المحاضرة الأولى

المحور الأول: الانـــحدار الخطـــي البســـــيط

Simple Linear Regression

1.2 مقدمـــــــــــــة:

ان دراسة العلاقة بين المتغيرات الاقتصادية يتطلب تحديد المتغيرات المؤثرة فـــي تلـــك العلاقة ومن ابسط واسهل أنواع العلاقات في التقدير والتحليل الإحصائي والاقتصادي، العلاقة بين متغيرين أحدهما المتغير التابع، Dependent Variable، والثاني المتغير المستقل، Independent Variable، وإذا رمزنا للمتغيرين بـــY و X وعلى التوالي، فان العلاقة الدالية التي تجمعهما تكون كالآتي:

$$Y = F(x) \qquad \ldots(1.2)$$

حيث يشير الرمز F الى كون المتغيرالتابع Y يعتمد على المتغير المستقل X. ولتحديد شـــكل العلاقـــة هـــذه – مـــا إذا كان خطياً أم غير خطي- يمكن الاستعانة بالنظرية الاقتصـــادية، كمـــا يستعان بالاقتصاد الرياضي والإحصاء لصياغة العلاقة واختبار المتغـــيرات، كما لابد من تحديد شكل العلاقة هذه إذ تحكم العلاقة بين المتغيرات بعدد من الأشكال (الصيغ) ابسطها وأكثرها شيوعاً الصيغة الخطية، وتسمى العلاقة الخطية بيـــن متغيرين بالانحدار الخطي البسيط، Simple Linear Regression، فالعلاقة الخطية بين X ولتكن دخل الأسرة وY ولتكن الأنفاق على سلعة معينة يمكن ان تكتب بالصيغة الرياضية الآتية:

$$Y_i = B_0 + B_1 X_i \qquad \ldots(2.2)$$

حيث :

B_1 و B_0 عبارة عن معلمات مجهولة القيم وثوابت يُشرحان من وجهة النظر الرياضية كالآتي:

B_0: تمثل تقاطع خط الانحدار مع المحور العمودي وهي عبارة عن القيمة التي تتخذها Y عندما تكون قيمة X مساوية للصفر.

B_1: تمثل الميل.

ومن وجهة النظر الاقتصادية تمثل (B_0) حالة الكفاف و(B_1) الميل الحدي للاستهلاك، وقيمة الميل عبارة عن مقدار الزيادة المتحققة في قيمة المتغير التابع Y نتيجة زيادة المتغير المستقل بمقدار وحدة واحدة.

غير ان العلاقة أعلاه (2.2) لا يمكن ان تشرح العلاقة بين المتغيرين بشكل دقيق، فهناك أسباب مهمة تجعل هذه المعادلة غير معبرة عن العلاقة بين X و Y تعبيراً كاملاً فقد يكون هناك انحراف بين العلاقة الحقيقية والمعادلة الإحصائية التي تمثلها نتيجة أخطاء في القياسات أو في اختيار المتغير المستقل، مما يتطلب إضافة متغير جديد يسمى بالحد العشوائي، Random Variable ويرمز له عادةً بالرمز (U) ودوره امتصاص العوامل غير القابلة للقياس، وكذلك أخطاء القياس، عليه فان العلاقة من الصيغة (2.2) يجب ان تعدل لكي تضم حد الخطأ العشوائي حيث يصبح:

$$Y_i = B_0 + B_1 X_i + U_i \qquad \ldots(3.2)$$

2.2 الفرضيات الخاصة بالمتغير العشوائي:

1- ان المتغير العشوائي (U_i) هو متغير تعتمد قيمته في أية فترة زمنية على عامل الصـــدفة، فقد تكون اكبر أو اصغر أو مساوية إلى الصفر، إلا أنها في المتوسط تساوي صفر، أي $E(U_i)=0$، ويمكن توضيح ذلك على النحو الآتي:

$$Y_i = B_0 + B_1 X_i + U_i \quad \ldots(4.2)$$

$$U_i = Y_i - B_0 - B_1 X_i \quad \ldots(5.2)$$

وبإدخال Σ على طرفي المعادلة 5.2:

$$\Sigma U_i = \Sigma(Y_i - B_0 - B_1 X_i)$$

$$\Sigma U_i = \Sigma Y_i - nB_0 - B_1 \Sigma X_i \quad \ldots(6.2)$$

$$\because B_0 = \overline{Y} - B_1 \overline{X}$$

نعوّض عن B_0 بما يساويها في المعادلة (6.2):

$$\Sigma U_i = \Sigma Y_i - n(\overline{Y} - B_1 \overline{X}) - B_1 \Sigma X_i \quad \ldots(7.2)$$

$$\because \overline{Y} = \frac{\Sigma Y_i}{n} \quad , \overline{X} = \frac{\Sigma X_i}{n}$$

وبحاصل ضرب الطرفين في الوسطين نحصل:

$$\Sigma Y_i = n\overline{Y} \quad , \Sigma X_i = n\overline{X}$$

وبالتعويض عن ذلك في المعادلة (7.2) تكون:

$$\sum U_i = \sum Y_i - \sum Y_i + B_1 \sum X_i - B_1 \sum X_i$$

$$\sum U_i = 0$$

$$E(U_i) = 0$$

2- ان المتغـــير العشوائي (U_i) يتوزع توزيعاً طبيعياً، Normally distributed، حـــول القيمة المتوقعة أو حول الوسط الحسابي المساوي للصفر عند كل قيمة من قيم المتغير المستقل X أي بشكل جرس.

3- ان تبايـــن، Variance، المتغـــير العشوائي (حد الخطأ)، حول الوسط الحسابي مقدار ثابت عند كل قيمة من قيم X أي:

$$var(U_i) = E[U_i - E(U_i)]^2$$

$$\because E(U_i) = 0$$

$$\therefore var(U_i) = E(U_i)^2 = 6^2$$

وإذا كان تباين الخطأ غير ثابت عندئذ تظهر مشكلة تسمى مشكلة عدم تجانس التباين، Heteroscedasticity، والتي سنتناولها بشيء من التفصيل لاحقاً.

الفرضيات الثلاث السابقة يمكن جمعها بشكل مختصر وتمثيلها كالآتي:

$$U_i \sim N(0, \sigma^2)$$

أي بمعـــنى ان الخطـــأ العشوائي، U_i، يتوزع، ~، توزيعاً طبيعياً، N، بوسط حسابي مساوي للصفر، 0، وتباين ثابت قيمته σ^2.

4- أن قـــيم U_i غير مرتبطة بأي من المتغيرات المستقلة، أي انعدام التباين المشترك Covariance بين U_i وX_i أي:

$$Cov(U_i X_i) = E(U_i X_i)$$

$$Cov(U_i X_i) = X_i E(U_i)$$

$$\because E(U_i) = 0$$

$$\therefore Cov(U_i X_i) = 0$$

ويمكن توضيح ذلك على النحو الآتي:

$$Y_i = B_0 + B_1 X_i + U_i \qquad \ldots(8.2)$$

$$U_i = Y_i - B_0 - B_1 X_i \qquad \ldots(9.2)$$

وبضرب طرفي المعادلة بـ $\sum X_i$:

$$\sum X_i U_i = \sum X_i Y_i - B_0 \sum X_i - B_1 \sum X_i^2$$

$$\because B_0 = \bar{Y} - B_1 \bar{X}$$

وعند تعويض ذلك:

$$\sum X_i U_i = \sum X_i Y_i - \sum X_i (\bar{Y} - B_1 \bar{X}) - B_1 \sum X_i^2 \qquad \ldots(10.2)$$

$$\because \bar{Y} = \frac{\sum Y_i}{n} \quad , \quad \bar{X} = \frac{\sum X_i}{n}$$

وبتعويض ذلك يكون:

$$\sum X_i U_i = \sum X_i Y_i - \sum X_i \left(\frac{\sum Y_i}{n} - B_1 \frac{\sum X_i}{n}\right) - B_1 \sum X_i^2$$

$$\sum X_i U_i = \sum X_i Y_i - \sum X_i \left(\frac{\sum Y_i - B_1 \sum X_i}{n}\right) - B_1 \sum X_i^2$$

$$\sum X_i U_i = \sum X_i Y_i - \sum X_i Y_i + B_1 \sum X_i^2 - B_1 \sum X_i^2$$

وبعد الحذف والتبسيط يكون:

$$\sum X_i U_i = 0$$

5- القيم المختلفة للمتغير العشوائي (U_i) تكون مستقلة عن بعضها البعض بعبارة أخرى التباين المشترك لـ U_i مع U_j مساوية للصفر، وعليه فان قيمة العنصر العشوائي في أي فترة لا تعتمد على قيمته في فترة أخرى أي:

$$Cov(U_i U_j) = E(U_i U_j) = 0 \qquad (i, j = 1,2,3,\ldots,n,\ i \neq j)$$

وإذا حدث وجود ارتباط بينها تظهر مشكلة تسمى مشكلة الارتباط الذاتي، Autocorrelation، وسيتم شرحها لاحقاً.

6- انعدام العلاقة بين المتغيرات المستقلة وفي حالة وجود علاقة قوية بينها تظهر مشكلة تسمى مشكلة الارتباط الخطي المتعدد، Multicollinearity، والتي سيتم تناولها فيما بعد.

3.2 طريقة المربعات الصغرى،

:The Ordinary Least Squares (OLS)

بالرجوع إلى العلاقة الخطية بين دخل الأسرة X وانفاقها على سلعة معينة Y.

$$Y_i = B_0 + B_1 X_i + U_i \qquad \ldots(11.2)$$

يتبين لنا بان تأثير الدخل في الأنفاق على السلعة موضوع البحث يتحدد من خلال العلاقة المنتظمة $(B_0 + B_1 X_1)$، أما تأثير العوامل الأخرى فانه متجسد في (U_i).

وعليه فانه لمعرفة العلاقة الحقيقية بين دخل الأسرة وانفاقها على السلعة في القطر يتطلب احتساب B_0 و B_1، إلا أن احتساب المعالم المذكورة لا يمكن ان يتم إلا في حالة الحصول على دخل وانفاق جميع الأسر في ذلك القطر وهذا أمر غير ممكن

بسبب صعوبة العملية الإحصائية اللازمة ولتسهيل العمل نسحب عينة من أسر القطر، ومن ثم نقدر قيم المعالم ويتم التقدير بواسطة المعادلة:

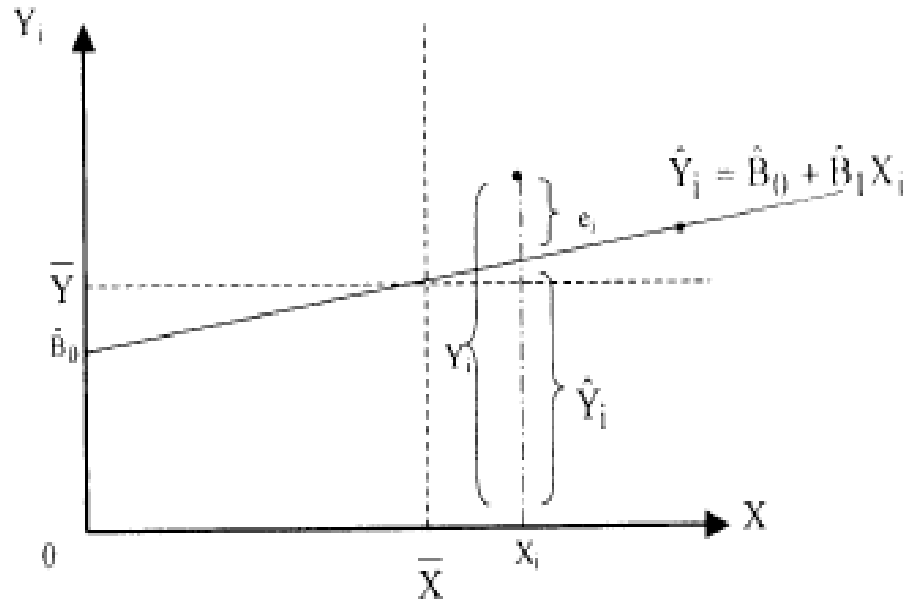
$$Y_i = \hat{B}_0 + \hat{B}_1 X_i + e_i \qquad \ldots(12.2)$$

ولتقدير تأثير الدخل بصورة مستقلة في الإنفاق فانه يتم بواسطة المعادلة التالية:

$$\hat{Y}_i = \hat{B}_0 + \hat{B}_1 X_i \qquad \ldots(13.2)$$

تسمى المعادلة (13.2) بمعادلة خط الانحدار، وتشير العلامة (^) إلى كون القيم تقديرية وليست حقيقية وكل نقطة من نقاطه ($\hat{Y}_i$) تمثل القيمة التقديرية لمتوسط إنفاق جميع العوائل ذات الدخل البالغ X. ويتبين من المعادلتين (12.2) و(13.2) بأن قيم المشاهدات الفعلية Y_i تنحرف عن القيم التقديرية ($\hat{Y}_i$) بمقدار e_i وكما مبين في الشكل الاتي:

شكل (1.2)



من الشكل يتبين ان:

$$e_i = Y_i - \hat{Y}_i$$

حيث يمكن للبواقي، e_i، ان تكون سالبة أو موجبة حسب موضع نقطة المشاهدة من الخط المقدر. ولإيجاد افضل خط مستقيم لعينة مشاهدات Y ، X من بين خطوط لا نهائية العدد تصف المعادلة الخطية، تستخدم طريقة المربعات الصغرى (OLS)، وتتضمن ذلك في محاولة جعل مجموع مربع انحرافات القيم الحقيقية Y_i عن القيم التقديرية $\hat{Y}_i$ اقل ما يمكن، أي جعل مجموع مربعات الأخطاء العشوائية عند نهايتها الصغرى وبما ان طريقة OLS تشترط تصغير القيمة (Σe_i^2) إلى الحد الأدنى فانها عبارة عن مشكلة النهايات الصغرى أي:

$$\min \rightarrow \sum_{i-1}^{n} e_i^2$$

حيث ان:

$$e_i = Y_i - \hat{Y}_i$$

إذن:

$$\Sigma e_i^2 = \Sigma(Y_i - \hat{Y}_i)^2$$

بما ان معادلة الخط المستقيم الحقيقية غير المعروفة هي:

$$Y_i = B_0 + B_1 X_i$$

فان معادلة الخط المستقيم التقديرية تكون:

$$\hat{Y}_i = \hat{B}_0 + \hat{B}_1 X_i$$

بالتعويض عن $\hat{Y}_i$ بما يساويها نحصل:

$$\Sigma e_i^2 = \Sigma(Y_i - \hat{B}_0 - \hat{B}_1 X_i)^2$$

وكشرط رياضي لتصغير Σe_i^2 نوخذ المشتقات الجزئية لكل من $\hat{B}_0$ و $\hat{B}_1$ ومساواة كل منها بالصفر.