R-bloggers

R news and tutorials contributed by (573) R bloggers

- Home
- About
- RSS
- add your blog!
- R jobs���
- Contact us

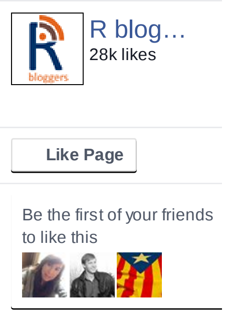
Welcome!

Follow @rbloggers 25K f

Here you will find daily **news and tutorials about R**, contributed by over 573 bloggers.
There are many ways to **follow us** -
By e-mail:

Your e-mail here
Subscribe
21406 readers
BY FEEDBURNER

On Facebook:


R blog…
28k likes

Like Page

Be the first of your friends to like this

If you are an R blogger yourself you are invited to add your own R content feed to this site (**Non-English** R bloggers should add themselves- here)

Jobs for R-users

- Looking for a Freelance R-Developer with Shiny experience
- Data Analyst @ New York
- R/Shiny App with d3 (small job, quick turnaround, $250 < 4hrs)
- Postdoctoral position (Duke’s River center) in data visualization and ecosystem science @ Durham, North Carolina, United States
- Software Engineer for The Computational Biology Program at Oregon Health and Science University @ Portland, Oregon, United States

Search & Hit Enter

Popular Searches

- heatmap
- web scraping
- maps
- hadoop
- twitter
- alt=
- boxplot
- shiny
- time series
- animation
- ggplot2
- ggplot
- how to import image file to R
- latex
- trading
- finance
- PCA
- excel
- googlevis
- eclipse
- quantmod
- rstudio
- market research
- rattle
- Tutorial
- coplot
- rcmdr
- knitr
- title=
- rbloggers

Recent Posts

- Partools, Recommender Systems and More
- Using htmlwidgets with knitr and Jekyll
- Wind in Netherlands
- Bioenergetics in R Workshop
- Correlation and Linear Regression
- Linear model with time series random component
- James Bond movies
- What it means to be a US Veteran Today
- Blog Post at Pluralsight
- The Lady Loves Statistics
- Annotables: R data package for annotating/converting Gene IDs
- Szkolenie z analizy sieciowej
- Applied Statistical Theory: Quantile Regression
- Let’s meet on SatRdays: the link between RUGs and conferences
- importance sampling with infinite variance

Other sites

- SAS blogs
- Statistics of Israel
- Jobs for R-users

Veterinary Epidemiologic Research: GLM – Evaluating Logistic Regression Models (part 3)

March 19, 2013
By denishaine

(This article was first published on **denis haine » R**, and kindly contributed to R-bloggers)

Third part on logistic regression (first here, second here).
Two steps in assessing the fit of the model: first is to determine if the model fits using summary measures of goodness of fit or by assessing the predictive ability of the model; second is to deterime if there's any observations that do not fit the model or that have an influence on the model.

Covariate pattern
A covariate pattern is a unique combination of values of predictor variables.

```
mod3 <- glm(casecont ~ dcpct + dneo + dc
+             family = binomial("logit"
summary(mod3)

Call:
glm(formula = casecont ~ dcpct + dneo +
    family = binomial("logit"), data =

Deviance Residuals:
    Min       1Q   Median       3Q
-1.9191  -0.7682   0.1874   0.5876   2.

Coefficients:
                  Estimate Std. Error z
(Intercept)      -3.776896   0.993251
dcpct             0.022618   0.007723
dneoYes           3.184002   0.837199
dcloxYes          0.445705   1.026026
dneoYes:dcloxYes -2.551997   1.205075
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01

(Dispersion parameter for binomial fami

    Null deviance: 149.72  on 107  degre
Residual deviance: 103.42  on 103  degre
AIC: 113.42

Number of Fisher Scoring iterations: 5

library(epiR)
Package epiR 0.9-45 is loaded
Type help(epi.about) for summary informa

mod3.mf <- model.frame(mod3)
(mod3.cp <- epi.cp(mod3.mf[-1]))
$cov.pattern
    id  n dcpct dneo dclox
1    1  7     0   No    No
2    2 38   100  Yes    No
3    3  1    25   No    No
4    4  1     1   No    No
5    5 11   100   No   Yes
8    6  1    25  Yes   Yes
10   7  1    14  Yes    No
12   8  4    75  Yes    No
13   9  1    90  Yes   Yes
14  10  1    30   No    No
15  11  3     5  Yes    No
17  12  9   100  Yes   Yes
22  13  2    20  Yes    No
23  14  8   100   No    No
25  15  2    50  Yes   Yes
26  16  1     7   No    No
27  17  4    50  Yes    No
28  18  1    50   No    No
```

```
31  19  1    30  Yes    No
34  20  1    99   No    No
35  21  1    99  Yes   Yes
40  22  1    80  Yes   Yes
48  23  1     3  Yes    No
59  24  1     1  Yes    No
77  25  1    10   No    No
84  26  1    83   No   Yes
85  27  1    95  Yes    No
88  28  1    99  Yes    No
89  29  1    25  Yes    No
105 30  1    40  Yes    No

$id
  [1]  1  2  3  4  5  1  1  6  5  7  5
 [26] 16 17 18  1  2 19  2 14 20 21 12
 [51] 14 12 11  5 15  2  8  2 24  2  2
 [76]  2 25  2 17  2  2  2  2 26 27 13
[101]  2  2  2  2 30  2  2  5
```

There are 30 covariate patterns in the dataset. The pattern dcpct=100, dneo=Yes, dclox=No appears 38 times.

Pearson and deviance residuals
Residuals represent the difference between the data and the model. The Pearson residuals are comparable to standardized residuals used for linear regression models. Deviance residuals represent the contribution of each observation to the overall deviance.

```
residuals(mod3) # deviance residuals
residuals(mod3, "pearson") # pearson res
```

Goodness-of-fit test
All goodness-of-fit tests are based on the premise that the data will be divided into subsets and within each subset the predicted number of outcomes will be computed and compared to the observed number of outcomes. The Pearson χ^2 and the deviance χ^2 are based on dividing the data up into the natural covariate patterns. The Hosmer-Lemeshow test is based on a more arbitrary division of the data.

The Pearson χ^2 is similar to the residual sum of squares used in linear models. It will be close in size to the deviance, but the model is fit to minimize the deviance and not the Pearson χ^2. It is thus possible even if unlikely that the χ^2 could increase as a predictor is added to the model.

```
sum(residuals(mod3, type = "pearson")^2)
[1] 123.9656
deviance(mod3)
[1] 103.4168
1 - pchisq(deviance(mod3), df.residual(mo
[1] 0.4699251
```

The p-value is large indicating no evidence of lack of fit. However, when using the deviance statistic to assess the goodness-of-fit for a nonsaturated logistic model, the χ^2 approximation for the likelihood ratio test is questionable. When the covariate pattern is almost as large as N, the deviance cannot be assumed to have a χ^2 distribution.
Now the Hosmer-Lemeshow test, usually dividing by 10 the data:

```
hosmerlem <- function (y, yhat, g = 10)
+   cutyhat <- cut(yhat, breaks = quant
+                        include.lowest = TRU
+   obs <- xtabs(cbind(1 - y, y) ~ cutyh
+   expect <- xtabs(cbind(1 - yhat, yha
+   chisq <- sum((obs - expect)^2 / exp
+   P <- 1 - pchisq(chisq, g - 2)
+   c("X^2" = chisq, Df = g - 2, "P(>Ch
+ }
hosmerlem(y = nocardia$casecont, yhat =
Erreur dans cut.default(yhat, breaks =
  'breaks' are not unique
```

The model used has many ties in its predicted probabilities (too few covariate values?) resulting in an error when running the Hosmer-Lemeshow test. Using fewer cut-points (g = 5 or 7) does not solve the problem. This is a typical example when not to use this test. A better goodness-of-fit test than Hosmer-Lemeshow and Pearson / deviance χ^2 tests is the le Cessie – van Houwelingen – Copas – Hosmer unweighted sum of squares test for global goodness of fit (also here) implemented in the rms package (but you have to implement your model with the lrm function of this package):

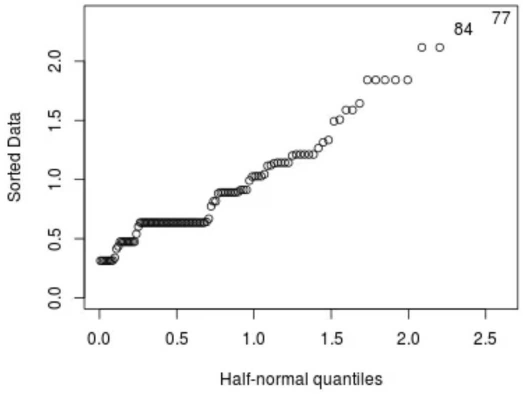
```
mod3b <- lrm(casecont ~ dcpct + dneo + dc
+                method = "lrm.fit", model
+                linear.predictors = TRUE,
residuals(mod3b, type = "gof")
Sum of squared errors     Expected value
           16.4288056            16.8235
                    Z
           -1.4219860             0.1550
```

The p-value is 0.16 so there's no evidence the model is incorrect. Even better than these tests would be to check for linearity of the predictors.

Overdispersion

Sometimes we can get a deviance that is much larger than expected if the model was correct. It can be due to the presence of outliers, sparse data or clustering of data. The approach to deal with overdispersion is to add a dispersion parameter σ^2. It can be estimated with: $\hat{\sigma}^2 = \frac{\chi^2}{n-p}$ (p = probability of success). A half-normal plot of the residuals can help checking for outliers:

```
library(faraway)
halfnorm(residuals(mod1))
```



Half-normal plot of the residuals

The dispesion parameter of model 1 can be found as:

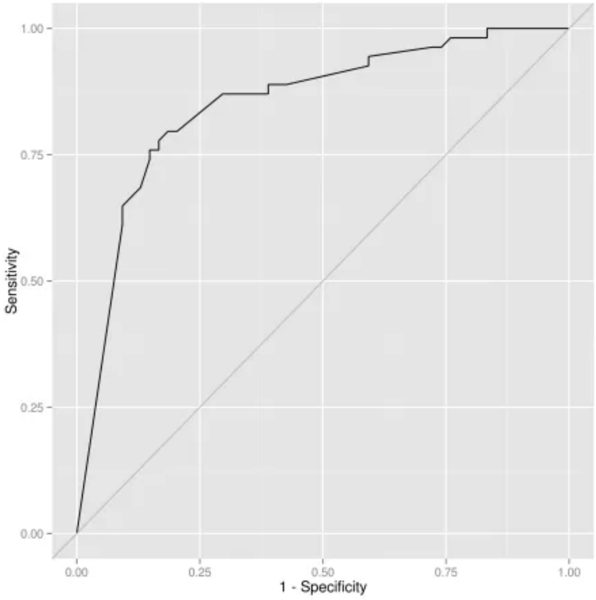
```
(sigma2 <- sum(residuals(mod1, type = "p
[1] 1.128778
drop1(mod1, scale = sigma2, test = "F")
Single term deletions

Model:
casecont ~ dcpct + dneo + dclox

scale:  1.128778

       Df Deviance    AIC F value    Pr
<none>      107.99 115.99
dcpct   1   119.34 124.05 10.9350  0.001
dneo    1   125.86 129.82 17.2166 6.834
dclox   1   114.73 119.96  6.4931  0.01
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01
Message d'avis :
In drop1.glm(mod1, scale = sigma2, test
  le test F implique une famille 'quasib
```

The dispersion parameter is not very different than one (no dispersion). If dispersion was present, you could use it in the F-tests for the predictors, adding scale to drop1.

Predictive ability of the model

A ROC curve can be drawn:

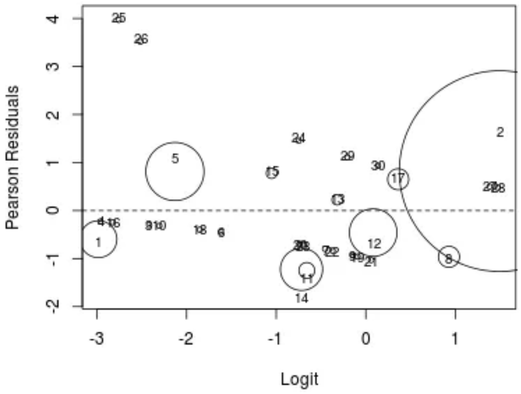
```
predicted <- predict(mod3)
library(ROCR)
prob <- prediction(predicted, nocardia$c
+                      label.ordering = c
tprfpr <- performance(prob, "tpr", "fpr"
tpr <- unlist(slot(tprfpr, "y.values"))
fpr <- unlist(slot(tprfpr, "x.values"))
roc <- data.frame(tpr, fpr)
ggplot(roc) + geom_line(aes(x = fpr, y =
+    geom_abline(intercept = 0, slope = 1
+   ylab("Sensitivity") +
+   xlab("1 - Specificity")
```



ROC curve

Identifying important observations
Like for linear regression, large positive or negative standardized residuals allow to identify points which are not well fit by the model. A plot of Pearson residuals as a function of the logit for model 1 is drawn here, with bubbles relative to size of the covariate pattern. The plot should be an horizontal band with observations between -3 and +3. Covariate patterns 25 and 26 are problematic.

```
nocardia$casecont.num <- as.numeric(noca
mod1 <- glm(casecont.num ~ dcpct + dneo
+             data = nocardia) # "logit'
mod1.mf <- model.frame(mod1)
mod1.cp <- epi.cp(mod1.mf[-1])
nocardia.cp <- as.data.frame(cbind(cpid
+                                   noc
+                                   fit
### Residuals and delta betas based on c
mod1.obs <- as.vector(by(as.numeric(noca
+                         as.factor(noc
mod1.fit <- as.vector(by(nocardia.cp$fit
+                         FUN = min))
mod1.res <- epi.cpresids(obs = mod1.obs,
+                         covpattern =

mod1.lodds <- as.vector(by(predict(mod1
+                           FUN = min)

plot(mod1.lodds, mod1.res$spearson,
+       type = "n", ylab = "Pearson Resid
text(mod1.lodds, mod1.res$spearson, labe
symbols(mod1.lodds, mod1.res$pearson, ci
```



Bubble plot of standardized residuals

The hat matrix is used to calculate leverage values and other diagnostic parameters. Leverage measures the potential impact of an observation. Points with high leverage have a potential impact. Covariate patterns 2, 14, 12 and 5 have the largest leverage values.

```
mod1.res[sort.list(mod1.res$leverage, de
cpid  leverage
2   0.74708052
14  0.54693851
12  0.54017700
5   0.42682684
11  0.21749664
1   0.19129427
...
```

Delta-betas provides an overall estimate of the effect of the j^{th} covariate pattern on the regression coefficients. It is analogous to Cook's distance in linear regression. Covariate pattern 2 has the largest delta-beta (and represents 38 observations).

```
mod1.res[sort.list(mod1.res$sdeltabeta, (
cpid sdeltabeta
2    7.890878470
14   3.983840529
...
```

Related

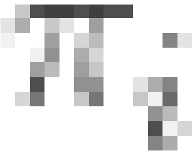
$(75 - 50) \times 0.022 = 0.55$

Veterinary Epidemiologic Research: GLM – Logistic Regression (part 2)

In "R bloggers"

Logistic Regression with R

In "R bloggers"



Veterinary Epidemiologic Research: GLM – Logistic Regression

In "R bloggers"

To **leave a comment** for the author, please follow the link and comment on their blog: **denis haine » R**.

Comments are closed.

Top 3 Posts from the past 2 days

- James Bond movies
- Correlation and Linear Regression
- Linear model with time series random component

Search & Hit Enter

Top 9 articles of the week

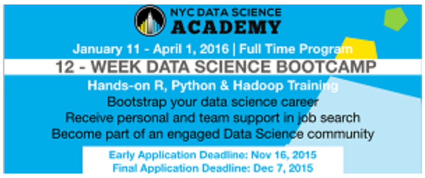
1. Installing R packages

2. The Data Science Industry: Who Does What (Infographic)
3. In-depth introduction to machine learning in 15 hours of expert videos
4. magrittr: The best thing to have ever happened to R?
5. Using apply, sapply, lapply in R
6. How to Make a Histogram with Basic R
7. Introducing Distributed Data-structures in R
8. James Bond movies
9. Adding a legend to a plot

Sponsors

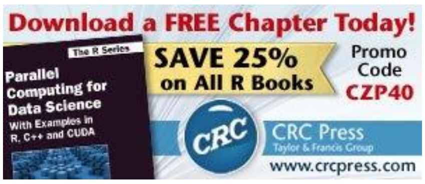
plotly online R graphing

STATISTICS VIEWS
Bringing Statistics Together


NYC DATA SCIENCE ACADEMY
January 11 - April 1, 2016 | Full Time Program
12 - WEEK DATA SCIENCE BOOTCAMP
Hands-on R, Python & Hadoop Training
Bootstrap your data science career
Receive personal and team support in job search
Become part of an engaged Data Science community
Early Application Deadline: Nov 16, 2015
Final Application Deadline: Dec 7, 2015

DataCamp
Learn R Interactively

#ODSC
50 Talks
14 Workshops
6 Pre-conference Workshops
SAN FRANCISCO | NOV. 14TH - 15TH


Download a FREE Chapter Today!
The R Series
Parallel Computing for Data Science
With Examples in R, C++ and CUDA
SAVE 25% on All R Books
Promo Code CZP40
CRC Press
Taylor & Francis Group
www.crcpress.com

STATWORX
Consulting
Schulung
Data Mining
Mehr erfahren

Try the FASTEST ML for R
Click for a Free Trial
YOTTAMINE ANALYTICS

Become a
Certified Data Scientist
ENROLL NOW
simplilearn

Jobs for R users

- Looking for a Freelance R-Developer with Shiny experience
- Data Analyst @ New York
- R/Shiny App with d3 (small job, quick turnaround, $250 < 4hrs)
- Postdoctoral position (Duke’s River center) in data visualization and ecosystem science @ Durham, North Carolina, United States
- Software Engineer for The Computational Biology Program at Oregon Health and Science University @ Portland, Oregon, United States
- Data Scientist @ Israel
- Jr Financial Engineer

Search & Hit Enter

Full list of contributing R-bloggers

R-bloggers was founded by Tal Galili, with gratitude to the R community.
Is powered by WordPress using a bavotasan.com design.