Lecture 4: Parameter estimation and diagnostics in logistic regression

Claudia Czado

TU München

Overview

- Parameter estimation
- Regression diagnostics

Parameter estimation in logistic regression

loglikelihood:

$$l(\boldsymbol{\beta}) := \sum_{i=1}^{n} \left[(y_i \ln \left(\frac{e^{\boldsymbol{x_i^t \beta}}}{1+e^{\boldsymbol{x_i^t \beta}}} \right) + (n_i - y_i) \ln \left(1 - \frac{e^{\boldsymbol{x_i^t \beta}}}{1+e^{\boldsymbol{x_i^t \beta}}} \right) \right] + \begin{array}{r} const\ ind. \\ of\ \boldsymbol{\beta} \end{array}$$

$$= \sum_{i=1}^{n} \left[(y_i \boldsymbol{x_i^t \beta}) - n_i \ln(1 + e^{\boldsymbol{x_i^t \beta}}) \right] + const$$

scores:

$$s_j(\boldsymbol{\beta}) := \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - n_i \frac{e^{\boldsymbol{x_i^t \beta}}}{1+e^{\boldsymbol{x_i^t \beta}}} x_{ij} = \sum_{i=1}^{n} x_{ij} \left(y_i - \underbrace{n_i \frac{e^{\boldsymbol{x_i^t \beta}}}{1 + e^{\boldsymbol{x_i^t \beta}}}}_{E(Y_i | \mathbf{X_i = x_i})} \right) \qquad j = 1, \ldots, p$$

$$\Rightarrow \mathbf{s}(\boldsymbol{\beta}) = X^t \left(\mathbf{Y} - E(\mathbf{Y} | \mathbf{X} = \mathbf{x}) \right)$$

Hessian matrix in logistic regression

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

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

where $D = diag(d_1, \ldots, d_n)$ 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 $l(\boldsymbol{\beta})$ in logistic regression is strict concave in $\boldsymbol{\beta}$ if $rank(X) = p$*

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

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

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

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

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

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

$\Rightarrow$ There is at most one solution to the score equations, i.e. if the MLE of $\boldsymbol{\beta}$ 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 $\exists \boldsymbol{\beta}^* \in \mathbb{R}^p$ with

$$\begin{aligned} \boldsymbol{x_i^t}\boldsymbol{\beta}^* &> 0 \quad \text{if} \quad Y_i = 1 \\ \boldsymbol{x_i^t}\boldsymbol{\beta}^* &\leq 0 \quad \text{if} \quad Y_i = 0 \end{aligned}$$

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

Consider $\alpha\boldsymbol{\beta}^*$ for $\alpha > 0 \to \infty$

$$\Rightarrow l(\alpha\boldsymbol{\beta}^*) = \sum_{i=1, Y_i=1}^{n} \{\alpha \boldsymbol{x_i^t}\boldsymbol{\beta}^* - \ln(1 + e^{\alpha \boldsymbol{x_i^t}\boldsymbol{\beta}^*})\} - \sum_{i=1, Y_i=0}^{n} \ln(1 + e^{\alpha \boldsymbol{x_i^t}\boldsymbol{\beta}^*}) \to 0$$

for $\alpha \to \infty$.

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

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

Asymptotic theory

(Reference: Fahrmeir and Kaufmann (1985))

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

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

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

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

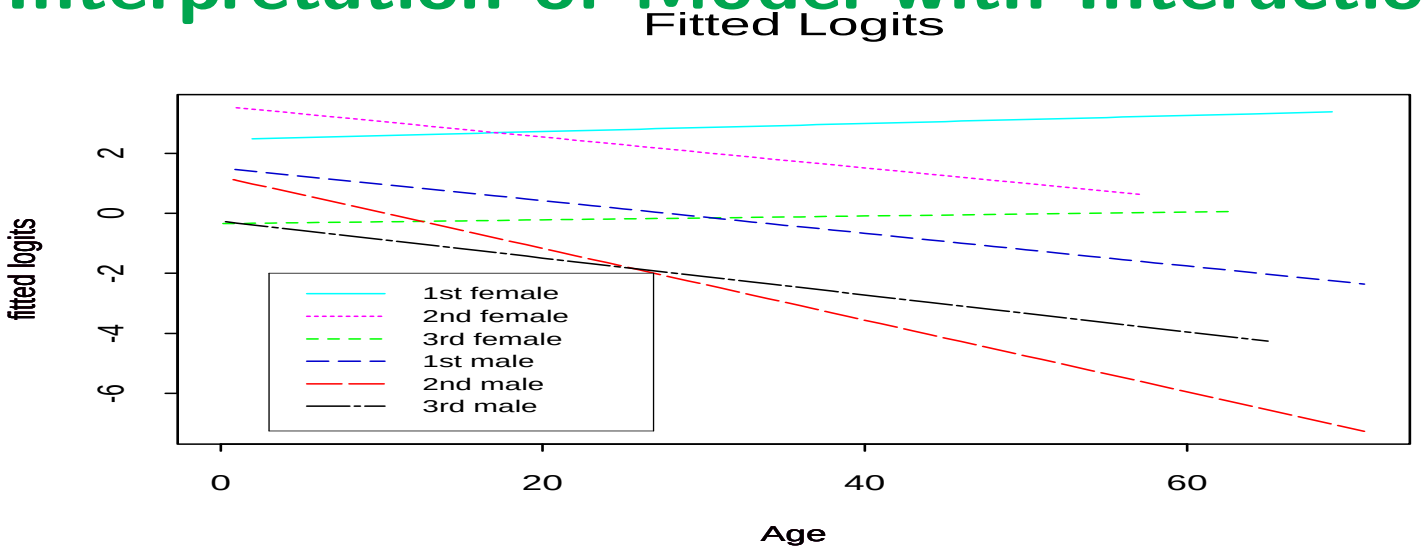
With Interaction Effects:

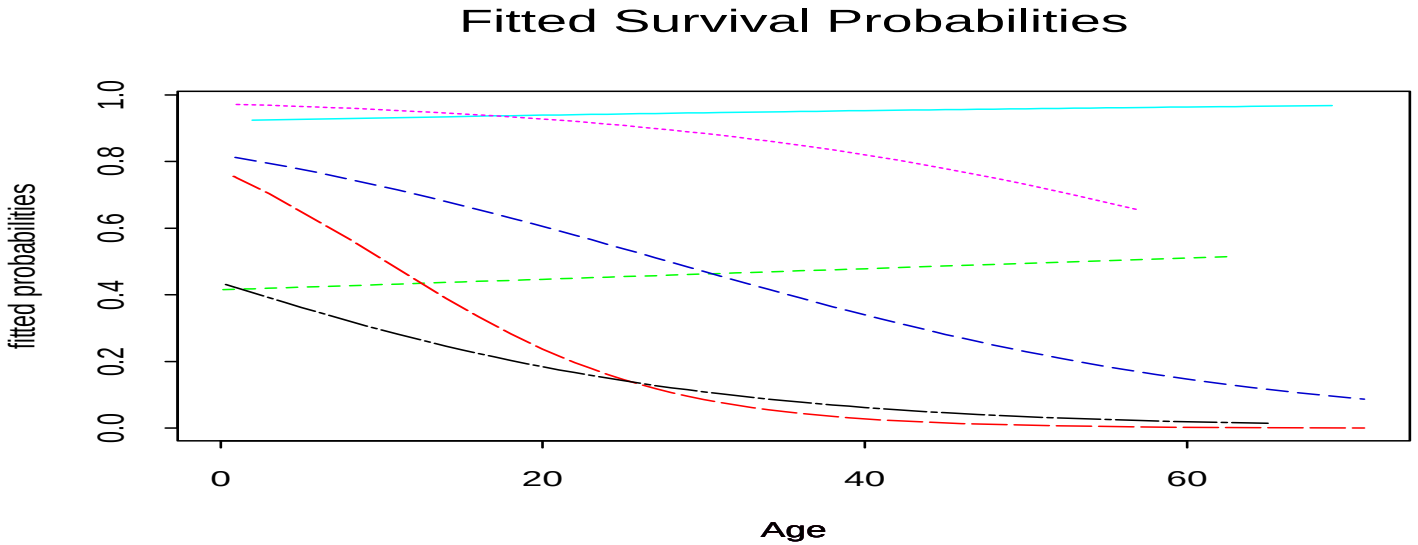
```
> f.group.inter_cbind(Survived, Not.Survived)
   ~(Age+Sex+PClass)^2
> r.group.inter_glm(f.group.inter,family=
  binomial,na.action=na.omit)
> summary(r.group.inter)[[3]]
                Value Std. Error t value
 (Intercept)  2.464      0.843    2.92
         Age  0.013      0.020    0.67
         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:





Diagnostics in logistic regression

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

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

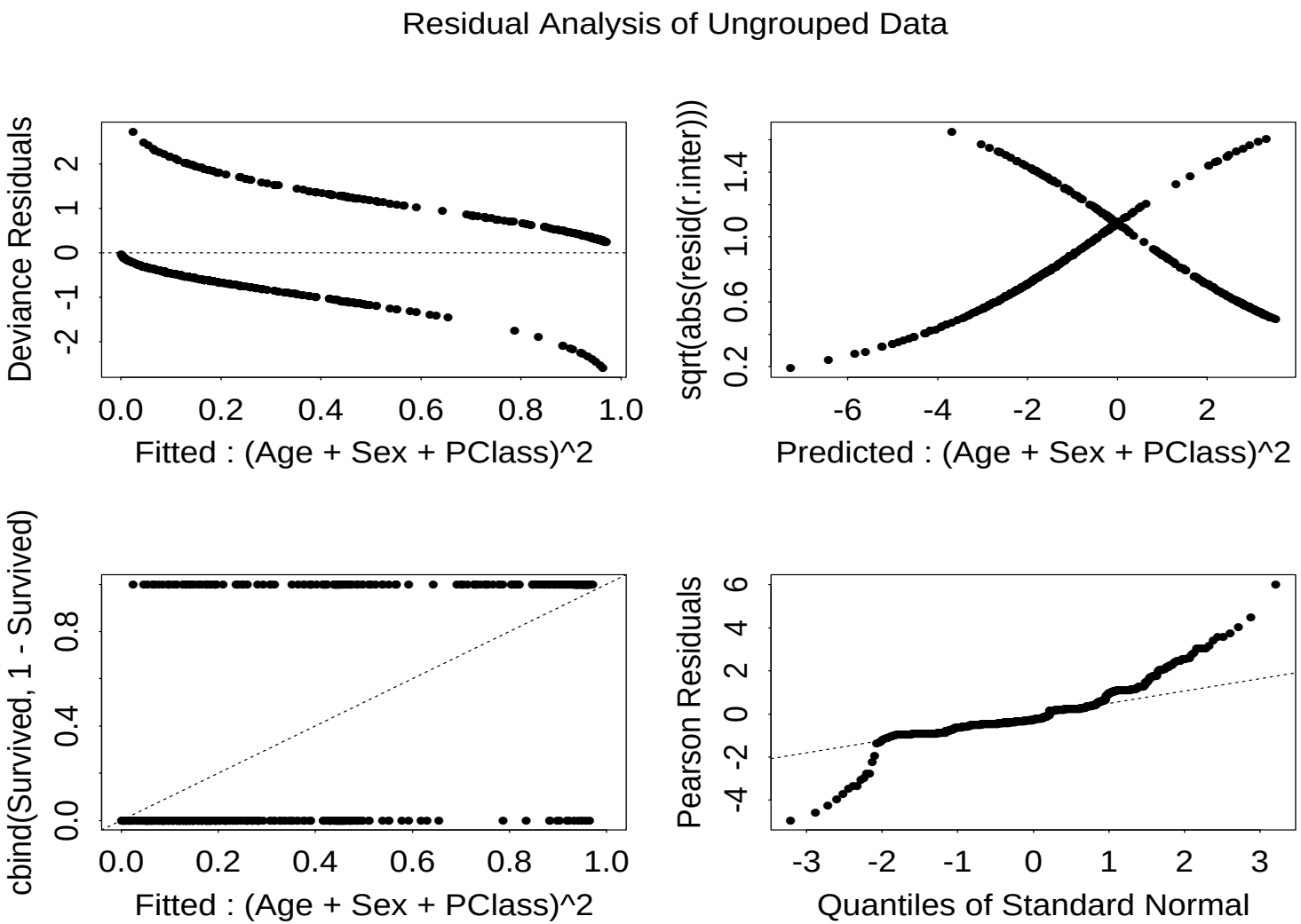
Pearson residuals: $e_i^P := \frac{Y_i - n_i\hat{p}_i}{(n_i\hat{p}_i(1-\hat{p}_i))^{1/2}}$

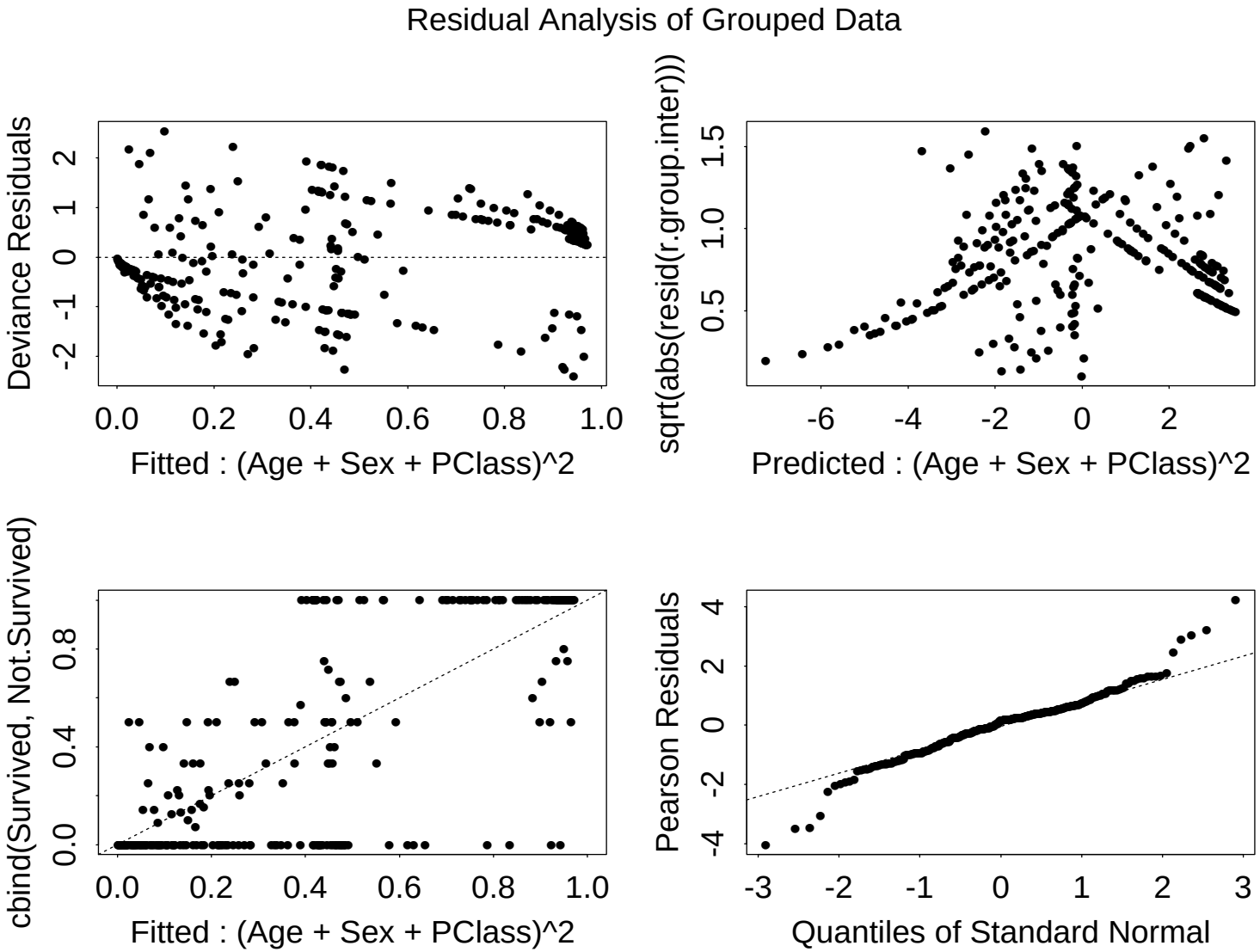
$$\Rightarrow \chi^2 = \sum_{i=1}^{n}(e_i^P)^2$$

deviance residuals: $e_i^d := sign(Y_i - n_i\hat{p}_i)(2y_i \ln\left(\frac{Y_i}{n_i\hat{p}_i}\right) + 2(n_i - y_i)\ln\left(\frac{n_i - Y_i}{n_i(1-\hat{p}_i)}\right))$

$$\Rightarrow D = \sum_{i=1}^{n}(r_i^D)^2$$

Residual Analysis:




Residual Analysis of Grouped Data
Deviance Residuals
-2
-1
0
1
2
0.0
0.2
0.4
0.6
0.8
1.0
Fitted : (Age + Sex + PClass)^2
sqrt(abs(resid(r.group.inter)))
0.5
1.0
1.5
-6
-4
-2
0
2
Predicted : (Age + Sex + PClass)^2
cbind(Survived, Not.Survived)
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Fitted : (Age + Sex + PClass)^2
Pearson Residuals
-4
-2
0
2
4
-3
-2
-1
0
1
2
3
Quantiles of Standard Normal

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

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

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

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

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

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

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

$D(\boldsymbol{\beta}) = diag(d_1(\boldsymbol{\beta}), \ldots, d_n(\boldsymbol{\beta})), \quad d_i(\boldsymbol{\beta}) := n_i p(\mathbf{x_i})(1 - p(\mathbf{x_i}))$

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

Recall from lec.2:
The next iteration $\boldsymbol{\beta}^{s+1}$ in the IWLS is given by

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

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

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

Define

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

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

$$\begin{aligned} E(\mathbf{e}) &= \left(I - X(X^t D(\hat{\boldsymbol{\beta}}) X)^{-1} X^t D(\hat{\boldsymbol{\beta}}) \right) \underbrace{E(\mathbf{Z}^{\hat{\beta}})}_{= X\hat{\beta}} = \mathbf{0} \\ Var(\mathbf{e}) &= [I - X(X^t D(\hat{\boldsymbol{\beta}}) X)^{-1} X^t D(\hat{\boldsymbol{\beta}})] \underbrace{Cov(\mathbf{Z}^{\hat{\beta}})}_{= D^{-1}(\hat{\beta})} [I - X(X^t D(\hat{\boldsymbol{\beta}}) X)^{-1} X^t D(\hat{\boldsymbol{\beta}})] \\ &= D^{-1}(\hat{\boldsymbol{\beta}}) - X(X^t D(\hat{\boldsymbol{\beta}}) X)^{-1} X^t \end{aligned}$$

Additionaly

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

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

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

$$e_i^a := \frac{e_i^r}{Var(e_i^r)^{1/2}} \qquad \text{“adjusted residuals”}$$

$$= \frac{e_i^r}{\left\{\{n_i\hat{p}_i(1-\hat{p}_i)\}^2[\frac{1}{n_i\hat{p}_i(1-\hat{p}_i)} - (X(X^tD(\hat{\boldsymbol{\beta}})X)^{-1}X^t)_{ii}]\right\}^{1/2}}$$

$$= \frac{e_i^r}{\left\{n_i\hat{p}_i(1-\hat{p}_i)[1 - n_i\hat{p}_i(1-\hat{p}_i)(X(X^tD(\hat{\boldsymbol{\beta}})X)^{-1}X^t)_{ii}]\right\}^{1/2}}$$

$$= \frac{e_i^P}{\left[1 - n_i\hat{p}_i(1-\hat{p}_i)(X(X^tD(\hat{\boldsymbol{\beta}})X)^{-1}X^t)_{ii}\right]^{1/2}}$$

$$= \frac{e_i^P}{[1-h_{ii}]^{1/2}} \quad \text{where } h_{ii} := n_i\hat{p}_i(1-\hat{p}_i)(X(X^tD(\hat{\boldsymbol{\beta}})X)^{-1}X^t)_{ii}$$

High leverage and influential points in logistic regression

(Reference: Pregibon (1981))

Linear models:

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

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

$$\Rightarrow \quad \hat{\mathbf{e}} := \mathbf{Y} - \hat{\mathbf{Y}} = (I - H)\mathbf{Y}$$

$$= (I - H)(\mathbf{Y} - \hat{\mathbf{Y}}) \quad \text{since } H\hat{\mathbf{Y}} = H(H\mathbf{Y}) = H^2\mathbf{Y} = H\mathbf{Y} = \hat{\mathbf{Y}}$$

$$= (I - H)\hat{\mathbf{e}}$$

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

logistic regression:

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

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

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

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

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

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

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

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, $X_{-j} \in \mathbb{R}^{n \times (p-1)}$.
$\mathbf{X_j} = (x_{1j}, \ldots, x_{nj})^t - j^{th}$ column of matrix X.

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

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

The partial residual plot is given by plotting $\mathbf{e_{X_j|X_{-j}}}$ versus $\mathbf{e_{Y|X_{-j}}}$.

$$\mathbf{Y} = X_{-j}\boldsymbol{\beta}_{-j} + \mathbf{X_j}\beta_j + \boldsymbol{\epsilon} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{-j} \\ \beta_j \end{pmatrix}$$

$$\Rightarrow \underbrace{(I - H_{-j})\mathbf{Y}}_{\mathbf{e_{Y|X_{-j}}}} = \underbrace{(I - H_{-j})X_{-j}}_{=\mathbf{0}}\boldsymbol{\beta}_{-j} + \underbrace{(I - H_{-j})\mathbf{X_j}}_{\mathbf{e_{X_j|X_{-j}}}}\beta_j + \underbrace{(I - H_{-j})\boldsymbol{\epsilon}}_{\boldsymbol{\epsilon}^* \text{ with } E(\boldsymbol{\epsilon}^*)=\mathbf{0}},$$

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

$$\Rightarrow \mathbf{e_{Y|X_{-j}}} = \beta_j\, \mathbf{e_{X_j|X_{-j}}} + \boldsymbol{\epsilon}^* \qquad \text{Model (*)}$$

The LSE of β_j in (*), denoted by $\hat{\beta}_j^*$ satisfies $\hat{\beta}_j^* = \hat{\beta}_j$ where $\hat{\beta}_j$ is LSE in $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ (exercise).
Since $\hat{\beta}_j^* = \hat{\beta}_j$ 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 $\mathbf{e_{Y|X_{-j}}}$. Same behavior if $\mathbf{X_j}$ does not depend on other covariates.
This follows from the fact, that ML estimators of $\boldsymbol{\beta}_{-j}$ for two models with and without j^{th} covariate coincide, if $\mathbf{X_j}$ is orthogonal to X_{-j}. Then

$$\mathbf{e} = \mathbf{Y} - X_{-j}\hat{\boldsymbol{\beta}}_{-j} - \mathbf{X_j}\hat{\beta}_j$$

$$\Rightarrow \quad \mathbf{e_{Y|X_{-j}}}(:= \mathbf{Y} - X_{-j}\hat{\boldsymbol{\beta}}_{-j}) = \hat{\beta}_j\mathbf{X_j} + \mathbf{e}, \qquad (*)$$

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

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

Heuristic derivation:
Justification: IWLS

Recall: $\mathbf{Z}^{\hat{\beta}} = X\hat{\boldsymbol{\beta}} + \hat{D}^{-1}\mathbf{e}^{\mathbf{r}}$ "obs. vector"

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

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

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

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

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

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

For binary data we need to smooth data.

Cook's distance in linear models

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

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

Cook's distance in logistic regression

Using LRT it can be shown, that

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

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

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

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

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

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

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

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

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

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.