

# Genetische Statistik

## WS 2021/2022 R-Übung 3: Regressionsmodelle in R

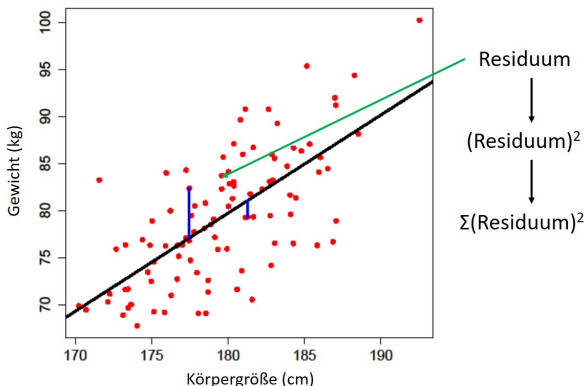
Dr. Janne Pott ([janne.pott@uni-leipzig.de](mailto:janne.pott@uni-leipzig.de))

November 23, 2021

# Aufgabe 1: Lineare Regression - Hintergrund (1)

- Regresswerte = Vorhergesagte Werte:  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residuen = Abweichung Vorhersage - Beobachtung:  $\hat{y}_i - y_i$
- Residual Sum of Square (RSS) = Summe der quadratischen Fehler:  
 $\sum_{i=1}^n (\hat{y}_i - y_i)^2$
- Mean Square Error = Residuale Varianz:  $\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$

# Aufgabe 1: Lineare Regression - Hintergrund (2)



**Figure 1:** Lineare Regression. Quelle: Vorlesung

# Aufgabe 1: Lineare Regression (Einschub - 1)

Größe (cm)	180	175	160	170	190
Gewicht (kg)	80	80	58	60	85

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n [(\beta_0 + \beta_1 x_i) - y_i]^2$$

- $\hat{\beta}_0$  und  $\hat{\beta}_1$ ?
- $\beta_0$  und  $\beta_1$  für den obigen Datensatz!
- Welches Gewicht können Sie für eine 176 cm große Person vorhersagen?

# Aufgabe 1: Lineare Regression (Einschub - 2)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n [(\beta_0 + \beta_1 x_i) - y_i]^2; \bar{x} = \frac{1}{n} \sum x_i \text{ und } \bar{y} = \frac{1}{n} \sum y_i$$

$$\begin{aligned} \frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} &= 2 \sum (\beta_0 + \beta_1 x_i - y_i) \\ &= 2n\beta_0 + 2\beta_1 \sum x_i - 2 \sum y_i \stackrel{!}{=} 0 \\ \Rightarrow \beta_0 &= \frac{1}{n} (\sum y_i - \beta_1 \sum x_i) = \bar{y} - \beta_1 \bar{x} \end{aligned}$$

(1)

# Aufgabe 1: Lineare Regression (Einschub - 3)

$$\begin{aligned}\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} &= 2 \sum [\beta_0 + \beta_1 x_i - y_i] x_i \\ &= 2\beta_0 \sum x_i + 2\beta_1 \sum x_i^2 - 2 \sum x_i y_i \stackrel{!}{=} 0 \\ \Rightarrow \beta_1 &= \frac{\sum x_i y_i - \beta_0 \sum x_i}{\sum x_i^2} = \frac{\sum x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum x_i}{\sum x_i^2} \\ &\Leftrightarrow \beta_1 (\sum x_i^2 - \bar{x} \sum x_i) = \sum x_i y_i - \bar{y} \sum x_i \\ &\Leftrightarrow \beta_1 (\sum x_i^2 - n \bar{x}^2) = \sum x_i y_i - n \bar{x} \bar{y}\end{aligned}\tag{2}$$

# Aufgabe 1: Lineare Regression (Einschub - 4)

$$\begin{aligned}\Rightarrow \beta_1 &= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \\&= \frac{\sum x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2 - n \bar{x}^2 + n \bar{x}^2} \\&= \frac{\sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n \bar{x} \bar{y}}{\sum x_i^2 - 2 \bar{x} \sum x_i + n \bar{x}^2} \\&= \frac{\sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y})}{\sum (x_i^2 - 2 x_i \bar{x} + \bar{x}^2)} \\&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\&= \frac{\text{empirischen Kovarianz von } x \text{ und } y}{\text{empirischen Varianz von } x}\end{aligned}$$

## Aufgabe 1: Lineare Regression (Einschub - 5)

```
groesse<-c(180,175,160,170,190)
gewicht<-c(80,80,58,60,85)

x_m<-mean(groesse)
y_m<-mean(gewicht)
var1<-sum((groesse-x_m)*(gewicht-y_m))
var2<-sum((groesse-x_m)^2)
b1<-var1/var2
b0<-y_m - b1*x_m

b1; b0; b0+b1*176

## [1] 1.01

## [1] -104.15

## [1] 73.61
```



# Aufgabe 1: Lineare Regression

- Univariate lineare Regression für *sex* und *SNP* auf *trait*!
- Multivariate lineare Regression & Interaktionsanalyse!
- Bestes Modell?
- Welches genetische Modell wird hier verwendet?
- Test der anderen genetischen Modelle
- Autosomal oder X-chromosomal?

# Aufgabe 1: Lösung (1)

```
mod4<-lm(trait~sex*SNP,data=myDat)
summary(mod4)
```

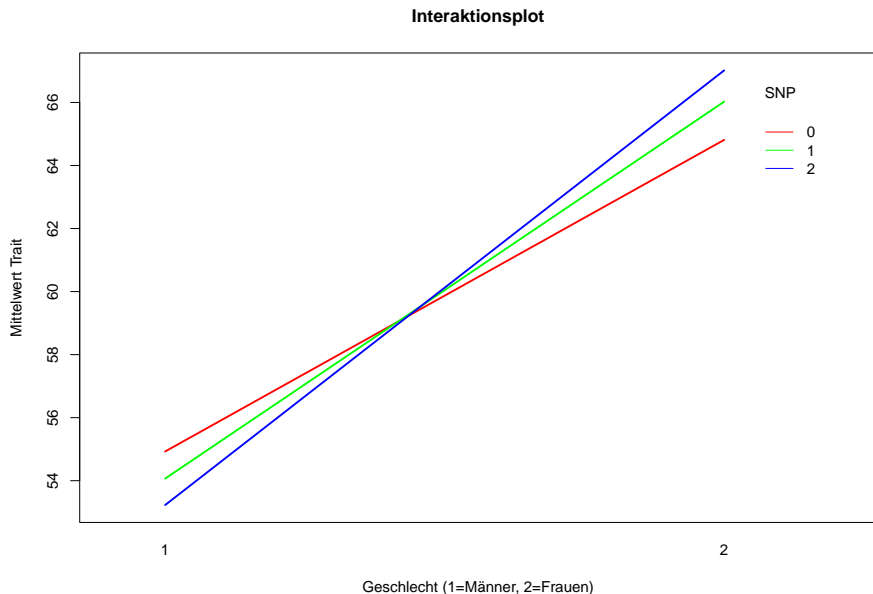
```
##
## Call:
## lm(formula = trait ~ sex * SNP, data = myDat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0476  -3.3349  -0.0202   3.3597  22.0749
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   44.9109     0.1842   243.76  <2e-16 ***
## sex           10.0145     0.1475    67.88  <2e-16 ***
## SNP           -2.7565     0.1821   -15.13  <2e-16 ***
## sex:SNP        1.9012     0.1157    16.43  <2e-16 ***
```

# Aufgabe 1: Lösung (2)

**Table 2:** Modellvergleich 1 (Modellgüte Adjustierung)

	adj. $r^2$	logLik	df	AIC
sex	0.5881264	-60649.90	3	121305.8
SNP	0.2416666	-66753.96	3	133513.9
sex+SNP	0.5881462	-60648.92	4	121305.8
sex*SNP	0.5936130	-60514.79	5	121039.6

# Aufgabe 1: Lösung (3)



# Aufgabe 1: Lösung (4)

```
add<-wskAB + 2*wskBB
```

```
dom<-wskAB + wskBB
```

```
rez<-wskBB
```

```
mod_add<-lm(myDat$trait ~ add*myDat$sex)
```

```
mod_dom<-lm(myDat$trait ~ dom*myDat$sex)
```

```
mod_rez<-lm(myDat$trait ~ rez*myDat$sex)
```

# Aufgabe 1: Lösung (5)

**Table 3:** Modellvergleich 2 (Modellgüte genetisches Modell)

	adj. $r^2$	logLik	df	AIC
additiv	0.5936130	-60514.79	5	121039.6
dominant	0.5915228	-60566.09	5	121142.2
rezessiv	0.5916271	-60563.54	5	121137.1

## Aufgabe 1: Lösung (6)

```
mod<-lm(myDat$trait ~ (wskAB + wskBB) * myDat$sex)
summary(mod)
```

```
##
## Call:
## lm(formula = myDat$trait ~ (wskAB + wskBB) * myDat$sex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0489  -3.3386  -0.0188   3.3584  22.0624
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    45.0402     0.2425  185.701  <2e-16 ***
## wskAB         -2.9341     0.3030  -9.684   <2e-16 ***
## wskBB         -5.6036     0.4737 -11.830   <2e-16 ***
## myDat$sex       9.8866     0.2141  46.171   <2e-16 ***
```

# Aufgabe 1: Lösung (7)

**Table 4:** Modellvergleich 3 (Effektschätzer genetisches Modell)

	beta	adj. $r^2$	logLik	df	AIC
additiv	-2.756529	0.5936130	-60514.79	5	121039.6
dominant	-3.757434	0.5915228	-60566.09	5	121142.2
rezessiv	-3.875292	0.5916271	-60563.54	5	121137.1
Effekt AB	-2.934067	0.5935862	-60514.45	7	121042.9
Effekt BB	-5.603627	0.5935862	-60514.45	7	121042.9



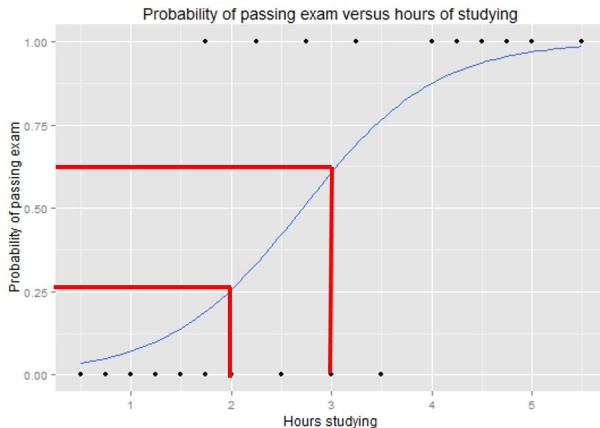
# Aufgabe 1: Lösung (8) - Zusammenfassung

- Interaktion am besten, wenn auch nur kleiner Zuwachs für  $r^2$ .
- Additives Modell am besten, bestätigt im multivariaten Modell:
  - beide genetischen Effektschätzer sind signifikant von 0 unterschiedlich
  - $\beta_1 \approx 0.5 * \beta_2$
- X-chromosom: Männer nur A oder B (keine Heterozygoten möglich)
  - Option 1: 3770 Männer als Genotypisierungsfehler unwahrscheinlich
  - Option 2: keine X-Inaktivierung, die 618 mit 2 sind Genotypisierungsfehler
  - Option 3: X-Chromosom, aber PAR (pseudo-autosomal-region), heterozygote Männer möglich
  - Fazit: Man sollte besser vorher wissen, was das für ein SNP ist, und dann ggf die Modelle wählen (Inaktivierung, Interaktion).

## Aufgabe 2: Logistische Regression - Hintergrund (1)

- Bsp. Fragestellung: Eine Gruppe von 20 Studierenden lernt zwischen 0 und 6 Stunden für eine Prüfung. Wie beeinflusst der Lernaufwand die Wahrscheinlichkeit die Prüfung zu bestehen?
- $\beta_1 = 1.5, \beta_0 = -4.1$
- Log-odds Bestehen =  $1.5 * x - 4.1 = 1.5 * (x - 2.7)$
- 1 Stunde mehr lernen erhöht die log-odds um 1.5; 50% Chance bei 2.7 h.
- $P(\text{bestehen} | \text{lernen} = 2h) = 0.25$
- $P(\text{bestehen} | \text{lernen} = 3h) = 0.61$
- $P(\text{bestehen}) = \frac{e^{\beta_0 + \beta_1 * x}}{1 + e^{\beta_0 + \beta_1 * x}}$

## Aufgabe 2: Logistische Regression - Hintergrund (2)

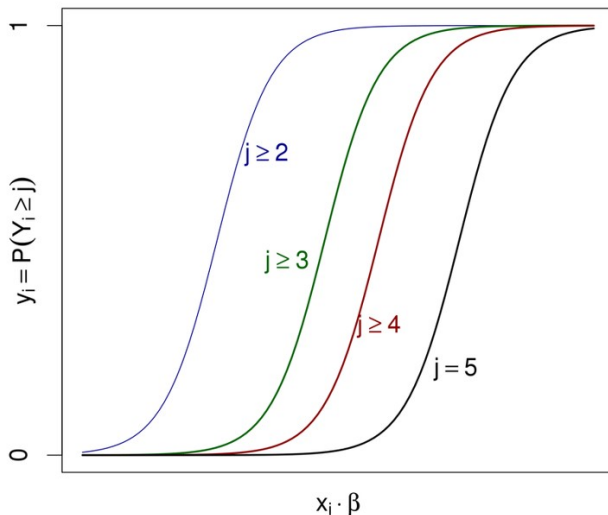


**Figure 2:** Logistische Regression. Quelle: Wikipedia

## Aufgabe 2: Proportional Odds - Hintergrund (1)

- Bsp. Fragestellung: Eine Gruppe von 20 Studierenden lernt zwischen 0 und 6 Stunden für eine Prüfung. Wie beeinflusst der Lernaufwand die Wahrscheinlichkeit die Prüfung mit Note 1, 2, 3, 4 oder nicht zu bestehen?
- Annahme: “equal slope”, die logistische Funktion für die Wahrscheinlichkeit, mindestens Note  $j$  zu erreichen, verläuft für jede Note parallel verschoben, aber mit der gleichen Steigung.
- $\beta_1 = 1.5$  wie vorher, aber pro Stufe anderer Intercept

## Aufgabe 2: Logistische Regression - Hintergrund (2)



**Figure 3:** Proportional Odds Regression. Quelle: Schlarmann, Galatsch, 2014

## Aufgabe 2: Logistische / Proportional Odds Regression

- **Median** und **Quartile** von *trait* & Definition je eines binären und kategorialen Phänotypen.
- Univariaten, multivariaten Effekte von *sex* und *SNP* auf *trait2* mittels logistischer Regression!
- Univariaten, multivariaten Effekte von *sex* und *SNP* auf *trait3* mittels proportional odds regression!
- Vergleichen mit linearer Regression

## Aufgabe 2: Lösung (1)

```
modB1<-glm(trait2~sex,family = "binomial",data = myDat)
modB2<-glm(trait2~SNP,family = "binomial",data = myDat)
modB3<-glm(trait2~SNP+sex,family = "binomial",data = myDat)
modB4<-glm(trait2~SNP*sex,family = "binomial",data = myDat)

modD1<-polr(as.factor(trait3)~sex,data = myDat,Hess = T)
modD2<-polr(as.factor(trait3)~SNP,data = myDat,Hess = T)
modD3<-polr(as.factor(trait3)~SNP+sex,data = myDat,Hess = T)
modD4<-polr(as.factor(trait3)~SNP*sex,data = myDat,Hess = T)
```

## Aufgabe 2: Lösung (2)

```
summary(modB4)
```

```
##
```

```
## Call:
```

```
## glm(formula = trait2 ~ SNP * sex, family = "binomial", data
```

```
##
```

```
## Deviance Residuals:
```

##	Min	1Q	Median	3Q	Max
##	-2.19797	-0.54185	0.02112	0.51875	2.28582

```
##
```

```
## Coefficients:
```

##		Estimate	Std. Error	z value	Pr(> z )
##	(Intercept)	-5.24240	0.10735	-48.834	<2e-16 ***
##	SNP	-1.07609	0.11971	-8.989	<2e-16 ***
##	sex	3.39801	0.08530	39.835	<2e-16 ***
##	SNP:sex	0.73013	0.07401	9.865	<2e-16 ***

```
## ---
```



## Aufgabe 2: Lösung (3)

```
summary(modD4)
```

```
## Call:
```

```
## polr(formula = as.factor(trait3) ~ SNP * sex, data = myDat,
```

```
##
```

```
## Coefficients:
```

```
##           Value Std. Error t value
```

```
## SNP      -0.9143    0.07104  -12.87
```

```
## sex       3.6291    0.06431   56.43
```

```
## SNP:sex   0.6197    0.04498   13.78
```

```
##
```

```
## Intercepts:
```

```
##      Value      Std. Error t value
```

```
## 0|1    3.4137    0.0753    45.3120
```

```
## 1|2    5.5537    0.0859    64.6222
```

```
## 2|3    7.7634    0.0977    79.4778
```

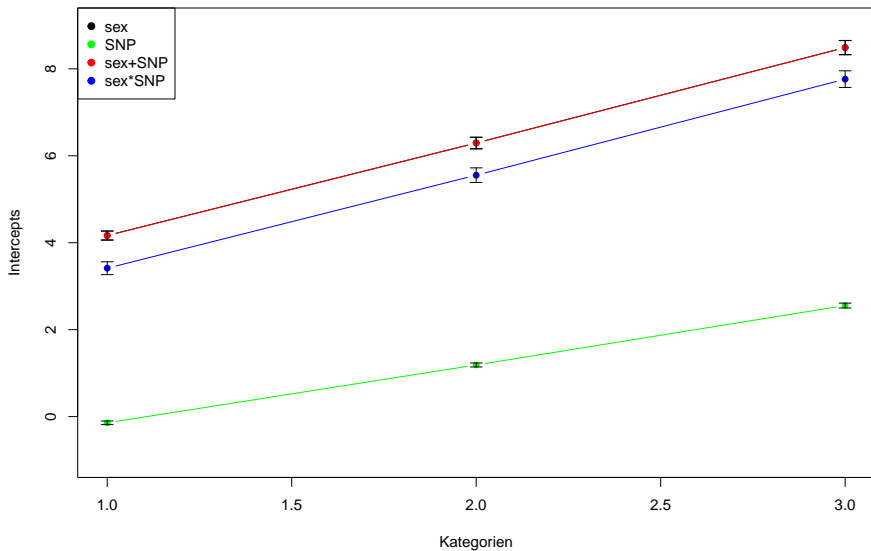
```
##
```

## Aufgabe 2: Lösung (4)

**Table 5:** Modellvergleich 4 (Modellgüte log & prop Odds)

	logLik - log	AIC - log	logLik - prop Odds	AIC - prop Odds
sex	-7093.055	14190.11	-19890.78	39789.56
SNP	-11251.702	22507.40	-25097.02	50202.03
sex+SNP	-7092.853	14191.71	-19890.66	39791.31
sex*SNP	-7043.272	14094.54	-19795.05	39602.11

## Aufgabe 2: Lösung (5)

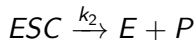
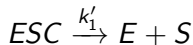
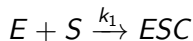


## Aufgabe 2: Lösung (6) - Zusammenfassung

- Das Interaktionsmodell bleibt das beste
- Die Intercepts bei der Prop. Odds Regression steigen quasi-linear an (=proportional)

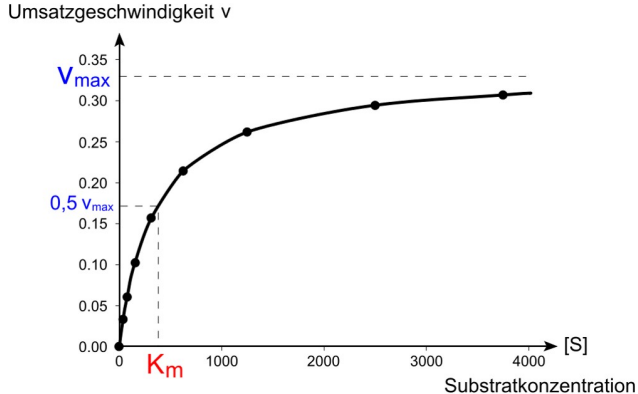
## Aufgabe 3: Nichtlineare Regression - Hintergrund (1)

- Sättigungsfunktion: Umsatzgeschwindigkeit einer enzymatischen Reaktion in Abhängigkeit der Substratkonzentration
- Annahme: Enzymkonzentration ist fix.
- $V_{max}$ : maximale Reaktionsgeschwindigkeit
- $K_m = \frac{k_1' + k_2}{k_1}$ : Michaelis-Menten-Konstante = Substratkonzentration bei halb-maximaler Geschwindigkeit, Abhängig von Reaktion und Milieu (z.B. pH-Wert)
- $K_d = \frac{k_1'}{k_1} =$  Dissoziationskonstante



(3)

## Aufgabe 2: Logistische Regression - Hintergrund (2)



**Figure 4:** Proportional Odds Regression. Quelle: wikipedia

# Aufgabe 3: Nichtlineare Regression

- $V_{max}$  und  $K_m$  für Erwachsene und Embryonen getrennt!
  - Startwerte:  $V_{max} = \max(v)$  und  $K_m = \frac{1}{2}\max(v)$
- Was passiert wenn man die Startwerte weglässt?
- Plot und Interpretation!

## Aufgabe 3: Lösung (1)

```
vmaxA<-max(myDat$vA)
vmaxE<-max(myDat$vE)
kmA<-vmaxA/2
kmE<-vmaxE/2

modA<-nls(vA ~ vmax*cS/(cS+km), data=myDat,
          start = list(km=kmA,vmax=vmaxA))
modA2<-nls(vA ~ vmax*cS/(cS+km), data=myDat)

modE<-nls(vE ~ vmax*cS/(cS+km), data=myDat,
          start = list(km=kmE,vmax=vmaxE))
modE2<-nls(vE ~ vmax*cS/(cS+km), data=myDat)
```



## Aufgabe 3: Lösung (2)

```
summary(modA)$coef
```

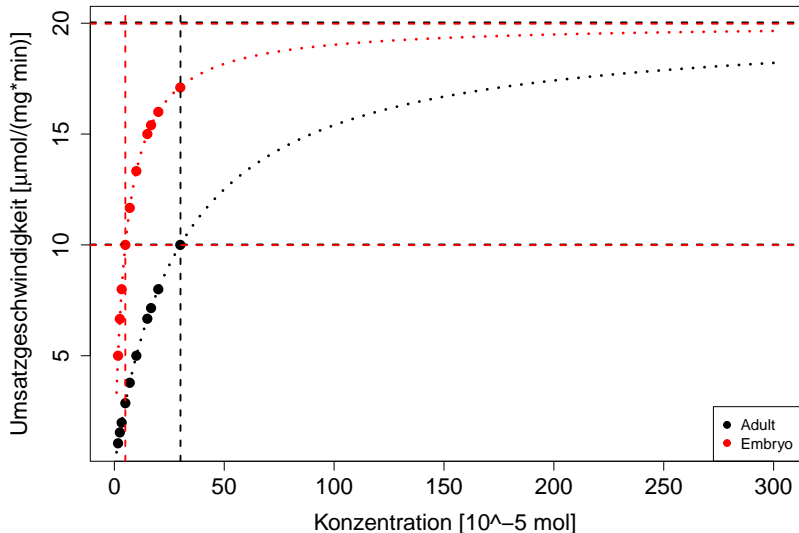
```
##           Estimate Std. Error  t value      Pr(>|t|)
## km      30.09194  0.10508271  286.3643  2.475794e-17
## vmax    20.03563  0.04408017  454.5270  6.147233e-19
```

```
summary(modA2)$coef
```

```
##           Estimate Std. Error  t value      Pr(>|t|)
## vmax    20.03563  0.04408017  454.5270  6.147236e-19
## km      30.09194  0.10508272  286.3643  2.475795e-17
```

# Aufgabe 3: Lösung (3)

## Michaelis-Menten Kinetik



## Aufgabe 3: Lösung (4) - Zusammenfassung

- Maximale Umsatzgeschwindigkeit ist in Embryonen und Erwachsenen fast gleich
- Die Affinität ( $K_m$ ) unterscheidet sich stark

## Blatt 3: - Zusammenfassung

- In der Genetik wird am häufigsten die lineare Regression mit einem additiven genetischen Effekt bestimmt (am einfachsten & am schnellsten)
- Ausnahme: es gibt Vorwissen!
- Kandidaten-Loki können nochmal genauer mit anderen Modellen getestet werden
- Lineare & logistische Regression sind in den meisten Analysesoftware enthalten, prop. Odds nicht (daher auch recht selten  $\rightarrow$  Annahme pseudo-linear  $\rightarrow$  lin. Regression)