Chapter 3. Models with Random Effects

Abstract. This chapter considers the Chapter 2 data structure yet now models the heterogeneity using random quantities in lieu of fixed parameters; these random quantities are known as random effects. By introducing linear random quantities, the analysis of longitudinal and panel data can now be cast in the mixed linear model framework.

Although mixed linear models are an established part of statistical methodology, their use is not as widespread as regression. Thus, the chapter introduces this modeling framework, beginning with the special case of a single random intercept known as the error components model and then focusing on the linear mixed effects model that is particularly important for longitudinal data. After introducing the models, this chapter describes estimation of regression coefficients and variance components, as well as hypothesis testing for regression coefficients.

3.1 Error components model

Sampling and inference

Suppose that you are interested in studying the behavior of individuals that are randomly selected from a population. For example, in Section 3.2 we will study the effects that an individual’s economic and demographic characteristics have on the amount of income tax paid. Here, the set of subjects that we will study is randomly selected from a larger database that is itself a random sample of the US taxpayers. In contrast, the Chapter 2 Medicare example dealt with a fixed set of subjects. That is, it is difficult to think of the 54 states as a subset from some “super-population” of states. For both situations, it is fruitful to use subject-specific parameters, \{\alpha_i\}, to represent the heterogeneity among subjects. Unlike Chapter 2, Chapter 3 discusses situations in which it is more reasonable to represent \{\alpha_i\} as random variables instead of fixed, yet unknown, parameters. By arguing that \{\alpha_i\} are draws from a distribution, we will have the ability to make inferences about subjects in a population that are not included in the sample.

Basic model and assumptions

The error components model equation is

\[ y_{it} = \alpha_i + x_{it}' \beta + \epsilon_{it} \]  

(3.1)

This portion of the notation is the same as the error representation of the basic fixed model. However, now the term \alpha_i is assumed to be a random variable, not a fixed, unknown parameter. The term \alpha_i is known as a random effect. Mixed effects models are ones that include random as well as fixed effects. Because equation (3.1) includes random effects and fixed effects, the error components model is a special case of the mixed linear model.

To complete the specification of the error components model, we assume that \{\alpha_i\} are identically and independently distributed with mean zero and variance \sigma^2_\alpha. Further, we assume that \{\alpha_i\} are independent of the error random variables, \{\epsilon_{it}\}. For completeness, we still assume
that \( \mathbf{x}_i \) is a vector of covariates, or explanatory variables, and that \( \mathbf{\beta} \) is a vector of fixed, yet unknown, global parameters. Note that because \( \mathbb{E} \alpha_i = 0 \), it is customary to include a constant within the vector \( \mathbf{x}_i \). This was not true of the fixed effects models in Chapter 2 where we did not center the subject-specific terms about 0.

Linear combinations of the form \( \mathbf{x}_i' \mathbf{\beta} \) quantify the effect of known variables that may affect the response. Additional variables, that are either unimportant or unobservable, comprise the “error term.” In the error component model, we may think of a regression model \( y_{it} = \mathbf{x}_{it}' \mathbf{\beta} + \eta_{it} \), where the error term \( \eta_{it} \) is decomposed into two components so that \( \eta_{it} = \alpha_i + \varepsilon_{it} \). The term \( \alpha_i \) represents the time-constant portion whereas \( \varepsilon_{it} \) represents the remaining portion. To identify the model parameters, we assume that the two terms are independent.

**Traditional ANOVA set-up**

In the error components model, the terms \( \{ \alpha_i \} \) account for the heterogeneity. To help interpret this feature, consider the special case where \( K = 1 \), \( x_{it} = 1 \) and denote \( \mu = \beta_1 \). In this case, equation (3.1) contains no explanatory variables and reduces to

\[
y_{it} = \mu + \alpha_i + \varepsilon_{it},
\]

the traditional random effects, one-way ANOVA model. Neter and Wasserman (1974G) describe this classic model. This model can be interpreted as arising from a two-stage sampling scheme.

**Stage 1.** Draw a random sample of \( n \) subjects from a population. The subject-specific parameter \( \alpha_i \) is associated with the \( i \)th subject.

**Stage 2.** Conditional on \( \alpha_i \), draw realizations of \( \{ y_{it} \} \), for \( t = 1, \ldots, T_i \) for the \( i \)th subject.

That is, in the first stage, we draw a sample from a population of subjects. In the second stage, we observe each subject over time. Because the first stage is considered a draw from a population of subjects, we represent characteristics that do not depend on time as random through the quantity \( \alpha_i \). Figure 3.1 illustrates the two-stage sampling scheme.

![Figure 3.1. Two-stage random effects sampling.](image-url)
Within this traditional model, questions of interest generally center about the distribution of the population of subjects. For example, the parameter \( \text{Var } \alpha_i = \sigma_a^2 \) summarizes the heterogeneity among subjects. In Chapter 2 on fixed effects models, we examined the heterogeneity issue through a test of the null hypothesis \( H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_n \). In contrast, under the random effects model, we examine the null hypothesis \( H_0: \sigma_a^2 = 0 \). As another example, because of the heterogeneity issue, estimates of \( \sigma_a^2 \) are of interest but require scaling to interpret. A more useful quantity to report is \( \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \), the *intra-class correlation*. As we saw in Section 2.5.1, this quantity can be interpreted as the correlation between observations within a subject. The correlation is constrained to lie between 0 and 1 and does not depend on the units of measurement for the response. Further, it can also be interpreted as the proportion of variability of a response that is due to heterogeneity.

### Sampling and model assumptions

The Section 2.1 basic fixed effects and error components models are similar in appearance yet, as will be discussed in Section 7.2, can lead to different substantive conclusions in the context of a specific application. As we have described, the choice between these two models is dictated primarily by the method in which the sample is drawn. On the one hand, selecting subjects based on a two-stage, or *cluster*, sample implies use of the random effects model. On the other hand, selecting subjects based on exogenous characteristics suggests a stratified sample and thus using a fixed effects model.

The sampling basis allows us to restate the error components model, as follows.

**Error Components Model Assumptions**

| R1 | \( E(y_{it} | \alpha_i) = \alpha_i + x_{it}' \beta \). |
| R2 | \( \{x_{it,1}, \ldots, x_{it,K}\} \) are nonstochastic variables. |
| R3 | \( \text{Var } y_{it} | \alpha_i = \sigma_e^2 \). |
| R4 | \( \{y_{it}\} \) are independent random variables, conditional on \( \{\alpha_1, \ldots, \alpha_n\} \). |
| R5 | \( y_{it} \) is normally distributed, conditional on \( \{\alpha_1, \ldots, \alpha_n\} \). |
| R6 | \( E \alpha_i = 0, \text{Var } \alpha_i = \sigma_a^2 \) and \( \{\alpha_1, \ldots, \alpha_n\} \) are mutually independent. |
| R7 | \( \{\alpha_i\} \) is normally distributed. |

Assumptions R1-R5 are similar to the fixed effects models assumptions F1-F5; the main difference is that we now condition on random subject-specific terms, \( \{\alpha_1, \ldots, \alpha_n\} \). Assumptions R6 and R7 summarize the sampling basis of the subject-specific terms. Taken together, these assumptions comprise our *error components model*.

However, assumptions R1-R7 do not provide an “observables” representation of the model because they are based on unobservable quantities, \( \{\alpha_1, \ldots, \alpha_n\} \). We summarize the effects of Assumptions R1-R7 on the observables variables, \( \{x_{it,1}, \ldots, x_{it,K}, y_{it}\} \).

**Observables Representation of the Error Components Model**

| RO1 | \( E y_{it} = x_{it}' \beta \). |
| RO2 | \( \{x_{it,1}, \ldots, x_{it,K}\} \) are nonstochastic variables. |
| RO3 | \( \text{Var } y_{it} = \sigma_e^2 + \sigma_a^2 \) and \( \text{Cov } (y_{ir}, y_{is}) = \sigma_a^2 \), for \( r \neq s \). |
| RO4 | \( \{y_{it}\} \) are independent random vectors. |
| RO5 | \( \{y_{it}\} \) is normally distributed. |

To reiterate, the properties RO1-5 are a consequence of R1-R7. As we progress into more complex situations, our strategy will consist of using sampling bases to suggest basic assumptions, such as R1-R7, and then convert them into testable properties such as RO1-5. Inference about the testable properties then provides information about the more basic
assumptions. When considering nonlinear models beginning in Chapter 9, this conversion will not be as direct. In some instances, we will focus on the observable representation directly and refer to it as a marginal or population-averaged model. The marginal version emphasizes the assumption that observations are correlated within subjects (Assumption RO3), not the random effects mechanism for inducing the correlation.

For more complex situations, it will be useful to describe these assumptions in matrix notation. As in equation (2.13), the regression function can be expressed more compactly as
\[ \mathbb{E}(y_i | a_i) = a_i \mathbf{I}_i + \mathbf{X}_i \mathbf{\beta} \]
and thus,
\[ \mathbb{E} y_i = \mathbf{X}_i \mathbf{\beta}. \] (3.2)

Recall that \( \mathbf{I}_i \) is a \( T_i \times 1 \) vector of ones and, from equation (2.14), that \( \mathbf{X}_i \) is a \( T_i \times K \) matrix of explanatory variables, \( \mathbf{X}_i = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{iT_i} \end{pmatrix} \). The expression for \( \mathbb{E}(y_i | a_i) \) is a restatement of Assumption R1 in matrix notation. Equation (3.2) is a restatement of Assumption RO1. Alternatively, equation (3.2) is due to the law of iterated expectations and assumptions R1 and R6, because \( \mathbb{E} y_i = \mathbb{E} (\mathbb{E}(y_i | a_i)) = \mathbb{E} a_i \mathbf{I}_i + \mathbf{X}_i \mathbf{\beta} = \mathbf{X}_i \mathbf{\beta} \). For Assumption RO3, we have
\[ \text{Var} y_i = \mathbf{V}_i = \sigma_a^2 \mathbf{J}_i + \sigma^2 \mathbf{I}_i. \] (3.3)

Here, recall that \( \mathbf{J}_i \) is a \( T_i \times T_i \) matrix of ones, and \( \mathbf{I}_i \) is a \( T_i \times T_i \) identity matrix.

**Structural models**

Model assumptions are often dictated by sampling procedures. However, we also wish to consider stochastic models that represent causal relationships suggested by a substantive field, known as structural models. Section 6.1 describes structural modeling in longitudinal and panel data analysis. To illustrate, in models of economic applications, it is important to consider carefully what one means by the “population of interest.” Specifically, when considering choices of economic entities, a standard defense for a probabilistic approach to analyzing economic decisions is that, although there may be a finite number of economic entities, there is an infinite range of economic decisions. For example, in the Chapter 2 Medicare hospital cost example, one may argue that each state faces a distribution of infinitely many economic outcomes and that this is the population of interest. This viewpoint argues that one should use an error components model. Here, we interpret \( \{a_i\} \) to represent those aspects of the economic outcome that are unobservable yet constant over time. In contrast, in Chapter 2 we implicitly used the sampling based model to interpret \( \{a_i\} \) as fixed effects.

This viewpoint is the standard rationale for studying stochastic economics. To illustrate, a quote from Haavelmo (1944E) is related to this point:

“... the class of populations we are dealing with does not consist of an infinity of different individuals, it consists of an infinity of possible decisions which might be taken with respect to the value of \( y \).”

This defense is well summarized by Nerlove and Balestra in a monograph edited by Mátyás and Sevestre (1996E, Chapter 1) in the context of panel data modeling.

**Inference**

When designing a longitudinal study and considering whether to use a fixed or random effects methods, keep in mind the purposes of the study. If you would like to make statements about a population larger than the sample, then use the random effects model. Conversely, if you are simply interested in controlling for subject-specific effects (treating them as nuisance parameters) or in making predictions for a specific subject, then use the fixed effects model.
Time-constant variables

When designing a longitudinal study and considering whether to use a fixed or random effects methods, also keep in mind the variables of interest. Often, the primary interest is in testing for the effect of a time-constant variable. To illustrate, in our taxpayer example, we may be interested in the effects that gender may have on an individual’s tax liability. (We assume that this variable does not change for an individual over the course of our study.) Another important example of a time-constant variable is a variable that classifies subjects into groups. Often, we wish to compare the performance of different groups, for example, a “treatment group” and a “control group.”

In Section 2.3, we saw that time-constant variables are perfectly collinear with subject-specific intercepts and hence are inestimable. In contrast, it will turn out that coefficients associated with time-constant variables are estimable in a random effects model. Hence, if a time-constant variable such as gender or treatment group is the primary variable of interest, one should design the longitudinal study so that a random effects model can be used.

Degrees of freedom

When designing a longitudinal study and considering whether to use a fixed or random effects methods, also keep in mind the size of the data set necessary for inference. In most longitudinal data studies, inference about the population parameters $\beta$ is the primary goal whereas the terms $\{\alpha_i\}$ are included to control for the heterogeneity. In the basic fixed effects model, we have seen that there are $n+K$ linear regression parameters plus 1 variance parameter. This is compared to only $1+K$ regression plus 2 variance parameters in the basic random effects model. Particularly in studies where the time dimension is small (such as $T=2$ or 3), a design suggesting a random effects model may be preferable because fewer degrees of freedom are necessary to account for the subject-specific parameters.

GLS estimation

Equations (3.2) and (3.3) summarize the mean and variance of the vector of responses. To estimate regression coefficients, this chapter uses generalized least squares (GLS) equations of the form:

$$\left( \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right) \hat{\beta} = \sum_{i=1}^{n} X_i' V_i^{-1} y_i.$$  

The solution of these equations yields generalized least square estimators that, in this context, we call the error components estimator of $\beta$. Additional algebra (Exercise 3.1) shows that this estimator can be expressed as

$$\mathbf{b}_{EC} = \left( \sum_{i=1}^{n} X_i' \left( I_i - \frac{\zeta_i}{T_i} J_i \right) X_i \right)^{-1} \sum_{i=1}^{n} X_i' \left( I_i - \frac{\zeta_i}{T_i} J_i \right) y_i.$$  

Here, the quantity $\zeta_i = \frac{T_i \sigma_a^2}{T_i \sigma_a^2 + \sigma^2}$ is a function of the variance components $\sigma_a^2$ and $\sigma^2$. In Chapter 4, we will refer to this quantity as the credibility factor. Further, the variance of the random effects estimator turns out to be

$$\text{Var} \mathbf{b}_{EC} = \sigma^2 \left( \sum_{i=1}^{n} X_i' \left( I_i - \frac{\zeta_i}{T_i} J_i \right) X_i \right)^{-1}.$$
To interpret $b_{EC}$, we give an alternative form for the corresponding Chapter 2 fixed effects estimator. That is, from equation (2.6) and some algebra, we have

$$b = \left( \sum_{i=1}^{n} X_i' \left( I_i - T_i^{-1} J_i \right) X_i \right)^{-1} \sum_{i=1}^{n} X_i' \left( I_i - T_i^{-1} J_i \right) y_i.$$ 

Thus, we see that the random effects $b_{EC}$ and fixed effects $b$ are approximately equal when the credibility factors are close to one. This occurs when $\sigma_{\alpha}^2$ is large relative to $\sigma^2$. Intuitively, when there is substantial separation among the intercept terms, relative to the uncertainty in the observations, we anticipate that the fixed and random effect estimators will behave similarly. Conversely, equation (3.3) shows that $b_{EC}$ is approximately equal to an ordinary least squares estimator when $\sigma^2$ is large relative to $\sigma_{\alpha}^2$ (so that the credibility factors are close to zero). Section 7.2 further develops the comparison among these alternative estimators.

**Feasible generalized least squares estimator**

The calculation of the GLS estimator in equation (3.4) assumes that the variance components $\sigma_{\alpha}^2$ and $\sigma^2$ are known.

**Procedure for computing a “feasible” generalized least squares estimator**

1. First run a regression assuming $\sigma_{\alpha}^2 = 0$, resulting in an ordinary least squares estimate of $\beta$.
2. Use the residuals from Step 1 to determine estimates of $\sigma_{\alpha}^2$ and $\sigma^2$.
3. Using the estimates of $\sigma_{\alpha}^2$ and $\sigma^2$ from Step 2, determine $b_{EC}$ using equation (3.4).

For Step 2, there are many ways of estimating the variance components. Section 3.5 provides details. This procedure could be iterated. However, studies have shown that iterated versions do not improve the performance of the one-step estimators. See, for example, Carroll and Rupert (1988G).

To illustrate, we consider some simple moment-based estimators of $\sigma_{\alpha}^2$ and $\sigma^2$ due to Baltagi and Chang (1994E). Define the residuals $e_{it} = y_{it} - (a_i + x_{it}' b)$ using $a_i$ and $b$ according to the Chapter 2 fixed effects estimators in equations (2.6) and (2.7). Then, the estimator of $\sigma^2$ is $s^2$ as given in equation (2.11). The estimator of $\sigma_{\alpha}^2$ is:

$$s_{\alpha}^2 = \frac{\sum_{i=1}^{n} T_i (a_i - \bar{a}_w)^2 - s^2 c_n}{N - \sum_{i=1}^{n} T_i^2 / N},$$

where $\bar{a}_w = N^{-1} \sum_{i=1}^{n} T_i a_i$ and

$$c_n = n - 1 + \text{trace} \left[ \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^{n} T_i (\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})' \right].$$

A potential drawback is that a particular realization of $s_{\alpha}^2$ may be negative; this feature is undesirable for a variance estimator.

**Pooling test**

As with the traditional random effects ANOVA model, the test for heterogeneity, or pooling test, is written as a test of the null hypothesis $H_0: \sigma_{\alpha}^2 = 0$. That is, under the null hypothesis, we do not have to account for subject-specific effects. Although this is a difficult issue for the general case, in the special case of error components, desirable test procedures have been developed. We discuss here a test that extends a Lagrange multiplier test statistic due to
Breusch and Pagan (1980E) to the unbalanced data case. (See Appendix C.7 for an introduction to Lagrange multiplier statistics.) This test is a simpler version of one developed by Baltagi and Li (1990E) for a more complex model (specifically, a two-way error component model that we will introduce in Chapter 6).

**Pooling test procedure**

1. Run the pooled cross-sectional regression model $y_{it} = x_{it}' \beta + \varepsilon_{it}$ to get residuals $e_{it}$.

2. For each subject, compute an estimator of $\sigma_\alpha^2$, $s_i = \frac{1}{T_i(T_i - 1)} \left( T_i^2 \overline{e}_i^2 - \sum_{t=1}^{T_i} e_{it}^2 \right)$, where

$$\overline{e}_i = T_i^{-1} \sum_{t=1}^{T_i} e_{it}.$$

3. Compute the test statistic, $TS = \frac{1}{2n} \left( \frac{n}{N} \right)^{-1} \left( \sum_{i=1}^{n} s_i \sqrt{T_i(T_i - 1)} \right)^2$.

4. Reject $H_0$ if $TS$ exceeds a percentile from a $\chi^2$ (chi-square) distribution with one degree of freedom. The percentile is one minus the significance level of the test.

Note that the pooling test procedure uses estimators of $\sigma_\alpha^2$, $s_i$, that may be negative with positive probability. Section 5.4 discusses alternative procedures where we restrict variance estimators to be nonnegative.

### 3.2 Example: Income tax payments

In this section, we study the effects that an individual’s economic and demographic characteristics have on the amount of income tax paid. Specifically, the response of interest is LNTAX, defined as the natural logarithm of the liability on the tax return. Table 3.1 describes several taxpayer characteristics that may affect tax liability.

The data for this study are from the Statistics of Income (SOI) Panel of Individual Returns, a part of the Ernst and Young/University of Michigan Tax Research Database. The SOI Panel represents a simple random sample of unaudited individual income tax returns filed for tax years 1979-1990. The data are compiled from a stratified probability sample of unaudited individual income tax returns, Forms 1040, 1040A and 1040EZ, filed by U.S. taxpayers. The estimates that are obtained from these data are intended to represent all returns filed for the income tax years under review. All returns processed are subjected to sampling except tentative and amended returns.

We examine a balanced panel from 1982-1984 and 1986-1987 taxpayers included in the SOI panel; a four percent sample of this comprises our sample of 258 taxpayers. These years are chosen because they contain the interesting information on paid preparer usage. Specifically, these data include line item tax return data plus a binary variable noting the presence of a paid tax preparer for years 1982-1984 and 1986-1987. These data are also analyzed in Frischmann and Frees (1999O).

The primary goal of this analysis is to determine whether tax preparers significantly affect tax liability. To motivate this question, we note that preparers have the opportunity to impact virtually every line item on a tax return. Our variables are selected because they appear consistently in prior research and are largely outside the influence of tax preparers (that is, they are “exogenous”). Briefly, our additional explanatory variables are as follows: MS, HH, AGE, EMP, and PREP are binary variables coded for married, head-of-household, at least 65 years of age, employed, and presence of a paid preparer, respectively.
age, self-employed, and paid preparer, respectively. Further, DEPEND is the number of dependents and MR is the marginal tax rate measure. Finally, LNTPI and LNTAX are the total positive income and tax liability as stated on the return in 1983 dollars, in logarithmic units.

Table 3.1 Taxpayer Characteristics

<table>
<thead>
<tr>
<th>Demographic Characteristics</th>
<th>Economic Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td>LNTPI</td>
</tr>
<tr>
<td>HH</td>
<td>is the natural logarithm of the sum of all positive income line items on the return, in 1983 dollars.</td>
</tr>
<tr>
<td>DEPEND</td>
<td>MR</td>
</tr>
<tr>
<td>AGE</td>
<td>is the marginal tax rate. It is computed on total personal income less exemptions and the standard deduction.</td>
</tr>
<tr>
<td>EMP</td>
<td>PREP</td>
</tr>
<tr>
<td>LNTAX</td>
<td>is a variable indicating the presence of a paid preparer.</td>
</tr>
</tbody>
</table>

Tables 3.2 and 3.3 describe the basic taxpayer characteristics used in our analysis. The binary variables in Table 3.2 indicate that over half the sample is married (MS) and approximately half the sample uses a paid preparer (PREP). Preparer use appears highest in 1986 and 1987, years straddling significant tax law change. Slightly less than ten percent of the sample is 65 or older (AGE) in 1982. The presence of self-employment income (EMP) also varies over time.

<table>
<thead>
<tr>
<th>YEAR</th>
<th>MS</th>
<th>HH</th>
<th>AGE</th>
<th>EMP</th>
<th>EMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>0.597</td>
<td>0.081</td>
<td>0.085</td>
<td>0.140</td>
<td>0.450</td>
</tr>
<tr>
<td>1983</td>
<td>0.597</td>
<td>0.093</td>
<td>0.105</td>
<td>0.159</td>
<td>0.442</td>
</tr>
<tr>
<td>1984</td>
<td>0.624</td>
<td>0.085</td>
<td>0.112</td>
<td>0.155</td>
<td>0.484</td>
</tr>
<tr>
<td>1986</td>
<td>0.647</td>
<td>0.081</td>
<td>0.132</td>
<td>0.147</td>
<td>0.508</td>
</tr>
<tr>
<td>1987</td>
<td>0.647</td>
<td>0.093</td>
<td>0.147</td>
<td>0.147</td>
<td>0.516</td>
</tr>
</tbody>
</table>

The summary statistics for the other, non-binary, variables are in Table 3.3. Further analyses indicate an increasing income trend, even after adjusting for inflation, as measured by total positive income (LNTPI). Both the mean and median marginal tax rates (MR) are decreasing, although mean and median tax liabilities (LNTAX) are increasing. These results are consistent with congressional efforts to reduce rates and expand the tax base through broadening the definition of income and eliminating deductions.
To explore the relationship between each indicator variable and logarithmic tax, Table 3.4 presents the average logarithmic tax liability by level of indicator variable. Table 3.4 shows that married filers pay greater tax, head of household filers pay less tax, taxpayers 65 or over pay less, taxpayers with self-employed income pay less and taxpayers that use a professional tax preparer pay more.

Table 3.5 summarizes basic relations among logarithmic tax and the other, non-binary, explanatory variables. Both LNTPI and MR are strongly correlated with logarithmic tax whereas the relationship between DEPEND and logarithmic tax is positive, yet weaker. Further, Table 3.5 shows that LNTPI and MR are strongly positively correlated.

Although not presented in detail here, exploration of the data revealed several other interesting relationships among the variables. To illustrate, a basic added variable plot in Figure 3.2 shows the strong relation between logarithmic tax liability and total income, even after controlling for subject-specific time-constant effects.
The error components model described in Section 3.1 was fit, using the explanatory variables described in Table 3.1. The estimated model appears in Display 3.1, from a fit using the statistical package SAS. Display 3.1 shows that HH, EMP, LNTPI and MR are statistically significant variables that affect LNTAX. Somewhat surprisingly, the PREP variable was not statistically significant.

To test for the importance of heterogeneity, the Section 3.1 pooling test was performed. A fit of the pooled cross-sectional model, with the same explanatory variables, produced residuals and an error sum of squares equal to $Error \ SS = 3599.73$. Thus, with $T = 5$ years and $n = 258$ subjects, the test statistic is $TS = 273.5$. Comparing this test statistic to a chi-square distribution with one degree of freedom indicates that the null hypothesis of homogeneity is rejected. As we will see in Chapter 7, there are some unusual features of this data set that cause this test statistic to be large.
### Display 3.1 Selected SAS Output

#### Iteration History

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>-2 Log Like</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>4984.68064143</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4791.25465804</td>
<td>0.00000001</td>
</tr>
</tbody>
</table>

Convergence criteria met.

#### Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>SUBJECT</td>
<td>0.9217</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>1.8740</td>
</tr>
</tbody>
</table>

#### Fit Statistics

-2 Log Likelihood: 4791.3
AIC (smaller is better): 4813.3
AICC (smaller is better): 4813.5
BIC (smaller is better): 4852.3

#### Solution for Fixed Effects

| Effect  | Estimate | Error  | DF | t Value | Pr > |t|   |
|---------|----------|--------|----|---------|-------|----|
| Intercept | -2.9604  | 0.5686 | 257| -5.21   | <.0001|
| MS      | 0.03730  | 0.1818 | 1024| 0.21    | 0.8375|
| HH      | -0.6890  | 0.2312 | 1024| -2.98   | 0.0029|
| AGE     | 0.02074  | 0.1993 | 1024| 0.10    | 0.9171|
| EMP     | -0.5048  | 0.1674 | 1024| -3.02   | 0.0026|
| PREP    | -0.02170 | 0.1171 | 1024| -0.19   | 0.8530|
| LNTPI   | 0.7604   | 0.06972| 1024| 10.91   | <.0001|
| DEPEND  | -0.1128  | 0.05907| 1024| -1.91   | 0.0566|
| MR      | 0.1154   | 0.007288| 1024| 15.83   | <.0001|
3.3 Mixed effects models

Similar to the extensions for the fixed effects model described in Section 2.5, we now extend the error components model to allow for variable slopes, serial correlation and heteroscedasticity.

3.3.1 Linear mixed effects model

We now consider conditional regression functions of the form

$$E(y_{it} | \alpha_i) = z_{it}' \alpha_i + x_{it}' \beta.$$  \hspace{1cm} (3.5)

Here, the term $z_{it}' \alpha_i$ comprises the random effects portion of the model. The term $x_{it}' \beta$ comprises the fixed effects portion. As with equation (2.15) for fixed effects, equation (3.5) is short-hand notation for

$$E(y_{it} | \alpha_i) = \alpha_{i1} z_{i1} + \alpha_{i2} z_{i2} + \ldots + \alpha_{iq} z_{iq} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{ik}.$$ 

As in equation (2.16), a matrix form of equation (3.5) is

$$E(y_i | \alpha_i) = Z_i \alpha_i + X_i \beta.$$ \hspace{1cm} (3.6)

We also wish to allow for serial correlation and heteroscedasticity. Similar to Section 2.5.1 for fixed effects, we can incorporate these extensions through the notation $\text{Var}(y_i | \alpha_i) = R_i$. We maintain the assumption that the responses between subjects are independent.

Further, we assume that the subject-specific effects $\{\alpha_i\}$ are independent with mean $E(\alpha_i) = 0$ and variance-covariance matrix $\text{Var}(\alpha_i) = D$, a $q \times q$ positive definite matrix. By assumption, the random effects are mean zero; thus, any nonzero mean for a random effect must be expressed as part of the fixed effects terms. Thus, the columns of $Z_i$ are usually a subset of the columns of $X_i$.

Taken together, these assumptions comprise what we term the linear mixed effects model.

### Linear Mixed Effects Model Assumptions

- **R1.** $E(y_i | \alpha_i) = Z_i \alpha_i + X_i \beta$.
- **R2.** $\{x_{it,1}, \ldots, x_{it,K}\}$ and $\{z_{it,1}, \ldots, z_{it,q}\}$ are nonstochastic variables.
- **R3.** $\text{Var}(y_i | \alpha_i) = R_i$.
- **R4.** $\{y_i\}$ are independent random vectors, conditional on $\{\alpha_i, \ldots, \alpha_n\}$.
- **R5.** $\{y_i\}$ is normally distributed, conditional on $\{\alpha_{i1}, \ldots, \alpha_{in}\}$.
- **R6.** $E(\alpha_i) = 0$, $\text{Var}(\alpha_i) = D$ and $\{\alpha_{i1}, \ldots, \alpha_{in}\}$ are mutually independent.
- **R7.** $\{\alpha_i\}$ is normally distributed.

With assumptions R3 and R6, the variance of each subject can be expressed as

$$\text{Var}(y_i) = Z_i D Z_i' + R_i = V_i(\tau) = V_i.$$ \hspace{1cm} (3.7)

The notation $V_i(\tau)$ means that the variance-covariance matrix of $y_i$ depends on variance components $\tau$. Section 2.5.1 provided several examples that illustrate how $R_i$ may depend on $\tau$; we will give special cases to show how $V_i$ may depend on $\tau$.

With this, we may summarize the effects of Assumptions R1-R7 on the observables variables, $\{x_{it,1}, \ldots, x_{it,K}, z_{it,1}, \ldots, z_{it,q}, y_{it}\}$.
Observables Representation of the Linear Mixed Effects Model

RO1. \( E y_i = X \beta \).

RO2. \( \{x_{it1}, \ldots, x_{itK}\} \) and \( \{z_{i01}, \ldots, z_{i0q}\} \) are nonstochastic variables.

RO3. \( \text{Var} y_i = Z_i D Z_i' + R_i = V(\tau) = V_i \).

RO4. \( \{y_i\} \) are independent random vectors.

RO5. \( \{y_i\} \) is normally distributed.

As in Chapter 2 and Section 3.1, the properties RO1-5 are a consequence of R1-R7. We focus on these properties because they are the basis for testing our specification of the model. The observable representation is also known as a marginal or population-averaged model.

Example – Trade localization

Feinberg, Keane and Bognano (1998E) studied \( n = 701 \) U.S. based multinational corporations over the period 1983-1992. Using firm-level data available from the Bureau of Economic Analysis of the U.S. Department of Commerce, they documented how large corporation’s allocation of employment and durable assets (property, plant and equipment) of Canadian affiliates changed in response to changes in Canadian and U.S. tariffs. Specifically, their model was

\[
\ln y_{it} = \beta_{1i} CT_{it} + \beta_{2i} UT_{it} + \beta_{3i} \text{Trend}_t + x_{it}^{**}' \beta^{**} + \epsilon_{it}
\]

\[
= (\beta_1 + \alpha_{1i}) CT_{it} + (\beta_2 + \alpha_{2i}) UT_{it} + (\beta_3 + \alpha_{3i}) \text{Trend}_t + x_{it}^{**}' \beta^{**} + \epsilon_{it}
\]

\[
= \alpha_{1i} CT_{it} + \alpha_{2i} UT_{it} + \alpha_{3i} \text{Trend}_t + x_{it}' \beta + \epsilon_{it}.
\]

Here, \( CT_{it} \) is the sum over all industry Canadian tariffs in which firm \( i \) belongs, and similarly for \( UT_{it} \). The vector \( x_{it} \) includes \( CT_{it}, UT_{it} \) and \( \text{Trend}_t \) (for the mean effects), as well as real U.S. and Canadian wages, gross domestic product, price–earnings ratio, real U.S. interest rates and a measure of transportation costs. For the response, they used both Canadian employment and durable assets.

The first equation emphasizes that response to changes in Canadian and U.S. tariffs, as well as time trends, is firm-specific. The second equation provides the link to the third expression that is in terms of the linear mixed effects model form. Here, we have included \( CT_{it}, UT_{it} \) and \( \text{Trend}_t \) in \( x_{it}^{**} \) to get \( x_{it} \). With this reformulation, the mean of each random slope is zero, that is, \( \text{E} \alpha_{1i} = \text{E} \alpha_{2i} = \text{E} \alpha_{3i} = 0 \). In the first specification, the means are \( \text{E} \beta_{1i} = \beta_1, \text{E} \beta_{2i} = \beta_2 \) and \( \text{E} \beta_{3i} = \beta_3 \). Feinberg, Keane and Bognano found that a significant portion of the variation was due to firm-specific slopes; they attribute this variation to idiosyncratic firm differences such as technology and organization. They also allowed for heterogeneity in the time trend. This allows for unobserved time-varying factors (such as technology and demand) that affect individual firms differently.

A major finding of this paper is that Canadian tariff levels were negatively related to assets and employment in Canada; this finding contradicts the hypothesis that lower tariffs would undermine Canadian manufacturing.
Special cases
To help interpret linear mixed effects models, we consider several important special cases. We begin by emphasizing the case where $q = 1$ and $z_{it} = 1$. In this case, the linear mixed effects model reduces to the error components model, introduced in Section 3.1. For this model, we have only subject-specific intercepts, no subject-specific slopes and no serial correlation.

Repeated measures design
Another classic model is the so-called repeated measures design. Here, several measurements are collected on a subject over a relatively short period of time, under controlled experimental conditions. Each measurement is subject to a different treatment but the order of treatments is randomized so that no serial correlation is assumed.

Specifically, we consider $i = 1, \ldots, n$ subjects. A response for each subject is measured based on each of $T$ treatments, with the order of treatments is randomized. The mathematical model is:

$$ y_{it} = \alpha_i + \beta_i x_{it} + \epsilon_{it} $$

The main research question of interest is $H_0: \beta_1 = \beta_2 = \ldots = \beta_T$, that of no treatment differences.

The repeated measures design is a special case of equation (3.5), taking $q = 1$, $z_{it} = 1$, $T_i = T$, $K = T$ and using the $t$th explanatory variable, $x_{it,t}$, to indicate whether the $t$th treatment has been applied to the response.

Random coefficients model
We now return to the linear mixed effects model and suppose that $q = K$ and $z_{it} = x_{it}$. In this case the linear mixed effects model reduces to a random coefficients model, of the form

$$ E(y_{it} | \alpha_i) = x_{it}'(\alpha_i + \beta_i) = x_{it}'\beta_i. \quad (3.8) $$

Here, $\{\beta_i\}$ are random vectors with mean $\beta$. The random coefficients model can be easily interpreted as a two-stage sampling model. In the first stage, one draws the $i$th subject from a population that yields a vector of parameters $\beta_i$. From the population, this vector has mean $E \beta_i = \beta$ and variance $Var \beta_i = D$. At the second stage, one draws $T_i$ observations for the $i$th observation, conditional on having observed $\beta_i$. The mean and variance of the observations are $E(y_i | \beta_i) = X_i \beta_i$ and $Var(y_i | \beta_i) = R_i$. Putting these two stages together yields

$$ E y_i = X_i E \beta_i = X_i \beta $$

and

$$ Var y_i = E (Var(y_i | \beta_i)) + Var(E(y_i | \beta_i)) = R_i + Var(X_i \beta_i) = R_i + X_i' D X_i = V_i. $$

Example – Taxpayer study – Continued
The random coefficients model was fit using the Taxpayer data with $K = 8$ variables. The model fitting was done using the statistical package SAS, with the MIVQUE(0) variance components estimation techniques, described in Section 3.5. The resulting fitting $D$ matrix appears in Table 3.6. Display 3.2 provides additional details of the model fit.
Table 3.6 Values of the Estimated D Matrix

<table>
<thead>
<tr>
<th></th>
<th>INTERCEPT</th>
<th>MS</th>
<th>HH</th>
<th>AGE</th>
<th>EMP</th>
<th>PREP</th>
<th>LNTPI</th>
<th>MR</th>
<th>DEPEND</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>47.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td>-0.40</td>
<td>20.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH</td>
<td>-1.26</td>
<td>1.25</td>
<td>23.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>18.48</td>
<td>2.61</td>
<td>-0.79</td>
<td>22.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMP</td>
<td>-8.53</td>
<td>0.92</td>
<td>0.12</td>
<td>0.21</td>
<td>20.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PREP</td>
<td>4.22</td>
<td>-0.50</td>
<td>-1.85</td>
<td>0.14</td>
<td>-0.21</td>
<td>-0.50</td>
<td>21.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LNTPI</td>
<td>-4.54</td>
<td>-0.17</td>
<td>0.15</td>
<td>-2.38</td>
<td>0.11</td>
<td>-0.38</td>
<td>21.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR</td>
<td>0.48</td>
<td>0.06</td>
<td>-0.03</td>
<td>0.14</td>
<td>-0.09</td>
<td>0.04</td>
<td>-0.09</td>
<td>20.68</td>
<td></td>
</tr>
<tr>
<td>DEPEND</td>
<td>3.07</td>
<td>0.41</td>
<td>0.29</td>
<td>-0.60</td>
<td>-0.40</td>
<td>-0.35</td>
<td>-0.34</td>
<td>0.01</td>
<td>20.68</td>
</tr>
</tbody>
</table>

Variations of the random coefficients model

Certain variations of the two-stage interpretation of the random coefficients models lead to other forms of the random effects model in equation (3.6). To illustrate, in equation (3.6), we may take the columns of $Z_i$ to be a strict subset of the columns of $X_i$. This is equivalent to assuming that certain components of $\beta_i$ associated with $Z_i$ are stochastic whereas other components that are associated with $X_i$ (but not $Z_i$) are nonstochastic.

Note that the convention in equation (3.6) is to assume that the mean of the random effects $\alpha_i$ are known and equal to zero. Alternatively, we could assume that they are unknown with mean, say, $\alpha$, that is, $E \alpha_i = \alpha$. However, this is equivalent to specifying additional fixed effects terms $Z, \alpha$ in equation (3.6). By convention, we absorb these additional terms into the “$X_i \beta$” potion of the model. Thus, it is customary to include those explanatory variables in the $Z_i$ design matrix as part of the $X_i$ design matrix.

Display 3.2 Selected SAS Output for the Random Coefficients Model

Fit Statistics

<table>
<thead>
<tr>
<th>Fit Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Res Log Likelihood</td>
<td>7876.0</td>
</tr>
<tr>
<td>AIC (smaller is better)</td>
<td>7968.0</td>
</tr>
<tr>
<td>AICC (smaller is better)</td>
<td>7971.5</td>
</tr>
<tr>
<td>BIC (smaller is better)</td>
<td>8131.4</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect   | Estimate | Error | DF | t Value | Pr > |t| |
|----------|----------|-------|----|---------|-------|---|
| Intercept| -9.5456  | 2.1475 | 253| -4.44   | <.0001|
| MS       | -0.3183  | 1.0664 | 41 | -0.30   | 0.7668|
| HH       | -1.0514  | 1.4418 | 16 | -0.73   | 0.4764|
| AGE      | -0.4027  | 1.1533 | 20 | -0.35   | 0.7306|
| EMP      | -0.1498  | 0.9019 | 31 | -0.17   | 0.8691|
| PREP     | -0.2156  | 0.6118 | 67 | -0.35   | 0.7257|
| LNTPI    | 1.6118   | 0.3712 | 257| 4.34    | <.0001|
| DEPEND   | -0.2814  | 0.4822 | 70 | -0.58   | 0.5613|
| MR       | 0.09303  | 0.2853 | 250| 0.33    | 0.7446|
Another variation of the two-stage interpretation uses known variables $B_i$ such that $E\beta_i = B_i\beta$. Then, we have, $E y_i = X_iB_i\beta$ and $\text{Var } y_i = R_i + X_i' \Sigma X_i$. This is equivalent to our equation (3.6) model replacing $X_i$ with $X_iB_i$ and $Z_i$ with $X_i$. Hsiao (1986E, Section 6.5) refers to this as a variable-coefficients model with coefficients that are functions of other exogenous variables. Chapter 5 describes this approach in greater detail.

**Example – Lottery sales**

Section 4.5 will describe a case in which we wish to predict lottery sales. The response variable $y_{it}$ is logarithmic lottery sales in week $t$ for a geographic unit $i$. No time-varying variables are available for these data, so a basic explanation of lottery sales is through the one-way random effects ANOVA model of the form $y_{it} = \alpha_i + \epsilon_{it}$. We can interpret to be the (conditional) mean lottery sales for the $i$th geographic unit. In addition, we will have available several time-constant variables that describe the geographic unit, including population, median household income, median home value and so forth. Denote this set of variables that describe the $i$th geographic unit as $B_i$. With the two-stage interpretation, we could use these variables to explain the mean lottery sales with the representation $\alpha_i = B_i' \beta + \alpha_i$. Note that the variable $\alpha_i$ is unobservable, so this model is not estimable by itself. However, when combined with the ANOVA model, we have

$$y_{it} = \alpha_i + B_i' \beta + \epsilon_{it},$$

our error components model. This combined model is estimable.

**Group effects**

In many applications of longitudinal data analysis, it is of interest to assess differences of responses from different groups. In this context, the term “group” refers to a category of the population. For example, in the Section 3.2 Taxpayer example, we may be interested in studying the differences in tax liability due to gender (male/female) or due to political party affiliation (democrat/republican/libertarian/ and so on).

A typical model that includes group effects can be expressed as a special case of the linear mixed effects model, using $q = 1$, $z_i = 1$ and the expression

$$E(y_{git} | \alpha_{gi}) = \alpha_{gi} + \delta_g + x_{git}' \beta.$$

Here, the subscripts range over $g = 1, ..., G$ groups, $i = 1, ..., n_g$ subjects in each group and $t = 1, ..., T_{gi}$ observations of each subject. The terms $\{\alpha_{gi}\}$ represent random, subject-specific effects and $\{\delta_g\}$ represent fixed differences among groups. An interesting aspect of random effects portion is that subjects need not change groups over time for the model to be estimable. To illustrate, if we were interested in gender differences in tax liability, we would not expect individuals to change gender over such a small sample. This is in contrast to the fixed effects model, where group effects are not estimable due to their collinearity with subject-specific effects.

**Time-constant variables**

The study of time-constant variables provides a powerful motivation for designing a panel, or longitudinal, study that can be analyzed as a linear mixed effects model. Within a linear mixed effects model, both the heterogeneity terms $\{\alpha_i\}$ and parameters associated with time-constant variables can be analyzed simultaneously. This was not the case for the fixed effects models, where the heterogeneity terms and time-constant variables are perfectly collinear. The group effect discussed above is a specific type of time-constant variable. Of course, it is also possible to analyze group effects where individuals switch groups over time, such as with
political party affiliation. This type of problem can be handled directly using binary variables to indicate the presence of absence of a group type, and represents no particular difficulties.

We may split the explanatory variables associated with the global parameters into those that vary by time and those that do not (time-constant). Thus, we can write our linear mixed effects conditional regression function as

$$E(y_{it} | a_i) = a_i' z_{it} + x_{1it}' \beta_1 + x_{2it}' \beta_2.$$ 

This model is a generalization of the group effects model.

### 3.3.2 Mixed linear models

In the Section 3.3.1 linear mixed effects models, we assumed independence among subjects (Assumption RO4). This assumption is not tenable for all models of repeated observations on a subject over time, so it is of interest to introduce a generalization known as the mixed linear model. This model equation is given by

$$y = Z \alpha + X \beta + \varepsilon. \quad (3.9)$$

Here, $y$ is a $N \times 1$ vector of responses, $\varepsilon$ is a $N \times 1$ vector of errors, $Z$ and $X$ are $N \times q$ and $N \times K$ matrices of explanatory variables, respectively, and $\alpha$ and $\beta$ are $q \times 1$ and $K \times 1$ vectors of parameters.

For the mean structure, we assume $E(y | \alpha) = Z \alpha + X \beta$ and $E \alpha = 0$, so that $E y = X \beta$. For the covariance structure, we assume $\text{Var}(y | \alpha) = R$, $\text{Var} \alpha = D$ and $\text{Cov}(\alpha, \varepsilon') = 0$. This yields $\text{Var} y = Z D Z' + R = V$.

Unlike the linear mixed effects model in Section 3.3.1, the mixed linear model does not require independence between subjects. Further, the model is sufficiently flexible so that several complex hierarchical structures can be expressed as special cases of it. To see how the linear mixed effects model is a special case of the mixed linear model, take

$$\varepsilon = (\varepsilon_1', \varepsilon_2', ..., \varepsilon_n'), \quad \alpha = (\alpha_1', \alpha_2', ..., \alpha_n'),$$

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{pmatrix} \quad \text{and} \quad Z = \begin{pmatrix} Z_1 & 0 & 0 & \cdots & 0 \\ 0 & Z_2 & 0 & \cdots & 0 \\ 0 & 0 & Z_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & Z_n \end{pmatrix}.$$ 

With these choices, the mixed linear model reduces to the linear mixed effects model.

The two-way error components model is an important panel data model that is not a specific type of linear mixed effects model, although it is a special case of the mixed linear model. This model can be expressed as

$$y_{it} = \alpha_i + \lambda_t + x_{it}' \beta + \varepsilon_{it}. \quad (3.10)$$

This is similar to the error components model but we have added a random time component, $\lambda_t$. We assume that $\{\lambda_t\}, \{\alpha_i\}$ and $\{\varepsilon_{it}\}$ are mutually independent. See Chapter 8 for additional details regarding this model.

To summarize, the mixed linear model generalizes the linear mixed effects model and includes other models that are of interest in longitudinal data analysis. Much of the estimation can be accomplished directly in terms of the mixed linear model. To illustrate, in this book many of the examples are analyzed using PROC MIXED, a procedure within the statistical package SAS specifically designed to analyze mixed linear models. The primary advantage of the linear mixed effects model is that it provides a more intuitive platform for examining longitudinal data.
Example – Income inequality


Specifically, the model was:

\[
\ln y_{c,i,t} = (1-z_{1,t}) \alpha_{c,1} + z_{1,t} \alpha_{c,2} + \lambda_t + x_{c,i,t}' \beta_1 + z_{1,t} x_{c,i,t}' \beta_2 + e_{c,i,t}.
\]

Here, \( y_{c,i,t} \) represents income for the \( i \)th subject in the \( c \)th city at time \( t \). The vector \( x_{c,i,t} \) represents several control variables that include gender, age, age squared, education, occupation and work organization (government, other public, and private firms). The variable \( z_{1,t} \) is a binary variable defined to be one if \( t \geq 1985 \) and zero otherwise. Thus, the vector \( \beta \) represents parameter estimates for the explanatory variables before 1985 and \( \beta_2 \) represents the differences after urban reform. The primary interest is in the change of the explanatory variable effects, \( \beta_2 \).

For the other variables, the random effect \( \lambda_t \) is meant to control for undetected time effects. There are two city effects: \( (1-z_{1,t}) \alpha_{c,1} \) is for cities before 1985 and \( z_{1,t} \alpha_{c,2} \) is for after 1984. Note that these random effects are at the city level and not at the subject \((i)\) level. Zhou used a combination of error components and autoregressive structure to model the serial relationships of the disturbance terms. Including these random effects accounted for clustering of responses within cities and within time periods, thus providing more accurate assessment of the regression coefficients \( \beta \) and \( \beta_2 \).

Zhou found significant returns to education and these returns increased in the post-reform era. Little change was found among organization effects, with the exception of significantly increased effects for private firms.

3.4 Regression coefficient inference

Estimation of the linear mixed effects model proceeds in two stages. In the first stage, we estimate the regression coefficients \( \beta \), assuming knowledge of the variance components \( \tau \). Then, in the second stage, the variance components \( \tau \) are estimated. Section 3.5 discusses variance component estimation whereas this section discusses regression coefficient inference, assuming that the variance components are known.

GLS estimation

From Section 3.3, we have that the vector \( y_i \) has mean \( X_i \beta \) and variance \( Z_i D Z_i' + R_i = V_i(\tau) = V_i \). Thus, direct calculations show that the generalized least squares (GLS) estimator of \( \beta \) is

\[
b_{GLS} = \left( \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i.
\]  

(3.11)

The GLS estimator of \( \beta \) takes the same form as in the error components model estimator in equation (3.4) yet with a more general variance covariance matrix \( V_i \). Furthermore, direct calculation show that the variance is

\[
\text{Var} b_{GLS} = \left( \sum_{i=1}^{n} X_i' V_i^{-1} X_i \right)^{-1}.
\]  

(3.12)
As with fixed effects estimators, it is possible to express \( \mathbf{b}_{GLS} \) as a weighted average of subject-specific estimators. To this end, for the \( i \)th subject, define the GLS estimator \( \mathbf{b}_{i, GLS} = (\mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{y}_i \), and the weight \( \mathbf{W}_{i, GLS} = \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \). Then, we can write

\[
\mathbf{b}_{GLS} = \left( \sum_{i=1}^{n} \mathbf{W}_{i, GLS} \right)^{-1} \sum_{i=1}^{n} \mathbf{W}_{i, GLS} \mathbf{b}_{i, GLS}.
\]

**Matrix inversion formula**

To simplify the calculations and to provide better intuition for our expressions, we cite a formula for inverting \( \mathbf{V}_i \). Note that the matrix \( \mathbf{V}_i \) has dimension \( T_i \times T_i \). From Appendix A.5, we have

\[
\mathbf{V}_i^{-1} = (\mathbf{R}_i + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i')^{-1} = \mathbf{R}_i^{-1} - \mathbf{R}_i^{-1} \mathbf{Z}_i \left( \mathbf{D}^{-1} + \mathbf{Z}_i' \mathbf{R}_i^{-1} \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i' \mathbf{R}_i^{-1}.
\]  

(3.13)

The expression on the right-hand side of equation (3.13) is easier to compute than the left-hand side when the temporal covariance matrix \( \mathbf{R}_i \) has an easily computable inverse and the dimension \( q \) is smaller than \( T_i \). Moreover, because the matrix \( \mathbf{D}^{-1} + \mathbf{Z}_i' \mathbf{R}_i^{-1} \mathbf{Z}_i \) is only a \( q \times q \) matrix, it is easier to invert than \( \mathbf{V}_i \), a \( T_i \times T_i \) matrix.

Some special cases are of interest. First, note that in the case of no serial correlation, we have \( \mathbf{R}_i = \sigma^2 \mathbf{I}_i \) and equation (3.13) reduces to

\[
\mathbf{V}_i^{-1} = \left( \sigma^2 \mathbf{I}_i + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' \right)^{-1} = \frac{1}{\sigma^2} \left( \mathbf{I}_i - \mathbf{Z}_i \left( \sigma^2 \mathbf{D}^{-1} + \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i' \right).
\]  

(3.14)

Further, in the error components model considered in Section 3.1, we have \( q = 1, \mathbf{D} = \sigma^2 \mathbf{I}_i, \mathbf{Z}_i = \mathbf{1}_i \), so that equation (3.13) reduces to

\[
\mathbf{V}_i^{-1} = \left( \sigma^2 \mathbf{I}_i + \sigma^2 \mathbf{Z}_i \mathbf{Z}_i' \right)^{-1} = \frac{1}{\sigma^2} \left( \mathbf{I}_i - \frac{\sigma^2}{T_i \sigma^2 + \sigma^2} \mathbf{Z}_i \mathbf{Z}_i' \mathbf{Z}_i' \right),
\]  

(3.15)

where \( \zeta_i = \frac{T_i \sigma^2}{T_i \sigma^2 + \sigma^2} \), as in Section 3.1. This demonstrates that equation (3.4) is a special case of equation (3.11).

For another special case, consider the random coefficients model \( (z_{it} = x_{it}) \) with no serial correlation so that \( \mathbf{R}_i = \sigma^2 \mathbf{I}_i \). Here, the weight \( \mathbf{W}_{i, GLS} \) takes on a simple form: \( \mathbf{W}_{i, GLS} = \left( \mathbf{D} + \sigma^2 \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \) (see Exercise 3.8). From this form, we see that subjects with “large” values of \( \mathbf{X}_i' \mathbf{X}_i \) have a greater effect on \( \mathbf{b}_{GLS} \) than subjects with smaller values.

**Maximum likelihood estimation**

With assumption RO5, the log-likelihood of a single subject is

\[
I_i(\beta, \tau) = -\frac{1}{2} \left( T_i \ln(2\pi) + \ln \det \mathbf{V}_i(\tau) + (\mathbf{y}_i - \mathbf{X}_i' \beta)' \mathbf{V}_i(\tau)^{-1} (\mathbf{y}_i - \mathbf{X}_i' \beta) \right).
\]  

(3.16)

Appendix B provides additional background on joint normality and the related likelihood function. Appendix C reviews likelihood estimation in a general context. With equation (3.16), the log-likelihood for the entire data set is

\[
L(\beta, \tau) = \sum_{i=1}^{n} I_i(\beta, \tau).
\]

The values of \( \beta \) and \( \tau \) that maximize \( L(\beta, \tau) \) are the maximum likelihood estimators (MLEs) which we denote as \( \mathbf{b}_{MLE} \) and \( \tau_{MLE} \).
The score vector is the vector of derivatives of the log-likelihood taken with respect to the parameters. We denote the vector of parameters by \( \theta = (\beta', \tau')' \). With this notation, the score vector is \( \partial L(\theta)/\partial \theta \). Typically, if this score has a root, then the root is a maximum likelihood estimator. To compute the score vector, first take derivatives with respect to \( \beta \) and find the root. That is, 

\[
\sum_{i=1}^{n} (y_i - X_i \beta)' V_i(\tau)^{-1} (y_i - X_i \beta) = 0
\]

Setting the score vector equal to zero yields \( b_{MLE} = \left( \sum_{i=1}^{n} X_i' V_i(\tau)^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i' V_i(\tau)^{-1} y_i = b_{GLS} \). (3.17)

That is, for fixed covariance parameters \( \tau \), the maximum likelihood estimator and the general least squares estimator are the same.

**Robust estimation of standard errors**

Even without the assumption of normality, the maximum likelihood estimator \( b_{MLE} \) has desirable properties. It is unbiased, efficient and asymptotically normal with covariance matrix given in equation (3.12). However, the estimator does depend on knowledge of variance components. As an alternative, it can be useful to consider an alternative, weighted least squares estimator

\[
b_{W} = \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1} \sum_{i=1}^{n} X_i' W_{i,RE} V_i X_i \]

(3.18)

where the weighting matrix \( W_{i,RE} \) depends on the application at hand. To illustrate, one could use the identity matrix so that \( b_{W} \) reduces to the ordinary least squares estimator. Another choice is \( Q \) from Section 2.5.3 that yields fixed effects estimators of \( \beta \). We explore this choice further in Section 7.2. The weighted least squares estimator is an unbiased estimator of \( \beta \) and is asymptotically normal, although not efficient unless \( W_{i,RE} = V_i^{-1} \). Basic calculations show that it has variance

\[
\text{Var } b_{W} = \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1} \sum_{i=1}^{n} X_i' W_{i,RE} V_i \sum_{i=1}^{n} X_i' W_{i,RE} X_i \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1}.
\]

As in Section 2.5.3, we may consider estimators that are robust to unsuspected serial correlation and heteroscedasticity. Specifically, following a suggestion made independently by Huber (1967G), White (1980E) and Liang and Zeger (1986B), we can replace \( V_i \) by \( e_i e_i' \), where \( e_i = y_i - X_i b_{W} \) is the vector of residuals. Thus, a robust standard error of \( b_{W,j} \), the \( j \)th element of \( b_{W} \), is

\[
se(b_{W,j}) = \sqrt{\text{diagonal element of } \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1} \sum_{i=1}^{n} X_i' W_{i,RE} e_i e_i' W_{i,RE} X_i \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1}}.
\]

**Testing hypotheses**

For many statistical analyses, testing the null hypothesis that a regression coefficient equals a specified value may be the main goal. That is, the interest may be in testing \( H_0: \beta_j = \beta_{j,0} \).
where the specified value $\beta_{j,0}$ is often (although not always) equal to 0. The customary procedure is to compute the relevant
\[
t - \text{statistic} = \frac{b_{j,\text{GLS}} - \beta_{j,0}}{se(b_{j,\text{GLS}})}.
\]

Here, $b_{j,\text{GLS}}$ is the $j$th component of $b_{\text{GLS}}$ from equation (3.17) and $se(b_{j,\text{GLS}})$ is the square root of the $j$th diagonal element of $\left(\sum_i X_i'V_i(\hat{\tau})^{-1}X_i\right)^{-1}$ where $\hat{\tau}$ is the estimator of the variance component that will be described in Section 3.5. Then, one assesses $H_0$ by comparing the $t$-statistic to a standard normal distribution.

There are two widely used variants of this standard procedure. First, one can replace $se(b_{j,\text{GLS}})$ by $se(b_{j,w})$ to get so-called “robust $t$-statistics.” Second, one can replace the standard normal distribution with a $t$-distribution with the “appropriate” number of degrees of freedom. There are several methods for calculating the degrees of freedom that depend on the data and the purpose of the analysis. To illustrate, in Display 3.2 you will see that the approximate degrees of freedom under the “DF” column is different for each variable. This is produced by the SAS default “containment method.” For the applications in this text, we typically will have large number of observations and will be more concerned with potential heteroscedasticity and serial correlation and thus will use robust $t$-statistics. For readers with smaller data sets interested in the second alternative, Littell et al. (1996S) describes the $t$-distribution approximation in detail.

For testing hypotheses concerning several regression coefficients simultaneously, the customary procedure is the likelihood ratio test. One may express the null hypothesis as $H_0: C \beta = d$, where $C$ is a $p \times K$ matrix with rank $p$, $d$ is a $p \times 1$ vector (typically $0$) and recall that $\beta$ is the $K \times 1$ vector of regression coefficients. Both $C$ and $d$ are user specified and depend on the application at hand. This null hypothesis is tested against the alternative $H_0: C \beta \neq d$.

**Likelihood ratio test procedure**

1. Using the unconstrained model, calculate maximum likelihood estimates and the corresponding likelihood, denoted as $L_{\text{MLE}}$.
2. For the model constrained using $H_0: C \beta = d$, calculate maximum likelihood estimates and the corresponding likelihood, denoted as $L_{\text{Reduced}}$.
3. Compute the likelihood ratio test statistic, $LRT = 2 (L_{\text{MLE}} - L_{\text{Reduced}})$.
4. Reject $H_0$ if $LRT$ exceeds a percentile from a $\chi^2$ (chi-square) distribution with $p$ degrees of freedom. The percentile is one minus the significance level of the test.

Of course, one may also use $p$-values to calibrate the significance of the test. See Appendix C.7 for more details on the likelihood ratio test.

The likelihood ratio test is the industry standard for assessing hypotheses concerning several regression coefficients. However, we note that better procedures may exist, particularly for small data sets. To illustrate, Pinheiro and Bates (2000S) recommend the use of “conditional $F$-tests” when $p$ is large relative to the sample size. As with testing individual regression coefficients, we shall be more concerned with potential heteroscedasticity for large data sets. In this case, a modification of the Wald test procedure is available.

For the case of no heteroscedasticity and/or serial correlation, the Wald procedure for testing $H_0: C \beta = d$ is to compute the test statistic
and compare this statistic to a chi-square distribution with $p$ degrees of freedom. Compared to the likelihood ratio test, the advantage of the Wald procedure is that the statistic can be computed with just one evaluation of the likelihood, not two. However, the disadvantage is that for general constraints such as $C \beta = d$, specialized software is required.

An advantage of the Wald procedure is that it is straightforward to compute robust alternatives. For a robust alternative, we use the regression coefficient estimator defined in equation (3.18) and compute

$$
(C_{W} - d)' C \left[ \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right]^{-1} \left( \sum_{i=1}^{n} X_i' W_{i,RE} e_i W_{i,RE} X_i \left( \sum_{i=1}^{n} X_i' W_{i,RE} X_i \right)^{-1} C' \right) (C_{W} - d).
$$

We compare this statistic to a chi-square distribution with $p$ degrees of freedom.

### 3.5 Variance components estimation

In this section, we describe several methods for estimating the variance components. The two primary methods entail maximizing a likelihood function, in contrast to moment estimators. In statistical estimation theory (Lehmann, 1991G), there are well-known trade-offs when considering moment compared to likelihood estimation. Typically, likelihood functions are maximized by using iterative procedures that require starting values. At the end of this section, we describe how to obtain reasonable starting values for the iteration using moment estimators.

#### 3.5.1 Maximum likelihood estimation

The log-likelihood was presented in Section 3.4. Substituting the expression for the generalized least squares estimator in equation (3.11) into the log-likelihood in equation (3.16) yields the concentrated or profile log-likelihood

$$
L(b_{GLS}, \tau) = -\frac{1}{2} \sum_{i=1}^{n} (T_i \ln(2\pi) + \ln \det V_i(\tau) + (Error SS)_i(\tau)),
$$

a function of $\tau$. Here, the error sum of squares for the $i$th subject is

$$(Error SS)_i(\tau) = (y_i - X_i b_{GLS})' V_i^{-1}(\tau) (y_i - X_i b_{GLS}).
$$

Thus, we now maximize the log-likelihood as a function of $\tau$ only. In only a few special cases can one obtain closed form expressions for the maximizing variance components. Exercise 3.10 illustrates one such special case.

**Special case – Error components model**

For this special case, the variance components are $\tau = (\sigma^2, \sigma^2_{\alpha})'$. Using equation (A.5) in Appendix A.5, we have that $\ln \det V_i = \ln \det (\sigma^2 J_i + \sigma^2 \mathbf{I}) = T_i \ln \sigma^2 + \ln (1 + T_i \sigma^2_{\alpha} / \sigma^2)$. From this and equation (3.9), we have that the concentrated likelihood is

$$
L(b_{GLS}, \sigma^2, \sigma^2_{\alpha}) = -\frac{1}{2} \sum_{i=1}^{n} \left[ T_i \ln(2\pi) + T_i \ln \sigma^2 + \ln \left( 1 + T_i \sigma^2_{\alpha} / \sigma^2 \right) \right]
$$
where $\mathbf{b}_{\text{GLS}}$ is given in equation (3.4). This likelihood can be maximized over $(\sigma^2, \sigma^2_\alpha)$ using iterative methods.

### Iterative estimation

In general, the variance components are estimated recursively. This can be done using either the Newton-Raphson or Fisher scoring method, see for example, Harville (1977S) and Wolfinger et al. (1994S).

**Newton-Raphson.** Let $L = L(\mathbf{b}_{\text{GLS}}(\tau), \tau)$, and use the iterative method:

$$
\tau_{\text{NEW}} = \tau_{\text{OLD}} - \left( \frac{\partial^2 L}{\partial \tau \partial \tau'} \right)^{-1} \frac{\partial L}{\partial \tau} \bigg|_{\tau = \tau_{\text{OLD}}}.
$$

Here, the matrix $-\frac{\partial^2 L}{\partial \tau \partial \tau'}$ is called the *sample information matrix*.

**Fisher scoring.** Define the expected information matrix $I(\tau) = -E \left( \frac{\partial^2 L}{\partial \tau \partial \tau'} \right)$ and use

$$
\tau_{\text{NEW}} = \tau_{\text{OLD}} + I(\tau_{\text{OLD}})^{-1} \frac{\partial L}{\partial \tau} \bigg|_{\tau = \tau_{\text{OLD}}}.
$$

### 3.5.2 Restricted maximum likelihood

As the name suggests, *restricted maximum likelihood* (REML) is a likelihood-based estimation procedure. Thus, it shares many of the desirable properties of maximum likelihood estimators (MLEs). Because it is based on likelihoods, it is not specific to a particular design matrix, as are analysis of variance estimators (Harville, 1977S). Thus, it can be readily applied to a wide variety of models. Like MLEs, REML estimators are translation invariant.

Maximum likelihood often produces biased estimators of the variance components, $\tau$. In contrast, estimation based on REML results in unbiased estimators of $\tau$, at least for many balanced designs. Because maximum likelihood estimators are negatively biased, they often turn out to be negative, an intuitively undesirable situation for many users. Because of the unbiasedness of many REML estimators, there is less of a tendency to produce negative estimators (Corbeil and Searle, 1976a,S). As with MLEs, REML estimators can be defined to be parameter values for which the (restricted) likelihood achieves a maximum value over a constrained parameter space. Thus, as with maximum likelihood, it is straightforward to modify the method to produce nonnegative variance estimators.

The idea behind REML estimation is to consider the likelihood of linear combinations of the responses that do not depend on the mean parameters. To illustrate, consider the mixed linear model. We assume that the responses, denoted by the vector $\mathbf{y}$, are normally distributed, have mean $E\mathbf{y} = \mathbf{X}\beta$ and variance-covariance matrix $\text{Var}\mathbf{y} = \mathbf{V} = \mathbf{V}(\tau)$. The dimension of $\mathbf{y}$ is $N \times 1$, and the dimension of $\mathbf{X}$ is $N \times p$. With this notation, define the projection matrix $\mathbf{Q} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and consider the linear combination of responses $\mathbf{Qy}$. Straightforward calculations show that $\mathbf{Qy}$ has mean $\mathbf{0}$ and variance-covariance matrix $\text{Var}(\mathbf{Qy}) = \mathbf{QVQ}$. Because (i) $\mathbf{Qy}$ has a multivariate normal distribution and (ii) the mean and variance-covariance matrix do not depend
on $\beta$, the distribution of $Qy$ does not depend on $\beta$. Further, Appendix 3A.1 shows that $Qy$ is independent of the generalized least squares estimator $b_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y$.

The vector $Qy$ is the residual vector from an ordinary least squares fit of the data. Hence, REML is also referred to as residual maximum likelihood estimation. Because the rank of $Q$ is $N - p$, we lose some information by considering this transformation of the data; hence the term restricted maximum likelihood. (There is some information about $\tau$ in the vector $b_{GLS}$ that we are not using for estimation.) Further, note that we could also use any linear transform of $Q$, such as $AQ$, in that $AQy$ has a multivariate normal distribution with a mean and variance-covariance matrix that do not depend on $\beta$. Patterson and Thompson (1971S) and Harville (1974S, 1977S) showed that the likelihood does not depend on the choice of $A$. They introduced the “restricted” log-likelihood:

$$L_{REML}(b_{GLS}(\tau), \tau) = -\frac{1}{2} \left[ \ln \det(V(\tau)) + \ln \det(X'V(\tau)^{-1}X) + (Error SS)(\tau) \right], \quad (3.21)$$

up to an additive constant. See Appendix 3A.2 for a derivation of this likelihood. REML estimators $\tau_{REML}$ are defined to be maximizers of the function $L_{REML}(b_{GLS}(\tau), \tau)$. Here, the error sum of squares is

$$(Error SS)(\tau) = (y - Xb_{GLS}(\tau))'V(\tau)^{-1}(y - Xb_{GLS}(\tau)). \quad (3.22)$$

Analogous to equation (3.19), the usual log-likelihood is

$$L(b_{GLS}(\tau), \tau) = -\frac{1}{2} \left[ \ln \det(V(\tau)) + (Error SS)(\tau) \right],$$

up to an additive constant. The only difference between the two likelihoods is the term $\ln \det(X'V(\tau)^{-1}X)$. Thus, iterative methods of maximization are the same—that is, using either Newton-Raphson or Fisher scoring. For linear mixed effects models, this additional term is $\ln \det(\sum_{i=1}^{n} X_i'V_i(\tau)^{-1}X_i)$.

For balanced analysis of variance data ($T_i = T$), Corbeil and Searle (1976a,S) established that the REML estimation reduces to standard analysis of variance estimators. Thus, REML estimators are unbiased for these designs. However, REML estimators and analysis of variance estimators differ for unbalanced data. REML estimators achieve their unbiasedness by accounting for the degrees of freedom lost in estimating the fixed effects $\beta$; MLEs do not account for this loss of degrees of freedom. When $p$ is large, the difference between REML estimators and MLEs is significant. Corbeil and Searle (1976b,S) showed that, in terms of mean square errors, MLEs outperform REML estimators for small $p$ (< 5), although the situation is reversed for large $p$ with a sufficiently large sample.

Harville (1974S) gave a Bayesian interpretation of REML estimators. He pointed out that using only $Qy$ to make inferences about $\tau$ is equivalent to ignoring prior information about $\beta$ and using all the data.

Some statistical packages present maximized values of restricted likelihoods, suggesting to users that these values can be used for inferential techniques, such as likelihood ratio tests. For likelihood ratio tests, one should use “ordinary” likelihoods, even when evaluated at REML estimators, not the “restricted” likelihoods that are used to determine REML estimators. Appendix 3A.3 illustrates the potentially disastrous consequences of using REML likelihoods for likelihood ratio tests.

**Starting values**

Both the Newton-Raphson and Fisher scoring algorithms and the ML and REML estimation methods involve recursive calculations that require starting values. We now describe two non-recursive methods due to Swamy (1970E) and Rao (1970S), respectively. One can use
the results of these non-recursive methods as starting values in the Newton-Raphson and Fisher
scoring algorithms.
Swamy’s moment-based procedure appeared in the econometrics panel data literature. We consider a random coefficients model; that is, equation (3.8) with \( x_i = z_i \) and \( R_i = \sigma^2 I_i \).

**Procedure for computing moment-based variance component estimators**

1. Compute an ordinary least squares estimator of \( \sigma_i^2 \),
   \[
   s_i^2 = \frac{1}{T_i - K} y_i' \left( I_i - X_i (X_i' X_i)^{-1} X_i' \right) y_i.
   \]
   This is an ordinary least squares procedure in that it ignores \( D \).
2. Next, calculate \( b_{i,OLS} = (X_i' X_i)^{-1} X_i' y_i \), a predictor of \( \beta + \alpha_i \).
3. Finally, estimate \( D \) using
   \[
   D_{SWAMY} = \frac{1}{n-1} \sum_{i=1}^n (b_{i,OLS} - \bar{b})(b_{i,OLS} - \bar{b})' - \frac{1}{n} \sum_{i=1}^n s_i^2 (X_i' X_i)^{-1},
   \]
   where \( \bar{b} = \frac{1}{n} \sum_{i=1}^n b_{i,OLS} \).

The estimator of \( D \) can be motivated by examining the variance of \( b_{i,OLS} \),
\[
\text{Var}(b_{i,OLS}) = \text{Var}\left((X_i' X_i)^{-1} X_i' (\beta + \alpha_i + \varepsilon_i)\right)
= \text{Var}\left(\beta + \alpha_i + (X_i' X_i)^{-1} X_i' \varepsilon_i\right) = D + \sigma_i^2 (X_i' X_i)^{-1}
\]
Using \( \frac{1}{n-1} \sum_{i=1}^n (b_{i,OLS} - \bar{b})(b_{i,OLS} - \bar{b})' \) and \( s_i^2 \) as estimators of \( \text{Var}(b_{i,OLS}) \) and \( \sigma_i^2 \) respectively, yields \( D_{SWAMY} \) as an estimator of \( D \).

Various modifications of this estimator are possible. One can iterate the procedure by using \( D_{SWAMY} \) to improve the estimators \( s_i^2 \), and so on. Homoscedasticity of the \( \varepsilon \) terms could also be assumed. Hsiao (1986E) recommends dropping the second term, \( n^{-1} \sum_{i=1}^n s_i^2 (X_i' X_i)^{-1} \), to ensure that \( D_{SWAMY} \) is non-negative definite.

### 3.5.3 MIVQUE estimators

Another non-recursive method is Rao’s (1970S) minimum variance quadratic unbiased estimator (MIVQUE). To describe this method, we return to the mixed linear model \( y = X \beta + \varepsilon \), in which \( \text{Var} \, y = V = V(\tau) \). We wish to estimate the linear combination of variance components, \( \sum_{k=1}^r c_k \tau_k \), where the \( c_k \) are specified constants and \( \tau = (\tau_1, \ldots, \tau_r)' \). We assume that \( V \) is linear in the sense that
\[
V = \sum_{k=1}^r \tau_k \frac{\partial}{\partial \tau_k} V.
\]
Thus, with this assumption, we have that the matrix of second derivatives (the Hessian) of \( V \) is zero (Graybill, 1969G). Although this assumption is generally viable, it is not satisfied by, for example, autoregressive models. It is not restrictive to assume that \( \frac{\partial}{\partial \tau_k} V \) is known even though
the variance component $\tau_k$ is unknown. To illustrate, consider an error components structure so that $V = \sigma_\alpha^2 J + \sigma^2 I$. Then, $\frac{\partial}{\partial \sigma_\alpha^2} V = J$ and $\frac{\partial}{\partial \sigma^2} V = I$ are both known.

Quadratic estimators of $\sum_{k=1}^r c_k \tau_k$ are based on $y' A y$, where $A$ is a symmetric matrix to be specified. The variance of $y' A y$, assuming normality, can easily be shown to be $2 \text{trace}(VAVA)$. We would like the estimator to be invariant to translation of $\beta$. That is, we require

$$y' A y = (y - X b_0)' A (y - X b_0)$$

for each $b_0$.

Thus, we restrict our choice of $A$ to those that satisfy $A X = 0$.

For unbiasedness, we would like $\sum_{k=1}^r c_k \tau_k = E(y'y)$. Using $A X = 0$, we have

$$E(y'Ay) = E(\epsilon'A\epsilon) = \text{trace}(E(\epsilon\epsilon'A)) = \text{trace}(VA)$$

Because this equality should be valid for all variance components $\tau_k$, we require that $A$ satisfy

$$c_k = \text{trace}\left(\frac{\partial}{\partial \tau_k} V \right) A,$$

for $k = 1, \ldots, r$. (3.23)

Rao showed that the minimum value of $\text{trace}(VAVA)$ satisfying $A X = 0$ and the constraints in equation (3.23) is attained at

$$A_*(V) = \sum_{k=1}^r \lambda_k V^{-1} Q \left( \frac{\partial}{\partial \tau_k} V \right) V^{-1} Q,$$

where $Q = Q(V) = I - X (X' V^{-1} X) X' V^{-1}$ and $(\lambda_1, \ldots, \lambda_r)$ is the solution of

$$S(\lambda_1, \ldots, \lambda_r)' = (c_1, \ldots, c_r)' .$$

Here, the $(i, j)$th element of $S$ is given by

$$\text{trace}\left( V^{-1} Q \left( \frac{\partial}{\partial \tau_i} V \right) V^{-1} Q \left( \frac{\partial}{\partial \tau_j} V \right) \right).$$

Thus, the MIVQUE estimator of $\tau$ is the solution of

$$S \tau_{\text{MIVQUE}} = G,$$ (3.24)

where the $k$th element of $G$ is given by

$$y' V^{-1} Q \left( \frac{\partial}{\partial \tau_k} V \right) V^{-1} Q y.$$
Further reading

When compared to regression and linear models, there are fewer textbook introductions to mixed linear models, although more are becoming available. Searle, Casella and McCulloch (1992S) give an early technical treatment. A slightly less technical is Longford (1993EP). McCulloch and Searle (2001S) give an excellent recent technical treatment. Other recent contributions integrate statistical software into their exposition. Little et al. (1996D) and Verbeke and Molenberghs (2000D) introduce mixed linear models using the SAS statistical package. Pinheiro and Bates (2000D) provide an introduction using the S and S-Plus statistical packages.

Random effects in ANOVA and regression models has been part of the standard statistical literature for quite some time; see, for example, Scheffé (1959G), Searle (1971G) or Neter and Wasserman (1974G). Balestra and Nerlove (1966E) introduced the error components model to the econometric literature. The random coefficients model was described early on by Hildreth and Houck (1968S).

As described in Section 3.5, most of the development of variance component estimators occurred in the 1970’s. More recently, Baltagi and Chang (1994E) compared the relative performance of several variance components estimators for the error components model.
Appendix 3A. REML Calculations

Appendix 3A.1 Independence of Residuals and Least Squares Estimators

Assume that \( \mathbf{y} \) has a multivariate normal distribution with mean \( \mathbf{X}\beta \) and variance-covariance matrix \( \mathbf{V} \), where \( \mathbf{X} \) has dimension \( N \times p \) with rank \( p \). Recall that the matrix \( \mathbf{V} \) depends on the parameters \( \tau \).

We use the matrix \( \mathbf{Q} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \). Because \( \mathbf{Q} \) is idempotent and has rank \( N - p \), we can find an \( N \times (N - p) \) matrix \( \mathbf{A} \) such that

\[
\mathbf{A}'\mathbf{A}' = \mathbf{Q} \quad \text{and} \quad \mathbf{A}'\mathbf{A} = \mathbf{I}_N.
\]

We also need \( \mathbf{G} = \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \), an \( N \times p \) matrix. Note that \( \mathbf{G}'\mathbf{y} = \mathbf{b}_{\text{GLS}} \), the generalized least squares estimator of \( \beta \).

With these two matrices, define the transformation matrix \( \mathbf{H} = (\mathbf{A} \mathbf{G}) \), an \( N \times N \) matrix.

Consider the transformed variables

\[
\mathbf{H}'\mathbf{y} = \begin{bmatrix} \mathbf{A}'\mathbf{y} \\ \mathbf{G}'\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{A}'\mathbf{y} \\ \mathbf{b}_{\text{GLS}} \end{bmatrix}.
\]

Basic calculations show that

\[
\mathbf{A}'\mathbf{y} \sim N(0, \mathbf{A}'\mathbf{V}\mathbf{A}) \quad \text{and} \quad \mathbf{G}'\mathbf{y} = \mathbf{b}_{\text{GLS}} \sim N(\mathbf{b}, (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}),
\]

in which \( \mathbf{z} \sim N(\mu, \mathbf{V}) \) denotes that a random vector \( \mathbf{z} \) has a multivariate normal distribution with mean \( \mu \) and variance \( \mathbf{V} \). Further, we have that \( \mathbf{A}'\mathbf{y} \) and \( \mathbf{b}_{\text{GLS}} \) are independent. This is due to normality and zero covariance matrix:

\[
\text{Cov}(\mathbf{A}'\mathbf{y}, \mathbf{b}_{\text{GLS}}) = \mathbf{E}(\mathbf{A}'\mathbf{y}'\mathbf{G}) = \mathbf{A}'\mathbf{V}\mathbf{G} = \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = 0.
\]

We have \( \mathbf{A}'\mathbf{X} = 0 \) because \( \mathbf{A}'\mathbf{X} = (\mathbf{A}'\mathbf{A})\mathbf{A}'\mathbf{X} = \mathbf{A}'\mathbf{Q}\mathbf{X} \) and \( \mathbf{Q}\mathbf{X} = 0 \). Zero covariance, together with normality, imply independence.

Appendix 3A.2 Restricted Likelihoods

To develop the restricted likelihood, we first check the rank of the transformation matrix \( \mathbf{H} \). Thus, with \( \mathbf{H} \) as in Appendix 3A.1 and equation (A.2) of Appendix A.5, we have

\[
\det(\mathbf{H}'\mathbf{H}) = \det(\mathbf{H}'\mathbf{H}) = \det\begin{bmatrix} \mathbf{A}' & \mathbf{G}' \\ \mathbf{G}' & \mathbf{G}' \end{bmatrix} = \det\begin{bmatrix} \mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{G} \\ \mathbf{G}'\mathbf{A} & \mathbf{G}'\mathbf{G} \end{bmatrix}
\]

\[
= \det(\mathbf{A}'\mathbf{A})\det(\mathbf{G}'\mathbf{G} - \mathbf{G}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{G})
\]

\[
= \det(\mathbf{G}'\mathbf{G} - \mathbf{G}'\mathbf{Q}\mathbf{G}) = \det(\mathbf{G}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}) = \det((\mathbf{X}'\mathbf{X})^{-1}),
\]

using \( \mathbf{G}'\mathbf{X} = \mathbf{I} \). Thus, the transformation \( \mathbf{H} \) is non-singular if and only if \( \mathbf{X}'\mathbf{X} \) is non-singular. In this case, no information is lost by considering the transformation \( \mathbf{H}'\mathbf{y} \).

We now develop the restricted likelihood based on the probability density function of \( \mathbf{A}'\mathbf{y} \). We first note a relationship used by Harville (1974S), concerning the probability density function of \( \mathbf{G}'\mathbf{y} \). We write \( f_{\mathbf{G}_Z}(\mathbf{z}, \beta) \) to denote the probability density function of the random vector \( \mathbf{G}'\mathbf{y} \), evaluated at the (vector) point \( \mathbf{z} \) with mean (vector) parameter \( \beta \). Because probability density functions integrate to 1, we have the relation
\[ 1 = \int f_{G'}(z, \beta)dz = \int \frac{1}{(2\pi)^{p/2} \det(X'V^{-1}X)^{-1/2}} \exp \left( -\frac{1}{2} (z - \beta)'X'V^{-1}X(z - \beta) \right) dz \]

\[ = \int \frac{1}{(2\pi)^{p/2} \det(X'V^{-1}X)^{-1/2}} \exp \left( -\frac{1}{2} (z - \beta)'X'V^{-1}X(z - \beta) \right) d\beta \]

\[ = \int f_{G'}(z, \beta)d\beta, \text{ for each } z, \]

with a change of variables.

Because of the independence of \(A'y\) and \(G'y = b_{GLS}\), we have \(f_{H'y} = f_{A'y}f_{G'y}\). Here, \(f_{H'y}, f_{A'y}\) and \(f_{G'y}\) are the density functions of the random vectors \(H'y, A'y\) and \(G'y\), respectively. For notation, let \(y^*\) be a potential realization of the random vector \(y\). Thus, the probability density function of \(A'y\) is

\[ f_{A'y}(A'y^*) = \int f_{A'y}(A'y^*)f_{G'y}(G'y^*, \beta)d\beta = \int f_{H'y}(H'y^*, \beta)d\beta = \int \det(H)^{-1} f_{y}(y^*, \beta)d\beta, \]

using a change of variables. Now, let \(b_{GLS}^*\) be the realization of \(b_{GLS}\) using \(y^*\). Then, from a standard equality from analysis of variance,

\[ (y^* - X\beta)'V^{-1}(y^* - X\beta) = (y^* - Xb_{GLS}^*)'V^{-1}(y^* - Xb_{GLS}^*) + (b_{GLS}^* - \beta)'X'V^{-1}X(b_{GLS}^* - \beta). \]

With this equality, the probability density function \(f_y\) can be expressed as

\[ f_{y}(y^*, \beta) = \frac{1}{(2\pi)^{N/2} \det(V)^{1/2}} \exp \left( -\frac{1}{2} (y^* - X\beta)'V^{-1}(y^* - X\beta) \right) \]

\[ = \frac{1}{(2\pi)^{N/2} \det(V)^{1/2}} \exp \left( -\frac{1}{2} (y^* - Xb_{GLS}^*)'V^{-1}(y^* - Xb_{GLS}^*) \right) \exp \left( -\frac{1}{2} (b_{GLS}^* - \beta)'X'V^{-1}X(b_{GLS}^* - \beta) \right). \]

\[ = \frac{(2\pi)^{p/2} \det(X'V^{-1}X)^{-1/2}}{(2\pi)^{N/2} \det(V)^{1/2}} \exp \left( -\frac{1}{2} (y^* - Xb_{GLS}^*)'V^{-1}(y^* - Xb_{GLS}^*) \right) f_{G'}(b_{GLS}^*, \beta). \]

Thus,

\[ f_{A'y}(A'y^*) = \frac{(2\pi)^{p/2} \det(X'V^{-1}X)^{-1/2}}{(2\pi)^{N/2} \det(V)^{1/2}} \det(H)^{-1} \exp \left( -\frac{1}{2} (y^* - Xb_{GLS}^*)'V^{-1}(y^* - Xb_{GLS}^*) \right) \int f_{G'}(b_{GLS}^*, \beta)d\beta. \]

\[ = (2\pi)^{-(N-p)/2} \det(V)^{-1/2} \det(X'X)^{1/2} \det(X'V^{-1}X)^{-1/2} \exp \left( -\frac{1}{2} (y^* - Xb_{GLS}^*)'V^{-1}(y^* - Xb_{GLS}^*) \right) \]

This yields the REML likelihood in Section 3.5, after taking logarithms and dropping constants that do not involve \(\tau\).
Appendix 3A.3 Likelihood Ratio Tests and REML

Recall the likelihood ratio statistic, \( LRT = 2 \left( L(\theta_{MLE}) - L(\theta_{Reduced}) \right) \). This is evaluated using the so-called “concentrated” or “profile” log-likelihood given in equations (3.19) and (3.20). For comparison, from equation (3.21), the “restricted” log-likelihood is

\[
L_{REML}(b_{GLS}, \tau) = -\frac{1}{2} \sum_{i=1}^{n} \left( T_i \ln(2\pi) + \ln \det V_i(\tau) + (Error SS_i)(\tau) \right) - \frac{1}{2} \ln \det \left( \sum_{i=1}^{n} X_i'V_i(\tau)^{-1}X_i \right).
\] (3A.1)

To see why a REML likelihood does not work for likelihood ratio tests, consider the following example.

**Special Case. Testing the Importance of a Subset of Regression Coefficients.**

For simplicity, we assume that \( V_i = \sigma^2 I_i \), so that there is no serial correlation. For this special case, we have the finite, and asymptotic, distribution of the partial F-test (Chow). Because the asymptotic distribution is well known, we can easily judge whether or not REML likelihoods are appropriate.

Write \( \beta = (\beta_1', \beta_2')' \) and suppose that we wish to use the null hypothesis \( H_0: \beta_2 = 0 \). Assuming no serial correlation, the generalized least square estimator of \( \beta \) reduces to the ordinary least squares estimator, that is, \( b_{GLS} = b_{OLS} = \left( \sum_{i=1}^{n} X_i'X_i \right)^{-1} \sum_{i=1}^{n} X_i'y_i \). Thus, from equation (3.19), the concentrated likelihood is:

\[
L(b_{OLS}, \sigma^2) = -\frac{1}{2} \sum_{i=1}^{n} \left( T_i \ln(2\pi) + T_i \ln \sigma^2 + \frac{1}{\sigma^2} (y_i - X_i'b_{OLS})' (y_i - X_i'b_{OLS}) \right)
= \frac{1}{2} \left( N \ln(2\pi) + N \ln \sigma^2 + \frac{1}{\sigma^2} (Error SS)_{Full} \right),
\]

where \( (Error SS)_{Full} = \sum_{i=1}^{n} (y_i - X_i'b_{OLS})' (y_i - X_i'b_{OLS}) \). The maximum likelihood estimator of \( \sigma^2 \) is \( \sigma^2_{MLE} = (Error SS)_{Full} / N \) so the maximum likelihood is

\[
L(b_{OLS}, \sigma^2_{MLE}) = -\frac{1}{2} \left( N \ln(2\pi) + N \ln (Error SS)_{Full} - N \ln N + N \right).
\]

Now, write \( X_i = (X_{i1}, X_{i2}) \) where \( X_{i1} \) has dimension \( T_i \times (K-r) \) and \( X_{i2} \) has dimension \( T_i \times r. \) Under \( H_0, \) the estimator of \( \beta_1 \) is \( b_{Reduced} = \left( \sum_{i=1}^{n} X_{i1}'X_{i1} \right)^{-1} \sum_{i=1}^{n} X_{i1}'y_i \). Thus, under \( H_0, \) the log-likelihood is:

\[
L(b_{OLS, Reduced}, \sigma^2_{MLE, Reduced}) = -\frac{1}{2} \left( N \ln(2\pi) + N \ln (Error SS)_{Reduced} - N \ln N + N \right),
\]

where \( (Error SS)_{Reduced} = \sum_{i=1}^{n} (y_i - X_{i1}'b_{Reduced})' (y_i - X_{i1}'b_{Reduced}) \). Thus, the likelihood ratio test statistic is:

\[
LRT_{MLE} = 2 \left( L(b_{OLS, \sigma^2_{MLE}}) - L(b_{OLS, Reduced, \sigma^2_{MLE, Reduced}}) \right) = N \ln \left( \frac{(Error SS)_{Reduced}}{(Error SS)_{Full}} \right).
\]

From a Taylor series approximation, we have \( \ln y = \ln x + \frac{y - x}{x} - \frac{1}{2} \frac{(y - x)^2}{x^2} + \ldots \). Thus, we have

\[
LRT_{MLE} = N \left( \frac{(Error SS)_{Reduced} - (Error SS)_{Full}}{(Error SS)_{Full}} \right) + \ldots
\]
which has an approximate chi-square distribution with \( r \) degrees of freedom.

For comparison, from equation (3A.1), the “restricted” log-likelihood is

\[
L_{REML} (b_{OLS}, \sigma^2) = -\frac{1}{2} \left( N \ln(2\pi) + (N - K) \ln \sigma^2 + \frac{1}{\sigma^2} (Error \ SS)_{Full} \right) - \frac{1}{2} \ln \det \left( \sum_{i=1}^{n} X'_i X_i \right).
\]

The restricted maximum likelihood estimator of \( \sigma^2 \) is

\[
\sigma^2_{REML} = \left( Error \ SS \right)_{Full} / (N - K).
\]

Thus, the restricted maximum likelihood is

\[
L_{REML} (b_{OLS}, \sigma^2_{REML}) = -\frac{1}{2} \left( N \ln(2\pi) + (N - K) \ln\left( Error \ SS \right)_{Reduced} \right) - \frac{1}{2} \ln \det \left( \sum_{i=1}^{n} X'_i X_i \right)
\]

\[
+ \frac{1}{2} \left( (N - K) \ln(N - K) - (N - K) \right).
\]

Under \( H_0 \), the restricted log-likelihood is:

\[
L_{REML} (b_{OLS, Reduced}, \sigma^2_{REML, Reduced})
\]

\[
= -\frac{1}{2} \left( N \ln(2\pi) + (N - (K - q)) \ln\left( Error \ SS \right)_{Reduced} \right) - \frac{1}{2} \ln \det \left( \sum_{i=1}^{n} X'_{i, li} X_{i, li} \right)
\]

\[
+ \frac{1}{2} \left( (N - (K - q)) \ln(N - (K - q)) - (N - (K - q)) \right).
\]

Thus, the likelihood ratio test statistic using a restricted likelihood is:

\[
LRT_{REML} = 2 \left[ L_{REML} (b_{OLS}, \sigma^2_{REML}) - L_{REML} (b_{OLS, Reduced}, \sigma^2_{REML, Reduced}) \right]
\]

\[
= (N - K) \left[ \ln\left( Error \ SS \right)_{Reduced} - \ln\left( Error \ SS \right)_{Full} \right] + q \ln\left( Error \ SS \right)_{Reduced}
\]

\[
+ \sum_{i=1}^{n} \left( \ln \det \left( X'_{i, li} X_{i, li} \right) - \ln \det \left( X'_{i} X_{i} \right) \right) + (N - K) \ln \frac{N - K}{N - (K - q)} - q \ln\left( N - (K - q) \right) - q
\]

\[=(N - K) LRT_{MLE} + \left( \ln \det \left( \sum_{i=1}^{n} X'_{i, li} X_{i, li} \right) - \ln \det \left( \sum_{i=1}^{n} X'_{i} X_{i} \right) \right)
\]

\[
+ q \ln \left( \frac{\left( Error \ SS \right)_{Reduced}}{N - (K - q)} - 1 \right) + (N - K) \ln \left( 1 - \frac{q}{N - (K - q)} \right).
\]

The first term is asymptotically equivalent to the likelihood ratio test, using “ordinary” maximized likelihoods. The third and fourth terms tend to constants. The second term, \( \ln \det \left( \sum_{i=1}^{n} X'_{i, li} X_{i, li} \right) - \ln \det \left( \sum_{i=1}^{n} X'_{i} X_{i} \right) \), may tend to plus or minus infinity, depending on the values of the explanatory variables. For example, in the special case that \( X'_{i} X_{i} = 0 \), we have

\[
\ln \det \left( \sum_{i=1}^{n} X'_{i, li} X_{i, li} \right) - \ln \det \left( \sum_{i=1}^{n} X'_{i} X_{i} \right) = (-1)^{n} \ln \det \left( \sum_{i=1}^{n} X'_{2i} X_{2i} \right).
\]

Thus, this term will tend to plus or minus infinity for most explanatory variable designs.
3. Exercises and Extensions

Section 3.1

3.1. Generalized least squares (GLS) estimators

For the error components model, the variance of the vector of responses is given as
\[ V_i = \sigma_\alpha^2 J_i + \sigma_\epsilon^2 I_i. \]

a. By multiplying \( V_i \) by \( V_i^{-1} \), check that
\[ -\begin{pmatrix} \zeta_i \sigma_\epsilon & J_i \end{pmatrix} = -\begin{pmatrix} \zeta_i \sigma_\epsilon & \frac{\zeta_i}{T_i} J_i \end{pmatrix}. \]

b. Use this form of \( V_i^{-1} \) and the expression for a GLS estimator,
\[ -\begin{pmatrix} \sum_{i=1}^n X_i' V_i^{-1} X_i \end{pmatrix}^{-1} \sum_{i=1}^n X_i' V_i^{-1} y_i, \]
to establish the formula in equation (3.3).

c. Use equation (3.4) to show that the basic random effects estimator can be expressed as:
\[ b_{EC} = \left( \sum_{i=1}^n X_i' \left( T_i^{-1} J_i \right) X_i \right)^{-1} \sum_{i=1}^n \left( X_i' \left( T_i^{-1} J_i \right) y_i \right). \]

d. Show that
\[ b = \left( \sum_{i=1}^n X_i' \left( I_i - T_i^{-1} J_i \right) X_i \right)^{-1} \sum_{i=1}^n X_i' \left( I_i - T_i^{-1} J_i \right) y_i \]
is an alternative expression for the basic fixed effects estimator given in equation (2.6).

e. Suppose that \( \sigma_\alpha^2 \) is large relative to \( \sigma_\epsilon^2 \) so that we assume that \( \sigma_\alpha^2 / \sigma_\epsilon^2 \to \infty \). Give an expression and interpretation for \( b_{EC} \).

f. Suppose that \( \sigma_\alpha^2 \) is small relative to \( \sigma_\epsilon^2 \) so that we assume that \( \sigma_\alpha^2 / \sigma_\epsilon^2 \to 0 \). Give an expression and interpretation for \( b_{EC} \).

3.2. GLS estimator as a weighted average

Consider the basic random effects model and suppose that \( K = 1 \) and that \( x_{it} = 1 \). Show that
\[ b_{EC} = \frac{\sum_{i=1}^n \zeta_i y_i}{\sum_{i=1}^n \zeta_i}. \]

3.3. Error components with one explanatory variable

Consider the error components model, \( y_{it} = \alpha_i + \beta_0 + \beta_1 x_{it} + \epsilon_{it} \). That is, consider the model in equation (3.2) with \( K = 2 \) and \( x_{it} = (1 x_{it})' \).

a. Show that \( \zeta_i = \frac{\sigma_\epsilon^2}{\sigma_\alpha^2} T_i (1 - \zeta_i) \).

b. Show that we may write the generalized least squares estimators of \( \beta_0 \) and \( \beta_1 \) as
\[ b_{1,EC} = \frac{\sum_{i,t} x_{it} y_{it} - \sum_{i} T_i \zeta_i x_i y_i - \left( \sum_{i} (1 - \zeta_i) T_i \right) \bar{x}_w \bar{y}_w}{\sum_{i,t} x_{it}^2 - \sum_{i} T_i \zeta_i x_i^2 - \left( \sum_{i} (1 - \zeta_i) T_i \right) \bar{x}_w^2}, \]
and
\[ b_{0,EC} = \bar{y}_w - \bar{x}_w b_{1,EC}, \]
where
\[ x_{w} = \frac{\sum_{i} \zeta_{i} x_{i}}{\sum_{i} \zeta_{i}} \text{ and } y_{w} = \frac{\sum_{i} \zeta_{i} y_{i}}{\sum_{i} \zeta_{i}}. \]

(Hint: use the expression of \( b_{1,EC} \) in Exercise 3.1c.)

c. Suppose that \( \sigma_{a}^{2} \) is large relative to \( \sigma^{2} \) so that we assume that \( \sigma_{a}^{2} / \sigma^{2} \rightarrow \infty \). Give an expression and interpretation for \( b_{1,EC} \).

d. Suppose that \( \sigma_{a}^{2} \) is small relative to \( \sigma^{2} \) so that we assume that \( 0 / \sigma_{a}^{2} \rightarrow \infty \). Give an expression and interpretation for \( b_{1,EC} \).

3.4. Two-population slope interpretation

Consider the basic random effects model and suppose that \( K = 1 \) and that \( x \) is binary variable. Suppose further that \( x \) takes on the value of 1 for those from population 1 and \(-1\) for those from population 2. Analogous to Exercise 2.4, let \( n_{1,i} \) and \( n_{2,i} \) be the number of ones and minus ones for the \( i \)th subject, respectively. Further, let \( y_{1,i} \) and \( y_{2,i} \) be the average response when \( x \) is one and minus one, for the \( i \)th subject, respectively.

Show that we may write the error components estimator as
\[ b_{1,2,1,1} = \frac{\sum_{i=1}^{n} (w_{1,i}y_{1,i} - w_{2,i}y_{2,i})}{\sum_{i=1}^{n} (w_{1,i} + w_{2,i})}, \]

with weights \( w_{1,i} = n_{1,i}(1 + \zeta_{i} - 2\zeta_{i}n_{1,i} / T_{i}) \) and \( w_{2,i} = n_{2,i}(1 + \zeta_{i} - 2\zeta_{i}n_{2,i} / T_{i}) \).

(Hint: use the expression of \( b_{EC} \) in Exercise 3.1c.)

3.5. Unbiased variance estimators

Perform the following steps to check that the variance estimators given by Baltagi and Chang (1994E) are unbiased variance estimators for the unbalanced error components model introduced in Section 3.1. For notational simplicity, assume the model follows the form \( y_{it} = \mu_{a} + \alpha_{i} + x_{it}' \beta + \epsilon_{it} \), where \( \mu_{a} \) is a fixed parameter, represent the model intercept. As described in Section 3.1, we will use the residuals \( e_{it} = y_{it} - (a_{i} + x_{i}' b) \) using \( a_{i} \) and \( b \) according to the Chapter 2 fixed effects estimators in equations (2.6) and (2.7).

a. Show that response deviations can be expressed as \( y_{it} - \overline{y}_{i} = (x_{it} - \overline{x}_{i})' \beta + \epsilon_{it} - \overline{\epsilon}_{i} \).

b. Show that the fixed effects slope estimator can be expressed as
\[ b = \beta + \left( \sum_{i=1}^{n} \sum_{t=1}^{T_{i}} (x_{it} - \overline{x}_{i})(x_{it} - \overline{x}_{i})' \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T_{i}} (x_{it} - \overline{x}_{i})e_{it}. \]

c. Show that the residual can be expressed as \( e_{it} = (x_{it} - \overline{x}_{i})(\beta - b) + \epsilon_{it} - \overline{\epsilon}_{i} \).

d. Show that the mean square error defined in equation (2.11) is an unbiased estimator for this model. That is, show that
\[ \text{EMSE} = E \left( 1 \bigg/ N - (n + K) \right) \sum_{i=1}^{n} \sum_{t=1}^{T_{i}} e_{it}^{2} = \sigma^{2}. \]

e. Show that \( a_{i} - \overline{a}_{w} = \alpha_{i} - \overline{\alpha}_{w} + (\overline{x}_{i} - \overline{x})(\beta - b) + \overline{\epsilon}_{i} - \overline{\epsilon} \), where \( \overline{a}_{w} = N^{-1} \sum_{i=1}^{n} T_{i} \alpha_{i} \).

f. Show that \( E(\alpha_{i} - \overline{\alpha}_{w})^{2} = \sigma_{a}^{2} \left( 1 + N^{-2} \sum_{i=1}^{n} T_{i}^{2} - 2T_{w} / N \right) \).

g. Show that \( E_{s_{a}^{2}} = \sigma_{a}^{2} \).
3.6. Ordinary least squares estimator

Perform the following steps to check that the ordinary least square estimator of the slope coefficient still performs well when the error components model is true. To this end:

a. Show that the ordinary least squares estimator for the model \( y_{it} = x_{it}' \beta + \epsilon_{it} \) can be expressed as:

\[
\beta_{OLS} = \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it} x_{it}' \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it} y_{it} \right).
\]

b. Assuming the error components model, \( y_{it} = \alpha_i + x_{it}' \beta + \epsilon_{it} \), show that the difference between the part (a) estimator and the vector of parameters is:

\[
\beta_{OLS} - \beta = \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it} x_{it}' \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it} (\alpha_i + \epsilon_{it}) \right).
\]

c. Use part (b) to argue that the estimator given in part (a) is unbiased.

d. Calculate the variance of \( \beta_{OLS} \).

e. For \( K=1 \), show that the variance calculated in part (d) is larger than the variance of the random effects estimator, \( \text{Var} \beta_{EC} \), given in Section 3.1.

3.7. Pooling test

Perform the following steps to check that the test statistic for the pooling test given in Section 3.1 has an approximate chi-square distribution under the null hypothesis of a homogeneous model of the form \( y_{it} = x_{it}' \beta + \epsilon_{it} \).

a. Check that the residuals can be expressed as \( e_{it} = \epsilon_{it} + x_{it}'(\beta - \beta_{OLS}) \), where \( \beta_{OLS} \) is the ordinary least squares estimator of \( \beta \) in Exercise 3.5a.

b. Check that \( \frac{1}{N-K} \sum_{i=1}^{n} \sum_{t=1}^{T_i} e_{it}^2 = \sigma^2 \).

c. Check that \( T_i (T_i - 1) s_i = T_i^2 \bar{e}_i^2 - \sum_{t=1}^{T_i} e_{it}^2 = \sum_{r \neq s} e_{ir} e_{is} \), where the latter sum is over \( \{(r, s) \text{ such that } s \neq r \text{ and } r=1, \ldots, T_i, s=1, \ldots, T_i\} \).

d. Check, for \( s \neq r \), that \( \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{T_i} \sum_{s=1}^{T_i} h_{ir, is} e_{ir} e_{is} = -h_{ir, is} \sigma^2 \), where \( h_{ir, is} = x_{ir}' \left( \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{it} x_{it}' \right)^{-1} x_{is} \) is an element of the hat matrix.

e. Establish conditions so that the bias is negligible. That is, check that

\[
n^{-1/2} \sum_{i=1}^{n} h_{ir, is} / \sqrt{T_i (T_i - 1)} \rightarrow 0.
\]

f. Determine the approximate variance of \( s_i \) by showing that

\[
\text{E} \left( \frac{1}{T_i (T_i - 1)} \sum_{r \neq s} e_{ir} e_{is} \right)^2 = \frac{2\sigma^4}{T_i (T_i - 1)}.
\]

g. Outline an argument to show that \( \frac{1}{\sigma^2 \sqrt{2n}} \sum_{i=1}^{n} s_i \sqrt{T_i (T_i - 1)} \) is approximately standard normal, thus, completing the argument for the behavior of the pooling test statistic under the null hypothesis.
Section 3.3
3.8. Nested models
Let \( y_{i,j,t} \) be the output of the \( j \)th firm in the \( i \)th industry for the \( t \)th time period. Assume that the error structure is given by
\[
y_{i,j,t} = E + \delta_{i,j,t},
\]
where \( \delta_{i,j,t} = \alpha_i + \nu_{i,j} + \epsilon_{i,j,t} \). Here, assume that each of \( \{\alpha_i\} \), \( \{\nu_{i,j}\} \) and \( \{\epsilon_{i,j,t}\} \) are independently and identically distributed and independent of one another.

a. Let \( y_i \) be the vector of responses for the \( i \)th industry. Write \( y_i \) as a function of \( \{y_{i,j,t}\} \).

b. Use \( 2\alpha\sigma_\alpha, 2\nu\sigma_\nu \) and \( 2\epsilon\sigma_\epsilon \) to denote the variance of each error component, respectively. Give an expression for \( \text{Var} y_i \) in terms of these variance components.

c. Consider the linear mixed effects model,
\[
y_i = Z_i \alpha_i + X_i \beta + \epsilon_i.
\]
Show how to write the quantity \( Z_i \alpha_i \) in terms of the error components \( \alpha_i \) and \( \nu_{i,j} \) and the appropriate explanatory variables.

Section 3.4
3.9. GLS estimator as a weighted average of subject-specific GLS estimators
Consider the random coefficients models and consider the weighted average expression for the GLS estimator
\[
b_{\text{GLS}} = \left( \sum_{i=1}^{n} W_{i,G} \right)^{-1} \sum_{i=1}^{n} W_{i,G} \mathbf{b}_{i,G}.
\]
a. Show that the weights can be expressed as \( W_{i,G} = \left(D + \sigma^2(X'X)^{-1}\right)^{-1} \).
b. Show that \( \text{Var} \mathbf{b}_{i,G} = D + \sigma^2(X'X)^{-1} \).

3.10. Matched pairs design
Consider a pair of observations that have been “matched” in some way. The pair may consist of siblings, firms with similar characteristics but from different industries, or the same entity observed before and after some event of interest. Assume that there is reason to believe that the pair of observations are dependent in some fashion. Let \( (y_{i1}, y_{i2}) \) be the set of responses and \( (x_{i1}, x_{i2}) \) be the corresponding set of covariates. Because of the sampling design, the assumption of independence between \( y_{i1} \) and \( y_{i2} \) is not acceptable.

a. One alternative is to analyze the difference between the responses. Thus, let \( y_i = y_{i1} - y_{i2} \). Assuming perfect matching, we might assume that \( x_{i1} = x_{i2} = x_i \), say, and use the model \( y_i = x_i' \gamma + e_i \). Without perfect matching, one could use the model \( y_i = x_{i1}' \beta_1 - x_{i2}' \beta_2 + \eta_i \). Calculate the ordinary least squares estimator \( \beta = (\beta_1', \beta_2')' \), and call this estimator \( \mathbf{b}_{PD} \) because the responses are paired differences.

b. As another alternative, form the vectors \( y_i = (y_{i1}, y_{i2})' \) and \( e_i = (e_{i1}, e_{i2})' \), as well as the matrix
\[
X_i = \begin{pmatrix} x_{i1}' & 0 \\ 0 & x_{i2}' \end{pmatrix}.
\]
Now, consider the linear mixed effects model \( y_i = X_i \beta + e_i \), where the dependence between responses is induced by the variance \( \mathbf{R} = \text{Var} e_i \).

i Under this model specification, show that \( \mathbf{b}_{PD} \) is unbiased.

ii Compute the variance of \( \mathbf{b}_{PD} \).

iii Calculate the generalized least squares estimator of \( \beta \), say \( \mathbf{b}_{GLS} \).

c. For yet another alternative, assume that the dependence is induced by a common latent random variable \( \alpha_i \). Specifically, consider the error components model \( y_{i1} = \alpha_i + x_i' \beta_1 + e_{i1} \).

i Under this model specification, show that \( \mathbf{b}_{PD} \) is unbiased.
ii Calculate the variance of \( b_{PD} \).

iii Let \( b_{EC} \) be the generalized least squares estimator under this model. Calculate the variance.

iv Show that if \( \sigma^2 \to \infty \), then the variance of \( b_{EC} \) tends to the variance of \( b_{PD} \). Thus, for highly correlated data, the two estimators have the same efficiency.

d. Continue with the model in part (c). We know that \( \operatorname{Var} b_{EC} \) is “smaller” than \( \operatorname{Var} b_{PD} \), because \( b_{EC} \) is the generalized least squares estimator of \( \beta \). To quantify this in a special case, assume asymptotically equivalent matching so that

\[
n^{-1} \sum_{i=1}^n x_i' x_i \to \Sigma,
\]

and assume that \( \Sigma \) is symmetric. Suppose that we are interested in differences of the respondents, so that the (vector of) parameters of interest are \( \beta_1 - \beta_2 \). Let \( b_{EC} = (b_{1,EC}', b_{2,EC}')' \) and \( b_{PD} = (b_{1,PD}', b_{2,PD}')' \).

i Show that

\[
\frac{n}{\sigma^2} \operatorname{Var}(b_{1,EC} - b_{2,EC}) = \frac{2}{(1 - \zeta/2)} (\Sigma + z\Sigma_{12})^{-1},
\]

where \( z = \frac{\zeta / 2}{(1 - \zeta / 2)} \) and \( \zeta = \frac{2\sigma^2}{\sigma^2 + \sigma^2} \).

ii Show that

\[
\frac{n}{\sigma^2} \operatorname{Var}(b_{1,PD} - b_{2,PD}) = 4(\Sigma + z\Sigma_{12})^{-1}
\]

iii Use parts d(i) and d(ii) to quantify the relative variances. For example, if \( \Sigma_{12} = 0 \), then the relative variances (efficiency) is \( \frac{1}{(2 - \zeta)} \) which is between 0.5 and 1.0.

### 3.11. Robust standard errors

To estimate the linear mixed effects model, consider the weighted least squares estimator, given in equation (3.18). The variance of this estimator, \( \operatorname{Var} b_W \), is also given in Section 3.4, along with the corresponding robust standard error of the \( j \)th component of \( b_W \), denoted as \( \operatorname{se}(b_{W,j}) \). In particular, explain:

a. How one goes from a variance-covariance matrix to a standard error and

b. What about this standard error makes it “robust?”

c. Let’s derive a new robust standard error. To keep things simple, assume the linear mixed effects model with \( n = 1 \) and drop the \( i \) subscripts. Now, define the “hat matrix”

\[
H_W = W^{-1/2} X (X' W^{-1/2} X)^{-1} X' W^{-1/2}.
\]

i Show that the weighted residuals can be expressed as a linear combination of weighted errors. That is, show

\[
W^{-1/2} e = (I - H_W) W^{-1/2} e.
\]

ii Show that

\[
E \left(W^{-1/2} ee' W^{-1/2} \right) = (I - H_W) W^{-1/2} V W^{-1/2} (I - H_W).
\]

iii Show that \( ee' \) is an unbiased estimator of a linear transform of \( V \). Specifically, show that

\[
E (ee') = (I - H_W') V (I - H_W')
\]

where \( H_W' = W^{-1/2} H_W W^{-1/2} \).

iv Use the result in c(iii) to suggest a new robust estimator, analogous to \( \operatorname{se}(b_{W,j}) \) in part (b) above. (Hint: If the question seems difficult, you may assume that the weight matrix, \( W_{RE} \), is an identity matrix. It makes the calculations simpler!)
Section 3.5

3.12. Bias of MLE and REML variance component estimators

Consider the basic random effects model and suppose that $T_i = T$, $K = 1$ and that $x_{it} = 1$. Further, do not impose boundary conditions so that the estimators may be negative.

a. Show that the maximum likelihood estimator of $\sigma^2$ may be expressed as:

$$\hat{\sigma}^2_{ML} = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2.$$

b. Show that $\hat{\sigma}^2_{ML}$ is an unbiased estimator of $\sigma^2$.

c. Show that the maximum likelihood estimator of $\sigma^2_{\alpha}$ may be expressed as:

$$\hat{\sigma}^2_{\alpha,ML} = \frac{1}{n} \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2 - \frac{1}{T} \hat{\sigma}^2_{ML}.$$

d. Show that $\hat{\sigma}^2_{\alpha,ML}$ is a biased estimator of $\sigma^2_{\alpha}$ and determine the bias.

e. Show that the restricted maximum likelihood estimator of $\sigma^2$ equals the corresponding maximum likelihood estimator, that is, show $\hat{\sigma}^2_{REML} = \hat{\sigma}^2_{ML}$.

f. Show that the restricted maximum likelihood estimator of $\sigma^2_{\alpha}$ may be expressed as:

$$\hat{\sigma}^2_{\alpha,REML} = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2 - \frac{1}{T} \hat{\sigma}^2_{ML}.$$

g. Show that $\hat{\sigma}^2_{\alpha,REML}$ is an unbiased estimator of $\sigma^2_{\alpha}$.

Empirical Exercises


a. Error components model

Run an error components model of CHARITY on INCOME, PRICE, DEPS, AGE and MS. State which variables are statistically significant and justify your conclusions.

b. Re-run the step in part (a) by including the supply-side measures as additional explanatory variables. State whether or not these variables should be included in the model. Explain your reasoning.

c. Incorporating temporal effects. Is there an important time pattern? For the model in part a(i):

i. re-run it excluding YEAR as an explanatory variable yet including an AR(1) serial component for the error.

ii. re-run it including YEAR as an explanatory variable and including AR(1) serial component for the error.

iii. re-run it including YEAR as an explanatory variable and including an unstructured serial component for the error. (This step may be difficult to achieve convergence of the algorithm!)

iv. Which model do you prefer, (i), (ii), or (iii)? Justify your choice. In your justification, discuss the nonstationarity of errors.

d. Variable slope models

i. Re-run the model in part (a) including a variable slope for INCOME. State which of the two models is preferred and state your reason.

ii. Re-run the model in part (a) including a variable slope for PRICE. State which of the two models is preferred and state your reason.

Final Part. Which model do you think is best? Do not confine yourself to the options that you tested in the preceding parts. Justify your choice.
a Run an error components model using state as the subject identifier and VEHCMILE, GSTATEP, POPDENSY, WCMPMAX, URBAN, UNEMPLOY and JSLIAB as explanatory variables.
b Re-run the error components model in part (a) and include the additional explanatory variables COLLRULE, CAPS and PUNITIVE. Test whether these additional variables are statistically significant using the likelihood ratio test. State your null and alternative hypotheses, your test statistic and decision-making rule.
c Notwithstanding your answer in part (b), re-run the model in part (a) but also include variable random coefficients associated with WCMPMAX. Which model do you prefer, the model in part (a) or this one?
d Just for fun, re-run the model in part (b) and including variable random coefficients associated with WCMPMAX.
e Re-run the error components model in part (a) but include an autoregressive error of order (1). Test for the significance of this term.
f Run the model in part (a) but with fixed effects. Compare this model to the random effects version.

3.15. Housing Prices – refer to Exercise 2.21 for the problem description.

a Basic summary statistics
   i Produce a multiple time series plot of NARSP.
   ii Produce a multiple time series plot of YPC.
   iii Produce a scatter plot of NARSP versus YPC.
   iv Produce an added variable plot of NARSP versus YPC, controlling for the effects of MSA.
   v Produce a scatter plot of NARSP versus YPC.

b Error components model
   i Run a one-way error components model of NARSP on YPC and YEAR. State which variables are statistically significant.
   ii Re-run the step in b(i) by including the supply-side measures as additional explanatory variables. State whether or not these variables should be included in the model. Explain your reasoning.

c Incorporating temporal effects. Is there an important time pattern?
   i Run a one-way error components model of NARSP on YPC. Calculate residuals from this model. Produce a multiple time series plot of residuals.
   ii Re-run the model in part c(i) and include an AR(1) serial component for the error. Discuss the stationarity of errors based on the output of this model fit and your analysis in part c(i).

d Variable slope models
   i Re-run the model in part c(i), including a variable slope for YPC. Assume that the two random effects (intercepts and slopes) are independent.
   ii Re-run the model in part d(i) but allow for dependence between the two random effects. State which of the two models you prefer and why.
   iii Re-run the model in part d(ii) but incorporate the time-constant supply-side variables. Estimate the standard errors using robust standard errors. State which variables are statistically significant; justify your statement.
   iv Given the discussion of non-stationarity in part (c), describe why robust variance estimators are preferred when compared to the model-based standard errors.