Generalized Linear Regression / Robust Estimation / Heteroskedasticity

Greene, Ch. 9; Kennedy Ch.8
R scripts `mod4s2a mods4b, mods4c, mods4d, mods4e`

Introduction

The Generalized Linear Regression Model (GLRM) differs from the CLRM in the structure of the variance-covariance matrix of the error vector $\boldsymbol{\varepsilon}$. Specifically, we no longer have spherical disturbances (independent errors that all share the same variance), but error that may be potentially correlated and / or follow distributions with different variances. The GLRM in its generic form can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad E(\boldsymbol{\varepsilon}) = \mathbf{0} \qquad E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \underset{nxn}{\boldsymbol{\Omega}} \neq \sigma^2 \underset{nxn}{\mathbf{I}} \tag{1}$$

Thus, the error vector still has a mean of zero, but its variance-covariance matrix now takes the a general form with theoretically up to *n(n+1)/2* unknown parameters. Naturally, with only *n* observations estimation of that many additional parameters is infeasible. We usually follow one of two strategies: (i) Claim complete ignorance about the structure of $\boldsymbol{\Omega}$ and use robust estimation methods to still derive consistent estimates of coefficients, or (ii) Assume that $\boldsymbol{\Omega}$ has a simple structure with only a few additional parameters (example: clusters of errors have separate variances, but not each individual error term). We'll return to these estimation strategies below.

Properties of the LS Estimator in Presence of Non-spherical errors

The above model violates CLRM assumption # 4 ("homoskedasticity, non-autocorrelation"). Let's first examine how this affects the finite sample properties of the LS estimator:

$$\begin{aligned} E(\mathbf{b}) &= \boldsymbol{\beta} + E\left(\left(\mathbf{X'X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} \\ V(\mathbf{b}) &= E\left(\left(\mathbf{X'X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}\left(\mathbf{X'X}\right)^{-1}\right) = \left(\mathbf{X'X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X'X}\right)^{-1} \end{aligned} \tag{2}$$

The LS estimator remains unbiased since the generalization of the errors' variance does not affect the uncorrelatedness of **X** with $\boldsymbol{\varepsilon}$. However, the finite variance of **b** now takes a very different form (compared to $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$), which implies that the OLS variance is biased in this case. This, in turn, produces biased standard errors of coefficients and renders any finite-sample hypothesis tests (t, F) unreliable.

The asymptotic properties of the LS estimator follow this pattern:

$$\mathbf{b} \overset{a}{\sim} n\left(\boldsymbol{\beta}, \tfrac{1}{n}\mathbf{Q}^{-1}\text{plim}\left(\tfrac{1}{n}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\right)\mathbf{Q}^{-1}\right) \qquad \text{where } \mathbf{Q} = \text{plim}\left(\tfrac{1}{n}\mathbf{X}'\mathbf{X}\right) \tag{3}$$

In other words, **b** is still consistent, but its asymptotic variance is biased. This results in misleading asymptotic standard errors and hypothesis tests.

The Generalized Least Squares (GLS) Estimator

Let's assume for a moment that $\mathbf{\Omega}$ is fully known. In that case we can use a simple extension of the CLRM estimation framework. Conceptually, the best way to show this is as follows:

For ease of comparison with the CLRM case first factor $\mathbf{\Omega}$ into $\sigma^2\tilde{\mathbf{\Omega}}$ (this is trivial – you can always factor out any scalar from any matrix or vector). Then factor the inverse of $\tilde{\mathbf{\Omega}}$ into $\mathbf{P'P}$ where **P** is itself an *n* by *n* symmetric square matrix, using results from spectral decomposition (see Matrix Algebra Tutorial, p. 14). Then pre-multiply the CLRM by **P** – this will reinstate the spherical properties of the disturbances - and use OLS on the transformed model. Formally:

$$\begin{aligned}
&\mathbf{y}^* = \mathbf{X}^{*\prime}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \quad \text{where} \quad \mathbf{y}^* = \mathbf{Py}, \quad \mathbf{X}^* = \mathbf{PX}, \quad \boldsymbol{\varepsilon}^* = \mathbf{P}\boldsymbol{\varepsilon}, \\
&E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \mathbf{\Omega} = \sigma^2\tilde{\mathbf{\Omega}} \qquad \mathbf{P'P} = \tilde{\mathbf{\Omega}}^{\mathbf{-1}} \\
&E(\boldsymbol{\varepsilon}^*) = \mathbf{P}E(\boldsymbol{\varepsilon}) = 0, \quad E\left(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}\right) = \mathbf{P}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{P} = \sigma^2\mathbf{P}\tilde{\mathbf{\Omega}}\mathbf{P} = \sigma^2\mathbf{PP^{-1}P^{-1}P} = \sigma^2\mathbf{I}
\end{aligned} \tag{4}$$

Then the GLS estimator is derived as

$$\mathbf{b_{GLS}} = \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^* = \left(\mathbf{X'PPX}\right)^{-1}\mathbf{X'PPy} = \left(\mathbf{X'\Omega^{-1}X}\right)^{-1}\mathbf{X'\Omega^{-1}y} \tag{5}$$

This estimator has the following general properties:

$$\begin{aligned}
&E(\mathbf{b_{GLS}}) = E\left(\left(\mathbf{X'\Omega^{-1}X}\right)^{-1}\mathbf{X'\Omega^{-1}X}\boldsymbol{\beta}\right) + E\left(\left(\mathbf{X'\Omega^{-1}X}\right)^{-1}\mathbf{X'\Omega^{-1}}\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} \\
&V(\mathbf{b_{GLS}}) = E\left(\left(\mathbf{X'\Omega^{-1}X}\right)^{-1}\mathbf{X'\Omega^{-1}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{\Omega^{-1}X}\left(\mathbf{X'\Omega^{-1}X}\right)^{-1}\right) = \left(\mathbf{X'\Omega}^{-1}\mathbf{X}\right)^{-1} \\
&\mathbf{b_{GLS}} \overset{a}{\sim} n\left(\boldsymbol{\beta}, \left(\mathbf{X'\Omega}^{-1}\mathbf{X}\right)^{-1}\right)
\end{aligned} \tag{6}$$

Thus, $\mathbf{b_{GLS}}$ is unbiased and consistent. It is also asymptotically efficient. It is also the minimum variance linear unbiased estimator, i.e. it is "BLUE" for the generalized regression model. Specifically, the "correct" variance of the OLS estimator in (2) will be less efficient than V(b_{GLS}) (see script `mod4s2a` for evidence)

<u>Again:</u> Keep in mind that for the above exposition we assumed that all elements in $\mathbf{\Omega}$ are known, i.e. they are not parameters to be estimated. Only the slope parameters (i.e. the elements of $\boldsymbol{\beta}$) had to be estimated.

The Feasible Generalized Least Squares (FGLS) Estimator

The full content of $\mathbf{\Omega}$ will be known only in very few applications. More often we have a general idea about the structure of $\mathbf{\Omega}$ and assume that it is a function of just a few additional parameters, i.e. $\mathbf{\Omega}=\mathbf{\Omega}(\boldsymbol{\theta})$. For example, in many time series applications we specify

$$\mathbf{\Omega}=\frac{\sigma^2}{1-\varphi^2}\begin{bmatrix} 1 & \varphi & \varphi^2 & \cdots & \varphi^{n-1} \\ \varphi & 1 & \varphi & \cdots & \varphi^{n-2} \\ \varphi^2 & \varphi & 1 & \cdots & \varphi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \varphi^{n-3} & \cdots & 1 \end{bmatrix}, \tag{7}$$

which adds only a single additional parameter (φ) to the model. Another example is group-wise heteroskedasticity, where a limited number of blocks (or "clusters") of errors share different variances. Alternatively, the elements of $\mathbf{\Omega}$ may be assumed to be a combination of observed data and a few unknown parameters. For example, under heteroskedasticity we often assume that the variance of observation *i*'s error term is itself a function of attributes associated with that observation, i.e.

$$\sigma_i^2 = f\left(\mathbf{z_i},\boldsymbol{\gamma}\right) \quad \text{and}$$
$$\mathbf{\Omega}=\begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_n^2 \end{bmatrix} \tag{8}$$

Whatever the assumed structure of $\mathbf{\Omega}$, FGLS proceeds in the following 2 steps:

1. Derive a consistent estimator for $\mathbf{\Omega}$, i.e. $\hat{\mathbf{\Omega}}=\mathbf{\Omega}(\hat{\boldsymbol{\theta}})$.
2. Use this estimator in the expression for $\mathbf{b_{GLS}}$ (see (5)).

Formally:

$$\mathbf{b_{FGLS}}=\left(\mathbf{X}'\hat{\mathbf{\Omega}}^{\mathbf{-1}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{\mathbf{-1}}\mathbf{y} \quad \text{where} \quad \hat{\mathbf{\Omega}}=\mathbf{\Omega}(\hat{\boldsymbol{\theta}}) \tag{9}$$

For most "standard settings" the resulting $\mathbf{b_{FGLS}}$ will have the same desirable asymptotic properties as $\mathbf{b_{GLS}}$.

Alternatively, we can use a *Maximum Likelihood* approach to simultaneous estimate $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Such a "full-information maximum likelihood" (FIML) approach is generally more efficient in a small sample context.

However, if our assumptions regarding the structure of $\mathbf{\Omega}$ are incorrect, the same problems as presented for the OLS estimator arise. Some researchers therefore prefer not to make any assumption on $\mathbf{\Omega}$, but instead use *robust estimation* to derive consistent results. This is our next topic.

Robust Estimation for the GLRM

The general intuition for robust estimation is to use an expression for $\mathbf{\Omega}$ that does not require knowledge or estimation of additional parameters, but still allows for the consistent estimation of V(**b**). There exist a variety of robust estimators for GLRM's. Their exact form depends on the nature of the "generalization" at hand. This immediately implies there is no "one fits all" robust estimator. For example, for the case of *heteroskedasticity* where $\mathbf{\Omega}$ takes the form of (8), White (1980) has shown that under very general conditions the term

$$\hat{\hat{V}}_a(\mathbf{b}) = (\mathbf{X'X})^{-1}\mathbf{S_0}(\mathbf{X'X})^{-1} = (\mathbf{X'X})^{-1}\left(\sum_{i=1}^{n} e_i^2\mathbf{x_i x_i'}\right)(\mathbf{X'X})^{-1} \tag{10}$$

is a consistent estimator for the true asymptotic V(**b**). Thus, using this estimator, which is based on the OLS residuals e_i's , corrects the shortcomings of the naive OLS approach. This works well in a large sample context, but the properties of this estimator under small sample sizes are still disputed.

Note that for programming purposes in ***R*** you can use

$$\sum_{i=1}^{n} e_i^2\mathbf{x_i x_i'} = \mathbf{X'EX} \qquad \text{where}$$

$$\mathbf{E} = \begin{bmatrix} e_1^2 & 0 & 0 & 0 \\ 0 & e_2^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & e_n^2 \end{bmatrix} \qquad \texttt{= diag(e*e)} \tag{11}$$

The analogous version of this *robust or "White" estimator* (sometimes also called "sandwich estimator" or Huber-White sandwich estimator) for MLE is

$$\hat{\hat{V}}_a\left(\hat{\boldsymbol{\beta}}\right) = \left(-\hat{\mathbf{H}}\right)^{-1}\hat{\mathbf{G}}'\hat{\mathbf{G}}\left(-\hat{\mathbf{H}}\right)^{-1}, \tag{12}$$

where $\hat{\mathbf{H}}$ is the Hessian matrix at the MLE solution, and **G** is the *n* by *k* matrix of individual gradients at the MLE solution (the same "**G**" we used earlier in this course).

The White estimator can be used in any context where it is suspected that the spherical distribution assumption of the CLRM is violated. It has been especially popular to control for heteroskedastic errors – our next big topic.

Dealing with Heteroskedasticity

Introduction

Heteroskedasticity (HSK) is a common occurrence in many real-world applications, where individual errors or clusters of error terms have different variances. In many cases this difference in variances is related to observed data, which, if known, helps address the problems that arise when OLS is used to estimate such data.

For example, letting regression residuals proxy for the unknown error terms, in a regression of credit card expenditures on some explanatory variables, the residuals show a wider spread as income increases, cet. par. (Greene's Example 9.1). In other words, *otherwise identical individuals* have similar expenditures when their income is low or moderate, but vastly different spending patterns when income is large. Intuitively, this makes sense as individuals with high income have the option to spend more, while those with lower income are somewhat constrained in their expenditures.

Other examples with notorious HSK issues are regressions of home prices on home features (larger, more expensive homes with otherwise identical features have a lot more "wiggle room" to include features not captured by the regression, leading to pronounced variability in price), and water or energy consumption by firms (where variability in consumption generally increases with firm characteristics that are not included in the demand function, usually related to firm "size").

However, any difference in error variances falls under HSK, even when there is no clear "pattern" or link to an observed variable. In the most extreme case, each error term has its own variance, i.e.

$$E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)=\boldsymbol{\Omega}=\begin{bmatrix}\sigma_1^2 & 0 & 0 & 0\\ 0 & \sigma_2^2 & 0 & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \sigma_n^2\end{bmatrix}=\sigma^2\begin{bmatrix}\omega_1 & 0 & 0 & 0\\ 0 & \omega_2 & 0 & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \omega_n\end{bmatrix}. \tag{13}$$

Writing σ_i^2 as $\sigma^2\omega_i$ is completely optional, but makes it easier to compare the resulting model to the basic CLRM. If we make this arbitrary scaling by a "common" σ^2, we usually also implement a second arbitrary scaling and assume that the resulting ω_i - weights sum to the sample size. We then have $\sum_{i=1}^{n}\omega_i = n$, such that $tr\left(\boldsymbol{\Omega}\right)=\sum_{i=1}^{n}\sigma_i^2=\sigma^2\sum_{i=1}^{n}\omega_i = n\sigma^2$, as would be the case for the CLRM.

Recall that under any violation of the spherical error assumption, such as HSK , the generic OLS estimator is unbiased and consistent, but its variance is biased and inconsistent (producing misleading standard errors, t-values, F-tests etc.).

Testing for HSK

Residual versus Predictor Plots

A simple first check to see if HSK might be present in your data is to plot the regression residuals against explanatory variables that could possibly be linked to HSK. For a well-behaved CLRM we would expect the plotted residuals to randomly vary around zero along the entire support of any predictor variable. Any systematic pattern in this plot (e.g. the often observed “funnel shape”) would be indicative of HSK (assuming we don’t have any other mis-specification issue such as O.V. problems etc).

See scripts `mod4s2b,c` for examples.

White’s General Test

This test is based on the following intuition: If $\sigma_i^2 = \sigma^2 \forall i$, the conventional estimated asymptotic variance for **b**, $s^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$, should be a consistent estimator for the correct asymptotic variance of **b** under HSK, i.e. $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{\Omega}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$. Thus, the test is based on H$_0$: $\sigma_i^2 = \sigma^2 \forall i$ vs. H$_A$: $\sigma_i^2 \neq \sigma^2$ for at least one “*i*”.

The advantage of this test is that it is very general, i.e. it simply tests for “any HSK” and not for a specific pattern of HSK (as will the next test). However, this is also a weak point as it leaves the alternative hypothesis very vague, i.e. the test is “*non-constructive*” by default. If we reject the null, there is no guidance as to which form of HSK is present and how we can efficiently correct for it. Also, a rejection of the null may simply be indicative of other mis-specification problems, such as O.V.s.

Traditionally, if the test rejects the null (and we don’t run any other more “prescriptive” tests), we would directly proceed to *robust estimation* (see below).

Practical Implementation of White’s test:

1. Run your regression of interest and capture the residuals.
2. Run an auxiliary model by regressing the *squared* residuals against all variables in the original **X**, plus all of these variables squared (except for the constant and any 0/1 indicators, of course), and all unique cross-terms. Capture the R^2 from that regression.
3. Compute nR^2, where *n* is the sample size. Under the null this statistic will follow a chi-squared distribution with *k*-1 degrees of freedom, where *k* is the total number of regressors in the auxiliary model.

See script `mod4s2c` for an example of the White-test

Which variables to include for the White test?

It can be tricky to construct the augmented data matrix for the White test, especially when the original **X** contains indicator ("dummy") variables, and / or squared terms, and / or interaction terms.

The basic rule is to include in the augmented data matrix (let's call it $\tilde{\mathbf{X}}$) the following:

1. The entire original **X**, including the constant term
2. The squared terms for all variables in the original X, *except:*

    a. The constant term
    b. Any indicator variables or interactions of indicator variables (So any variable that can only take values of 0 or 1, regardless how they were constructed in the original model)

3. Any permissible 2-way interaction that can be constructed from the original **X**.

The tricky part is the definition of "permissible" for the interaction terms. The notes on working with indicator variables under "additional topics" on our course web site will provide some guidance in this respect. Basically, we cannot include any interaction terms that would lead to perfect collinearities in $\tilde{\mathbf{X}}$. This immediately implies the following restrictions:

a) No interactions with the constant term, of course…
b) No interactions that are already included in the original **X**
c) No interactions for implicit indicators (see web notes), i.e. 0/1 variables in the original **X** that indicate exclusive categories (such as "month1", "month2", "month3", etc, or "freshman", "sophomore", "junior", etc.). Naturally one cannot be a "freshman" and "sophomore" at the same time, so such an interaction would result in a vector of zeros, which, in turn, would introduce perfect collinearity.
d) No interactions of any other original indicators that would lead to an all-zero column (i.e. for which there simply aren't any observations in the data). For example, if one of your original variables is "female" (0=male, 1=female), and another is ("military training" – 0=no, 1=yes), and you don't have any females with military training in your data, the resulting interaction would produce an all-zero variable – a can't-do.

All other interactions should be included – see `mod4s2c`.

Breusch-Pagan / Godfrey LM Test

This test has a more specific alternative hypothesis than White’s general test. It examines if individual variance terms are related to a specific set of observed variables (that may or may not be included in the original **X**). The stipulated HSK model and the hypotheses for the test are as follows:

$$\begin{aligned}&\sigma_i^2 = \sigma^2 \cdot f\left(\mathbf{z}_\mathbf{i}'\boldsymbol{\alpha}\right) \qquad \mathbf{z}_\mathbf{i} = \begin{bmatrix}1 & z_{2,i} & \cdots & z_{k,i}\end{bmatrix}' \\ &H_0 : \alpha_2 = \alpha_3 = \cdots \alpha_k = 0 \text{ (and, therefore } \sigma_i^2 = \sigma^2\ \forall i) \\ &H_a : \text{at least one of the } \alpha\text{'s is not zero}\end{aligned} \tag{14}$$

A convenient feature of this test is that the function $f(.)$ does not need to be explicitly defined, i.e. the test is invariant to its explicit form. Technically the test is a Lagrange Multiplier (LM) test. Greene (p. 166) gives a convenient form for the test statistic:

$$LM = \tfrac{1}{2}\left(\mathbf{g}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{g}\right) \sim \chi_J^2 \quad \text{where} \quad \mathbf{g} = \left(\frac{n}{\mathbf{e}'\mathbf{e}}\begin{bmatrix}e_1^2 \\ e_2^2 \\ \vdots \\ e_n^2\end{bmatrix}\right) - 1 \qquad \text{and} \qquad J = k_z - 1 \tag{15}$$

As before, **e** denotes the vector of residuals from the original OLS model. If the null is rejected a consistent model incorporating this HSK pattern can be estimated via FGLS or MLE (see examples below). See script `mod4s2c` for an example of the BP-test.

The main drawback of the BP test is that it's sensitive to the normality assumption for the error terms under small sample sizes. In other words, if normality is violated, test results can be misleading. Koenker (1981) and Koenker and Bassett (1982) propose a more robust version of the test, based on the following more robust estimator of the error variance:

$$\hat{V}(\varepsilon_i)=\tfrac{1}{n}\sum_{i=1}^{n}\left(e_i^2-\frac{\mathbf{e}'\mathbf{e}}{n}\right)^2 \tag{16}$$

The test statistic then takes the form of

$$LM=\left(\tfrac{1}{V}\right)\left(\mathbf{e^2}-i\left(\tfrac{\mathbf{e}'\mathbf{e}}{n}\right)\right)'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\left(\mathbf{e^2}-i\left(\tfrac{\mathbf{e}'\mathbf{e}}{n}\right)\right) \tag{17}$$

where $\mathbf{e^2}=\begin{bmatrix}e_1^2 & e_2^2 & \dots & e_n^2\end{bmatrix}$. As stated in Greene, p. 166, this modified test statistic will follow the same asymptotic distribution as the PB statistic under normality. It provides a more powerful test in absence of normality. It is algebraically identical to the White test if the elements in $\mathbf{z_i}$ are the same as those used in the White test (i.e. original **X**, squared terms, and cross-terms).

Robust Estimation

If these tests suggest clear evidence of HSK the analyst has two basic options: (i) remain ignorant about the actual form or cause of HSK (i.e. the structure of $\mathbf{\Omega}$) and switch to *robust* OLS or MLE, or (ii) assume a specific form of HSK (with guidance from the BP test perhaps) and estimate a HSK-adjusted model via FGLS or MLE.

For option (ii), if the assumption on the underlying form of HSK is correct, the HSK-adjusted model will be more efficient than a robust model. However, if the assumption is incorrect (and HSK takes a different form) $\mathbf{\Omega}$ will be mis-specified, leading again to an inconsistent estimate of $V\left(\mathbf{b_{FGLS}}\right)$ or $V\left(\hat{\boldsymbol{\beta}}\right)$.

For this reason analysts often prefer to work with a less efficient but consistent *robust* model. For the GLRM robust variance estimators for the general HSK case are given in equations (10)-(12) in the previous Lecture notes. Script `mod4s2c` provides an example for the OLS case.

HSK Case Studies

Multiplicative HSK

This follows closely Greene pp. 170-175. The GLRM with *multiplicative HSK (mHSK)* was first proposed by Harvey (1976). We will distinguish between the main regression of interest and the *skedastic function*, i.e. the expression that links individual variance terms to observed data. For the mHSK model, the skedastic function can be written as

$$\sigma_i^2=\exp\left(\mathbf{z_i'}\boldsymbol{\gamma}\right)\rightarrow\log\left(\sigma_i^2\right)=\mathbf{z_i'}\boldsymbol{\gamma} \tag{18}$$

As before we assume that $\mathbf{z_i}$ includes a constant term, so that $\sigma^2 = \exp(\gamma_1)$ can be (conveniently) interpreted as “baseline variance”.

Using FGLS, we perform the following steps:

1. Estimate the main regression via OLS and capture the residuals
2. Using again OLS, regress $\log(e_i^2)$ against $\mathbf{z_i}$. This yields an estimate of $\boldsymbol{\gamma}$ (call it $\hat{\boldsymbol{\gamma}}$).
3. Estimate $\hat{\sigma}_i^2 = \exp(\mathbf{z_i}'\hat{\boldsymbol{\gamma}}) + 1.2704$ (the last number is Harvey’s proposed correction term for the estimated constant)
4. Define $\hat{\boldsymbol{\Omega}} = \begin{bmatrix} \hat{\sigma}_1^2 & & & \\ & \hat{\sigma}_2^2 & & \\ & & \ddots & \\ & & & \hat{\sigma}_n^2 \end{bmatrix}$, and use it in the expression for the GLS estimator $\mathbf{b_{GLS}}$.

See script `mod4s2d` for an example.

Groupwise HSK

Assume you have an application where your data can be divided into two or more groups (similar to what we had for the “Chow” test, but with possibly more than 2 groups). For example, you may have monthly observations on water consumption for a set of hotels. Each subset of observations for a specific hotel would then qualify as a “group”. Another example would be sets of observations on treatment outcomes associated with different hospitals. All observations for a given hospital would constitute a group.

Generically, assume there are *g = 1...G* groups represented in your data, with n_1, n_2,.. .n_G denoting the corresponding number of observations. Group-wise HSK results if all error terms associated with a single group share the same variance, but these variances differ across groups. Formally:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = diag\begin{bmatrix} \sigma_1^2 & \underset{n_1 \text{ terms}}{\sigma_1^2} & \cdots & \sigma_1^2 & \sigma_2^2 & \underset{n_2 \text{ terms}}{\sigma_2^2} & \cdots & \sigma_2^2 & \cdots & \sigma_G^2 & \underset{n_G \text{ terms}}{\sigma_G^2} & \cdots & \sigma_G^2 \end{bmatrix} \quad (19)$$

As with any form of HSK you have the option to use robust OLS with White-corrected standard errors. Alternatively, there is a particularly convenient FGLS estimator available for this case:

A straightforward FGLS version is as follows:

1. Estimate the main regression via OLS and capture the residuals.
2. For each group, compute $\hat{\sigma}_g^2 = \dfrac{\mathbf{e_g}'\mathbf{e_g}'}{n_g}$ where $\mathbf{e_g}$ is the OLS residual vector associated with group *g*.
3. Define $\hat{\boldsymbol{\Omega}} = diag\begin{bmatrix} \hat{\sigma}_1^2 & \underset{n_1 \text{ terms}}{\hat{\sigma}_1^2} & \cdots & \hat{\sigma}_1^2 & \hat{\sigma}_2^2 & \underset{n_2 \text{ terms}}{\hat{\sigma}_2^2} & \cdots & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_G^2 & \underset{n_G \text{ terms}}{\hat{\sigma}_G^2} & \cdots & \hat{\sigma}_G^2 \end{bmatrix}$, and use it in the expression for the GLS estimator $\mathbf{b_{GLS}}$.

To test for groupwise HSK, you can use the PB test with $\mathbf{Z} = [\mathbf{i} \quad \mathbf{d_1} \quad \mathbf{d_2} \quad \cdots \quad \mathbf{d_{G-1}}]$, where $\mathbf{d_g}$ is an indicator variable taking the value of one if observation *i* belongs to group g, and zero otherwise. The use of "*G-1*" subscript for the last term simply serves as a reminder that one group needs to be omitted from **Z** to avoid perfect collinearity.

See script `mod4s2e` for an example.

Working with aggregate dependent variables – the case of known HSK

In Macroeconomics and regional economics we often work with aggregate data for both dependent and explanatory variables. For example, assume that you are interested in relating disposable income to a set of micro-and macroeconomic regressors. Assume you have data for County *j* = 1…J. At the individual level, the model is given by

$$y_{ij} = \mathbf{x}'_{\mathbf{ij}}\boldsymbol{\beta} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim n\left(0,\sigma^2\right) \tag{20}$$

However, we often don't have individual-level data. Instead, we often work with aggregate data. Consider a regression of average per-capita disposable income for each county in the U.S. on a set of explanatory variables (which themselves may or may not be aggregated):

$$\bar{y}_j = \bar{\mathbf{x}}'_{\mathbf{j}}\boldsymbol{\beta} + \bar{\varepsilon}_j \qquad\qquad \bar{\varepsilon}_j = \tfrac{1}{n_j}\sum_{i=1}^{n_j}\varepsilon_{ij} \tag{21}$$

This model is heteroskedastic by default, since we have

$$V\left(\bar{\varepsilon}_j\right) = \tfrac{1}{n_j^2} n_j \sigma^2 = \tfrac{1}{n_j}\sigma^2 \tag{22}$$

If we know each County's total population (which we usually do) we can derive a consistent and efficient estimator via *Weighted Least Squares* (WLS). This is similar to our FGLS from above, except the individual variance weights are not a function of explanatory variables and additional parameters, but simply a function of individual "weights", here the County population.

We know that

$$\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \frac{1}{n_1} & & & \\ & \frac{1}{n_2} & & \\ & & \ddots & \\ & & & \frac{1}{n_J} \end{bmatrix} \tag{23}$$

We than simply use this variance matrix in our GLS formula.