منهجية دمج طريقتي الانحدار وتحليل المركبات الأساسية في اختيار طاقم المتغيرات المستقلة

A Methodology of Emerging Regression And Principal Component Analyses Methods For Selecting Set of Variables

د. عبدالحميد عبدالمجيد البلداوي
استـــاذ مساعـــد
كليـــة المنصـــور الجامعـــة
بغداد - العراق

تمهيد :

تتوفر عدة طـرق ومنهجيـات احـصائية لغـرض اختيـار مجموعـة المتغيـــرات (التفسيرية) التي يمكن تضمينها في النموذج الذي يتم بنـاءه ، ويأتي هذا الاختيار لاجل تحقيق هدفين حاسمين وهما :

- تحاشي مشكلة العلاقات الخطية المتداخلة بـين المتغيـرات التـي يـتم ترشـيحها واخـضاعها للتحليـل وهـي مـا يطلـق عليهـا بمـسألة " Multicolinearity " .
- والاخر هو لتقليص عدد المتغيرات التي يضمها النموذج بغية تقليل الكلفة من جهة وتسهيل عملية احتسابه واستخدامه من جهة اخرى .

ومن اهم هذه الطرق (التي سيرد التطرق اليها لاحقا) هي التـي تعتمد الانحدار المتعـدد Multiple Regression ; (Draper and smith 1990) ، وتحليـل المركبـات الاساسـية Principal Component Analysis (Hocking 1976 Jeffers 1978, Kendall 1980, Al- Beldawi 1978)

الا ان الطريقة المقترح تطبيقها موضوع هذا البحث هي حـصيلة دمج بين كلا الطـريقتين اعـلاه (Daling and Tamura, 1970) والتي يمكن اعتبارها الحالة العمليـة لطريقـة Principal Component Regression (Kendall, 1959) . والتي من ميزاتها هو تـوفير مرونة عالية للباحث في اختيار المتغير الاكثر اهميـة لتـضمينه فـي النموذج ، بالاضافة الى النتائج المعنوية التي يمكن ان تـستمخض عنهـا عملية التحليل بأستخدام المنهجية المقترحة وكما افصحت عن ذلك نتائج هذا البحث .

وللوقوف على مدى اهمية المنهجية مقارنة بـــالطرق التقليديـــة الاخرى سيتم اخضاع المتغيرات المرشحة لعملية التحليل بأستخدام الطرق الثلاث : الانحدار المتعدد ، وتحليـــل المركبـــات الاساســـية ، والطريقـــة المقترحة وهي الدمج بين هاتين الطريقتين ، ومن ثم توظيـــف معـــايير تقييم المعنوية لنتائج كل منها .

1. مفهوم اختيار افضل طاقم من المتغيرات المستقلة

عند صياغة النموذج الخطي العام ، يفترض وجود علاقة خطيـــة بين المتغيـــر المعتمـــد Y و(K-1 مـــن المتغيـــــرات المـــستقلـــــة (التفسيريـــــــة)) ($X_1, X_2, \ldots\ldots\ldots\ldots, X_{K-1}$) والمتغيـــر العشوائي μ وان هـــذه العلاقـــة فـــي حالـــة اخـــذ عينـــة مـــن المـــشـــــــاهدات Y_i ($i = 1,2,3,\ldots\ldots, n$) فـــأن الـــصيغة العامـــة للنموذج يكون :

$$Y_i = B_o + B_1 X_1 + B_2 X_2 + \text{-------} + B_{K-1} X_{K-1} + \mu$$

وان هذا النموذج يجب ان يحقق بالاضافة للفرضـــيات المتعلقـــة بالمتغير العشوائي فرضية عدم وجود علاقة خطية م بـــين المتغيـــرات المستقلة (التفسيرية) ، اى ان هذه المتغيرات تكون مستقلة عن بعـــضها كما وان عدد المشاهـــدات التـــي
هي n يجب ان لا تزيد على عدد المعلمات المجهولة في النموذج اى :

$$K \leq n$$

وفي حالة عدم تحقق هذه الافتراضات فأن النتـائج التـي يـتم التوصل اليها لا تكون صحيحة ، لذلك ينبغي عـدم اسـتخدام النمـوذج الخطي العام الا بعد التأكد اولا من مدى انطباق الافتراضات التـي علـى اساسها تم بناء النموذج على واقع المشاهدات . ومن بـين اهـم هـذه الفرضيات هي المتعلقة بمشكلة العلاقات الخطية ، والتي تبـرز عنـدما تكون هناك علاقات تامة او شبه تامة بين المتغيرات المستقلة ، لان قيمة محدد المصفوفـة $|X'X|$ تساوي عندئـذ صفر مما يتعذر معه ايجاد $(X'X)^{-1}$ وبالتالي يتعذر تقدير قيم المعلمات المجهولة .

2. نموذج تحليل المركبات الاساسية **Principal Component Analysis**

وهي طريقة احصائية وصفية تستخدم مع البيانات ذات المتغيرات المتعددة . وتقوم بتجميع كل مجموعة من البيانات المترابطة خطيا فـي احد المركبات الاساسية . وتعتمد فـي ذلـك علـى مـصفوفة الارتبـاط للمتغيرات التي تعتبر المرحلة الاساس في عملية تحليل المركبات.

والمركبات الاساسية $C\rho\ 'S$ هـي متغيـرات عـشوائية غيـر مترابطة ، وكل منهـا تضـم مجموعـة متغيـرات عـشـوائية ($X_1, X_2, ----------, X_p$) مترابطة وتشترك بأتجاه خطـي ، بحيـث تأخذ الصيغة التالية :

$$C_j = \sum_{i=1}^{p} a_{ij} \quad , \quad j = 1,2, -------- , p$$

وان a_{ij} هي عبـارة عـن معـاملات عناصـر (elements) مصفوفة الموجهات الذاتية الطبيعية (Normalised eigen vectors) لمصفوفة الارتباط للمتغيرات $X_i's$

وبواسطة مصفوفة قيم الموجهـات الذاتيـة (eigen values) نحصل على الحجم النسبي للتغاير او التباين المفسر للبيانات الاحـصائية بواسطة كل من المركبات الاساسية التي يتم احتسابها مـن المتغيـرات $X_i's$ بأستخدام التحويل المتعامـد (orthogonal transformation) ويطلق على المعاملات a_{ij} عـادة بتحمـيلات المركبـة (component loadings) وهذه التحميلات تشير الى وزن العلاقة بـين المتغيـرات $X_i's$ والمركبات الاساسية $C_p's$ بشرط اخذ الجذر التربيعي لمـصفوفة ارتباط تباينات الشيوع لقـيم التبـاين الـذاتي (Morrison, 1967) . ويطلق على قيم الجذر التربيعي لتباينات الـشيوع (Communality) ويمكن الرمز له بـ :

$$h_j = a_j^2{}_1 + a_j^2{}_2 + \text{----------} + a_j^2{}_p$$

3. نموذج الانحدار المتعدد Multiple Regression

ان جوهر الافكار التي تعتمد عليها جميع طرق الاختيار التي سيلي ذكرها هي تضمين المتغير الذي يضيف اكبر زيادة ممكنة الـى قـوة التفـسير للنموذج ، واذا كان على المتغير ان يحذف فيجب ان يكون تأثير حذفـه اقل ما يمكن على قدرة النموذج التفسيرية .

اما اهم طرق نموذج الانحدار المتعدد المستخدمة لاختيـار افـضل طـاقم متغيرات مستقلة فهي :

1) طريقة شـمول كافـة المتغيـرات (All Possible Regression) :- وتستخدم اذا كان عدد المتغيرات ليس كبيـرا ، وابـرز عيوبهـا حاجتهـا لعمليات حسابية ووقت كبيرين .

2) طريقة الاضافة المتتاليـة (Forward Selection Regression) :- وفيها اذا كانت قيمة F المجدولة هي اقل من المحتسبة عنـدها يتوقـف البحث عن متغير ، وبعكسه يتم ادخال متغير جديد الـى المعادلـة واعـادة الاحتسـاب أي : $H_o : \beta_i = o$

$vs\ H_1 : \beta_i \neq o$

طريقـة الحـذف التنـازلي (Backward Elimination Selection Regression):- وهذا اذا كانت قيمة F المحتسبة لكافة المتغيرات اكبر من قيمة F الجدولية ، عندها يحذف متغير من المعادلـة والرجـوع لمعرفة قيمة F المحتسبة من جديد وهكذا لغاية تفوق قيمة F الجدولية .

3) طريقة الخطوات المتتاليـة (Stepwise Selection Regression) :- تجمع بين طريقتي الاضافة المتتالية (FS) والحذف التنازلي (BE)، وفي كل خطوة يتم اختيار متغير ابتداء من الاكثر اهمية ولغاية عدم هبوط قيمـة F المحتسبة عن قيمة F الجدولية بكلمة اخرى اجـراء اختبـار معـاملات المتغيرات لمعرفة معنويتها من عدمها .

وتعتبر طريقة الخطوات Stepwise Regression هي اكثـــر الطـــرق استخداما وانتشارا من الناحية العملية لقلة الوقت الذي تحتاجه في عملية الاحتساب بالاضافة الى انها تعرض النتائج في كل خطوة بصورة واضحة ومرضية ومبكرة من دون الحاجة لاجراء الخطوات غير المعنوية .

4. منهجية الدمج بين طريقتي المركبات الاساسية والانحدار :

Principal Component Regression Analysis

وهي الطريقة المقترحة من قبل Kendall, 1957 ، وتمتاز بتوفيرهـــا لمرونة كافية لاختيار المتغير الاكثر اهمية وذلك بالاعتماد على خبـــرة الباحـــث العملية ومثل هذه المرونة لا توفرها الطرق التقليدية التي تعتمد حـــصرا علـــى الاساليب الميكانيكية بغض النظر عن مدى اهمية المتغير .

والفكرة التي تقوم عليها الطريقة هي : عند افتراض ان R هي مصفوفة الارتباط لـــ P من المتغيرات التوضيحية او التفسيرية Xi وكما في الـــسابق حيث ان(P=I , 2,-------,1) وافترضنا بأن C التي هي مـــصفوفة المركبـــات الاساسية المتعامدة orthogonal matrix و D هـــي المـــصفوفة القطريـــة Diagonal Matrix لقيم التباين الذاتي λi بحيث :

$$D= C^{T}RC$$

فأن المركبات الاساسية هي عناصر للموجه (Vector)

$$gi = C^T X$$

وبذلك فأن انحدار المتغير التابع (الاستجابة) Y على المركبات gi تعطي انحدار تعامدي

ويمكن تلخيص اهم الخطوات العملية للطريقة بما يلي :

1) اختيار عدد المركبات الاساسية التي تفسر اعلى نسبة من التباين ولنقل ما يزيد على 90% مثلا .
2) تحليل انحدار المتغير التابع (الاستجابة) Y مع كل مركبة يتم اختيارها في الخطوة 1 اعلاه واحتساب قيمة R^2 مع التحقق من صحة اشارة المعاملات .
3) اختيار المركبات التي تحقق اعلى R^2 .
4) اختيار متغير مستقل من كل مركبة تم اختيارها في الخطوة 3 اعلاه على الاسس التالية:
   - الاهمية العملية للمتغير بالنسبة للمتغير التابع (الاستجابة)
   - سهولة قياسه (كلفته) .
   - درجة ارتباطه مع المتغير التابع (الاستجابة) .
5) اجراء عملية تحليل الانحدار للمتغير المعتمد مع كل من المتغيرات المستقلة التي يتم اختيارها في الخطوة 4 اعلاه .

5. البيانات الاحصائية :

لاجل اختبار مستوى اداء المنهجية المقترحة ومقارنتهـــا بمـــستوى اداء الطـــــــرق التقليدية فقد تم ترشيح ثلاثة عشر متغيرا توضيحيا تم الحـــصول عليها من مسح احصائي لعينة شملت 842 مسافرا من مسافري النقل بين مراكز البلديات في العراق والتي تم توفيرها لغرض بناء نموذج احصائي لتفسير الطلب على النقل بين المدن . وسيكون المطلوب في هذا البحث هو استخدام كـــل مـــن النماذج الثلاث لاختيار افضل مجموعة متغيرات معنوية من بـــين هـــذه لـــــ 13 متغيرا . والمتغيرات المرشحة هي :

1) مسافة الطريق (بالكم) X1 قيمة مطلقة
2) يوم السفر X2 متغير هيكلي (Dumy Variable) (1, 2, ----, 7)
3) نوع واسطة النقل (باص متوسط الحجم ، باص كبير الحجم) X3 متغيـــر هيكلي 1,2
4) وقت السفر(قبل الظهر ، بعد الظهر) X4 متغير هيكلي 1،2
5) الغرض مـــــن السفر (رحلـــــة عمل ، رحلة غير عمل) X5 متغيـــر هيكلي 1،2
6) عمر المسافر X6 قيمة مطلقة (عدد السنين)
7) جنس المسافر X7 متغير هيكلي 1،2
8) مهنة المسافر X8 متغير هيكلي 1,2,3,--,5
9) معدل دخل اسرة المسافر الشهري X 9 قيمة مطلقة (بالدينار)
10) طول زمن الرحلة X10 قيمة مطلقة (بالدقائق)
11) معدل وقت انتظار المسافر في المحطة X11 قيمة مطلقة (بالدقائق)

12) اجور السفر X12 قيمة مطلقة (بالدينار)

13) معدل عدد رحلات واسطة النقل الواحدة يوميا X13 قيمة مطلقة (عدد)

1. نتائج التحليل :

6ـ1 نموذج الدمج بين طريقتي المركبات الاساسية والانحدار

طبقا لمتطلبات التحليل التي تطرقنا اليها في (الفقـرة 4) اعـلاه ، ففـي الخطوة الاولى وكما مبين في الجدول رقم (1) حصلنا على 10 مركبات اساسـية معنوية استطاعت هذه المركبات ان تقوم بتفسير اكثر من 99% من التباين .

وعند اجراء عملية تحليل الانحدار للخطوة التالية مع كل من المركبـات العشرة التي حصلنا عليها ، واحتساب قيمة R^2 لكل منها ، نستدل على معنوية 7 مركبات وهي ، 1،4،5،6،7،9،10 وكما مبين في الجدول رقم (1) .

وبأعتماد الاسس المشار اليها في الخطوة الرابعة لاجل تحديد متغير واحد من كل مركبة ، نجد ان مجموعة المتغيرات المختارة تشمل على التوالي كل من :

X2 , X5 , X7 , X9 ,X12 ,X11 ,X4 ,

وعند بناء النموذج الذي سيضم هذه المجموعة من المتغيرات , سيحقق المعايير الاحصائية التالية :

R^2 = 0.84

S.E = 232

F-ratio = 609 Sig at 0.000

عدد المتغيرات = 7

2-6 نموذج الانحـــدار بطريقـــة كافـــة المتغيـــرات (All. Possible Regression)وبأفتراض ان البيانات خاليـــة مـــن مـــسألة العلاقـــات المتداخلة (Multicollinearity)وتضمين المتغيرات المستقلة البـــالغ عددها 13 متغيرا في النموذج ، فأن النموذج سيؤول الى ما يلي :

$R^2 = 0.84$

S.E. = 231

F = 329 Sig at 0.000

عدد المتغيرات = 13

Table (1)

Results of Variables Selection Using The Principal Component Regression .

Component	R^2 *	Variable Selected
1	.05	X11
2	.02	---
3	.00	---
4	.05	X9
5	.06	X7
6	.04	X4
7	.04	X2
8	.00	----
9	.05	X5
10	.05	X12

* مع حجم عينة $n = 842$ فأن قيم R^2 في الجدول اعـلاه هـي عاليـة المعنويـــة أي $\alpha = 0.01$

3-6 نموذج تحليل المركبات الاساسية

وبأستخدام طريقة تحليل المركبات الاساسية ، نستدل وكما موضـح فـي الجدول (2) بأن هناك 10 مركبات (عوامل) معنوية ، استطاعت تفـسير 99.4%

من مجموع التباين . وبأعتماد حجـم التحميـل (size of loading) واهميـة المتغيرات بالنسـبة للمتغير المعتمد فأن هذه المركبات العشرة يمكن تـسميتها كما يلي :

1. طول زمن الرحلة . وان اعلى معامل تحميل في هذه المركبة يعود لمتغيري مسافة الطريق (0.98) وطول زمن الرحلة (0.97) وان نسبة التباين التي قامت مجموعة متغيرات هذه المركبة بتفسيرها هي 33.9% .
2. واسطة النقل . وقد بلغ معامل التحميل لهذا المتغير هو 99% وساهمت
3. المركبة الثانية بتفسير 13.7% من مجموع التباين .
4. عمر المسافر . بمعامل تحميل مقداره 97% وتفسير 8.8% من مجموع التباين .
5. معدل دخل الاسرة الشهري . بمعامل تحميل مقداره 98% وساهمت هذه المركبة بتفسير 7.9% من مجموع التباين .
6. جنس المسافر . بمعامل التحميل مقداره 99% وبتفـسير 7.4% مـن مجموع التباين .
7. الغرض من السفر . بمعامل تحميل مقداره 99% وبتفسير 6.6% من مجموع التباين .
8. يوم السفر . بمعامل تحميل مقداره 99% وبتفسير 6.5% من مجموع التباين .
9. مهنة المسافر . بمعامل تحميل مقداره 95% وبتفسير 5.6% مـن مجموع التباين .

10. وقت السفر . بمعامل تحميل مقـداره 95% وبتفـسير 5.1% مـن مجموع التباين .

11. معدل وقت الانتظار في المحطة . بمعامل مقداره 93% وبتفسير 3.9% من مجموع التباين .

Table (2)
Principal Component Loading for 13 Situat6ional Independent Variables

Variable	Components									
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
X1	.98									
X2							.99			
X3		.99								
X4	.31									
X5						.99			.95	
X6			.97							
X7					.99					
X8								.95		
X9				.98						
X10	.97									
X11	.35									
X12	.94									
X13	.95									
Eigenvalue	4.4	1.76	1.15	1.03	.97	.86	.84	.72	.66	.51
% Variation	33.9	13.7	8.8	7.9	7.4	6.6	6.5	5.6	5.1	3.9
Cum.% Variation	33.9	47.9	56.4	64.3	71.7	87.3	84.8	90.4	95.5	99.4

وان معايير تقييم نتائج عملية بناء النموذج الذي يتضمن مجموعة لـــ 10 متغيرات المختارة في اعلاه ذات اعلى تحميل Loading في كل مركبـــة هي :

$R^2 = 0.84$
S.E = 231
F =0 609 Sig. at 0.000

عدد المتغيرات = 10

4-6 المقارنة بين النتائج

لاجل اعطاء صورة واضحة من خلال المقارنة فأن الجدول التالي يعطـــي حصيلة كل طريقة ومستوى معنوية كل نموذج .

المعـــــــايير الاحصائية	النموذج النهـائي لكافة المتغيـرات بطريقة الانحدار	النموذج النهـائي بالاعتمـاد علـى طريقـــة تحليـــل المركبات	النمـــوذج النهـــائي بالاعتمـــاد علـــى طوٍيقة الدمج بـين تحليـــل المركبـــات والانحدار
عدد المتغيرات	13	10	7
R^2	0.84	0.84	0.84
S.E	231	233	232
F-ratio	329	609	856
At	0.000	0.000	0.000

ومنه نلاحظ بأنه رغم ما حققته المنهجية المقترحة وهـي دمـج بـين تحليـل المركبات والانحدار في اعداد المتغيرات التي يتـضمنها النمـوذج الا ان درجـة المعنوية قد ارتفعت بالنسبة لمعيار F-ratio مع عدم حصول اى انخفاض فـي درجة معنوية باقي المعايير الاحصائية

2. الاستنتاجات والتوصيات:

7-1 الاستنتاجات :

1. ان النتائج التي تمخضت عن عملية التحليل وفقا للمعايير الاحصائيـــــة والاقتصادية تشير الى ان طريقة نموذج الدمج بين طريقتي تحليل المركبات وتحليل الانحدار ، تمتاز بتقليص عدد المتغيرات المستقلة الى 7 متغيرات فقط مقابل عدم الانخفاض في قدرة النموذج التفسيرية للتباين .
2. ان منهجية الدمج ساعدت في تحديد اهم المتغيرات وفقا للاسباب المنطقية التي لها علاقة في تفسير الطلب مع شفافية الكشف عن المتغيرات المهمة التي لم يتضمنها النموذج ولكنها ممثلة Proxy بواسطة المتغيرات الداخلة الداخلة في النموذج من خلال النظر على تحميلات المتغيرات المسطرة في مركبة عند تحليل المركبات الاساسية ، استنادا الى العلاقة الخطية التي تربط هذه المتغيرات للسركبة الواحدة .
3. ان المنهجية تزداد فائدتها ومردوداتها الاقتصادية عندما تكون امام عدد كبير من المتغيرات وذلك لما تيسره في الكشف عن تلك الاكثر اهمية من خلال التأمل بحجم معاملات التحميل لكل متغير .

4. وفقا للأسس النظرية لتحليل المركبات الاساسية القائلة بوجـود علاقـات خطية مترابطة بين متغيرات كل مركبة وعدم وجود العلاقة بين المركبات . فأن عملية التخلص من العلاقات المتداخلة بين المتغيرات هـي اكثـر وضوحا وفعالية وتيسرا من خلال اختيار الممثل الانسب في كل مركبة .

7-2 التوصيـات

1. بالرجوع عما تحقق من توظيف المنهجية المقترحة من نتائج مهمة ، فقد يكون من المفيد استخدام هذه المنهجية في عملية بنـاء النمـاذج الاقتـصادية والاجتماعية . وعلى الاخص عند مجهولية المتغيرات التي لها تأثير فعلي على الظاهرة المدروسة بسبب عـدم وجود دراسـات سـابقةعنها او بـسبب مستجدات قد طرأت على الظاهرة مما يستدعـي ترشيح اعداد كبيـرة من المتغيـرات لتحديـد المجموعة التـي لها تأثيـر فعلي وهـام علـى الظاهـرة الاقتصاديـة والاجتماعية . فأستخـدام هـذه المنهجيـة مـن شأنـه تيسيـر عمليـة التحليـل وللخروج بشفافيـة اكبـر في اتخاذ قرار اختبار المتغيرات .

2. ان اللجوء الى المنهجية موضوعة البحث من شأنها التخلص مـن دوامـة اختيار الطريقة الانسب في عملية اختبار افـضل طـاقم متغيـرات مـستقلة ، والذي من شأنه محاولة اللجوء الى اغلب الطرق لاختبار الافـضل وهـو مـا يتطلب كلفة اضافية في الوقت والاموال .

3. من المفيد ، اجراء دراسات لاحقة بأجراء تجربة محاكاة (Simulation) للوقوف على نتائج وتوصيات متقدمة ايضا بهذا الصدد.

المصـــــــادر

1. A.H.Al-Beldawi, Methods of Scientific Research using SPSS, Dar Al-Shrouk, Aman, 2004 .

2. A.H. Al-Beldawi, Statistics for Business and Applied Sciences, Dar Al-Shrouk, Aman, 1997 .

3. Daling , J.R. and Tomura, Use of Orthogonal Factors for selection of Variables in a Regression Equation .
   An illustration , Applied Statistics, PP 260-68, 19,1970 .

4. Draper & Smith, Applied Regression Analysis, John Witley and son Inc. , London 1990 .
5. Hocking, R.R. , The Analysis and Selection of Variables in Linear Regression Biometrics, 32, PP.1-49, 1976 .
6. Jeffers, J.P An Introduction to system Analysis : with ecological applications, William Clowes and sons LTD, London , 1978 .
7. Kendall , M. Multivariate Analysis, Second edition, Charts Griffin and LTD. , London, 1980.
8. Kendall , M. (1957) , A course in Mutivariat Analysis, Hafner New York , 1957 .

9. Morrison, D.F. , Multivariate Statistical Methods, Mc. Graw-Hill, New York, 1967 .

10. Morrison, D.F., Multivariate Statistical Methods, Mc. Graw-Hill, New York, 1967 .

11. W.J. Krzanowski, Principles of Multivariate Analysis, Oxford University Press, 1988 .