

# Nichtlineare Effekte in der linearen Regression

## Visualisierung Regression

Jan-Philipp Kolb

Freitag, 20.06.2014



# Inhalt

Einführung

Graphiken mit R

Ergebnisse mit LaTeX

Graphische Analyse zur Vorbereitung der Regression

Datenanalyse

Regressionsdiagnostik

Konstante Varianz

Normalität

# Gliederung

Einführung

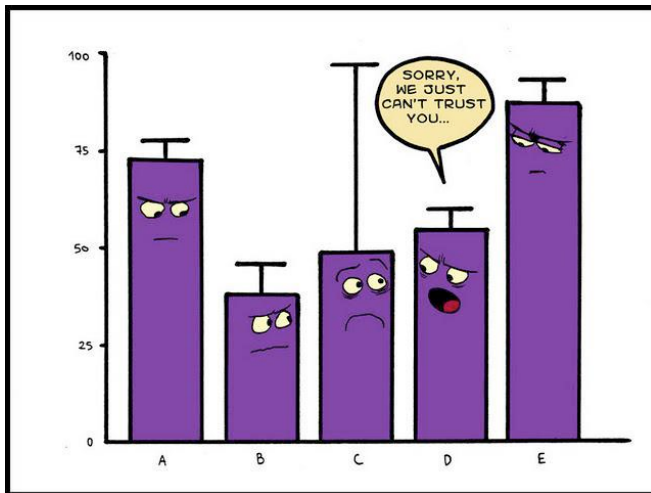
Graphiken mit R

Ergebnisse mit LaTeX

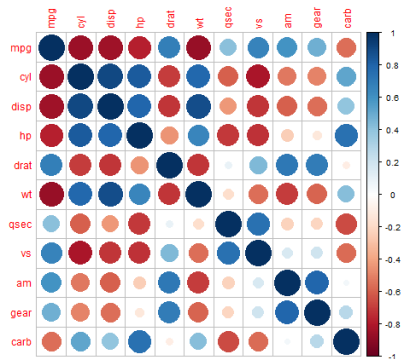
Graphische Analyse zur Vorbereitung der Regression

Datenanalyse

Regressionsdiagnostik



## Worum geht es in diesem Abschnitt?



### ► Visualisierung und linearen Regression

# Exkurs - R und LaTeX

Table 1: The Importance of Clustering Standard Errors

	M1	M2	M3
(Intercept)	4.35*** (1.32)	4.35*** (0.76)	4.35*** (1.32)
x	2.26* (1.27)	2.26*** (0.73)	2.26* (1.32)
R <sup>2</sup>	0.00	0.00	0.00
Adj. R <sup>2</sup>	0.00	0.00	0.00
Num. obs.	1000	3000	3000

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

Quelle: <http://diffuseprior.wordpress.com/tag/stargazer/>

## Das R-Paket xtable

- ▶ R bietet tolle Möglichkeiten mit LaTeX zu interagieren
- ▶ Zusammen mit word funktioniert das leider weniger gut (höchstens Paket R2wd)

---

```
N <- 1000
```

```
a <- sample(1:5, N, replace=T)
```

```
b <- sample(1:3, N, replace=T)
```

```
tabAB <- table(a, b)
```

```
xtable(tabAB)
```

---

## Das R-Paket stargazer

---

```
stargazer(attitude)
```

---

Table :

Statistic	N	Mean	St. Dev.	Min	Max
rating	30	64.633	12.173	40	85
complaints	30	66.600	13.315	37	90
privileges	30	53.133	12.235	30	83
learning	30	56.367	11.737	34	75
raises	30	64.633	10.397	43	88
critical	30	74.767	9.895	49	92
advance	30	42.933	10.289	25	72



## Mehr zu reproducible research mit R

- ▶ Das R-Paket `library(texreg)`
- ▶ Das R-Paket `library(R2wd)`
- ▶ <http://knutur.at/wsmt/slides/long-slides.pdf>
- ▶ R Task View zu reproducible research:

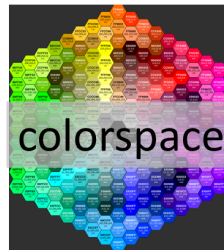
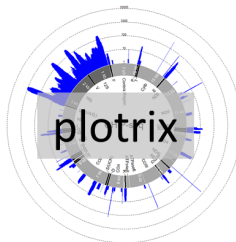
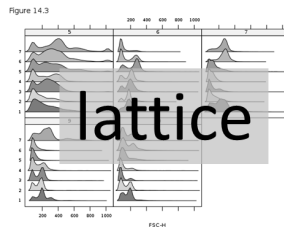
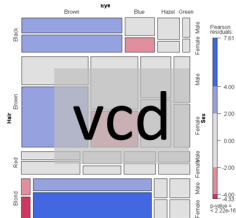
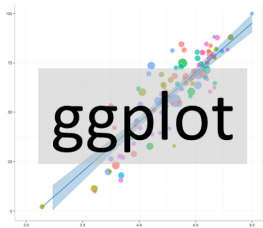
**CRAN Task View: Reproducible Research**

**Maintainer:** Max Kuhn

**Contact:** max.kuhn at pfizer.com

**Version:** 2014-01-20

## Pakete - deskriptive Datenanalyse

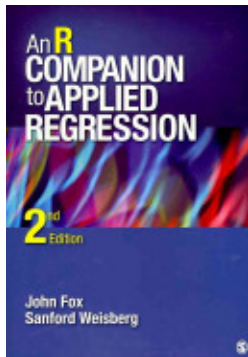


## Vorbereitung einer Regression

Vor jeder Analyse, sollten die Daten auf folgende Punkte untersucht werden:

- ▶ Dateneingabe Fehler
- ▶ Fehlende Werte
- ▶ Ausreißer
- ▶ Unübliche (e.g. unsymmetrische) Verteilungen
- ▶ Veränderungen in der Variabilität
- ▶ Cluster-Effekte
- ▶ Nicht-linear bivariate Zusammenhänge
- ▶ Unerwartete Muster

## Der Prestige Datensatz



- ▶ Benötigte library: `car`
- ▶ Als Datenmaterial werden die *Prestige* Daten verwendet.
- ▶ Zugrunde liegen 102 Beobachtungen.
- ▶ Variablen des Datensatzes:
  - ▶ **income**
  - ▶ **education**
  - ▶ **women**
  - ▶ **prestige**
  - ▶ **census**
  - ▶ **type**

# Der Prestige Datensatz

---

```
library(car)
data(Prestige)
attach(Prestige)
```

---

Prestige {car}

R Documentation

## Prestige of Canadian Occupations

### Description

The `Prestige` data frame has 102 rows and 6 columns. The observations are occupations.

## Der Prestige Datensatz

education	Average education of occupational incumbents, years (1971)
income	Average income of incumbents, dollars, in 1971.
women	Percentage of incumbents who are women.
prestige	Prestige score for occupation
census	Canadian Census occupational code.
type	Type of occupation. A factor with levels

## Eine Beispielsession in R

Ausschnitt des Datensatzes:

```
> Prestige
```

	education	income	women	prestige	census	type
GOV.ADMINISTRATORS	13.11	12351	11.16	68.8	1113	prof
GENERAL.MANAGERS	12.26	25879	4.02	69.1	1130	prof
ACCOUNTANTS	12.77	9271	15.70	63.4	1171	prof
PURCHASING.OFFICERS	11.42	8865	9.11	56.8	1175	prof
CHEMISTS	14.62	8403	11.68	73.5	2111	prof
PHYSICISTS	15.64	11030	5.13	77.6	2113	prof
BIOLOGISTS	15.09	8258	25.65	72.6	2133	prof
...						
TAXI.DRIVERS	7.93	4224	3.59	25.1	9173	bc
LONGSHOREMEN	8.37	4753	0.00	26.1	9313	bc
TYPESETTERS	10.00	6462	13.58	42.2	9511	bc
BOOKBINDERS	8.55	3617	70.87	35.2	9517	bc

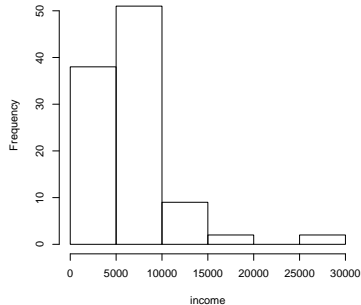
```
>
```

# Univariate Datenanalyse

## Histogramm der Variable *income*

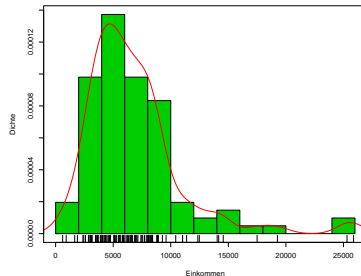
```
hist(income)
```

Histogram of income



```
hist(income, probability=T)  
points(density(income),  
type="l", col="red")
```

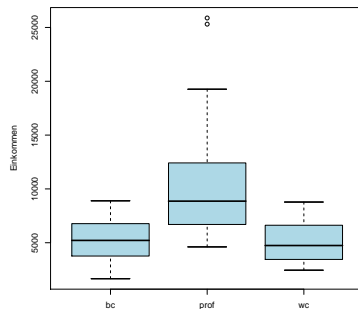
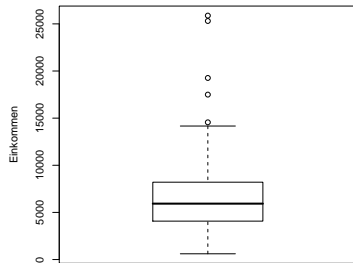
Histogramm: Data Prestige





# Univariate Datenanalyse

## Boxplot der Variable *income*



## Boxplot

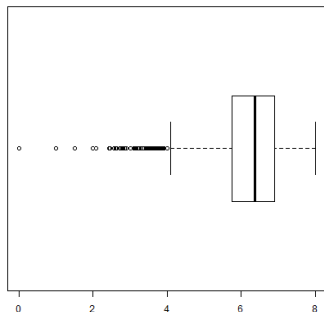
- ▶ Einen einfachen Boxplot erstellt man mit `boxplot()`
- ▶ Auch `boxplot()` muss mindestens ein Beobachtungsvektor übergeben werden

---

?`boxplot`

```
boxplot(Chem97$gcsescore,  
horizontal=TRUE)
```

---



## Gruppierte Boxplots

- ▶ Ein sehr einfacher Weg, einen ersten Eindruck über bedingte Verteilungen zu bekommen ist über sog. Gruppierte notched Boxplots
- ▶ Dazu muss der Funktion `boxplot()` ein sog. Formel-Objekt übergeben werden
- ▶ Die bedingende Variable steht dabei auf der rechten Seite einer Tilde

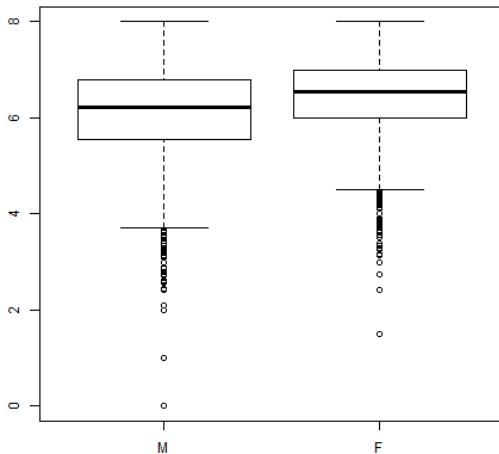
### Die Funktion `boxplot()`

---

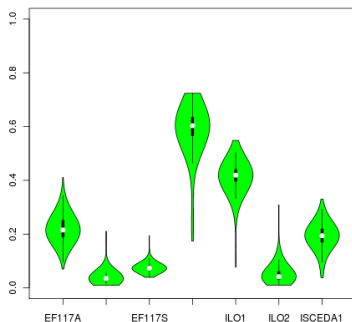
```
boxplot(Chem97$gcsescore~Chem97$gender)
```

---

## Gruppierte Boxplots



## Violinplot - *library(vioplot)*



- ▶ Baut auf Boxplot auf
- ▶ Zusätzlich Informationen über Dichte der Daten
- ▶ Dichte wird über Kernel Methode berechnet.
- ▶ weißer Punkt - Median
- ▶ Je weiter die Ausdehnung, desto größer ist die Dichte an dieser Stelle.

# Edgar Anderson's Iris Daten



WIKIPEDIA  
The Free Encyclopedia

[Main page](#)

[Contents](#)

[Featured content](#)

[Current events](#)

[Random article](#)

[Donate to Wikipedia](#)

[Wikimedia Shop](#)

▼ Interaction

[Help](#)

[About Wikipedia](#)

[Community portal](#)

[Recent changes](#)

[Contact page](#)

Article [Talk](#)

[Read](#)

[Edit](#)

[View history](#)

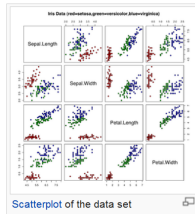
[Create account](#) [Log in](#)

## Iris flower data set

From Wikipedia, the free encyclopedia

The **Iris flower data set** or **Fisher's Iris data set** is a [multivariate data set](#) introduced by [Sir Ronald Fisher](#) (1936) as an example of [discriminant analysis](#).<sup>[1]</sup> It is sometimes called **Anderson's Iris data set** because [Edgar Anderson](#) collected the data to quantify the morphologic variation of [Iris](#) flowers of three related species.<sup>[2]</sup> Two of the three species were collected in the [Gaspé Peninsula](#) "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".<sup>[3]</sup>

The data set consists of 50 samples from each of three species of *Iris* (*Iris setosa*, *Iris virginica* and *Iris versicolor*). Four [features](#) were measured from each sample: the length and the width of the [sepals](#) and [petals](#), in centimetres. Based on the combination of these four features, Fisher developed a [linear discriminant model](#) to distinguish the species from each other.



## Edgar Anderson's Iris Daten

petal length and width	Blütenblatt Länge und Breite
sepal length and width	Kelchblatt Länge und Breite

```
> head(iris)
```

	sepal.Length	sepal.width	petal.Length	petal.width	species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

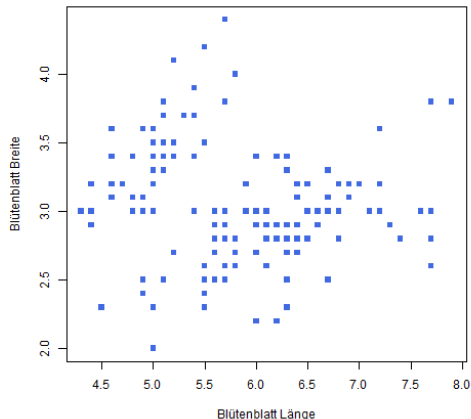
## Scatterplots

- ▶ Ein einfacher two-way scatterplot kann mit der Funktion `plot()` erstellt werden
- ▶ `plot()` muss mindestens ein `x` und ein `y` Beobachtungsvektor übergeben werden
- ▶ Um die Farbe der Plot-Symbole anzupassen gibt es die Option `col` (Farbe als character oder numerisch)
- ▶ Die Plot-Symbole selbst können mit `pch` (plotting character) angepasst werden (character oder numerisch)
- ▶ Die Achsenbeschriftungen (labels) werden mit `xlab` und `ylab` definiert



# Scatterplot

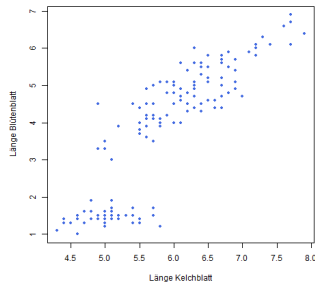
```
plot(iris$Sepal.Length, iris$Sepal.Width, xlab='Blattlaenge',  
     ylab='Blattbreite', pch=15, col='royalblue')
```



## Pearson Korrelationskoeffizient

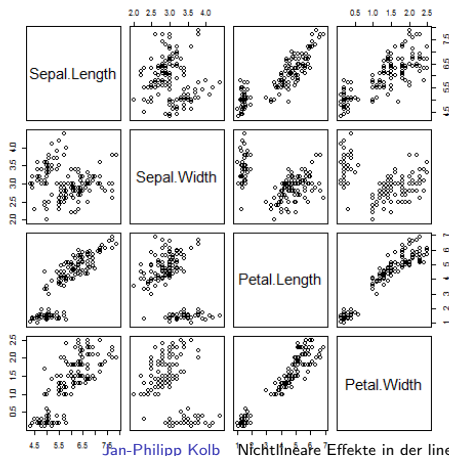
```
cor(iris$Sepal.Length, iris$Petal.Length)
```

- ▶ Korrelation zwischen Länge Kelchblatt und Blütenblatt 0,87
- ▶ Der Pearson'sche Korrelationskoeffizient ist die default methode in `cor()`:



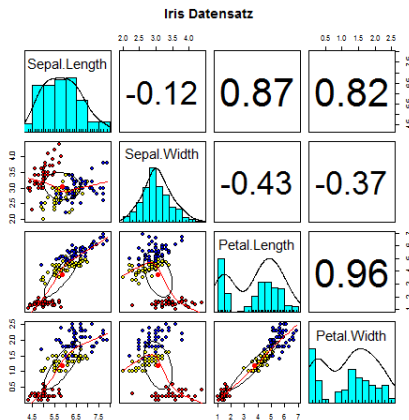
## Zusammenhang zwischen mehreren Variablen

```
pairs(iris[,1:4])
```

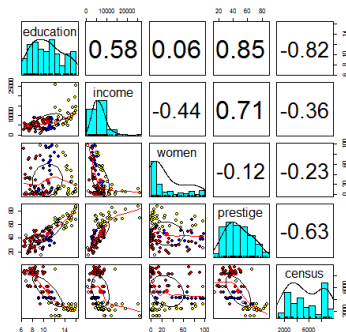
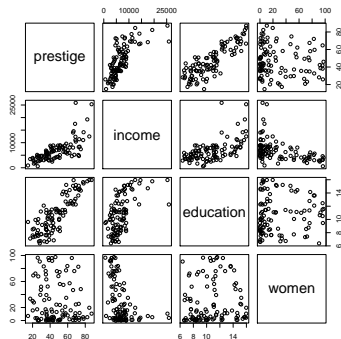


## Zusammenhang zwischen mehreren Variablen

```
pairs.panels(iris[1:4], bg=c("red", "yellow", "blue")  
[iris$Species], pch=21, main="Iris Datensatz")
```

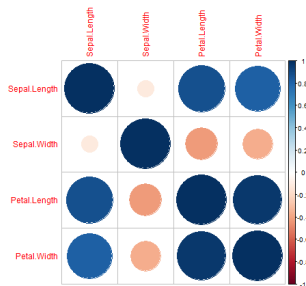


# Multivariate Datenanalyse



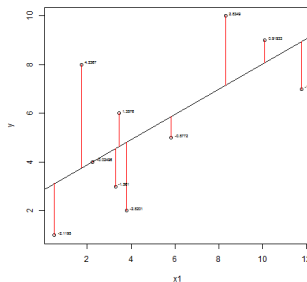
## Zusammenhang - corrplot

```
library(corrplot)
M <- cor(iris[,1:4])
corrplot(M)
```



# Regressionsdiagnostik

```
plot(x1,y)  
abline(mod1)  
segments(x1, y, x1, pre, col="red")  
textxy(x1,y, res, cx=0.7)
```



# Annahmen der Linearen Einfachregression

## Annahmen bezüglich des Störterms

1. Die Störterme haben den Erwartungswert 0:

$$E(\varepsilon_i | x_i) = 0$$

2. Die Störterme weisen eine konstante Varianz auf (Homoskedastizität)

$$\text{var}(\varepsilon_i | x_i) = \sigma_\varepsilon^2$$

3. Die Störvariablen sind unkorreliert

$$\text{cov}(\varepsilon_i, \varepsilon_j | x_i) = 0 \quad \text{für alle } i \neq j$$

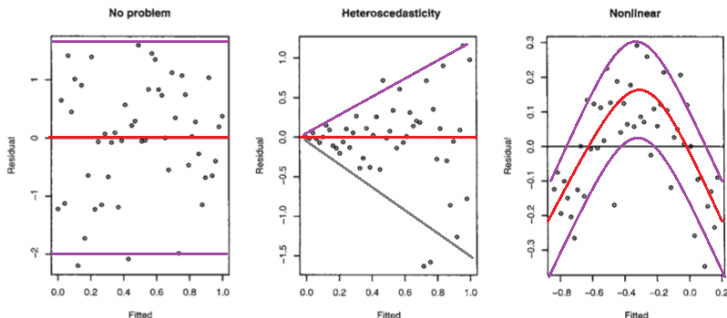
4. Normalverteilungsannahme

$$\varepsilon_i | x_i \sim N(0; \sigma_\varepsilon^2)$$



## Regressionsdiagnostik - konstante Varianz

- ▶ Wenn man nur auf die Residuen schaut kann man die Annahme der konstanten Varianz nicht überprüfen
- ▶ Man muss die Residuen zu etwas in Relation setzen
- ▶ → Plot der Residuen gegen die gefitteten Werte



<http://stats.stackexchange.com/questions/76226/>

interpreting-the-residuals-vs-fitted-values-plot-for-verifying-the-assumptions

Jan-Philipp Kolb

Nichtlineare Effekte in der linearen Regression

## Regressionsdiagnostik - konstante Varianz

Zur Verdeutlichung wird folgendes Regressionsmodell gerechnet:

---

```
data(savings)
g <- lm(sr ~ pop15+pop75+dpi+ddpi, savings)
```

---

Überprüfung anhand von Graphiken:

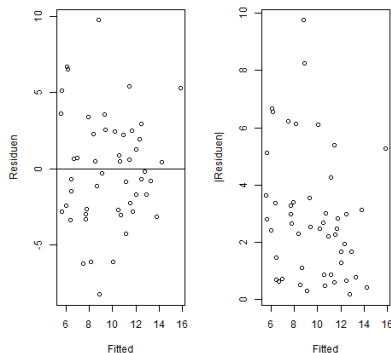
---

```
par(mfrow=c(1,2))
plot(fitted(g), residuals(g), xlab="Fitted",
     ylab="Residuen")
abline(h=0)
plot(fitted(g), abs(residuals(g)), xlab="Fitted",
     ylab="|Residuen|")
```

---

## Regressionsdiagnostik - konstante Varianz

- ▶ **Linker Plot** - Nichtlinearität?
- ▶ **Rechter Plot** - Nichtkonstante Varianz?



## Regressionsdiagnostik - konstante Varianz

- Folgende Regression bietet eine einfache Möglichkeit zu testen.

---

```
summary(lm(abs(residuals(g)) ~ fitted(g)))
```

---

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	4.8398	1.1865	4.079	0.00017	***
fitted(g)	-0.2035	0.1185	-1.717	0.09250	.

## Regressionsdiagnostik - Normalität

- ▶ Tests und KI die bei der Regression verwendet werden basieren auf der Annahme der Normalität.
- ▶ Die Residuen können mit dem Q-Q plot auf Normalität untersucht werden

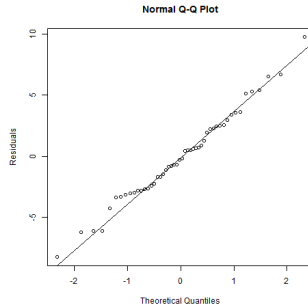
---

```
qqnorm(residuals(g), ylab="Residuals")  
qqline(residuals(g))
```

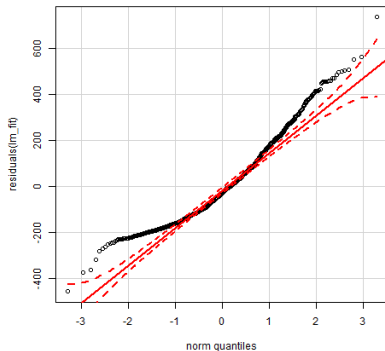
---

## Q-Q plot

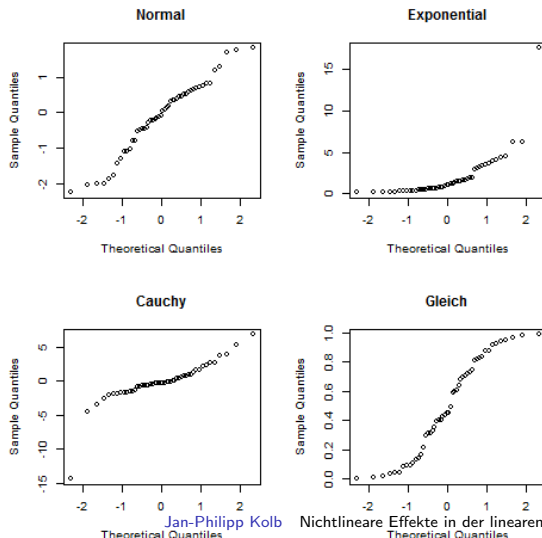
- ▶ Es scheint keine Ausreißer zu geben
- ▶ Wenn die Residuen normalverteilt sind sollten sie auf einer Linie liegen.



## Regressionsdiagnostik - Normalverteilung der Residuen



# Q-Q plot - verschiedene Verteilungen





## Regressionsdiagnostik - Normalität

- ▶ Wenn die Residuen nicht normalverteilt sind, ist die KQ-Regression evtl. nicht die richtige Wahl.
- ▶ Tests und KI's sind nicht exakt
- ▶ Leichte Nichtnormalität kann ignoriert werden
- ▶ Je größer der zu Grunde liegende Datensatz, je geringer ist das Problem mit Nichtlinearität

Handlungsmöglichkeiten:

- ▶ Evtl. können robuste Regressionen eingesetzt werden
- ▶ Transformation der abhängigen Variable

## Regressionsdiagnostik

Artikel, der die wichtigsten Annahmen aufzählt und darstellt, wie sie überprüft werden können.



**Testing the assumptions of linear regression**

<http://people.duke.edu/~rnau/testing.htm>

## Aufgabe C1 - Visualisierung Regression

- ▶ Laden Sie den **Prestige** Datensatz aus dem Paket **car** ein.
- ▶ Finden Sie das beste Regressionsmodell für die Variable **Prestige** als abhängige Variable.
- ▶ Testen Sie ob die Residuen normalverteilt sind.