Package 'hcc'

February 20, 2015

Version 0.54

Date 2013-03-23

Title Hidden correlation check

Author Yun Shi and A. I. McLeod

Maintainer A. I. McLeod <aimcleod@uwo.ca>

Depends R (>= 2.1.0)

Description A new diagnostic check for model adequacy in regression and generalized linear models is implemented.

LazyLoad yes

License GPL (>= 2)

URL http://www.stats.uwo.ca/faculty/aim

NeedsCompilation no

Repository CRAN

Date/Publication 2013-03-23 17:01:19

R topics documented:

---

hcc-package *Hidden correlation check*

Description

A new diagnostic check for model adequacy in regression and generalized linear models is implemented.

Details

Package:	hcc
Type:	Package
Version:	0.54
Date:	2013-03-23
License:	GPL (>=2)

This package provides a new diagnostic test that will be discussed in a future paper.

Author(s)

Yun Shi and A.I. McLeod Maintainer: A. I. McLeod <aimcleod@uwo.ca>

Examples

```
#Example 1
#an example with hidden correlation
set.seed(313477)
data <- simer(50, 5)
ans <- lm(y~x, data=data)
summary(ans)
#the usual regression plots
par(mfrow=c(2,2))
plot(ans)
par(mfrow=c(1,1))
#hidden correlation signficance test
res <- resid(ans)
hctest(data$x, res)
#Poincare plot
PoincarePlot(data$x, res)
#residual dependency test
rdplot(data$x, res)

#Example 2
data(birthwt)
ans<-glm(low~., data=birthwt[,-10], family="binomial")
ans2<-step(ans)
```

```
#only lwt (mother's weight) is a numeric variable
res <- resid(ans2)
hctest(birthwt$lwt, res)
#the test for hidden correlation is signficant
PoincarePlot(birthwt$lwt, res)
#the Poincare plot confirms the lack of fit
rdplot(birthwt$lwt, res)
#the residual-dependency plot does not give a clear signal
```

---

beams *The strength of ten wood beams are effected by the specific gravity and moisture content*

Description

The data was collected on the specific gravity, moisture content and strength of ten wood beams.

Usage

```
data(beams)
```

Format

A data frame with 10 observations on the following 3 variables.

`y` the strength of ten wood beams

`x1` the specific gravity

`x2` the moisture content

Source

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1993). A Handbook of Small Datasets. Chapman and Hall.

References

Draper, N.R. and Stoneman, D.M. (1966) Testing for the inclusion of variables in linear regression by a randomisation technique. Technometrics, 8, 695-699.

Examples

```
data(beams)
mod <-lm(y ~ x1+x2, data=beams)
x <- beams[, 2]
res <- resid(mod)
hctest(x, res)
```

`birthwt` *Risk Factors Associated with Low Infant Birth Weight*

Description

The birthwt data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

Usage

```
data(birthwt)
```

Format

A data frame with 189 observations on the following 10 variables.

`low` indicator of birth weight less than 2.5 kg.

`age` mother's age in years.

`lwt` mother's weight in pounds at last menstrual period.

`race` mother's race (1 = white, 2 = black, 3 = other).

`smoke` smoking status during pregnancy.

`ptl` number of previous premature labours.

`ht` history of hypertension.

`ui` presence of uterine irritability.

`ftv` number of physician visits during the first trimester.

`bwt` birth weight in grams.

Source

Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression. New York: Wiley

References

Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

Examples

```
data(birthwt)
attach(birthwt)
race <- factor(race, labels=c("white","black","other"))
ptd <- factor(ptl >0)
ftv <- factor(ftv)
levels(ftv)[-(1:2)] <- "2+"
bwt <- data.frame(low=factor(low), age, lwt, race, smoke=(smoke>0), ptd, ht=(ht>0), ui=(ui>0), ftv)
birthwt.glm <- glm(low ~ ., family=binomial, data=bwt )
res<-resid(birthwt.glm)
hctest(age, res)
```

cats *Anatomical data from domestic cats*

Description

The heart and body weights of samples of male and female cats used for digitalis experiments. The cats were all adult, over 2 kg body weight.

Usage

```
data(cats)
```

Format

A data frame with 144 observations on the following 3 variables.

`Sex` sex:Factor with evels "F" and "M".

`Bwt` body weight in kg.

`Hwt` heart weight in g.

References

R. A. Fisher (1947) The analysis of covariance method for the relation between a part and the whole, Biometrics 3, 65-68.

Examples

```
data(cats)
attach(cats)
mod<-lm(Hwt~Sex+Bwt+Sex:Bwt,data=cats)
res <- resid(mod)
hctest(Bwt, res)
```

dicentric *Radiation dose effects on chromosomal abnormality*

Description

An experiment was conducted to determine the effect of gamma radiation on the numbers of chromosomal abnormalities observed

Usage

```
data(dicentric)
```

Format

A data frame with 27 observations on the following 4 variables.

`cells` Number of cells in hundreds

`ca` Number of chromosomal abnormalities

`doseamt` amount of dose in Grays

`doserate` rate of dose in Grays/hour

Source

Purott R. and Reeder E. (1976) The effect of changes in dose rate on the yield of chromosome aberrations in human lymphocytes exposed to gamma radiation. Mutation Research. 35, 437-444.

References

Frome E. and DuFrain R. (1986) Maximum Likelihood Estimation for Cytogenic Dose-Response Curves. Biometrics. 42, 73-84 and Extending the linear model with R. Chapman & Hall/CRC Taylor & Francis Group, 2006.

Examples

```
data(dicentric)
dicentric$dosef <- factor(dicentric$doseamt)
rmod <- glm(ca ~ offset(log(cells))+log(doserate)*dosef, family=poisson,dicentric)
x <- dicentric[,4]
res <- resid(rmod)
hctest(x, res)
```

---

gala *Species diversity on the Galapagos Islands*

Description

There are 30 Galapagos islands and 7 variables in the dataset. The relationship between the number of plant species and several geographic variables is of interest. The original dataset contained several missing values which have been filled for convenience.

Usage

```
data(gala)
```

Format

A data frame with 30 observations on the following 7 variables.

`Species` the number of plant species found on the island

`Endemics` the number of endemic species

`Area` the area of the island (km^2)

`Elevation` the highest elevation of the island (m)

`Nearest` the distance from the nearest island (km)

`Scruz` the distance from Santa Cruz island (km)

`Adjacent` the area of the adjacent island (square km)

Source

M. P. Johnson and P. H. Raven (1973) "Species number and endemism: The Galapagos Archipelago revisited" Science, 179, 893-895

References

J. J. Faraway. Linear Models with R. Chapman & Hall/CRC, 2005 and J. J.Faraway. Extending the linear model with R. Chapman & Hall/CRC Taylor & Francis Group, 2006.

Examples

```
data(gala)
gala <- gala[,-2]
modt <- lm(sqrt(Species) ~ . , gala)
res <- resid(modt)
hctest(gala[,3], residuals(modt))
```

---

hctest *Test for hidden correlation with one input*

Description

Statistical signficance test for hidden correlation given an input x and residuals.

Usage

```
hctest(x, res)
```

Arguments

`x`	the predictor variable
`res`	residuals, the same length as x

Value

p-value

Author(s)

Yun Shi and A.I. McLeod

Examples

```
#Example 1
#in this example, there is no hidden correlation
set.seed(313477)
n <- 50
err <- rnorm(n)
x <- rnorm(n)
y <- 1+2*x+err
res <- resid(lm(y~x))
hctest(x, res)
```

---

ozone *Ozone readings in LA*

Description

Example Dataset from Practical Regression and Anova

Usage

```
data(ozone)
```

Format

A data frame with 330 observations on the following 10 variables.

`O3` the molecular formula for ozone

`vh` the molecular formula for ozone

`wind` the flow of gases on a large scale

`humidity` the amount of water vapor in the air

`temp` the physical quantity that is a measure of hotness and coldness on a numerical scale

`ibh` the molecular formula for ozone

`dpg` the molecular formula for ozone

`ibt` the molecular formula for ozone

`vis` the molecular formula for ozone

`doy` the molecular formula for ozone

References

J. J.Faraway. Extending the linear model with R. Chapman & Hall/CRC Taylor & Francis Group, 2006.

Examples

```
data(ozone)
alm <- lm(O3 ~ vis+doy+ibt+humidity+temp,data=ozone)
res <- resid(alm)
hctest(ozone[,10], res)
```

`PoincarePlot` *Poincare plot*

Description

Scatter plot check for hidden correlation given an input x and residuals.

Usage

```
PoincarePlot(x, res)
```

Arguments

`x`	an input variable in the regression
`res`	residuals, the same length as x

Details

Plot the ordered lagged one residuals along with a loess smooth to help visualize whether there is a correlation in the residuals.

Value

plot produced

Author(s)

Yun Shi and A. I. McLeod

See Also

`hctest`

Examples

```
data(trees)
ans<-lm(Volume~Girth+Height, data=trees)
x <- trees$Girth
res <- resid(ans)
PoincarePlot(x, res)
```

---

rdplot *Residual dependency plot*

Description

Plots the residuals vs an input variable. The loess smoother is shown.

Usage

```
rdplot(x, res, f = 0.8)
```

Arguments

x	input variable
res	residuals
f	smoothing parameter for loess

Value

plot produced as a side-effect

Author(s)

A.I. McLeod

References

W.S. Cleveland, Visualizing Data.

Examples

```
x <- runif(50, 0, 50)
res <- rnorm(50)
rdplot(x, res)
```

rubber *Abrasion loss for various hardness and tensile strength*

Description

The data come from an experiment to investigate how the resistance of rubber to abrasion is affected by the hardness of the rubber and its tensile strength.

Usage

```
data(rubber)
```

Format

A data frame with 30 observations on the following 3 variables.

`hardness` hardness in degree Shore

`tensile.strength` tensile strength in kg per square meter

`abrasion.loss` abrasion loss in gram per hour

`ts.low` tensile strength minus the breakpoint 180 km/m^2

`ts.high` tensile strength minus the breakpoint 180 km/m^2

Source

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1993). A Handbook of Small Datasets. Chapman and Hall.

References

Cleveland, W. S. (1993). Visualizing data. Hobart Press, Summit: New Jersey. Davies, O.L. and Goldsmith, P.L.(1972) Statistical methods in Research and Production.

Examples

```
data(rubber)
rmod <- lm(abrasion.loss~hardness+tensile.strength, data=rubber)
res <- resid(rmod)
hctest(rubber[,1], res)
```

simer *Simulation of simple linear regression with hidden correlation*

Description

Simulates n observations, (y, x) from a simple linear regression, y = a+b*x+e, where x is uniformly distributed on (0, n), e are normally distributed with mean 0 and variance 1. The parameeters a and b are zero. The error term is not independent but has an exponential correlation with parameter r so that observations close together in the x-space are positively correlated. When r=0, the correlation is zero but as r increases the correlation gets stronger.

Usage

```
simer(n, r)
```

Arguments

n	number of observations
r	correlation parameter

Value

A data frame with components:

x	input variable
y	output variable

Author(s)

A. I. McLeod and Yun Shi

Examples

```
data <- simer(50, 5)
```

tensile *The tensile strength of Kraft paper measure against the percentage of hardwood*

Description

The tensile strength of Kraft paper was measured against the percentage of hardwood in the batch of pulp from which the paper was produce.

Usage

```
data(tensile)
```

Format

A data frame with 19 observations on the following 2 variables.

`Y` the tensile strength of Kraft paper

`x` hardwood in the batch of pulp

Source

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1993). A Handbook of Small Datasets. Chapman and Hall.

References

Joglekar, G., Schuenemeyer, J.H. and LaRiccia, V. (1989) Lack-of-fit testing when replicates are not available. American Statistician, 43, 135-143.

Examples

```
data(tensile)
tmod1 <- lm(Y~x+I(x^2), tensile)
x<-tensile[,2]
res <- resid(tmod1)
hctest(x, res)
```

---

ustemp *U.S. winter temperatures for various latitudes and longitudes*

Description

The data collect from 56 U.S. cities winter temperatures for various latitudes and longitudes from 1931 to 1960.

Usage

```
data(ustemp)
```

Format

A data frame with 56 observations on the following 3 variables.

`y` winter temperature (deg F)

`x1` latitude

`x2` longitude

Source

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1993). A Handbook of Small Datasets. Chapman and Hall.

References

J. L. Peixoto. A property of well-formulated polynomial regression models. The American Statistician, 44:26, 1990.

Examples

```
data(ustemp)
lmod<-lm(y~x1+x2, data=ustemp)
x1<-ustemp[,"x1"]
x2<-ustemp[,"x2"]
res<-resid(lmod)
hctest(x1, res)
```

---

windmill *Direct current output was measured against wind velocity*

Description

Data collect from direct current output and wind velocity.

Usage

```
data(windmill)
```

Format

A data frame with 25 observations on the following 2 variables.

`Y` direct current output

`x` wind velocity

Source

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1993). A Handbook of Small Datasets. Chapman and Hall.

References

Joglekar, G., Schuenemeyer, J.H. and LaRiccia, V. (1989) Lack-of-fit testing when replicates are not available. American Statistician, 43, 135-143.

Examples

```
data(windmill)
g1<-lm(Y~x,data=windmill)
res<- resid(g1)
x<- windmill[,2]
hctest(x,res)
```

Index