MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 9, spring 2019
Model validation in logistic regression

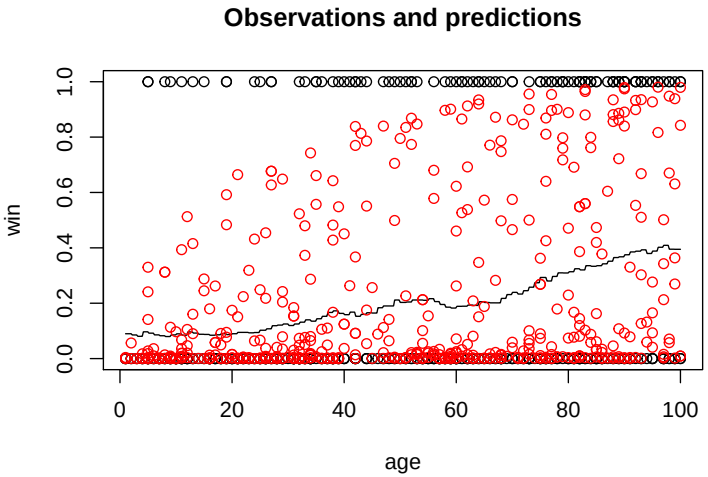
Mathematical Statistics / Centre for Mathematical Sciences
Lund University

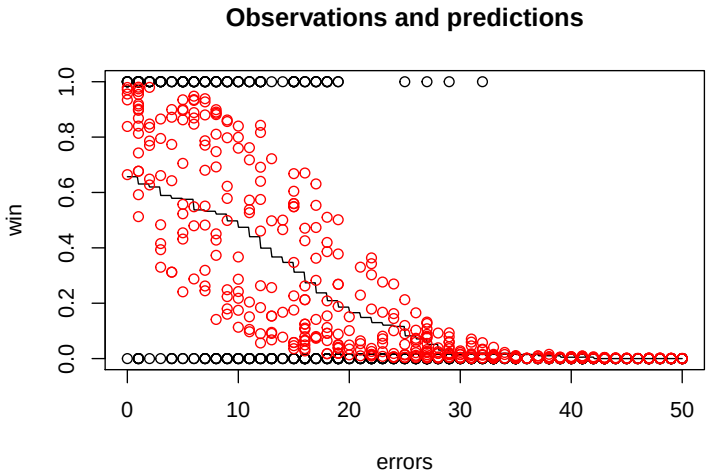
8/5-19

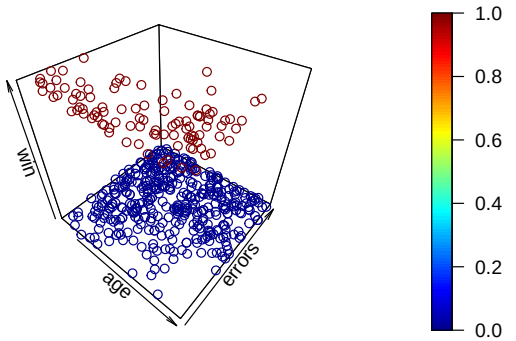
Example: winning depending on age and errors made

$p_i = Pr(\text{win a game})$ where

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{errors}_i$$







Leverage

We have the linear predictors

$$\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}$$

The leverage values in logistic regression are the diagonal values v_{ii} of the hat-matrix

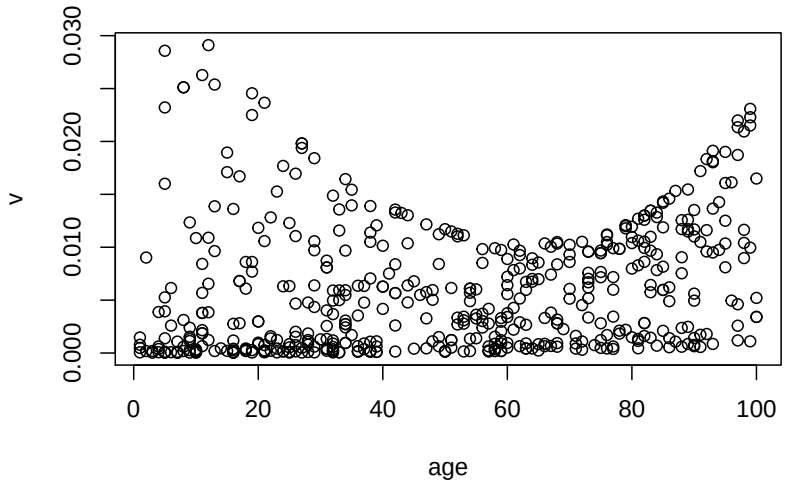
$$\mathbf{P} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$

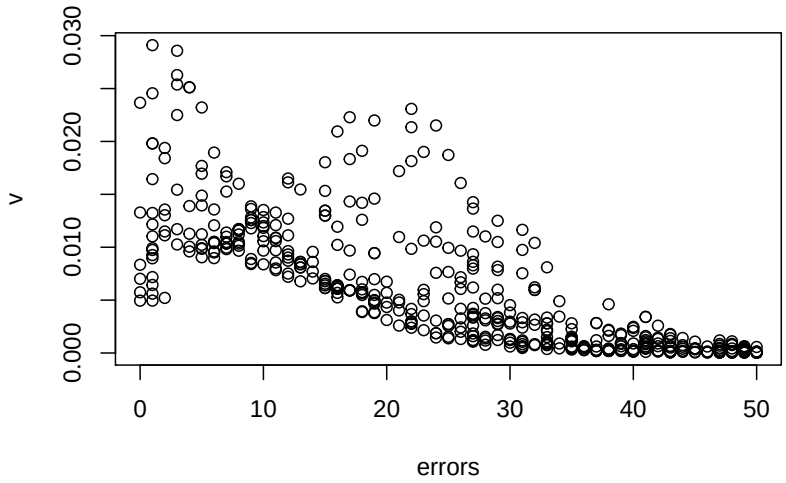
where $\mathbf{W}^{1/2}$ is a diagonal matrix with elements $\sqrt{w_{ii}} = \sqrt{\hat{p}_i(1-\hat{p}_i)}$.

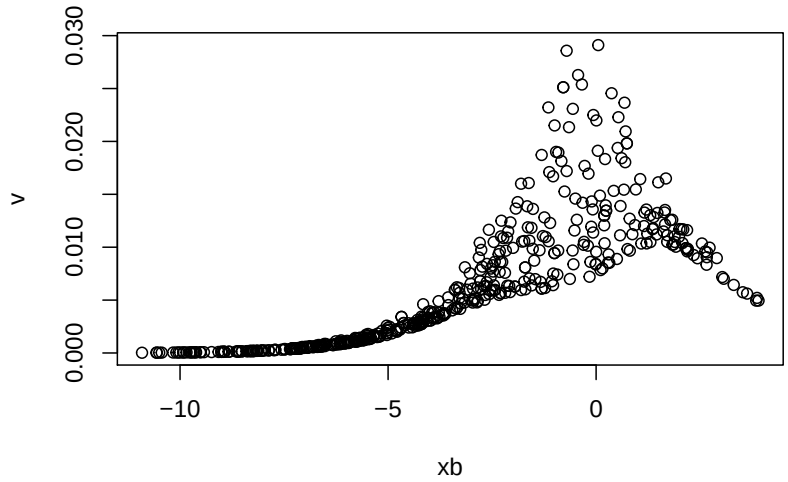
The Leverage values are now depending both on $\mathbf{X}$ and $\mathbf{Y}$ and as such these are no longer indicators of outliers w.r.t. $\mathbf{X}$.

However, they can still be used to standardize residuals.

In R, v_{ii} can be obtained using `influence(model)$hat`.


v
0.000
0.010
0.020
0.030
0
20
40
60
80
100
age


v
0.000
0.010
0.020
0.030
0
10
20
30
40
50
errors


v
0.000
0.010
0.020
0.030
−10
−5
0
xb

Residuals in logistic regression

Pearson residuals

Simple standardization, since $Y_i \sim Bin(1, p_i)$ with $E(Y_i) = p_i$ and $V(Y_i) = p_i(1 - p_i)$:

$$\tilde{r}_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}} \qquad (\nsim N(\cdot, \cdot) \quad \text{not even asymptotically!})$$

In R: `influence(model)$pear.res`.

However, in general the problem with residual analysis for logistic regression is that such plots are not very revealing because of the binary nature of Y.

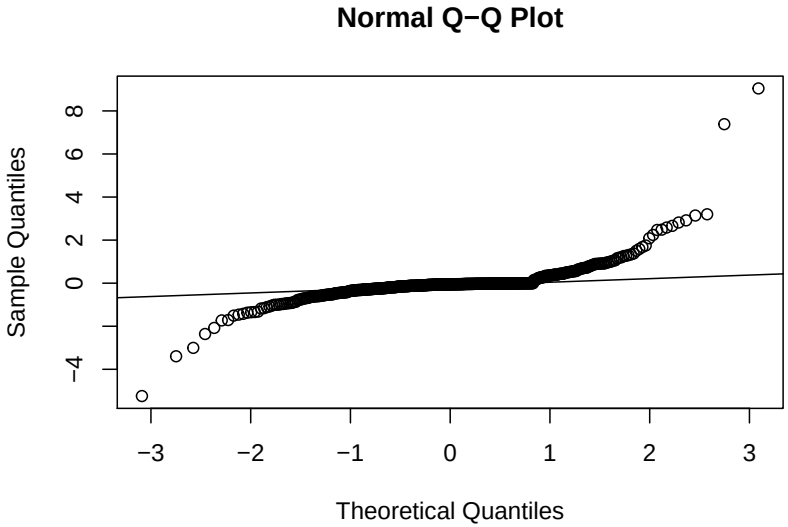
Standardized residuals

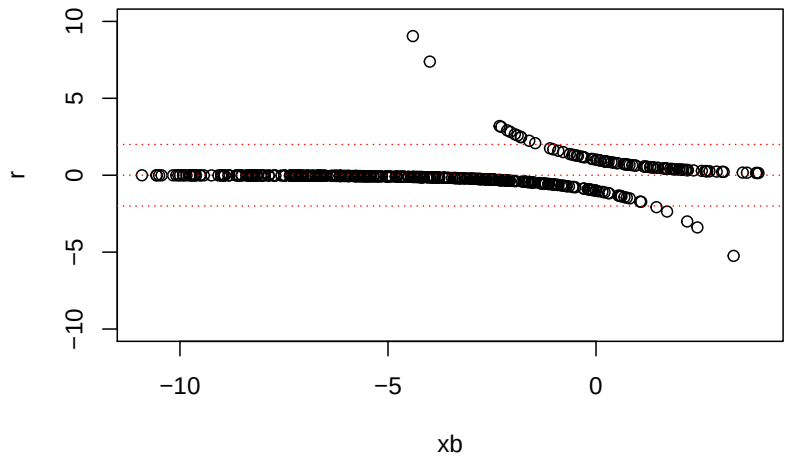
As in linear regression, we can standardize the residuals using the leverage:

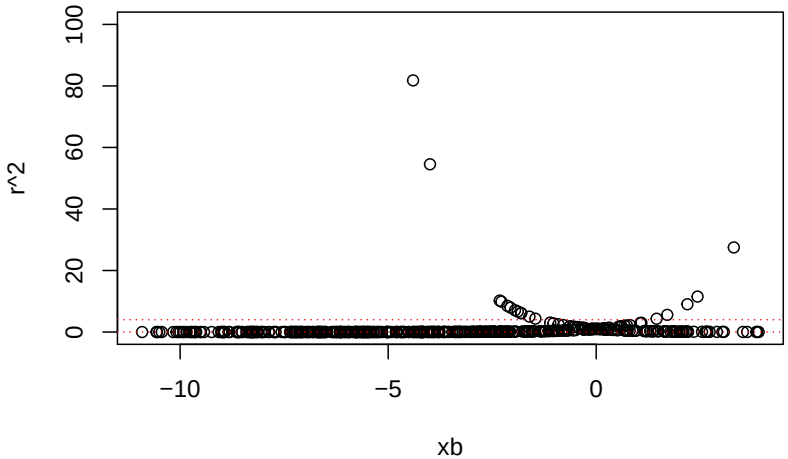
$$r_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)(1-v_{ii})}} \approx N(0,1) \qquad \text{(for large } n\text{)}$$

If $|r_i| > |\lambda_{\alpha/2}| \approx 2$ it might be considered suspiciously large.

Plots of r_i vs $\mathbf{x}_i\hat{\boldsymbol{\beta}}$ can be useful, although it's sometimes more revealing to plot their squares, e.g. r_i^2 vs $\mathbf{x}_i\hat{\boldsymbol{\beta}}$.


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
−4
0
2
4
6
8


r
xb
−10
−5
0
5
10


r^2
xb
0
20
40
60
80
100
−10
−5
0

Deviance residuals

Use the contribution to the deviance, $D = \sum_{i=1}^{n} d_i^2$ where

$$d_i = \pm\sqrt{2\left(y_i \ln\frac{y_i}{\hat{p}_i} + (1-y_i)\ln\frac{1-y_i}{1-\hat{p}_i}\right)}$$

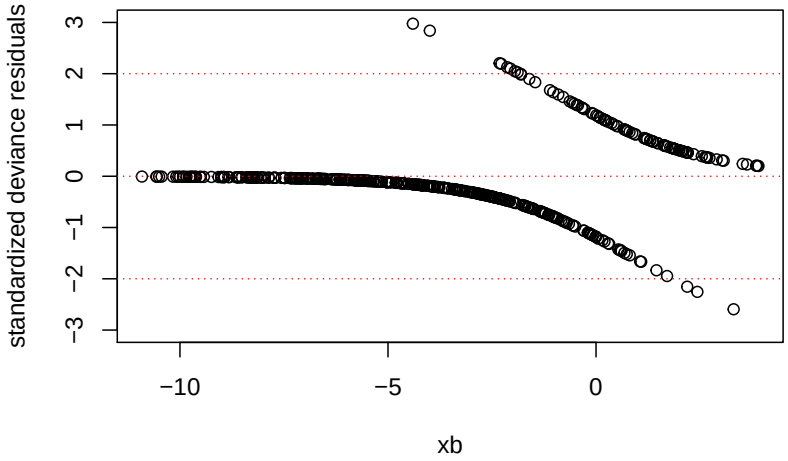
using the sign of $Y_i - \hat{p}_i$.

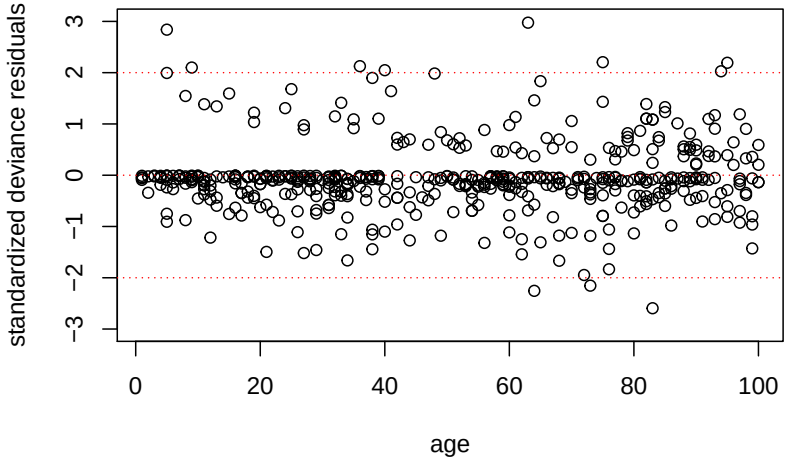
$$d_i = \begin{cases} -\sqrt{2\ln\frac{1}{1-\hat{p}_i}} & \text{if } Y_i = 0 \\ +\sqrt{2\ln\frac{1}{\hat{p}_i}} & \text{if } Y_i = 1 \end{cases}$$

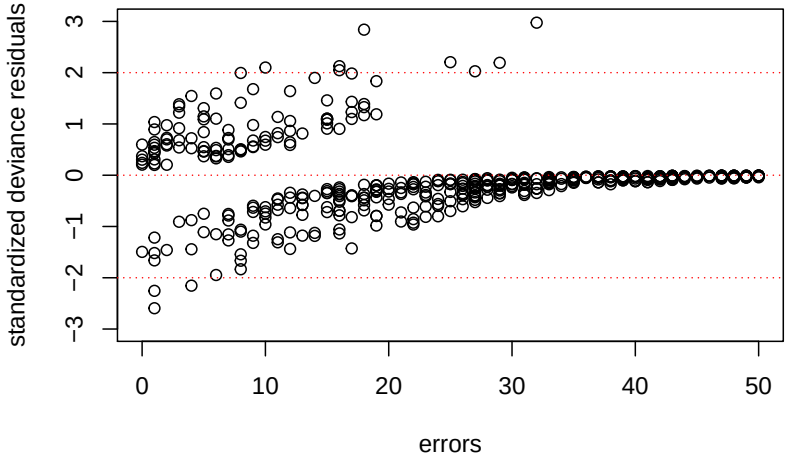
The deviance residual will be small if $Y_i = 0$ and $\hat{p}_i$ is close to zero, or if $Y_i = 1$ and $\hat{p}_i$ is close to one. Otherwise it will be large. If $|d_i| > 2$ it can be considered to be too large.

In R: `influence(model)$dev.res`.

The deviance residuals can be standardized as $d_i/\sqrt{1-v_{ii}}$.


standardized deviance residuals
−3
−2
−1
0
1
2
3
−10
−5
0
xb


standardized deviance residuals
−3
−2
−1
0
1
2
3
0
20
40
60
80
100
age


standardized deviance residuals
−3
−2
−1
0
1
2
3
0
10
20
30
40
50
errors

Influential observations

We can measure the influence of individual observations on the β-estimates in a similar way as in linear regression.

Cook's distance

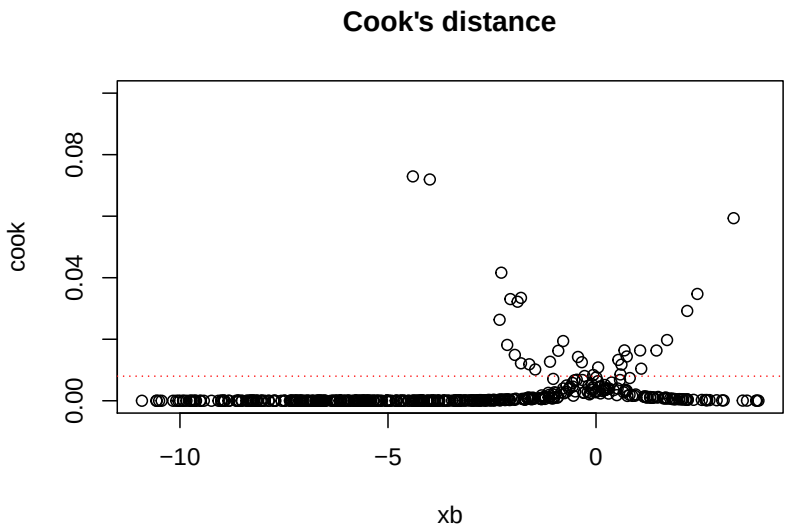
There is a version of Cook's distance for logistic regression:

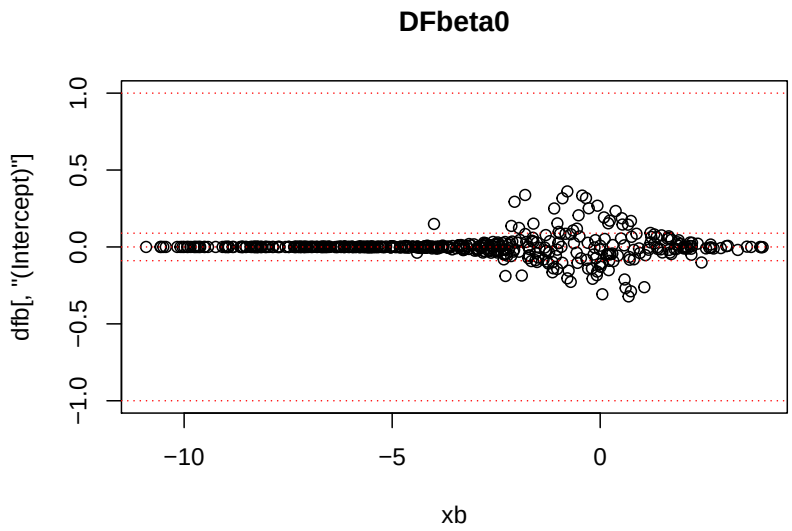
$$D_i^{Cook} = \frac{r_i^2}{p+1} \cdot \frac{v_{ii}}{1-v_{ii}}$$

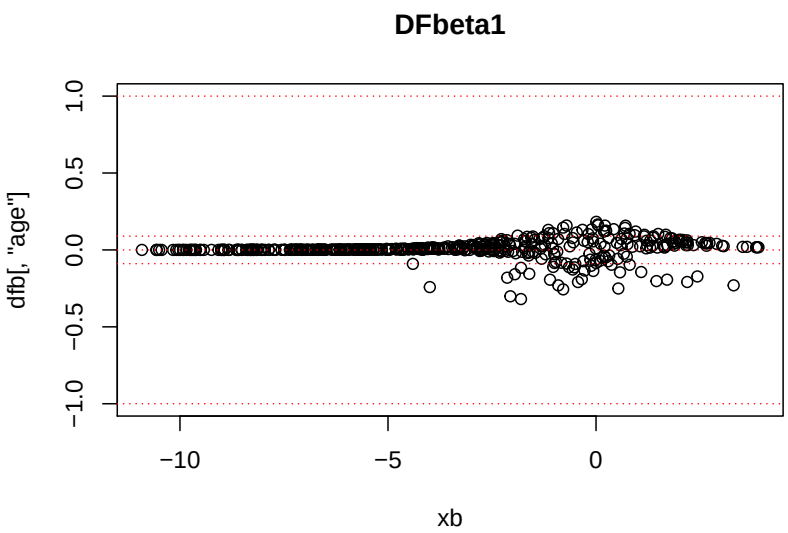
We might consider influential cases those with $D_i^{Cook} > 1$.

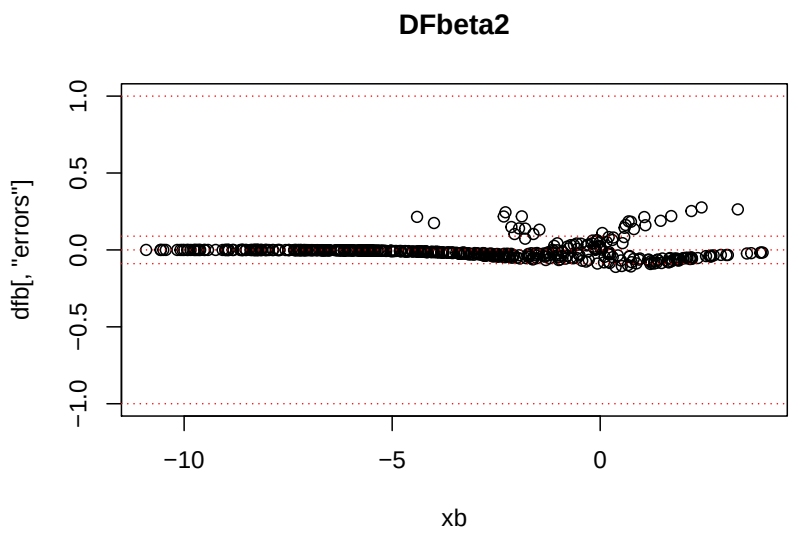
dfbetas

We also have similar versions of DFBETA.


Cook's distance
cook
0.00
0.04
0.08
−10
−5
0
xb


DFbeta0
dfb[, "(Intercept)"]
−1.0
−0.5
0.0
0.5
1.0
−10
−5
0
xb


DFbeta1
dfb[, "age"]
−1.0
−0.5
0.0
0.5
1.0
−10
−5
0
xb


DFbeta2
dfb[, "errors"]
−1.0
−0.5
0.0
0.5
1.0
−10
−5
0
xb

Goodness of fit

Sometimes we want to use our model to classify future objects as "success" or "failure", depending on the probabilies given by their x-values. We then classify the predicted values using

$$\hat{Y}_i = \begin{cases} \text{failure} & \text{if } \hat{p}_i \leq 0.5 \text{ or } \mathbf{x}_i\hat{\boldsymbol{\beta}} \leq 0, \\ \text{success} & \text{if } \hat{p}_i > 0.5 \text{ or } \mathbf{x}_i\hat{\boldsymbol{\beta}} > 0 \end{cases}$$

and compare them to the observed values Y_i.

- Sensitivity is the proportion of the true successes that have been correctly classified as successes (true positive).
- Specificity is the proportion of the true failures that have been correctly classified as failures (true negatives).

Both sensitivity and specificity should be large.

Note, this is not very useful when looking at a rare event when almost no observations have $\hat{p}_i > 0.5$.

Example

- The "best" model $\beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{errors}_i$:

	Predicted			Correctly	
Observed	lose	win	total	classified	
lose	379	20	399	95 %	specificity
win	26	75	101	74 %	sensitivity

- The slightly worse model $\beta_0 + \beta_2 \cdot \text{errors}_i$:

	Predicted			Correctly	
Observed	lose	win	total	classified	
lose	369	30	399	92 %	specificity
win	38	63	101	62 %	sensitivity

Both sensitivity and specificity are higher for the "best" model.