# Data Analysis: Robust Estimation

Mark Hendricks

August Review

UChicago Financial Mathematics

# Outline

Robust Estimators

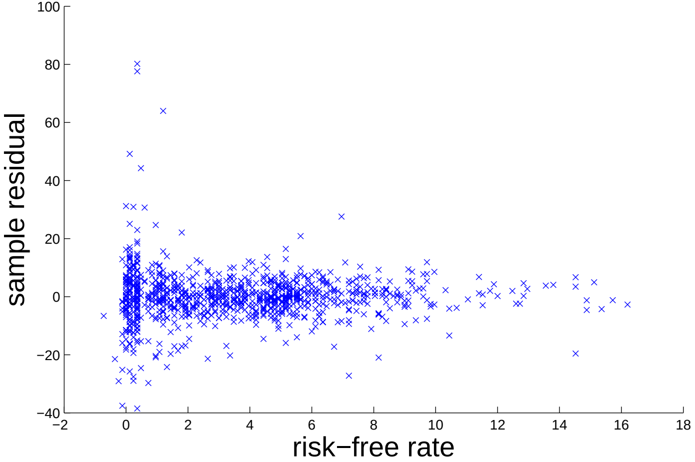
Serial Correlation

# Checking for heteroscedasticity

Consider a regression of excess stock returns on the risk-free rate.

- To see if heteroscedasticity is a problem, try plotting the residuals against some conditioning variable.
- If the range of the sample residuals seems to change across the values of the conditioning variable, this may indicate heteroscedasticity.

# Residuals: Excess return on risk-free rate



# Lagrange Multipler test

The **Lagrange Multiplier Test** (Breusch-Pagan) tests the hypothesis that

$$\sigma_i^2 = \sigma^2 \left[\mathbf{z}_i' \boldsymbol{\alpha}\right]$$

where $\mathbf{z}_i$ is a vector of conditioning variables for observation $i$.

- If the model is homoscedastic, then $\boldsymbol{\alpha} = \mathbf{0}$.
- One might try using a subset of $x$ for the variables $z$.
- This tests a certain form of heteroscedasticity. In fact, it need not be linear, but even tests

$$\sigma_i^2 = \sigma^2 f\left(\left[\mathbf{z}_i' \boldsymbol{\alpha}\right]\right)$$

# Computing the LM test

Regress sample estimates of variances on $\mathbf{x}$ (or subset of $\mathbf{x}$),

$$e_i^2 = \mathbf{x}'\gamma + \nu_i$$

LM test stat is $R^2$ from this regression multiplied by the sample size:

$$LM = n\,R^2, \quad LM \sim^a \chi^2(k)$$

- For the example above, the LM test rejects homoscedasticity at the 1% level.
- The LM test can perform poorly with nonnormal data, but the simple adjustments are available.

# Other tests of heteroscedasticity

The LM tests again a certain parametrization of heteroscedasity. This gives the test power, but means it may be mispecified.

- **White's test** is quite general: it makes no assumption about the nature of the heteroscedasticity. It examines the R-squared from regressing the squared errors on $\mathbf{X}$ along with quadratic terms in $\mathbf{X}$.
- The **Goldfeld-Quandt** test simply tests one subset of the data against another subset. It looks for statistical difference in the variances of the subsets.

# Correcting for heteroscedasticity

With heteroscedasticty,

$$\text{var}\left[\boldsymbol{b} \mid \mathbf{x}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\Sigma\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

The key is how to estimate $\Sigma$. There are two approaches:

- Use nonparametric estimation of $\mathbf{X}'\Sigma\mathbf{X}$.
- Make parametric assumptions about the form of $\Sigma$, and estimate these.

# Nonparametric estimation of $\Sigma$

Recall that $\Sigma = \mathbb{E}\left[\epsilon\epsilon' \,|\mathbf{x}\right]$

- This is an $n \times n$ matrix. There is no hope of estimating it using a sample of size $n$.
- This is one reason that a parametric assumption on $\Sigma$ is useful.
- But using just the data, one can get an estimate of $\mathbf{X}'\Sigma\mathbf{X}$, a $(k+1) \times (k+1)$ matrix.

# White estimator

Write out

$$\mathbf{X}'\Sigma\mathbf{X} = \sum_{i=1}^{n} \sigma_i^2 \mathbf{x}_i \mathbf{x}_i'$$

noting that we are assuming $\Sigma$ is diagonal (no autocorrelation.)

Then the **White estimator** is

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i'\right) \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

# Parametric Estimation

If we know the form of the serial correlation and heteroscedasticity, we can form efficient estimators.

- Recall that heteroscedasticity means that some observations have more statistical noise (epsilon shocks) than others.
- Efficient estimation would simply put less weight on these observations.
- Similarly, if we know which observations have correlated errors, we can put relatively less weight on these observations given that they do not contain as much new information.

# Generalized Least Squares

Suppose that we know the covariance matrix of $\epsilon$, denoted $\Sigma$.

- Weight the observations by the inverted covariance matrix. (Pay more attention to the more precise data.)
- This yields the following, efficient estimator:

$$\boldsymbol{b} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y}$$

- The covariance of the GLS estimator is

$$\text{var}\left(\boldsymbol{b}\right) = \Omega = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}$$

# Non-parametric v.s. parametric estimation

There is a tradeoff in model assumptions and estimation precision.

- The White estimator is impressive in that it makes no assumption about the form of heteroscedasticity.
- However, sample estimates of $\mathbf{X\Sigma X}$ can perform quite poorly.
- Further, the White estimator reveals nothing about the underlying heteroscedasticity which is useful for forecasting or studying the variance process.

# Non-parametric estimation

The goal is to estimate

$$\text{var}\left[\boldsymbol{b} \mid \mathbf{x}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\Sigma\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

which depends on estimating $\mathbf{X}'\Sigma\mathbf{X}$.

$$\mathbf{X}'\Sigma\mathbf{X} = \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{i,j}\mathbf{x}_i\mathbf{x}_j'$$

One might estimate $\mathbf{X}'\Sigma\mathbf{X}$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} e_i e_j \mathbf{x}_i\mathbf{x}_j'$$

# Trouble in estimation

Unfortunately, this sample estimate is not guaranteed to be positive definite.

- The common way to deal with this is to put less weight on observations further separated by time.
- Several different weighting schemes have been employed.

# Newey-West estimator

The **Newey-West estimator** of $\mathbf{X}'\Sigma\mathbf{X}$ is popular:

$$\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i' + \sum_{\ell=1}^{L} \sum_{t=\ell+1}^{n} w_\ell e_t e_{t-\ell} \left( \mathbf{x}_t \mathbf{x}_{t-\ell}' + \mathbf{x}_{t-\ell} \mathbf{x}_t' \right)$$

$$w_\ell = 1 - \frac{\ell}{L+1}$$

for some number of lags $L$.

- First term is the same as the heteroscedasticity-consistent estimator.
- Second term estimates autocorrelations of errors.

# Outline

Robust Estimators

Serial Correlation

# Serial correlation

As with heteroscedasticity, serial correlation changes the inference of the OLS estimate compared to the classic case where $\Sigma = \sigma^2\mathcal{I}$.

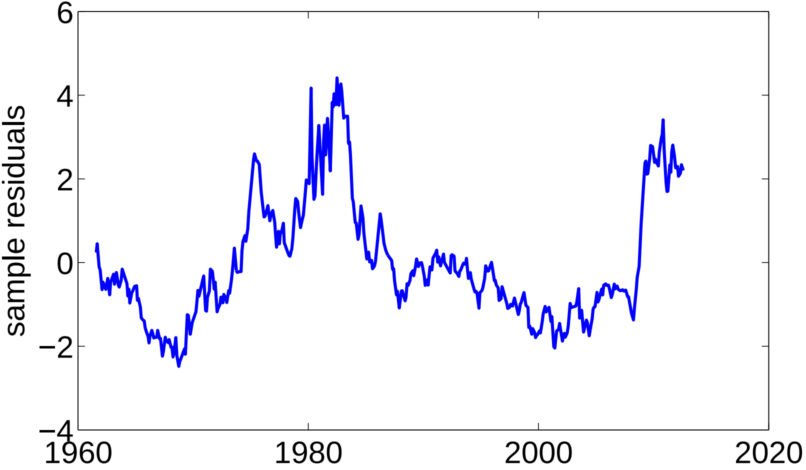
- With serial correlation, there are off-diagonal elements in $\Sigma$.
- As mentioned, OLS is still valid, given that one uses the more complicated equation for the variance of $\boldsymbol{b}$.

# Example of residual autocorrelation

In many time-series regressions, the errors exhibit autocorrelation.

- This is the idea that a shock to the variable at time $t$ may still be affecting the value at time $t + 1$.
- Correlation of residuals invalidates the finite-sample inference results of OLS.
- Consider the previous example of regressing unemployment on U.S. Treasury yields.

# Residuals: Unemployment on the yield spread



# Autocorrelated series

The autocorrelation of the error series at a monthly frequency is 97%.

- This essentially says that the regression has much less data than the classic formulas understand.
- With highly correlated data, there is little true sample variation for OLS to use in estimation.
- Consider regressing one very persistent data series on another persistent data series.
- The levels of such persistent $X$ and $Y$ may track closely together just due to the persistence.

# Model mispecification

Often, autocorrelated errors are a sign that the model is mispecified.

- This is commonly caused by having a time-trend in the data.
- This also may be a sign that the model should use the differenced data.
- Much of time-series statistics deals with examining whether the data has a time-trend, a random-walk, or cointegration.
- This is beyond the scope of the notes.

# Non-parametric v.s. parametric estimation

Like with heteroscedasticity, one can use parametric assumptions to simplify the estimation.

- Time-series statistics often makes assumptions about a linear model having autoregressive (AR) or moving average (MA) components.
- Again, this will be discussed more later in the program.

# AR(1) serial correlation

Consider the AR(1) model for $\epsilon$.

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

where $u_t$ is homoscedastic, uncorrelated, with variance $\sigma_u^2$.

This implies that

$$\Sigma = \frac{\sigma_u^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ & & \vdots & & & \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho^2 & \rho & 1 \end{bmatrix}$$

This is a widely used model for time-series correlation.

# References


- Cochrane, John. *Asset Pricing.* 2001.
- Greene, William. *Econometric Analysis.* 2011.
- Hamilton, James. *Time Series Analysis.* 1994.
- Wooldridge, Jeffrey. *Econometric Analysis of Cross Section and Panel Data.* 2011.