MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: . . . with Data Gathering, 9 hp
Lecture 9, spring 2018
Model validation in logistic regression

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

3/5-18
Comparing non-nested models

AIC and BIC again

Information for a model with $p + 1$ parameters:

$$AIC(p + 1) = 2(p + 1) - 2 \ln L(\hat{\beta}) = 2(p + 1) + D$$

$$BIC(p + 1) = (p + 1) \ln n - 2 \ln L(\hat{\beta}) = (p + 1) \ln n + D$$

Tradeoff between small deviance and large number of parameters: $SS(\text{Error})_{p+1}$ decreases and $p + 1$ increases with $p$.

The "best" model is the one with the smallest AIC/BIC.
R\textsuperscript{2} for linear regression (again)

For linear regression we could calculate the fraction of the variability in Y that was explained by our model by

\[ R^2 = 1 - \frac{SS(\text{Error})}{SS(\text{Total}_{corr})}, \quad R_{\text{adj}}^2 = 1 - \frac{MS(\text{Error})}{MS(\text{Total}_{corr})} \]

Since we do not use the sums of squares this is no longer possible.

Cox-Snell pseudo R\textsuperscript{2} for logistic regression

For a logistic regression we can use the likelihood function to create something similar:

\[ R_{\text{Cox-Snell}}^2 = 1 - \left( \frac{L(\hat{\beta}_0)}{L(\hat{\beta})} \right)^{2/n} \]

with \( 0 \leq R_{\text{Cox-Snell}}^2 \leq 1 - (L(\hat{\beta}_0))^{2/n}. \)
Nagelkerke pseudo $R^2$ for logistic regression

Since the we would like a model with a perfect fit, $L(\hat{\beta}) = 1$, to give $R^2 = 1$ we can rescale it as

$$R^2_{\text{Nagelkerke}} = \frac{R^2_{\text{Cox-Snell}}}{1 - (L(\hat{\beta}_0))^{2/n}} = \frac{1 - \left(\frac{L(\hat{\beta})}{L(\hat{\beta}_0)}\right)^{2/n}}{1 - (L(\hat{\beta}_0))^{2/n}}$$

which has $0 \leq R^2_{\text{Nagelkerke}} \leq 1$.

Note, the value of the pseudo $R^2$ is not really interpretable but it can be used to compare models. A model with a larger pseudo $R^2$ is ”better”. Does not compensate for using more covariates!

In R you can

- get the likelihood values using the deviances from `anova(model)` and calculate the pseudo $R^2$ yourself, or
- install the package pscl (once), then activate it with `library(pscl)` and run `pR2(model)`.
Example

Full model with interaction:

$$\ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

Best model according to AIC and BIC:

$$\ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$
Leverage

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

\[ P = W^{1/2}X(X'WX)^{-1}X'W^{1/2} \]

where $W^{1/2}$ is a diagonal matrix with elements $\sqrt{w_{ii}}$.

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

However, they can still be used to standardize residuals.

In R, $v_{ii}$ can be obtained using `influence(model)$hat`. 
The leverage is high where the slope of the estimated probability is high.
Residuals in logistic regression

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

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

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:

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

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

Plots of \(r_i\) vs \(x_i\hat{\beta}\) can be useful, although it’s sometimes more revealing to plot their squares, e.g. \(r_i^2\) vs \(x_i\hat{\beta}\).
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 \( d_i/\sqrt{1 - v_{ij}} \).
Residual plots

- Standardized residuals
- Squared Standardized residuals
- Standardized deviance residuals
- Normal Q–Q Plot: standardized deviance residuals

Anna Lindgren - anna@maths.lth.se
Influential observations

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

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

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

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

dfbetas
We also have similar versions of DFBETA.
Goodness of fit

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

$$\hat{Y}_i = \begin{cases} 
\text{failure} & \text{if } \hat{p}_i \leq 0.5, \\
\text{success} & \text{if } \hat{p}_i > 0.5
\end{cases}$$

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 $\hat{p}_i > 0.5$.  

Anna Lindgren - anna@maths.lth.se
Example

- The "best" model $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$:

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Correctly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>failure</td>
<td>63</td>
<td>94 % specificity</td>
</tr>
<tr>
<td>success</td>
<td>6</td>
<td>82 % sensitivity</td>
</tr>
</tbody>
</table>

- The slightly worse model $\beta_0 + \beta_1 x_{1i}$:

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Correctly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>failure</td>
<td>58</td>
<td>87 % specificity</td>
</tr>
<tr>
<td>success</td>
<td>10</td>
<td>70 % sensitivity</td>
</tr>
</tbody>
</table>

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