Regressionsanalyse

Inhaltsverzeichnis

Regressionsanalyse
Lernhinweise

Benötigte Vorkenntnisse
Für diesen Lernschritt sollten Sie wissen, was eine Zufallsauswahl ist, wie Stichproben zu Stande kommen, welche Bedeutung Messniveaus haben, was abhängige und unabhängige Variablen sind, was es mit den Freiheitsgraden auf sich hat und was eine Korrelationsanalyse ist.

Lernziele
Sie lernen anhand von Übungen und Beispielen, wie Sie die Regressionskoeffizienten b0 und b1 bestimmen können. Dazu sind folgende Schritte notwendig:

- Verständnis der Idee der Regression;
- Berechnung der Regressionskoeffizienten;
- Bestimmung der Güte des Modells;
- Bestimmung der Standardfehler und der Signifikanz;
- Verständnis der Anforderungen an die Datenqualität.

Einführung

Übersicht
Der Pearson'sche Produktmoment-Korrelationskoeffizient r gibt an, ob ein Zusammenhang zwischen zwei metrischen Variablen besteht (vgl. Lernschritt Korrelationsanalyse). Oft möchte man jedoch wissen, wie die *genaue mathematische* Beziehung zwischen einer abhängigen Variablen Y und einer unabhängigen Variablen X aussieht.
X ———> Y
Wenn X das BSP eines Landes und Y die Zuwanderungsrate eines Landes ist, möchten wir wissen, welche Zuwanderungsrate Y wir bei bestimmten Pro-Kopf-Einkommen X erwarten können.
Unsere Hypothese lautet dabei: Wohlstand zieht Zuwanderung an
Der Einfachheit halber gehen wir dabei von einer linearen Beziehung aus.

Theorie (1-8)

Inhaltsübersicht:

1. Allgemeine Beziehungen

Nehmen wir an, es längen nur Daten für zwei Fälle vor, etwa für Deutschland und Frankreich. Die Beziehung zwischen X und Y wäre in diesem Falle durch eine Linie, die sog. Regressionslinie, gegeben, die genau durch die beiden Datenpunkte verläuft.
Die mathematische Form dieser Linie ist:
$Y = b_0 + b_1 \cdot X$
Man bezeichnet diese Formel als Modellgleichung oder kurz als **Modell**.

- b_0 ist der Punkt, in dem sich die Gerade mit der Y-Achse schneidet (das sog. Absolutglied oder **intercept** im Englisch der Statistik);
- b_1 ist die Steigung der Gerade (*slope*).

b_0 und b_1 sind die sog. *Regressionskoeffizienten*.
Setzen Sie im nachfolgenden Diagramm zwei Punkte, und beobachten Sie, wie sich b_0 und b_1 verändern, wenn Sie die Regressionslinie über den *intercept* und den **slope** verschieben.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

2. 'Best Fit'

Üblicherweise hat man es mit mehr als zwei Datenpunkten zu tun. Meist gibt es dann keine Linie mehr, die durch alle Datenpunkte verläuft. Ob man dann noch Vorhersagen über erwartete Zusammenhänge machen kann, hängt dann davon ab, wie *nahe* die Linie an den einzelnen Datenpunkten vorbeiläuft.
Um möglichst genaue Vorhersagen zu machen, muss die Gerade also so durch die Datenpunkte gelegt werden, dass sie so gut wie möglich zu den beobachteten Daten "passt" (im Englischen spricht man von einem *best fit*). Dabei sollten möglichst kleine Abweichungen **e** (für *error*) zwischen den (aufgrund der Geraden) erwarteten Werten
$\hat{Y}$
und den beobachteten Werten *Y* auftreten.
Erstellen Sie im folgenden Beispiel zusätzliche Datenpunkte, indem Sie auf das Diagramm klicken. Verschieben Sie den *intercept* und den *slope*, um die Gerade möglichst nah an die einzelnen Punkte zu legen. Behalten Sie dabei die Abweichungen *aller* Punkte von der Regressionsgeraden im Auge. Lassen Sie sich dann den *best fit* anzeigen.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

3. 'Ordinary Least Squares'

In der Statistik gilt als Regel, dass die Summe der quadrierten Fehler e^2 (SSE für *sum of squared errors*), d.h. die Summe der quadrierten senkrechten Abstände zwischen den auf der Y-Achse gemessenen Datenpunkten *Y* und den auf der Regressionslinie gelegenen (erwarteten) Punkten

$\hat{Y}$

auf der Regressionslinie, möglichst klein sein sollten.
Mathematisch lässt sich das folgendermassen ausdrücken:

$$\min SSE = \min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Man spricht in diesem Sinne vom (normalen) Verfahren der kleinsten Quadrate (kurz: OLS, für *ordinary least squares*).
Erzeugen Sie im folgenden Beispiel eine Regressionslinie (indem Sie per Mausklick zwei Punkte bestimmen), und probieren Sie, von Hand einen *best fit* zu erreichen, indem Sie die Steigung (*slope*) und den Schnittpunkt der Geraden mit der Y-Achse (*intercept*) verändern, und zwar so, dass sich die Summe der quadrierten Fehler minimiert. Lassen Sie sich dann den *best fit* anzeigen.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

4. Formel der Regressionskoeffizienten

Sie haben an den letzten Beispielen gesehen: "Probieren" reicht nicht, um die richtige Lage der Regressionslinie zu ermitteln. Zum Glück lassen sich die Regressionskoeffizienten b_1 (*intercept*) und b_0 (*slope*) jedoch berechnen, und zwar nach den Formeln:

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Auf die Herleitung dieser Formeln wird hier verzichtet; vgl. dazu Bohley (2000) Kapitel 7. Vergleichen Sie aber die Formel für b_1 mit jener für den Korrelationskoeffizienten

Entsprechend lässt sich auch die Formel für b_1 weiter vereinfachen: Über dem Bruchstrich handelt es sich um die sog. Produktsumme SP, unter dem Bruchstrich um die Quadratsumme (*sum of squares*) SS der abhängigen Variablen, also um SS_x.

Man erhält also folgende vereinfachte Formel:

$$b_1 = \frac{SP_{xy}}{SS_x}$$

5. Berechnung der Regressionskoeffizienten

Ähnlich wie bei der Berechnung des Korrelationskoeffizienten lassen sich auch b_0 und b_1 mit Hilfe einer einer Tabelle berechnen. Im nachfolgenden Beispiel können Sie die vorgegebenen X- und Y- Werte überschreiben, und sich die gesamte Tabelle neu berechnen lassen.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Die hier mitberechneten Werte für die t- und F-Statistiken dienen zur Bestimmung der Signifikanz. Wir werden darauf noch zurückkommen.

6. Erklärte Varianz - Qualität des Modells

Auf dem vorherigen Rechenblatt wurde neben dem Korrelationskoeffizienten r der sog. **Determinationskoeffizient** (oder auch Bestimmtheitskoeffizient bzw. Bestimmtheitsmass) berechnet, der als R^2 definiert ist.

Im Falle der bivariaten Regression (nur eine unabhängige Variable) ist der Determinationskoeffizient schlicht das Quadrat des Korrelationskoeffizienten r.

Allgemein gibt der Determinationskoeffizient an, welcher Anteil der Abweichungsquadrate der abhängigen Variable SS_y oder auch der Varianz[1] von *Y* durch die Modellgleichung (kurz: Modell) **erklärt** wird, und zwar in Bruchteilen von 1. Multipliziert man diesen Wert mit 100, kann man diesen Anteil auch in % ausdrücken, also etwa:

$R^2 = 0.9 = 90\%$

R^2 sagt also etwas aus über die **Güte des Modells** aus.

Für die im Modell "erklärten" Abweichungsquadrate (SS_{expl}) gilt dabei folgendes:

$SS_{expl} = R^2 \cdot SS_y = SS_y - SSE$

Entsprechendes gilt für die Varianzen:

$var_{expl} = R^2 \cdot var_y = var_y - var_e$

Anmerkung:
Die Varianz ist (im Falle einer Stichprobe von Daten) definiert als Quadratsumme SS geteilt durch die Anzahl Fälle minus eins (n-1). Für die Varianz von Y gilt also $var_y = SS_y /(n-1)$. Die erklärte Varianz ist entsprechend $var_{expl} = SS_{expl} /(n-1)$. Für die Varianz der Fehler *e* gilt $var_e = SSE/(n-1)$.

7. Standardfehler und Signifikanz

Die Bestimmung des Standardfehlers und der Signifikanz dient der Feststellung der Verlässlichkeit des Modells. Dies ist deshalb notwendig, weil der Regressionskoeffizient b_1 und der Determinationskoeffizient R^2 üblicherweise anhand von Stichproben berechnet werden.

Wie weit darf man den Resultaten trauen?
Für die Signifikanz von R^2 gilt dasselbe wie für die

Signifikanz von r.

Sie haben Ihre Daten nach Zusammenhängen untersucht und einen bestimmten Wert für r berechnet. Was bedeutet dieser Wert nun aber? Ist 0.82 ein "guter" Wert? Oder wäre 0.62 besser? Oder -0.42?
Um das zu entscheiden, müssen Sie zuerst in der Lage sein, die Signifikanz von r zu bestimmen.

Grundüberlegung:
Korrelationskoeffizienten werden üblicherweise nicht für eine Grundgesamtheit von Fällen berechnet, sondern nur für eine Stichprobe. Je grösser die Anzahl Fälle n der Stichprobe, umso näher kommt der berechnete Korrelationskoeffizient r dem "echten" Korrelationskoeffizienten
ρ
(rho), der aber natürlich nicht bekannt ist.
Da eine Stichprobe nur begrenzt aussagekräftig ist stellt sich die Frage nach der Verlässlichkeit des von ihnen berechneten Korrelationskoeffizienten: beschreibt er die tatsächliche Korrelation zwischen den untersuchten Variablen in der Grundgesamtheit, oder muss man davon ausgehen, dass er zufällige Zusammenhänge einer Stichprobe aufzeigt?
Zur Beantwortung dieser Vorgehen: Frage wird überprüft, ob der berechnete Korrelationskoeffizient signifikant ist, d.h. ob man die *Null-Hypothese*: „es besteht keine Korrelation“ im vorliegenden Fall verwerfen kann oder nicht.

Anmerkung
Wir gehen von einem sog. *einseitigen Test* aus, d.h. uns interessiert nur das Risiko in einer Richtung (zu geringe Korrelation).

Vorgehen:

- Zunächst wird das sog. Signifikanzniveau festgelegt. Dabei handelt es sich um die Irrtumswahrscheinlichkeit, die man bei der Rückweisung der Null-Hypothese „keine Korrelation" zu akzeptieren gewillt ist. Üblich sind 5%, 1%, 0.1% oder 0.01%. Bei einer kleinen Anzahl von Fällen n wird man z.B. 1% wählen. Die Prozentwerte werden üblicherweise als Wahrscheinlichkeiten p (Teile von 1) angegeben. 1% entspricht z.B. p=0.01.
- Danach wird mit der Teststatistik F die Varianz der Stichproben berechnet. Der F-Test ist definiert als:

$$F_{1,n-2} = \frac{r^2}{(1-r^2)}(n-2)$$

Im Index von F stehen neben den Freiheitsgraden (*degrees of freedom*, df) der Variablen df1 (in unserem Fall 1) die Freiheitsgrade df2 der n-2 Fälle. Anhand der Freiheitsgrade kann in einer entsprechenden statistischen Tafel überprüft werden, welcher Wert von F erreicht werden muss, damit der zugehörige r-Wert bei der gewählten Irrtumswahrscheinlichkeit als Signifikant gelten kann.

Beispiel:
Bei 13 Freiheitsgraden (entspricht n=15) muss der F-Wert auf dem Signifikanzniveau von 1% (entspricht p=0.01) F=9.07 erreichen, damit die Null-Hypothese verworfen werden kann. Bei F-Werten, die kleiner als 9.07 sind, kann die Null-Hypothese *nicht* mit ausreichender Sicherheit verworfen werden. Man muss dann also davon ausgehen, dass zwischen den Variablen *kein* (signifikanter) Zusammenhang besteht.
Klicken Sie hier für einen Auszug aus der Tabelle mit F-Werten.

Auch hier kann die F-Statistik zum testen der Signifikanz verwendet werden:

$$F_{1,n-2} = \frac{r^2}{(1-r^2)}(n-2)$$

Wieder wird F für eine unabhängige Variable und n-2 Freiheitsgrade indiziert. Und wieder kann anhand der Tabelle der F-Werte geprüft werden, ob der kritische Wert für das gewählte Signifikanzniveau erreicht wird oder nicht.
Was den sog. Standardfehler des Regressionskoeffizienten b_1 betrifft, sind dagegen einige weitere Überlegungen notwendig:

- Üblicherweise liegen nicht alle Datenpunkte auf der Regressionslinie. Vielmehr gibt es Fehler e (für *error*). Dies sind die senkrechten Abstände zwischen den Datenpunkten *Y* und den korrespondierenden Punkten

  

  auf der Regressionslinie, deren aufsummierte Quadrate SSE (*sum of squared errors*) bei der Positionierung der Regressionslinie zu minimieren sind.
- Wenn wir SSE durch die um 2 reduzierte Anzahl der Fälle der Untersuchung (n-2) teilen, erhalten wir den sog. **Standardfehler der Schätzung MSE** (*mean squared error*):

$$MSE = \sqrt{\frac{SS_e}{(n-2)}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{(n-2)}} = \sqrt{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 / (n-2)}$$

- Den Standardfehler des Regressionskoeffizienten $SE(b_1)$ erhält man nun, wenn man den Standardfehler der Schätzung MSE durch die Wurzel der Abweichungsquadrate von X (also SS_x) teilt. Als Gleichung:

- Der Standardfehler von b_1 gibt zunächst einen Anhalt zur Streuung, d.h. zur Bandbreite der erwarteten Werte, wenn dieselben Berechnungen mit anderen Stichproben durchgeführt würden. Er sollte also **klein** sein.
- Darüber hinaus geht es bei Signifikanztests aber immer um die Frage, mit welcher Berechtigung sich die sog. Null-Hypothese (kein Zusammenhang zwischen X und Y) verwerfen lässt. "Kein Zusammenhang" würden bedeuten, dass der Regressionskoeffizient null wäre ($b_1=0$). Eine Faustregel besagt, dass b_1 am besten drei oder mehr Standardfehler grösser als null sein sollte:
  $b_1 > 3SE(b_1)$
- Genauer arbeitet der sog. t-Test: Hier wird der berechnete Wert des Regressionskoeffizienten in Beziehung zu seinem Standardfehler gesetzt.

$$t_{n-2} = \frac{b}{SE(b_1)}$$

Auch hier kann man aufgrund einer Tafel kontrollieren, ob der sog. kritische Wert für das gewählte Signifikanzniveau erreicht wird.

Beispiel:
b=9, $SE(b_1)$=3, t=3 wäre bei n=16 auf dem Niveau von 0.1% bei 14 Freiheitsgraden (df=n-2=14) für einen zweiseitigen Test noch knapp signifikant. Kritischer Wert ist t=2.624; vgl. dazu Blalock (1960) S. 559.

8. Voraussetzungen der Regression

Die Anwendung der Regressionsanalyse stellt einige Anforderungen an die Qualität der Daten und die Gültigkeit der getroffenen Annahmen. Die *wichtigsten* dieser Anforderungen werden hier kurz skizziert.

- **Normalverteilung:** Sowohl die X-Werte als auch die Y-Werte sollten für sich genommen annähernd normal verteilt sein. Ob das der Fall ist, lässt sich mit Histrogrammen überprüfen.
- **Linearität:** Die lineare univariate Regression, wie sie hier vorgestellt wurde, unterstellt die Linearität der Beziehung zwischen den Variablen, d.h. man sollte zumindest annäherungsweise von einer Linearität der Beziehung ausgehen können. Kontrollieren lässt sich das durch ein Streudiagramm.

- **Homoskedastizität:** Die Varianz der Verteilung der abhängigen Variablen muss für alle Werte der unabhängigen Variablen konstant sein (Annahme der sog. Homoskedastizität), d.h. mit steigenden Werten der unabhängigen Variablen sollten die Werte der abhängigen Variablen nicht weiter streuen. Ist dies der Fall, liegt eine sog, **Heteroskedastizität** vor. Auch dies lässt sich mit Hilfe eines Streudiagramms überprüfen.
- **Unabhängigkeit der Daten und der Fehler *e*:** Alle Daten sollten unabhängig voneinander sein, d.h. die Fälle sollten nicht untereinander korrelieren. Der Wert X_4 sollte also nicht einfach von X_3 abgeleitet werden können. Das gilt auch für die Fehler oder Residuen *e*. Ob eine sog. Autokorrelation vorliegt, kann mit dem Durbin-Watson-Koeffizienten geprüft werden, der von den meisten Statistik-Programmen berechnet wird.

SPSS-Kochbuch

Kochbuch: Regressionsanalyse in SPSS, *.pdf, 93 KB

Übungsdaten:

- im SPSS-eigenen *.sav Format, ready to use, 1 KB
- als Excel Tabelle, die umkodiert werden muss, 14 KB