# الاقتصاد القياسي لمعطيات البانل

محاضرات مقياس لطلبة الماستر 1 تخصص أقتصاد كمي – قسم الاقتصاد

bouabdallah salah
جامعة محمد بوضياف المسيلة |

# فصل 1. البيانات المقطعية المجمعة

**مفاهيم أساسية في البيانات المجمعة والمتغيرة الوهمية – المعالجة الإحصائية بتغير الثابت أو بالتفاعل وفرق الفروق**

**توطئة.**

ندرس في هذا الفصل نوعا من البيانات يدمج بين البيانات "العادية" أو "المقطعية"، وهي بيانات مأخوذة في لحظة زمنية ما، وبين السلاسل الزمنية التي تستمر مع الزمن. تستخدم هذه المزاوجة بين بعدين في الكثير من الدراسات التي تعني بالظواهر الاقتصادية وسيرورتها لقياس واختبار تطور ظاهرة ما مع الزمن أو تغير تأثرها بعوامل معينة أو أيضا قياس واختبار أثر سياسة عمومية ما أو حدث اقتصادي معين. قسمنا هذا الفصل إلى مبحثين: نتطرق في الأول لمفاهيم أساسية عن البيانات المقطعية وبيانات البانل والمتغيرة الوهمية، وفي المبحث الثاني نتطرق للمعالجة الإحصائية لنموذج البيانات المجمعة من خلال طريقة السماح بتغير الثابت، طريقة تفاعل المتغيرة المفسرة مع الزمن وطريقة فرق الفروق.

الفي فقرات هذا الدرس سنتعرض لأمثلة عن كيفية استخدام طرق تحليل الانحدار لدراسة هذا النوع من المسائل. الأمثلة والشروح مستقاة أساسا من كتاب وولدرج[1] Wooldridge, 2015.

## 1. مفاهيم أساسية

تعريف البيانات المقطعية المستقلة المجمعة
تعريف بيانات البانل
طريقة تنظيم البيانات
ترميز المتغيرة الإسمية

### 1-1. البيانات المقطعية المجمعة (المستقلة)

**يقصد بالبيانات المقطعية** time)تلك البيانات المأخوذة مرة واحدة، في لحظة زمنية واحدة، فهي لا تتضمن بعد الزمن.

**يقصد بالبيانات المقطعية المجمعة المستقلة**[2] (pooled cross sections) تلك البيانات المأخوذة كعينات عشوائية من مجتمع كبير في فترات مختلفة، غالبا سنوات، مع استقلال بيانات كل عينة (سنة) عن بيانات العينات الأخرى. في هذه الحالة فإن التوزيع الاحتمالي للبيانات يختلف من فترة لأخرى، مثلا توزيع الأجر يختلف من فترة لأخرى وكذلك المؤهل العلمي، وهكذا. **لأخذ هذا الاختلاف في الحسبان في نموذج تحليل الانحدار المتعدد، يمكن إما أن نسمح للثابت بأن يتغير من فترة لأخرى، أو تتغير معاملات الانحدار الجزئية، أو كلاهما.**

[1] Wooldridge Jeffrey M., Introduction à l'économétrie, une approche modèrne, de boeck, Belgique, 2015
[2] Données en coupes transversalles empilées indépendantes.

**أمثلة:**

- بيانات تعداد السكان: نستخرج في 2010 عينة من 300 عائلة، ونأخذ بياناتها للدخل وعدد الأولاد، وعدد غرف السكن، المستوى الدراسي للأم. في 2020 نستخرج مجددا عينة من 320 عائلة أخرى، ونأخذ ذات البيانات المذكورة. نقوم بتجميع البيانات في جدول بيانات واحد، نقول عنها أنها بيانات مقطعية مجمعة (مستقلة). حجم العينة الإجمالي هو 620 عائلة.

| Obs | Year | Income | kids | rooms | mothlevel |
|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... |
| **298** | **2010** | **55000** | **2** | **3** | **8** |
| **299** | **2010** | **28000** | **3** | **4** | **12** |
| **300** | **2010** | **32000** | **4** | **4** | **6** |
| 301 | 2020 | 31000 | 2 | 2 | 14 |
| 302 | 2020 | 29000 | 4 | 5 | 11 |
| 303 | 2020 | 28000 | 5 | 6 | 12 |
| ... | ... | ... | ... | ... | ... |
| 620 | 2020 | 46000 | 6 | 1 | 8 |

- نأخذ بيانات: 'الأجر' و'عدد ساعات العمل' و'عدد العطل المرضية' و'المؤهل المهني' لعينة من العمال في سنة 2005، ثم بعد 5 سنوات نكرر العملية على عينة جديدة من العمال، **لدراسة تطور تأثر الأجر بالمتغيرات المستقلة المذكورة**. يمكن أن نكرر العملية أكثر من مرة.
- نكون عينة من المؤسسات من قطاع معين في 2020 ونسجل بيانات 'الربح' لقياس تأثره ب'رقم الأعمال' و'الإنفاق الإشهاري' وبعد سنتين نكرر العملية مع عينة جديدة. الهدف هو قياس **تأثر علاقة 'الربح' بالمتغيرتين المذكورتين بعد إدخال ضريبة جديدة مثلا.**

## 1-2. بيانات البانل

يقصد ببيانات البانل في الأصل عينة ثابتة من الأفراد أو العائلات التي يتم استجوابها على فترات. على غرار البيانات المقطعية المجمعة، تزاوج بيانات البانل بين البعدين: الفردي والزمني، لكن العينات ليست مستقلة، وذلك لأن المفردات تبقى نفسها. في بيانات البانل (تسمى أيضا البيانات الطولية Longitudinal) نتتبع نفس المفردات (مؤسسات، بلدان، مدن، عائلات، مواطنون ...) لنستقي بياناتها كل فترة. العينة المكررة تنقض الاستقلالية؛ فأجر عامل ما في سنة ما لن يكون مستقلا عن أجره في السنة السابقة، وأداء مؤسسة في سنة ما ليس مستقلا عن أدائها قبل سنتين، إلخ.

**مثال 2.** في دراسة شهيرة لجرينفيلد[1] حول اعتماد النمو الاستثماري الحقيقي للمؤسسة على القيمة الحقيقية للمؤسسة وأسهم راس المال الحقيقية، حصل الباحث على بيانات عدة شركات نأخذ منها أربعة هي (General GE

[1] Y. Grunfeld, 'The dterminants of Corporate Investement,

(General Motors) GM ،Electric)، وUS للصلب، وويستنجهاوس. البيانات مستمرة على 20 سنة، من 1935 إلى 1954.

| Observation | I | $F_{-1}$ | $C_{-1}$ |
|---|---|---|---|
| | GE | | |
| 1935 | 33.1 | 1170.6 | 97.8 |
| 1936 | 45.0 | 2015.8 | 104.4 |
| 1937 | 77.2 | 2803.3 | 118.0 |
| 1938 | 44.6 | 2039.7 | 156.2 |
| 1939 | 48.1 | 2256.2 | 172.6 |
| 1940 | 74.4 | 2132.2 | 186.6 |
| 1941 | 113.0 | 1834.1 | 220.9 |
| 1942 | 91.9 | 1588.0 | 287.8 |
| 1943 | 61.3 | 1749.4 | 319.9 |
| 1944 | 56.8 | 1687.2 | 321.3 |
| 1945 | 93.6 | 2007.7 | 319.6 |
| 1946 | 159.9 | 2208.3 | 346.0 |
| 1947 | 147.2 | 1656.7 | 456.4 |
| 1948 | 146.3 | 1604.4 | 543.4 |
| 1949 | 98.3 | 1431.8 | 618.3 |
| 1950 | 93.5 | 1610.5 | 647.4 |
| 1951 | 135.2 | 1819.4 | 671.3 |
| 1952 | 157.3 | 2079.7 | 726.1 |
| 1953 | 179.5 | 2371.6 | 800.3 |
| 1954 | 189.6 | 2759.9 | 888.9 |
| | GM | | |
| 1935 | 317.6 | 3078.5 | 2.8 |
| 1936 | 391.8 | 4661.7 | 52.6 |
| 1937 | 410.6 | 5387.1 | 156.9 |
| 1938 | 257.7 | 2792.2 | 209.2 |
| 1939 | 330.8 | 4313.2 | 203.4 |
| 1940 | 461.2 | 4643.9 | 207.2 |
| 1941 | 512.0 | 4551.2 | 255.2 |
| 1942 | 448.0 | 3244.1 | 303.7 |
| 1943 | 499.6 | 4053.7 | 264.1 |
| 1944 | 547.5 | 4379.3 | 201.6 |
| 1945 | 561.2 | 4840.9 | 265.0 |
| 1946 | 688.1 | 4900.0 | 402.2 |
| 1947 | 568.9 | 3526.5 | 761.5 |
| 1948 | 529.2 | 3245.7 | 922.4 |
| 1949 | 555.1 | 3700.2 | 1020.1 |
| 1950 | 642.9 | 3755.6 | 1099.0 |
| 1951 | 755.9 | 4833.0 | 1207.7 |
| 1952 | 891.2 | 4924.9 | 1430.5 |
| 1953 | 1304.4 | 6241.7 | 1777.3 |
| 1954 | 1486.7 | 5593.6 | 2226.3 |

| Observation | I | $F_{-1}$ | $C_{-1}$ |
|---|---|---|---|
| | US | | |
| 1935 | 209.9 | 1362.4 | 53.8 |
| 1936 | 355.3 | 1807.1 | 50.5 |
| 1937 | 469.9 | 2673.3 | 118.1 |
| 1938 | 262.3 | 1801.9 | 260.2 |
| 1939 | 230.4 | 1957.3 | 312.7 |
| 1940 | 361.6 | 2202.9 | 254.2 |
| 1941 | 472.8 | 2380.5 | 261.4 |
| 1942 | 445.6 | 2168.6 | 298.7 |
| 1943 | 361.6 | 1985.1 | 301.8 |
| 1944 | 288.2 | 1813.9 | 279.1 |
| 1945 | 258.7 | 1850.2 | 213.8 |
| 1946 | 420.3 | 2067.7 | 232.6 |
| 1947 | 420.5 | 1796.7 | 264.8 |
| 1948 | 494.5 | 1625.8 | 306.9 |
| 1949 | 405.1 | 1667.0 | 351.1 |
| 1950 | 418.8 | 1677.4 | 357.8 |
| 1951 | 588.2 | 2289.5 | 341.1 |
| 1952 | 645.2 | 2159.4 | 444.2 |
| 1953 | 641.0 | 2031.3 | 623.6 |
| 1954 | 459.3 | 2115.5 | 669.7 |
| | WEST | | |
| 1935 | 12.93 | 191.5 | 1.8 |
| 1936 | 25.90 | 516.0 | 0.8 |
| 1937 | 35.05 | 729.0 | 7.4 |
| 1938 | 22.89 | 560.4 | 18.1 |
| 1939 | 18.84 | 519.9 | 23.5 |
| 1940 | 28.57 | 628.5 | 26.5 |
| 1941 | 48.51 | 537.1 | 36.2 |
| 1942 | 43.34 | 561.2 | 60.8 |
| 1943 | 37.02 | 617.2 | 84.4 |
| 1944 | 37.81 | 626.7 | 91.2 |
| 1945 | 39.27 | 737.2 | 92.4 |
| 1946 | 53.46 | 760.5 | 86.0 |
| 1947 | 55.56 | 581.4 | 111.1 |
| 1948 | 49.56 | 662.3 | 130.6 |
| 1949 | 32.04 | 583.8 | 141.8 |
| 1950 | 32.24 | 635.2 | 136.7 |
| 1951 | 54.38 | 732.8 | 129.7 |
| 1952 | 71.78 | 864.1 | 145.5 |
| 1953 | 90.08 | 1193.5 | 174.8 |
| 1954 | 68.60 | 1188.9 | 213.5 |

العينة التي نحصل عليها بإدخال البيانات هي 80 مفردة. النموذج يأتي كما يلي:

$$y_{it} = \beta_0 + \beta_1 F_{it} + \beta_2 F_{it} + u_{it}, \quad t = 1, 2, 3, 4 \; ; \; i = 1, \ldots, 20.$$

**مثال 2:** ليكن لدينا بيانات 'معدل الضريبة' و'البطالة' ل 90 دولة نامية، وهذه البيانات ممتدة من 2001 إلى 2005، ونريد أن نستخدم هذه البيانات لتفسير البطالة. إذا أدخلنا بيانات بلد واحد خلال الفترة فلن نعرف إن كانت العلاقة تتعمم على البلدان الأخرى، وإذا إدخلنا بيانات البلدان في لحظة واحدة فلن نعرف إن كانت العلاقة معممة على مر الزمن. الحل هو أخذ بيانات مختلف البلدان على مدى الزمن، وهي البيانات المجمعة.

| Obs | State | Year | TaxRate | Unemployment |
|---|---|---|---|---|
| ... | ... | ... | ... | ... |
| 80 | Gabon | 2001 | .35 | 0.23 |
| 81 | Gabon | 2002 | .37 | 0.20 |
| 82 | Gabon | 2003 | .36 | 0.21 |

| | | | | |
|---|---|---|---|---|
| 83 | Togo | 2001 | .14 | 0.18 |
| 84 | Togo | 2002 | .16 | 0.19 |
| 85 | Togo | 2003 | .11 | 0.14 |
| ... | ... | ... | ... | ... |

الحصول على بيانات بانل قد يكون سهلا في وحدات مستقرة، يمكن الرجوع إليها، مثل مؤسسات، مدارس، بلدان...، على عكس الأفراد، فالفرد قد لا يمكن الرجوع إليه بسبب تغيير العنوان أو الموت أو غيره.

**مثال3**. نهتم بتأثير الدخل الوطني، ولدينا بيانات عينة من البلدان تم تتبعها على عدة سنوات.

**مثال4**. نريد اختبار تأثير مستوى التعليم على دخل الفرد، ولدينا بينات عينة من الأفراد تم تتبعهم على عدة سنوات.

**خلاصة:** استخدام بيانات البانل بالمزاوجة بين البعدين الزمني والفردي يحسن تحليل السياسات الاقتصادية، **لأنه يسمح بعزل تأثير عوامل خفية عديدة** لم يتم قياسها لاستحالة أو صعوبة القياس، مثل العوامل الثقافية والاجتماعية وغيرها[1].

بيانات البانل يمكن أن تكون "**متوازنة**" (balanced) إذا كانت كل المفردات لدينا مشاهداتها في كل الفترات ($T_i$ = T for all i)، أو تكون "**غير متوازنة**" (unbalanced) إذا كانت هناك قيم مفقودة بحيث أن بعض المفردات ليس لدينا مشاهداتها لكل الفترات($T_i \neq T$).

في بيانات البانل يمكن أن يكون هناك ارتباط مشاهدات الفرد الواحد عبر الزمن، لكن ليس هناك ارتباط بين الافراد. مثلا: دخل الفرد مرتبط ذاتيا عبر الزمن، لكن دخول الافراد المختلفين مستقلة عن بعضها البعض.

بيانات البانل أنواع: يمكن أن تكون **قصيرة** (short panel data) بمعنى بيانات لمفردات كثيرة على فترات زمنية قليلة، وهي الحالة الغالبة. يمكن أن تكون **طويلة** (long) بمعنى فترات طويلة زمنيا ومفردات قليلة، وهي حالة قليلا ما نجدها في الدراسات الاقتصادية. والحالة الثالثة تدعى (both) وهي بيانات كثير من الافراد على فترة طويلة، وهي حالة نادرة.

المتغيرات المفسرة للنموذج أنواع: مفسرات متغيرة ($x_{it}$ varying regressors)، مثلا الدخل، عدد أولاد العائلة، الحالة العائلية ... فهي متغيرة من فرد لآخر ومن فترى لأخرى. هناك مفسرات ثابتة في الزمن ($x_{it}$ = time-invariant regressors) ($x_i$ for all t)، مثل الجنس، والعرق، ... وهناك مفسرات ثابتة بين الأفراد ($x_{it} = x_t$ for all i individual-invariant) مثل متغيرات التوجه الاقتصادية معدل البطالة أو التضخم.

## 1-3. طريقة تنظيم بيانات البانل

بفرض أن لدينا 50 مدينة، مشاهدة على مرحلتين. هذا يعني أنه سيكون لدينا 100 سطر. الطريقة الأولى والشائعة لتنظيم البيانات هي تخصيص سطرين لكل مدينة، السطرين الأولين للمدينة الأولى في العينة: الأول للمشاهدة الأولى والثاني

[1] تسمى هذه الطريقة في المربعات الصغرى نموذج الأثر المثبت (fixed effects)، في هذه الحالة سيكون لدينا نموذج الأثر المثبت للأفراد ونموذج الأثر المثبت للزمن ويمكن استخدام نموذج للأثر المثبت لكل منهما. عيب هذه الطريقة أنها تحتاج إلى عدد كبير من المتغيرات الوهمية، وهذا يقلل من درجات الحرية للنموذج، وبالتالي فهي لا تصلح إلا إذا كان عدد الأنماط في المتغيرات النوعية ضئيلا. الطريقة الأخرى لمعالجة البيانات ذات البعدين (بيانات البانل) هي نموذج الأثر العشوائي (Random effects model).

للمشاهدة الثانية، والسطرين المواليين للمدينة الثانية، وهكذا. بهذه الطريقة يسهل حساب الفروق بين السنتين، كما يسهل استخدام طريقة البيانات المجمعة للمقارنة بين الطريقتين: التقدير بالبيانات المجمعة والفروق الأولى.

في حالة وجود أكثر من فترتين، يمكن استخدام ذات الطريقة: المشاهدات الخاصة بكل مفردة (مدينة) مرتبة تصاعديا بجوار بعضها.

الطريقة الثانية هي تخصيص سطر واحد لكل مفردة (مدينة) وهذا يعني إدخال كل متغيرة مرتين، مرة لكل فترة. في مثال معدل الضريبة والبطالة سيكون لدينا لكل عامل المتغيرات التالية:

TaxRate 2001; Taxrate2002 ;TaxRate2003 ; Unemployment2001 ;
Unemployment2002 ; Unemployment2003 ;

بحيث تحسب الفروق أفقيا لا عموديا. هذه الطريقة لا تسمح باستخدام طريقة التقدير بالبيانات المجمعة، كما أنها لا تصلح عندما تكون الفترات أكثر من 2.

**18-02-2020**

## 1-4. ترميز المتغيرة المفسرة الإسمية بمتغيرات ثنائية

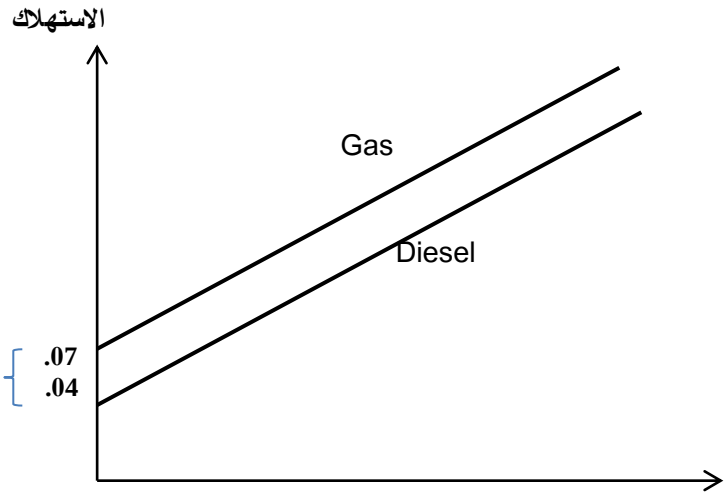
ماهي المتغيرة الإسمية وكيف ندرجها كمفسرات في النموذج؟ المتغيرة الإسمية (وتسمى أيضا نوعية) تأخذ أنماطا (مثلا متغيرة الحالة العائلية تأخذ الأنماط التالية: 'متزوج'، 'أعزب'، 'مطلق'، ...) بينما المتغيرة الكمية تأخذ قيما عددية. نعبر عن المتغيرة النوعية بمجموعة من المتغيرات الثنائية المصطنعة أو الوهمية (المتغيرة الثنائية تأخذ لكل مشاهدة إحدى القيمتين 0 أو 1: تأخذ 0 إذا لم يتحقق النمط في المشاهدة وتأخذ 1 إذا تحقق). عدد المتغيرات الوهمية (dummy variables) الضرورية لتعويض المتغيرة النوعية يساوي عدد أنماطها ناقصا الواحد (أحد الأنماط يترك كمرجع حتى لا يتكون ارتباط مشترك تام).

**مثال 1.** نريد ربط كمية الوقود المستهلك (باللتر) بمتغيرة كمية هي المسافة المقطوعة Dist (بالكلم) ومتغيرة نوعية هي نوع الوقود Type، أنماطها "ديزل" و "بنزين". لإدراج Type نحتاج إلى متغيرة وهمية واحدة (لأن عدد الأنماط هو 2: نستخدم الرمز "0" للديزل و "1" للبنزين). نفترض أن الدالة جاءت كما يلي:

$$Cons\hat{} = 0.4 + 0.3\ Type + 0.6\ Dist$$

معامل الانحدار الجزئي للمتغيرة Type يمثل الفرق بين استهلاك البنزين والديزل: المركبة التي تستخدم الوقود من نوع "بنزين" (رمز النمط هو 1) تستهلك في المتوسط 0.3 لتر أكثر في الكلم الواحد مقارنة مع المركبات العاملة بالديزل، وذلك لنفس المسافة المقطوعة.

لو مثلنا الانحدار البسيط لاستهلاك الوقود على المسافة المقطوعة بمركبات البنزين لوحدها، وفعلنا الأمر نفسه مع مركبات الديزل، سيكون خط المربعات الصغرى لمركبات البنزين أعلى بمقدار 0.3 من خط مركبات الديزل.




**رسم توضيحي 1. إدراج متغيرة ثنائية في النموذج. لذات المسافة يستهلك محرك البنزين وقودا أكثر.**

**مثال 2:** في المثال السابق، نفترض أن المتغيرة الإسمية Type تأخذ ثلاثة أنماط: "Gas"، "Diesel"، "Super"؛ لترميز هذه المتغيرة الإسمية نحتاج إلى:

- متغيرة D تأخذ القيمة 1 للسيارة العاملة ب "Diesel"، و0 إذا لم تكن كذلك،
- متغيرة G تأخذ 1 للسيارات العاملة ب "Gas"، و0 إذا لم تكن كذلك،

جدول 1 ترميز متغيرة نوعية ذات ثلاثة أنماط

| المتغيرات | | الأنماط | | |
|---|---|---|---|---|
| | | "Gas" | "Diesel" | "Super" |
| Gas | 1 | 1 | 0 | 0 |
| Diesel | 0 | 0 | 1 | 0 |

يكتب النموذج كالتالي:

$$Cons = \beta_0 + \delta_0\, Gas + \delta_1\, Diesel + \beta_1\, Distance + \varepsilon$$

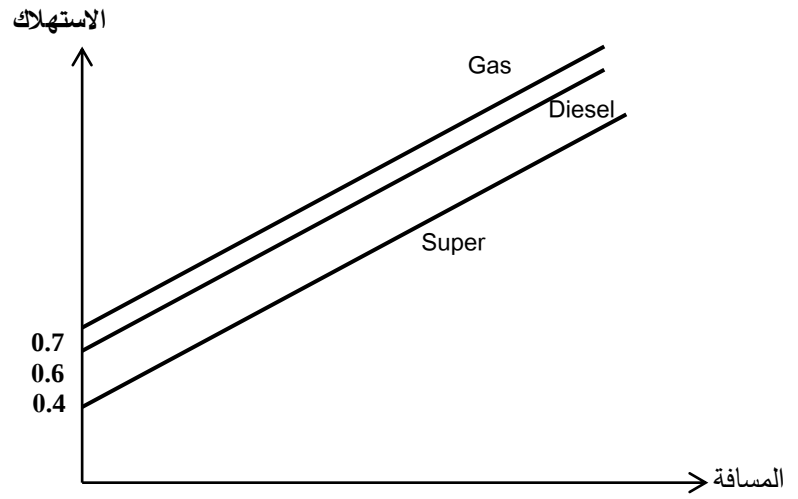
نفترض أن الدالة كانت:

$$\widehat{Cons} = 0.4 + 0.3\, Gas + 0.2\, Diesel + 0.6\, Dist$$

في هذه الحالة فإن معامل الانحدار الجزئي للمتغيرة Gas يمثل الفرق بين استهلاك البنزين العادي واستهلاك البنزين الممتاز، أي أنه يأخذ التفسير التالي: المركبة التي تستخدم "بنزين" (Gas) تستهلك في المتوسط 0.3 لتر أكثر في الكلم الواحد مقارنة مع المركبات العاملة ب "الممتاز"، وذلك لنفس المسافة المقطوعة.

معامل الانحدار الجزئي للمتغيرة Diesel يمثل الفرق بين استهلاك الديزل واستهلاك الممتاز، أي أنه يأخذ التفسير التالي: المركبة التي تستخدم الوقود من نوع Diesel تستهلك في المتوسط 0.2 لتر أكثر في الكلم الواحد مقارنة مع المركبات العاملة بالممتاز، وذلك لنفس المسافة المقطوعة.

لو مثلنا الانحدار البسيط لاستهلاك الوقود على المسافة المقطوعة بمركبات البنزين لوحدها، وفعلنا الأمر نفسه مع مركبات الديزل والسوبر، سيكون خط المربعات الصغرى لمركبات البنزين أعلى بمقدار 0.3 من خط المربعات الصغرى لمركبات الديزل.




**رسم توضيحي 2. إدراج متغيرة ثنائية في النموذج. لذات المسافة يستهلك محرك النزين وقودا أكثر.**

سؤال: قم بصياغة الدالة في حالة اتخاذ النمط Gas كمرجع.

**مثال 3[1]:** بفرض أننا أمام نموذج يتوقع كمية الوقود المستهلك بناءا على نوع السيارة، وتأخذ أربعة أنماط: "رياضية"، "تجارية"، "سياحية"، و"عائلية". الترميز سيكون على غرار المثال السابق ولكن بثلاث متغيرات ثنائية.

جدول 2 ترميز متغيرة نوعية ذات أربعة أنماط

| أنماط المتغيرة الاسمية | | "رياضية" | "تجارية" | "عائلية" | "سياحية" |
|---|---|---|---|---|---|
| المتغيرات الثنائية | $D_1$ | 1 | 0 | 0 | 0 |
| | $D_2$ | 0 | 1 | 0 | 0 |
| | $D_3$ | 0 | 0 | 1 | 0 |

**أكتب الدالة، وأستخرج كمية الوقوج المستهلك لمختلف أنواع السيارات الأربع.**

الجواب:

$$\hat{y} = b_0 + b_1D_1 + b_2D_2 + b_3D_3$$

النمط الرابع "سياحية" اتخذ مرجعا ولم يدرج في الدالة.

يظهر هذا المثال العلاقة بين تحليل الانحدار البسيط بمتغيرة إسمية وتحليل التباين الأحادي. التباين المتبقي هو نفسه في النموذجين، والتباين المفسر هونفسه أيضا، وكذلك اختبار F لمجمل النموذج، أما المعاملات فتفسر كما يلي:

- **السيارات السياحية:** $D_1 = D_2 = D_3 = 0$

إذن تأخذ الدالة القيمة $b_0$ ويمثل الاستهلاك المتوقع للسيارا السياحية:

$$\hat{y} = b_0 + b_1(0) + b_2(0) + b_3(0) = b_0$$

- **للسيارات الرياضية:**

[1] Malhotra ،Décaudin ،Bouguerra, et al. étude marketing avec SPSS, Pearson, Paris, 2007

$D_1 = 1, D_2 = D_3 = 0 \Rightarrow \hat{y} = b_0 + b_1(1) + b_2(0) + b_3(0) = b_0 + b_1$

$b_1$ يمثل الفرق المتوقع بين السيارات السياحية والسيارات الرياضية في استهلاك الوقود.

- **السيارات التجارية:**

$$D_1 = 0, D_2 = 1, D_3 = 0 \Rightarrow \hat{y} = b_0 + b_2$$

المعامل $b_2$ يفسر بالفرق المتوقع بين السيارات التجارية والسياحية.

- **السيارات العائلية:**

$$D_1 = 0, D_2 = 0, D_3 = 1 \Rightarrow \hat{y} = b_0 + b_3$$

المعامل $b_3$ يفسر بالفرق بين القيمة المتوقعة بين السيارات العائلية والسياحية.

عرفنا إذن في هذا المبحث أن البيانات المقطعية المجمعة – أو بيانات البانل البسيط – هي بيانات لعينات مستقلة من المفردات تأخذ في فترات زمنية متتالية، فترتين أو أكثر، بينما بيانات البانل هي بيانات عينة من المفردات يتم تتبعها على فترات متتالية، فترتين أو أكثر. عرفنا أيضا كيفية الترميز للمتغيرة الاسمية باستخدام متغيرة وهمية ثنائية أو عدد من المتغيرات الثنائية الوهمية وتفسير معاملات هذه المتغيرات الوهمية. في المبحث الموالي نتطرق للمعالجة الإحصائية لنموذج البيانات المقطعية المجمعة وبيانات البانل، حيث نرى أن هناك أكثر من طريقة لإدراج عامل الزمن في النموذج.

25-02-2020

## 2. المعالجة الإحصائية لنموذج البيانات المجمعة

السماح بتغير الثابت

تفاعل المفسؤات مع الزمن

فرق الفروق

من أسباب استخدام البيانات المقطعية المجمعة المستقلة (أو اختصارا البيانات المجمعة pooled data) الرغبة في زيادة حجم العينة. **البيانات المجمعة عن نفس المجتمع على فترات تعطي مقدرات أكثر دقة واختبارات أقوى**. في هذه الحالة بالذات، **يكون التجميع مفيدا إذا كانت علاقة المتغيرة التابعة ثابتة عبر الزمن على الأقل ببعض المتغيرات المستقلة** (وإلا إذا كانت كل العلاقات تتغير يصبح استخدام نماذج مختلفة هو الحل بدون الحاجة للتجميع).

يترتب على تجميع بيانات من فترات مختلفة بعض التعقيدات الإحصائية لكي يمكن احتساب اختلاف التوزيع الاحتمالي للمتغيرات. **يمكن أن يتم ذلك إما بالسماح بتغير الثابت، أو بالسماح بتداخل الزمن مع معاملات المتغيرات المستقلة**.

### 2-1. السماح بتغير الثابت (أثر الزمن)

في هذه الطريقة ندرج الزمن كمتغيرة مفسرة في النموذج ولكن باعتباره متغيرة إسمية وليس كمية. يكون ذلك بإدراج متغيرة مفسرة ثنائية لكل سنة إلا سنة الأساس أو "المرجع"، وغالبا ما تكون هي السنة الأولى (لكن ليس بالضرورة). أحيانا يكون تطور المعاملات السنوية في حد ذاته مهما.

النموذج الذي يمكن من احتساب أثر الزمن (بفرض أن لدينا بيانات K فترة و J عامل) يكون كالتالي:

$$y = \beta_0 + \delta_0 year_2 + \delta_1 year_3 + \ldots + \delta_{K-2} year_K + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_J x_J + \varepsilon$$

المعامل $\beta_0$ هو الثابت (في حالة السنة 1).

$\beta_0 + \delta_0$ يمثل الثابت بالنسبة لمشاهدات السنة الثانية (السنوات الأخرى تختفي لأن معاملاتها تنعدم، وتبقى العوامل $X_J$)

$\beta_0 + \delta 1$ يمثل الثابت بالنسبة لمشاهدات السنة الثالثة (السنوات الأخرى تختفي لأن معاملاتها تنعدم، وتبقى العوامل $X_J$)

…

المعامل $\beta_1$ يقيس الأثر الجزئي للعامل الأول $X_1$ عند تثبيت العوامل الأخرى

…

**$\delta_0$ يمثل التغير في التابع من السنة الأولى (سنة المرجع) إلى السنة الثانية عند تثبيت العوامل الأخرى**

**$\delta_1$ يقيس التغير في التابع من السنة الأولى إلى السنة الثالثة عند تثبيت العوامل الأخرى**

…

المعامل $\beta_J$ يقيس الأثر الجزئي للعامل $X_J$ عند تثبيت العوامل الأخرى

**مثلا[1] (الخصوبة):** في "دراسة عامة اجتماعية" قام بها "مركز البحث في الرأي الوطني" في الولايات المتحدة، عولجت المسألة التالية: هل **اختلفت الخصوبة** (kids) لدى النساء الأكثر من 35 سنة بين سنة 1972 وسنة 1984؟ وهذا بعد احتساب العوامل الديمغرافية التالية: سنوات التعليم educ، العمر age، العمر عند الزواج age2، العرق Black، المدينة والجهة، وعوامل أخرى. كيف يكون النموذج وتفسيره؟ المعاينة تمت كل سنتين.

**جدول3 جدول بيانات مقطعية مجمعة (pooled data) . هنا الصورة وهمية لتوضيح شكل جدول البيانات.**

| Obs | individuals | Y74 | Y76 | Y78 | Y80 | Y82 | Y84 | kids | educ | age | age2 | black | …. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| 199 | Pamila | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 10 | 36 | 30 | 0 | … |
| 200 | Linda | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 12 | 36 | 22 | 1 | … |
| 201 | Sofia | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 11 | 37 | 18 | 1 | … |
| 202 | Lilia | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 9 | 35 | 28 | 0 | … |
| 203 | Kati | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 35 | 38 | 0 | … |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| 401 | Saly | 0 | 1 | 0 | 0 | 0 | 0 | 5 | 5 | 40 | 41 | 1 | … |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| 1129 | Margerit | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 14 | 42 | 39 | 0 | … |

[1] Wooldridge J. M., 2015, op. cit. p. 644.

| Xj | Coefficients | SE |
|---|---|---|
| Educ | −0.128 | 0.018 |
| Age | 0.532 | 0.138 |
| Age2 | −0.0058 | 0.0016 |
| Black | 1.076 | 0.174 |
| East | 0.217 | 0.133 |
| Northen | 0.363 | 0.121 |
| West | 0.198 | 0.167 |
| Farm | −0.053 | 0.147 |
| Othrural | −0.163 | 0.175 |
| Town | 0.084 | 0.124 |
| Smcity | 0.212 | 0.160 |
| Y74 | 0.268 | 0.173 |
| Y76 | −0.097 | 0.179 |
| Y78 | −0.069 | 0.182 |
| Y80 | −0.071 | 0.183 |
| Y82 | −0.522 | 0.172 |
| Y84 | −0.545 | 0.175 |
| **Constant** | −7.742 | 3.052 |
| N = 1129, $R^2$ = 0.1295, $R^2$aj = 0.1162 | | |

- **ما هو مقدار التغير في الخصوبة (عدد الأولاد) للمرأة من سنة 72 إلى 74، من 72 إلى 76، من 72 إلى 78، من 72 إلى 80، من 72 إلى 82، من 72 إلى 84، عند تثبيت العوامل الأخرى؟**

ج: من 72 إلى 84 انخفضت الخصوبة في المتوسط ب 0.545، أي في المتوسط، لكل مئة امرأة 55 طفل أقل.

- **هل أسباب تغير الخصوبة المحسوبة بمعاملات Y74 إلى Y84 هي المتغيرات المذكورة قبل في الجدول؟**

ج: لا. المعاملات تحتسب الأثر الجزئي، أي مع تثبيت العوامل الأخرى.

- **س: هل النساء الذين زاولن التعليم لمدة أطول كان لهن أولاد أقل؟ ما هو مقدار الفرق في الخصوبة بين ذوات التعليم الثانوي وذوات التعليم الجامعي (3 سنوات فرق)، عند تثبيت العوامل الأخرى؟**

الجواب هو نعم، في ثلاث سنوات انخفضت الخصوبة بمعدل (0.384 = 3 x 0.128).

- **هل النساء ذوات العرق الأسود هن أكثر خصوبة، بكم؟** ج: نعم، بطفل تقريبا.

- **علق على قيمة $R^2$= 0.1295**. ج: قيمة ضئيلة، وهذا لأن عدد الأطفال تاثر عليه عوامل كثيرة أخرى، لم يتم احتسابها ويصعب احتسابها جميعا، ولذلك ليس من النادر رأية قيمة ضئيلة ل $R^2$ في الدراسات الاقتصادية والاجتماعية والإنسانية عموما.

رأينا في المثال أعلاه أن نموذج تغير الثابت يعطينا إذن تطور المتغيرة التابعة مع الزمن، فهذا هو الهدف الأساسي لهذه الطريقة، لكنه لا يبين تغير العوامل مع الزمن، لأنه في الواقع يفترض أن العوامل الأخرى ثابتة مع الزمن. في الفقرة الموالية نتطرق لحالة تغير أثر العوامل عبر الزمن وكيفية حسابه.

## 2-2. تفاعل المتغيرات المفسرة مع الزمن (تغير أثر العوامل عبر الزمن)

في المثال السابق استخدم تغير الثابت لقياس أثر الزمن على المتغيرة التابعة. يمكن أيضا قياس تغير الأثر نفسه مع الزمن، أي تغير أثر المغيرة أو المتغيرات المفسرة على التابعة مع الزمن. يكون ذلك بإدراج تفاعل معاملات المتغيرات المفسرة مع الزمن.

بفرض أن لدينا بيانات K فترة و J عامل، يمكن أن نكتب النموذج كالتالي:

$$y = \beta_0 + \delta_0 year_2 + \delta_1 year_3 + \ldots + \delta_{K-2} year_K + \beta_1 x_1 + \ldots + \beta_J x_J + \ldots + \delta_k year_k x_j + \ldots + e$$

المعامل $\beta_0$ هو الثابت بالنسبة لمشاهدات سنة الأساس.

$\beta_0 + \delta_0$ هو الثابت بالنسبة لمشاهدات السنة الثانية (لأن السنوات الأخرى معدومة).

$\delta_0$: يقيس التغير في التابع من السنة الأولى إلى السنة الثانية عند تثبيت العوامل الأخرى

$\delta_1$: يقيس التغير في التابع من السنة الأولى إلى السنة الثالثة عند تثبيت العوامل الأخرى

...

**$\delta_k$ في الحد $\delta_k year_k x_j$ يمثل تغير أثر العامل $x_j$ بين سنة الأساس وسنة المقارنة k.**

يمكن بهذه الطريقة قياس تغير أثر العامل أو العوامل $x_j$ مع الزمن. عمليا يتطلب هذا إنشاء متغيرة جديدة هي $year_k x_j$.

**مثال[1] 2 (الأجر):** في دراسة على تغير تأثير كل من عدد سنوات التعليم educ، والنوع female، على الأجر wage بين سنتي 1978 و1985، مع تثبيت أثر عدد سنوات الخبرة exper والانتماء إلى النقابة union جاء، النموذج كما يلي:

$$\text{Log(wage)} = 0.459 + 0.118\ y85 + 0.0747\ educ + 0.0185\ y85.educ + 0.0296\ exper - 0.0004\ exper^2 + 0.202\ union - 0.317\ female + 0.085\ y85.female + e$$

| | | | | | |
|---|---|---|---|---|---|
| SE = | (0.093) | (0.124) | (0.0067) | (0.0094) | (0.0036) |
| | (0.00008) | (0.03) | (0.037) | (0.051) | |

[1] Wooldridge J. M., op. cit. 2015, p. 646.

| $N = 1084; R^2 = 0.426; R^2_{aj} = 0.422$ |
|---|

y85 هي متغيرة ثنائية تأخذ 1 إذا كانت المشاهدة مسجلة في عينة 85، وتأخذ 0 بخلاف ذلك. Union تأخذ 1 إذا كان العامل ينتمي لنقابة، وتأخذ 0 بخلاف ذلك و female تأخذ 1 للنساء و0 للرجال.

- **ما هو الثابت بالنسبة لمشاهدات 78؟** الجواب: 0.459
- **ما هو الثابت بالنسبة لمشاهدات 85؟** الجواب: 0.118 + 0.459
- **ماذا تمثل** 0.118 + ؟ ج: تمثل زيادة أو توجه عام زائد في لغ الأجر من 78 إلى 85 لفائدة الجميع بصرف النظر عن الجنس والدراسة والخبرة والانتماء للنقابة...
- **ماهو مردود التعليم (أثر سنة تعليم إضافية على الاجر) في 78؟** الجواب: هو معامل educ أي: 0.0747
- **ماهو مردود التعليم في سنة 85؟** الجواب: 0.0185 + 0.0747
- **ماذا يقيس معامل y85.educ؟** الجواب: القيمة **0.0185** تقيس **التغير في مردود سنة تعليم إضافية**، بين الفترتين 78 و 85.
- **ماذا يقيس معامل female؟** ج: المعامل **0.3.17−** **يقيس فارق الأجر بين الجنسين.**
- **ماذا يقيس معامل y85.female؟** ج: المعامل **0.085** يقيس **التغير في فارق الأجر بين الجنسين**، بين الفترتين 78 و85 (كونه عكس إشارة المعامل female يعني أن الفارق تقلص).
- **كيف يمكن اختبار فرضية أن فارق الأجر بين النساء والرجال لم يتغير بين 78 و85؟**

الجواب: باختبار فرضية أن معامل (y85.female) يساوي الصفر: نقسم المعامل 0.085 على الخطأ المعياري لنحصل على t ونقارنها مع القيمة الجدولية، أو مع 1.96، إذا جاءت أكبر نقول إن المعامل دال، عند مستوى معنوية 5% (اختبار ثنائي).

- **ماهو تأثير اللغرتم على تفسير المعاملات؟**

الجواب: **إدخال اللغرتم على التابع يجعل معاملات الانحدار الجزئية تمثل نسبا** (عند ضربها في مئة تصبح نسبا مئوية، أي نسب الزيادة في التابع عند زيادة المستقل بوحدة واحدة) **تقريبا**. مثلا: معامل عدد سنوات التكوين educ يفسر هكذا: في المتوسط فإن سنة تكوين إضافية تؤدي إلى زيادة الأجر ب 7.47 بالمائة تقريبا، أي أن مردود سنة تكوين في 78 هو 7.5% تقريبا، مردود سنة تكوين في 85 هو 9.35% تقريبا أي(1.85 + 7.5).

- **قم باختبار تغير مردود التعليم educ بين 78 و85.**

الجواب: إحصائية الاختبار t هي نسبة المعامل إلى خطئه المعياري: 1.97 = 0.0185/0.0094، لذلك فهي دالة عند مستوى معنوية 5% (اختبار ثنائي). نعم هناك تغير في مردود التكوين، وبما أن المعامل 0.0185 موجب فالتغير هو بالزيادة.

- **بكم كان يختلف أجر النساء عن الرجال في 78 في المتوسط؟** الجواب: في 78 كان أجر المرأة يقل عن أجر الرجل في المتوسط بنسبة 31.7% تقريبا (التقدير الحقيقي هو 27.2%= 1 − (0.317−)exp)،
- **قم بقياس واختبار إنخفاض فارق الأجر بسبب الجنس بين الفترتين.**

الجواب: في 85، أجر المرأة باللغرتم يقل عن أجر الرجل في المتوسط ب 0.232 (−0.317 + 0.085).

إحصائية الاختبار تساوي 1.67 = 0.085/0.51، فهي دالة عند مستوى معنوية 5%، ($t_{tab}$ = 1.64) في الإختبار الأحادي من اليمين، حيث الفرضية البديلة هي أن معامل التفاعل أكبر من 0).

- **ماذا يحدث إذا فاعلنا كل المتغيرات المفسرة بالمتغيرة y85 في الدالة أعلاه؟**

الجواب: هذا يماثل استخدام دالتين، أحداهما لسنة 78 والأخرى لسنة 85. يمكن عند ذلك المقارنة بين الدالتين، واستخدام اختبار شاو Le test de Chow لوجود اختلاف بين الدالتين. كما ذكرنا سابقا، لا يعود التجميع مفيدا إذا كانت علاقة المتغيرة التابعة ثابتة عبر الزمن على الأقل ببعض المتغيرات المستقلة، وإلا (إذا كانت كل العلاقات تتغير) يصبح استخدام نموذج لكل سنة هو الحل بدون الحاجة للتجميع (~إضافة إلى أن استخدام نماذج متعددة، واحد لكل سنة، يعني تقليص حجم العينة لكل نموذج، تماما كما أن تعدد المتغيرات بإدراج التداخلات يتطلب عينة أكبر).

- **كيف يمكن احتساب التضخم في هذا النموذج (قيمة العملة بعد 7 سنوات لا تبقى على حالها، فكيف نتأكد أن التضخم لا يؤثر على المعاملات)؟**

الجواب: الطريقة المباشرة هي أن نستخدم القيمة الحقيقية لا الإسمية للأجر، وذلك باحتساب بيانات أجر عينة 85 بعملة 78، وذلك عن طريق التكميشdéflateur (في الدراسة كان معامل التكميش هو 1.65 = 107.6/65.2)، أي قسمة كل بيانات الأجر ل 85 على المعامل 1.65. لكن هذا ليس ضروريا طالما أن لدينا المفسرة y85 (تغيير الثابت بين الفترتين) وأن المتغيرة التابعة هي لغرتم الأجر. إذا لم نستخدم تغيير الثابت فإن استخدام الأجر الإسمي يمكن أن يعطي نتائج خاطئة، وإذا لم نستخدم اللغرتم يتعين استخدام متغيرة تمثل السنة واستخدام بيانات الأجر الحقيقي لا الإسمي. التفسير هو أن استخدام الأجر الحقيقي (مكمش) لا يؤثر هنا على المعاملات وإنما فقط على الثابت:

$$Log(wage/1.65) = log(wage) - log(1.65)$$

## 2-3. فرق الفروق (أثر السياسة أو الحدث)

التفاعل بين الزمن والمتغيرة له تطبيقات عديدة، خاصة عندما تأتي البيانات من **تجربة طبيعية (أو شبه طبيعية)**. الهدف هو قياس أثر سياسة أو حدث ما على المفردات (أفراد، مؤسسات، عائلات، مدن، ...)، وللقيام بذلك يتم تكوين عينتين، عينة مراقبة échantillon de contrôle، لا تخضع للحدث، ولنرمز لها ب C، وعينة معالجة échantillon de traitement تخضع للحدث، نرمز لها ب T. في الوقت نفسه، نحتاج إلى بيانات كل من العينتين قبل وبعد المعالجة (أو الحدث أو السياسة). بهذا يكون لدينا أربع عينات: **عينة المراقبة قبل؛ عينة المراقبة بعد؛ عينة المعالجة قبل؛ وعينة المعالجة بعد**.

**مثلا**، لاختبار تأثير عملية تأهيل المؤسسات على أدائها، نأخذ أداء عينة من المؤسسات التي خضعت للتأهيل (عينة معالجة) وعينة من المؤسسات التي لم تخضع للتأهيل (عينة مراقبة). في الوقت نفسه، نحتاج أن تقسم البيانات إلى عينتين قبل وبعد فترة التأهيل لاحتساب الفروق التي تميز العينتين، أي قبل الحدث وبعد الحدث. في

المجموع سيكون لدينا أربع مجموعات من المؤسسات: عينة المراقبة قبل؛ عينة المراقبة بعد؛ عينة المعالجة قبل؛ وعينة المعالجة بعد.

لتمييز هذه البيانات نستحدث متغيرتين:

- نستحدث متغيرة dT: تساوي 1 لمشاهدات عينة المعالجة، و0 لغيرها،
- نستحدث متغيرة d2: تساوي 1 لمشاهدات العينة الثانية في الزمن، أي التي تلي الحدث، و0 للأولى.

النموذج الذي يمكن من احتساب أثر الحدث على المفردات التي خضعت له يكون كالتالي(difference in differences):

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1\ d2.dT + \beta_2 x_2 + \ldots + \beta_J x_J + e$$

المعامل $\beta_0$ يقيس متوسط عينة المراقبة قبل الحدث.

$\beta_0 + \delta_0$ يعطي متوسط عينة المراقبة بعد الحدث.

المعامل $\delta_0$ يقيس الفرق بين الفترتين لدى **عينة المراقبة**.

المعامل $\beta_1$ يقيس الفرق بين عينتي المراقبة والمعالجة قبل الحدث (لماذا قبل؟ ج: تذكر أنه معامل جزئي، فهو يدرج المعالجة دون الزمن).

المعامل $\delta_1$: **يقيس أثر السياسة أو الحدث على المفردات التي خضعت له** (وهذا الذي نبحث عنه، لأنه يقيس تغير الفرق بين عينة المعالجة وعينة المراقبة بعد خضوع الأولى للحدث).

**في حالة عدم وجود عوامل أخرى**: إذا لم يدرج في النموذج عوامل أخرى، فقط الفترة والمعالجة، فإن $\delta\hat{}_1$ يقدر فرق الفروق، أو ما يسمى مقدر الفروق المضاعفة، ويمكن حسابه كما يلي:

$$\delta\hat{}_1 = (My_{2.T} - My_{2.C}) - (My_{1.T} - My_{1.C})$$

المقدر $\delta\hat{}_1$ يسمى أحيانا متوسط أثر المعالجة (effet moyen du traitement)، هو الفرق بعد المعالجة مطروحا منه الفرق قبل المعالجة (بين متوسطي عينتي المراقبة والمعالجة)، وهو يحسب كما يلي:

- نحسب في كل فترة الفرق بين متوسطي عينتي المعالجة والمراقبة
- نحسب الفرق بين الفرقين.

يمكن أيضا إجراء الحساب كما يلي:

$$\delta\hat{}_1 = (My_{2.T} - My_{1.T}) - (My_{2.C} - My_{1.C})$$

- أي نحسب الانتقال الزمني لكل من عينة المراقبة والمعالجة
- ثم نحسب الفرق بين الفرقين.

بتعويض متغيرتي التفسير بقيمهما في الدالة يمكن تمثيل نتائج التقدير وعادة ما يتم في جدول كالتالي:

| | قبل | بعد | بعد ناقص قبل |
|---|---|---|---|
| عينة المراقبة | $\beta_0$ | $\beta_0 + \delta_0$ | $\delta_0$ |
| عينة المعالجة | $\beta_0 + \beta_1$ | $\beta_0 + \delta_0 + \beta_1 + \delta_1$ | $\delta_0 + \delta_1$ |

| | | | |
|---|---|---|---|
| $\delta_1$ | $\beta_1 + \delta_1$ | $\beta_1$ | المعالجة – المراقبة |

في حالة إدراج عوامل أخرى في النموذج، لاحتساب الفروق الموجودة في العينة عبر الزمن، فإن قيمة المعامل $\delta_1$ لا تحسب بالطريقة البسيطة (الفرق بين الفرقين) لكن تفسيره يبقى مشابها.

**مثال[1] 1 (مركز الردم)**. درس كايل وماكلاين (Kiel et McClain, 1995) أثر وجود مركز ردم نفايات على سعر الشقق في نورث أندوفر North Andover، ماساشوستس، الولايات المتحدة. نشأت الإشاعة عن إنشاء المركز في 1978، وبدأ البناء في 1981، وبدأ المركز في العمل في 1985. البيانات المستخدمة هي لأسعار الشقق المباعة في 78 و82، وهما عينتان مستقلتان. الفرضية المختبرة هي أن الشقق المأجرة بقرب المركز (أقل من 3 أميال) أقل سعرا من الأبعد. سوف نستخدم السعر الحقيقي rprice عن طريق مؤشر السعر للعقار.

سوف نستخدم عدة دوال لكي نرى الاختلاف في التحليل وفائدة استخدام البيانات المجمعة، وتفسير المعاملات.

**<u>الدالة الأولى</u>: استخدام بيانات 1981** فقط (بداية البناء)، وnearinc متغيرة وهمية ثنائية تأخذ 1 للشقق القريبة من المركز و0 للبعيدة.

**^rprice = 101307.5 – 30688.27 nearinc**

(3092) (5827.71) ;

N = 142; $R^2$ = 0.165.

- **فسر الثابت في الدالة أعلاه. ج:** الثابت هو متوسط سعر الشقق البعيدة عن المركز سنة 81، أي عند بداية البناء (هذا التفسير "الجيد" للثابت ممكن لأن النموذج لا يتضمن متغيرات مفسرة أخرى).

- **فسر معامل الانحدار الجزئي في الدالة.**

معامل الانحدار الجزئي: 30688.27– هو متوسط فرق السعر للشقق القريبة من المركز مقارنة بالبعيدة منه.

- **إختبر وجود فرق في السعر بين الشقق القريبة والبعيدة (سنة 81).**

الجواب: t = 30688.27/5827.71 = 5.26، وهي قيمة عالية، أكبر من القيمة الجدولية (لاختبار ثنائي القيمة الجدولية هي تقريبا 1.96)، وبالتالي يمكن أن نرفض فرضية تساوي متوسط سعر الشقق من الفئتين (ونستدل بالتالي على وجود فرق في السعر بين الشقق القريبة والبعيدة عن المركز).

- **هل تدل النتيجة السابقة على أن خبر بناء المركز أدى إلى إنخفاض أسعار الشقق القريبة منه؟ وكيف يمكن التحقق من ذلك؟**

الجواب: ليس بالضرورة، يمكن أن يكون المركز بني بالقرب من الشقق منخفضة السعر، وفي هذه الحالة ليس هو المسؤول عن تدني أسعارها لأنها كانت متدنية أصلا قبل بنائه. للتحقق من ذلك يتعين اختبار الفرق بين الأسعار قبل وجود إشاعة البناء، أي استخدام بيانات 78.

- **<u>الدالة الثانية</u>: بيانات سنة 78** أعطت الدالة التالية:

[1] Wooldridge J. M., op. cit. 2015, p. 649.

**^rprice = 82517.23 – 18824.37 nearinc**

(2653.79) (4744.59)

N = 179; $R^2$ = 0.082.

• **فسر الثابت في الدالة أعلاه. ج:** الثابت هو متوسط سعر الشقق البعيدة عن المركز سنة 78، أي قبل بداية البناء (مرة أخرى هذا التفسير "الجيد" للثابت ممكن فقط لأن النموذج لا يتضمن متغيرات أخرى).

• **هل كانت أسعار الشقق القريبة من المركز منخفضة مقارنة مع البعيدة عنه حتى قبل إنتشار إشاعة إنشائه؟ كيف يمكن اختبار ذلك؟**

الجواب: نعم، حتى قبل إشاعة بناء المركز كانت أسعار الشقق المعنية منخفضة بفارق 18824.37$−

(وهو معامل الانحدار في الدالة أعلاه) لاختبار الفرق، نقسم المعامل على الخطأ المعياري، فنجد: t = −3.97 = 18824.37/4744.59−، وهي قيمة بعيدة عن 1.64−، إذن الفرق دال بمستوى معنوية 5% (اختبار أحادي وليس ثنائي لأننا نختبر الانخفاض وليس الاختلاف).

• **هل ارتفعت أسعار الشقق البعيدة عن المركز أم انخفضت بين 78 و81؟**

الجواب: إرتفعت من متوسط **82517.23 إلى 101307.5 أي بفارق 18790.27**

• **هل ارتفعت أسعار الشقق القريبة من المركز أم انخفضت بين 78 و81؟**

الجواب: أسعار الشقق **القريبة من المركز** أيضا ارتفعت:

في 78 متوسط سعرها: (82517.23 – 18824.37 = 63692.86)

في 81 متوسط سعرها: (101307.5 – 30688.27 = 70619).

لكن مقدار ارتفاع الشقق القريبة (70619 – 63692.86 = 6926.14) هو أقل من الارتفاع الذي عرفته الشقق البعيدة عن مركز الردم (18790.27).

• **كيف يمكن من الدالتين أعلاه حساب فرق السعر العائد لبناء المركز معزولا عن فرق المنطقة وفرق الزمن؟**

الجواب هو فرق الفروق، ويمكن حسابه بطريقتين:

ط1. متوسط الفرق بين ارتفاع الشقق القريبة وارتفاع الشقق البعيدة:

$$\hat{d}=6926.14 - 18790.27 = -11863.9$$

ارتفاع الشقق القريبة اقل ب 11863.9 من ارتفاع الشقق البعيدة.

تسمى هذه القيمة **مقدر فرق الفروق** estimateur de la différence des différences

هذا يعني أن وجود المركز ربما أثر على أسعار الشقق القريبة منه بأن حد من ارتفاع أسعارها مقارنة مع الشقق البعيدة. المشكلة هي أننا لا نستطيع أن نعرف إن كان هذا التأثير دالا إحصائيا، لأن الطريقة لا تعطينا الخطأ المعياري للمقدر $\hat{d}$ .

ط2: متوسط الفرق بين اختلاف السعر في 78 واختلافه في 81:

في 78 الاختلاف بين المنطقتين كان 18824.37 وفي 81 صار 30688.27، نحسب الفارق بين الفرقين،

$$\hat{d} = 18824.37 - 30688.27 = -11863.9$$

من الواضح أن الفارق في السعر بين المنطقتين زاد عما كان عليه قبل بناء المركز، أي أن مركز الردم كان له أثر على أسعار الشقق القريبة منه. المشكة في الطريقة هي دائما عدم القدرة على الاختبار الإحصائي للهذا الأثر، وهنا تظهر فائدة نموذج البيانات المقطعية المجمعة.

- **كيف يمكن اختبار إن كان هذا الفرق (الفرق العائد للمركز) دالا إحصائيا؟**

الجواب. حصلنا على الفرق الذي يقيس الأثر الصافي لبناء المركز. بقي أن نختبر هل هو دال، نحتاج إلى إدراج متغيرة جديدة في الدالة هي التفاعل بين الزمن والمنطقة، ومن ثم نقسم معامل هذه المتغيرة على خطئه المعياري.

- **أكتب الدالة التي تعطي تأثير بناء المركز على أسعار الشقق بمعزل عن المنطقة والزمن، وفسر معاملاتها، واختبر تأثير بناء المركز.**

الدالة الثالثة: ندخل **بيانات مجمعة 78 و81**، وندرج التفاعل بين الزمن والمنطقة كمتغيرة جديدة:

$$\hat{rprice} = \beta_0 + \delta_0\, y81 + \beta_1\, nearinc + \delta_1\,(y81 * nearinc)$$

$\beta_0$ : متوسط سعر الشقق في 78 في المنطقة البعيدة عن المركز.

$\delta_0$ : الزيادة في أسعار الشقق البعيدة عن المركز خلال الفترة بين 78 و 81.

$\beta_0 + \delta_0$ : **متوسط أسعار الشقق البعيدة** عن المركز بعد البناء أي في 81.

$\beta_1$: الفرق في أسعار الشقق بسبب المنطقة (القرب من المركز) بصرف النظر عن البدء في بناء المركز (independamment de l'existance de l'incinérateur).

**$\delta_1$: التغير في أسعار الشقق بسبب المنطقة (وهو المطلوب)** عند افتراض أن أسعار الشقق القريبة والبعيد لم تتطور بشكل مختلف لأسباب أخرى (لأننا لم ندخل مفسرات أخرى). الدالة جاءت كالتالي:

$$\hat{rprice} = 82517.23 + 18790.29\, y81 - 18824.37\, nearinc - 11863.9\,(y81*nearinc)$$
$$(2727.91) \quad (4050.07) \quad (4874.32) \quad (7456.65)$$

$N = 321,\ R^2 = 0.174,$

لاحظ أولا أن حجم العينة هو مجموع حجمي عينتي النموذجين السابقين $142 + 179 = 321$

لاحظ أيضا أن قيمة المقدر $\delta_1 = -11863.9$ هي ذاتها قيمة $\hat{d}$. الفائدة هنا هي أنه لدينا الخطأ المعياري لهذا المقدر، أي $SE(\delta_1)$ ، وبالتالي يمكن استخدامها لاختبار تأثير بناء المركز على أسعار الشقق القريبة منه:

$$t = \delta_1 / SE(\delta_1) = -11863.9/7456.65 = -1.59$$

هذه القيمة لإحصائية الاختبار 1.59− هي أقل من القيمة الجدولية لاختبار أحادي عند مستوى معنوية 5 بالمائة (1.64−)، وبالتالي نقول بأن المعامل 11863.9غير دال في اختبار وجود تأثير سلبي لوجود المركز على أسعار الشقق القريبة منه (pvalue = 0.057). بعبارة أخرى، حسب نتيجة هذا الاختبار، لا يمكن الاستدلال على وجود تأثير سلبي لوجود المركز على أسعار الشقق القريبة منه (ولا نقول "نستدل على عدم وجود تأثير" ! ليس الأمر سيان !، ولذلك يمكن إجراء اختبارات أخرى، من خلال تعديل النموذج مثلا بإدراج مفسرات أخرى).

- **هل يمكن أن تتغير نتيجة الاختبار في حالة إدارج متغيرات أخرى؟ وهل إدارج متغيرات أخرى مهم؟**

قام كايل وماكلاين (1995) بإدخال خصائص متنوعة لسوق العقار (عدد الغرف، المساحة، ...)، هذه الإضافة مبررة لسببين: أولا لأن أنواع الشقق المباعة في 78 قد تكون مختلفة عن الأنواع المباعة في 81، وبالتالي وجب أخذ هذا في الحسبان. ثانيا، حتى إن لم تكن خصائص الشقق المباعة تغيرت (تأثيرها غير دال)، فإن إدخالها يمكن أن **يقلص تباين الخطأ**، مما يمكن أن يؤدي إلى **تقليص الخطأ المعياري للمعامل** $\delta_1$.

- **علق على نتائج كايل وماكلاين في النماذج الثلاث أدناه.**

في النموذج الثاني تم إدخال عمر الشقق، وفي النموذج الثالث تم إدخال: العمر، عدد الغرف، عدد غرف الماء، المساحة المنزل، مساحة الأرض، المسافة عن حدود الولاية. المتغيرة التابعة هي السعر الحقيقي rprice

| المتغيرات | النموذج(1) | النموذج(2) | النموذج(3) |
|---|---|---|---|
| Constante | 82517.23<br>(2727.91) | 89116.54<br>(2406.05) | 13807.67<br>(11166.59) |
| Y81 | 18790.29<br>(4050.07) | 21321.04<br>(3443.63) | 13928.48<br>(2789.75) |
| Nearinc | −18824.37<br>(4874.32) | −9397.97<br>(4812.22) | −3780.34<br>(4453.42) |
| Y81 * nearinc | −11863.90<br>(7456.65) | −21920.27<br>(6359.75) | −14177.93<br>(4987.27) |
| Autres facteurs | Non | age, age² | age, room, bath, surf, land, dist. |
| N | 321 | 321 | 321 |
| $R^2$ | 0.174 | 0.414 | 0.660 |

Source: Cengage learning. 2013, in Wooldridge J. M., 2015.

**كيف تغير $R^2$ في الدالتين المواليتين؟**

**الجواب:** كما هو متوقع، الدالتين الثانية والثالثة أعطتا نسبة تباين مفسر أكبر.

- **ما هي نتيجة اختبار تأثير المركز في النموذجين الثاني والثلاث؟**

**ج:** في النموذج الثاني، على عكس النموذج الأول، جاء اختبار تأثير التداخل دالا:

$$t = -21920.27/6359.75 = -3.45$$

في النموذج الثالث كذلك: $t = -14177.93/4987.27 = -2.84$

**فسر التغير في المعامل nearinc!**

**ج:** معامل nearinc (أثر المنطقة بصرف النظر عن إنشاء المركز) انخفض بشدة من النموذج الأول إلى النموذجين الثاني والثالث، وهذا يعني أن العوامل الإضافية التي تم إدراجها لها تأثيرها.

- **أي النماذج أفضل؟**

النموذج الثالث أفضل لأنه يأخذ في الحسبان عوامل أكثر ولأن الخطأ المعياري لمعاملاته أقل (تقير دقيق)، ما عدا بالنسبة للثابت، والذي ليس له أهمية هنا (ليس له بعد معنى اقتصادي مثل المعني الذي يكتسبه في النموذج الأول: متوسط سعر ... فهو مجرد ثابت طالما أن النموذج يتضمن متغيرات الخصائص مثل المساحة وغيرها).

- **عودة إلى الدالة الثالثة: كيف يصبح التفسير لو حولنا المتغيرة التابعة price إلى اللغرتم بدلا من استخدام السعر الحقيقي rprice:**

الجواب: تفسير $\delta_1$ يصبح كما يلي: $(100)\delta_1$ تعطي تقريبا نسبة الفرق في سعر الشقق العائد للمركز.

باستخدام نفس المعطيات تأتي الدالة كما يلي:

$$\widehat{Log(price)} = 11.29 + 0.457\ y81 + 0.340\ nearinc + 0.063\ y81 * nearinc$$
$$(0.31)\quad (0.045)\quad (0.055)\quad (0.083)$$

$N=321,\ R^2 = 0.409.$

معامل حد التداخل يظهر أنه بسبب بناء المركز إنخفضت أسعار الشقق القريبة منه ب 6.3%، وهي غير دالة (أي نفس نتيجة الاختبار مع rprice. تم أيضا إدخال المتغيرات الأخرى في الدالة الأخيرة (الخصائص)، فأصبحت النسبة 13.2%.

**ماذا لو استخدمنا log(rprice)؟**

النتيجة هي ذاتها، الفرق فقط في الثابت: إستخدام لغ السعر الإسمي log(price) أو لغ السعر الحقيقي log(rprice) لا يغير إلا الثابت فقط.

**مثال2**. (مدة النقاهة[1]). درس ماير، فيسكوسي، ودوربن (Meyer, Viscusi, et Durbin ; 1995) المدة بالأسبوع للنقاهة التي يحصل عليها العامل (تعويض) بعد حادث عمل (durat). رفعت ولاية كانتوكي في 1980 سقف الدخل الأسبوعي الذي يغطيه نظام تعويضات حوادث العمل للموظفين ذوي الدخل العالي، علما أن الإجراء لم يمس ذوي الدخل المنخفض. للقيام بالدراسة تم تكوين عينة مراقبة (ممن لا يمسهم الإجراء)، وعينة معالجة (ممن يعنيهم الإجراء أي من ذوي الدخل العالي). من خلال استطلاع عينة عشوائية من الموظفين قبل وبعد الإجراء، أراد الباحثون اختبار ما إذا كان المستفيدون منه استخدموا مدة نقاهة أطول بسبب الإجراء. قام الباحثون أولا بتقدير نموذج فرق فروق باستخدام لغرتم المدة log(durat) كمتغيرة تابعة. المتغيرة Afchnge هي متغيرة ثنائية تعبر عن الفترة قبل أم بعد السياسة، والمتغيرة highearn تعبر عن فئة الدخل (1 = دخل عالي، 0 = لا).

النتائج كانت كالتالي:

$$\mathbf{\widehat{log(durat)} = 1.126 + 0.0077\ aftchnge + 0.256\ highearn + 0.191\ aftchnge * highearn}$$
$$SE = (0.031)\quad (0.0447)\quad (0.047)\quad (0.069)$$

$N = 5626;\ R^2 = 0.021.$

1. **كم هي نسبة التغير في مدة النقاهة للفئة المستفيدة نتيجة للإجراء؟ هل التغير دال إحصائيا؟**
2. **كم هي نسبة التغير في مدة النقاهة لغير المعنيين بالإجراء؟ علق على قيمتها؟**
3. **كيف يفسر معامل highearn؟**
4. **كم هي نسبة التباين المفسر بالنموذج؟ علق عليها.**

[1] Wooldridge J. M., op. cit. 2015, p. 654.

**5. علق على حجم العينة.**

الجواب:

- $\delta\hat{}_1 = 0.191$ تمثل نسبة التغير في مدة النقاهة لدى الفئة المستفيدة وهي 19% تقريبا، وهي نسبة معتبرة. بقسمة هذه القيمة على الخطأ المعياري نحصل على: ($t = 0.191/0.069 = 2.77$)، الزيادة إذن دالة إحصائيا. (النسبة الحقيقية هي $\exp(0.191) - 1 = 0.21$ أي 21 بالمائة).
- معامل aftchnge يمثل أثر السياسة على غير ذوي الدخل العالي أي الذين لا يعنيهم الإجراء (عينة المراقبة)؛ نسبته 0.77%، وهي نسبة شبه معدومة، وذلك طبيعي لأن الإجراء لم يمسهم.
- معامل highearn يمثل الفرق في مدة الاجازة عند ذوي الدخل العالي، ويعني أن ذوي الدخل العالي يأخذون إجازات نقاهة أطول حتى قبل الإجراء المتخذ لفائدتهم (aftchnge =0)، حوالي 29.2% في المتوسط: $\exp(0.256) - 1 = 0.2917$
- نسبة التباين المفسر بالنموذج ضعيفة، 2.1%، وهذا طبيعي لأن العوامل التي تأثر في مدة النقاهة عديدة، أهمها طبيعة وشدة الحادث، ... لحسن الحظ، كبر حجم العينة سمح بالحصول على نتائج دالة، رغم ضعف نسبة التباين المفسر. لاحظ أن الدراسة سمحت باستكشاف تأثير السياسة على التابع (مدة العطل المرضية) رغم أن معظم تباين هذه الأخيرة غير محتسب بالنموذج.
- للإشارة، أخذ الباحثون في نموذج آخر عدة عوامل أخرى، مثل الجنس، والعمر، والحالة العائلية، والقطاع، ونوع الجروح. إدراج هذه العوامل لم يكن له إلا تأثير خفيف على المعامل $\delta_1$.

رأينا في هذا الفصل أهمية البيانات المجمعة وكيف تستخدم لإظهار تطور المتغيرة التابعة في الزمن من خلال إدراج متغيرات وهمية ثنائية، وكيفية قياس تطور تأثير العوامل الأخرى من خلال تفاعل العوامل مع الزمن.

## خلاصة

تطرقنا للمفاهيم الأساسية في موضوع بيانات البانل وشرحنا الفرق بين البيانات المقطعية المجمعة وبيانات البانل وكيفية تنظيمها في جدول البيانات. تطرقنا أيضا لترميز المتغيرة الاسمية من خلال متغيرات وهمية ثنائية، وهو أمر نستخدمه في باقي الفصل والمقياس عموما.

درسنا في هذا الفصل أيضا طرق معالجة البيانات المقطعية المجمعة مستقلة. طريقة OLS على البيانات المجمعة هي الطريقة الأساسية للتقدير والعمليات المعتادة للاستدلال صحيحة حتى في حالة عدم التجانس (الارتباط الذاتي ليس مشكلة لأن العينات مستقلة خلال الزمن). كثيرا ما يتم إدراج عدة ثوابت ويمكن أيضا إدراج تفاعل الزمن مع بعض المتغيرات المفسرة الرئيسية لستطلاع كيف تغيرت خلال الزمن، وهو أمر مهم في تقييم السياسات العمومية.