الانحدار الخطي المتعدد

يعد الانحدار الخطي المتعدد من الأساليب الإحصائية المتقدمة والتي تضمن دقة الاستدلال من اجل تحسين نتائج البحث عن طريق الاستخدام الأمثل للبيانات في إيجاد علاقات سببية بين ظواهر موضوع البحث. هو عبارة عن إيجاد معادلة رياضية تعبر عن العلاقة بين متغير تابع واحد Y وعدة متغيرات مستقلة $x_1, x_2, \dots, x_k$. يستعمل الانحدار الخطي المتعدد لتقدير قيم سابقة والتنبؤ بقيم مستقبلية. لذا فهو يستخدم في التنبؤ بتغيرات المتغير التابع الذي يؤثر فيه عدة متغيرات مستقلة. اذ يستخدم الانحدار المتعدد في شرح العلاقة بين متغير تابع ومتغيران مستقلان او اكثر.

أهمية الانحدار الخطي المتعدد:

1. معرفة اكثر المتغيرات المستقلة تأثيرا على المتغير التابع
2. معرفة مدى تأثير كل متغير مستقل على المتغير التابع واتجاه هذا التأثير ، هل هو سلبي ام إيجابي؟

شروط استخدام الانحدار الخطي المتعدد:

1. ان تكون العلاقة خطية بين المتغيرات المستقلة والمتغير التابع، وذلك من خلال إيجاد واستخدام مصفوفة الارتباط وذلك بحساب معامل الارتباط بين المتغيرات المستقلة والمتغير التابع، فاذا كانت قيمة معامل الارتباط بين كل متغير مستقل على حده والمتغير التابع مرتفعة، فهذا يؤكد على وجود علاقة خطية بين المتغير المستقل والمتغير التابع وهذا ما يؤكد ان لهذا المتغير تأثيرا على المتغير التابع.
2. العشوائية في اختيار العينة، واستقلالية درجات كل فرد عن الافراد الاخرين في العينة المختارة.
3. ان يكون المتغير التابع كمي، بينما المتغيرات المستقلة ممكن ان تكون كمية او رتبيه.
4. يجب ان تكون المتغيرات المستقلة ذات تأثير على المتغير التابع
5. ان يكون تباين أي متغير اكبر من الصفر $(S^2 > 0)$ ، والغرض من هذا ان يسهم كل متغير مستقل في تفسير التباين في درجات المتغير التابع.
6. ان يكون حجم العينة مساويا على الأقل لأربعة اضعاف عدد المتغيرات المستقلة. مثلا ان كان لدينا ثلاث متغيرات مستقلة، عليه يكون حجم العينة يساوي 12 ، $(3 \times 4 = 12)$.
7. ان يكون متوسط البواقي او الأخطاء العشوائية $(y_i - \hat{y})$ يساوي صفرا وتباينه يساوي $\left(S_e{}^2\right)$ والذي يمكن حسابه من الصيغة التالية:

$$S_e{}^2 = \frac{\sum(y_i - \hat{y})^2}{n-1}$$

8. عدم وجود علاقة خطية بين المتغيرات المستقلة او ما يسمى بعدم وجود علاقة متداخلة بين المتغيرات المستقلة، ويمكن معرفة ذلك من خلال حساب عامل تضخم التباين variance inflation factor VIF، ويتم حسابه من خلال الصيغة التالية:

$$VIF = \frac{1}{1 - {R_i}^2}$$

حيث ان :

VIF : عامل تضخم التباين للمتغير المستقل

R^2 : معامل التحديد للمتغير المستقل

أي يتم حساب VIF لكل متغير مستقل، فاذا كانت قيمة VIF لاحد المتغيرات المستقلة اكبر او يساوي من 10 ، فان هذا يدل على وجود ازدواج خطي بين المتغيرات المستقلة. اما اذا كانت قيمة VIF اقل من 10 ، فهذا يدل على عدم وجود ازدواج خطي بين المتغيرات المستقلة. لعلاج مشكلة الازدواج او التداخل الخطي ، يمكن اتخاذ الإجراءات الاتية:

i. اذا كانت العلاقة الارتباطية بين المتغيرين المستقلين قوية فيجب حذف احدهم.
ii. ربط المتغير التابع بأحد ابعاد او مكونات المتغيرات المستقلة
iii. زيادة حجم العينة

9. ان تكون بيانات المتغير التابع المتنبئ بها $(\hat{Y})$ موزعة توزيعا طبيعيا. يتم معرفة ذلك من خلال رسم شكل الانتشار الذي تكون فيه البواقي المعيارية (على المحور العمودي)، وقيم المتغير التابع غير المعيارية المتنبأ بها (على المحور الافقي) ، فاذا اظهر شكل الانتشار ان النقاط تتوزع بشكل افقي متساوي حول الصفر، وان 95% من البواقي تقع داخل المدى (+2 الى -2) ، فان ذلك يعني ان بيانات المتغير التابع المتنبأ بها $(\hat{Y})$ موزعة توزيعا طبيعيا.
10. تجانس او ثبات البواقي (الأخطاء العشوائية)، ذلك انه في حالة عدم ثبات تباين البواقي فانه يتم الحصول على تباينات كبيرة للتقديرات مما يؤدي الى اتساع فترة الثقة لمعالم نموذج الانحدار ، كذلك اختبارات الدلالة الإحصائية (T, F) لن تكون صحيحة . يمكن علاج مشكلة عدم ثبات تباين البواقي من خلال الطرق التالية:

i. شكل الانتشار الذي تكون فيه البواقي المعيارية (على المحور العمودي) واحد المتغيرات المستقلة (على المحور الافقي)، فاذا اظهر شكل الانتشار ان تشتت قيم البواقي المعيارية يزداد او ينقص بزيادة فيم احد المتغيرات المستقلة، فهذا يعني وجود عدم ثبات في تباين البواقي.
ii. شكل الانتشار الذي تكون فيه البواقي المعيارية (على المحور العمودي) وقيم المتغير التابع غير المعيارية المتنبأ بها (على المحور الافقي) ، فاذا اظهر شكل الانتشار ان النقاط تتوزع بشكل افقي متساوي حول الصفر، وان 95% من البواقي تقع داخل المدى (من +2 الى -2) فان ذلك يعني ان النموذج لا يعاني من مشكلة عدم ثبات تباين البواقي.
iii. شكل الانتشار الذي تكون فيه البواقي المعيارية (على المحور العمودي) وقيم المتغير التابع الحقيقية (على المحور الافقي)، فاذا اظهر شكل الانتشار ان النقاط تتوزع بشكل افقي حول الصفر، وان 95% من البواقي تقع داخل المدى (من +2 الى -2) فان ذلك يعني ان النموذج لا يعاني من مشكلة عدم الثبات لتباين البواقي.

iv. حساب معامل ارتباط الرتب لسبيرمان بين البواقي المعيارية وقيم المتغير المستقل، فاذا كان معامل الارتباط غير دال احصائيا ، ذل ذلك على عدم وجود مشكلة اختلاف تباين البواقي.

11. ان تكون البواقي (الأخطاء العشوائية) لها توزيع طبيعي، ويمكن اختيار ذلك من خلال رسم شكل الانتشار الذي تكون فيه البواقي المعيارية (على المحور العمودي) وقيم المتغير التابع غير المعيارية المتنبأ بها (على المحور الافقي) ، فاذا اظهر شكل الانتشار ان النقاط تتوزع بشكل افقي متساوي حول الصفر، وان 95% من البواقي تقع داخل المدى من (+2 الى -2) ، فان ذلك يعني ان البواقي موزعة توزيعا طبيعيا.

12. عدم وجود ارتباط تسلسلي بين البواقي (الأخطاء العشوائية) ، أي ان البواقي مستقلة عن بعضها البعض ، وتظهر هذه المشكلة عند تحليل بيانات السلاسل الزمنية، لان هذا يعني ان المتغير التابع يعتمد على المتغيرات المستقلة والبواقي معا، ويمكن معرفة ذلك من خلال تطبيق اختبار (**دوربن – واتسون**) والذي يتم حسابه من خلال الصيغة التالية:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum e_i{}^2}$$

حيث ان:

e_i : بواقي نموذج الانحدار

n : حجم العينة

e_{i-1} : بواقي نموذج الانحدار السابق للقيمة i

ان قيمة اختبار دوربن واتسون (d) تكون بين 0 الى 4 $(0 \leq d \leq 4)$. جدول دوربن واتسون يستخدم لمعرفة وجود او عدم وجود ارتباط متسلسل بين البواقي، حيث توجد قيمتان في الجدول هما الحد الأعلى (Upper Bound) يرمز له بالرمز d_u ، والحد الأدنى (Lower Bound) ويرمز له بالرمز d_L. يتم اتخاذ قرار وجود او عدم وجود ارتباط متسلسل بين البواقي حسب الجدول الاتي:

الشرط	$0 < d < d_L$	$d_L \leq d \leq d_u$	$4 - d_L < d < 4$	$d_L - 4 \leq d \leq d_u - 4$	$d_u < d < d_u - 4$
القرار	وجود ارتباط موجب بين البواقي	عدم اتخاذ القرار	وجود ارتباط سالب بين البواقي	عدم اتخاذ القرار	لا يوجد ارتباط بين البواقي

طرق الانحدار الخطي المتعدد:

هناك عدة طرق للانحدار الخطي المتعدد ، حيث يتم استعمالها لمعرفة افضل المتغيرات المستقلة تأثيرا وادخالها لنموذج الانحدار ، ومن هذه الطرق:

1. **طريقة ادخال كافة المتغيرات**

في هذه الطريقة يتم ادخال جميع المتغيرات المستقلة ودراستها مع بعض لمعرفة مدى تاثيرها على المتغير التابع دون ابعاد أي متغير

2. **طريقة إضافة المتغيرات على التوالي**

في هذه الطريقة نقوم أولا حساب جميع معاملات الارتباط لبيرسون بين كل متغير مستقل ومتغير تابع، بعدها نقوم بإدخال المتغير المستقل الذي يتحصل على اعلى معدل لمعامل الارتباط بينه وبين المتغير التابع ، ثم يتم إضافة متغير مستقل اخر في كل مرة للنموذج الذي يؤدي الى زيادة ملحوظة في معامل التحديد ، مع مراعاة شرط الدلالة الإحصائية لمعامل التحديد في كل مرة تقوم فيها بإضافة متغير مستقل جديد (والتي يتم حسابها من خلال اختبار F) حتى نصل الى اخر متغير مستقل نقوم بإضافته الى نموذج الانحدار يكون له اثر على المتغير التابع.

3. **طريقة حذف المتغيرات على التوالي**

في هذه الطريقة نقوم بتضمين جميع المتغيرات المستقلة في نموذج الانحدار ، وحساب معامل التحديد بين كل هذه المتغيرات مع المتغير التابع في كل مرة ، ويتم حذف المتغير المستقل في كل مرة الذي لا يؤدي حذفه الى انقاص قيمة معامل التحديد ، فهذا يعني انه ليس له تأثير على المتغير التابع.

4. **طريقة الخطوات المتتالية**

هذه الطريقة تجمع بين طريقة الإضافة المتتالية وطريقة الحذف المتتالية، حيث يتم في كل خطوة اختيار متغير مستقل ابتداء من المتغير الأكثر أهمية وتأثيرها الى غاية الوصول الى المتغير الأقل أهمية وتأثير ، ويتم الاستعانة باختبار F لاختبار الدلالة الإحصائية لمعاملات الانحدار الخطي المتعدد من اجل معرفة دلالة معنوية كل معامل واهميته في نموذج الانحدار.

حساب الانحدار الخطي المتعدد

يمكن حساب الانحدار الخطي المتعدد بطريقة المصفوفات او بطريقة المربعات الصغرى ، الا ان هذه الأخيرة شائعة الاستعمال اكثر في حالة دراسة المجتمع وهي:

$$\boldsymbol{Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e_i}$$

لكن في كثير من الأحيان ، نلجأ الى الاستعانة بالعينة لصعوبة التعامل مع قيم المجتمع، لذا فانه عند تقدير معاملات انحدار المجتمع $\alpha, \beta_1, \beta_2, \dots, \beta_k$ ، فان افضل معاملات لتقديرها بنقطة هي $a, \beta_1, \beta_2, \dots, \beta_k$ على التوالي ، وبهذا تكون معادلة الانحدار الخطي المتعدد للعينة ، كما يالي:

1. تقدير معادلة انحدار Y على $X_1, X_2, \dots, X_K$

اذا كان لدينا متغير تابع واحد Y وعدة متغيرات مستقلة k، فانه يمكن التنبؤ بدرجات المتغير التابع من درجات المتغيرات المستقلة عن طريق استخدام معادلة الانحدار الخطي المتعدد التالية:

$$\boldsymbol{\hat{y}_i = a + b_1X_{1i} + b_2X_{2i} + \cdots + b_kX_{ki} + e_i}$$

حيث ان:

$\hat{y}$: المتغير التابع (المتنبأ به)

b_1: ميل انحدار y على المتغير المستقل الأول X_1

b_2 : ميل انحدار y على المتغير المستقل الثاني X_2

b_k: ميل انحدار y على المتغير المستقل X_k

X_1 : المتغير المستقل الاول

X_2:المتغير المستقل الثاني

X_k: المتغير المستقل k

e_i: البواقي او الأخطاء العشوائية $e_i = (y_i - \hat{y}_i)$

n: حجم العينة

المعادلة أعلاه هي اختصار لعدد n من المعادلات

$$\hat{y}_1 = a + b_1X_{11} + b_2X_{21} + \cdots + b_kX_{k1} + e_1$$

$$\hat{y}_2 = a + b_2X_{12} + b_2X_{22} + \cdots + b_kX_{k2} + e_2$$

.

.

.

$$\hat{y}_k = a + b_nX_{1n} + b_2X_{2n} + \cdots + b_kX_{kn} + e_n$$

2. تقدير معادلة انحدار Y على X_2, X_1

اذا كان لدينا متغير تابع واحد Y ومتغيرين مستقلين (X_1, X_2)، فانه لا يمكن التنبؤ بدرجات المتغير التابع من درجات المتغيرين المستقلين وفق طريقة المربعات الصغرى ، ومن خلال استخدام معادلة الانحدار الخطي المتعدد التالية:

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

حيث ان b_2, b_1 تمثلان معاملات الانحدار ويتم حسابها وفق للطرق التالية:

الطريقة الأولى:

$$b_1 = \frac{r_{yx_1} - [(r_{yx_2}) \times (r_{x_1x_2})]}{1 - (r_{x_1x_2})^2} \times \frac{S_y}{S_{x_1}}$$

الطريقة الثانية:

$$b_2 = \frac{r_{yx_2} - [(r_{yx_1}) \times (r_{x_1x_2})]}{1 - (r_{x_1x_2})^2} \times \frac{S_y}{S_{x_2}}$$

Where

r_{yx_1}: Correlation coefficient between the dependent variable Y and the first independent variable x_1

r_{yx_2}: Correlation coefficient between the dependent variable Y and the second independent variable x_2

$r_{x_1x_2}$: Correlation coefficient between the first independent variable x_1 and the second independent variable x_2

S_yThe standard deviation of the dependent variable: $S_y = \sqrt{\frac{\sum y^2 - n(\bar{Y})^2}{n-1}}$

S_{x_1}: The standard deviation of the first independent variable: $S_{x_1} = \sqrt{\frac{\sum (x_1)^2 - n(\overline{x_1})^2}{n-1}}$

S_{x_2}:The standard deviation of the second independent variable:

$$S_2 = \sqrt{\frac{\sum (x_2)^2 - n(\overline{x_2})^2}{n-1}}$$

اما a فيمكن حسابها وفق الطرق التالية:

الطريقة الأولى:

$$a = \bar{Y} - [(\overline{x_1}) \times (b_1)] - [(\overline{x_2}) \times (b_2)]$$

الطريقة الثانية:

$$a = \frac{\sum Y - b_1 X_1 - b_2 X_2}{n}$$

Where

$\bar{Y}$: the mean of the dependent variable

$\overline{x_1}$: The mean of the first independent variable

$\overline{x_2}$: The mean of the second independent variable

اما β_1, β_2 اللتان تمثلان معاملات الانحدار المعيارية والتي تستعملان في عقد المقارنة بين من هو اكثر المتغيرات المستقلة تأثيرا على المتغير التابع ، حيث ان المتغير المستقل الذي تكون فيه قيمة β كبيرة هو اكثر تأثيرا عن باقي المتغيرات وهكذا بالترتيب مع باقي المتغيرات المستقلة ، ويتم حسابهما وفق الطرق التالية:

الطريقة الأولى (في حالة متغيرين مستقلين ومتغير تابع)

$$\beta_1 = \frac{r_{yx_1} - [(r_{yx_2}) \times (r_{x_1x_2})]}{1 - (r_{x_1x_2})^2}, \qquad \beta_2 = \frac{r_{yx_2} - [(r_{yx_1}) \times (r_{x_1x_2})]}{1 - (r_{x_1x_2})^2}$$

الطريقة الثانية (في حالة اكثر من متغيرين مستقلين):

$$\beta_1 = b_1 \times \frac{S_{x_1}}{S_y}$$

$$\beta_2 = b_2 \times \frac{S_{x_2}}{S_y},$$

$$\beta_k = b_k \times \frac{S_{x_k}}{S_y}$$

تفسير نتائج الانحدار الخطي البسيط

يمكن تفسير قيمة معامل الانحدار الناتج من المعادلة التالية:

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

اذا كانت قيمة معامل الانحدار b_2 موجبة فهذا يعني ، انه كلما كانت هناك زيادة في المتغير المستقل X_1 بدرجة واحدة كلما تبعه زيادة في المتغير التابع Y. ونفس الشيء بالنسبة لتفسير قيمة b_2. اما اذا كانت قيمتي b_2و b_1 مساوية للصفر فانه في هذه الحالة لا يوجد تأثير للمتغير المستقل X_1 و X_2 على المتغير التابع Y.

البواقي او الأخطاء العشوائية:

هي انحراف كل قيمة حقيقية للمتغير التابع y_i عن القيمة التقديرية لها $\hat{y}_i$ ويرمز للبواقي بالرمز e_i ويتم حسابها من خلال العلاقة التالية:

$$e_i = (y_i - \hat{y})$$

حيث متوسط البواقي يساوي صفر. اما تباين البواقي فيتم حسابه وفق العلاقة التالية:

$$S_e{}^2 = \frac{\sum(y_i - \hat{y})^2}{n-1}$$

معامل التحديد:

عبارة عن النسبة بين التغير بالقيم المقدرة على التغير الكلي ، ويرمز له بالرمز R^2 ، ويتم حسابه في الصيغة التالية:

1. $R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$

2. $R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$

3. $R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i)^2}$

4. $R^2 = 1 - \frac{\sum(\widehat{y_\iota} - \overline{y})^2}{\sum(y_i)^2}$

5. في حالة متغيرين مستقلين ومتغير تابع

$$R^2 = \left[(b_1) \times \left(\frac{S_{x_1}}{S_y}\right) \times (r_{x_1y})\right] + \left[(b_2) \times \left(\frac{S_{x_2}}{S_y}\right) \times (r_{x_2y})\right]$$

6. في حالة اكثر من متغيرين مستقلين

$$R^2 = \left[(b_1) \times \left(\frac{S_{x_1}}{S_y}\right) \times (r_{x_1y})\right] + \left[(b_2) \times \left(\frac{S_{x_2}}{S_y}\right) \times (r_{x_2y})\right] + \left[(b_i) \times \left(\frac{S_{x_i}}{S_y}\right) \times (r_{x_iy})\right]$$

7. في حالة متغيرين مستقلين ومتغير تابع

$$R^2 = \frac{[(r_{x_1y}{}^2)+(r_{x_2y}{}^2)]-2[(r_{x_1y})\times(r_{x_2y})\times(r_{x_1x_2})]}{1-r_{x_1x_2}{}^2}$$

القيمة الناتجة يتم ضربها في 100% كي نحصل على نسبة تفسير المتغيرات المستقلة $x_1, x_2, \dots, x_k$ لتباين درجات المتغير التابع Y . اما نسبة التباين الغير مفسر (غير متنبأ به) فيتم حسابه من خلال العلاقة التالية

$$S^2 = 1 - R^2$$

بعدها يتم اختبار الدلالة الإحصائية لقيمة معامل التحديد من خلال تطبيق اختبار F وفق الطرق الاتية:

1. $F = \dfrac{\frac{R^2}{k}}{\frac{1 - R^2}{n - k - 1}}$

2. $F = \dfrac{R^2}{k} \times \dfrac{n - k - 1}{1 - R^2}$

$$3.\ \ F = \frac{\frac{\sum(\hat{y}_i - \bar{y})^2}{k}}{\frac{\sum(y_i - \hat{y})^2}{n-k-1}}$$

Where

R^2: The value of the coefficient of determination

K: number of independent variables

n: sample size

ويتم استخراج قيمة f الجدولية عند درجة حرية البسط (df=k) ودرجة حرية المقام (df=n-k-1) عند مستوى الدلالة ،$(\alpha = 0.05)$ ، ويتم مقارنة قيمة f المحسوبة مع قيمة f الجدولية ، فاذا كانت قيمة f المحسوبة اكبر من او تساوي قيمة f الجدولية دل ذلك على ان معامل التحديد دال احصائيا ، ما يعني ان النسبة المفسرة للتباين دالة احصائيا.

اما اذا كانت قيمة f المحسوبة اقل من قيمة f الجدولية دل ذلك على ان معامل التحديد غير دال احصائيا. ما يعني ان النسبة المفسرة للتباين غير دالة احصائيا.

مثال: نفرض ان قيمة معامل التحديد $(R^2 = 0.67)$ فان هذه القيمة تعني ان المتغيرات المستقلة في نموذج الانحدار تفسر نسبة 67% من الاختلافات (التباين) المشاهدة في قيم المتغير التابع Y. اما 33% النسبة المتممة لها، فهي تمثل تأثير بعض العوامل او المتغيرات الأخرى التي تربطها علاقة بالمتغير التابع Y والتي لم يتم إدخالها بالدراسة.
