Package ‘paraconformal’

October 18, 2018

Version 0.1

Date 2018-10-17

Title Conformal Prediction for Generalized Linear Regression Models

Author Daniel J. Eck <daniel.eck@yale.edu>

Maintainer Daniel J. Eck <daniel.eck@yale.edu>

Depends R (>= 3.0.0)

Imports stats, MASS, statmod, conformalInference, parallel

Suggests

ByteCompile FALSE

Description Compute and compare prediciton regions for the normal, Gamma, and inverse Gaussian families in the \{}code{glm} package. There is functionality to construct the usual prediction region that one obtains from maximum likelihood estimation and the delta method, the parametric conformal prediction region, the nonparametric conformal prediction region, and prediction regions from conformalization of residuals.

License MIT + file LICENSE

URL https://bitbucket.org/forrestcrawford/conformal/branches/compare/

R topics documented:

conformalprediction *Prediction Regions for Generalized Linear Regression Models*

Description

Compute and compare prediciton regions for the normal, Gamma, and inverse Gaussian families in the `glm` package. There is functionality to construct the usual prediction region that one obtains from maximum likelihood estimation and the delta method, the parametric conformal prediction region, the nonparametric conformal prediction region, and prediction regions from conformalization of residuals.

Usage

```
conformalprediction(object, ..., newdata = NULL, alpha = 0.10,
 cores = 6, bins = NULL, parametric = TRUE, LS = FALSE, intercept = TRUE,
 nonparametric = FALSE)
```

Arguments

`object` an object of class `"glm"`.

`...` further arguments passed to or from other methods.

`newdata` an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing new observations for which a prediction is desired. If missing, then prediction regions will be provided at the observed data.

`alpha` the error tolerance desired for the prediction region. The default is set at 0.10.

`cores` the number of cores used to compute the conformal prediction regions. The default is set at 6 cores. Users calling this function on machines with fewer than 6 cores are encouraged to change the default.

`bins` an optional argument for specifying the desired number of bins to use along one dimension of the predictor space. If missing, the theoretical large sample optimal bin width is used (width = `O(log(n)/n)^(1/(d+1))` where `n` is the sample size and `d` is the dimension of the main effects).

`parametric` a Boolean variable corresponding to whether or not the parametric conformal region is to be computed. The default is set at TRUE.

`LS` a Boolean variable corresponding to whether or not the prediction region by conformalization of residualsis is to be computed. The default is set at TRUE.

`intercept` a Boolean variable corresponding to whether or not the intercept is included in the regression equation. This is only relevant for the computation of the prediction region by conformalization of residuals (when `LS = TRUE`). The default is set at TRUE.

`nonparametric` a Boolean variable corresponding to whether or not the nonparametric conformal region is to be computed. The default is set at TRUE.

Details

This function calls on the regions function to compute all of the prediction regions outlined in the description. This function is easier to use than the regions function since it can be called directly on an object of class glm.

Value

regions has functionality to return the usual prediction region that one obtains from maximum likelihood estimation and the delta method, the parametric conformal prediction region, the nonparametric conformal prediction region, and prediction regions from conformalization of residuals.

paraconformal The parametric conformal prediction region which is returned when parametric = TRUE.

nonparaconformal
The nonparametric conformal prediction region which is returned when nonparametric = TRUE.

LSconformal The parametric prediction region from conformalization of residuals which is returned when LS = TRUE.

interval.plugin
The usual prediction region that one obtains from maximum likelihood estimation and the delta method.

References

Eck, D.~J., Crawford, F.~W., and Aronow, P.~M. (2018+) Conformal prediction for exponential families and generalized linear models. Preprint available on request (email daniel.eck@yale.edu).

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2016) Distribution-Free Predictive Inference for Regression. https://arxiv.org/abs/1604.04173

Lei, J. and Wasserman, L. (2014) Distribution-Free Prediction Bands for Non-parametric Regression. Journal of the Royal Statistical Society: Series B, 76(1), 71-96.

Lei, J., Robins, J., and Wasserman, L. (2013) Distribution Free Prediction Sets. Journal of the American Statistical Association, 108(501), 278-287.

See Also

regions, glm

Examples

```
# example of section 2.4 in Geyer (2009)
# data(sports)
# out <- glmdr(cbind(wins, losses) ~ 0 + ., family = "binomial", data = sports)
#summary(out)
```

`insurance` *Insurance cost data for nonsmokers*

Description

Total health insurance costs for the nonsmokers in a simulated study.

Usage

```
insurance
```

Format

The data consists of the response variable which is total healthcare cost paid by an insurer measured in thousands of dollars (charges) and two predictors which are age in years (age) and body mass index (bmi). The predictor variables are rescaled so that the support of the predictor space is [0,1]^2.

References

Lantz, Brett (2013) *Machine learning with R*, Packt Publishing Ltd. https://www.kaggle.com/lbronchal/explanatory-models-for-healthcare-costs

`regions` *Prediction Regions for Generalized Linear Regression Models*

Description

Compute and compare prediciton regions for the normal, Gamma, and inverse Gaussian families in the `glm` package. There is functionality to construct the usual prediction region that one obtains from maximum likelihood estimation and the delta method, the parametric conformal prediction region, the nonparametric conformal prediction region, and prediction regions from conformalization of residuals.

Usage

```
regions(formula, data, newdata, family = "gaussian", link, alpha = 0.10,
  cores = 6, bins = NULL, intercept = TRUE, parametric = TRUE,
  LS = FALSE, nonparametric = FALSE)
```

Arguments

`formula` an object of class `"formula"` (or one that can be coerced to that class): a symbolic description of the model to be fitted. See `glm` and `formula` for description of the R formula mini-language.

`data` a data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`, typically the environment from which `regions` is called.

`newdata` an optional matrix, list or environment (or object coercible by `as.data.frame` to a data frame) containing new observations for which a prediction is desired. If missing, then prediction regions will be provided at the observed data.

`family` a character string specifying the family, must be one of `"gaussian"` (default), `"Gamma"`, or `"inverse.gaussian"`. May be abbreviated.

`link` the function which takes the conditional expectation of the response variable given predictors as its argument and has the linear regression equation as its output. If missing then the default link function in `glm` will be specified.

`alpha` the error tolerance desired for the prediction region. The default is set at 0.10.

`cores` the number of cores used to compute the conformal prediction regions. The default is set at 6 cores. Users calling this function on machines with fewer than 6 cores are encouraged to change the default.

`bins` an optional argument for specifying the desired number of bins to use along one dimension of the predictor space. If missing, the theoretical large sample optimal bin width is used (width = `O(log(n)/n)^(1/(d+1))` where `n` is the sample size and `d` is the dimension of the main effects).

`intercept` a Boolean variable corresponding to whether or not the intercept is included in the regression equation. This is only relevant for the computation of the prediction region by conformalization of residuals (when `LS = TRUE`). The default is set at TRUE.

`parametric` a Boolean variable corresponding to whether or not the parametric conformal region is to be computed. The default is set at TRUE.

`LS` a Boolean variable corresponding to whether or not the prediction region by conformalization of residualsis is to be computed. The default is set at TRUE.

`nonparametric` a Boolean variable corresponding to whether or not the nonparametric conformal region is to be computed. The default is set at TRUE.

Details

The function which computes all of the prediction regions outlined in the description. It is an internal function of the `conformalprediction` function which can be fit directly to objects of class `glm`.

Value

`regions` has functionality to return the usual prediction region that one obtains from maximum likelihood estimation and the delta method, the parametric conformal prediction region, the nonparametric conformal prediction region, and prediction regions from conformalization of residuals.

paraconformal The parametric conformal prediction region which is returned when parametric = TRUE.

nonparaconformal The nonparametric conformal prediction region which is returned when nonparametric = TRUE.

LSconformal The parametric prediction region from conformalization of residuals which is returned when LS = TRUE.

interval.plugin The usual prediction region that one obtains from maximum likelihood estimation and the delta method.

References

Eck, D.~J., Crawford, F.~W., and Aronow, P.~M. (2018+) Conformal prediction for exponential families and generalized linear models. Preprint available on request (email daniel.eck@yale.edu).

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2016) Distribution-Free Predictive Inference for Regression. https://arxiv.org/abs/1604.04173

Lei, J. and Wasserman, L. (2014) Distribution-Free Prediction Bands for Non-parametric Regression. Journal of the Royal Statistical Society: Series B, 76(1), 71-96.

Lei, J., Robins, J., and Wasserman, L. (2013) Distribution Free Prediction Sets. Journal of the American Statistical Association, 108(501), 278-287.

See Also

conformalprediction, glm

Examples

```
# example of section 2.4 in Geyer (2009)
# data(sports)
# out <- glmdr(cbind(wins, losses) ~ 0 + ., family = "binomial", data = sports)
#summary(out)
```

Index