

Lecture 4: Parameter estimation and diagnostics in logistic regression

Claudia Czado

TU München



Overview

- Parameter estimation
- Regression diagnostics

Parameter estimation in logistic regression

loglikelihood:

$$\begin{aligned} l(\boldsymbol{\beta}) &:= \sum_{i=1}^n \left[(y_i \ln \left(\frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}} \right) + (n_i - y_i) \ln \left(1 - \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}} \right) \right] + \text{const ind.} \\ &= \sum_{i=1}^n \left[(y_i \mathbf{x}_i^t \boldsymbol{\beta}) - n_i \ln(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}) \right] + \text{const} \end{aligned} \quad \text{of } \boldsymbol{\beta}$$

scores:

$$\begin{aligned} s_j(\boldsymbol{\beta}) &:= \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n y_i x_{ij} - n_i \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}} x_{ij} = \sum_{i=1}^n x_{ij} \left(y_i - \underbrace{n_i \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}}}_{E(Y_i | \mathbf{X}_i = \mathbf{x}_i)} \right) \quad j = 1, \dots, p \\ \Rightarrow \mathbf{s}(\boldsymbol{\beta}) &= \mathbf{X}^t (\mathbf{Y} - E(\mathbf{Y} | \mathbf{X} = \mathbf{x})) \end{aligned}$$

Hessian matrix in logistic regression

$$\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = - \sum_{i=1}^n n_i \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}})^2} x_{is} x_{ir} = - \sum_{i=1}^n n_i p(\mathbf{x}_i) (1 - p(\mathbf{x}_i)) x_{is} x_{ir}$$

$$\Rightarrow H(\boldsymbol{\beta}) = \left[\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} \right]_{r,s=1,\dots,p} = -X^t D X \in \mathbb{R}^{p \times p}$$

where $D = \text{diag}(d_1, \dots, d_n)$ and $d_i := n_i p(\mathbf{x}_i) (1 - p(\mathbf{x}_i))$.

$H(\boldsymbol{\beta})$ independent of $\mathbf{Y}$ (since canonical link) $\Rightarrow E(H(\boldsymbol{\beta})) = H(\boldsymbol{\beta})$

Existence of MLE's in logistic regression

Proposition: *The log likelihood $l(\beta)$ in logistic regression is strict concave in β if $\text{rank}(X) = p$*

Proof: $H(\beta) = -X^t D X$

$$\Rightarrow \mathbf{Z}^t H(\beta) \mathbf{Z} = -\mathbf{Z}^t X^t D^{1/2} D^{1/2} X \mathbf{Z} = -\|D^{1/2} X \mathbf{Z}\|^2$$

$$\Rightarrow \|D^{1/2} X \mathbf{Z}\|^2 = 0 \Leftrightarrow D^{1/2} X \mathbf{Z} = \mathbf{0}$$

$$\Leftrightarrow X^t D^{1/2} D^{1/2} X \mathbf{Z} = \mathbf{0}$$

$$\stackrel{D^{1/2} X \text{ full rank}}{\Leftrightarrow} \mathbf{Z} = (X^t D X)^{-1} \mathbf{0}$$

$$\Leftrightarrow \mathbf{Z} = \mathbf{0} \quad q.e.d.$$

$\Rightarrow$ There is **at most one solution** to the score equations, i.e. **if the MLE of β exists, it is unique and solution to the score equations.**

Warning: MLE in logistic regression does not need to exist.

Example: $n_i = 1$. Assume there $\exists \beta^* \in \mathbb{R}^p$ with

$$\begin{aligned} \mathbf{x}_i^t \beta^* &> 0 & \text{if } Y_i = 1 \\ \mathbf{x}_i^t \beta^* &\leq 0 & \text{if } Y_i = 0 \end{aligned}$$

$$\begin{aligned} \Rightarrow l(\beta^*) &= \sum_{i=1}^n [y_i \mathbf{x}_i^t \beta^* - \ln(1 + e^{\mathbf{x}_i^t \beta^*})] \\ &= \sum_{i=1, Y_i=1}^n \{ \mathbf{x}_i^t \beta^* - \ln(1 + e^{\mathbf{x}_i^t \beta^*}) \} - \sum_{i=1, Y_i=0}^n \ln(1 + e^{\mathbf{x}_i^t \beta^*}) \end{aligned}$$

Consider $\alpha \beta^*$ for $\alpha > 0 \rightarrow \infty$

$$\Rightarrow l(\alpha \beta^*) = \sum_{i=1, Y_i=1}^n \{ \alpha \mathbf{x}_i^t \beta^* - \ln(1 + e^{\alpha \mathbf{x}_i^t \beta^*}) \} - \sum_{i=1, Y_i=0}^n \ln(1 + e^{\alpha \mathbf{x}_i^t \beta^*}) \rightarrow 0$$

for $\alpha \rightarrow \infty$.

We know that $L(\beta) = \prod_{i=1}^n \underbrace{p(\mathbf{x}_i)^{Y_i} (1 - p(\mathbf{x}_i))^{1-Y_i}}_{\leq 1} \leq 1 \Rightarrow l(\beta) \leq 0$

Therefore we found $\alpha \beta^*$ such that $l(\alpha \beta^*) \rightarrow 0 \Rightarrow$ **no MLE exists.**

Asymptotic theory

(Reference: Fahrmeir and Kaufmann (1985))

Under **regularity conditions** for $\hat{\beta}_n$ the MLE in logistic regression we have

1) $\hat{\beta}_n \rightarrow \beta$ a.s. for $n \rightarrow \infty$ (consistency)

2) $V(\beta)^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{D} N_p(0, I_p)$ where

$V(\beta) = [X^t D(\beta) X]^{-1}$ (asymptotic normality)

Logistic models for the Titanic data

Without Interaction Effects:

```
> options(contrasts = c("contr.treatment", "contr.poly"))
> f.main_cbind(Survived, 1 - Survived) ~ poly(Age,2) +Sex + PClass
> r.main_glm(f.main,family=binomial,na.action=na.omit,x=T)
> summary(r.main)
```

```
Call: glm(formula = cbind(Survived, 1 - Survived) ~ poly(Age,
  2) + Sex + PClass, family = binomial, na.action = na.omit)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.8	-0.72	-0.38	0.62	2.5

Coefficients:

	Value	Std. Error	t value
(Intercept)	2.5	0.24	10.4
poly(Age, 2)1	-14.9	2.97	-5.0
poly(Age, 2)2	3.7	2.53	1.4
Sex	-2.6	0.20	-13.0
PClass2nd	-1.2	0.26	-4.7
PClass3rd	-2.5	0.28	-8.9

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 1026 on 755 degrees of freedom

Residual Deviance: 693 on 750 degrees of freedom

Correlation of Coefficients:

	(Intercept)	poly(Age, 2)1	poly(Age, 2)2
poly(Age, 2)1	-0.41		
poly(Age, 2)2	-0.09	0.07	
Sex	-0.66	0.11	-0.02
PClass2nd	-0.66	0.41	0.13
PClass3rd	-0.76	0.52	0.09

	Sex	PClass2nd
poly(Age, 2)1		
poly(Age, 2)2		
Sex		
PClass2nd	0.16	
PClass3rd	0.30	0.61

```
> r.main$x[1:4,] # Designmatrix
```

	(Intercept)	poly(Age, 2)1	poly(Age, 2)2	Sex
1	1	-0.0036	-0.026	0
2	1	-0.0725	0.100	0
3	1	-0.0010	-0.027	1
4	1	-0.0138	-0.019	0

	PClass2nd	PClass3rd
1	0	0
2	0	0
3	0	0
4	0	0

Analysis of Deviance:

```
> anova(r.main)
```

```
Analysis of Deviance Table
```

```
Binomial model
```

```
Response: cbind(Survived, 1 - Survived)
```

```
Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			755	1026
poly(Age, 2)	2	12	753	1013
Sex	1	225	752	789
PClass	2	95	750	693

With Interaction Effects:

Linear Age Effect

```
> f.inter_cbind(Survived, 1 - Survived)
~ (Age + Sex + PClass)^2
> r.inter_glm(f.inter,family=binomial,na.action=na.omit)
> summary(r.inter)[[3]]
```

	Value	Std. Error	t value
(Intercept)	2.464	0.835	2.95
Age	0.013	0.020	0.67
Sex	-0.946	0.824	-1.15
PClass2nd	1.116	1.002	1.11
PClass3rd	-2.807	0.825	-3.40
Age:Sex	-0.068	0.018	-3.70
AgePClass2nd	-0.065	0.024	-2.67
AgePClass3rd	-0.007	0.020	-0.35
SexPClass2nd	-1.411	0.715	-1.97
SexPClass3rd	1.032	0.616	1.67

```
> anova(r.inter)
```

Analysis of Deviance Table

Binomial model

Response: cbind(Survived, 1 - Survived)

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			755	1026
Age	1	3	754	1023
Sex	1	227	753	796
PClass	2	100	751	695
Age:Sex	1	28	750	667
Age:PClass	2	5	748	662
Sex:PClass	2	21	746	641

A drop in deviance of $54=28+5+21$ on 5 df is highly significant ($p - value = 2.1e - 10$), therefore **strong interaction effects** are present.

Quadratic Age Effect:

```
> Age.poly1_poly(Age,2)[,1]
> Age.poly2_poly(Age,2)[,2]
> f.inter1_cbind(Survived, 1 - Survived) ~ Sex + PClass +Age.poly1+Age.poly2+
  Sex*Age.poly1+ Sex*PClass+Age.poly1*PClass+Age.poly2*PClass
> r.inter1_glm(f.inter1,family=binomial,na.action=na.omit)
> summary(r.inter1)[[3]]
```

	Value	Std. Error	t value
(Intercept)	2.92	0.47	6.27
Sex	-3.12	0.51	-6.14
PClass2nd	-0.72	0.58	-1.24
PClass3rd	-3.02	0.53	-5.65
Age.poly1	3.28	7.96	0.41
Age.poly2	-4.04	5.63	-0.72
Sex:Age.poly1	-21.94	7.10	-3.09
SexPClass2nd	-1.23	0.71	-1.74
SexPClass3rd	1.27	0.62	2.04
PClass2ndAge.poly1	-14.02	10.39	-1.35
PClass3rdAge.poly1	3.72	9.34	0.40
PClass2ndAge.poly2	23.07	9.50	2.43
PClass3rdAge.poly2	10.99	7.85	1.40

```
> anova(r.inter1)
```

Analysis of Deviance Table

Binomial model

Response: cbind(Survived, 1 - Survived)

Terms added sequentially (first to last)

		Df	Deviance	Resid. Df	Resid. Dev
	NULL			755	1026
	Sex	1	229	754	797
	PClass	2	73	752	724
	Age.poly1	1	28	751	695
	Age.poly2	1	2	750	693
	Sex:Age.poly1	1	29	749	664
	Sex:PClass	2	17	747	646
	Age.poly1:PClass	2	7	745	639
	Age.poly2:PClass	2	6	743	634

The **quadratic effect in Age in the interaction between Age and PClass** is **weakly significant**, since a drop in deviance of $641-634=7$ on 2 df gives a $p - value = .03$.

Are there any 3 Factor Interactions?

```
> f.inter3_cbind(Survived, 1 - Survived) ~ (Age + Sex + PClass)^3  
> r.inter3_glm(f.inter3,family=binomial,na.action=na.omit)  
> anova(r.inter3)
```

Analysis of Deviance Table

Binomial model

Response: cbind(Survived, 1 - Survived)

Terms added sequentially (first to last)

		Df	Deviance	Resid. Df	Resid. Dev
	NULL			755	1026
	Age	1	3	754	1023
	Sex	1	227	753	796
	PClass	2	100	751	695
	Age:Sex	1	28	750	667
	Age:PClass	2	5	748	662
	Sex:PClass	2	21	746	641
	Age:Sex:PClass	2	2	744	640

A drop of 2 on 2 df is nonsignificant, therefore **three factor interactions are not present.**

Grouped models:

The residual deviance is difficult to interpret for binary responses. When we group the data to get binomial responses the residual deviance can be approximated better by a Chi Square distribution under the assumption of a correct model. The data frame `titanic.group.data` contains the group data set over age.

Without Interaction Effects:

```
> attach(titanic.group.data)
> dim(titanic.group.data)
[1] 274    5          # grouping reduces obs from 1313 to 274
> titanic.group.data[1,]
  ID Age Not.Survived Survived PClass    Sex
  8  2             1         0    1st  female
> f.group
cbind(Survived, Not.Survived) ~ poly(Age, 2) + Sex +
    PClass
> summary(r.group)[[3]]
              Value Std. Error t value
(Intercept)   2.5      0.24    10.5
poly(Age, 2)1 -11.1     2.17    -5.1
poly(Age, 2)2  2.7     1.90     1.4 #nonsignificant
Sex           -2.6     0.20   -13.0
PClass2st    -1.2     0.26    -4.7
PClass3st    -2.5     0.28    -8.9
```

```
> anova(r.group)
```

```
Analysis of Deviance Table
```

```
Binomial model
```

```
Response: cbind(Survived, Not.Survived)
```

```
Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			273	645
poly(Age, 2)	2	12	271	632
Sex	1	225	270	407
PClass	2	95	268	312

```
> 1-pchisq(312,268)
```

```
[1] 0.033      # model not very good for  
              # grouped data
```

```
> 1-pchisq(693,750)
```

```
[1] 0.93      # unreliable in binary case
```

With Interaction Effects:

```
> f.group.inter_cbind(Survived, Not.Survived)
~(Age+Sex+PClass)^2
> r.group.inter_glm(f.group.inter,family=
  binomial,na.action=na.omit)
> summary(r.group.inter)[[3]]
```

	Value	Std. Error	t value
(Intercept)	2.464	0.843	2.92
Age	0.013	0.020	0.67
Sex	-0.947	0.830	-1.14
PClass2st	1.117	1.010	1.11
PClass3st	-2.807	0.831	-3.38
Age:Sex	-0.068	0.019	-3.68
AgePClass2st	-0.065	0.025	-2.65
AgePClass3st	-0.007	0.020	-0.35
SexPClass2st	-1.410	0.722	-1.95
SexPClass3st	1.033	0.622	1.66

```
> anova(r.group.inter)
Analysis of Deviance Table
```

Binomial model

Response: cbind(Survived, Not.Survived)

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			273	645
Age	1	3	272	642
Sex	1	227	271	415
PClass	2	100	269	314
Age:Sex	1	28	268	286
Age:PClass	2	5	266	281
Sex:PClass	2	21	264	260

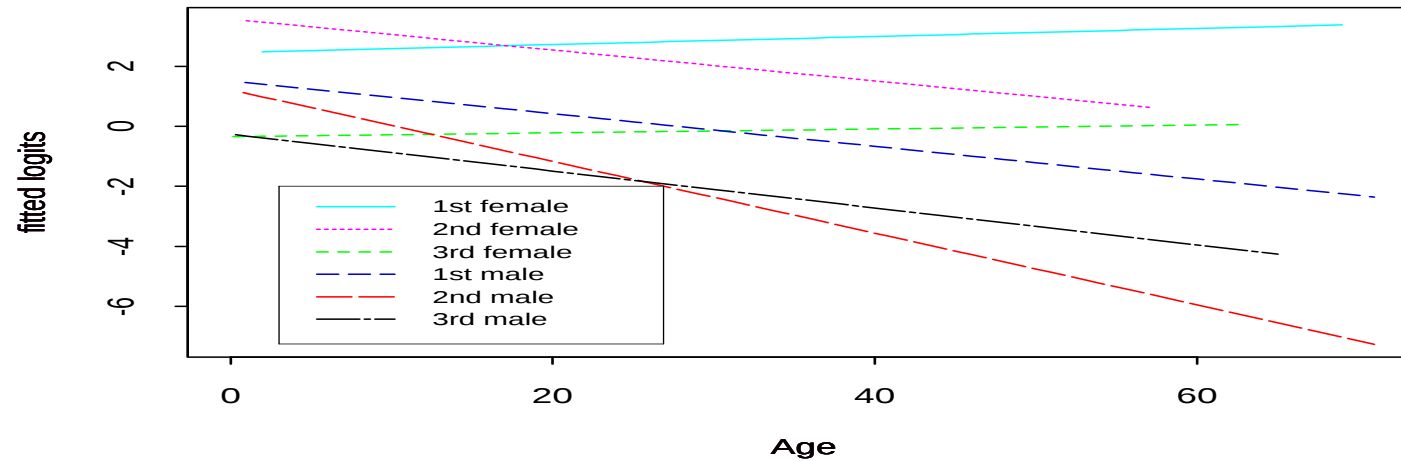
```
> 1-pchisq(264,260)
```

```
[1] 0.42 # residual deviance test p-value
```

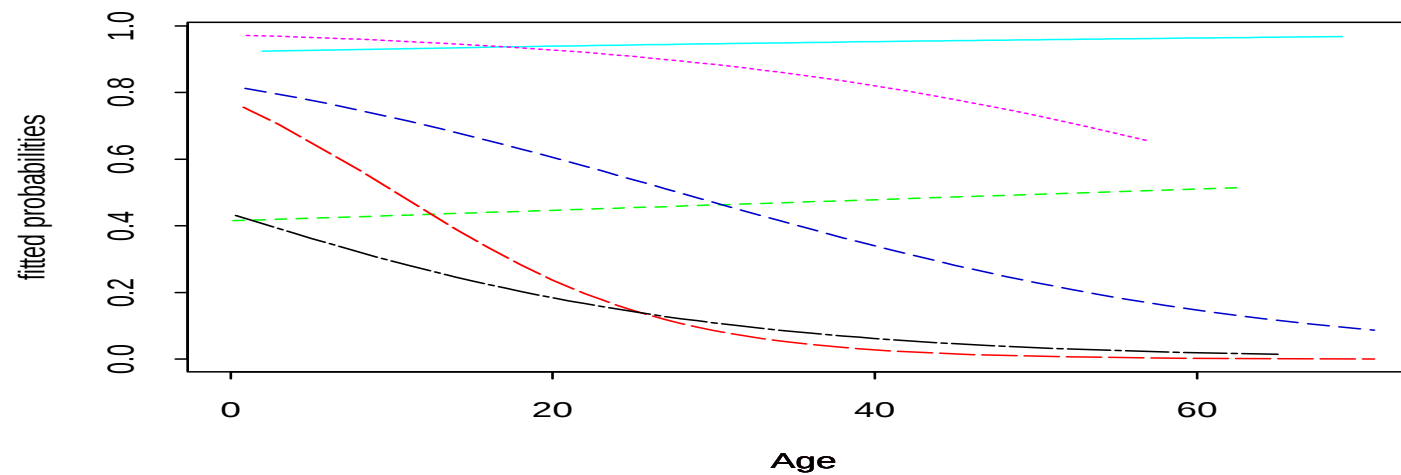
The residual deviance test gives $p - value = .42$, i.e. **Interactions** improve fit **strongly**.

Interpretation of Model with Interaction:

Fitted Logits



Fitted Survival Probabilities



Diagnostics in logistic regression

Residuals: $Y_i \sim \text{bin}(n_i, p_i) \quad p_i = \frac{e^{x_i^t \beta}}{1 + e^{x_i^t \beta}} \quad \hat{p}_i = \frac{e^{x_i^t \hat{\beta}}}{1 + e^{x_i^t \hat{\beta}}}$

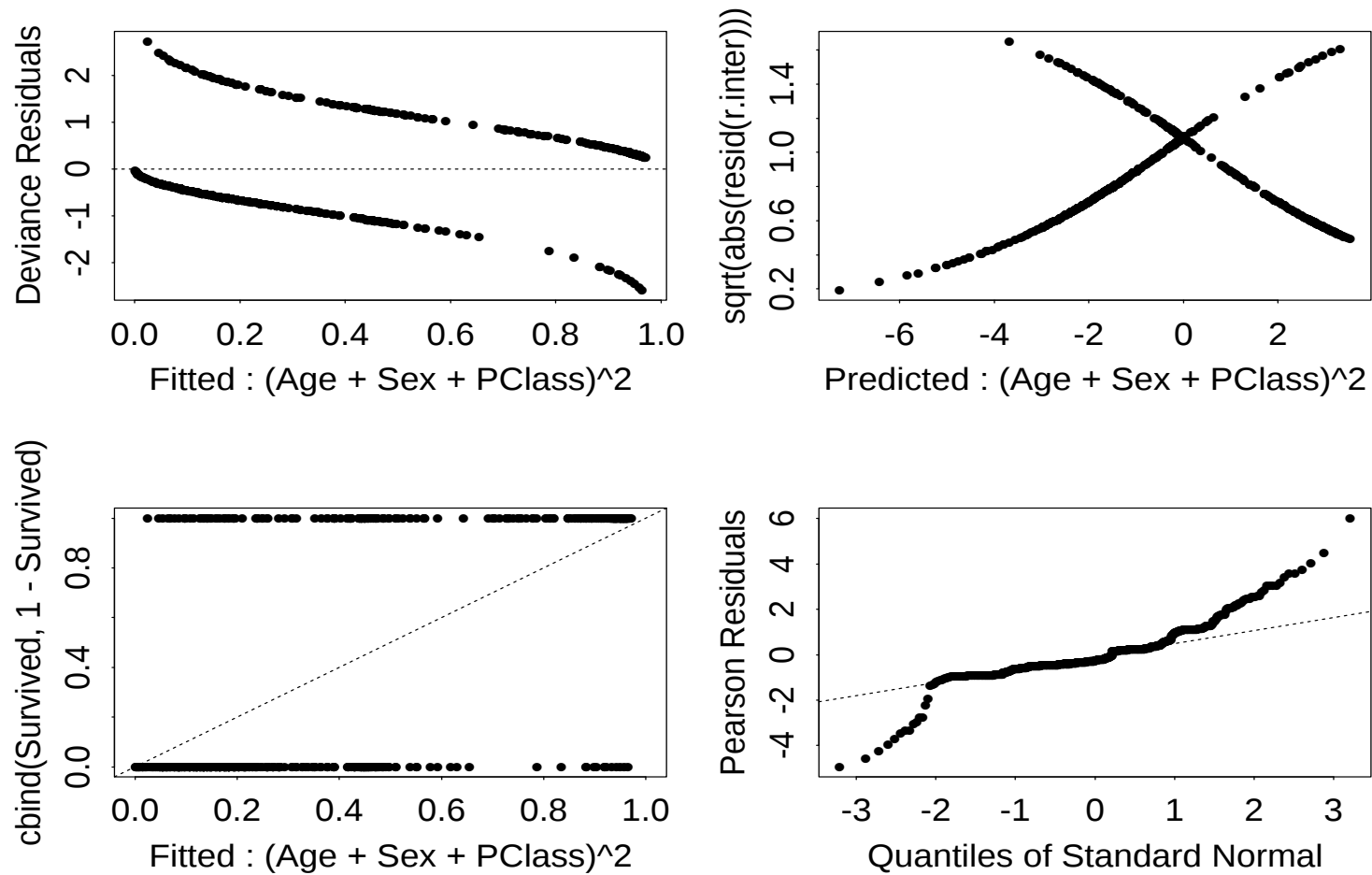
raw residuals: $e_i^r := Y_i - n_i \hat{p}_i$

Pearson residuals: $e_i^P := \frac{Y_i - n_i \hat{p}_i}{(n_i \hat{p}_i (1 - \hat{p}_i))^{1/2}}$
 $\Rightarrow \chi^2 = \sum_{i=1}^n (e_i^P)^2$

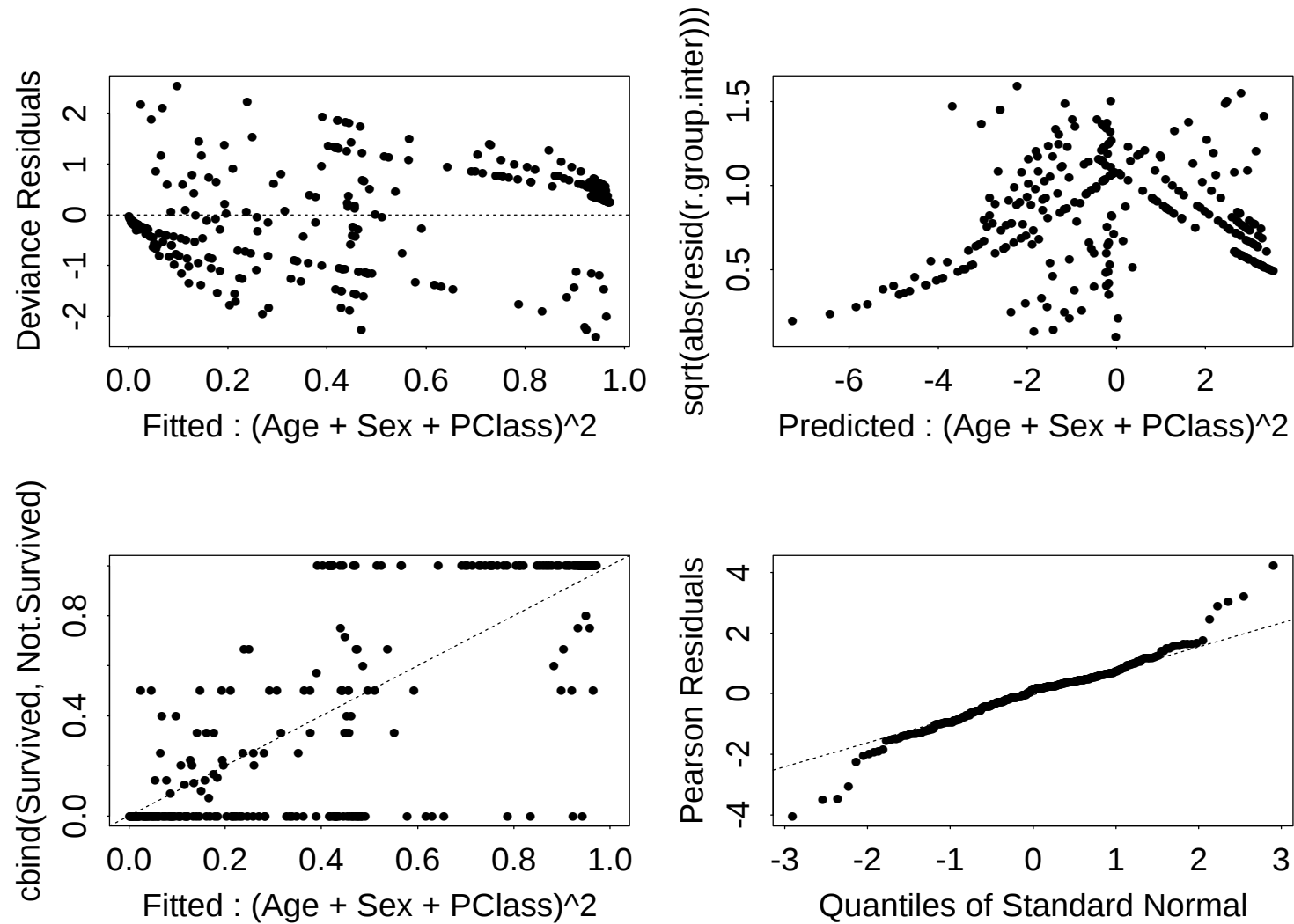
deviance residuals: $e_i^d := \text{sign}(Y_i - n_i \hat{p}_i) \left(2Y_i \ln \left(\frac{Y_i}{n_i \hat{p}_i} \right) + 2(n_i - Y_i) \ln \left(\frac{n_i - Y_i}{n_i (1 - \hat{p}_i)} \right) \right)$
 $\Rightarrow D = \sum_{i=1}^n (e_i^d)^2$

Residual Analysis:

Residual Analysis of Ungrouped Data



Residual Analysis of Grouped Data



Adjusted residuals

Want to adjust raw residuals such that adjusted residuals have unit variance.

Heuristic derivation: We have shown that $\hat{\beta}$ can be calculated as a weighted *LSE* with response

$$Z_i^\beta := \eta_i + (y_i - \mu_i) \frac{d\eta_i}{d\mu_i} \quad i = 1, \dots, n$$

Here $\eta_i = \mathbf{x}_i^t \boldsymbol{\beta} = \ln \left(\frac{p_i}{1-p_i} \right) = \ln \left(\frac{n_i p_i}{n_i - n_i p_i} \right) = \ln \left(\frac{\mu_i}{n_i - \mu_i} \right)$

$$\Rightarrow \frac{d\eta_i}{d\mu_i} = \frac{n_i - \mu_i}{\mu_i} \left[\frac{(n_i - \mu_i - (-1)\mu_i)}{(n_i - \mu_i)^2} \right] = \frac{n_i}{\mu_i(n_i - \mu_i)} = \frac{1}{n_i p_i (1 - p_i)}$$

$$\Rightarrow Z_i^\beta = \mathbf{x}_i^t \boldsymbol{\beta} + (y_i - n_i p_i) \frac{1}{n_i p_i (1 - p_i)} \quad \text{in logistic regression}$$

$$\Rightarrow \mathbf{Z}^\beta = \mathbf{X} \boldsymbol{\beta} + \mathbf{D}^{-1}(\boldsymbol{\beta}) \boldsymbol{\epsilon} \quad \text{where}$$

$$\boldsymbol{\epsilon} := \mathbf{Y} - (n_1 p_1, \dots, n_n p_n)^t$$

$$\mathbf{D}(\boldsymbol{\beta}) = \text{diag}(d_1(\boldsymbol{\beta}), \dots, d_n(\boldsymbol{\beta})), \quad d_i(\boldsymbol{\beta}) := n_i p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))$$

$$\Rightarrow E(\mathbf{Z}^\beta) = \mathbf{X} \boldsymbol{\beta}; \quad \text{Cov}(\mathbf{Z}^\beta) = \mathbf{D}^{-1}(\boldsymbol{\beta}) \underbrace{\text{Cov}(\boldsymbol{\epsilon})}_{=\mathbf{D}(\boldsymbol{\beta})} \mathbf{D}^{-1}(\boldsymbol{\beta}) = \mathbf{D}^{-1}(\boldsymbol{\beta})$$

Recall from lec.2:

The next iteration β^{s+1} in the IWLS is given by

$$\beta^{s+1} = (X^t D(\beta^s) X)^{-1} X^t D(\beta^s) \mathbf{Z}^{\beta^s}$$

At convergence of the IWLS algorithm ($s \rightarrow \infty$), the estimate $\hat{\beta}$ satisfies:

$$\hat{\beta} = (X^t D(\hat{\beta}) X)^{-1} X^t D(\hat{\beta}) \mathbf{Z}^{\hat{\beta}}$$

Define

$$\mathbf{e} := \mathbf{Z}^{\hat{\beta}} - X\hat{\beta} = \left(I - X(X^t D(\hat{\beta}) X)^{-1} X^t D(\hat{\beta}) \right) \mathbf{Z}^{\hat{\beta}}$$

If one considers $D(\hat{\beta})$ as a **non random** constant quantity, then we have

$$E(\mathbf{e}) = \left(I - X(X^t D(\hat{\beta}) X)^{-1} X^t D(\hat{\beta}) \right) \underbrace{E(\mathbf{Z}^{\hat{\beta}})}_{= X\hat{\beta}} = \mathbf{0}$$

$$\begin{aligned} Var(\mathbf{e}) &= [I - X(X^t D(\hat{\beta}) X)^{-1} X^t D(\hat{\beta})] \underbrace{Cov(\mathbf{Z}^{\hat{\beta}})}_{= D^{-1}(\hat{\beta})} [I - X(X^t D(\hat{\beta}) X)^{-1} X^t D(\hat{\beta})] \\ &= D^{-1}(\hat{\beta}) - X(X^t D(\hat{\beta}) X)^{-1} X^t \end{aligned}$$

Additionally

$$\begin{aligned}\mathbf{e} &:= \mathbf{Z}^{\hat{\beta}} - X\hat{\beta} = D^{-1}(\hat{\beta})\mathbf{e}^r \\ \Rightarrow e_i^r &= D_i(\hat{\beta})e_i = n_i\hat{p}_i(1 - \hat{p}_i)e_i\end{aligned}$$

If one considers $n_i\hat{p}_i(1 - \hat{p}_i)$ as **nonrandom constant** we have

$$\text{Var}(e_i^r) = (n_i\hat{p}_i(1 - \hat{p}_i))^2 \text{Var}(e_i)$$

$$\begin{aligned}e_i^a &:= \frac{e_i^r}{\text{Var}(e_i^r)^{1/2}} \quad \text{“adjusted residuals”} \\ &= \frac{e_i^r}{\left\{ \{n_i\hat{p}_i(1 - \hat{p}_i)\}^2 \left[\frac{1}{n_i\hat{p}_i(1 - \hat{p}_i)} - (X(X^t D(\hat{\beta})X)^{-1}X^t)_{ii} \right] \right\}^{1/2}} \\ &= \frac{e_i^r}{\left\{ n_i\hat{p}_i(1 - \hat{p}_i) [1 - n_i\hat{p}_i(1 - \hat{p}_i)(X(X^t D(\hat{\beta})X)^{-1}X^t)_{ii}] \right\}^{1/2}} \\ &= \frac{e_i^P}{\left[1 - n_i\hat{p}_i(1 - \hat{p}_i)(X(X^t D(\hat{\beta})X)^{-1}X^t)_{ii} \right]^{1/2}} \\ &= \frac{e_i^P}{[1 - h_{ii}]^{1/2}} \quad \text{where } h_{ii} := n_i\hat{p}_i(1 - \hat{p}_i)(X(X^t D(\hat{\beta})X)^{-1}X^t)_{ii}\end{aligned}$$

High leverage and influential points in logistic regression

(Reference: Pregibon (1981))

Linear models:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y} \quad \hat{\mathbf{Y}} = \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{H} \mathbf{Y},$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \quad \mathbf{H}^2 = \mathbf{H}$$

$$\begin{aligned} \Rightarrow \hat{\mathbf{e}} &:= \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H}) \mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{Y} - \hat{\mathbf{Y}}) \quad \text{since } \mathbf{H} \hat{\mathbf{Y}} = \mathbf{H}(\mathbf{H} \mathbf{Y}) = \mathbf{H}^2 \mathbf{Y} = \mathbf{H} \mathbf{Y} = \hat{\mathbf{Y}} \\ &= (\mathbf{I} - \mathbf{H}) \hat{\mathbf{e}} \end{aligned}$$

$\Rightarrow$ raw residuals satisfy $\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H}) \hat{\mathbf{e}}$

logistic regression:

Define $H := \hat{D}^{1/2} X (X^t \hat{D} X)^{-1} X^t \hat{D}^{1/2}$ with $\hat{D} = D(\hat{\beta})$

$$\text{Lemma: } \mathbf{e}^P = (I - H)\mathbf{e}^P, \quad \text{where } e_i^P := \frac{Y_i - n_i \hat{p}_i}{(n_i \hat{p}_i (1 - \hat{p}_i))^{1/2}}$$

Proof: $D(\hat{\beta}) = \text{diag}(\dots n_i \hat{p}_i (1 - \hat{p}_i) \dots)$ nonsingular

$$\Rightarrow \mathbf{e}^P = \hat{D}^{-1/2}(\mathbf{Y} - \hat{\mathbf{Y}})$$

$$H\mathbf{e}^P = [\hat{D}^{1/2} X (X^t \hat{D} X)^{-1}] X^t (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0} \quad (*)$$

since $s(\hat{\beta}) = X^t (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}$

$$\Rightarrow \mathbf{e}^P = (I - H)\mathbf{e}^P \quad q.e.d.$$

Note that $H^2 = H$ as in linear models (exercise).

High leverage points in logistic regression

$$\mathbf{e}^P = \underbrace{(I - H)}_M \mathbf{e}^P$$

M spans residual space $\mathbf{e}^P$. This suggests that small m_{ii} (or large h_{ii}) should be useful in detecting extreme points in the design space X .

We have $\sum_{i=1}^n h_{ii} = p$ (exercise), therefore we consider $h_{ii} > \frac{2p}{n}$ as “high leverage points”.

Partial residual plot

Linear models:

Consider $X = [\mathbf{X}_j; X_{-j}]$,

$X_{-j} = X$ with j^{th} column removed, $X_{-j} \in \mathbb{R}^{n \times (p-1)}$.

$\mathbf{X}_j = (x_{1j}, \dots, x_{nj})^t$ – j^{th} column of matrix X .

$$\begin{aligned} \mathbf{e}_{\mathbf{Y}|\mathbf{X}_{-j}} &:= \mathbf{Y} - \underbrace{X_{-j}(X_{-j}^t X_{-j})^{-1} X_{-j}^t}_{=: H_{-j}} \mathbf{Y} = (I - H_{-j}) \mathbf{Y}, \\ &= \text{raw residuals in model with } j^{th} \text{ covariable removed} \end{aligned}$$

$$\begin{aligned} \mathbf{e}_{\mathbf{X}_j|\mathbf{X}_{-j}} &:= \mathbf{X}_j - X_{-j}(X_{-j}^t X_{-j})^{-1} X_{-j}^t \mathbf{X}_j = (I - H_{-j}) \mathbf{X}_j \\ &= \text{raw residuals in model } \mathbf{X}_j = X_{-j} \boldsymbol{\beta}_{-j}^* + \boldsymbol{\epsilon}_x \quad \boldsymbol{\epsilon}_x \sim N(0, \sigma_x^2) \text{ i.i.d.} \\ &= \text{measure of linear dependency of } \mathbf{X}_j \text{ on the remaining covariates} \end{aligned}$$

The **partial residual plot** is given by plotting $\mathbf{e}_{\mathbf{X}_j|\mathbf{X}_{-j}}$ versus $\mathbf{e}_{\mathbf{Y}|\mathbf{X}_{-j}}$.

$$\mathbf{Y} = \mathbf{X}_{-j}\boldsymbol{\beta}_{-j} + \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\epsilon} \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{-j} \\ \boldsymbol{\beta}_j \end{pmatrix}$$

$$\Rightarrow \underbrace{(\mathbf{I} - \mathbf{H}_{-j})\mathbf{Y}}_{\mathbf{e}_{\mathbf{Y}|\mathbf{X}_{-j}}} = \underbrace{(\mathbf{I} - \mathbf{H}_{-j})\mathbf{X}_{-j}}_{=0}\boldsymbol{\beta}_{-j} + \underbrace{(\mathbf{I} - \mathbf{H}_{-j})\mathbf{X}_j}_{\mathbf{e}_{\mathbf{X}_j|\mathbf{X}_{-j}}}\boldsymbol{\beta}_j + \underbrace{(\mathbf{I} - \mathbf{H}_{-j})\boldsymbol{\epsilon}}_{\boldsymbol{\epsilon}^* \text{ with } E(\boldsymbol{\epsilon}^*)=0},$$

since $(\mathbf{I} - \mathbf{X}_{-j}(\mathbf{X}_{-j}^t\mathbf{X}_{-j})^{-1}\mathbf{X}_{-j}^t)\mathbf{X}_{-j} = \mathbf{X}_{-j} - \mathbf{X}_{-j} = \mathbf{0}$

$$\Rightarrow \mathbf{e}_{\mathbf{Y}|\mathbf{X}_{-j}} = \boldsymbol{\beta}_j \mathbf{e}_{\mathbf{X}_j|\mathbf{X}_{-j}} + \boldsymbol{\epsilon}^* \quad \text{Model (*)}$$

The **LSE of $\boldsymbol{\beta}_j$ in (*)**, denoted by $\hat{\boldsymbol{\beta}}_j^*$ satisfies $\hat{\boldsymbol{\beta}}_j^* = \hat{\boldsymbol{\beta}}_j$ where $\hat{\boldsymbol{\beta}}_j$ is LSE in $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ (exercise).

Since $\hat{\boldsymbol{\beta}}_j^* = \hat{\boldsymbol{\beta}}_j$ we can **interpret the partial residual plot** as follows:

If partial residual plot scatters

- **around 0** $\Rightarrow \mathbf{X}_j$ has **no influence** on $\mathbf{Y}$
- **linear** $\Rightarrow \mathbf{X}_j$ should **be linear** in model
- **nonlinear** $\Rightarrow \mathbf{X}_j$ should **be included with this nonlinear form**.

Simpler plot: Plot $\mathbf{X}_j$ versus $\mathbf{e}_{\mathbf{Y}|\mathbf{X}_{-j}}$. Same behavior if $\mathbf{X}_j$ does not depend on other covariates.

This follows from the fact, that ML estimators of β_{-j} for two models with and without j^{th} covariate coincide, if $\mathbf{X}_j$ is orthogonal to $\mathbf{X}_{-j}$. Then

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}_{-j}\hat{\beta}_{-j} - \mathbf{X}_j\hat{\beta}_j$$

$$\Rightarrow \mathbf{e}_{\mathbf{Y}|\mathbf{X}_{-j}}(:= \mathbf{Y} - \mathbf{X}_{-j}\hat{\beta}_{-j}) = \hat{\beta}_j\mathbf{X}_j + \mathbf{e}, \quad (*)$$

where $\mathbf{e}$ is distributed around 0 and components of $\mathbf{e}$ are assumed nearly independent.

Partial residual plot

Logistic regression:

Landwehr, Pregibon, and Shoemaker (1984) propose to use

$$e_{(\mathbf{Y}|\mathbf{X}_{-j})_i} := \frac{Y_i - n_i \hat{p}_i}{n_i \hat{p}_i (1 - \hat{p}_i)} + \hat{\beta}_j x_{ij} \quad \text{as partial residual}$$

Heuristic derivation:

Justification: IWLS

$$\text{Recall: } \mathbf{Z}^{\hat{\beta}} = \mathbf{X} \hat{\beta} + \hat{D}^{-1} \mathbf{e}^r \quad \text{“obs. vector”}$$

$$\text{cov}(\hat{D}^{-1} \mathbf{e}^r) = \hat{D}^{-1} \Rightarrow \hat{\beta} = (\mathbf{X}^t \hat{D} \mathbf{X})^{-1} \mathbf{X}^t \hat{D} \mathbf{Z}^{\hat{\beta}}$$

Consider $\text{logit}(\mathbf{p}) = X\boldsymbol{\beta} + \mathbf{L}\boldsymbol{\gamma}$, $\mathbf{L} \in \mathbb{R}^n$ new covariable with $\mathbf{L} \perp \mathbf{X}$.

$$\begin{aligned}\mathbf{Z}_{\mathbf{L}} &:= X\hat{\boldsymbol{\beta}} + \mathbf{L}\hat{\boldsymbol{\gamma}} + \hat{D}_L^{-1}\mathbf{e}_{\mathbf{L}}^r \quad \text{with} \\ e_{Li}^r &:= y_i - n_i \underbrace{\frac{e^{\mathbf{x}_i^t \hat{\boldsymbol{\beta}} + l_i \hat{\boldsymbol{\gamma}}}}{1 + e^{\mathbf{x}_i^t \hat{\boldsymbol{\beta}} + l_i \hat{\boldsymbol{\gamma}}}}}_{\hat{p}_{Li}} \quad \mathbf{L} = (l_1, \dots, l_n)^t \\ \hat{D}_{\mathbf{L}} &:= \text{diag}(\dots, n_i \hat{p}_{Li}(1 - \hat{p}_{Li}), \dots)\end{aligned}$$

As in linear models (see (*)) partial residuals can be defined as

$$e_{(\mathbf{Y}|\mathbf{X}_{-j})_i} := (\hat{D}_L^{-1})_{ii} e_{Li}^r + \hat{\gamma} l_i = \frac{y_i - n_i \hat{p}_{Li}}{n_i \hat{p}_{Li}(1 - \hat{p}_{Li})} + \hat{\gamma} l_i$$

$\Rightarrow$ partial residual plot in logistic regression: plot x_{ij} versus $e_{(\mathbf{Y}|\mathbf{X}_{-j})_i}$.

For binary data we need to smooth data.

Cook's distance in linear models

$$\begin{aligned} D_i &:= \|X\hat{\beta} - X\hat{\beta}_{-i}\|^2 / p\hat{\sigma}^2 \\ &= (\hat{\beta} - \hat{\beta}_{-i})^t (X^t X) (\hat{\beta} - \hat{\beta}_{-i}) / p\hat{\sigma}^2 \\ &= \frac{\hat{e}_i^2 h_{ii}}{p\hat{\sigma}^2 (1 - h_{ii})^2} \end{aligned}$$

Measures change in confidence ellipsoid when i^{th} obs. is removed.

Cook's distance in logistic regression

Using LRT it can be shown, that

$\{\beta : -2 \ln \left\{ \frac{L(\beta)}{L(\hat{\beta})} \right\} \leq \chi_{1-\alpha, p}^2\}$ is an approx. $100(1 - \alpha)$ % CI for β

$\Rightarrow D_i := -2 \ln \left\{ \frac{L(\hat{\beta}_{-i})}{L(\hat{\beta})} \right\}$ measures change when i^{th} obs. removed; difficult to calculate. Using Taylor expansion we have:

$$\{\beta : -2 \ln \left\{ \frac{L(\beta)}{L(\hat{\beta})} \right\} \leq \chi_{1-\alpha, p}^2\} \approx \{\beta : (\beta - \hat{\beta})^t X^t \hat{D} X (\beta - \hat{\beta}) \leq \chi_{1-\alpha, p}^2\}$$

$$\Rightarrow D_i \approx (\hat{\beta}_{-i} - \hat{\beta})^t X^t \hat{D} X (\hat{\beta}_{-i} - \hat{\beta})$$

Approximate $\hat{\beta}_{-i}$ by a single step Newton Rapson starting from $\hat{\beta}$:

$$\Rightarrow \hat{\beta}_{-i} \approx \hat{\beta} - \frac{(X^t \hat{D} X)^{-1} \mathbf{x}_i (Y_i - n_i \hat{p}_i)}{1 - h_{ii}} \quad (\text{exercise})$$

where $h_{ii} = n_i \hat{p}_i (1 - \hat{p}_i) \{X (X^t \hat{D} X)^{-1} X^t\}_{ii}$

$$\Rightarrow D_i \approx \frac{(e_i^a)^2 h_{ii}}{(1 - h_{ii})} = (e_i^P)^2 \frac{h_{ii}}{(1 - h_{ii})^2} =: D_i^a$$

where $e_i^a = \frac{e_i^P}{(1 - h_{ii})^{1/2}}, \quad e_i^P := \frac{Y_i - n_i \hat{p}_i}{(n_i \hat{p}_i (1 - \hat{p}_i))^{1/2}}$

In general D_i^a underestimates D_i , but shows influential observations.

References

- Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* 13, 342–368.
- Landwehr, J. M., D. Pregibon, and A. C. Shoemaker (1984). Graphical methods for assessing logistic regression models. *JASA* 79, 61–83.
- Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics* 9, 705–724.