

Goodness of Fit

Wie gut passt das Modell? Deviance = Lack of fit.

$\hat{L}_c = L_c(\hat{\beta}) =$ maximale Likelihood des betrachteten Modells (c= current)

$\hat{L}_f =$ maximale Likelihood des vollen Modells (perfekter Fit: $y_i/n_i = \hat{p}_i$)

(skalierte) Devianz:

$$D^* = -2 \log\left(\frac{\hat{L}_c}{\hat{L}_f}\right) = -2[\log \hat{L}_c - \log \hat{L}_f]$$

Devianz für Binomialdaten

Likelihood-Funktion:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

Devianz:

$$D^* = 2 \sum_{i=1}^n \left\{ y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left(\frac{y_i}{\hat{y}_i} \right) \right\}$$

Bei p unbekannten Parametern ist $D^* \stackrel{as}{\sim} \chi_{n-p}^2$, falls n_i genügend gross. D. h. wenn D^* ca. $n - p$, dann ist das Modell gut. Für $D^* > \chi_{n-p,0.95}^2$ sign. Lack of Fit.

Devianz für Binärdaten

Devianz:

$$D^* = -2 \sum_{i=1}^n \{ \hat{p}_i \log \hat{p}_i + (1 - \hat{p}_i) \log(1 - \hat{p}_i) \}$$

Die Devianz D^* ist für binäre Daten kein sinnvolles Mass für die Diskrepanz zwischen Beobachtungen (Daten) und angepassten Werten (Modell), da sie nicht abhängt von den Beobachtungen.

Vergleich von Modellen

Modell (1): q Parameter D_1^* mit $df = n - q$

Modell (2): p Parameter D_2^* mit $df = n - p$

$q < p$, Modell (1) ist im Model (2) enthalten.

Modellvergleich heisst $H_0 : \beta_{q+1} = \dots = \beta_p = 0$

$$D_1^* - D_2^* = -2 \left[\log \hat{L}_{c_1} - \log \hat{L}_{c_2} \right] \sim \chi_{p-q}^2$$

Falls $D_1^* - D_2^* > \chi_{p-q,0.95}^2$ ist, genügt das kleinere Modell (1) nicht.

AIC Kriterium: $AIC = D^* + 2p$.

Prognose bei Prostatakrebs

Haben das Alter, Phosphatasesäure, Röntgenbefund (0=negativ, 1=positiv), Tumorgrösse (0= klein, 1=gross) und Tumorgrad (0= weniger ernst, 1= ernst) einen Zusammenhang mit der Ausbreitung des Krebses in die Lymphknoten?

Pat	Age	Acid	X-ray	Size	Grade	Nodal Involv.
1	66	0.48	0	0	0	0
2	68	0.56	0	0	0	0
...	

R-Output für Prostata

```
glm(formula= y ~age + log(acid) + xray + size + grade,  
     family = binomial,data = prostata)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	2.45977	3.52218	0.698	0.4849	
age	-0.06370	0.05874	-1.085	0.2781	
log(acid)	2.57250	1.19700	2.149	0.0316	*
xray1	2.04009	0.82885	2.461	0.0138	*
size1	1.54664	0.78113	1.980	0.0477	*
grade1	0.83447	0.78895	1.058	0.2902	

Null deviance: 70.252 on 52 degrees of freedom

Residual deviance: 46.560 on 47 degrees of freedom

AIC: 58.56

R-Output für Prostata (Fort. I)

```
> add1(glm3, ~.^2, test="Chisq")
```

```
Single term additions
```

```
Model:
```

```
y ~ age + log(acid) + xray + size + grade
```

	Df	Deviance	AIC	LRT	Pr(Chi)	
<none>		46.560	58.560			
age:log(acid)	1	46.035	60.035	0.525	0.468550	
age:xray	1	45.518	59.518	1.042	0.307274	
age:size	1	46.485	60.485	0.075	0.783925	
age:grade	1	45.915	59.915	0.645	0.421786	
log(acid):xray	1	45.593	59.593	0.967	0.325529	
log(acid):size	1	46.540	60.540	0.020	0.887510	
log(acid):grade	1	42.518	56.518	4.042	0.044385	*
xray:size	1	46.303	60.303	0.257	0.612271	
xray:grade	1	46.514	60.514	0.046	0.829807	
size:grade	1	38.431	52.431	8.129	0.004356	**

R-Output für Prostata (Fort. II)

```
> stepAIC(glm5) #library(MASS)
Start:  AIC= 51.34
y~age + log(acid) + xray + size + grade + grade:size + log(acid):grade
      Df Deviance    AIC
- age      1   36.287 50.287
<none>           35.338 51.338
- log(acid):grade 1   38.431 52.431
- xray      1   41.104 55.104
- size:grade 1   42.518 56.518

Step:  AIC= 50.29
y ~ log(acid) + xray + size + grade + size:grade + log(acid):grade
      Df Deviance    AIC
<none>           36.287 50.287
- log(acid):grade 1   40.454 52.454
- xray      1   41.982 53.982
- size:grade 1   43.157 55.157
```


R-Output für Prostata (Fort.III)

```
Call: glm(formula = y ~ log(acid) + xray + size + grade +  
size:grade + log(acid):grade, family = binomial, data = prost
```

Coefficients:

(Intercept)	log(acid)	xray1	size1	grade1	size1:grade1
-2.553	1.709	2.340	3.138	9.961	-5.648
log(acid):grade1					
10.426					

Degrees of Freedom: 52 Total (i.e. Null); 46 Residual

Null Deviance: 70.25

Residual Deviance: 36.29 AIC: 50.29

Definition von Residuen

Sei $\hat{y}_i = n_i \hat{p}_i$.

Pearson Residuen:

$$\frac{y_i - \hat{y}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}$$

Devianz Residuen:

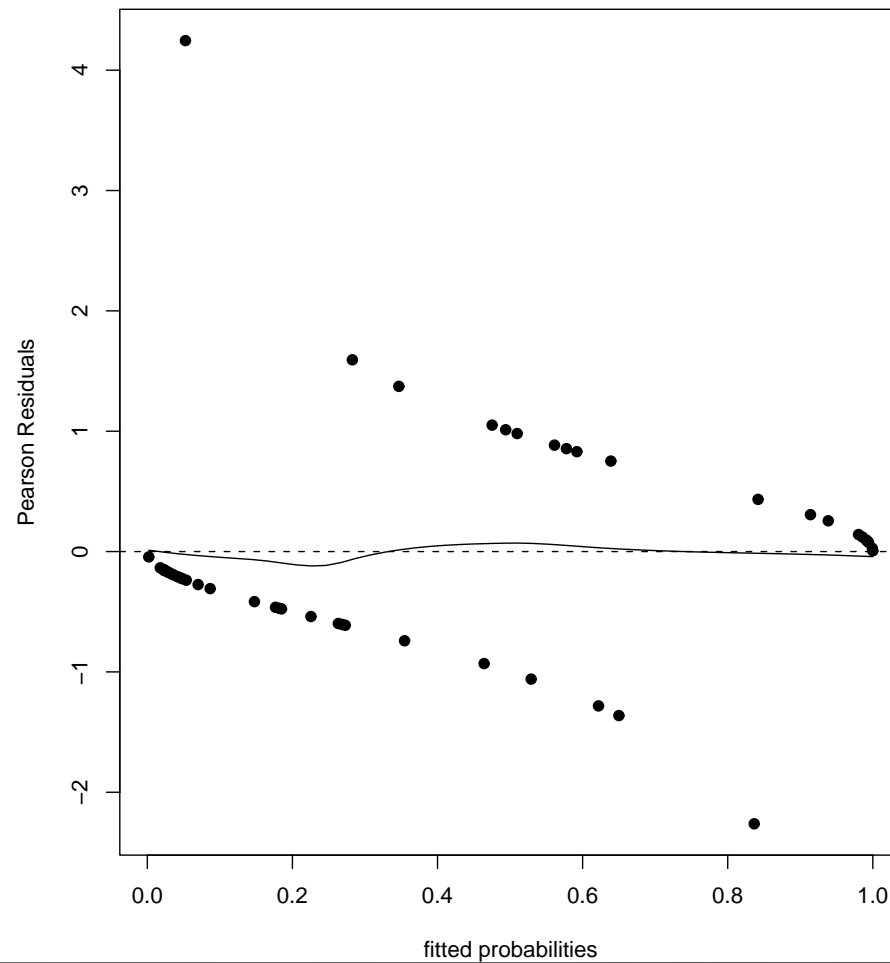
$$\text{sign}(y_i - \hat{y}_i) \sqrt{d_i}$$

wobei d_i die i-te Komponente der Devianz ist.

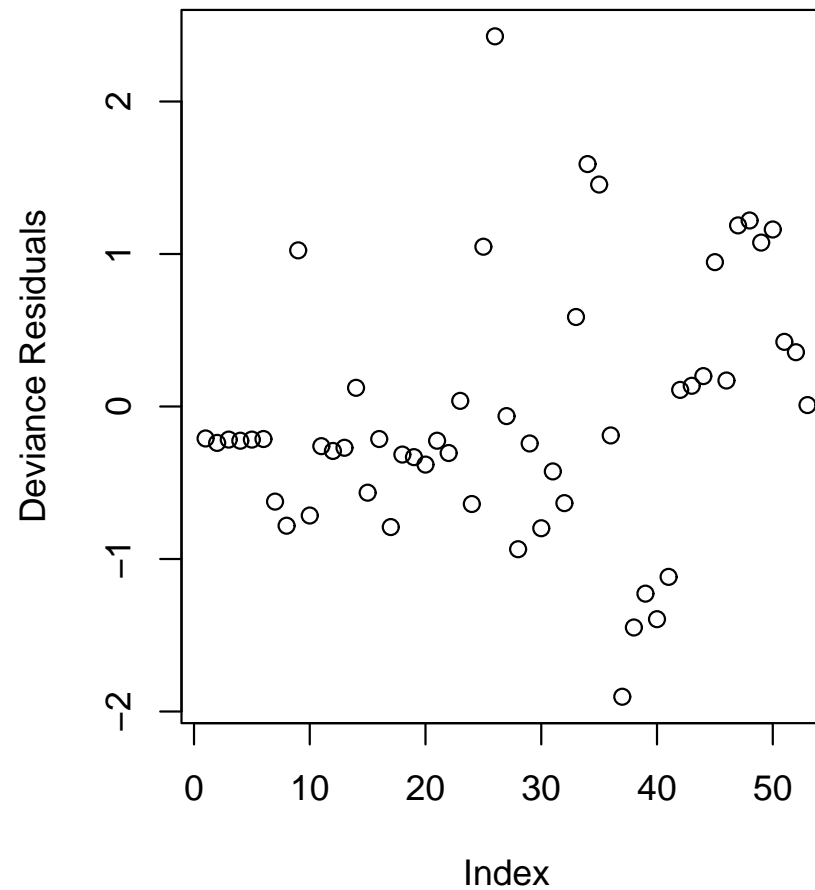
Residuenanalyse

- Für Binärdaten nehmen die Residuen für gegebenes i nur 2 Werte an. Der Plot Res vs. \hat{p}_i ist deshalb schwer interpretierbar. Glättung hilft.
- Normalplot macht keinen Sinn, ausser wenn die n_i gross sind.
- Indexplot der Residuen zur Identifikation von Ausreissern (Fehlklassifikationen)
- Leverages und Cook's Distances

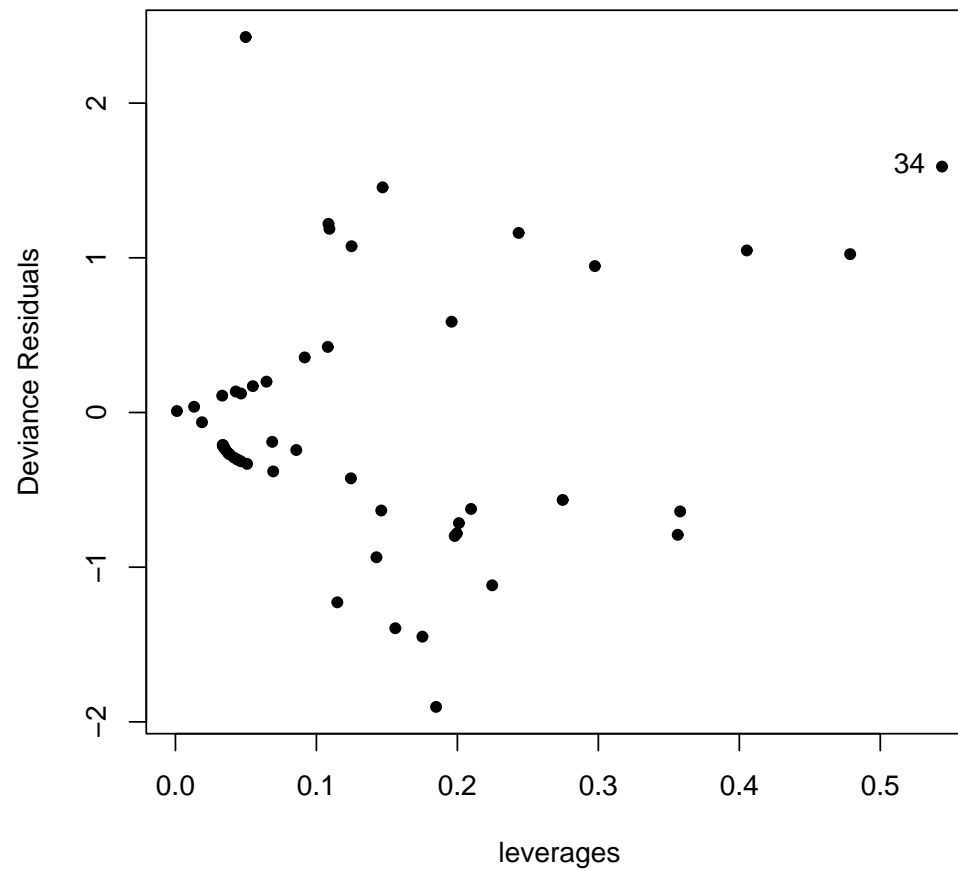
Pearson Residuen vs. \hat{p}_i



Indexplot Devianz-Residuen



Devianz-Residuen vs. Leverages



Cook's Distances

