تأثير فقدان إحدى مشاهدات المتغير المستقل وموقعها في تحليل الانحدار البسيط

الباحث علي درب كسار الحيالي الباحثة عفاف صالح الحاني
كلية الزراعة / قسم الاقتصاد الزراعي- جامعة بغداد

المستخلص

يهدف البحث إلى تسليط الضوء على أهم معالجات مشكلة فقدان بيانات السلسلة الزمنية وتأثيرها في تحليل الانحدار الخطي البسيط. ويتناول البحث تأثير فقدان المشاهدات في المتغير المستقل فقط دون المتغير التابع. ويتم ذلك من خلال افتراض فقدان إحدى المشاهدات من السلسلة الزمنية المتكاملة أصلا ومن ثم اختبار اثر هذا الفقدان على تحليل الانحدار البسيط لبيانات تجربة تتعلق بأثر كميات العليقة المستهلكة على وزن الدجاج ولمدة 15 أسبوعا. وتوصل البحث إلى إن فقدان مشاهدة واحدة لم يكن له ذلك التأثير الواضح إذ كان الأنموذج المقدر بعد فقدان مشاهدة متماسكا في معاملاته ومعنويا ضمن المستويات الإحصائية المقبولة، فضلا عن إن المنهجية المتبعة أعطت قيمة اكبر للمشاهدة المفقودة من القيمة الأصلية عندما كان الفقد في المشاهدات الواقعة في منتصف ونهاية السلسلة الزمنية في حين كانت الإشارة سالبة أي القيمة المتوقعة اقل من القيمة الأصلية عندما وقع الفقدان في بداية السلسلة الزمنية. وكل ذلك سيؤثر في القيم المتنبأ بها خارج السلسلة الزمنية تبعا للقيمة المستخرجة للمشاهدة المفقودة. وأوصى البحث بضرورة أن يتم البحث في تحليل اثر فقدان أكثر من مشاهدة فضلا عن دراسة الحالة عندما يكون الفقد في المتغير التابع فقط وفي المتغيرين التابع والمستقل معا.

ABSTRACT

The objective of the research , is to shed light on the most important treatment of the problem of missing values of time series data and its influence in simple linear regression. This research deals with the effect of the missing values in independent variable only. This was carried out by proposing missing value from time series data which is complete originally and testing the influence of the missing value on simple regression analysis of data of an experiment related with the effect of the quantity of consumed ration on broilers weight for 15 weeks. The results showed that the missing value had not a significant effect as the estimated model after missing value was consistent and significant statistically. The results also showed that the estimated missing value was larger than the original value when the missing value situated either in the middle or at the end of the series while the sign was negative or the estimated value was less than the original value when the missing value situated in the beginning of the time series. All of that would affect the estimated values outside the time series data according to estimated value of missing value. The research recommended to work on the analysis of the effect of missing more than one value and also when the missing is in the dependent variable only and in both dependent and independent variables.

مقدمة

يعاني كثير من الباحثين من مشكلة فقدان بعض المشاهدات لبعض المتغيرات قيد البحث لاسيما عندما تكون تلك المشاهدات لبيانات سلسلة زمنية. وتعد هذه القيم المفقودة عقبة تواجه تطبيق أسلوب إحصائي معين أو الوصول إلى نتائج معينة. فإذا أريد تقدير العلاقة بين متغير معتمد (Y) ومتغير أو مجموعة متغيرات مستقلة (X,s) تصبح هذه المشكلة عائقا عند القيام بتحليل الانحدار. وتأخذ المشكلة اهتماما اكبر عندما يكون الفقد في جانبي الانحدار أي في المتغير التابع والمتغيرات المستقلة أو بعضها. وتخف حدة المشكلة عندما يفتقد احد جانبي الانحدار إحدى المشاهدات. ويظهر تأثير القيم المفقودة بشكل واضح عند القيام بعملية التنبؤ في السلاسل الزمنية لاسيما المستقرة منها.

مشكلة البحث

انطلق البحث من مشكلة البيانات المفقودة في السلاسل الزمنية وتأثير ذلك في التحليل سواء كان رياضيا أم قياسيا فضلا عن تأثيره في التنبؤ.

فرضية البحث

يفترض البحث إن فقدان إحدى مشاهدات المتغير المستقل في تحليل الانحدار البسيط لن يكون تأثيره معنويا وكذلك الحال فيما يتعلق بموقع المشاهدة المفقودة في بيانات السلسلة الزمنية وتأثيرها المعنوي في التنبؤ.

هدف البحث

يهدف البحث إلى تحليل الانحدار الخطي البسيط بعد فقدان إحدى مشاهدات المتغير المستقل فقط واختبار تأثير ذلك الفقدان، فضلا عن دراسة تأثير موقع فقدان المشاهدة على التحليل.

الدراسات والبحوث السابقة

لقد تصدت الكثير من البحوث والدراسات إلى هذه المشكلة وعالجتها من نواح عديدة. فمن المعالجات، تلك التي تناولت ما يسمى (طرائق الرتبة الأولى) (First order methods) ، والتي اعتمدت على تقدير البيانات المفقودة بمجموعة طرائق مختلفة للانحدار ومن ثم تقدير تلك المعاملات عند اكتمال المشاهدات بواسطة طريقة المربعات الصغرى (10).

أشار (Kayaalp) إلى إن تأثير الفقد في بيانات السلسلة الزمنية سيؤثر في بعض الخصائص الواجب توافرها في معلمات الأنموذج المقدر بواسطة طريقة المربعات الصغرى (9).

قام (المحيميد) بتقدير الدرجات المفقودة في الاختبارات التي لا يتسنى للطالب تقديمها نتيجة لعذر مبرر، حيث تعتمد على تقدير أداءه بناء على متوسط القيمة المعيارية لدرجاته في الاختبارات الأولى، ومن ثم تقدر درجة الطالب في الاختبار المفقودة درجته آخذا في الاعتبار متوسط درجات باقي الطلبة وانحرافه المعياري في هذا الاختبار(1).

أما (Pascal) فقد توصل إلى إن تأثير الفقد في بيانات السلسلة الزمنية سينسحب أثره على التنبؤ بهذه البيانات، وان درجة الدقة في التنبؤ ستتأثر كلما كان عدد المشاهدات المفقودة اكبر (12).

تناول كل من (Valicer and Suzanna) مقارنة مجموعة من الطرق لتقدير المشاهدات المفقودة في تحليل (ARIMA) (Auto regressive integrated moving average)للسلاسل الزمنية، وتوصل البحث إلى إن طريقة الإمكان الأعظم (Maximum Likelihood) قد أعطت أفضل التقديرات للبيانات المفقودة(13).

أما (Honaker) فقد قدم محاولة بناء أنموذج يسمى بتمهيد الاتجاه الزمني فضلا عن التحولات في البيانات المقطعية وصولا إلى دقة أكثر في البيانات المكونة من السلاسل الزمنية والبيانات المقطعية وان بناء هذا الأنموذج وتطويره أسهم في زيادة تماسك البيانات وعدم تفكيكها أو كسرها (7).

أما (Chi fung) فقد قام باستعراض الطرق المستعملة لتقدير المشاهدات المفقودة في بيانات السلسلة الزمنية ومن ثم اختبار تلك الطرق بمجموعة اختبارات (4).

استعرض (David) مجموعة الطرائق التي تعالج الفقد في المشاهدات لبيانات السلسلة الزمنية، وأشار إلى إن أول طريقة تم استخدامها هي طريقة المتغيرات الوهمية غير أنها لم تعتمد كثيرا بسبب المشاكل التي يسببها فقدان بعض المشاهدات (5).

أما (Ahmed) فقد اعتمد طريقة معينة في تقدير البيانات المفقودة تسمى طريقة الترشيح(Faltering process) باستخدام منهجية (بوكس وجينكنز) (B&J) (3).

درس (Gevorgayan) مشكلة البيانات المفقودة باستعمال طريقتين تعتمد الأولى على مايسمى مدى التمثيل(Space representation) أما الثانية فتعتمد على القياسات المترابطة المتسلسلة (Sequencing correlated measures) وتم اختبار كلا الطريقتين على البيانات المتولدة(6).

درس (Mohsen) تقدير وتوليد البيانات المفقودة في السلاسل الزمنية المستقرة من خلال إيجاد تنبؤات متعددة باستعمال الحزم الإحصائية للكومبيوتر فضلا عن إمكانية استخدام بعض الحلول لإيجاد القيم المفقودة في تطوير منهجية المعادلات الآنية لتقدير المعاملات وكذلك القيم المفقودة في السلاسل الزمنية (11).

تناول (Kaiser) تقدير القيم المفقودة بواسطة ما يسمى النقاط المتنبأ بها والتي اعتمدت على افتراض إن القيم المفقودة هي عشوائية، كما إن نوعية النقاط المتنبأ بها كمقدرات للقيم المفقودة تعتمد على عدد التنبؤات في معادلة الانحدار ودرجة ترابطها (8).

الإطار النظري

إن القيام بتقدير القيم المفقودة في السلاسل الزمنية يستلزم معرفة وجود هذا الفقدان هل هو في جانب المتغير المستقل أم التابع، فضلا عن ذلك، هل يشمل التحليل الانحدار المتعدد أم البسيط، فلكل حالة لها معالجاتها. وسيتناول البحث تقدير القيم المفقودة في جانب المتغير المستقل على افتراض إن البيانات المعتمدة متكاملة العدد وسيتم افتراض إن إحدى المشاهدات قد فقدت ومن خلالها سيتم عرض تأثير فقدان هذه المشاهدة على تحليل الانحدار، والذي سيتم اعتماده هنا وهو تحليل الانحدار البسيط المعتمد على وجود متغيرين احدهما تابع والآخر مستقل. وقبل تطبيق ذلك لابد من عرض المنهجية التي سوف تستند إليها طريقة التقدير (9).

سنفترض هنا إن لدينا عينة ذات حجم N من $a(P+1)$ التوزيعات المتعددة وكما يأتي:-

$$f(Y/X_1,X_2,.....,X_p)\sim N(\beta_0+\beta_1X_1+....+\beta_pX_p,\sigma^2e)$$

وقد أشارت طريقة الحل لغرض تقدير المعلمات $(b_1,b_2,....b_n)$ بمنهجية الإمكان الأعظم (Maximum Likelihood) في حالة الانحدار الاعتيادي إلى إن جميع المتغيرات المستقلة تعد ثابتة في حالة عدم وجود فقد في البيانات ويمكن أن يعبر عنها كما يأتي:-

$$\hat{\beta}_j=\sum_{k=1}^{p}\left(\hat{\sigma}_{jk}\right)^{-1}\hat{\sigma}_{k0}\qquad (j\neq k).....(1)$$

$$\hat{\beta}_0=\overline{Y}-\beta_j\overline{X}_j$$

$$\hat{\sigma}_{k0}=\sum\left(X_{ij}-\overline{X}_j\right)\left(Y_i-\overline{Y}\right)/N-1$$

$$\hat{\sigma}_{jk}=\sum\left(X_{ij}-\overline{X}_j\right)\left(X_{ik}-\overline{X}_k\right)/N-1.....(2)$$

$$\left(\hat{\sigma}_{jk}\right)^{-1}=1/\hat{\sigma}_{jk}\qquad (j,k\neq 0)$$

X_{ij} = تمثل قيمة X_j لكل قيمة مفردة من i.

$$\overline{X}_j=\sum X_{ij}/N\quad ,\quad \overline{Y}=\sum Y_i/N$$

سيتم الآن عرض جدول تحليل التباين في حالة عدم فقدان بعض المشاهدات وفي حالة فقدانها في المتغير المستقل فقط (9).

جدول تحليل التباين في حالة عدم وجود فقدان في المشاهدات

s.o.v	*D.F*	S.S
Regression	***P***	$\sum S_{xiy}\beta_i[2]$
Residual	***n-P-1***	$[1]-[2]$
Total	***n-1***	$S_{yy}[1]$

***P*:عدد المتغيرات المستقلة**

***n*:عدد المشاهدات**

جدول تحليل التباين في حالة وجود فقدان في المشاهدات

s.o.v	*D.F*	S.S
Regression	***P***	$Wij\sum S_{xjy}\beta_j[2]$
Residual	***n-1-k-P***	$[1]-[2]$
Total	***n-1-k***	$S_{yy}[1]$

***k*:عدد المشاهدات المفقودة**

Wij **دالة المؤشر العشوائي: (Random indicator function)**

تحليل الجانب التطبيقي

لغرض معرفة تأثير فقدان إحدى المشاهدات سنأخذ بيانات التجربة الآتية والتي تمثل تأثير كمية العليقة المستهلكة (X) على وزن الدجاجة (Y) لمدة 15 أسبوعا.

جدول1. الكميات المستهلكة من العليقة (X) (باوند) ووزن الدجاج (Y) (باوند) (2)

(Y)	0.13	0.16	0.43	0.61	0.90	1.29	1.66	2.04	2.46	2.89	2.31	3.31	4.03	4.28	4.48
(X)	0.29	0.58	0.97	1.40	2.10	3.10	4.09	5.19	6.50	7.98	9.59	9.59	13.16	14.96	16.92

ولقد تم تقدير معاملات الأنموذج في حالتين:-

الحالة الأولى:- عدم وجود فقدان في البيانات.

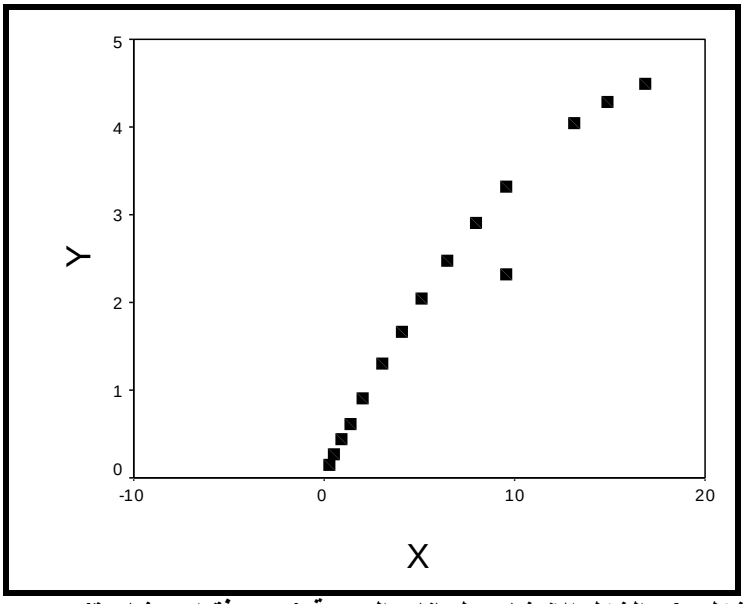
من خلال القيام بتحليل الانحدار البسيط تم الحصول على المعاملات الآتية:-

$$Y = b_0 + b_1 X$$

$$Y = 0.354 + 0.267X$$

وأدناه جدول تحليل التباين للأنموذج المقدر في الحالة الأولى

S.V	*D.F*	S.S	*M.S*	*F*	*Sig*
Regression	**1**	**29.605**	**29.605**	**283.87**	**0.00**
Residual	**13**	**1.356**	**0.104**		
Total	**14**	**30.961**			



شكل -1- الشكل الانتشاري لبيانات التجربة (بدون فقدان مشاهدة)
X= كمية العليقة المستهلكة Y= وزن الدجاجة

الحالة الثانية: حالة فقدان في إحدى مشاهدات المتغير المستقل.

سنفترض هنا إن المشاهدة رقم 13 في المتغير المستقل قد فقدت وعليه سيتم تقدير أنموذج الانحدار للبيانات الآتية:-

جدول2. الكميات المستهلكة من العليقة (X) (باوند) ووزن الدجاج (Y) (باوند) (فقدان مشاهدة)

(Y)	0.13	0.16	0.43	0.61	0.90	1.29	1.66	2.04	2.46	2.89	2.31	3.31	4.03	4.28	4.48
(X)	0.29	0.58	0.97	1.40	2.10	3.10	4.09	5.19	6.50	7.98	9.59	9.59	*	14.96	16.92
W_{ij}	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1

*** مشاهدة مفقودة**

W_{ij}=1 في حالة X_{ij} موجودة أو (ملاحظة) (وجود مشاهدة لكل مفردات X)

W_{ij}=0 في حالة X_{ij} غير موجودة (فقدان مشاهدة من مشاهدات X)

$N_j = \sum W_{ij}$ تمثل عدد المفردات لكل X_j (عند وجود قيمة لـ X)

$N_{jk} = \sum W_{ij} W_{ik}$ تمثل عدد المفردات لكل من X_j و X_k الملاحظتين

$\bar{X}_{j(j)} = \sum W_{ij} X_{ij} / N_j$ تمثل المتوسط الحسابي لقيم X المعتمدة على القيم المثبتة أو الملاحظة لقيم X عندما تكون كلا من X_j و X_k موجودتين(ملاحظتين).

$\bar{Y}_{(j)} = \sum W_{ij} Y_i / N_j$ يمثل المتوسط الحسابي لقيم Y المقابلة لقيم X الموجودة (غير المفقودة).

من هنا سيظهر لنا انه من المعقول أن يتم تطبيق نفس التقديرات السابقة باستخدام جميع البيانات المتاحة.

أما:-

$$\hat{\sigma}_{jk} = \sum W_{ij} W_{ik} \left(X_{ij} - \bar{X}_{j\,(jk)}\right)\left(X_{ik} - \bar{X}_{k\,(jk)}\right)/(N_{jk} - 1) \qquad (j,k \neq 0)$$

$$\hat{\sigma}_{j0} = \sum W_{ij} \left(X_{ij} - \bar{X}_{j(j)}\right)\left(Y_i - \bar{Y}_{(j)}\right)/ N_j - 1$$

ومن هنا فان تقديراتنا يمكن أن تكتب على النحو الآتي:-

$$\hat{\beta}_j = \sum \left(\hat{\sigma}_{jk}\right)^{-1} \hat{\sigma}_{k0} \qquad (j \neq 0)$$

$$\hat{\beta}_0 = \bar{Y} - \sum \beta_j \bar{X}_{j(j)} \text{.....(3)}$$

$$N_1 = \sum W_{i1} = 14$$

$$\bar{X}_{1(1)} = \sum W_{i1} X_{i1} / N_1 = \left[(1)(0.29) + ... + (1)(16.92)\right]/14 = 5.9471$$

$$\bar{Y}_{(1)} = \sum W_{i1} Y_i / N_1 = \left[(1)(0.13 + ... + (1)(4.48)\right]/14 = 1.9321$$

$$\hat{\sigma}_{11} = \sum W_{i1} W_{i1} (X_{i1} - \bar{X}_{1(1)})^2 /(N_1 - 1)$$

$$= \left[(1)(1)(0.29 - 5.9471) + ... + (1)(1)(16.92 - 5.9471)^2\right]/14 - 1 = 26.655$$

$$\hat{\sigma}_{10} = \sum W_{i1} \left(X_{i1} - \bar{X}_{1(1)}\right)\left(Y_i - \bar{Y}_{(1)}\right)/(N_1 - 1)$$

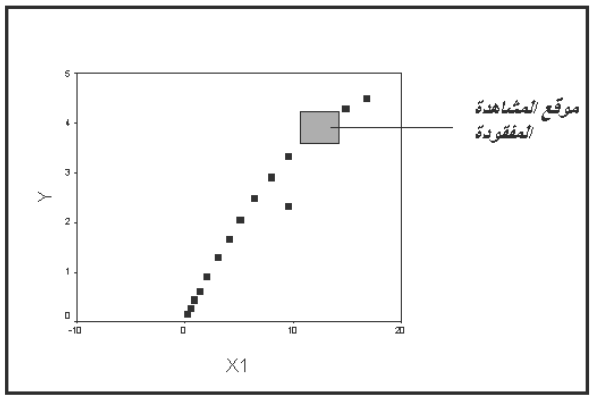
$$= \left[(-5.7)(-1.8) + ... + (10.97)(2.5)\right]/13 = 7.36$$

$$\hat{\beta}_1 = \sum \left(\hat{\sigma}_{11}\right)^{-1} \hat{\sigma}_{10} = (1/26.655)(7.36) = 0.276$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}_{1(1)} = 1.9321 - (0.276)(5.9471) = 0.2921$$

وأدناه جدول تحليل التباين للأنموذج المقدر في الحالة الثانية(فقدان مشاهدة)

S.V	*D.F*	S.S	*M.S*	*F*	*Sig*
Regression	**1**	**25.528**	**25.528**	**231.50**	**0.00**
Residual	**12**	**1.325**	**0.110**		
Total	**13**	**26.853**			



شكل-2- الشكل الانتشاري لبيانات التجربة عند فقدان مشاهدة

وبعد القيام بتحليل الانحدار بمشاهدة مفقودة بقي أن نعرف إن قيمة المشاهدة المتوقعة ستكون

$$X_{(13)} = 14.495$$

ولغرض معرفة تأثير موقع فقدان المشاهدة، تم افتراض إن المشاهدة ذات التسلسل (2) في بداية السلسلة الزمنية (0.58) قد فقدت، وتم الحصول على النتائج الآتية بإتباع نفس المنهجية عند افتراض فقدان المشاهدة رقم 13.

$$\hat{\sigma}_{11} = 29.055 \ , \quad \hat{\sigma}_{10} = 7.41$$

$$\hat{\beta}_1 = 0.255$$

$$\hat{\beta}_0 = 0.45$$

$$Y = 0.45 + 0.255X$$

وأدناه جدول تحليل التباين للأنموذج المقدر بعد فقدان المشاهدة الثانية (0.58)

S.V	*D.F*	S.S	*M.S*	*F*	*Sig*
Regression	**1**	**26.160**	**26.160**	**244.677**	**0.000**
Residual	**12**	**1.283**	**0.107**		
Total	**13**	**27.443**			

وبعد تقدير القيمة المتوقعة للمشاهدة المفقودة الثانية كانت القيمة =1.137- أي إن الإشارة سالبة وهي تبتعد كثيرا عن القيمة الأصلية (0.58) وهذا ما يؤكد أهمية موقع المشاهدة المفقودة عند تطبيق هذه المنهجية.

أما في حالة افتراض فقدان المشاهدة الثامنة (5.19) فقد جاءت النتائج معنوية لمعلمات الأنموذج وكذلك للأنموذج المقدر وكانت القيمة المتوقعة للمشاهدة المفقودة =6.41.

الاستنتاجات والتوصيات

يلاحظ من النتائج إن جميع المقدرات هي مقدرات متماسكة وغير متحيزة ويمكن ملاحظة ذلك بسبب إن توليفات أو أزواج قيم المتغير المستقل $(X_{ij}\ ,X_{ik})$ هي من مجتمع محدود ويتحقق هذا الشرط طالما أنها اختيرت بشكل مستقل. كما إن التباين σ_{jk} هي مقدر متماسك لـ σ_{jk} ويبقى هذا صحيحا طالما إن التباينات تدنو من الصفر. واعتمادا على ذلك فان الاستنتاج الذي تم التوصل إليه فيما يخص σ_{jk} سيكون منطبقا على σ_{k0}. وكذلك الحال بالنسبة إلى $X_{j(j)}$ فهي مقدرات ثابتة لـ X_j. وتعتمد كفاءة هذه الطريقة على الارتباط بين المتغيرات المستقلة وكذلك على معدل المشاهدات المفقودة.كما يلاحظ إن الأنموذج المقدر بعد فقدان المشاهدة 13 والمشاهدة الثانية (0.58) والمشاهدة الثامنة (5.19) لم تتأثر معنويته كثيرا وظلت معلماته متماسكة وقد يعزى سبب ذلك إلى إن فقدان مشاهدة واحدة لم يؤثر بشكل واضح. أما فيما يتعلق بتأثير موقع فقدان المشاهدة فقد تبين انه عندما يكون موقع المشاهدة المفقودة في بداية السلسلة الزمنية كانت القيمة المتوقعة سالبة في حين إن القيم كانت موجبة في حالة وقوع المشاهدة المفقودة في المنتصف أو في نهاية السلسلة الزمنية. ويستتبع ذلك بالضرورة أن تختلف القيم المتنبأ بها خارج السلسلة الزمنية لاسيما إذا تم التقدير بطريقة الاتجاه العام والتي تتأثر بشكل واضح بالمسار الزمني لقيم المشاهدات. فضلا عن ذلك يلعب طول السلسلة الزمنية دورا في تقدير المشاهدات المفقودة في السلسلة الزمنية.

مما تقدم يوصي البحث بدراسة تأثير فقدان أكثر من مشاهدة على تحليل الانحدار فضلا عن القيام بتحليل الانحدار عندما يكون الفقد في جانبي الانحدار، أي في جانب المتغير التابع والمستقل معا. أو دراسة تأثير فقدان مشاهدة أو أكثر في المتغير التابع فقط. وكذلك محاولة معرفة تأثير فقدان المشاهدات في حالة كونها كانت عشوائية أم غير عشوائية.

المصادر

1. محمد عبد الهادي المحيميد. 1990. تقدير الدرجات المفقودة في الاختبارات. جامعة الكويت. مجلة العلوم الاقتصادية والإدارية. العدد15. ص ص 75-94.
2. هاشم علوان السامرائي. 1982. إدارة الأعمال المزرعية. مطابع دار السياسة. الكويت. ص36.
3. Ahmed,M.R. and A.M.H. Al-Khazaleh.2008. Estimation of missing data by using the filtering process in time series modeling. www.yahoo.com
4. Chi fun, D.sh.2006.Motheds of the estimation of missing values in thme series.M.Sc.thesis, Dept. of Mathematics and planning., Faculty of communications, Health and Science.,Univ.of Edith Cowan,.www.google.com
5. David, C.H.2008.Treatment of missing data. www.uvm.edu
6. Gevorgyan.R, and N.Melikyan.2008.Missing data problem and the empirical yield curve analysis. An example of T-bills market in Armenia. www.google.com
7. Honaker,J.,and G.king.2006. What to do about missing values in time series cross-sectin data. http://gking.harvard.edu/Amelia
8. Kaiser,J.and D.B Tracy.2008. Estimation of missing values by predicted score. www.google.com
9. Kayaalp,G.T.,1999.Linear regression analysis with missing observation among the independent variables in Animal breeding.Tr.J.of Veterinary and Animal Sciences. www.google.com
10. Kelejian,H.H.1990. Missing observations in multivariate regression:efficiency of A first order method. www.google.com
11. Mohsen,P.2008.Estimation and interpolation of missing values of a stationary time series. Journal of time series analysis. 10(2):149-196. www.google.com
12. Pascal, B.2005.Influence of missing values on the prediction of stationary time series. Journal of time series analysis.26(4):519-525. www.ssrn.com
13. Valicer.W.F. and C.M.Suzanne.2005. A comparison of missing-data procedures for Arima time series analysis. Educational and pcychological measurement, 65(4):596-615. www.google.com.