CHAPTER 2

Homogeneity Tests for Linear Regression Models (Analysis of Covariance)

2.1 INTRODUCTION

Suppose we have sample observations of characteristics of N individuals over T time periods denoted by $y_{it}, x_{kit},\ i = 1, \ldots N, t = 1, \ldots, T, k = 1, \ldots, K$. Conventionally, observations of y are assumed to be the random outcomes of some experiment with a probability distribution conditional on vectors of the characteristics $\mathbf{x}$ and a fixed number of parameters $\boldsymbol{\theta}$, $f(y \mid \mathbf{x}, \boldsymbol{\theta})$. When panel data are used, one of the ultimate goals is to use all available information to make inferences on $\boldsymbol{\theta}$. For instance, a simple model commonly postulated is that y is a linear function of $\mathbf{x}$. Yet to run a least-squares regression with all NT observations, we need to assume that the regression parameters take value common to all cross-sectional units for all time periods. If this assumption is not valid, as shown in Chapter 1, Section 1.2, the pooled least-squares estimates may lead to false inferences. Thus, as a first step toward full exploitation of the data, we often test whether or not parameters characterizing the random outcome variable y stay constant across all i and t.

In the case of linear regression model, a widely used procedure to identify the source of sample variation preliminarily is the analysis of covariance (ANCOVA) test. The name "analysis of variance" (ANOVA) is often reserved for a particular category of linear hypotheses that stipulate that the expected value of a random variable y depends only on the class (defined by one or more factors) to which the individual considered belongs, but excludes tests relating to regressions. On the other hand, ANCOVA models are of a mixed character involving genuine exogenous variables, as do regression models, and at the same time allowing the true relation for each individual to depend on the class to which the individual belongs, as do the usual ANOVA models.

A linear model commonly used to assess the effects of both quantitative and qualitative factors is postulated as

$$y_{it} = \alpha_{it}^* + \boldsymbol{\beta}_{it}'\mathbf{x}_{it} + u_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{2.1.1}$$

where α_{it}^* and $\boldsymbol{\beta}_{it}' = (\beta_{1it}, \beta_{2it}, \ldots, \beta_{Kit})$ are 1×1 and $1 \times K$ vectors of constants that vary across i and t, respectively, $\mathbf{x}_{it}' = (x_{1it}, \ldots, x_{Kit})$ is a $1 \times K$

https://doi.org/10.1017/CBO9781139839327.005 Published online by Cambridge University Press

vector of exogenous variables, and u_{it} is the error term. Two aspects of the estimated regression coefficients can be tested: first, the homogeneity of regression slope coefficients; second, the homogeneity of regression intercept coefficients. The test procedure as suggested by Kuh (1963) has three main steps:

1. Test whether or not slopes and intercepts simultaneously are homogeneous among different individuals at different times.
2. Test whether or not the regression slopes collectively are the same.
3. Test whether or not the regression intercepts are the same.

It is obvious that if the hypothesis of overall homogeneity (step 1) is accepted, the testing procedure will go no further. However, should the overall homogeneity hypothesis be rejected, the second step of the analysis is to decide if the regression slopes are the same. If this hypothesis of homogeneity is not rejected, one then proceeds to the third and final test to determine the equality of regression intercepts. In principle, step 1 is separable from steps 2 and 3.[1]

Although this type of analysis can be performed on several dimensions, as described by Scheffé (1959) or Searle (1971), only one-way ANCOVA has been widely used. Therefore, here we present only the procedures for performing one-way ANCOVA.

2.2 ANALYSIS OF COVARIANCE

Model (2.1.1) only has descriptive value. It can neither be estimated nor used to generate prediction because the available degrees of freedom, NT, is less than the number of parameters, $NT(K+1)$ number of parameters, characterizing the distribution of y_{it}. A structure has to be imposed on (2.1.1) before any inference can be made. To start with, we assume that parameters are constant over time, but can vary across individuals. Thus, we can postulate a separate regression for each individual:

$$y_{it} = \alpha_i^* + \boldsymbol{\beta}_i'\mathbf{x}_{it} + u_{it}, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T. \end{aligned} \tag{2.2.1}$$

Three types of restrictions can be imposed on (2.2.1), namely:

H_1: Regression slope coefficients are identical, and intercepts are not. That is,

$$y_{it} = \alpha_i^* + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}. \tag{2.2.2}$$

H_2: Regression intercepts are the same, and slope coefficients are not. That is,

$$y_{it} = \alpha^* + \boldsymbol{\beta}_i'\mathbf{x}_{it} + u_{it}. \tag{2.2.3}$$

[1] Note that even if the homogeneity hypothesis is rejected, some useful information can be found in pooling the data, as long as the source of sample variability can be identified. For details, see later chapters.

https://doi.org/10.1017/CBO9781139839327.005 Published online by Cambridge University Press

H_3: Both slope and intercept coefficients are the same. That is,

$$y_{it} = \alpha^* + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}. \tag{2.2.4}$$

Because it is seldom a meaningful question to ask if the intercepts are the same when the slopes are unequal, we shall ignore the type of restrictions postulated by (2.2.3). We shall refer to (2.2.1) as the unrestricted model, (2.2.2) as the individual-mean or cell-mean corrected regression model, and (2.2.4) as the pooled regression.

Let

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \tag{2.2.5}$$

$$\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}, \tag{2.2.6}$$

be the means of y and $\mathbf{x}$, respectively, for the ith individual. The least-squares estimates of $\boldsymbol{\beta}_i$ and α_i^* in the unrestricted model (2.2.1) are given by[2]

$$\hat{\boldsymbol{\beta}}_i = W_{xx,i}^{-1} W_{xy,i}, \quad \hat{\alpha}_i = \bar{y}_i - \hat{\boldsymbol{\beta}}_i'\bar{\mathbf{x}}_i, \quad i = 1, \ldots, N, \tag{2.2.7}$$

where

$$\begin{aligned} W_{xx,i} &= \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)', \\ W_{xy,i} &= \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i), \\ W_{yy,i} &= \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2. \end{aligned} \tag{2.2.8}$$

In the ANCOVA terminology, equations (2.2.7) are called within-group estimates. The ith-group residual sum of squares is $\mathrm{RSS}_i = W_{yy,i} - W_{xy,i}' W_{xx,i}^{-1} W_{xy,i}$. The unrestricted residual sum of squares is

$$S_1 = \sum_{i=1}^{N} \mathrm{RSS}_i. \tag{2.2.9}$$

The least-squares regression of the individual mean corrected model yields parameter estimates

$$\begin{aligned} \hat{\boldsymbol{\beta}}_w &= W_{xx}^{-1} W_{xy}, \\ \hat{\alpha}_i^* &= \bar{y}_i - \hat{\boldsymbol{\beta}}_w'\bar{\mathbf{x}}_i, \qquad i = 1, \ldots, N, \end{aligned} \tag{2.2.10}$$

[2] We assume that $T > K + 1$. For details of this, see Chapter 3, Section 3.2.

https://doi.org/10.1017/CBO9781139839327.005 Published online by Cambridge University Press

where

$$W_{xx} = \sum_{i=1}^{N} W_{xx,i} \quad \text{and} \quad W_{xy} = \sum_{i=1}^{N} W_{xy,i}.$$

Let $W_{yy} = \sum_{i=1}^{N} W_{yy,i}$; the residual sum of squares of (2.2.2) is

$$S_2 = W_{yy} - W'_{xy} W_{xx}^{-1} W_{xy}. \tag{2.2.11}$$

The least-squares regression of the pooled model (2.2.4) yields parameter estimates

$$\hat{\boldsymbol{\beta}} = T_{xx}^{-1} T_{xy}, \qquad \hat{\alpha}^* = \bar{y} - \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}, \tag{2.2.12}$$

where

$$T_{xx} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}})(\mathbf{x}_{it} - \bar{\mathbf{x}})',$$

$$T_{xy} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}})(y_{it} - \bar{y}),$$

$$T_{yy} = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \bar{y})^2,$$

$$\bar{y} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it}, \qquad \bar{\mathbf{x}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it}.$$

The (overall) residual sum of squares is

$$S_3 = T_{yy} - T'_{xy} T_{xx}^{-1} T_{xy}. \tag{2.2.13}$$

Under the assumption that the u_{it} are independently normally distributed over i and t with mean 0 and variance σ_u^2, F tests can be used to test the restrictions postulated by (2.2.2) and (2.2.4). In effect, (2.2.2) and (2.2.4) can be viewed as (2.2.1) subject to various types of linear restrictions. For instance, the hypothesis of heterogeneous intercepts but homogeneous slopes [equation (2.2.2)] can be reformulated as (2.2.1) subject to $(N-1)K$ linear restrictions:

$$H_1 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_N.$$

The hypothesis of common intercept and slopes can be viewed as (2.2.1) subject to $(K+1)(N-1)$ linear restrictions:

$$H_3 : \alpha_1^* = \alpha_2^* = \cdots = \alpha_N^*,$$
$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_N.$$

Thus, application of the ANCOVA test is equivalent to the ordinary hypothesis test based on the sums of squared residuals from linear regression outputs.

The unrestricted residual sum of squares S_1 divided by σ_u^2 has a chi square (χ^2) distribution with $NT - N(K+1)$ degrees of freedom. The increment in the explained sum of squares due to allowing for the parameters to vary across i is measured by $(S_3 - S_1)$. Under H_3, the restricted residual sum of squares S_3 divided by σ_u^2 has a χ^2 distribution with $NT - (K+1)$ degrees of freedom, and $(S_3 - S_1)/\sigma_u^2$ has a χ^2 distribution with $(N-1)(K+1)$ degrees of freedom. Because $(S_3 - S_1)/\sigma_u^2$ is independent of S_1/σ_u^2, the F statistic,

$$F_3 = \frac{(S_3 - S_1)/[(N-1)(K+1)]}{S_1/[NT - N(K+1)]}, \tag{2.2.14}$$

can be used to test H_3. If F_3 with $(N-1)(K+1)$ and $N(T-K-1)$ degrees of freedom is not significant, we pool the data and estimate a single equation of (2.2.4). If the F ratio is significant, a further attempt is usually made to find out if the nonhomogeneity can be attributed to heterogeneous slopes or heterogeneous intercepts.

Under the hypothesis of heterogeneous intercepts but homogeneous slopes (H_1), the residual sum of squares of (2.2.2), $S_2 = W_{yy} - W'_{xy} W_{xx}^{-1} W_{xy}$, divided by σ_u^2 has a χ^2 distribution with $N(T-1) - K$ degrees of freedom. The F test of H_1 is thus given by

$$F_1 = \frac{(S_2 - S_1)/[(N-1)K]}{S_1/[NT - N(K+1)]}. \tag{2.2.15}$$

If F_1 with $(N-1)K$ and $NT - N(K+1)$ degrees of freedom is significant, the test sequence is naturally halted and model (2.2.1) is treated as the maintained hypothesis. If F_1 is not significant, we can then determine the extent to which nonhomogeneities can arise in the intercepts.

If H_1 is accepted, one can also apply a conditional test for homogeneous intercepts, namely,

$$H_4 : \alpha_1^* = \alpha_2^* = \cdots = \alpha_N^* \quad \text{given} \quad \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_N.$$

The unrestricted residual sum of squares now is S_2, and the restricted residual sum of squares is S_3. The reduction in the residual sum of squares in moving from (2.2.4) to (2.2.2) is $(S_3 - S_2)$. Under H_4, S_3 divided by σ_u^2 is χ^2 distributed with $NT - (K+1)$ degrees of freedom, and S_2 divided by σ_u^2 is χ^2 distributed with $N(T-1) - K$ degrees of freedom. Because S_2/σ_u^2 is independent of $(S_3 - S_2)/\sigma_u^2$, which is χ^2 distributed with $N-1$ degrees of freedom, the F test for H_4 is

$$F_4 = \frac{(S_3 - S_2)/(N-1)}{S_2/[N(T-1) - K]} \tag{2.2.16}$$

We can summarize these tests in an ANCOVA table (Table 2.1).

Alternatively, we can assume that coefficients are constant across individuals at a given time, but can vary over time. Hence, a separate regression can be

Table 2.1. *Covariance tests for homogeneity*

Source of variation	Residual sum of squares	Degrees of freedom	Mean squares
Within group with heterogeneous intercept and slope	$S_1 = \sum_{i=1}^{N} \left(W_{yy,i} - W'_{xy,i} W^{-1}_{xx,i} W_{xy,i} \right)$	$N(T - K - 1)$	$S_1/N(T - K - 1)$
Constant slope: heterogeneous intercept	$S_2 = W_{yy} - W'_{xy} W^{-1}_{xx} W_{xy}$	$N(T - 1) - K$	$S_2/[N(T - 1) - K]$
Common intercept and slope	$S_3 = T_{yy} - T'_{xy} T^{-1}_{xx} T_{xy}$	$NT - (K + 1)$	$S_3/[NT - (K + 1)]$

Notation:

Cells or groups (or individuals)	$i = 1, \ldots, N$
Observations within cell	$t = 1, \ldots, T$
Total sample size	NT
Within-cell (group) mean	$\bar{y}_i, \bar{\mathbf{x}}_i$
Overall mean	$\bar{y}, \bar{\mathbf{x}}$
Within-group covariance	$W_{yy,i}, W_{yx,i}, W_{xx,i}$
Total variation	T_{yy}, T_{yx}, T_{xx}

postulated for each cross section:

$$y_{it} = \alpha_t^* + \boldsymbol{\beta}_t'\mathbf{x}_{it} + u_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{2.2.17}$$

where we again assume that u_{it} is independently normally distributed with mean 0 and constant variance σ_u^2. Analogous ANCOVA can then be performed to test the homogeneities of the cross-sectional parameters over time. For instance, we can test for overall homogeneity ($H_3' : \alpha_1^* = \alpha_2^* = \cdots = \alpha_T^*, \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T$) by using the F statistic

$$F_3' = \frac{(S_3 - S_1')/[(T-1)(K+1)]}{S_1'/[NT - T(K+1)]}. \tag{2.2.18}$$

with $(T-1)(K+1)$ and $NT - T(K+1)$ degrees of freedom, where

$$\begin{aligned}
S_1' &= \sum_{t=1}^{T}(W_{yy,t} - W_{xy,t}' W_{xx,t}^{-1} W_{xy,t}), \\
W_{yy,t} &= \sum_{i=1}^{N}(y_{it} - \bar{y}_t)^2, \qquad \bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}, \\
W_{xx,t} &= \sum_{i=1}^{N}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)', \qquad \bar{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{it}, \\
W_{xy,t} &= \sum_{i=1}^{N}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(y_{it} - \bar{y}_t).
\end{aligned} \tag{2.2.19}$$

Similarly, we can test the hypothesis of heterogeneous intercepts, but homogeneous slopes ($H_1' : \alpha_1^* \neq \alpha_2^* \neq \cdots \neq \alpha_T^*, \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T$), by using the F statistic

$$F_1' = \frac{(S_2' - S_1')/[(T-1)K]}{S_1'/[NT - T(K+1)]}, \tag{2.2.20}$$

with $(T-1)K$ and $NT - T(K+1)$ degrees of freedom, where

$$S_2' = \sum_{t=1}^{T} W_{yy,t} - \left(\sum_{t=1}^{T} W_{xy,t}'\right)\left(\sum_{t=1}^{T} W_{xx,t}\right)^{-1}\left(\sum_{t=1}^{T} W_{xy,t}\right), \tag{2.2.21}$$

or test the hypothesis of homogeneous intercepts conditional on homogeneous slopes $\beta_1 = \beta_2 = \cdots = \beta_T$ (H_4') by using the F statistic

$$F_4' = \frac{(S_3 - S_2')/(T-1)}{S_2'/[T(N-1) - K]}, \tag{2.2.22}$$

https://doi.org/10.1017/CBO9781139839327.005 Published online by Cambridge University Press

with $(T - 1)$ and $T(N - 1) - K$ degrees of freedom. In general, unless both cross section and time series ANCOVAs indicate the acceptance of homogeneity of regression coefficients, unconditional pooling (i.e., a single least-squares regression using all observations of cross-sectional units through time) may lead to serious bias.

It should be noted that the foregoing tests are not independent. For example, the uncomfortable possibility exists that according to F_3 (or F_3') we might find homogeneous slopes and intercepts, and yet this finding could be compatible with opposite results according to $F_1(F_1')$ and $F_4(F_4')$, because the alternative or null hypotheses are somewhat different in the two cases. Worse still, we might reject the hypothesis of overall homogeneity using the test ratio $F_3(F_3')$, but then find according to $F_1(F_1')$ and $F_4(F_4')$ that we cannot reject the null hypothesis, so that the existence of heterogeneity indicated by F_3 (or F_3') cannot be traced. This outcome is quite proper at a formal statistical level, although at the less formal but important level of interpreting test statistics it is an annoyance.

It should also be noted that the validity of the F-tests are based on the assumption that the errors of the equation, u_{it}, are independently, identically distributed (i.i.d.) and are independent of the regressors, $\mathbf{x}_{it}$ (i.e., the conditional variables, $\mathbf{x}_{it}$, are strictly exogenous (or are fixed constants). In empirical analysis, the errors of the equation could be heteroscedastic or serially correlated, or even correlated with the regressors due to simultaneity or joint dependence. Interpreting F-test statistics ignoring these issues could be seriously misleading. Nevertheless, the idea of F-tests continue to serve as the basis for developing more robust interference procedures (e.g., the robust standard errors of Stock and Watson 2008). Moreover, given the availability of F-test statistics in practically all statistical software packages, it could be considered as a useful first and preliminary step to explore the source of sample variability. We shall discuss some more sophisticated exploratory diagnostic statistics in later chapters when we relax the assumption of "classical" regression model one by one.

2.3 AN EXAMPLE

With the aim of suggesting certain modifications to existing theories of investment behavior and providing estimates of the coefficients of principal interest, Kuh (1963) used data on 60 small and middle-sized firms in capital-goods-producing industries from 1935 to 1955, excluding the war years (1942–1945), to probe the proper specification for the investment function. He explored various models based on capacity accelerator behavior or internal funds flows, with various lags. For ease of illustration, we report here only functional specifications and results based on profit theories, capacity-utilization theories, financial restrictions, and long-run growth theories in arithmetic form (Table 2.2, part A), their logarithmic transformations (part B), and several ratio models (part C). The equations are summarized in Table 2.2.

Table 2.2. *Investment equation forms estimated by Kuh (1963)*

Part A

$$\Delta I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 \Delta K_{it} + \beta_3 \Delta S_{it} \quad (2.3.1)$$
$$\Delta I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 \Delta K_{it} + \beta_4 \Delta P_{it} \quad (2.3.2)$$
$$\Delta I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 \Delta K_{it} + \beta_3 \Delta S_{it} + \beta_4 \Delta P_{it} \quad (2.3.3)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 S_{it} \quad (2.3.4)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_4 P_{it} \quad (2.3.5)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 S_{it} + \beta_4 P_{it} \quad (2.3.6)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 S_{i,t-1} \quad (2.3.7)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_4 P_{i,t-1} \quad (2.3.8)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 S_{i,t-1} + \beta_4 P_{i,t-1} \quad (2.3.9)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 [(S_{it} + S_{i,t-1}) \div 2] \quad (2.3.10)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_4 [(P_{it} + P_{i,t-1}) \div 2] \quad (2.3.11)$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 [(S_{it} + S_{i,t-1}) \div 2] + \beta_4 [(P_{it} + P_{i,t-1}) \div 2] \quad (2.3.12)$$
$$[(I_{it} + I_{i,t-1}) \div 2] = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 [(S_{it} + S_{i,t-1}) \div 2] \quad (2.3.13)$$
$$[(I_{it} + I_{i,t-1}) \div 2] = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_4 [(P_{it} + P_{i,t-1}) \div 2] \quad (2.3.14)$$
$$[(I_{it} + I_{i,t-1}) \div 2] = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 [(S_{it} + S_{i,t-1}) \div 2] + \beta_4 [(P_{it} + P_{i,t-1}) \div 2] \quad (2.3.15)$$

Part B

$$\Delta \log I_{it} = \alpha_0 + \beta_1 \log C_i + \beta_2 \Delta \log K_{it} + \beta_3 \Delta \log S_{it} \quad (2.3.16)$$
$$\log I_{it} = \alpha_0 + \beta_1 \log C_i + \beta_2 \log K_{it} + \beta_3 \log S_{it} \quad (2.3.17)$$
$$\log I_{it} = \alpha_0 + \beta_1 \log C_i + \beta_2 \log K_{it} + \beta_3 \log S_{i,t-1} \quad (2.3.18)$$
$$\log I_{it} = \alpha_0 + \beta_1 \log C_i + \beta_2 \log[(K_{it} + K_{i,t-1}) \div 2] + \beta_3 \log[(S_{it} + S_{i,t-1}) \div 2] \quad (2.3.19)$$

Part C

$$\frac{I_{it}}{K_{it}} = \alpha_0 + \beta_1 \frac{P_{it}}{K_{it}} + \beta_2 \frac{S_{i,t-1}}{C_i \cdot K_{i,t-1}} \quad (2.3.20)$$
$$\frac{I_{it}}{K_{it}} = \alpha_0 + \beta_1 \frac{P_{it}}{K_{it}} + \beta_2 \frac{S_{i,t-1}}{C_i \cdot K_{i,t-1}} + \beta_3 \frac{S_{it}}{C_i \cdot K_{it}} \quad (2.3.21)$$
$$\frac{I_{it}}{K_{it}} = \alpha_0 + \beta_1 \frac{P_{it} + P_{i,t-1}}{K_{it} \cdot 2} + \beta_2 \frac{S_{i,t-1}}{C_i \cdot K_{i,t-1}} \quad (2.3.22)$$
$$\frac{I_{it}}{K_{it}} = \alpha_0 + \beta_1 \frac{P_{it} + P_{i,t-1}}{K_{it} \cdot 2} + \beta_2 \frac{S_{i,t-1}}{C_i \cdot K_{i,t-1}} + \beta_3 \frac{S_{it}}{C_i \cdot K_{it}} \quad (2.3.23)$$

Note: I = gross investment; C = capital-intensity index; K = capital stock; S = sales; P = gross retained profits.

There were two main reasons that Kuh resorted to using individual-firm data rather than economic aggregates. One was the expressed doubt about the quality of the aggregate data, together with the problems associated with estimating an aggregate time series model when the explanatory variables are highly correlated. The other was the desire to construct and test more complicated behavioral models that require many degrees of freedom. However, as stated in Section 1.2, a single regression using all observations through time

https://doi.org/10.1017/CBO9781139839327.005 Published online by Cambridge University Press

makes sense only when individual observations conditional on the explanatory variables can be viewed as random draws from the same universe. Kuh (1963) used the ANCOVA techniques discussed in Section 2.2. to test for overall homogeneity (F_3 or F_3'), slope homogeneity (F_1 or F_1'), and homogeneous intercept conditional on acceptance of homogeneous slopes (F_4 or F_4') for both cross-sectional units and time series units.[3] The results for testing homogeneity of time series estimates across cross-sectional units and homogeneity of cross-sectional estimates over time are reproduced in Tables 2.3 and 2.4, respectively.

A striking fact recorded from these statistics is that except for the time series results for equations (2.3.1) and (2.3.3) (which are in first-difference form), all other specifications failed the overall homogeneity tests.[4] Furthermore, in most cases, with the exception of cross-sectional estimates of (2.3.17) and (2.3.18) (Table 2.4), the intercept and slope variabilities cannot be rigorously separated. Nor do the time series results correspond closely to cross-sectional results for the same equation. Although ANCOVA, like other statistics, is not a mill that will grind out results automatically, these results do suggest that the effects of excluded variables in both time series and cross sections may be very different. It would be quite careless not to explore the possible causes of discrepancies that give rise to the systematic interrelationships between different individuals at different periods of time.[5]

Kuh explored the sources of estimation discrepancies through decomposition of the error variances, comparison of individual coefficient behavior, assessment of the statistical influence of various lag structures, and so forth. He concluded that sales seem to include critical time-correlated elements common to a large number of firms and thus have a much greater capability of annihilating systematic, cyclical factors. In general, his results are more favorable to the acceleration sales model than to the internal liquidity/profit hypothesis supported by the results obtained using cross-sectional data (e.g., Meyer and Kuh 1957). He found that the cash flow effect is more important some time before the actual capital outlays are made than it is in actually restricting the outlays during the expenditure period. It appears more appropriate to view internal liquidity flows as a critical part of the budgeting process that later is modified, primarily in light of variations in levels of output and capacity utilization.

The policy implications of Kuh's conclusions are clear. Other things being equal, a small percentage increase in sales will have a greater effect on

[3] See Johnston (1972, Chapter 6) for an illustration of the computation of analysis of covariance.

[4] If the firm differences stay constant over time, heterogeneity among firms can be absorbed into the intercept term. Because intercepts are eliminated by first-differencing, the first-difference model (such as (2.3.1) or (2.3.3)) will be more likely to display homogeneous responses. See Chapter 3 and Chapter 4.

[5] For further discussion of this issue, see Chapter 11, Section 11.3 and Mairesse (1990).

Table 2.3. *Covariance tests for regression-coefficient homogeneity across cross-sectional units*[a]

Equation	F_3 overall test: Degrees of freedom: Numerator	Denominator	Actual Fs	F_1 slope homogeneity: Degrees of freedom: Numerator	Denominator	Actual Fs	F_4 cell mean significance: Degrees of freedom: Numerator	Denominator	Actual Fs
(2.3.1)	177	660	1.25	118	660	1.75[c]	57	660	0.12
(2.3.2)	177	660	1.40[b]	118	660	1.94[c]	57	660	0.11
(2.3.3)	236	600	1.13	177	600	1.42[b]	56	600	0.10
(2.3.4)	177	840	2.28[c]	118	840	1.58[c]	57	840	3.64[c]
(2.3.5)	177	840	2.34[c]	118	840	1.75[c]	57	840	3.23[c]
(2.3.6)	236	780	2.24[c]	177	780	1.76[c]	56	780	3.57[c]
(2.3.7)	177	720	2.46[c]	118	720	1.95[c]	57	720	3.57[c]
(2.3.8)	177	720	2.50[c]	118	720	1.97[c]	57	720	3.31[c]
(2.3.9)	236	660	2.49[c]	177	660	2.11[c]	56	660	3.69[c]
(2.3.10)	177	720	2.46[c]	118	720	1.75[c]	57	720	3.66[c]
(2.3.11)	177	720	2.60[c]	118	720	2.14[c]	57	720	3.57[c]
(2.3.12)	236	660	2.94[c]	177	660	2.49[c]	56	660	4.18[c]
(2.3.16)	177	720	1.92[c]	118	720	2.59[c]	57	720	0.55
(2.3.17)	177	840	4.04[c]	118	840	2.70[c]	57	840	0.39
(2.3.18)	177	720	5.45[c]	118	720	4.20[c]	57	720	6.32[c]
(2.3.19)	177	720	4.68[c]	118	720	3.17[c]	57	720	7.36[c]
(2.3.20)	177	720	3.64[c]	118	720	3.14[c]	57	720	3.66[c]
(2.3.21)	236	660	3.38[c]	177	660	2.71[c]	56	660	4.07[c]
(2.3.22)	177	600	3.11[c]	118	600	2.72[c]	57	600	3.22[c]
(2.3.23)	236	540	2.90[c]	177	540	2.40[c]	56	540	3.60[c]

[a] Critical F values were obtained from A.M. Mood, *Introduction to Statistics*, Table V, pp. 426–427. Linear interpolation was employed except for degrees of freedom exceeding 120. The critical F values in every case have been recorded for 120 degrees of freedom for each denominator sum of squares even though the actual degrees of freedom were at least four times as great. The approximation error in this case is negligible.

[b] Significant at the 5 percent level.

[c] Significant at the 1 percent level.

Source: Kuh (1963, pp. 141–142).

Table 2.4. *Covariance tests for homogeneity of cross-sectional estimates over time*[a]

	F_3' overall test			F_1' slope homogeneity			F_4' cell mean significance		
	Degrees of freedom			Degrees of freedom			Degrees of freedom		
Equation	Numerator	Denominator	Actual Fs	Numerator	Denominator	Actual Fs	Numerator	Denominator	Actual Fs
(2.3.1)	52	784	2.45[b]	39	784	2.36[b]	10	784	2.89[b]
(2.3.2)	52	784	3.04[b]	39	784	2.64[b]	10	784	4.97[b]
(2.3.3)	65	770	2.55[b]	52	770	2.49[b]	9	770	3.23[b]
(2.3.4)	64	952	2.01[b]	48	952	1.97[b]	13	952	2.43[b]
(2.3.5)	64	952	2.75[b]	48	952	2.45[b]	13	952	3.41[b]
(2.3.6)	80	935	1.91[b]	64	935	1.82[b]	12	935	2.66[b]
(2.3.7)	56	840	2.30[b]	42	840	2.11[b]	11	840	3.66[b]
(2.3.8)	56	840	2.83[b]	42	840	2.75[b]	11	840	3.13[b]
(2.3.9)	70	825	2.25[b]	56	825	2.13[b]	10	825	3.53[b]
(2.3.10)	56	840	1.80[b]	42	840	1.80[b]	11	840	1.72[d]
(2.3.11)	56	840	2.30[b]	42	840	2.30[b]	11	840	1.79[d]
(2.3.12)	70	825	1.70[b]	56	825	1.74[b]	10	825	1.42
(2.3.13)	56	840	2.08[b]	42	840	2.11[b]	11	840	2.21[c]
(2.3.14)	56	840	2.66[b]	42	840	2.37[b]	11	840	2.87[b]
(2.3.15)	70	825	1.81[b]	56	825	1.76[b]	10	825	2.35[c]
(2.3.16)	56	840	3.67[b]	42	840	2.85[b]	11	840	3.10[b]
(2.3.17)	64	952	1.51[c]	48	952	1.14	13	952	0.80
(2.3.18)	56	840	2.34[b]	42	840	1.04	11	840	1.99[c]
(2.3.19)	56	840	2.29[b]	42	840	2.03[b]	11	840	2.05[c]
(2.3.20)	42	855	4.13[b]	28	855	5.01[b]	12	855	2.47[b]
(2.3.21)	56	840	2.88[b]	42	840	3.12[b]	11	840	2.56[b]
(2.3.22)	42	855	3.80[b]	28	855	4.62[b]	12	855	1.61[b]
(2.3.23)	56	840	3.51[b]	42	840	4.00[b]	11	840	1.71[b]

[a] Critical F values were obtained from A.M. Mood, *Introduction to Statistics*, Table V, pp. 426–427. Linear interpolation was employed except for degrees of freedom exceeding 120. The critical F values in every case have been recorded for 120 degrees of freedom for each denominator sum of squares even though the actual degrees of freedom were at least four times as great. The approximation error in this case is negligible.
[b] Significant at the 1 percent level.
[c] Significant at the 5 percent level.
[d] Significant at the 10 percent level.

Source: Kuh (1963, pp. 137–138).

investment than will a small percentage increase in internal funds. If the government seeks to stimulate investment and the objective is magnitude, not qualitative composition, it inexorably follows that the greatest investment effect will come from measures that increase demand rather than from measures that increase internal funds.[6]

[6] For further discussion of investment expenditure behavior, see Chapter 6 or Hsiao and Tahmiscioglu (1997).