# Vignette for package blm

*Stephanie A. Kovalchik*

**Abstract**
The blm package provides functions for fitting flexible binomial models for cohort data with a binary outcome. 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. Estimation of the model parameters is based on constrained maximum likelihood, which ensures that the fitted model yields feasible risk estimates. In this vignette, the BLM and LEXPIT model classes and their methods are demonstrated in risk models of type II diabetes among Pima Indians.

## Binomial linear model

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

$$\pi_i = x_i'\beta$$

Each $\beta$ 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 sample cohort defines the *feasible region* for $\beta$. To ensure that the estimates for $\beta$ are within the region of feasibility, constrained maximum likelihood is used. The default algorithm employed is an augmented Lagrangian method (Madsen et al., 2004) which is implemented with the `auglag` function of the package `alabama` (Varadhan, 2011). An adaptive barrier method can also be used by setting the argument `augmented` to `FALSE` (Lange, 2010). In this case, the function `constrOptim.nl` of `alabama` performs the optimization. The function `blm` provides a wrapper for each method in fitting the linear model.

As an illustration of the model syntax we consider a model to estimate the risk of diabetes among Pima indians based on the `Pima.te` dataset of the `MASS` package.

We begin the `R` session by loading the packages with the binomial model fitters (`blm`) and the dataset for the analysis (`MASS`). The dataset `Pima.te` is loaded.

```
> library(blm)
> library(MASS)
> data(Pima.te)
> head(Pima.te)

  npreg glu bp skin  bmi   ped age type
1     6 148 72   35 33.6 0.627  50  Yes
2     1  85 66   29 26.6 0.351  31   No
3     1  89 66   23 28.1 0.167  21   No
4     3  78 50   32 31.0 0.248  26  Yes
5     2 197 70   45 30.5 0.158  53  Yes
6     5 166 72   19 25.8 0.587  51  Yes
```

The sample consists of 332 adult women of the Pima tribe in Phoenix, Arizona. There are eight demographic/anthropometric measures. The outcome of the analysis is `type`, which is a Yes/No indicator for WHO criteria of diabetes.

The fitted model will examine the risk association of age and body mass index on the probability of type II diabetes. The syntax for `blm` is much like `lm`, consisting of `formula` and `data` arguments.

```
> Pima.te$diabetes <- ifelse(Pima.te$type == "Yes", 1, 0
> fit <- blm(diabetes ~ scale(age) + scale(bmi), Pima.te
> fit

Call:  diabetes ~ scale(age) + scale(bmi)

Coefficients:
             estimates  t-value    std. err   p-value
(Intercept)  3.280e-01  1.429e+01  2.296e-02  0.000e+00
scale(age)   1.173e-01  5.433e+00  2.159e-02  1.082e-07
scale(bmi)   1.186e-01  7.396e+00  1.603e-02  1.174e-12

Degrees of Freedom: 329
Run time (sec): 0.137
LogLik:            -178.7         AIC: 363.5
```

The `scale` function standardizes each continuous measure, subtracting each observation by its mean and dividing by its standard deviation. For a normally distributed variable, this standardization will result in a covariate that is $\sim N(0,1)$; zero corresponds to the mean and one unit change corresponds to a standard deviation change from the mean.

Showing the result of `fit` returns point estimates, t-values, standard errors, and p-values for $\beta$. We find that all of the factors are statistically significant. The average age of the Pima Indians in the sample cohort was 31.3 years and the average BMI was 33.2. The model suggests that a female Pima Indian of this age and BMI has a 32.8% chance of being diabetic. The risk of diabetes for a Pima woman that is a standard deviation older than a Pima woman of the same BMI, is increased by an absolute risk of 11.7%. The risk difference for diabetes between Pima Indians of the same age but who differ by a standard deviation in

BMI is 11.9%, with the risk increasing with higher BMI.

The log-likelihood, AIC, and degrees of freedom are also reported, which can be useful for model comparison.

For more information about the convergence properties of the fit, we use the `summary` function.

```
> summary(fit)

$est
                 [,1]
(Intercept) 0.3280491
scale(age)  0.1173087
scale(bmi)  0.1185589

$gradient
[1]  45.24306 -35.80706 -79.38399

$feasible
[1] TRUE

$active
(Intercept)  scale(age)  scale(bmi)
  1.0000000  -0.8758996  -1.9003085

$convergence
[1] 0

$message
NULL

$loglik
[1] -178.7356

$df
[1] 329

$AIC
[1] 363.4712

$null.deviance
[1] 420.2973

$seconds.to.run
[1] 0.137
```

This returns a list with elements with the following elements. The element `est` are the regression coefficients, which could also be obtained by applying `coef`. The element `gradient` is the first derivative of the objective function with respect to $\beta$, where the objective function for `auglag` is the log-likelihood in addition to a first-order barrier term and a second order penalty term of the inequality constraints. The gradient should be close to zero at the maximum likelihood solutions. But, if the boundary is reached, the gradient values could be large. The element `feasible` is a logical value indicating whether all of the predicted risks in the sample are true probabilities. The `active` element fives the covariate classes whose risks are at the boundary of the parameter space. If all constraints are inactive then `active` is `NULL`. Here, we find that there is one active constraint which is associated with younger age and low BMI. The exact predicted risk for this subject type can be obtained as follows.

```
> fit@active.constraints$active %*% coef(fit)

             [,1]
[1,] 2.694218e-09
```

The element `convergence` is a numerical value indicating whether the algorithm successfully converged. A value of 0 indicates success. Any other number indicates a failure to converge and `message` provides some description of the type of failure.

The remaining elements provide some assessments of the model fit. As an exact test of the model fit we can use a likelihood ratio test.

```
> LR <- summary(fit)$null.deviance - 2 * summary(fit)$lo
> df <- length(coef(fit))
> 1 - pchisq(LR, df = df)

[1] 0
```

The global test is highly significant. As a further diagnostic of the model fit, a Hosmer-Lemeshow type test is appropriate, given that the model includes continuous covariates.

```
> gof(fit)

$chisq
[1] 8.811154

$p.value
[1] 0.3584765
```

There is no evidence that the linear model is a poor fit. To investigate this further, a lattice plot of the observed incidence of diabetes, binned by BMI and age groups, against the predicted mean risk, is useful. Figure 1 shows linear risk effects for the youngest age groups but there is some discrepancy among the oldest Pima Indians.

```
> library(lattice)
> y <- tapply(Pima.te$diabetes, list(cut(Pima.te$bmi, c
+     30, 35, 40, 50)), cut(Pima.te$age, c(20, 30, 40, 5
+     mean)
> pre <- tapply(predict(fit), list(cut(Pima.te$bmi, c(18
+     35, 40, 50)), cut(Pima.te$age, c(20, 30, 40, 50, 7
> plot.data <- data.frame(obs = as.vector(y), pre = as.v
+     age = rep(colnames(y), each = 5), bmi = rep(row.na
+         4))
> xyplot(obs + pre ~ bmi | age, data = plot.data, auto.k
+     type = "b")
```

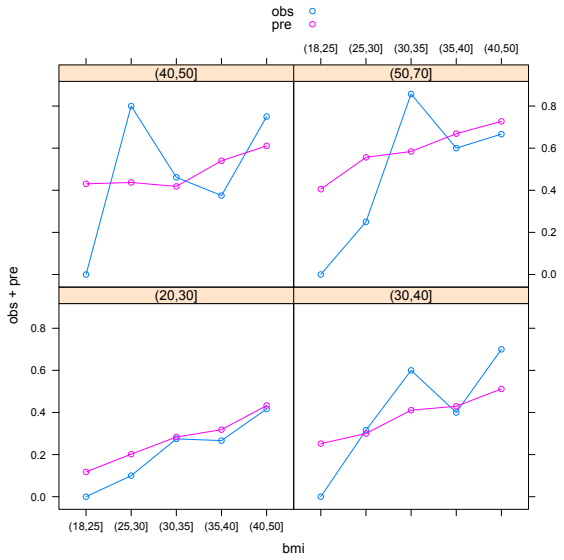


Figure 1: Graphical inspection of linearity assumption for BLM model `fit`

Had the model consisted of only categorical variables, we would assess goodness of fit with deviance, Pearson chi-square statistics, and a comparison of the observed and expected event counts in each covariate category.

```
> fit.categorical <- blm(diabetes ~ I(age > 50) + I(bmi > 30),
+     Pima.te)
> fit.categorical

Call:  diabetes ~ I(age > 50) + I(bmi > 30)

Coefficients:
                  estimates  t-value    std. err   p-value
(Intercept)       1.573e-01  4.768e+00  3.298e-02  2.801e-06
I(age > 50)TRUE   2.168e-01  1.875e+00  1.156e-01  6.165e-02
I(bmi > 30)TRUE   2.513e-01  5.277e+00  4.763e-02  2.389e-07

Degrees of Freedom: 329
Run time (sec): 0.018
LogLik:             -195.7          AIC: 397.4

> dispersion(fit.categorical)

$observed
100 101 110 111
 19  78   1  11

$expected
      100       101       110       111
18.083969 79.670266  2.618417  9.380598

$deviance
[1] 391.4493

$pearson
[1] 1.361308

$pearson.df
[1] 1

$deviance.df
[1] 329

$pearson.p
[1] 0.243311

$deviance.p
[1] 0.01013852
```

The `dispersion` function provides the observed and expected events in each covariate class, whose pattern is indicated by the binary sequence corresponding to the three parameters of the model: intercept, age, BMI. Degrees of freedom and p-values are given for the deviance and chi-square statistics.

Returning to our starting model, if we wanted to use an adaptive barrier method for the optimization, we would use the `augmented` argument.

```
> fit.barrier <- blm(diabetes ~ scale(age) + scale(bmi),
+     augmented = FALSE)
> fit.barrier

Call:  diabetes ~ scale(age) + scale(bmi)

Coefficients:
             estimates  t-value    std. err   p-value
(Intercept)  3.263e-01  1.210e+01  2.697e-02  0.000e+00
scale(age)   1.169e-01  5.047e+00  2.317e-02  7.440e-07
scale(bmi)   1.178e-01  4.041e+00  2.915e-02  6.641e-05

Degrees of Freedom: 329
Run time (sec): 0.041
LogLik:             -178.7          AIC: 363.5
```

The point estimates are quite similar but we note that the t-value for `bmi` is nearly twice that of `fit`, indicating that the standard error is $\approx 50\%$ of what was found for the fit with the augmented Lagrangian method.

```
> sqrt(diag(vcov(fit)))/sqrt(diag(vcov(fit.barrier)))

[1] 0.8513155 0.9321175 0.5498353
```

The models differ because they use a different approach to determining the covariance-variance matrix of the model estimates. In the augmented Lagrangian, the active inequality constraints are included in the objective function, its gradient, and Hessian. For the barrier method, the standard unconstrained Hessian is used. The models will give an equivalent standard error when no constraints are active. But when the boundary is hit, the unconstrained Hessian might be inaccurate and a warning is thrown to caution users against its use.

# Linear-Expit (LEXPIT) model

Suppose we expanded the BLM model to include the effects of plasma glucose concentration $> 100$ mg/dl.

```
> fit <- blm(diabetes ~ scale(age) + scale(bmi) + I(glu > 100),
+     Pima.te)
> fit

Call:  diabetes ~ scale(age) + scale(bmi) + I(glu > 100)

Coefficients:
                  estimates  t-value    std. err  p-value
(Intercept)       2.022e-01  5.001e+00  4.043e-02  9.291e-07
scale(age)        8.755e-02  3.589e+00  2.439e-02  3.828e-04
scale(bmi)        7.412e-02  3.458e+00  2.144e-02  6.166e-04
I(glu > 100)TRUE  2.089e-01  3.628e+00  5.757e-02  3.306e-04

Degrees of Freedom: 328
Run time (sec): 0.131
LogLik:            -171         AIC: 350

> summary(fit)

$est
                       [,1]
(Intercept)       0.20218407
scale(age)        0.08754867
scale(bmi)        0.07412256
I(glu > 100)TRUE 0.20890254

$gradient
[1]   75.142631  -56.984052 -114.478044    4.495058

$feasible
[1] TRUE

$active
    (Intercept) scale(age) scale(bmi) I(glu > 100)TRUE
234           1 -0.9699179  -1.570769
271           1 -0.7818813  -1.804193

$convergence
[1] 9

$message
[1] "Convergence due to lack of progress in parameter updates"

$loglik
[1] -171.0021

$df
[1] 328

$AIC
[1] 350.0042

$null.deviance
[1] 449.231

$seconds.to.run
[1] 0.131
```

The introduction of the additional parameter results in two boundary cases and the augmented Lagrangian algorithm failed to converge. Although we could consider adjusting the algorithm settings, if we are unsure whether linearity applies to all of the risk factors, we can fit a more flexible LEXPIT model that allows us to consider a mixture of linear and nonlinear effects.

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), $\text{expit}(x) = \exp(x)/(1+\exp(x))$.

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

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, $\beta$ represent risk differences for unit changes in $x_i$. The coefficients $\gamma$ are odds ratios after baseline adjustment for the effects of $x_i'\beta$.

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.

To illustrate the syntax of the `lexpit` function and its potential utility, we fit the expanded model for type II diabetes in Pima Indians with linear effects for age and logistic effects for BMI and plasma glucose concentration.

```
> fit.lexpit <- lexpit(f.linear = diabetes ~ scale(age),
+     scale(bmi) + I(glu > 100), Pima.te)
> fit.lexpit

Linear Call:  diabetes ~ scale(age)
Expit Call:  diabetes ~ scale(bmi) + I(glu > 100)
     0
     0
Coefficients:
                   estimates   t-value     std. err    p-
scale(age)          9.293e-02   4.874e+00   1.907e-02   1
(Intercept)        -1.596e+00  -7.295e+00   2.187e-01   2
scale(bmi)          4.560e-01   4.115e+00   1.108e-01   4
I(glu > 100)TRUE    1.297e+00   4.581e+00   2.831e-01   6

Degrees of Freedom: 328
Run time (sec): 0.183
LogLik:            -168.2         AIC: 344.4

> summary(fit.lexpit)

$est.linear
[1] 0.0929253

$est.expit
     (Intercept)       scale(bmi) I(glu > 100)TRUE
       -1.595716         0.456040         1.297065

$baseline.risk
```

```
(Intercept)
  0.1685812

$OR
      scale(bmi) I(glu > 100)TRUE
        1.577813         3.658542

$gradient
[1] -37.270160  12.472267 -10.467269   5.860206

$feasible
[1] TRUE

$active
                      (Intercept)       scale(bmi) I(glu > 100)TRUE
      -0.9699179        1.0000000       -1.5707695        0.0000000

$convergence
[1] 0

$message
NULL

$loglik
[1] -168.222

$df
[1] 328

$AIC
[1] 344.444

$null.deviance
[1] 470.4549

$seconds.to.run
[1] 0.183
```

The LEXPIT model meets the criteria for convergence with only one active constraint. Similar methods as shown for BLM are available for the `lexpit` class which provide measures of model fit, methods to compute confidence intervals, and make predictions.

```
> LR <- summary(fit.lexpit)$null.deviance - 2 * summary(fit.lexpit)$loglik
> df <- length(coef(fit.lexpit))
> 1 - pchisq(LR, df = df)

[1] 0

> gof(fit.lexpit)

$chisq
[1] 11.90896

$p.value
[1] 0.1553122
```

To estimate a confidence interval for the risk difference associated with 2 standard deviation difference in age we could use the `ci` function and specify the vector for the linear effects with the argument `C`.

```
> ci(fit.lexpit, C = 2, baseline = FALSE)

$est
          [,1]
[1,] 0.1858506

$se
          [,1]
[1,] 0.0381302

$lower
          [,1]
[1,] 0.1111168

$upper
          [,1]
[1,] 0.2605844

$CI
[1] "0.1859, (0.1111, 0.2606)"
```

The argument `FALSE` specifies that the expit components are not included in the confidence interval determination. If we wanted a confidence interval for the absolute risk of diabetes for an Pima woman of average age, BMI, and with a plasma glucose $> 100$ mg/dl, we would specify the expit components as follows.

```
> diabetic.risk <- ci(fit.lexpit, C = 0, C.expit = c(1,
> diabetic.risk

$est
          [,1]
[1,] 0.4258872

$se
           [,1]
[1,] 0.03387777

$lower
         [,1]
[1,] 0.359488

$upper
          [,1]
[1,] 0.4922864

$CI
[1] "0.4259, (0.3595, 0.4923)"
```

We find that the estimated risk of type II diabetes for a Pima woman of this type is 42.6%.

## 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 cohort data. The instantiation and methods for the `blm` and

`lexpit` classes are in keeping with other linear models in `R`. The LEXPIT provides additional flexibility that can be useful when estimates of the linear model are near the boundary.

# Bibliography

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.

Varadhan, R. (2011). *alabama: Constrained nonlinear optimization*. R package version 2011.3-1.