المعهد العربي للتدريب والبحوث الإحصائية
ورشة عمل ضمن التعليم عن بعد
في

الاساليب الاحصائية في معالجة القيم المفقودة في التعداد والمسوح

د. حسان أبو حسان

١٣ــــ١٥ نيسان \ابريل ٢٠٢٠

الاساليب الاحصائية في معالجة القيم المفقودة

إعداد

د. حسان أبو حسان

دكتوراه إحصاء: جامعة Southern Illinois University 2003-2007

ماجستير اقتصاد قياسي: Southern Illinois University 2007-2008

ماجستير رياضيات: الجامعة الأردنية ١٩٨٩-١٩٩١

خبير العينات في الجهاز المركزي للإحصاء الفلسطيني: ١٩٩٤ – ٢٠٠١

مدير برنامج ماجستير الإحصاء التطبيقي وعلم البيانات: ٢٠٠٨ - تاريخه

جامعة بير زيت – فلسطين

hmhassan@birzeit.edu, qaishassan@yahoo.com

Cell phone: 0097059983004

Tel: 0097022962069

المحاضرة الأولى

مصادر (أنواع) البيانات المفققودة

- Non-Responseعدم الاستجابة
  ١. Unit Non-Response عدم الاستجابة الكلية
  ٢. Item Non-Responseعدم الاستجابة الجزئية

- Dropoutالانسحاب
  - عادة يحصل في الدراسات الطولية (الزمنية)
- Researcherالباحث
- عدم رغبة أو قدرة الحكومات على توفير إحصاءات ذات طابع حساس لسنة أو أكثر

أهمية معالجة القيم المفقودة

- تحيز التقديرات المبنية على الحالات المكتملة.

التحيز يعتمد على:

١. نسبة عدم الاستجابة

٢. الاختلاف بين الوسطين الحسابيين (أو التوزيعين) في كل من مجتمع المستجيبين ومجتمع غير المستجيبين

- صعوبة تحديد مدى اختلاف التوزيع الاحتمالي للمتغيرات بين كل من المستجيبين وغير المستجيبين
  - غالباً لا تتوفر معلومات عن غير المستجيبين

أهمية معالجة القيم المفقودة-٢

- ضرورة فحص سبب عدم الاستجابة
  - هل آلية وجود عدم الاستجابة ترتبط بالمتغير قيد التحليل analysis variable؟
  - إطار العوامل المسببة لعدم الاستجابة
    ١. الباحث (جامع البيانات)
    ٢. المبحوث (المستجوب)
    ٣. محتوى المسح (الاستبانة، آليات العمل الميداني)

أهمية معالجة القيم المفقودة-٣

- نقص حجم العينة يقلل دقة التقديرات
  - نسبة استجابة منخفضة ⟵ حجم عينة منخفض ⟵ تباين عال
  - زيادة حجم العينة لا يحل مشكلة التحيز (مثال PECS)
  - حل مشكلة نقص حجم العينة أسهل من حل مشكلة التحيز (مثال:لو أردنا حجم عينة ١٠٠٠، وتوقعنا نسبة ١٠% عدم استجابة، يمكن اختيار عينة أكبر من ١٠٠٠ بحيث تصبح بعد حذف عدم الاستجابة ١٠٠٠ كما يلي:)

$$n_d = \frac{n_t}{1 - NR} = \frac{1000}{1 - 0.1} = 1111$$

أمثلة

- **مثال١**: في المسوح الأسرية والتي يكون الفرد فيها هو وحدة الدراسة، قد يكون هناك عدم توازن في نسبة الجنس (نسبة المستجيبات من النساء أعلى من نسبة المستجيبين من الرجال، بسبب عدم تواجد كثير من الرجال لحظة الزيارة)
- ينتج عن ذلك تحيز في تقديرات المجاميع Totals في المجتمع لأي متغير، مثل دخل الفرد، يكون نصيب الرجال فيه أعلى من نصيب النساء
- هناك طرق تقدير من شأنها تقليل هذا التحيز (لاحقاً)

أمثلة - ٢

- مثال٢: في المسوح الأسرية، تكون عادة نسبة الاستجابة من الأسر صغيرة الحجم أقل من نسبة الاستجابة من بقية الأسر
- ينتج عن ذلك عينة "زائدة" oversampling في الأسر متوسطة وكبيرة الحجم نسبة إلى الأسر صغيرة الحجم
- ينتج عن ذلك تيحز في تقدير متوسط حجم الأسرة في المجتمع
- قد ينتج أيضاً تحيزاً في تقديرات لمتغيرات مرتبطة بحجم الأسرة مثل متوسط إنفاق الأسرة على مجموعات افاق معينة (مثل: الخبز،.....)

أمثلة - ٣

- في المسوح الاقتصادية: لنفرض أن المنشآت غير المستجيبة هي في الأغلب من المنشآت الكبيرة
- عندئذ سيكون هناك نحيز كبير في التقديرات إذا لم يتم إجراء معالجة جيدة للقيم المفقودة

أنواع البيانات المفقودة
Missing Completely At Random

- MCAR مفقودة بشكل عشوائي تام:
- العينة المستجيبة هي عبارة عن عينة عشوائية جزئية من العينة المختارة (عادةً هذا غير صحيح)
- ميل المبحوث للاستجابة ليس مرتبطاً ب:
  ١. بيانات معروفة عن المبحوث Covariates (X)
  ٢. متغير الدراسة Response Variable (Y)
  - مثال: عندما تفقد استبانة

أنواع البيانات المفقودة- ٢

- ينتج عن ذلك:

١. لو اخترنا عينة عشوائية بحجم n فإن المجموعة المستجيبة تعتبر عينة عشوائية من المجتمع بحجم n_R

- يعتبر $\overline{Y}_R$ (متوسط العينة للأفراد المستجيبين) تقديراً غير متحيزٍ ل $\overline{Y}_U$ (متوسط المجتمع)

Missing At Random given covariates

- MAR بيانات مفقودة بشكل عشوائي مشروط
- ميل المبحوث للاستجابة :

١.يرتبط ببيانات معروفة عن المبحوث Covariates (X)

٢.لا يرتبط بمتغير الدراسة Response Variable (Y)

– مثال: X = جنس المبحوث في مسح القوى العاملة ؟

X = حجم الأسرة في مسح ميزانية الأسرة

- طرق التعديل تعتمد على نماذج لعدم الاستجابة Non-response models (لاحقاً)

بيانات مفقودة بشكل غير عشوائي
Missing Not At Random

- ميل المبحوث لعدم الاستجابة:
  ١. يرتبط ب Y
  ٢. لا يمكن تفسيره بشكل كامل من خلال عوامل أخرى X
- مثال: عند سؤال المبحوث عن دخله، من المرجح أن تكون قيمة الدخل مفقودة إذا كان الدخل مرتفعاً
- عندما تكون البيانات مفقودة بشكل غير عشوائي MNAR فإنه تكون قد فقدت بيانات مهمة، ولا يوجد طريقة متفق عليها لمعالجة القيم المفقودة بشكل جيد.

طرق واستراتيجيات معالجة البيانات المفقودة

استراتيجية ١: الوقاية خير من العلاج

- ضرورة مراعاة عبئ المستجوب عند تصميم المسح
- محتوى المسح: تجنب الأسئلة ذات الحساسية لدى المبحوث
- التوقيت: في المسوح الزراعية تجنب أوقات الذروة
- التوقيت: الأعياد والعطل الرسمية ترتبط بارتفاع نسبة عدم الاستجابة
- الباحثين: تدريب، موظفين لإقناع الرافضين من المبحوثين
- طرق وآليات جمع البيانات: البريد\الفاكس\الانترنت لها أعلى نسبة عدم استجابة، يليها التلفون، تليها المقابلة الشخصية

طرق واستراتيجيات معالجة البيانات المفقودة

استراتيجية ١: الوقاية خير من العلاج

- CAPI يقلل من item NR الناتج عن خطأ الباحث
- تصميم الاستبانة: تقليل العبئ على المبحوث- استبانة قصيرة نسبياً، استخدام مفاهيم بسيطة وغير مركبة، تسلسل منطقي وسهل للأسئلة.
- تصميم العينة: تقليل عبئ المبحوث، استخدام تصميم وطرق تقدير تزيد الدقة بدون زيادة حجم العينة، مثل استخدام الطبقات، واستخدام مقدرات النسبة ومعادلة الانحدار
- متابعة عدم المستجيبين
- محفزات

استراتيجية٢: إعادة المحاولة مع عينة من غير المستجيبين
Call-backs and double sampling

- **الخطوات:**

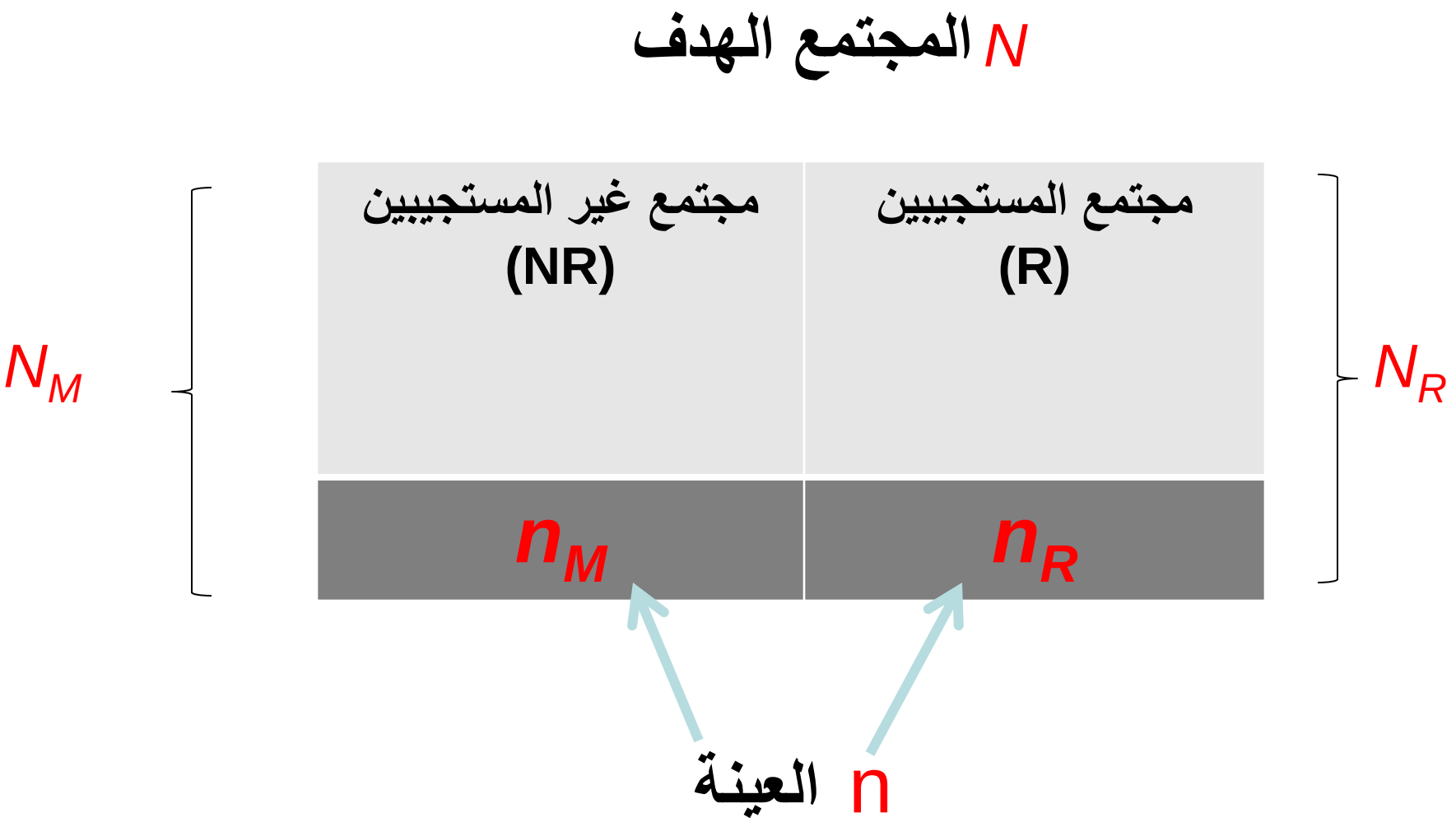
١. يتم اختيار عينة جزئية من غير المستجيبين

٢. يتم جمع البيانات من هذه العينة

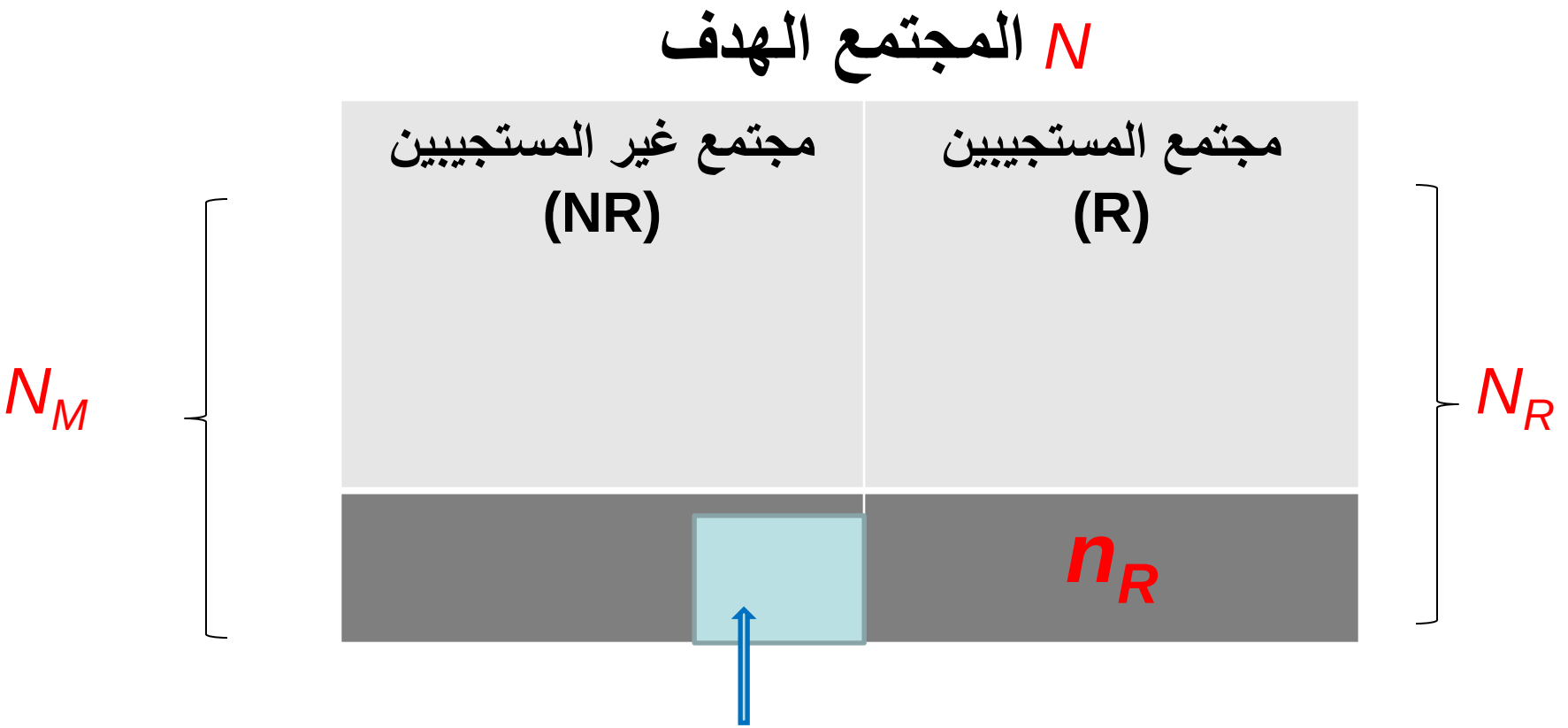
٣. يتم استخدام هذه البيانات لتقدير متوسط مجتمع غير المستجيبين

- **يعتبر هذا التصميم كمثال على تصميم "double" or "2-phase" sampling**
- **سوف نستخدم الفرضية (غير الواقعية) بأن جميع عينة غير المستجيبين سوف تزودنا بالبيانات المطلوبة عند إعادة المحاولة.**

Framework


المجتمع الهدف N
مجتمع غير المستجيبين
(NR)
مجتمع المستجيبين
(R)
N_M
N_R
n_M
n_R
العينة n

Subsample the non-responding portion of population



Sample 100 v % of the non-responding part of sample

$=n_{MCB} = v\ n_M$

حساب التقديراتEstimation

- متوسط العينة من مجتمع المستجيبين:

$$\bar{y}_R = \frac{1}{n_R} \sum_{i=1}^{n_R} y_i$$

- متوسط العينة من غير المستجيبين الذين تم إعادة مقابلتهم

$$\bar{y}_M = \frac{1}{n_{MCB}} \sum_{i=1}^{n_{MCB}} y_i$$

Estimation - 2حساب التقديرات

- تقدير متوسط المجتمع Population Mean

$$\hat{\bar{y}} = \hat{\mu} = \frac{n_R}{n}\bar{y}_R + \frac{n_M}{n}\bar{y}_M$$

- تقدير المجاميع في المجتمع Population Totals

$$\hat{t} = N\hat{\bar{y}} \ldots = N\frac{n_R}{n}\bar{y}_R + N\frac{n_M}{n}\bar{y}_M$$

$$\ldots\ldots\ldots\ldots = N\frac{n_R}{n}\frac{1}{n_R}\sum_{i\in R} y_i + N\frac{n_M}{n}\frac{1}{n_{MCB}}\sum_{i\in MCB} y_i$$

$$\ldots\ldots\ldots\ldots = \frac{N}{n}\sum_{i\in R} y_i + \frac{N}{n}\frac{1}{v}\sum_{i\in MCB} y_i$$

Estimation - 3 حساب التقديرات

- حساب الأوزان
  - أوزان المستجيبين في العينة الأصلية $W_i = \frac{N}{n}$
  - أوزان المستجيبين في إعادة المقابلة Call-Backs $W_i = \frac{N}{n}\frac{1}{v}$
- حساب التباين لتقدير المتوسط $\hat{\bar{y}}$

$$\hat{V}(\hat{\bar{y}}) = \frac{n_R - 1}{n-1}\frac{S_R^2}{n} + \frac{n_M - 1}{n-1}\frac{S_M^2}{vn} + \frac{1}{n-1}\left[\frac{n_R}{n}(\bar{y}_R - \hat{\bar{y}}) + \frac{n_M}{n}(\bar{y}_M - \hat{\bar{y}})\right]$$

استراتيجية ٣: تعديل الأوزان لعدم الاستجابة

- الطريقة
  - تعديل الأوزان على مستوى الطبقة
  - الأوزان (أوزان التصميم) هي معكوس احتمال الاختيار

$$W_i = 1/\pi_i$$

  - تفسير الأوزان
    - عدد وحدات المجتمع التي تمثلها الوحدة i في العينة
  - الاحتمال الثاني = احتمال الاستجابة للوحدة i

$$\phi_i = pr\{R_i = 1\}$$

تعديل الأوزان لعدم الاستجابة

- وزن الوحدة *i*

$$\tilde{W}_i = \frac{1}{\pi_i}\frac{1}{\phi_i} = \frac{1}{\pi_i \phi_i}$$

  - التفسير
    - عدد الوحدات في المجتمع التي تمثلها الوحدة المستجيبة *i*
  - بافتراض أن البيانات مفقودة بشكل عشوائي مشروط MAR
  - يتم إنشاء مجموعة من "فئات التوزين" بحيث يمكننا افتراض أن احتمال الاستجابة ثابت ضمن كل فئة.
  - يتم تقدير احتمال الاستجابة ضمن كل فئة باستخدام أوزان التصميم كما يلي:

$$\hat{\phi}_i = \frac{sum\,of\,weights\,for\,respondents\,\,belong\,to\,class\,c}{sum\,of\,weights\,for\,selected\,units\,belong\,to\,class\,c}$$

تعديل الأوزان لعدم الاستجابة - ٢

- الأوزان المعدلة لعينة المستجيبين

$$\tilde{w}_i = \frac{1}{\pi_i \hat{\phi}_i}$$

- تقدير المجاميع t_U والمتوسطات $\bar{y}_U$ في المجتمع

$$\hat{t}_{WC} = \sum_{i \in responding\ sample} \tilde{w}_i y_i$$

$$\hat{\bar{y}}_{wc} = \frac{\hat{t}_{wc}}{\sum_{i \in responding \quad sample} \tilde{w}_i}$$

استراتيجية ٤: الإسناد (التعويض)

- البيانات المفقودة على مستوى البند (السؤال) أمر مؤلوف في المسوح الإحصائية (لكنه طبعاً غير محبب)
  - الرفض، أخطاء الباحثين (جامعي البيانات)، أخطاء إدخال البيانات
- الإسناد هي طريقة إحصائية ل"ملىء" الفراغات في البيانات
  - عند إجراء عملية الإسناد لجميع البيانات المفقودة، يصبح لدينا مستطيل كامل من البيانات (الأسطر = الوحدات، العمدة = المتغيرات)
  - ينبغي إنشاء مؤشر (indicator variable) لتمييز البيانات الأصلية من تلك التي تم إجراء اسناد لها

طرق الإسناد

- Deductive imputation
  - طريقة معروفة، لكنها نادراً ما تكون قابلة للتطبيق
- الإسناد باستخدام المتوسط الحسابي للفئة
  - تؤدي إلى توزيع غير صحيح (متحيز) للمتغير Y
- Hot-deck imputation (random)
  - الأكثر شيوعاً، وعادة ما يكون قابلاً للتطبيق
- الإسناد باستخدام معادلة الإنحدار
  - يتراوح بين الطريقتين الثانية (المتوسط) والثالثة (Hot-deck)
- الإسناد المتعدد Multiple imputation
  - يأخذ بعين الاعتبار التباين الناتج عن عملية الاسناد نفسها

Deductive imputation

- يوجد معلومات كافية لتحديد قيمة البيانات المفقودة
  - نسبياً طريقة غير شائعة (خاصة مع وجود الكمبيوتر والحزم البرمجية الإحصائية)

الإسناد باستخدام متوسط الفئة

- الطريقة
  - قسم الوحدات المستجيبة إلى فئات إسناد
  - ضمن كل فئة إسناد:
    - إحسب المتوسط الحسابي للبيانات المتوفرة
    - إملأ الفراغات (البيانات المفقودة) بالمتوسط الحسابي للفئة

الإسناد باستخدام متوسط الفئة

- ميزاتها
  - تفترض أن البيانات مفقودة بشكل عشوائي مشروط (MAR) حيث أن (Covariates = Classes)
  - يحافظ على المتوسط الحسابي لفئة الإسناد دون تغيير
  - Under-estimates variance, distorts distribution of Y
    - تستبدل جميع البيانات المفقودة في فئة الاسناد بنفس القيمة (المتوسط الحسابي للفئة)

الإسناد باستخدام (Random) hot deck

- الطريقة
  - قسم الوحدات المستجيبة إلى فئات إسناد (فئات تعويض)
  - ضمن فئة التويض:
    - إختر عشوائياً متبرع من الوحدات المستجيبة
    - إملأ الفراغ للوحدة غير المستجيبة بقيمة من الوحدة المتبرعة

(Random) hot deck الإسناد باستخدام

- ميزاتها

  - يتم ملىء الفراغ عشوائياً من سجل بيانات مشابه (وحدة مشابهة)
  - Assumes MAR (imputation class = covariate)
  - يمكن ملأ الفراغات لعدة متغيرات من نفس المتبرع
  - لا تستخدم معلومات وعلاقات بين Y و X

الإسناد باستخدام Regression Imputation
معادلة الإنحدار

- الطريقة

  - استخدم نموذج الانحدار لربط متغرات ما X بالمتغير الذي يحتوي بيانات مفقودة
  - احسب تقدير معاملات معادلة الانحدار من بيانات المستجيبين
  - استبدل القيمة المفقودة بالقيمة المتنبأ بها من خلال معادلة الانحدار، وفي حالة النحدار اللوجيستي (if pr > 0.5, binary y = 1)

الإسناد باستخدام Regression Imputation
معادلة الإنحدار

- **ميزاتها**

- تعتبر هذه الطريقة مفيدة بشكل خاص عندما يكون عدد المشاهدات المستجيبة في فئة الاسناد قليل جداً
- يفترض أن البيانات مفقودة بشكل عشوائي مشروط MAR
- مفيدة في حالة كان هناك علاقة قوية بين المتغيرات X والمتغير Y ذي البيانات المفقودة
- تحتاج نموذج منفص لكل متغير يحتوي بيانات مفقودة

طريقة الإسناد المتعدد

- الطريقة
  - إختر طريقة إسناد (تعويض)
  - إحصل على $m > 1$ قيمة اسناد (تعويض) لكل قيمة مفقودة
  - وبذلك تكون قد حصلت على m مجموعة بيانات (مختلفة) لا تحتوي قيم مفقودة

طريقة الإسناد المتعدد

ميزاتها

- التباين في التقديرات الناتجة عن استخدام مجموعات بيانات مختلفة يزودنا بتقدير للتباين الناتج عن الإسناد (التعويض)
- تحل مشاكل موجودة في طرق تعويض (اسناد) أحرى
  - معظم طرق التحليل تعامل البيانات المعوضة وكأنها بيانات حقيقية وليست تقديرات
  - تقلل من التباين الحقيقي للتقديرات

ملخص وتقييم طرق التعويض (الإسناد)

- معظم الطرق تفترض MAR given covariates
  - التباين بين هذه الطرق يكون في النموذج المستخدم للاستفادة من هذه المتغيرات المشتركة X
- الطرق الجيدة هي تلك التي لا تؤدي إلى تغيير (تشويه) التوزيع الاحتمالي للمتغير Y
  - تجنب طريقة الاسناد (التعويض) بالمتوسط الحسابي

ملخص وتقييم طرق التعويض (الإسناد)

- طريقة Hot-deck تسمح لنا بإجراء تعويض لعدة متغيرات مرة واحدة، لكنها تحتاج إلى إمكانيات برمجية.
- عند حساب التباين، فإن معظم طرق الاسناد (التعويض) لا تأخذ بعين الاعتبار حقيقة أننا نقوم بعملية "تقدير" للبيانات
  - وهذا هو الحافز وراء طريقة التعويض (الاسناد) المتعدد
  - في حالة التعويض المتعدد، فإننا نحتاج إلى طرق خاصة لتقدير التباين

نتيجة

- يمكن حساب نسبة عدم الاستجابة إذا ما تم تسجيل نتيجة المقابلة لكل العناصر التي اختيرت في العينة.
- هذه العملية هامة ل:
  - معرفة وفهم مصادر وأنواع عدم الاستجابة
  - للسيطرة على وتقليل عدم الاستجابة
  - للتنبؤ بنسب عدم الاستجابة في المسوح القادمة (بالمستقبل)
  - ولتقدير الأثر المحتمل لعدم الاستجابة على نتائج المسح

المراجع

- Kish, Leslie (1965), “Survey Sampling”.
- Lohr, Sharon (1999), “Sampling Techniques”, Duxbury.
- Schafer. J.L. (1997), “Analysis of Incomplete Multivariate Data”, Chapman & Hall
- Sarndal, Carl-Errik (2006), “Estimation in Surveys with Nonresponse”, John Wiley & Sons
- Rubin, Donald B. (1987), “Multiple Imputation”, John Wiley & Sons
- Donders, A. Rogier T.; Van Der Heijden, Geert J.M.G; Stijnen, Theo; Moons, Karel G.M. (2006). “Review: A gentile introduction to imputation of missing values”, Journal of Clinical Epidemiology.

شكرا لحسن إصغائكم