Wissenschaftliche Methodik I

Varianzanalyse II

Beispieldatensätze zu dieser VL (enthalten im Workspace "ANOVA_2.Rdata") :

- ***mice.csv***
- ***circulation.csv***
- ***chol.csv***
- ***Telekom.csv***
- ***glassrod.csv***

Einfaktorielle ANOVA:

```
> summary(aov(Punkte~Jahrgang, data=points))
             Df  Sum Sq Mean Sq F value Pr(>F)
Jahrgang      4    447  111.64   2.161  0.074 .
Residuals   243  12554   51.66
---
Signif. codes:  0 '***': p < 0.001;**:          *:p < 0.05
```

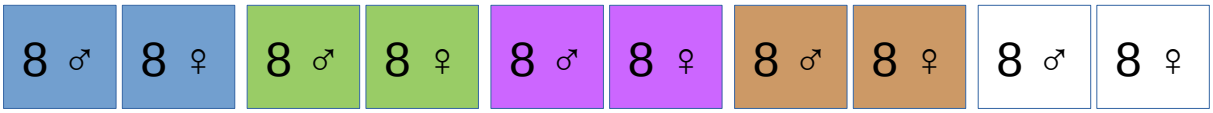
Abhängige Variable

Faktor

- **Einfluss einer unabhängigen Variablen (Faktor) mit k verschiedenen Ausprägungen (level) auf eine abhängige Variable**
  - **Nullhypothese:** $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$
  - **Alternativhypothese:** $H_1: \mu_{kx} \neq \mu_{ky}$

Zweifaktorielle ANOVA:

- **Beispiel:**
  - Vier Zytostatika sollen an 80 Mäusen getestet werden.
  - Es ist unklar, ob die Wirkung bei weiblichen und männlichen Tieren gleich ist.



  - Zwei Faktoren mit fünf bzw. zwei Faktorstufen

```
> model = aov(wellness ~ drug + gender, data=mice)
```

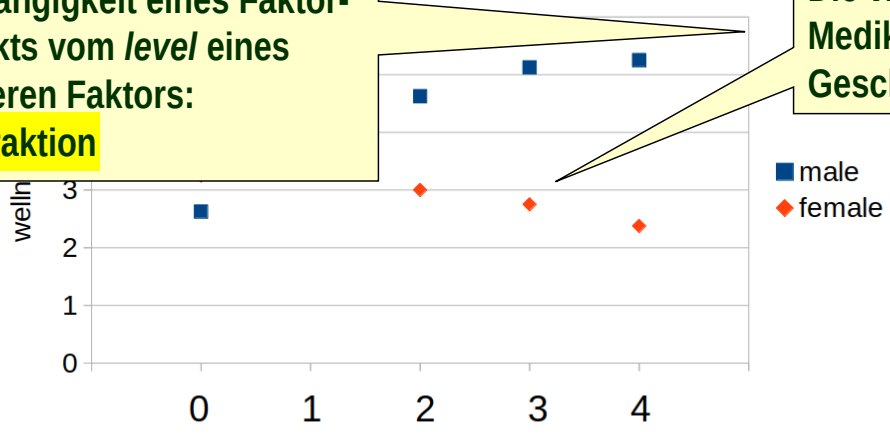
$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$ mit i=1,..,5; j = 1,2, k=1,...,8

Zweifaktorielle ANOVA:

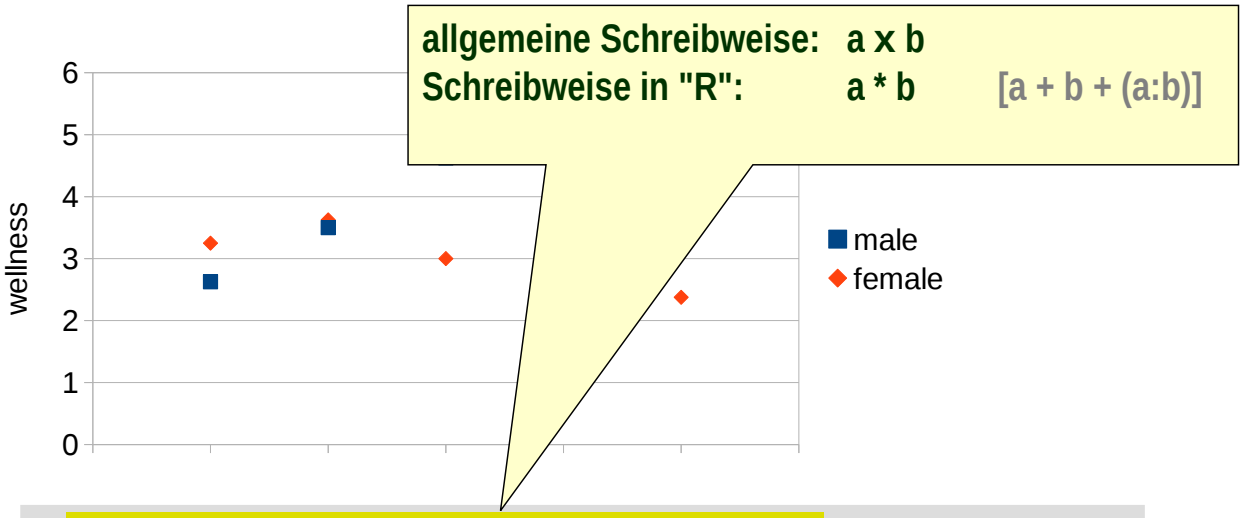
```
> summary(aov(wellness~drug+gender, data=mice))
            Df Sum Sq Mean Sq F value   Pr(>F)
drug         4  10.30   2.575   2.056   0.0952 .
gender       1  30.01  30.013  23.965 5.59e-06 ***
Residuals   74  92.67   1.252
```

Abhängigkeit eines Faktor-Effekts vom *level* eines anderen Faktors: Interaktion

Die Wirkung des Medikaments hängt vom Geschlecht ab

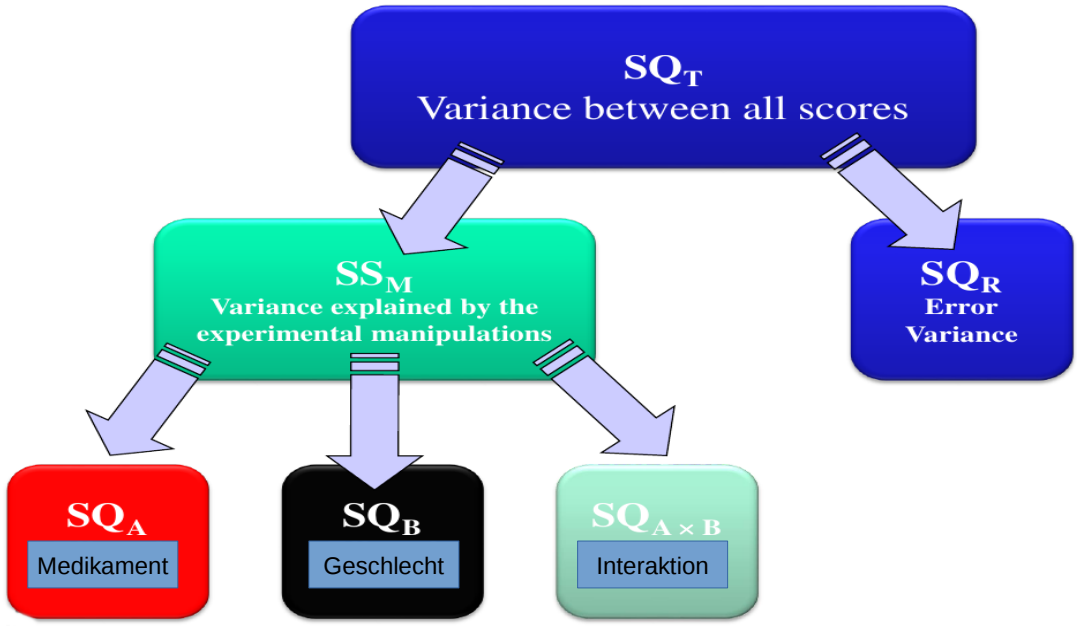


Interaktion:




```
> summary(aov(wellness~drug*gender, data=mice))
             Df Sum Sq Mean Sq F value   Pr(>F)
drug          4  10.30   2.575   3.285   0.0159 *
gender        1  30.01  30.013  38.285 3.65e-08 ***
drug:gender   4  37.80   9.450  12.055 1.65e-07 ***
Residuals    70  54.87   0.784
```

"Varianz-Zerlegung":



Interaktion:

```
> mice.aov = aov(wellness~drug*gender, data=mice)
> summary(mice.aov)
             Df Sum Sq Mean Sq F value   Pr(>F)
drug          4  10.30   2.575   3.285   0.0159 *
gender        1  30.01  30.013  38.285 3.65e-08 ***
drug:gender   4  37.80   9.450  12.055 1.65e-07 ***
Residuals    70  54.87   0.784
```

Wer unterscheidet sich von wem?
Post-Hoc Test!

```
TukeyHSD(mice.aov, "drug:gender", ordered=T)

$`drug:gender`
                                  diff        lwr       upr     p adj
control:male-Drug4:female     0.250 -1.19665684 1.[illegible]657 0.9999034
Drug3:female-Drug4:female     0.375 -1.07165684 1.82[illegible] 0.9974718
Drug2:female-Drug4:female     0.625 -0.82165684 2.07165[illegible]194725
control:female-Drug4:female 0.875 -0.57165684 2.321657 0.[illegible]371
Drug1:male-Drug4:female       1.125 -0.32165684 2.571657 0.28[illegible]
Drug1:female-Drug4:female     1.250 -0.19665684 2.696657 0.14879[illegible]
Drug2:male-Drug4:female       2.250  0.80334316 3.696657 0.0001237
Drug3:male-Drug4:female       2.750  1.30334316 4.196657 0.0000014
Drug4:male-Drug4:female       2.875  1.42834316 4.321657 0.0000005
Drug3:female-control:male     0.125 -1.32165684 1.571657 0.9999998
Drug2:female-control:male     0.375 -1.07165684 1.821657 0.9974718
control:female-control:male 0.625 -0.82165684 2.071657 0.9194725
Drug1:male-control:male       0.875 -0.57165684 2.321657 0.6178371
Drug1:female-control:male     1.000 -0.44665684 2.446657 0.4282084
Drug2:male-control:male       2.000  0.55334316 3.446657 0.0009769
Drug3:male-control:male       2.500  1.05334316 3.946657 0.0000139
Drug4:male-control:male       2.625  1.17834316 4.071657 0.0000045
Drug2:female-Drug3:female     0.250 -1.19665684 1.696657 0.9999034
control:female-Drug3:female 0.500 -0.94665684 1.946657 0.9798644
Drug1:male-Drug3:female       0.750 -0.69665684 2.196657 0.7949252
Drug1:female-Drug3:female     0.875 -0.57165684 2.321657 0.6178371
Drug2:male-Drug3:female       1.875  0.42834316 3.321657 0.0025951
Drug3:male-Drug3:female       2.375  0.92834316 3.821657 0.0000420
Drug4:male-Drug3:female       2.500  1.05334316 3.946657 0.0000139
control:female-Drug2:female 0.250 -1.19665684 1.696657 0.9999034
```

TukeyHSD(mice.aov, "drug:gender", ordered=T)

... der größere Wert steht vorne

... es genügt die Interaktion!

... das aov-Objekt

Interaktion:

```
> library("agricolae")
> model = with(mice, interaction(drug, gender))
> mice.aov = aov(wellness~model, data=mice)
> HSD.test(mice.aov, "model", group=T, console=T)

$groups
                trt means  M
1  Drug4.male      5.250  a
2  Drug3.male      5.125  a
3  Drug2.male      4.625 ab
4  Drug1.female    3.625 bc
5  Drug1.male      3.500 bc
6  control.female 3.250 bc
7  Drug2.female    3.000  c
8  Drug3.female    2.750  c
9  control.male    2.625  c
10 Drug4.female    2.375  c
```

Reduktion: "einfaktoriell" (Interaktion)

HSD.test: TukeyHSD

"gleiche" Gruppen erhalten gleiche Buchstaben

Varianzanalyse: Post Hoc Tests

- H_1: $\mu_x \neq \mu_{total}$ » **mindestens zwei μ verschieden!**
- **Anzahl möglicher Vergleiche: *C* = k(k - 1) / 2**
- **Post Hoc Tests berücksichtigen "*error inflation*"**
- **Tests mit Annahme von Varianz-Homogenität:**
  - **Least Squared Difference (LSD)** — *Keine Korrektur α Fehler*
  - **Duncan**
  - **Ryan-Einot-Gabriel-Welsh**
  - **Tukey HSD**
  - **Scheffe** — *Zu konservativ!*
  - **Bonferroni**

Power

Diverse Post-Hoc Tests:

Der Duncan-Test ist unter Statistikern umstritten!

```
> HSD.test(mice.aov,"model",group=T)

$groups
              trt means  M
1  Drug4.male     5.250  a
2  Drug3.male     5.125  a
3  Drug2.male     4.625 ab
4  Drug1.female   3.625 bc
5  Drug1.male     3.500 bc
6  control.female 3.250 bc
7  Drug2.female   3.000  c
8  Drug3.female   2.750  c
9  control.male   2.625  c
10 Drug4.female   2.375  c
```

```
> duncan.test(mice.aov,"model",group=T,console=T)

                wellness groups
Drug4.male         5.250      a
Drug3.male         5.125      a
Drug2.male         4.625      a
Drug1.female       3.625      b
Drug1.male         3.500     bc
control.female     3.250    bcd
Drug2.female       3.000    bcd
Drug3.female       2.750    bcd
control.male       2.625     cd
Drug4.female       2.375      d
```

Varianzanalyse: Post Hoc Tests

Tests ohne Annahme von Varianz-Homogenität:

- **Dunnett's C**
- **Tamhane T2 (sehr konservativ)**
- **Dunnett T3**
- **Games Howell (nicht immer genau)**

Welches *level* soll mit allen übrigen verglichen werden?

```
> library(DescTools)

> DunnettTest(Punkte~as.factor(Jahrgang), data=points, control="J2015")

  Dunnett's test for comparing several treatments with a control :
    95% family-wise confidence level

$J2015
                 diff    lwr.ci   upr.ci   pval
J2013-J2015  1.551524 -2.182902 5.285950 0.7021
J2014-J2015  2.279938 -1.228100 5.787977 0.3185
J2016-J2015 -1.015311 -4.339225 2.308603 0.8710
J2017-J2015 -1.048811 -4.615793 2.518172 0.8851
```

PostHoc Test für Kruskal-Wallis:

```
> library(FSA)

> dunnTest(Punkte~Jahrgang, data=points)
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Holm method.

      Comparison          Z    P.unadj     P.adj
1  J2013 - J2014 -0.3522129 0.72467859 1.0000000
2  J2013 - J2015  0.9845540 0.32484320 1.0000000
3  J2014 - J2015  1.4294662 0.15287028 0.9172217
4  J2013 - J2016  1.6848544 0.09201666 0.6441166
5  J2014 - J2016  2.1946482 0.02818885 0.2818885
6  J2015 - J2016  0.7334179 0.46330360 1.0000000
7  J2013 - J2017  1.7045975 0.08826952 0.7061561
8  J2014 - J2017  2.1754764 0.02959443 0.2663499
9  J2015 - J2017  0.8105063 0.41764927 1.0000000
10 J2016 - J2017  0.1311243 0.89567694 0.8956769
Warnmeldung:
Jahrgang was coerced to a factor.
```

PostHoc Test für Modelle mit random effects:

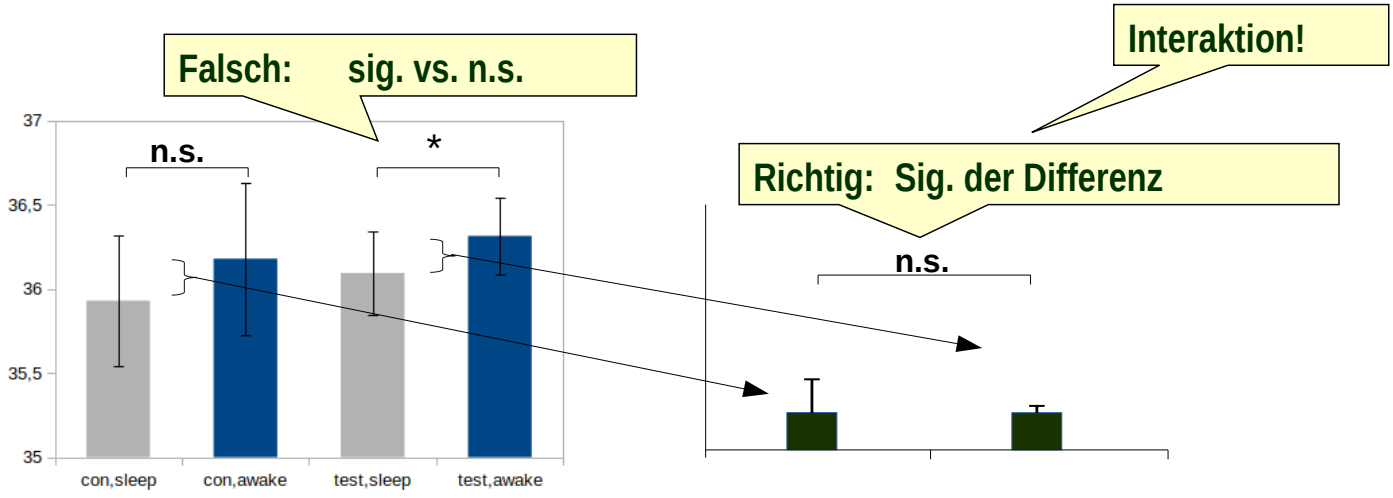
```
> model <- with(CRISPR, interaction(guideRNA, FCS)
> c.aov <- aov(transfection~model + Error(rep/FCS), data=CRISPR)
> HSD.test(c.aov, "model", group=T, console=T)

Fehler in as.data.frame.default(x[[i]], optional = TRUE, stri
stringsAsFactors) :
  kann Klasse ‘c("aovlist", "listof")’ nicht in data.frame umwandeln
```

Estimated Marginal Means

```
> library(emmeans)
> c.aov <- aov(transfection~guideRNA*FCS + Error(rep/FCS), data=CRISPR)
> emm <- emmeans(c.aov, ~ guideRNA | FCS)
Note: re-fitting model with sum-to-zero contrasts
> pairs(emm)
FCS = A:
 contrast estimate   SE df t.ratio p.value
 a - b       2.633 1.56 20   1.686  0.5558
 a - c      -5.567 1.56 20  -3.563  0.0207
 a - d      -6.233 1.56 20  -3.990  0.0081
 ...
FCS = B:
 contrast estimate   SE df t.ratio p.value
 a - b       2.767 1.56 20   1.771  0.5047
 a - c       0.433 1.56 20   0.277  0.9997
 a - d       0.600 1.56 20   0.384  0.9987
 ...
P value adjustment: tukey method for comparing a family of 6 estimates
```

Interaktionsanalyse:

- **Beispiel:** Medikament zur Durchblutungsförderung
  - Messgröße: Hauttemperatur in Kontroll- u. Test-Gruppe jeweils schlafend (grau) und wach (blau)



Interaktionsanalyse:

- **Beispiel:** Medikament zur Durchblutungsförderung

```
> summary(aov(Temp~sleep, data=subset(circulation, group=="test")))
            Df Sum Sq Mean Sq F value  Pr(>F)
sleep        1  0.484  0.4840   8.379 0.00626 **
Residuals   38  2.195  0.0578
```

```
> summary(aov(Temp~sleep, data=subset(circulation, group=="control")))
            Df Sum Sq Mean Sq F value Pr(>F)
sleep        1  0.625  0.6250   3.496 0.0692 .
Residuals   38  6.794  0.1788
```

```
> summary(aov(Temp~sleep*group, data=circulation))
            Df Sum Sq Mean Sq F value Pr(>F)
sleep        1  1.104  1.1045   9.338 0.0031 **
group        1  0.450  0.4500   3.805 0.0548 .
sleep:group  1  0.004  0.0045   0.038 0.8459
Residuals   76  8.989  0.1183
```

Ko-Varianz: ANCOVA

- **Beeinflussung der abhängigen Variablen durch eine weitere metrische Größe (Ko-Variate)**
  - Fragestellung zielt auf Faktoreffekte (sonst: Korrelationsanalyse)
  - Effekt der Kovariate ist selbst nicht relevant: "Störgröße"
  - Ziel ist Reduktion des Fehlerterms
- **Beispiel:**
  - Blut-Cholesterin in EU und USA
  - Ko-Variate: Alter

Ko-Varianz: ANCOVA

```
> summary(aov(cholesterol~age*region, data=chol))
            Df Sum Sq Mean Sq F value   Pr(>F)
age          1  75292   75292  53.327 1.85e-08 ***
region       1  11722   11722   8.302  0.00681 **
age:region   1     11      11   0.008  0.93011
Residuals   34  48004    1412
```

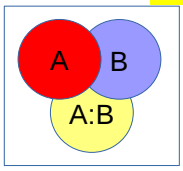
```
> summary(aov(cholesterol~region*age, data=chol))
            Df Sum Sq Mean Sq F value   Pr(>F)
region       1   6818    6818   4.829   0.0349 *
age          1  80196   80196  56.800 9.45e-09 ***
region:age   1     11      11   0.008   0.9301
Residuals   34  4800    1412
```

Problem: unbalanciertes Design;
19 x region; age ist metrisch – jeder Wert 1 mal!

Type I ANOVA ist hierfür nicht geeignet!

Typ I, II, III ANOVA:

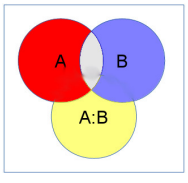
- **Typ I: sequentielle Berechnung der Sum of Squares**



  - SS(A) für Faktor A
  - SS(B | A) für Faktor B
  - SS(A:B | A, B) für Interaktion

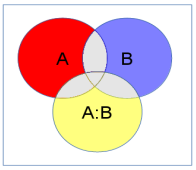
Äquivalent für balancierte Designs

- **Typ II: separate Berechnung für Haupteffekte**

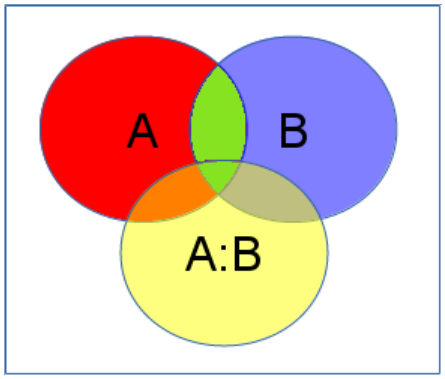


  - SS(A | B) für Faktor A
  - SS(B | A) für Faktor B
  - SS(A:B | A, B) für Interaktion

- **Typ III: separat für Haupteffekte, Korrektur**



  - SS(A | B, A:B) für Faktor A
  - SS(B | A, A:B) für Faktor B
  - SS(A:B | A, B) für Interaktion



Ko-Varianz: ANCOVA

```
> summary(aov(cholesterol~age*region, data=chol))
            Df Sum Sq Mean Sq F value   Pr(>F)
age          1  75292   75292  53.327 1.85e-08 ***
region       1  11722   11722   8.302  0.00681 **
age:region   1     11      11   0.008  0.93011
Residuals   34  48004    1412
```

```
> model = lm(cholesterol~age*region, data=chol)
> library("car")
> Anova(m      l)
Anova Table      II tests)

Response: choles

age
region
age:region     11  1  0.0078  0.930111
Residuals   48004 34
```

lm( ) : linear model

note: Objekte vom Typ "lm" können nicht mit TukeyHSD() analysiert werden – aber mit HSD.test()!

Lineare Regressions-Modelle:

- Mathematisches Konstrukt zur Beschreibung der Beziehung zwischen Variable und Faktor
- Nimmt linearen Zusammenhang an, d.h. Lösung ist mit Geradengleichung möglich
- Erlaubt das Kombinieren von metrischen und nominalen Einflussgrößen

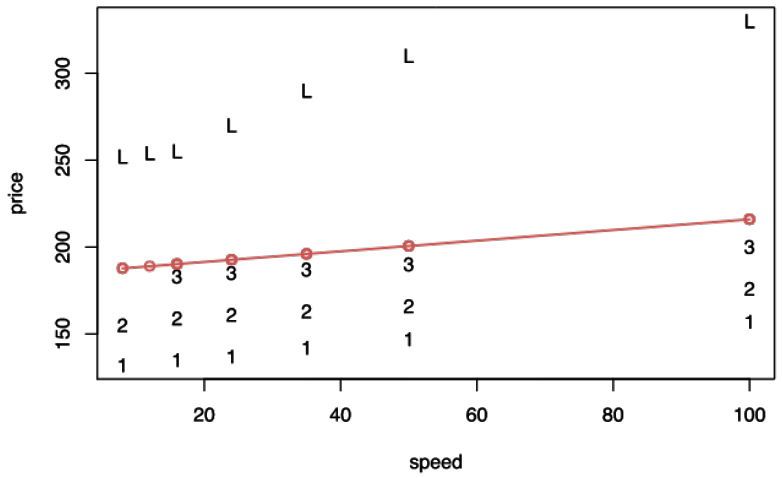
$$Yi = b_0 + b_iX_i + \varepsilon_i$$

- Abh. Wert
- Intercept (Y für x=0)
- Regressions-koeffizient
- Fehler

Lineare Regression:

- **Beispiel: Preise für Internetanbindung der türkischen Telekom***

speed	limit	price
100	Limitsiz	330
100	300GB	200
100	200GB	176
100	100GB	157
50	Limitsiz	310
50	300GB	190
50	200GB	166
50	100GB	147
35	Limitsiz	290



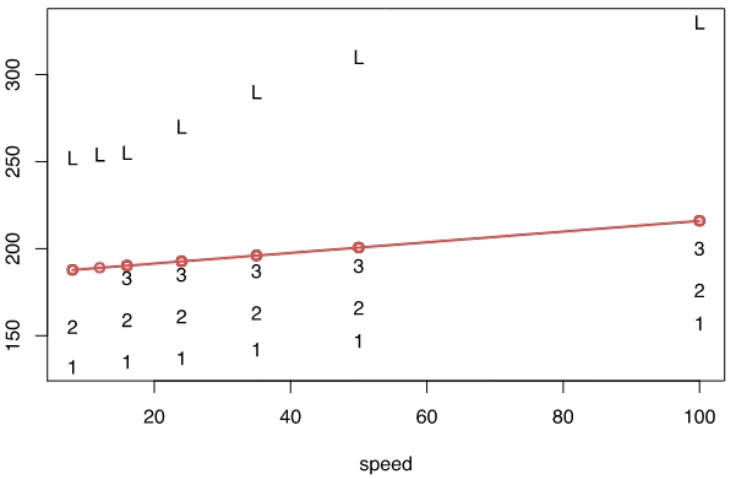
*) Daten und Konzept von Andrés Aravena, https://anaraven.bitbucket.io/blog/2020/msr/linear-models-with-factors.html

Lineare Regression:

- **Beispiel: Preise für Internetanbindung der türkischen Telekom***

```
> summary(lm(price~speed, data=telekom))

Call:
lm(formula = price ~ speed, data = telekom)

Residuals:
   Min     1Q Median     3Q    Max
-59.03 -43.45 -23.66  64.30 113.97

Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept) 185.3823   19.8484   9.340 4.12e-09 ***
speed         0.3064    0.4018   0.763    0.454
---

Residual standard error: 59.68 on 22 degrees of
freedom. Multiple R-squared:  0.02576,
Adjusted R-squared:  -0.01853
F-statistic: 0.5816 on 1 and 22 DF,  p-value: 0.4538
```

...dürftig!



*) Daten und Konzept von Andrés Aravena, https://anaraven.bitbucket.io/blog/2020/msr/linear-models-with-factors.html

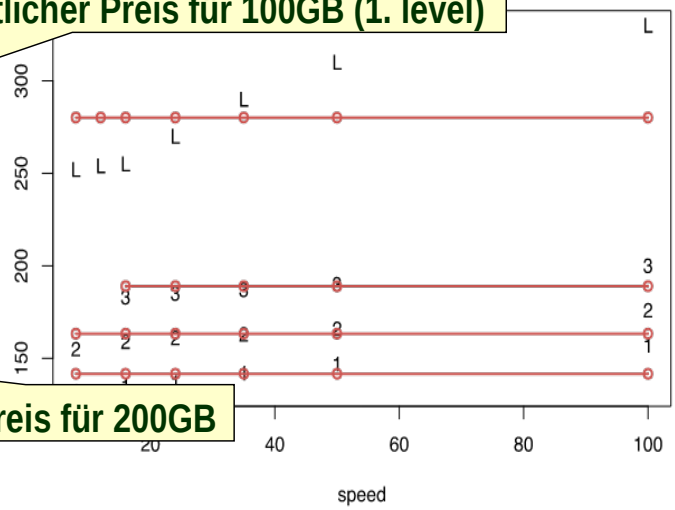
Lineare Regression:

- **Beispiel:** Preise für Internetanbindung der türkischen Telekom*

```
> summary(lm(price~limit, data=telekom))

Call:
lm(formula = price ~ limit, data = telekom)

Residuals:
    Min      1Q  Median      3Q     Max
-28.143  -7.083  -2.167   6.[illegible]  49.857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    141.667      7.387  19.179 2.40e-14 ***
limit200GB      21.667     10.446   2.074 0.051196 .
limit300GB      47.333     10.956   4.[illegible] 0.000333 ***
limitLimitsiz  138.476     10.066  13.756 1.17e-11 ***
---

Residual standard error: 18.09 on 20 degrees of freedom
Multiple R-squared:  0.9186,   Adjusted R-squared:  0.9064
F-statistic: 75.22 on 3 and 20 DF,  p-value: 4.554e-11
```

...durchschnittlicher Preis für 100GB (1. level)

...Aufpreis für 200GB



*) Daten und Konzept von Andrés Aravena, https://anaraven.bitbucket.io/blog/2020/msr/linear-models-with-factors.html

Lineare Regression:

- **Beispiel:** Preise für Internetanbindung der türkischen Telekom*

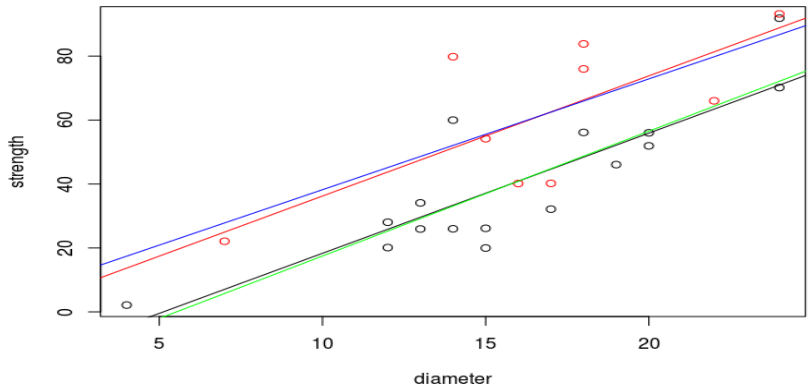
```
> summary(lm(price~limit+speed, data=telekom))

Call:
lm(formula = price ~ limit + speed, data = telekom)

Residuals:
    Min      1Q  Median      3Q     Max
-17.079  -6.763   1.795   4.793  23.491

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   125.18437    5.71069  21.921 5.97e-15 ***
limit200GB     21.66667    6.78770   3.192   0.0048 **
limit300GB     44.71597    7.13594   6.266 5.12e-06 ***
limitLimitsiz 140.10320    6.54792  21.397 9.30e-15 ***
speed           0.42444    0.07969   5.326 3.85e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.76 on 19 degrees of freedom
Multiple R-squared:  0.9673,   Adjusted R-squared:  0.9605
F-statistic: 140.7 on 4 and 19 DF,  p-value: 7.771e-14
```

*) Daten und Konzept von Andrés Aravena, https://anaraven.bitbucket.io/blog/2020/msr/linear-models-with-factors.html

Lineare Regressions-Modelle:

Beispiel: Stabilität von Glas-Rohren abhängig von Herstellungsverfahren und Ø

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.2460     9.9774  -1.929  0.06674 .
diameter      3.7583     0.5885   6.387    2e-06 ***
finishS      17.9057     5.8642   3.053  0.00583 **
```



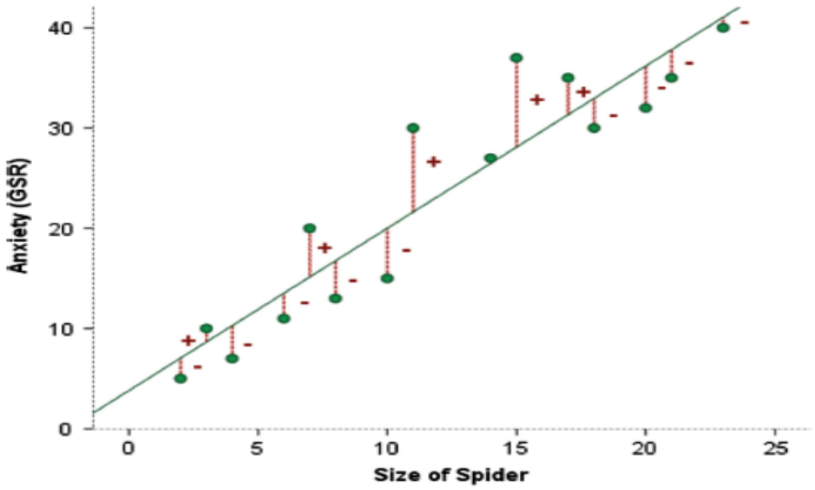
Lineare Regressions-Modelle:

```
> summary(lm(strength~diameter+finish, data=glassrod))
Call:
lm(formula = strength ~ diameter + finish, data = glassrod)
Residuals:
     Min       1Q   Median       3Q      Max
-22.3551  -7.3878  -0.8737   6.3454  28.5375
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.2460     9.9774  -1.929  0.06674 .
diameter      3.7583     0.5885   6.387    2e-06 ***
finishS      17.9057     5.8642   3.053  0.00583 **
```

```
> Anova(lm(strength~diameter+finish, data=glassrod))
Anova Table (Type II tests)

Response: strength
          Sum Sq Df F value    Pr(>F)
diameter  8012.9  1 40.7878 1.996e-06 ***
finish    1831.6  1  9.3232  0.005825 **
Residuals 4322.0 22
```

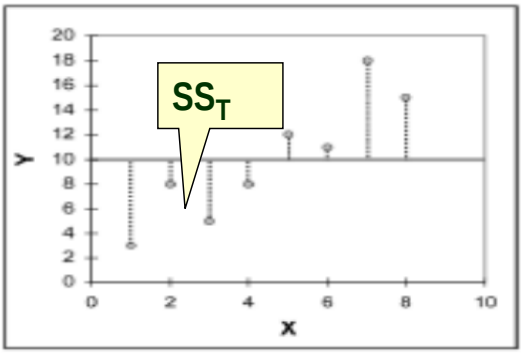
Lineare Regressions-Modelle:

- **Beispiel: hängt die Angst vor Spinnen von der Größe ab?**

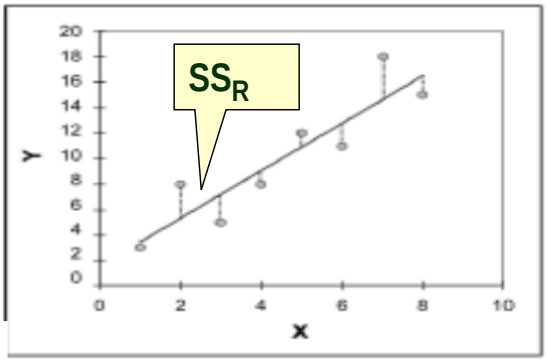


Lineare Regressions-Modelle:

- **Beispiel: hängt die Angst vor Spinnen von der Größe ab?**



H_0: Angst ist immer gleich



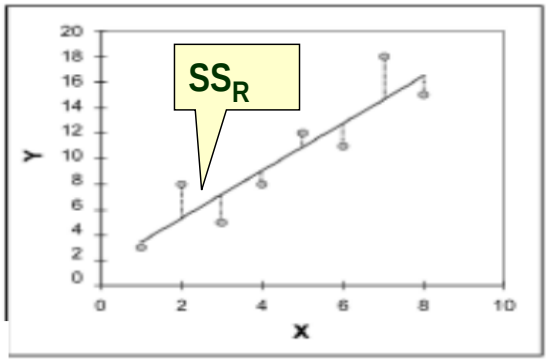
H_1: es gibt einen linearen Zusammenhang

$SS_M = SS_T - SS_R$: Maß für die Güte des Modells

Lineare Regressions-Modelle:

- **Beispiel: hängt die Angst vor Spinnen von der Größe ab?**

$$R^2 = \frac{SS_M}{SS_T}$$



R^2 = 0.6:
Die Angst geht zu 60%
auf Größe der Spinne zurück

$$F = \frac{R^2 / (k-1)}{(1 - R^2) / (N - k)}$$

df_{model}

df_{error}