Lineare Regression mit Faktoren

Wdh: Einfache Lineare Regression

- $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)\ i.i.d.$
  Linear in Koeffizienten
- Schätzer $\widehat{\beta_j}$ für β_j minimieren Residuenquadratsumme (RSS):

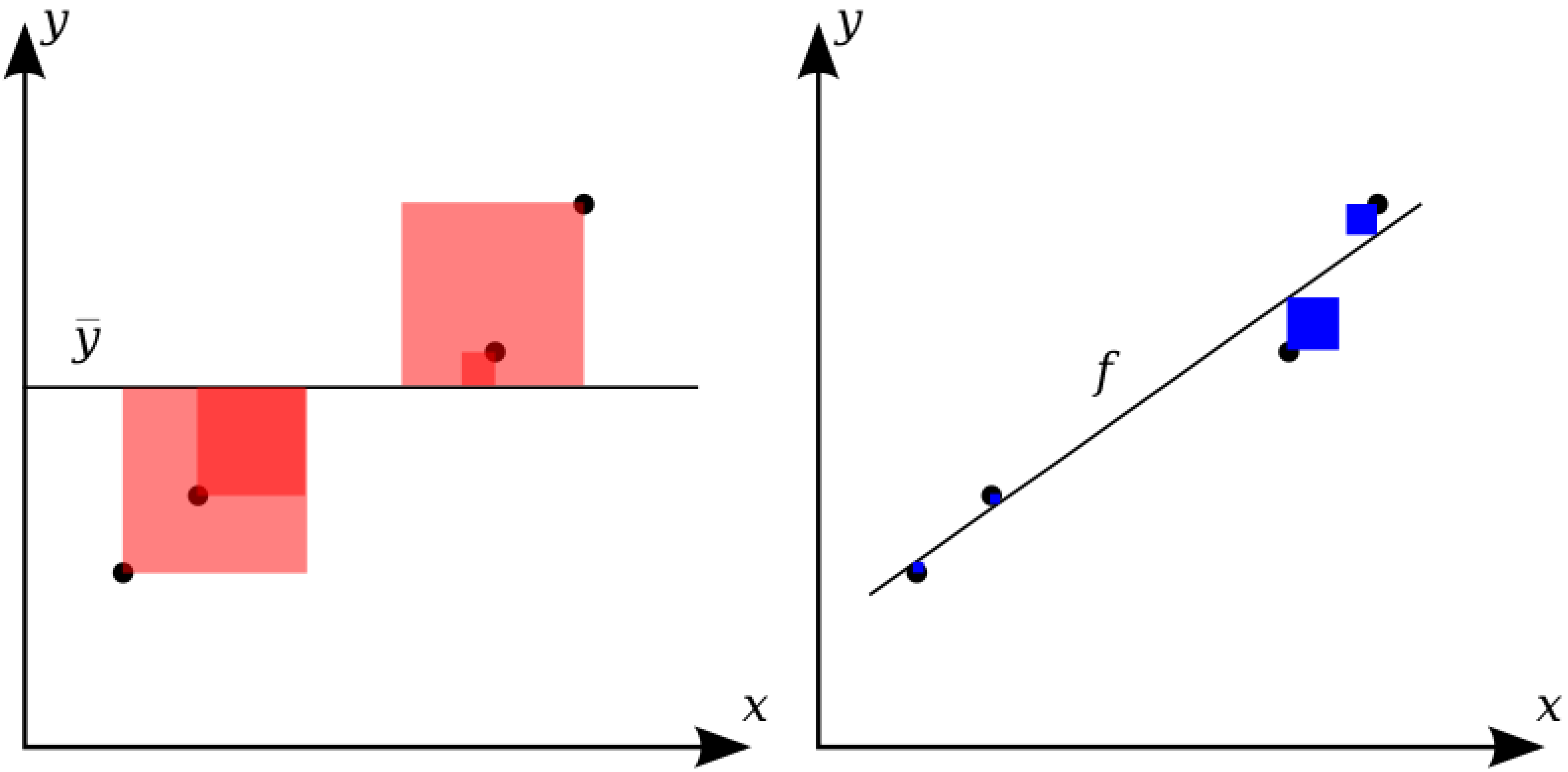
$$\widehat{\beta_0}, \widehat{\beta_1}\ \ minimieren\ \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

- Unter obigen Annahmen und $H_0: \beta_j = 0$:

$$\mathrm{t} = \frac{\widehat{\beta_j} - 0}{SE\left(\widehat{\beta_j}\right)} \sim t_{n-2}$$

$\rightarrow$ t-Test in der Linearen Regression

Intuition: Einfache Lineare Regression



Verkaufszahlen

Beispiel in R: Marketing Daten

```
Call:
lm(formula = Sales ~ Newspaper, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.2272  -3.3873  -0.8392   3.5059  12.7751 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
Newspaper    0.05469    0.01658    3.30  0.00115 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,	Adjusted R-squared:  0.04733 
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

$R^2 = 0.052 \rightarrow$ Modell erklärt nur kleinen Anteil der Streuung in den Daten

$$Sales_i = 12.35 + 0.055 * Newspaper_i + \varepsilon_i \, ; \varepsilon_i \sim N(0, 5.09^2)$$

95%-VI: $0.055 \pm 2 * 0.017 = [0.021; 0.089]$
Effekt von 'Newspaper' ist signifikant
(p-Wert = 0.001)

Wdh: Multiple Lineare Regression

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)\ i.i.d.$
- Schätzer $\widehat{\beta}_j$ für β_j minimieren Residuenquadratsumme (RSS):

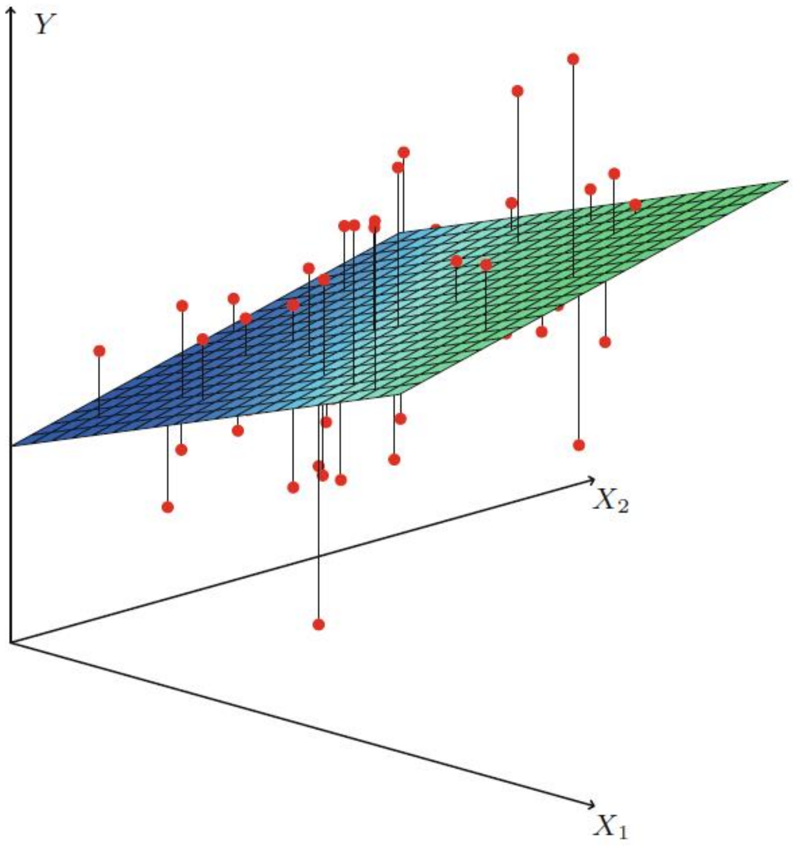
$$\widehat{\beta}_j\ minimieren \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1}\right)\right)^2$$

- Unter obigen Annahmen und $H_0: \beta_j = 0$:

$$\mathrm{t} = \frac{\widehat{\beta}_j - 0}{SE\left(\widehat{\beta}_j\right)} \sim t_{n-p}$$

$\rightarrow$ t-Test in der Linearen Regression

Intuition: Multiple Lineare Regression



Beispiel in R: Marketing Daten

```
Call:
lm(formula = Sales ~ Newspaper + TV + Radio, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908  0.2418  1.1893  2.8292 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86    
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Gegeben TV und Radio
ist Newspaper
nicht mehr signifikant

Wenn die Radio-Ausgaben um
eine Einheit erhöht werden
und die TV- und Newspaper-Ausgaben
konstant bleiben,
erhöhen sich die Sales um
0.189 (95%-VI: [0.171; 0.206]).

$R^2 = 0.897 \rightarrow$ relativ viel Streuung
wird durch Modell erklärt

Faktoren als erklärende Variable

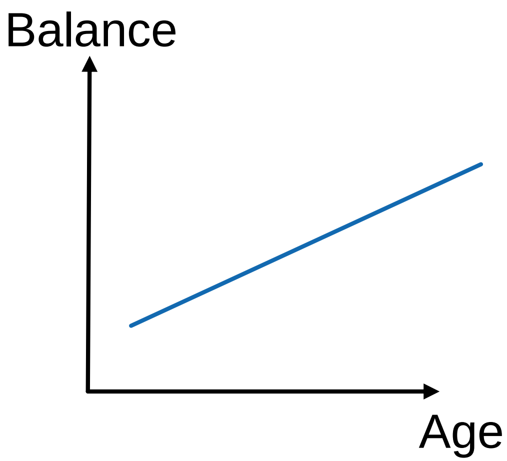
- Faktor = Diskrete erklärende Variable
  Bsp 1: Geschlecht
  Bsp 2: Haarfarbe
- Level = Werte, die ein Faktor annehmen kann
  Bsp 1: Der Faktor ‘Geschlecht’ hat 2 Levels: ‘Mann’ und ‘Frau’
  Bsp 2: Der Faktor ‘Haarfarbe’ hat 4 Levels: ‘Rot’, ‘Blond’, ‘Braun’, ‘Schwarz’

Schulden mit Kreditkarte

Datensatz ‘credit’: Schulden erklären durch Geschlecht und Alter

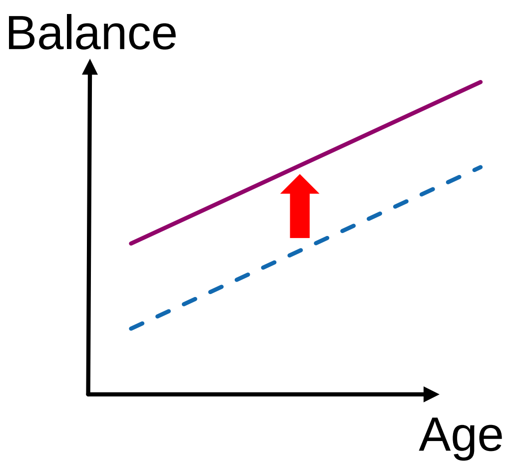
Faktoren: Intuition

"***Referenzlevel***" z.B. "Männer"



Männer:

$$Balance_i = \beta_0 + \beta_1 * Age_i + \varepsilon_i$$



Frauen:

$$Balance_i = \underbrace{\beta_0 + \beta_2}_{\text{Neuer } \textbf{Achsenabschnitt} \text{ für Frauen}} + \beta_1 * Age_i + \varepsilon_i$$

Faktoren: Technik

- Dummy Variable
- Zwei levels: Eine binäre Dummy Variable
  $x_i = 0$, falls Person i männlich ist
  $x_i = 1$, falls Person i weiblich ist
  $$\rightarrow Balance_i = \beta_0 + \beta_2 * x_i + \beta_1 * Age_i + \varepsilon_i$$
- Mehr als zwei levels (CH, D, USA):
  - ein Referenzlevel (CH)
  - eine binäre Dummy Variablen für jedes andere level (D, USA)
- Software regelt das im Detail

Beispiel in R: Faktoren

```
Call:
lm(formula = Balance ~ Age + Gender, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-530.76 -455.90  -61.05  335.05 1487.22

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  507.21157   81.41578   6.230 1.19e-09 ***
Age            0.04661    1.33738   0.035    0.972
GenderFemale  19.72667   46.10947   0.428    0.669
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.8 on 397 degrees of freedom
Multiple R-squared:  0.0004642, Adjusted R-squared:  -0.004571
F-statistic: 0.09218 on 2 and 397 DF,  p-value: 0.912
```

Achsenabschnitt in der Gruppe mit dem Referenzlevel (Männer)

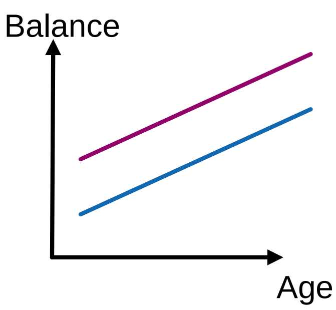
Steigung ist in beiden Gruppen gleich

Veränderung des Achsenabschnitts, wenn man von der Referenzgruppe (Männer) in die andere Gruppe (Frauen) wechselt. Achsenabschnitt für Frauen ist also:
$$507.2 + 19.7 = 526.9$$

$$\rightarrow Balance_i = 507.2 + 19.7 * Gender_i + 0.047 * Age_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, 460.8^2)$$

Männer: $Balance_i = 507.2 + 0.047 * Age_i + \varepsilon_i,\ \varepsilon_i \sim N(0, 460.8^2)$

Frauen: $Balance_i = 526.9 + 0.047 * Age_i + \varepsilon_i,\ \varepsilon_i \sim N(0, 460.8^2)$



Beispiel in R: Schlussfolgerung

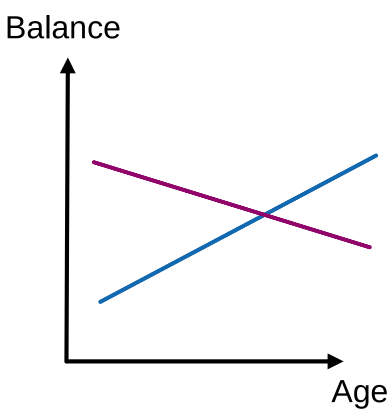
```
Call:
lm(formula = Balance ~ Age + Gender, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-530.76 -455.90  -61.05  335.05 1487.22

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  507.21157   81.41578   6.230 1.19e-09 ***
Age            0.04661    1.33738   0.035    0.972
GenderFemale  19.72667   46.10947   0.428    0.669
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.8 on 397 degrees of freedom
Multiple R-squared:  0.0004642, Adjusted R-squared:  -0.004571
F-statistic: 0.09218 on 2 and 397 DF,  p-value: 0.912
```

Es gibt keinen Hinweis darauf,
dass Alter oder Geschlecht
einen Einfluss auf die Schulden haben

Wechselwirkung (WW; Interaktion)

- WW ist zwischen zwei (oder mehr) Variablen
  Bsp: WW zwischen Age und Gender
- WW zwischen Age und Gender:
  Age hat je nach Gender einen unterschiedlichen Einfluss auf die Zielgrösse (Balance)
- Falls WW vorhanden: Steigungen in verschiedenen Gruppen sind unterschiedlich
- Praxis: Prüfen, ob WW vorhanden ist

Wechselwirkung: Intuition



Ist β_3 sign. verschieden von 0?

Modell ohne Interaktion:
$$Balance_i = (\beta_0 + \beta_2 * x_i) + \beta_1 * Age_i$$

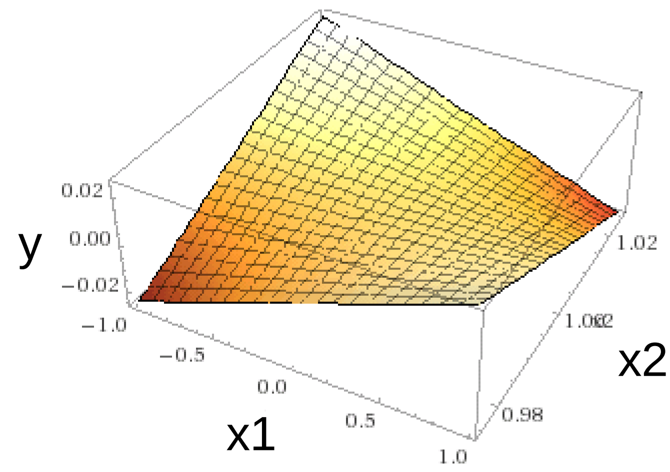
Modell mit Interaktion:
$$Balance_i = (\beta_0 + \beta_2 * x_i) + (\beta_1 + \beta_3 * x_i) * Age_i$$

Geraden parallel

Geraden nicht parallel

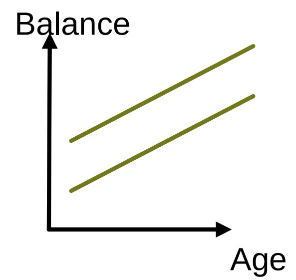
Wechselwirkung

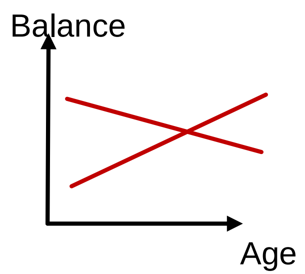
- Effekt von einer Variable hängt von dem Wert einer anderen Variable ab
- Meistens: Wechselwirkung zwischen Faktor und kontinuierlicher Variable
- WW zw. zwei kontinuierlichen Variablen auch möglich



Wechselwirkung: Notation & Konvention

- Notation in R:
  $$\text{Balance} \sim \underbrace{\text{Age} + \text{Gender}}_{\text{“Haupteffekte”}} + \underbrace{\text{Age:Gender}}_{\text{“Wechselwirkung”}} = \text{Age} * \text{Gender}$$





- Konvention: Falls eine Wechselwirkung im Modell ist, müssen auch die beteiligten Haupteffekte im Modell sein

Beispiel in R: Wechselwirkung

```
Call:
lm(formula = Balance ~ Age * Gender, data = dat2)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     478.6139   113.8643   4.203 3.25e-05 ***
Age               0.5610     1.9590   0.286    0.775
GenderFemale     73.4491   156.3361   0.470    0.639
Age:GenderFemale -0.9652     2.6835  -0.360    0.719

Residual standard error: 461.3 on 396 degrees of freedom
Multiple R-squared:  0.0007906, Adjusted R-squared:  -0.006779
F-statistic: 0.1044 on 3 and 396 DF,  p-value: 0.9575
```

Achsenabschnitt: Männer

Steigung: Männer

Änderung Achsenabschnitt: Frauen

Änderung Steigung: Frauen

$$Balance_i = (478.6 + 73.4 * Gender_i) + (0.56 - 0.97 * Gender_i) * Age_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, 461.3^2)$$

Männer: $Balance_i = 478.6 + 0.56 * Age_i + \varepsilon_i,\ \varepsilon_i \sim N(0, 461.3^2)$

Frauen: $Balance_i = 552.0 - 0.41 * Age_i + \varepsilon_i,\ \varepsilon_i \sim N(0, 461.3^2)$

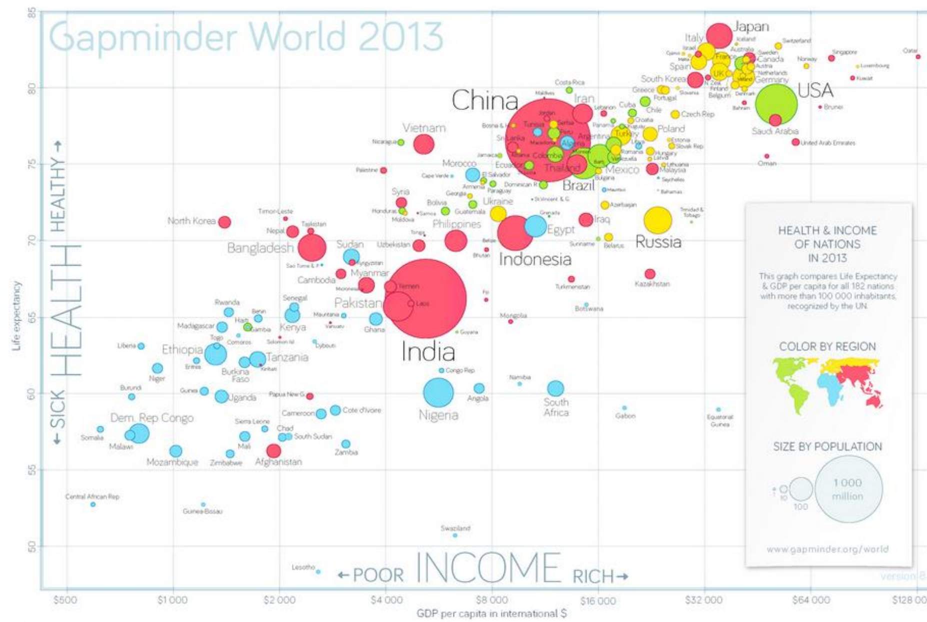
Beispiel in R: Schlussfolgerung

```
Call:
lm(formula = Balance ~ Age * Gender, data = dat2)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      478.6139   113.8643   4.203 3.25e-05 ***
Age                0.5610     1.9590   0.286    0.775
GenderFemale      73.4491   156.3361   0.470    0.639
Age:GenderFemale  -0.9652     2.6835  -0.360    0.719

Residual standard error: 461.3 on 396 degrees of freedom
Multiple R-squared:  0.0007906, Adjusted R-squared:  -0.006779
F-statistic: 0.1044 on 3 and 396 DF,  p-value: 0.9575
```

Wechselwirkung ist nicht signifikant verschieden von 0.
Der Einfachheit halber bevorzugen wir dann ein Modell ohne WW (parallele Geraden).

Es gibt keinen Hinweis darauf, dass der
Effekt von Alter auf die Schulden vom Geschlecht abhängt

Reales Beispiel: Globale Lebenserwartung

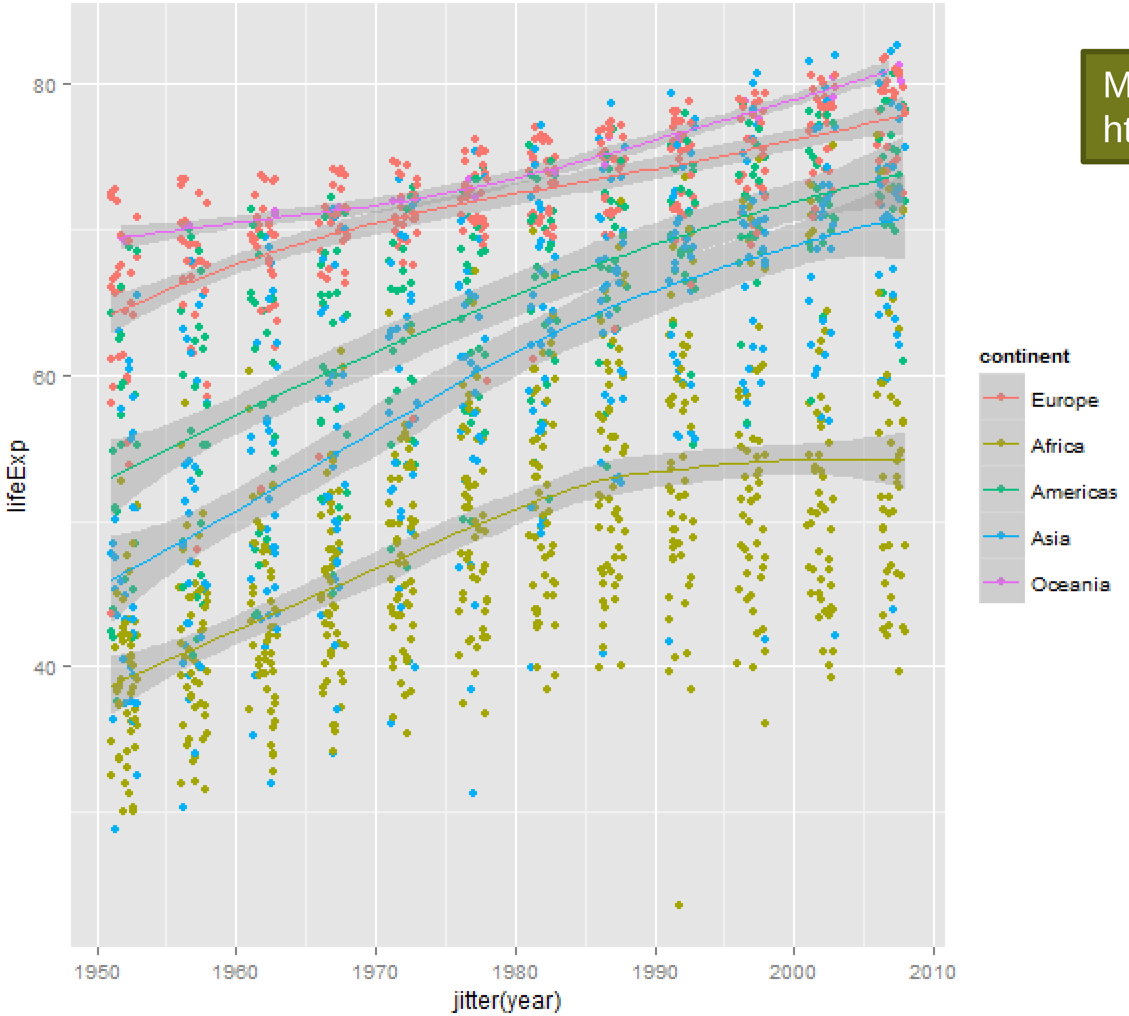


R Paket: ‘gapminder’

- Paket ‘gapminder’ mit Datensatz ‘gapminder’
- Lebenserwartung:174 Ländern gemessen alle 5 Jahre über die letzten ca. 60 Jahre

- Gibt es einen globalen Trend ?
- Ist dieser Trend von Kontinent zu Kontinent unterschiedlich ?

Globale Lebenserwartung mit Paket ggplot2



Mehr zu ggplot2:
http://ggplot2.org/book/

Mögliche Prüfungsfragen

- Lade Daten aus csv-File; verschaffe Überblick (kont. Zielgrösse, eine kont. erklärende Var., ein Faktor oder eine zweite kont. erklärende Var.)
- Fitte Lineare Regression;
  - Ist WW nötig?
  - Interpretation der Parameter?
- Verständnisfragen:
  Z.B.: Empfehlen sie für die Daten im Plot ein Modell mit oder ohne WW?