Journal of Statistical Software

August 2013, Volume 54, Issue 1. *http://www.jstatsoft.org/*

Fitting Additive Binomial Regression Models with the R Package blm

Stephanie Kovalchik
National Cancer Institute

Ravi Varadhan
Johns Hopkins University

Abstract

The R package **blm** provides functions for fitting a family of additive regression models to binary data. The included models are the binomial linear model, in which all covariates have additive effects, and the linear-expit (lexpit) model, which allows some covariates to have additive effects and other covariates to have logisitc effects. Additive binomial regression is a model of event probability, and the coefficients of linear terms estimate covariate-adjusted risk differences. Thus, in contrast to logistic regression, additive binomial regression puts focus on absolute risk and risk differences. In this paper, we give an overview of the methodology we have developed to fit the binomial linear and lexpit models to binary outcomes from cohort and population-based case-control studies. We illustrate the **blm** package's methods for additive model estimation, diagnostics, and inference with risk association analyses of a bladder cancer nested case-control study in the NIH-AARP Diet and Health Study.

Keywords: constrained optimization, logistic regression, binary outcome, absolute risk, risk difference.

1. Introduction

Logistic regression is the default approach for studying how explanatory factors are associated with a binary outcome (Hosmer and Lemeshow 2000). In the logistic model, the log-odds are expressed as a linear function of the regression coefficients, and the model coefficients estimate adjusted odds ratios. In an additive regression model of binary data, the effects of covariates are linearly related to risk, and the model coefficients estimate adjusted risk differences. The binomial linear model (BLM) – the generalized linear model for the binomial family with an identity link – is one example (Cox 1970; Wacholder 1986). Despite the relevance of absolute risks and risk differences to epidemiology, finance, and other fields, few methods or software for absolute risk and risk difference estimation exist. As with survival data (Aalen 1989;

Scheike and Zhang 2003), convenient tools for additive modeling of binary data have lagged behind tools for log-linear models because reliable estimation of additive models is technically more challenging (Austin 2010; Spiegelman and Hertzmark 2005; Newcombe 2006; Greenland 1987).

In this paper, we introduce the R (R Core Team 2013) package **blm** (Kovalchik 2013), available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=blm`. The package provides methods to fit two types of additive regression models for binary data: BLM, a strictly additive model, and *lexpit*, a more flexible model that consists of additive and multiplicative effects, where multiplicative effects are modeled through an inverse-logit (expit) term (Kovalchik, Varadhan, Fetterman, Poitras, Wacholder, and Katki 2013). Sections 2.1.1 and 2.1.2 detail each model and their interpretation. Section 2.2 describes the data sets to which the models can be applied. The methods for estimation and inference are presented in Section 2.3 and Section 2.4. We overview the **blm** package in Section 3. In Section 4, we demonstrate the main uses of the package with risk association analyses of an NIH-AARP bladder cancer case-control study.

2. Methods

2.1. Models

Binomial linear model (BLM)

Let Y_τ be a Bernoulli random variable taking the value 1 if the event occurs within the time interval τ and 0 otherwise. Under the binomial linear model, the probability of an event is a linear function of a set of p time-independent covariates $\mathbf{x}$,

$$\mathsf{P}(Y_\tau = 1|\mathbf{x}) = \mathbf{x}^\top \beta. \tag{1}$$

Under the BLM, each coefficient is the risk difference associated with a unit increase in the corresponding covariate, adjusted for all other covariates of the model. As a specific example, consider a model with a single covariate, x_1, that is a zero-one indicator of exposure, $\mathsf{P}(Y_\tau = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1$. In this case, β_0 is the expected risk of an event in the unexposed, $\beta_0 + \beta_1$ is the expected risk for the exposed, and β_1 the excess risk due to exposure.

Linear-expit model (lexpit)

The lexpit model is a generalization of BLM, which incorporates a multiplicative component that is a function of q covariates $\mathbf{z}$,

$$\mathsf{P}(Y_\tau = 1|\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \beta + \text{expit}\{\mathbf{z}^\top \gamma\}. \tag{2}$$

In (2), $\text{expit}(x) = \exp(x)/(1+\exp(x))$ is the inverse-logit function. When there are no additive terms, $\mathsf{P}(Y_\tau = 1|\mathbf{x}, \mathbf{z}) = \text{expit}\{\mathbf{z}^\top \gamma\}$ becomes a conventional logistic model, which shows that this model is also a special case of the lexpit model.

In the lexpit, the intercept is included in the expit term so that the background risk of the model – the risk when all remaining covariates of $\mathbf{z}$ and $\mathbf{x}$ are zero – is $\text{expit}\{\gamma_0\}$. Like the

BLM, the additive coefficients of the lexpit estimate the adjusted risk difference measures of association for the corresponding covariate of $\mathbf{x}$. The parameter $\exp(\gamma_j)$ is the odds ratio association between the residual risk $\mathsf{P}(Y_\tau = 1|\mathbf{x}, \mathbf{z}) - \mathbf{x}^\top\beta$ and the jth covariate z_j. As with logistic regression, the exponentiated regression coefficient is the odds ratio association between two individuals with different z_j exposure, fixing all other factors of the model.

2.2. Data

BLM and lexpit can be fit to binary data collected from a cohort study or from a population-based case-control study with sufficient sampling information. In what follows, we assume that the binary variable of interest is based on an underlying time-to-event variable and represents the occurrence of event within a specified time interval τ, $Y = I(T \in \tau)$. For each study type, the covariates $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$ and $(\mathbf{z}_1, \mathbf{z}_2, \ldots)$ are the observed values at the start of the interval τ.

Cohort study

Given a cohort study of n observations, the outcomes for the additive binomial model are the $y_1, \ldots, y_n$ indicators of an event occurring during the time interval τ. The binary outcomes can be defined in terms of the corresponding time-to-event variables $t_1, \ldots, t_n$ and event indicators $\delta_1, \ldots, \delta_n$, as $y_i = \delta_i I(t_i \in \tau)$.

Population-based case-control study

A population-based case-control study identifies all cases of an event occurring in a well-defined population during a specified period of time τ. The population is divided into J strata, each consisting of N_j individuals, and m_j controls are sampled from each stratum with simple random sampling. In addition to case status y_{ij}, each observation has a sampling weight, which is the inverse inclusion probability, $w_{ij} = N_j/m_j$ for controls and $w_{ij} = 1$ for cases, assuming $m_j << N_j$ for all j. In additive risk modeling, sampling information is needed to weigh-back to the underlying cohort and, thereby, obtain estimates for the absolute risk in the source population.

2.3. Estimation

Estimates for the parameters of the BLM and lexpit model are obtained by constrained maximization of a pseudo log-likelihood using a block relaxation algorithm (de Leeuw 1994). We describe the estimation methodology for the lexpit model. Fitting for the BLM is essentially equivalent to fitting a lexpit model with a constant expit term.

The estimates for the regression parameters $\Theta = (\beta, \gamma)$ are the solutions to the maximization problem,

$$\hat{\Theta} = \text{argmax}_\Theta\{\sum_i \sum_j w_{ij} l_{ij}(\Theta)\},\ \Theta \in \mathcal{F} \tag{3}$$

with the constraints

$$\mathcal{F} = \{0 \leq \mathbf{x}^\top\beta + \text{expit}(\mathbf{z}^\top\gamma) \leq 1\},\ \forall x, z. \tag{4}$$

1. *Initialization.* Set the intercept term

$$\text{expit}(\hat{\gamma}_1^{(0)}) = \sum_i \sum_j w_{ij} y_{ij} / \sum_i \sum_j w_{ij},$$

and all other parameters to zero.

2. *Linear update.* For the kth iteration, fix $q_{ij} = \text{expit}(\mathbf{z}_{ij}^\top \gamma^{(k)})$ and obtain

$$\hat{\beta}^{(k+1)} = \text{argmax}_\beta \{\sum_i \sum_j w_{ij} [y_{ij} \text{logit}(\mathbf{x}_{ij}^\top \beta + q_{ij}) + \log(1 - \mathbf{x}_{ij}^\top \beta - q_{ij})]\},$$

subject to the constraint that $\hat{\beta}^{(k+1)} \in \mathcal{F}$, where

$$\mathcal{F} = \{\beta : -q_{ij} \leq \mathbf{x}_{ij}^\top \beta \leq 1 - q_{ij} \; \forall x_{ij}\}.$$

3. *Expit update.* At the kth iteration, fix $p_{ij} = \mathbf{x}_{ij}^\top \hat{\beta}^{(k+1)}$ and with IRLS obtain

$$\hat{\gamma}^{(k+1)} = \text{argmax}_\gamma \{\sum_i \sum_j w_{ij} [y_{ij} \text{logit}(p_{ij} + \text{expit}(\mathbf{z}_{ij}^\top \gamma)) + \log(1 - p_{ij} - \text{expit}(\mathbf{z}_{ij}^\top \gamma))]\},$$

using iterative reweighted least squares.

4. Iterate between Steps 2 and 3 until convergence.

Table 1: Optimization procedure for lexpit model.

The objective function in Equation 3 is the weighted sum of the log-likelihood components for binomial data, with each probability following the lexpit model,

$$l_{ij}(\Theta) = y_{ij} \log(\pi_{ij}(\Theta)) + (1 - y_{ij}) \log(1 - \pi_{ij}(\Theta)),$$

where $\pi_{ij}(\Theta) = \mathbf{x}_{ij}^\top \beta + \text{expit}(\mathbf{z}_{ij}^\top \gamma)$.
The solution to (3) would be a standard maximum likelihood problem if it were not for the constraint that all estimated probabilities of the model be within the (0, 1) range. The space $\mathcal{F}$ is termed the *feasible region* because it ensures the feasibility of all the fitted values of the model. Although any covariate patterns could conceivably be specified in $\mathcal{F}$, our practice is to use an empirically-based region that is defined by the observed covariate patterns in the study sample.

Optimization algorithm

The constrained maximization procedure uses a two-stage block relaxation approach (de Leeuw 1994), which is summarized in Table 1. In the first stage, expit terms are considered fixed and the maximizing values for $\hat{\beta}$ are determined with an adaptive barrier algorithm

(Lange 1994, 2010) that, in the **blm** package, is implemented with the `constrOptim` function of the **stats** package. In the second stage, the linear terms are treated as fixed, using an offset term, and an iterative reweighted least squares algorithm with risk offset is used to update $\hat{\gamma}$. The block relaxation procedure is monotonic so convergence to a stationary point is guaranteed.

Optimization for the BLM does not require Step 3, and there is no offset term ($q_{ij} = 0$) in the updating of the $\hat{\beta}$. In this case, the intercept term is incorporated into the linear part and is initialized to $\hat{\beta}_0 = \sum_i \sum_j w_{ij} y_{ij} / \sum_i \sum_j w_{ij}$.

2.4. Inference

Variances for $\hat{\Theta}$ are estimated using an influence-based method. Several authors have previously described influence methods for variance estimation of complex survey statistics (Demnati and Rao 2010; Graubard and Fears 2005), and the influence operator is well-known for its use in the study of robustness (Hampel 1974). Further details of the influence function and its use with variance estimation are given by Deville (1999).

When the influence operator, $\Delta\{.\}$, is applied to an estimator, it yields an estimate of the Gâteaux derivative and each component of this derivative is an analytic jackknife deviate – the estimated deviation in the estimator when one observation is omitted. The variation in the deviates generated by the influence operator can therefore estimate a statistic's variance in the same way as the deviates generated from jackknife resampling. In the case of the lexpit model, using the index $k = (0, 1)$ to denote case status, the influences for $\hat{\beta}$ are

$$\Delta_{ijk}\{\hat{\beta}\} = [-\mathcal{H}(\hat{\beta})]^{-1} \mathbf{x}_{ijk} w_{ijk} (y_{ijk} - \mathbf{x}_{ijk}^{\top} \beta - \text{expit}(\mathbf{z}_{ijk}^{\top} \gamma))$$

and

$$\Delta_{ijk}\{\hat{\gamma}\} = [-\mathcal{H}(\hat{\gamma})]^{-1} \mathbf{z}_{ijk} w_{ijk} (y_{ijk} - \mathbf{x}_{ijk}^{\top} \beta - \text{expit}(\mathbf{z}_{ijk}^{\top} \gamma))$$

where $\mathcal{H}(\theta)$ is the second derivative of the objective function given in Equation 3 under the constraints $\mathcal{F}$. Letting $\Delta_{ijk}\{\hat{\Theta}\}' = (\Delta_{ijk}\{\hat{\beta}\}, \Delta_{ijk}\{\hat{\gamma}\})$, be the combined influences of the ijth observation on the parameters $\hat{\Theta}$, the variance estimate for $\hat{\Theta}$ is

$$\widehat{\text{Var}}(\hat{\Theta}) = \sum_k \sum_j n_{jk}/(n_{jk} - 1) \sum_{i=1}^{n_{jk}} (\Delta_{ijk}\{\hat{\Theta}\} - \bar{\Delta}_{.jk}\{\hat{\Theta}\})(\Delta_{ijk}\{\hat{\Theta}\} - \bar{\Delta}_{.jk}\{\hat{\Theta}\})^{\top} \quad (5)$$

with n_{jk} the number of k types in the jth stratum and $\bar{\Delta}_{.jk}\{\hat{\Theta}\}$ the average influence over the n_{jk} observations. The approximate large-sample distribution for $(\hat{\Theta} - \Theta)$ is $MVN(0, \widehat{\text{Var}}(\hat{\Theta}))$, and this result is the basis for the package's Wald tests and confidence interval construction.

When some fitted values are at the boundary of the feasible region (either 0 or 1), large-sample normality may not hold (Self and Liang 1987; Andrews 2000). Since the boundary cases in lexpit affect individual fitted values, we believe standard inference should apply when the number of constrained observations is few. However, because standard inference is not guaranteed, active constraints should be closely monitored (as we describe in Section 4.1) and caution taken with the interpretation of the fitted model when active constraints are present.

3. Package description

3.1. Overview

The **blm** package (Version 2013.2.4.4) consists of two model classes, `blm` and `lexpit`, supporting class methods, and additional functions to help diagnose the fitted model. Table 2 summarizes the main features of the package.

3.2. Model classes

The `blm` and `lexpit` are `S4` class objects whose constructers and methods have been designed to emulate the `lm` class. The basic syntax for fitting a `blm` model with cohort data is

```
blm(y ~ x, data)
```

where `y ~ x` is a `formula` and `data` is a `data.frame`. The syntax for the lexpit model has separate formulae for the linear and expit terms of the model

```
lexpit(formula.linear = y ~ x, formula.expit = y ~ z, data)
```

but its usage is otherwise the same as `blm`. The slots of the modeling objects, which can be accessed with the `@` operator or the named method, contain a similar set of attributes as the `lm` class. The accessor method for the model formula, `model.formula`, is unique

Function	Description
Model Classes	
`blm`	Fits a binomial linear model
`lexpit`	Fits a lexpit model
Class methods	
`coef`	Extractor for model coefficients
`confint`	Compute confidence intervals for model coefficients
`predict`	Estimate risks for specified covariates
`resid`	Extractor for residuals
`logLik`	Extractor for log-likelihood
`summary`	Table of coefficients, standardd errors, t values, p values
`vcov`	Variance-covariance of coefficients
`model.formula`	Extractor for model formula
Diagnostic functions	
`EO`	Expected to observed within subgroups
`crude.risk`	Crude risk estimates by a continuous covariate
`gof`	Hosmer-Lemeshow goodness-of-fit test
`LRT`	Likelihood ratio test
`Rsquared`	R^2 measures
`which.at.boundary`	Index of observations at boundary (i.e., risk of 0 or 1)

Table 2: Functions of the **blm** package.

to the `blm`/`lexpit` classes. In addition to the class methods listed in Table 2, the slots include the initialization parameters of the algorithm (`par.init`), the log-likelihood for the fitted (`loglik`) and null model (`loglik.null`), and the barrier.value for the constrained optimization algorithm (`barrier.value`).

When the models are fit to population-based case-control data, the function call should also include a vector of `weights` containing the sampling weights for each observation in the data set and a `factor` for the `strata` argument, if the control sampling used stratification.

4. Application: Bladder cancer in the NIH-AARP Study

The NIH-AARP Diet and Health Study is the largest study of diet and health ever conducted (Schatzkin, Subar, Thompson, Harlan, Tangrea, Hollenbeck, Hurwitz, Coyle, Schussler, Michaud, Freedman, Brown, Midthune, and Kipnis 2001). Between 1995 and 1996, over half a million members of the American Association of Retired Persons (AARP) responded to a detailed questionnaire about their dietary habits and other health behaviors and all participants were followed for cancer incidence and mortality outcomes. Instructions for researchers interested in submitting a proposal to study the NIH-AARP Diet and Health Study cohort are available at `http://dietandhealth.cancer.gov/resource`.

The present analysis was based on a nested case-control study of bladder cancer within the NIH-AARP cohort. Cases were 292 study participants over the age of 60 years at the time of the baseline questionnaire who were diagnosed with bladder cancer (ICD-O3 C67.0-67.9) by age 70 years. Thus, the time interval of the analysis is $\tau = (60, 70]$. 292 controls were randomly sampled from all individuals between ages 60 and 70 years at the time of the baseline questionnaire who at age 70 years had never been diagnosed with bladder cancer.

4.1. Gender, smoking, and bladder cancer

Relative risk analyses have previously suggested that gender and smoking are associated with the risk of developing bladder cancer (Freedman, Silverman, Hollenbeck, Schatzkin, and Abnet 2011). The first model fit examines the absolute risk differences for each gender and smoking-status subgroup.

```
R> library("blm")
R> data("aarp")
R> fit <- blm(bladder70 ~ female * smoke_status, data = aarp,
+    weights = aarp$w)
```

Here we fit a BLM with main effects for gender, smoking status categories, and each interaction using the pre-loaded data set `aarp`. The variable `smoke_status` is a factor with levels for Never, Curent, Former, and Unknown smoking statuses. The outcome variable `bladder70` is a zero-one indicator of bladder cancer case status by age 70 years. The weights `aarp$w` are the sampling fractions for each observation, which are needed to weigh the risk estimates back to the underlying AARP cohort. Stratification was not used in this case-control study so `strata` is left to take its default `NULL` value.

The object `fit` is of the `blm` class. One of the methods for this class is `coef`, which can be used to extract the baseline risk and risk differences associated with each parameter.

```
R> coef(fit) * 1000
```

```
               (Intercept)                    female
                 0.1946805                 0.6215476
        smoke_statusFormer       smoke_statusCurrent
                 0.3220126                 1.6890026
       smoke_statusUnknown female:smoke_statusFormer
                 0.3794089                 0.2214473
female:smoke_statusCurrent female:smoke_statusUnknown
                 2.6367932                 0.1427371
```

This shows, for example, that the baseline absolute risk of bladder cancer by age 70, the risk in the reference group of male never smokers, is 0.2 per 1,000 persons. The excess risk for male current smokers is 1.7 per 1,000, corresponding to an overall absolute risk for male current smokers is $0.2 + 1.7 = 1.9$ per 1,000.

Both `summary` and `confint` can be used to assess the significance of the estimated effects.

```
R> summary(fit)
```

```
                              Estimate    Std.Err t value  Pr(>|t|)
(Intercept)                1.9468e-04 3.4955e-05  5.5694 3.925e-08 ***
female                     6.2155e-04 1.7731e-04  3.5054 0.0004915 ***
smoke_statusFormer         3.2201e-04 1.2333e-04  2.6110 0.0092619 **
smoke_statusCurrent        1.6890e-03 7.6346e-04  2.2123 0.0273366 *
smoke_statusUnknown        3.7941e-04 5.4332e-04  0.6983 0.4852611
female:smoke_statusFormer  2.2145e-04 2.8608e-04  0.7741 0.4391990
female:smoke_statusCurrent 2.6368e-03 2.0856e-03  1.2643 0.2066348
female:smoke_statusUnknown 1.4274e-04 1.0915e-03  0.1308 0.8960017
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Converged: TRUE
```

```
R> confint(fit) * 1000
```

```
                                 Est.       Lower     Upper
(Intercept)                0.1946805  0.12616976 0.2631912
female                     0.6215476  0.27401928 0.9690759
smoke_statusFormer         0.3220126  0.08029532 0.5637299
smoke_statusCurrent        1.6890026  0.19265701 3.1853481
smoke_statusUnknown        0.3794089 -0.68547754 1.4442954
female:smoke_statusFormer  0.2214473 -0.33925281 0.7821473
female:smoke_statusCurrent 2.6367932 -1.45086954 6.7244559
female:smoke_statusUnknown 0.1427371 -1.99656247 2.2820367
```

The significance levels of `summary` are based on a Wald test. The confidence intervals for `confint` are at the 95% level and are constructed with a large-sample approximation based

on Student's t distribution. Both methods of inference suggest that the main effects of gender, and former and current smoking status are significant risk factors for bladder cancer.

To obtain the fitted values for each covariate of interest, we can use the `predict` method. When `predict` is supplied with the fitted `blm`, it returns the fitted absolute risk for each observation of the data frame used in the model's estimation. One can also provide a data frame with the `newdata` argument to compute fitted values for any covariate pattern of interest. The inclusion of standard errors is specified by the logical argument `se`. In the following code, we create a data frame containing the eight possible covariate types for the gender and smoking model and obtain fitted values and standard errors for these risk types.

```
R> all.vars(model.formula(fit))

[1] "bladder70"    "female"       "smoke_status"

R> risk.types <- unique(subset(aarp, select = all.vars(model.formula(fit))))
R> risk.types <- subset(risk.types, bladder70 == 0)
R> risk.types

      bladder70 female smoke_status
358           0      0       Former
489           0      0        Never
4656          0      0      Current
12193         0      0      Unknown
12922         0      1        Never
34758         0      1      Current
53309         0      1       Former
68611         0      1      Unknown

R> predict(fit, risk.types, se = TRUE) * 1000

            fit        se
358   0.5166931 0.1154913
489   0.1946805 0.0349551
4656  1.8836831 0.7610958
12193 0.5740894 0.5413172
12922 0.8162281 0.1709099
34758 5.1420238 1.9157903
53309 1.3596879 0.1727547
68611 1.3383741 0.9254869
```

Three functions for assessing the fit of the model are `which.at.boundary`, `logLik`, and `Rsquared`. The method `which.at.boundary` provides a matrix of covariate patterns whose predicted risks are at the boundary of the feasible region (0 or 1) according to a specified criterion. The default criterion is a risk within `1e-6` of the lower or upper bounds of this region. Although not a direct assessment of fit, the evaluation of the number and types of boundary cases can be indicative of a poorly specified model and each of these observations should be treated like potential points of influence.

The `logLik` method returns an object of the class `logLik` and is registered with the **stats4** package. Thus, the returned value can be used with applicable methods, such as `AIC`. However, when the `blm` or `lexpit` object is fit with weights, it is important to keep in mind that the returned value is a pseudo-log-likelihood. Although χ^2 testing does not necessarily apply to pseudo-log-likelihoods, the measures can still be useful for informal comparisons of improvement in fit between nested models, and the `AIC` for informal comparisons between nested and non-nested models, for example, between a `blm` and `lexpit` model fit to the same binary outcome.

The `Rsquared` method returns McFadden's pseudo unadjusted and adjusted R^2 statistics (McFadden 1974). Binomial regression models do not have equivalent measure for explained variation as the R^2 of logistic regression based on ordinary least squares (OLS) . Still, these measures that attempt to mimic the R^2 of OLS can be useful for comparing the fit between models that have been applied to the same data set, with better-fitting models having an R^2 value closer to 1.

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

No boundary constraints using the given criterion.

R> AIC(fit)

[1] 5493.509

R> Rsquared(fit)

$R2
[1] 0.04318212

$R2adj
[1] 0.03965591
```

There are no concerns regarding cases at the boundary. We have used the `logLik` method to obtain the pseudo-AIC of the model, which we can compare to any later extensions we consider. The low R^2 measures for the current model suggest that we have not greatly improved the fit of the model over a null model and an expanded model should be considered.

4.2. Mode of effects

We next consider some simple strategies for assessing the possible functional relationship between a continuous covariate and absolute risk. A graphical method provided by the **blm** package is the `risk.exposure.plot`. The `risk.exposure.plot` is a loess scatter plot of the unadjusted risk in subgroups defined by the covariate. The function `crude.risk` creates the data frame with the estimates of the crude risk in ordered bins defined by the covariate, which consists of overlapping groups of 20% of the supplied data set and a sliding window of 1% of the sample size. When the output of `crude.risk` is plotted with `risk.exposure.plot`, it provides a visual representation of the continuous relationship between absolute risk and the continuous covariate that is not influenced by any model assumptions.

In the code below, we use `crude.risk` to obtain a data frame of the unadjusted risk estimates for bladder cancer by age 70 by dietary fiber. This returns a data frame with the population-based risk estimates, `risk`, and the mean covariate value in each overlapping subgroup, `x`.

```
R> risk <- crude.risk(bladder70 ~ fiber.centered, data = aarp,
+    weights = aarp$w)
R> head(risk)
```

```
          risk         x
1 0.001433959 -9.364749
2 0.001293353 -9.110063
3 0.001249824 -8.864049
4 0.001167258 -8.497228
5 0.001229360 -8.227736
6 0.001229360 -7.979171
```

We then plot the resulting data using the `risk.exposure.plot`, using the argument `scale` to change the y-axis to units of risk per 1,000. Additional arguments are passed to the function `scatter.smooth`.

```
R> risk.exposure.plot(object = risk, scale = 1000, las = 1,
+    col = "royalblue", pch = 19, ylab = "Crude risk (per 1,000)",
+    xlab = "Avg. Fiber Consumption (Centered)")
```

Figure 1 shows the results of the plot of the crude risks. Because this gives a sense of the functional relationship between risk and the continuous covariate, it can be useful for guiding the choice of representation of the covariate in the `blm` or `lexpit` model. For dietary fiber, we see a general decline in risk with greater fiber consumption, but there is an increase in risk the intermediate range of consumed fiber. This suggests that a higher-order polynomial for fiber on the multiplicative scale may be more appropriate than a simple linear effect for fiber.

There appears to be a strong relationship between bladder cancer and fiber but of a non-linear nature. We therefore expand the absolute risk model using lexpit regression. The linear term of the model will include the same gender and smoking effects as we specified with the BLM. The expit term will have a main effect for the continuous variable `redmeat`, while fiber consumption will have a linear and quadratic term, centering fiber consumed on the median value of the sixth category of the factor (`fiber.centered`). The following script fits the described lexpit model.

```
R> formula.linear <- bladder70 ~ female * smoke_status
R> formula.expit <- bladder70 ~ redmeat + fiber.centered +
+    I(fiber.centered^2)
R> fit <- lexpit(formula.linear, formula.expit, data = aarp, weight = aarp$w)
```

The results of `summary` indicate that `redmeat` and `fiber.centered` are both significantly associated with bladder cancer but suggest that the quadratic term for `fiber.centered` might not be necessary.

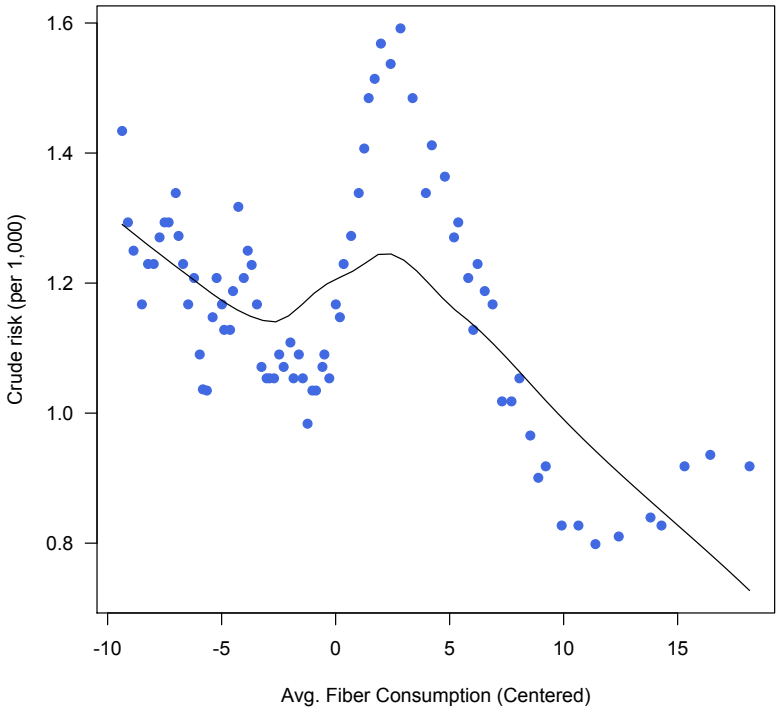

Figure 1: Plot of crude absolute risk of bladder cancer by age 70 (per 1,000) against dietary fiber (centered) using `risk.exposure.plot`.

```
R> summary(fit)

Linear effects:
                              Estimate    Std.Err t value Pr(>|t|)
female                     0.00042769 0.00017732  2.4120  0.01618 *
smoke_statusFormer         0.00029654 0.00012333  2.4045  0.01651 *
smoke_statusCurrent        0.00155339 0.00076348  2.0346  0.04235 *
smoke_statusUnknown        0.00033553 0.00054331  0.6176  0.53710
female:smoke_statusFormer  0.00026832 0.00028608  0.9379  0.34867
female:smoke_statusCurrent 0.00266586 0.00208577  1.2781  0.20173
female:smoke_statusUnknown 0.00015327 0.00109153  0.1404  0.88838
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Expit effects:
                       Estimate     Std.Err  t value  Pr(>|t|)
(Intercept)         -9.30754468  0.20059714 -46.3992 < 2.2e-16 ***
redmeat              0.01958238  0.00293749   6.6664 6.175e-11 ***
fiber.centered      -0.05197903  0.01582348  -3.2849  0.001082 **
I(fiber.centered^2)  0.00080416  0.00104696   0.7681  0.442750
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Converged: TRUE
```

The `risk.exposure.plot` provided a means of looking at a continuous covariates possible functional relationship to the crude (unadjusted) risk. If we wanted to consider the functional relationship *after* adjustment for other covariates, we could use testing approach. A test for the inclusion of a factor in the linear or expit term can be done directly when more than one additional covariate is included in the expit term. When this is the case, the lexpit regression can include a linear *and* multiplicative term for the covariate of interest. Testing the significance of each term provides a comparative assessment of the strength of the information of each mode of effect. Fitting both linear and multiplicative terms is possible because the expit transformation removes collinearity between each term. The code below shows how to use this procedure for the variable `fiber.centered`.

```
R> fit.both <- lexpit(update(formula.linear,
+    ~ . + fiber.centered + I(fiber.centered^2)),
+    formula.expit, data = aarp, weight = aarp$w)
R> summary(fit.both)

Linear effects:
                              Estimate     Std.Err t value Pr(>|t|)
female                      4.3435e-04  1.7627e-04  2.4641  0.01403 *
smoke_statusFormer          2.9354e-04  1.1683e-04  2.5125  0.01226 *
smoke_statusCurrent         1.5583e-03  7.7170e-04  2.0193  0.04393 *
smoke_statusUnknown         3.4461e-04  5.2059e-04  0.6620  0.50826
fiber.centered              6.0702e-06  8.0784e-06  0.7514  0.45272
I(fiber.centered^2)        -9.5411e-08  5.0429e-07 -0.1892  0.85000
female:smoke_statusFormer   2.7139e-04  2.8171e-04  0.9634  0.33576
female:smoke_statusCurrent  2.6586e-03  2.0833e-03  1.2761  0.20243
female:smoke_statusUnknown  1.5230e-04  1.0483e-03  0.1453  0.88453
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Expit effects:
                      Estimate    Std.Err  t value  Pr(>|t|)
(Intercept)         -9.3810342  0.2005971 -46.7655 < 2.2e-16 ***
redmeat              0.0195877  0.0029375   6.6682 6.122e-11 ***
fiber.centered      -0.0763240  0.0158235  -4.8235 1.812e-06 ***
I(fiber.centered^2)  0.0010515  0.0010470   1.0043    0.3156
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Converged: TRUE
```

Both the linear and quadratic additive terms for `fiber.centered` are not significant. We

therefore conclude that the simpler model with only multiplicative effects for `fiber.centered` may adequately describe the risk association for this dietary variable and bladder cancer.

The overall fit of the simpler model `fit` can be assessed with `Rsquared`, `EO`, and the `gof` functions. We have already described `Rsquared`. The `EO` function computes the ratio of expected and observed counts and its 95% confidence interval within subgroups of a specified categorical factor. Ratios that are not significantly different from one indicate that the model has good internal (within the training data) calibration, while ratios significantly below (above) suggest that the model is under-predicting (over-predicting) for those subgroups. In the script below, we look at the internal calibration in groups defined by education level.

```
R> Rsquared(fit)

$R2
[1] 0.04587181

$R2adj
[1] 0.04102327

R> AIC(fit)

[1] 5475.305

R> EO(fit, aarp$educ)

                      E   O      EtoO   lowerCI  upperCI
< 8 yrs       16.759585  21 0.7980755 0.5203512 1.224028
8-11 yrs      57.380071  46 1.2473928 0.9343303 1.665352
High School   32.546785  35 0.9299081 0.6676683 1.295148
Some college  66.618248  69 0.9654819 0.7625556 1.222410
College+     112.660377 114 0.9882489 0.8225154 1.187377
Unknown        6.008502   7 0.8583574 0.4092081 1.800496
```

In comparison to the BLM, the lexpit model has improved the pseudo R^2 and AIC measures of fit, and the model is well calibrated for all educational categories.

The function `gof` assesses the overall fit of the model. This function performs the Hosmer-Lemeshow goodness-of-fit test across deciles of risk. For cohort data, this statistic is compared to a χ^2 distribution, with large values suggesting a lack of fit. For case-control data, the function employs the adjustment proposed by Archer, Lemeshow, and Hosmer (2007) for use with weighted estimators.

```
R> gof(fit)

$table
$table$cases
                      O         E
[4.94e-05,0.000348]   7  8.659499
(0.000348,0.000551]  16 16.854658
```

```
(0.000551,0.000829] 21 21.950468
(0.000829,0.00112]  25 30.560476
(0.00112,0.00124]   28 30.700164
(0.00124,0.00136]   31 30.361227
(0.00136,0.00156]   37 28.015231
(0.00156,0.00184]   38 29.604923
(0.00184,0.00501]   42 38.987318
(0.00501,0.00588]   47 56.279603

$table$controls
                            O        E
[4.94e-05,0.000348] 45264.04 45262.38
(0.000348,0.000551] 36559.42 36558.56
(0.000551,0.000829] 32207.11 32206.16
(0.000829,0.00112]  29595.72 29590.16
(0.00112,0.00124]   26113.87 26111.17
(0.00124,0.00136]   23502.48 23503.12
(0.00136,0.00156]   19150.17 19159.16
(0.00156,0.00184]   17409.25 17417.64
(0.00184,0.00501]   13927.40 13930.41
(0.00501,0.00588]   10445.55 10436.27

$X2
[1] 0.8589446

$p.value
[1] 0.562016
```

The goodness-of-fit statistic suggests that the lexpit model's fit is generally good across the observed distribution of risk for bladder cancer.

Given that the good fit of current model, we can draw some preliminary conclusions about the risk associations for bladder cancer by age 70 in the AARP population. We do this by considering the absolute risk estimates and their 95% confidence intervals using the `confint` method. First, we consider the linear terms, which are reported first in the matrix returned by the `confint` method.

```
R> CIs <- confint(fit)
R> CIs[1:7, ] * 1000
```

```
                                  Est.       Lower     Upper
female                       0.4276926  0.08015648 0.7752288
smoke_statusFormer           0.2965384  0.05482000 0.5382567
smoke_statusCurrent          1.5533889  0.05700385 3.0497739
smoke_statusUnknown          0.3355341 -0.72934148 1.4004096
female:smoke_statusFormer    0.2683246 -0.29238295 0.8290322
female:smoke_statusCurrent   2.6658558 -1.42217844 6.7538900
female:smoke_statusUnknown   0.1532681 -1.98609447 2.2926306
```

Smoking had the largest effect of all categorical risk factors. Among male members of the AARP over 60 years old, current smokers had a 1.5 per 1,000 greater risk (95% CI 0.06 to 3.05 per 1,000) of bladder cancer by age 70 than never smokers. Among women members, the excess risk increased by 2.7 per 1,000 as compared to male smokers, but this was not a statistically significant difference (95% CI -1.42 to 6.75 per 1,000). Gender was also associated with a greater risk of bladder cancer in never smokers. Female gender was associated with a significant excess risk of 0.4 per 1,000 risk (95% CI 0.08 to 0.78 per 1,000) of bladder cancer among never smokers.

```
R> CIs[8:11, ]
```

```
                          Est.       Lower        Upper
(Intercept)         -9.3075446811 -9.70070786 -8.914381505
redmeat              0.0195823750  0.01382500  0.025339746
fiber.centered      -0.0519790296 -0.08299247 -0.020965585
I(fiber.centered^2)  0.0008041563 -0.00124784  0.002856152
```

```
R> expit(CIs[8, ]) * 10000
```

```
     Est.     Lower     Upper
0.9072883 0.6123638 1.3442341
```

Terms from the '`Intercept`' down of the `confint` output are variables in the expit term. The 'Intercept' is the logit of the background risk. The reference group for the fitted model was male never smokers, with no consumption of red meat, who 18 grams of fiber intake per day (the centering value). The risk of bladder cancer by age 70 for this subpopulation was 0.9 per 10,000 persons (95% CI 0.6 to 1.3 per 10,000). The remaining expit terms represent log-odd ratios conditional on all other factors in the model. Thus, for two individuals of the same gender, smoking status, and fiber intake, the person who consumed an additional one gram per day of red meat had a 2% greater odds (95% 1.4 to 2.6) of bladder cancer.

5. Summary

The R package **blm** provides easy-to-use tools to fit additive regression models for binary data from observational studies. The **blm** and `lexpit` models directly estimate absolute risks and adjusted risk differences for cohort and some case-control studies, making them an important addition to the statistician's toolbox. By complementing conventional multiplicative modeling, the tools of the **blm** package can help clarify how covariates affect a binary outcome.

Acknowledgments

This research was supported by the intramural research program of the NIH/NCI. We thank Dr. Hormuzd A. Katki and Dr. Sholom Wacholder for their input on this work. We are grateful to the participants of the NIH-AARP Diet and Health Study. Dr. Ravi Varadhan is a Brookdale Leadership in Aging Fellow at the Johns Hopkins University.

References

Aalen OO (1989). "A Linear Regression Model for the Analysis of Life Times." *Statistics in Medicine*, **8**(8), 907–25.

Andrews DWK (2000). "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space." *Econometrica*, **68**(2), 399–405.

Archer KJ, Lemeshow S, Hosmer DW (2007). "Goodness-of-fit Tests for Logistic Regression Models when Data Are Collected Using a Complex Sampling Design." *Computational Statistics & Data Analysis*, **51**(9), 4450–4464.

Austin PC (2010). "Absolute Risk Reductions, Relative Risks, Relative Risk Reductions, and Numbers Needed to Treat Can Be Obtained from a Logistic Regression Model." *Journal of Clinical Epidemiology*, **63**(1), 2–6.

Cox DR (1970). *The Analysis of Binary Data*. Methuen, London.

de Leeuw J (1994). "Block Relaxation Algorithms in Statistics." In HH Bock, W Lenski, MM Richter (eds.), *Information Systems and Data Analysis*, pp. 308–325. Springer-Verlag, Berlin.

Demnati A, Rao JNK (2010). "Linearization Variance Estimators for Model Parameters from Complex Survey Data." *Survey Methodology*, **36**(2), 193–201.

Deville JC (1999). "Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques." *Survey Methodology*, **25**, 193–203.

Freedman ND, Silverman DT, Hollenbeck AR, Schatzkin A, Abnet CC (2011). "Association Between Smoking and Risk of Bladder Cancer Among Men and Women." *Journal of the American Medical Association*, **306**(7), 737–745.

Graubard BI, Fears TR (2005). "Standard Errors for Attributable Risk for Simple and Complex Sample Designs." *Biometrics*, **61**(3), 847–855.

Greenland S (1987). "Interpretation and Choice of Effect Measures in Epidemiologic Analyses." *American Journal of Epidemiology*, **125**(5), 761–768.

Hampel FR (1974). "Influence Curve and Its Role in Robust Estimation." *Journal of the American Statistical Association*, **69**(346), 383–393.

Hosmer DW, Lemeshow S (2000). *Applied Logistic Regression*. 2nd edition. John Wiley & Sons, New York.

Kovalchik SA (2013). ***blm**: Binomial Linear and Linear-Expit Regression*. R package version 2013.2.4.4. URL http://CRAN.R-project.org/package=blm.

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." *Statistics in Medicine*, **32**(5), 808–21.

Lange K (1994). "An Adaptive Barrier Method for Convex Programming." *Methods and Applications of Analysis*, **1**(4), 392–402.

Lange K (2010). *Numerical Analysis for Statisticians*. 2nd edition. Springer-Verlag, New York.

McFadden D (1974). "Conditional Logit Analysis of Qualitative Choice Behavior." In P Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York.

Newcombe RG (2006). "A Deficiency of the Odds Ratio as a Measure of Effect Size." *Statistics in Medicine*, **25**(24), 4235–4240.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schatzkin A, Subar AF, Thompson FE, Harlan LC, Tangrea J, Hollenbeck AR, Hurwitz PE, Coyle L, Schussler N, Michaud DS, Freedman LS, Brown CC, Midthune D, Kipnis V (2001). "Design and Serendipity in Establishing a Large Cohort with Wide Dietary Intake Distributions: The National Institutes of Health-American Association of Retired Persons Diet and Health Study." *American Journal of Epidemiology*, **154**(12), 1119–1125.

Scheike TH, Zhang MJ (2003). "Extensions and Applications of the Cox-Aalen Survival Model." *Biometrics*, **59**(4), 1036–1045.

Self SG, Liang KY (1987). "Asymptotic Properties of Maximum-Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions." *Journal of the American Statistical Association*, **82**(398), 605–610.

Spiegelman D, Hertzmark E (2005). "Easy SAS Calculations for Risk or Prevalence Ratios and Differences." *American Journal of Epidemiology*, **162**(3), 199–200.

Wacholder S (1986). "Binomial Regression in GLIM: Estimating Risk Ratios and Risk Differences." *American Journal of Epidemiology*, **123**(1), 174–184.

Affiliation:

Stephanie Kovalchik
Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute
Rockville, MD, 20852, United States of America
E-mail: kovalchiksa@mail.nih.gov

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 54, Issue 1 *Submitted:* 2011-11-22
August 2013 *Accepted:* 2013-05-30