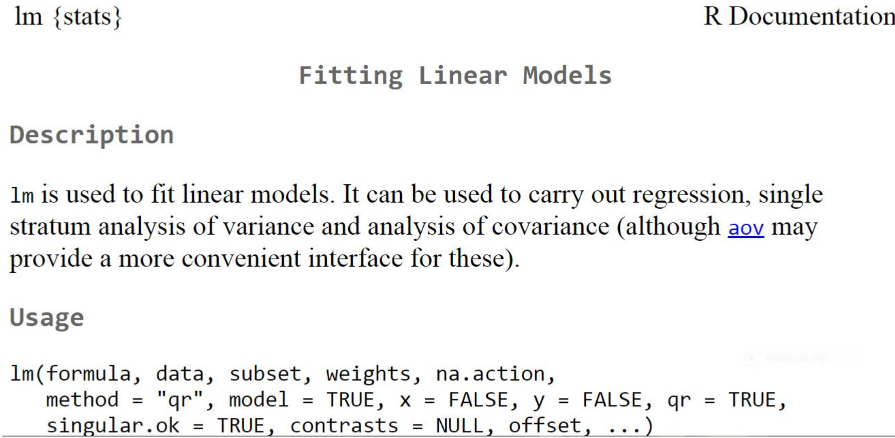
Beyond Linear Regression



lm {stats} R Documentation

Fitting Linear Models

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

```
lm(formula, data, subset, weights, na.action,
method = "qr", model = TRUE, x = FALSE, y = FALSE,
qr = TRUE, singular.ok = TRUE, contrasts = NULL,
offset, ...)
```

`weights` :

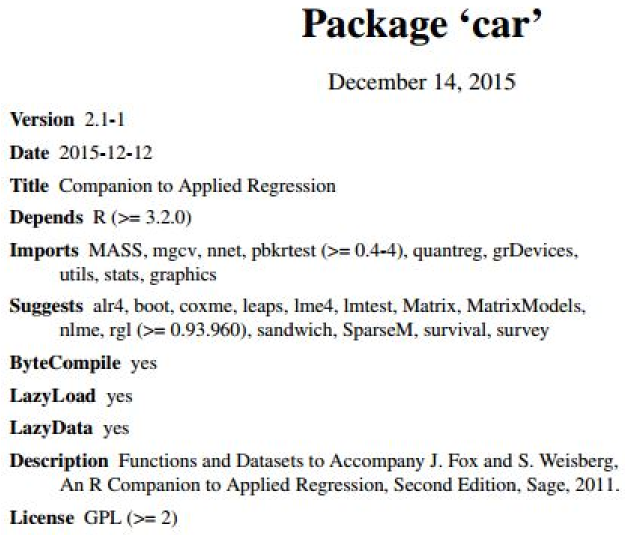
an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. If non-NULL, weighted least squares is used with weights weights

As an aside: Some useful Commands to know

- AIC and BIC
- predict
- confint
- coef
- influence
- dfbetas, dffits, covratio, cooks.distance

Regression Model Diagnostics

John Fox - Companion to Applied Regression



Package ‘car’

December 14, 2015

Version 2.1-1

Date 2015-12-12

Title Companion to Applied Regression

Depends R (>= 3.2.0)

Imports MASS, mgcv, nnet, pbkrtest (>= 0.4-4), quantreg, grDevices, utils, stats, graphics

Suggests alr4, boot, coxme, leaps, lme4, lmtest, Matrix, MatrixModels, nlme, rgl (>= 0.93.960), sandwich, SparseM, survival, survey

ByteCompile yes

LazyLoad yes

LazyData yes

Description Functions and Datasets to Accompany J. Fox and S. Weisberg, An R Companion to Applied Regression, Second Edition, Sage, 2011.

License GPL (>= 2)

Residual Diagnostics - car package

Outliers

```
# Assessing Outliers

# Bonferonni p-value for most extreme obs
outlierTest(fit)

#qq plot for studentized resid
qqPlot(fit, main="QQ Plot")


# leverage plots
leveragePlots(fit)
```

Residual Diagnostics

```
# Influential Observations
# added variable plots
av.Plots(fit)

# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2
plot(fit, which=4, cook.levels=cutoff)
```

Residual Diagnostics

```
# Influence Plot
influencePlot(fit, id.method="identify",
   main="Influence Plot")
```

Residual Diagnostics

Non-constant Error Variance

```
# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(fit)

# plot studentized residuals vs. fitted values
spreadLevelPlot(fit)
```

Residual Diagnostics

Multi-collinearity

```
# Evaluate Collinearity

vif(fit) # variance inflation factors

sqrt(vif(fit)) > 2 # problem?
```

Residual Diagnostics

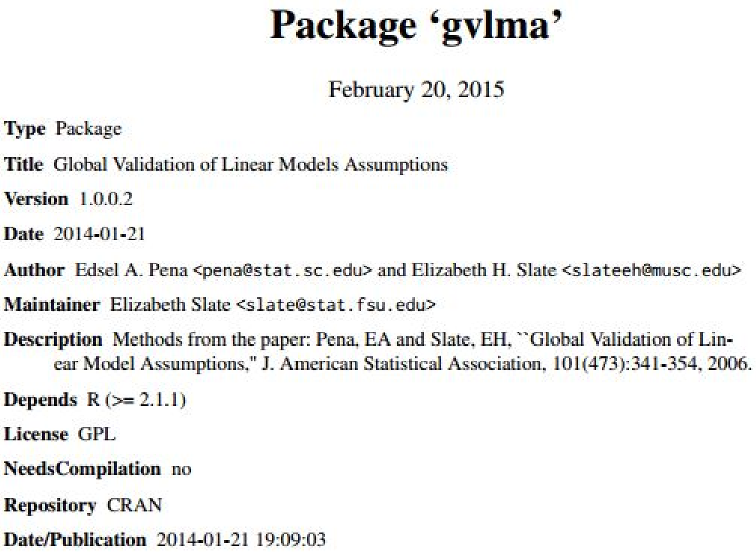
Nonlinearity

```
# Evaluate Nonlinearity
# component + residual plot
crPlots(fit)

# Ceres plots
ceresPlots(fit)
```

Residual Diagnostics

Autocorrelation : Non-independence of Errors

```
# Test for Autocorrelated Errors
durbinWatsonTest(fit)
```



Package ‘gvlma’

February 20, 2015

Type Package

Title Global Validation of Linear Models Assumptions

Version 1.0.0.2

Date 2014-01-21

Author Edsel A. Pena <pena@stat.sc.edu> and Elizabeth H. Slate <slateeh@musc.edu>

Maintainer Elizabeth Slate <slate@stat.fsu.edu>

Description Methods from the paper: Pena, EA and Slate, EH, ``Global Validation of Linear Model Assumptions,'' J. American Statistical Association, 101(473):341-354, 2006.

Depends R (>= 2.1.1)

License GPL

NeedsCompilation no

Repository CRAN

Date/Publication 2014-01-21 19:09:03

Figure:

Residual Diagnostics

Additional Diagnostic Help

The `gvlma( )` function in the gvlma package, performs a global validation of linear model assumptions as well separate evaluations of skewness, kurtosis, and heteroscedasticity.

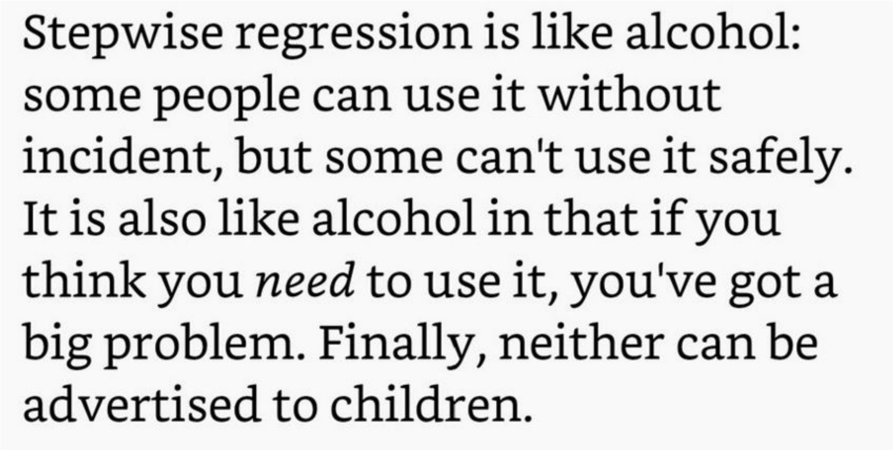
```
# Global test of model assumptions
library(gvlma)
gvmodel <- gvlma(fit)
summary(gvmodel)
```

By The Way...

- The spellings homoskedasticity and heteroskedasticity are also frequently used.
- J. Huston McCulloch argued that there should be a "k" in the middle of the word and not a "c".
- His argument was that the word had been constructed in English directly from Greek roots rather than coming into the English language indirectly via the French.

See McCulloch, J. Huston (March 1985). "Miscellanea: On Heteros∗edasticity". Econometrica 53 (2): 483. JSTOR 1911250.

Stepwise Regression



Stepwise regression is like alcohol: some people can use it without incident, but some can't use it safely. It is also like alcohol in that if you think you *need* to use it, you've got a big problem. Finally, neither can be advertised to children.

Figure: Tony Fischetti

Stepwise Regression

- Stepwise regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention.
- This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables.
- Stepwise regression basically fits the regression model by adding/dropping predictor variables one at a time based on a specified criterion.
- It is one of the method to handle higher dimensionality of data set.

Stepwise Regression

- **Standard stepwise regression** does two things. It adds and removes predictors as needed for each step.
- Forward selection starts with most significant predictor in the model and adds variable for each step.
- **Backward elimination** starts with all predictors in the model and removes the least significant variable for each step.
- The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables.

Stepwise Regression

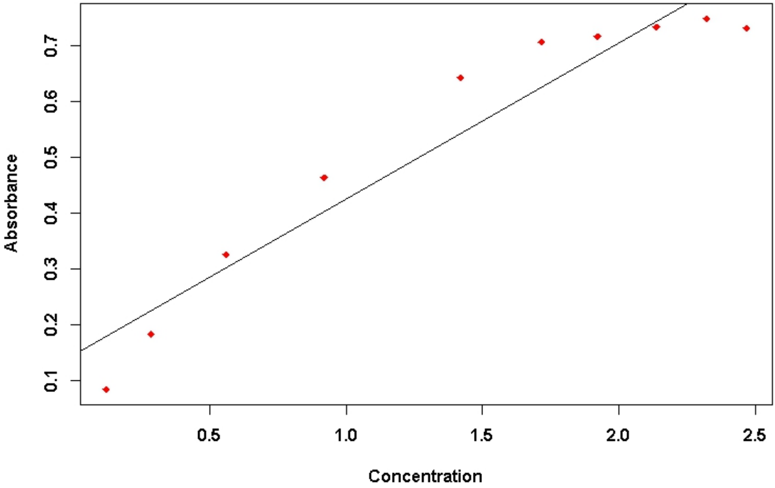
```
#BACKWARD SELECTION
FitBS = lm(mpg ~ . ,data=mtcars)
step(FitAll, direction = "backward")

#FORWARD SELECTION
FitFS = lm(mpg ~ 1)
step(FitAll, direction = "forward")
```

Stepwise Regression

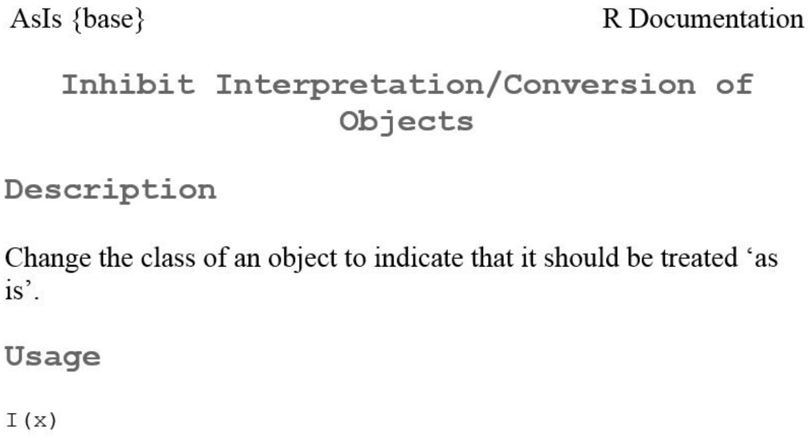
```
library(MASS)
fit <- lm(y~x1+x2+x3,data=mydata)
stepwise <- stepAIC(fit, direction="both")

stepwise$anova # display results
```

Polynomial Regression

Curvilinear Relationship



Specifying Polynomial Models



AsIs {base} R Documentation

Inhibit Interpretation/Conversion of Objects

Description

Change the class of an object to indicate that it should be treated ‘as is’.

Usage

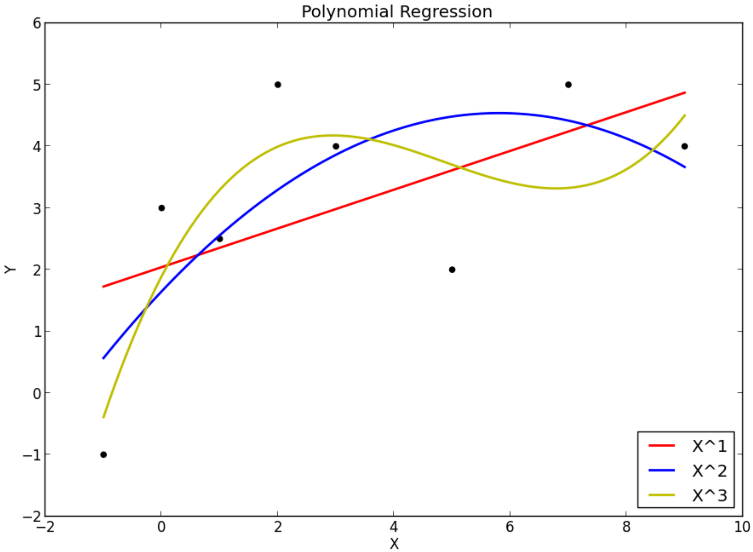
```
I(x)
```

Polynomial Regression

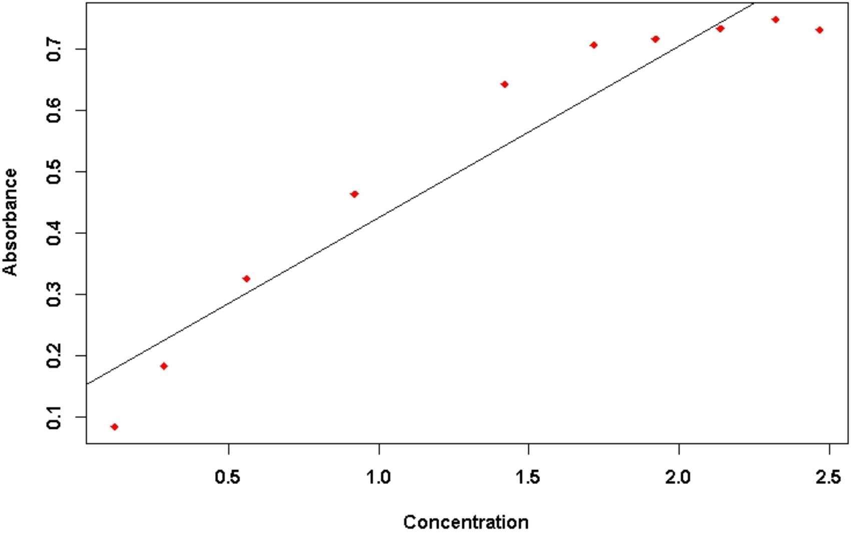
- Suppose that, when you inspect the data, the best fit line is not a straight line, rather a curve that fits into the data points.
- A regression equation is a polynomial regression equation if the power of independent variable is more than 1.
- The equation below represents a quadratic equation:

$$y = b_0 + +b_1 x + b_2 x^2$$

Polynomial Regression



Polynomial Regression



```
# Absorbance
x<-c(0.084, 0.183, 0.326, 0.464, 0.643,
0.707, 0.717, 0.734, 0.749, 0.732)

# Concentration
y<-c(0.123, 0.288, 0.562, 0.921, 1.420,
1.717, 1.921, 2.137, 2.321, 2.467)
```

- Compare linear, quadratic and cubic fit.

Polynomial Regression

Fitting a polynomial of degree 3.

```
lm(y ~ x + I(x^2) + I(x^3))


lm(y ~ poly(x, 3))
```

Polynomial Regression

Important Points:

- While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in **over-fitting**.
- Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem.
- Here is an example of how plotting can help: underfitting-overfitting
- Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing weird results on extrapolation.

Segmented Regression

Segmented Regression

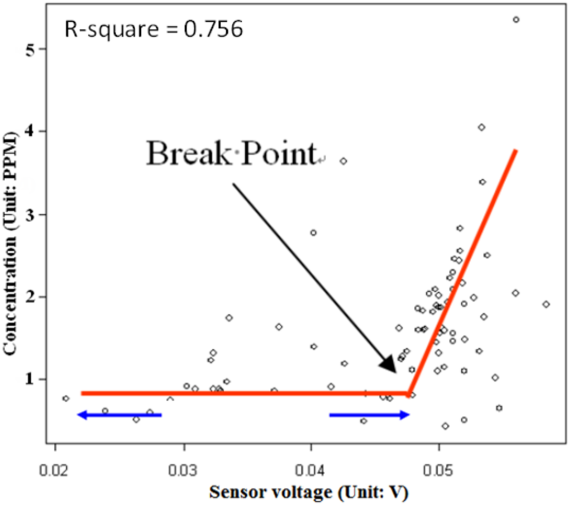
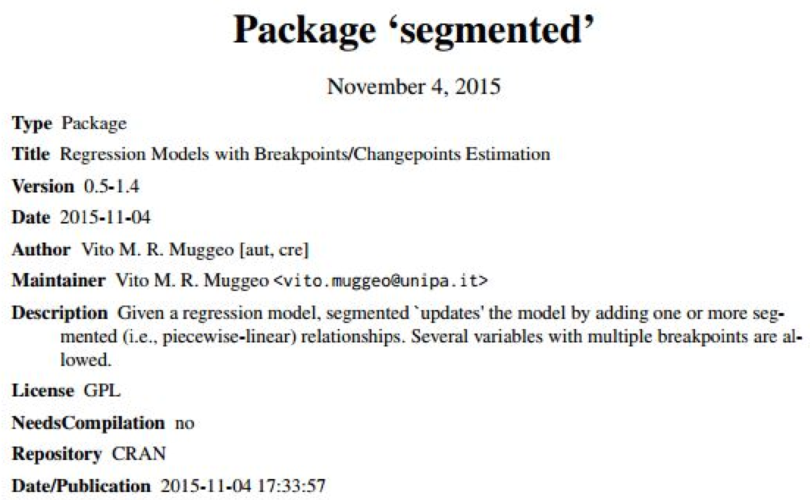


Figure:

Segmented Regression

- **Segmented regression** is a method in regression analysis in which the independent variable is partitioned into intervals and a separate line segment is fit to each interval.
- Segmented regression analysis can also be performed on multivariate data by partitioning the various independent variables.
- Segmented regression is useful when the independent variables, clustered into different groups, exhibit different relationships between the variables in these regions.
- The boundaries between the segments are **breakpoints**.

Segmented Regression



Package ‘segmented’

November 4, 2015

Type Package

Title Regression Models with Breakpoints/Changepoints Estimation

Version 0.5-1.4

Date 2015-11-04

Author Vito M. R. Muggeo [aut, cre]

Maintainer Vito M. R. Muggeo <vito.muggeo@unipa.it>

Description Given a regression model, segmented `updates' the model by adding one or more segmented (i.e., piecewise-linear) relationships. Several variables with multiple breakpoints are allowed.

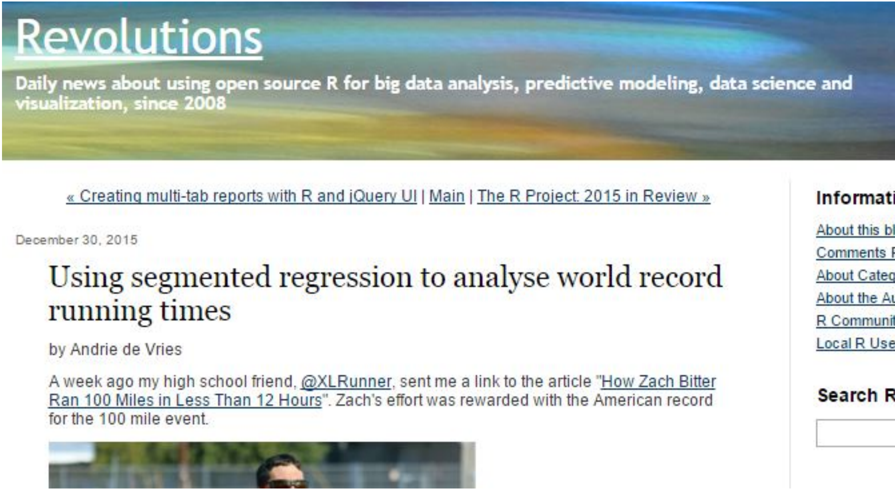
License GPL

NeedsCompilation no

Repository CRAN

Date/Publication 2015-11-04 17:33:57

Segmented Regression



Revolutions

Daily news about using open source R for big data analysis, predictive modeling, data science and visualization, since 2008

« Creating multi-tab reports with R and jQuery UI | Main | The R Project: 2015 in Review »

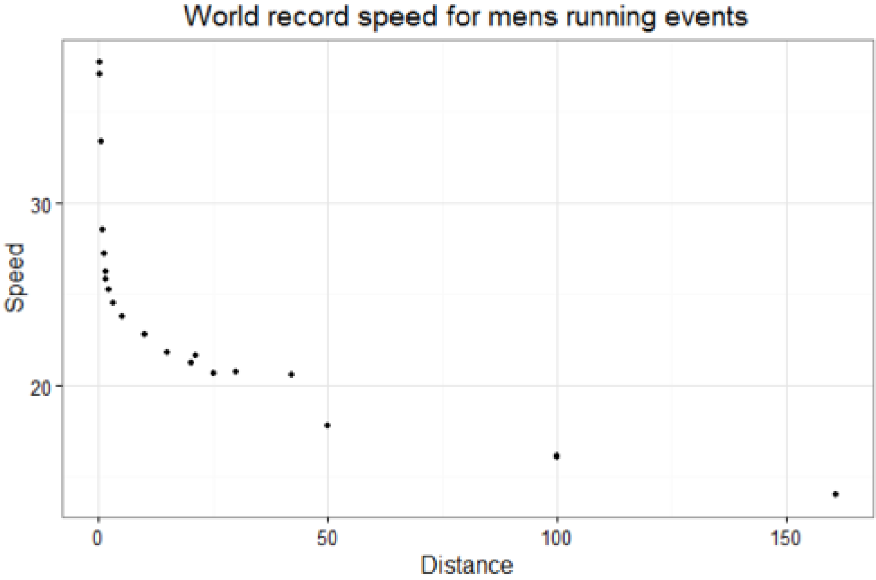
December 30, 2015

Using segmented regression to analyse world record running times

by Andrie de Vries

A week ago my high school friend, @XLRunner, sent me a link to the article "How Zach Bitter Ran 100 Miles in Less Than 12 Hours". Zach's effort was rewarded with the American record for the 100 mile event.

Segmented Regression



Segmented Regression

- The `segmented()` function allows you to modify a fitted object of class `lm` or `glm`, specifying which of the independent variables should have segments (kinks).
- In my case, I fitted a linear model with a single variable (log of distance), and allowed `segmented()` to find a single kink point.

Segmented Regression

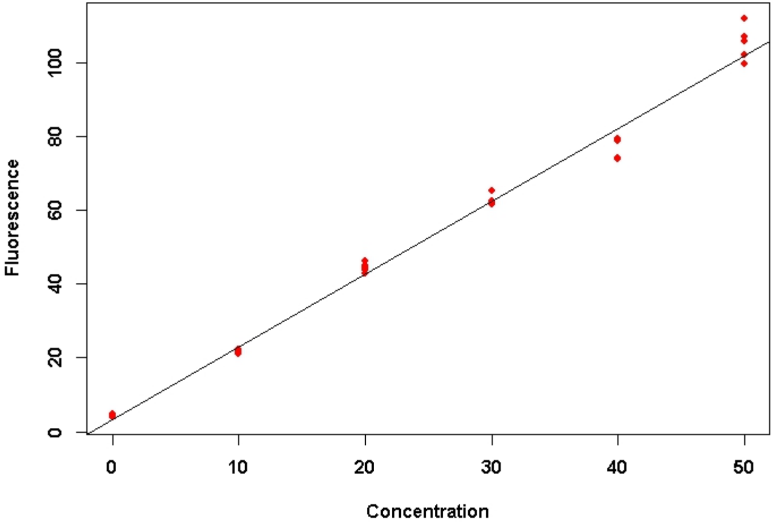
- First fit a generic linear model, then use the `segmented()` function to fit the piecewise regression.
- The `segmented()` function takes for its arguments the generic linear model, `seg.Z` which is a one sided formula describing the predictor with a segment
- `psi` is a starting value of the breakpoint.

Segmented Regression

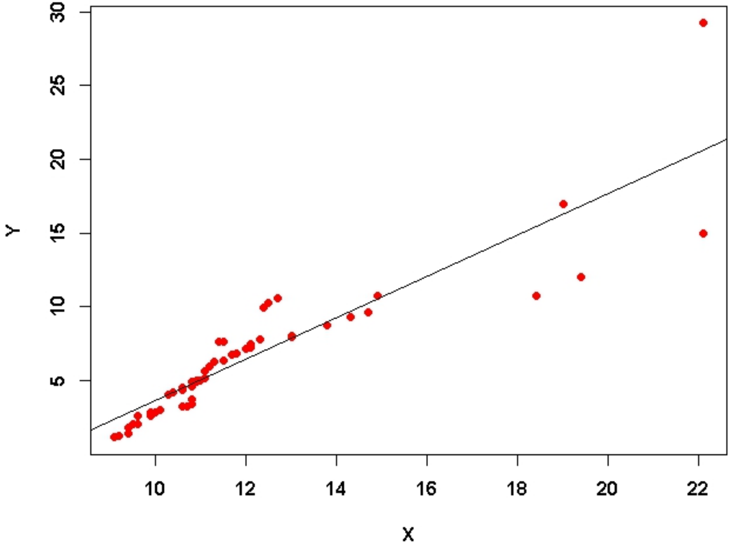
```
#Andrie De Vries’s Model

# Fit linear model
lfit <- lm(Speed ~ logDistance, data = modeldata)

# Fit segmented model
sfit <- segmented(lfit, seg.Z = ~ logDistance)

#Identify Breakpoints
exp10(sfit$psi)
summary(sfit)
```

Weighted Regression

Heteroskedascity



Heteroskedascity



Robust Regression

Robust Regression

- When fitting a least squares regression, we might find some outliers or high leverage data points.
- However data points are not data entry errors, neither they are from a different population than most of our data.
- No proper reason to exclude them from the analysis.

Robust Regression

- Robust regression might be a good strategy since it is a compromise between excluding these points entirely from the analysis and including all the data points and treating all them equally in OLS regression.
- The idea of robust regression is to weigh the observations differently based on how well behaved these observations are.

Robust Regression

- When fitting a least squares regression, we might find some outliers or high leverage data points.
- We have decided that these data points are not data entry errors, neither they are from a different population than most of our data. So we have no proper reason to exclude them from the analysis.

Robust Regression

- Robust regression might be a good strategy since it is a compromise between excluding these points entirely from the analysis and including all the data points and treating all them equally in OLS regression.
- The idea of robust regression is to weigh the observations differently based on how well behaved these observations are.

Robust Regression

There are several weighting functions that can be used for Robust Regression.

- Huber
- Hampel
- Bisquare

Robust Regression

Huber Weighting
In Huber weighting, observations with small residuals get a weight of 1 and the larger the residual, the smaller the weight.

This is defined by the weight function

$$w(e) = \begin{cases} 1 & \text{for } |e| <= k \\ \frac{k}{|e|} & \text{for } |e| > k \end{cases} \tag{1}$$

Robust Regression

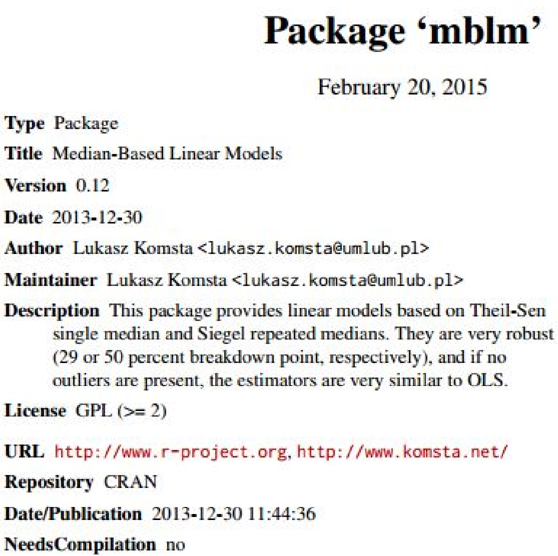
Tuning Constant

- The value k for the Huber and bisquare estimators is called a **tuning constant**
- Smaller values of k produce more resistance to outliers, but at the expense of lower efficiency when the errors are normally distributed.
- The tuning constant is generally picked to give reasonably high efficiency in the normal case; in particular, $k = 1.345\sigma$ for the Huber and $k = 4.685\sigma$ for the bisquare (where σ is the standard deviation of the errors) produce 95-percent efficiency when the errors are normal, and still offer protection against outliers.

Robust Regression

- The idea of robust regression is to weigh the observations differently based on how well behaved these observations are.
- Roughly speaking, it is a form of weighted and reweighted least squares regression (i.e. a two step process , first fitting a linear model, then a robust model to correct for the influence of outliers).
- Robust regression is done by **iterated re-weighted least squares (IRLS)**.
- The rlm command in the MASS package command implements several versions of robust regression.

Median Based Linear Models



Package ‘mblm’

February 20, 2015

Type Package

Title Median-Based Linear Models

Version 0.12

Date 2013-12-30

Author Lukasz Komsta <lukasz.komsta@umlub.pl>

Maintainer Lukasz Komsta <lukasz.komsta@umlub.pl>

Description This package provides linear models based on Theil-Sen single median and Siegel repeated medians. They are very robust (29 or 50 percent breakdown point, respectively), and if no outliers are present, the estimators are very similar to OLS.

License GPL (>= 2)

URL http://www.r-project.org, http://www.komsta.net/

Repository CRAN

Date/Publication 2013-12-30 11:44:36

NeedsCompilation no

Censored Regression

Censored Models: Tobit Regression

Examples of Tobit Analysis
In the 1980s there was a federal law restricting speedometer readings to no more than 85 mph. So if you wanted to try and predict a vehicle's top-speed from a combination of horse-power and engine size, you would get a reading no higher than 85, regardless of how fast the vehicle was really traveling. This is a classic case of right-censoring (censoring from above) of the data. The only thing we are certain of is that those vehicles were traveling at least 85 mph.

Censored Models: Tobit Regression

Examples of Tobit Analysis

- A research project is studying the level of lead in home drinking water as a function of the age of a house and family income.
- The water testing kit cannot detect lead concentrations below 5 parts per billion (ppb).
- The EPA considers levels above 15 ppb to be dangerous. These data are an example of left-censoring (censoring from below).

Censored Models: Tobit Regression

Examples of Tobit Analysis

- Consider the situation in which we have a measure of academic aptitude (scaled 200-800) which we want to model using reading and math test scores, as well as, the type of program the student is enrolled in (academic, general, or vocational).
- The problem here is that students who answer all questions on the academic aptitude test correctly receive a score of 800, even though it is likely that these students are not "truly" equal in aptitude.
- The same is true of students who answer all of the questions incorrectly. All such students would have a score of 200, although they may not all be of equal aptitude.

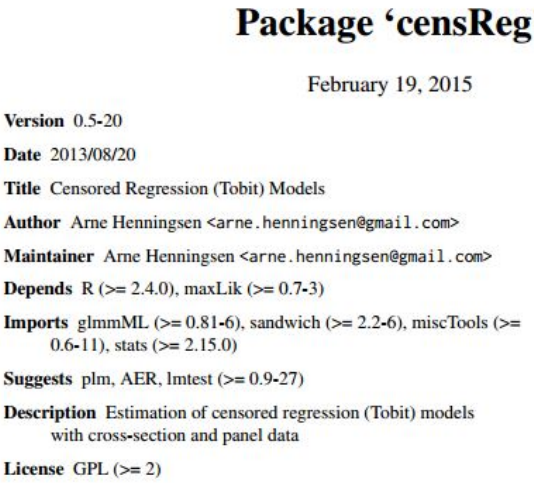
Censored Models: Tobit Regression

Tobit Regression

The tobit model, also called a censored regression model, is designed to estimate linear relationships between variables when there is either left- or right-censoring in the dependent variable (also known as censoring from below and above, respectively).

Censored Models: Tobit Regression

Censoring
Censoring from above takes place when cases with a value at or above some threshold, all take on the value of that threshold, so that the true value might be equal to the threshold, but it might also be higher. In the case of censoring from below, values those that fall at or below some threshold are censored.



Package ‘censReg’

February 19, 2015

Version 0.5-20

Date 2013/08/20

Title Censored Regression (Tobit) Models

Author Arne Henningsen `<arne.henningsen@gmail.com>`

Maintainer Arne Henningsen `<arne.henningsen@gmail.com>`

Depends R (>= 2.4.0), maxLik (>= 0.7-3)

Imports glmmML (>= 0.81-6), sandwich (>= 2.2-6), miscTools (>= 0.6-11), stats (>= 2.15.0)

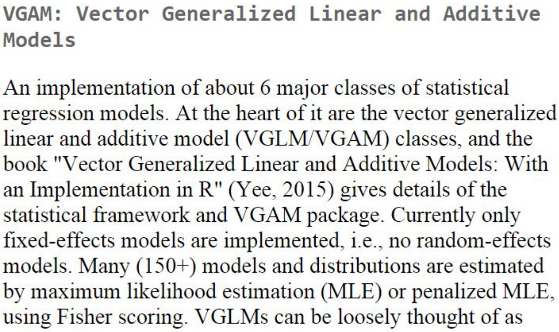
Suggests plm, AER, lmtest (>= 0.9-27)

Description Estimation of censored regression (Tobit) models with cross-section and panel data

License GPL (>= 2)

Censored Models: Tobit regression

We run the tobit model, using the `vglm()` function of the **VGAM** package.



VGAM: Vector Generalized Linear and Additive Models

An implementation of about 6 major classes of statistical regression models. At the heart of it are the vector generalized linear and additive model (VGLM/VGAM) classes, and the book "Vector Generalized Linear and Additive Models: With an Implementation in R" (Yee, 2015) gives details of the statistical framework and VGAM package. Currently only fixed-effects models are implemented, i.e., no random-effects models. Many (150+) models and distributions are estimated by maximum likelihood estimation (MLE) or penalized MLE, using Fisher scoring. VGLMs can be loosely thought of as

```
library(VGAM)
summary(m <- vglm(apt ~ read + math + prog,
             tobit(Upper = 800), data = dat))
```

```
## Call:
## vglm(formula = apt ~ read + math + prog, family = to
##     data = dat)
##
## Pearson Residuals:
##          Min    1Q Median   3Q Max
## mu      -2.6 -0.76 -0.051 0.79 4.1
## log(sd) -1.1 -0.62 -0.369 0.25 5.4
## Coefficients:
##                 Estimate Std. Error z value
## (Intercept):1      209.6     32.457     6.5
## (Intercept):2        4.2      0.053    79.4
## read                 2.7      0.618     4.4
## math                 5.9      0.705     8.4
## proggeneral        -12.7     12.355    -1.0
## progvocational     -46.1     13.770    -3.4
## Number of linear predictors:  2
##
## Names of linear predictors: mu, log(sd)
##
## Dispersion Parameter for tobit family:   1
##
## Log-likelihood: -1041 on 394 degrees of freedom
##
## Number of iterations: 4
```

Truncated Regression

Truncated and Censored Regression

- Censored regression models are often confused with truncated regression models.
- Truncated regression models are used for data where whole observations are missing so that the values for the dependent and the independent variables are unknown.
- Censored regression models are used for data where only the value for the dependent variable is unknown while the values of the independent variables are still available.

Truncated Regression

Case Studies 1

- One example of truncated samples come from historical military height records. Many armies imposed a minimum height requirement on soldiers.
- This implies that men shorter than the MHR are not included in the sample.
- This implies that samples drawn from such records are statistically incomplete, in as much as a substantial portion of the underlying population's height distribution is unavailable for analysis.
- Consequently, without proper statistical correction, any results obtained from such deficient samples, such as means, correlations, or regression coefficients are wrong (biased).

Truncated Regression

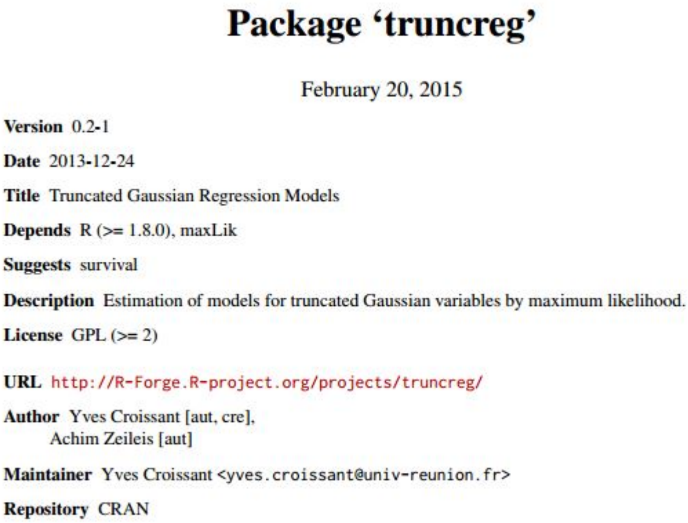
Case Studies 2

- A study of students in a special GATE (gifted and talented education) program wishes to model achievement as a function of language skills and the type of program in which the student is currently enrolled.
- A major concern is that students are required to have a minimum achievement score of 40 to enter the special program.
- Thus, the sample is truncated at an achievement score of 40.

Truncated Regression

- In such a case truncated regression has the considerable advantage of immediately providing consistent and unbiased estimates of the coefficients of the independent variables, as well as their standard errors, thereby allowing for further statistical inference, such as the calculation of the t-values of the estimates.

Truncated Regression



Package 'truncreg'

February 20, 2015

Version 0.2-1

Date 2013-12-24

Title Truncated Gaussian Regression Models

Depends R (>= 1.8.0), maxLik

Suggests survival

Description Estimation of models for truncated Gaussian variables by maximum likelihood.

License GPL (>= 2)

URL `http://R-Forge.R-project.org/projects/truncreg/`

Author Yves Croissant [aut, cre],
Achim Zeileis [aut]

Maintainer Yves Croissant `<yves.croissant@univ-reunion.fr>`

Repository CRAN

Figure:

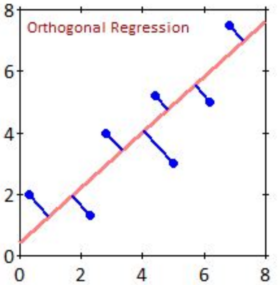
Truncated regression

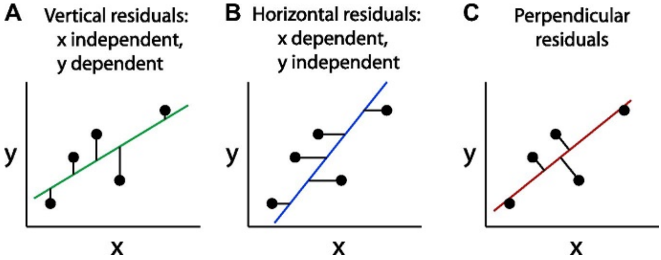
- Use the truncreg function in the **truncreg** package to estimate a truncated regression model.
- The point argument indicates where the data are truncated, and the direction indicates whether it is left or right truncated.

```
m <- truncreg(achiv ~ langscore + prog,
   data = dat, point = 40, direction = "left")
```

Deming Regression

Deming Regression



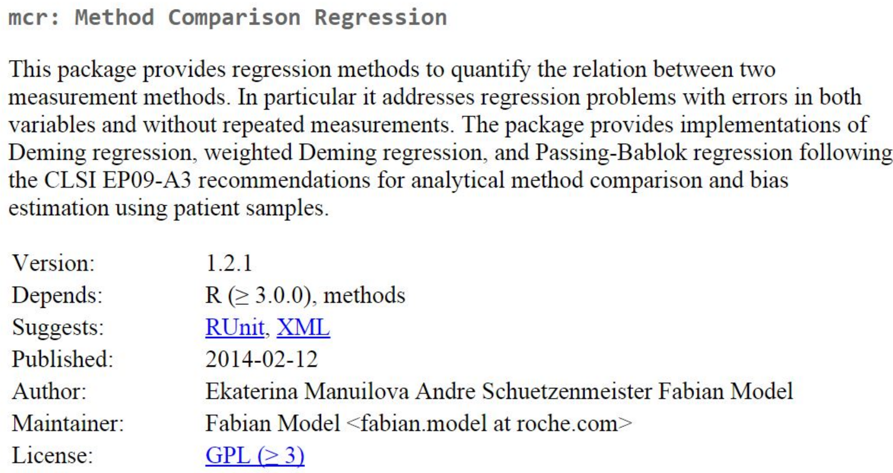


Deming Regression

- An ion-selective electrode (ISE) determination of sulphide from sulphate-reducing bacteria was compared with a gravimetric determination.
- Each pair of determinations were taken from the same sample.
- The results obtained by both methods are expressed in milligrams of sulphide, and are tabulated below.

ISE method	108	102	152	73	106	114	128
gravimetry	105	96	113	91	108	101	141
ISE method	85	106	114	128	142	160	128
gravimetry	91	108	101	141	161	182	118

Deming Regression

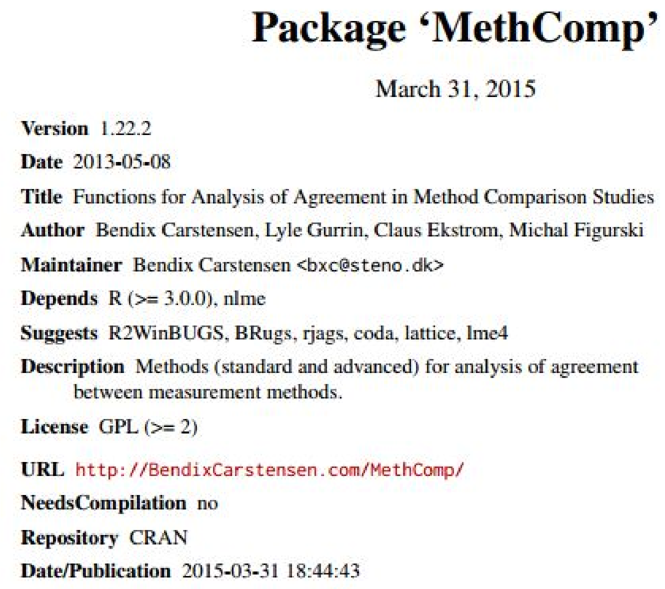


mcr: Method Comparison Regression

This package provides regression methods to quantify the relation between two measurement methods. In particular it addresses regression problems with errors in both variables and without repeated measurements. The package provides implementations of Deming regression, weighted Deming regression, and Passing-Bablok regression following the CLSI EP09-A3 recommendations for analytical method comparison and bias estimation using patient samples.

Version:	1.2.1
Depends:	R (≥ 3.0.0), methods
Suggests:	RUnit, XML
Published:	2014-02-12
Author:	Ekaterina Manuilova Andre Schuetzenmeister Fabian Model
Maintainer:	Fabian Model <fabian.model at roche.com>
License:	GPL (≥ 3)

Deming Regression



Package ‘MethComp’

March 31, 2015

Version 1.22.2

Date 2013-05-08

Title Functions for Analysis of Agreement in Method Comparison Studies

Author Bendix Carstensen, Lyle Gurrin, Claus Ekstrom, Michal Figurski

Maintainer Bendix Carstensen <bxc@steno.dk>

Depends R (>= 3.0.0), nlme

Suggests R2WinBUGS, BRugs, rjags, coda, lattice, lme4

Description Methods (standard and advanced) for analysis of agreement between measurement methods.

License GPL (>= 2)

URL http://BendixCarstensen.com/MethComp/

NeedsCompilation no

Repository CRAN

Date/Publication 2015-03-31 18:44:43