计量经济学

第六讲：一元线性回归（二）

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室　粤海校区汇文楼2613
E-mail　huangjp@szu.edu.cn
Website　https://huangjp.com

主要内容

- 假设检验和置信区间
- X 为二值变量时的回归
- 异方差和同方差
- Gauss-Markov 定理

假设检验和置信区间

回归系数的大样本抽样分布

- 在 OLS 假设成立，且满足大样本条件时，回归系数的估计量 $\widehat{\beta}_0$ 和 $\widehat{\beta}_1$ 服从联合正态分布。其边缘分布分别是

$$\widehat{\beta}_0 \sim N(\beta_0, \sigma^2_{\widehat{\beta}_0}),$$

$$\sigma^2_{\widehat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ 其中 } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right]X_i$$

$$\widehat{\beta}_1 \sim N(\beta_1, \sigma^2_{\widehat{\beta}_1}),$$

$$\sigma^2_{\widehat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}$$

β_1 的假设检验

- t 统计量的一般形式：

$$t = \frac{\text{估计量} - \text{假设值}}{\text{估计量的标准误}}$$

- β_1 的双边假设：

$$H_0 : \beta_1 = \beta_{1,0}; \quad H_1 : \beta_1 \neq \beta_{1,0}$$

- t 统计量为

$$t = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\widehat{\beta}_1)},$$

其中 $\mathrm{SE}(\widehat{\beta}_1) = \sqrt{\widehat{\sigma}^2_{\widehat{\beta}_1}}$，$\widehat{\sigma}^2_{\widehat{\beta}_1} = \dfrac{1}{n} \times \dfrac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \widehat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}$

$\widehat{\beta}_0$ 的标准误见附录 5.1

STAR 数据的回归结果

Model 1: OLS, using observations 1–420
Dependent variable: testscr

	coefficient	std. error	t-ratio	p-value	
const	698.933	9.46749	73.82	6.57e-242	***
str	-2.27981	0.479826	-4.751	2.78e-06	***

Mean dependent var	654.1565	S.D. dependent var	19.05335
Sum squared resid	144315.5	S.E. of regression	18.58097
R-squared	0.051240	Adjusted R-squared	0.048970
F(1, 418)	22.57511	P-value(F)	2.78e-06
Log-likelihood	-1822.250	Akaike criterion	3648.499
Schwarz criterion	3656.580	Hannan-Quinn	3651.693

回归结果的表述方式

- 我们的回归模型是

$$\text{testscr} = \beta_0 + \beta_1 \text{str} + u_i$$

- 通常可以将回归结果用等式形式表述

$$\widehat{\text{testscr}} = \underset{(9.47)}{698.9} - \underset{(0.48)}{2.28} \times \text{str}, \quad R^2 = 0.051,\ \text{SER} = 18.6$$

此时，回归系数估计值下方的括号内为该系数估计量的**标准误**。

注：在括号内填写 t 值也是可以接受的，因为通常原假设都为 $\beta_k = 0$。但填写标准误的优点更大，且被更广泛地采用。无论填写哪个量，都需要在文章中适当的地方准确告知读者你填写的是什么。

β_1 的置信区间

- β_1 的 95% 双侧置信区间为：

$$[\hat{\beta}_1 - 1.96\,\mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 1.96\,\mathrm{SE}(\hat{\beta}_1)]$$

- X 的变化引起的预期效应的置信区间

  当 X 的变化量为 Δx 时，对应的 Y 的预期变化为 $\beta_1 \Delta x$，其 95% 置信区间为

$$[\hat{\beta}_1 \Delta x - 1.96\,\mathrm{SE}(\hat{\beta}_1)\Delta x,\ \hat{\beta}_1 \Delta x + 1.96\,\mathrm{SE}(\hat{\beta}_1)\Delta x]$$

X 为二值变量时的回归

二值变量

Binary variable

- 二值变量（binary variable）即为只取两个值的变量，通常为分类变量，因此去值可以设为 0 或 1。二值变量也称作指示变量（indicator variable）或虚拟变量（dummy variable）。

- 当唯一的回归变量为二值变量时，回归分析等价于对均值之差的分析：

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \quad D_i \in \{0,1\}$$

则 $Y_i = \beta_0 + u_0 \ (D_i = 0)$，$Y_i = \beta_0 + \beta_1 + u_0 \ (D_i = 1)$，因此 $E(Y_i \mid D_i = 0) = \beta_0$，$E(Y_i \mid D_i = 1) = \beta_0 + \beta_1$，即

$$\beta_1 = E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0)$$

课后练习（不需提交）

- 使用 STAR 数据集，根据 str 变量定义二值变量 D_i

$$D_i = \begin{cases} 1 & \text{if str} < 20 \\ 0 & \text{if str} \geq 20 \end{cases}$$

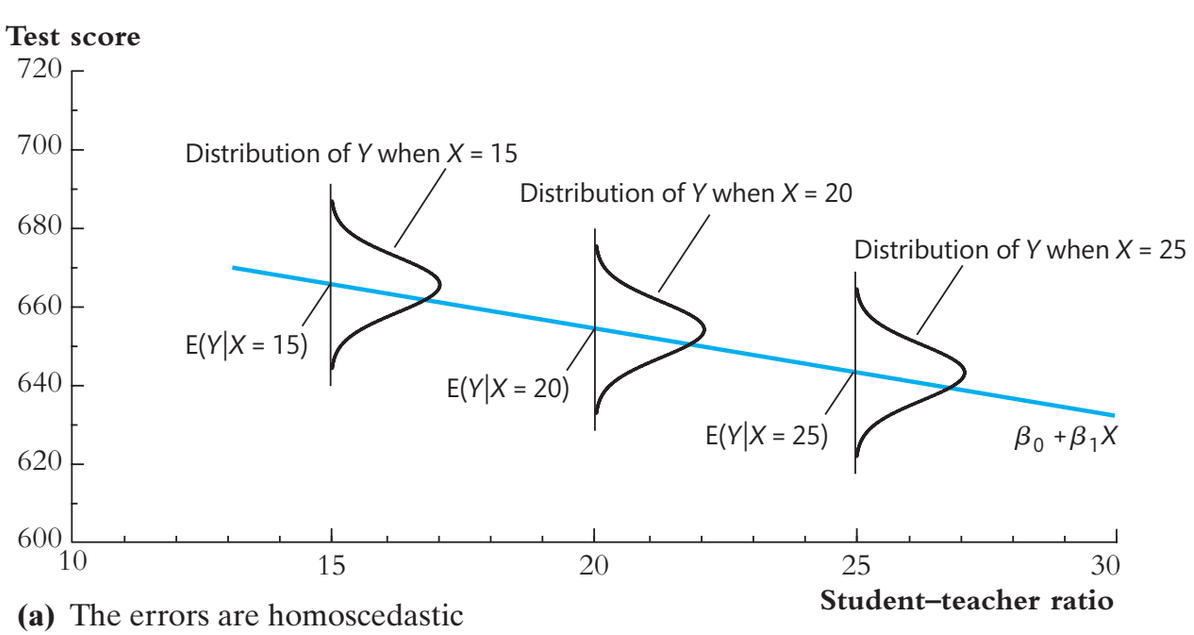
- 用 gretl 对模型 $Y_i = \beta_0 + \beta_1 D_i + u_i$ 进行回归分析，找到系数 β_1 的统计值和标准误。

- 令 str_L 为 $\text{str} < 20$ 的子集，str_H 为 $\text{str} \geq 20$ 的子集。针对两个子集中 testscr 的均值进行比较，并对均值差为零的原假设进行检验（参考3.4节）。比较检验结果和回归分析结果。

异方差和同方差

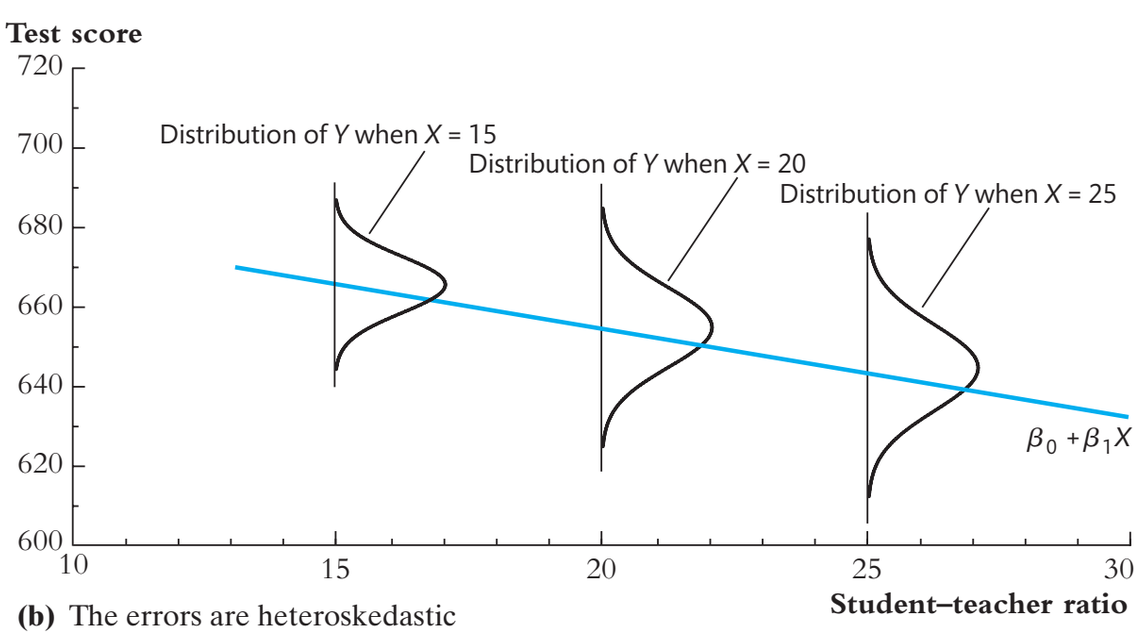
异方差的定义

- 如果对于任意的 $i = 1,2,\ldots,n$，$\text{var}(u_i \mid X_i = x)$ 为常数且不依赖于 x 时，误差项 u_i 是**同方差的（homoskedastic）**，否则，误差项是**异方差的（heteroskedastic）**。



(a) The errors are homoscedastic

同方差



(b) The errors are heteroskedastic

异方差

同方差和 OLS 假设

- OLS 估计量的无偏性、一致性、服从渐近正态分布的性质只要求 OLS 假设，与同方差或异方差无关。

- 如满足同方差和 OLS 假设，则 OLS 估计量是 BLUE 的。（5.5 节 Gauss-Markov 定理）

- 同方差和异方差时，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准误的表达式不同：
  - 异方差时：$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$，$\hat{\sigma}^2_{\hat{\beta}_1} = \dfrac{1}{n} \times \dfrac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}$

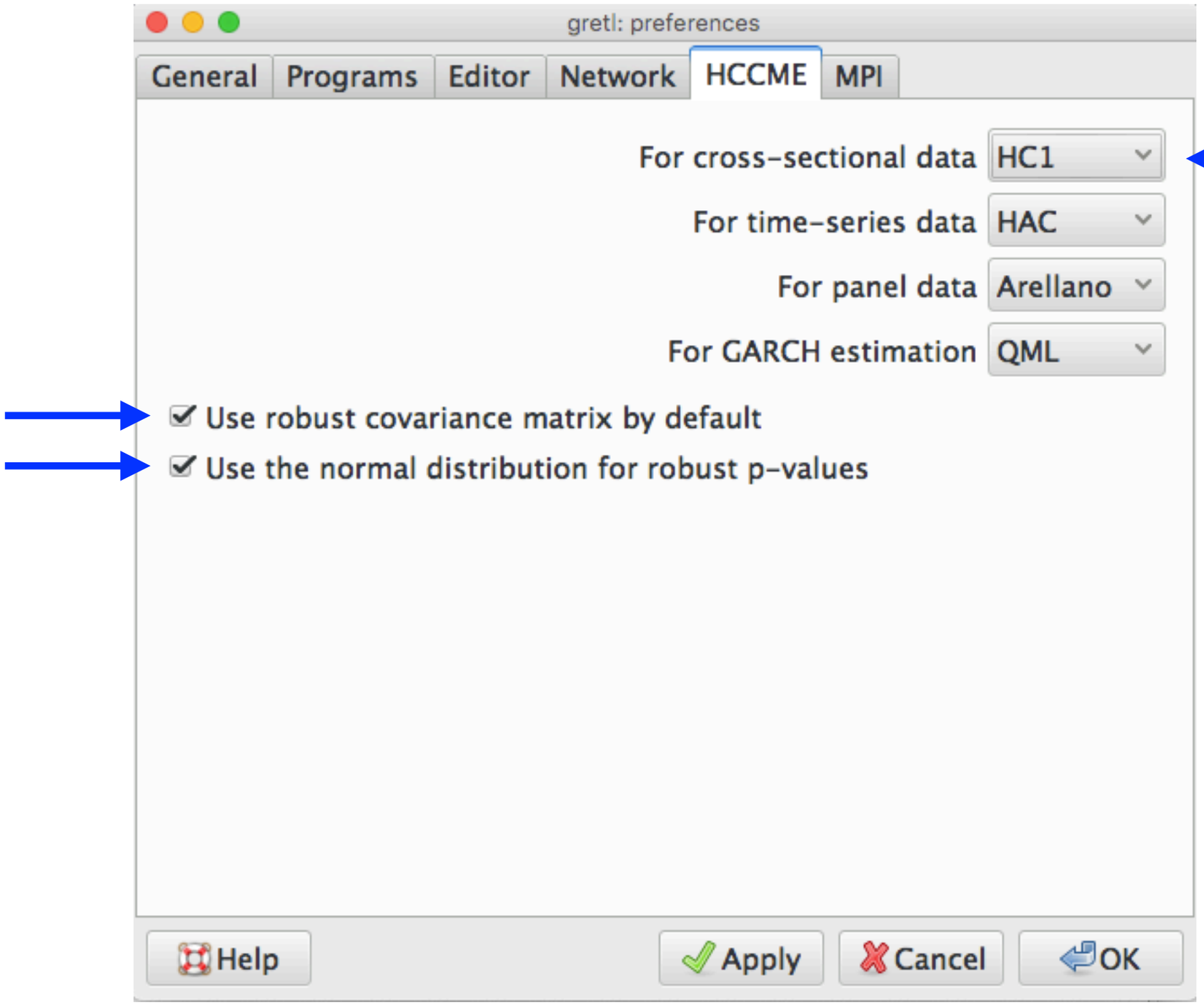
    异方差稳健标准误（heteroskedasticity-robust SE, or HC1）
    See, e.g., MacKinnon & White (1985).
  - 同方差时：$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\tilde{\sigma}^2_{\hat{\beta}_1}}$，$\tilde{\sigma}^2_{\hat{\beta}_1} = \dfrac{\mathrm{SER}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

我们应该假设同方差吗?

- 异方差出现在很多现实经济现象和数据中，如收入、性别。
- 假设同方差的优点
  - 实践上：令回归系数的 OLS 统计量标准误的计算变的略简单；
  - 理论上：可推导出 OLS 统计量具有 BLUE 性质（后述）。
- 假设同方差的缺点
  - 如果事实是异方差，则 OLS 统计量标准误将会计算错误，导致错误的检验结果。
- 假设同方差或异方差，都不影响 OLS 统计量的估计值、无偏性、一致性和服从渐进正态分布。
- 因此，我们应该承认异方差的普遍存在，并默认使用异方差稳健标准误。

在 gretl 中使用异方差稳健标准误

偏好设置 Tools > Preferences > General...



HC0: White (1980)
HC1: HC0 + 调整自由度
即本书中的公式

在 gretl 中使用异方差稳健标准误

编程模式

- 在整个程序开始时运行以下命令，可适用于整个程序

```
set force_hc on
set hc_version 1
    # 0 (the original White's) is the default
set robust_z on
```

- 对于单一 OLS 回归，则可以采用以下写法

```
ols yvar xvar --robust
```

（当然，你还是需要提前设置 hc_version）

STAR 数据的回归结果

同方差

```
Model 1: OLS, using observations 1-420
Dependent variable: testscr

             coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        698.933      9.46749     73.82     6.57e-242 ***
  str           -2.27981    0.479826    -4.751    2.78e-06  ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    144315.5   S.E. of regression   18.58097
R-squared            0.051240   Adjusted R-squared   0.048970
F(1, 418)            22.57511   P-value(F)           2.78e-06
Log-likelihood      -1822.250   Akaike criterion     3648.499
Schwarz criterion    3656.580   Hannan-Quinn         3651.693
```

STAR 数据的回归结果

异方差

```
Model 1: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error      z      p-value
  ------------------------------------------------------------
  const        698.933      10.3644      67.44    0.0000   ***
  str           -2.27981     0.519489    -4.389   1.14e-05 ***

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    144315.5   S.E. of regression   18.58097
R-squared            0.051240   Adjusted R-squared   0.048970
F(1, 418)            19.25943   P-value(F)           0.000014
Log-likelihood      -1822.250   Akaike criterion     3648.499
Schwarz criterion    3656.580   Hannan-Quinn         3651.693
```

Gauss-Markov 定理

线性条件无偏估计量

Linear conditional unbiased estimator

- 线性条件估计量：在 $X_1, X_2, \ldots, X_n$ 的条件下，β_1 的估计量

$$\widetilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$$

被称为线性条件估计量。其中权重 $a_1, a_2, \ldots, a_n$ 可以依赖于 $X_1, X_2, \ldots, X_n$ 但不能依赖于 $Y_1, Y_2, \ldots, Y_n$。

- 若 $E(\widetilde{\beta}_1 \mid X_1, X_2, \ldots, X_n) = \beta_1$，则估计量 $\widetilde{\beta}_1$ 是条件无偏的。
- OLS 估计量是线性的、条件无偏的（证明见附录5.2）。

Gauss-Markov 定理

若三个 OLS 假设成立，且误差同方差，则 OLS 估计量 $\hat{\beta}_1$ 是最佳（最有效的）线性条件无偏估计量（BLUE）。

- 证明见附录 5.2。由附录 5.2 也可知 Gauss-Markov 定理所需要的条件，即 Gauss-Markov 条件（弱于 OLS 假设 + 同方差）。

- Gauss-Markov 定理的局限性：

  - 其条件在实际应用中可能不成立，如同方差性。

  - 存在其他非线性的条件无偏估计量，且又可能比 OLS 估计量更有效。如加权最小二乘估计量（weighted least squares, WLS）、最小绝对变差估计量（least absolute deviations, LAD）。

关于异方差稳健标准误的参考文献

- White, H., (1980). A Heteroskedasticity-consistent Covariance Matrix and a Direct Test for Heteroskedasticity, *Econometrica*, 48, 817-838.

- MacKinnon, J., and White, H., (1985). Some Heteroskedasticity-consistent Covariance Matrix Estimators with Improved Finite Sample Properties, *Journal of Econometrics*, 29, 305-325.

- MacKinnon, J., (2013). Thirty Years of Heteroskedasticity-Robust Inference, in X. Chen and N. R. Swanson (eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, Springer, New York.

- Cottrell, A, and Lucchetti, R., (2021). *Gretl User's Guide*, Chapter 22. GNU.