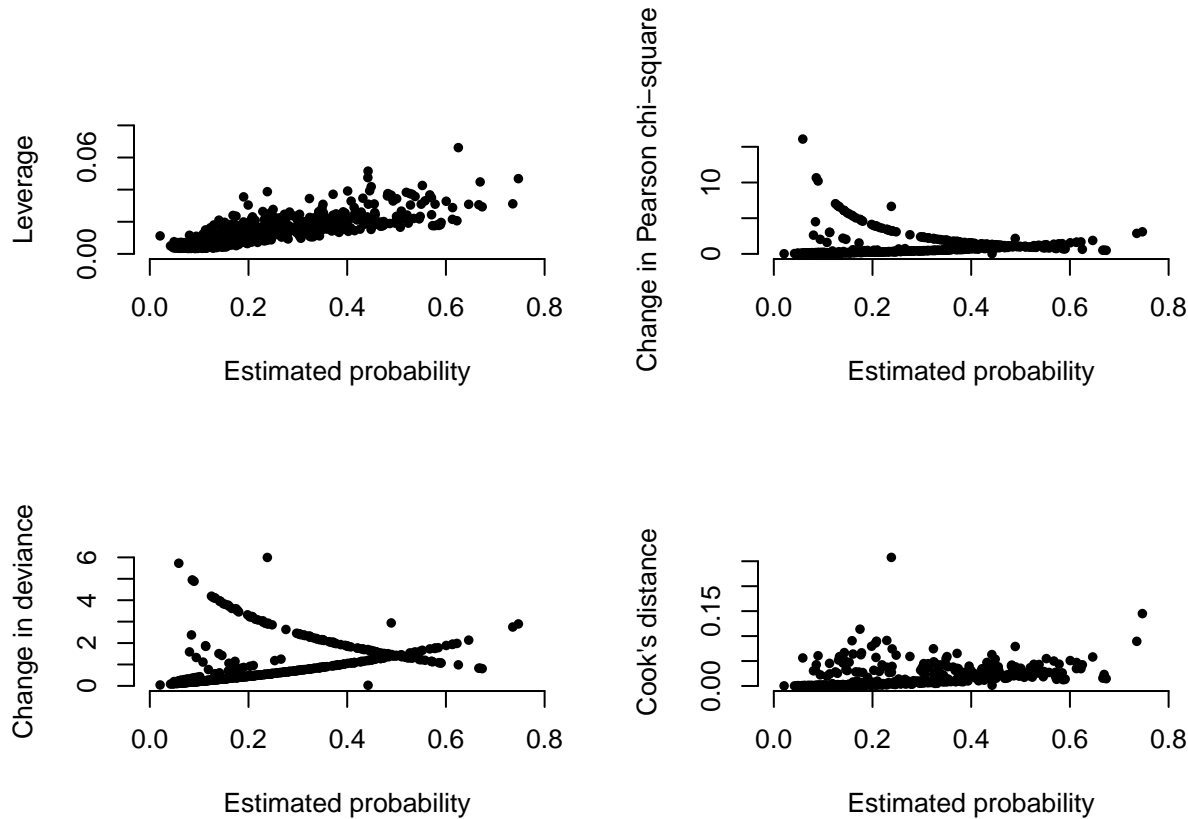


Inconsistencies in the reproduction of figures 5.12 and 5.15 of Hosmer et al. 2013

Johann Popp

29 October 2017

Hosmer et al¹ have suggested several graphics to identify and investigate extreme and influencing covariate patterns in logistic regression models. These plots were reproduced using the computer software R², taking data from package `aplore3`³ and using package `epiR`⁴ for the calculation of statistics based on covariate patterns rather than single data rows.



This works fine for the graphs showing change of Pearson chi-square and change of deviance but the plots of leverage vs. fitted values and Cook's distance vs. fitted values differ substantially from figures 5.12 and 5.15 of the book.

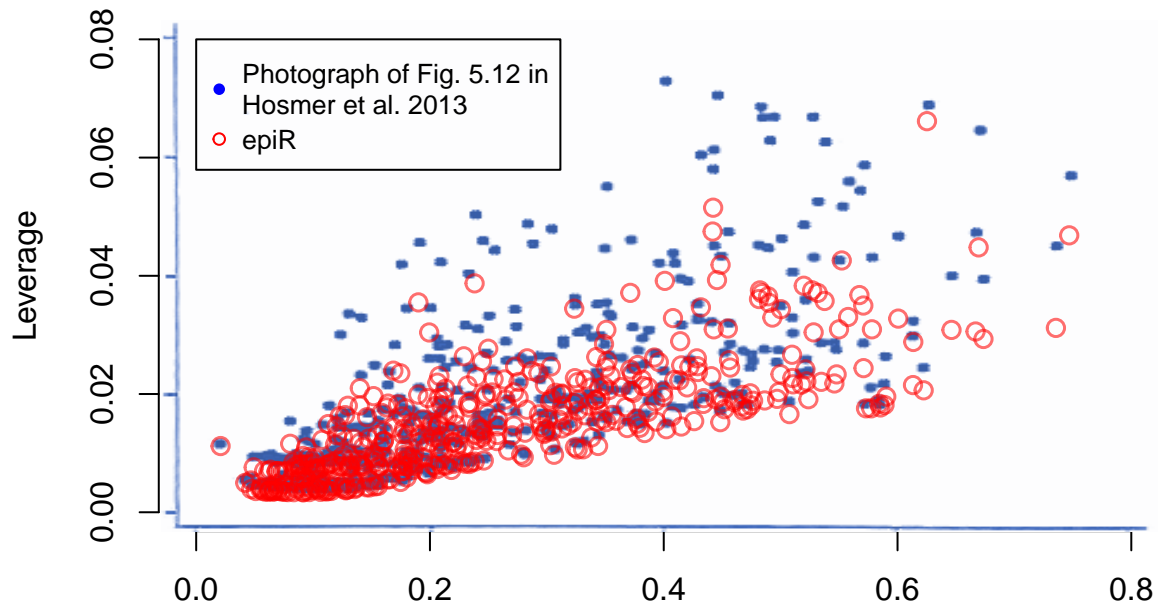
¹Hosmer, David W., Stanley Lemeshow, und Rodney X. Sturdivant. Applied logistic regression. 3rd Ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley, 2013, p 186 ff.

²R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

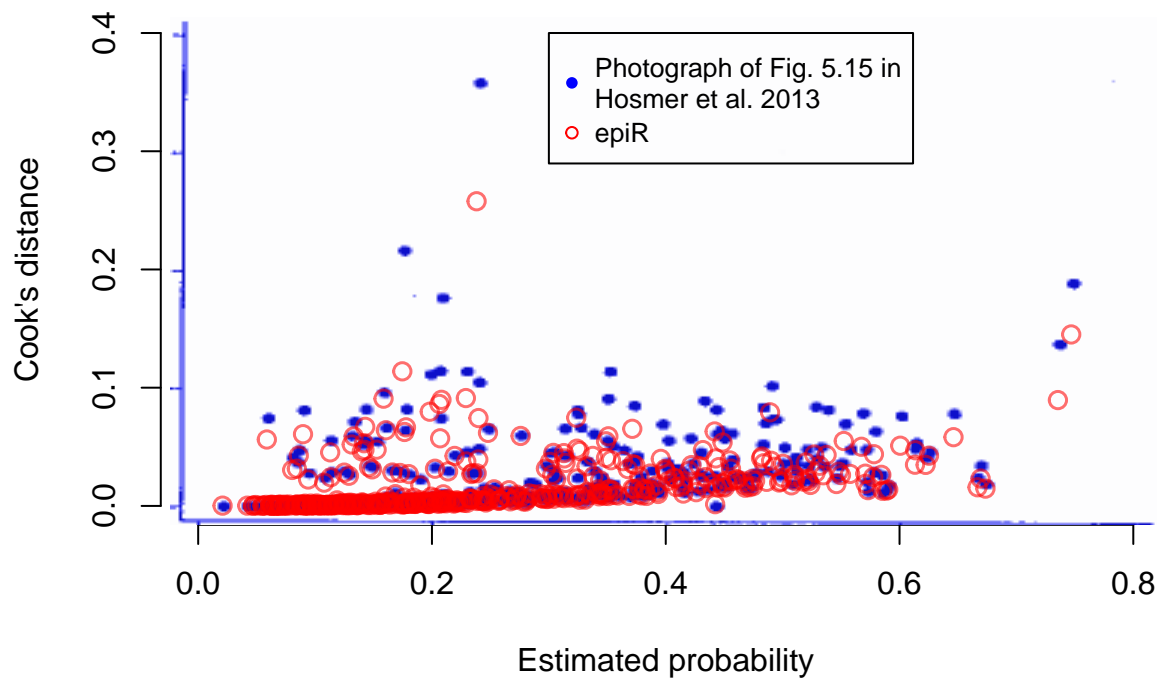
³Luca Braglia (2016). `aplore3`: Datasets from Hosmer, Lemeshow and Sturdivant, "Applied Logistic Regression" (3rd Ed., 2013). R package version 0.9. <https://CRAN.R-project.org/package=aplore3>

⁴Mark Stevenson with contributions from Telmo Nunes, Cord Heuer, Jonathon Marshall, Javier Sanchez, Ron Thornton, Jenő Reiczig, Jim Robison-Cox, Paola Sebastiani, Peter Solymos, Kazuki Yoshida, Geoff Jones, Sarah Pirikahu, Simon Firestone and Ryan Kyle. (2017). `epiR`: Tools for the Analysis of Epidemiological Data. R package version 0.9-87. <https://CRAN.R-project.org/package=epiR>

Leverage based on package epiR vs. published figure 5.12

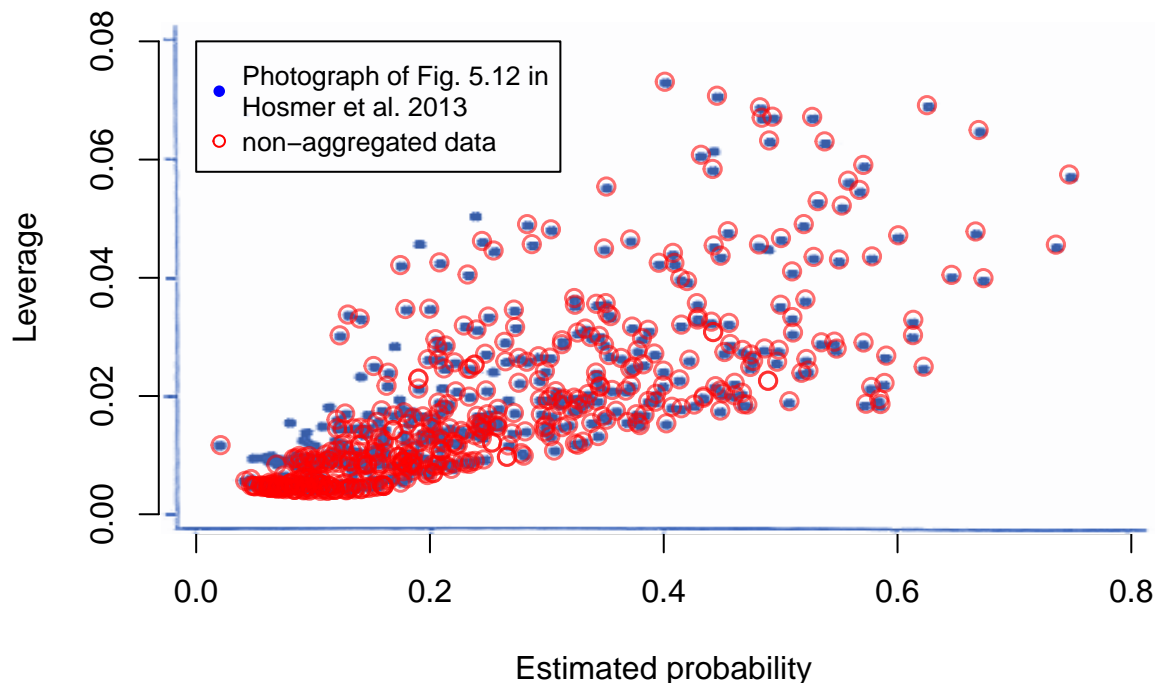


Cook's distance based on package epiR vs. published figure 5.15



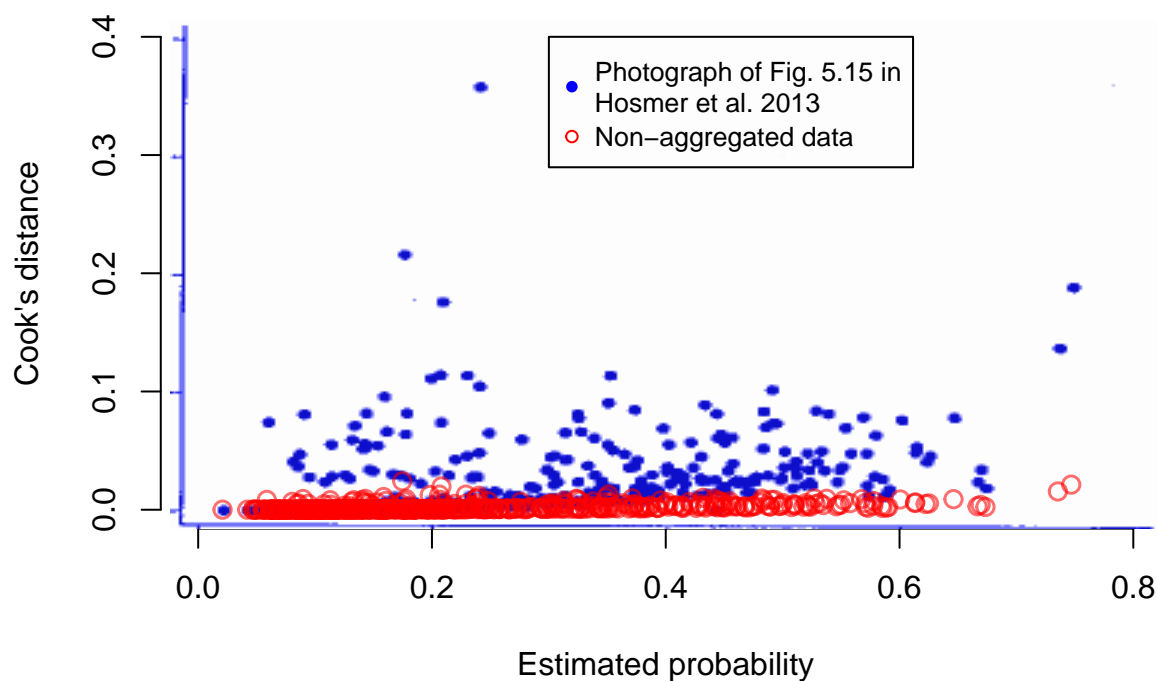
With leverages based on non-aggregated n-statistics instead of covariate patterns (m-statistics) the plot can be roughly reproduced, but there are still some points that do not match.

Leverage based on non-aggregated data vs. published figure 5.12



But this remedy does not work for Cook's distance. The values based on n-statistics are completely different from those printed in the book and those calculated by epiR in magnitude and pattern.

Cook's distance based non-aggregated vs. published figure 5.15

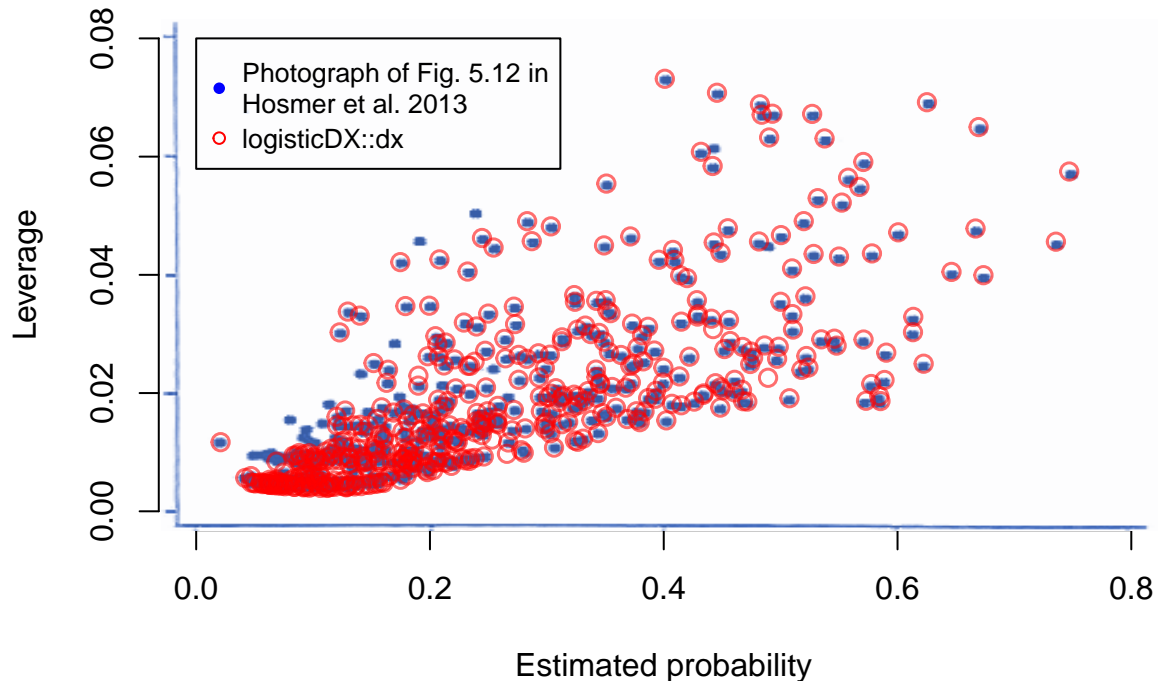


I tried out the newer package `logisticDx`⁵ to calculate the diagnostic statistics. This package specializes on

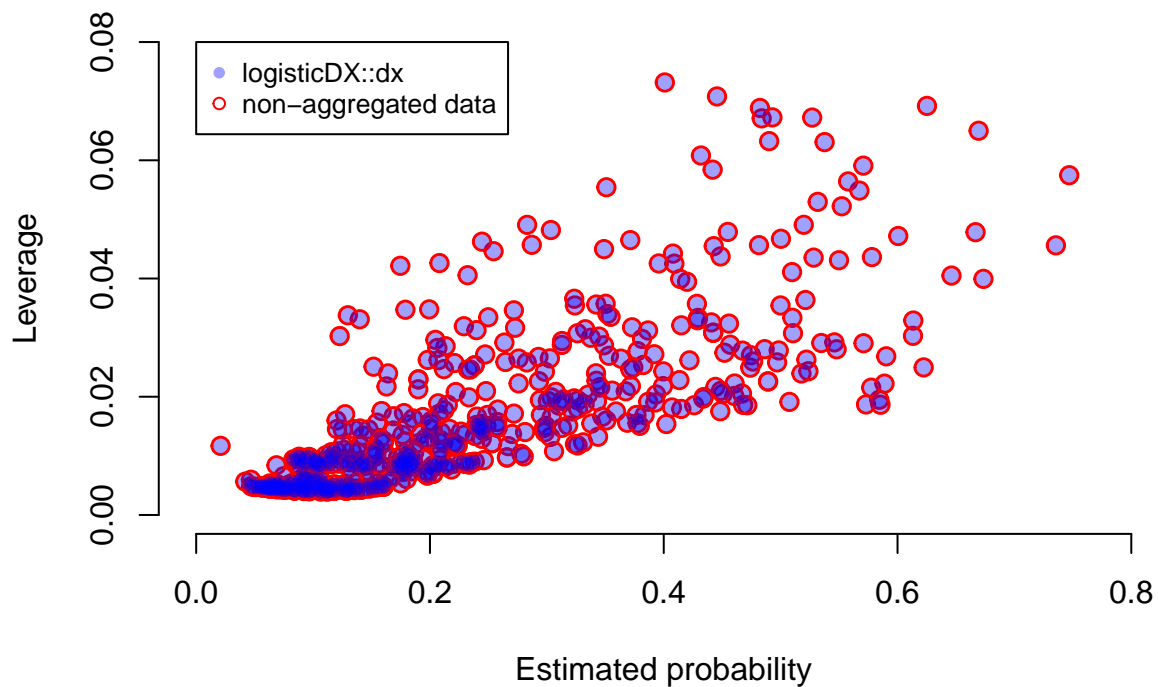
⁵Chris Dardis (2015). `LogisticDx`: Diagnostic Tests for Models with a Binomial Response. R package version 0.2. <https://github.com/cdardis/logisticDx>

diagnostic tests for regression models with binomial response and is explicitly based on the book of Hosmer et al. albeit in its second edition from 2000.

Leverage based on package `logisticDX` vs. published figure 5.12



Leverage based on `logisticDX::dx` vs. original data



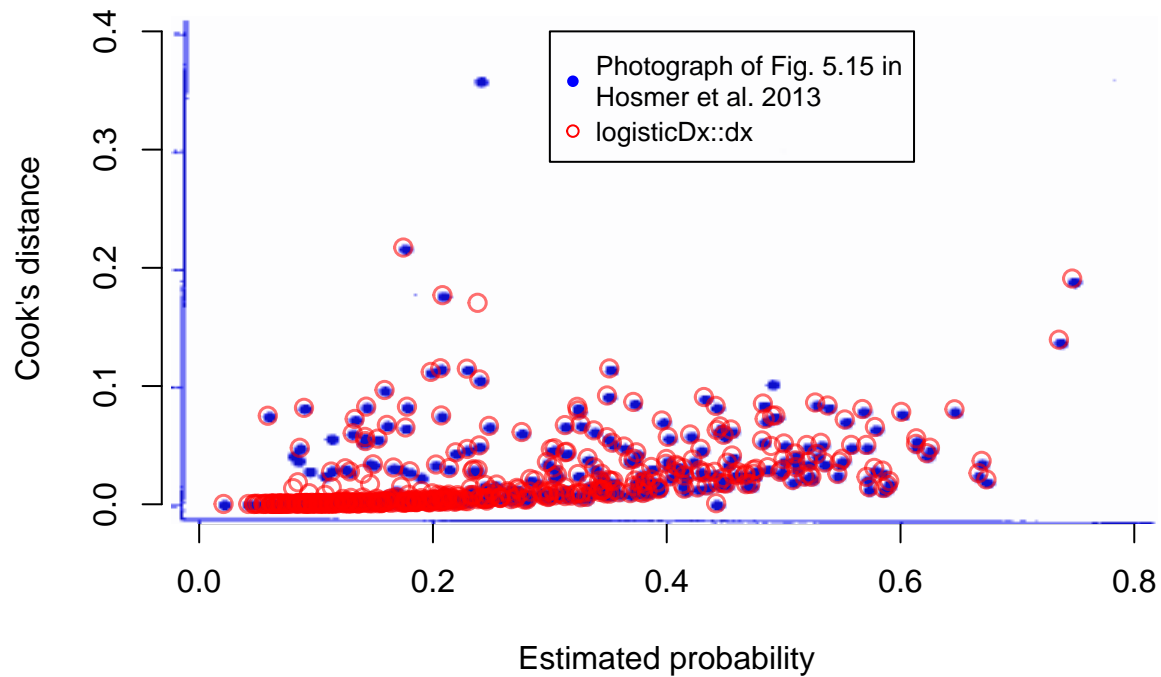
You can see, that the leverages of `logisticDX::dx` are exactly the same as those calculated from non-aggregated data (n-statistics). We have seen before that these are matching figure 5.12 quite well but not with all the

[//CRAN.R-project.org/package=LogisticDx](https://CRAN.R-project.org/package=LogisticDx)

points.

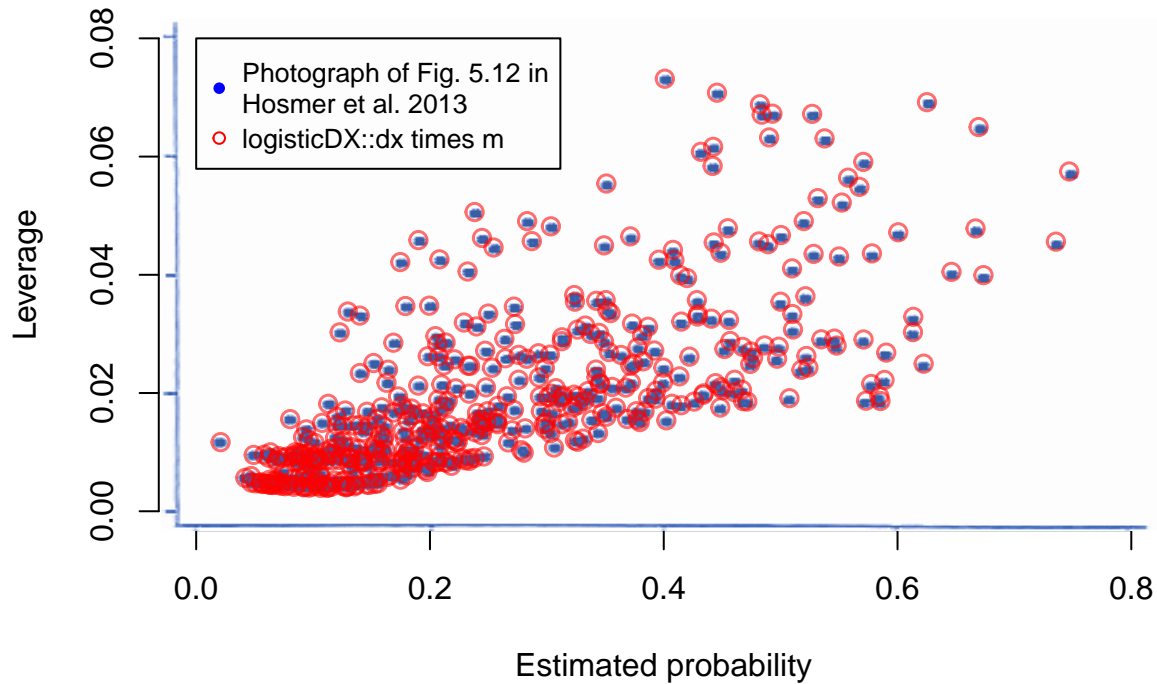
When they are calculated with `logisticDX::dx`, also the values of Cook's distance are fitting well with figure 5.15 despite of some single cases. I guess these are the same as those that do not fit in leverage.

Cook's distance based on package `logisticDx` vs. published figure 5.



Let's see what happens if we multiply the leverage times m , the number of cases in the covariate pattern:

Leverage based on package `logisticDX` times `m` vs. published figure 5.12



This fits perfectly. But the Stata manual⁶ gives an other formula:

The diagonal $h_j = (XVX')_{jj}$ times $M_j p_j(1 - p_j)$

where X is the design matrix for the j covariate patterns and V is the covariance matrix, M is the number of cases in covariate pattern j , p is the estimated outcome probability for covariate pattern j .

Leverage following Stata formula

Extract the design matrix

```
X <- model.matrix(model)
```

Extract data used by the model

```
datN <- model$model
```

aggregate data to covariate patterns

```
cp <- unique(datN[, -1])
```

Extract an indicator of covariate patterns by pasting the values of all the x-values

```
cp$cpid <- apply(cp, 1, function(x) paste(x, collapse = ""))
```

Get the estimated y-probability for each covariate pattern

```
cp$fit <- model$fitted.values[rownames(cp)]
```

Get n-based hat values for each covariate pattern

```
cp$H <- hatvalues(model)[rownames(cp)]
```

create a covariate pattern indicator for the non-aggregated data (Same method as for covariate pattern)

```
cpid <- apply(datN[, -1], 1, function(x) paste(x, collapse = ""))
```

calculate group size for each covariate pattern

```
m <- tapply(datN[, 1], cpid, length)[rank(cp$cpid)]
```

(tapply sorts the entries by cpid; [rank(cp\$cpid)] sorts back to the original order)

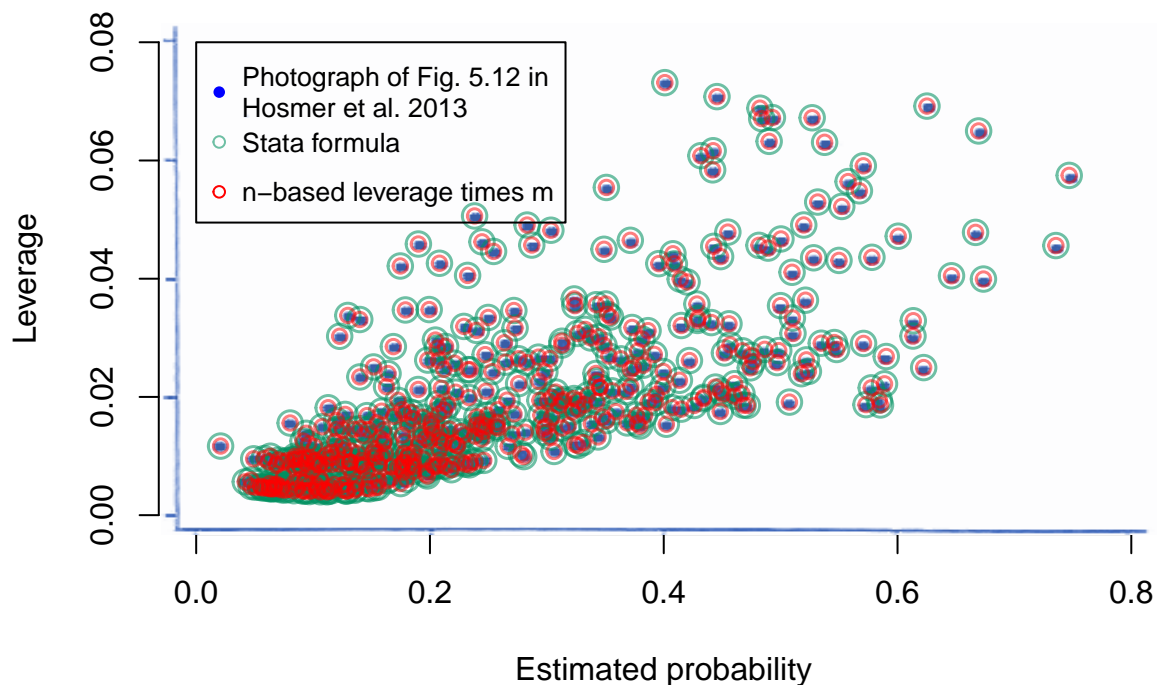
⁶“Logistic Postestimation - Postestimation Tools for Logistic.” In Base Reference Manual, Release 15. College Station, Texas: Stata Press, 2017. <https://www.stata.com/bookstore/base-reference-manual/>.

```
# Aggregate the design matrix to covariate patterns
Xu <- unique(X)
# Extract the variance-covariance matrix of the model
V <- vcov(model)
```

```
### This is the formula described in the Stata reference manual
```

```
# Calculate the raw hat values
rawH <- diag(Xu %*% V %*% t(Xu))
# Adjust for covariate pattern
H <- rawH * m * cp$fit * (1-cp$fit)
```

Leverage based on Stata formula vs. published figure 5.12

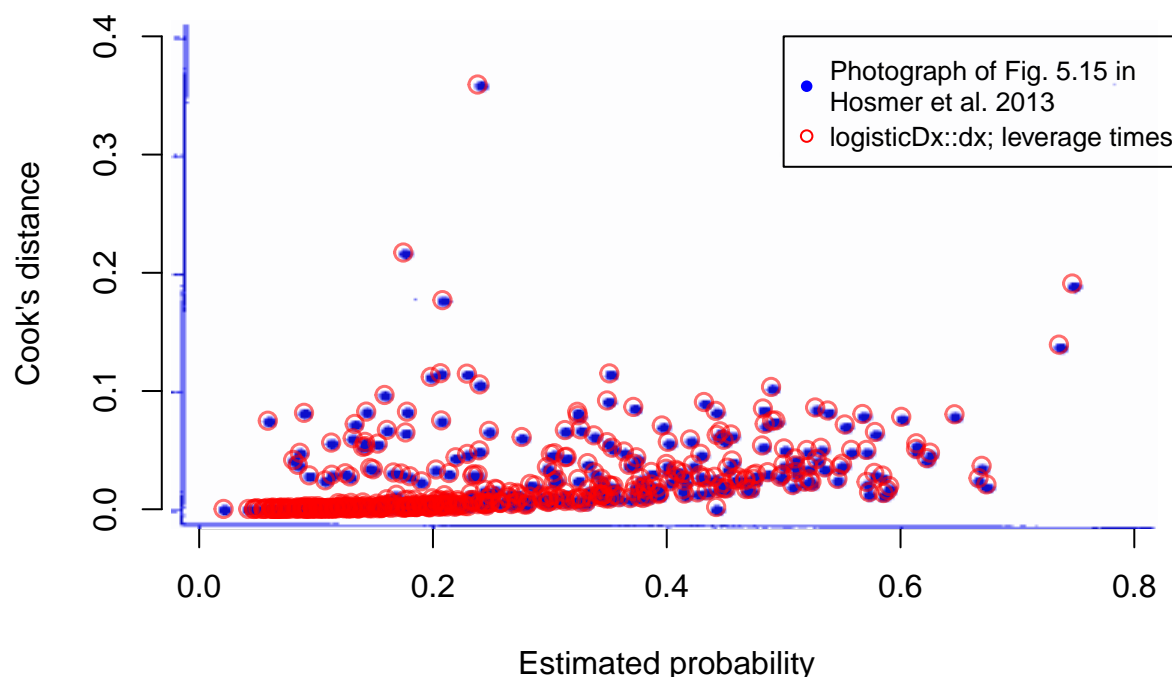


You can see that the result is exactly the same the n-based leverages times the group size for each covariate pattern. I would be delighted if somebody who knows more about matrix algebra could prove this instead of only showing it by example as I did.

I also calculated Cook's distance based on these leverages:

```
nLev <- ldx$h*ldx$n # Leverage times n ("ldx is the dx-object i have created from the
nsPr <- ldx$Pr / sqrt(1 - nLev) # New standardized Pearson residuals
nCook <- (nsPr^2 * nLev) / (1 - nLev) # New Cook's distance
```

Cook's distance based on package logisticDx with leverage times m vs. published figure 5.15



This looks pretty good as well.

I will work out the same for change in Pearson chi-square and change in deviance soon.

Summary

There are two R packages `epiR` and `logisticDx` that provide functions to calculate diagnostics of outliers and influential covariate patterns for logistic regression models. None of them exactly reproduces the plots recommended by Hosmer et al. that were produced with Stata. Plots based on `epiR` differ substantially from the book, those based on `logisticDx` only marginally. Stata - the program used to produce the plots in the book - uses a different formula but it seems that you can multiply the single case leverages (that are routinely calculated by R) with the group size of the covariate pattern to get the same result.