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

Lecture 8, spring 2019
Maximum likelihood-estimates and deviance

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

6/5-19

Logistic regression model

$$Y_i = \text{"success"}\ (=1)\ \text{or}\ \text{"failure"}\ (=0)$$

$$P(Y_i = 1) = 1 - P(Y_i = 0) = p_i$$

$$Y_i \sim Bin(1, p_i), \quad i = 1, \ldots, n, \ \text{(independent)}$$

$$E(Y_i) = p_i = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}}$$

$$\text{logit}(p_i) = \ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} = \mathbf{x}_i\boldsymbol{\beta}$$

How should we estimate β

Least squares estimates?

- Minimize $Q(\beta) = \sum_{i=1}^{n}(\ln \frac{Y_i}{1-Y_i} - \mathbf{x}_i\beta)^2$?
  No, $\ln \frac{Y_i}{1-Y_i} = \ln 0 = -\infty$ or $\ln \infty = \infty$. Useless!
- Minimize $Q(\beta) = \sum_{i=1}^{n}(Y_i - p_i)^2 = \sum_{i=1}^{n}(Y_i - \frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}})^2$?
  No, since $V(Y_i) = p_i(1-p_i)$ is not constant. We would need to do a weighted least squares but the weights $1/V(Y_i)$ are unknown.
- Minimize $Q(\beta) = \sum_{i=1}^{n}\frac{(Y_i-p_i)^2}{p_i(1-p_i)} = \sum_{i=1}^{n}\frac{(Y_i-\frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}})^2}{\frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}}(1-\frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}})}$
  using iteratively re-weighted least squares?
  No, it can be done but it is a very inefficient method with a slow convergence rate.

Totally different method? Yes!

Maximum likelihood-method

Since we know what type of distribution our data come from, $Y_i \in Bin(1, p_i)$, we can find the β-values that maximize the probability of getting exactly the observation values that we got. That means that we should maximize the likelihood function

$$L(\beta) = Pr(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} Pr(Y_i = y_i)$$

$$= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}}\right)^{y_i}\left(1-\frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}}\right)^{1-y_i}$$

It is easier to maximize the log-likelihood function

$$\ln L(\beta) = \sum_{i=1}^{n} \left(y_i \mathbf{x}_i \beta - \ln(1+e^{\mathbf{x}_i\beta}) \right)$$

ML-estimate for the Null model, $\ln \frac{p_i}{1-p_i} = \beta_0$

For the simplest model, having only an intercept, we have

$$p_i = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

and the ML-estimate can easily be derived as

$$\ln L(\beta_0) = \sum_{i=1}^{n} \left(y_i \beta_0 - \ln(1 + e^{\beta_0}) \right) = \beta_0 \sum_{i=1}^{n} y_i - n \ln(1 + e^{\beta_0})$$

$$\frac{d \ln L(\beta_0)}{d\beta_0} = \sum_{i=1}^{n} y_i - \frac{n e^{\beta_0}}{1 + e^{\beta_0}} = 0 \Rightarrow$$

$$\hat{\beta}_0 = \ln \frac{\bar{y}}{1 - \bar{y}} \Rightarrow \hat{p}_i = \bar{y} = \frac{\text{number of successes}}{\text{number of observations}}$$

ML-estimate for the full model: $\ln \frac{p_i}{1-p_i} = \mathbf{x}_i\boldsymbol{\beta}$

Find the $\boldsymbol{\beta}$ that maximizes the log-likelihood. This means setting all the partial derivatives equal to 0:

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} \left(y_i - \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}} \right) = 0$$

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left(x_{ji} \cdot y_i - x_{ji} \cdot \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}} \right) = 0$$

This gives us the following relationships:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p_i, \quad \sum_{i=1}^{n} x_{ji} \cdot y_i = \sum_{i=1}^{n} x_{ji} \cdot p_i, \quad j = 1, \ldots, p$$

Matrix formulation: $\mathbf{X}'\mathbf{p} = \mathbf{X}'\mathbf{Y}$

Nonlinear in $\boldsymbol{\beta}$ so no closed form solutions. We need an iterative method, e.g. Newton-Raphson algorithm.

Estimates via Newton-Raphson (a.k.a. Fisher-scoring)*

- Start from an arbitrary "guess" $\hat{\boldsymbol{\beta}}^{(0)}$ then iterate until $\| \hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)} \|$ is "small enough".
- A generic iteration k of Newton-Raphson/Fisher-scoring is:

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + (\mathbf{X}'\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{p}}^{(k)}), \quad k = 0, 1, ...$$

- Here $\hat{\mathbf{p}}^{(k)}$ are estimated using the current $\hat{\boldsymbol{\beta}}^{(k)}$
- $\mathbf{W}^{(k)}$ diagonal matrix with elements $(w_{11}^{(k)}, ..., w_{nn}^{(k)})$ and $w_{ii}^{(k)} = \hat{p}_i^{(k)}(1 - \hat{p}_i^{(k)})$.
- at convergence (k large) we write $\mathbf{W}^{(k)} \equiv \mathbf{W}$ and $\hat{\mathbf{p}}^{(k)} \equiv \hat{\mathbf{p}}$.

ML-estimates

At convergence the ML-estimates of $\boldsymbol{\beta}$ become

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

where $\mathbf{W}$ is a diagonal matrix with elements

$$w_{ii} = \hat{p}_i(1 - \hat{p}_i), \quad i = 1, \ldots, n,$$

$\mathbf{Z}$ is a column vector with elements

$$Z_i = \ln \frac{\hat{p}_i}{1 - \hat{p}_i} + \frac{Y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}, \quad i = 1, \ldots, n$$

and

$$\hat{p}_i = \frac{e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}, \quad i = 1, \ldots, n.$$

Asymptotics from likelihood estimation

For all maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\mathbf{0}, \mathbf{I}_{\text{Fish}}^{-1}) \qquad (n \to \infty)$$

where $\mathbf{I}_{\text{Fish}}$ is the Fisher information matrix (see any reference in inference theory and some numerical analysis).
In this case, it means that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}) \qquad (n \to \infty)$$
$$\mathbf{x}_0\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_0\boldsymbol{\beta}, \mathbf{x}_0(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_0') \qquad (n \to \infty)$$

Motivates the Wald test and confidence interval for β_j and constructing intervals for p_0 based on the log odds $\mathbf{x}_0\boldsymbol{\beta}$ in Lecture 7.

Deviance

In logistic regression (and in general for nonlinear models) we do not have sums of squares and cannot do an ANOVA decomposition. We can thus not use either the global or the partial F-test to test our model.
Solution: use the maximized likelihood function, $L(\hat{\boldsymbol{\beta}})$ instead!

Deviance

The *Deviance*, D, of a model is defined as

$$D = -2 \ln L(\hat{\boldsymbol{\beta}}) \sim \chi^2(n - (p+1)) \text{ when } n \to \infty$$

(p is the number of covariates, not a probability)

A large likelihood means a large probability for the estimated model to produce our data, which is good. So a small deviance is good.

The deviance for the null model with only the intercept and $\hat{p}_i = \bar{y}$ then becomes

$$D_0 = -2 \ln L(\hat{\beta}_0) = -2 \sum_{i=1}^{n} (y_i \ln \hat{p}_i + (1 - y_i) \ln(1 - \hat{p}_i))$$
$$= -2n (\bar{y} \ln \bar{y} + (1 - \bar{y}) \ln(1 - \bar{y})) .$$

For a general model with $\hat{p}_i = \frac{e^{\mathbf{x}_i \hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i \hat{\boldsymbol{\beta}}}}$ the deviance becomes

$$D = -2 \ln L(\hat{\boldsymbol{\beta}}) = -2 \sum_{i=1}^{n} (y_i \ln \hat{p}_i + (1 - y_i) \ln(1 - \hat{p}_i)).$$

Deviance in R: Green fruit

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6931     0.1225  -5.660 1.52e-08 ***
typeapple     0.9808     0.2708   3.622 0.000292 ***
typemelon     0.6931     0.3391   2.044 0.040956 *
---
Null deviance: 548.46  on 409  degrees of freedom
Residual deviance: 532.97  on 407  degrees of freedom
AIC: 538.97
```

`Null deviance` is D_0 for a model having intercept only.
`Residual deviance` is D for our model.
Can we use this to test if our model is significantly different from the null model (similar to a global F-test)?

Likelihood ratio test (global)

For nested models we can compare the likelihoods through the deviance. If H_0: $\beta_1 = \cdots = \beta_p = 0$ is false the full model is better than the null model and we should get

$$L(\hat{\boldsymbol{\beta}}) > L(\hat{\beta}_0) \Leftrightarrow \ln L(\hat{\boldsymbol{\beta}}) > \ln L(\hat{\beta}_0) \Leftrightarrow -2 \ln L(\hat{\beta}_0) > -2 \ln L(\hat{\boldsymbol{\beta}})$$

$$\Leftrightarrow D_0 - D = \sum_{i=1}^{n} \left(y_i \ln(\frac{\hat{p}_i}{\bar{y}})^2 + (1 - y_i) \ln(\frac{1 - \hat{p}_i}{1 - \bar{y}})^2 \right) > 0$$

If H_0 is true it can be proven that $D_0 - D \sim \chi^2(p)$, asymptotically, and we should reject H_0 at significance level α if

$$D_0 - D > \chi^2_\alpha(p)$$

Likelihood ratio test (partial)

The likelihood ratio test also does the job of a partial F-test:

- Test H_0: k specific β-parameters (e.g. the last k) $= 0$.
- If H_0 is false then

$$D_{\text{red}} - D_{\text{full}} = \sum_{i;y_i=1} \ln(\frac{\hat{p}_{i,\text{full}}}{\hat{p}_{i,\text{red}}})^2 + \sum_{i;y_i=0} \ln(\frac{1-\hat{p}_{i,\text{full}}}{1-\hat{p}_{i,\text{red}}})^2 > 0.$$

- For the successes, when $y_i = 1$, the probability p_i should be close to 1 and we want $\hat{p}_{i,\text{full}} > \hat{p}_{i,\text{red}}$.
- For the failures, when $y_i = 0$, the probability p_i should be close to 0 and we want $1 - \hat{p}_{i,\text{full}} > 1 - \hat{p}_{i,\text{red}}$.
- If H_0 is true then $D_{\text{red}} - D_{\text{full}} \sim \chi^2(k)$, asymptotically, and we should reject H_0 if

$$D_{\text{red}} - D_{\text{full}} > \chi^2_\alpha(k).$$

Example: Fruit

Are there differences in the proportion of green among the fruit?

- H_0: $\beta_1 = \beta_2 = 0$.
- Full model: $\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$.
- Reduced (Null) model: $\ln \frac{p_i}{1-p_i} = \beta_0$.
- $D_{\text{diff}} = D_{\text{red}} - D_{\text{full}} = 548.5 - 533.0 = 15.5 > \chi^2_{0.05}(2) = 5.99$. Reject H_0 at $\alpha = 0.05$ significance level. There are differences between at least two of the fruit.

R is intelligent and we can use `anova(model.red, model.full)`. When `model` is estimated using `glm()`, the `anova()`-function knows to produce a deviance table instead of an ANOVA table.

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.

```
              df       AIC df       BIC
model.0        1 550.4628  1 554.4789
model.banana   2 537.4905  2 545.5228  Bananas are different!
model.apple    2 541.0960  2 549.1283
model.melon    2 550.2676  2 558.3000
model.full     3 538.9674  3 551.0159
```

R^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

$$R^2 = 1 - \frac{SS(\text{Error})}{SS(\text{Total}_{\text{corr}})}, \qquad R^2_{\text{adj}} = 1 - \frac{MS(\text{Error})}{MS(\text{Total}_{\text{corr}})}$$

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

Cox-Snell pseudo R^2 for logistic regression

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

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

with $0 \leq R^2_{\text{Cox-Snell}} \leq 1 - (L(\hat{\beta}_0))^{2/n}$.

Nagelkerke pseudo R^2 for logistic regression

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

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

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 `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: Fruit

Pseudo R^2 in percentages. Maximum value for $R^2_{\text{Cox-Snell}} = 73.8\,\%$.

	R2CS	R2N
model.0	0.00	0.00
model.banana	3.59	4.86
model.apple	2.73	3.71
model.melon	0.53	0.72
model.full	3.71	5.03

The full model has, of course, the largest pseudo R^2-value but the banana-model, favoured by AIC and BIC, is only slightly worse.
Best model:

$$\ln \frac{p_{\text{green}}}{1 - p_{\text{green}}} = \begin{cases} \beta_0 & \text{for bananas} \\ \beta_0 + \beta_1 & \text{for apples and melons} \end{cases}$$