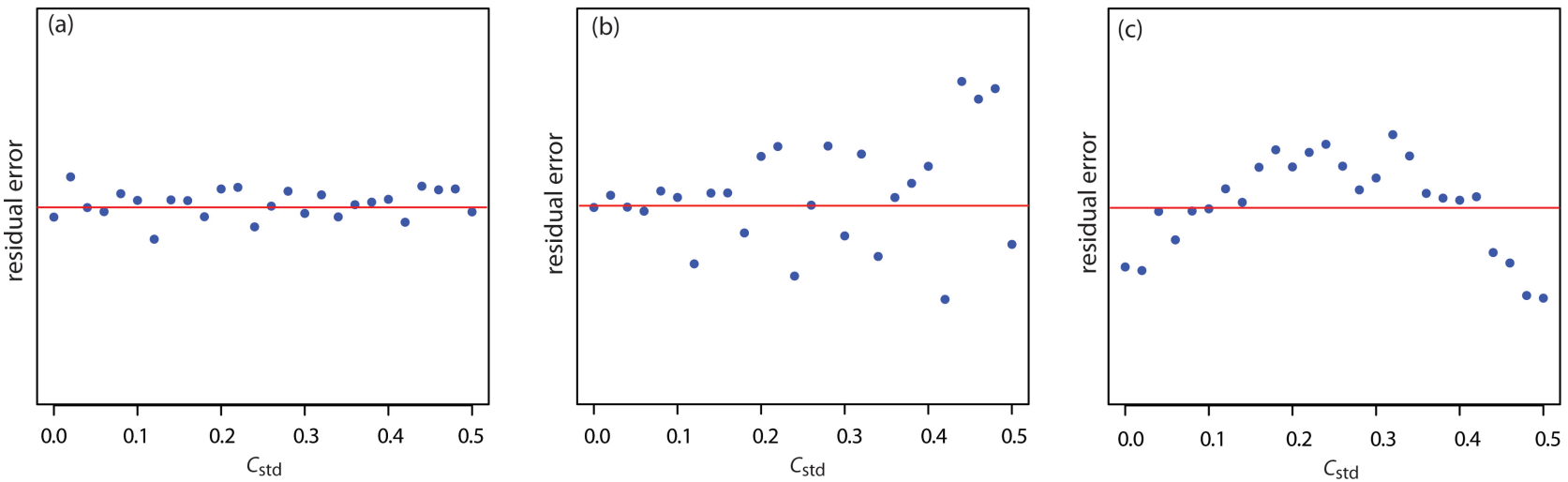
Nichtlineare Effekte in der linearen Regression

Transformationen

Jan-Philipp Kolb

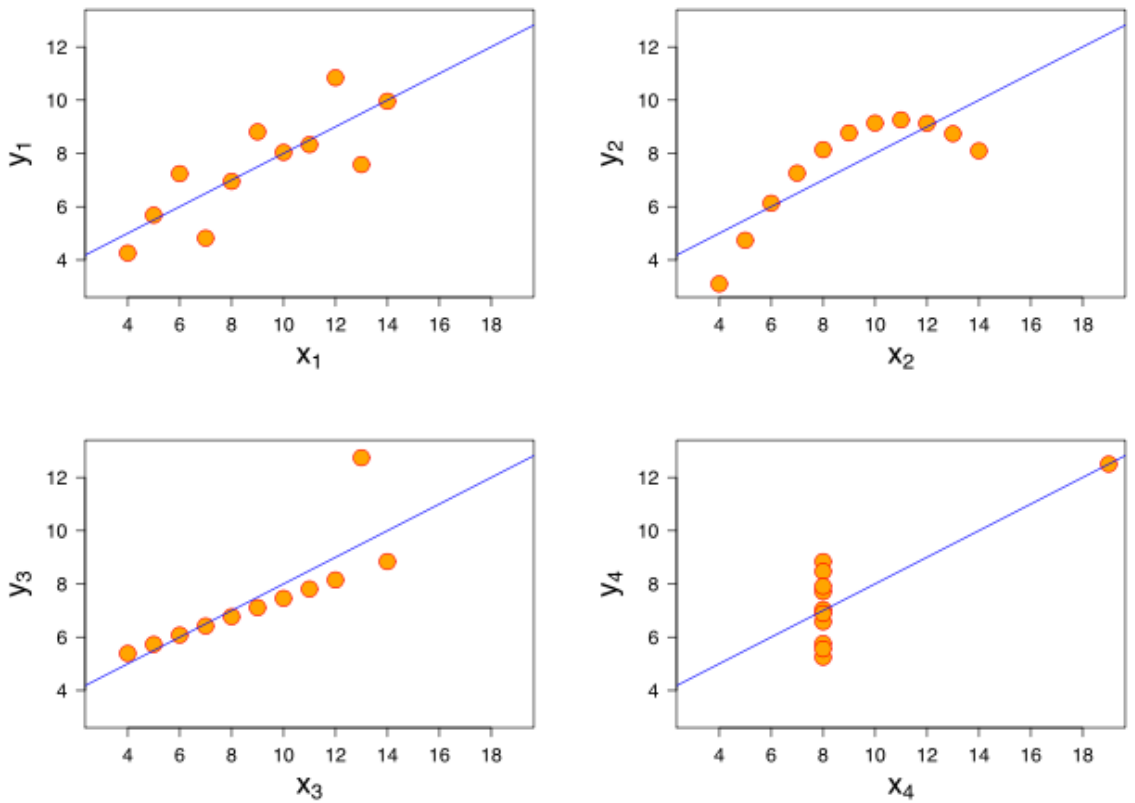
Freitag, 20.06.2014

Regression - nicht ganz linear



Quelle: `http://chemwiki.ucdavis.edu/@api/deki/files/12883/Figure5.13.jpg`

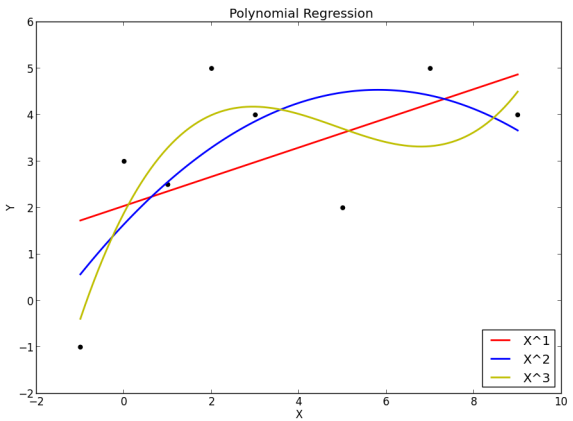
Worum gehts?



Überall die gleiche Regressionsgerade aber sehr unterschiedliche zu Grunde liegende Datensätze

Quelle: `http://en.wikipedia.org/wiki/Linear_regression`

Die polynomiale Regression

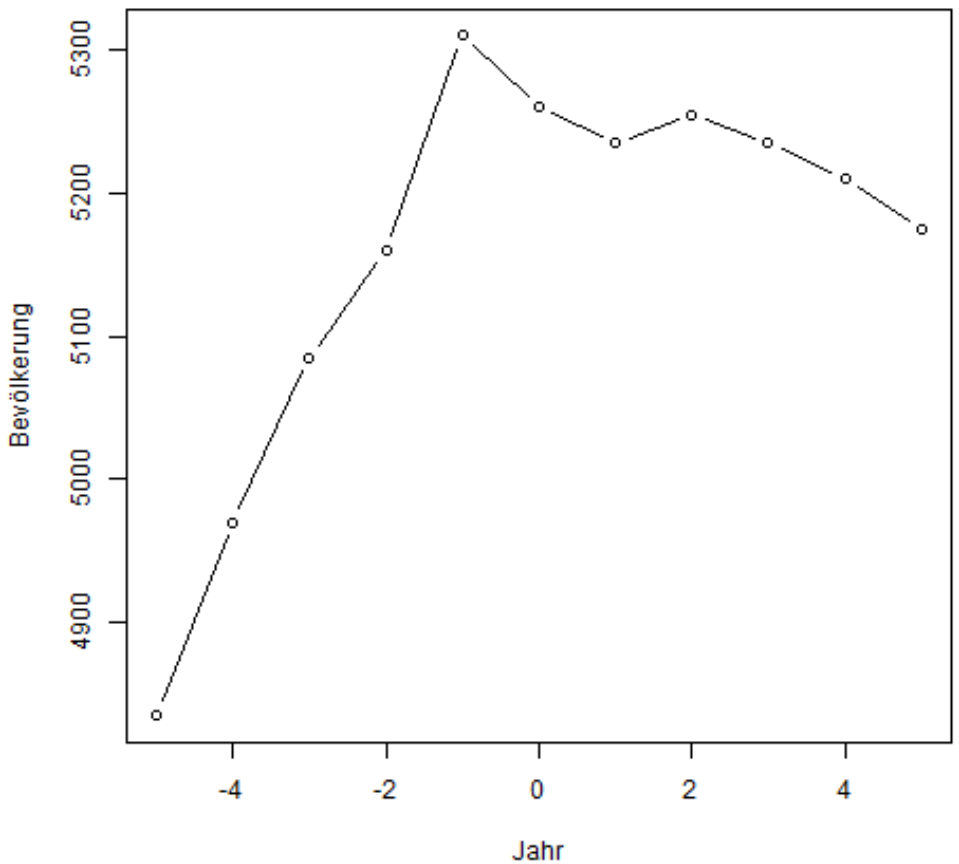


Quelle: `hhttp://www.libresoft.es/node/323`

Bevölkerung einer italienischen Stadt über 10 Jahre

Jahr	Bevölkerung
1959	4835
1960	4970
1961	5085
1962	5160
1963	5310
1964	5260
1965	5235
1966	5255
1967	5235
1968	5210
1969	5175

Bevölkerung einer italienischen Stadt über 10 Jahre

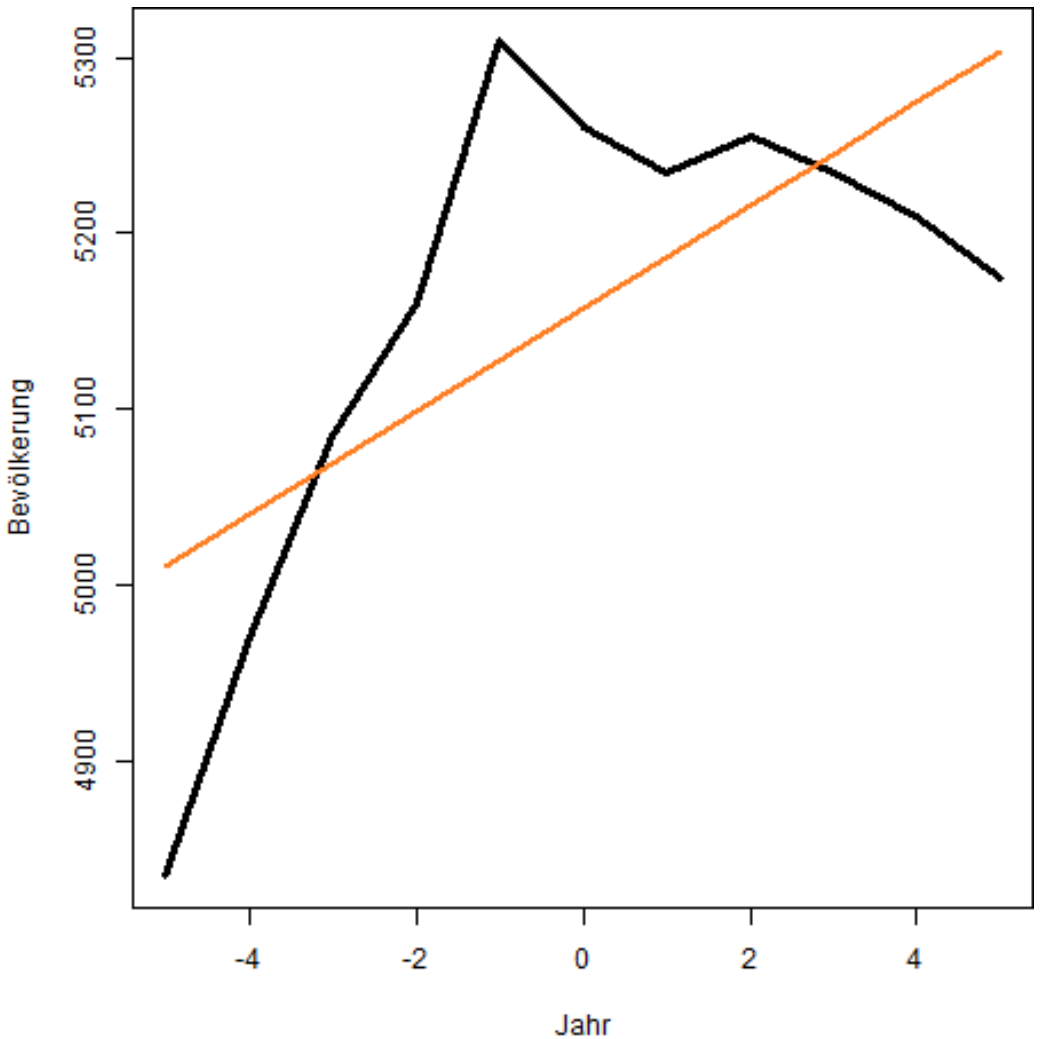


Lineares Modell

```
fit1 <- with(sample1,lm(Population ~ Year))
summary(fit1)
```

```
Call:
lm(formula = Population ~ Year)

Residuals:
    Min      1Q  Median      3Q     Max
-175.68  -67.27   15.68   54.89  182.04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5157.27      33.06 155.988   <2e-16 ***
Year           29.32      10.46   2.804   0.0206 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109.7 on 9 degrees of freedom
Multiple R-squared:  0.4663, Adjusted R-squared:  0.407
F-statistic: 7.863 on 1 and 9 DF,  p-value: 0.02057
```

Lineares Modell - Anpassung an die Daten



Polynome

Polynomiales Modell, das nahe an die Daten heran kommt:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_n x^n$$

- Für eine hohe Genauigkeit muss ein hoher Grad des Polynoms gewählt werden.
- Je größer die Genauigkeit, desto komplizierter Berechnungen sind nötig...

Geschätzte β-Koeffizienten ersten, zweiten und dritten Grades:

```
fit1 <- with(sample1,lm(Population ~ Year))
fit2 <- with(sample1,lm(Population ~ Year + I(Year^2)))
fit3 <- with(sample1,lm(Population ~ Year + I(Year^2)
    + I(Year^3)))
```

Schneller geht's so:

```
fit2b <- lm(sample1$Population ~ poly(sample1$Year, 2,
raw=TRUE))
fit3b <- lm(sample1$Population ~ poly(sample1$Year, 3,
raw=TRUE))
```

	Dependent variable:
	Population
Year	29.318^{***}
	(3.696)
I(Year^2)	-10.589^{***}
	(1.323)
Constant	$5{,}263.159^{***}$
	(17.655)
Observations	11
R^2	0.941
Adjusted R^2	0.926
Residual Std. Error	38.762 (df = 8)
F Statistic	63.478^{***} (df = 2; 8)
Note:	*p<0.1; **p<0.05; ***p<0.01

Gleichung des Polynoms

Die Gleichung des Polynoms des zweiten Grads für unser Modell ist:

$$f(x) = 5263.1597 + 29.318x - 10.589x^2$$

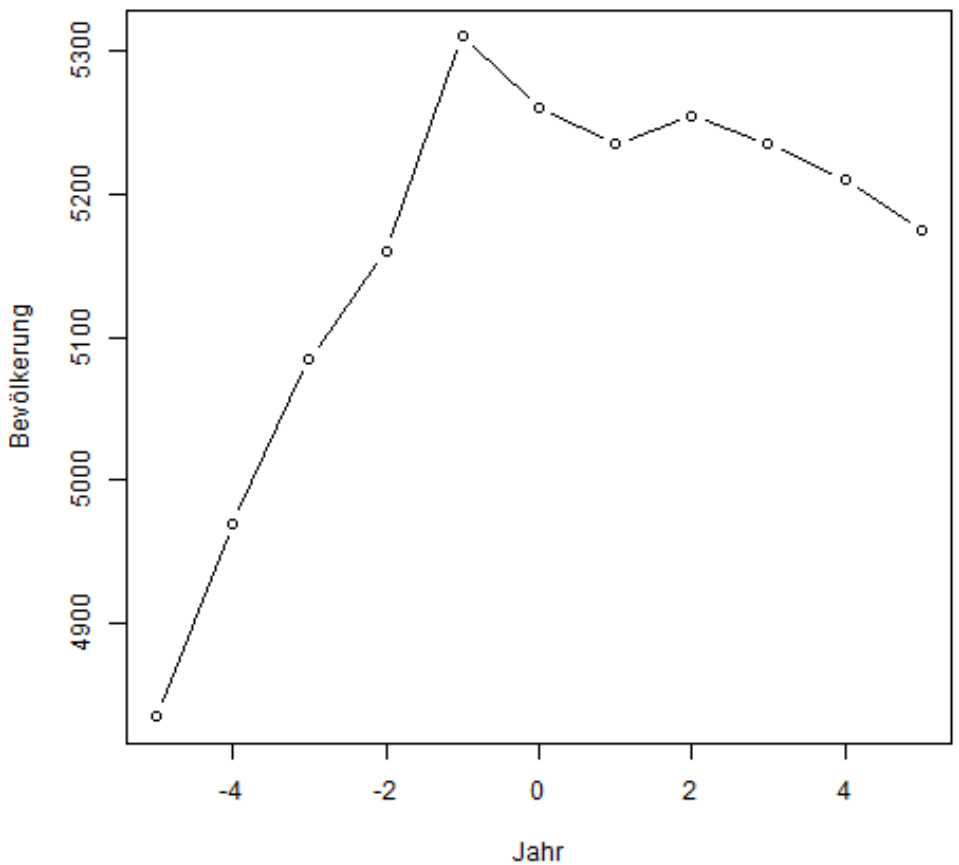
Die Modelle können mittels Anova verglichen werden:

```
anova(fit2, fit3)
```

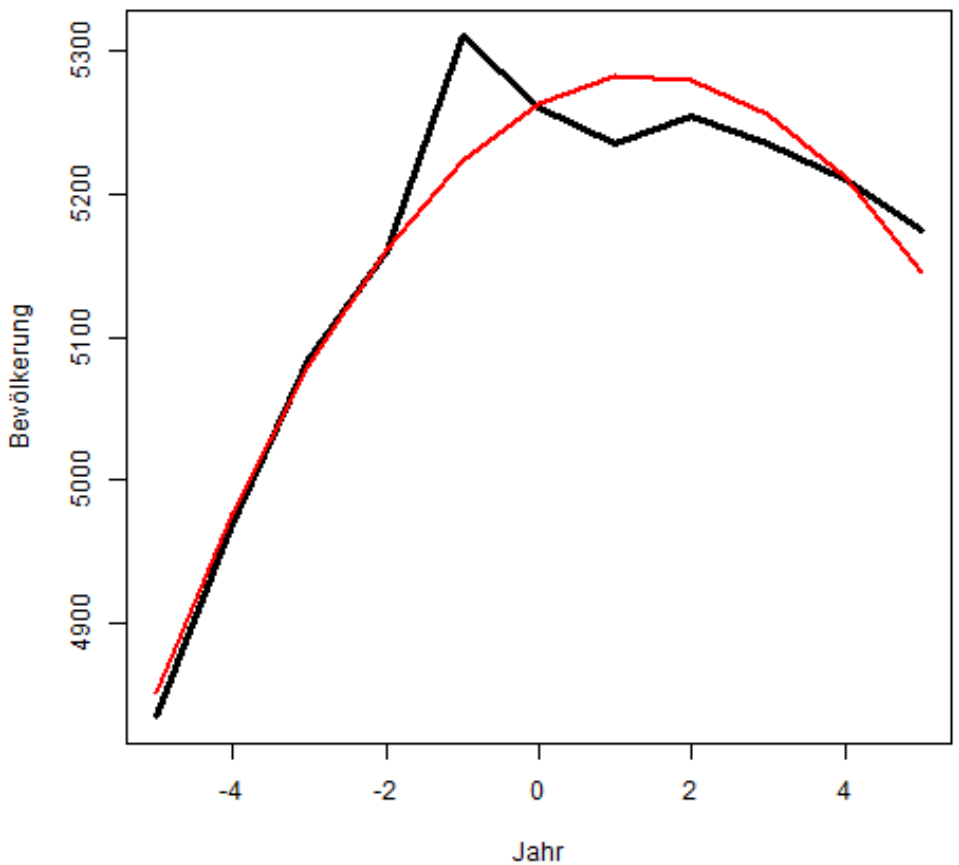
```
Analysis of Variance Table

Model 1: Population ~ Year + I(Year^2)
Model 2: Population ~ Year + I(Year^2) + I(Year^3)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      8 12019.8
2      7  7659.5  1    4360.3 3.9848 0.0861 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-Wert $> 0,05 \rightarrow$ Nullhypothese kann nicht abgelehnt werden, also keine Verbesserung

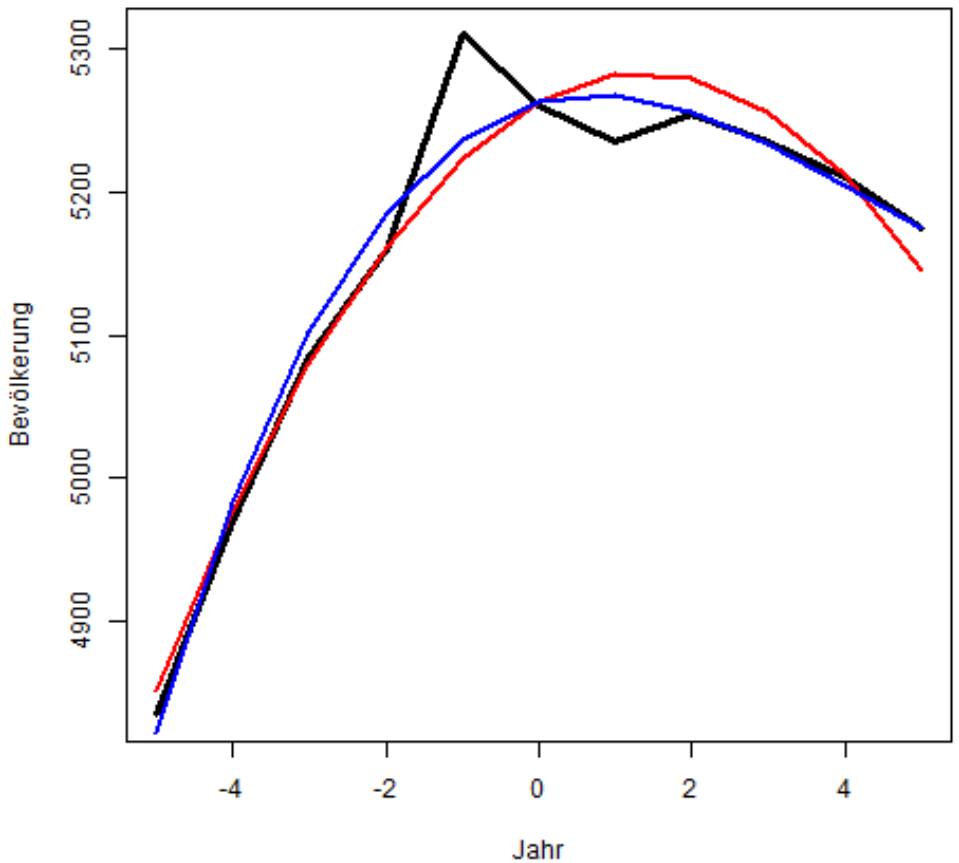
```
plot(sample1$Year, sample1$Population, type="l", lwd=3)
points(sample1$Year, predict(fit2), type="l", col="red", lwd
```



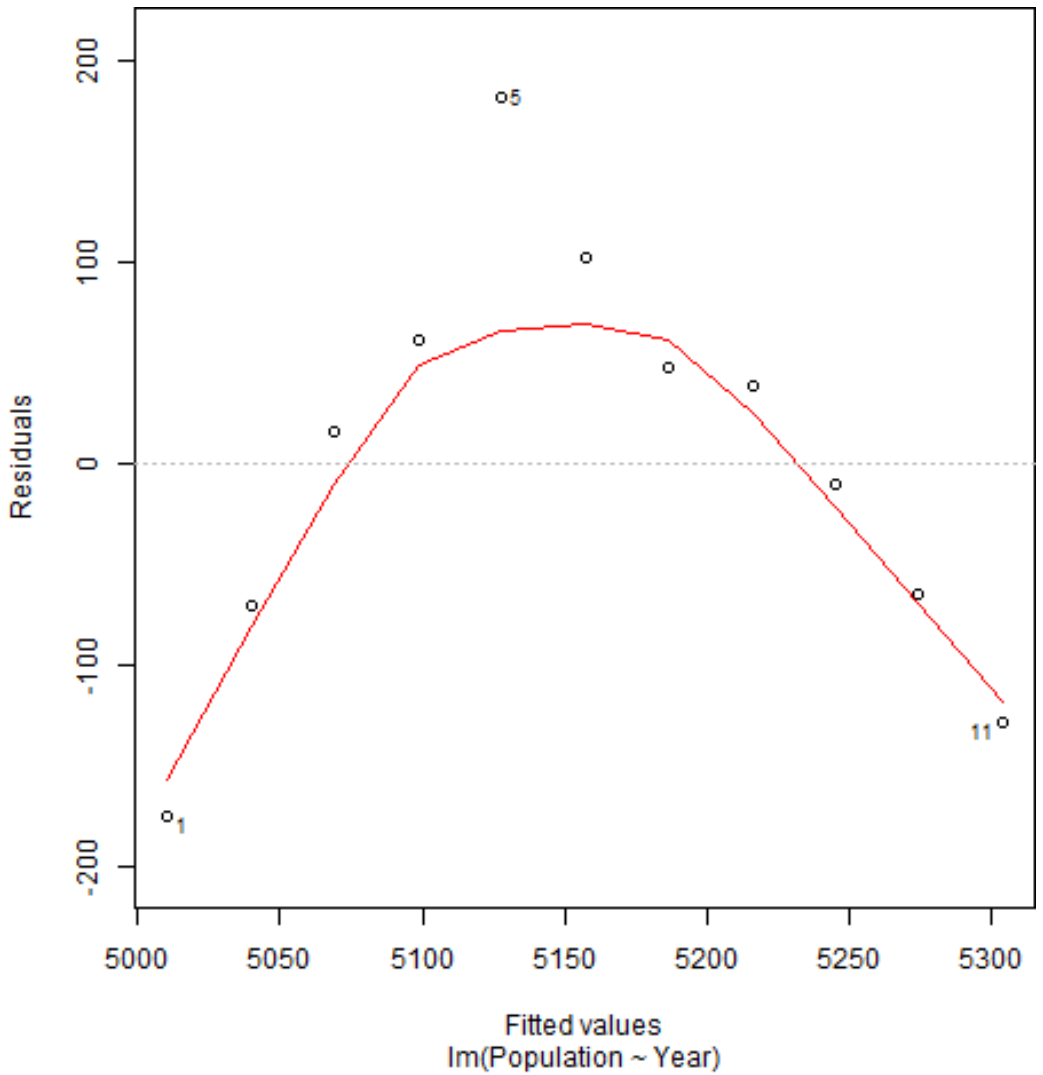
```
plot(sample1$Year, sample1$Population, type="l", lwd=3)
points(sample1$Year, predict(fit2), type="l", col="red", lwd
```



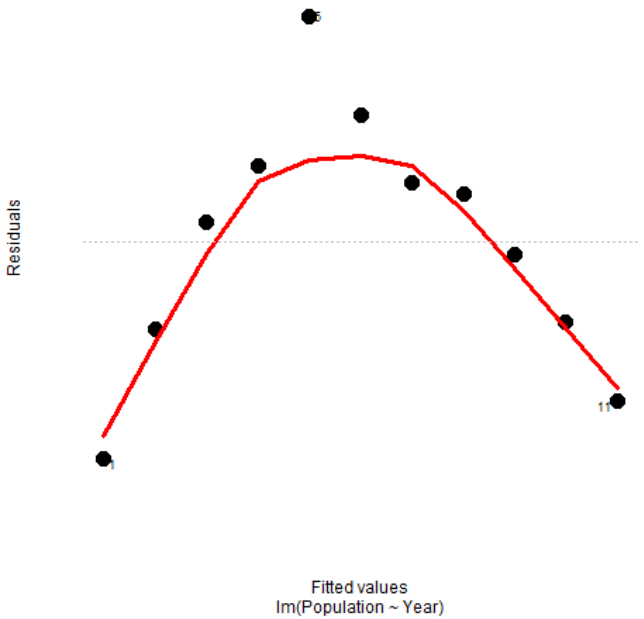
```
plot(sample1$Year, sample1$Population, type="l", lwd=3)
points(sample1$Year, predict(fit2), type="l", col="red", lwd
```



Plot - Residuen gegen gefittete Werte



Plot - Residuen gegen gefittete Werte



Bei kurvenförmigem Verlauf:

→ BLUE-Annahme einer linearen Beziehung zwischen Variablen verletzt

→ Transformation einer der Variablen notwendig

Anmerkungen zur Polinomialregression

- Orthogonale Polynome sollten verwendet werden
  - numerisch stabil
  - korrekter Grad lässt sich einfacher wählen
- Mit höherem Grad wird der polynomiale Fit immer unattraktiver

Bemerkungen zu Regressionsmodellen

Probleme, die im Rahmen der Regressionsrechnung beachtet werden sollten:

- Lineare Restriktionen;
- Annahmeverletzungen:
  - Heteroskedastizität;
  - Serielle Korrelation;
  - Mangelnde Unabhängigkeit zwischen exogener Variablen und Störterm;
- Nichtlineare Regressionsverfahren;
- Glättungsverfahren;

Überblick Modelle

Parametrische Modelle

$f(x|\beta) = \beta_0 + \beta_1 x$ — Modell mit linearem Prädiktor

$f(x|\beta) = \beta_0 + \beta_1 log(x)$ — Transformationen

Nichtparametrische Modelle

Überblick Modelle

Parametrische Modelle

$f(x|\beta) = \beta_0 + \beta_1 x$ — Modell mit linearem Prädiktor

$f(x|\beta) = \beta_0 + \beta_1 log(x)$ — Transformationen

$f(x|\beta) = \beta_0 + \beta_1 x + \beta_1 x^2$ — Modell mit polynomialen Term

Nichtparametrische Modelle

Überblick Modelle

Parametrische Modelle

$f(x|\beta) = \beta_0 + \beta_1 x$ — Modell mit linearem Prädiktor

$f(x|\beta) = \beta_0 + \beta_1 log(x)$ — Transformationen

$f(x|\beta) = \beta_0 + \beta_1 x + \beta_1 x^2$ — Modell mit polynomialen Term

Nichtparametrische Modelle

Überblick Modelle

Parametrische Modelle

$f(x|\beta) = \beta_0 + \beta_1 x$ — Modell mit linearem Prädiktor

$f(x|\beta) = \beta_0 + \beta_1 log(x)$ — Transformationen

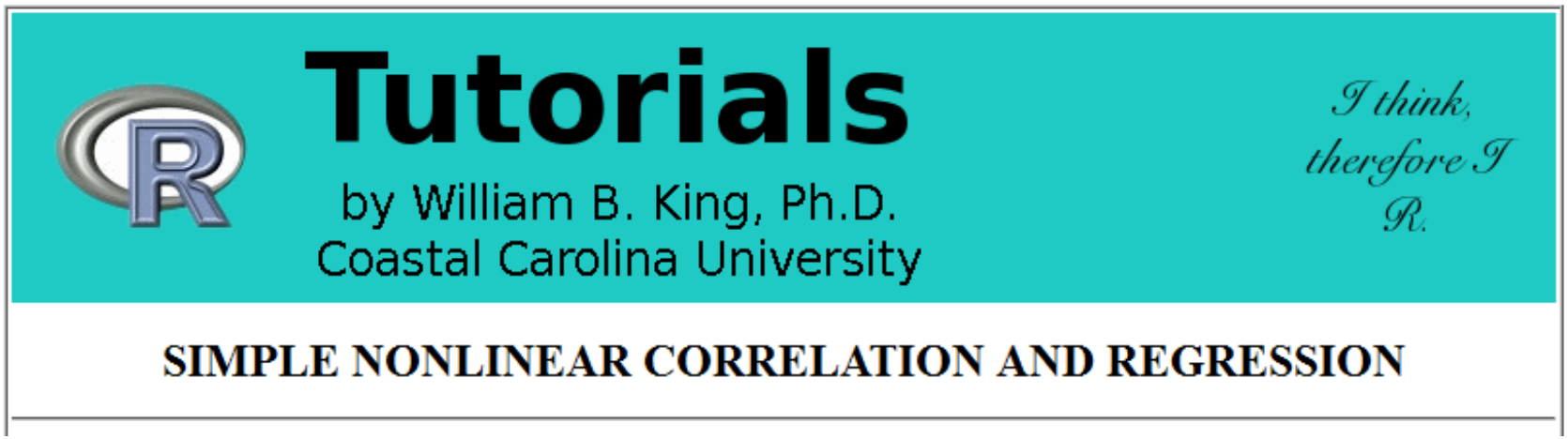
$f(x|\beta) = \beta_0 + \beta_1 x + \beta_1 x^2$ — Modell mit polynomialen Term

$f(x|\beta) = \beta_0 + \beta_1 x^{\beta 2}$ — Nichtlineraes Modell

Nichtparametrische Modelle

Splines → nächster Abschnitt

Basisliteratur



`http://ww2.coastal.edu/kingw/statistics/R-tutorials/simplenonlinear.html`