Statistische Datenanalyse mit R, Teil 4 online, Korrelation und Regression

Dr. Andrea Denecke
Leibniz Universität IT-Services

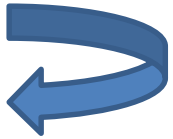
Korrelationsanalyse

Eine Korrelationsanalyse soll herausfinden

- ob ein linearer Zusammenhang zwischen zwei metrischen Variablen besteht
- die Stärke des Zusammenhangs
- die Richtung des Zusammenhangs (positiv, negativ)

Erster Schritt: Graphische Darstellung (z.B. Streudiagramm), um einen (*subjektiven!*) Eindruck über den Zusammenhang zu erhalten, Abweichungen von der Linearität lassen sich erkennen

Was tun, wenn nicht-linear? Siehe Folie 12



Zweiter Schritt: Berechnung des Korrelationskoeffizienten

Pearson‘s correlation (setzt normalverteilte Variablen voraus→ Feststellung z.B. über Boxplots, Q-Q-plots)
Spearman rank correlation oder **Kendalls Tau** für ordinale oder nicht normalverteilte Variablen, letzterer wird eher für kleine Stichprobenumfänge empfohlen.

2 Korrelationsanalyse

Beispiel: Wir erwarten einen linearen Zusammenhang zwischen „nitrate“ und „livestock“ des „nitrate“ Datensatzes.

Zuerst das Streudiagramm:

```
scatterplot(nitrate$nitrate, nitrate$livestock, smooth=F)
```

Dieses bestätigt subjektiv unsere Vermutung. Zur Auswahl des geeigneten Verfahrens zur Berechnung der Korrelation muss man noch untersuchen, ob die Variablen normalverteilt sind (Methode: Boxplots und Shapiro-Wilk Test).

```
boxplot(nitrate$nitrate)
shapiro.test(nitrate$nitrate)
```

Das Gleiche dann entsprechend für „livestock“.

Normalverteilung ist nicht gegeben, man würde in diesem Fall die nicht-parametrische Variante wählen.

3

Korrelationsanalyse

Dann berechnen wir den Korrelationskoeffizienten:

```
cor.test(nitrate$nitrate, nitrate$livestock, method=„spearman“)
```

Man erhält ein rho von 0.655, also liegt eine „mittelstarke“ positive Korrelation zwischen dem Viehbestand und dem Nitratgehalt vor, die statistisch signifikant ist ($p<0.001$).

Übung 5: Berechnen Sie den Korrelationskoeffizienten von der Bodenart (soil) und dem Nitratgehalt. (Die Bodenart muss zuvor als numerische Variable angelegt werden: Reihenfolge ist SULT)

4 Regressionsanalyse

Eine Regression ist eine mathematische Beschreibung einer Korrelation/ eines Zusammenhangs.
Häufig werden lineare Regressionen verwendet

$$y=mx+b+\varepsilon$$

mit m=Steigung
b= Achsenabschnitt
x= Regressor
ε= Residuum

Die „Least-Square-Methode“ minimiert die Quadratsumme der Residuen. Die Residuen sollten voneinander unabhängig und normalverteilt sein sowie gleiche Varianzen haben. Die Güte der Anpassung des Modells an die Daten wird durch R^2 bestimmt.

Regressionsanalyse

Beispiel: Wir möchten den Nitratgehalt anhand des Viehbestandes „vorhersagen“.

Die lineare Regression wird durchgeführt über:

```
reg.nitrate <- lm(nitrate$nitrate ~ nitrate$livestock)
```

und erhalten durch

```
reg.nitrate
```

folgende Ausgabe:

```
Coefficients: (Intercept) nitrate$livestock
                   -1.7463890  0.0007602
```

also ergibt sich für die Geradengleichung (y=mx+b)
nitrate= 0.00076 * livestock -1.7464

Wenn nicht in den Basispaketen vorhanden, finden sich viele Funktionen zu Regressionen in den Paketen {car} und {MASS}.

Regressionsanalyse

Im nächsten Schritt lassen wir uns eine Zusammenfassung des Modells geben:

```
summary(reg.nitrate)
```

Man erhält den folgenden Output:

```
Call:
lm(formula = nitrate$nitrate ~ nitrate$livestock)

Residuals:
     Min       1Q   Median       3Q      Max
-24.8269  -9.9921  -0.6046   7.8602  30.7168

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.746e+00  6.709e+00  -0.260    0.797
nitrate$livestock  7.602e-04  9.703e-05   7.835 1.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.48 on 28 degrees of freedom
Multiple R-squared:  0.6868,    Adjusted R-squared:  0.6756
F-statistic: 61.39 on 1 and 28 DF,  p-value: 1.556e-08
```

Regressionsanalyse

Diese Ausgabe gibt uns die Information, dass das R^2 0.68 beträgt, das heißt, 68% der Gesamtvarianz wird durch dieses Modell erklärt. Die erklärende Variable livestock erklärt einen signifikanten Anteil des Regressionsmodells ($p<0.001$).

Um herauszufinden, ob das Regressionsmodell geeignet ist um den Zusammenhang zwischen den zwei Variablen zu beschreiben, sollten die Residuen
1) voneinander unabhängig (keine Autokorrelation) und
2) normalverteilt sein und
3) die gleiche Varianz haben.

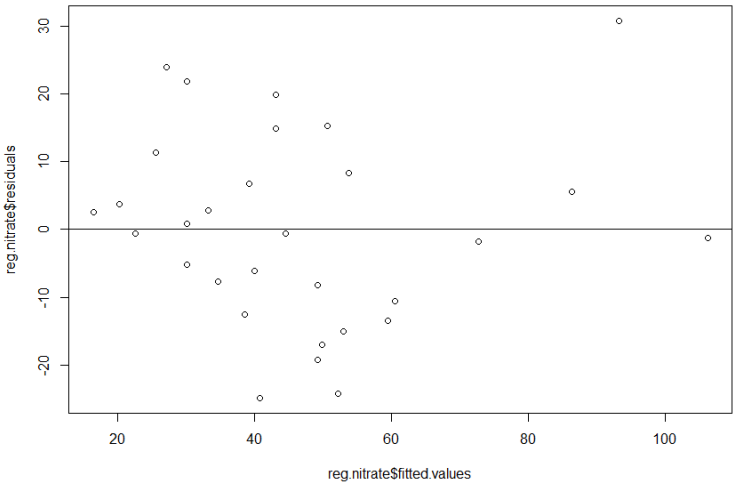
Um 1) zu testen, erzeugt man ein Residuendiagramm und überprüft visuell, ob ein Muster erkennbar ist.

```
plot(reg.nitrate$fitted.values, reg.nitrate$residuals)
abline(h=0)
```
erzeugt eine Nulllinie

Regressionsanalyse

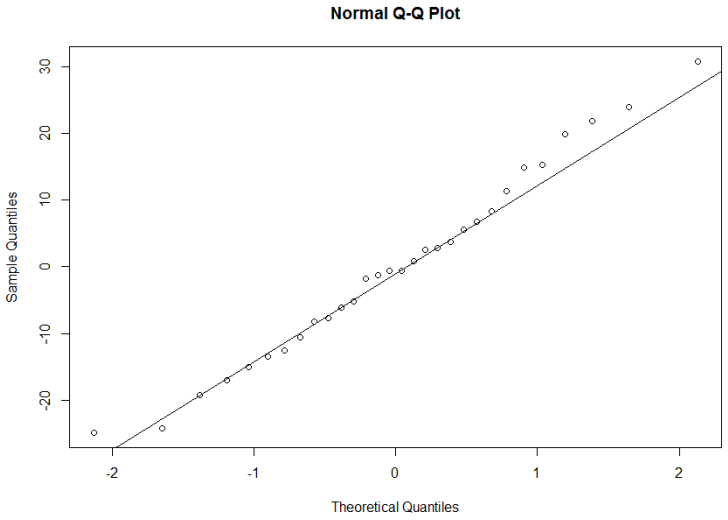
Anhand der Grafik lässt sich kein Muster in den Residuen erkennen, daher ist 1) erfüllt.
Man kann zusätzlich einen Test anwenden:
`durbinWatsonTest(reg.nitrate){car}`

Der erhaltene Wert für die Durbin-Watson Statistik von 1.69 weist nicht auf Autokorrelation der Werte hin (Werte zwischen 1.5 und 2.5 gelten als unauffällig).



Um 2)(Normalverteilung) zu beweisen, kann ein Q-Q-plot erzeugt werden
`qqnorm(reg.nitrate$residuals)`
`qqline(reg.nitrate$residuals)`

Man kann keine nennenswerte Abweichung von der Normalverteilung erkennen. Der Shapiro Wilk Test
`shapiro.test(reg.nitrate$residuals)`
bestätigt dies mit einem p-Wert von 0.9026.



Regressionsanalyse

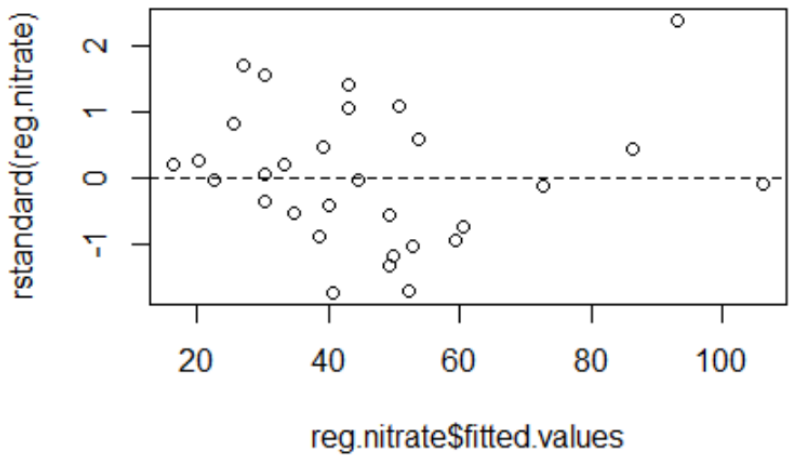
Um 3) (gleiche Varianzen) zu beweisen, werden die standardisierten Residuen (mean=0, sd=1) in einem Diagramm verwendet:

```
plot(reg.nitrate$fitted.values, rstandard(reg.nitrate))
```

Zusätzlich fügen wir eine horizontale Orientierungslinie bei der Höhe 0 ein:

```
abline(h=0, lty=2)
```

lty=2 erzeugt eine gestrichelte Linie



Eine gleichmäßige Punktwolke deutet auf homogene Varianzen der Residuen hin. Kritisch wären ein deutliches Muster oder eine trichterförmige Form sowie zu viele Punkte außerhalb +2/ -2 Standardabweichung (y-Achse).

Diese (bzw. ähnliche) diagnostische Plots erhält man auch über

```
plot(reg.nitrate)
```

Regressionsanalyse

Unser Modell kann also im Prinzip als geeignet betrachtet werden, den Zusammenhang zwischen dem Nitratgehalt und dem Viehbestand zu beschreiben. Das R^2 von 0,68 ist allerdings etwas unbefriedigend (für dieses Fachgebiet!!!)!

Man könnte noch die Bodenart als erklärende Variable in das Modell aufnehmen (würde aber hier den Rahmen sprengen).

Eine recht gute weiterführende Anleitung zu Regressionen findet sich bei QuickR:
https://www.statmethods.net/stats/regression.html

Übung 6

Laden Sie den Testdatensatz „Davis“ (Paket {car}). Inspizieren Sie die Daten sorgfältig!

1. Zeigen Sie, dass ein linearer Zusammenhang zwischen der Größe und dem Gewicht besteht.
   - Streudiagramm
   - Normalverteilte Variablen?
   - p-Wert

2. Führen Sie anhand der Daten eine lineare Regression durch („height“ ist vorherzusagende Variable, „weight“ die erklärende).
   - Streudiagramm (wie oben)
   - Regressionsmodell
   - Residuenanalyse auf Unabhängigkeit
     Normalverteilung
     Homogene Varianzen

12 Exkurs: Transformation von Daten

Transformationen sind ein gängiges Mittel, um z.B. nicht-lineare Beziehungen zu linearisieren oder nicht-normalverteilte Daten zu „normalisieren“.

Transformation ist kein „Hinbiegen“ der Daten, wo kein Zusammenhang ist, kann auch keiner transformiert werden. Die Reihenfolge der Punkte bleibt erhalten, lediglich die Abstände dazwischen werden verändert!
Bei Nicht-Linearität, positiver Schiefe, positiver Kurtosis, ungleichen Varianzen hilft häufig die log-Transformation (z.B. `log10(x)`) oder die Wurzeltransformation (`sqrt()`), tw. auch die Kehrwert-Transformation (`1/x`). Bei Langschwänzigkeit der Verteilung/ Ausreißern eher getrimmten Mittelwert nehmen (`mean(x, trim=0.1)`).