

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.

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 Leverage and Influence

2.1 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.

2.2 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.

2.3 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.

2.4 Summary of Influence Statistics

- **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.

3 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} .

3.0.1 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.

$$DFBETA_a = \hat{\beta} - \hat{\beta}_{(a)} \quad (1)$$

$$= B(Y - Y_{\bar{a}}) \quad (2)$$

For k predictors in the model, there are $k + 1$ dfbetas.

3.0.2 DFFITS

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

$$DFFITS = \frac{\hat{y}_i - \widehat{y}_{i(k)}}{s_{(k)}\sqrt{h_{ii}}}$$

3.0.3 PRESS

The prediction residual sum of squares (PRESS) is a value associated with this calculation. When fitting linear models, PRESS can be used as a criterion for model selection, with smaller values indicating better model fits.

$$PRESS = \sum (y - y^{(k)})^2 \tag{3}$$

- $e_{-Q} = y_Q - x_Q\hat{\beta}_{-Q}$
- $PRESS_{(U)} = y_i - x_i\hat{\beta}_{(U)}$

3.1 Influential Observations : DFBeta and DFBetas

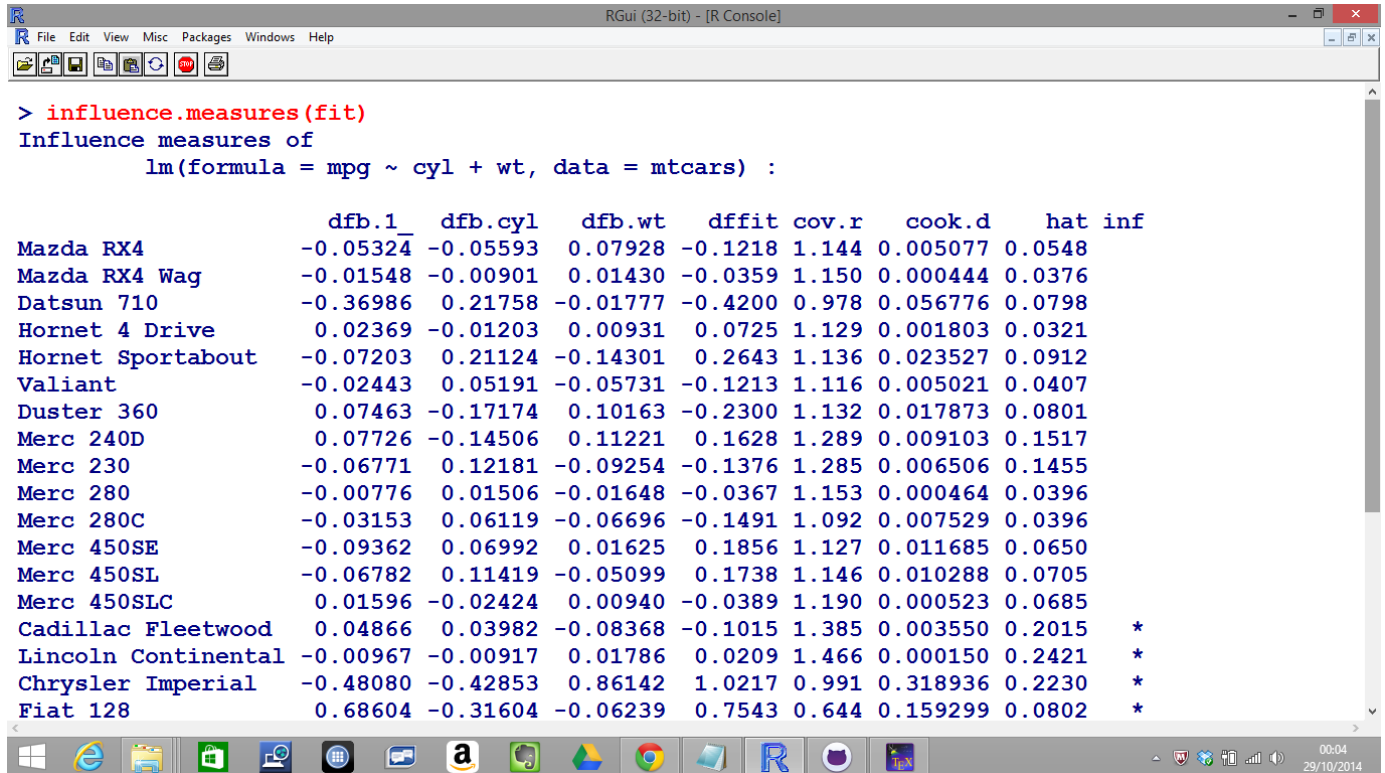


Figure 1:

4 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 latter 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.

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.