Regression analysis: Part II

Daniele Bianchi[1]

whitesphd.com

[1] School of Economics and Finance
Queen Mary, University of London

Summary

This week provides an overview of the large-sample properties of the OLS estimator as well as large-sample hypothesis testing. Finally, we cover advance topics in model selection and specification.

Contents

1. Large-sample properties
2. Large-sample hypothesis testing
3. Violations of the large-sample assumptions
4. Model selection and specification checking
5. Exercises

Large-sample properties

Large-sample assumptions

While the small-sample assumptions allow the exact distribution of the OLS estimator and test statistics to be derived, these assumptions are not realistic in applications using financial data.

Asset returns are non-normal (both skewed and leptokurtic), heteroskedastic, and correlated.

The large-sample framework allows for inference on β without making strong assumptions about the distribution of error covariance structure.

Four new assumptions are needed to analyze the asymptotic behavior of the OLS estimator: stationarity and ergodicity, rank, martingality and moment existence.

Large-sample assumptions

Assumption (4.1 Stationary Ergodicity)

$\{\mathbf{x}_i, \epsilon_i\}$ *is a strictly stationary and ergodic sequence.*

This is a technical assumption needed for consistency and asymptotic normality. It implies two properties about the joint density of $\{\mathbf{x}_i, \epsilon_i\}$:

1. The joint distribution of $\{\mathbf{x}_i, \epsilon_i\}$ and $\{\mathbf{x}_{i+j}, \epsilon_{i+j}\}$ depends on the time between observations j and **not** the observation index i.
2. The averages will converge to their expected value (as long as they exist).

The concept of stationarity will be considered more deeply next week.

Large-sample assumptions

Assumption (4.2 Rank)

$E\left[\mathbf{x}_i'\mathbf{x}_i\right] = \Sigma_{XX}$ *is non-singular and finite.*

This assumption, like $\text{rank}(X) = k$, is needed to ensure identification.

Assumption (4.3 Martingale difference)

$\{\mathbf{x}_i'\boldsymbol{\epsilon}_i, \mathcal{F}_i\}$ *is a martingale difference sequence,*

$$E\left[\left(X_{j,i}\boldsymbol{\epsilon}_i\right)^2\right] < \infty, j = 1, 2, \ldots, k, i = 1, 2, \ldots,$$

and $\mathbf{S} = V\left[n^{-1/2}\mathbf{X}'\boldsymbol{\epsilon}\right]$ *is finite and non singular.*

A martingale difference sequence has the property that its mean is unpredictable using the information contained in the information set $(\mathcal{F}_i)$. For instance, in the CAPM, the return on the market portfolio can be thought of as being determined independently of the idiosyncratic shock affecting individual assets.

Large-sample assumptions

Assumption (4.4 Moment existence)

$E\left[X_{j,i}^4\right] < \infty, \qquad i = 1, 2, \ldots, \qquad j = 1, 2, \ldots, k$ *and*
$E\left[\epsilon_i^2\right] = \sigma^2 < \infty, i = 1, 2, \ldots$

This final assumption requires that the fourth moment of any regressor exists and the variance of the error is finite. This is needed to derive a consistent estimator of the parameter covariance.

Large-sample properties

These assumptions lead to two theorems that describe the asymptotic behavior of $\hat{\beta}$: it is consistent and asymptotically normally distributed.

First, some new notation is needed. Let

$$\hat{\beta} = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \left(\frac{\mathbf{X}'\mathbf{y}}{n}\right),$$

be the regression coefficient using n realisations $\mathbf{y}_i, \mathbf{x}_i$.

Theorem (4.1 Consistency of $\hat{\beta}$)

Under linearity and Assumptions 4.1-4.3, then $\hat{\beta}_n \xrightarrow{p} \beta$

N.B., Consistency is a weak property of the OLS estimator, but it is important. This result relies crucially on the implication of assumption 4.3, i.e., that $n^{-1}\mathbf{X}'\boldsymbol{\epsilon} \xrightarrow{p} 0$.

Large-sample properties

Under the same assumptions, the OLS estimator is also asymptotically normally distributed.

Theorem (4.2 Asymptotic Normality of $\hat{\beta}$)

Under linearity and assumptions 4.1-4.3, then

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} N\left(0, \Sigma_{XX}^{-1}\mathbf{S}\Sigma_{XX}^{-1}\right),$$

where $\Sigma_{XX} = E\left[\mathbf{x}_i'\mathbf{x}_i\right]$ *and* $\mathbf{S} = V\left[n^{-1/2}\mathbf{X}'\boldsymbol{\epsilon}\right]$

Asymptotic normality provides the basis for hypothesis tests on β. However, Σ_{XX} and $\mathbf{S}$ are unknown, and so must be estimated.

Large-sample properties

Theorem (4.3 Consistency of the OLS covariance estimator)

Under linearity and assumptions 4.1-4.3,

$$\begin{aligned}\hat{\Sigma}_{XX} &= n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \Sigma_{XX},\\ \hat{\mathbf{S}} &= n^{-1}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i'\mathbf{x}_i \xrightarrow{p} \mathbf{S},\\ &= n^{-1}\left(\mathbf{X}'\hat{\mathbf{E}}\mathbf{X}\right),\end{aligned}$$

and

$$\hat{\Sigma}_{XX}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{XX}^{-1} \xrightarrow{p} \Sigma_{XX}^{-1}\mathbf{S}\Sigma_{XX}^{-1},$$

where $\hat{\mathbf{E}} = diag\left(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_n^2\right)$ *is a matrix with the estimated residuals squared along its diagonal.*

Large-sample properties

The usual estimator of the residual variance is also consistent for the variance of the innovations under the same conditions.

Theorem (4.4 Consistency of the OLS variance estimator)

Under assumptions 4.1-4.3, then $\hat{\sigma}_n^2 = n^{-1}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \xrightarrow{p} \sigma^2$

Further, if homoskedasticity is assumed, then the parameter covariance estimator can be simplified.

Theorem (4.5 Homoskedastic errors)

Under linearity, homoskedasticity, normality and assumptions 4.1-4.3, then $\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} N\left(0, \sigma^2\Sigma_{XX}^{-1}\right)$

Combining this theorem with the results above, a consistent estimator of $\sigma^2\Sigma_{XX}^{-1}$ is given by $\hat{\sigma}_n^2\hat{\Sigma}_{XX}^{-1}$.

Large-sample hypothesis testing

Large-sample hypothesis testing

Combining the theorems above, the OLS estimator turns out to be:

- Consistent
- Asymptotically normal
- Asymptotic variance can be consistently estimated

These three properties provide the tools necessary to conduct hypothesis testing in the asymptotic framework.

Both the **Wald** and the **Likelihood Ratio** (LR) test have large-sample equivalents that exploit the estimated parameters' asymptotic normality.

Wald test

Recall that $\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} N\left(0, \Sigma_{XX}^{-1}\mathbf{S}\Sigma_{XX}^{-1}\right)$. Following from Theorem 4.2, if $H_0 : \mathbf{R}\boldsymbol{\beta} - \mathbf{r} = 0$ is true, then[1]

$$\sqrt{n}\left(\mathbf{R}\hat{\beta}_n - \mathbf{r}\right) \xrightarrow{d} N\left(0, \mathbf{R}\Sigma_{XX}^{-1}\mathbf{S}\Sigma_{XX}^{-1}\mathbf{R}'\right),$$

and

$$\Gamma^{-1/2}\sqrt{n}\left(\mathbf{R}\hat{\beta}_n - \mathbf{r}\right) \xrightarrow{d} N\left(0, \mathbf{I}_k\right),$$

where $\Gamma = \mathbf{R}\Sigma_{XX}^{-1}\mathbf{S}\Sigma_{XX}^{-1}\mathbf{R}'$. Under the null that $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$,

$$n\left(\mathbf{R}\hat{\beta}_n - \mathbf{r}\right)'\left[\mathbf{R}\Sigma_{XX}^{-1}\mathbf{S}\Sigma_{XX}^{-1}\mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\beta}_n - \mathbf{r}\right) \xrightarrow{d} \chi_m^2,$$

where m is the $\operatorname{rank}(\mathbf{R})$.

[1] Applying the properties of a normal random variable, if $\mathbf{z} \sim N\left(\mu, \Sigma\right), \mathbf{c}'\mathbf{z} \sim N\left(\mathbf{c}'\mu, \mathbf{c}'\Sigma\mathbf{c}\right)$ and that if $w \sim N\left(\mu, \sigma^2\right)$ then $\frac{(w-\mu)^2}{\sigma^2} \sim \chi_1^2$.

Wald test

By applying the results of Theorem 4.3, we can substitute the unknown quantities $\Gamma, \mathbf{S}$ and Σ_{XX} with their empirical counterparts, i.e.,

$$\hat{\Sigma}_{XX} = n^{-1}\mathbf{X}'\mathbf{X}, \qquad \hat{\mathbf{S}} = n^{-1}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i' \mathbf{x}_i, \qquad \hat{\Gamma} = \hat{\Sigma}_{XX}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{XX}^{-1}, \quad (1)$$

The feasible Wald statistic is then defined as

$$W = n\left(\mathbf{R}\hat{\beta}_n - \mathbf{r}\right)'\left[\mathbf{R}\hat{\Sigma}_{XX}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{XX}^{-1}\mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\beta}_n - \mathbf{r}\right) \xrightarrow{d} \chi_m^2,$$

Test statistic values can be compared to the critical value C_α from a χ_m^2 at the α-significance level and the null is rejected if W is greater than C_α.

Wald test

To summarise, the Wald test in large samples is implemented as follows:

1. Estimate the unrestricted model $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$
2. Estimate the parameter covariance using $\hat{\Sigma}_{XX}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{XX}^{-1}$ where

$$\hat{\Sigma}_{XX} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i, \qquad \hat{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2\mathbf{x}_i'\mathbf{x}_i,$$

3. Construct the restriction matrix, $\mathbf{R}$ and the value of the restriction $\mathbf{r}$, from the null hypothesis.
4. Compute $W = n\left(\mathbf{R}\hat{\beta}_n - \mathbf{r}\right)'\left[\mathbf{R}\hat{\Sigma}_{XX}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{XX}^{-1}\mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\beta}_n - \mathbf{r}\right)$.
5. Reject the null if $W > C_\alpha$ where C_α is the critical value from a χ^2_m using a size of α.

Likelihood ratio test

A critical distinction between small-sample and large-sample hypothesis testing is the omission of conditional normality. Without this assumption, the distribution of the errors is left unspecified.

It may be tempting to think the LR is asymptotically valid. It is not.

However, there is a feasible LR-like test available. The functional form of the test is given by

$$LR = n\tilde{\mathbf{s}}'\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m,$$

where $\hat{S}$ is estimated using the scores of the unrestricted model (under the alternative)

$$\hat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2\mathbf{x}_i'\mathbf{x}_i,$$

Likelihood ratio test

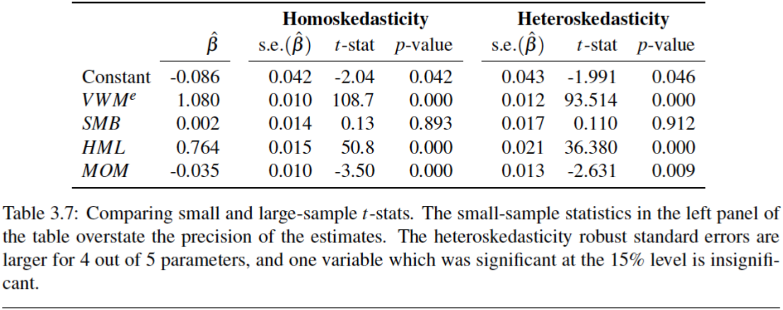
To summarise, the LR test in large samples is defined as

1. Estimate the unrestricted model $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$.
2. Impose the null on the unrestricted model and estimate the restricted model $\tilde{Y}_i = \tilde{\mathbf{X}}_i$
3. Compute the residuals from the restricted regression, $\tilde{\epsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$, and from the unrestricted regression, $\hat{\epsilon} = Y_i - \mathbf{x}_i\hat{\beta}$.
4. Construct the score from both models, $\tilde{\mathbf{s}}_i = \mathbf{x}_i\tilde{\epsilon}_i$ and $\hat{\mathbf{s}} = \mathbf{x}_i\hat{\epsilon}_i$, where in both cases $\mathbf{x}_i$ are the regressors from the unrestricted model.
5. Estimate the average score and the covariance of the score

$$\tilde{\mathbf{s}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{s}}_i, \qquad \hat{\mathbf{S}} = n^{-1}\sum_{i-1}^{n}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i,$$

6. Compute the LR test statistic as $LR = n\tilde{\mathbf{s}}\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}}'$.
7. Reject the null if $LR > C_\alpha$ where C_α is the critical value from a χ^2_m using a size of α.

Example on factor pricing models



		Homoskedasticity			Heteroskedasticity		
	$\hat{\beta}$	s.e.$(\hat{\beta})$	t-stat	p-value	s.e.$(\hat{\beta})$	t-stat	p-value
Constant	-0.086	0.042	-2.04	0.042	0.043	-1.991	0.046
VWM^e	1.080	0.010	108.7	0.000	0.012	93.514	0.000
SMB	0.002	0.014	0.13	0.893	0.017	0.110	0.912
HML	0.764	0.015	50.8	0.000	0.021	36.380	0.000
MOM	-0.035	0.010	-3.50	0.000	0.013	-2.631	0.009

Table 3.7: Comparing small and large-sample t-stats. The small-sample statistics in the left panel of the table overstate the precision of the estimates. The heteroskedasticity robust standard errors are larger for 4 out of 5 parameters, and one variable which was significant at the 15% level is insignificant.

Source: Financial Econometrics Notes by Kevin Sheppard.

Violations of the large-sample assumptions

Violations of the large-sample assumptions

The large-sample assumptions are just that: assumptions. While this set of assumptions is more general than the finite-sample setup, they may be violated in a number of ways.

Few common violations will be investigated:

- Omitted variable bias
- Inclusion of irrelevant variables
- Heteroskedasticity

Omitted variable bias

Suppose that the model is linear but misspecified, and a subset of the relevant regressors are excluded. The model can be specified as

$$Y_i = \beta_1 \mathbf{X}_{1i} + \beta_2 \mathbf{X}_{2i} + \epsilon_i,$$

where $\mathbf{X}_{1i}$ is 1 by k_1 vector of included regressors and $\mathbf{X}_{2i}$ is a 1 by k_2 vector of excluded but relevant regressors.

Omitting $\mathbf{x}_{2i}$ from the fit model, the least-squares estimator is

$$\hat{\beta}_{1n} = \left(\frac{\mathbf{X}_1' \mathbf{X}_1}{n} \right)^{-1} \frac{\mathbf{X}_1' \mathbf{y}}{n},$$

This misspecified estimator is biased, and the bias depends on the magnitude of the coefficients on the omitted variables and the correlation between the omitted and the excluded regressors.

Omitted variable bias

Theorem (4.6. Misspecified regression)

Under linearity and assumptions 4.1-4.3, if $\mathbf{X}$ *can be partitioned* $[\mathbf{X}_1,\ \mathbf{X}_2]$ *where* $\mathbf{X}_1$ *correspond to included variables while* $\mathbf{X}_2$ *correspond to the excluded variables with non-zero coefficients, then*

$$\hat{\beta}_{1n} \xrightarrow{p} \beta_1 + \delta\beta_2,$$

where $\delta = \Sigma_{x_1x_1}^{-1}\Sigma_{x_1x_2}$*, with* $\Sigma_{x_1x_1} = \frac{1}{n}\sum_{i=1}^{n} x_{1i}'x_{1i}$ *and* $\Sigma_{x_1x_2} = \frac{1}{n}\sum_{i=1}^{n} x_{1i}'x_{2i}$.

The bias term, $\delta\beta_2$ is composed of two elements:

- The first, δ is a matrix of regression coefficients where the jth column is the probability limit of the least-squares estimator in the regression $\mathbf{X}_{2j} = \mathbf{X}_1\delta_j + v$, where $\mathbf{X}_{2j}$ is the jth column of $\mathbf{X}_2$.
- The second component is the original regression coefficient.

Omitted variable bias

As should be expected, larger coefficients on omitted variables lead to larger bias; that is $\beta_{1n} \xrightarrow{p} \beta_1$ under one of three conditions:

- $\hat{\delta}_n \xrightarrow{p} 0$, i.e., no correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$.
- $\beta_2 = 0$, i.e., the model is correctly specified.
- The product $\hat{\delta}_n \beta_2 \xrightarrow{p} 0$.

β_2 has been assumed to be non-zero.

Including irrelevant variables

Let now consider the case where some irrelevant variables are included. The correct model specification is

$$Y_i = \mathbf{X}_{1i}\beta_1 + \epsilon_i,$$

and the model estimated is

$$Y_i = \mathbf{X}_{1i}\beta_1 + \mathbf{X}_{2i}\beta_2 + \epsilon_i,$$

As long as the assumptions of the asymptotic framework are satisfied, the least-squares estimator is consistent under theorem 4.1 and

$$\hat{\beta}_n \xrightarrow{p} \left[\begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right] = \left[\begin{array}{c} \beta_1 \\ 0 \end{array} \right],$$

However, the variance of $\hat{\beta}_{1n}$ is larger than $\Sigma_{_{11}}$ (the variance of β_1 is $\mathbf{X}_2$ was not included). That is, including irrelevant predictors increase the noise, meaning parameter estimates are less precise.

Heteroskedasticity

The assumption of stationarity and ergodicity does not require data to be homoskedastic. This is key since heteroskedasticity is the rule rather than the exception in financial data.

If the data are homoskedastic, the asymptotic covariance of $\hat{\beta}$ can be consistently estimated by

$$\hat{\mathbf{S}} = \hat{\sigma}^2 \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1},$$

However, heteroskedastic errors require the use of a more complicated covariance estimator, and the asymptotic variance can be consistently estimated using

$$\hat{\Sigma}_{XX}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{XX}^{-1} = n \left(\mathbf{X}'\mathbf{X} \right)^{-1} \left(\mathbf{X}' \hat{E} \mathbf{X} \right) \left(\mathbf{X}'\mathbf{X} \right)^{-1},$$

where $\hat{\mathbf{E}} = \operatorname{diag}\left(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_n^2\right)$ is a matrix with the estimated residuals along its diagonal.

Heteroskedasticity

Despite the presence of heteroskedasticity or not, it may be tempting to rely exclusively on the robust estimator which operates under minimal assumptions.

The covariance estimator $\hat{\Sigma}_{XX}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{XX}^{-1}$ is know as the White heteroskedasticity consistent covariance estimator and standard errors computed based on such estimator are called heteroskedasticity robust standard errors or White standard errors (White 1980).

White (1980) also provides a test to determine if a heteroskedasticity robust covariance estimator is required. The White's test is simply formulated as a regression of *squared* estimated residuals on all unique squares and cross products of $\mathbf{x}_i$.

White's test

The White's heteroskedasticity test works as follows:

1. Fit the model $Y_i = \mathbf{X}_i\beta + \epsilon_i$
2. Construct the fit residuals $\hat{\epsilon}_i = Y_i - \mathbf{X}_i\hat{\beta}$
3. Construct the auxiliary regressors $\mathbf{Z}_i$ where the $k(k+1)/2$ elements of $\mathbf{z}_i$ are computed from $X_{io}X_{ip}$ for
   $o = 1, 2, \ldots, k, \qquad p = o, o+1, \ldots, k$.
4. Estimate the auxiliary regression $\hat{\epsilon}_i^2 = \mathbf{Z}_i\gamma + \eta_i$
5. Compute White's test statistic as nR^2 where the R^2 is from the auxiliary regression
6. Compare the test statistic to the critical value at size α from a $\chi^2_{k(k+1)/2-1}$.

Generalised Least Squares

An alternative to model heteroskedastic data is to transform the data so that it is homoskedastic using the Generalised Least Squares (GLS).

GLS extends the OLS estimator to allow for arbitrary weighting matrices, and it is defined as

$$\hat{\beta}^{GLS} = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y},$$

for some positive matrix $\mathbf{W}$.

The full value of GLS is only realised when $\mathbf{W}$ is wisely chosen. In particular, assuming $\mathbf{W} = \mathbf{V} = V\left[\epsilon|\mathbf{X}\right]$, the GLS is BLUE.

Generalised Least Square

Theorem (4.6 Variance of $\hat{\beta}^{GLS}$)

Under linearity and assumption 4.6

$$V\left[\hat{\beta}^{GLS}|\mathbf{X}\right] = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1},$$

and $V\left[\hat{\beta}^{GLS}|\mathbf{X}\right] \leq V\left[\tilde{\beta}|\mathbf{X}\right]$ *where* $\tilde{\beta} = \mathbf{C}\mathbf{y}$ *is any other linear unbiased estimator with* $E\left[\tilde{\beta}\right] = \beta$*.*

The Feasible GLS (FGLS) is thus estimated in few steps:

1. Estimate $\hat{\beta}$ using OLS.
2. Using the estimated residuals $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$ estimate an auxiliary model by regressing the squared residuals on the variables of the variance model.
3. Using the estimated variance parameters $\hat{\omega}$ produce a fit variance $\hat{\mathbf{V}}$.
4. Compute $\tilde{\mathbf{y}} = \hat{\mathbf{V}}^{-1/2}\mathbf{y}$ and $\hat{\mathbf{X}} = \hat{\mathbf{V}}^{-1/2}\mathbf{X}$ compute $\hat{\beta}^{FGLS}$ using the OLS estimator on the transformed regressors and regressand.

Generalised Least Square

Hypothesis testing can be performed on $\hat{\beta}^{FGLS}$ using the standard test statistics with the FGLS variance estimator,

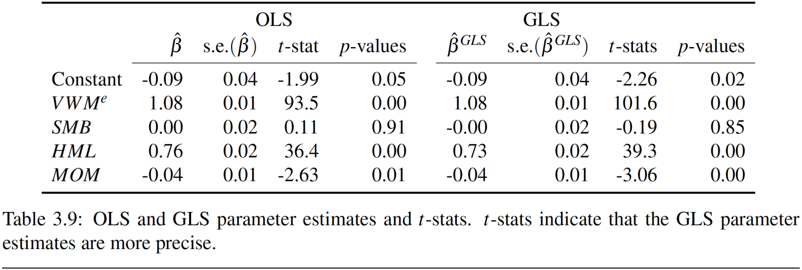
$$\tilde{\sigma}^2 \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1} = \tilde{\sigma}^2 \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1},$$

where $\tilde{\sigma}^2$ is the sample variance of the FGLS regression errors $\left(\tilde{\epsilon} - \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta}^{GLS}\right)$.

Two comments:

- While FGLS is only formally asymptotically justified, FGLS estimates are often much more precise in finite samples, especially if the data is very heteroskedastic.
- Even when estimated with a diagonal weighting matrix that may be slightly misspecified, the FGLS can produce substantially more precise estimates.

Example on factor pricing models



	OLS				GLS			
	$\hat{\beta}$	s.e.$(\hat{\beta})$	t-stat	p-values	$\hat{\beta}^{GLS}$	s.e.$(\hat{\beta}^{GLS})$	t-stats	p-values
Constant	-0.09	0.04	-1.99	0.05	-0.09	0.04	-2.26	0.02
VWM^e	1.08	0.01	93.5	0.00	1.08	0.01	101.6	0.00
SMB	0.00	0.02	0.11	0.91	-0.00	0.02	-0.19	0.85
HML	0.76	0.02	36.4	0.00	0.73	0.02	39.3	0.00
MOM	-0.04	0.01	-2.63	0.01	-0.04	0.01	-3.06	0.00

Table 3.9: OLS and GLS parameter estimates and t-stats. t-stats indicate that the GLS parameter estimates are more precise.

Source: Financial Econometrics Notes by Kevin Sheppard.

Model selection and specification checking

Model selection and specification

Econometric problems often begin with a variable whose dynamics are of interest and a relatively large set of candidate explanatory variables.

The process by which the set of regressors is reduced is known as model selection or building.

Model selection is often more an art than a science and some lessons can only be taught through experience.

One principle that should be universally applied when selecting a model is to rely on economic theory and, failing that, common sense.

There are few variable selection methods which can be examined for their properties. For now, we are going to focus on two methods:

- Information criteria (IC)
- Cross-validation

Information criteria

Information Criteria (IC) reward the model for producing smaller errors while pushing it for the inclusion of additional regressors. The two most used criteria are:

- Akaike Information Criterion (AIC)
- Schwarz/Bayesian Information Criterion (SIC)

Most information criteria are of the form

$$-2l + P$$

where l is the log-likelihood value at the parameter estimatrs and P is a penalty term.

In the case of least squares, where the log-likelihood is not known, IC's take the form $\ln \hat{\sigma}^2 + P$.

Information criteria

Definition (Akaike Information Criterion)

For likelihood-based models tha AIC is defined

$$AIC = -2l + 2k,$$

and its least squares application is

$$AIC = \ln \hat{\sigma}^2 + \frac{2k}{n},$$

Definition (Schwarz/Bayesian Information Criterion)

For likelihood-based models tha BIC (SIC) is defined

$$BIC = -2l + k \ln n,$$

and its least squares application is

$$BIC = \ln \hat{\sigma}^2 + k \frac{\ln n}{n},$$

Information criteria

The obvious difference between these two IC is that the AIC has a constant penalty term while the BIC has a penalty term that increases with the number of observations.

The effect of the sharper penalty in the BIC is that for larger data sizes, the marginal increase in the likelihood (or decrease in the variance) must be greater.

Information criteria to choose the model specification can be used in two ways:

General-to-Specific: If the number of regressors is not too large (e.g., less than 20) it is possible to try every possible combination and choose the smallest IC. This requires 2^L regressions where L is the number of available regressors.

Specific-to-General: Begin with the smallest model (usually a constant) and increase the model space. If any *reduce* the IC, extend the specification to include the variable that produced the smallest IC.

Cross-validation

Cross-validation uses pseudo-out-of-sample prediction performance to assess model specification.

It is most commonly used to select a preferred model from a set of candidate models.

The idea is that variables with robust predictive power should be useful both in- and out-of-sample.

Cross-validation estimates parameters using a random subset of the data and then computes the pseudo-out-of-sample sum of squared residuals on the observations not used in the estimation.

Such criterion rewards models include variables with good predictive power and exclude models that incorporate variables with small coefficients that do not improve the out-of-sample prediction.

Cross-validation

A k-fold cross-validation is made as follows:

1. Split the data randomly into the k-equal-sized bins.
2. For each model $m = 1, \dots M$ under construction
   - for $i = 1, \dots, k$
     - Estimate model parameters excluding the observations in the block i,

       $$\hat{\beta}_{m,i} = \arg \underset{\beta}{m, i} \sum_{j=1, j \notin \mathcal{B}_i} \left(Y_j - \mathbf{x}_{m,j} \beta_{m,i} \right)^2 ,$$

       where $\mathbf{x}_m$ are the regressors included in model m and $\mathcal{B}_i$ is the set of observation indices in block i.
     - Compute the block i SSE as $SSE_{m,i} = \sum_{j \in \mathcal{B}_i} \left(Y_j - \mathbf{x}_{m,j} \hat{\beta}_{m,i} \right)^2$
   - Compute the overall cross-validated SSE as $SSE_{m,CV} = \sum_{i=1}^{k} SSE_{m,i}$.
3. Select the model that produces the smallest cross-validates SSE.

Specification checking

Once a model has been selected, the final step is to examine the specification, where a number of issues may rise.

For example,:

- A model may have been neglected some nonlinear features in the data
- A few outliers may be determining the parameter estimates
- The data may be heteroskedastic (see previous slides).

Residuals form the basis of most specification checks, although the first step in assessing model fit is always to plot the residuals.

Residuals plots and non-linearity

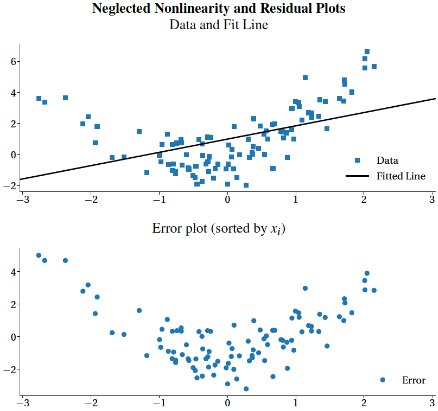


Figure 3.7: The top panel contains data generated according to $Y_i = X_i + X_i^2 + \varepsilon_i$ and a fit from a model $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$. The nonlinearity should be obvious, but is even clearer in the ordered (by X_i) residual plot where a distinct "U“ shape can be seen (bottom panel).

Source: Financial Econometrics Notes by Kevin Sheppard.

Residuals plots and non-linearity

A statistical test for detecting neglected nonlinearity is Ramsey's RESET test.

Suppose the model $Y_i = \mathbf{X}_i\beta + \epsilon_i$ is fit and one desires to test whether there is a neglected nonlinearity present.

The RESET test uses powers of the fit data, $\hat{Y}_i$ as additional regressors to test whether there is evidence of nonlinearity in the data.

Definition (Ramsey's RESET test)

The RESET test is a test of the null $H_0 : \gamma_1 = \ldots = \gamma_R = 0$ in an auxiliary regression.

$$Y_i = \mathbf{X}_i\beta + \gamma_1\hat{Y}_i^2 + \gamma_2\hat{Y}_i^3 + \ldots + \gamma_R\hat{Y}_i^{R-1} + \epsilon_i,$$

where $\hat{Y}_i$ are the fit values of Y_i generated in the initial regression. The test statistic has an asymptotic χ^2_R distribution.

The biggest problem with the RESET test is that rejection of the null is not informative about the changes needed in the original model specification.

Parameter stability

Parameter instability is a common problem in actual data.

For example, let us consider a simple CAPM

$$R_i^e = \beta_0 + \beta_1 \left(R_i^M - R_i^f \right) + \epsilon_i,$$

where R_i^M is the return on the market, R_i^f is the return on the risk free asset and R_i^e is the excess return on the dependent asset.

For example, recent evidence suggests that the market β in a CAPM may be different across up and down markets (see Ang, Chen and Xing 2006).

As a result, a model fit assuming the market beta in a CAPM is constant would be mispecified.

Parameter stability

A simple test to check whether the slope is different across different market cycles work as follows:

- Define a dummy variable $\mathbf{I}_i$ that takes value 1 if $\left(R_i^M - R_i^f\right) < 0$, and 0 otherwise.
- Estimate the regression

$$R_i^e = \beta_0 + \beta_1 \left(R_i^M - R_i^f\right) + \beta_2 I_i \left(R_i^M - R_i^f\right) + \epsilon_i,$$

- Test the null hypothesis $H_0 : \beta_2 = 0$ in the regression.

Alternatives could be testing if the parameters are different in recessions vs expansions, high vs low volatility periods, etc.

Parameter stability

Rolling and recursive parameter estimates are also useful tools for detecting parameter instability in regression analysis.

The idea of rolling regressions is simple; we use a **fixed-length sample** of data to estimate β and then “roll” the sampling window to produce a sequence of estimates.

Definition (m-sample Rolling Regression Estimates)

The m-sample rolling regression estimates are defined as the sequence:

$$\hat{\beta}_j = \left(\sum_{i=j}^{j+m-1} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \mathbf{x}_i' Y_i,$$

for $j = 1, 2, \ldots, n - m + 1$.

The rolling window m should be large enough so that parameter estimates are reasonably well approximated by a CLT but not so long as to smooth out any variation in β. 60-months (3-months and 2-year) is a common window length in applications using monthly (daily) asset price data.

Parameter stability

An alternative to rolling regressions is to recursively estimate parameters which uses an **expanding window** of observations to estimate $\hat{\beta}$

Definition (Recursive Regression Estimates)

Recursive regression estimates are defined as the sequence:

$$\hat{\beta}_j = \left(\sum_{i=1}^{j} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \mathbf{x}_i' Y_i,$$

for $j = l, 2, \ldots, n$ where $l > k$ is the smallest window used.

Documenting evidence of parameter instability using recursive estimates is often more difficult than with rolling since OLS estimates tend to be smoother as the sample size increases.

Example on factor pricing models

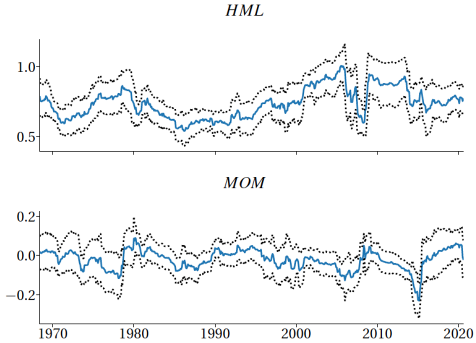


Figure 3.8: 60-month rolling parameter estimates from the model $BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$. Approximate confidence intervals were constructed by scaling the full sample parameter covariance. These rolling estimates indicate that the market loading of the Big-High portfolio varied substantially at the beginning of the samplefixed-length sample and that the loadings on both SMB and HML may be time-varying.

Source: Financial Econometrics Notes by Kevin Sheppard.

Normality

Normality may be a concern if the validity of the small-sample assumption is important.

The standard method to test for normality of estimated residuals is the Jarque-Bera (JB) test, which is based on two higher order moments, namely **skewness** and **kurtosis**.

Skewness and kurtosis are defined as

$$sk = \frac{n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^3}{\left(\hat{\sigma}^2\right)^{3/2}}, \qquad k = \frac{n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^4}{\left(\hat{\sigma}^2\right)^2},$$

The JB test is then defined as

$$JB = \frac{n}{6}\left(sk^2 + \frac{1}{4}\left(k-3\right)^2\right),$$

and is distributed χ_2^2, i.e., if $JB > C_\alpha$ we reject the null of normality (with C_α the critical value from a χ_2^2.

Exercises

Exercises

Problem 4.1: What are the information criteria and how they are used?

Problem 4.2: Discuss the White's covariance estimator, and in particular when should White's covariance estimator be used? What are the consequences to using White's covariance estimator when it is not needed? How can one determine if White's covariance estimator is needed?

Problem 4.3: Describe the steps to implement k-fold cross-validation in a regression to select a model.

Exercises

Problem 4.4: Consider the simple regression model $Y_i = \beta X_{1,i} + \epsilon_i$ where the random error terms are i.i.d. with mean zero and variance σ^2 and are uncorrelated with the $X_{1,i}$;

1. Show that the OLS estimator of β is consistent.
2. Now suppose the data generating process is

   $$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$

   derive the OLS estimator of β_1 and β_2.
3. What would be the bias of β?

Problem 4.5: Suppose an unrestricted model is

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + +\beta_3 X_{3,i} + \epsilon_i$$

Sketch the steps required to test a null $H_0 : \beta_1 = \beta_2 = 0$ in the large-sample framework using a Wald test, a t-test and an LR test.