Package 'glmfitmiss'

May 8, 2026

Title Fitting GLMs with Missing Data in Both Responses and Covariates

Description
Fits generalized linear models (GLMs) when there is missing data in both the response and categorical covariates. The functions implement likelihood-based methods using the Expectation and Maximization (EM) algorithm and optionally apply Firth's bias correction for improved inference. See Pradhan, Nychka, and Bandyopadhyay (2025) <https:>, Maiti and Pradhan (2009) <doi:10.1111/j.1541-0420.2008.01186.x>, Maity, Pradhan, and Das (2019) <doi:10.1080/00031305.2017.1407359> for further methodological details.

Version 2.1.0

Depends R (>= 4.0.0)

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.3.1

Imports data.table (>= 1.12.8), dplyr (>= 1.0.0), abind (>= 1.4-5), MASS (>= 7.3-53), brglm2 (>= 0.7.1)

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author Vivek Pradhan [aut, cre], Douglas Nychka [aut], Soutir Bandyopadhyay [aut]

Maintainer Vivek Pradhan <vpradhan2009@gmail.com>

Repository CRAN

Date/Publication 2025-04-22 14:10:02 UTC

Contents

`glmfitmiss-package` *glmfitmiss: Fitting Binary Regression Models with Missing Data*

Description

The glmfitmiss package provides functions for fitting binary regression models in the presence of missing data in both response variable level and covariate levels. The package includes likelihood-based methods, primarily based on the EM algorithm by Ibrahim (1990) for handling missing data mechanisms. Bias-reducing adjusted score approaches introduced by Firth (1993) are also incorporated in all the supported methods.

Details

This package enhances the accuracy of binary regression modeling in the presence of missing data by incorporating Ibrahim (1990) EM algorithm and Firth (1993) bias-reducing adjusted score methods.

The main functions in this package are:

- `emBinRegMAR`: Fits a binary regression model with missing categorical covariates. Assumes missing data are Missing at Random (MAR).
- `emBinRegNonIG`: Fits a binary regression model with missing responses that are nonignorable based on Ibrahim and Lipsitz (1996).
- `emBinRegMixedMAR`: Fits a binary regression model with missing responses and covariates, accounting for the non-ignorable missing responses assumption and Missing at Random (MAR) missing covariates.

- logRegMAR: Fits a logistic regression model (binary regression with a link=logit) with missing categorical covariates that are Missing at Random (MAR).

The other functions and data included in this package are

- emforbeta: The function to fit binary regression models with missing categorical covariates is implemented using a likelihood-based method, specifically the EM algorithm proposed by Ibrahim (1990).
- est: Example using Eastern Cooperative Oncology Group clinical trial.
- meningitis: Example using Meningococcal Disease Data.
- metastmelanoma: Example from a cancer clinical trial metastatic melanoma – Kirkwood et al. (1996).
- emyxmiss: The main function fits binary regression models while accounting for missing responses and missing categorical covariates. This function implements a novel likelihood-based method using the EM algorithm. For more information, refer to the work by Pradhan, Nychka, and Bandyopadhyay (2025).
- meningitis60ymis: Meningococcal Disease Data with missing response variable.
- llkmiss: Log-likelihood function for models with missing data with out using EM-algorithms.
- est45: Example using Eastern Cooperative Oncology Group clinical trial – a subset of the 'est' data.
- simulateData: Function to simulate response data.
- simulateCovariateData: Function to simulate covariate data.
- felinedata: Sykes et al. (1999) data, the risk factors for Chlamy, a chlamydial infection in cats.
- simulateMissDfYorX: This function generates missing covariate or missing responses data. The missing data generation in the last two supplied covariates will be generated based on a predefined mechanisms. Missing data generation in the response variable will be based on the suppilied true alpha.
- sixcitydata: Longitudinal study of health effects of air pollution using data from six cities Ware et al. (1984).
- ibrahim: Example dataset used in Ibrahim (1990, JASA).
- testyxm: Function for testing models with missing data.

Author(s)

Maintainer: Vivek Pradhan <vpradhan2009@gmail.com>

Authors:

- Douglas Nychka <nychka@mines.edu>
- Soutir Bandyopadhyay <bsoutir@gmail.com>

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Ibrahim, J. G., and Lipsitz, S. R. (1996). Parameter Estimation from Incomplete Data in Binomial Regression when the Missing Data Mechanism is Nonignorable, Biometrics, 52, 1071–1078.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Louis, T. A. (1982). Finding the observed information when using the EM algorithm. Proceedings of the Royal Statistical Society, Ser B, 44, 226-233.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Pradhan, V., Nychka, D., and Bandyopadhyay, S. (2025). Addressing Missing Responses and Categorical Covariates in Binary Regression Modeling: An Integrated Framework (to be submitted).

Pradhan, V., Nychka, D., and Bandyopadhyay, S. (2025). Bridging Gaps in Logistic Regression: Tackling Missing Categorical Covariates with a New Likelihood Method (to be submitted).

Pradhan, V., Nychka, D., and Bandyopadhyay, S. (2025). glmFitMiss: Binary Regression with Missing Data in R (to be submitted).

See Also

emBinRegMAR, emBinRegMixedMAR, logRegMAR, meningitis, emforbeta, meningitis60ymis, emyxmiss, est, metastmelanoma, simulateCovariateData, est45, simulateData, felinedata, sixcity-data, ibrahim, testyxm, llkmiss

---

`emBinRegMAR` *Fitting binary regression with missing categorical covariates using Expectation-Maximisation (EM) based method*

Description

This function allows users to fit generalized linear models with incomplete predictors that are categorical. The model is fitted using a likelihood-based method, which ensures reliable parameter estimation even when dealing with missing data. For more information on the underlying methodology, please refer to Pradhan, Nychka, and Bandyopadhyay (2025).

Usage

```
emBinRegMAR(
  formula,
  data,
  conflev = 0.95,
  vcorctn = TRUE,
  family = binomial(link = "logit"),
  biascorrectn = TRUE,
  verbose = TRUE
)
```

Arguments

formula	a formula expression as for regression models, of the form response ~ predictors. The response should be a numeric binary variable with missing values, and predictors can be any variables. A predictor with categorical values with missing can be used in the model. See the documentation of formula for other details.
data	Input data for fitting the model
conflev	a value for the confidence interval, the default is 0.95
vcorctn	a variance-covariance matrix computation using Louis (1982). Defualt is TRUE.
family	A character string specifying the type of model family. The default is family=binomial (lin=logit)
biascorrectn	a TRUE or FALSE value, an option for bias reduced estimates due to Firth (1993). The default is TRUE
verbose	a TRUE or FALSE value, default is verbose = TRUE

Details

The family parameter in the emBinRegMAR function allows you to specify the probability distribution and link function for the response variable in the linear model. It determines the nature of the relationship between the predictors and the response variable. The family argument is particularly important when working with binary data, where the response variable has only two possible outcomes. In such cases, you typically want to fit a logistic regression model.

Currently family=binomial is supported for binary data:

You can also specify different link functions within binomial family. The default link function is the logit function, which models the log-odds of success. Other available link functions include:

- "probit" for the probit link function, which models the cumulative standard normal distribution.
- "cloglog" for the complementary log-log link function, which models the complementary log-log of the survival function.

It is important to choose the appropriate link function based on the specific characteristics and assumptions of your binary data. The default "binomial" family with the logit link function is often a good starting point, but alternative link functions might be more appropriate depending on the

research question and the nature of the data. Note that, this function uses the function 'emforbeta' function. For more details of the function and corresponding different output objects, review the 'emforbeta' function.

Value

return the glm estimates

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Louis, T. A. (1982). Finding the observed information when using the EM algorithm. Proceedings of the Royal Statistical Society, Ser B, 44, 226-233.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
data(ibrahim)
#Fits a logistic regression mode with missing categorical covariates using Ibrahim (1990)

fit <- emBinRegMAR(y~x1+x2+x3, data=ibrahim)
fit

data(est45)
f_fit <- emBinRegMAR (resp ~ Fetoprtn + Antigen + Jaundice + Age, data = est45, biascorrectn=FALSE)
f_fit

data(est45)
f_fit <- emBinRegMAR (resp ~ Fetoprtn + Antigen + Jaundice + Age, data = est45, biascorrectn=FALSE)
f_fit

# -----------------Bias reduced estimates due to Firth (1993) --------------
f_fit1 <- emBinRegMAR (resp ~ Fetoprtn + Antigen + Jaundice + Age, data = est45, biascorrectn=TRUE)
f_fit1
```

`emBinRegMixedMAR` *Fits binary regression models with both nonignorable missing responses and missing categorical covariates.*

Description

This function allows users to fit generalized linear models with presence of both missing responses that are nonignorable and incomplete predictors that are categorical. The model is fitted using an EM-based method, which ensures reliable parameter estimation even when dealing with missing data. For more information on the underlying methodology, please refer to Pradhan, Nychka, and Bandyopadhyay (2025).

Usage

```
emBinRegMixedMAR(
  formula,
  data,
  conflev = 0.95,
  adtnlCovforR = NULL,
  vcorctn = TRUE,
  family = binomial(link = "logit"),
  biascorrectn = TRUE,
  verbose = TRUE
)
```

Arguments

`formula`	a formula expression as for regression models, of the form response ~ predictors. The response should be a numeric binary variable with missing values, and predictors can be any variables. A predictor with categorical values with missing can be used in the model. See the documentation of formula for other details.
`data`	Input data for fitting the model.
`conflev`	a value for the confidence interval, the default is 0.95
`adtnlCovforR`	an optional list of covariates to be used to fit the logistic regression `logit(R) ~ response+predictors+adtnlCovforR`. `adtnlCovforR` has to be supplied as a vector. Default is `NULL`.
`vcorctn`	a TRUE or FALSE value, by default it is FALSE. If TRUE, it calculates a variance and standard error using Louis (1982). The default is `vcorctn= TRUE`.
`family`	A character string specifying the type of model family. The default is `family=binomial (lin=logit)`.
`biascorrectn`	a TRUE or FALSE value, an option for bias reduced estimates due to Firth (1993). The default is `biascorrectn=TRUE`.
`verbose`	a TRUE or FALSE value, default is verbose = TRUE

Details

The `family` parameter in the `emBinRegMixedMAR` function allows you to specify the probability distribution and link function for the response variable in the linear model. It determines the nature of the relationship between the predictors and the response variable. The `family` argument is particularly important when working with binary data, where the response variable has only two possible outcomes. In such cases, you typically want to fit a logistic regression model.

Currently family=binomial is supported for binary data:

You can also specify different link functions within binomial family. The default link function is the logit function, which models the log-odds of success. Other available link functions include:

- "probit" for the probit link function, which models the cumulative standard normal distribution.
- "cloglog" for the complementary log-log link function, which models the complementary log-log of the survival function.

It is important to choose the appropriate `link` function based on the specific characteristics and assumptions of your binary data. The default "binomial" family with the logit link function is often a good starting point, but alternative link functions might be more appropriate depending on the research question and the nature of the data.

Value

return the glm estimates

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Louis, T. A. (1982). Finding the observed information when using the EM algorithm. Proceedings of the Royal Statistical Society, Ser B, 44, 226-233.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Addressing Missing Responses and Categorical Covariates in Binary Regression Modeling: An Integrated Framework (submitted).

Examples

```
data(testyxm) # testyxm is a list called dt
dataWithMiss <- testyxm$dataMissing
fit <- emBinRegMixedMAR(Wheeze ~ city + soc + cond,
                        data = dataWithMiss, adtnlCovforR = c("age"),
                        biascorrectn=TRUE)
#display summary of the beta estimates of the model
```

```
fit$beta

#display summary of the alpha estimates of the model used
#for non-ignorability setting of the missing responses
fit$alpha

# Examples using Firth (1993) type bias reduction. Complete case analysis or
# biascorrection=FALSE encounters separation
fit <- emBinRegMixedMAR(resp~Numnill+Numsleep+Smoke+Set+Reftime,
                        data=meningitis60ymis, biascorrectn=TRUE)
#display summary of the beta estimates of the model
fit$beta

#display summary of the alpha estimates of the model used
#for non-ignorability setting of the missing responses
fit$alpha
```

`emBinRegNonIG` *Fitting binary regression with missing responses that are nonignorable based on Ibrahim and Lipsitz (1996)*

Description

This function allows users to fit binary regression models with nonignorable missing responses. The model is fitted using a likelihood-based method, which ensures reliable parameter estimation even when dealing with missing data. For more information on the underlying methodology, please refer to Pradhan, Nychka, and Bandyopadhyay (2025).

Usage

```
emBinRegNonIG(
  formula,
  data,
  conflev = 0.95,
  vcorctn = TRUE,
  biascorrectn = TRUE,
  verbose = TRUE
)
```

Arguments

`formula`	a formula expression as for regression models, of the form response ~ predictors. The response should be a numeric binary variable with missing values, and predictors can be any variables. A predictor with categorical values with missing can be used in the model. See the documentation of formula for other details.
`data`	Input data for fitting the model

conflev	a value for the confidence interval, the default is 0.95
vcorctn	a variance-covariance matrix computation using Louis (1982). Default is TRUE.
biascorrectn	a TRUE or FALSE value, an option for bias reduced estimates due to Firth (1993). The default is TRUE
verbose	a TRUE or FALSE value, default is verbose = TRUE

Details

The family parameter in the emBinRegNonIG function allows you to specify the binomial distribution and link function for the response variable in the linear model. It determines the nature of the relationship between the predictors and the response variable. The family argument is particularly important when working with binary data, where the response variable has only two possible outcomes. In such cases, you typically want to fit a binary regression model.

You can also specify different link functions within binomial family. The default link function is the logit function, which models the log-odds of success. Other available link functions include:

- "probit" for the probit link function, which models the cumulative standard normal distribution.
- "cloglog" for the complementary log-log link function, which models the complementary log-log of the survival function.

It is important to choose the appropriate link function based on the specific characteristics and assumptions of your binary data. The default "binomial" family with the logit link function is often a good starting point, but alternative link functions might be more appropriate depending on the research question and the nature of the data. Note that, this function uses the function 'emforbeta' function. For more details of the function and corresponding different output objects, review the 'emforbeta' function.

Value

return the glm estimates

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Ibrahim, J. G., and Lipsitz, S. R. (1996). Parameter Estimation from Incomplete Data in Binomial Regression when the Missing Data Mechanism is Nonignorable, Biometrics, 52, 1071–1078.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Louis, T. A. (1982). Finding the observed information when using the EM algorithm. Proceedings of the Royal Statistical Society, Ser B, 44, 226-233.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
data(incontinence)
#Fits a binary regression mode with nonignorable missing responses using Ibrahim and Lipsitz (1996)
#biascorrectn=TRUE enables Firth type bias correction of the parameter estimates
fit <- emBinRegNonIG(y~x1+x2+x3, data=incontinence, biascorrectn=TRUE)
fit$beta
#prints the nonignorable missing mechanism
summary(fit$alpha)
```

---

emforbeta *Fitting binary regression with missing categorical covariates using likelihood based method*

Description

This function allows users to fit generalized linear models with incomplete predictors that are categorical. The model is fitted using a likelihood-based method, which ensures reliable parameter estimation even when dealing with missing data. For more information on the underlying methodology, please refer to Pradhan, Nychka, and Bandyopadhyay (2025).

Usage

```
emforbeta(
  formula,
  data,
  family = "binomial",
  vcorctn = FALSE,
  method = "glm.fit",
  NIterations = 50,
  verbose = FALSE,
  theta = NULL,
  convergenceCriterion = 1e-04,
  augmented = NULL,
  VarWithMissingVal = NULL
)
```

Arguments

formula	a formula expression as for regression models, of the form response ~ predictors. The response should be a numeric binary variable with missing values, and predictors can be any variables. A predictor with categorical values with missing can be used in the model. See the documentation of formula for other details.
data	Input data for fitting the model
family	a character string specifying the type of model family. The default is family=binomial (lin=logit)

vcorctn	a TRUE or FALSE value, by default it is FALSE. If TRUE, it calculates a variance and standard error using Louis (1982)
method	a method="brglmFit" or method="glm.fit" will be used for fitting model. The method="brglmFit" fits generalized linear models using bias reduction methods (Kosmidis, 2014), and other penalized maximum likelihood methods.The deafult option method="glm.fit" fits regression with generalized linear models.
NIterations	is the number of iterations to be used for convergence. The default is NIterations=50
verbose	a TRUE or FALSE value, by default it is FALSE. A value TRUE prints all intermediate computational details
theta	a vector containing multinomial parameters that sums to 1, default is NULL
convergenceCriterion	Convergence criteria to be used for convergence. The default is 1e-4
augmented	is the name of an augmented data. The default is NULL
VarWithMissingVal	is a vector of variables including missing values. The default is NULL

Details

The family parameter in the emforbeta function allows you to specify the probability distribution and link function for the response variable in the linear model. It determines the nature of the relationship between the predictors and the response variable. The family argument is particularly important when working with binary data, where the response variable has only two possible outcomes. In such cases, you typically want to fit a logistic regression model.

The following commonly used families are supported for binary data:

- "binomial" for a binomial distribution, suitable for binary or dichotomous response variables.

You can also specify different link functions within binomial family. The default link function is the logit function, which models the log-odds of success. Other available link functions include:

- "probit" for the probit link function, which models the cumulative standard normal distribution.
- "cloglog" for the complementary log-log link function, which models the complementary log-log of the survival function.

It is important to choose the appropriate link function based on the specific characteristics and assumptions of your binary data. The default "binomial" family with the logit link function is often a good starting point, but alternative link functions might be more appropriate depending on the research question and the nature of the data. See also the function 'emBinRegMAR' function.

Value

return the glm estimates

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Louis, T. A. (1982). Finding the observed information when using the EM algorithm. Proceedings of the Royal Statistical Society, Ser B, 44, 226-233.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
data(sixcitydata)
f_fit <- emforbeta(Wheeze~city+soc+cond,
                   data=sixcitydata,
                   vcorctn= TRUE,
                   family=binomial(link="logit"),
                   method="glm.fit")
summary(f_fit$mfit) #creates the summary like glm using the return object mfit
vcov_beta<-f_fit$cvcov #creates variance using Louis (1982)

# Computes the standard error of the estimates
se_beta_em<-sqrt(diag(vcov_beta))
se_beta_em

# Firth correction
f_fit <- emforbeta(Wheeze~city+soc+cond,
                   data=sixcitydata,
                   family=binomial(link="logit"),
                   method="brglmFit")
# creates the summary like glm using the return object mfit

data(ibrahim)
f_fit2 <- emforbeta(y~x1+x2+x3,
                    data=ibrahim,
                    family="binomial")
summary(f_fit2$mfit) #creates the summary like glm using the return object mfit

f_fit2 <- emforbeta(y~x1+x2+x3,
                    data=ibrahim,
                    family=binomial (link="probit"),
                    method="brglmFit")
# creates the summary like glm using the return object mfit
summary(f_fit2$mfit) #

data(est)
```

```
f_fit <- emforbeta(survive~Fetoprtn+Antigen+Jaundice+Age,
                   data=est,
                   family=binomial,
                   method="glm.fit")
summary(f_fit$mfit)

f_fit <- emforbeta(survive~Fetoprtn+Antigen+Jaundice+Age,
                   data=est,
                   family=binomial,
                   method="brglmFit")
# Firth corrected estimates with out Louis (1982) correction (see Maiti and Pradhan (2009))
summary(f_fit$mfit)

data(metastmelanoma)
f_fit <- emforbeta(failcens~size+type+nodal+age+sex+trt,
                   data=metastmelanoma,
                   family=binomial,
                   method="glm.fit")
summary(f_fit$mfit)

f_fit <- emforbeta(failcens~size+type+nodal+age+sex+trt,
                   data=metastmelanoma,
                   family=binomial,
                   method="brglmFit")
# Firth corrected estimates with out Louis (1982) correction (see Maiti and Pradhan (2009))
summary(f_fit$mfit)

data(felinedata)
f_fit <- emforbeta(chlamy~Season+Agegrp+Conj+FHV1,
                   data=felinedata,
                   family=binomial,
                   method="glm.fit")
summary(f_fit$mfit)

f_fit <- emforbeta(chlamy~Season+Agegrp+Conj+FHV1,
                   data=felinedata,
                   family=binomial,
                   method="brglmFit")
# Firth corrected estimates with out Louis (1982) correction
summary(f_fit$mfit)
```

---

emil *Fitting binary regression model with missing responses based on Ibrahim and Lipsitz (1996)*

Description

This function enables users to fit generalized linear models when handling incomplete data in the response variable. The missing responses are assumed to be nonignorable. The model is fitted using

a novel likelihood-based method proposed by Ibrahim and Lipsitz(1996).

Usage

```
emil(
  formula,
  data,
  adtnlCovforR = NULL,
  eps0 = 1e-05,
  maxit = 75,
  family = "binomial",
  method = "brglmFit"
)
```

Arguments

formula	a formula expression as for regression models, of the form response ~ predictors. The response should be a numeric binary variable with missing values, and predictors can be any variables. A predictor with categorical values with missing can be used in the model. See the documentation of formula for other details.
data	an optional data frame in which to interpret the variables occurring in formula.
adtnlCovforR	an optional list of covariates to be used to fit the logistic regression logit(R) ~ response+predictors+adtnlCovforR. adtnlCovforR has to be supplied as a vector. Default is NULL.
eps0	arguments to be used to for the convergence criteria of the maximum likelihood computation of the joint likelihood function. The default is 1e-3.
maxit	arguments to be used to for the maximization of the joint likelihood function. The default is 50.
family	A character string specifying the type of model family.
method	a method="brglmFit" or method="glm.fit" will be used for fitting model. The method="brglmFit" fits generalized linear models using bias reduction methods (Kosmidis, 2014), and other penalized maximum likelihood methods.

Details

The family parameter in the emil function allows you to specify the probability distribution and link function for the response variable in the linear model. It determines the nature of the relationship between the predictors and the response variable. The family argument is particularly important when working with binary data, where the response variable has only two possible outcomes. In such cases, you typically want to fit a binary regression model with an appropriate link.

Currently the package only supports family=binomial for binary or dichotomous response variables.

You can also specify different link functions within the family=binomial. The default link function is the logit function, which models the log-odds of success. Other available link functions include:

- "probit" for the probit link function, which models the cumulative standard normal distribution.

- "cloglog" for the complementary log-log link function, which models the complementary log-log of the survival function.

It is important to choose the appropriate `family` and `link` function based on the specific characteristics and assumptions of your binary data. The default "binomial" family with the logit link function is often a good starting point, but alternative link functions might be more appropriate depending on the research question and the nature of the data.

Value

return the generalized linear model estimates

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Ibrahim, J. G., and Lipsitz, S. R. (1996). Parameter Estimation from Incomplete Data in Binomial Regression when the Missing Data Mechanism is Nonignorable, Biometrics, 52, 1071–1078.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Louis, T. A. (1982). Finding the observed information when using the EM algorithm. Proceedings of the Royal Statistical Society, Ser B, 44, 226-233.

Maity, A., Pradhan, V., Das U (2019). Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. The American Statistician, (73) 340-349.

Pradhan V, Nychka DW, Bandyopadhyay S (2025). Addressing Missing Responses and Categorical Covariates in Binary Regression Modeling: An Integrated Framework (to be submitted).

Examples

```
# using incontinence data
fit <- emil(y~x1+x2+x3,
                   data=incontinence,
                   family=binomial,
                   method="brglmFit")
summary(fit$fit_y)
```

`emyxmiss` *Fitting generalized linear models with Incomplete data*

Description

This function enables users to fit generalized linear models when handling incomplete data in both the response variable and categorical covariates. The missing responses are assumed to be nonignorable, while missing categorical covariates are assumed to be missing at random. The model is fitted using a novel likelihood-based method proposed by Pradhan, Nychka, and Bandyopadhyay (2025).

Usage

```
emyxmiss(
  formula,
  data,
  adtnlCovforR = NULL,
  eps0 = 0.001,
  maxit = 75,
  family = "binomial",
  method = "glm.fit"
)
```

Arguments

formula	a formula expression as for regression models, of the form response ~ predictors. The response should be a numeric binary variable with missing values, and predictors can be any variables. A predictor with categorical values with missing can be used in the model. See the documentation of formula for other details.
data	an optional data frame in which to interpret the variables occurring in formula.
adtnlCovforR	an optional list of covariates to be used to fit the logistic regression logit(R) ~ response+predictors+adtnlCovforR. adtnlCovforR has to be supplied as a vector. Default is NULL.
eps0	arguments to be used to for the convergence criteria of the maximum likelihood computation of the joint likelihood function. The default is 1e-3.
maxit	arguments to be used to for the maximization of the joint likelihood function. The default is 50.
family	A character string specifying the type of model family.
method	a method="brglmFit" or method="glm.fit" will be used for fitting model. The method="brglmFit" fits generalized linear models using bias reduction methods (Kosmidis, 2014), and other penalized maximum likelihood methods.

Details

The family parameter in the emyxmmiss function allows you to specify the probability distribution and link function for the response variable in the linear model. It determines the nature of the relationship between the predictors and the response variable. The family argument is particularly important when working with binary data, where the response variable has only two possible outcomes. In such cases, you typically want to fit a logistic regression model.

The following commonly used families are supported for binary data:

- "binomial" for a binomial distribution, suitable for binary or dichotomous response variables.

You can also specify different link functions within each family. For the "binomial" family, the default link function is the logit function, which models the log-odds of success. Other available link functions include:

- "probit" for the probit link function, which models the cumulative standard normal distribution.

- "cloglog" for the complementary log-log link function, which models the complementary log-log of the survival function.

It is important to choose the appropriate `family` and `link` function based on the specific characteristics and assumptions of your binary data. The default "binomial" family with the logit link function is often a good starting point, but alternative link functions might be more appropriate depending on the research question and the nature of the data.

Value

return the generalized linear model estimates

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Ibrahim, J. G., and Lipsitz, S. R. (1996). Parameter Estimation from Incom- plete Data in Binomial Regression when the Missing Data Mechanism is Nonignorable, Biometrics, 52, 1071–1078.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Louis, T. A. (1982). Finding the observed information when using the EM algorithm. Proceedings of the Royal Statistical Society, Ser B, 44, 226-233.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan V, Nychka DW, Bandyopadhyay S (2025). Addressing Missing Responses and Categorical Covariates in Binary Regression Modeling: An Integrated Framework (to be submitted).

Examples

```
data(testyxm) # testyxm is a list called dt
dataWithMiss <- testyxm$dataMissing
# Binary regression with link=logit
fit_yx <- emyxmiss(Wheeze ~ city + soc + cond,
                   data = dataWithMiss,
                   adtnlCovforR = c("age"),
                   family = binomial(link = "logit"),
                   method = "brglmFit")
fit_yx

# Binary regression with link=probit
fit_yx <- emyxmiss(Wheeze ~ city + soc + cond,
                   data = dataWithMiss,
                   adtnlCovforR = c("age"),
                   family = binomial(link = "probit"))
fit_yx


# Firth correction and link=probit
```

```
fit_yx <- emyxmiss(Wheeze ~ city + soc + cond,
                   data = dataWithMiss,
                   adtnlCovforR = c("age"),
                   family = binomial(link = "probit"),
                   method = "brglmFit")
fit_yx

# on simulated data
demo_df <- simulateCovariateData(50, nCov=6)
simulated_df <- simulateData(demo_df)
testMissData <- simulated_df$dataMissing
fit_yx <- emyxmiss(y~x2+x3+x4,
                   data=testMissData,
                   adtnlCovforR=c("x1"),
                   family=binomial,
                   method="glm.fit")
fit_yx
summary(fit_yx$fit_y)
```

`est` *EST data – Eastern Cooperative Oncology Group clinical trials, EST 2282*

Description

The dataset `est` is from the Eastern Cooperative Oncology Group clinical trials, specifically EST 2282 (Falkson, Cnaan, and Simson, 1990) and EST 1286 (Falkson et al., 1995). The dataset consists of 191 observations. It includes several covariates: Fetoprtn (alpha fetoprotein), Antigen (antihepatitis B antigen), Jaundice (a biochemical marker; coded as 1 if present, 0 otherwise), and Age (age in years). The response variable Y represents the number of cancerous liver cells present at the start of the clinical trial.

To assess the impact of these covariates on the likelihood of survival, a new variable called "survive" is created. "survive" is dichotomized based on Y: it is set to 1 if the number of cancerous liver cells is less than or equal to 8, and 0 otherwise.

Maiti and Pradhan (2009) fitted a logistic regression using the model `survive ~ Fetoprtn + Antigen + Jaundice + Age`. This model explores the relationship between the covariates and the likelihood of survival for patients in the clinical trials.

Usage

```
est
```

Format

A data frame with 191 rows and several variables:

Y Response variable representing the number of cancerous liver cells

Fetoprtn Alpha fetoprotein

BMI Body Mass Index

Antigen Anti-hepatitis B antigen

Jaundice Jaundice indicator, coded as 1 if present, 0 otherwise

Age Age in years

Weeks times in weeks

survive Dichotomized survival variable based on Y

Source

Generated for example purposes

References

Cytel Inc (2010). LogXact 9 User Manual: Discrete Regression Analysis. Cambridge, Massachusetts: Cytel Inc.

Falkson, G., Lipsitz, S., Borden, E., Simson, I., W., and Haller, D. (1995). A ECOG randomized phase II study of beta interferon and Menogoril. American Journal of Clinical Oncology 18, 287–292.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
data(est)
f_fit <- emforbeta(survive ~ Fetoprtn + Antigen + Jaundice + Age,
                   data = est,
                   family = binomial, method = "glm.fit")
summary(f_fit$mfit)

f_fit <- emforbeta(survive ~ Fetoprtn + Antigen + Jaundice + Age,
                   data = est,
                   family = binomial, method = "brglmFit")
summary(f_fit$mfit)
```

`est45` *EST data – Eastern Cooperative Oncology Group clinical trials, EST 2282*

Description

The dataset est45 is from the Eastern Cooperative Oncology Group clinical trials, specifically EST 2282 (Falkson, Cnaan, and Simson, 1990) and EST 1286 (Falkson et al., 1995) containing 45 observations. The dataset consists of 191 observations. It includes several covariates: Fetoprtn (alpha fetoprotein), Antigen (antihepatitis B antigen), Jaundice (a biochemical marker; coded as 1 if present, 0 otherwise), and Age (age in years). The response variable Y represents the number of cancerous liver cells present at the start of the clinical trial.

To assess the impact of these covariates on the likelihood of survival, a new variable called "survive" is created. "survive" is dichotomized based on Y: it is set to 1 if the number of cancerous liver cells is less than or equal to 8, and 0 otherwise.

Usage

```
est45
```

Format

A data frame with 45 rows and 9 variables:

Y Response variable

Weeks Time in weeks

Fetoprtn Alpha fetoprotein

Antigen Anti-hepatitis B antigen

Jaundice Jaundice indicator

BMI Body mass index

Age Age in years

grp Group identifier

resp Response variable dichotomized

Source

Generated for example purposes

References

Cytel Inc (2010). LogXact 9 User Manual: Discrete Regression Analysis. Cambridge, Massachusetts: Cytel Inc.

Falkson, G., Lipsitz, S., Borden, E., Simson, I., W., and Haller, D. (1995). A ECOG randomized phase II study of beta interferon and Menogoril. American Journal of Clinical Oncology 18, 287–292.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
data(est45)
f_fit <- emforbeta(resp ~ Fetoprtn + Antigen + Jaundice + Age,
                   data = est45, family = binomial, method = "glm.fit")
summary(f_fit$mfit)

#Bias-reduced estimates due to Firth (1993)
f_fit <- emforbeta(resp ~ Fetoprtn + Antigen + Jaundice + Age,
                   data = est45, family = binomial, method = "brglmFit")
summary(f_fit$mfit)
```

`felinedata` *felinedata – Chlamydial Infection in Cats*

Description

In a study conducted by Sykes et al. (1999), the risk factors for Chlamy, a chlamydial infection in cats, were investigated. The analysis considered important variables such as FHV1 (Herpes virus infection), Season, Conjunctivitis (Conj), and Age group. Season was coded from 1 to 4 to represent the seasons, FHV1 was binary (1 for infected cats, 0 for non-infected cats), Conj was binary (1 if present, 0 if absent), and Age group was categorized into specific ranges. The original dataset had 462 observations, with around 20% missing values. After removing missing values, the analysis was conducted with a sample size of 371. The fitted model included Chlamy as the outcome variable, with FHV1, Season, Conj, and Age group as predictors, treating Age group and Season as class variables with a base value of 1.

Usage

```
felinedata
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 120 rows and 6 columns.

References

Cytel Inc (2010). LogXact 9 User Manual: Discrete Regression Analysis. Cambridge, Massachusetts: Cytel Inc.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2024). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Sykes, J. E., Anderson, G. A., Studdert, V. P., and Browning, G. F. (1999). Prevalence of feline Chlamydia psittaci and feline her- pesvirus 1 in cats with upper respiratory tract disease. Journal of Veterinary Internal Medicine 13, 153–162.

Examples

```
data("felinedata")
expanded_data <- felinedata[rep(seq_len(nrow(felinedata)), felinedata$GrpSize), ]
fit <- glm(chlamy ~ FHV1+Season+Conj+Agegrp, data=expanded_data, family="binomial")
# High Std. Error values indicate the model did not converge for complete case analysis
summary(fit)

#Fitting the model with emforbeta using Ibrahim (1990)
fit2 <- emforbeta(chlamy ~ FHV1+Season+Conj+Agegrp, data=expanded_data, family="binomial")
# High Std. Error values indicate the model did not converge for complete case analysis
summary(fit2$mfit)

#Fitting the model with Ibrahim (1990) and Firth correction (Maiti and Pradhan (2009))
fit2 <- emforbeta(chlamy ~ FHV1+Season+Conj+Agegrp,
                  data=expanded_data, family="binomial", method = "brglmFit")
summary(fit2$mfit)
```

---

ibrahim *ibrahim data – Ibrahim (1990) JASA*

Description

The dataset ibrahim is from Ibrahim, IG (1990, 85, 765–769, JASA). The data contains a response variable y and predictors x1 x2 x3, and the total number of observations is 82.

Usage

```
ibrahim
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 82 rows and 4 columns.

References

Cytel Inc (2010). LogXact 9 User Manual: Discrete Regression Analysis. Cambridge, Massachusetts: Cytel Inc.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
data(ibrahim)
f_fit <- emBinRegMAR(y ~ x1+x2+x3, data=ibrahim, family="binomial", biascorrectn=FALSE)
f_fit$beta
#Firth type bias correction
f_fit <- emBinRegMAR(y ~ x1+x2+x3, data=ibrahim, family="binomial", biascorrectn=TRUE)
f_fit$beta
```

`incontinence` *incontinence- incontinence Data taken from brlrmr pacakge*

Description

The dataset incontinence is from. The dataset is available in the brlrmr pacakge. Pradhan, Nychka and Bandyopadhyay (2024) fitted the model `y~x1+x2+x3`.

Usage

```
incontinence
```

Format

A data frame with several rows and columns representing various variables:

y response variable

x1 is a covariate

x2 is a covariate

x3 is a covariate

References

Maity, A., Pradhan, V., Das, U. (2019). Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. The American Statistician, (73) 340-349.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2024). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
fit <- emil(y~x1+x2+x3, data=incontinence,family=binomial, method="brglmFit")
# display summary of the beta estimates of the model
summary(fit$fit_y)
# for non-ignorability setting of the missing responses
summary(fit$fit_r)
```

`llkmiss` *Fitting binary regression with missing categorical covariates using new likelihood based method that does not require EM algorithm*

Description

This function allows users to fit logistic regression models with incomplete predictors that are categorical. The model is fitted using a new likelihood-based method, which ensures reliable parameter estimation even when dealing with missing data. For more information on the underlying methodology, please refer to Pradhan, Nychka, and Bandyopadhyay (2025).

Usage

```
llkmiss(par, data, formula, augData, biasCorr = TRUE)
```

Arguments

`par`	A vector including a list of parameters to be estimated. This include the beta (the regression parameters) and theta, the multinomial paraters for observing a missing covaraite pattern.
`data`	Input data for fitting the model
`formula`	A formula expression as for regression models, of the form `response ~ predictors`. The response should be a numeric binary variable with missing values, and predictors can be any variables. A predictor with categorical values with missing can be used in the model. See the documentation of formula for other details.
`augData`	An augmented data including all possible covarites that could have been observed.
`biasCorr`	a TRUE or FALSE value, by default it is TRUE.

Value

return the regression estimates

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Bridging Gaps in Logistic Regression: Tackling Missing Categorical Covariates with a New Likelihood Method (to be submitted).

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). glmFitMiss: Binary Regression with Missing Data in R (to be submitted)

`logRegMAR` *Fitting binary regression with missing categorical covariates using new likelihood based method*

Description

This function allows users to fit logistic regression models with incomplete predictors that are categorical. The model is fitted using a new likelihood-based method, which ensures reliable parameter estimation even when dealing with missing data. For more information on the underlying methodology, please refer to Pradhan, Nychka, and Bandyopadhyay (2024).

Usage

```
logRegMAR(formula, data, conflev = 0.95, correctn = TRUE, verbose = TRUE)
```

Arguments

`formula`	A formula expression as for regression models, of the form `response ~ predictors`. The response should be a numeric binary variable with missing values, and predictors can be any variables. A predictor with categorical values with missing can be used in the model. See the documentation of formula for other details.
`data`	Input data for fitting the model
`conflev`	Confidence level, the default is 0.95
`correctn`	a TRUE or FALSE value, by default it is TRUE.
`verbose`	a TRUE or FALSE value, default is verbose = TRUE

Value

return the logistic regression estimates

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38. doi:10.2307/2336755.

Kosmidis, I., Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. Biometrika, 108, 71-82. doi:10.1093/biomet/asaa052.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Bridging Gaps in Logistic Regression: Tackling Missing Categorical Covariates with a New Likelihood Method (to be submitted).

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). glmFitMiss: Binary Regression with Missing Data in R (to be submitted)

Examples

```
# -----------------Example 1: Metastatic Melanoma --------------------------

est1 <- logRegMAR (failcens ~ size+type+nodal+age+sex+trt,
                   data = metastmelanoma, conflev = 0.95, correctn = FALSE)

est1
# -----------------Bias reduced estimates due to Firth (1993) --------------
est2 <- logRegMAR (failcens ~ size+type+nodal+age+sex+trt,
                   data = metastmelanoma, conflev = 0.95, correctn = TRUE)

est2
# -----------------Bias reduced estimates due to Firth (1993) --------------
est2 <- logRegMAR (CaseCntrl ~ Numnill+Numsleep+Smoke+Set+Reftime,
                   data=meningitis, conflev = 0.95, correctn = TRUE)
est2
```

`meningitis` *meningitis- Meningococcal Disease Data with missing data in the response variable*

Description

The dataset meningitis is from a brief outbreak of meningococcal disease at the University of Illinois, Urbana-Champaign campus in the years 1991 and 1992. The dataset is available in the LogXact software and also analyzed in Imrey et al. (1996). Maiti and Pradhan (2009) fitted a logistic regression using the model CaseCntrl ~ Numill + Numsleep + Smoke + Set + Reftime.

Usage

```
meningitis
```

Format

A data frame with several rows and columns representing various variables:

CaseCntrl Case control status

Numnill Number of illnesses

Numsleep Number of sleep disturbances

Smoke Smoking status

Set Set variable

Reftime Reference time

References

Cytel Inc (2010). LogXact 9 User Manual: Discrete Regression Analysis. Cambridge, Massachusetts: Cytel Inc.

Imrey, P. B., Jackson, L. A., Ludwinski, P. H., England, A. C. II, Fox, B. C., Isdale, L. B., Reeves, M. W., and Wenger, J. D. (1996). Outbreak of serogroup C meningococcal disease associated with campus bar patronage. American Journal of Epidemiology 143, 624–630.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2024). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
# Examples using Firth (1993) type bias reduction. Complete case analysis or
# biascorrection=FALSE encounters separation
fit <- emBinRegMAR(CaseCntrl~Numnill+Numsleep+Smoke+Set+Reftime,
                        data=meningitis, biascorrectn=TRUE)
# display summary of the beta estimates of the model
fit$beta
# display summary of the alpha estimates of the model used
# for non-ignorability setting of the missing responses
fit$alpha
```

`meningitis60ymis` *meningitis60ymis- Meningococcal Disease Data with missing data in the response variable*

Description

The dataset meningitis is from a brief outbreak of meningococcal disease at the University of Illinois, Urbana-Champaign campus in the years 1991 and 1992. The dataset is available in the LogXact software and also analyzed in Imrey et al. (1996). Pradhan, Nychka and Bandyopadhyay (2024) fitted the model `resp~Numnill+Numsleep+Smoke+Set+Reftime`, where the the response variable resp included missing value.

Usage

```
meningitis60ymis
```

Format

A data frame with several rows and columns representing various variables:

CaseCntrl Case control status

Numnill Number of illnesses

Numsleep Number of sleep disturbances

R an indicator variable for missing CaseCntrl

Smoke Smoking status

Set Set variable

m indicator variable for observations with missing values

Reftime Reference time

resp Response variable with missing data for the varaible CaseCntrl

References

Cytel Inc (2010). LogXact 9 User Manual: Discrete Regression Analysis. Cambridge, Massachusetts: Cytel Inc.

Imrey, P. B., Jackson, L. A., Ludwinski, P. H., England, A. C. II, Fox, B. C., Isdale, L. B., Reeves, M. W., and Wenger, J. D. (1996). Outbreak of serogroup C meningococcal disease associated with campus bar patronage. American Journal of Epidemiology 143, 624–630.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
fit <- emBinRegMixedMAR(resp~Numnill+Numsleep+Smoke+Set+Reftime,
                        data=meningitis60ymis, biascorrectn=TRUE)
# display summary of the beta estimates of the model
fit$beta
# display summary of the alpha estimates of the model used
# for non-ignorability setting of the missing responses
fit$alpha
```

`metastmelanoma` *metastmelanoma - metastatic melanoma trial data*

Description

The dataset data from a cancer clinical trial and the results are published in Kirkwood et al. (1996). In this study following surgery for deep primary or metastatic melanoma, the overall survival and the disease-free effect of Interferon alpha-2b (IFN) was investigated on 285 patients. Maiti and Pradhan (2009) fitted a logistic regression considering failcens as the response variable, where failcens is 1 if the subject relapses and 0 otherwise. We fit the model including six important predictors size type nodal age sex trt; where size is the size of primary in cm2, which is dichotomized at the median; type is the type of primary containing two levels— superficial spreading and other; nodal is the presence or absence of microscopic nonpalpable and palpable regional lymph node metastasis—1 for node positive and 0 otherwise; age is the age of a subject in years; sex is a variable indicating the gender (male or female); and finally trt is the treatment containing two levels (1 if treated with IFN and 0 otherwise).

Usage

```
metastmelanoma
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 285 rows and 11 columns.

References

Cytel Inc (2010). LogXact 9 User Manual: Discrete Regression Analysis. Cambridge, Massachusetts: Cytel Inc.

Kirkwood, J. M., Strawderman, M. H., Ernstoff, M. S., Smith, T. J., Borden, E. C., and Blum, R. H. (1996). Interferon alfa-2b ad- juvant therapy of high-risk resected cutaneous melanoma: The Eastern Cooperative Oncology Group trial EST 1684. Journal of Clinical Oncology 14, 7–17.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
data(metastmelanoma)
f_fit <- emforbeta(failcens ~ size+type+nodal+age+sex+trt,
                   data=metastmelanoma,
                   family=binomial, method="glm.fit")
summary(f_fit$mfit)
vcov_beta<-f_fit$cvcov # variance-covariance calculation using Louis (1982)
vcov_beta
se_beta_em<-sqrt(diag(vcov_beta))
se_beta_em

# Firth Correction
f_fit <- emforbeta(failcens ~ size+type+nodal+age+sex+trt,
                   data=metastmelanoma,
                   family=binomial, method="brglmFit")
summary(f_fit$mfit)
vcov_beta<-f_fit$cvcov # variance-covariance calculation using Louis (1982)
vcov_beta
se_beta_em<-sqrt(diag(vcov_beta))
se_beta_em
```

---

simulateCovariateData *Simulate data with independent categorical covariates*

Description

This function generates a simulated data with independent categorical covariates. The first two covariates namely x1 and x2 are generated using random normal rnorm(n, 40, 20) and random poisson rpois(n, lambda = 4).The remaining covariates are generated at random with categories 0,1,2.

Usage

```
simulateCovariateData(n, nCov = 2)
```

Arguments

n	numner of observations to be generated for the data
nCov	4+ number of covariates to be generated for the data, the first 4 covariates generated based on pre-specified distributions

Value

returns a data frame with covariates x1, x2, ...

Examples

```
simulateCovariateData(10, nCov=15)
```

---

simulateData *Simulate data based on an input covariate data*

Description

This function generates missing data both in the response variables as well as in the predictors. The missing data generation in the last two supplied covariates will be generated based on a predefined mechanisms. Missing data generation in the response variable will be based on the suppilied true alpha.

Usage

```
simulateData(
  dataCov,
  truebeta = c(1, -1, 1, 5),
  truealpha = c(-1, 5, -1, -1, -1, 0.01),
  nsim = 2
)
```

Arguments

dataCov	input data, the default number of covariates is 7 (5+2)
truebeta	the beta parameter to be used to generate binary response values 1/0 s logit(y=1)=x1+x2+x3
truealpha	to be used to generate nonignorable missing values based on the model logit(R=1)=y+x1+x2+x3+x4+..
nsim	number of simulated dataset, default is 2

Value

returns a list with original data called originalData and a data with imputed missing values dataMissing

Examples

```
demo_df <- simulateCovariateData(100, nCov=6)
simulated_df <- simulateData(demo_df, nsim=2)
testMissData <- simulated_df$dataMissing
head(testMissData)
```

---

simulateMissDfYorX *Simulate missing covariate or missing responses data based on an input covariate data*

Description

This function generates missing covariate or missing responses data. The missing data generation in the last two supplied covariates will be generated based on a predefined mechanisms. Missing data generation in the response variable will be based on the suppilied true alpha.

Usage

```
simulateMissDfYorX(
  dataCov,
  truebeta = c(1, -1, 1, 5),
  truealpha = c(-1, 5, -1, -1, -1, 0.01),
  x2Mar = c(1, -1, -1),
  ymiss = FALSE,
  nsim = 1
)
```

Arguments

dataCov	input data, the default number of covariates is 7 (5+2)
truebeta	the beta parameter to be used to generate binary responses 1/0 s logit(y=1)=x1+x2+x3
truealpha	to be used to generate nonignorable missing values based on the model logit(R=1)=y+x1+x2+x3+x4+..
x2Mar	to be used to generate missing values in x2 based on the model logit(x2=missing)=x1+y
ymiss	to be used for missing responses, default is FALSE
nsim	number of simulated dataset, default is 2

Value

returns a list with original data called originalData and a data with imputed missing values dataMissing

Examples

```
demo_df <- simulateCovariateData(100, nCov=6)
simulated_df <- simulateMissDfYorX(demo_df, nsim=2)
testMissData <- simulated_df$dataMissing
head(testMissData)
```

`sixcitydata` *sixcitydata – A very well published Six city data published in many articles including Ware et al (1984), Ibrahim and Lipsitz (1996). Also avaialble in LogXact User Manual. The dataset is a longitudinal study of the health effects of air pollution (ware et al., 1984).*

Description

The 'sixcitydata' dataset contains information on wheezing status, city of residence, maternal smoking habits, socioeconomic status, and medical condition of children at age 11.

The dataset includes the following variables:

- Wheeze: Binary response variable indicating wheezing status of children at age 11 (1 for wheeze, 0 for no wheeze).
- city: Categorical variable indicating city of residence (1 for polluted city, 0 for Portage or Wisconsin).
- smoke: Binary variable indicating mother's smoking habits (1 for >20 cigarettes a day, 0 otherwise).
- soc: Binary variable indicating high socioeconomic status of subject (1 for high socioecnomic status, 0 otherwise).
- cond: Binary variable indicating previous medical condition of subject (1 for previous medical condition, 0 otherwise).

Usage

```
sixcitydata
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 2106 rows and 5 columns.

References

Cytel Inc (2010). LogXact 9 User Manual: Discrete Regression Analysis. Cambridge, Massachusetts: Cytel Inc.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85, 765–769.

Maiti, T., Pradhan, V. (2009). Bias reduction and a solution of separation of logistic regression with missing covariates. Biometrics, 65, 1262-1269.

Pradhan, V., Nychka, D. and Bandyopadhyay, S. (2025). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Ware, JH., Dockery, DW., Spiro, A III., Speizer, FE., and Ferris, BG Jr. (1984). Passive smoking, gas cooking, and respiratory health of children living in sex cities. American Review of Respiratory Disease, 129, 366-374.

Examples

```
data(sixcitydata)
f_fit <- emforbeta(Wheeze ~ city+soc+cond,
                   data=sixcitydata,
                   family=binomial(link="logit"), method="glm.fit")
#creates the summary like glm using the return object mfit
summary(f_fit$mfit)
vcov_beta<-f_fit$cvcov #creates variance using Louis (1982)
se_beta_em<-sqrt(diag(vcov_beta))
se_beta_em
```

`testyxm` *Simulated Test Data* – `testyxm`

Description

A test list data that returns a list called `testyxm` with two components:

- `dataOriginal`: The original data set without missing values.
- `dataMissing`: The data set with artificially introduced missing values.

Usage

```
data(testyxm)
```

Format

A list with the following components:

dataOriginal A data frame with several rows and columns representing various variables.

dataMissing A data frame with missing values corresponding to the same structure as `dataOriginal`.

Details

Simulated Test Data

This dataset is a list called `testyxm` that contains two data frames: `dataOriginal` and `dataMissing`.

References

Pradhan, V., Nychka, D., and Bandyopadhyay, S. (2024). Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples (submitted).

Examples

```
data(testyxm)
Fulldata <- testyxm$dataOriginal
Missdata <- testyxm$dataMissing
```

Index