6 Regression with non-i.i.d. errors

As discussed in section 4.2.2, if the regression errors are i.i.d., OLS produces consistent estimates; the large-sample distribution in large samples is normal with a mean at the true coefficient values, and the VCE is consistently estimated by (4.15). If the zero-conditional-mean assumption holds but the errors are not i.i.d., OLS produces consistent estimates whose sampling distribution in large samples is still normal with a mean at the true coefficient values but whose VCE cannot be consistently estimated by (4.15).

We have two options when the errors are not i.i.d. First, we can use the consistent OLS point estimates with a different estimator of the VCE that accounts for non-i.i.d. errors. Or, if we can specify how the errors deviate from i.i.d. in our regression model, we can use a different estimator that produces consistent and more efficient point estimates.

The tradeoff between these two methods is *robustness* versus *efficiency*. A robust approach places fewer restrictions on the estimator: the idea is that the consistent point estimates are good enough, although we must correct our estimator of their VCE to account for non-i.i.d. errors. The efficient approach incorporates an explicit specification of the non-i.i.d. distribution into the model. If this specification is appropriate, the additional restrictions it implies will produce a more efficient estimator than that of the robust approach.

The i.i.d. assumption fails when the errors are either not *identically* distributed or not *independently* distributed (or both). When the variance of the errors, conditional on the regressors, changes over the observations, the identically distributed assumption fails. This problem is known as *heteroskedasticity* (unequal variance), with its opposite being *homoskedasticity* (common variance). The i.i.d. case assumes that the errors are *conditionally homoskedastic*: there is no information in the regressors about the variance of the disturbances.

When the errors are correlated with each other, they are not *independently* distributed. In this chapter, we allow the errors to be correlated with *each other* (violating the i.i.d. assumption) but not with the regressors. We still maintain the zero-conditional-mean assumption, which implies that there is no correlation between the regressors and the errors. The case of nonindependent errors is different from the case in which the regressors are correlated with the errors.

After introducing some common causes for failure of the assumption of i.i.d. errors, we present the robust approach. We then discuss the general form of the efficient approach, the estimation and testing in the most common special cases, and the testing for i.i.d. errors in these subsections because efficient tests require that we specify the form of the deviation.

6.1 The generalized linear regression model

The popularity of the least-squares regression technique is linked to its generality. If we have a model linking a response variable to several regressors that satisfy the zero-conditional-mean assumption of (5.1), OLS will yield consistent point estimates of the β parameters. We need not make any further assumptions on the distribution of the u process and specifically need not assume that it is distributed multivariate normal.[1]

Here I present the *generalized linear regression model* (GLRM) that lets us consider the consequences of non-i.i.d. errors on the estimated covariance matrix of the estimated parameters $\widehat{\boldsymbol{\beta}}$. The GLRM is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{u} \\ E[\mathbf{u}|\mathbf{X}] &= \mathbf{0} \\ E[\mathbf{u}\mathbf{u}'|\mathbf{X}] &= \boldsymbol{\Sigma}_u \end{aligned}$$

where $\boldsymbol{\Sigma}_u$ is a positive-definite matrix of order $N \times N$.[2] This is a generalization of the i.i.d. error model in which $\boldsymbol{\Sigma}_u = \sigma^2 I_N$.

When $\boldsymbol{\Sigma}_u \neq \sigma^2 I_N$ the OLS estimator of β is still unbiased, consistent, and normally distributed in large samples, but it is no longer efficient, as demonstrated by

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ E[\widehat{\boldsymbol{\beta}} - \beta] &= 0 \end{aligned}$$

given the assumption of zero-conditional mean of the errors. That assumption implies that the sampling variance of the linear regression estimator (conditioned on $\mathbf{X}$) will be

$$\begin{aligned} \mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Sigma}_u\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{6.1}$$

The VCE computed by `regress` is merely $\sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$, with σ_u^2 replaced by its estimate, s^2.

When $\boldsymbol{\Sigma}_u \neq \sigma^2 I_N$, this simple estimator of the VCE is not consistent and the usual inference procedures are inappropriate. Hypothesis tests and confidence intervals using the simple estimator of the VCE after `regress` will not be reliable.

6.1.1 Types of deviations from i.i.d. errors

The GLRM lets us consider models in which $\boldsymbol{\Sigma}_u \neq \sigma^2 I_N$. Three special cases are of interest. First, consider the case of pure *heteroskedasticity* in which $\boldsymbol{\Sigma}_u$ is a diagonal

1. We do need to place some restrictions on the higher moments of u. But we can safely ignore those technical regularity conditions.

2. $\mathbf{y}$ is an $N \times 1$ vector of observations on y, $\mathbf{X}$ is an $N \times K$ matrix of observations on $\mathbf{x}$, and $\mathbf{u}$ is an $N \times 1$ disturbance vector.

matrix. This case violates the *identically distributed* assumption. When the diagonal elements of $\mathbf{\Sigma}_u$ differ, as in

$$\mathbf{\Sigma}_u = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \sigma_N^2 \end{pmatrix}$$

the model allows the variance of u, conditional on $\mathbf{X}$, to vary across the observations. For instance, using a household survey, we could model consumer expenditures as a function of household income. We might expect the error variance of high-income individuals to be much greater than that of low-income individuals because high-income individuals have much more discretionary income.

Second, we can separate the observations into several groups or *clusters* within which the errors are correlated. For example, when we are modeling households' expenditures on housing as a function of their income, the errors may be correlated over the households within a neighborhood.

Clustering—correlation of errors within a cluster of observations—causes the $\mathbf{\Sigma}_u$ matrix to be *block-diagonal* because the errors in different groups are independent of one another. This case drops the *independently distributed* assumption in a particular way. Since each cluster of observations may have its own error variance, the *identically distributed* assumption is relaxed as well.

$$\mathbf{\Sigma}_u = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} & & & \mathbf{0} \\ \mathbf{0} & \ddots & & & \\ & & \mathbf{\Sigma}_m & & \\ & & & \ddots & \mathbf{0} \\ \mathbf{0} & & & \mathbf{0} & \mathbf{\Sigma}_M \end{pmatrix} \tag{6.2}$$

In this notation, $\mathbf{\Sigma}_m$ represents an intracluster covariance matrix. For cluster (group) m with τ_m observations, $\mathbf{\Sigma}_m$ will be $\tau_m \times \tau_m$. Zero covariance between observations in the M different clusters gives the covariance matrix $\mathbf{\Sigma}_u$ a block-diagonal form.

Third, the errors in time-series regression models may show *serial correlation*, in which the errors are correlated with their predecessor and successor. In the presence of serial correlation, the error covariance matrix becomes

$$\mathbf{\Sigma}_u = \sigma_u^2 \begin{pmatrix} 1 & \rho_1 & \dots & \rho_{N-1} \\ \rho_1 & 1 & \dots & \rho_{2N-3} \\ \vdots & & \ddots & \\ \rho_{N-1} & \rho_{2N-3} & \dots & 1 \end{pmatrix} \tag{6.3}$$

where the unknown parameters $\rho_1, \rho_2, \dots, \rho_{\{N(N-1)\}/2}$ represent the correlations between successive elements of the error process. This case also drops the *independently distributed* assumption but parameterizes the correlations differently.

6.1.2 The robust estimator of the VCE

If the errors are conditionally heteroskedastic and we want to apply the robust approach, we use the Huber–White–sandwich estimator of the variance of the linear regression estimator. Huber and White independently derived this estimator, and the *sandwich* aspect helps you understand the robust approach. We need to estimate $\mathrm{Var}[\widehat{\boldsymbol{\beta}}|X]$, which according to (6.1) is of the form

$$\begin{aligned}\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Sigma}_u\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{6.4}$$

The term that we must estimate, $\{\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X}\}$, is sandwiched between the $(\mathbf{X}'\mathbf{X})^{-1}$ terms. Huber (1967) and White (1980) showed that

$$\widehat{S}_0 = \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{x}_i'\mathbf{x}_i \tag{6.5}$$

consistently estimates $(\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X})$ when the u_i are conditionally heteroskedastic. In this expression, $\widehat{u}_i$ is the ith regression residual, and $\mathbf{x}_i$ is the ith row of the regressor matrix: a $1 \times k$ vector of sample values. Substituting the consistent estimator from (6.5) for its population equivalent in (6.4) yields the robust estimator of the VCE[3]

$$\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{N}{N-k}(\mathbf{x}'\mathbf{x})^{-1}\left(\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{X}_i'\mathbf{X}_i\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{6.6}$$

The `robust` option available with most Stata estimation commands, including the `regress` command, implements the sandwich estimator described above. Calculating robust standard errors affects only the coefficients' standard errors and interval estimates and does not affect the point estimates $\widehat{\boldsymbol{\beta}}$. The ANOVA F table will be suppressed, as will the adjusted R^2 measure because neither is valid when robust standard errors are being computed. The title `Robust` will be displayed above the standard errors of the coefficients to remind you that a robust estimator of the VCE is in use. After `regress`, Wald tests produced by `test` and `lincom`, which use the robust estimator of the VCE, will be robust to conditional heteroskedasticity of unknown form.[4] See [U] **20.14 Obtaining robust variance estimates** for more detail.

If the assumption of homoskedasticity is valid, the simple estimator of the VCE is more efficient than the robust version. If we are working with a sample of modest size and the assumption of homoskedasticity is tenable, we should rely on the simple estimator of the VCE. But because the robust estimator of the VCE is easily calculated in

3. There is no error in (6.6). As in the appendix to chapter 4, we define $\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$ to be a large-sample approximation to the variance of our estimator. The large-sample calculations cause the $1/N$ factor in (6.5) to be dropped from (6.6). The factor $N/(N-k)$ improves the approximation in small samples.

4. Davidson and MacKinnon (1993) recommend using a different divisor that improves the performance of the robust estimator of the VCE estimator in small samples. Specifying the `hc3` option on `regress` will produce this robust estimator of the VCE.

Stata, it is simple to estimate both VCEs for a particular equation and consider whether inference based on the nonrobust standard errors is fragile. In large datasets, it has become increasingly common to report results using the robust estimator of the VCE.

To illustrate the use of the robust estimator of the VCE, we use a dataset (`fertil2`) that contains data on 4,361 women from a developing country. We want to model the number of children ever born (`ceb`) to each woman based on their `age`, their age at first birth (`agefbrth`), and an indicator of whether they regularly use a method of contraception (`usemeth`).[5] I present the descriptive statistics for the dataset with `summarize` based on those observations with complete data for a regression:

```
. use http://www.stata-press.com/data/imeus/fertil2, clear
. quietly regress ceb age agefbrth usemeth
. estimates store nonRobust
. summarize ceb age agefbrth usemeth children if e(sample)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ceb	3213	3.230003	2.236836	1	13
age	3213	29.93931	7.920432	15	49
agefbrth	3213	19.00498	3.098121	10	38
usemeth	3213	.6791161	.4668889	0	1
children	3213	2.999378	2.055579	0	13

The average woman in the sample is 30 years old, first bore a child at 19, and has had 3.2 children, with just under three children in the household. We expect that the number of children ever born is increasing in the mother's current age and decreasing in her age at the birth of her first child. The use of contraceptives is expected to decrease the number of children ever born.

For later use, we use `estimates store` to preserve the results of this (undisplayed) regression. We then fit the same model with a robust estimator of the VCE, saving those results with `estimates store`. We then use the `estimates table` command to display the two sets of coefficient estimates, standard errors, and t statistics.

```
. regress ceb age agefbrth usemeth, robust
Linear regression                               Number of obs =    3213
                                                F(  3,  3209) =  874.06
                                                Prob > F      =  0.0000
                                                R-squared     =  0.5726
                                                Root MSE      =   1.463
```

ceb	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.2237368	.0046619	47.99	0.000	.2145962	.2328775
agefbrth	-.2606634	.0095616	-27.26	0.000	-.2794109	-.2419159
usemeth	.1873702	.0606446	3.09	0.002	.0684642	.3062762
_cons	1.358134	.1675624	8.11	0.000	1.029593	1.686674

```
. estimates store Robust
```

5. Since the dependent variable is an integer, this model would be properly fitted with Poisson regression. For pedagogical reasons, we use linear regression.

```
. estimates table nonRobust Robust, b(%9.4f) se(%5.3f) t(%5.2f)
> title(Estimates of CEB with OLS and Robust standard errors)
Estimates of CEB with OLS and Robust standard errors
```

Variable	nonRobust	Robust
age	0.2237	0.2237
	0.003	0.005
	64.89	47.99
agefbrth	-0.2607	-0.2607
	0.009	0.010
	-29.64	-27.26
usemeth	0.1874	0.1874
	0.055	0.061
	3.38	3.09
_cons	1.3581	1.3581
	0.174	0.168
	7.82	8.11

legend: b/se/t

Our prior results are borne out by the estimates, although the effect of contraceptive use appears to be marginally significant. The robust estimates of the standard errors are similar to the nonrobust estimates, suggesting that there is no conditional heteroskedasticity.

6.1.3 The cluster estimator of the VCE

Stata has implemented an estimator of the VCE that is robust to the correlation of disturbances within groups and to not identically distributed disturbances. It is commonly referred to as the cluster–robust–VCE estimator, because these groups are known as clusters. Within-cluster correlation allows the $\boldsymbol{\Sigma}_u$ in (6.2) to be *block-diagonal*, with nonzero elements within each block on the diagonal. This block-diagonal structure allows the disturbances within each cluster to be correlated with each other but requires that the disturbances from difference clusters be uncorrelated.

If the within-cluster correlations are meaningful, ignoring them leads to inconsistent estimates of the VCE. Since the `robust` estimate of the VCE assumes independently distributed errors, its estimate of $(\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X})$ is not consistent. Stata's `cluster()` option, available on most estimation commands including `regress`, lets you account for such an error structure. Like the `robust` option (which it encompasses), application of the `cluster()` option does not affect the point estimates but only modifies the estimated VCE of the estimated parameters. The `cluster()` option requires you to specify a group- or cluster-membership variable that indicates how the observations are grouped.

The cluster–robust–VCE estimator is

$$\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{N-1}{N-k}\frac{M}{M-1}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{j=1}^{M}\widetilde{\mathbf{u}}_j'\widetilde{\mathbf{u}}_j\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{6.7}$$

where M is the number of clusters, $\widetilde{\mathbf{u}}_j = \sum_{i=1}^{N_k}\widehat{u}_i\mathbf{x}_i$, N_j is the number of observations in the jth cluster, $\widehat{u}_i$ is the ith residual from the jth cluster, and $\mathbf{x}_i$ is the $1 \times k$ vector of regressors from the ith observation in the jth cluster.

Equation (6.7) has the same form as (6.6). Aside from the small-sample adjustments, the (6.7) differs from (6.6) only in that the "meat" of the sandwich is now the cluster–robust estimator of $(\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X})$.

The goal of the robust and the cluster–robust–VCE estimators is to consistently estimate the $\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$ in the presence of non-i.i.d. disturbances. Different violations of the i.i.d. disturbance assumption simply require distinct estimators of $(\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X})$.

To illustrate the use of the cluster estimator of the covariance matrix, we revisit the model of fertility in a developing country that we estimated above via nonrobust and robust methods. The clustering variable is `children`: the number of living children in the household. We expect the errors from households of similar size to be correlated, while independent of those generated by households of different size.

```
. regress ceb age agefbrth usemeth, cluster(children)
Linear regression                               Number of obs =    3213
                                                F(  3,    13) =   20.91
                                                Prob > F      =  0.0000
                                                R-squared     =  0.5726
Number of clusters (children) = 14              Root MSE      =   1.463

------------------------------------------------------------------------------
             |               Robust
         ceb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2237368   .0315086     7.10   0.000     .1556665    .2918071
    agefbrth |  -.2606634   .0354296    -7.36   0.000    -.3372045   -.1841224
    usemeth |   .1873702   .0943553     1.99   0.069     -.016472    .3912125
       _cons |   1.358134   .4248589     3.20   0.007     .4402818    2.275985
------------------------------------------------------------------------------
```

The cluster estimator, allowing for within-cluster correlation of errors, results in much more conservative standard errors (and smaller t statistics) than those displayed in the previous example.

6.1.4 The Newey–West estimator of the VCE

In the presence of heteroskedasticity and autocorrelation, we can use the *Newey–West* estimator of the VCE. This heteroskedastic and autocorrelation consistent (HAC) estimator of the VCE has the same form as the robust and cluster–robust estimators, but it uses a distinct estimator for $(\mathbf{X}'E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\mathbf{X})$. Rather than specifying a cluster variable,

the Newey–West estimator requires that we specify the maximum order of any significant autocorrelation in the disturbance process—known as the maximum lag, denoted by L.

In addition to the term that adjusts for heteroskedasticity, the estimator proposed by Newey and West (1987) uses weighted cross products of the residuals to account for autocorrelation:

$$\widehat{\mathbf{Q}} = \widehat{\mathbf{S}}_0 + \frac{1}{T}\sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l\, \widehat{u}_t\widehat{u}_{t-l}(\mathbf{x}_t'\mathbf{x}_{t-l} + \mathbf{x}_{t+l}'\mathbf{x}_t)$$

Here $\widehat{\mathbf{S}}_0$ is the robust estimator of the VCE from (6.5), $\widehat{u}_t$ is the tth residual, and $\mathbf{x}_t$ is the tth row of the regressor matrix. The Newey–West formula takes a specified number (L) of the sample autocorrelations into account, using the Bartlett kernel estimator,

$$w_l = 1 - \frac{l}{L+1}$$

to generate the weights.

The estimator is said to be HAC, allowing for any deviation of $\mathbf{\Sigma}_u$ from $\sigma_u^2 I$ up to Lth-order autocorrelation. The user must specify her choice of L, which should be large enough to encompass any likely autocorrelation in the error process. One rule of thumb is to choose $L = \sqrt[4]{N}$. This estimator is available in the Stata command `newey` (see [TS] **newey**), which you can use as an alternative to `regress` to estimate a regression with HAC standard errors. This command has the following syntax,

`newey` *depvar* [*indepvars*] [*if*] [*in*] , `lag(#)`

where the number given for the `lag()` option is L above. Like the `robust` option, the HAC estimator does not modify the point estimates; it affects only the estimator of the VCE. Test statistics based on the HAC VCE are robust to arbitrary heteroskedasticity and autocorrelation.

We illustrate this estimator of the VCE by using a time-series dataset of monthly short-term and long-term interest rates on U.K. government securities (Treasury bills and gilts), 1952m3–1995m12. The descriptive statistics for those series are given by `summarize`:

```
. use http://www.stata-press.com/data/imeus/ukrates, clear
. summarize rs r20
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          rs |       526    7.651513    3.553109   1.561667      16.18
         r20 |       526    8.863726    3.224372       3.35      17.18
```

The model expresses the monthly change in the short rate `rs`, the Bank of England's monetary policy instrument, as a function of the prior month's change in the long-term

rate r20. The regressor and regressand are created on the fly by Stata's time-series operators D. and L. The model represents a monetary policy reaction function.

We fit the model with and without HAC standard errors by using regress and newey, respectively, using estimates store to save the results and estimates table to juxtapose them. Since there are 524 observations, the rule of thumb for lag selection recommends five lags, which we specify in newey's lag() option.

```
. quietly regress D.rs LD.r20
. estimates store nonHAC
. newey D.rs LD.r20, lag(5)
Regression with Newey-West standard errors          Number of obs   =       524
maximum lag: 5                                      F(  1,   522)   =     36.00
                                                    Prob > F        =    0.0000
```

D.rs	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
r20						
LD.	.4882883	.0813867	6.00	0.000	.3284026	.648174
_cons	.0040183	.0254102	0.16	0.874	-.0459004	.0539371

```
. estimates store NeweyWest
. estimates table nonHAC NeweyWest, b(%9.4f) se(%5.3f) t(%5.2f)
> title(Estimates of D.rs with OLS and Newey--West standard errors)
Estimates of D.rs with OLS and Newey--West standard errors
```

Variable	nonHAC	NeweyWest
LD.r20	0.4883	0.4883
	0.067	0.081
	7.27	6.00
_cons	0.0040	0.0040
	0.022	0.025
	0.18	0.16

legend: b/se/t

The HAC standard error estimate of the slope coefficient from newey is larger than that produced by regress, although the coefficient retains its significance.

Two issues remain with this HAC VCE estimator. First, although the Newey–West estimator is widely used, there is no justification for using the Bartlett kernel. We might use several alternative kernel estimators, and some may have better properties in specific instances. The only requirement is that the kernel delivers a positive-definite estimate of the VCE. Second, if there is no reason to question the assumption of homoskedasticity of u, we may want to deal with serial correlation under that assumption. We may want the AC without the H. The standard Newey–West procedure as implemented in newey does not allow for this. The ivreg2 routine (Baum, Schaffer, and Stillman 2003) can estimate robust, AC, and HAC standard errors for regression models, and it provides a choice of alternative kernels. See chapter 8 for full details on this routine.

6.1.5 The generalized least-squares estimator

This section presents a class of estimators for estimating the coefficients of a GLRM when the zero-conditional-mean assumption holds, but the errors are not i.i.d. Known as *feasible generalized least squares* (FGLS), this technique relies on the insight that if we knew $\mathbf{\Sigma}_u$, we could algebraically transform the data so that the resulting errors were i.i.d. and then proceed with linear regression on the transformed data. We do not know $\mathbf{\Sigma}_u$, though, so this estimator is infeasible. The *feasible* alternative requires that we assume a structure that describes how the errors deviate from i.i.d. errors. Given that assumption, we can consistently estimate $\mathbf{\Sigma}_u$. Any consistent estimator of $\mathbf{\Sigma}_u$ may be used to transform the data to generate observations with i.i.d. errors.

Although both the robust estimator of the VCE approach and FGLS estimators account for non-i.i.d. disturbances, FGLS estimators place more structure on the estimation method to obtain more efficient point estimates and consistent estimators of the VCE. In contrast, the robust estimator of the VCE approach uses just the OLS point estimates and makes the estimator of the VCE robust to the non-i.i.d. disturbances.

First, consider the infeasible GLS estimator of

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ E[\mathbf{u}\mathbf{u}'|\mathbf{X}] &= \mathbf{\Sigma}_u \end{aligned}$$

The known $N \times N$ matrix $\mathbf{\Sigma}_u$ is symmetric and positive definite, which implies that it has an inverse $\mathbf{\Sigma}_u^{-1} = \mathbf{P}\mathbf{P}'$, where $\mathbf{P}$ is a triangular matrix. Premultiplying the model by $\mathbf{P}'$ yields

$$\begin{aligned} \mathbf{P}'\mathbf{y} &= \mathbf{P}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}'\mathbf{u} \\ \mathbf{y}_* &= \mathbf{X}_*\boldsymbol{\beta} + \mathbf{u}_* \end{aligned} \tag{6.8}$$

where[6]

$$\text{Var}[\mathbf{u}_*] = E[\mathbf{u}_*\mathbf{u}_*'] = \mathbf{P}'\mathbf{\Sigma}_u\mathbf{P} = \mathbf{I}_N$$

With a known $\mathbf{\Sigma}_u$ matrix, regression of $\mathbf{y}_*$ on $\mathbf{X}_*$ is asymptotically efficient by the Gauss–Markov theorem presented in section 4.2.3. That estimator merely represents standard linear regression on the transformed data:

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}(\mathbf{X}_*'\mathbf{y}_*)$$

The VCE of the GLS estimator $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is

$$\text{Var}[\widehat{\boldsymbol{\beta}}_{\text{GLS}}|\mathbf{X}] = (\mathbf{X}'\mathbf{\Sigma}_u^{-1}\mathbf{X})^{-1}$$

6. $E[\mathbf{P}'\mathbf{u}\mathbf{u}'\mathbf{P}] = \mathbf{P}'E[\mathbf{u}\mathbf{u}']\mathbf{P} = \mathbf{P}'\mathbf{\Sigma}_u\mathbf{P}$. But that expression equals $\mathbf{P}'(\mathbf{P}\mathbf{P}')^{-1}\mathbf{P} = \mathbf{P}'(\mathbf{P}')^{-1}\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}_N$. See Davidson and MacKinnon (2004, 258).

The FGLS estimator

When $\mathbf{\Sigma}_u$ is unknown, we cannot apply the GLS estimator of (6.8). But if we have a consistent estimator of $\mathbf{\Sigma}_u$, denoted $\widehat{\mathbf{\Sigma}}_u$, we may apply the FGLS estimator, replacing $\mathbf{P}'$ with $\widehat{\mathbf{P}}'$ in (6.8). The FGLS estimator has the same large-sample properties as its infeasible counterpart.[7] That result does not depend on using an efficient estimator of $\mathbf{\Sigma}_u$, but merely any consistent estimator of $\mathbf{\Sigma}_u$.

The challenge in devising a consistent estimator of $\mathbf{\Sigma}_u$ lies in its dimension. $\mathbf{\Sigma}_u$ is a square symmetric matrix of order N with $\{N(N+1)\}/2$ distinct elements. Fortunately, the most common departures from i.i.d. errors lead to parameterizations of $\widehat{\mathbf{\Sigma}}_u$ with many fewer parameters. As I discuss in the next sections, heteroskedasticity and autocorrelation can often be modeled with a handful of parameters. All we need for consistency of these estimates is a fixed number of parameters in $\widehat{\mathbf{\Sigma}}_u$ as $N \to \infty$.

The gain from using FGLS depends on the degree to which $\mathbf{\Sigma}_u$ diverges from $\sigma^2 I_N$, the covariance matrix for i.i.d. errors. If that divergence is small, the FGLS estimates will be similar to those of standard linear regression, and vice versa.

The following two sections discuss the most common violations of the i.i.d. errors assumption—heteroskedasticity and serial correlation—and present the FGLS estimator appropriate for each case.

6.2 Heteroskedasticity in the error distribution

In cross-sectional datasets representing individuals, households, or firms, the disturbance variances are often related to some measure of scale. For instance, in modeling consumer expenditures, the disturbance for variance of high-income households is usually larger than that of poorer households. For the FGLS estimator described above, the diagonal elements of the $\mathbf{\Sigma}_u$ matrix for these errors will be related to that scale measure.

We may instead have a dataset in which we may reasonably assume that the disturbances are homoskedastic *within* groups of observations but potentially heteroskedastic *between* groups. For instance, in a labor market survey, self-employed individuals or workers paid by salary and commission (or salary and tips) may have a greater variance around their conditional-mean earnings than salaried workers. For the FGLS estimator, there will be several distinct values of σ_u^2, each common to those individuals in a group but differing between groups.

As a third potential cause of heteroskedasticity, consider the use of *grouped data*, in which each observation is the average of microdata (e.g., state-level data for the United States, where the states have widely differing populations). Since means computed from larger samples are more accurate, the disturbance variance for each observation is known up to a factor of proportionality. Here we are certain (by the nature of grouped

7. See Davidson and MacKinnon (2004).

data) that heteroskedasticity exists, and we can construct the appropriate $\widehat{\mathbf{\Sigma}}_u$. In the two former cases, we are not so fortunate.

We may also find heteroskedasticity in time-series data, especially *volatility clustering*, which appears in high-frequency financial-market data. I will not discuss this type of conditional heteroskedasticity at length, but the use of the autoregressive conditional heteroskedasticity (ARCH) and generalized ARCH (GARCH) models for high-frequency time-series data is based on the notion that the errors in these contexts are *conditionally* heteroskedastic and that the evolution of the conditional variance of the disturbance process may be modeled.[8]

6.2.1 Heteroskedasticity related to scale

We often use an economic rationale to argue that the variance of the disturbance process is related to some measure of *scale* of the individual observations. For instance, if the response variable measures expenditures on food by individual households, the disturbances will be denominated in dollars (or thousands of dollars). No matter how well the estimated equation fits, the dollar dispersion of wealthy households' errors around their predicted values will likely be much greater than those of low-income households.[9] Thus a hypothesis of

$$\sigma_i^2 \propto z_i^\alpha \tag{6.9}$$

is often made, where z_i is some scale-related measure for the ith unit. The notion of proportionality comes from the definition of FGLS: we need only estimate $\widehat{\mathbf{\Sigma}}_u$ up to a factor of proportionality. It does not matter whether z is one of the regressors or merely more information we have about each unit in the sample.

We write z_i^α in (6.9) since we must indicate the nature of this proportional relationship. For instance, if $\alpha = 2$, we are asserting that the *standard deviation* of the disturbance process is proportional to the level of z_i (e.g., to household income or a firm's total assets). If $\alpha = 1$, we imply that the *variance* of the disturbance process is proportional to the level of z_i, so that the standard deviation is proportional to $\sqrt{z_i}$.

Given a plausible choice of z_i, why is the specification of α so important? If we are to use FGLS to deal with heteroskedasticity, our choices of z_i and α in (6.9) will define the FGLS estimator to be used. Before I discuss correcting for heteroskedasticity related to scale, you must understand how to detect the presence of heteroskedasticity.

8. The development of ARCH models was a major factor in the award of the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel to Robert F. Engle in 2003. He shared the prize with fellow time-series econometrician Clive Granger. A bibliography of Engle's published and unpublished works may be found at http://ideas.repec.org/e/pen9.html.

9. With firm data, the same logic applies. If we are explaining a firm's capital investment expenditures, the degree to which spending differs from the model's predictions could be billions of dollars for a huge multinational but much smaller for a firm of modest size.

Testing for heteroskedasticity related to scale

After fitting a regression model, we can base a test for heteroskedasticity on the regression residuals. Why is this approach reasonable, if the presence of heteroskedasticity renders the standard errors unusable? The consistent point estimates $\widehat{\boldsymbol{\beta}}$ produce estimated residuals that may be used to make inferences about the distribution of u. If the assumption of homoskedasticity conditional on the regressors holds, it can be expressed as follows:

$$H_0 : \text{Var}\,[u|\mathbf{X}] = \sigma_u^2 \tag{6.10}$$

Under this null hypothesis the conditional variance of the error process does not depend on the explanatory variables. Given that $E[u] = 0$, this null hypothesis is equivalent to requiring that $E[u^2|\mathbf{X}] = \sigma_u^2$. The conditional mean of the squared disturbances should not be a function of the regressors, so a regression of the squared residuals on any candidate $\mathbf{z}_i$ should have no meaningful explanatory power.[10,11]

One of the most common tests for heteroskedasticity is derived from this line of reasoning: the *Breusch–Pagan* (BP) test (Breusch and Pagan 1979).[12] The BP test, an LM test, involves regressing the squared residuals on a set of variables in an auxiliary regression:[13]

$$\widehat{u}_i^2 = d_1 + d_2 z_{i2} + d_3 z_{i3} + \ldots\ d_\ell z_{i\ell} + v_i \tag{6.11}$$

We could use the original regressors from the fitted model as the $\mathbf{z}$ variables,[14] use a subset of them, or add measures of scale as discussed above. If the magnitude of the squared residual is not systematically related to any of the $\mathbf{z}$ variables, then this auxiliary regression will have no explanatory power. Its R^2 will be small, and its ANOVA F statistic will indicate that it fails to explain any meaningful fraction of the variation of $\widehat{u}_i^2$ around its own mean.[15]

The BP test can be conducted by using either the F or LM statistic from the auxiliary regression (6.11). Under the null hypothesis of (6.10), LM $\sim \chi^2_\ell$, where there are ℓ regressors in the auxiliary regression. We can obtain the BP test with `estat hettest` after `regress`. If no regressor list (of z's) is provided, `hettest` uses the fitted values from the previous regression (the $\widehat{y}_i$ values). As mentioned above, the variables specified in the set of z's could be chosen as measures that did not appear in the original regressor list.

10. $\mathbf{z}_i$ must be a function of the regressor.
11. $\mathbf{z}_i$ has been generalized to be a vector.
12. The Stata manuals document this test as that of Cook and Weisberg. Breusch and Pagan (1979), Godfrey (1978), and Cook and Weisberg (1983) separately derived (and published) the same test statistic. It should not be confused with a different test devised by Breusch and Pagan implemented in `sureg`.
13. An LM test statistic evaluates the results of a restricted regression model. In the BP test, the restrictions are those implied by homoskedasticity, which implies that the squared regression disturbances should be uncorrelated with any measured characteristics in the regression. For more details, see Wooldridge (2006, 185–186).
14. Although the residuals are uncorrelated by construction with each of the regressors of the original model, that condition need not hold for their squares.
15. Although the regression residuals from a model with a constant term have mean zero, the mean of their squares must be positive unless $R^2 = 1$.

The BP test with $\mathbf{z} = \mathbf{x}$ is a special case of *White's general test* (White 1980) for heteroskedasticity, which takes the list of regressors $(x_2, x_3, \ldots, x_k)$ and augments it with squares and cross products of each of these variables. The White test then runs an auxiliary regression of $\widehat{u}_i^2$ on the regressors, their squares, and their cross products, removing duplicate elements. For instance, if `crime` and `crime`-squared were in the original regression, only one instance of the squares term will enter the list of Zs. Under the null hypothesis, none of these variables should have any explanatory power for the squared residual series. The White test is another LM test of the $N \times R^2$ form but involves many more regressors in the auxiliary regression (especially for a regression in which k is sizable). The resulting test may have relatively low power because of the many degrees of freedom devoured by a lengthy regressor list. An alternate form of White's test uses only the fitted values of the original regression and their squares. We can compute both versions of White's test with `whitetst` as described in Baum, Cox, and Wiggins (2000), which you can install by using `ssc`. The original version of White's test may also be computed by the `estat imtest` command, using the `white` option.

All these tests rest on the specification of the disturbance variance expressed in (6.9). A failure to reject the tests' respective null hypotheses of homoskedasticity does not indicate an absence of heteroskedasticity but implies that the heteroskedasticity is not likely to be of the specified form. In particular, if the heteroskedasticity arises from group membership (as discussed in section 6.2.2), we would not expect tests based on measures of scale to pick it up unless there was a strong correlation between scale and group membership.[16]

We consider the potential scale-related heteroskedasticity in our model of median housing prices where the scale can be thought of as the average size of houses in each community, roughly measured by number of rooms. After fitting the model, we calculate three test statistics: that computed by `estat hettest, iid` without arguments, which is the BP test based on fitted values; `estat hettest, iid` with a variable list, which uses those variables in the auxiliary regression; and White's general test statistic from `whitetst`.[17]

```
. use http://www.stata-press.com/data/imeus/hprice2a, clear
(Housing price data for Boston-area communities)
. regress lprice rooms crime ldist
```

Source	SS	df	MS
Model	47.9496883	3	15.9832294
Residual	36.6325827	502	.072973272
Total	84.5822709	505	.167489645

Number of obs	= 506
F(3, 502)	= 219.03
Prob > F	= 0.0000
R-squared	= 0.5669
Adj R-squared	= 0.5643
Root MSE	= .27014

16. Many older textbooks discuss the Goldfeld–Quandt test, which is based on forming two groups of residuals defined by high and low values of one z variable. Because there is little to recommend this test relative to the BP or White test approaches, which allow for multiple z's, I do not discuss it further.

17. By default, `estat hettest` produces the original BP test, which assumes that the u_i are normally distributed. Typing `estat hettest, iid` yields the Koenker (1981) LM test, which assumes the u_i to be i.i.d. under the null hypothesis.

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rooms	.3072343	.0178231	17.24	0.000	.2722172	.3422514
crime	-.0174486	.001591	-10.97	0.000	-.0205744	-.0143228
ldist	.074858	.0255746	2.93	0.004	.0246115	.1251045
_cons	7.984449	.1128067	70.78	0.000	7.762817	8.20608

```
. estat hettest, iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lprice

         chi2(1)      =    44.67
         Prob > chi2  =   0.0000

. estat hettest rooms crime ldist, iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: rooms crime ldist

         chi2(3)      =    80.11
         Prob > chi2  =   0.0000

. whitetst

White's general test statistic :  144.0052  Chi-sq( 9)  P-value =  1.5e-26
```

Each of these tests indicates that there is a significant degree of heteroskedasticity in this model.

FGLS estimation

To use FGLS on a regression equation in which the disturbance process exhibits heteroskedasticity related to scale, we must estimate the $\mathbf{\Sigma}_u$ matrix up to a factor of proportionality. We implement FGLS by transforming the data and running a regression on the transformed equation. For FGLS to successfully deal with the deviation from i.i.d. errors, the transformations must purge the heteroskedasticity from the disturbance process and render the disturbance process in the transformed equation i.i.d.

Say that we test for this form of heteroskedasticity and conclude, per (6.9), that the disturbance variance of the ith firm is proportional to z_i^2, with z defined as a measure of scale related to the covariates in the model. We assume that z_i is strictly positive or that it has been transformed to be strictly positive. The appropriate transformation to induce homoskedastic errors would be to divide each variable in $(y, \mathbf{X})$ (including $\boldsymbol{\iota}$, the first column of $\mathbf{X}$) by z_i. That equation will have a disturbance term u_i/z_i, and since z_i is a constant, $\text{Var}[u_i/z_i] = (1/z_i^2)\text{Var}[u_i]$. If the original disturbance variance is proportional to z_i, dividing it by z_i^2 will generate a constant value: homoskedasticity of the transformed equation's error process.

We could implement FGLS on the equation

$$y_i = \beta_1 + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + u_i \tag{6.12}$$

by specifying the transformed equation

$$\frac{y_i}{z_i} = \frac{\beta_1}{z_i} + \frac{\beta_2 x_{i,2}}{z_i} + \cdots + \frac{\beta_k x_{i,k}}{z_i} + \frac{u_i}{z_i} \tag{6.13}$$

or

$$y_i^* = \beta_1 \iota^* + \beta_2 x_{i,2}^* + \cdots + \beta_k x_{i,k}^* + u_i^* \tag{6.14}$$

where $i^* = 1/z_i$. The economic meaning of the coefficients in the transformed equation has not changed; β_2 and its estimate $\widehat{\beta}_2$ still represent $\partial y/\partial x_2$. Since we have changed the dependent variable, measures such as R^2 and `Root MSE` are not comparable to those of the original equation. In particular, the transformed equation does not have a constant term.

Although we could do these transformations by hand with `generate` statements followed by `regress` on the transformed (6.14), that approach is cumbersome. For instance, we will normally want to evaluate measures of goodness of fit based on the original data, not the transformed data. Furthermore, the transformed variables can be confusing. For example, if z_i were also regressor x_2 in (6.12),[18] the $\mathbf{x}^*$ variables would include $1/z_i$ and $\boldsymbol{\iota}$, a units vector. The coefficient on the former is really an estimate of the constant term of the equation, whereas the coefficient labeled as `_cons` by Stata is actually the coefficient on z, which could become confusing.

Fortunately, we need not perform FGLS by hand. FGLS in a heteroskedastic context can be accomplished by *weighted least squares*. The transformations we have defined above amount to *weighting* each observation (here by $1/z_i$). Observations with smaller disturbance variances receive a larger weight in the computation of the sums and therefore have greater weight in computing the weighted least-squares estimates. We can instruct Stata to perform this weighting when it estimates the original regression by defining $1/z_i^2$ as the so-called *analytical weight*. Stata implements several kinds of weights (see [U] **11 Language syntax** and [U] **20.16 Weighted estimation**), and this sort of FGLS involves the analytical weight (`aw`) variety. We merely estimate the regression specifying the weights,

```
. generate rooms2 = rooms^2
. regress lprice rooms crime ldist [aw=1/rooms2]
(sum of wgt is   1.3317e+01)
```

Source	SS	df	MS		Number of obs =	506
					F(3, 502) =	159.98
Model	39.6051883	3	13.2017294		Prob > F =	0.0000
Residual	41.426616	502	.082523139		R-squared =	0.4888
					Adj R-squared =	0.4857
Total	81.0318042	505	.160459018		Root MSE =	.28727

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rooms	.2345368	.0194432	12.06	0.000	.1963367	.272737
crime	-.0175759	.0016248	-10.82	0.000	-.0207682	-.0143837
ldist	.0650916	.027514	2.37	0.018	.0110349	.1191483
_cons	8.450081	.1172977	72.04	0.000	8.219626	8.680536

18. We assume that z_i is strictly positive.

which indicates that the regression is to be performed using `1/rooms2` as the analytical weight,[19] where `rooms2` = rooms^2. These estimates are qualitatively similar to those obtained with `robust`, with slightly weaker measures of goodness of fit.

The coefficient estimates and standard errors from this weighted regression will be identical to those computed by hand if the $y^*, \mathbf{x}^*$ variables are generated. But unlike the regression output from (6.14), the regression with analytical weights produces the desired measures of goodness of fit (e.g., R^2 and `Root MSE`) and `predict` will generate predicted values or residuals in the units of the untransformed dependent variable. The FGLS *point estimates* differ from those generated by `regress` from the untransformed regression; see (6.12). However, both the standard regression and FGLS point estimates are consistent estimates of $\boldsymbol{\beta}$.

The series specified as the analytical weight (`aw`) must be the *inverse of the observation variance*, not its standard deviation, and the original data are multiplied by the analytical weight, not divided by it. Some other statistical packages that provide facilities for FGLS differ in how they specify the weighting variable, for instance, requiring you to provide the value that appears as the divisor in (6.13).

We often see empirical studies in which a regression equation has been specified in some ratio form. For instance, per capita dependent and independent variables for data on states or countries are often used, as are financial ratios for firm- or industry-level data. Although the study may not mention heteroskedasticity, these ratio forms probably have been chosen to limit the potential damage of heteroskedasticity in the fitted model. Heteroskedasticity in a per capita form regression on country-level data is much less likely to be a problem in that context than it would be if the levels of GDP were used in that model. Likewise, scaling firms' values by total assets, total revenues, or the number of employees can mitigate the difficulties caused by extremes in scale between large corporations and corner stores. Such models should still be examined for their errors' behavior, but the popularity of the ratio form in these instances is an implicit consideration of potential heteroskedasticity related to scale.

6.2.2 Heteroskedasticity between groups of observations

Between-group heteroskedasticity is often associated with *pooling* data across what may be nonidentically distributed sets of observations. For instance, a consumer survey conducted in Massachusetts (MA) and New Hampshire (NH) may give rise to a regression equation predicting the level of spending as a function of several likely factors. If we merely pool the sets of observations from MA and NH into one dataset (using `append`), we may want to test that any fitted model is *structurally stable* over the two states' observations: that is, are the same $\boldsymbol{\beta}$ parameters appropriate?[20] Even if the two states' observations share the same population parameter vector $\boldsymbol{\beta}$, they may have different σ_u^2 values. For instance, spending in MA may be more sensitive to the presence of sales tax on many nonfood items, whereas NH shoppers do not pay a sales tax. This difference

19. This is one of the rare instances in Stata syntax when the square brackets (`[]`) are used.
20. A discussion of testing for structural stability appears in section 7.4.

may affect not only the slope parameters of the model but also the error variance. If so, then the assumption of homoskedasticity is violated in a particular manner. We may argue that the intrastate (or more generally, intragroup) disturbance variance is constant but that it may differ *between* states (or groups).

This same situation may arise, as noted above, with other individual-level series. Earnings may be more variable for self-employed workers, or those who depend on commissions or tips than salaried workers. With firm data, we might expect that profits (or revenues or capital investment) might be much more variable in some industries than others. Capital-goods makers face a much more cyclical demand for their product than do, for example, electric utilities.

Testing for heteroskedasticity between groups of observations

How might we test for groupwise heteroskedasticity? With the assumption that each group's regression equation satisfies the classical assumptions (including that of homoskedasticity), the s^2 computed by `regress` is a consistent estimate of the group-specific variance of the disturbance process. For two groups, we can construct an F test, with the larger variance in the numerator; the degrees of freedom are the residual degrees of freedom of each group's regression. We can easily construct such a test if both groups' residuals are stored in one variable, with a group variable indicating group membership (here 1 or 2). We can then use the third form of `sdtest` (see [R] **sdtest**), with the `by(`*groupvar*`)` option, to conduct the F test.

What if there are more than two groups across which we wish to test for equality of disturbance variance: for instance, a set of 10 industries? We may then use the `robvar` command (see [R] **sdtest**), which like `sdtest` expects to find one variable containing each group's residuals, with a group membership variable identifying them. The `by(`*groupvar*`)` option is used here as well. The test conducted is that of Levene (1960), labeled as w_0, which is robust to nonnormality of the error distribution. Two variants of the test proposed by Brown and Forsythe (1992), which uses more robust estimators of central tendency (e.g., median rather than mean), w_{50} and w_{10}, are also computed.

I illustrate groupwise heteroskedasticity with state-level data from the `NEdata.dta`. These data comprise one observation per year for each of the six U.S. states in the New England region for 1981–2000. Descriptive statistics are generated by `summarize` for `dpipc`, state disposable personal income per capita.

```
. use http://www.stata-press.com/data/imeus/NEdata, clear
. summarize dpipc
```

Variable	Obs	Mean	Std. Dev.	Min	Max
dpipc	120	18.15802	5.662848	8.153382	33.38758

We fit a linear trend model to `dpipc` by regressing that variable on `year`. The residuals are tested for equality of variances across states with `robvar`.

```
. regress dpipc year
```

Source	SS	df	MS		
Model	3009.33617	1	3009.33617	Number of obs	= 120
Residual	806.737449	118	6.83675804	F(1, 118)	= 440.17
				Prob > F	= 0.0000
Total	3816.07362	119	32.0678456	R-squared	= 0.7886
				Adj R-squared	= 0.7868
				Root MSE	= 2.6147

dpipc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	.8684582	.0413941	20.98	0.000	.7864865	.9504298
_cons	-1710.508	82.39534	-20.76	0.000	-1873.673	-1547.343

```
. predict double eps, residual
. robvar eps, by(state)
```

state	Summary of Residuals Mean	Std. Dev.	Freq.
CT	4.167853	1.3596266	20
MA	1.618796	.86550138	20
ME	-2.9841056	.93797625	20
NH	.51033312	.61139299	20
RI	-.8927223	.63408722	20
VT	-2.4201543	.71470977	20
Total	-6.063e-14	2.6037101	120

```
W0  = 4.3882072   df(5, 114)     Pr > F = .00108562
W50 = 3.2989849   df(5, 114)     Pr > F = .00806752
W10 = 4.2536245   df(5, 114)     Pr > F = .00139064
```

The hypothesis of equality of variances is soundly rejected by all three `robvar` test statistics, with the residuals for Connecticut possessing a standard deviation considerably larger than those of the other three states.

FGLS estimation

If different groups of observations have different error variances, we can apply the GLS estimator using analytical weights, as described above in section 6.2.1. In the groupwise context, we define the analytical weight (`aw`) series as a constant value for each observation in a group. That value is calculated as the estimated variance of that group's OLS residuals. Using the residual series calculated above, we construct an estimate of its variance for each New England state with `egen` and generate the analytical weight series:

```
. by state, sort: egen sd_eps = sd(eps)
. generate double gw_wt = 1/sd_eps^2
. tabstat sd_eps gw_wt, by(state)
Summary statistics: mean
  by categories of: state
```

state	sd_eps	gw_wt
CT	1.359627	.5409545
MA	.8655014	1.334948
ME	.9379762	1.136623
NH	.611393	2.675218
RI	.6340872	2.48715
VT	.7147098	1.957675
Total	.8538824	1.688761

The `tabstat` command reveals that the standard deviations of New Hampshire and Rhode Island's residuals are much smaller than those of the other four states. We now reestimate the regression with FGLS, using the analytical weight series:

```
. regress dpipc year [aw=gw_wt]
(sum of wgt is   2.0265e+02)
```

Source	SS	df	MS		
Model	2845.55409	1	2845.55409	Number of obs	= 120
Residual	480.921278	118	4.07560405	F(1, 118)	= 698.19
				Prob > F	= 0.0000
Total	3326.47537	119	27.9535745	R-squared	= 0.8554
				Adj R-squared	= 0.8542
				Root MSE	= 2.0188

dpipc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	.8444948	.0319602	26.42	0.000	.7812049	.9077847
_cons	-1663.26	63.61705	-26.14	0.000	-1789.239	-1537.281

Compared with the unweighted estimates' Root MSE of 2.6147, FGLS yields a considerably smaller value of 2.0188.

6.2.3 Heteroskedasticity in grouped data

In section 6.2, I addressed a third case in which heteroskedasticity arises in cross-sectional data, where our observations are grouped or aggregated data, representing different numbers of microdata records. This situation arises when the variables in our dataset are averages or standard deviations of groups' observations, for instance, a set of 50 U.S. state observations. Because we know the population of each state, we know precisely how much more accurate California's observation (based on more than 30 million individuals) is than Vermont's (based on fewer than a million). This situation would also arise in the context of observations representing average attainment scores for individual schools or school districts, where we know that each school's (or school district's) student population is different. In these cases we know that heteroskedastic-

ity will occur in the grouped or aggregated data, and we know Ω because it depends only on the N_g underlying each observation.

You could consider this a problem of nonrandom sampling. In the first example above, when 30 million California records are replaced by one state record, an individual has little weight in the average. In a smaller state, each individual would have a greater weight in her state's average values. If we want to conduct inference for a national random sample, we must equalize those weights, leading to a heavier weight being placed on California's observation and a lighter weight being placed on Vermont's. The weights are determined by the relative magnitudes of the states' populations. Each observation in our data stands for an integer number of records in the population (stored, for instance, in pop).

FGLS estimation

We can deal with the innate heteroskedasticity in an OLS regression on grouped data by considering that the precision of each group mean (i.e., its standard error) depends on the size of the group from which it is calculated. The analytical weight, proportional to the inverse of the observation's variance, must take the group size into account. If we have state-level data on per capita saving and per capita income, we could estimate

```
. regress saving income [aw=pop]
```

in which we specify that the analytical weight is pop. The larger states will have higher weights, reflecting the greater precision of their group means.

I illustrate this correction with a dataset containing 420 public school districts' characteristics. The districts' average reading score (read_scr) is modeled as a function of their expenditures per student (expn_stu), computers per student (comp_stu), and the percentage of students eligible for free school lunches (meal_pct, an indicator of poverty in the district). We also know the enrollment per school district (enrl_tot). The descriptive statistics for these variables are given by summarize:

```
. use http://www.stata-press.com/data/imeus/pubschl, clear
. summarize read_scr expn_stu comp_stu meal_pct enrl_tot
```

Variable	Obs	Mean	Std. Dev.	Min	Max
read_scr	420	654.9705	20.10798	604.5	704
expn_stu	420	5312.408	633.9371	3926.07	7711.507
comp_stu	420	.1359266	.0649558	0	.4208333
meal_pct	420	44.70524	27.12338	0	100
enrl_tot	420	2628.793	3913.105	81	27176

First, we estimate the parameters by using regress, ignoring the total enrollment per school district, which varies considerably over the districts. We expect that districts' average reading scores will be positively related to expenditures per student and computers per student and negatively related to poverty.

```
. regress read_scr expn_stu comp_stu meal_pct
```

Source	SS	df	MS		Number of obs =	420
					F(3, 416) =	565.36
Model	136046.267	3	45348.7558		Prob > F =	0.0000
Residual	33368.3632	416	80.2124115		R-squared =	0.8030
					Adj R-squared =	0.8016
Total	169414.631	419	404.330861		Root MSE =	8.9561

read_scr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
expn_stu	.0046699	.0007204	6.48	0.000	.0032538	.006086
comp_stu	19.88584	7.168347	2.77	0.006	5.795143	33.97654
meal_pct	-.635131	.0164777	-38.54	0.000	-.667521	-.602741
_cons	655.8528	3.812206	172.04	0.000	648.3592	663.3464

Our prior results on the relationship between reading scores and these factors are borne out. We reestimate the parameters, using enrollment as an analytical weight.

```
. regress read_scr expn_stu comp_stu meal_pct [aw=enrl_tot]
(sum of wgt is   1.1041e+06)
```

Source	SS	df	MS		Number of obs =	420
					F(3, 416) =	906.75
Model	123692.671	3	41230.8903		Prob > F =	0.0000
Residual	18915.9815	416	45.4711093		R-squared =	0.8674
					Adj R-squared =	0.8664
Total	142608.652	419	340.354779		Root MSE =	6.7432

read_scr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
expn_stu	.0055534	.0008322	6.67	0.000	.0039176	.0071892
comp_stu	27.26378	8.197228	3.33	0.001	11.15063	43.37693
meal_pct	-.6352229	.013149	-48.31	0.000	-.6610696	-.6093762
_cons	648.988	4.163875	155.86	0.000	640.8031	657.1728

Including the weights modifies the coefficient estimates and reduces the `Root MSE` of the estimated equation. Equally weighting very small and very large school districts places too much weight on the former and too little on the latter. For instance, the effect of increases in the number of computers per student is almost 50% larger in the weighted estimates, and the effect of expenditures per student is smaller in the OLS estimates. The weighting also yields more precise coefficient estimates.

6.3 Serial correlation in the error distribution

Our discussion of heteroskedasticity in the error process focused on the first i in i.i.d.: the notion that disturbances are *identically* distributed over the observations. As in the discussion of the cluster estimator, we also may doubt the second i, that the disturbances are *independently* distributed. With cross-sectional data, departures from independence may reflect neighborhood effects, as accounted for by the cluster–VCE estimator. Observations that are similar in some way share a correlation in their disturbances.

When we turn to time-series data, we see a similar rationale for departures from independence. Observations that are close in time may be correlated, with the strength of that correlation increasing with proximity. Although there is no natural measure of proximity in cross-sectional data, time-series data by its nature defines *temporal* proximity. The previous and subsequent observations are those closest to y_t chronologically. When correlations arise in a time series, we speak of the disturbance process exhibiting *serial correlation* or *autocorrelation*; it is literally correlated with itself.

We must be wary of specification issues, as apparent serial correlation in the errors may be nothing more than a reflection of one or more systematic factors mistakenly excluded from the regression model. As discussed in section 5.2, inadequate specification of dynamic terms may cause such a problem. But sometimes errors will be, by construction, serially correlated rather than independent across observations. Theoretical schemes such as partial-adjustment mechanisms and agents' adaptive expectations can give rise to errors that cannot be serially independent. Thus we also must consider this sort of deviation of $\mathbf{\Sigma}_u$ from $\sigma^2 I_N$, one that is generally more challenging to deal with than is pure heteroskedasticity.

6.3.1 Testing for serial correlation

How might we test for the presence of serially correlated errors? Just as for pure heteroskedasticity, we base tests of serial correlation on the regression residuals. In the simplest case, autocorrelated errors follow the AR(1) model: an *autoregressive process* of order one, also known as a first-order Markov process:

$$u_t = \rho u_{t-1} + v_t, \ |\rho| < 1 \tag{6.15}$$

where the v_t are uncorrelated random variables with mean-zero and constant variance. We impose the restriction that $|\rho| < 1$ to ensure that the disturbance process u is stationary with a finite variance. If $\rho = 1$, we have a *random walk*, which implies that the variance of u is infinite, and u is termed a *nonstationary* series, or an integrated process of order one [often written as $I(1)$]. We assume that the u process is *stationary*, with a finite variance, which will imply that the effects of a shock, v_t, will dissipate over time.[21]

The larger (in absolute value) ρ is, the greater will be the *persistence* of that shock to u_t and the more highly *autocorrelated* will be the sequence of disturbances u_t. In fact, in the AR(1) model, the *autocorrelation function* of u will be the geometric sequence $\rho, \rho^2, \rho^3, \ldots$, and the correlation of disturbances separated by τ periods will be ρ^τ. In Stata, the autocorrelation function for a time series may be computed with the `ac` or `corrgram` commands ([TS] **corrgram** refers to the correlogram of the series).

If we suspect that there is autocorrelation in the disturbance process of our regression model, we could use the estimated residuals to diagnose it. The empirical counterpart

21. If there is reason to doubt the stationarity of a time series, a *unit root test* should be performed: see, for example, [TS] **dfgls**.

to u_t in (6.15) will be the $\widehat{u}_t$ series produced by `predict`. We estimate the auxiliary regression of $\widehat{u}_t$ on $\widehat{u}_{t-1}$ without a constant term because the residuals have mean zero. The resulting slope estimate is a consistent estimator of the first-order autocorrelation coefficient ρ of the u process from (6.15). Under the null hypothesis $\rho = 0$, so a rejection of this null hypothesis by this LM test indicates that the disturbance process exhibits AR(1) behavior.

A generalization of this procedure that supports testing for higher-order autoregressive disturbances is the LM test of Breusch and Godfrey (Godfrey 1988). In this test, the regression is augmented with p lagged residual series. The null hypothesis is that the errors are serially independent up to order p. The test evaluates the *partial correlations* of the regressors $\mathbf{x}$ partialled off.[22] The residuals at time t are orthogonal to the columns of $\mathbf{x}$ at time t, but that need not be so for the lagged residuals. This is perhaps the most useful test for nonindependence of time-series disturbances, since it allows the researcher to examine more than first-order serial independence of the errors in one test. The test is available in Stata as `estat bgodfrey` (see [R] **regress postestimation time series**).

A variation on the Breusch–Godfrey test is the Q test of Box and Pierce (1970), as refined by Ljung and Box (1979), which examines the first p sample autocorrelations of the residual series:

$$Q = T(T+2)\sum_{j=1}^{p}\frac{r_j^2}{T-j}$$

where r_j^2 is the jth autocorrelation of the residual series. Unlike the Breusch–Godfrey test, the Q test does not condition the autocorrelations on a particular x. Q is based on the simple correlations of the residuals rather than their partial correlations. Therefore, it is less powerful than the Breusch–Godfrey test when the null hypothesis (of no serial correlation in u up to order p) is false. However, the Q test may be applied to any time series whether or not it contains residuals from an estimated regression model. Under the null hypothesis, $Q \sim \chi^2(p)$. The Q test is available in Stata as `wntestq`, named such to indicate that it may be used as a general test for so-called *white noise*, a property of random variables that do not contain autocorrelation.

The oldest test (but still widely used and reported, despite its shortcomings) is the Durbin and Watson (1950) d statistic:

$$d = \frac{\sum_{t=2}^{T}(\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_{t=1}^{T}\widehat{u}_t^2} \simeq 2(1-\rho)$$

The Durbin–Watson (D–W) test proceeds from the principle that the numerator of the statistic, when expanded, contains twice the variance of the residuals minus twice the (first) autocovariance of the residual series. If $\rho = 0$, that autocovariance will be near zero, and d will equal 2.0. As $\rho \to 1$, $d \to 0$, whereas as $\rho \to -1$, $d \to 4$. However,

22. The partial autocorrelation function of a time series may be calculated with the `pac` command; see [TS] **corrgram**.

the exact distribution of the statistic depends on the regressor matrix (which must contain a constant term and must not contain a lagged dependent variable). Rather than having a set of critical values, the D–W test has two, labeled d_L and d_U. If the d statistic falls below d_L, we reject the null; above d_U, we do not reject; and in between, the statistic is inconclusive. (For negative autocorrelation, you test $4 - d$ against the same tabulated critical values.) The test is available in Stata as `estat dwstat` (see [R] **regress postestimation time series**) and is automatically displayed in the output of the `prais` estimation command.

In the presence of a lagged dependent variable or generally, predetermined regressors, the d statistic is biased toward 2.0, and Durbin's *alternative* (or h) test (Durbin 1970) must be used.[23] That test is an LM test, which is computed by regressing residuals on their lagged values and the original $\mathbf{X}$ matrix. The test is asymptotically equivalent to the Breusch–Godfrey test for $p = 1$ and is available in Stata as command `estat durbinalt` (see [R] **regress postestimation time series**).

I illustrate the diagnosis of autocorrelation with a time-series dataset of monthly short-term and long-term interest rates on U.K. government securities (Treasury bills and gilts), 1952m3–1995m12. `summarize` gives the descriptive statistics for these series:

```
. use http://www.stata-press.com/data/imeus/ukrates, clear
. summarize rs r20
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rs	526	7.651513	3.553109	1.561667	16.18
r20	526	8.863726	3.224372	3.35	17.18

The model expresses the monthly change in the short rate `rs`, the Bank of England's monetary policy instrument, as a function of the prior month's change in the long-term rate `r20`. The regressor and regressand are created on the fly by Stata's time-series operators `D.` and `L.`. The model represents a monetary policy reaction function. We save the model's residuals with `predict` so that we can use `wntestq`.

```
. regress D.rs LD.r20
```

Source	SS	df	MS		
Model	13.8769739	1	13.8769739	Number of obs	= 524
Residual	136.988471	522	.262430021	F(1, 522)	= 52.88
				Prob > F	= 0.0000
Total	150.865445	523	.288461654	R-squared	= 0.0920
				Adj R-squared	= 0.0902
				Root MSE	= .51228

D.rs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
r20						
LD.	.4882883	.0671484	7.27	0.000	.356374	.6202027
_cons	.0040183	.022384	0.18	0.858	-.0399555	.0479921

23. A variable x is predetermined if $E[x_t u_{t+s}] = 0$ for all t and s. See Davidson and MacKinnon (1993).

```
. predict double eps, residual
(2 missing values generated)
. estat bgodfrey, lags(6)
Breusch-Godfrey LM test for autocorrelation
```

lags(p)	chi2	df	Prob > chi2
6	17.237	6	0.0084

H0: no serial correlation

```
. wntestq eps
Portmanteau test for white noise
---------------------------------------------------------------------------
 Portmanteau (Q) statistic =    82.3882
 Prob > chi2(40)           =     0.0001
. ac eps
```

The Breusch–Godfrey test performed here considers the null of serial independence up to sixth order in the disturbance process, and that null is soundly rejected. That test is conditioned on the fitted model. The Q test invoked by `wntestq`, which allows for more general alternatives to serial independence of the residual series, confirms the diagnosis. To further analyze the nature of the residual series' lack of independence, we compute the autocorrelogram (displayed in figure 6.1). This graph indicates the strong presence of first-order autocorrelation—AR(1)—but also signals several other empirical autocorrelations outside the Bartlett confidence bands.

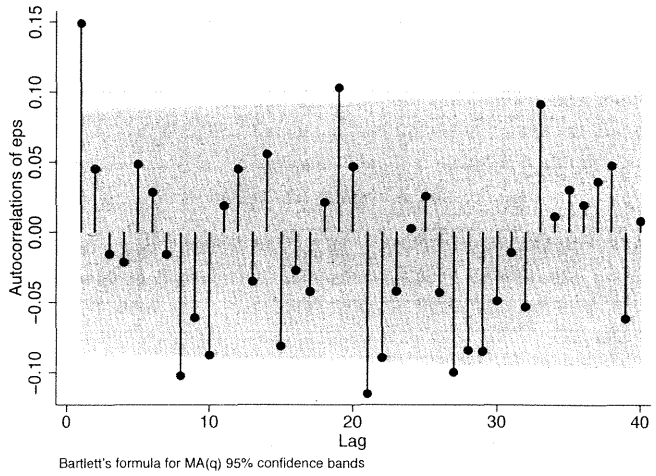

Figure 6.1: Autocorrelogram of regression residuals

6.3.2 FGLS estimation with serial correlation

For AR(1) disturbances of (6.15), if ρ were known, we could estimate the coefficients by GLS. The form of $\mathbf{\Sigma}_u$ displayed in (6.3) is simplified when we consider first-order serial correlation with one parameter ρ. An analytical inverse of $\mathbf{\Sigma}_u$ may be derived as

$$\mathbf{\Sigma}_u^{-1} = \sigma_u^{-2} \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 \\ -\rho & 1 & \dots & 0 \\ & & \vdots & \\ 0 & -\rho & 1 & 0 \\ 0 & \dots & -\rho & 1 \end{pmatrix} \tag{6.16}$$

As with heteroskedasticity, we do not explicitly construct and apply this matrix. Rather, we can implement GLS by transforming the original data and running a regression on the transformed data. For observations 2, ..., T, we *quasidifference* the data: $y_t - \rho y_{t-1}$, $x_{j,t} - \rho x_{j,t-1}$, and so on. The first observation is multiplied by $\sqrt{1-\rho^2}$.

The GLS estimator is not feasible because ρ is an unknown population parameter just like $\boldsymbol{\beta}$ and σ_u^2. Replacing the unknown ρ values above with a consistent estimate and computing $\widehat{\mathbf{\Sigma}}_u$ yields the FGLS estimator. As with heteroskedasticity, the OLS residuals from the original model may be used to generate the necessary estimate. The Prais and Winsten (1954) estimator uses an estimate of ρ based on the OLS residuals to estimate $\widehat{\mathbf{\Sigma}}_u^{-1}$ by (6.16). The closely related Cochrane and Orcutt (1949) variation on that estimator differs only in its treatment of the first observation of the transformed data, given the estimate of ρ from the regression residuals. Either of these estimators may be iterated to convergence: essentially they operate by ping-ponging back and forth between estimates of $\boldsymbol{\beta}$ and ρ. Optional iteration refines the estimate of ρ, which is strongly recommended in small samples. Both estimators are available in Stata with the `prais` command.

Other approaches include maximum likelihood, which simultaneously estimates one parameter vector $(\boldsymbol{\beta}', \sigma^2, \rho)$, and the grid search approach of Hildreth and Lu (1960). Although you could argue for the superiority of a maximum likelihood approach, Monte Carlo studies suggest that the Prais–Winsten estimator is nearly as efficient in practice as maximum likelihood.

I illustrate the Prais–Winsten estimator by using the monetary policy reaction function displayed above. FGLS on this model finds a value of ρ of 0.19 and a considerably smaller coefficient on the lagged change in the long-term interest rate than that of our OLS estimate.

```
. prais D.rs LD.r20, nolog

Prais-Winsten AR(1) regression -- iterated estimates
```

Source	SS	df	MS		Number of obs =	524
					F(1, 522) =	25.73
Model	6.56420242	1	6.56420242		Prob > F =	0.0000
Residual	133.146932	522	.25507075		R-squared =	0.0470
					Adj R-squared =	0.0452
Total	139.711134	523	.2671341		Root MSE =	.50505

D.rs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
r20						
LD.	.3495857	.068912	5.07	0.000	.2142067	.4849647
_cons	.0049985	.0272145	0.18	0.854	-.0484649	.0584619
rho	.1895324					

```
Durbin-Watson statistic (original)    1.702273
Durbin-Watson statistic (transformed) 2.007414
```

In summary, although we may use FGLS to deal with autocorrelation, we should always be aware that this diagnosis may reflect misspecification of the model's dynamics or omission of one or more key factors from the model. We may mechanically correct for first-order serial correlation in a model, but we then attribute this persistence to some sort of clockwork in the error process rather than explaining its existence. Applying FGLS as described here is suitable for AR(1) errors but not for higher-order AR(p) errors or *moving-average* (MA) error processes, both of which may be encountered in practice. Regression equations with higher-order AR errors or MA errors can be modeled by using Stata's `arima` command.

Exercises

1. Use the `cigconsump` dataset, retaining only years 1985 and 1995. Regress `lpackpc` on `lavgprs` and `lincpc`. Use the Breusch–Pagan test (`hettest`) for variable `year`. Save the residuals, and use `robvar` to compute their variances by `year`. What do these tests tell you?
2. Use FGLS to refit the model, using analytical weights based on the residuals from each year. How do these estimates differ from the OLS estimates?
3. Use the `sp500` dataset, applying `tsset date`. Regress the first difference of `close` on two lagged differences and lagged `volume`. How do you interpret the coefficient estimates? Use the Breusch–Godfrey test to evaluate the errors' independence. What do you conclude?
4. Refit the model with FGLS (using `prais`). How do the FGLS estimates compare to those from OLS?