2.3) Croux Estimator:

In order to overcome the problem of Inconsistency of Pregibon estimator and biasedness of Copas estimator, Bianco and Yohai (1996) proposed a consistent and more robust version of Pregibon estimator by adding a bias correction term .

Instead of minimizing $\sum_{i=1}^n \rho(d_i)$, they suggested to solve:

$$\text{Min} \sum_{i=1}^n \rho(d_i) + C_i \quad (2.9)$$

where d_i is the negative log-likelihood function and C_i is a bias correction term given by:

$$C_i = G(\pi(x_i)) - G(1 - \pi(x_i))$$

The bias correction term can be written as:

$$C(s, y) = G(F(s)) - G(1 - F(s)) \text{ , where:}$$

$$F(s) = \frac{1}{1 + e^{-s}} \text{ and } s = Z^T \gamma \text{ as previously defined}$$

$$\text{And : } G(t) = \int_0^t \rho'(-\ln u) du \quad (2.10)$$

where this choice of G is made to make the estimate consistent.

So, the estimator of Bianco and Yohai, simply defined by γ_{BY} is a special case of (2.1) with the function ρ given by:

$$\varphi_{BY}(s) = \rho(-\log(1-F(s))) + G(F(s)) + G(1-F(s)) \quad (2.11)$$

So, it is clear that the construction of Bianco and Yohai estimator depends on the choice of ρ function. Therefore, they suggested to use the following function:

$$\rho(t) = \begin{cases} t - \frac{t^2}{2k} & t \leq k \\ \frac{k}{2} & o.w \end{cases}$$

Where k is a tuning robustness constant .

Croux and Haesbroeck(2003) stated that the Bianco and Yohai estimator is consistent, asymptotically normal and have good bias and robustness properties. However, by numerical experiments carried by them, they confirmed inconveniency of this estimator(i.e:the Bianco and Yohai estimator most frequently does not exist). One major disadvantage of the BY estimator as pointed above is that the estimator sometimes may not exist .Even with samples without any outliers , it may happen that the estimator explodes to infinity.

A second disadvantage is that the derivative of the ρ function , given by:

$$\rho'(t) = \begin{cases} 1 - \frac{t}{k}, k \neq 0 & t \leq k \\ 0 & o.w \end{cases}$$

Vanish for large values of t and misclassified observations are severely downweighted while those observations that contain most of the information in logistic regression .

They suggested a new form of the loss function as follows:

$$\rho(t) = \begin{cases} te^{-\sqrt{d}} & t \leq d \\ -2e^{-\sqrt{t}}(1+\sqrt{t}) + e^{-\sqrt{d}}(2(1+\sqrt{d})+d) & o.w \end{cases}$$

Where k is a tuning robustness constant and it should be chosen to reach a compromise between robustness and efficiency. As k increase, the function gains in efficiency but loses in robustness. As a compromiser, they suggested $k = 0.5$

The estimator of Croux and Haesbroeck is given by :

$$\varphi_{CH}(s) = \rho(-\log(1-F(s)) + G(F(u)) + G(1-F(u)))$$

we derive Croux and Haesbroeck function as follows:

$$\ell^{ch}(u) = \rho(\ell^{ch}(u)) + G(F(u)) + G(1-F(u)) \quad (2.12)$$

$$\text{Where: } F(u) = \frac{1}{1+e^{-u}}, \quad \rho(u) = \begin{cases} u e^{-\sqrt{d}} & u \leq d \\ -2e^{-\sqrt{u}}(1+\sqrt{u}) + e^{-\sqrt{d}}(2(1+\sqrt{d})+d) & u > d \end{cases}$$

And $\rho(\ell^{CH}(u))$ is obtained as follows:

$$\rho(\ell^{CH}(u)) = \begin{cases} \log(1+e^{-u})e^{-\sqrt{d}} & \log(1+e^{-u}) \leq d \\ -2e^{-\sqrt{\log(1+e^{-u})}}(1+\sqrt{\log(1+e^{-u})}) + e^{-\sqrt{d}}(2(1+\sqrt{d})+d) & \log(1+e^{-u}) > d \end{cases}$$

$\rho(\ell^{CH}(u))$ can be written as:

$$\rho(\ell^{CH}(u)) = \begin{cases} \log(1+e^{-u})e^{-\sqrt{d}} & u \geq -\log(e^d - 1) \\ -2e^{-\sqrt{\log(1+e^{-u})}}(1+\sqrt{\log(1+e^{-u})}) + e^{-\sqrt{d}}(2(1+\sqrt{d})+d) & u < -\log(e^d - 1) \end{cases}$$

$$G(F(u)) = \begin{cases} \frac{1}{1+e^{-u}} e^{-\sqrt{-\log(\frac{1}{1+e^{-u}})}} - e^{-\frac{1}{4}}\sqrt{\pi} + e^{\frac{1}{4}}\sqrt{\pi}\Phi(\sqrt{2}(\frac{1}{2} + \sqrt{-\log(\frac{1}{1+e^{-u}})})) & \frac{1}{1+e^{-u}} \leq -\log(e^d - 1) \\ e^{-\sqrt{d}} \frac{1}{1+e^{-u}} - e^{-\frac{1}{4}}\sqrt{\pi} + e^{\frac{1}{4}}\sqrt{\pi}\Phi(\sqrt{2}(\frac{1}{2} + \sqrt{d})) & \frac{1}{1+e^{-u}} > -\log(e^d - 1) \end{cases}$$

$G(F(u))$ can be written as follows:

$$G(F(u)) = \begin{cases} \frac{1}{1+e^{-u}} e^{-\sqrt{\log(1+e^{-u})}} - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{\log(1+e^{-u})})) & u \leq -\log(e^d - 1) \\ e^{-\sqrt{d}} \frac{1}{1+e^{-u}} - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{d})) & u > -\log(e^d - 1) \end{cases}$$

$$1-G(F(u)) = \begin{cases} 1 - \frac{1}{1+e^{-u}} e^{-\sqrt{\log(1-\frac{1}{1+e^{-u}})}} - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{-\log(1-\frac{1}{1+e^{-u}})})) & 1 - \frac{1}{1+e^{-u}} \leq e^{-d} \\ e^{-\sqrt{d}} (1 - \frac{1}{1+e^{-u}}) - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{d})) & 1 - \frac{1}{1+e^{-u}} > e^{-d} \end{cases}$$

$1-G(F(u))$ can be written as follows:

$$1-G(F(u)) = \begin{cases} \frac{1}{1+e^u} e^{-\sqrt{\log(1+e^u)}} - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{\log(1+e^u)})) & u \geq \log(e^d - 1) \\ e^{-\sqrt{d}} (\frac{1}{1+e^u}) - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{d})) & u < \log(e^d - 1) \end{cases}$$

Now, combining similar terms together, we get:

$$\begin{aligned} \ell^{CH}(u) = & I_{\{u \geq -\log(e^d - 1)\}} \left[\log(1+e^{-u}) e^{-\sqrt{d}} + e^{-\sqrt{d}} \frac{1}{1+e^{-u}} - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{d})) \right] \\ & + I_{\{u < -\log(e^d - 1)\}} \left[-2e^{-\sqrt{\log(1+e^{-u})}} (1 + \sqrt{\log(1+e^{-u})}) + e^{-\sqrt{d}} (2(1 + \sqrt{d}) + d) + e^{-\sqrt{d}} \frac{1}{1+e^{-u}} - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{d})) \right] \\ & + I_{\{u \geq \log(e^d - 1)\}} \left[\frac{1}{1+e^u} e^{-\sqrt{\log(1+e^u)}} - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{\log(1+e^u)})) \right] \\ & + I_{\{u < \log(e^d - 1)\}} \left[e^{-\sqrt{d}} (\frac{1}{1+e^u}) - e^{-\frac{1}{4}} \sqrt{\pi} + e^{\frac{1}{4}} \sqrt{\pi} \Phi(\sqrt{2}(\frac{1}{2} + \sqrt{d})) \right] \end{aligned}$$

(2.13)

where I denote the indicator function.

So, In view of loss function, we get the equivalent problem of log-likelihood maximization :

$$\text{MIN} \sum_{i=1}^n \ell^{CH}(y_i f(x_i)) \quad , \text{ where } \ell^{CH}(y_i f(x_i)) \text{ is as defined in (2.12)}$$

Croux and Haesbroeck estimator has the property that it exists whenever the ML-estimator exists. They provided a fast and stable algorithm for its computation and showed in a simulation study the good performance of that estimator with respect to other proposals.

Also, they provided in (2008) The weighted version of the Bianco and Yohai estimator which is defined as:

$$\hat{\gamma}_n = \arg \text{Min} \sum_{i=1}^n w_i Q(Z_i^T \gamma, y_i)$$

where the weights depend on the Robust Distance of the observation x_i . This robust distance RD_i is equal to the Mahalanobis distance of x_i to the center of the data cloud in the covariate space, with the center and covariance-matrix robustly estimated. The weights are generated as:

$$W(t) = I(t^2 \leq \chi_{p,0.975})$$

(2.6) Maximum Estimated Likelihood (MEL) and Weighted Maximum Estimated Likelihood(WEMEL) estimators:

Rousseeuw and Christmann(2003) presented the weighted Estimated Maximum likelihood (WEML) estimator.

Consider the following maximum likelihood estimator of T given (Y = y)

$$\hat{T}_{ML}(Y = 0) = f$$

$$\hat{T}_{ML}(Y = 1) = s$$

Where f and s are the state of failure and success, respectively

If we consider the conditional probability that $Y = 1$ given $\hat{T}_{ML}$, then:

$$\begin{aligned} p(Y = 1 | \hat{T}_{ML}) &= \delta_0 \quad \text{if } y = 0 \\ p(Y = 1 | \hat{T}_{ML}) &= \delta_1 \quad \text{if } y = 1 \end{aligned} \quad (2.13)$$

Where y is the observed value of Y . Denoting (3.13) by $\tilde{Y}$, then we write:

$$\tilde{Y}_i = (1 - \tilde{Y}_i) \delta_0 + \tilde{Y}_i \delta_1$$

Which is a weighted average of δ_0 and δ_1 with weights $1 - Y$ and Y .

In a model with n observations, $\tilde{y}_i$ is obtained analogously as:

$$\tilde{y}_i = (1 - y_i) \delta_0 + y_i \delta_1 \quad (2.14)$$

Which we call pseudo observations.

Choice of δ_0 and δ_1 :

If δ_0 and δ_1 are known from the context (e.g. from the type I and type II error probabilities), then we can use these values. But in many cases, but in many cases, δ_0 and δ_1 are not given in advance. Copas (1988) found that accurate estimation of δ_0 and δ_1 itself is very difficult, if not possible unless n is extremely large. Note that Rousseeuw and Christmann approach is somewhat similar to Copas approach. The 'symmetric' approach used by Copas is to choose a single constant $\Pi = \delta_0$ and $1 - \Pi = \delta_1$. On the other hand, there is something to be said about the 'asymmetric' choice which takes into account how many y 's are 0 and 1 in the data set. When we consider the marginal distribution of the y_i (that is unconditional on x_i), then we can construct some estimate $\hat{\pi}$ of the marginal success probability $P(Y = 1)$. In this case, the best estimate of $P(Y = 1)$ is $\hat{\pi} = (\text{number of } y_i = 1) / n$ (i.e., the average of the y_i 's ($\bar{y}$) equal $\hat{\pi}$). So, in a same manner, we want to choose δ_0 and δ_1

Such that the average of pseudo observations equal $\hat{\pi}$.

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^n \tilde{y}_i = (1 - \hat{\pi})\delta_0 + \hat{\pi}\delta_1$$

$$\hat{\pi} - \hat{\pi}\delta_1 = \delta_0 - \hat{\pi}\delta_0 \rightarrow \hat{\pi}(1 - \delta_1) = \delta_0(1 - \hat{\pi})$$

Which can be written as :

$$\frac{1 - \delta_1}{\delta_1 - \hat{\pi}} = \frac{\delta_0}{\hat{\pi} - \delta_0} \quad (2.15)$$

Now, if we let the latter ratios equal a small positive number, say, δ . Then, we can write both

δ_0 and δ_1 as follows:

$$1 - \delta_1 = \delta(\delta_1 - \hat{\pi}) \rightarrow \delta_1 = \frac{1 + \hat{\pi}\delta}{1 + \delta} \rightarrow 1 + \hat{\pi}\delta = \delta_1(1 + \delta)$$

From which we get: $\delta_1 = \frac{1 + \hat{\pi}\delta}{1 + \delta}$. Similarly, $\delta_0 = \delta(\hat{\pi} - \delta_0) \rightarrow \delta_0(1 + \delta) = \hat{\pi}\delta$ which

$$\text{leads to } \delta_0 = \frac{\hat{\pi}\delta}{1 + \delta}$$

Estimation of probability:

From (2.15), since the denominator cannot be negative, we conclude that:

$$\hat{\pi} \leq \delta_1 \quad \text{and} \quad \hat{\pi} \geq \delta_0 \quad (\text{i.e.: } \delta_0 \leq \hat{\pi} \leq \delta_1)$$

$$\text{- When all } y_i \text{'s} = 1 \rightarrow \hat{\pi} = \bar{\pi} = 1$$

$$\text{- When all } y_i \text{'s} = 0 \rightarrow \hat{\pi} = \bar{\pi} = 0 \rightarrow \hat{\pi} = \bar{\pi} = 1$$

So, $\hat{\pi}$ could be zero or 1. So, a natural idea is to bound $\hat{\pi}$ away from 0 and 1 by putting:

$$\hat{\pi} = \max(\delta, \min(1 - \delta, \bar{\pi})) \text{ or equivalently:}$$

$$\hat{\pi} = \min(\delta, \max(1 - \delta, \bar{\pi}))$$

With the last formula , we conclude that $\hat{\pi}$ is estimated robustly with the Minimax property of M-estimators.

Missclassification probabilities(classification errors):

When $Y_i = 1$, we have:

$$p(Y_i = 1 | T = s) = \delta_1 \quad (\text{correct})$$

$$p(Y_i = 0 | T = s) = 1 - \delta_1 \quad (\text{Missclassification})$$

Similarly,

When $Y_i = 0$

$$p(Y_i = 0 | T = f) = 1 - \delta_0 \quad (\text{correct})$$

$$p(Y_i = 1 | T = f) = \delta_0 \quad ((\text{Missclassification}))$$

So, the misclassification probabilities are:

$$1 - \delta_1 \quad \text{and} \quad \delta_0$$

For the first type of Missclassification: δ_0

$$\delta_0 = \frac{\hat{\pi}\delta}{1 + \delta} < \frac{\delta}{1 + \delta} < \delta$$

For the first type of Missclassification: δ_1

$$1 - \delta_1 = 1 - \frac{1 + \hat{\pi}\delta}{1 + \delta} = \frac{(1 - \hat{\pi})\delta}{1 + \delta} < \frac{\delta}{1 + \delta} < \delta$$

So, both types of errors are less than δ .

By simulation studies, Christmann showed an appropriate choice of δ is $\delta = 0.01$ which will lead to smaller classification errors.

To downweight leverage points in X-space, Christmann suggested the weighted maximum estimated likelihood with weight function given by:

$$w_i = \frac{M}{\max\{RD(x_i^*), M\}} \quad \text{Where: } x_i^* = (x_{i,2}, \dots, x_{i,p})^T, \quad RD(x_i^*) \text{ is the robust}$$

Mahalanobis distance that depend on the minimum covariance determinant (MCD) estimator and M is the 75th percentile of all $RD(x_j^*)$, $j=1,2,\dots,n$.

Advantages of Wiegthed Estimated Maximum likelihood estimator:

1- WEMEL always exist and id does not require the condition of overlap(i.e: it can always be computed ,even in case of complete separation of the two groups)

2-As, we have shown , WEMEL can be computed when all values of y are 0 or 1.

3- WEMEL is x-affine equivariant

This property means that when the x_i are replaced by $\tilde{x}_i = Ax_i$, where A is a non-singular $p \times p$ matrix, then the regression coefficients transforms accordingly.

4- WEMEL is equivariant with respect to recoding the response variable

MEL estimate is the value that maximizes the log-likelihood:

$$\begin{aligned} \log L &= \sum_{i=1}^n \{ \tilde{y}_i \log \hat{\pi}(x_i) + (1 - \tilde{y}_i) \log(1 - \hat{\pi}(x_i)) \} \\ &= \sum_{i=1}^n \{ [(1 - y_i)\delta_0 + y_i\delta_1] \log(\max(\delta, \min(1 - \delta), \bar{\pi})) + [(1 - (1 - y_i)\delta_0 + y_i\delta_1)] (\log(\min(\delta, \min(1 - \delta), \bar{\pi}))) \} \\ &= \sum_{i=1}^n \{ [\delta_0 + (\delta_1 - \delta_0)y_i] \log(\max(\delta, \min(1 - \delta), \bar{\pi})) + [(1 - \delta_0) + (\delta_0 - \delta_1)y_i] (\log(\min(\delta, \min(1 - \delta), \bar{\pi}))) \} \\ &= \sum_{i=1}^n \log \left\{ \frac{(\max(\delta, \min(1 - \delta), \bar{\pi}))^{[\delta_0 + (\delta_1 - \delta_0)y_i]}}{(\min(\delta, \min(1 - \delta), \bar{\pi}))^{[1 - \delta_0 + (\delta_0 - \delta_1)y_i]}} \right\} \end{aligned} \quad (2.16)$$

Where:

$$\bar{\pi} = p(Y_i = 1)$$

And in view of loss function, we get:

$$\ell^{MEL}(u) = \frac{e^{(\min(\delta, \min(1-\delta), \bar{\pi}))^{[1-\delta_0+(\delta_0-\delta_1)u_i]}}}{e^{(\max(\delta, \min(1-\delta), \bar{\pi}))^{[\delta_0+(\delta_1-\delta_0)u_i]}}}$$

So, the problem becomes: $\text{Min } \ell^{MEL}(y_i)$, where $\ell^{MEL}(y_i)$ is as given in (2.16)

References

- 1- **Bianco, A. M. and Yohai, V. J. (1996)**, “Robust estimation in the logistic regression model,” in *Robust Statistics, Data Analysis and Computer Intensive Methods*, 17–34, Springer Verlag, New York
- 2- **Cox(1966),D.R.**, “Some procedures associated with the logistic qualitative response curve”., *Research papers on statistics* , 55–77. John Wiley, New York.
- 3- **Croux, C. and Haesbroeck, G. (2003)**, “Implementing the Bianco and Yohai estimator for logistic regression,” *Computational Statistics and Data Analysis*, **44**, 273–295.
- 4- **Copas, J. B. (1988)**,”Binary regression models for contaminated data (with discussion)" *J. Royal Statist. Soc. B*, **50**, 225-265.
- 5- **Croux, C., Haesbroeck, G., and Joossens, K. (2008)**, ``Logistic Discrimination using Robust Estimators: an Influence Function Approach," *The Canadian Journal of Statistics*, **36**,157-174
- 6- **Hampel,F.R.(1974).**,”The influence curve and its role in estimation”, *American Statistics Association*,383-393.
- 7- **Pregibon, D. (1981)**, “Logistic Regression Diagnostics,” *Annals of Statistics*, **9**,705–24.
- 8- **Rousseeuw, P. J.and Christmann, A. (2003)**, “Robustness against separation and outliers in logistic regression,” *Computational Statistics and Data Analysis*, **43**, 315–332.
- 9- **Sanizah Ahmad, Norazan Mohamed Ramli, Habshah Midi."**(2010),” Robust estimators in logistic regression: A comparative simulation study”, *Modern Applied Statistical Methods*,**9**,502-511.