Statistik
08. Lineare Modelle

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: `https://github.com/rsling/VL-Deutsche-Syntax`

Inhalt

LMs

Literatur

- Gravetter & Wallnau 2007
- Zuur u. a. 2009
- Maxwell & Delaney 2004

Übersicht

- Wiederholung der Pearson-Korrleation (r, r^2)
- Siginifikanztests mit Korrelationen
- Unterschied von Pearsons r zu Spearmans Rang-Korrelation
- Unterschiede zwischen Korrelation und Regression
- Berechnung linearer Regressionsmodelle
- Signifikanztests für Modell und Koeffizienten

Pearson-Korrelation (Wh.)

$$r(x_1, x_2) = \frac{cov(x_1,x_2)}{s(x_1) \cdot s(x_2)}$$

In Gravetter & Wallnau, Kap. 16 lautet die Formel:

$$r = \frac{SP}{\sqrt{SQ_x \cdot SQ_y}}$$

Die Formeln sind äquivalent, weil (mit x, y statt x_1, x_2):

$$r(x, y) = \frac{cov(x,y)}{s(x) \cdot s(y)} = \frac{\frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{x})}{n-1}}{\sqrt{\frac{\sum(x_i - \bar{x})}{n-1} \cdot \frac{\sum(y_i - \bar{y})}{n-1}}} = \frac{\frac{SP(x,y)}{n-1}}{\sqrt{\frac{\sum(x_i - \bar{x}) \cdot \sum(y_i - \bar{y})}{n-1}}} =$$

$$\frac{\frac{SP(x,y)}{n-1}}{\frac{\sqrt{\sum(x_i - \bar{x}) \cdot \sum(y_i - \bar{y})}}{n-1}} = \frac{SP(x,y)}{n-1} \cdot \frac{n-1}{\sqrt{SQ(x) \cdot SQ(y)}} = \frac{SP(x,y)}{\sqrt{SQ(x) \cdot SQ(y)}}$$

r^2 und Siginifikanztests

- Maß der Varianzerklärung durch r: r^2 (vgl. t-Test)
- Signifikanztest möglich: Schluss auf Korrelation in der Grundgesamtheit
- $df_r = n - 2$
- Unter der H0 (keine Korrelation) t-verteilt:
  $t = r\sqrt{\frac{n-2}{1-r^2}}$
- ...oder Tabellen (z. B. G&W, B.6)

Voraussetzungen

- Intervallskalierung
- lineare Abhängigkeit
- bei kleinen n: Normalverteilung für x und y

- wenn nicht: Spearmans Rang-Korrelation

Spearmans Rang-Korrelation

- mathematisch nicht andere als eine Pearson-Korrleation
- vorher: Umrechnung der rohen x,y-Werte in Ränge
- bei gleichen Werten: alle gleichen Werte bekommen Rang-Mittel

Werte in Ränge umrechnen

Ein Beispiel zur Umwandlung in Ränge:

Index:	1	2	3	4	5
Messwerte x:	4	7	3	1	3
Messwerte y:	9	12	11	2	8

Statt der Messwerte arbeitet man mit den Rängen der Messwerte an den jeweiligen Indexen.

Index:	1	2	3	4	5
Ränge der Messwerte x:	4	5	2.5	1	2.5
Ränge der Messwerte y:	3	5	4	1	2

Abkürzung der Berechnung

Wenn $Rang(x_i)$ der Rang für x_i in x ist:

Spearmans Rang-Korrelation:

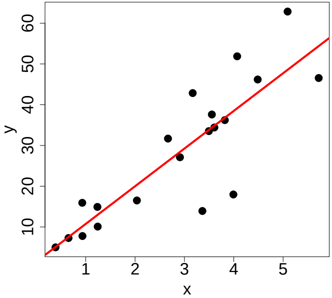
$$r_S = 1 - \frac{6\sum_{i=1}^{n}(Rang(x_i) - Rang(y_i))^2}{n(n^2 - 1)}$$

Unterschiede zwischen Korrelation und Regression

- Korrelation: Stärke des Zusammenhangs
- Regression: genaue Funktion zur Modellierung des Zusammenhangs

- Korrelation: Diagnostik/Test
- Regression: Vorhersage (und Test)

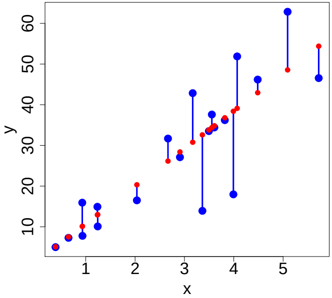
Spezifikation der Funktion für die Regressionsgerade



- Schnittpunkt mit der y-Achse (Intercept): a
- Steigung (Slope): b (b heißt auch Koeffizient)
- Regressiongleichung (=Modell): $\hat{y} = b \cdot x + a$
- Für jeden beobachteten Wert: $y_i = b \cdot x_i + a + e_i$ (e_i als Fehlerterm)

Idee der kleinsten Quadrate

Die vom Modell vorhergesagten Werte (rot, auf der Regressionsgerade) sollen insgesamt einen so geringen Abstand wie möglich zu den Beobachtungen (blau) haben.



Die Summe der quadrierten negativen und positiven Differenzen (blau) soll minimiert werden (=kleinste Quadrate): Minimierung von $\sum e^2$

Berechnung der Regressionsgleichung

- Slope/Steigung: $b = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SP(x,y)}{SQ(x)}$
- Intercept: $a = \bar{y} - b \cdot \bar{x}$
- Der Beweis, dass dies die Gerade mit den kleinsten Quadraten schätzt, erfordert bereits erheblichen mathematischen Aufwand, den wir uns sparen.
- Determinationskoeffizient: $r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$

Standardfehler für die Gleichung

- Wie stark variiert der Fehler für Stichproben einer Größe?
- $SF_{residual} = \sqrt{\frac{\sum e^2}{n-2}}$
- Je kleiner $SF_{residual}$, desto besser das Modell.
- Beachte: n wird größer (größere Stichprobe): $SF_{residual}$ wird kleiner.
- Und: Fehler e werden kleiner: $SF_{residual}$ wird kleiner.

F-Test für Model

- Wie bei ANOVA: $F = \frac{erklaerte\ Varianz}{zufaellige\ Varianz} = \frac{s^2_{regression}}{s^2_{residual}}$

- zufällige Varianz: $s^2_{residual} = \frac{(1-r^2) \cdot SQ(y)}{1}$
- erklärte Varianz: $s^2_{regression} = \frac{r^2 \cdot SQ(y)}{n-2}$

- Freiheitsgrade sind immer $df_1 = 1$ und $df_2 = n - 1$.
- Beachte: r^2 ist in $[0..1]$ und teilt die Varianz von y auf.

Standardfehler und t-Test für Koeffizienten

- Für b und a kann je ein Standardfehler angegeben werden.
- $SF(b) = \frac{\sqrt{\frac{\sum e^2}{n-1}}}{\sqrt{SQ(x)}}$
- Unter der H0: $b = 0$ ist dann t-verteilt:
  $t = \frac{b}{SF(b)}$

Mehrere unabhängige

- Design bei einfachem LM:
  - eine intervallskalierte Abhängige
  - eine Unabhängige
- wie bei mehrfaktorieller ANOVA:
  - oft interessiert mehrfaktorielle Abhängigkeit

Multivariate Modellgleichung

Mehrere Koeffizienten im allgemeinen linearen Modell:

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 \ldots b_n \cdot x_n + a$$

Konzeptuell bleibt die Berechnung aller Werte und Tests gleich,
die Mathematik wird ungleich komplizierter.

Man schreibt R^2 statt r^2.

Normalitätsannahme

Die Residuen müssen normalverteilt sein.
(als Diagnostik für: Die Messwerte müssen normalverteilt sein.)

- Missverständnis: Test aller Residuen auf Normalität
- denn: Für jedes x_i müssen die e normalverteilt sein.
- erfordert mehrere Messungen pro x_i oder Intervallbildung
- größere Stichproben, kleinere Probleme
- visuelle Diagnose: Q-Q-Plots (hier nicht behandelt)

Unabhängigkeit

Jedes y_i darf nur von x_i abhängen,
niemals zusätzlich von x_j mit $i \neq j$.

- mathematisch: nicht-lineare Abhängigkeit
- konzeptuell: Zeitserien
- konzeptuell: Sequenzen in Texten
- Lösung: andere Modellspezifikation

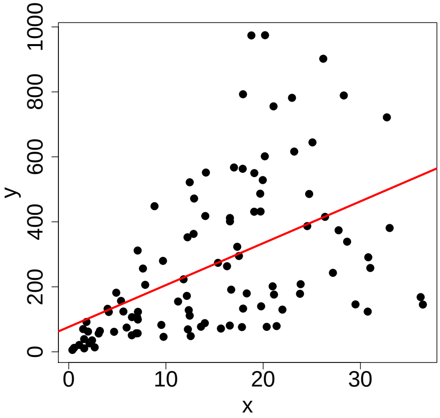
Homoskedastizität

Die Residuen müssen homoskedastisch verteilt sein.

- Bedeutung: Die Varianz der e muss über alle x homogen sein.
- vgl. die Forderung der „Varianzhomogenität“ bei t-Test und ANOVA

Darstellung heteroskedastischer Residuen

Hier wird die Varianz der Residuen mit steigendem x immer größer.
Ein lineares Modell versagt hier wegen verletzter Verteilungsannahmen.



Lösung von LM-Krisen

- mehr Daten ziehen, Daten transformieren
- generalisierte lineare Modelle (GLM)
  legen andere Verteilungsannahmen zugrunde
- (generalisiert) additive Modelle (GAM)
  schätzen Smoothingfunktionen für Koeffizienten

ANOVA als Modell mit kategorialen Regressoren

n Gruppen der ANOVA können als n dichotome Variablen dargestellt werden:

		ANOVA-Gruppen		
		A_1	A_2	A_3
Regressor	$x_1 =$	$\mathbf{1}$	0	0
	$x_2 =$	0	$\mathbf{1}$	0
	$x_3 =$	0	0	$\mathbf{1}$

Lineares Modell mit solchen „Dummy-Variablen“

Normale Modellspezifikation:

$$\hat{y} = b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + a$$

Da jeweils nur eins der $x_i = 1$ und alle anderen immer 0 werden,
wird einfach der Wert des entsprechenden β_i (plus a) vorhergesagt.

Spearmans Rang-Korrleation in R

Die Funktion `cor()` hat ein Argument `method`,
das als `"spearman"` angegeben werden kann.

```
> cor(x, y, method = "spearman")
```

Lineare Modelle in R

- Modellformeln: `y~x`
  „y abhängig von x“
- Mehrere Unabhängige: `y~x1+x2`
- Mehrere Unabhängige mit Interaktion: `y~x1*x2`
- Mehrere Unabhängige nur Interaktion: `y~x1:x2`

- Lineares Modell schätzen und speichern:
  ```
  > m <- lm(y~x)
  ```
- Ausgabe Evaluation:
  ```
  > summary(m)
  ```

Ausgabe LM

Interpretieren Sie diese Ausgabe anhand der Folien:

```
Call:
lm(formula = y ~ x)

Residuals:
Min        1Q    Median       3Q       Max
-20.4298  -2.4920  -0.2625   3.8038  14.2922

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.513      4.321   0.350     0.73
x              9.242      1.333   6.933 1.77e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.008 on 18 degrees of freedom
Multiple R-squared:  0.7275, Adjusted R-squared:  0.7124
F-statistic: 48.06 on 1 and 18 DF,  p-value: 1.768e-06
```

Nächste Woche | Überblick

Einzelthemen

Literatur I

Gravetter, Frederick J. & Larry B. Wallnau. 2007. *Statistics for the Behavioral Sciences*. 7. Aufl. Belmont: Thomson.

Maxwell, Scott E. & Harold D. Delaney. 2004. *Designing experiments and analyzing data: a model comparison perspective*. Mahwa, New Jersey, London: Taylor & Francis.

Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Berlin etc.: Springer.

Autor

Kontakt

Prof. Dr. Roland Schäfer
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena
Fürstengraben 30
07743 Jena

https://rolandschaefer.net
roland.schaefer@uni-jena.de

Lizenz