Vignette for package `blm`

Stephanie A. Kovalchik

April 3, 2013

SUMMARY

The `blm` package provides functions for fitting flexible binomial models for cohort studies of a binary outcome and population-based case-control studies. The binomial linear model (BLM) is a strictly linear model. The linear-expit (lexpit) model allows risk to be expressed as a function of linear and nonlinear effects, where nonlinear effects take the form of the inverse logit function (Kovalchik, 2013). Estimation of the model parameters is based on constrained maximum likelihood, which ensures that the fitted model yields feasible risk estimates. In this vignette, BLM and lexpit model fitting is demonstrated with analyses of a simulated population-based case-control study.

1 Binomial linear model (BLM)

1.1 Model

Given the binary event y_i, the probability that $Y_i = 1$ under a binomial linear model (BLM) (Kovalchik, 2013) is a linear function of covariates x_i,

$$\pi_i = x_i'\beta \tag{1}$$

Each β of nonconstant covariates represents the risk difference associated with a unit change in the given covariate, when all other factors are fixed.

Suppose that $\tilde{x}$ is the covariate pattern for a subject from the target population of the model whose risk we want to estimate. To be a valid risk, $\tilde{x}'\beta \in (0, 1)$. In general, we might not be able to specify all of the possible $\tilde{x}$ of our population. Instead, we make use of the x_i from our sample and require that all $x_i'\beta \in (0, 1)$. Thus, the set of covariate patterns of the study sample defines the *feasible region* for β.

To ensure that the estimates for β are within the region of feasibility, constrained maximum likelihood is used. Since the system of constraints are linear in the parameters, an adpative barrier algorithm (Lange, 2010) can be used to perform the constrained optimization as implemented by `constrOptim`. For cohort studies, the objective function is a penalized binomial log-likelihood with each π_i defined by Equation (1). For population-based case-control studies, the objective function is a penalized pseudo-likelihood where each control subject's contribution to the binomial likelihood is weighted by a sampling weight w_i that reflects their representativeness of the target population. By definition, each case's weight is $w_i = 1$.

1.2 Model fitting

As an illustration of the model syntax we consider a model to estimate the risk of disease based on a simulated population-based case control study. We begin the `R` session by loading the package, `blm`, and the dataset `ccdata`.

```
> library(blm)
> data(ccdata)
> names(ccdata)

[1] "female"   "packyear" "strata"   "y"        "w"

> table(ccdata$y)

  0   1
378 378
```

The sample consists of 756 subjects and case status is indicated by the variable `y`. There are two design variables, the `strata` and inverse sampling fractions `w` , and two candidate explanatory variables, which are an indicator for female gender, `female`, and a discrete variable, `packyear`, ndicating the number of pack-years smoked.

The syntax for `blm` is much like `lm`, consisting of `formula` and `data` arguments. For population-based estimates, we need to additionally include the design information on sample stratification and the sampling weights. The following code fits a population-based linear risk model with additive effects due to female gender and packyears.

```
> fit <- blm(y˜female+packyear, data = ccdata,
+                weight = ccdata$w,
+                strata = ccdata$strata)
> fit

y ˜ female + packyear
(Intercept)      female    packyear
 0.07048229  0.01110223  0.01588852

> summary(fit)

            Est.     Std. Err t-value  p-value
(Intercept) 0.070482 1.124622 0.062672 0.950044
female      0.011102 1.612392 0.006886 0.994508
packyear    0.015889 0.095005 0.167239 0.867227

Converged: TRUE
```

The method summary provides measures of variance and Wald tests of significance for each fitted parameter. Also, a logical object indicates whether convergence of the optimizaiton algorithm was achieved.

The variance-covariance is estimated using Taylor-linearization Deville (1999). The coefficients and variance-covariance can be extracted directly using coef and vcov.

```
> coef(fit)
```

```
(Intercept)      female    packyear
 0.07048229  0.01110223  0.01588852
```

```
> vcov(fit)
```

```
            (Intercept)       female     packyear
(Intercept)  1.26477489 -1.288921908 -0.032739888
female      -1.28892191  2.599807633  0.008376159
packyear    -0.03273989  0.008376159  0.009025932
```

Each regression coefficient for an explanatory variable of the BLM model provides an estimate of the adjusted risk difference associated with a unit increase in the given variable. Thus, the fitted model suggests that there is a 1.1% increased risk for females and a 15.9% increased absolute risk for every 10-year increase in cumulative pack-years smoked. To obtain confindence intervals for these parameters, we can use the confint method.

```
> confint(fit)
```

```
                  Est.      Lower     Upper
(Intercept) 0.07048229 -2.1337365 2.2747011
female      0.01110223 -3.1491278 3.1713323
packyear    0.01588852 -0.1703177 0.2020947
```

```
> confint(fit, parm="female")
```

```
           Est.     Lower    Upper
[1,] 0.01110223 -3.149128 3.171332
```

To assess the constraints imposed on the model, we can examine the barrier.value, which is one of the slots of the blm class.

```
> fit@barrier.value
```

```
[1] 0.04078781
```

Risk estimates near the boundary could be an indication of influential points or a poor-fitting model with BLM or lexpit. Boundary estimates for a given distance criterion can be obtained with the function which.at.boundary.

```
> which.at.boundary(fit)

No boundary constraints using the given criterion.

> which.at.boundary(fit, criter = 1e-3)

No boundary constraints using the given criterion.
```

In the above, we first use a default criterion of 1e-06 or 0.999999. In the second case, we provide a user-specified criterion. With either criterion, no estimate was at the boundary.

There are several functions to evaluate the BLM model fit. McFadden's R-squared, adjusted and unadjusted, provides a measure of the variability explained by the explanatory variables.

```
> Rsquared(fit)

$R2
[1] 0.1643529

$R2adj
[1] 0.1606472
```

Comparisons of observed and expected counts for the target population can be made with the function `EO` for expected and observed. A factor can also be supplied to compare the expected to observed within subgroups defined by the categorical variable.

```
> EO(fit)

               E   O     EtoO   lowerCI  upperCI
Overall 387.2257 378 1.024407 0.9261713 1.133062

> EO(fit, ccdata$female)

         E   O     EtoO   lowerCI  upperCI
1 192.5145 183 1.051992 0.9101015 1.216004
2 194.7112 195 0.998519 0.8677618 1.148979
```

When the number of covariate classes represented by the model are few, we can perform a goodness-of-fit test with Pearson's chi-squared statistic. This compares the observed to expected within each unique risk type defined by the model.

```
> gof.pearson(fit)
```

```
      E  O
1 17.22 22
2 23.27 16
3 79.42 59
4 72.60 86
5 23.25 24
6 18.30 13
7 83.50 78
8 69.65 80

 Chi-squared:  42.24285

 P-value:  1.646512e-07
```

This suggests a lack of fit at the population level. When a model has more than 10 covariate classes, the Hosemer-Lemeshow test should be used. For the blm function, this test is implemented by the function gof.

1.3 Mode of exposure's effect

Because of the lack of fit of the additive model, we might suspect the linear assumption for pack-years. We could assess the functional relationship between risk and pack-years graphically by plotting risk against exposure. Because we cannot observe risk, we use estimates of the crude risk within groups defined by the ordered covariate. To have a fairly continuous assessment, we allow overlap in the covariate groups as we move from low to high covariate values. Each group consists of 20% of the study sample and we use a sliding window of 1% of the study sample. For each grouping we calculate a crude risk $\bar{r}$ and the population covariate mean $\bar{x}$. We call the scatter plot of $\bar{r}$ against $\bar{x}$ a risk-exposure scatter plot.

The function risk.exposure.plot implements the procedure. It has three basic arguments: the vector indicator of case/event status y, the vector of the covariate (should be a continuous exposure) group, and an optional argument for the vector of sampling weights weights, if a case-control study is being analyzed. Additional arguments are passed to scatter.smooth. Below is an example of how to construct a risk-exposure scatter plot for crude risk and pack-years.

```
> plot <- risk.exposure.plot(ccdata$y, ccdata$packyear, ccdata$w,
+                               xlab="Pack-Years", col="red")
```

There are only seven points in the scatter plot in Figure 1 because these were the number of unique values of pack-years reported. There is some suggestion of a non-linear relationship between risk and pack-years smoked.

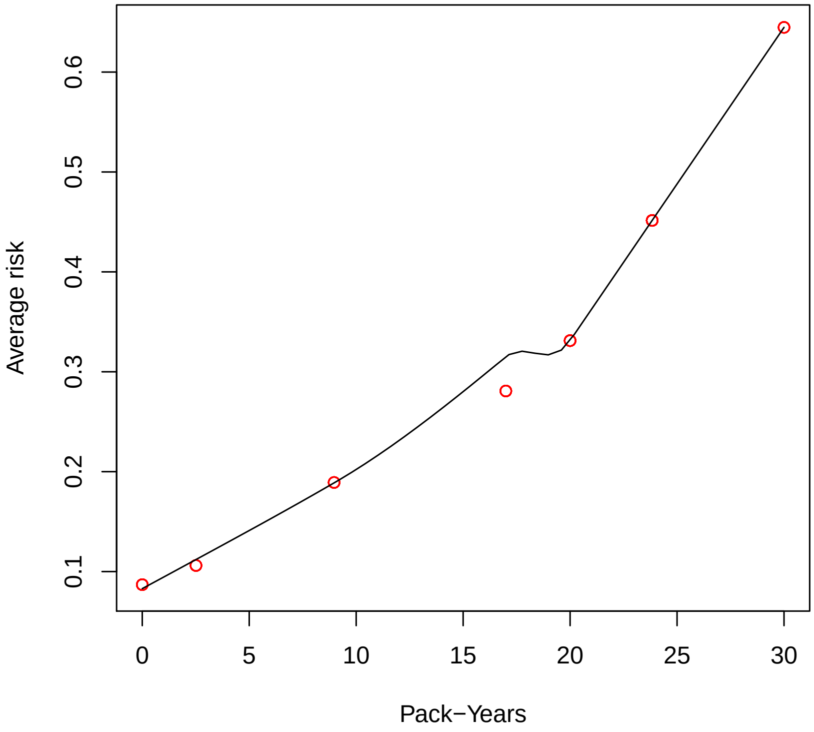

Figure 1: Risk-exposure scatter plot of crude risk against pack-years.

2 Linear-Expit (lexpit) model

2.1 Model

Suppose we were concerned that the linear assumption for pack-years might not be valid. If we thought that linearity on the relative risk scale was a more plausible model for the effect of pack-years, we could consider a lexpit model (Kovalchik, 2013). The lexpit model describes the probability of $Y_i = 1$ as a function of linear and nonlinear effects, where the nonlinear effects are the expit function (the inverse of the logit), $\text{expit}(x) = \exp(x)/(1 + \exp(x))$.

$$\pi_i = x_i'\beta + \text{expit}(z_i'\gamma) \tag{2}$$

The x_i variables are linear effects and z_i are the logistic effects. The first component of z_i is an intercept term, so that when the remaining components are 0, $\text{expit}(\gamma_0)$ is the baseline risk. As in BLM, β represent risk differences for unit changes in x_i. The coefficients γ are odds ratios after baseline adjustment for the effects of $x_i'\beta$, what we can think of as 'excess odds ratios'.

The lexpit model provides a more flexible way to estimate risk differences since it imposes fewer parameter constraints. This is possible because any $z_i'\gamma$ yields a probability measure.

Estimation for the lexpit model proceeds in two stages. The first stage fixes the expit parameters and estimates the linear coefficients with constrained maximization as described for the BLM in Section 1.1. Thus, in this stage, the expit term can be thought of as an offset in a BLM model. The second stage maximizes the expit parameters treating the linear term as fixed. Maximization at this stage uses a standard iterative reweighted least squares algorithm with modified weights that incorporate the linear risk offset.

2.2 Model fitting

The syntax for the `lexpit` takes two formula arguments: one for the linear components and one for the expit components. Note that the intercept is always included in the expit term. Otherwise, the syntax is identical to `blm`.

```
> fit.lexpit <- lexpit(y~female, y~packyear,
+                       data = ccdata,
+                       weight = ccdata$w,
+                       strata = ccdata$strata)
> summary(fit.lexpit)

Linear effects:
       Est.    Std. Err t-value p-value
female 0.01923 0.03385  0.56804 0.57018

Expit effects:
             Est.       Std. Err   t-value    p-value
(Intercept) -2.639e+00  1.458e-01 -1.811e+01  4.269e-61
packyear     1.015e-01  8.105e-03  1.253e+01  7.577e-33
```

```
Converged: TRUE
```

All of the methods described for the `blm` class are also available for `lexpit` objects.

```
> which.at.boundary(fit.lexpit)

No boundary constraints using the given criterion.

> confint(fit.lexpit)

                   Est.       Lower       Upper
female       0.01922563 -0.04711068  0.08556194
(Intercept) -2.63944371 -2.92517669 -2.35371073
packyear     0.10154727  0.08566217  0.11743236

> gof.pearson(fit.lexpit)

      E  O
1 16.28 22
2 16.70 16
3 72.09 59
4 79.67 86
5 24.47 24
6 13.99 13
7 77.70 78
8 77.31 80

 Chi-squared:  11.99811

 P-value:  0.06201106
```

The goodness-of-fit has improved with the lexpit model, suggesting that multiplicative rather than linear risk effects might be a more suitable model for the effect of continuous pack-years.

3 Conclusion

The `blm` package provides two models, BLM and lexpit, that can be used to obtain direct estimates of absolute risk and risk differences for binary data obtained from observational study designs. Fitting the `blm` and `lexpit` models will be straight-forward for `R` users because they provide similar syntax and methods as the `lm` class. The BLM and lexpit and models provide alternatives to logistic regression analysis of binary data that are appealing for epidemiological interpretation because they allow for the assessment of risk associations on an absolute rather than relative risk scale.

References

Kovalchik SA, Varadhan R, Fetterman B, Poitras NE, Wacholder S, Katki HA (2013). A general binomial regression model to estimate standardized risk differences from binary response data *Stat Med* 32:808–21.

Deville JC (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques *Survey methodology* 25(2):193–204

Lange K (2010). *Numerical Analysis for Statisticians.* Springer-Verlag, New York.

Madsen K, Nielsen H, and Tingleff O (2004). *Optimization with constraints.* IMM, Technical University of Denmark.