The Comparison of Some Methods in Analysis of Linear Regression Using *R* Software

Ilir Palla
Department of Mathematics and Physics, Faculty of Natural and Human Sciences, University of Korça, Albania

Abstract

This article contains the OLS method, WLS method and bootstrap methods to estimate coefficients of linear regression and their standard deviation. If regression holds random errors with constant variance and if those errors are independent normally distributed we can use least squares method, which is accurate for drawing inferences with these assumptions. If the errors are heteroscedastic, meaning that their variance depends from explanatory variable, or have different weights, we can't use least squares method because this method cannot be safe for accurate results. If we know weights for each error, we can use weight least squares method. In this article we have also described bootstrap methods to evaluate regression parameters. The bootstrap methods improved quantile estimation. We simulated errors with non constant variances in a linear regression using R program and comparison results. Using this software we have found confidence interval, estimated coefficients, plots and results for any case.

Keywords: homoscedasticity, heteroscedasticity, studentized errors, ncvTest.

Introduction

The processes of model fitting are: model selection, parameters estimation, checking the adequacy of the model, calculating confidence intervals and testing hypotheses about the parameters in the model and interpreting the results. In the chosen model we have to consider: the assumption of observation independence or at least unrelated, a single error term in the model, the choice of scale for the analysis, the choice of variables that will be included in the model.

If the variability comes only from variability of errors and variance of the errors term does not depend from the values of the explanatory variable X, thus have equal variance. This property is called homoscedasticity.

A test that controls the dispersion of errors in a regression is called "Breusch-Pagan test" by Breusch, and Pagan, 1979, which control the hypothesis if the variance of errors is constant versus the alternative the error variance changes with the level of response. With command **"**ncvTest"**,** which is into the R program we compute a score test of the hypothesis of constant error variance.

In issue 1.1 is treated least squares method to estimate the parameters of regression, hat matrix, leverages, studentized residuals, variance of parameters and the standard error for a particular component.

In issue 1.2 is treated weight least squares method. In this case the errors have unequal variance where the form is known. A test that controls the dispersion of errors in a regression is called "**Breusch-Pagan test**" by Breusch, and Pagan, 1979, which control the hypothesis if the variance of errors is constant versus the alternative the error variance changes with the level of response.

In issue 1.3 is treated bootstrap methods to estimate parameters of linear regression. Finally in issue 1.4 is simulated a linear regression with heteroscedasticity errors. Using R software we have found confidence intervals, estimated coefficients, plots and results for any case. In this issue are shown the commands that are used in R to get the results.

1.1 Least squares linear regression

A vector of observations $y = (y_1, ..., y_n)$ having *n* component is assumed to be a realization of a random variable *Y* whose components are independently. If we repeat the observation for a value given by the explanatory variable, we will not necessarily have the same value from the explained variable, so we may have another value because they explained variable will be considered a random variable (Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, "Probability and Statistics for engineers and Scientists").

Explanatory variable *X* will be considered as fixed values, matrix with size $n \times p$.

The linear regression model is:

$$Y = X\beta + e, \tag{1}$$

where: $C(Y / X) = C(e) = \sigma^2 I_n$, $E(e) = 0$. $C(e)$ is covariance matrix, *I_n* is matrix unit $n \times n$,

$E(Y / X) = X\beta$.

To estimate the parameters $\beta_0, \beta_1, ..., \beta_p$, we use the least squares method. The sum of the squared errors (*SSE*) is:

$$SSE(\beta) = \sum_{i=1}^{n}(Y_i - \sum_{j=0}^{p}\beta_j x_{ij})^2 = (Y - X\beta)^T (Y - X\beta). \tag{2}$$

The least squares estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$, which minimizes $RSS(\beta)$ are:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

if $(X^T X)^{-1}$ exists. The equation $Y = X\beta$ is called the theoretical regression equation, while $\hat{y} = X\hat{\beta}$ the fitted least squares regression equation.

$$X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy \tag{4}$$

$H = X(X^T X)^{-1} X^T$ is called hat matrix. Residuals: $\hat{e} = y - \hat{y} = (I_n - H)Y$.

Studentized residuals are:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_i}}, \tag{5}$$

which if the model is assumed correctly they have variance equal to 1 and $corr(r_i, r_j)$ tend to be small. The quantities $h_i = h_{ii}$ are known as leverages, the main diagonal elements of hat matrix *H*. Studentization of residuals can only correct the natural non-constant variance in residuals when the errors have constant variance. If there is some underlying heteroscedascity in the errors, studentization can not correct for it.

$E(r_i) = E(\frac{e_i}{\sigma\sqrt{(1-h_i)}}) = 0$, and $\text{var}(r_i) = \frac{1}{\sigma^2(\sqrt{1-h_i})^2}\text{var}(e_i) = \frac{\sigma^2(1-h_i)}{\sigma^2(1-h_i)} = 1.$

Residual sum of squares:

$RSS(\hat{\beta}) = \min_{\beta}[RSS(\beta)] = \hat{e}^T\hat{e} = (Y - X\hat{\beta})^T(Y - X\hat{\beta})y^T = (I_n - H)Y$.

If $E(e)=0$ and $C(e)=\sigma^2 I_n$, then

$$E(\hat{\beta})=\beta\,,\ \operatorname{var}(\hat{\beta})=\sigma^2(X^T X)^{-1} \tag{6}$$

$$\hat{\sigma}^2=\frac{RSS(\hat{\beta})}{n-(p+1)}. \tag{7}$$

The standard error for a particular component is: $se(\hat{\beta}_i)=\sqrt{(X^T X)_{ii}^{-1}}\,\hat{\sigma}$.

If errors are independent and identically normally distributed with mean 0 and variance σ^2, which means $y\sim N(X\beta,\sigma^2 I_n)$, then using the fact that linear combinations of normally distributed values are also normal:

$$\hat{\beta}\sim N(\beta,(X^T X)^{-1}\sigma^2). \tag{8}$$

We can use these results to test hypothesis about parameters and to construct confidence intervals for β. The estimators have normal distribution, if the errors have normal distribution. If the errors do not have normal distribution, then this method cannot be safe for accurate conclusions.

1.2 Weighted least squares method

The basic tool for examining the fit is the residuals, and we have already looked for patterns in residuals and assessed the normality of their distribution. Where the errors are uncorrelated, but have unequal variance where the form is known we can use weighted least squares (WLS). A simple case:

$$E(Y/X=x_i)=\beta^T x_i, \tag{9}$$

$$\operatorname{var}(Y/X=x_i)=\operatorname{var}(e_i)=\frac{\sigma^2}{w_i}, \tag{10}$$

where w_i for *i* = 1,...,n are known. In this situation the model is:

$$Y=X\beta+e\,,\ \operatorname{var}(e)=\sigma^2 W^{-1}. \tag{11}$$

The sum of the squared errors in this case is:

$$RSS(\beta)=(Y-X\beta)^T W(Y-X\beta)=\sum w_i(Y_i-x_i^T\beta)^2. \tag{12}$$

Weighted least squares estimations of parameters are:

$$\hat{\beta}=(X^T WX)^{-1}X^T WY. \tag{13}$$

Weighted least squares is appropriate when the form of the non-constant variance is either known exactly. A test that controls the dispersion of errors in a regression is called "**Breusch-Pagan test**" by Breusch, and Pagan, 1979, which control the hypothesis if the variance of errors is constant versus the alternative the error variance changes with the level of response (fitted values), or with a linear combination of predictors.

Suppose we have a regression model:

$$Y=\beta_0+\beta_1 X_1+\cdots\beta_p X_p+e,\quad i=1,\dots,n \tag{14}$$

$$\operatorname{var}(e\,|\,\underline{X}) = \sigma^2 h(\underline{X}). \tag{15}$$

If the function h (.) is linear in relation to independent variables, then the last equation can be written in the form: $\operatorname{var}(e\,|\,\underline{X}) = \sigma^2(\alpha_0 + \alpha_1 X_1 + \cdots \alpha_p X_p).$

We build auxiliary regression: $\hat{e}^2 = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p$. Estimate the regression coefficients with the least square method and check the hypothesis that all coefficients of the least regression equation, in addition to the first coefficient, get the zero value.

$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0$ *versus* $H_a : \textit{there is at least one } i \textit{ where } \gamma_i \neq 0, i = 1,\ldots, p.$

1.3 Re-sampling errors and re-sampling cases in regression

Re-sampling errors are a bootstrap method to provide more accurate analysis. The application of bootstrap in linear regressions was originally made by Bradley Efron and Robert J Tibshirani (1993). The algorithm for model based re-sampling in simple linear regression (A.C.Davison and D.V.Hinkley 1997, p. 262). Another bootstrap method is re-sampling cases (A.C.Davison and D.V.Hinkley 1997, p. 264). In this method we make no assumption if variance is constant.

1.4 Simulation: True regression, OLS regression, WLS regression and bootstrap methods. Using R software

We start with regression where we know the true linear regression and simulation the errors of regression: $Y = 20 + 10X$,

where Y- define the response variable. X - predictor variable (explanatory variable).

We simulate the errors in regression such that their variances are linearly dependent on variable X . The aim here is to see the difference between results derived from different methods. At first download packages in R program: library(car), library(MASS), library(nlme), library(boot).

The blue and bold words written below are the commands we used in the program R.

```
x<-seq(0,29,by=1) # explanatory variable
y = 20+10*x + rnorm(30,0,sapply(x,function(x){1+5*x})) # response variable
sim.data<-data.frame(y,x)
```

OLS regression

```
fit.ols = lm(y~x, data=sim.data) # OLS regression
summary(fit.ols)
```

Coefficients (OLS regression):

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept β_0	9.348	25.969	0.360	0.722
β_1	10.024	1.538	6.518	0.00013***

Residual standard error: 74.56 on 28 degrees of freedom. Multiple R-squared: 0.6028, Adjusted R-squared: 0.5886. F-statistic: 42.48 on 1 and 28 DF, p-value: 0.0001305.

We find the confidence intervals for the coefficients of the OLS regression using: **confint(fit.ols).**

WLS regression

With command **ncvTest** computes a score test of the hypothesis of constant error variance against the alternative that the error variance changes with the level of the response (fitted values), or with a linear combination of predictors. Breusch, T. S. and Pagan, A. R.1979.

ncvTest(fit.ols) # test for heteroscedasticity

Non-constant Variance Score Test. Chisquare = 5.75913, Df = 1, p = 0.01640319. The p-values suggestion to reject the hypothesis of homoscedasticity and we accept heteroscedasticity random errors. We expected this result because such as we designed simulated errors. Now we use weight least square method. Try fitting the regression with weights and see what the difference is.

fit.wls = lm(y~x, weights=1/(1+5*x), data=sim.data)

summary(fit.wls)

Coefficients (WLS regression):

		Estimate	Std. Error	t value	Pr(>|t|)
Intercept	β_0	19.5242	6.5710	2.971	0.00245**
	β_1	9.3218	0.9343	9.998	1.01e-10***

Residual standard error: 7.674 on 28 degrees of freedom. Multiple R-squared: 0.7805. Adjusted R-squared: 0.7726. F-statistic: 99.55 on 1 and 28 DF, p-value: 1.012e-10. We find the confidence intervals for the coefficients of the WLS regression using: **confint (fit.wls).**

We use the following commands to make the graphics of the regressions:

```
name <-c("True regression","OLS regression", "WLS regression")
colors.reg <- c("red", "blue", "black")
type.line <- c(1, 2, 4)
line.width <- c(3, 2, 4)
plot(x,y, main = "True regression, OLS regression and WLS regression",lwd=1, bty="o")
abline(a=20,b=10,col=colors.reg[1],lty = type.line[1],lwd =line.width [1])
abline(fit.ols$coefficients, col=colors.reg[2], lty = type.line[2], lwd =line.width[2])
abline(fit.wls$coefficients,col=colors.reg[3],lty = type.line[3],lwd =line.width[3])
legend ("topleft", legend = name, col = colors.reg, lty = type.line,lwd = line.width ,bty ="n")
```

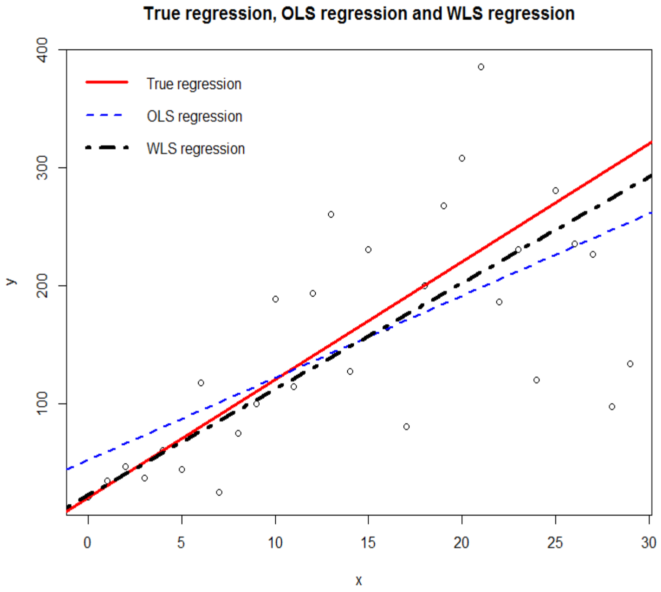

Figure 1 Truth regression, OLS regression and WLS regression for the simulate data.

Bootstrap method: re-sampling errors.

Another method for to estimating regression coefficients and standard deviation is the bootstrap method of re-sampling errors.

```
sim.lm<-glm(y~x,data=sim.data) #  the regression
sim.fit<-function(data) coef(glm(data$y~data$x))
sim.diag<-glm.diag.plots(sim.lm,ret=T) # Diagnostics plots
sim.res<-sim.diag$res*sim.diag$sd
sim.res<-sim.res-mean(sim.res)
sim.df<-data.frame(sim.data,res=sim.res,fit=fitted(sim.lm))
sim.model<-function(data,i){
            d<-data
            d$y<-d$fit+d$res[i]
            sim.fit(d)
}
fit.boot<-boot(sim.df,sim.model,R=9999)
```

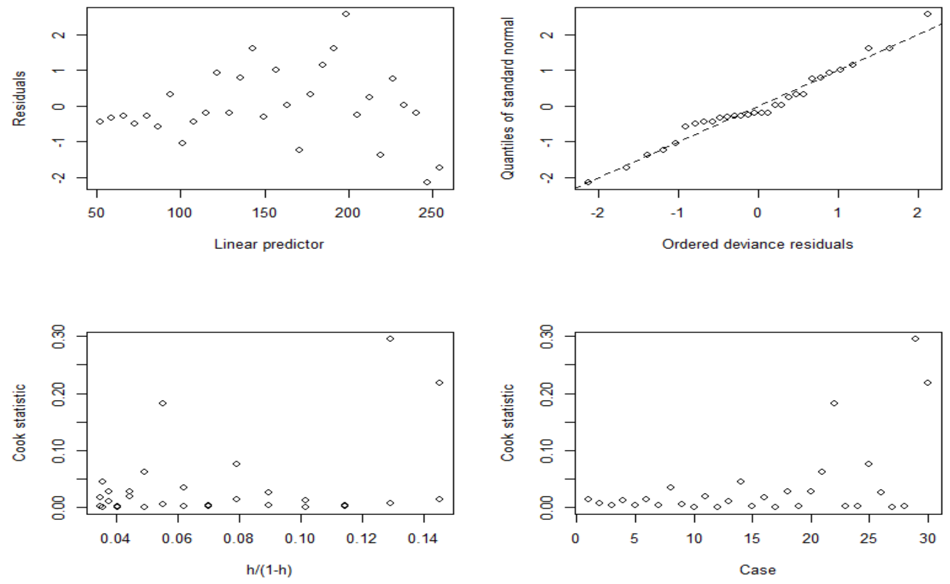

Figure 2 *Diagnostics plots for sim.lm regression.*

The plot on the top left is a plot of the jackknife deviance residuals against the fitted values. The plot on the top right is a normal QQ plot of the standardized deviance residuals. The dotted line is the expected line if the standardized residuals are normally distributed. The bottom two panels are plots of the Cook statistics. On the left is a plot of the Cook statistics against the standardized leverages. The final plot again shows the Cook statistic this time plotted against case number enabling us to find which observations are influential (Davison, A.C. and Snell, E.J., 1991).

summary (fit.boot)

```
Bootstrap Statistics (re-sampling errors without weight):
        R       Original          bias              std.error
1       9999    9.038062          0.01855348        26.3335
2       9999    10.045030         0.00809159        1.5606
```

Bootstrap percentile interval and adjusted bootstrap percentile (BCa) interval.

Calculations based on 9999 bootstrap replicates.

boot.ci(fit.boot, index = 1:min(2,length(fit.boot$t0)),type = c('perc', 'bca'))

```
        Level       Percentile              BCa

        95%         (-41.597, 62.66)        (-39.150, 65.582)
```

boot.ci(fit.boot, index = 2:min(2,length(fit.boot$t0)),type = c('perc','bca'))

```
        Level       Percentile              BCa

        95%         (6.98, 13.08)  (7.00, 13.11)
```

Bootstrap method: re-sampling case.

```
sim.fit<-function(data)coef(glm(data$y~data$x))
sim.case<-function(data,i)sim.fit(data[i,])
fit.boot.case<-boot(sim.data,sim.case,R=9999)
fit.boot.case
```

```
 Bootstrap Statistics: (re-sampling cases without weight):

        Original bias      std. error

 t1*    9.3481021    0.3387152    18.40179

 t2*    10.023602    -0.0594927   1.64067
```

Bootstrap method: weight error resamling.

```
plot(fit.boot.case,index=2,jack=T)
sim.lm.w<-glm(y~x,weight=1/(1+5*x),data=sim.data)
sim.fit.w<-function(data) coef(glm(data$y~data$x))
sim.diag.w<-glm.diag.plots(sim.lm.w,ret=T)
```

The last command makes plot of jackknife deviance residuals against linear predictor, normal scores plots of standardized deviance residuals, plot of approximate Cook statistics against $\frac{h_i}{1-h_i}$ where h_i are the leverages.

```
sim.res.w<-sim.res-mean(sim.res.w)
sim.df.w<-data.frame(sim.data,res=sim.res.w,fit=fitted(sim.lm.w))
sim.model<-function(data,i){ d<-data
d$y<-d$fit+d$res[i]
sim.fit(d)}
fit.boot.w<-boot(sim.df.w,sim.model,R=9999); fit.boot.w
```

```
 Bootstrap Statistics (re-sampling errors with weights):

        Original bias            std. error

 t1*    19.289300    -0.50796443   26.0353

 t2*    9.3432310    0.03159331    1.5364
```

We find bootstrap percentile interval and adjusted bootstrap percentile (BCa) interval only for the second coefficient of the regression fit.boot.w.

```
Level       Percentile                BCa
            95%        (6.318, 12.342)         (6.471, 12.518)
```

Bootstrap methods: weight errors and re-sampling case

```
sim.fit.w<-function(data)coef(glm(data$y~data$x,weight=1/(1+5*x)))
sim.case.w<-function(data,i)sim.fit(data[i,])
fit.boot.case.w<-boot(sim.data,sim.case,R=9999); summary(fit.boot.case.w)
```

Summary table: Estimate errors and standard deviations.

Method	Coefficients	Std. Error
True regression	β_0 =20	-
	β_1**=10**	-
OLS regression	$\hat{\beta}_0 = 9.348$	25.969
	$\hat{\beta}_1 =$ **10.024**	**1.538**
WLS regression	$\hat{\beta}_0 = 19.5242$	6.5710
	$\hat{\beta}_1 =$ **9.3218**	**0.9343**
Bootstrap method resampling errors without weight	$\hat{\beta}_0 = 9.038062$	26.3335
	$\hat{\beta}_1 =$ **10.045030**	**1.5606**
Bootstrap method resampling cases without weight	$\hat{\beta}_0 = 9.3481021$	18.40179
	$\hat{\beta}_1 =$ **10.023602**	**1.64067**
Bootstrap method resampling errors with weights	$\hat{\beta}_0 = 19.289300$	26.0353
	$\hat{\beta}_1 =$ **8.957616**	**1.5364**
Bootstrap method resampling case with weights	$\hat{\beta}_0 = 9.348102$	18.5540
	$\hat{\beta}_1 =$ **10.02364**	**1.6552**

Conclusions

For linear regression with normal random errors having constant variance, the least squares method of the coefficients estimators and standard deviation are accurate, but when the random errors having non constant variance, the estimators with this method are not accurate. The bootstrap methods improved quantile estimation. The WLS method is the most accurate where we know weights errors. Where we don't know completely the form of the variance of error, we can use bootstrap method resampling errors. We can use packages inside R software to get results.

References

[1] Bradley Efron and Robert J Tibshirani (1993): *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

[2] Breusch, T. S. and Pagan, A. R., (1979) *A simple test for heteroscedasticity and random coefficient variation*. Econometrica 47, 1287–1294.

[3] Cook, R. Dennis; Weisberg, Sanford (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall. ISBN 041224280X.

[4] D. A. Freedman and S. C. Peters: *Bootstrapping a Regression Equation: Some Empirical Results*. Jurnal of the American Statistical Association, Vol.79.385. (1984), pp: 97-106.

[5] Davison, A. C. and Hinkley, D. V. (1997). "Bootstrap Methods and Their Application". Cambridge University Press.

[6] Davison, A.C. and Snell, E.J., (1991). Residuals and diagnostics. In Statistical Theory and Modelling: In Honour of Sir David Cox D.V. Hinkley, N. Reid, and E.J. Snell (editors), 83–106. Chapman and Hall.

[7] P. J. Bickel and D. A. Freedman: *Asymptotic Normality and the bootstrap in stratified sampling*. The Annals of Statistics. 1984, Vol12, No2, pp: 470-482.

[8] R. D. Cook and S. Weisberg: *Diagnostics for heteroscedasticity in regression,* (1983). Oxford University Press, Biometrika, Vol.70, No.1, pp.1-10.

[9] Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, "Probability and Statistics for engineers and Scientists", 2002, pp. 400-425. Prentice Hall. ISBN 0-13-041529-4.

[10] White, Halbert (1980). *"A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity*". Econometrica. 48 (4): 817–838. doi:10.2307/1912934.