1 Regression Deletion Diagnostics

This suite of functions can be used to compute some of the regression (leave-one-out deletion) diagnostics for linear and generalized linear models discussed in Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), etc.

Details

- The primary high-level function is `influence.measures` which produces a class "infl" object tabular display showing the DFBETAS for each model variable, DFFITS, covariance ratios, Cook's distances and the diagonal elements of the hat matrix. Cases which are influential with respect to any of these measures are marked with an asterisk.
- The functions `dfbetas`, `dffits`, `covratio` and `cooks.distance` provide direct access to the corresponding diagnostic quantities.
- Functions `rstandard` and `rstudent` give the standardized and Studentized residuals respectively.
- (These functions re-normalize the residuals to have unit variance, using an overall and leave-one-out measure of the error variance respectively.)
- Values for generalized linear models are approximations, as described in Williams (1987) (except that Cook's distances are scaled as F rather than as chi-square values). The approximations can be poor when some cases have large influence.
- The optional `infl`, `res` and `sd` arguments are there to encourage the use of these direct access functions, in situations where, e.g., the underlying basic influence measures (from `lm.influence` or the generic influence) are already available.
- Note that cases with `weights == 0` are dropped from all these functions, but that if a linear model has been fitted with `na.action = na.exclude`,

suitable values are filled in for the cases excluded during fitting.

- The function `hat()` exists mainly for S (version 2) compatibility; we recommend using `hatvalues()` instead.

```
Usage
influence.measures(model)

rstandard(model, ...)

## S3 method for class ’lm’
rstandard(model, infl = lm.influence(model, do.coef = FALSE),
sd = sqrt(deviance(model)/df.residual(model)), ...)

## S3 method for class ’glm’
rstandard(model, infl = influence(model, do.coef = FALSE),
type = c("deviance", "pearson"), ...)
```

```
rstudent(model, ...)

## S3 method for class ’lm’
rstudent(model, infl = lm.influence(model, do.coef = FALSE),
res = infl$wt.res, ...)

## S3 method for class ’glm’
rstudent(model, infl = influence(model, do.coef = FALSE), ...)

dffits(model, infl = , res = )
```

2 Standardized and Studentized Residuals

The standardized residual is the residual divided by its standard deviation.

Plot the standardized residual of the simple linear regression model of the data set faithful against the independent variable waiting.

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm. Then we compute the standardized residual with the rstandard function.

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
> eruption.stdres = rstandard(eruption.lm)
```

We now plot the standardized residual against the observed values of the variable waiting.

```
> plot(faithful$waiting, eruption.stdres,
+     ylab="Standardized Residuals",
+     xlab="Waiting Time",
+     main="Old Faithful Eruptions")
> abline(0, 0)                  # the horizon
```

3 Leverage and Influence

3.1 Influence

The influence of an observation can be thought of in terms of how much the predicted scores for other observations would differ if the observation in question were not included.

3.2 Influential Points

An influential point is an outlier that greatly affects the slope of the regression line. One way to test the influence of an outlier is to compute the regression equation with and without the outlier.

This type of analysis is illustrated below. The scatter plots are identical, except that the plot on the right includes an outlier. The slope is flatter when the outlier is present (-3.32 vs. -4.10), so this outlier would be considered an influential point.

3.3 Without Outlier

- Regression equation: $\hat{y} = 104.78 - 4.10x$
- Coefficient of determination: $R^2 = 0.94$
- Regression equation: $\hat{y} = 97.51 - 3.32x$
- Coefficient of determination: $R^2 = 0.55$

The charts below compare regression statistics for another data set with and without an outlier. Here, the chart on the right has a single outlier, located at the high end of the X axis (where x = 24). As a result of that single outlier, the slope of the regression line changes greatly, from -2.5 to -1.6; so the outlier would be considered an influential point.

Sometimes, an influential point will cause the coefficient of determination to be bigger; sometimes, smaller. In the first example above, the coefficient of determination is smaller when the influential point is present (0.94 vs. 0.55). In the second example, it is bigger (0.46 vs. 0.52).

If your data set includes an influential point, here are some things to consider.

- An influential point may represent bad data, possibly the result of measurement error. If possible, check the validity of the data point.

- Compare the decisions that would be made based on regression equations defined with and without the influential point. If the equations lead to contrary decisions, use caution.

3.4 Summary of Influence Statistics

3.5 rstudent

The studentized residual RSTUDENT is estimated by $s(i)^2$ without the ith observation, not by s^2. For example,

$$\text{RSTUDENT} = \frac{r_i}{s_{(i)}\sqrt{(1-h_i)}}$$

Observations with RSTUDENT larger than 2 in absolute value may need some attention.

- **Studentized Residuals** Residuals divided by their estimated standard errors (like t-statistics). Observations with values larger than 3 in absolute value are considered outliers.
- **Leverage Values (Hat Diag)** Measure of how far an observation is from the others in terms of the levels of the independent variables (not the dependent variable). Observations with values larger than $2(k+1)/n$ are considered to be potentially highly influential, where k is the number of predictors and n is the sample size.
- **DFFITS** Measure of how much an observation has effected its fitted value from the regression model. Values larger than $2\sqrt{(k+1)/n}$ in absolute value are considered highly influential.
- **DFBETAS** Measure of how much an observation has effected the estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). Values larger than 2/sqrt(n) in absolute value are considered highly influential.

The measure that measures how much impact each observation has on a particular predictor is DFBETAs The DFBETA for a predictor and for a particular observation is the difference between the regression coefficient calculated for all of the data and the regression coefficient calculated with the observation deleted, scaled by the standard error calculated with the observation deleted.

- **Cooks D** Measure of aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. Values larger than 4/n are considered highly influential.

4 Other Measures of Influence

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. The measure DFBETA is the studentized value of this difference.

Influence arises at two stages of the LME model. Firstly when V is estimated by $\hat{V}$, and subsequent estimations of the fixed and random regression coefficients β and u, given $\hat{V}$.

4.0.1 DFBETA

A group of measures that measures how much impact each observation has on a particular predictor are the DFBETAs. The DFBETA for a predictor and for a particular observation is the difference between the regression coefficient calculated for all of the data and the regression coefficient calculated with the observation deleted, scaled by the standard error calculated with the observation deleted.

$$\begin{aligned} DFBETA_a &= \hat{\beta} - \hat{\beta}_{(a)} & (1) \\ &= B(Y - Y_{\bar{a}} & (2) \end{aligned}$$

For k predictors in the mode, there ar $k+1$ dfbetas.

4.0.2 DFFITS

DFFITS is a statistical measured designed to a show how influential an observation is in a statistical model. It is closely related to the studentized residual.

$$DFFITS = \frac{\widehat{y_i} - \widehat{y_{i(k)}}}{s_{(k)}\sqrt{h_{ii}}}$$

4.0.3 COVRATIO

The COVRATIO statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the ith observation:

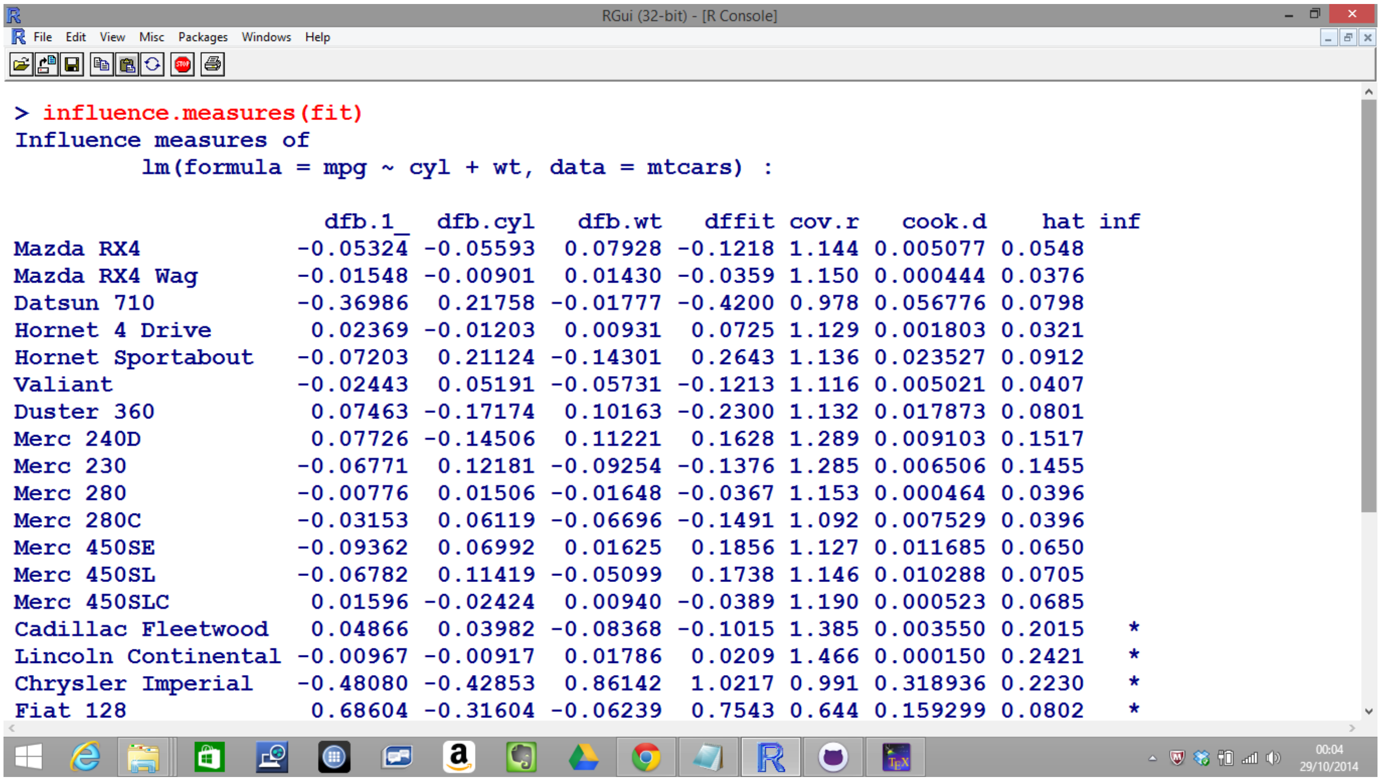
$$COVRATIO = \frac{det(s^2_{(i)}(X'_{(i)}X_{(i)})^{-1}))}{det(s^2(X'X)^{-1})}$$

Observations with

$$|COVRATIO - 1| \geq \frac{3k}{n}$$

where k is the number of parameters in the model and n is the number of observations used to fit the model, are worth further investigation.

4.1 Influential Observations : DFBeta and DFBetas



```
> influence.measures(fit)
Influence measures of
	 lm(formula = mpg ~ cyl + wt, data = mtcars) :

                      dfb.1_  dfb.cyl   dfb.wt   dffit cov.r   cook.d    hat inf
Mazda RX4           -0.05324 -0.05593  0.07928 -0.1218 1.144 0.005077 0.0548
Mazda RX4 Wag       -0.01548 -0.00901  0.01430 -0.0359 1.150 0.000444 0.0376
Datsun 710          -0.36986  0.21758 -0.01777 -0.4200 0.978 0.056776 0.0798
Hornet 4 Drive       0.02369 -0.01203  0.00931  0.0725 1.129 0.001803 0.0321
Hornet Sportabout   -0.07203  0.21124 -0.14301  0.2643 1.136 0.023527 0.0912
Valiant             -0.02443  0.05191 -0.05731 -0.1213 1.116 0.005021 0.0407
Duster 360           0.07463 -0.17174  0.10163 -0.2300 1.132 0.017873 0.0801
Merc 240D            0.07726 -0.14506  0.11221  0.1628 1.289 0.009103 0.1517
Merc 230            -0.06771  0.12181 -0.09254 -0.1376 1.285 0.006506 0.1455
Merc 280            -0.00776  0.01506 -0.01648 -0.0367 1.153 0.000464 0.0396
Merc 280C           -0.03153  0.06119 -0.06696 -0.1491 1.092 0.007529 0.0396
Merc 450SE          -0.09362  0.06992  0.01625  0.1856 1.127 0.011685 0.0650
Merc 450SL          -0.06782  0.11419 -0.05099  0.1738 1.146 0.010288 0.0705
Merc 450SLC          0.01596 -0.02424  0.00940 -0.0389 1.190 0.000523 0.0685
Cadillac Fleetwood   0.04866  0.03982 -0.08368 -0.1015 1.385 0.003550 0.2015   *
Lincoln Continental -0.00967 -0.00917  0.01786  0.0209 1.466 0.000150 0.2421   *
Chrysler Imperial   -0.48080 -0.42853  0.86142  1.0217 0.991 0.318936 0.2230   *
Fiat 128             0.68604 -0.31604 -0.06239  0.7543 0.644 0.159299 0.0802   *
```

Figure 1:

5 Influential Points in Regression

```
inflm.fit <- influence.measures(fit)
which(apply(inflm.fit$is.inf, 1, any))
```

```
dfbeta(model, ...)
## S3 method for class ’lm’
dfbeta(model, infl = lm.influence(model, do.coef = TRUE), ...)

dfbetas(model, ...)
## S3 method for class ’lm’
dfbetas(model, infl = lm.influence(model, do.coef = TRUE), ...)

covratio(model, infl = lm.influence(model, do.coef = FALSE),
res = weighted.residuals(model))
```

```
cooks.distance(model, ...)
## S3 method for class ’lm’
cooks.distance(model, infl = lm.influence(model, do.coef = FALSE),
res = weighted.residuals(model),
sd = sqrt(deviance(model)/df.residual(model)),
hat = infl$hat, ...)
## S3 method for class ’glm’
cooks.distance(model, infl = influence(model, do.coef = FALSE),
res = infl$pear.res,
dispersion = summary(model)$dispersion,
hat = infl$hat, ...)
```

```
hatvalues(model, ...)
## S3 method for class ’lm’
```

```
hatvalues(model, infl = lm.influence(model, do.coef = FALSE), ...)

hat(x, intercept = TRUE)
```

Arguments

model an R object, typically returned by lm or glm.

infl influence structure as returned by lm.influence or influence (the la

res (possibly weighted) residuals, with proper default.

sd standard deviation to use, see default.

dispersion dispersion (for glm objects) to use, see default.

hat hat values H[i,i], see default.

type type of residuals for glm method for rstandard.

x the X or design matrix.

intercept should an intercept column be prepended to x?

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