Package 'svydiags'

June 4, 2015

Type Package

Title Linear Regression Model Diagnostics for Survey Data

Version 0.1

Date 2015-01-21

Author Richard Valliant

Maintainer Richard Valliant <rvallian@umd.edu>

Description svydiags contains functions for computing 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.

Suggests doBy, foreign, NHANES, sampling

Depends MASS, survey

License GPL (>= 2)

LazyLoad yes

R topics documented:

`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. (2014). survey: analysis of complex survey samples. R package version 3.30.

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. (2014). survey: analysis of complex survey samples. R package version 3.30.

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. (2014). survey: analysis of complex survey samples. R package version 3.30.

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. (2014). survey: analysis of complex survey samples. R package version 3.30.

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, ρ

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. (2014). survey: analysis of complex survey samples. R package version 3.30.

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")
```

Index