Linear models and time series

Gabriela Ciołek

March 22, 2019

The Heteroskedasticity Function Must Be Estimated: Feasible GLS

- Consider linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

- Last time we saw some examples of where the heteroskedasticity is known up to a multiplicative form.
- In most cases, the exact form of heteroskedasticity is not obvious.
- In practice: it is difficult to find the function $h(x_i)$.
- In this case we can model the function h and use the data to estimate $\hat{h}_i$.

FGLS method

Using $\hat{h}_i$ instead of h_i in the GLS transformation yields an estimator called **the feasible GLS (FGLS)** estimator.

- We model the heteroskedasticity in a following way:

$$Var(u|x) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k),$$

  where $x_1, \cdots, x_k$ are explanatory variables and δ_j are uknown parameters.
- Other functions of the x_j can appear, but we will focus primarily on

$$h(x) = \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k)$$

Why the form

$$h(x) = \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k)$$

- We assumed that heteroskedasticity was a linear function of the x_j (see TP4).
- Linear alternativesre fine when testing for heteroskedasticity, but they can be problematic when correcting for heteroskedasticity using weighted least squares.
- Linear models do not ensure that predicted values are positive, and our estimated variances must be positive in order to perform WLS.

- We assume that
$$h(x) = \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k).$$
- Under this hypothesis we write
$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k)\nu$$
with $\mathbb{E}(\nu | x_1, \cdots, x_k) = 1$.
- If we assume that ν is independent of $x_1, \cdots, x_k$ we write
$$\log(u^2) = \alpha_0 + \delta_1 x_1 + \cdots + \delta_k x_k + e,$$
where e is zero mean random variable and independent of $x_1, \cdots, x_k$. The intercept in this equation is different from δ_0, but this is not important.
- Run the regression of $log(u^2)$ on $x_1, x_2, \cdots, x_k$.
- We need from this regression the fitted values; call these $\hat{g}_i$. Then, the estimates of h_i are simply $\hat{h}_i = \exp(\hat{g}_i)$.

A FEASIBLE GLS PROCEDURE TO CORRECT HETEROSKEDASTICITY:

1. Run the regression of y on $x_1, x_2, \cdots, x_k$ and obtain the residuals $\hat{u}^2$.
2. Create $log(\hat{u}^2)$ by first squaring the OLS residuals and then taking the natural log.
3. Run the regression $log(u^2)$ on $x_1, x_2, \cdots, x_k$ and obtain the fitted values $\hat{g}$.
4. Exponentiate the fitted values from, i.e. $\hat{h} = exp(\hat{g})$.
5. Estimate the equation

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

by WLS, using weights $1/\hat{h}$.

Exercise 1

- Consider a linear model $sav = \beta_0 + \beta_1 inc + u$.

  load saving.raw

- Conduct a correction by the method FLS

We perform regression $sav = \beta_0 + \beta_1 inc + u$ and obtain residuals u. Next:

$$lu2 = log(u.^2);$$

$$y = lu2;$$

$$beta = inv(X' * X) * X' * y$$

$$u = y - X * beta;$$

Recall the estimated values, called $\hat{g}_i$, and the estimated values of h_i are simply $\hat{h}_i = \exp(\hat{g}_i)$

$$g = X * beta$$

$$new\_weight = sqrt(exp(g))$$

Finally, we apply the least weighted squares with weights $1/\hat{h}_i$ instead of $1/h_i$.

$$y = saving(:, 1);$$

$$inc = saving(:, 2);$$

$$X = [ones(n, 1), inc];$$

$$ys = y./new\_weight;$$

$$Xs = [ones(n, 1)./new\_weightinc./new_weight];$$

$$beta = inv(Xs' * Xs) * Xs' * ys$$

$$u = ys - Xs * beta;$$

Exercise 2

- Consider a model:

$$cigs|log(income), log(cigpric), educ, age, age^2, restaurn,$$

  where *cigs* is a demand function for daily cigarette consumption.
- load smoke.raw
- Test the homoskedasticity
- Conduct a correction by the method FLS

Time series

- A time series is a series of data points indexed in time order. An obvious characteristic of time series data which distinguishes it from cross-sectional data is that a time series data set comes with a temporal ordering.
- The autoregressive process AR(p) is of the form:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t$$

  with $\mathbb{E}[u_t] = 0, \mathbb{E}[u_t u_s] = \sigma_u^2$ if $t = s$ and 0 else
- The moving average process Ma(q) is of the form:

$$y_t = u_t + \psi_1 u_{t-1} + \cdots + \psi_q u_{t-q}$$

- The process ARMA(p,q) is of the form:

$$y_t - \sum_{k=1}^{p} \phi_k y_{t-k} = \epsilon_t + \sum_{j=1}^{q} \psi_j u_{t-j}$$

The autocorrelation (ACF) and partial autocorrelation (PACF) functions of these processes make it possible to identify the time series. ACF and PACF plots graphically summarize the strength of a relationship with an observation in a time series with observations at prior time steps.

- autocorrelation, also known as serial correlation, is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. Autocorrelation is the linear dependence of a variable with itself at two points in time. For stationary processes, autocorrelation between any two observations only depends on the time lag h between them.
- the partial autocorrelation function (PACF) gives the partial correlation of a time series with its own lagged values, controlling for the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags. Partial autocorrelation is the autocorrelation between y_t and y_{t-1} after removing any linear dependence on $y_1, y_2, \cdots, y_{t-h+1}$.

- Autoregressive processes (**AR (p)**): The ACF decreases exponentially. Concerning the **PACF**, the coefficients are null or not significant for $|h| > p$.
- Moving average processes (**MA(q)**): The PACF decreases exponentially. Concerning the **ACF**, the coefficients are null or not significant for $|h| > q$.
- Processes (**ARMA(p,q)**): The ACF and the PACF are decreasing, but they do not necessarily become zero after a certain delay. It is therefore more difficult to identify an ARMA model than a pure autoregressive or a moving average model.

Exercise 3

- Simulate $AR(1)$

$$y_n = 0.6y_{n-1} + u_n$$

  and plot its trajectory
- Compute ACF and PACF

$$y = zeros(1000, 1);$$

$$for \;\; i = 2 : 1000;$$

$$y(i) = 0.6 * y(i-1) + randn;$$

$$end;$$

$$plot(y)$$

$$acf = autocorr(y, 20);$$

$$pacf = parcorr(y, 20);$$

Exercise 4

- Simulate $MA(1)$

$$z_n = e_n + 0.8e_{n-1}$$

  and plot its trajectory
- Compute ACF and PACF

$$n = 1000; z = zeros(n, 1)$$

$$e = randn(n, 1);$$

$$for \;\; i = 2 : 1000;$$

$$z(i) = e(i) + 0.8 * e(i-1);$$

$$end;$$

$$plot(z);$$

$$acf = autocorr(z, 20);$$

$$pacf = parcorr(z, 20);;$$

Exercise 5

- Simulate $AR(2)$

$$y_n = 0.6y_{n-1} + 0.2y_{n-2} + u_n$$

  and plot its trajectory
- Compute ACF and PACF

Exercise 6

- Simulate $ARMA(2, 2)$

$$y_n = 0.5y_{n-1} - 0.8y_{n-2} + e_n + 0.6e_{n-1} + 0.2e_{n-2}$$

  and plot its trajectory
- Compute ACF and PACF

The model of the ordinary least squares is considered under the following hypotheses:

- H1 : The model is linear in $x_{i,t}$
- H2 : the values $x_{i,t}$ are observed without error
- H3 : $\mathbb{E}[u] = 0$ the mean of the error is equal to zero
- H4: $\mathbb{E}[u^t u] = \sigma^2 I_n$, the variance of the error is constant
- H5 : $\mathbb{E}[u_t u_{t+1}] = 0$, the errors are not correlated
- H6 : $Cov(x_{i,t}, u_t)$, the error is independent of the explanatory variable

- The violation of hypothesis H5 concerns time series where the off-diagonal elements of the covariance matrix of the errors are nonzero. If $H5$ fails we say that the errors in

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

  suffer autocorrelation because they are correlated across time.
- In this case, the obtained OLS estimators are unbiased but no longer have a minimal variance.
- The detection of a potential dependency of the errors can only be carried out through the analysis of the residuals.

- Consider an equation for determining three-month, T-bill rates (i3t) based on the inflation rate (inft) and the federal deficit as a percentage of gross domestic product (deft):

$$i3t = \beta_0 + \beta_1 inft + \beta_2 deft + u_t$$

for $t = 1, \cdots, n$.

$$load \quad INTDEF.RAW$$

The data in INTDEF.RAW come from the 1997 Economic Report of the President and span the years 1948 through 1996. The variable i3 is the three-month T-bill rate, inf is the annual inflation rate based on the consumer price index (CPI), and def is the federal budget deficit as a percentage of GDP.

- Test the autocorrelaton of residuals in the model.

$$y = intdef(:, 2);$$

$$[n, k] = size(intdef)$$

$$X = [ones(n, 1), intdef(:, [3, 6])];$$

$$[n, k] = size(X)$$

$$beta = inv(X' * X) * X' * y$$

$$u = y - X * beta;$$

The residuals of the regression can be correlated in series. The most popular and simple model to be tested is the AR (1) model. We will therefore test the presence of correlations in series of type AR (1).
We assume that the residuals are given by

$$u_t = \rho u_{t-1} + e_t$$

We assume that $|\rho| < 1$ (statbility conditon), and that e_t are independent, zero mean random variables with variance σ_e^2.
In the model AR(1), the null hypothesis H_0 assumes that the errors are not correlated in the series

$$H_0 : \rho = 0.$$

We perform regression $\hat{u}_t = \rho \hat{u}_{t-1} + e_t$ for $t = 2$, recover $\hat{\rho}$,and calculate the statistic $t_{\hat{\rho}}$.

$$u\_ = [u(2 : n)]$$

$$u\_lag = [u(1 : n-1)]$$

$$y = u\_$$

$$X = [u\_lag];\ \textit{without constant}$$

$$[n, k] = size(X)$$

$$rho = inv(X' * X) * X' * y$$

$$u = y - X * rho;$$

$$sig2 = u' * u/(n-k)$$

$$std = sqrt(diag(sig2 * inv(X' * X)))$$

$$t = rho./std$$

Use t-test to test H_0.