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

Johann Popp

11 September 2017

Hosmer et al[1] have suggested several graphics to identify and investigate extreme and influencing covariate patterns in logistic regression models. These plots where reproduced using the computer software R[2], taking data from package aplore3[3] and using package epiR[4] for the calculation of statistics based on covariate patterns rather then single data rows.

```
par(mfrow = c(2,2))

# Load example data
glow <- aplore3::glow500

# Recode RATERISK
glow$raterisk3 <- cut(as.numeric(glow$raterisk), 2)
levels(glow$raterisk3) <- c("less/same", "greater")


# Logistic model from Table 4.16
model <- glm(fracture ~ age + height + priorfrac + momfrac + armassist + raterisk3 + age:priorfrac + mo

### Convert to covariate patterns
library(epiR)

# aggregate to covariate pattern
cp <- epi.cp(model.frame(model)[-1])
# Number of outcome events per covariate pattern
obs <- as.vector(by(as.numeric(as.factor(model$model[,1]))-1, as.factor(cp$id), sum))
# Estimated outcome probability per covariate pattern
fit <- as.vector(by(model$fitted.values, as.factor(cp$id), min))
# Calculate residuals and influence measures per covariate pattern
res <- epi.cpresids(obs, fit, cp)
# Calculate Change in deviance
deltaDeviance <- res$deviance^2 + ((res$pearson^2*res$leverage)/(1 - res$leverage))
# Put it together in one data.frame
dat <- data.frame(res, deltaDeviance, fit, cp$cov.pattern)


# Plot leverage vs. fit
plot(dat$fit, dat$leverage, pch = 16, cex = 0.8, xlab = "Estimated probability", ylab = "Leverage", xli
```

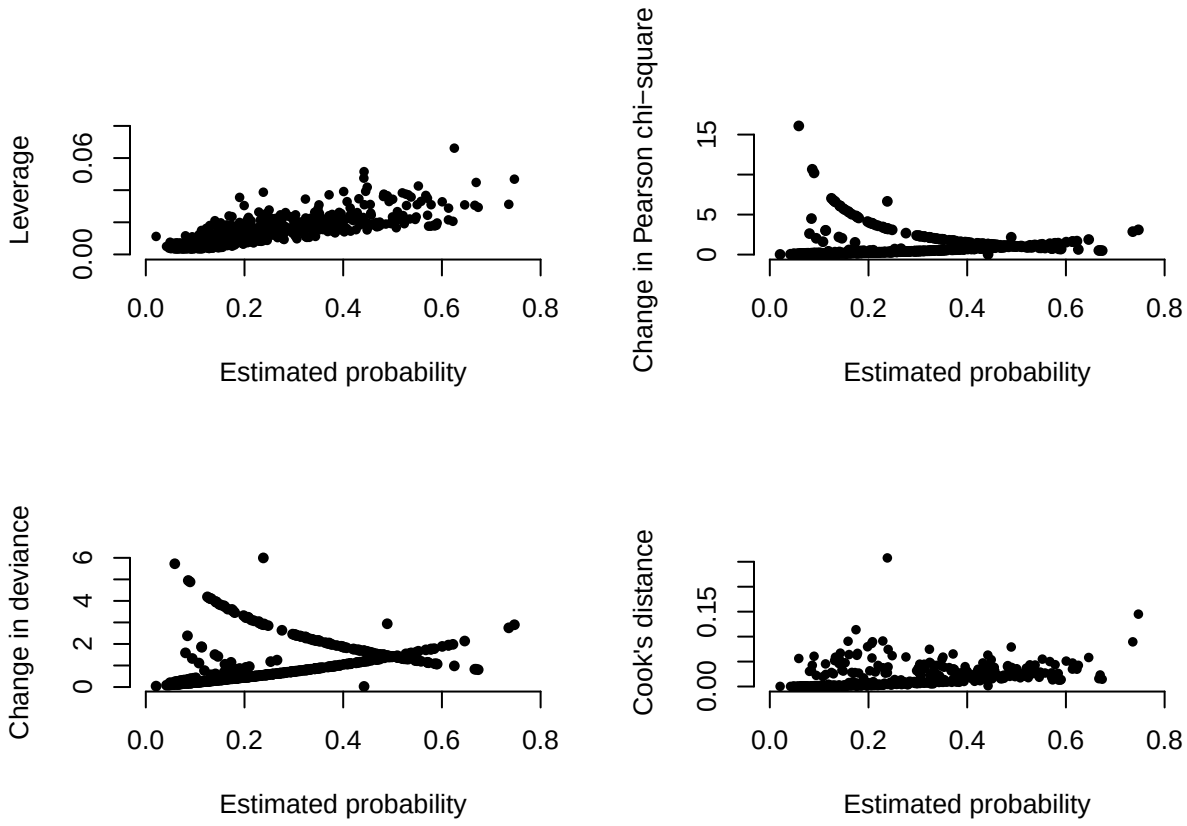
[1]Hosmer, David W., Stanley Lemeshow, und Rodney X. Sturdivant. Applied logistic regression. 3rd Ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley, 2013, p 186 ff.

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

[3]Luca 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

[4]Mark Stevenson with contributions from Telmo Nunes, Cord Heuer, Jonathon Marshall, Javier Sanchez, Ron Thornton, Jeno Reiczigel, 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

```r
# Plot change in Chi-square vs. fit.
plot(dat$fit, dat$deltachi, pch = 16, cex = 0.9, xlab = "Estimated probability", ylab = "Change in Pears

# Plot change in deviance vs. fit.
plot(dat$fit, dat$deltaDeviance, pch = 16, cex = 0.9, xlab = "Estimated probability", ylab = "Change in

# Plot Cook's distans vs. fit
plot(dat$fit, dat$deltabeta, pch = 16, cex = 0.8, xlab = "Estimated probability", ylab = "Cook's distan
```

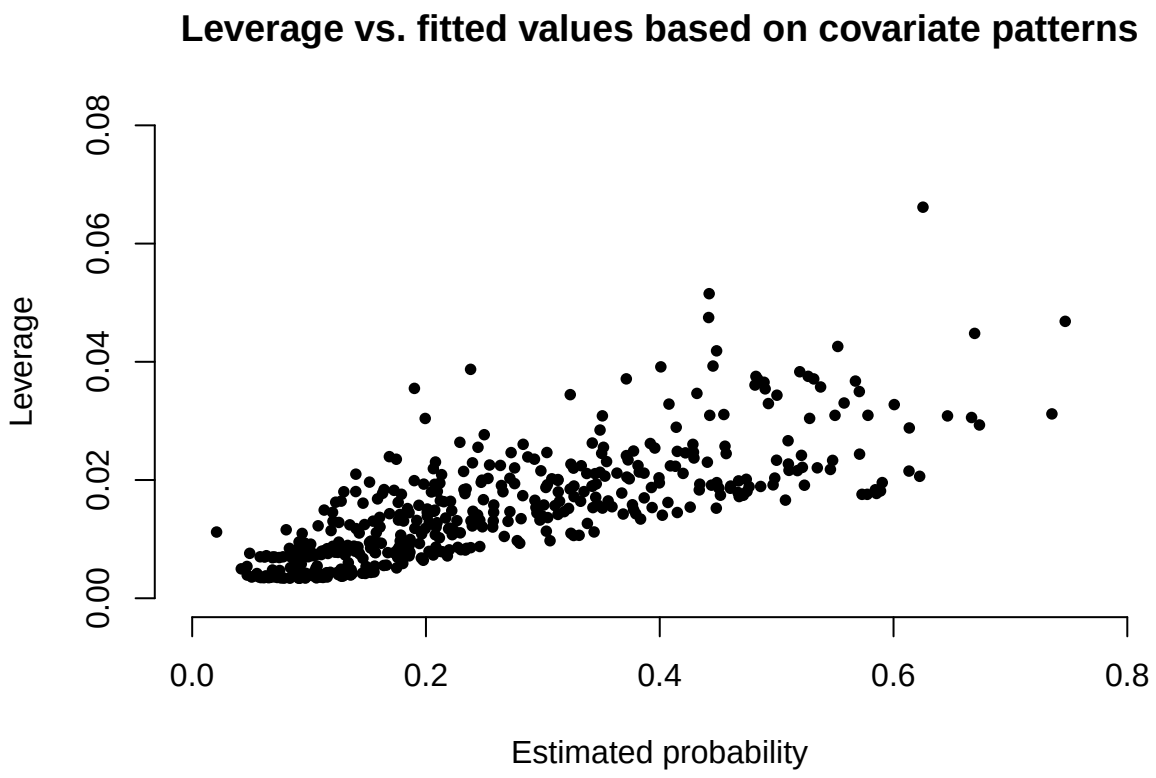



```r
par(mfrow = c(1,1))
```

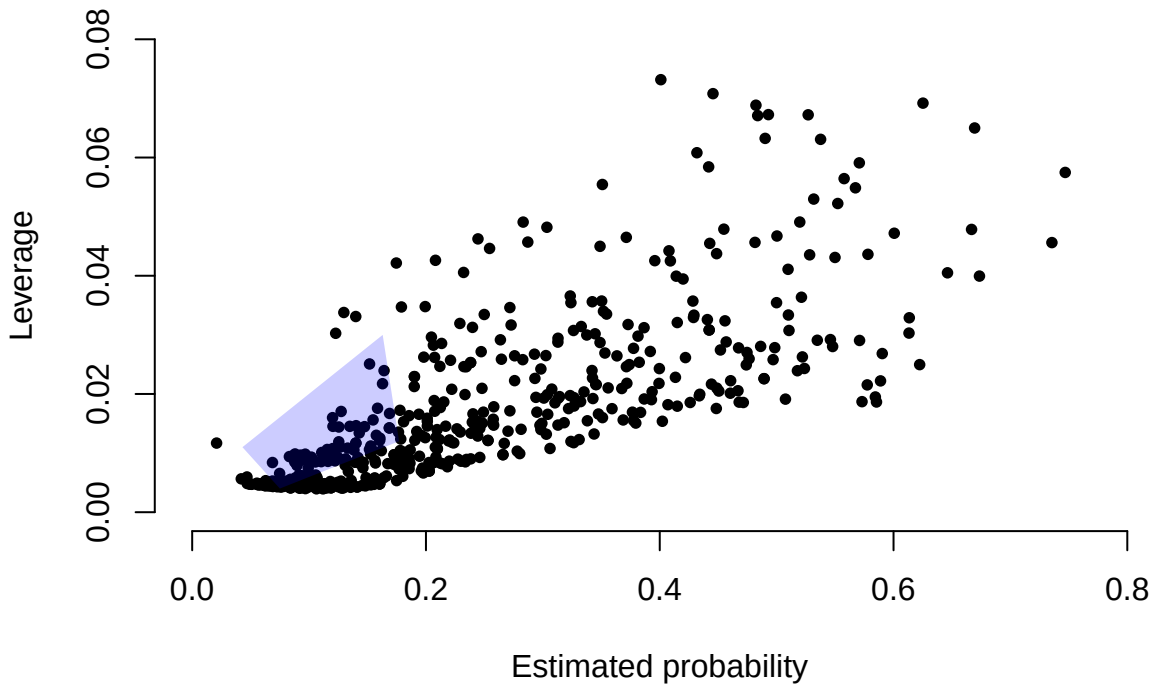
This works fine for the graphs showings change of Pearson chi-square and change of deviance but the plots of leverage vs. fitted values and Cook's distance vs. fitted values differ substantially from figures 5.12 and 5.15 of the book.

In contrast to what is said in the book, figure 5.12 of the book can be reproduced with a plot based on n-statistics (all the cases) instead of m-statistics (covariate patterns).

```r
# Plot leverage vs. fit for covariate patterns
plot(dat$fit, dat$leverage, pch = 16, cex = 0.8, xlab = "Estimated probability", ylab = "Leverage", mai
```



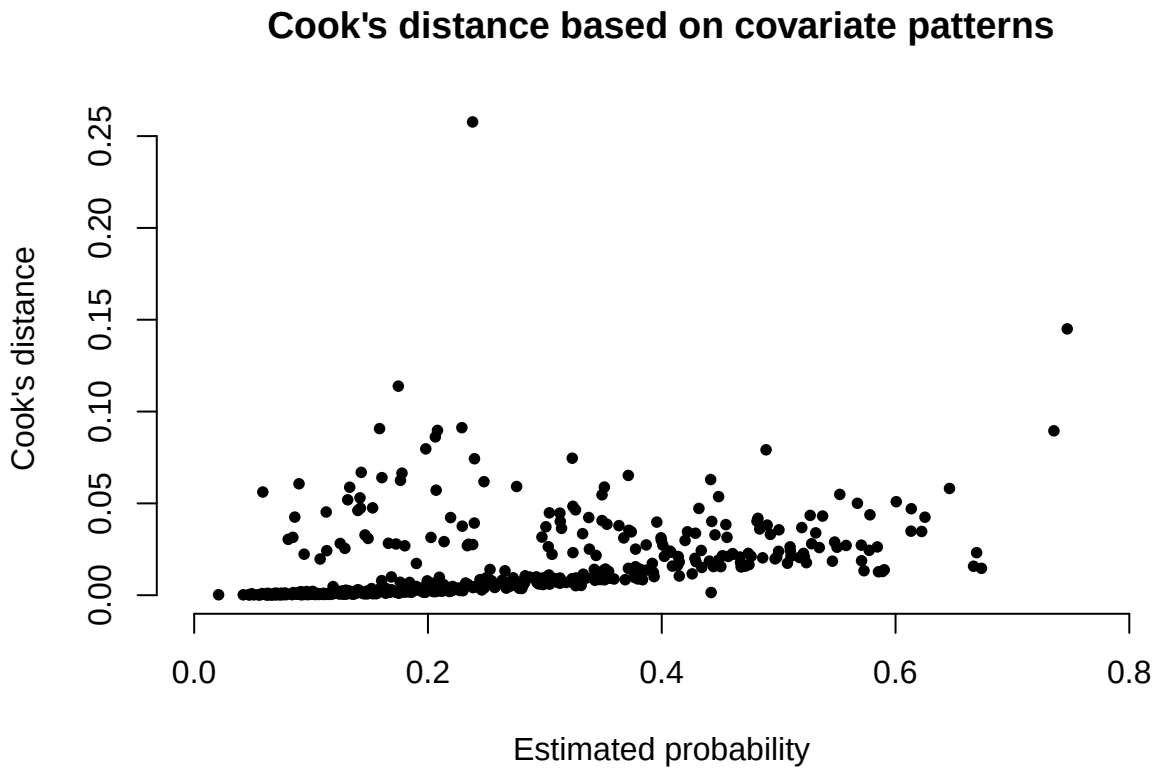

```
# Plot leverage vs. fit for each case
plot(fitted.values(model), hatvalues(model), pch = 16, cex = 0.8, xlab = "Estimated probability", ylab =

polygon(x = c(0.043, 0.075, 0.177, 0.163), y = c(0.011, 0.004, 0.012, 0.030), col = rgb(0,0,1,0.2), bor
```



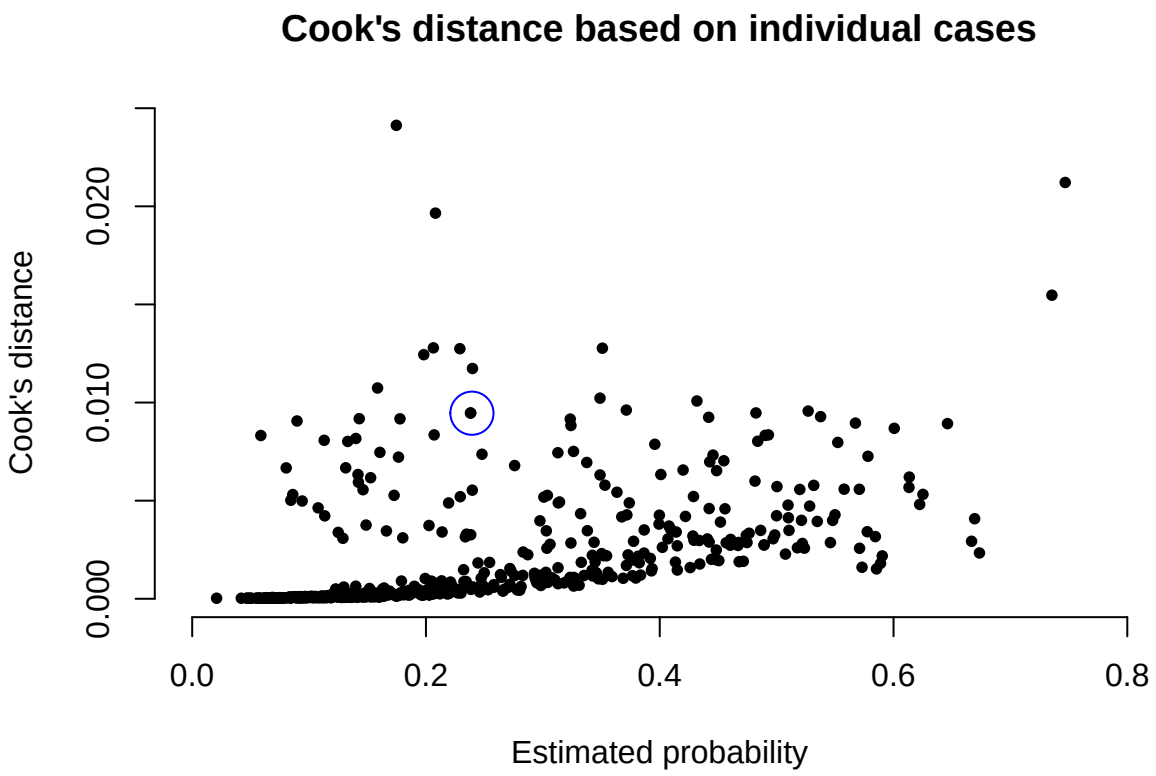
But still a close inspection of the highlighted area shows, that even this plot is not exactly the same as figure 5.12 in the book.

Regarding figure 5.15 some parts seem to be reproduced by a plot based on covariate patterns and other parts are quite well fitted by the reproduction based on individual cases. Overal the reproduced values of Cook's distance are substantially smaller than those shown in figure 5.15. Especially the highlightet point seems to be the one with a value of about 3.8 in the book.

```
# Plot Cook's distans vs. fit
plot(dat$fit, dat$deltabeta, pch = 16, cex = 0.8, xlab = "Estimated probability", ylab = "Cook's distanc
```




```
# Plot Cook's distans vs. fit
plot(fitted.values(model), cooks.distance(model), pch = 16, cex = 0.8, xlab = "Estimated probability", y

points(x = 0.2393311, y = 0.009456836, col = "blue", cex = 3)
```



As Cook's distance is based on leverage values it seems to me, that the calculation of those values in the package epiR differ from the calculation underlying the the figures of the book. As I am unfortunately not really into matrix algebra I can not effectively compare the syntax of the function epiR::epi.cpresids with formulas given in the book.