Lecture 4: Parameter estimation and diagnostics in logistic regression

Claudia Czado

TU München
Overview

- Parameter estimation
- Regression diagnostics
Parameter estimation in logistic regression

loglikelihood:

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

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

scores:

\[ s_j(\beta) := \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - n_i \frac{e^{x_i^t \beta}}{1+e^{x_i^t \beta}} x_{ij} = \sum_{i=1}^{n} x_{ij} \begin{pmatrix} y_i - n_i \frac{e^{x_i^t \beta}}{1+e^{x_i^t \beta}} \\ \frac{E(Y_i|X_i=x_i)}{E(Y|X=x)} \end{pmatrix} \quad j = 1, \ldots, p \]

\[ \Rightarrow s(\beta) = X^t \left( Y - E(Y|X=x) \right) \]
Hessian matrix in logistic regression

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

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

where \(D = \text{diag}(d_1, \ldots, d_n)\) and \(d_i := n_ip(x_i)(1 - p(x_i))\).

\(H(\beta)\) independent of \(Y\) (since canonical link) \(\Rightarrow E(H(\beta)) = H(\beta)\)
Existence of MLE’s in logistic regression

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

Proof: \( H(\beta) = -X^tDX \)

\[
\Rightarrow Z^tH(\beta)Z = -Z^tX^tD^{1/2}D^{1/2}XZ = -\|D^{1/2}XZ\|^2
\]

\[
\Rightarrow \|D^{1/2}XZ\|^2 = 0 \iff D^{1/2}XZ = 0
\]

\[
\iff \begin{cases} X^tD^{1/2}D^{1/2}XZ = 0 \\
D^{1/2}X \text{ full rank} \Rightarrow Z = (X^tDX)^{-1}0 
\end{cases}
\]

\[
\iff Z = 0 \quad q.e.d.
\]

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

\[
\begin{align*}
    x_i^t \beta^* &> 0 \quad \text{if } Y_i = 1 \\
    x_i^t \beta^* &\leq 0 \quad \text{if } Y_i = 0
\end{align*}
\]

\[\Rightarrow l(\beta^*) = \sum_{i=1}^{n} \left[ y_i x_i^t \beta^* - \ln(1 + e^{x_i^t \beta^*}) \right] = \sum_{i=1}^{n} \left[ x_i^t \beta^* - \ln(1 + e^{x_i^t \beta^*}) \right] - \sum_{i=1, Y_i = 0}^{n} \ln(1 + e^{x_i^t \beta^*}) \]

Consider \( \alpha \beta^* \) for \( \alpha > 0 \to \infty \)

\[\Rightarrow l(\alpha \beta^*) = \sum_{i=1, Y_i = 1}^{n} \left\{ \alpha x_i^t \beta^* - \ln(1 + e^{\alpha x_i^t \beta^*}) \right\} - \sum_{i=1, Y_i = 0}^{n} \ln(1 + e^{\alpha x_i^t \beta^*}) \to 0 \]
for \( \alpha \to \infty \).

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

Therefore we found \( \alpha \beta^* \) such that \( l(\alpha \beta^*) \to 0 \Rightarrow \) 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:

```r
> 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

<table>
<thead>
<tr>
<th></th>
<th>poly(Age, 2)1</th>
<th>poly(Age, 2)2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.41</td>
<td>0.07</td>
</tr>
<tr>
<td>poly(Age, 2)1</td>
<td>-0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>poly(Age, 2)2</td>
<td>0.07</td>
<td>-0.02</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.66</td>
<td>0.11</td>
</tr>
<tr>
<td>PClass2nd</td>
<td>-0.66</td>
<td>0.41</td>
</tr>
<tr>
<td>PClass3rd</td>
<td>-0.76</td>
<td>0.52</td>
</tr>
</tbody>
</table>

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

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-0.0036</td>
<td>-0.026</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-0.0725</td>
<td>0.100</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>-0.0010</td>
<td>-0.027</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>-0.0138</td>
<td>-0.019</td>
</tr>
</tbody>
</table>

PClass2nd PClass3rd

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Analysis of Deviance:

> anova(r.main)
Analysis of Deviance Table

Binomial model

Response: cbind(Survived, 1 - Survived)

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td>755</td>
<td>1026</td>
<td></td>
</tr>
<tr>
<td>poly(Age, 2)</td>
<td>2</td>
<td>12</td>
<td>753</td>
<td>1013</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>225</td>
<td>752</td>
<td>789</td>
</tr>
<tr>
<td>PClass</td>
<td>2</td>
<td>95</td>
<td>750</td>
<td>693</td>
</tr>
</tbody>
</table>
With Interaction Effects:

Linear Age Effect

```r
> 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]]

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.464</td>
<td>0.835</td>
<td>2.95</td>
</tr>
<tr>
<td>Age</td>
<td>0.013</td>
<td>0.020</td>
<td>0.67</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.946</td>
<td>0.824</td>
<td>-1.15</td>
</tr>
<tr>
<td>PClass2nd</td>
<td>1.116</td>
<td>1.002</td>
<td>1.11</td>
</tr>
<tr>
<td>PClass3rd</td>
<td>-2.807</td>
<td>0.825</td>
<td>-3.40</td>
</tr>
<tr>
<td>Age:Sex</td>
<td>-0.068</td>
<td>0.018</td>
<td>-3.70</td>
</tr>
<tr>
<td>AgePClass2nd</td>
<td>-0.065</td>
<td>0.024</td>
<td>-2.67</td>
</tr>
<tr>
<td>AgePClass3rd</td>
<td>-0.007</td>
<td>0.020</td>
<td>-0.35</td>
</tr>
<tr>
<td>SexPClass2nd</td>
<td>-1.411</td>
<td>0.715</td>
<td>-1.97</td>
</tr>
<tr>
<td>SexPClass3rd</td>
<td>1.032</td>
<td>0.616</td>
<td>1.67</td>
</tr>
</tbody>
</table>
```
> anova(r.inter)

Analysis of Deviance Table

Binomial model

Response: cbind(Survived, 1 - Survived)

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td></td>
<td></td>
<td>1026</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>3</td>
<td>754</td>
<td>1023</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>227</td>
<td>753</td>
<td>796</td>
</tr>
<tr>
<td>PClass</td>
<td>2</td>
<td>100</td>
<td>751</td>
<td>695</td>
</tr>
<tr>
<td>Age:Sex</td>
<td>1</td>
<td>28</td>
<td>750</td>
<td>667</td>
</tr>
<tr>
<td>Age:PClass</td>
<td>2</td>
<td>5</td>
<td>748</td>
<td>662</td>
</tr>
<tr>
<td>Sex:PClass</td>
<td>2</td>
<td>21</td>
<td>746</td>
<td>641</td>
</tr>
</tbody>
</table>

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:

\[ \text{Age.poly1} = \text{poly}(\text{Age}, 2)[,1] \]
\[ \text{Age.poly2} = \text{poly}(\text{Age}, 2)[,2] \]
\[ f \text{.inter1}_\text{cbind}(\text{Survived}, 1 - \text{Survived}) \sim \text{Sex} + \text{PClass} + \text{Age.poly1} + \text{Age.poly2} + \text{Sex} \times \text{Age.poly1} + \text{Sex} \times \text{PClass} + \text{Age.poly1} \times \text{PClass} + \text{Age.poly2} \times \text{PClass} \]
\[ r \text{.inter1}_\text{glm}(f \text{.inter1}, \text{family=binomial}, \text{na.action}=\text{na.omit}) \]
\[ \text{summary}(r \text{.inter1})[[3]] \]

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.92</td>
<td>0.47</td>
<td>6.27</td>
</tr>
<tr>
<td>Sex</td>
<td>-3.12</td>
<td>0.51</td>
<td>-6.14</td>
</tr>
<tr>
<td>PClass2nd</td>
<td>-0.72</td>
<td>0.58</td>
<td>-1.24</td>
</tr>
<tr>
<td>PClass3rd</td>
<td>-3.02</td>
<td>0.53</td>
<td>-5.65</td>
</tr>
<tr>
<td>Age.poly1</td>
<td>3.28</td>
<td>7.96</td>
<td>0.41</td>
</tr>
<tr>
<td>Age.poly2</td>
<td>-4.04</td>
<td>5.63</td>
<td>-0.72</td>
</tr>
<tr>
<td>Sex:Age.poly1</td>
<td>-21.94</td>
<td>7.10</td>
<td>-3.09</td>
</tr>
<tr>
<td>Sex:PClass2nd</td>
<td>-1.23</td>
<td>0.71</td>
<td>-1.74</td>
</tr>
<tr>
<td>Sex:PClass3rd</td>
<td>1.27</td>
<td>0.62</td>
<td>2.04</td>
</tr>
<tr>
<td>PClass2nd:Age.poly1</td>
<td>-14.02</td>
<td>10.39</td>
<td>-1.35</td>
</tr>
<tr>
<td>PClass3rd:Age.poly1</td>
<td>3.72</td>
<td>9.34</td>
<td>0.40</td>
</tr>
<tr>
<td>PClass2nd:Age.poly2</td>
<td>23.07</td>
<td>9.50</td>
<td>2.43</td>
</tr>
<tr>
<td>PClass3rd:Age.poly2</td>
<td>10.99</td>
<td>7.85</td>
<td>1.40</td>
</tr>
</tbody>
</table>
> anova(r.inter1)
Analysis of Deviance Table

Binomial model
Response: cbind(Survived, 1 - Survived)

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td>755</td>
<td>1026</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>229</td>
<td>754</td>
<td>797</td>
</tr>
<tr>
<td>PClass</td>
<td>2</td>
<td>73</td>
<td>752</td>
<td>724</td>
</tr>
<tr>
<td>Age.poly1</td>
<td>1</td>
<td>28</td>
<td>751</td>
<td>695</td>
</tr>
<tr>
<td>Age.poly2</td>
<td>1</td>
<td>2</td>
<td>750</td>
<td>693</td>
</tr>
<tr>
<td>Sex:Age.poly1</td>
<td>1</td>
<td>29</td>
<td>749</td>
<td>664</td>
</tr>
<tr>
<td>Sex:PClass</td>
<td>2</td>
<td>17</td>
<td>747</td>
<td>646</td>
</tr>
<tr>
<td>Age.poly1:PClass</td>
<td>2</td>
<td>7</td>
<td>745</td>
<td>639</td>
</tr>
<tr>
<td>Age.poly2:PClass</td>
<td>2</td>
<td>6</td>
<td>743</td>
<td>634</td>
</tr>
</tbody>
</table>

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?

```r
> 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)

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td>755</td>
<td>1026</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>1023</td>
<td>754</td>
<td>1023</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>796</td>
<td>753</td>
<td>796</td>
</tr>
<tr>
<td>PClass</td>
<td>2</td>
<td>695</td>
<td>751</td>
<td>695</td>
</tr>
<tr>
<td>Age:Sex</td>
<td>1</td>
<td>667</td>
<td>750</td>
<td>667</td>
</tr>
<tr>
<td>Age:PClass</td>
<td>2</td>
<td>662</td>
<td>748</td>
<td>662</td>
</tr>
<tr>
<td>Sex:PClass</td>
<td>2</td>
<td>641</td>
<td>746</td>
<td>641</td>
</tr>
<tr>
<td>Age:Sex:PClass</td>
<td>2</td>
<td>640</td>
<td>744</td>
<td>640</td>
</tr>
</tbody>
</table>

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:

```r
> 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]]

<table>
<thead>
<tr>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.5</td>
<td>0.24</td>
</tr>
<tr>
<td>poly(Age, 2)1</td>
<td>-11.1</td>
<td>2.17</td>
</tr>
<tr>
<td>poly(Age, 2)2</td>
<td>2.7</td>
<td>1.90</td>
</tr>
<tr>
<td>Sex</td>
<td>-2.6</td>
<td>0.20</td>
</tr>
<tr>
<td>PClass2st</td>
<td>-1.2</td>
<td>0.26</td>
</tr>
<tr>
<td>PClass3st</td>
<td>-2.5</td>
<td>0.28</td>
</tr>
</tbody>
</table>
```
> anova(r.group)
Analysis of Deviance Table

Binomial model

Response: cbind(Survived, Not.Survived)

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>273</td>
<td>645</td>
<td></td>
</tr>
<tr>
<td>poly(Age, 2)</td>
<td>2</td>
<td>12</td>
<td>271</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>225</td>
<td>270</td>
</tr>
<tr>
<td>PClass</td>
<td>2</td>
<td>95</td>
<td>268</td>
</tr>
</tbody>
</table>

> 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:

```r
> 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]]

<table>
<thead>
<tr>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.464</td>
<td>2.92</td>
</tr>
<tr>
<td>Age</td>
<td>0.013</td>
<td>0.67</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.947</td>
<td>-1.14</td>
</tr>
<tr>
<td>PClass2st</td>
<td>1.117</td>
<td>1.11</td>
</tr>
<tr>
<td>PClass3st</td>
<td>-2.807</td>
<td>-3.38</td>
</tr>
<tr>
<td>Age:Sex</td>
<td>-0.065</td>
<td>-3.68</td>
</tr>
<tr>
<td>AgePClass2st</td>
<td>-0.065</td>
<td>-2.65</td>
</tr>
<tr>
<td>AgePClass3st</td>
<td>-0.007</td>
<td>-0.35</td>
</tr>
<tr>
<td>SexPClass2st</td>
<td>-1.410</td>
<td>-1.95</td>
</tr>
<tr>
<td>SexPClass3st</td>
<td>1.033</td>
<td>1.66</td>
</tr>
</tbody>
</table>
```
> anova(r.group.inter)
Analysis of Deviance Table

Binomial model

Response: cbind(Survived, Not.Survived)

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td></td>
<td></td>
<td>273</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>3</td>
<td>272</td>
<td>642</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>227</td>
<td>271</td>
<td>415</td>
</tr>
<tr>
<td>PClass</td>
<td>2</td>
<td>100</td>
<td>269</td>
<td>314</td>
</tr>
<tr>
<td>Age:Sex</td>
<td>1</td>
<td>28</td>
<td>268</td>
<td>286</td>
</tr>
<tr>
<td>Age:PClass</td>
<td>2</td>
<td>5</td>
<td>266</td>
<td>281</td>
</tr>
<tr>
<td>Sex:PClass</td>
<td>2</td>
<td>21</td>
<td>264</td>
<td>260</td>
</tr>
</tbody>
</table>

> 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:
  \[ e_i^d := \text{sign}(Y_i - n_i \hat{p}_i) (2y_i \ln \left( \frac{Y_i}{n_i \hat{p}_i} \right) + 2(n_i - y_i) \ln \left( \frac{n_i - Y_i}{n_i (1 - \hat{p}_i)} \right)) \]
  \[ \Rightarrow D = \sum_{i=1}^{n} (r_i^D)^2 \]
Residual Analysis:

Residual Analysis of Ungrouped Data

![Deviance Residuals](image1)

Fitted: \((\text{Age} + \text{Sex} + \text{PClass})^2\)

Predicted: \((\text{Age} + \text{Sex} + \text{PClass})^2\)

![Pearson Residuals](image2)

Fitted: \((\text{Age} + \text{Sex} + \text{PClass})^2\)

Quantiles of Standard Normal
Residual Analysis of Grouped Data

- Deviance Residuals
  - Fitted: \((\text{Age} + \text{Sex} + \text{PClass})^2\)
  - Predicted: \((\text{Age} + \text{Sex} + \text{PClass})^2\)

- Pearson Residuals
  - Fitted: \((\text{Age} + \text{Sex} + \text{PClass})^2\)
  - 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 \textit{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 \( \eta_i = \mathbf{x}_i^t \beta = \ln \left( \frac{p_i}{1-p_i} \right) = \ln \left( \frac{n_ip_i}{n_i-n_ip_i} \right) = \ln \left( \frac{\mu_i}{n_i-\mu_i} \right) \)

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

\[\Rightarrow Z_i^\beta = \mathbf{x}_i^t \beta + (y_i - n_ip_i) \frac{1}{n_ip_i(1-p_i)} \] in logistic regression

\[\Rightarrow \mathbf{Z}^\beta = \mathbf{X} \beta + D^{-1}(\beta) \epsilon \] where

\[\epsilon := \mathbf{Y} - (n_1p_1, \ldots, n_n p_n)^t \]

\[D(\beta) = \text{diag}(d_1(\beta), \ldots, d_n(\beta)), \quad d_i(\beta) := n_ip(\mathbf{x}_i)(1-p(\mathbf{x}_i)) \]

\[\Rightarrow E(\mathbf{Z}^\beta) = \mathbf{X} \beta; \quad \text{Cov}(\mathbf{Z}^\beta) = D^{-1}(\beta) \text{Cov}(\epsilon) D^{-1}(\beta) = D^{-1}(\beta) = D(\beta) \]
Recall from lec.2:
The next iteration $\beta^{s+1}$ in the IWLS is given by

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

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

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

Define

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

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

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

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

$$
= D^{-1}(\hat{\beta}) - X (X^t D(\hat{\beta}) X)^{-1} X^t
$$
Additionally
\[
e := Z\hat{\beta} - X\hat{\beta} = D^{-1}(\hat{\beta})e^r
\]

\[
\Rightarrow e^r_i = D_i(\hat{\beta})e_i = n_i\hat{p}_i(1 - \hat{p}_i)e_i
\]

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

\[
Var(e^r_i) = (n_i\hat{p}_i(1 - \hat{p}_i))^2 Var(e_i)
\]

\[
e^a_i := \frac{e^r_i}{Var(e^r_i)^{1/2}} \quad \text{“adjusted residuals”}
\]

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

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

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

\[
= \frac{e_i^P}{\left[1 - h_{ii}\right]^{1/2}} \quad \text{where} \quad h_{ii} := n_i\hat{p}_i(1 - \hat{p}_i)(X(X^tD(\hat{\beta})X)^{-1}X^t)_{ii}
\]
High leverage and influential points in logistic regression

(Reference: Pregibon (1981))

Linear models:

\[
\begin{align*}
Y &= X\beta + \epsilon \\
\hat{\beta} &= (X^tX)^{-1}X^tY \\
\hat{Y} &= X\hat{\beta} = HY,
\end{align*}
\]

\[
H = X(X^tX)^{-1}X^t \\
H^2 = H
\]

\[
\Rightarrow \hat{e} := Y - \hat{Y} = (I - H)Y
\]

\[
= (I - H)(Y - \hat{Y}) \quad \text{since } H\hat{Y} = H(HY) = H^2Y = HY = \hat{Y}
\]

\[
= (I - H)\hat{e}
\]

\[
\Rightarrow \text{raw residuals satisfy } \hat{e} = (I - H)\hat{e}
\]
logistic regression:

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

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

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

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

$He^P = [\hat{D}^{1/2} X (X^t \hat{D} X)^{-1}] X^t (Y - \hat{Y}) = 0 \quad (\ast)$

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

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

Note that $H^2 = H$ as in linear models (exercise).
High leverage points in logistic regression

\[ e^P = (I - H) e^P \]

\( M \) spans residual space \( e^P \). 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 $X = [X_j; X_{-j}]$,

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

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

$e_{Y|X_{-j}} := Y - X_{-j}(X_{-j}^tX_{-j})^{-1}X_{-j}^tY = (I - H_{-j})Y,$

$= \text{raw residuals in model with } j^{th} \text{ covariable removed}$

$e_{X_j|X_{-j}} := X_j - X_{-j}(X_{-j}^tX_{-j})^{-1}X_{-j}^tX_j = (I - H_{-j})X_j$

$= \text{raw residuals in model } X_j = X_{-j}\beta_{-j}^* + \epsilon_x \quad \epsilon_x \sim N(0, \sigma_x^2) \ i.i.d.$

$= \text{measure of linear dependency of } X_j \text{ on the remaining covariates}$
The partial residual plot is given by plotting $e_{X_j|X_{-j}}$ versus $e_{Y|X_{-j}}$.

$$Y = X_{-j} \beta_{-j} + X_j \beta_j + \epsilon \quad \beta = \begin{pmatrix} \beta_{-j} \\ \beta_j \end{pmatrix}$$

$$\Rightarrow (I - H_{-j})Y = (I - H_{-j})X_{-j} \beta_{-j} + (I - H_{-j})X_j \beta_j + (I - H_{-j})\epsilon$$

$$= 0 \quad \text{since} \quad (I - X_{-j}(X_{-j}^tX_{-j})^{-1}X_{-j})X_{-j} = X_{-j} - X_{-j} = 0$$

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

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

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

If partial residual plot scatters
- around 0 $\Rightarrow X_j$ has no influence on $Y$
- linear $\Rightarrow X_j$ should be linear in model
- nonlinear $\Rightarrow X_j$ should be included with this nonlinear form.
**Simpler plot:** Plot $X_j$ versus $e_{Y|X_{-j}}$. Same behavior if $X_j$ does not depend on other covariates.
This follows from the fact, that ML estimators of $\beta_{-j}$ for two models with and without $j^{th}$ covariate coincide, if $X_j$ is orthogonal to $X_{-j}$. Then

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

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

where $e$ is distributed around 0 and components of $e$ are assumed nearly independent.
Partial residual plot

Logistic regression:
Landwehr, Pregibon, and Shoemaker (1984) propose to use

\[ e(Y|X_{-j})_i := \frac{Y_i - n_i \hat{p}_i}{n_i \hat{p}_i (1 - \hat{p}_i)} + \hat{\beta}_j x_{ij} \]

as partial residual

Heuristic derivation:
Justification: IWLS

Recall: \( Z^\hat{\beta} = X^\hat{\beta} + \hat{D}^{-1}e^r \) “obs. vector”

\[ \text{cov}(\hat{D}^{-1}e^r) = \hat{D}^{-1} \Rightarrow \hat{\beta} = (X^t \hat{D} X)^{-1} X^t \hat{D} Z^\hat{\beta} \]
Consider \( \text{logit}(p) = X\beta + L\gamma, \) \( L \in \mathbb{R}^n \) new covariable with \( L \perp X. \)

\[
Z_L := X\hat{\beta} + L\hat{\gamma} + \hat{D}_L^{-1}e_{L}^r \quad \text{with}
\]

\[
e_{L_i}^r := y_i - n_i \frac{e^{x_i^t\hat{\beta} + l_i\hat{\gamma}}}{1 + e^{x_i^t\hat{\beta} + l_i\hat{\gamma}}} \quad L = (l_1, \ldots, l_n)^t
\]

\[
\hat{D}_L := \text{diag}(\ldots, n_i\hat{p}_{L_i}(1 - \hat{p}_{L_i}), \ldots)
\]

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

\[
e(Y|X_{-j})_i := (\hat{D}_L^{-1})_{ii}e_{L_i}^r + \hat{\gamma}_l = \frac{y_i - n_i\hat{p}_{L_i}}{n_i\hat{p}_{L_i}(1 - \hat{p}_{L_i})} + \hat{\gamma}_l
\]

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

For binary data we need to smooth data.
Cook’s distance in linear models

\[ D_i \ := \ ||X\hat{\beta} - X\hat{\beta}_{-i}||^2 / p\hat{\sigma}^2 \]

\[ = \ (\hat{\beta} - \hat{\beta}_{-i})^t (X^t X)(\hat{\beta} - \hat{\beta}_{-i}) / p\hat{\sigma}^2 \]

\[ = \ \frac{\hat{\varepsilon}_i^2 h_{ii}}{p\hat{\sigma}^2 (1 - h_{ii})^2} \]

Measures change in confidence ellipsoid when \( i^{th} \) obs. is removed.
Cook’s distance in logistic regression

Using LRT it can be shown, that

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

is an approx. $100(1 - \alpha)$ % CI for $\beta$

$$\Rightarrow D_i := -2 \ln \left\{ \frac{L(\hat{\beta}_{-i})}{L(\hat{\beta})} \right\}$$

measures change when $i^{th}$ obs. removed; difficult to calculate. Using Taylor expansion we have:

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

$$\Rightarrow D_i \approx (\hat{\beta}_{-i} - \hat{\beta})^t X^t \hat{D} X (\hat{\beta}_{-i} - \hat{\beta})$$
Approximate $\hat{\beta}_{-i}$ by a single step Newton Rapson starting from $\hat{\beta}$:

$$
\Rightarrow \quad \hat{\beta}_{-i} \approx \hat{\beta} - \frac{(X^t DX)^{-1}x_i(y_i - n_i \hat{p}_i)}{1-h_{ii}} \quad \text{(exercise)}
$$

where

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

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

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

