Regression Diagnostics

April 28, 2016

Contents

1 Model misspecification

If the model is not "correctly" specified, we encounter the problem of model specification error or model specification bias.

TYPES OF SPECIFICATION ERRORS

1. Omission of a relevant variable(s)
2. Inclusion of an unnecessary variable(s)
3. Adopting the wrong functional form
4. Errors of measurement
5. Incorrect specification of the stochastic error term

1.1 Including Irrelevant Variables in a Regression Model

Bais of estimator

One issue that we can dispense with fairly quickly is that of inclusion of an irrelevant variable or overspecifying the model in multiple regression analysis. This means that one (or more) of the independent variables is included in the model even though it has no partial effect on y in the population. (That is, its population coefficient is zero.)

To illustrate the issue, suppose we specify the model as

$$y_i = \beta_0 + \beta_1 x_{1i} + v_i$$

However, β_0 is no effect on y after x_1 have been controlled for, which means that $\beta_0 = 0$. Thus, the true model is

$$y_i = \beta_1 x_{1i} + u_i$$

In this situation, we have the OLS estimator

$$b_1 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{1i} - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x})y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x})^2}$$

Thus, the expectation of b_1 is

$$E(b_1) = E(\frac{\sum_{i=1}^{n}(x_{1i} - \bar{x})y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x})^2}) == E(\frac{\sum_{i=1}^{n}(x_{1i} - \bar{x})(\beta_1 x_{1i} + u_i)}{\sum_{i=1}^{n}(x_{1i} - \bar{x})^2}) == \beta_1 + \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x})E(u_i)}{\sum_{i=1}^{n}(x_{1i} - \bar{x})^2} = \beta_1$$

The conclusion of the preceding example is much more general: including one or more irrelevant variables in a multiple regression model, or overspecifying the model, does not affect the unbiasedness of the OLS estimators. Does this mean it is harmless to include irrelevant variables? No. Including irrelevant variables can have undesirable effects on the variances of the OLS estimators.

Variance of estimator

Now let us assume that

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

is the truth, but we fit the following model:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i$$

and thus commit the specification error of including an unnecessary variable in the model. The consequences of this specification error are as follows:

1. The OLS estimators of the parameters of the "incorrect" model are all unbiased and consistent, that is, $\mathrm{E}(\hat{\alpha}_1) = \beta_1$, $\mathrm{E}(\hat{\alpha}_2) = \beta_2$, and $\mathrm{E}(\hat{\alpha}_3) = \beta_3 = 0$.
2. The error variance σ^2 is correctly estimated.
3. The usual confidence interval and hypothesis-testing procedures remain valid.
4. However, the estimated α's will be generally inefficient, that is, their variances will be generally larger than those of the $\hat{\beta}$'s of the true model.

From the usual OLS formula we know that

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2}$$

and

$$\mathrm{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{23}^2)}$$

Therefore,

$$\frac{\mathrm{var}(\hat{\alpha}_2)}{\mathrm{var}(\hat{\beta}_2)} = \frac{1}{1 - r_{23}^2}$$

Since $0 \leq r_{23}^2 \leq 1$, it follows that var $(\hat{\alpha}_2) \geq$ var $(\hat{\beta}_2)$; that is, the variance of $\hat{\alpha}_2$ is generally greater than the variance of $\hat{\beta}_2$ even though, on average, $\hat{\alpha}_2 = \beta_2$ [i.e., $\mathrm{E}(\hat{\alpha}_2) = \beta_2$].

If we exclude a relevant variable, the coefficients of the variables retained in the model are generally biased as well as inconsistent, the error variance is incorrectly estimated, and the usual hypothesis-testing procedures become invalid. On the other hand, including an irrelevant variable in the model still gives us unbiased and consistent estimates of the coefficients in the true model, the error variance is correctly estimated, and the conventional hypothesis-testing methods are still valid; the only penalty we pay for the inclusion of the superfluous variable is that the estimated variances of the coefficients are larger, and as a result our probability inferences about the parameters are less precise. An unwanted conclusion here would be that it is better to include irrelevant variables than to omit the relevant ones. But this philosophy is not to be espoused because addition of unnecessary variables will lead to loss in efficiency of the estimators and may also lead to the problem of multicollinearity.

1.2 Underfitting of model (Omitting a relevant variable)

Now suppose that, rather than including an irrelevant variable, we omit a variable that actually belongs in the true (or population) model. This is often called the problem of excluding a relevant variable or underspecifying the model.

Deriving the bias caused by omitting an important variable is an example of misspecification analysis. We begin with the case where the true population model has two explanatory variables and an error term:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

However, we perform a simple regression of y on x_1 only, obtaining the equation

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i$$

The consequences of omitting variable X_3 are as follows:

1. If the left-out, or omitted, variable X_3 is correlated with the included variable X_2, that is, r_{23}, the correlation coefficient between the two variables, is nonzero, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are biased as well as inconsistent. That is, $E(\hat{\alpha}_1) \neq \beta_1$ and $E(\hat{\alpha}_2) \neq \beta_2$, and the bias does not disappear as the sample size gets larger.
2. Even if X_2 and X_3 are not correlated, $\hat{\alpha}_1$ is biased, although $\hat{\alpha}_2$ is now unbiased.

3. The disturbance variance σ^2 is incorrectly estimated.
4. The conventionally measured variance of $\hat{\alpha}_2$ $(= \sigma^2/\sum x_{2i}^2)$ is a biased estimator of the variance of the true estimator $\hat{\beta}_2$.
5. As another consequence, the forecasts based on the incorrect model and the forecast (confidence) intervals will be unreliable.

In this situation,

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha_2 x_{2i})^2$$

and

$$\frac{\partial RSS}{\partial \alpha_2} = -2 \sum_{i=1}^{n} (y_i - \alpha_2 x_{2i}) x_{2i} = 0$$

we have the OLS estimator of α_2 is

$$\hat{\alpha}_2 = \frac{\sum_{i=1}^{n} x_{2i} y_i}{\sum_{i=1}^{n} x_{2i}^2}$$

Thus, the expectation of $\hat{\alpha}_2$ is

$$\begin{aligned} E(\hat{\alpha}_2) = E(\frac{\sum_{i=1}^{n} x_{2i} y_i}{\sum_{i=1}^{n} x_{2i}^2}) &= E(\frac{\sum_{i=1}^{n} x_{2i}(\beta_1 x_{2i} + \beta_2 x_{3i} + u_i)}{\sum_{i=1}^{n} x_{2i}^2}) \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{2i} x_{3i}}{\sum_{i=1}^{n} x_{2i}^2} + \frac{\sum_{i=1}^{n} x_{2i} E(u_i)}{\sum_{i=1}^{n} x_{2i}^2} \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{3i} x_{2i}}{\sum_{i=1}^{n} x_{2i}^2} \end{aligned}$$

As aboved shows, $\hat{\alpha}_2$ is biased, unless β_2 or the second term or both are zero. We rule out β_3 being zero, because in that case we do not have specification error to begin with. The second term will be zero if X_2 and X_3 are uncorrelated, which is unlikely in most economic data.
Now let us examine the variances of $\hat{\alpha}_2$ and $\hat{\beta}_2$

$$\begin{aligned} \text{var}(\hat{\alpha}_2) &= \frac{\sigma^2}{\sum x_{2i}^2} \\ \text{var}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} = \frac{\sigma^2}{\sum x_{2i}^2} \text{VIF} \end{aligned}$$

where VIF (a measure of collinearity) is the variance inflation factor $[= 1/(1 - r_{23}^2)]$ and r_{23} is the correlation coefficient between variables X_2 and X_3. In general, var $(\hat{\alpha}_2)$ will be different from var $(\hat{\beta}_2)$. But we know that var $(\hat{\beta}_2)$ is unbiased, therefore, var $(\hat{\alpha}_2)$ is biased. Since $0 < r_{23}^2 < 1$, it would seem that in the present case var $(\hat{\alpha}_2)$ < var $(\hat{\beta}_2)$. So, there is a tradeoff involved here. However, since $\hat{\sigma}^2 = RSS/df$, which depends on the number of regressors included in the model as well as the df (= n, number of parameters estimated). Now if we add variables to the model, the RSS generally decreases (recall that as more variables are added to the model, the R^2 increases), but the degrees of freedom also decrease because more parameters are estimated. The net outcome depends on whether the RSS decreases sufficiently to offset the loss of degrees of freedom due to the addition of regressors. It is quite possible that if a regressor has a strong impact on the regressand—for example, it may reduce RSS more than the loss in degrees of freedom as a result of its addition to the model—inclusion of such variables will not only reduce the bias but will also increase precision (i.e., reduce standard errors) of the estimators.

1.3 Ramsey's RESET

Ramsey's RESET (regression specification error test; Ramsey 1969) takes powers of the fitted values ^y and tests whether they have a significant influence when added to the regression model. Alternatively, powers of the original regressors or of the first principal component of X can be used. All three versions are implemented in the function resettest(). It defaults to using second and third powers of the fitted values as auxiliary variables. With only one real regressor in the model matrix X (excluding the intercept), all three strategies yield equivalent results. Hence we use

install.packages("lmtest")

```
library("lmtest")

## Loading required package:  zoo
##
## Attaching package:  'zoo'
## The following objects are masked from 'package:base':
##
##    as.Date, as.Date.numeric

x <- c(1:30)
y1 <- 1 + x + x^2 + rnorm(30)
y2 <- 1 + x + rnorm(30)
resettest(y1 ~ x , power=2, type="regressor")

##
##  RESET test
##
## data:  y1 ~ x
## RESET = 148580, df1 = 1, df2 = 27, p-value < 2.2e-16

resettest(y2 ~ x , power=2, type="regressor")

##
##  RESET test
##
## data:  y2 ~ x
## RESET = 2.1008, df1 = 1, df2 = 27, p-value = 0.1587
```

2 Heteroscedasticity

2.1 The problem of heteroscedasticity

An important assumption of the classical linear regression model (Assumption 4) is that the disturbances ui appearing in the population regression function are homoscedastic; Cross-sectional data are often plagued by the problem of heteroscedasticity; that is, they all have the same variance. Symbolically,

$$E(u_i^2) = \sigma^2$$

Diagrammatically, in the two-variable regression model heteroscedastic can be shown as in the following figure.

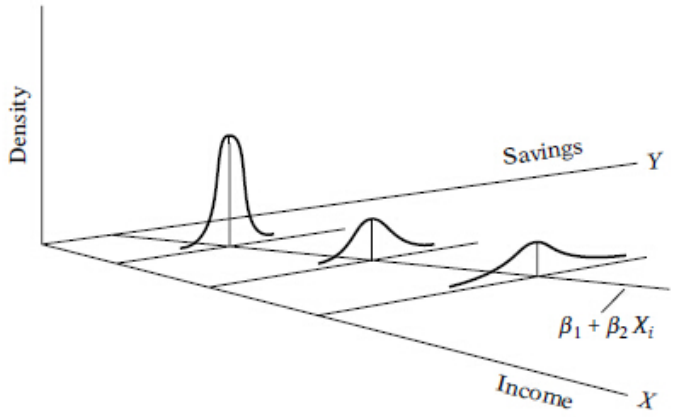


Figure 1: Heteroscedastic disturbances

In contrast, consider aboved figure, which shows that the conditional variance of Y_i increases as X increases. Here, the variances of Y_i are not the same. Hence, there is heteroscedasticity. Symbolically,

$$E(u_i^2) = \sigma_i^2$$

Notice the subscript of σ^2, which reminds us that the conditional variances of u_i (= conditional variances of Y_i) are no longer constant. Consider the following model:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Applying the usual formula, the OLS estimator of β_2 is

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

but its variance is now given by the following expression

$$\text{var}(\hat{\beta}_2) = \frac{\sum(X_i^2 - \bar{X})^2\sigma_i^2}{(\sum(X_i^2 - \bar{X})^2)^2}$$

which is obviously different from the usual variance formula obtained under the assumption of homoscedasticity, namely,

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum(X_i^2 - \bar{X})^2}$$

It is easy to prove that $\hat{\beta}_2$ is still linear and unbiased. However, in this situation, $\hat{\beta}_2$ is no longer best and the minimum variance, namely, $\hat{\beta}_2$ is not the most efficient estimator in unbiased estimator.

2.2 Detection of heteroscedasticity

The first important practical question is: How does one know that heteroscedasticity is present in a specific situation?

2.2.1 Graphical Method

If there is no a priori or empirical information about the nature of heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no heteroscedasticity and then do a postmortem examination of the residual squared $\hat{u}_i^2$ to see if they exhibit any systematic pattern.

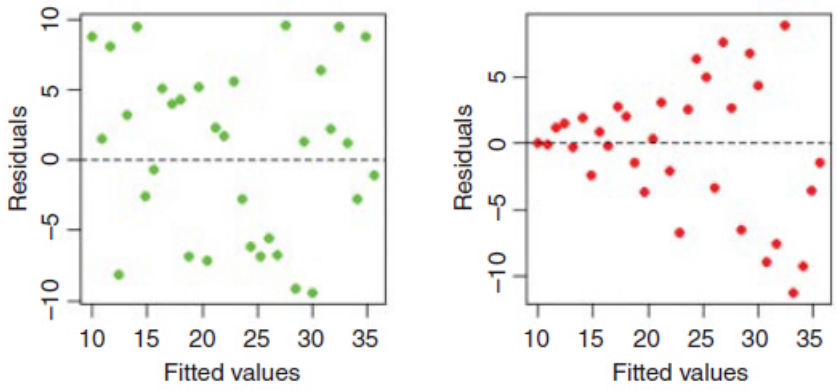


Figure 2: Heteroscedasticity

2.2.2 Goldfeld-Quandt Test

We consider the following model:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Suppose σ_i^2 is positively related to X_i as

$$\sigma_i^2 = \sigma^2 X_i^2 \tag{1}$$

where σ^2 is a constant. The estimator of variance for u_i is

$$\hat{\sigma}_u^2 = \frac{\sum u_i^2}{n-k-1}$$

If 1 is appropriate, it would mean σ_i^2 would be larger, the larger the values of X_i. If u_i are assumed to be normally distributed and there is no autocorrelation, Goldfeld and Quandt suggest the following steps:

$$H_0 : u_i \text{ is homoscedastic}; H_A : u_i \text{ is heteroscedastic}$$

- Step 1. Order or rank the observations according to the values of X_i, beginning with the lowest X value.
- Step 2. Omit c central observations, where c is specified a priori, and divide the remaining $(n-c)$ observations into two groups each of $(n-c)/2$ observations.
- Step 3. Fit separate OLS regressions to the first $(n-c)/2$ observations and the last $(n-c)/2$ observations, and obtain the respective residual sums of squares RSS1 and RSS2, RSS1 representing the RSS from the regression corresponding to the smaller X_i values (the small variance group) and RSS2 that from the larger X_i values (the large variance group).
- Step 4. Compute the ratio

$$F = \frac{\sum u_{1i}^2/(\frac{(n-c)}{2} - k)}{\sum u_{2i}^2/(\frac{n-c-2k}{2})} \qquad \sim F(\frac{(n-c)}{2} - k, \frac{n-c-2k}{2})$$

If in an application the computed $\lambda(= F)$ is greater than the critical F at the chosen level of significance, we can reject the hypothesis of homoscedasticity, that is, we can say that heteroscedasticity is very likely. But the ability of the Goldfeld–Quandt test to do this successfully depends on how c is chosen. For the two-variable model the Monte Carlo experiments done by Goldfeld and Quandt suggest that c is about 8 if the sample size is about 30, and it is about 16 if the sample size is about 60.

2.2.3 White's General Heteroscedasticity Test

White's General Heteroscedasticity Test does not rely on the normality assumption and is easy to implement. Consider the following three-variable regression model:

$$y_i = \beta_0 + \beta_1 x_{2i} + \beta_2 x_{3i} + u_i \tag{2}$$

The White test proceeds as follows:

- Step 1. Given the data, we estimate 2 and obtain the residuals, $\hat{u}_i$.
- Step 2. We then run the following (auxiliary) regression:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + \nu_i$$

- Step 3. Under the null hypothesis that there is no heteroscedasticity, it can be shown that sample size (n) times the R^2 obtained from the auxiliary regression asymptotically follows the chi-square distribution with df equal to the number of regressors (excluding the constant term) in the auxiliary regression.

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$$

$$nR^2 \sim \chi^2_{\text{df}} \tag{3}$$

where df is as defined previously. In our example, there are 5 df since there are 5 regressors in the auxiliary regression.

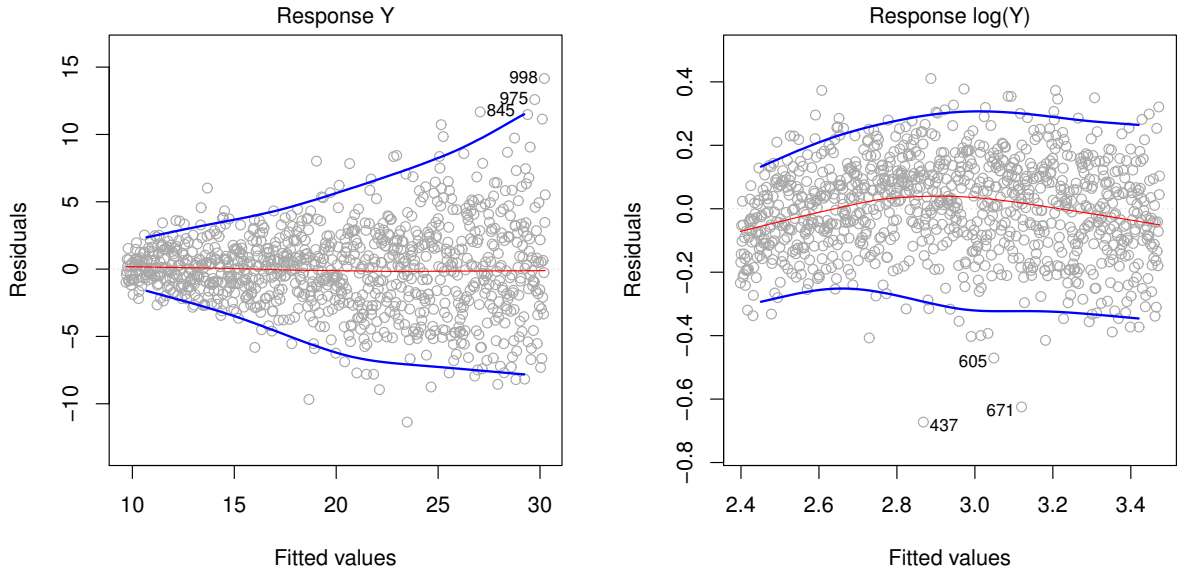
- Step 4. If the chi-square value obtained in 3 exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroscedasticity. If it does not exceed the critical chi-square value, there is no heteroscedasticity.

Other test methods: Breusch and Pagan Test; Godfrey and Koenker Test; Glejser Test; Park Test; Spearman Test.

2.3 Remedial measures

2.3.1 Using a concave function such as log Y

An example is shown in the left-hand panel of the following Figure, in which the magnitude of the residuals tends to increase with the fitted values. When faced with this problem, one possible solution is to transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity. The right-hand panel of the Figure displays the residual plot after transforming the response using $\log Y$. The residuals now appear to have constant variance, though there is some evidence of a slight non-linear relationship in the data.



As we have seen, heteroscedasticity does not destroy the unbiasedness and consistency properties of the OLS estimators, but they are no longer efficient, not even asymptotically (i.e., large sample size). This lack of efficiency makes the usual hypothesis-testing procedure of dubious value. Therefore, remedial measures may be called for: There are two approaches to remediation: when σ_i^2 is known and when σ_i^2 is not known.

2.3.2 The method of generalized least squares (GLS)

We consider the following model:

$$y = X\beta + u \tag{4}$$

and $E(u) = 0$, $E(uu') = \Sigma = \sigma^2\Omega$, Ω is $n \times n$ positive matrix.

$$\Omega = \begin{bmatrix} \omega_{11} & 0 & \dots & 0 \\ 0 & \omega_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \omega_{mn} \end{bmatrix}$$

which trace is

$$tr(\Omega) = n$$

Since Ω is $n \times n$ positive matrix, we have

$$P\Omega P' = I \text{ and } P'P = \Omega^{-1} \tag{5}$$

From 4 and 5, we have

$$Py = PX\beta + Pu$$

Notate $y^* = Py$, $X^* = PX$, $u^* = Pu$, we obtain

$$y^* = X^*\beta + u^*$$

The expectation of $u^*$is

$$E(u^*) = E(Pu) = PE(u) = 0$$

The variance is

$$E(u^* u^{*'}) = E(Puu'P') = PE(uu')P' = P\sigma^2 \Omega P' = \sigma^2 I$$

We obtain the GLS estimator:

$$\begin{aligned} \hat{\beta} &= (X^{*'} X^*)^{-1} X^* y^* \\ &= (X'P'PX)^{-1} X'P'Py \\ &= (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} y \end{aligned}$$

The variance of GLS estimator:

$$\begin{aligned} V(\hat{\beta}) &= \sigma^2 (X^{*'} X^*)^{-1} \\ &= \sigma^2 (X'P'PX)^{-1} \\ &= \sigma^2 (X'\Omega^{-1}X)^{-1} = (X'\Sigma^{-1}X)^{-1} \end{aligned}$$

2.3.3 The Method of Weighted Least Squares

Let us continue with the now-familiar two-variable model:

$$y_i = \beta_0 x_{0i} + \beta_1 x_i + u_i$$

where $X_{0i} = 1$ for each i. Now assume that the heteroscedastic variances σ_i^2 are known. We have

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{x_{0i}}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i}$$

We obtain

$$\frac{y_i}{\sigma_i} = \hat{\beta}_0^* \frac{x_{0i}}{\sigma_i} + \hat{\beta}_1^* \frac{x_i}{\sigma_i} + \frac{\hat{u}_i}{\sigma_i}$$

Now, to obtain the WLS estimators, we minimize

$$\sum (\frac{\hat{u}_i}{\sigma_i})^2 = \sum [(\frac{y_i}{\sigma_i}) - \hat{\beta}_0^* (\frac{x_{0i}}{\sigma_i}) - \hat{\beta}_1^* (\frac{x_i}{\sigma_i})]^2$$

As shown there, the WLS estimator of β_2^* is

$$\hat{\beta}_2^* = \frac{(\sum w_i)(\sum w_i x_i y_i) - (\sum w_i x_i)(\sum w_i y_i)}{(\sum w_i)(\sum w_i x_i^2) - (\sum w_i x_i)^2}$$

and its variance is given by

$$\text{var}(\hat{\beta}_2) = \frac{\sum w_i}{(\sum w_i)(\sum w_i x_i^2) - (\sum w_i x_i)^2}$$

where $w_i = 1/\sigma_i^2$.

Thus, in GLS we minimize a weighted sum of residual squares with $w_i = 1/\sigma_i^2$ acting as the weights, but in OLS we minimize an unweighted or (what amounts to the same thing) equally weighted RSS.

2.3.4 White's Heteroscedasticity-Consistent Variances and Standard Errors

White has shown that this estimate can be performed so that asymptotically valid (i.e., large-sample) statistical inferences can be made about the true parameter values. Incidentally, White's heteroscedasticity-corrected standard errors are also known as robust standard errors.
In this case, the variance of β_2 is now given by the following expression

$$\text{var}(\hat{\beta}_2) = \frac{\sum(X_i^2 - \bar{X})^2\sigma_i^2}{(\sum(X_i^2 - \bar{X})^2)^2}$$

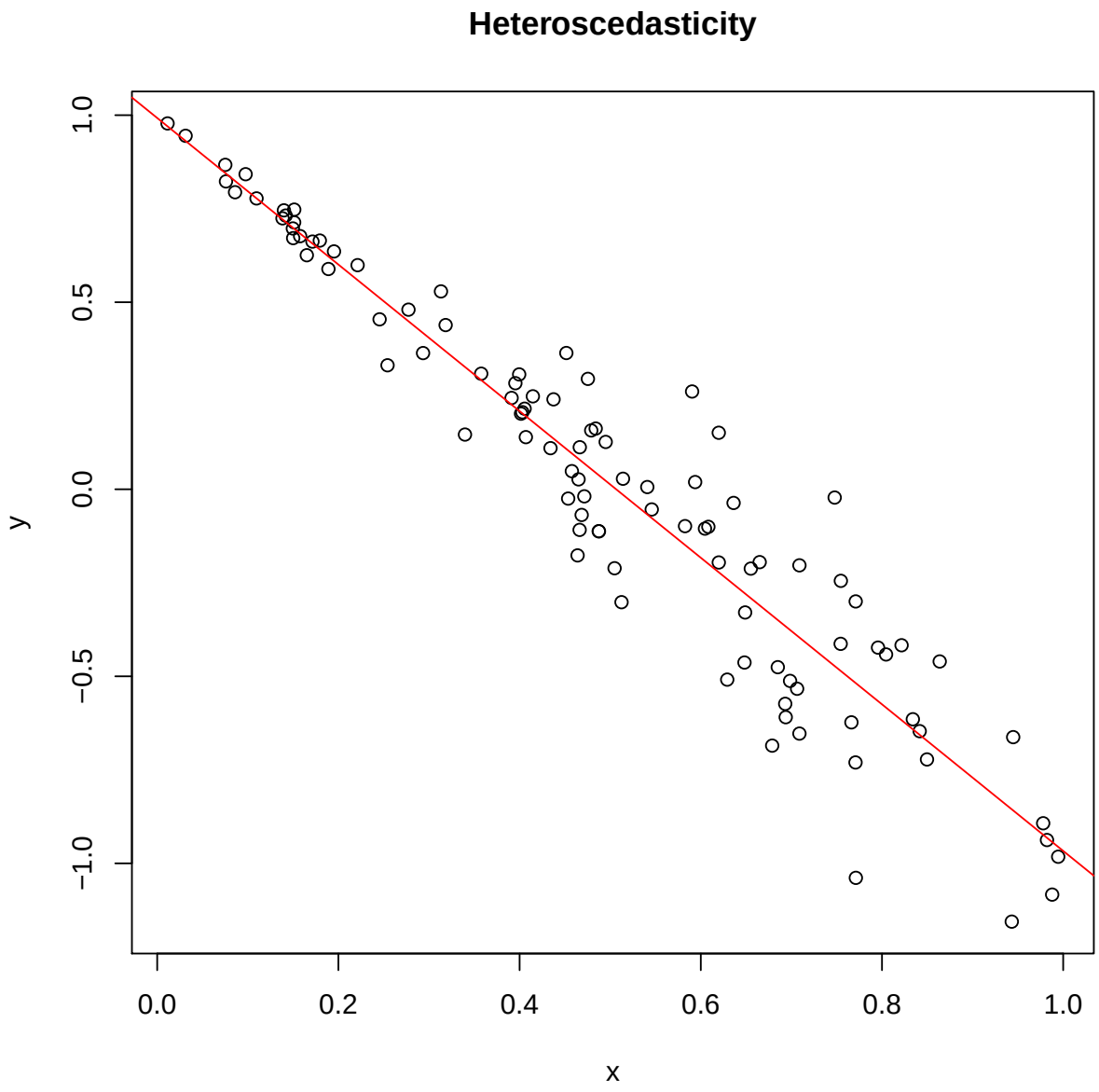
Since there is an unknown parameter σ_i, we use the $\hat{u}_i(= y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$ instead of σ. Then we have Heteroscedasticity Consistent Standard Error (HCSE)

$$HCSE = \sqrt{\frac{\sum(X_i^2 - \bar{X})^2\hat{u}_i^2}{(\sum(X_i^2 - \bar{X})^2)^2}}$$

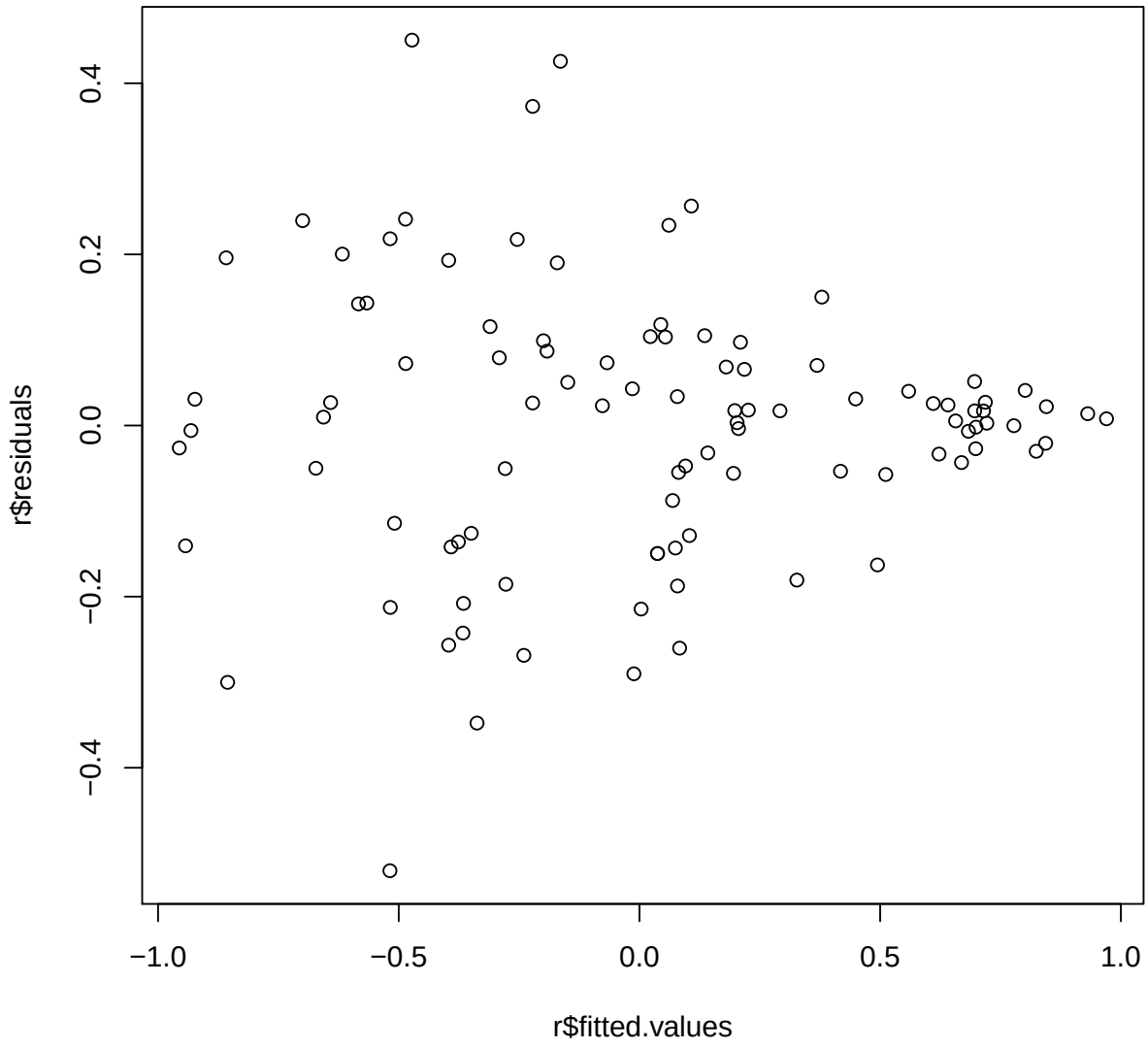
we can use OLS but correct the standard errors for autocorrelation by a procedure developed by Newey and West. This is an extension of White's heteroscedasticity-consistent standard errors that we discussed in the previous chapter. The corrected standard errors are known as HAC (heteroscedasticity- and autocorrelation-consistent) standard errors or simply as Newey–West standard errors.

2.4 An Example

```
library(lmtest)   #gqtest
x <- runif(100)
y <- 1 - 2*x + .3*x*rnorm(100)
plot(y~x)
r <- lm(y~x)
abline(r, col="red")
title(main="Heteroscedasticity")
```




```
plot(r$residuals ~ r$fitted.values)
```




```
gqtest(r)
```

```
##
##  Goldfeld-Quandt test
##
## data:  r
## GQ = 1.0935, df1 = 48, df2 = 48, p-value = 0.379
```

```
bptest(r)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  r
## BP = 12.5464, df = 1, p-value = 0.000397
```

```r
#white test
aux <- r$residuals^2
aux_lm <- lm(aux ~ x + I(x^2))
pchisq(length(x)*summary(aux_lm)$r.squared,df=2,lower.tail=F)
```

```
## [1] 0.0008080956
```

```r
#HCSE
library(car)
hccm(r)
```

```
##               (Intercept)            x
## (Intercept)  0.0003859213 -0.000896916
## x           -0.0008969160  0.003101375
```

```r
#NeweyWest
library(sandwich)
NeweyWest(r)
```

```
##               (Intercept)             x
## (Intercept)  0.0002776838 -0.0006441432
## x           -0.0006441432  0.0024447774
```

```r
nerlove<-read.csv("C:\\Users\\XXXHHF\\Documents\\R\\workfile\\#R lecture\\11\\nerlove.csv")
head(nerlove)
```

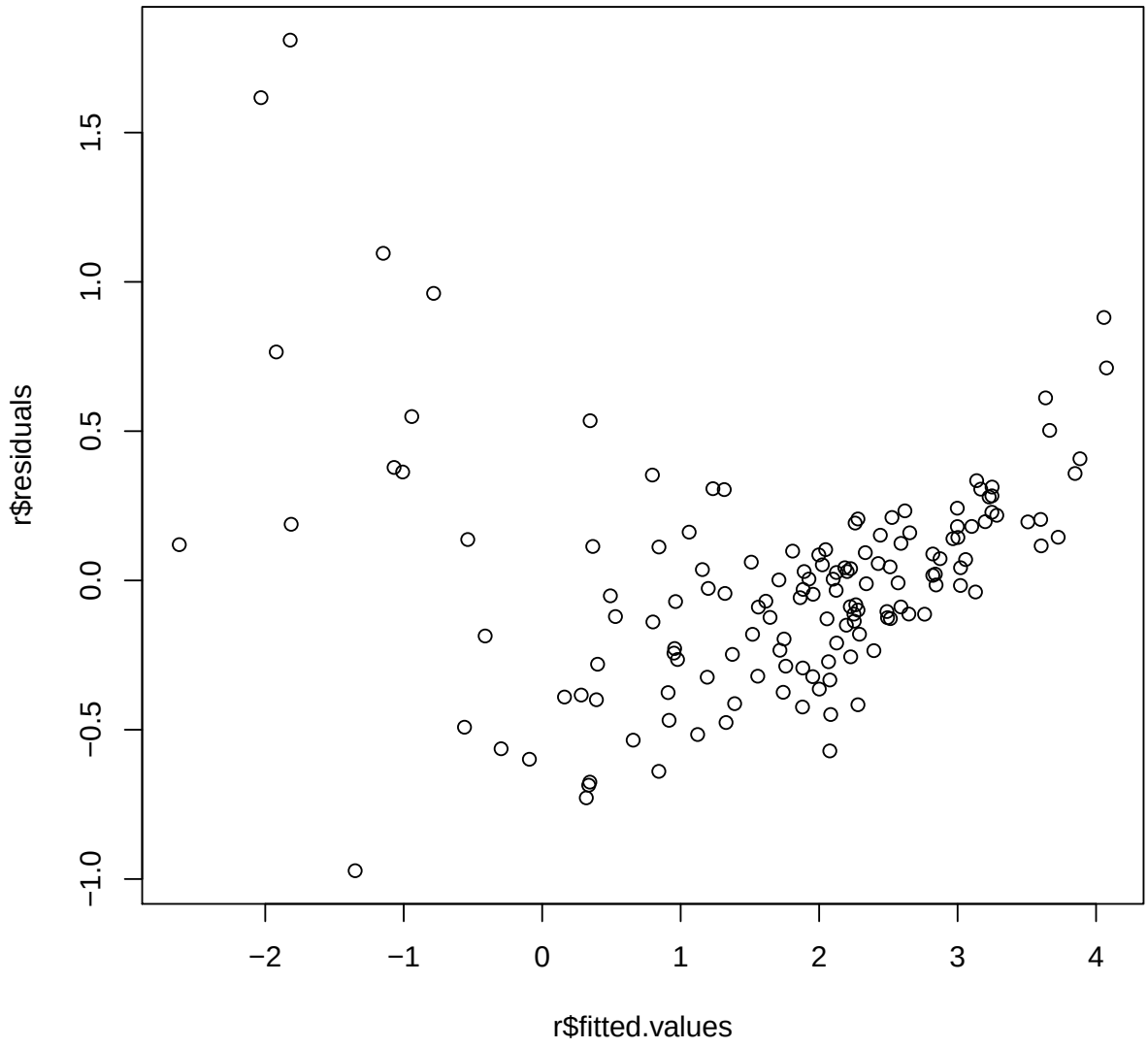
```
##      tc q  pl   pf  pk       lntc       lnq      lnpl     lnpf     lnpk
## 1 0.082 2 2.1 17.9 183 -2.5010360 0.6931472 0.7419373 2.884801 5.209486
## 2 0.661 3 2.1 35.1 174 -0.4140014 1.0986120 0.7419373 3.558201 5.159055
## 3 0.990 4 2.1 35.1 171 -0.0100503 1.3862940 0.7419373 3.558201 5.141664
## 4 0.315 4 1.8 32.2 166 -1.1551830 1.3862940 0.5877866 3.471967 5.111988
## 5 0.197 5 2.1 28.6 233 -1.6245520 1.6094380 0.7419373 3.353407 5.451038
## 6 0.098 9 2.1 28.6 195 -2.3227880 2.1972250 0.7419373 3.353407 5.273000
##   lntc_q_pf   lnpl_pf  lnpk_pf
## 1 -6.078984 -2.142863 2.324685
## 2 -5.070815 -2.816264 1.600854
## 3 -4.954546 -2.816264 1.583462
## 4 -6.013443 -2.884180 1.640021
## 5 -6.587396 -2.611469 2.097632
## 6 -7.873419 -2.611469 1.919593
```

```r
r<-lm(lntc~lnq+lnpl+lnpk+lnpf,data=nerlove)
summary(r)
```

```
##
## Call:
## lm(formula = lntc ~ lnq + lnpl + lnpk + lnpf, data = nerlove)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97203 -0.23377 -0.01091  0.16185  1.80985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.56651    1.77938  -2.004    0.047 *
```

```
## lnq          0.72091    0.01743  41.352  < 2e-16 ***
## lnpl         0.45596    0.29980   1.521    0.131
## lnpk        -0.21515    0.33983  -0.633    0.528
## lnpf         0.42581    0.10032   4.244 3.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3923 on 140 degrees of freedom
## Multiple R-squared:  0.926,Adjusted R-squared:  0.9239
## F-statistic: 437.9 on 4 and 140 DF,  p-value: < 2.2e-16
```

```
plot(r$residuals ~ r$fitted.values)
```



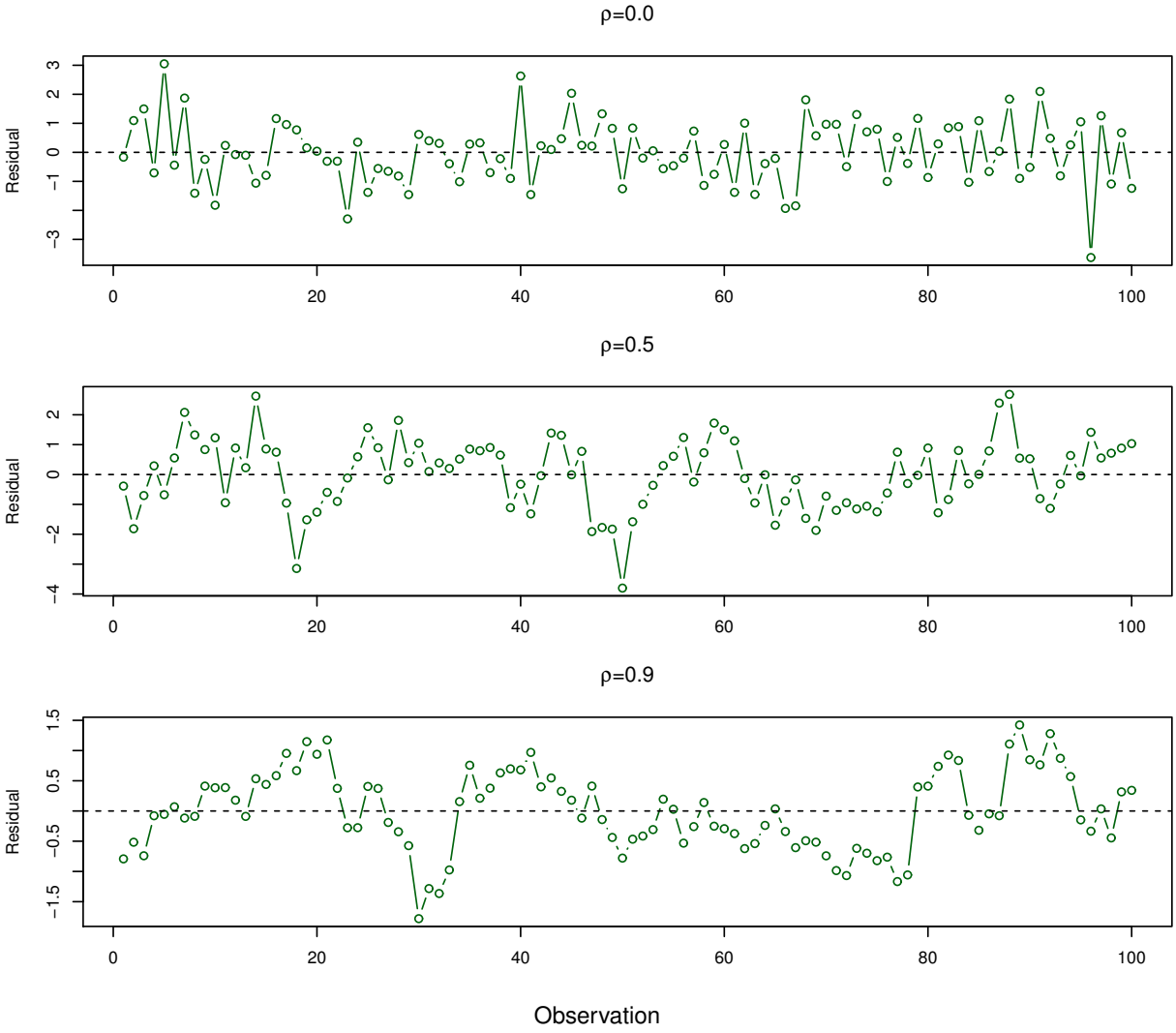
```
library(lmtest)
bptest(r)

##
##  studentized Breusch-Pagan test
##
## data:  r
## BP = 36.1635, df = 4, p-value = 2.678e-07

#white test
aux <- r$residuals^2
aux_lm <- lm(aux ~ (lnq+lnpl+lnpk+lnpf)^2+I(lnq^2)+I(lnpl^2)
             +I(lnpk^2)+I(lnpf^2),data=nerlove)
summary(aux_lm)

##
## Call:
## lm(formula = aux ~ (lnq + lnpl + lnpk + lnpf)^2 + I(lnq^2) +
##     I(lnpl^2) + I(lnpk^2) + I(lnpf^2), data = nerlove)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91988 -0.08957 -0.00977  0.07473  1.84624
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -53.318407  55.141552  -0.967  0.33537
## lnq          -1.071327   0.658943  -1.626  0.10641
## lnpl          8.292634  11.255178   0.737  0.46258
## lnpk         24.839592  20.394057   1.218  0.22544
## lnpf         -6.380571   5.340893  -1.195  0.23439
## I(lnq^2)      0.032475   0.005213   6.230    6e-09 ***
## I(lnpl^2)    -0.357911   1.838472  -0.195  0.84595
## I(lnpk^2)    -2.892902   1.927122  -1.501  0.13574
## I(lnpf^2)    -0.217969   0.185446  -1.175  0.24199
## lnq:lnpl     -0.007767   0.137959  -0.056  0.95519
## lnq:lnpk      0.213300   0.122527   1.741  0.08408 .
## lnq:lnpf     -0.150873   0.046131  -3.271  0.00137 **
## lnpl:lnpk    -2.087954   2.124647  -0.983  0.32757
## lnpl:lnpf     1.036419   0.878694   1.179  0.24035
## lnpk:lnpf     1.579700   1.061430   1.488  0.13910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2829 on 130 degrees of freedom
## Multiple R-squared:  0.5095,Adjusted R-squared:  0.4567
## F-statistic: 9.645 on 14 and 130 DF,  p-value: 2.136e-14

pchisq(length(nerlove$lnq)*summary(aux_lm)$r.squared,df=14,lower.tail=F)

## [1] 3.800068e-10

#NeweyWest
library(car)
linearHypothesis(r, "lnpk = 0", vcov=NeweyWest(r))
```

```
## Linear hypothesis test
##
## Hypothesis:
## lnpk = 0
##
## Model 1: restricted model
## Model 2: lntc ~ lnq + lnpl + lnpk + lnpf
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1    141
## 2    140  1 0.4574    0.5

#
summary(r)

##
## Call:
## lm(formula = lntc ~ lnq + lnpl + lnpk + lnpf, data = nerlove)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97203 -0.23377 -0.01091  0.16185  1.80985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.56651    1.77938  -2.004    0.047 *
## lnq          0.72091    0.01743  41.352  < 2e-16 ***
## lnpl         0.45596    0.29980   1.521    0.131
## lnpk        -0.21515    0.33983  -0.633    0.528
## lnpf         0.42581    0.10032   4.244 3.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3923 on 140 degrees of freedom
## Multiple R-squared:  0.926,Adjusted R-squared:  0.9239
## F-statistic: 437.9 on 4 and 140 DF,  p-value: < 2.2e-16

names(r)

##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"

r$coefficients

## (Intercept)         lnq        lnpl        lnpk        lnpf
##  -3.5665130   0.7209135   0.4559638  -0.2151477   0.4258140

library(sandwich)
vv<-NeweyWest(r)
(r$coefficients[4]/sqrt(vv[4,4]))^2

##      lnpk
## 0.4573648
```

```
#coeftest
coeftest(r, vcov = NeweyWest(r))

##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -3.566513   1.948686 -1.8302   0.06934 .
## lnq          0.720914   0.066717 10.8056 < 2.2e-16 ***
## lnpl         0.455964   0.275933  1.6524   0.10069
## lnpk        -0.215148   0.318131 -0.6763   0.49997
## lnpf         0.425814   0.083301  5.1118 1.029e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

sqrt(vv[4,4])

## [1] 0.3181305
```

3 Autocorrelation

Under both heteroscedasticity and autocorrelation the usual OLS estimators, although linear, unbiased, and asymptotically (i.e., in large samples) normally distributed, are no longer minimum variance among all linear unbiased estimators. In short, they are not efficient relative to other linear and unbiased estimators. Put differently, they may not be BLUE. As a result, the usual, t, F, and χ^2 may not be valid.



3.1 The problem of autocorrelation

The term autocorrelation may be defined as "correlation between members of series of observations ordered in time. In the regression context, the classical linear regression model assumes that such autocorrelation does not exist in the disturbances u_i. Symbolically,

$$E(u_i u_j) = 0, \qquad i \neq j$$

However, if there is such a dependence, we have autocorrelation. Symbolically,

$$E(u_i u_j) \neq 0, \qquad i \neq j$$

Specification Bias: Excluded Variables Case

In empirical analysis the researcher often starts with a plausible regression model that may not be the most "perfect" one. For example, suppose we have the following demand model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t \tag{6}$$

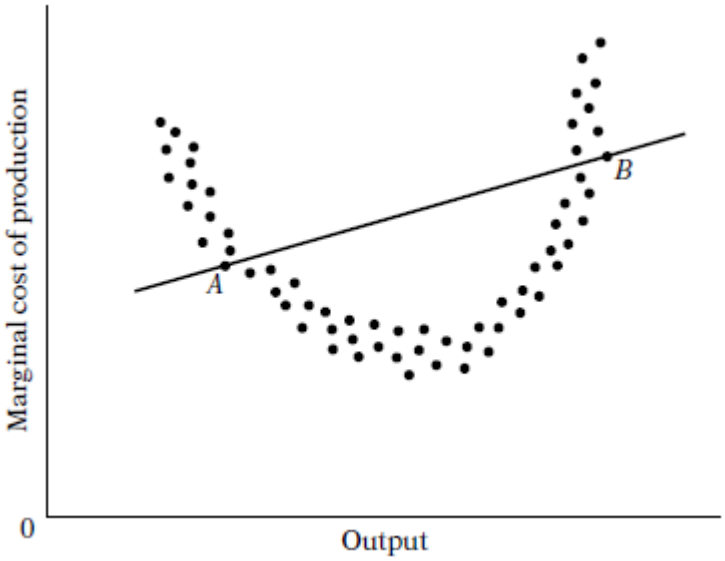


Figure 3: Specification bias: incorrect functional form.

where Y = quantity of beef demanded, X_2 = price of beef, X_3 = consumer income, X_4 = price of pork, and t = time. However, for some reason we run the following regression:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \tag{7}$$

Now if 7 is the "correct" model or the "truth" or true relation, running 6 is tantamount to letting $v_t = \beta_4 X_{4t} + u_t$. And to the extent the price of pork affects the consumption of beef, the error or disturbance term v_i will reflect a systematic pattern, thus creating (false) autocorrelation.

Specification Bias: Incorrect Functional Form

Suppose the "true" or correct model in a cost-output study is as follows:

$$\text{Marginal cost}_i = \beta_1 + \beta_2 \text{output}_i + \beta_3 \text{output}_i^2 + u_i$$

but we fit the following model:

$$\text{Marginal cost}_i = \beta_1 + \beta_2 \text{output}_i + u_i$$

This result is to be expected, because the disturbance term v_i is, in fact, equal to output$^2 + u_i$, and hence will catch the systematic effect of the output^2 term on marginal cost. In this case, v_i will reflect autocorrelation because of the use of an incorrect functional form.

Lags

In a time series regression of consumption expenditure on income, it is not uncommon to find that the consumption expenditure in the current period depends, among other things, on the consumption expenditure of the previous period.

$$\text{Consumption}_t = \beta_1 + \beta_2 \text{income}_t + \beta_3 \text{Consumption}_{t-1} + u_t \tag{8}$$

Consumers do not change their consumption habits readily for psychological, technological, or institutional reasons. Now if we neglect the lagged term in 8, the resulting error term will reflect a systematic pattern due to the influence of lagged consumption on current consumption.

"Manipulation" of Data

In empirical analysis, the raw data are often manipulated. For example

$$c_t = \beta_0 + \beta_1 y_t + u_t, \quad t = 1, 2, \ldots, T$$

where c is consumption, y is income. We calculate the average of four quarterly as

$$\begin{aligned} c_t^s &= (c_t + c_{t-1} + c_{t-2} + c_{t-3})/4 \\ y_t^s &= (y_t + y_{t-1} + y_{t-2} + y_{t-3})/4 \end{aligned}$$

Then, we have

$$\begin{aligned} c_t^s &= (\beta_0 + \beta_1 y_t + u_t)/4 + (\beta_0 + \beta_1 y_{t-1} + u_{t-1})/4 \\ &\quad +(\beta_0 + \beta_1 y_{t-2} + u_{t-2})/4 + (\beta_0 + \beta_1 y_{t-3} + u_{t-3})/4 \\ &= \beta_0 + \beta_1 (y_t + y_{t-1} + y_{t-2} + y_{t-3})/4 + (u_t + u_{t-1} + u_{t-2} + u_{t-3})/4 \\ &= \beta_0 + \beta_1 y_t^s + v_t \end{aligned}$$

where

$$v_t = (u_t + u_{t-1} + u_{t-2} + u_{t-3})/4$$

Then, we have a autocorrelation problem:

$$E(v_t) = 0, \quad E(v_t^2) = (1/4)\sigma^2, \quad E(v_t v_{t-s}) \begin{cases} (3/16)\sigma^2 & s = 1 \\ (1/8)\sigma^2 & s = 2 \\ (1/16)\sigma^2 & s = 3 \\ 0 & s > 3 \end{cases}$$

3.2 Detecting autocorrelation

The wage_2 gives data on indexes of real compensation per hour (Y) and output per hour (X) in the business sector of the U.S. economy for the period 1959–1998, the base of the indexes being 1992 = 100. Since the relationship between real compensation and labor productivity is expected to be positive, it is not surprising that the two variables are positively related. What is surprising is that the relationship between the two is almost linear, although there is some hint that at higher values of productivity the relationship between the two may be slightly nonlinear. Therefore, we decided to estimate a linear as well as a log–linear model, with the following model:

$$Y = \beta_0 + \beta_1 X_t + u_t$$

```
wage<-read.csv("C:\\Users\\XXXHHF\\Documents\\R\\workfile\\#R lecture\\11\\wage_2.csv",header=T)
head(wage)
```

```
##    obs    Y    X
## 1 1959 58.5 47.2
## 2 1960 59.9 48.0
## 3 1961 61.7 49.8
## 4 1962 63.9 52.1
## 5 1963 65.3 54.1
## 6 1964 67.8 56.6
```

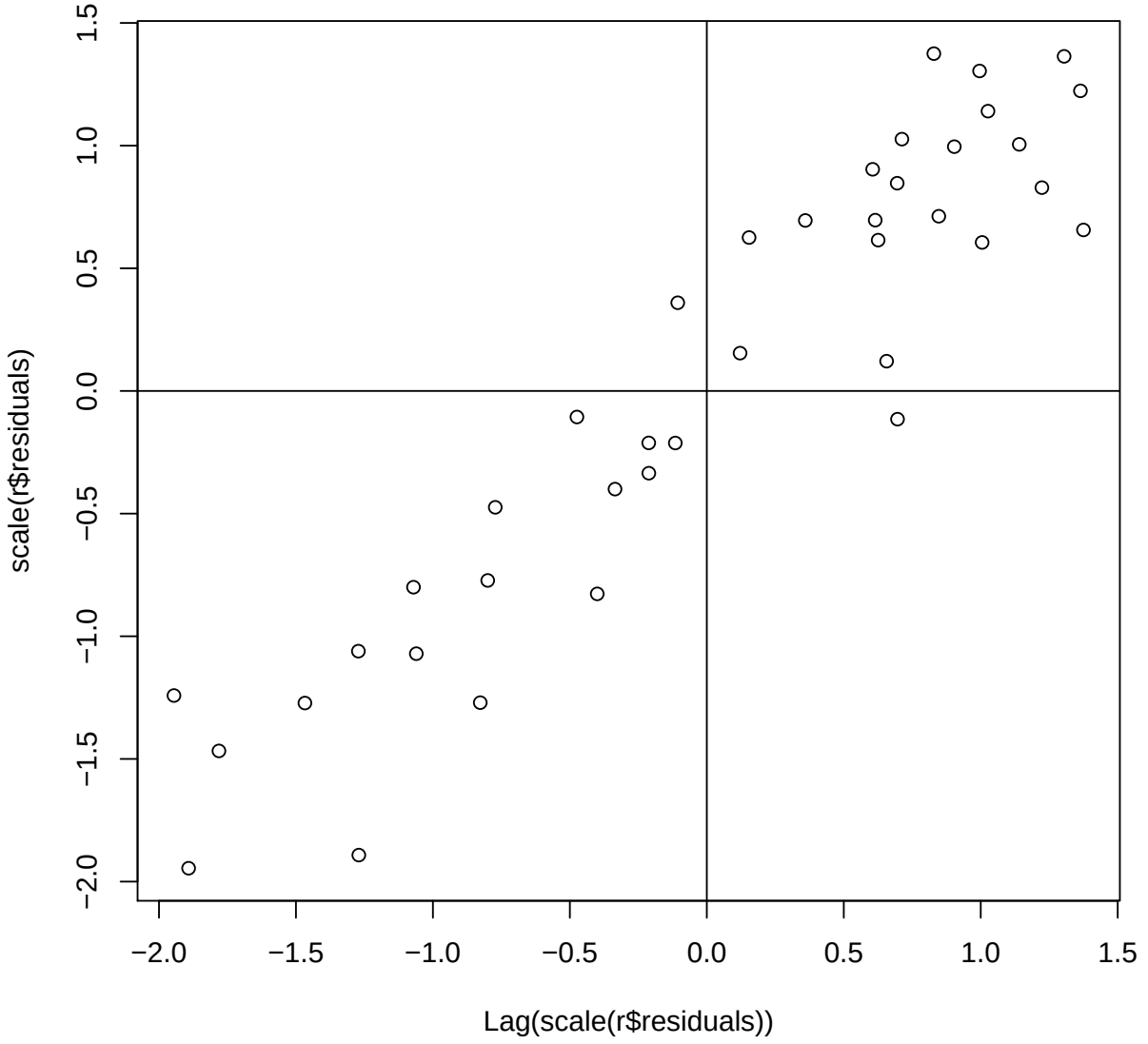
```
r<-lm(Y~X,data=wage)
summary(r)
```

```
##
## Call:
## lm(formula = Y ~ X, data = wage)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.138 -2.130  0.364  2.201  3.632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.5192     1.9424   15.20   <2e-16 ***
## X             0.7137     0.0241   29.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.676 on 38 degrees of freedom
## Multiple R-squared:  0.9584,Adjusted R-squared:  0.9574
## F-statistic: 876.5 on 1 and 38 DF,  p-value: < 2.2e-16
```

3.2.1 Graphical Method

We can plot the standardized residuals against time. The standardized residuals are simply the residuals ($\hat{u}_t$) divided by the standard error of the regression ($\sqrt{\hat{\sigma^2}}$), that is, they are ($\hat{u}_t/\hat{\sigma}$). Notice that $\hat{u}_t$ and $\hat{\sigma}$ are measured in the units in which the regress and Y is measured. The values of the standardized residuals will therefore be pure numbers (devoid of units of measurement) and can be compared with the standardized residuals of other regressions. Moreover, the standardized residuals, like $\hat{u}_t$, have zero mean and approximately unit variance. We can plot $\hat{u}_t$ against $\hat{u}_{t-1}$, that is, plot the residuals at time t against their value at time (t − 1).

```
#residuals
head(r$residuals)

##         1         2         3         4         5         6
## -4.703979 -3.874907 -3.359494 -2.800911 -2.828229 -2.112378

#standardized residuals
head(scale(r$residuals))

##         [,1]
## 1 -1.7811303
## 2 -1.4672076
## 3 -1.2720499
## 4 -1.0605460
## 5 -1.0708901
## 6 -0.7998378

#plot
library(Hmisc)
plot(scale(r$residuals)~Lag(scale(r$residuals)))
abline(h=0,v=0)
```



3.2.2 Durbin–Watson d Test

The most celebrated test for detecting serial correlation is that developed by statisticians Durbin and Watson. It is popularly known as the Durbin– Watson d statistic, which is defined as

$$DW = \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T}\hat{u}_{t-1}^2} \tag{9}$$

which is simply the ratio of the sum of squared differences in successive residuals to the RSS. Note that in the numerator of the d statistic the number of observations is n− 1 because one observation is lost in taking successive differences.

Warning

1. The regression model includes the intercept term.

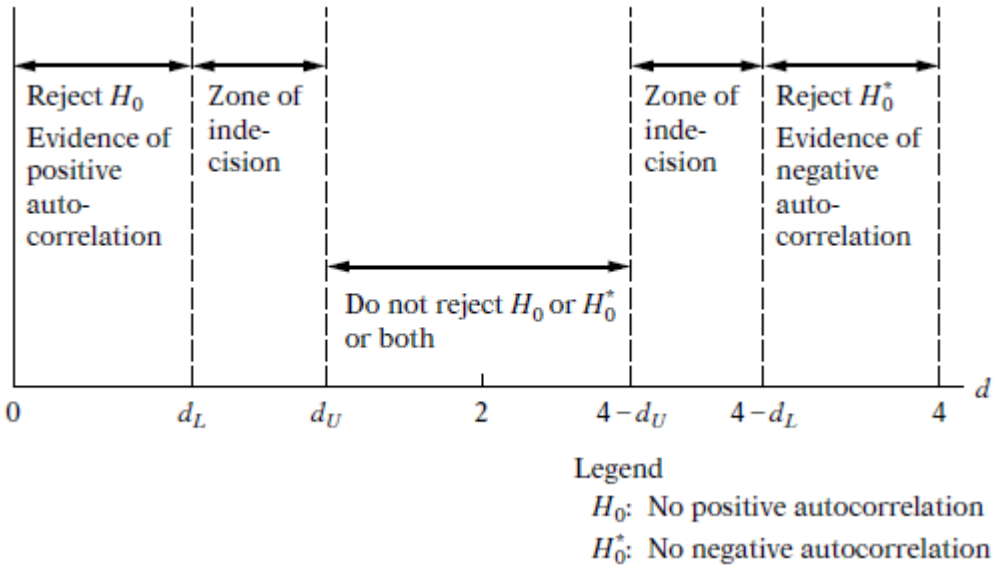


Figure 4: Durbin–Watson d statistic

2. The disturbances u_t are generated by the first-order autoregressive scheme: $u_t = \rho u_{t-1} + \varepsilon_t$. Therefore, it cannot be used to detect higher-order autoregressive schemes.
3. The error term u_t is assumed to be normally distributed.
4. The regression model does not include the lagged value(s) of the dependent variable as one of the explanatory variables.
5. There are no missing observations in the data.
6. The number of sample is more than 15.

Expand 9 to obtain

$$DW = \frac{\sum_{t=2}^{T} \hat{u}_t^2 + \sum_{t=2}^{T} \hat{u}_{t-1}^2 - 2\sum_{t=2}^{T} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^{T} \hat{u}_t^2}$$

Since $\sum_{t=2}^{T} \hat{u}_t^2$ and $\sum_{t=2}^{T} \hat{u}_{t-1}^2$ differ in only one observation, they are approximately equal.

$$DW \approx 2\left(\frac{\sum_{t=2}^{T} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^{T} \hat{u}_t^2}\right) \tag{10}$$

Now let us define

$$r = \frac{\sum_{t=2}^{T} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^{T} \hat{u}_t^2}$$

as the sample first-order coefficient of autocorrelation. We can express 10 as

$$DW \simeq 2(1 - r)$$

But since $-1 \leq \rho \leq 1$, it implies that

$$0 \leq d \leq 4$$

```
library(car)
durbinWatsonTest(r)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.8781174     0.1229044       0
##  Alternative hypothesis: rho != 0
```

3.2.3 The Breusch–Godfrey (BG) Test

To avoid some of the pitfalls of the Durbin–Watson d test of autocorrelation, statisticians Breusch and Godfrey have developed a test of autocorrelation that is general in the sense that it allows for (1) nonstochastic regressors, such as the lagged values of the regressand; (2) higher-order autoregressive schemes, such as AR(1), AR(2), etc.; and (3) simple or higher-order moving averages of white noise error terms.

$$Y = \beta_0 + \beta_1 X_t + u_t$$

Assume that the error term ut follows the pth-order autoregressive, AR(p), scheme as follows:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t$$

where ε_t is a white noise error term as discussed previously. As you will recognize, this is simply the extension of the AR(1) scheme.

The null hypothesis H_0 to be tested is that

$$H_0 : H_0 = \rho_1 = \rho_2 = \cdots = \rho_p = 0$$

That is, there is no serial correlation of any order. The BG test involves the following steps:

1. Estimate original model by OLS and obtain the residuals.
2. Run the following regression: $\hat{u}_t = \beta_0 + \beta_1 X_t + \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + \cdots + \rho_p \hat{u}_{t-p} + \varepsilon_t$, and obtain R2 from this (auxiliary) regression.
3. If the sample size is large (technically, infinite), Breusch and Godfrey have shown that $(n-p)R^2 \sim \chi^2_p$

A drawback of the BG test is that the value of p, the length of the lag, cannot be specified a priori. Some experimentation with the p value is inevitable. Sometimes one can use the so-called Akaike and Schwarz information criteria to select the lag length.

```
library(lmtest)
bgtest(r)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  r
## LM test = 32.2046, df = 1, p-value = 1.388e-08
```

```
bgtest(r,order=3)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 3
##
## data:  r
## LM test = 32.2706, df = 3, p-value = 4.589e-07
```

3.3 Remedial measures

If after applying one or more of the diagnostic tests of autocorrelation discussed in the previous section, we find that there is autocorrelation, we have three options:

1. Try to find out if the autocorrelation is pure autocorrelation and not the result of mis-specification of the model. Sometimes we observe patterns in residuals because the model is misspecified— that is, it has excluded some important variables—or because its functional form is incorrect.

2. If it is pure autocorrelation, one can use appropriate transformation of the original model so that in the transformed model we do not have the problem of (pure) autocorrelation. As in the case of heteroscedasticity, we will have to use some type of generalized least-square (GLS) method.

3. In large samples, we can use the Newey–West method to obtain standard errors of OLS estimators that are corrected for autocorrelation. This method is actually an extension of White's heteroscedasticity-consistent standard errors method.

3.3.1 The method of generalized least squares (GLS)

Knowing the consequences of autocorrelation, especially the lack of efficiency of OLS estimators, we may need to remedy the problem. As a starter, consider the two-variable regression model:

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

and assume that the error term follows the AR(1) scheme, namely

$$u_t = \rho_1 u_{t-1} + \varepsilon_t$$

Now we consider two cases: (1) ρ is known and (2) ρ is not known but has to be estimated.

When ρ Is Known

If the coefficient of first-order autocorrelation is known, the problem of autocorrelation can be easily solved. If the regression model holds true at time t, it also holds true at time (t − 1).

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + u_{t-1} \tag{11}$$

Multiplying the aboved model by ρ on both sides, we obtain

$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{t-1} + \rho u_{t-1} \tag{12}$$

Subtracting 11 from 12 gives

$$Y_{t-1} - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_{t-1} - \rho X_{t-1}) + \varepsilon_t \tag{13}$$

where $\varepsilon_t = (u_t - \rho u_{t-1})$

Since the error term in 13 satisfies the usual OLS assumptions, we can apply OLS to the transformed variables Y* and X* and obtain estimators with all the optimum properties, namely, BLUE.

Regression 13 is known as the generalized, or quasi, difference equation. It involves regressing Y on X, not in the original form, but in the difference form, which is obtained by subtracting a proportion (= ρ) of the value of a variable in the previous time period from its value in the current time period. In this differencing procedure we lose one observation because the first observation has no antecedent. To avoid this loss of one observation, the first observation on Y and X is transformed as follows: $Y_1\sqrt{1-\rho^2}$ and $X_1\sqrt{1-\rho^2}$. This transformation is known as the Prais– Winsten transformation.

When ρ Is Not Known

Although conceptually straightforward to apply, the method of generalized difference given in 13 is difficult to implement because ρ is rarely known in practice. Therefore, we need to find ways of estimating ρ.

ρ Estimated from the Residuals

If the AR(1) scheme $u_t = \rho_1 u_{t-1} + \varepsilon_t$ is valid, a simple way to estimate rho is to regress the residuals $\hat{u}_t$ on $\hat{u}_{t-1}$, for the $\hat{u}_t$ are consistent estimators of the true u_t.

$$\hat{u}_t = \rho_1 \hat{u}_{t-1} + \varepsilon_t$$

where $\hat{u}_t$ are the residuals obtained from the original (level form) regression and where v_t are the error term of this regression. Note that there is no need to introduce the intercept term in the aboved model, for we know the OLS residuals sum to zero.

```
library(Hmisc)
e<-lm(r$residual~Lag(r$residual)-1)
summary(e)

##
## Call:
## lm(formula = r$residual ~ Lag(r$residual) - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9846 -0.3282  0.1143  0.5853  1.6304
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## Lag(r$residual)  0.91425    0.05634   16.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9106 on 38 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8739,Adjusted R-squared:  0.8706
## F-statistic: 263.4 on 1 and 38 DF,  p-value: < 2.2e-16

X1<-wage$X-e$coefficients*Lag(wage$X)
Y1<-wage$Y-e$coefficients*Lag(wage$Y)
rr<-lm(Y1~X1)
summary(rr)

##
## Call:
## lm(formula = Y1 ~ X1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93110 -0.53410  0.08462  0.44761  2.09922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.10820    0.65693   6.254 2.85e-07 ***
## X1           0.52890    0.07742   6.831 4.76e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8581 on 37 degrees of freedom
##   (1 observation deleted due to missingness)
```

```
## Multiple R-squared:  0.5578,Adjusted R-squared:  0.5458
## F-statistic: 46.66 on 1 and 37 DF,  p-value: 4.763e-08

library(car)
durbinWatsonTest(rr)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.1076249      1.620601   0.212
##  Alternative hypothesis: rho != 0
```

3.3.2 The NEWEY–WEST method of correcting

We can still use OLS but correct the standard errors for autocorrelation by a procedure developed by Newey and West. This is an extension of White's heteroscedasticity-consistent standard errors. The corrected standard errors are known as HAC (heteroscedasticity- and autocorrelation-consistent) standard errors or simply as Newey–West standard errors. if a sample is reasonably large, one should use the Newey–West procedure to correct OLS standard errors not only in situations of autocorrelation only but also in cases of heteroscedasticity, for the HAC method can handle both, unlike the White method, which was designed specifically for heteroscedasticity.

```
r<-lm(Y~X,data=wage)
summary(r)

##
## Call:
## lm(formula = Y ~ X, data = wage)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.138 -2.130  0.364  2.201  3.632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.5192     1.9424   15.20   <2e-16 ***
## X             0.7137     0.0241   29.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.676 on 38 degrees of freedom
## Multiple R-squared:  0.9584,Adjusted R-squared:  0.9574
## F-statistic: 876.5 on 1 and 38 DF,  p-value: < 2.2e-16

library(sandwich) #Newey & West (1987, 1994)
library(lmtest) #coeftest
coeftest(r, vcov = NeweyWest(r))

##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.51925   28.13875  1.0491  0.30078
## X            0.71366    0.41179  1.7331  0.09119 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(r, vcov = vcovHAC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 29.519255   4.238850   6.964 2.751e-08 ***
## X            0.713659   0.054111  13.189 9.277e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4 An other example

```
library(Hmisc)
library(car)
wage2<-read.csv("C:\\Users\\XXXHHF\\Documents\\R\\workfile\\#R lecture\\11\\Exercise_au.csv",header=T
head(wage2)
```

```
##   Year   X   Y
## 1 1970 239 300
## 2 1971 248 311
## 3 1972 258 329
## 4 1973 272 351
## 5 1974 268 354
## 6 1975 280 364
```

```
r<-lm(Y~X,data=wage2)
summary(r)
```

```
##
## Call:
## lm(formula = Y ~ X, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.872  -2.986   1.371   3.474  12.355
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68.16026   15.26513  -4.465 0.000177 ***
## X             1.52971    0.05098  30.008  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.986 on 23 degrees of freedom
## Multiple R-squared:  0.9751,Adjusted R-squared:  0.974
## F-statistic: 900.5 on 1 and 23 DF,  p-value: < 2.2e-16
```

```
#acf(r$residual)
#pacf(r$residual)
durbinWatsonTest(r)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7549386     0.3482883       0
##  Alternative hypothesis: rho != 0
```

```r
e<-lm(r$residual~Lag(r$residual)-1)
summary(e)

##
## Call:
## lm(formula = r$residual ~ Lag(r$residual) - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9661 -1.8012  0.0091  2.7594  9.5074
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## Lag(r$residual)   0.8738     0.1297   6.735  7.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.04 on 23 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6635,Adjusted R-squared:  0.6489
## F-statistic: 45.35 on 1 and 23 DF,  p-value: 7.196e-07

X1<-wage2$X-e$coefficients*Lag(wage2$X)
Y1<-wage2$Y-e$coefficients*Lag(wage2$Y)
rr<-lm(Y1~X1)
summary(rr)

##
## Call:
## lm(formula = Y1 ~ X1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7350 -2.9575  0.2045  3.1063  6.1288
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1981     7.7907   0.410    0.685
## X1            1.2520     0.1878   6.667 1.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.925 on 22 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6689,Adjusted R-squared:  0.6539
## F-statistic: 44.45 on 1 and 22 DF,  p-value: 1.056e-06

durbinWatsonTest(rr)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.2741419      1.322343   0.088
##  Alternative hypothesis: rho != 0
```

Table A-2
Models with an intercept (from Savin and White)

Durbin-Watson Statistic: 5 Per Cent Significance Points of dL and dU

	k'*=1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.700	1.356	0.467	1.896	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.763	1.332	0.559	1.777	0.367	2.287	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	-----	-----	-----	-----	-----	-----	-----	-----
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	-----	-----	-----	-----	-----	-----
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	-----	-----	-----	-----
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	-----	-----
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.838	1.369	1.874	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.747	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.846	1.608	1.862	1.593	1.877
200	1.758	1.779	1.748	1.789	1.738	1.799	1.728	1.809	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

*k' is the number of regressors excluding the intercept