6.1.4 Weitere Korrelationskoeffizienten

Anwendung des Korrelationskoeffizienten nach Bravais-Pearson auf dichotome nominale Merkmale

Liegen *dichotome* nominale Merkmale, d.h. Merkmale mit nur zwei ungeordneten Ausprägungen vor (z.B. ja/nein), *und* kodiert man die Ausprägung mit 0 und 1, so kann man die Formel des Korrelationskoeffizienten nach Bravais-Pearson sinnvoll anwenden. Man erhält den sogenannten *Punkt-Korrelationskoeffizienten*, der identisch zu Φ_s aus Kapitel 5.3 ist.

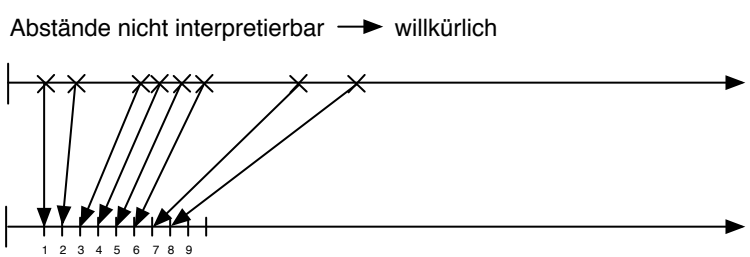
Im Fall einer dichotomen und einer metrischen Variablen ergibt sich bei Anwendung des Korrelationskoeffizienten nach Bravais-Pearson die sogenannte *Punkt-biseriale Korrelation*. (vgl. etwa Jann (2002, S.90f) oder Wagschal (1999, Kap 10.8).)

Rangkorrelationskoeffizient nach Spearman

- Wir betrachten ein bivariates Merkmal (X, Y), wobei X und Y nur *ordinalskaliert* sind, aber viele unterschiedliche Ausprägungen besitzen.
- Der Korrelationskoeffizient von Bravais-Pearson darf nicht verwendet werden, da hier die Abstände nicht interpretierbar sind. $(\bar{x}, \bar{y})$ wären willkürliche Zahlen, ebenso $(x_i - \bar{x}), (y_i - \bar{y})$.

Andererseits macht eine Darstellung und Analyse über Kontingenztafeln wenig Sinn, da wegen der vielen unterschiedlichen Ausprägungen die meisten Zellen bestenfalls spärlich besetzt sind.

Bsp. 6.7.



- Liegen keine $Bindungen$ vor, dann rechnet man bei ordinalskalierten Merkmalen statt mit $(x_i, y_i)_{i=1,\ldots,n}$ mit den zugehörigen Rängen $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))$ $i = 1, \ldots, n$. Dabei ist

$$\mathrm{rg}(x_i) = j :\Longleftrightarrow x_i = x_{(j)},$$

d.h. der Rang $\mathrm{rg}(x_i)$ ist die Nummer, die x_i in der geordneten Urliste $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ einnimmt (analog für $\mathrm{rg}(y_i)$). Der kleinsten Beobachtung wird also der Wert 1 zugeordnet, der zweitkleinsten der Wert 2, usw., der größten der Wert n.

Bsp. 6.8.

x_i	1	7	2	5.3	16
$\mathrm{rg}(x_i)$					

- Liegen sogenannte Bindungen vor, d.h. hier: Haben mehrere Einheiten dieselbe Ausprägung der Variablen X oder der Variablen Y, so nimmt man den Durchschnittswert der in Frage kommenden Ränge (Achtung: etwas anderer Begriff der Bindung als in Kapitel 5).

Bsp. 6.9.

x_i	1	7	7	3	10
Rang					
$\mathrm{rg}(x_i)$					

- Wende nun den Korrelationskoeffizienten nach Bravais-Pearson auf die Rangdaten an. Nach Umformung ergibt sich unter Benutzung von

$$\sum_{i=1}^{n} rg(x_i) = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \sum_{i=1}^{n} rg(y_i)$$

folgende Formel:

Definition:

$$\varrho_S(X,Y) := \frac{\displaystyle\sum_{i=1}^{n} \mathrm{rg}(x_i) \cdot \mathrm{rg}(y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\displaystyle\sum_{i=1}^{n} (\mathrm{rg}(x_i))^2 - n\left(\frac{n+1}{2}\right)^2}\sqrt{\displaystyle\sum_{i=1}^{n} (\mathrm{rg}(y_i))^2 - n\left(\frac{n+1}{2}\right)^2}}$$

heißt (empirischer) *Rangkorrelationskoeffizient nach Spearman*.

Bem. 6.10.

- Liegen keine Bindungen vor, so gilt (Beweis durch Nachrechnen: z.B. Ferschl (1985^3). Deskriptive Statistik, S. 285)

$$\varrho_{S,XY} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

wobei $d_i := \mathrm{rg}(x_i) - \mathrm{rg}(y_i)$.

- Wichtig für Interpretation: Da $\varrho_S(X, Y)$ sich aus der Anwendung von $\varrho(X, Y)$ auf Rangdaten ergibt, behalten die entsprechenden Bemerkungen zum Bravais-Pearson-Korrelationskoeffizienten – auf die Ränge bezogen – ihre Gültigkeit. Insbesondere gilt $-1 \leq \varrho_{S,XY} \leq 1$, und $\varrho_{S,XY}$ ist analog zu interpretieren.

- Im Gegensatz zum Korrelationskoeffizienten von Bravais-Pearson misst der Rangkorrelationskoeffizient nicht nur lineare, sondern allgemeiner monotone Zusammenhänge. Die Anwendung der Rangtransformation bewirkt in gewisser Weise eine Linearisierung monotoner Zusammenhänge.

- Die Bildung von Rängen ist deutlich unempfindlicher gegenüber Ausreißern, so dass auch der Rangkorrelationskoeffizient ausreißerresistenter ist.

Bsp. 6.11.

(fiktiv, Zahlen aus Jann, 2002/2005)

Zwei Gutachter sollen das *autoritäre* Verhalten von 5 Gruppenmitgliedern vergleichen, indem sie Scores auf einer Skala zwischen 0 und 100 vergeben. (Dies ist ein typischer Fall einer Ordinalskala; die Abstände sind nicht direkt interpretierbar, sondern nur die Reihenfolge!)

Man berechne den Rangkorrelationskoeffizienten nach Spearman für die Merkmale X und Y mit

X	Einstufung durch Gutachter 1
Y	Einstufung durch Gutachter 2

Person i	1	2	3	4	5
X: Gutachter 1	10	15	20	20	30
Y: Gutachter 2	20	10	30	40	60
$\mathrm{rg}(x_i)$					
$\mathrm{rg}(y_i)$					

- Die Bedeutung von Rangverfahren wird in der Survey-Statistik sicher zunehmen; bis vor kurzem waren sie rechentechnisch für sehr große Datensäze kaum bewältigbar. (Sortieren der Größe nach gehört zu den aufwändigsten Aufgaben.)

Bem. 6.12.

- Analog zur punkt-biserialen Korrelation gibt es auch eine *biseriale Rangkorrelation* zur Beschreibung des Zusammenhangs zwischen einer $0-1$-kodierten dichotomen nominalen und einer quasi-stetigen ordinalen Variable (vgl. Wagschal, 1999, Kap 10.7).

Ergebnisse bei Mietspiegeldaten:
Korrelation nach Spearman

- (Fläche, Miete) 0.5616114
- (Baujahr, Miete) 0.1993477
- (Baujahr, Fläche) -0.1683651

6.2 Regressionsanalyse I: Die lineare Einfachregression

6.2.1 Grundbegriffe und Hintergrund

Bedeutung der Regression:

- Eines der am häufigsten verwendeten statistischen Verfahren. Vielfache Anwendung in den Sozialwissenschaften $\rightarrow$ Analoge Ausdehnung auf viele Variablen möglich!
- Grundidee der Interpretation bleibt in verwandter Weise bei vielen allgemeineren Modellen erhalten, die hier nicht betrachtet werden (können).

Motivation:

- Wir betrachten zunächst zwei metrische Variablen X und Y.
- Der Korrelationskoeffizient nach Bravais-Pearson misst die Stärke des linearen Zusammenhangs zwischen X und Y, beantwortet also die Frage „Wie gut lassen sich Ausprägungen (x_i, y_i), $i = 1, \ldots, n$, durch eine Gerade beschreiben?
- Die Regression geht nun einen Schritt weiter:
  * Wie sieht die am besten passende Gerade aus?
  * $\Rightarrow$ Analyse und Beschreibung des Zusammenhangs.

* Dadurch zusätzliche Möglichkeiten: „Modell“: Y als ungefähre Funktion von X ausdrückbar
  * „individuelle“ Prognose basierend auf dem x-Wert: gegeben sei eine Einheit mit Ausprägung x^* von X. Wo liegt dem Modell nach das dazugehörige $\hat{y}^*$?
  * Elastizität: Wie stark wirkt sich eine Änderung von X um eine Einheit auf Y aus?

  Entscheidende Grundlage für Maßnahmenplanung

- Die Regression ist ein erster Schritt in die höhere Statistik. Fast alle gängigen Verfahren sind im weiteren Sinne Regressionsmodelle.
  * Zum einen wird sich zeigen, dass die Bedingung der Linearität des Zusammenhangs in der Praxis bei weitem nicht so restriktiv ist, wie es Anschein hat. Viele auf den ersten Blick nichtlineare Modelle erweisen sich als linear im später verwendeten Sinn.
  * Zudem wird sich zeigen: Die Einschränkung auf lineare Zusammenhänge läßt sich relativ leicht lockern. Viele Grundideen zur Interpretation gelten in verwandter Form auch für andere Regressionsmodelle.
- Die Regressionsanalyse erlaubt es auch, mehrere Variablen gemeinsam zu untersuchen, die Interpretation der komplexeren Modelle ist aber sehr ähnlich zur „linearen Einfachregression“ zweier Variablen.
- Bei der Regressionsanalyse wird die Symmetrie des Zusammenhangs i.A. aufgegeben, d.h. nun wird ein gerichteter Zusammenhang der Form $X \longrightarrow Y$ betrachtet.

Bezeichnungen:

X	Y
unabhängige Variable	abhängige Variable
exogene Variable	endogene Variable
erklärende Variable	zu erklärende Variable
Stimulus	Response
Einflußgröße	Zielgröße
	Outcome
Prädiktor	
Kovariable	

6.2.2 Lineare Einfachregression: Grundmodell und Kleinste-Quadrate-Prinzip

Idee: Versuche, Y als einfache Funktion f von X zu beschreiben:

$$Y \approx f(X).$$

Einfachste Möglichkeit: f linear, also

$$Y \approx \beta_0 + \beta_1 \cdot X.$$

Für die beobachteten Datenpunkte soll also für jedes $i = 1, \ldots, n$ gelten

$$y_i \approx \beta_0 + \beta_1 \cdot x_i$$

Normalerweise besteht kein perfekter linearer Zusammenhang, so dass ein unerklärter Rest ε_i in die Modellgleichung mit aufgenommen wird (In Statistik 2 werden wir ε_i als zufälligen Fehler interpretieren):

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i.$$

Dies ist das Modell der linearen Einfachregression.

β_0 und β_1 sind unbekannte Größen, die sogenannten *Regressionsparameter* oder *Regressionskoeffizienten*, die anhand der Daten bestimmt werden müssen. Man bezeichnet die aus den Daten gewonnenen Schätzungen dann mit $\hat{\beta}_0$ und $\hat{\beta}_1$.
Man beachte hierbei, dass β_0 und β_1 bzw. $\hat{\beta}_0$ und $\hat{\beta}_1$ keinen Index tragen; sie werden hier als interindividuell konstant betrachtet und beschreiben den Zusammenhang, der für alle Beobachtungen gelten soll.

Methode der kleinsten Quadrate: Bestimme $\hat{\beta}_0, \hat{\beta}_1$ so, dass alle Abweichungen der Daten von der Gerade „möglichst klein“ werden, d.h. so, dass die Summe der quadratischen Differenzen zwischen den Punkten y_i und der Gerade $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$ minimiert wird. D.h. minimiere das *Kleinste Quadrate Kriterium* (KQ-Kriterium):

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

bezüglich $\hat{\beta}_0$ und $\hat{\beta}_1$.

Definition: Gegeben seien zwei metrische Merkmale X und Y und das Modell der linearen Einfachregression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Dann bestimme man $\hat{\beta}_0$ und $\hat{\beta}_1$ so, dass mit

$$\begin{aligned} \hat{\varepsilon}_i &:= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \end{aligned}$$

das Kleinste-Quadrate-Kriterium

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

minimal wird. Die optimalen Werte $\hat{\beta}_0$ und $\hat{\beta}_1$ heißen <u>KQ-Schätzer</u>, $\hat{\varepsilon}_i$ bezeichnet das i-te (geschätzte) <u>Residuum</u>.

Bem. 6.13.

- Durch das Quadrieren tragen sowohl positive als auch negative Abweichungen von der Regressionsgeraden in gleicher Weise zum KQ-Kriterium bei.
- Diese Vorgehensweise entspricht genau dem Ansatz, Modelle über Verlustfunktionen anzupassen (vgl. „Exkurs: Herleitung von Lagemaßen als Lösung eines Optimierungsproblems“ in Kapitel 3)
- Wie dort besprochen, bewirkt das Quadrieren außerdem, dass große Abweichungen überproportional stark berücksichtigt werden. (Die KQ-Schätzer sind in diesem Sinne ausreißeranfällig, da mit aller Macht versucht wird, große Abweichungen zu vermeiden. Es gibt robustere Alternativen, die z.B. die Summe der absoluten Residuen minimieren (“$\mathcal{L}^1$-Regression“)

Satz: Für die KQ-Schätzer gilt

i) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{\tilde{s}_X^2} =$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \varrho_{X,Y}\frac{\tilde{s}_Y}{\tilde{s}_X},$$

ii) $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

iii) $$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

Bem. 6.14.

- Hat man standardisierte Variablen X und Y (gilt also $\tilde{s}_X = \tilde{s}_Y = 1$), so ist $\hat{\beta}_1$ genau $\rho_{X,Y}$.

- Die mittlere Abweichung von der Regressionsgeraden ist Null.
  Diese Eigenschaft kann auch verwendet werden, um die korrekte Berechnung der KQ-Schätzer zu überprüfen.

- Basierend auf den Schätzern $\hat{\beta}_0$ und $\hat{\beta}_1$ kann der Wert der abhängigen Variablen Y auch für neue, unbeobachtete Werte x^* der Kovariablen X berechnet werden (Prognose):
  $$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

- Weiß man, dass $\beta_1 = 0$ ist, und setzt daher $\hat{\beta}_1 = 0$, so lautet die KQ-Schätzung $\bar{y}$.
  In der Tat: $\bar{y}$ minimiert $\sum_{i=1}^n (y_i - \beta_0)^2$, vergleiche den eben angesprochenen Exkurs im Kapitel bei den Lagemaßen.

Interpretation der Regressionsgeraden:

- $\hat{\beta}_0$ ist der Achsenabschnitt, also der Wert der Gerade, der zu $x = 0$ gehört. Er lässt sich oft als „Grundniveau“ interpretieren.
- $\hat{\beta}_1$ ist die Steigung (Elastizität): Um wieviel erhöht sich y bei einer Steigerung von x um eine Einheit?
- $\hat{y}^*$ (Punkt auf der Gerade) ist der Prognosewert zu x^*.

Bsp. 6.15. *Fiktives „ökonomisches Beispiel“ zur Klärung: Kaffeeverkauf auf drei Flohmärkten*

X Anzahl verkaufter Tassen Kaffee
Y zugehöriger Umsatz (Preis Verhandlungssache)

Man bestimme die Regressionsgerade und interpretiere die erhaltenen KQ-Schätzungen! Welcher Umsatz ist bei zwölf verkauften Tassen zu erwarten?

i	y_i					x_i
1	9					10
2	21					15
3	0					5

Bem. 6.16.

Manchmal macht es Sinn, aus inhaltlichen Gründen zu fordern, dass $\beta_0 = 0$ ist, also die Regressionsgerade durch den Ursprung geht. Man setzt also an

$$y_i = \beta_1 x_i + \epsilon_i, \quad i = 1, \cdots, n$$

und erhält

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

6.2.3 Modellanpassung: Bestimmtheitsmaß und Residualplots

- Wie gut lässt sich die abhängige Variable Y durch die Kovariable X erklären?
- Wie gut passt der lineare Zusammenhang zwischen X und Y?

PRE-Ansatz:

Modell 1: Vorhersage von Y ohne X.

Dabei gemachter Gesamtfehler:

$$SQT :=$$

(Gesamtstreuung / Gesamtvariation der y_i: „<u>sum of squares total</u>“).

Modell 2: Vorhersage von Y mit X.

Dabei gemachter Gesamtfehler:

$$SQR :=$$

(Residualstreuung / Residualvariation: „<u>sum of squared residuals</u>“).

Die Differenz

$$SQE := SQT - SQR$$

nennt man die durch das Regressionsmodel erklärte Streuung („sum of squares explained“).

Man kann zeigen, dass gilt

$$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

Streuungszerlegung:

$$SQT = SQR + SQE$$

(analog zur Streuungszerlegung bei geschichteten Daten).

Bestimmtheitsmaß: Der PRE-Ansatz liefert das Gütekriterium

$$\frac{SQT - SQR}{SQT} = \frac{SQE}{SQT}.$$

Diese Größe bezeichnet man als Bestimmtheitsmaß. In der Tat gilt (nach etwas längerer Rechnung):

$$\frac{SQE}{SQT} = R_{XY}^2$$

d.h. dies ist genau das Bestimmtheitsmaß aus Definition 6.24.

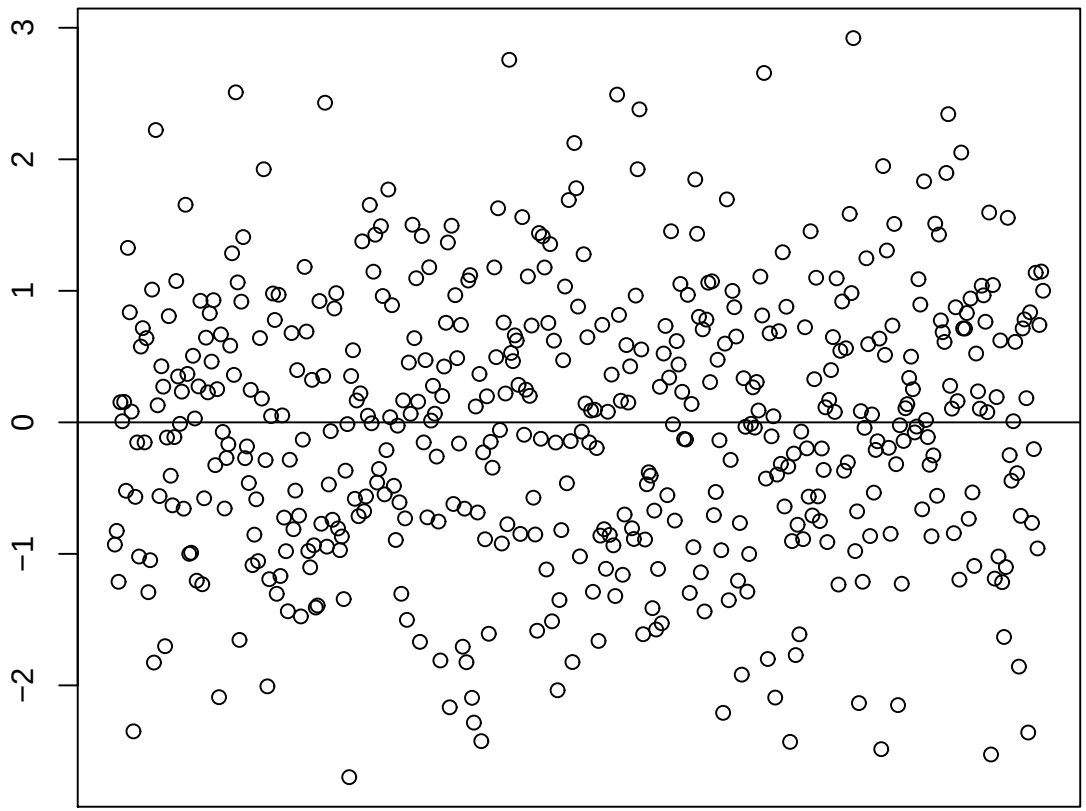
Es gibt also drei Arten, R^2_{XY} zu verstehen:

Eigenschaften:

- Es gilt: $0 \leq R^2_{XY} \leq 1$.
- $R^2_{XY} = 0$: Es wird keine Streuung erklärt, d.h. es gibt keinen (linearen) Zusammenhang zwischen X und Y.
- $R^2_{XY} = 1$: Die Streuung wird vollständig erklärt. Alle Beobachtungen liegen tatsächlich auf einer Geraden.

Residualplots

Eine wichtige optische Möglichkeit, die Anpassung zu beurteilen, beruht auf dem Studium der geschätzten Residuen $\hat{\varepsilon}_i$. Sie sollen unsystematisch um 0 streuen.



Zeigt sich eine Systematik, so war der lineare Ansatz unangemessen, und es ist größte Vorsicht bei der Interpretation geboten!