# Package 'svydiags'

October 14, 2022

**Type** Package

**Title** Linear Regression Model Diagnostics for Survey Data

**Version** 0.4

**Date** 2022-04-27

**Author** Richard Valliant

**Maintainer** Richard Valliant <valliant@umich.edu>

**Description** Diagnostics for fixed effects linear regression models fitted with survey data. Extensions of standard diagnostics to complex survey data are included: standardized residuals, leverages, Cook's D, dfbetas, dffits, condition indexes, and variance inflation factors as found in Li and Valliant (Surv. Meth., 2009, 35(1), pp. 15-24; Jnl. of Off. Stat., 2011, 27(1), pp. 99-119; Jnl. of Off. Stat., 2015, 31(1), pp. 61-75); Liao and Valliant (Surv. Meth., 2012, 38(1), pp. 53-62; Surv. Meth., 2012, 38(2), pp. 189-202). Variance inflation factors are also computed for some general linear models (logistic and poisson) as described in Liao (U. Maryland thesis, 2010).

**Suggests** doBy, foreign, NHANES, sampling

**Depends** MASS, Matrix, survey

**License** GPL (>= 2)

**LazyLoad** yes

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2022-04-27 22:10:02 UTC

## R topics documented:

---

`nhanes2007` *National Health and Nutrition Examination Survey data, 2007-2008*

---

## Description

Demographic and dietary intake variables from a U.S. national household survey

## Usage

```
data(nhanes2007)
```

## Format

A data frame with 4,329 person-level observations on the following 26 variables measuring 24-hour dietary recall. See https://wwwn.cdc.gov/nchs/nhanes/2013-2014/DR2IFF_H.htm for more details about the variables.

`SEQN` Identification variable

`SDMVSTRA` Stratum

`SDMVPSU` Primary sampling unit, numbered within each stratum (1,2)

`WTDRD1` Dietary day 1 sample weight

`GENDER` Gender (0 = female; 1 = male)

`RIDAGEYR` Age in years at the time of the screening interview; reported for survey participants between the ages of 1 and 79 years of age. All responses of participants aged 80 years and older are coded as 80.

`RIDRETH1` Race/Hispanic origin (1 = Mexican American; 2 = Other Hispanic; 3 = Non-Hispanic White; 4 = Non-Hispanic Black; 5 = Other Race including multiracial)

`BMXWT` Body weight (kg)

`BMXBMI` Body mass Index ((weight in kg) / (height in meters)**2)

`DIET` On any diet (0 = No; 1 = Yes)

`CALDIET` On a low-calorie diet (0 = No; 1 = Yes)

`FATDIET` On a low-fat diet (0 = No; 1 = Yes)

`CARBDIET` On a low-carbohydrate diet (0 = No; 1 = Yes)

`DR1DRSTZ` Dietary recall status that indicates quality and completeness of survey participant's response to dietary recall section. (1 = Reliable and met the minimum criteria; 2 = Not reliable or not met the minimum criteria; 4 = Reported consuming breast-milk (infants and children only))

`DR1TKCAL` Energy (kcal)

`DR1TPROT` Protein (gm)

`DR1TCARB` Carbohydrate (gm)

`DR1TSUGR` Total sugars (gm)

DR1TFIBE Dietary fiber (gm)

DR1TTFAT Total fat (gm)

DR1TSFAT Total saturated fatty acids (gm)

DR1TMFAT Total monounsaturated fatty acids (gm)

DR1TPFAT Total polyunsaturated fatty acids (gm)

DR1TCAFF Caffeine (mg)

DR1TALCO Alcohol (gm)

DR1_320Z Total plain water drank yesterday (gm)

## Details

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. The `nhis2007` data set contains observations for 4,329 persons collected in 2007-2008.

## Source

National Health and Nutrition Examination Survey of 2007-2008 conducted by the U.S. National Center for Health Statistics. https://www.cdc.gov/nchs/nhanes.htm

## Examples

```
data(nhanes2007)
str(nhanes2007)
summary(nhanes2007)
```

---

`svycollinear` *Condition indexes and variance decompositions in linear models fitted with complex survey data*

---

## Description

Compute condition indexes and variance decompositions for diagnosing collinearity in fixed effects, linear regression models fitted with data collected from one- and two-stage complex survey designs.

## Usage

```
svycollinear(mod, intcpt=TRUE, w, Vcov, sc=TRUE, svyglm.obj, rnd=3, fuzz=0.3)
```

## Arguments

| | |
|---|---|
| mod | Either (i) an $n \times p$ matrix of real-valued covariates used in fitting a linear regression; $n$ = number of observations, $p$ = number of covariates in model, excluding the intercept; the matrix mod should not contain columns for the strata and cluster identifiers (unless those variables are part of the model). No missing values are allowed. Or, (ii) a model object produced by svyglm in the survey package. |
| intcpt | TRUE if the model contains an intercept; FALSE if not. |
| w | $n$-vector of survey weights used in fitting the model. No missing values are allowed. |
| Vcov | Variance-covariance matrix of the estimated slopes in the regression model; component cov.unscaled in an svyglm object. This matrix can also be extracted with vcov(m) where m is an svyglm object. |
| sc | TRUE if the columns of the weighted model matrix $\tilde{\mathbf{X}}$ (defined in Details) should be scaled for computing condition indexes; FALSE if not. If TRUE, each column of $\tilde{\mathbf{X}}$ is divided by its Euclidean norm, $\sqrt{\tilde{\mathbf{x}}^T\tilde{\mathbf{x}}}$. |
| svyglm.obj | Is mod a svyglm.obj object? TRUE or FALSE. |
| rnd | Round the output to rnd decimal places. |
| fuzz | Replace any variance decomposition proportions that are less than fuzz by '.' in the output. |

## Details

svycollinear computes condition indexes and variance decomposition proportions to use for diagnosing collinearity in a linear model fitted from complex survey data as discussed in Liao and Valliant (2012). All measures are based on $\widetilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$ where $\mathbf{W}$ is the diagonal matrix of survey weights and $\mathbf{X}$ is the $n \times p$ matrix of covariates. In a full-rank model with $p$ covariates, there are $p$ condition indexes, defined as the ratio of the maximum eigenvalue of $\widetilde{\mathbf{X}}$ to its minimum eigenvalue. Before computing condition indexes, as recommended by Belsley (1991), the columns are normalized by their individual Euclidean norms, $\sqrt{\tilde{\mathbf{x}}^T\tilde{\mathbf{x}}}$, so that each column has unit length. The columns are not centered around their means because that can obscure near-dependencies between the intercept and other covariates (Belsley 1984).

Variance decompositions are for the variance of each estimated regression coefficient and are based on a singular value decomposition of the variance formula. Proportions of the model variance, $Var_M(\hat{\beta}_k)$, associated with each column of $\widetilde{\mathbf{X}}$ are displayed in an output matrix described below.

## Value

$p \times (p+1)$ data frame, $\mathbf{\Pi}$. The first column gives the condition indexes of $\widetilde{\mathbf{X}}$. Values of 10 or more are usually considered to potentially signal collinearity of two or more columns of $\widetilde{\mathbf{X}}$. The remaining columns give the proportions (within columns) of variance of each estimated regression coefficient associated with a singular value decomposition into $p$ terms. Columns $2, \ldots, p+1$ will each approximately sum to 1. Note that some 'proportions' can be negative due to the nature of the variance decomposition. If two proportions in a given row of $\mathbf{\Pi}$ are relatively large and its associated condition index in that row in the first column of $\mathbf{\Pi}$ is also large, then near dependencies between the covariates associated with those elements are influencing the regression coefficient estimates.

## Author(s)

Richard Valliant

## References

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley-Interscience.

Belsley, D.A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, 38(2), 73-77.

Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity, and Weak Data in Regression*. New York: John Wiley & Sons, Inc.

Liao, D, and Valliant, R. (2012). Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data. *Survey Methodology*, 38, 189-202.

Lumley, T. (2010). *Complex Surveys*. New York: John Wiley & Sons.

Lumley, T. (2021). survey: analysis of complex survey samples. R package version 4.1-1.

## See Also

svyvif

## Examples

```
require(survey)
    # example from svyglm help page
data(api)
dstrat <- svydesign(id=~1,strata=~stype, weights=~pw, data=apistrat, fpc=~fpc)
m1 <- svyglm(api00~ell+meals+mobility, design=dstrat)
    # send model object from svyglm
CI.out <- svycollinear(mod = m1, w=apistrat$pw, Vcov=vcov(m1), sc=TRUE, svyglm.obj=TRUE,
        rnd=3, fuzz= 0.3)

    # send model matrix from svyglm
svycollinear(mod = m1$model, w=apistrat$pw, Vcov=vcov(m1), sc=TRUE, svyglm.obj=TRUE,
        rnd=3, fuzz=0.3)

    # use model.matrix to create matrix of covariates in model
data(nhanes2007)
newPSU <- paste(nhanes2007$SDMVSTRA, nhanes2007$SDMVPSU, sep=".")
nhanes.dsgn <- svydesign(ids = ~newPSU,
                         strata = NULL,
                         weights = ~WTDRD1, data=nhanes2007)
m1 <- svyglm(BMXWT ~ RIDAGEYR + as.factor(RIDRETH1) + DR1TKCAL +
    DR1TTFAT + DR1TMFAT + DR1TSFAT + DR1TPFAT, design=nhanes.dsgn)
X <- model.matrix(~ RIDAGEYR + as.factor(RIDRETH1) + DR1TKCAL + DR1TTFAT + DR1TMFAT
            + DR1TSFAT + DR1TPFAT,
        data = data.frame(nhanes2007))
CI.out <- svycollinear(mod = X, w=nhanes2007$WTDRD1, Vcov=vcov(m1), sc=TRUE, svyglm.obj=FALSE,
        rnd=2, fuzz=0.3)
```

svyCooksD *Modified Cook's D for models fitted with complex survey data*

## Description

Compute a modified Cook's D for fixed effects, linear regression models fitted with data collected from one- and two-stage complex survey designs.

## Usage

```
svyCooksD(mobj, stvar=NULL, clvar=NULL, doplot=FALSE)
```

## Arguments

| | |
|---|---|
| mobj | model object produced by svyglm in the survey package |
| stvar | name of the stratification variable in the svydesign object used to fit the model |
| clvar | name of the cluster variable in the svydesign object used to fit the model |
| doplot | if TRUE, plot the modified Cook's D values vs. their sequence number in data set. Reference lines are drawn at 2 and 3 |

## Details

svyCooksD computes the modified Cook's D (m-cook; see Atkinson (1982) and Li & Valliant (2011, 2015)) which measures the effect on the vector of parameter estimates of deleting single observations when fitting a fixed effects regression model to complex survey data. The function svystdres is called for some of the calculations. Values of m-cook are considered large if they are greater than 2 or 3. The R package MASS must also be loaded before calling svyCooksD. The output is a vector of the m-cook values and a scatterplot of them versus the sequence number of the sample element used in fitting the model. By default, svyglm uses only complete cases (i.e., ones for which the dependent variable and all independent variables are non-missing) to fit the model. The rows of the data frame used in fitting the model can be retrieved from the svyglm object via as.numeric(names(mobj$y)). The data for those rows is in mobj$data.

## Value

Numeric vector whose names are the rows of the data frame in the svydesign object that were used in fitting the model

## Author(s)

Richard Valliant

## References

Atkinson, A.C. (1982). Regression diagnostics, transformations and constructed variables (with discussion). *Journal of the Royal Statistical Society*, Series B, Methodological, 44, 1-36.

Cook, R.D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, 19, 15-18.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. London:Chapman & Hall Ltd.

Li, J., and Valliant, R. (2011). Linear regression diagnostics for unclustered survey data. *Journal of Official Statistics*, 27, 99-119.

Li, J., and Valliant, R. (2015). Linear regression diagnostics in cluster samples. *Journal of Official Statistics*, 31, 61-75.

Lumley, T. (2010). *Complex Surveys*. New York: John Wiley & Sons.

Lumley, T. (2021). survey: analysis of complex survey samples. R package version 4.1-1.

## See Also

svydfbetas, svydffits, svystdres

## Examples

```
require(MASS)   # to get ginv
require(survey)
data(api)
    # unstratified design single stage design
d0 <- svydesign(id=~1,strata=NULL, weights=~pw, data=apistrat)
m0 <- svyglm(api00 ~ ell + meals + mobility, design=d0)
mcook <- svyCooksD(m0, doplot=TRUE)

    # stratified clustered design
require(NHANES)
data(NHANESraw)
dnhanes <- svydesign(id=~SDMVPSU, strata=~SDMVSTRA, weights=~WTINT2YR, nest=TRUE, data=NHANESraw)
m2 <- svyglm(BPDiaAve ~ as.factor(Race1) + BMI + AlcoholYear, design = dnhanes)
mcook <- svyCooksD(mobj=m2, stvar="SDMVSTRA", clvar="SDMVPSU", doplot=TRUE)
```

---

# svydfbetas *dfbetas for models fitted with complex survey data*

---

## Description

Compute the dfbetas measure of the effect of extreme observations on parameter estimates for fixed effects, linear regression models fitted with data collected from one- and two-stage complex survey designs.

## Usage

```
svydfbetas(mobj, stvar=NULL, clvar=NULL, z=3)
```

## Arguments

| | |
|---|---|
| mobj | model object produced by svyglm in the survey package |
| stvar | name of the stratification variable in the svydesign object used to fit the model |
| clvar | name of the cluster variable in the svydesign object used to fit the model |
| z | numerator of cutoff for measuring whether an observation has an extreme effect on its own predicted value; default is 3 but can be adjusted to control how many observations are flagged for inspection |

## Details

svydfbetas computes the values of dfbetas for each observation and parameter estimate, i.e., the amount that a parameter estimate changes when the unit is deleted from the sample. The model object must be created by svyglm in the R survey package. The output is a vector of the dfbeta and standardized dfbetas values. By default, svyglm uses only complete cases (i.e., ones for which the dependent variable and all independent variables are non-missing) to fit the model. The rows of the data frame used in fitting the model can be retrieved from the svyglm object via as.numeric(names(mobj$y)). The data for those rows is in mobj$data.

## Value

List object with values:

| | |
|---|---|
| Dfbeta | Numeric vector of unstandardized dfbeta values whose names are the rows of the data frame in the svydesign object that were used in fitting the model |
| Dfbetas | Numeric vector of standardized dfbetas values whose names are the rows of the data frame in the svydesign object that were used in fitting the model |
| cutoff | Value used for gauging whether a value of dffits is large. For a single-stage sample, cutoff=$z/\sqrt{n}$; for a 2-stage sample, cutoff=$z/\sqrt{n[1+\rho(\bar{m}-1)]}$ |

## Author(s)

Richard Valliant

## References

Li, J., and Valliant, R. (2011). Linear regression diagnostics for unclustered survey data. *Journal of Official Statistics*, 27, 99-119.

Li, J., and Valliant, R. (2015). Linear regression diagnostics in cluster samples. *Journal of Official Statistics*, 31, 61-75.

Lumley, T. (2010). *Complex Surveys*. New York: John Wiley & Sons.

Lumley, T. (2021). survey: analysis of complex survey samples. R package version 4.1-1.

## See Also

svydffits, svyCooksD

## Examples

```
require(survey)
data(api)
    # unstratified design single stage design
d0 <- svydesign(id=~1,strata=NULL, weights=~pw, data=apistrat)
m0 <- svyglm(api00 ~ ell + meals + mobility, design=d0)
svydfbetas(mobj=m0)

    # stratified cluster
require(NHANES)
data(NHANESraw)
dnhanes <- svydesign(id=~SDMVPSU, strata=~SDMVSTRA, weights=~WTINT2YR, nest=TRUE, data=NHANESraw)
m2 <- svyglm(BPDiaAve ~ as.factor(Race1) + BMI + AlcoholYear, design = dnhanes)
yy <- svydfbetas(mobj=m2, stvar= "SDMVSTRA", clvar="SDMVPSU")
apply(abs(yy$Dfbetas) > yy$cutoff,1, sum)
```

---

`svydffits` *dffits for models fitted with complex survey data*

---

## Description

Compute the dffits measure of the effect of extreme observations on predicted values for fixed effects, linear regression models fitted with data collected from one- and two-stage complex survey designs.

## Usage

```
svydffits(mobj, stvar=NULL, clvar=NULL, z=3)
```

## Arguments

| | |
|---|---|
| `mobj` | model object produced by `svyglm` in the `survey` package |
| `stvar` | name of the stratification variable in the `svydesign` object used to fit the model |
| `clvar` | name of the cluster variable in the `svydesign` object used to fit the model |
| `z` | numerator of cutoff for measuring whether an observation has an extreme effect on its own predicted value; default is 3 but can be adjusted to control how many observations are flagged for inspection |

## Details

`svydffits` computes the value of dffits for each observation, i.e., the amount that a unit's predicted value changes when the unit is deleted from the sample. The model object must be created by `svyglm` in the R `survey` package. The output is a vector of the dffit and standardized dffits values. By default, `svyglm` uses only complete cases (i.e., ones for which the dependent variable and all independent variables are non-missing) to fit the model. The rows of the data frame used in fitting the model can be retrieved from the `svyglm` object via `as.numeric(names(mobj$y))`. The data for those rows is in `mobj$data`.

## Value

List object with values:

| | |
|---|---|
| `Dffit` | Numeric vector of unstandardized dffit values whose names are the rows of the data frame in the `svydesign` object that were used in fitting the model |
| `Dffits` | Numeric vector of standardized dffits values whose names are the rows of the data frame in the `svydesign` object that were used in fitting the model |
| `cutoff` | Value used for gauging whether a value of dffits is large. For a single-stage sample, `cutoff`=$z/\sqrt{n}$; for a 2-stage sample, `cutoff`=$z\sqrt{p/n\bar{m}[1+\rho(\bar{m}-1)]}$ |

## Author(s)

Richard Valliant

## References

Li, J., and Valliant, R. (2011). Linear regression diagnostics for unclustered survey data. *Journal of Official Statistics*, 27, 99-119.

Li, J., and Valliant, R. (2015). Linear regression diagnostics in cluster samples. *Journal of Official Statistics*, 31, 61-75.

Lumley, T. (2010). *Complex Surveys*. New York: John Wiley & Sons.

Lumley, T. (2021). survey: analysis of complex survey samples. R package version 4.1-1.

## See Also

svydfbetas, svyCooksD

## Examples

```
require(survey)
data(api)
    # unstratified design single stage design
d0 <- svydesign(id=~1,strata=NULL, weights=~pw, data=apistrat)
m0 <- svyglm(api00 ~ ell + meals + mobility, design=d0)
yy <- svydffits(mobj=m0)
yy$cutoff
sum(abs(yy$Dffits) > yy$cutoff)

require(NHANES)
data(NHANESraw)
dnhanes <- svydesign(id=~SDMVPSU, strata=~SDMVSTRA, weights=~WTINT2YR, nest=TRUE, data=NHANESraw)
m2 <- svyglm(BPDiaAve ~ as.factor(Race1) + BMI + AlcoholYear, design = dnhanes)
yy <- svydffits(mobj=m2, stvar= "SDMVSTRA", clvar="SDMVPSU", z=4)
sum(abs(yy$Dffits) > yy$cutoff)
```

svyhat *Leverages for models fitted with complex survey data*

## Description

Compute leverages for fixed effects, linear regression models fitted from complex survey data.

## Usage

```
svyhat(mobj, doplot=FALSE)
```

## Arguments

| | |
|---|---|
| mobj | model object produced by svyglm in the survey package |
| doplot | if TRUE, plot the standardized residuals vs. their sequence number in data set. A reference line is drawn at 3 times the mean leverage |

## Details

svyhat computes the leverages from a model fitted with complex survey data. The model object mobj must be created by svyglm in the R survey package. The output is a vector of the leverages and a scatterplot of them versus the sequence number of the sample element used in fitting the model. By default, svyglm uses only complete cases (i.e., ones for which the dependent variable and all independent variables are non-missing) to fit the model. The rows of the data frame used in fitting the model can be retrieved from the svyglm object via as.numeric(names(mobj$y)). The data for those rows is in mobj$data.

## Value

Numeric vector whose names are the rows of the data frame in the svydesign object that were used in fitting the model.

## Author(s)

Richard Valliant

## References

Belsley, D.A., Kuh, E. and Welsch, R. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons, Inc.

Li, J., and Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, 35, 15-24.

Lumley, T. (2010). *Complex Surveys*. New York: John Wiley & Sons.

Lumley, T. (2021). survey: analysis of complex survey samples. R package version 4.1-1.

## See Also

svystdres

## Examples

```
require(survey)
data(api)
dstrat <- svydesign(id=~1,strata=~stype, weights=~pw, data=apistrat)
m1 <- svyglm(api00 ~ ell + meals + mobility, design=dstrat)
h <- svyhat(mobj = m1, doplot=TRUE)
100*sum(h > 3*mean(h))/length(h)    # percentage of leverages > 3*mean

require(NHANES)
data(NHANESraw)
dnhanes <- svydesign(id=~SDMVPSU, strata=~SDMVSTRA, weights=~WTINT2YR, nest=TRUE, data=NHANESraw)
m1 <- svyglm(BPDiaAve ~ as.factor(Race1) + BMI + AlcoholYear, design = dnhanes)
h <- svyhat(mobj = m1, doplot=TRUE)
```

---

# svystdres *Standardized residuals for models fitted with complex survey data*

---

## Description

Compute standardized residuals for fixed effects, linear regression models fitted with data collected from one- and two-stage complex survey designs.

## Usage

```
svystdres(mobj, stvar=NULL, clvar=NULL, doplot=FALSE)
```

## Arguments

| | |
|---|---|
| mobj | model object produced by svyglm in the survey package |
| stvar | name of the stratification variable in the svydesign object used to fit the model |
| clvar | name of the cluster variable in the svydesign object used to fit the model |
| doplot | if TRUE, plot the standardized residuals vs. their sequence number in data set. Reference lines are drawn at +/-3 |

## Details

svystdres computes the standardized residuals, i.e., the residuals divided by an estimate of the model standard deviation of the residuals. Residuals are used from a model object created by svyglm in the R survey package. The output is a vector of the standardized residuals and a scatterplot of them versus the sequence number of the sample element used in fitting the model. By default, svyglm uses only complete cases (i.e., ones for which the dependent variable and all independent variables are non-missing) to fit the model. The rows of the data frame used in fitting the model can be retrieved from the svyglm object via as.numeric(names(mobj$y)). The data for those rows is in mobj$data.

## Value

List object with values:

| | |
|---|---|
| `stdresids` | Numeric vector whose names are the rows of the data frame in the `svydesign` object that were used in fitting the model |
| `n` | number of sample clusters |
| `mbar` | average number of non-missing, sample elements per cluster |
| `rtsighat` | estimate of the square root of the model variance of the residuals, $\sqrt{(\sigma^2)}$ |
| `rhohat` | estimate of the intracluster correlation of the residuals, $\rho$ |

## Author(s)

Richard Valliant

## References

Li, J., and Valliant, R. (2011). Linear regression diagnostics for unclustered survey data. *Journal of Official Statistics*, 27, 99-119.

Li, J., and Valliant, R. (2015). Linear regression diagnostics in cluster samples. *Journal of Official Statistics*, 31, 61-75.

Lumley, T. (2010). *Complex Surveys*. New York: John Wiley & Sons.

Lumley, T. (2021). survey: analysis of complex survey samples. R package version 4.1-1.

## See Also

svyhat, svyCooksD

## Examples

```
require(survey)
data(api)
    # unstratified design single stage design
d0 <- svydesign(id=~1,strata=NULL, weights=~pw, data=apistrat)
m0 <- svyglm(api00 ~ ell + meals + mobility, design=d0)
svystdres(mobj=m0, stvar=NULL, clvar=NULL)

    # stratified cluster design
require(NHANES)
data(NHANESraw)
dnhanes <- svydesign(id=~SDMVPSU, strata=~SDMVSTRA, weights=~WTINT2YR, nest=TRUE, data=NHANESraw)
m1 <- svyglm(BPDiaAve ~ as.factor(Race1) + BMI + AlcoholYear, design = dnhanes)
svystdres(mobj=m1, stvar= "SDMVSTRA", clvar="SDMVPSU")
```

svyvif *Variance inflation factors (VIF) for general linear models fitted with complex survey data*

## Description

Compute a VIF for fixed effects, general linear regression models fitted with data collected from one- and two-stage complex survey designs.

## Usage

```
svyvif(mobj, X, w, stvar=NULL, clvar=NULL)
```

## Arguments

| | |
|---|---|
| mobj | model object produced by svyglm. The following families of models are allowed: binomial, gaussian, poisson, quasibinomial, and quasipoisson. Other families allowed by svyglm will produce an error in svyvif. |
| X | $n \times p$ matrix of real-valued covariates used in fitting a linear regression; $n$ = number of observations, $p$ = number of covariates in model, excluding the intercept. A column of 1's for an intercept should not be included. X should not contain columns for the strata and cluster identifiers (unless those variables are part of the model). No missing values are allowed. |
| w | $n$-vector of survey weights used in fitting the model. No missing values are allowed. |
| stvar | field in mobj that contains the stratum variable in the complex sample design; use stvar = NULL if there are no strata |
| clvar | field in mobj that contains the cluster variable in the complex sample design; use clvar = NULL if there are no clusters |

## Details

svyvif computes a variance inflation factor (VIF) appropriate for linear models and some general linear models (GLMs) fitted from complex survey data (see Liao & Valliant 2012). A VIF measures the inflation of a slope estimate caused by nonorthogonality of the predictors over and above what the variance would be with orthogonality (Theil 1971; Belsley, Kuh, and Welsch 1980). The standard VIF equals $1/(1 - R_k^2)$ where $R_k$ is the multiple correlation of the $k^{th}$ column of X regressed on the remaining columns. The complex sample value of the VIF for a linear model consists of the standard VIF multiplied by two adjustments denoted in the output as zeta and varrho. The VIF for a GLM is similar (Liao 2010, chap. 5). There is no widely agreed-upon cutoff value for identifying high values of a VIF, although 10 is a common suggestion.

## Value

$p \times 5$ matrix with columns:

| | |
|---|---|
| svy.vif | complex sample VIF |

| | |
|---|---|
| `reg.vif` | standard VIF, $1/(1-R_k^2)$, that omits the factors, `zeta` and `varrho`; $R_k^2$ is an R-square from a weighted least squares regression of the $k^{th}$ x on the other x's in the regression |
| `zeta` | 1st multiplicative adjustment to `reg.vif` |
| `varrho` | 2nd multiplicative adjustment to `reg.vif` |
| `zeta.x.varrho` | product of the two adjustments to `reg.vif` |
| `R.square` | R-square in the regression of the $k^{th}$ x on the other x's, including the intercept |

## Author(s)

Richard Valliant

## References


Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley-Interscience.

Liao, D. (2010). Collinearity Diagnostics for Complex Survey Data. PhD thesis, University of Maryland. http://hdl.handle.net/1903/10881.

Liao, D, and Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, 38, 53-62.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons, Inc.

Lumley, T. (2010). *Complex Surveys*. New York: John Wiley & Sons.

Lumley, T. (2018). survey: analysis of complex survey samples. R package version 3.34.


## See Also

`Vmat`

## Examples

```
require(survey)
data(nhanes2007)
X1 <- nhanes2007[order(nhanes2007$SDMVSTRA, nhanes2007$SDMVPSU),]
    # eliminate cases with missing values
delete <- which(complete.cases(X1)==FALSE)
X2 <- X1[-delete,]
nhanes.dsgn <- svydesign(ids = ~SDMVPSU,
                         strata = ~SDMVSTRA,
                         weights = ~WTDRD1, nest=TRUE, data=X2)
    # linear model
m1 <- svyglm(BMXWT ~ RIDAGEYR + as.factor(RIDRETH1) + DR1TKCAL
            + DR1TTFAT + DR1TMFAT, design=nhanes.dsgn)
summary(m1)
    # construct X matrix using model.matrix from stats package
X3 <- model.matrix(~ RIDAGEYR + as.factor(RIDRETH1) + DR1TKCAL + DR1TTFAT + DR1TMFAT,
        data = data.frame(X2))
    # remove col of 1's for intercept with X3[,-1]
svyvif(mobj=m1, X=X3[,-1], w = X2$WTDRD1, stvar=NULL, clvar=NULL)
```

```
    # Logistic model
X2$obese <- X2$BMXBMI >= 30
nhanes.dsgn <- svydesign(ids = ~SDMVPSU,
                         strata = ~SDMVSTRA,
                         weights = ~WTDRD1, nest=TRUE, data=X2)
m2 <- svyglm(obese ~ RIDAGEYR + as.factor(RIDRETH1) + DR1TKCAL
             + DR1TTFAT + DR1TMFAT, design=nhanes.dsgn, family="quasibinomial")
summary(m2)
svyvif(mobj=m2, X=X3[,-1], w = X2$WTDRD1, stvar = "SDMVSTRA", clvar = "SDMVPSU")
```

---

`Vmat` *Compute covariance matrix of residuals for general linear models fitted with complex survey data*

---

## Description

Compute a covariance matrix using residuals from a fixed effects, general linear regression model fitted with data collected from one- and two-stage complex survey designs.

## Usage

```
Vmat(mobj, stvar = NULL, clvar = NULL)
```

## Arguments

| | |
|---|---|
| `mobj` | model object produced by `svyglm` |
| `stvar` | field in `mobj` that contains the stratum variable in the complex sample design; use `stvar = NULL` if there are no strata |
| `clvar` | field in `mobj` that contains the cluster variable in the complex sample design; use `clvar = NULL` if there are no clusters |

## Details

`Vmat` computes a covariance matrix among the residuals returned from `svyglm` in the `survey` package. `Vmat` is called by `svyvif` when computing variance inflation factors. The matrix that is computed by `Vmat` is appropriate under these model assumptions: (1) in single-stage, unclustered sampling, units are assumed to be uncorrelated but can have different model variances, (2) in single-stage, stratified sampling, units are assumed to be uncorrelated within strata and between strata but can have different model variances; (3) in unstratified, clustered samples, units in different clusters are assumed to be uncorrelated but units within clusters are correlated; (3) in stratified, clustered samples, units in different strata or clusters are assumed to be uncorrelated but units within clusters are correlated.

## Value

$n \times n$ matrix where $n$ is the number of cases used in the linear regression model

## Author(s)

Richard Valliant

## References

Liao, D, and Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, 38, 53-62.

Lumley, T. (2010). *Complex Surveys*. New York: John Wiley & Sons.

Lumley, T. (2021). survey: analysis of complex survey samples. R package version 4.1-1.

## See Also

svyvif

## Examples

```
require(Matrix)
require(survey)
data(nhanes2007)
black <- nhanes2007$RIDRETH1 == 4
X <- nhanes2007
X <-  cbind(X, black)
X1 <- X[order(X$SDMVSTRA, X$SDMVPSU),]

    # unstratified, unclustered design
nhanes.dsgn <- svydesign(ids = 1:nrow(X1),
                         strata = NULL,
                         weights = ~WTDRD1, data=X1)
m1 <- svyglm(BMXWT ~ RIDAGEYR + as.factor(black) + DR1TKCAL, design=nhanes.dsgn)
summary(m1)

V <- Vmat(mobj = m1,
          stvar = NULL,
          clvar = NULL)

    # stratified, clustered design
nhanes.dsgn <- svydesign(ids = ~SDMVPSU,
                         strata = ~SDMVSTRA,
                         weights = ~WTDRD1, nest=TRUE, data=X1)
m1 <- svyglm(BMXWT ~ RIDAGEYR + as.factor(black) + DR1TKCAL, design=nhanes.dsgn)
summary(m1)
V <- Vmat(mobj = m1,
          stvar = "SDMVSTRA",
          clvar = "SDMVPSU")
```

# Index