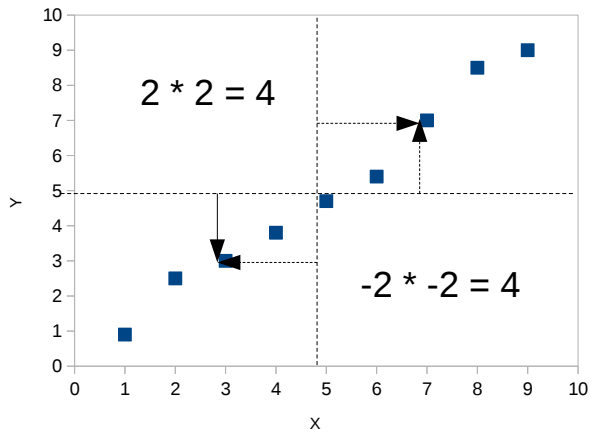
Wissenschaftliche Methodik I

Korrelationsanalyse

Beispieldatensatz zu dieser VL (enthalten im Workspace "Wq.Rdata") :
- ***Wq.csv***

Verwendete Pakete:
- ***ppcor***
- ***corrplot***
- ***PerformanceAnalytics***
- ***Hmisc***
- ***igraph***

Covarianz:
Lineare Beziehung metrischer Daten



X-Werte:
- Mittelwert: 5.0
- Standardabweichung: 2.74

Y-Werte:
- Mittelwert: 4.98
- Standardabweichung: 2.76

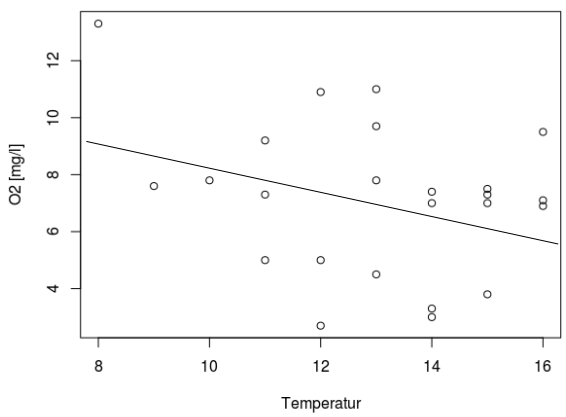
Co-Varianz:

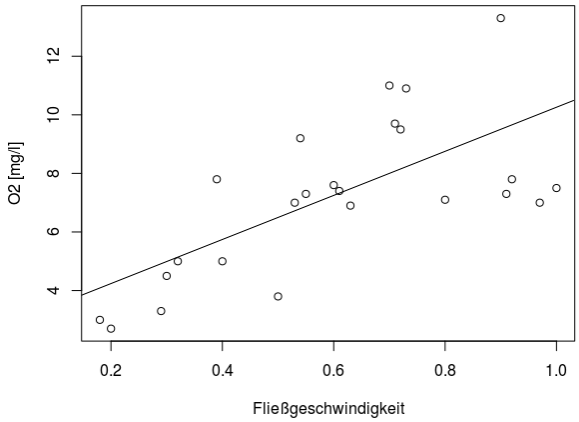
$$Cov_{XY} = \frac{\text{Summe der Abweichungsprodukte}}{(N - 1)}$$

Covarianz:
lineare Beziehung metrischer Daten

Beispiel: Untersuchung der Wasserqualität von Binnengewässern (24 Flüsse)

Sauerstoffkonzentration in mg/l





Covarianz:
lineare Beziehung metrischer Daten

Beispiel: Untersuchung der Wasserqualität von Binnengewässern (24 Flüsse)

Sauerstoffkonzentration in mg/l

$$Cov_{XY} = E\,[\,(X - E(X) * ((Y - E(Y))]$$

$$= \frac{\Sigma^{N}_{i=1}\,(x_i - \bar{x}\,) * (y_i - \bar{y}\,)}{(N - 1)}$$

Korrelation: standardisiertes Maß für lineare Assoziation

Beispiel: Untersuchung der Wasserqualität von Binnengewässern (24 Flüsse)

Sauerstoffkonzentration in mg/l

$$r_{XY} = Cov_{XY} / S_X * S_Y \qquad \text{S: SD}$$

$$= \frac{\Sigma^N_{i=1} (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\Sigma^N_{i=1} (x_i - \bar{x})^2 * \Sigma^N_{i=1} (y_i - \bar{y})^2}}$$

Covarianz: lineare Beziehung metrischer Daten

Beispiel: Covarianz von Nitratgehalt und Phosphat bzw. Ort der Probennahme

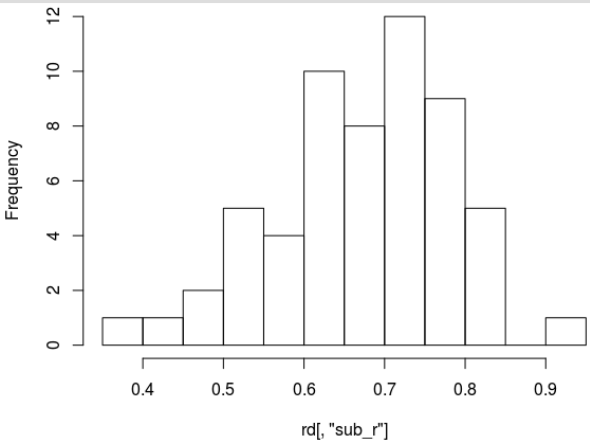
```
> Wq =
read.table("Wq.csv",header=T,sep=",")


> cor(Wq$s,Wq$t)
[1] -0.291754


> cor(Wq$s,Wq$v)
[1] 0.6790226
```

Wertebereich:

- ***-1 – 0.x : negative Korrelation***
- ***0 : keine Korrelation***
- ***0.x - 1 : positive Korrelation***

r ist nicht intervallskaliert: linksschief

```
cor.test(~S+V, data=Wq[sample(1:nrow(Wq), 12),])
```



Z-Transformation nach Fisher:

$$Z_{xy} = \frac{1}{2} \ln \frac{(1 + r)}{(1 - r)}$$

Signifikanz des Korrelationskoeffizienten

- Null-Hypothese: $\rho_{xy} = 0$
- Die Test-Statistik der Z-korrigierten r ist eine t-Statistik: hängt vom Wert von r und vom Probenumfang ab

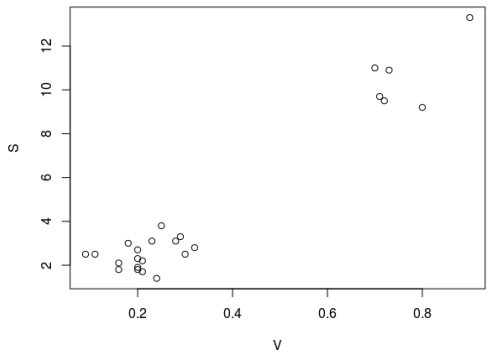
$$t = \frac{|r|}{\sqrt{\frac{1 - r^2}{N - 2}}} = |r| * \sqrt{\frac{N - 2}{1 - r^2}}$$

```
> cor.test(~s+v, data=Wq)
```

Einfluss des Probenumfangs:

N	80% Grenze
5	-0.69 – 0.69
15	-0.35 – 0.35
25	-0.26 – 0.26
50	-0.18 – 0.18
100	-0.13 – 0.13
200	-0.09 – 0.09

Scheinkorrelationen

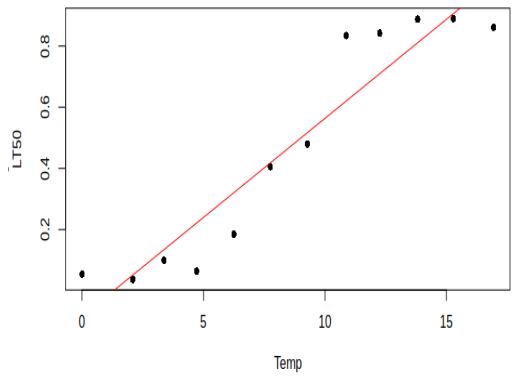


"Daten-Inseln" können Korrelation vortäuschen:

r = 0.97
P = 1.936e-14

Beachte: Pearson-Korrelationen sind nur bestimmbar für zweidimensional normalverteilte (bivariat normalverteilte) Wertepaare!

Nicht-lineare Beziehungen

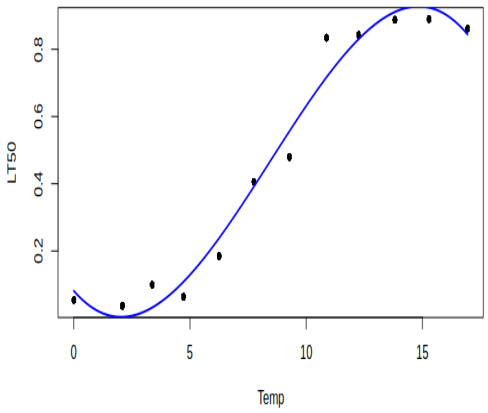


Nicht lineare Beziehungen werden nicht "erkannt":

r = -0.21
P = 0.3159

r = -0.95
P = 3*10^{-}

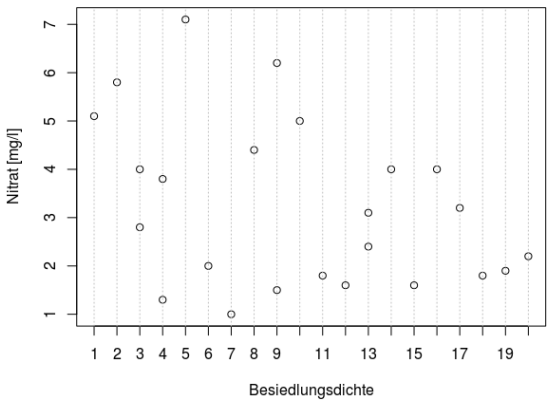
Nicht-lineare Beziehungen: "best fit" testen




```
> lt50 <- mutate(lt50,
                     temp2=temp*temp,
                     temp3=temp*temp*temp,
                     temp4=temp*temp*temp*temp)
```

```
> model.1 <- lm(lt50~temp, data=lt50)
> model.2 <- lm(lt50~temp+temp2, data=lt50)
> model.3 <- lm(lt50~temp+temp2+temp3, data=lt50)
> model.4 <- lm(lt50~temp+temp2+temp3+temp4, data=lt50)

> anova(model.1, model.2, model.3, model.4)

  Res.Df      RSS Df Sum of Sq       F   Pr(>F)
1     10 0.157994
2      9 0.157894  1  0.000099  0.0215 0.887509
3      8 0.033144  1  0.124751 27.0090 0.001257 **
4      7 0.032332  1  0.000812  0.1757 0.687659
```

Korrelation ordinal skalierter Merkmale



Der Korrelationskoeffizient nach Pearson hat 2 Voraussetzungen:

- Metrische Daten
- Beide normalverteilt

Bei ordinalskalierten Daten:

- Rang-Korrelation
- z.B. Spearman, Kendall

Spearman's rho:
Produkt-Moment-Korrelation der Rangreihen beider Werte

```
> cor.test(~n+b, method="spearman", data=Wq)

      Spearman's rank correlation rho

data:  n and b
S = 2991.5, p-value = 0.1534
alternative hypothesis: true rho is not
equal to 0
sample estimates:
       rho
-0.3006536

Warnmeldung:
In cor.test.default(x = c(1, 1.3, 1.5, 1.6,
1.6, 1.8, 1.8, 1.9,  :
Kann exakten p-Wert bei Bindungen nicht
berechnen
```

- Für X und Y werden die Ränge ermittelt
- Die Rangnummern werden korreliert

$$rho^{S}_{XY} = Cov_{r_{XY}} / S_{r_X} * S_{r_Y} \qquad S: SD$$

$$= \frac{\Sigma^{N}_{i=1} (r_{x_i} - \bar{r}_x) * (r_{y_i} - \bar{r}_y)}{\sqrt{\Sigma^{N}_{i=1} (r_{x_i} - \bar{r}_x)^2 * \Sigma^{N}_{i=1} (r_{y_i} - \bar{r}_y)^2}}$$

- Verfahren auch für nicht normalverteilte Daten!

Kendall's tau:
Wahrscheinlichkeit der Übereinstimmung von Rangfolgen

Person i	1	2	3	4	5
x: höchster Abschluss	Abitur	Studium	Haupt-schule	ohne Ab-schluss	Real-schule
y: Eltern	Real-schule	Studium	ohne Ab-schluss	Abitur	Haupt-schule

Person i	2	1	5	3	4
X: höchster Abschluss	1	2	3	4	5
y: Eltern	1	3	4	5	2

Rang 1 (möglich: 1,2,3,4,5)

Rang 3 (möglich: 2,3,4,5)

- Wertepaare nach Rang X sortiert
- Wie häufig sind Ränge konkordant?
- Quotient

$$\frac{(\text{Entsprechung} - \text{Durchbrechung})}{\text{Gesamt}}$$

entspricht Kendall's τ

- Wertebereich: -1 bis 1

Kendall's tau:
Wahrscheinlichkeit der Übereinstimmung von Rangfolgen

Person i	1	2	3	4	5
x: höchster Abschluss	Abitur	Studium	Haupt-schule	ohne Ab-schluss	Real-schule
y: Eltern	Real-schule	Studium	ohne Ab-schluss	Abitur	Haupt-schule

Person i	2	1	5	3	4
X: höchster Abschluss	1	2	3	4	5
y: Eltern	1	3	4	5	2

1-2	1-3	1-4	1-5	3-2	3-4	3-5	4-2	4-5	5-2
+	+	+	+	–	+	+	–	+	–

+ wenn realisierter Rang kleiner

- wenn realisierter Rang größer

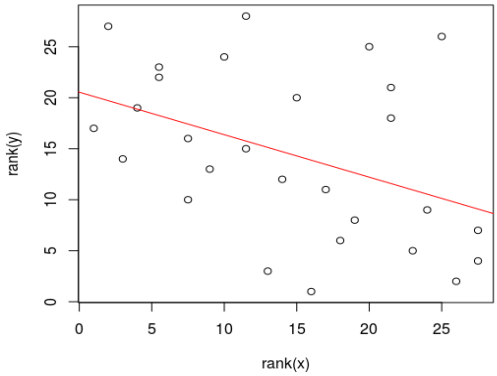
- Berechnung von τ:
- Alle "+" abzüglich aller "-"
  - S = 7 – 3 = 4
- Division durch Anzahl der möglichen Ränge (10):
  - τ = S / (n(n-1)/2) = 4/10

Kendall's τ vs. Spearman's ρ

- Kendall vergleicht den Rang von Y_i mit ALLEN möglichen
- Spearman vergleicht nur den Rang von X_i mit dem von Y_i
- In der Praxis ist ρ oft größer als τ (aber nicht immer!)
- τ ist vorzuziehen, wenn Ränge mehrfach besetzt sind (wenn unterschiedliche Y-Werte denselben X-Wert haben)
- Für kleine N liefert Kendall exaktere p-Werte
- τ ist weniger empfindlich für *outlier*

Kendall, Spearman – oder doch Pearson?

Überlegungen für nicht normal verteilte metrische Daten

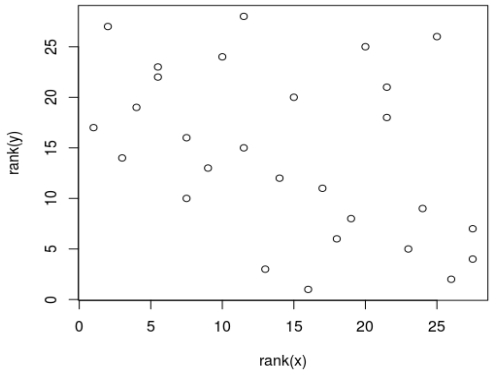


- Bindungen können die Berechnung von p-Werten verzerren
- Dadurch können fälschlich Korrelationen angenommen werden
- Die Pearson-Korrelation transformierter Daten ist dann vorzuziehen!

```
> cor.test(~y+x, method = "spearman")
        Spearman's rank correlation rho
S = 5175, p-value = 0.02757
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.4162674
Warnmeldung:
In cor.test.default(x = c(10.4, 10.3, 8.1, 6.8, 6.7, 7.2, 10.1,  :
  Kann exakten p-Wert bei Bindungen nicht berechnen
```


Kendall, Spearman – oder doch Pearson?

Überlegungen für nicht normal verteilte metrische Daten

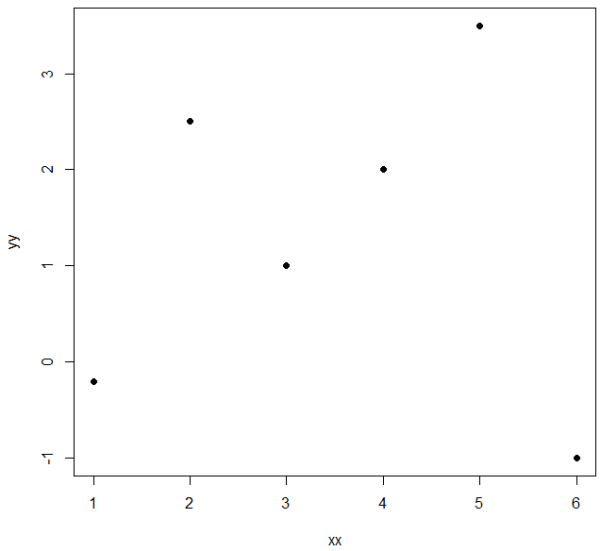


- Bindungen können die Berechnung von p-Werten verzerren
- Dadurch können fälschlich Korrelationen angenommen werden
- Die Pearson-Korrelation transformierter Daten ist dann vorzuziehen!


```
> cor.test(~log(y)+x)
        Pearson's product-moment correlation
data:  log(y) and x
t = -2.0001, df = 26, p-value = 0.05604
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.649726328  0.009160377
sample estimates:
       cor
-0.3651645
```

Kendall, Spearman – oder doch Pearson?

```
> xx <- 1:6
> yy <- c(-0.2,2.5,1,2,3.5,-1)
> plot(xx,yy)

> cor(xx,yy,method="spearman")
[1] -0.02857143
> cor(xx,yy,method="kendall")
[1] 0.06666667
> cor(xx,yy,method="pearson")
[1] 1.748437e-18
```



Korrelation nominal skalierter Merkmale

	oligotroph	eutroph	Σ Zeilen
Napf-schnecke vorhanden	4	9	13
Nicht vorhanden	8	3	11
Σ Spalten	12	12	24

- Ein Zusammenhang nominaler Merkmale kann sich nur auf deren Häufigkeit beziehen!
- Ob die Häufigkeiten zweier Merkmale zusammen hängen, prüft der Chi-Quadrat Test

$$(n_{1,1} - e_{1,1})^2 / e_{1,1} = (4 - \frac{13 * 12}{24})^2 / 6.5 = 0.96$$

Korrelation nominal skalierter Merkmale

	oligotroph	eutroph	Σ Zeilen
Napf-schnecke vorhanden	0.96	0.96	13
Nicht vorhanden	1.14	1.14	11
Σ Spalten	12	12	24

- $\chi^2 = 0.96 + 0.96 + 1.14 + 1.14 = 4.2$
- $p = P(\chi^2 \geq 4.2) = 0.041$

```
> chisq.test(Wq$f, Wq$e, correct=F)

        Pearson's Chi-squared test

data:  Wq$f and Wq$e
X-squared = 4.1958, df = 1,
p-value = 0.04052
```

Korrelation nominal skalierter Merkmale

Voraussetzungen für χ^2 Test:

- alle Häufigkeiten > 5
- nur absolute Häufigkeiten (nie: %)
- Stichproben müssen zufällig sein

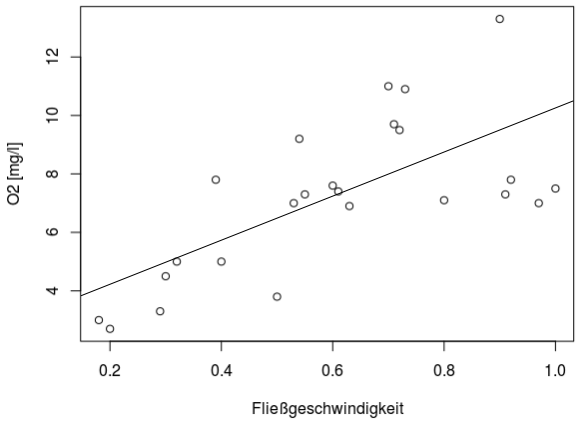
```
> chisq.test(Wq$f, Wq$e)

Pearson's Chi-squared test with
Yates' continuity correction

data:  Wq$F and Wq$E
X-squared = 2.6853, df = 1,
p-value = 0.1013
```

Fisher-Yates Test oder exakter χ^2 Test

Alternativ:

```
>fisher.test(Wq$f, Wq$e)

Fisher's Exact Test for Count Data

data:  Wq$f and Wq$e
p-value = 0.09953
```

Lineare Regression



- vor Beginn wird wird Kausalität festgelegt:
- Prädiktor (unabhängige Variable)
- Zielvariable (abhängige Variable)

- Jeder Messwert der Zielvariablen ist gegeben durch:

  $y_i = f(x_i) + e_i$

- wobei: $f(x_i) = b_0 + b_1 * x$

Lineare Regression

Voraussetzungen:

- Werte von x vorgegeben; [kein] Meßfehler
- es liegt linearer Zusammenhang vor
- Fehler e_i sind unabhängig voneinander
- Fehler e_i normalverteilt, unabhängig von x_i

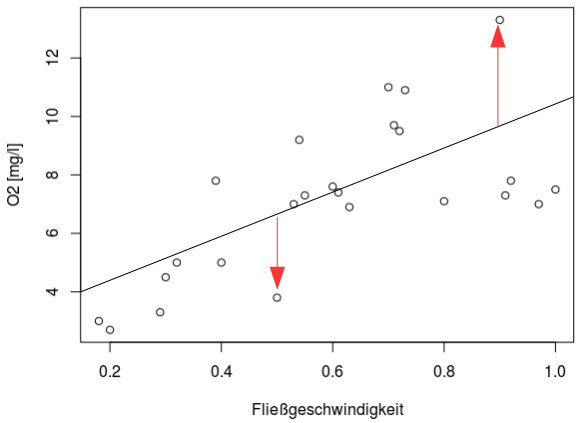
- vor Beginn wird wird Kausalität festgelegt:
- Prädiktor (unabhängige Variable)
- Zielvariable (abhängige Variable)

- Jeder Messwert der Zielvariablen ist gegeben durch:

  $y_i = f(x_i) + e_i$

- wobei: $f(x_i) = b_0 + b_1 * x$

Lineare Regression



Methode der kleinsten Fehlerquadrate

$\Sigma^{N}_{i=1} (y_i - \hat{y}_i)^2 = \Sigma^{N}_{i=1} (y_i - (b_o + b_1 * x_i))^2 \rightarrow \min$

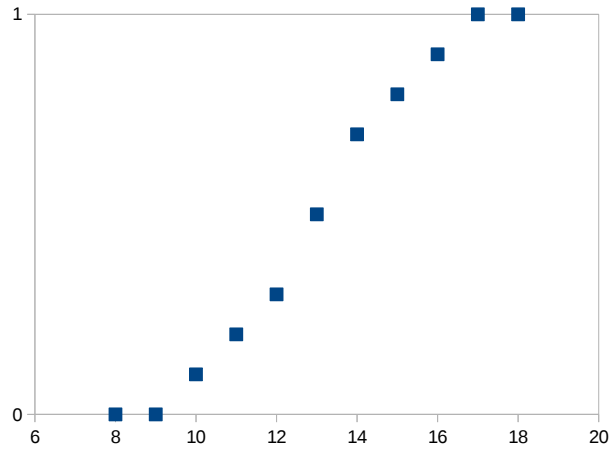
Güte der Regression:

- Anteil der durch $f(x_i)$ erklärten Varianz

Bestimmtheitsmaß:

- (erklärte Varianz) / (gesamte Varianz)
- $R^2 = \Sigma^{N}_{i=1} (\hat{y}_i - \bar{y})^2 / \Sigma^{N}_{i=1} (y_i - \bar{y})^2$

Logit Analyse



- Ein oder mehrere metrische Prädiktoren bestimmen eine „ja/nein“-Variable („Boolesche Variable“)
- Beeinflusst Temperatur den Eutrophiegrad?
- Haben Nitrat, Phosphat oder der O2-Gehalt Einfluss?

Logit Analyse

- LOGIT in R:

>Wq.logit <-glm(e~t,data=Wq,family="binomial")

>summary(Wq.logit)

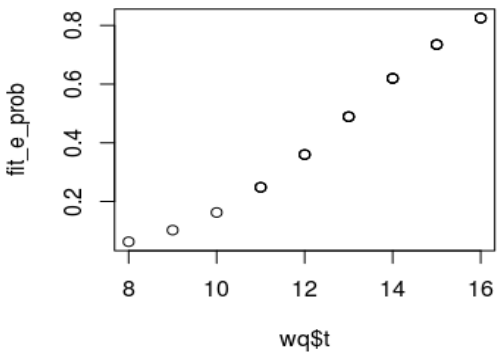
```
> fit_e = predict(Wq.logit, Wq)
> fit_e_prob = exp(fit_e)/(1+exp(fit_e))
> plot(fit_e_prob~Wq$t)
```

Deviance Residuals:

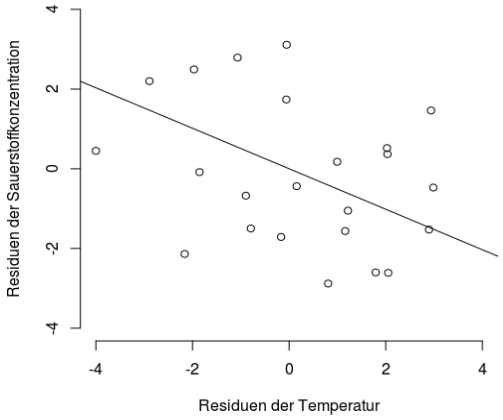
Min	1Q	Median	3Q	Max
-1.8671	-0.9445	0.1300	0.9787	1.6691

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.9555	3.4521	-2.015	0.0439	*
t	0.5316	0.2593	2.051	0.0403	*



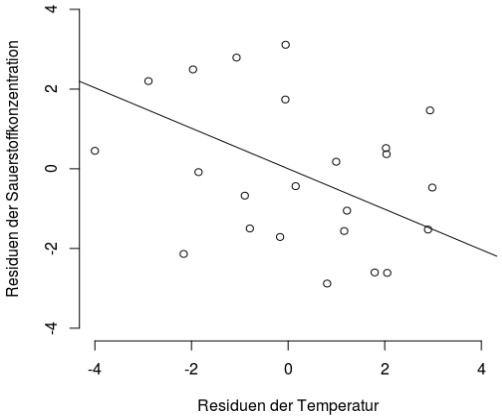
Partielle Korrelationsanalyse



- Zusammenhänge können von Störvariablen überlagert sein
- Ausschluss durch Berechnung der Korrelation von Regressionsresiduen
- Schätzwerte der Störgröße abziehen:

  $x_{res_i} = x_i - \hat{x}_i$; $y_{res_i} = y_i - \hat{y}_i$
- Residuen miteinander Korrelieren

Partielle Korrelationsanalyse



$$r_{XY.Z} = \frac{r_{XY} - r_{XZ} * r_{YZ}}{\sqrt{(1 - r_{XZ}^2) + (1 - r_{YZ}^2)}}$$

Signifikanztest:

$$T_{XY.Z} = \frac{r_{XY.Z} * \sqrt{n - 3}}{\sqrt{1 - r_{XY.Z}^2}}$$

```
> library(ppcor)
> attach(Wq)
> pcor.test(s,t,v, method="pearson")
  estimate    p.value     n
 -0.4516318  0.03051553   24
```

Multiple Regression

```
> fit = lm(s~v+t+q)
> summary(fit)

Call:
lm(formula = s ~ v + t + q)


Coefficients:
            Estimate  Pr(>|t|)
(Intercept)  7.35556  0.019814 *
v            7.84799  0.000322 ***
t           -0.35634  0.140056
q           -0.01440  0.766532
---
```

- Zielvariable wird von mehreren Prädiktoren beeinflusst
- Bsp.: Sauerstoffgehalt hängt von Fließgeschwindigkeit, Temperatur und Quell-Entfernung ab

- Aus einer Vielzahl von Prädiktoren sollen die relevanten identifiziert werden

$$y_i = f(x_{1i}, x_{2i}, ..., x_{ki}) + e_i$$

$$f(X_1, X_2, ..., X_k) = b_o + b_1X_1 + b_2X_2 + ... + b_kX_k$$

Multiple Regression: Beta-Gewichte

```
> fit = lm(s~v+t+q)
> summary(fit)

Call:
lm(formula = s ~ v + t + q)


Coefficients:
            Estimate  Pr(>|t|)
(Intercept)  7.35556  0.019814 *
v            7.84799  0.000322 ***
t           -0.35634  0.140056
q           -0.01440  0.766532
---
```

```
> scS = scale(Wq$s)
> scV = scale(Wq$v)
> scT = scale(Wq$t)
> scQ = scale(Wq$q)
> scwq <- data.frame(scS,scV,scT,scQ)
> lm(scS ~ scV+scT+scQ, data=scwq)


Coefficients:

(Intercept)  -2.101e-16
scV           0.7222
scT          -0.2948
scQ          -0.06328
```

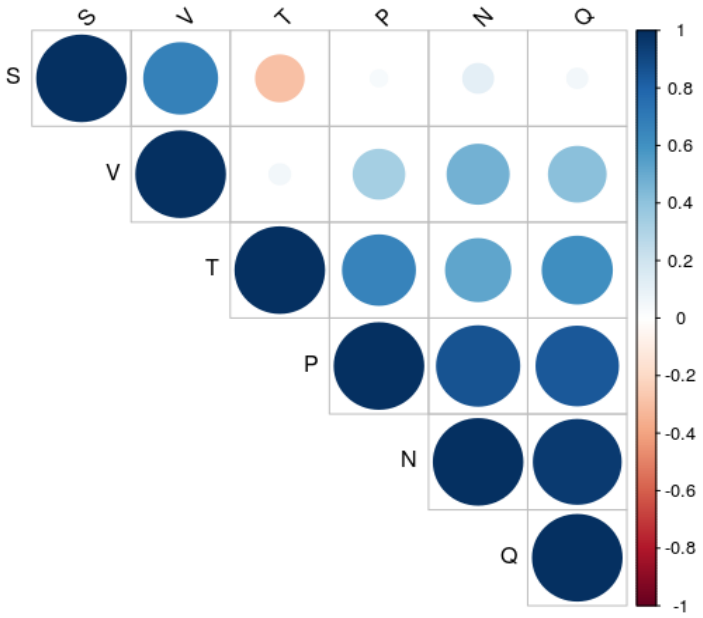
Multikollinearität
Redundanz und Suppression

r_{XY}	S	V	T	Q
S	1	0.68**	-0.29	0.05
V		1	0.06	0.41*
T			1	0.61**
Q				1

- Wenn Prädiktoren miteinander **korreliert** sind, beeinflusst das ihre Korrelation mit der Zielvariablen:
- **gleichsinnig**: die Prädiktoren sind redundant (→β-Gewicht niedrig)
- **gegensinnig**: nicht korrelierter Prädiktor supprimiert Varianzen des anderen (z.B. $r > 0$; $\beta < 0$)

Verarbeiten großer Datensätze

Korrelationsmatrix, Correlogram



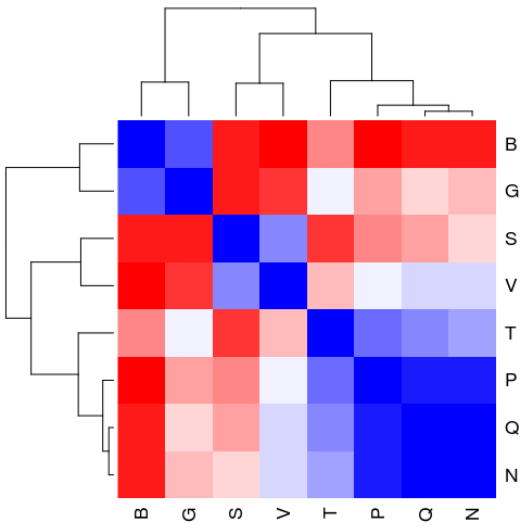

```
> res <- cor(Wq[,2:7])
> res
```

```
> library(corrplot)
> corrplot(res, type = "upper",
  order = "hclust", tl.col = "black",
  tl.srt = 45)
```

```
>  res <- cor(Wq[,2:9],
   method="spearman")
```

Verarbeiten großer Datensätze

Korrelationsmatrix, Correlogram, Heatmap



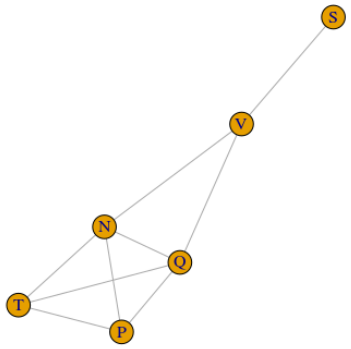

```
>  res <- cor(Wq[,2:9],
   method="spearman")
```

```
> library("PerformanceAnalytics")
> col<- colorRampPalette(c("red",
  "white", "blue"))(20)
> heatmap(x = res, col = col,
  symm = TRUE)
```

Netzwerk Analyse

```
> library(igraph)
> net <- graph_from_data_frame(d=res5, directed=F)
> plot(net)
```

```
> plot(net, layout=layout_with_fr(net))
```




```
> library(Hmisc)
> res2<-rcorr(as.matrix(wq[,2:7]))
```

```
> flattenCorrMatrix <-
  function(cormat, pmat) {
    ut <- upper.tri(cormat)
    data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor  =(cormat)[ut],
    p = pmat[ut]
    )
   }
```

```
> res3 <- flattenCorrMatrix(res2$r, res2$P)
> res4 <- subset(res3, p<0.05)
> res5 <- data.frame(res4[,1:3])
> colnames(res5) <- c("from", "to", "weight")
```