University of Tartu

Faculty of Science and Technology

Institute of Mathematics and Statistics

Indrek Künnapas

Predicting Annual Returns of Individual Stocks on the Baltic Stock Exchange

Master’s thesis (30 EAP)
Actuarial and Financial Engineering

Supervisor: Raul Kangro, PhD

Tartu 2025

Predicting Annual Returns of Individual Stocks on the Baltic Stock Exchange

Master's thesis

Indrek Künnapas

Abstract. The objective of this master's thesis is to investigate possibilities of deploying statistical models in predicting annual returns of stocks on the Baltic Stock Exchange. Both, theoretical and practical aspects of linear regression, fixed effects, random effects, mixed effects and auto regressive model, are illustrated.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics.
Keywords: linear regression, fixed effects, random effects, mixed effects, auto regressive, stock market.

Balti Börsi aktsiate aastase tulumäära prognoosimine

Magistritöö

Indrek Künnapas

Lühikokkuvõte. Magistritöö uurib võimalusi statistiliste mudelite kasutamiseks Balti Börsi aktsiate aastase tulumäära prognoosimisel. Illustreeritakse nii teoreetilisi kui ka praktilisi aspekte, mis seotud lineaarse regressiooni, fikseeritud mõjudega, juhuslike mõjudega, sega-mõjudega ja autoregressiivse mudeli kasutamisega.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.
Märksõnad: lineaarne regressioon, fikseeritud mõjud, juhuslikud mõjud, sega-mõjud, autoregressiivne, aktsia turg.

Table of Contents

Introduction

Most of the stock market analyses, is relaying on either the stochastic models, or even more commonly, standalone ratios indicating the relative price of individual shares (like price-to-equity ratio). As alternative, the current master's thesis, analyses possibilities to rely on statistical regression models when predicting annual returns of individual stocks. The analyses is carried out on all listed companies in the main markets of the Tallinn, Riga and Vilnius stock exchange. All-together, 16 companies from Estonia, 3 of Latvia and 13 from Lithuanian were analysed.

The total list of Estonian companies included Arco Vara, Baltika, Coop Pank, Ekspress Grupp, Eften Real Estate, Harju Elekter, LHV Grupp, Merko Ehitus, Nordecon, Pro Kapital Grupp, PRFoods, Silvano Fashion Grupp, Tallink Grupp, Tallinna Kaubamaja, Tallinna Sadam and Tallinna Vesi. For Latvia, enterprises SAF Tehnika, Hansa Matrix and Olainfarm were analysed. The companies included for Lithuania were Apranga, Auga group, Grigeo, Ignitis grupė, Klaipėdos nafta, Linas Agro Group, Novaturas, Panevėžio statybos trestas, Pieno žvaigždės, Rokiškio sūris, Šiaulių bankas, Telia Lietuva and Vilkyškių pieninė.

In order to increase the number of observations, each company was included up to 7 times into analyses:

1) using data from 2024 to predict returns during 2025 (until March);
2) using data from 2023 to predict returns during 2024;
3) using data from 2022 to predict returns during 2023;
4) using data from 2021 to predict returns during 2022;
5) using data from 2020 to predict returns during 2021;
6) using data from 2019 to predict returns during 2020;
7) using data from 2018 to predict returns during 2019.

The relatively long outcome prediction horizon, of one year, should enable usability of traditional prediction techniques, like regression model. At the same time inclusion of one company more than once, which is typical to panel survey, raises the need analyse possible panel effects. Such is investigated using the fixed and mixed effects models.

Based on the author experience and the conclusions from data analyses, the decision was taken to exclude from the final model construction data-set such observations, where company made loss during the given calendar year. Often, companies in financial difficulties, go through substantial re-organizations, which together with volatility in share prices, introduces major obstacles in using traditional statistical modelling techniques. Also, as current thesis carries the practical aim, of building a model, which could help an average investor while making long term investment decisions, then focus on profitable companies seems reasonable.

1. General theory behind chosen statistical models
1.1 Linear regression model

Mathematical Formulation

Linear regression models illustrate the relationship between a dependent variable and one or more independent variables using a linear function. In the general multiple linear regression form (with n observations and k predictors), the model can be written for each observation i as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where y_i is the outcome, x_{ij} is the j-th predictor for observation i, β_j where j = 1, …, k are regression coefficients, and ε_i is the error term for observation i. In matrix notation, this is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of outcomes, $\mathbf{X}$ is an $n \times (k+1)$ design matrix (including a column of ones for the intercept), $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of coefficients (including β_0), and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors. The ordinary least squares (OLS) estimator for $\boldsymbol{\beta}$ is obtained by solving the normal equations $\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$. The solution is:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

assuming $\mathbf{X}'\mathbf{X}$ is invertible. This $\widehat{\boldsymbol{\beta}}$ minimizes the sum of squared residuals $\sum_i \hat{\varepsilon}_i^2$ and provides the best linear fit to the data in a least-squares sense (Möls, M. 2019).

Key Assumptions

For the linear regression model to yield unbiased and efficient estimates (under the Gauss–Markov theorem), several key assumptions (often called the Classical Linear Model or Gauss-Markov conditions) must hold (Searle, R.S., Casella, G., McCulloch, C.E. 2006):

- **Linearity:** The model is linear in parameters. We assume $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, meaning the expected value of y is a linear combination of the predictors. (Nonlinear relationships can often be accommodated by transforming variables, as long as the model remains linear in the $\boldsymbol{\beta}$ coefficients)
- **Full Rank (No Perfect Multicollinearity):** The matrix $\mathbf{X}$ has full column rank $k + 1$. No independent variable is an exact linear combination of others. This condition ensures $(\mathbf{X}'\mathbf{X})^{-1}$ exists and coefficients are identifiable.
- **Zero Conditional Mean (Exogeneity):** The error term has an expectation of zero given any value of the independent variables: $E(\varepsilon) = 0$. Equivalently, the regressors are uncorrelated with the error. This assumption is crucial for $\hat{\beta}$ to be unbiased.
- **Homoskedasticity and No Autocorrelation:** The error term has constant variance and no serial correlation: $\mathrm{Var}(\varepsilon) = \sigma^2 \mathbf{I}_n$. All disturbances have the same variance σ^2 and are uncorrelated with each other (no systematic patterns in residuals). This is sometimes called the

spherical errors assumption. If errors are heteroskedastic or autocorrelated, OLS is still unbiased but no longer efficient (there exist better linear unbiased estimators).

- **Independence:** Observations are independent of each other (especially important in cross-sectional data). This is not explicitly listed in Gauss-Markov assumptions but is usually assumed for valid standard errors and inference.
- **Normality:** For inference (hypothesis tests, confidence intervals), one often assumes the errors are normally distributed $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Normality is not required for OLS to be unbiased or Best Linear Unbiased Estimator (BLUE), but it simplifies derivation of t/F-tests and ensures $\hat{\beta}$ is normally distributed. By the Central Limit Theorem, with large n this assumption becomes less critical.

Estimation (Ordinary Least Squares)

Under the above assumptions, the OLS method provides the BLUE for the regression coefficients (Gauss–Markov theorem). The OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is unbiased ($E(\hat{\beta}) = \beta$) so long as the regressors are exogenous (uncorrelated with errors). The intuition is that, on average, the errors cancel out and do not systematically push $\hat{\beta}$ away from the true β. The covariance matrix of $\hat{\beta}$ is $\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ under homoskedastic errors, which can be estimated by $\widehat{\sigma^2}(\mathbf{X}'\mathbf{X})^{-1}$ in practice. If heteroskedasticity is present, one may use generalized least squares or heteroskedasticity-robust standard errors to get correct inference (since OLS remains unbiased but is no longer efficient in that case). Estimation in matrix form can also be derived via maximum likelihood if $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$, in which case OLS coincides with the maximum likelihood estimator for β (Möls, M. 2019).

When to Use Linear Regression

Linear regression is appropriate for cross-sectional data or any dataset where the goal is to model an outcome as a linear combination of explanatory variables without complex grouping or hierarchical structure. It assumes all relevant predictors (especially those correlated with the outcome) can be included or controlled, and there is no need to account for unobserved group-specific effects or repeated measure's structure. In practice, multiple linear regression is a foundational tool used whenever the relationship is approximately linear, and the researcher can assume (or test for) the validity of the linear model assumptions. It can handle one or many independent variables (hence "multiple regression" for $k > 1$ predictors). The model is flexible in that predictors can be numeric or categorical (with dummy variables encoding categories) and can be extended with interaction terms or polynomial terms to capture more complex patterns while remaining linear in the parameters.

This model is suitable for a single-level data structure. For example, it is ideal for a one-time survey of individuals where everyone's income is regressed on their education, experience, and age. If each observation is independent and there are no inherent clusters or repeated measurements, a linear regression is appropriate. It is less appropriate if the data have grouped structure (e.g., students within schools, or repeated observations of the same individual over time), because then the independence assumption is violated and more advanced models (like fixed, random, or mixed effects) may be needed to account for that structure.

Interpretation of Coefficients and Errors

In a multiple linear regression, each coefficient β_j represents the partial effect of the j-th predictor on the outcome, holding all other variables constant. Specifically, β_j is the expected change in y for a one-unit increase in x_j, ceteris paribus. The intercept β_0 is the expected value of y when all predictors equal zero (which may or may not be meaningful depending on the variables and their scales). Interpretation should always respect the data's context and range (if $\mathrm{x_j} = 0$ is outside the observed range, β_0 is an extrapolated intercept).

The error term ε_i captures all other factors affecting $\mathrm{y_i}$ that are not included in the model. By assumption, ε_i has mean zero (so it doesn't systematically increase or decrease y) and variance σ^2. Each observed outcome can thus be seen as $\mathrm{y_i} = \hat{y}_i + \hat{\varepsilon}_i$, where $\hat{y}_i = \mathrm{x_i}'\hat{\beta}$ is the fitted value and $\hat{\varepsilon}_i = \mathrm{y_i} - \hat{y}_i$ is the residual. If assumptions hold, $\widehat{\varepsilon_\mathrm{i}}$ has mean 0 and provides an estimate of the random noise in $\mathrm{y_i}$. The coefficient estimates $\hat{\beta}_j$ can be used to infer the direction and magnitude of relationships, and if error normality is assumed, one can construct confidence intervals and perform t-tests to assess if coefficients are significantly different from zero (or any hypothesized value). For instance, a 95% CI for β_j is $\hat{\beta}_j \pm t_{\mathrm{n-k-1,0.975}} \cdot \mathrm{SE}(\hat{\beta}_j)$, where $\mathrm{SE}(\hat{\beta}_j)$ is the standard error derived from the variance-covariance matrix of the estimates, and $t_{\mathrm{n-k-1,0.975}}$ is the critical value from the t-distribution with $n - k - 1$ degrees of freedom.

Diagnostics and Tests

After fitting a linear regression, it is crucial to check model diagnostics to validate the assumptions:

- **Gauss–Markov Conditions:** Verify linearity and correct specification (no obvious non-linear patterns in residual vs. fitted plots), and check for omitted variables. Ensure no perfect multicollinearity (this can be checked via variance inflation factors or by inspecting if $(\mathbf{X}'\mathbf{X})$ is invertible). Assess the zero conditional mean assumption by looking for any pattern in residuals against each predictor; any systematic pattern might indicate model misspecification or endogeneity (e.g. an omitted variable or simultaneity).
- **Homoskedasticity:** Plot residuals versus fitted values or each predictor to see if the spread of residuals is roughly constant. Formal tests like the Breusch–Pagan test or White's test can be used to detect heteroskedasticity. If heteroskedasticity is detected, one might use robust standard errors or transform the model (or use weighted least squares).
- **No Autocorrelation:** If data are temporal or spatial, check residual autocorrelation (e.g., Durbin–Watson test for first-order autocorrelation in time series). For independent cross-sections this isn't an issue, but for time series or panel data, autocorrelation in errors violates OLS assumptions. Remedies include adding lags, using generalized least squares, or explicitly modelling the correlation.
- **Normality of Errors:** Check residual distribution (histogram or Q–Q plot) to see if it's approximately normal. The Jarque–Bera test or Shapiro–Wilk test are formal tests for normality. Mild deviations from normality are usually fine for large samples (by CLT), but strong skewness or heavy tails might suggest using a transformation or a different modelling approach if it affects inference.
- **Influential Observations:** Use leverage and influence diagnostics (like Cook's distance) to identify if any single observation unduly influences the regression results. Large outliers or high-leverage points can violate assumptions or distort the fit, in which case you may consider robust regression or verify data quality.

Overall, if all classical assumptions hold, OLS provides BLUE estimates. If some assumptions are violated, there are corresponding fixes: e.g., generalized least squares (GLS) can be used in place of OLS to handle non-spherical errors and restore efficiency. The linear model's simplicity makes it easy to interpret and a good baseline; subsequent models (fixed, random, mixed effects) build on this by relaxing or modifying assumptions to suit more complex data structures.

1.2 Fixed effects model

Mathematical Formulation

Fixed effects models address one of the key limitations of OLS by allowing for unobserved heterogeneity across groups or individuals when using panel data. In a fixed effects (FE) model, we control for all group-specific factors - even unmeasured ones - by estimating separate intercepts for each group. The mathematical form is:

$$y_{it} = \alpha_i + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_k x_{it,k} + u_{it}$$

where y_{it} is the outcome for unit i at time t , α_i is the unit-specific intercept capturing all time-invariant characteristics of unit, $x_{it,j}$ where j = 1,…, k are the covariates, β_j are the common slopes across units, and u_{it} is the idiosyncratic error. The crucial feature of fixed effects is that it removes the bias caused by any omitted variables that are constant over time but differ between units - this is achieved either through the inclusion of dummy variables for each unit (least squares dummy variable approach) or through transformation methods (such as within-group or de-meaned transformations).

As such, a fixed effects (FE) model extends linear regression to panel or longitudinal data by allowing each observational unit (e.g. person, firm, country) to have its own intercept. In this formulation, α_i are treated as unknown fixed parameters to be estimated (one for each entity). The model can equivalently be written in matrix form as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \mathbf{u},$$

where $\mathbf{D}$ is an $n \times N$ design matrix of dummies (indicator variables) for each entity and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$. Including these dummies explicitly is the Least Squares Dummy Variable (LSDV) approach. An alternative but algebraically identical approach is the within transformation: subtract each entity's time average from the data to remove α_i. Define $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\bar{x}_{ij} = \frac{1}{T}\sum_{t=1}^{T} x_{it,j}$ for each predictor j. Subtracting these from the model gives

$$y_{it} - \bar{y}_i = (\mathbf{x_{it}} - \bar{\mathbf{x}}_\mathbf{i})'\boldsymbol{\beta} + (u_{it} - \bar{u}_i),$$

Here, $\bar{y}_i$ and $\bar{x}_i$ denote the time-averaged outcome and covariates for unit i, and $\bar{u}_i$ is the average residual for that unit. The fixed intercept α_i drops out because $\alpha_i - \alpha_i = 0$. The fixed-effects estimator is then obtained by OLS on these demeaned variables (this is called the within estimator). It yields the same $\widehat{\boldsymbol{\beta}}$ as LSDV but without needing to estimate a separate α_i for each unit explicitly (which can be inefficient if N is large).

Typically, one also allows for a time fixed effect if needed (a λ_t for each time period) to capture shocks common to all units in a given time – this leads to a two-way fixed effects model: $y_{it} = \alpha_i + \lambda_t + \mathbf{x_{it}}'\boldsymbol{\beta} + u_{it}$. The principle is similar, but our focus here is on entity fixed effects as the core idea.

Key Assumptions

The fixed effects model relies on assumptions similar to OLS, but adapted for panel structure (Torres-Reyna, O. 2007):

- **Fixed individual heterogeneity:** Each entity has an arbitrary intercept α_i that does not vary over time and is treated as a fixed constant to be estimated. Crucially, this model allows α_i to be correlated with the regressors x_{it}. This is the defining assumption of FE: any time-invariant differences across entities (observed or not) are absorbed by α_i, so they do not bias the β estimates. There is no requirement that $E(\alpha_i) = 0$; α_i can be anything (for example, a person's innate ability or a firm's culture that might well correlate with their observed inputs).
- **Within-group exogeneity:** The idiosyncratic errors u_{it} satisfy $E(u_{it}) = 0$ for each i. In other words, after controlling for the fixed α_i and all included regressors (over all time periods), there is no remaining endogeneity. This is a strict exogeneity assumption in the panel context – no feedback from past, present, or future u_{it} to current or past x_{it} within the same entity. It rules out lagged dependent variables or other dynamics unless specially handled.
- **No perfect multicollinearity (within variation):** For each predictor, there must be variation within entities over time. If a regressor is constant for each entity (time-invariant), it will be perfectly collinear with α_i and thus cannot be separately identified (this is why fixed effects cannot estimate coefficients on time-invariant variables). This assumption means, for example, you cannot include a dummy for a person's gender in a person fixed-effects model if gender doesn't change over time, because that effect is absorbed by the person-specific intercept.
- **Homoskedasticity and No Serial Correlation (optional):** The classical assumption would be $\mathrm{Var}(u_{it}) = \sigma^2$ and no correlation between u_{it} and u_{is} for $t \neq s$. In practice, panel data often violate these (e.g. errors might be serially correlated within each entity). The fixed-effects estimator $\widehat{\boldsymbol{\beta}}$ remains consistent even if errors have within-panel autocorrelation or heteroskedasticity, but standard errors should then be adjusted (e.g., cluster-robust standard errors by entity) to get valid inference.

Other assumptions include the usual large sample requirements: as N (and/or T) grows, Law of Large Numbers and CLT apply for the estimator. If T is fixed and $N \to \infty$, $\widehat{\boldsymbol{\beta}}$ is consistent under the above conditions (this is the typical asymptotic panel scenario).

Estimation Approach

Estimation of a fixed effects model is most commonly done via OLS on the transformed data (within estimator). Concretely, one can:

- **Within Estimator:** Transform the data by demeaning: for each i, compute $\tilde{y}_{it} = y_{it} - \bar{y}_i$ and $\tilde{x}_{it,j} = x_{it,j} - \bar{x}_{i,j}$ for each regressor. Then run an OLS regression of $\tilde{y}_{it}$ on $\tilde{\mathbf{x}}_{it}$(with no intercept). This will yield $\widehat{\boldsymbol{\beta}}_{FE}$. The fixed intercepts α_i themselves can be recovered as $\widehat{\alpha}_i = \bar{y}_i - \overline{\mathbf{x}_i'}\widehat{\boldsymbol{\beta}}$ (the average residual for each entity).
- **LSDV Estimator:** Include $N - 1$ entity dummies in a pooled OLS regression (omitting one category as baseline to avoid perfect collinearity). This OLS will directly estimate one α_i (or $\widehat{\alpha}_i$) for each entity (relative to the omitted reference) along with $\widehat{\boldsymbol{\beta}}$. Algebraically, this gives the same $\widehat{\boldsymbol{\beta}}$ as the within estimator. However, LSDV becomes cumbersome when N is large (many dummies). In short panels (small T) LSDV is feasible, but in panels with very large N, the within approach or specialized routines are more efficient.

Under the hood, the within estimator solves normal equations on the transformed data, which is equivalent to solving $\sum_{i,t} x_{it,j}(y_{it} - \alpha_i - \mathbf{x}'_{it}\boldsymbol{\beta}) = 0$ for each j and an additional condition $\sum_i \hat{\alpha}_i = 0$ (if an intercept is included) to make the system solvable. The α_i act as nuisance parameters. Notably, as $N \rightarrow \infty$ and fixed T, $\hat{\beta}$ is consistent but $\hat{\alpha}_i$ (with finite T) are not consistently estimated (this is the incidental parameters problem – although α_i consistency is usually not a concern since they are not of primary interest). In practice, one often does not even report the individual α_i estimates, focusing instead on $\hat{\beta}$.

When to Use Fixed Effects (Panel Data Context)

Fixed effects models are appropriate when you have panel (longitudinal) data or any scenario with repeated observations for the same units, and you suspect there are unobserved, time-invariant characteristics of those units that influence the outcome. By using an FE model, you control for all such characteristics, even if you can't measure them, as long as they stay constant over time (Torres-Reyna, O. 2007). This makes FE particularly useful to address omitted variable bias due to unobserved heterogeneity.

Examples: In an economics context, if one has data on various countries over years and want to assess the effect of policy on GDP, a fixed country effect can control for innate differences between countries (geography, culture, etc.) that don't change year-to-year. In a personnel context, if you have multiple performance evaluations per worker, a fixed effect per worker can control for each worker's intrinsic ability or work ethic that is constant across evaluations. Essentially, any time-invariant factor (observed or unobserved) is differenced out by the fixed effects. This is powerful for causal inference: the model effectively compares each entity's outcome to its own history, attributing changes to the included regressors rather than to static traits.

The FE model is the default choice when you suspect α_i might be correlated with the regressors (violating the random effects assumption). If each α_i is thought of as an unobserved confounder that does not change over time, FE is the way to consistently estimate β. However, FE comes at a cost: you cannot identify the effects of any variable that does not vary within an entity. All between-entity variation is swept away with α_i. If time-invariant regressors are of interest (say, ethnicity in a wage regression), a pure fixed-effects approach can't estimate their coefficients. In such cases, one might use random effects or other methods (or include those time-invariant variables as controls in a pooled model if appropriate).

In summary, use FE for panel data when you want to control for all stable differences across units and focus on within-unit variation. It is especially appropriate when N (number of entities) is large, each with moderate T, and you worry about omitted bias from unmeasured stable factors.

Interpretation of Coefficients and Effects

The coefficients $\boldsymbol{\beta}$ in a fixed effects model have a within-unit interpretation. β_j tells you: for a given entity i, if x_{ij} changes by one unit from its usual level (i.e. relative to i's own mean), by how much does y change on average? It's essentially the effect of changing x over time within the same entity. Thus, $\hat{\boldsymbol{\beta}}$ in FE is often called the "within estimator" effect. If one takes an example of individuals and wages, and x is years of education, β would be identified by cases where individuals get more education over time (e.g., finish a degree) and see how their wage changes – the effect is inferred from those within-person changes rather than comparing different people.

The fixed effects α_i themselves capture the time-invariant baseline level of y for entity i, after controlling for the included x variables. For instance, $\hat{\alpha}_i$ might represent that person i tends to earn 20k more than the regression would predict from their observed characteristics alone – presumably due to some fixed ability or trait. These α_i are often not reported individually (especially if N is large), but they are important conceptually. We typically don't interpret each α_i's numeric value substantively (since they soak up unobserved factors), but we interpret their presence as having controlled for all time-invariant differences across entities.

The idiosyncratic error u_{it} now represents shocks or unobserved factors that vary within an entity over time. One can assume these u_{it} are uncorrelated with the regressors (after accounting for α_i). If that holds, the within-unit interpretation of $\boldsymbol{\beta}$ is valid as a causal or at least unbiased association. It's also worth noting that R^2 in fixed-effects models can be broken into "within" and "between" components. The within- R^2 measures how well the model explains deviations of y from entity means, whereas the between- R^2 (from a regression on entity averages) is usually not explained by the FE model except via α_i. The FE model focuses on the within-entity fit.

One must be cautious: if a regressor varies little within each entity (e.g., a slow-changing variable), its effect may be imprecisely estimated by FE because it's hard to separate from the intercept. Conversely, any effect that primarily operates across entities (between variation) cannot be detected by FE. For policy interpretation, FE coefficients imply the effect of a change in policy for a given unit relative to its baseline, rather than differences across units.

Diagnostics and Tests

Key diagnostics for fixed effects models often revolve around whether a fixed effects approach is necessary and appropriate:

- **Redundant Fixed Effects (F-test):** A standard test is to check whether the fixed effects are jointly significant. In an LSDV framework, this is an F-test of all $\alpha_i = 0$. Equivalently, compare the FE model to a pooled OLS model without α_i. This is sometimes called the **Chow test for panel data** or implemented via a pF-test. A significant result (p-value < 0.05) indicates that the fixed effects model provides a better fit (i.e., there are significant differences across entities that need to be modelled), so pooled OLS is inappropriate. If not significant, one might prefer the simpler pooled model.
- **Hausman Test (FE vs RE):** The Durbin–Wu–Hausman test is widely used to decide between fixed and random effects estimators. It tests whether the unique errors (the α_i in a random effects interpretation) are correlated with regressors. Under the null hypothesis, the random effects estimator is consistent and efficient, and FE and RE should not differ systematically. Under the alternative, RE is inconsistent (due to correlation), and FE is consistent and preferred. A Hausman test takes the difference $\widehat{\boldsymbol{\beta}}_{RE} - \widehat{\boldsymbol{\beta}}_{FE}$ and checks if it's statistically different from zero. A significant test ($p < 0.05$) suggests rejecting RE in favour of FE. In other words, it provides evidence that the fixed effect assumption (correlation present) holds, so FE is necessary. If the test is not significant, it suggests RE might be acceptable and more efficient. (One must ensure both models use the same specifications, and typically robust covariance is used for the test to handle any violations.)
- **Time Fixed Effects:** If two-way FE (adding time dummies λ_t) might be needed, one can include them and test jointly if all $\lambda_t = 0$. This checks for common time shocks. Oftentimes researchers include time dummies by default in panel FE to absorb macro-level trends. A joint test can confirm if they significantly improve the model.
- **Serial Correlation and Clustered SEs:** After FE estimation, it's common to check if residuals u_{it} are auto correlated (e.g., via Wooldridge test for autocorrelation in panel models). If so, one should use clustered standard errors at the entity level (which are robust to any within-

entity correlation and heteroskedasticity) to avoid underestimating SE's. Clustered SE's are usually a good idea in fixed-effects analysis because u_{it} often has some persistence.

- **Interpretation of R²:** It is normal to see a low overall R² in FE models even if the model is correct, because a lot of variance is due to fixed differences across entities which the FE model does not attribute to $\boldsymbol{\beta}$ (those are captured by α_i but typically not counted in the variance explained by regressors). Instead, focus on within-R² for goodness of fit in FE.

In sum, diagnostics involve confirming that fixed effects are needed (vs pooled OLS) and that one is choosing the right model (FE vs RE) and satisfying assumptions about error structure. If the Hausman test favours FE, it validates the usage of FE over RE. One should also ensure sufficient within-variation in key regressors; if not, the FE estimates could be very noisy or even inestimable (if truly time-invariant). When the FE model is appropriate, it provides a robust way to control for time-invariant confounders and yield credible estimates of the effects of interest within entities.

1.3 Random effects model

Mathematical Formulation

Random effects (RE) models provide an alternative approach to handling panel or hierarchical data structures, where the unobserved heterogeneity is assumed to be random and uncorrelated with the explanatory variables. Unlike fixed effects, which estimate a separate parameter for each unit, random effects models treat these unit-specific effects as random draws from a common distribution. This allows random effects models to use both within-unit and between-unit variation for estimation, making them more efficient (lower variance) than fixed effects when the RE assumptions hold. The model for observation i at time t is:

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_k x_{it,k} + u_i + \epsilon_{it}$$

Here, u_i is a random variable drawn from a distribution (usually assumed to be normally distributed with mean zero and variance σ_u^2) that captures the unit-specific random intercept. The RE model for panel data also acknowledges unobserved heterogeneity across entities, but treats it differently. Instead of a fixed intercept α_i for each entity, we assume each entity has a random intercept u_i drawn from a population distribution. While the β_0 is an overall intercept, u_i is the entity-specific random effect (also called random intercept), and ϵ_{it} is the idiosyncratic error. We typically assume $u_i \sim \mathcal{N}(0, \sigma_u^2) \text{and} \epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, with u_i independent of ϵ_{it} for all t. This is essentially a variance-components model: the composite error $y_{it} - \beta_0 - \boldsymbol{x_{it}}'\boldsymbol{\beta} = u_i + \epsilon_{it}$ has two components – one constant within i (u_i) and one varying within i over time (ϵ_{it}). In matrix terms, we can write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon},$$

where **b** is the vector of random effects (here just u_i for each i) and **Z** the design matrix mapping each observation to the appropriate u_i. For a balanced panel, **Z** could be thought of as block-diagonal with an $T \times 1$ column of ones for each entity i. The key difference from fixed effects is that we do not estimate a parameter for each i; instead we estimate the variance σ_u^2.

The RE model can also be seen as a special case of a linear mixed-effects model with a random intercept for each group (and possibly random coefficients, though standard RE usually only has random intercept). It's sometimes called the Error Components Model in econometrics, since u_i and

ϵ_{it} are two error components. The combined error term $v_{it} = u_i + \epsilon_{it}$ has a specific covariance structure: $\text{Var}(v_{it}) = \sigma_u^2 + \sigma_\epsilon^2$ and $\text{Cov}(v_{it}, v_{is}) = \sigma_u^2$ for $t \neq s$ (within the same i). This implies the observations are not independent within an entity (they are correlated due to the shared u_i). The degree of correlation is $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\epsilon^2}$, often called the intra-class correlation (ICC). It measures the proportion of variance in y attributable to differences between entities. When ρ is zero ($\sigma_u^2 = 0$), there is no panel effect and the model reduces to ordinary OLS; when ρ is high, within-entity observations are highly correlated.

Key Assumptions

The pivotal assumption of the random effects model is (Torres-Reyna, O. 2007):

- **Exogeneity of random effects:** The random intercept u_i is uncorrelated with all regressors $\mathbf{x}_{it}$ for all time periods. Formally, $E(u_i) = 0$. This means the unobserved factors that make entity i consistently high or low on y are unrelated to that entity's x values. This is a strong assumption; violation (i.e., if u_i correlates with x) leads to inconsistency in RE estimates. It is exactly the scenario where fixed effects would be needed.
- **Random effects distribution:** Usually u_i is assumed i.i.d. normally distributed (though in principle, the RE estimator (GLS) only needs second-moment assumptions). The normality assumption is used if one uses maximal likelihood method for estimation. For generalized least squares (GLS) estimation of β, we mainly need correct specification of $\text{Var}(u_i) = \sigma_u^2$. Often ϵ_{it} is also assumed i.i.d. normal and independent of u_i. So, the classical setup is $u_i \sim N(0, \sigma_u^2), \epsilon_{it} \sim N(0, \sigma_\epsilon^2)$, independent of each other. Under these, y_{it} given X is also normally distributed by mixture.
- **Homoskedasticity:** $\text{Var}(\epsilon_{it}) = \sigma_\epsilon^2$ same for all i, t, and $\text{Var}(u_i) = \sigma_u^2$ same for all i. Also u_i has no autocorrelation structure beyond being constant per i. Essentially, the composite error v_{it} has the variance components structure described. If there is additional heteroskedasticity or serial correlation beyond that induced by u_i, the basic RE GLS is suboptimal (one would need more complex random effects structure or use cluster-robust SEs).
- **Independence across entities:** u_i is independent of u_j for $i \neq j$, and ϵ_{it} independent of ϵ_{js} for $(i, t) \neq (j, s)$ except that within i, ϵ_{it} can correlate with ϵ_{is} only via u_i. In simpler terms, data for different entities are independent (like a random sample of entities from a population). This is often justifiable by assuming the individuals or firms were randomly selected. Indeed, one way to think of RE vs FE is: if your dataset can be seen as a random sample from a larger population of entities, RE is conceptually appropriate (with u_i as a random draw), whereas FE is more agnostic about the entities (just those specific ones).

If these assumptions hold, RE allows us to exploit both within-entity and between-entity variation for estimating $\boldsymbol{\beta}$. If the exogeneity assumption is false, RE estimates will be biased (for same reason OLS on pooled data would be biased – an omitted variable u_i correlating with x). The Hausman test directly checks this assumption.

Estimation Approach

Random effects models are typically estimated by Generalized Least Squares (GLS) or Feasible GLS. The idea is that OLS on the original model would suffer from omitted u_i, which causes serial correlation within i. GLS takes into account the covariance structure of v_{it} to produce efficient and unbiased estimates of $\boldsymbol{\beta}$. In practice:

- **GLS Derivation:** The variance of y (for a given entity i) can be written as $\Omega_i = \sigma_\epsilon^2 \mathbf{I_T} + \sigma_u^2 \mathbf{1_T} \mathbf{1_T}'$, where $\mathbf{1_T}$ is a T-vector of ones. GLS would weight observations to account for this intra-entity correlation. The GLS estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}$, where $\boldsymbol{\Omega} = \mathrm{Var}(\mathbf{Y})$ is block-diagonal consisting of each Ω_i. Since $\sigma_u^2, \sigma_\epsilon^2$ are generally unknown, in practice one uses Feasible GLS: first estimate the variance components (e.g., via an Analyses of Variance (ANOVA) method or Maximum Likelihood (ML)), then plug them in to form $\boldsymbol{\Omega^{-1}}$. Many software implementations of random effects use a variant of iterative GLS or maximum likelihood to simultaneously estimate $\boldsymbol{\beta}$ and the variances.
- **Swamy-Arora (Two-step GLS):** In older econometric literature, formulas by Swamy or Wallace-Hussain provide explicit estimates for σ_u^2 and σ_ϵ^2 (via pooling and between-group differences) which then give a GLS estimator. Modern packages usually just do ML/REML.
- **Equivalent transformation:** One can transform the data by subtracting a fraction of the entity means. Define $\theta = 1 - \sqrt{\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T\sigma_u^2}}$. It can be shown that the GLS estimator is obtained by running OLS on the transformed model: $y_{it} - \theta\bar{y}_i = \beta_0(1-\theta) + (\boldsymbol{x_{it}} - \theta\bar{\mathbf{x}}_{\boldsymbol{i}})'\boldsymbol{\beta} + (\text{transformed error})$. Here $\bar{y}_i$, where $i = 1, \dots, n$ are entity averages. If $\theta = 1$, this is the fixed-effects (within) transformation; if $\theta = 0$, it's OLS on the original data (pooled). For $0 < \theta < 1$, it partially pools the data. Thus, RE can be seen as a weighted intermediary between FE and pooled OLS. It uses some of the between-entity variation (unlike FE) but also discounts it depending on the noise. If σ_u^2 is small relative to σ_ϵ^2 (entities are nearly identical), θ is small and RE is close to pooled OLS; if σ_u^2 is large, θ is close to 1 and RE approaches FE.
- **Maximum Likelihood / REML:** One can also estimate via ML by writing down the likelihood of the data under normality of u_i and ϵ_{it}. Maximizing it yields estimates for $\boldsymbol{\beta}$, $\sigma_u^2, \sigma_\epsilon^2$. Often Restricted ML (REML) is used to get unbiased estimates of variance components. In balanced designs, ML/REML and GLS give essentially the same $\boldsymbol{\beta}$. With unbalanced panels, ML/REML is a common approach (e.g. implemented in mixed model procedures in statistical software).

Given consistent estimates of variance components, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}$ is consistent and asymptotically efficient (under assumptions). Standard errors for $\widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}$ are obtained from the covariance formula (from GLS or the inverse of Fisher information in ML). One also often reports the estimate of $\rho = \widehat{\sigma_u^2}/\left(\widehat{\sigma_u^2} + \widehat{\sigma_\epsilon^2}\right)$ to summarize how important the panel effect is.

When to Use Random Effects

Random effects models are most appropriate when you have panel/hierarchical data and you believe the unobserved differences across entities are not correlated with the regressors. This often aligns with the scenario where your entities (e.g., individuals, firms, countries) are a random sample from a larger population, and your goal is to make inferences that generalize beyond just those entities in the sample. The RE model "borrows strength" from both within- and between-entity variation. Because it does not consume degrees of freedom to estimate N separate intercepts, it can be more efficient than FE, especially when T is small and N is large.

One big advantage of RE is that you can include time-invariant covariates. Since u_i is assumed random, you do not partial out the entire between-entity variation. If a variable z_i is constant over time for each entity (like gender, race, or location), you can directly include z_i in a random effects regression and estimate its coefficient. By contrast, in FE that effect would be unidentifiable because it's absorbed by α_i. Thus, if key predictors are time-invariant and you are willing to assume exogeneity of u_i, RE is attractive. For example, in a wage panel, if you want to estimate returns to education (which may not change for many individuals after a certain age), RE would allow using the cross-person variation in education levels, whereas FE could not.

Random effects are also appropriate in multilevel/hierarchical settings, such as students within schools, where schools can be considered a random sample from a population of schools. If you assume the school-specific effect is random and independent of student-level covariates, a random intercept for schools is justified. Similarly, in longitudinal data, if you think individual-specific baseline differences are random (say, due to random sampling of individuals) and uncorrelated with included covariates, RE is an efficient approach.

However, one should use RE only if the assumption of no correlation between u_i and x_{it} is credible. In many economic or social contexts, that assumption is questionable (hence FE is often favored). If the Hausman test indicates no significant difference, it lends support to RE. Another scenario for RE is when the variation in key regressors is mostly between-entity rather than within. If X barely changes within each entity, FE would struggle to estimate β precisely (large standard errors), whereas RE can utilize the cross-sectional variation. Provided exogeneity is okay, RE will give more precise estimates in that case.

In summary, use RE for panel data when: (a) you do not find evidence of correlation between unobserved entity traits and regressors (either by subject-matter reasoning or tests), (b) you have important time-invariant variables to include, and (c) you view your entities as drawn from a larger population and want to model their effects as random. RE is also the building block for more complex mixed models (with multiple random effects), so conceptually it's a simpler special case of those.

Interpretation of Coefficients and Variance Components

The coefficient β_j in a random effects model is interpreted similarly to a regular OLS coefficient: it's the population-average effect of a one-unit increase in x_j on y, holding other variables constant. Because RE uses both within and between variation, β_j effectively represents a weighted average of the within-entity and between-entity effects (with more weight on within if σ_u^2 is small, more on between if σ_u^2 is large). If the RE assumptions hold, this distinction isn't problematic, and β_j consistently estimates the true common effect. If in truth within and between effects differ (a sign of potential model misspecification), RE would be estimating some mixture. In practice, one can compute both FE and RE to see if coefficients differ; divergence might indicate violation of RE assumptions or heterogeneous effects.

The intercept β_0 in the RE model represents the overall mean outcome when all $\mathbf{x_{it}} = 0$ and $u_i = 0$. Since u_i is zero on average, β_0 is the mean intercept across the population. Each entity has its own intercept $(\beta_0 + u_i)$, but we don't directly estimate those u_i values – instead we estimate their variance. We can obtain predicted random effects ûi after estimation (these are the BLUPs – Best Linear Unbiased Predictions – of the random intercepts). For interpretation, the variance component σ_u^2 tells us how much entities differ in their baseline levels, and σ_ϵ^2 tells us the variance of idiosyncratic shocks. The intra-class correlation $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2)$ is often interpreted as: the fraction of total variance in y attributable to differences in entities (rather than fluctuations within entities). For instance, if $\rho = 0.8$, most of the variability in y is due to persistent differences between individuals (high u_i variance), whereas if $\rho = 0.1$, most variability is within-individual over time.

Because u_i is random, how do we interpret $\boldsymbol{\beta}$ vis-à-vis an individual? One way is: β_j is the effect if we increased x_{ij} by one unit for all individuals while keeping their u_i the same. Since u_i is independent of x_i, on average individuals with higher x will have proportionally higher y. Some literature distinguishes population-averaged vs subject-specific effects in panel models: here β is subject-specific (conditional on a given u_i) effect, but due to linearity it's also equal to the population-average effect for random intercept models.

The random effects u_i themselves (if one looks at an individual entity) indicate how much entity i's expected outcome differs from the population average given its x's. If ûi is positive, that entity tends to have higher y than predicted by the fixed part $\beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta}$; if negative, lower. We usually don't over-interpret any single u_i unless we have interest in outliers; they're mainly a device to improve estimation of $\boldsymbol{\beta}$ and account for correlation. If the model is used for prediction, one could say: given the data, the best prediction for entity i's intercept is $\beta_0 + \hat{u}_i$. This is BLUP (Searle, R.S., Casella, G., McCulloch, C.E. 2006).

In summary, one can interpret $\boldsymbol{\beta}$ in RE much like OLS coefficients (effect of x on y on average) and interpret σ_u^2 and ρ to understand the structure of the data (how important are entity-level effects). If ρ is extremely low, it means differences across entities are negligible – a pooled OLS would have been fine. If ρ is high, RE model acknowledges strong entity effects; however, if those effects correlate with x, then β might be biased (which is why one would prefer FE in that case).

Diagnostics and Tests

To validate and properly use a random effects model, consider the following diagnostics and tests:

- **Hausman Test:** As noted earlier, the Hausman test is critical when deciding on RE. In the context of RE diagnostics, a non-significant Hausman test supports the RE assumption that $E(u_i) = 0$ (since FE and RE estimates are not statistically different). A significant test implies RE is giving inconsistent estimates, and one should use FE or otherwise model the correlation. Thus, always perform a Hausman test between RE and FE models for the same data. One should keep in mind the Hausman test requires FE to be consistent (which in turn requires strict exogeneity of x wrt u_{it}) and might not be valid if those assumptions fail or if one has clustered SEs (robust versions of Hausman exist). If the test favors RE (p-value high), you gain confidence in interpreting $\boldsymbol{\beta}$ from RE. If it favors FE, you likely should not trust RE results.
- **Lagrange Multiplier (Breusch–Pagan) Test for Random Effects:** This test checks whether σ_u^2 is zero. It compares a random effects model to pooled OLS (which is $u_i = 0$ case). The Breusch–Pagan LM test uses residuals from OLS to see if there's significant variance between entities. A significant test means there is evidence of panel effect (non-zero σ_u^2), so a random intercept is warranted. If not significant, one might as well use simple OLS since no evidence of random intercept. In practice, if ρ estimated is near zero, this test will reflect that.
- **Normality of Random Effects:** If one has fitted the model via ML, you might check if the distribution of estimated u_i (or BLUPs) appears roughly normal. Also check residuals ϵ_{it} for normality and homoskedasticity. Some robust methods are less sensitive to normality, but large deviations might suggest model misspecification (e.g., maybe a heavy-tailed random effect, or need a random slope). Graphical checks like Q–Q plots of u_i can be used. There isn't a simple hypothesis test for normality of random effects in standard output, but extreme skew or outliers in u_i could hint that a fixed effect (for outlier entities) might have been better, or a transformation is needed.
- **Within–Between Consistency:** One can explicitly compare the within (FE) estimates and the between (entity-average OLS) estimates. Another approach is the Mundlak approach: include the entity means $\overline{x_i}$ as additional regressors in a random effects model. If the coefficients on

$\bar{x}_i$ are significant, it indicates a difference between cross-sectional and within effects, implying correlation with u_i. Mundlak's device basically embeds a check of the RE assumption within the RE framework (if all between effects can be captured by $\bar{x}_i$, RE is effectively bias-corrected). This is a more regression-based diagnostic rather than a formal test.

- **Predicted Effects vs Observed Averages:** Sometimes one can compare the BLUP $\hat{u}_i$ to the actual entity mean of residuals. They should correlate, but the BLUP shrinks extreme values towards zero (because RE "partial pools"). If an entity's observed mean outcome is not well predicted by the model (including its random effect), that might indicate outliers or model issues.
- **Heteroskedasticity / Serial Correlation:** Standard RE assumes homoscedastic within each entity and no extra serial correlation beyond u_i. If in reality ϵ_{it} has serial correlation (maybe an AR(1) process) or variance changing over time, one might extend the random effects model to a more general GLS or use random effects with an autocorrelated error structure. Misspecified error structure can lead to inefficient estimates and misestimated SEs. Residual plots over time or an AC plot of residuals can diagnose leftover autocorrelation. If found, consider using a random effects model with a AR(1) error or using robust (clustered) standard errors to not underestimate error.
- **Outliers at entity level:** If a few entities have a very large $|u_i|$ compared to others, one wonders if those are truly random draws or if something systematic (maybe a covariate missing) differentiates them. In such cases, sometimes a fixed effect for those particular units or investigating missing variables can be considered.

In conclusion, a random effects model should always be subjected to the Hausman test to check its core assumption. Additionally, confirm the presence of significant random intercept via LM test, and inspect residuals. When assumptions are satisfied, RE provides a parsimonious and efficient way to model panel data, leveraging both within and between information while producing easily interpretable coefficients for both time-varying and time-invariant covariates.

1.4 Mixed effects model

Mathematical Formulation

Mixed effects models, also known as multilevel models or hierarchical linear models, generalize the random effects approach by allowing for both fixed effects (shared across all observations) and random effects (varying across groups or clusters). The general form of a mixed effects model is:

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \cdots + \beta_k x_{it,k} + b_{i0} + b_{i1} x_{it,1} + \cdots + \beta_{ik} x_{it,k} + \epsilon_{it}$$

where b_{i0} is the random intercept for unit i, b_{ij} is the random slope for predictor x_j for unit i, and ϵ_{it} is the residual error. Mixed effects models can include multiple levels of random effects (e.g., students within classes within schools) and allow for complex correlation structures within the data.
These models are highly flexible, enabling the analyst to capture both the average effects of covariates (fixed effects) and the variability of these effects across groups (random effects).

One of the main advantages of mixed effects models is their ability to handle unbalanced data, missing observations, and complex hierarchical structures, making them ideal for many applied research settings in the social sciences, biomedical sciences, and education. They can model both random intercepts and random slopes, which allows researchers to examine not only differences in baseline outcomes across groups but also differences in the strength of associations across groups. For example, in a study of student performance, a mixed effects model can account for differences in average performance across schools (random intercepts) and differences in how socioeconomic status affects performance across schools (random slopes).

A mixed effects model (also known as a multilevel model or hierarchical linear model) generalizes the idea of random effects by allowing multiple sources of random variation and possibly random slopes. It mixes fixed effects (population-level average effects) and random effects (group-specific deviations). The general form of a linear mixed-effects model can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ include fixed-effect coefficients as before, and $\mathbf{b}$ is a vector of random effects with $\mathbf{b} \sim \mathrm{N}(0, \mathbf{G})$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \mathbf{R})$ (Robinson, G. K. 1991). Here $\mathbf{Z}$ is the design matrix for random effects (analogous to $\mathbf{X}$ for fixed effects). The simplest case: a random intercept model is as we discussed for RE (with $\mathbf{Z}$ selecting an intercept per group). More complex: one could have random slopes, meaning some coefficients vary by group. For example, in a two-level model with individuals j nested in groups i, a random-intercept-and-slope model for a predictor x looks like:

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})x_{ij} + \epsilon_{ij}$$

Here b_{0i}, b_{1i} are random effects for group i (intercept shift and slope deviation), usually assumed $b_{0i} \sim \mathrm{N}(0, \tau_0^2), b_{1i} \sim \mathrm{N}(0, \tau_1^2)$ and possibly $\mathrm{Cov}(b_{0i}, b_{1i}) = \tau_{01}$. This model allows the relationship between x and y to vary across groups, not just the intercept. The fixed part β_1 is the average slope across groups, while b_{1i} tells how much group i's slope differs from the average. The matrix $\mathbf{Z}$ would contain the appropriate columns (e.g., a column for each random intercept and x_{ij} for each random slope). The variance components in $\mathbf{G}$ would include τ_0^2, τ_1^2, etc., and $\mathbf{R} = \sigma^2\mathbf{I}$ if we assume residuals ϵ_{ij} are i.i.d. with variance σ^2.

Mixed models can have multiple levels of nesting (e.g., students in classes in schools: random effects at class and school levels) or crossed random effects (e.g., each observation has two random factors like person and stimulus in a psychology experiment). The formulation scales to those scenarios with appropriate design matrices $\mathbf{Z}$ and structured $\mathbf{G}$.

Thus, the mixed model subsumes:

- **Random intercept models:** one random effect per group (like our earlier RE model).
- **Random intercept + slope models:** random coefficient models, where certain effects vary by group.
- **Two-way random effects, etc.:** multiple grouping factors each with random effects.

Each grouping gets its set of random effects. For instance, a two-level model has random effects b_i for each level-2 unit i. A three-level model might have b_i for level-3, c_{ij} for level-2 units within level-3, etc. The model structure should reflect the data hierarchy.

Key Assumptions

Mixed effects models share many assumptions with the simpler random effects case, but now for each random effect term:

- **Random effects are independent of covariates:** For each type of random effect, it's assumed $E(b_{qi}) = 0$. That is, the random deviations b_{qi} (for random effect q in group i) are not correlated with the covariates in the model. This is analogous to the RE exogeneity assumption but can be more complex if random slopes (one must assume, for example, that the distribution of slopes is independent of the level-1 predictors x_{ij} beyond what's captured by fixed effects). Some advanced models allow for correlated random effects and covariates (via "Mundlak" adjustments or correlated random effects models), but the standard linear multivariate model assumes independence.
- **Distributional assumptions:** The random effects **b** are usually assumed multivariate normal with some covariance matrix $\mathbf{G}$. The residuals ϵ_{ij} are assumed normal with variance σ^2 (or sometimes different variance per level-1 unit if modeling heteroskedasticity) and independent of $\mathbf{b}$. The normality of random effects might be approximately true or just a working assumption; in large samples the fixed effect estimates can still be consistent under misspecification, but for likelihood-based inference it's taken as true. Diagnostics can later assess if this was reasonable.
- **Linearity and additivity:** The model is linear in both fixed and random effects. Random effects enter linearly (e.g., random slope multiplies x_{ij} just like a fixed slope would). If relationships are fundamentally nonlinear, one might need to either transform variables or use nonlinear mixed models. Also, one assumes the correct hierarchy/nesting is specified (e.g., if data are nested, one doesn't inadvertently treat them as crossed random effects, etc.).
- **Independence given random effects:** Observations are independent conditional on the random effects. This means that once you account for shared group-level effects, any remaining correlation among observations is explained. For example, students in the same class are correlated because of the class random effect; given that effect, their residuals are independent (no further within-class correlation). If there's still structure (like spatial or temporal correlation beyond random effects), one may need to add those structures as well (e.g., an AR(1) residual structure).
- **Sufficient data for each random effect:** To reliably estimate a random effect variance, you need multiple observations per level of that effect. For example, a random intercept per person requires each person to have multiple observations; a random slope requires enough observations per group to distinguish slope variation. If not, the model might not identify those variances well (or might converge to zero variance).

Under these assumptions, the mixed model is in theory correctly specified. Violations, such as non-normal random effects or correlation with covariates, can lead to bias or incorrect inferences (though the fixed effects can be somewhat robust to moderate deviations due to BLUP shrinkage properties).

Estimation (ML / REML)

Unlike simple OLS or GLS closed-form solutions, mixed models typically require iterative algorithms to estimate parameters, especially when random slope(s) are present. Two common estimation approaches (Searle, R.S., Casella, G., McCulloch, C.E. (2006)):

- **Maximum Likelihood (ML):** We write the joint density of all responses given the model parameters $(\boldsymbol{\beta}, \mathbf{G}, \sigma^2)$. Integrating out the random effects $\mathbf{b}$ (using their normal distribution) gives a likelihood in terms of $\boldsymbol{\beta}$ and variance components. This likelihood is maximized to find estimates $\widehat{\beta}_{ML}, \widehat{G}_{ML}, \widehat{\sigma^2}_{ML}$. The ML uses all degrees of freedom (including those that would be used to estimate fixed effects) in estimating variance, which can lead to bias in variance component estimates (especially in small samples, ML tends to underestimate variance components).

- **Restricted Maximum Likelihood (REML):** REML is an approach that maximizes the likelihood of a transformed set of observations that does not depend on $\boldsymbol{\beta}$ (often done by first "differencing out" the fixed effects). It effectively accounts for the loss of degrees of freedom from estimating fixed effects, leading to less biased (often slightly larger) estimates of the variance components. REML is generally preferred for variance components, while ML is used when comparing models with different fixed effects (since REML likelihoods are only comparable between models with the same fixed structure). Many software default to REML for reporting variance components.
- **Iterative algorithms:** Common algorithms include Expectation-Maximization (EM) for variance component estimation, or Newton-Raphson and other gradient-based optimizers to solve the likelihood equations. The lme4 package in R, for instance, uses optimized Laplace approximations to maximize the likelihood, given the high dimensional integrals. The output provides $\widehat{\boldsymbol{\beta}}$ (fixed effects) and $\widehat{\sigma^2}, \hat{\tau}$ (random effects std devs) and possibly the covariance of random effects.
- **Inference:** Standard errors for $\boldsymbol{\beta}$ are obtained from the inverse Hessian of the fixed effects (or from the generalized least squares perspective as $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ is the marginal covariance of $\mathbf{Y}$). For variance components, one often uses likelihood ratio tests or approximations (because variances are on boundary of parameter space, their CI is asymmetric).

Because mixed models can be complex, verifying convergence and checking if the optimizer found a global maximum is important. It's also important to ensure the model is identifiable (e.g., not trying to fit more random effects than the data support). Sometimes simplifying the random effect structure (removing negligible components) is necessary for stability.

In summary, estimation of mixed models leverages ML/REML; the user often doesn't need to manually implement this, but should understand that unlike OLS, these are iterative and based on distributional assumptions. The result is estimates of both fixed effects (with SEs, p-values) and random effects (usually reported as variance or standard deviation of each random term, and correlation between random effects if multiple in same grouping).

When to Use Mixed Effects (Hierarchical Data)

Mixed effects models are highly useful when data have a hierarchical or clustered structure beyond a simple panel, or when you suspect random variation in some coefficients. They are appropriate in scenarios such as:

- **Multilevel/hierarchical data:** e.g., students nested in classes nested in schools (three-level model), or patients treated by doctors in hospitals, or repeated measures of patients in different clinics. Mixed models can handle multiple nested random effects (random intercepts at each level) to properly model correlations. This avoids the aggregation or omission of higher-level effects and allows one to quantify variability at each level. For example, one can estimate what proportion of outcome variation is at the student level vs class vs school.
- **Longitudinal data with individual trajectories:** If each subject is measured over time and you expect each subject to have their own intercept and slope (say growth rate), a random intercept and slope model is ideal. It acknowledges individuals start at different levels and progress at different rates. A fixed effect model could handle different intercepts (via dummy per individual), but it cannot easily allow different slopes without treating those slopes as fixed parameters too (which becomes infeasible if each individual has their own slope parameter). Mixed model treats slopes as random draws from a distribution, needing only a few parameters (mean slope, variance of slopes, etc.) rather than one parameter per individual.

- **Cases with grouped experimental units:** In experimental sciences, if experimental units are grouped (e.g., multiple measurements per animal, or plots within blocks in agriculture), mixed models account for these grouping factors as random effects, improving the generalizability of conclusions. They are often used in split-plot designs or similar where certain factors are applied at group level.
- **Unequal cluster sizes or missing data in clusters:** Mixed models naturally handle unbalanced designs (different T_i per entity, or some missing observations) without issue, under MAR (missing at random) assumptions. Traditional repeated measures ANOVA requires balanced data; mixed models do not. They use all available data and appropriately weight each cluster.
- **Relaxing independence assumption:** Unlike simple regression which assumes all observations independent, mixed models allow dependency within clusters. For example, repeated measures on the same person are likely correlated; mixed models explicitly model that correlation via random effects, thus avoiding the independence assumption for observations within a cluster. This often yields more accurate standard errors and inferences than pretending data points are independent.
- **Cross-classified or nested factors:** E.g., students are nested in schools but also in neighbourhoods, which are not nested within schools (crossed random effects). Mixed models can include a random intercept for school and another for neighbourhood to capture both effects.

In short, whenever data have multi-level structure or parameters that vary by context, and you have no reason to fix those context effects as constants, a mixed model is the appropriate choice. It reduces potential biases (compared to ignoring the structure) and usually improves efficiency by pooling information. It's essentially a compromise between complete pooling (one global model) and no pooling (separate model per group); mixed models do partial pooling: common effects plus deviations.

Interpretation of Coefficients and Random Effects

Interpreting fixed effects in a mixed model is similar to interpretation in an ordinary regression, with the understanding that these are average effects across the population or groups. For example, if β_2 is the fixed coefficient for a treatment variable, it represents the average treatment effect on y across all groups (assuming random slopes are not varying that treatment effect). It's what you expect the effect to be for a new group not in the data (hence "population effect").

If random slopes are present, β_j usually represents the mean slope and the random effect variance indicates how much individual slopes differ. So, you might say: "On average, increasing x_1 by one unit increases y by β_1, however this effect varies by group: the standard deviation of the group-specific effects is τ_1, so most groups have an effect in the range roughly [$\beta_1 \pm 2\tau_1$]." This gives a sense of heterogeneity. If the random slope variance is zero, it means the effect is consistent across groups (and the model might revert to a simpler one).

Random intercepts b_{0i} are interpreted as the deviation of group i's mean outcome from the overall mean, after controlling for x_i. They are not directly "interpreted" in the sense of hypothesis tests (since they are random draws, not fixed quantities of interest), but one can discuss the distribution of these intercepts. For instance, τ_0 (std dev of b_{0i}) might tell us that there's substantial variability between groups – e.g., a high τ_0 relative to σ implies outcomes vary more across groups than within.

The covariance between random effects (if modelled) indicates, say, whether groups with higher intercepts tend to have higher or lower slopes (a positive intercept-slope covariance would mean groups that are generally high also have a steeper increase with x). That can be substantively interesting (e.g., do higher-performing students also gain more from an intervention? If random intercept and slope for intervention are positively correlated, that's the case).

The residual ϵ_{ij} is the remaining deviation for individual j in group i after accounting for both fixed effects and that group's random effects. It's interpreted as usual residual noise or unmodeled idiosyncrasy.

Overall, mixed models allow rich interpretation: to talk about fixed effects as population-level relationships, and random effects in terms of variability and correlations at the group level. For example, in a random intercepts model, $\boldsymbol{\beta}$ are like the regression coefficients you'd get if you could somehow "average out" the differences between groups, and b_{0i} just shift each group up or down. In a random slopes model, β is the average slope, and each group i has slope $(\beta + b_{1i})$; we describe that distribution of slopes via its variance.

When communicating results, one might say: "Holding other factors constant, increasing x by one unit increases the response by β_1 on average ($p < 0.01$). However, this effect varies by group: about 95% of groups have an effect between $\beta_1 \pm 1.96\tau_1$." If random effect variability is small relative to residual, one might note that while there are group differences, most of the variation is at the individual level. Conversely, large random effect variance indicates an important grouping effect.

Diagnostics and Tests

Model checking for mixed models involves both the fixed and random parts:

- **Likelihood Ratio Tests for Random Effects:** To assess whether a particular random effect is needed, one can compare the log-likelihood (or REML criterion) of models with and without that random effect. For example, test if a random slope variance is zero by comparing the model with the random slope to one without (where that effect is fixed=common). The likelihood ratio test (LRT) is used, though one must remember that testing a variance at boundary (zero) requires a mixture chi-square distribution (a common approach is to use a conservative p by halving the usual value). A significant LRT indicates the random effect term significantly improves model fit. This is commonly done to justify complex random structure (e.g., do we need random slopes or just random intercepts?).
- **Examine Random Effects Distribution:** Extract the BLUPs for random effects and examine their distribution. Are they roughly normal (as assumed)? A normal Q–Q plot of the b_{0i} or histogram can reveal skewness or outliers. If a few groups have random intercepts far from others, maybe those groups have something special not captured by model (one could consider adding a fixed effect indicator for them if justified). If distribution is very non-normal, the inference might be affected (though fixed effects are usually robust, the random effect estimates might not be). Sometimes transforming the outcome can normalize things.
- **Residual Diagnostics:** After accounting for random effects, look at the level-1 residuals $\widehat{\epsilon_{ij}}$. Plot residuals vs fitted values to check homoskedasticity at the observation level. If residual variance seems to depend on some predictor or on group, one might need to model heterogeneous residual variance (some software allow different σ^2 per group or as function of covariates). Check autocorrelation of residuals if data are longitudinal with time order (a mixed model assumes no residual autocorrelation unless explicitly modeled; if present, consider adding an AR(1) or similar structure for R).
- **Influence and Outliers:** Identify if any data points or whole groups have undue influence. There are metrics for mixed models (like Cook's distance generalized) or one can do sensitivity: remove a group and re-fit to see if results change notably. If one group's removal

changes fixed effects a lot, maybe that group had leverage - possibly indicating a need for a group-specific fixed effect or that the random effect distribution is not capturing it well.

- **Check model fit:** For mixed models, traditional R^2 is less straightforward, but there are analogues: e.g., marginal R^2 (variance explained by fixed effects) and conditional R^2 (variance explained by fixed + random) have been proposed. While not tests, they help gauge how much variance is being captured at each level. If conditional R^2 is much higher than marginal, it means a lot of variance is explained by random effects (group structure) rather than the fixed covariates.
- **Cross-validation/prediction:** One can also assess predictive performance – do the random effects actually improve predictive accuracy (especially for new observations from known groups vs new groups)? This might be more advanced, but a well-specified mixed model should predict new data better than a naive model.
- **Hausman-type checks for random slopes:** If you have random slopes, one might do a Hausman test comparing a model where that slope is fixed vs treated as random (similar logic: if random slope is consistent). However, often one relies on the idea that if random slope covariance with random intercept is modelled, it covers potential correlation with baseline.

Given the complexity, the main formal test is usually the LRT for whether to include certain random effects. Graphical checks for normality and homogeneity at each level are critical as well. For example, a plot of group means vs group sizes might reveal if variance depends on size (which might be artifact). If any assumption seems violated, one can refine the model (e.g., add a covariate to explain a pattern, or allow a different variance structure).

Finally, ensure interpretation aligns with model: if a random slope is significant, emphasize variability in effect; if not, maybe simplify the model. Mixed models are powerful but can be overfit if too many random terms with little data per level – a diagnostic is to see if variance component estimates are zero or boundary, indicating possibly not enough info to estimate them (in which case simpler model is preferred by parsimony).

In conclusion, mixed effects models provide a comprehensive framework for regression with complex data structures, and their methodological overview spans understanding the mathematical formulation, verifying assumptions (like linearity, normality, independence of random effects) and using appropriate estimation (typically ML/REML). By choosing among linear regression, fixed effects, random effects, or fully mixed models, researchers match the model to the data structure: from a single-level OLS model to a fixed-effects within estimator controlling for latent heterogeneity, to a random effects model that assumes benign unobserved heterogeneity, up to a mixed model that captures multilevel randomness in parameters. Each has its place, and diagnostics like the Gauss–Markov conditions for OLS and Hausman tests between FE/RE help ensure the chosen model is appropriate and the results are reliable for inference at the MSc research level.

1.5 Mixed effects model with (first order) autoregressive term

Mathematical Formulation

In longitudinal or repeated-measures data, successive observations within the same experimental unit are often serially correlated in time. Ignoring such serial correlation can lead to inefficient estimates and biased standard errors. In this chapter, we extend ME model to include an autoregressive of order 1 (AR(1)) correlation structure for the residuals. It was conscious choice of the author to exclude more complex and higher order autoregressive model structures, while such would not be intuitive in the context of the specific research area of the current study. The AR(1) residual model assumes that within each cluster, errors closer in time are more highly correlated, with correlation decaying exponentially as the time lag increases.

When using notations from the previous paragraph on mixed effect models, then the term $\boldsymbol{\epsilon_i}$ is the $n_i \times 1$ vector of residual errors for cluster i. By assumption, the random effects $\mathbf{b_i}$ and residuals $\boldsymbol{\epsilon_i}$ are independent and follow multivariate normal distributions:

$$\mathbf{b_i} \sim \mathrm{N}(0,\ \mathbf{G}), \qquad \boldsymbol{\epsilon_i} \sim \mathrm{N}(0,\ \mathbf{R_i}),$$

with $\mathbf{G}$ a $q\ \times q$ covariance matrix (the random-effects covariance) and $\mathbf{R}_i$ an $n_i \times n_i$ covariance matrix for the residuals. We further assume independence across clusters, meaning b_i, $b_{i'}$, ϵ_i, and $\epsilon_{i'}$ are independent for $i \neq i'$. This implies the covariance between different subjects is zero and the total covariance matrix of all observations is block-diagonal by cluster. Under these assumptions (including b_i independent of ϵ_i), the marginal distribution of y_i is multivariate normal $\mathrm{N}(\mathbf{X_i\beta}, ; \mathbf{V_i})$ with

$$\mathbf{V_i} = \mathrm{Var}(\mathbf{Y_i}) = \mathbf{Z_i}\,\mathbf{G}\,\mathbf{Z_i'} + \mathbf{R_i},$$

since $\mathrm{Var}(\mathbf{Y_i}) = \mathbf{R_i}$ and $\mathrm{Var}(\mathbf{b_i}) = \mathbf{G}$. The matrix $\mathbf{R_i}$ captures the within-cluster residual covariance structure, which we now specify as an AR(1) structure.

AR(1) Residual Structure: To model serial correlation in the residuals, we assume that the within-cluster errors ϵ_i follow a first-order autoregressive process. In scalar form, for cluster i and time points $t = 1,2,\dots,n_i$, the residual ϵ_{it} satisfies the recursion:

$$\epsilon_{it} = \rho\,\epsilon_{i,t-1} + \eta_{it}, \qquad \text{for } t = 2,\dots,n_i,$$

where $|\rho| < 1$ is the autoregressive parameter and η_{it} is an independent innovation term (white noise) with $\eta_{it} \sim \mathrm{N}(0, ; \sigma^2)$. By construction η_{it} are uncorrelated over time within each cluster (and also independent across clusters), and act as the new information or shock at time t. The AR(1) equation implies that each residual is correlated with the immediately preceding residual by the factor ρ. Repeated substitution shows ϵ_{it} can be expressed in terms of past innovations: for example, $\epsilon_{i3} = \rho\epsilon_{i2} + \eta_{i3} = \rho^2\epsilon_{i1} + \rho\eta_{i2} + \eta_{i3}$, and so on. Under the AR(1) model, the correlation between residuals h time units apart is ρ^h. Intuitively, ρ describes the persistence of the residuals: if ρ is close to 1, an error at time t carries over strongly to $t + 1$ (high autocorrelation), whereas if ρ is near 0, the errors are essentially uncorrelated from one time to the next. A negative ρ would indicate an alternating pattern (one timepoint above expectation tends to be followed by next timepoint below expectation if $\rho < 0$). In most applications, we expect $0\ <\ \rho <\ 1$ for positive serial correlation in longitudinal data.

One assumes the process is stationary, meaning the statistical properties of ϵ_{it} do not depend on absolute time, only on intervals. Consequently, we take the variance of ϵ_{it} to be constant over time (homoscedastic), denoted $\mathrm{Var}(\epsilon_{it}) = \sigma^2$ for all t. For the first observation in each cluster (at $t = 1$), one can treat ϵ_{i1} as an initial draw from the stationary distribution of the AR(1) process. Under stationarity, ϵ_{i1} is simply $\mathrm{N}(0, \sigma^2)$ in our parameterization (since we have set the marginal variance at each time to σ^2). This yields a consistent covariance structure for all time points (McCulloch, C.E., Searle, R.S. 2001).

Estimation via ML and REML

The parameters to be estimated in a mixed model with AR(1) errors include: the fixed effects $\boldsymbol{\beta}$; the variance components in $\mathbf{G}$ (e.g. random effect variances and any covariances); and the residual variance and autocorrelation (σ^2, ρ) in $\mathbf{R_i}$. Let $\boldsymbol{\theta}$ denote the full set of variance/correlation parameters (everything except $\boldsymbol{\beta}$). Estimation is typically performed by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML), using iterative optimization since closed-form solutions are generally not available for θ. We outline the likelihood formulation and the two approaches:

Maximum Likelihood (ML): Under the Gaussian assumptions, the joint density of all observations $\mathbf{Y} = (\mathbf{Y}_1', \dots, \mathbf{Y}_m')'$ can be written in terms of the model parameters. Because clusters are independent, the log-likelihood is a sum of cluster-specific contributions. For a given parameter vector $(\boldsymbol{\beta}, \boldsymbol{\theta})$, the log-likelihood (ML) is:

$$\mathbf{l}(\boldsymbol{\beta}, \boldsymbol{\theta}) \ = \ -\frac{1}{2}\sum_{i=1}^{m}[n_i \ln(2\pi) + \ln|\mathbf{V_i}(\boldsymbol{\theta})| + (\mathbf{Y_i} - \mathbf{X_i}\boldsymbol{\beta})'\mathbf{V_i}(\boldsymbol{\theta})^{-1}(\mathbf{Y}_i - \mathbf{X_i}\boldsymbol{\beta})].$$

This is the log-likelihood of a multivariate normal with mean X_i and covariance $\mathbf{V_i} = \mathbf{Z_i G Z_i'} + \mathbf{R_i}(\boldsymbol{\rho}, \boldsymbol{\sigma^2})$. The ML estimates are obtained by maximizing $\mathbf{l}(\boldsymbol{\beta}, \boldsymbol{\theta})$ with respect to both β and θ. In practice, this optimization is done in two stages conceptually: for any trial values of θ, the likelihood is maximized w.r.t. β by the generalized least squares (GLS) solution

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\sum_i \mathbf{X}_i' \, \mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_i \mathbf{X}_i' \, \mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{Y}_i\right),$$

since the log-likelihood is quadratic in β. Substituting $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ back into l yields the profile likelihood $\mathbf{l}(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ that depends only on θ. The remaining variance parameters θ (including ρ) are typically estimated by numerical optimization of this profile log-likelihood. The optimization algorithms used can be Newton-Raphson, Fisher scoring, or derivative-free methods (Pinheiro, J.P, Bates, D.M., 2000). The result of ML estimation is $\hat{\theta}_{ML}$ and $\hat{\beta}_{ML} = \hat{\beta}(\hat{\theta}_{ML})$.

While ML yields consistent and asymptotically efficient estimates, a known issue is that ML can produce biased estimates of variance components in finite samples (tending to underestimate them, because it does not account for the loss of degrees of freedom from estimating β). This motivates Restricted Maximum Likelihood (REML) estimation.

Restricted Maximum Likelihood (REML): REML is an alternative likelihood-based approach that maximizes a likelihood after removing the information about the fixed effects. In essence, REML operates on a transformed set of data that are free of the fixed effects, focusing only on the variance parameters. One way to understand REML is that it maximizes the likelihood of certain linear combinations of the observations (often called error contrasts or residuals) that depend only on random effects and errors, not on β. Algebraically, the REML criterion can be derived by integrating the fixed effects out of the likelihood or by projecting the data onto the orthogonal complement of the column space of X. The REML log-likelihood can be written as:

$$\begin{aligned}\mathbf{l}_{\mathbf{REML}}(\boldsymbol{\theta}) \ = \ -\frac{1}{2}[(N - p)\ln(2\pi) + \ln|\mathbf{V}(\boldsymbol{\theta})| + \ln|\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}| \\ + \mathbf{Y}'(\mathbf{V}(\boldsymbol{\theta})^{-1} - \ (\boldsymbol{\theta})^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1})\mathbf{Y}],\end{aligned}$$

where $N = \sum_i n_i$ is the total number of observations and p is the number of fixed-effect parameters. The additional term $\ln\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right|$ (and the $(N - p)\ln(2\pi)$ term instead of $N\ln(2\pi)$) distinguishes REML from ML. Maximizing $\mathbf{l}_{\mathbf{REML}}(\boldsymbol{\theta})$ gives the REML estimate $\widehat{\boldsymbol{\theta}}_{\mathbf{REML}}$. Notably, $\boldsymbol{\beta}$ is typically estimated after θ in REML, usually by the same GLS formula using $\mathbf{V}(\widehat{\boldsymbol{\theta}}_{\mathbf{REML}})$. REML estimators of variance components are in many cases less biased (approximately unbiased in finite samples) than ML estimators. In fact, REML can be shown to produce unbiased estimators of variance in balanced linear mixed models and is generally preferred for variance inference.

Both ML and REML are asymptotically normal, and standard errors for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ can be obtained from the observed Fisher information matrix or via bootstrapping if needed. In summary, ML uses the full likelihood of y (treating $\boldsymbol{\beta}$ as any other parameter), whereas REML maximizes a likelihood that has eliminated $\boldsymbol{\beta}$ as a nuisance parameter. One practical consequence is that -2 log-likelihood values from REML cannot be directly compared between models with different fixed effects, since the REML criterion depends on $\mathbf{X}$. Model comparisons for fixed-effects (e.g., testing whether certain $\boldsymbol{\beta}$ are zero) should thus be made using ML, while comparisons involving only variance structures can use REML.

In the context of our AR(1) residual model, the choice of ML vs REML typically affects the estimates of $\mathbf{G}$, σ^2, and ρ, but not the fitted $\boldsymbol{\beta}$ (except very slightly through the difference in $\widehat{\boldsymbol{\theta}}$). The AR(1) correlation parameter ρ is usually of primary interest for describing the error process, and REML provides a more reliable estimate of ρ and σ^2 especially when m or N is not very large. Several authors (e.g., Pinheiro, J.P, Bates, D.M., 2000), advocate REML for variance component estimation in mixed models, noting its unbiasedness and better performance in simulations, unless ML is required for comparing fixed effect structures.

Model Assumptions

When incorporating an AR(1) residual structure in a mixed model, we make several important assumptions that ensure the validity of the model and the interpretations:

- **Linearity and Correct Specification:** The mean structure $\mathbf{X}_\mathbf{i}\boldsymbol{\beta} + \mathbf{Z}_\mathbf{i}\mathbf{b}_\mathbf{i}$ is assumed to be correctly specified. That is, the fixed effects $\mathbf{X}_\mathbf{i}$ adequately capture the population-average trends and the random effects $\mathbf{Z}_\mathbf{i}\mathbf{b}_\mathbf{i}$ capture the appropriate cluster-specific deviations. Any systematic trends over time should be included in $\mathbf{X}_\mathbf{i}$ (e.g. time effects, time-varying covariates) so that ϵ_{it} can be treated as a zero-mean random fluctuation.
- **Normality:** We assume the random effects b_i are multivariate normally distributed and independent of the residual errors ϵ_i, which are themselves multivariate normal. Normality of ϵ_i implies that marginally y_i is normal. Moderate deviations from normality may not severely bias β estimates (due to the Central Limit Theorem), but inference for θ relies on normal theory. If needed, transformations or robust methods should be considered if residuals appear grossly non-normal (this is usually checked in diagnostics).
- **Homoscedasticity:** The AR(1) residual structure in our formulation assumes constant residual variance over time (homogeneous σ^2). This is part of the stationarity assumption – the variance of ϵ_{it} does not depend on t. In some cases, one might extend to a heterogeneous AR(1) model (often denoted ARH(1)) where the variance at each time can differ while retaining AR(1)-like correlations. Here we assume homogeneity for simplicity; any large differences in variability at different times would violate this assumption and suggest using a variance function or heterogeneous covariance model.
- **Autocorrelation Structure:** We assume an autoregressive of order 1 form for within-cluster correlation. This means only the immediately preceding residual influences the current one directly. In other words, after conditioning on $\epsilon_{i,t-1}$, there is no additional correlation with

$\epsilon_{i,t-2}$ beyond that already induced via $\epsilon_{i,t-1}$. The resulting implied correlations ρ^h for lag h define a very specific decay pattern (geometric decay. We assume this pattern is correct. If the true residual correlation decays faster or slower than geometric, or has a different form (e.g. an oscillatory or moving-average pattern), then an AR(1) model may be misspecified. In practice, AR(1) is a reasonable choice when we expect a smooth decline in correlation with time lag and no long-range correlations beyond what an AR(1) would generate.

- **Stationarity:** As noted, $|\rho| < 1$ is required for a stationary covariance structure. Stationarity implies the correlation between ϵ_{it} and $\epsilon_{i,t+h}$ depends only on h (time difference) and not on t itself. It also implies the process has no unit root (i.e. we are not modeling a trend or random walk in the residuals) – any persistent time trends should be part of the fixed or random effects. We assume the data were collected under conditions that make stationarity plausible (e.g. steady-state or no structural breaks in the measurement process). If ρ were estimated very close to 1, that might indicate a violation of stationarity (near unit-root behavior) or insufficient data length to distinguish it; one should be cautious in interpreting such cases.
- **Independent Clusters:** Observations from different clusters (subjects) are independent. This is a standard multilevel model assumption. It means that we do not expect any residual correlation across different individuals or experimental units. The grouping factor that defines the clusters is presumed to partition the data into mutually independent blocks. In longitudinal studies, this is usually valid if different subjects do not influence each other. If there is reason to believe in cross-subject correlation (e.g. related individuals, shared environment), the model would need to be extended (for instance, by including higher-level random effects or a cross-correlation term), but that is beyond our current scope.
- **Random Effects Independence:** The random effects b_i are assumed independent across i, and also uncorrelated with the residuals within the same cluster. The latter was already stated, and it ensures that $Z_i b_i$ captures all the between-subject consistent variability, while ϵ_i captures independent random noise around the subject-specific trajectory. If this assumption is violated (say, if there were feedback where the random intercept and AR(1) error are correlated), the model likelihood would be misspecified. Standard mixed model theory relies on b_i and ϵ_i being independent.
- **Equal Spacing (for AR(1) interpretation):** Many implementations of AR(1) assume measurements are taken at equal time intervals. If the actual observation times are irregular, the model is assuming that it is the order of observations (1,2,3,...) that matters and treats successive observations as if evenly spaced in a unit time step. For moderately varying intervals, this may be a reasonable approximation, but if intervals vary widely, a continuous-time correlation model (or specifying the actual time differences in a spatial-power correlation structure) would be more appropriate. In our development, we treat time as indexed by an integer t, effectively assuming equal spacing or that $\rho^{|t-s|}$ is a suitable correlation for the given lag irrespective of actual gap length.
- **Cluster Size and Missing Data:** We assume each cluster i has a series of n_i measurements. If some data are missing, the model can still be applied (maximum likelihood estimation under missing at random assumptions will be valid as long as the correlation structure is correctly specified). The covariance R_i is defined for whatever times are observed for cluster i (it will be a smaller Toeplitz matrix if n_i is smaller). We assume that any missingness does not violate stationarity or independence assumptions (typically assuming missing at random so that the likelihood inference remains valid, as discussed in standard texts).
- **No Additional Residual Structure:** Aside from the AR(1) correlation, we assume no other residual patterns (e.g. no seasonal effects or periodicity in residuals, which would violate the AR(1) form). We also assume the innovation variance σ^2 is constant; if heteroscedasticity (non-constant residual variance) is present (e.g. increasing measurement error over time), one might need to incorporate a separate variance function or use ARH(1). For our purpose, the residuals η_{it} have common variance σ^2 and are independent across t and i.

These assumptions collectively define the model's validity. It is always good practice to perform diagnostics to check if these assumptions are approximately met, which we will discuss later. Authoritative sources like Diggle, Heagerty, Liang & Zeger (2002) provide extensive discussion on covariance structures and the importance of stationarity and proper model specification.

Interpretation of Model Parameters

After fitting a mixed model with AR(1) residuals, we interpret the parameters in a manner similar to standard mixed models, with additional insight into the correlation parameter:

- **Fixed Effects (β):** The fixed-effect coefficients β represent the population-average effects of the covariates on the response. For example, if β_1 is associated with a treatment indicator, β_1 is the estimated average treatment effect on the response (difference in mean response between treatment groups), holding other covariates constant. Importantly, in the presence of the AR(1) residual correlation, the interpretation of β is still as a marginal (population-level) effect, not to be confused with subject-specific instantaneous effects (the random effects adjust each subject's trajectory but β moves the overall mean). Because we have modeled the residual correlation explicitly, the fixed effects can be interpreted as if each observation provides effective information according to the correlation structure. For instance, standard errors for $\hat{\beta}$ will be larger than in an independent model if ρ is large, reflecting that there is less independent information in closely spaced observations. Nonetheless, β estimates themselves are typically similar to those from a model that ignores correlation (provided that model is generalized least squares), but their precision is correctly accounted for by our model. In summary, a fixed effect coefficient is the expected change in the response for a unit change in that predictor, averaged over the population, and conditional on the random effects being zero (baseline) for that cluster.
- **Random Effects (b_i) and G:** The random effects b_i represent cluster-specific deviations from the population mean trajectory. For example, if we include a random intercept b_{0i} for subject i, then b_{0i} indicates how much subject i's overall response level differs from the average response (with $b_{0i} \sim N(0, \sigma_b^2)$). Similarly, a random slope b_{1i} for time would mean subject i has its own slope in addition to the average slope β_1. The covariance matrix G contains the variance of each random effect and their covariances (if multiple random effects per cluster). These can be interpreted as measures of between-cluster variability. For instance, $\sqrt{\mathrm{Var}(b_{0i})}$ is the standard deviation of the random intercepts across subjects – a large value implies subjects vary widely in their baseline levels. A covariance like $\mathrm{Cov}(b_{0i}, b_{1i})$ would indicate whether clusters with higher intercepts tend to have higher or lower slopes. The presence of b_i in the model usually captures long-term or systematic differences among clusters, whereas the AR(1) residuals capture short-term correlations of the noise. In our model, the BLUP (Best Linear Unbiased Predictor) of b_i after fitting gives the estimated random effect for each cluster (e.g., each subject's intercept and slope adjustment). These b_i estimates, combined with fixed effects, yield the subject-specific fitted curves. We interpret random effects cautiously: they are not fixed parameters but latent variables – however, their estimated variance components tell us about population heterogeneity. For example, if σ_b^2 (random intercept variance) is small relative to σ^2 (residual variance), it means most of the variation is within subjects over time rather than persistent differences between subjects.
- **Residual Variance (σ^2):** The parameter σ^2 is the variance of the innovation term η_{it} (and also the variance of ϵ_{it} at each time, in our model formulation). It quantifies the within-cluster noise after accounting for both fixed effects and random effects, and after considering correlation. In practical terms, σ is the standard deviation of the unpredictable fluctuation in the outcome from one time point to the next (for a given subject). A larger σ^2 indicates more

erratic residual behavior (less consistency even after accounting for autocorrelation and random effects). It's worth noting that σ^2 here is analogous to the residual variance in an ordinary linear model, but interpreted at the level of the innovation process; because of correlation, the total variability of an observation also depends on neighboring residuals through ρ. However, σ^2 is the variance of the one-step-ahead prediction error, essentially.

- **Autoregressive Parameter (ρ):** The AR(1) parameter ρ is key to understanding the temporal correlation in the data. It represents the correlation between residuals one time unit apart within the same cluster (lag-1 autocorrelation). If $\rho = 0$, the model reduces to one with independent residuals (no autocorrelation). If $\rho = 0.5$, it means the correlation between adjacent residuals is 0.5 (moderate persistence), and between residuals two time units apart it would be $\rho^2 = 0.25$, etc. A value of ρ close to 1 indicates strong positive autocorrelation: residuals remain similar from one measurement to the next, implying a highly smooth residual process (perhaps the outcome changes very slowly over time after accounting for fixed/random effects). On the other hand, a negative ρ (e.g. $\rho = -0.3$) would imply that if one residual is positive, the next tends to be negative, creating an oscillating pattern around the fitted trajectory. In many biological and social applications, we expect ρ to be positive because most processes have inertia (e.g. if you are above average today, you might still be above average tomorrow). In terms of model fit, ρ close to 1 means observations within cluster convey redundant information (many repeated measurements give only marginal new information), whereas ρ near 0 means each new time point is almost like a new independent observation (given the random effects).
- **Total Variability and Marginal Correlation:** It can also be useful to interpret the combined effect of random effects and residual correlation on the marginal correlations of the response. For example, if we have a random intercept with variance σ_b^2 and AR(1) residuals with ρ, the marginal correlation between two observations y_{it} and y_{is} (for $t \neq s$ in the same cluster) can be derived from V_i. It will be $\mathrm{Cor}(y_{it}, y_{is}) = \frac{\sigma_b^2+\sigma^2\rho^{|t-s|}}{\sqrt{(\sigma_b^2+\sigma^2)(\sigma_b^2+\sigma^2)}}$ (for a random intercept model).

  This shows that at lag 1, correlation comes from both sharing the same intercept and the AR(1) linkage, whereas at longer lags the AR(1) contribution decays. The random intercept provides a baseline correlation to all pairs (i.e. compound symmetry component), whereas ρ gives an additional correlation that is highest for adjacent times and decreases for distant times. Practically, one might report $\widehat{\sigma_b}$ and $\hat{\sigma}$ and $\hat{\rho}$ to characterize the correlation: e.g., "Between-subject standard deviation was 5 (random intercept), residual day-to-day SD was 2, and residual autocorrelation was 0.6," which immediately communicates that there is substantial subject heterogeneity and also a strong autocorrelation in daily fluctuations.

In summary, the fixed effects tell us the average trend, the random effects tell us how much individuals or clusters deviate in systematic ways, and the AR(1) ρ tells us how the random fluctuations are structured in time (the memory of the error term). All these parameters together give a rich picture of the data structure. It is often useful to complement numerical estimates with plots: for instance, one could simulate or plot the implied correlation function ρ^h to visualize how quickly correlation drops off, or examine individual subject fitted trajectories versus the population trajectory to see the role of b_i. Several sources provide further discussions on interpreting linear mixed model parameters in longitudinal settings, emphasizing the distinction between population-averaged effects and subject-specific effects.

Diagnostic Techniques for AR(1) Mixed Models

After fitting a mixed-effects model with AR(1) residuals, it is crucial to assess whether the model assumptions hold and whether the AR(1) structure is indeed appropriate. The following diagnostic techniques and checks are recommended:

- **Residual Autocorrelation Function (ACF) Plot:** A primary diagnostic for the AR(1) assumption is examining the autocorrelation of residuals. One typically computes the sample autocorrelation function of the model residuals within clusters. There are different types of residuals one can use: raw residuals $y_{it} - \widehat{y_{it}}$, or normalized (Pearson) residuals which account for the variance structure. In practice, plotting the ACF of raw residuals by cluster and averaging them, or directly the Pearson residuals' ACF, can reveal remaining correlation. Under a correctly specified AR(1) model, the residuals should show little to no autocorrelation beyond lag 1, because the model has accounted for it. If the AR(1) model is adequate, the sample ACF of residuals will have no significant spike at lag 1 (or a much reduced one compared to a model without AR(1)) and no systematic pattern at higher lags. If instead one still see a slowly decaying ACF or significant correlations at lags 1, 2, etc., this suggests the AR(1) structure might not have fully captured the correlation. For example, if a fitted AR(1) model still exhibits a strong lag-1 residual ACF, it could mean that either ρ was under-estimated or the correlation structure is different (e.g. an ARMA or other pattern). Autocorrelation checks can be formalized with statistical tests (Ljung-Box test for whiteness of residuals), but graphical ACF is usually more informative to detect model inadequacy. We might also examine the partial autocorrelation function (PACF) of residuals; for a true AR(1) process, the PACF should show a sharp cutoff after lag 1. If the residual PACF decays more gradually, it might hint at a higher-order AR process needed.
- **Residual vs. Time Plots:** Plotting residuals over time (for each cluster or pooled) can help check for any non-stationarity or time-varying variance. Under the AR(1) model, we assumed stationarity and homoscedasticity. A residual-vs-time plot should show residuals fluctuating around 0 with no obvious trend and roughly constant spread over time. If there are, for instance, increasing spread of residuals as time increases, that violates the constant σ^2 assumption (suggesting a need for a time-dependent variance function). If you see a systematic oscillation or seasonality in residuals, that violates the AR(1) assumption (maybe an AR(2) or periodic model would be needed). Also, check if residuals at the beginning vs end of the series have similar behavior; sometimes the first observation can have a different distribution under certain AR(1) treatments, but in our approach we set it to the same σ^2 so it should not stand out.
- **Normal Q-Q Plots for Residuals and Random Effects:** Normality assumptions can be checked by Q-Q (quantile-quantile) plots of the residuals and of the empirical BLUPs of random effects. The residual Q-Q plot (especially of standardized residuals) should ideally follow a straight line if residuals are normal. Moderate departures (mild tails) are often acceptable, but severe deviations might require a different error distribution or a transformation of the response. For random effects, one can look at the distribution of b_i estimates (BLUPs) – they should appear roughly normal (centered at 0). Given the relatively small number of clusters often, this is a weaker check, but any gross outliers in random effects might indicate an outlier cluster or a missing covariate that explains that cluster's deviation.
- **Model Fit Criteria and Comparison:** One can compare the AR(1) model to alternative models using information criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion), or likelihood ratio tests (if models are nested). For instance, compare the AR(1) residual model to a simpler model with independent residuals (i.e. $\rho = 0$) or to a more complex model (like ARMA(1,1) or an unstructured covariance). A substantially lower

AIC for the AR(1) model versus the independent model would confirm that capturing autocorrelation improved the model. Conversely, if the AR(1) model has similar AIC to the independent model, $\hat{\rho}$ might be essentially zero and the AR(1) structure unnecessary. Authors like, Pinheiro, J.P, Bates, D.M., (2000), suggest using likelihood ratio tests to test $\rho = 0$ (though strictly, testing on the boundary $\rho = 0$ is nonstandard, one can treat it cautiously). If the data are very short series, sometimes estimating ρ is difficult and one might not see a big improvement.

- **Lag Plots or Variograms:** A lag plot of residuals (plotting $\epsilon_{i,t}$ vs $\epsilon_{i,t-1}$) can visually reveal autocorrelation: points should cluster around a line with slope ρ. After fitting, if one computes the normalized residuals (which ideally are uncorrelated if model is correct), a lag-1 residual plot should show no pattern. Another tool is the **variogram** of residuals, which plots semi-variance against lag. For a correctly specified AR(1), the semi-variance (which is roughly $\frac{1}{2}E\left(\epsilon_{i,t} - \epsilon_{i,t+h}\right)^2$) should increase with lag in the pattern consistent with the covariance ρ^h. Empirical variograms can be overlaid with the model-implied variogram to see if they match (Pinheiro, J.P, Bates, D.M., (2000) discuss this approach for evaluating correlation models).
- **Checking Stationarity Assumption:** While strict stationarity is hard to test with short sequences, one can check if the implied correlation structure is consistent with the data. For example, if you estimate separate empirical correlations for each lag (say, compute sample correlations of residuals at lag 1, 2, 3 from the data without assuming a model), do they roughly follow a ρ, ρ^2, ρ^3 pattern? If not – for instance, if empirical lag 2 correlation is higher than ρ^2 – the AR(1) model might be inadequate and a Toeplitz with different lag-1 and lag-2 correlations might fit better. Diggle et al. (2002) advocate looking at such empirical summaries as a guide. Another symptom of non-stationarity would be if residuals in early time periods have systematically different variance than later periods; this can be checked by splitting the series and computing variances.
- **Influential Points and Outliers:** Temporal correlation models can sometimes be sensitive to outliers (an extreme outlier can induce spurious autocorrelation). It's important to identify if any single observation or cluster is overly influencing the estimate of ρ. One approach is to refit the model excluding a suspect outlier or cluster and see if $\hat{\rho}$ changes markedly. Alternatively, examine studentized residuals: any residual that is extreme (say >3 in absolute value) might distort the correlation estimate. Ideally, the residuals after fitting AR(1) should all be relatively small and homogeneous.

In summary, thorough residual analysis and model comparison are vital for verifying the AR(1) mixed model. Authoritative texts (e.g., Diggle, P.J. 2002) illustrate checking fitted vs empirical correlations. These diagnostics ensure that our rigorous theoretical model translates into an accurate representation of the data.

The inclusion of an AR(1) residual structure makes the mixed model substantially more flexible for longitudinal data, capturing short-term correlations that random effects alone cannot. This leads to more accurate standard errors and often a better fit, as evidenced by improvements in likelihood criteria when AR(1) is justified. However, it also requires careful checking of assumptions; an AR(1) model is parsimonious and will be mis specified if the true correlation is more complex.

2. Practical implementation of chosen methods

The relatively long outcome prediction horizon, of one year, should enable usability of traditional prediction techniques, like regression model in forecasting stock returns. At the same time inclusion of one company more than once, which is typical to panel survey, raises the need analyse possible panel effects. Such is investigated via the fixed, random and mixed effects modelling. Finally, possible autoregression is modelled with auto-regressive parameter of the first order.

2.1 Time series and input variables to be analysed

The analyses is carried out on all listed companies in the main markets of the Tallinn, Riga and Vilnius Stock Exchange. All-together, 16 companies from Estonia, 3 of Latvia and 13 from Lithuanian were analysed.

The total list of Estonian companies inluded Arco Vara, Baltika, Coop Pank, Ekspress Grupp, Eften Real Estate, Harju Elekter, LHV Grupp, Merko Ehitus, Nordecon, Pro Kapital Grupp, PRFoods, Silvano Fashion Grupp, Tallink Grupp, Tallinna Kaubamaja, Tallinna Sadam and Tallinna Vesi. For Latvia, enterprises SAF Tehnika, Hansa Matrix and Olainfarm were analysed. The companies included for Lithuania were Apranga, Auga group, Grigeo, Ignitis grupė, Klaipėdos nafta, Linas Agro Group, Novaturas, Panevėžio statybos trestas, Pieno žvaigždės, Rokiškio sūris, Šiaulių bankas, Telia Lietuva and Vilkyškių pieninė.

In order to increase the number of observations, each company is included up to 7 times into analyses:
1) using data from 2024 to predict returns during 2025 (until March);
2) using data from 2023 to predict returns during 2024;
3) using data from 2022 to predict returns during 2023;
4) using data from 2021 to predict returns during 2022;
5) using data from 2020 to predict returns during 2021;
6) using data from 2019 to predict returns during 2020;
7) using data from 2018 to predict returns during 2019.

To counterbalance for shorter, i.e., 3 months outcome period of 2025, the 0.25 weight is assigned to observations from given period (instead of usual 1). Also, the actual return of those 3 months is annualized as $((1 + 3\,m\,return)^4 - 1)$, i.e. adjusted approximately 4 times higher.

Based on the author experience and the conclusions from data analyses, the decision was taken to exclude from the final model construction dataset such observations, where company made accounting loss during the given calendar year. Often, companies in financial difficulties, go through substantial re-organizations, which together with volatility in share prices, introduces major obstacles in using traditional statistical modelling techniques. Also, as current thesis carries the practical aim, of building a model, which could help an average investor while making long term investment decisions, then focus on profitable companies seems reasonable.

The set of the input variables, used in the analyses, covered all main ones used by the industry:

Revenue Size and Revenue Growth estimate the basic ability of the company to generate sales. Typically, analysts compare ability to generate sales within activity sector, as multiples might not be comparable otherwise. Revenue growth is especially important in the current inflationary environment.

Income and Profitability assess the ability of the company to generate profit. Sales without any profit could be adequate only in start up phase, but in longer horizon company needs to generate profit in order to survive and attract investors. There are different profitability ratios used by analysts, but one of the most commonly used, is the return-on-equity (ROE), which estimates size of income to size of equity.

Price to Equity (P/E) is the ratio for valuing a company that measures its current share price relative to its per-share earnings (EPS). The price-to-earnings ratio is also sometimes known as the price multiple or the earnings multiple. P/E ratios are used by investors and analysts to determine the relative value of a company's shares in an apples-to-apples comparison. It can also be used to compare a company against its own historical record or to compare aggregate markets against one another or over time. P/E may be estimated on a trailing (backward-looking) or forward (projected) basis (www.investopedia.com).

Price to Sales (P/S) ratio ratio is a valuation ratio that compares a company's stock price to its revenue. It is an indicator of the value that financial markets have placed on each dollar of a company's sales or revenues. Like all ratios, the P/S ratio is most relevant when used to compare companies in the same sector. A low ratio may indicate the stock is undervalued, while a ratio that is significantly above the average may suggest overvaluation (www.investopedia.com).

Price to Book Value (P/B) ratio compare a firm's market capitalization to its book value. It's calculated by dividing the company's stock price per share by its book value per share (BVPS). An asset's book value is equal to its carrying value on the balance sheet, and companies calculate it netting the asset against its accumulated depreciation. Book value is also the tangible net asset value of a company calculated as total assets minus intangible assets (.e.g. patents, goodwill) and liabilities. For the initial outlay of an investment, book value may be net or gross of expenses, such as trading costs, sales taxes, and service charges. Some people may know this ratio by its less common name, the price-equity ratio (www.investopedia.com).

Industry categorizes the main activity of the company. During the analyzes, based on the average return and logic behind different sectors, eventually four homogeneous categories were introduced:

1) Financial services;
2) Telecommunication, Manufacturing, Agriculture and Real Estate development;
3) Trade and Tourism.

Competition estimate barriers in entering the segment and relative intensity of competition. Typical low competition sectors are financial services (as a result of minimum capital requirement, supervisory approval etc), utilities (often natural monopolies), telecommunication (capital intense, limited network capacity), but also few manufacturing areas (like pharmaceutical and precise electronics). Very high international competition is encountered in production and sale of clothes. All other sectors were assessed as average competition.

Management estimate both the competence of the management team in the respective area of activity, but also historical success and track record from leading and carrying out different business ideas. The track record is estimated as "above average", "average" or "below average". The author fully comprehends the ambiguity and certain subjectivity of the variable but still believe it as very informative. Also, to keep the subjectivity low, the "above" and "below" are only used in case of sufficient evidence.

Country flag illustrates the stock market on which the respective company is registered. It is typically, but not always, also the country where majority of entity's income is generated. Country indicator is valuable input to capture the differences in both economical, but also in investment culture. One latest example of reasons behind the differences, originates from 2021, when second pillar was made voluntary in Estonia and lot of respective funds inflated local stock market.

In model construction it was decided to aggregate, based on the low number of observations available for Latvia, the Latvian and Lithuanian entities into one and Estonia as separate category. Such differentiation is supported among other by several economical studies. Researcher's have found both the corruption level (please see corruption index published by Transparency International at https://www.transparency.org/en/cpi/2021) and also shadow economy as being lowest in Estonia compared to Latvia and Lithuania (please see research by Stockholm School of Economics at https://www.sseriga.edu/shadow-economy-index-baltic-countries). Also, as additional consideration, the share of population investing into stocks is highest in Estonia. Altogether, the relation between the input and the output variables should vary between Estonia and other two countries.

Relative Return estimates the return of an individual company, above or under the average return on the respective stock exchange for the same period. Typically, companies with high returns, deliver constantly above market average (and vice versa). In order to increase the differentiating power of the variable, it was categorized between 3 categories: being able to deliver below -20%, in between or above +35% compared to market average during last 2 previous years.

Market cycle indicator illustrates the expected average return in outcome period. The parameter is not differentiated across the three markets at least for the model construction period, as it is somewhat difficult to imagine different market cycles in three Baltic economies. One could, ideally, construct separate ARIMA model to predict the cycle. However, it is also possible to build the forward-looking view, based on the latest available return estimate and publicly available macro forecasts (e.g., by ECB, please refer to https://economy-finance.ec.europa.eu/economic-surveillance-eu-economies_en).

Figure 1. Average return on Baltic Stock Exchange during 2022 (OMXTGI – Tallinn, OMXRGI – Riga, OMXVGI – Vilnius)

Respective variable is categorized into four buckets, with values from „0“ to „3“. The weakest, value

„0“, indicate negative returns below -5% on average and was assigned to, e.g., predicted outcome of 2022 in EE/LV (please see Figure 1 above). The strongest possible value „3“, was assigned to, e.g., year 2021 outcome, when average returns were above 15% (please see Figure 2 below). Years 2019 and 2020 were both estimated as „2“ while average returns 5-15%. It is important to emphasize, however, that the value of respective variable is not of great importance in the practical use of the model (until countries are not differentiated). Investor should just pick certain amount of companies, from the ones for which the highest returns predicted, while staying invested in the stock market across at all times. As for any regression model, one needs to take into account, that respective models are strong in rank ordering of observations, but cannot help in capturing the average risk correctly.

Figure 2. Average return on Baltic Stock Exchange during 2021 (OMXTGI – Tallinn, OMXRGI – Riga, OMXVGI – Vilnius)

2.2 Linear regression model

The linear regression is one of the most widely used statistical methods for practical applications (function „lm“ in R). Firstly, all the variables are run through the stepwise method (function „stepAIC“ in R), both the forward and the backward direction, in order not to miss any important predictors. As next step, the regression model is built trying to follow the results from stepwise analyses and utilizing only significant variables. The model below was concluded to fit the data the best:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                -29.270     15.667  -1.868 0.063596 .
factor(Management)2         14.248      7.691   1.853 0.065825 .
factor(Management)3         27.169      8.343   3.257 0.001383 **
P_B                         -7.716      3.628  -2.127 0.035019 *
factor(Industry)Other      -13.501      8.055  -1.676 0.095712 .
```

```
factor(Industry)Trade    -11.631     10.388  -1.120 0.264556
factor(Market_cycle)1     19.901      9.237   2.154 0.032736 *
factor(Market_cycle)2     36.355      9.642   3.770 0.000231 ***
factor(Market_cycle)3     49.621      9.831   5.047 1.24e-06 ***
factor(Rel_return_2y)2    17.471      9.483   1.842 0.067315 .
factor(Rel_return_2y)3    29.019     13.633   2.129 0.034854 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.77 on 156 degrees of freedom
Multiple R-squared:  0.2974, Adjusted R-squared:  0.2524
F-statistic: 6.605 on 10 and 156 DF,  p-value: 1.702e-08
```

The Adjusted R-Square of 0.2524 can be considered reasonable for stock market analyses, as stock market movements are typically more difficult to predict compared to more stable paradigm, like e.g., the credit scoring. The stochastic component and the market speculations will, always, play a part in the stock market movements. The AIC and the BIC were assessed to be 1651 and 1688 respectively, which are both impact by adjusted variance of 3 months outcome period in last year.

Final model includes variables of P/B ratio, industry of main activity, market cycle, management experience and relative return during previous 2 years. It can be seen, that most influential differentiation between the companies, is driven by variables of market cycle, management quality, P/B and relative return over market average. The industry of main activity plays somewhat a smaller role within the final model.

It was somewhat surprising for the author, that irrespectively of the several try's, the final model did not include the P/E ratio, which is one of the most used estimates by the industry. The balance sheet based P/B ratio, turned out as much more accurate predictor of the return.

As next, let's analyse the residuals of the model predictions. Just to recall that, because of using differently weighted observations for 2025 outcome, one should analyse weighted residuals rather than unweighted. As R software produces standardized residuals already as weighted (by the square root of the respective weight), then output for standardised residuals was analysed.

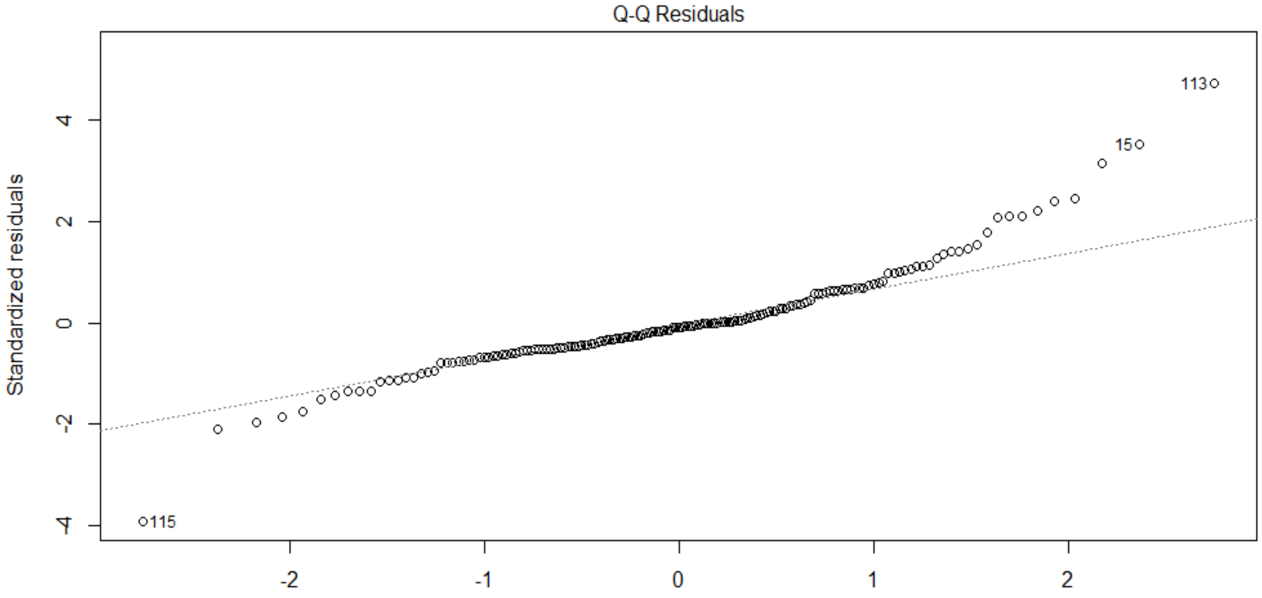

Figure 3. Standardized residuals for linear regression model

When analysing the model residuals, it can be seen, that three observations stand out:

1) observation number 113, which is SAF Tehnika return of 169% during 2021, while reorganized since last crises business model finally paid off;
2) observation number 15, which is Coop Pank return of 156% in 2021, where brake through of relatively new company coincided with second pillar pension reform in Estonia;
3) observation number 115, which is SAF Tehnika return of -66% during 2023, while sales in North America declined rapidly.

As was also expected, the model struggles to capture such extreme values, partly driven by subjectivity or changes in company business model.

Rather than pulling the observations from all four years together, one should analyse accuracy of the predictions within different years. Such is reasoned by the practical use of the model, while the investment decision can be made only at the concrete time point (typically latest). Theoretically, one can also try to capture the market cycle and withdraw investments just before the downturn, but only few investors typically manage that. Therefore, let's assume that model user wants to stay invested at all times. Table 1, below, lists across last four outcome periods the eight companies, for which highest returns were predicted by the model for given year, and compare to rank of actual return.

Model Rank	**For 2025**	**Return Rank**	**For 2024**	**Return Rank**	**For 2023**	**Return Rank**	**For 2022**	**Return Rank**
1	Coop Pank	13	Šiaulių bankas	4	SAF Tehnika	23	Šiaulių bankas	11
2	LHV Group	7	Vilkyškių pieninė	1	Coop Pank	15	Vilkyškių pieninė	2
3	Ekspress Grupp	11	Coop Pank	21	Ekspress Grupp	17	Rokiškio sūris	6
4	Šiaulių bankas	5	Ekspress Grupp	24	Vilkyškių pieninė	6	Linas Agro Group	1
5	Eften Real Estate	14	Ignitis grupė	8	LHV Group	8	Ignitis grupė	10
6	Harju Elekter	21	Rokiškio sūris	6	Šiaulių bankas	10	Apranga	3
7	Vilkyškių pieninė	6	Linas Agro Group	10	Eften Real Estate	12	Grigeo	17
8	TKM Grupp	12	LHV Group	17	TKM Grupp	7	Enefit Green	4
Sum		89		91		98		54

Table 1. Ranking of the companies by the linear regression model

First of all, it stands out, that model proposes to invest into relatively fixed set of companies. Names of two companies can be found in each year's list (Šiaulių bankas and Vilkyškių pieninė). Additonally, eight companies are found at least in two year lists out of four (Coop Pank, LHV Group, Ekspress Group, Eften Real Estate, TKM Group, Ignitis grupė, Rokiškio sūris and Linas Agro Group). Remianing five companies end up in lists only once (Harju Elekter, SAF Tehnika, Apranga, Grigeo and Enefit Green).

Secondly, it stands out, that model captures market movements pretty close also in downturn year like 2022, when returns turn negative. Such is logical, given model construction data included both upward and downward market cycle. Also, model variables include set of variables which should work in both. In market upward trend, the growth and historical returns above market average drive the return, while in downward years the multiples in valuations and management experience plays are large role. One needs to keep in mind, that model user is expected to stay invested at all times therefore, stable returns across different market cycles would be beneficial.

2.3 Fixed effects model

In general, the ANOVA and regression analyzes are built on assumption that input variables indicate fixed effects, i.e., they represent adequately the values found in the whole population. Random effect variables at the same time, represent only a certain set of possible variable values found in the whole population, and as such carry a bias. It is not a trivial exercise to determine on possible random effects, and typically also somewhat linked to the final use of the model. For example, if the model built on companies at Baltic stock exchange, is implemented in practice to predict the returns for US companies, then one could most certainly argue that country flag is indicating random effects rather.

In this chapter the fixed effects model is tested on the data set (function „plm“ in R). Model frame is set by company name and observation year. As such, one expects the fixed effects model also to construct company specific intercepts. Firstly, we let model to indicate fixed effects variables, using function („within“):

```
Coefficients:
                        Estimate Std. Error t-value  Pr(>|t|)    
P_B                     -26.2750     7.2299 -3.6342 0.0004003 ***
factor(Market_cycle)2    19.6412    10.2501  1.9162 0.0575345 .  
factor(Market_cycle)3    42.1391    11.2547  3.7441 0.0002708 ***
factor(Rel_return_2y)2   20.5405    12.0060  1.7109 0.0894926 .  
factor(Rel_return_2y)3   26.3017    16.7629  1.5690 0.1190687    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    197330
Residual Sum of Squares: 124060
R-Squared:      0.20089
Adj. R-Squared: -0.020404
F-statistic: 8.69005 on 5 and 130 DF, p-value: 3.9399e-07
```

It can be seen from model output that like expected, then only the variables, which values vary in time were included as model inputs. As a result, low number of regressors are included: price to book value, relative return in 2 years and market cycle. Also, the R-Square value ends up lower (0.20089). That does not mean necessarily, however, that model is poor. As an additional factor, each company carries a distinguishable return level, i.e., fixed effect. Let's also analyze individual intercepts assigned within groups (different companies).

```
                  Apranga                  Arco Vara                 AUGA group 
                23.203528                   0.542159                 -33.581005
                Coop Pank          EfTEN Real Estate             Ekspress Grupp
                28.112538                  -4.706920                 -11.695107
             Enefit Green                     Grigeo                HansaMatrix
                14.248819                  12.986191                  28.778343
            Harju Elekter              Ignitis grupė            Klaipėdos nafta
                -2.183506                 -11.258050                  -8.355727
                LHV Group           Linas Agro Group               Merko Ehitus
                55.277084                 -16.689951                  37.928683
                 Nordecon                  Novaturas                  Olainfarm
                24.665080                 -25.205047                  -4.660301
Panevėžio statybos trestas           Pieno žvaigždės                    PRFoods
               -43.370741                  17.843405                 -60.928709
        Pro Kapital Grupp             Rokiškio sūris                SAF Tehnika
                -2.730307                 -20.303098                  36.490184
    Silvano Fashion Group             Šiaulių bankas              Tallink Grupp
                 0.047111                   3.210579                  -7.146830
```

```
Tallinna Kaubamaja Grupp           Tallinna Sadam            Tallinna Vesi
               15.822325                -1.105236                22.147574
           Telia Lietuva        Vilkyškių pieninė
               41.465333                28.713038
```

High historical returns in Estonian banking sector (LHV Group and Coop Pank), but also Lithuanian telecom (Telia Lietuva), are clearly indicated by high fixed effects also. However, as market has not appreciated Lithuanian banking sector, until 2025, then the fixed effect calibration for Šiaulių bankas is low. The market has been historically clearly questioning the business idea and economic strength of PRFoods.

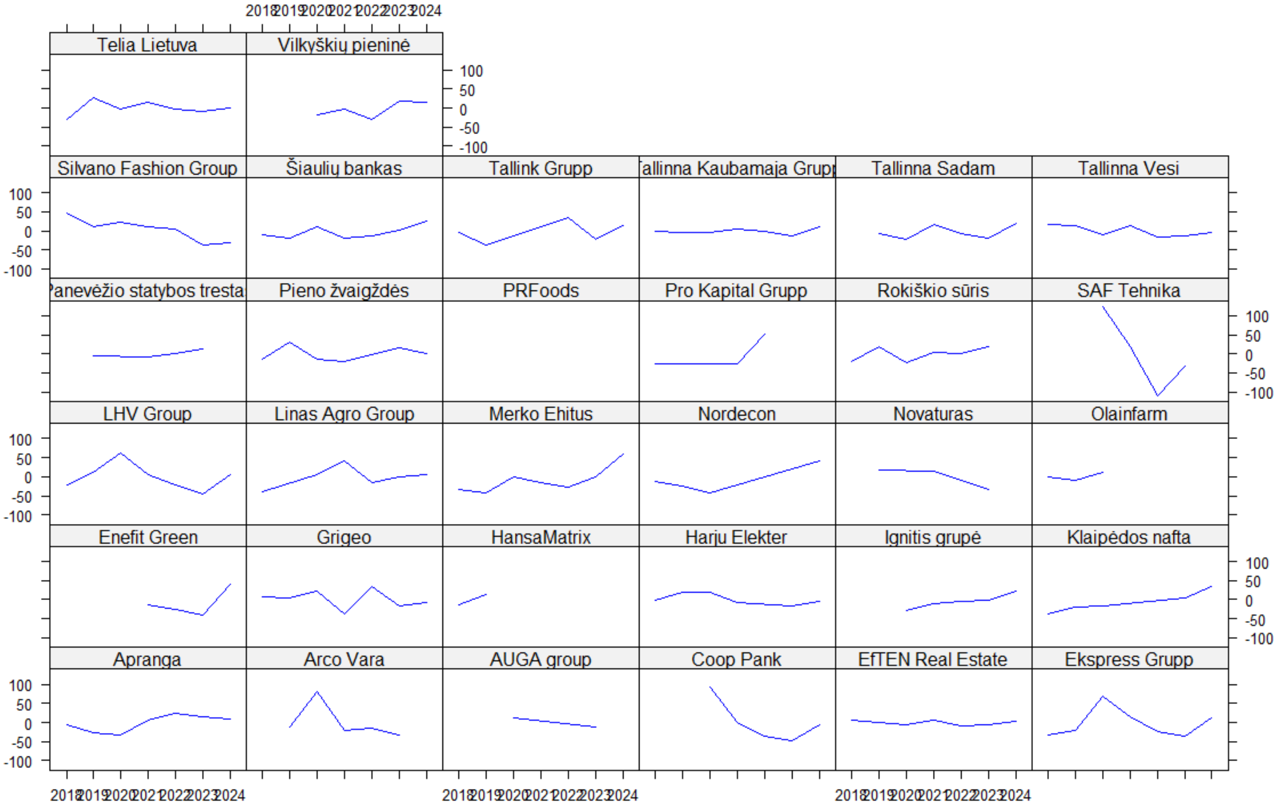

Figure 4. Residuals for fixed effects model

Similarly to linear regression model (please see Figure 3 and added reasoning), there are no extensive outliers other than SAF Tehnika and Coop Pank. As such, compared to the linear regression model above, the fixed effects model does not capture extremes any better. Let’s check next statistically if the fixed effects model should be preferred.

As was reasoned in chapter 1.2, there is typically F-test used, to test if the fixed effects are jointly significant. A significant result (p-value < 0.05) indicates that the fixed effects model provides a better fit, so pooled OLS is inappropriate. In case of comparing ordinary regression model and fixed effect model, constructed in chapter 2, the p-value was estimated as 0.0021. According to p-value the fixed effects add significant explanatory power.

At the end, one needs to consider both statistical and practical aspects of possible random effects. As next, let's analyse ranking of the companies by the model. Table 2, below, lists for last four outcome periods the eight companies, for which highest returns were predicted by the model for given year.

Compared to ordinary regression model, the results are less stable across years, i.e., number of different companies ending up in the table of predicted highest returns is larger.

Model Rank	**For 2025**	**Return Rank**	**For 2024**	**Return Rank**	**For 2023**	**Return Rank**	**For 2022**	**Return Rank**
1	Merko Ehitus	1	Merko Ehitus	2	SAF Tehnika	23	Vilkyškių pieninė	2
2	Nordecon	2	Vilkyškių pieninė	1	Vilkyškių pieninė	6	Enefit Green	4
3	Enefit Green	3	LHV Group	17	Merko Ehitus	5	Apranga	3
4	Klaipėdos nafta	4	Coop Pank	21	Coop Pank	15	Grigeo	17
5	Vilkyškių pieninė	6	Šiaulių bankas	4	Grigeo	2	Merko Ehitus	9
6	LHV Group	7	Silvano Fashion Grupp	22	LHV Group	8	Šiaulių bankas	11
7	Šiaulių bankas	5	Grigeo	12	Pro Kapital Grupp	1	Ignitis grupė	10
8	Ignitis grupė	8	Enefit Green	23	Šiaulių bankas	10	Linas Agro Group	1
Sum		36		102		70		57

Table 2. Ranking of the companies by the linear fixed effects model

Names of three companies can be found in each year's list (Merko Ehitus, Vilkyškių pieninė and Šiaulių bankas). Additionally, six companies are found at least in two year lists out of four (Enefit Green, LHV Group, Ignitis grupė, Coop Pank, Grigeo and SAF Tehnika). Remaining six companies end up in lists only once (Nordecon, Klaipėdos nafta, Silvano Fashion Grupp, Pro Kapital Grupp, Apranga and Linas Agro Group). When adding up the ranks of companies with highest predicted returns, the fixed effects model seems to outperform the regression model in all years, except for 2022 outcome.

The market cycle typically lasts for 10 years, out of which 1-2 years are correction or downturn years. During the market correction, other regressors than usual growth and profitability levels start to play an important role. One obvious example is management quality. As fixed effects model doesn't include management quality, it seems to somewhat struggle in correction years. Still, on cycle average and especially in good years, the fixed effects model seems to outperform the ordinary regression model. Also, it seems logical that company specific fixed effects are empowered in good years, while differences in return widen and continue for longer period. In downturn, however, the company specific corrections are typically quick and unique.

To conclude on the fixed effects model, then even so the F-test didn't suggest strong benefit from fixed effect model, then the analyses carried out on individual Baltic stock companies still suggested clear benefits, at least in good years of the market cycle.

2.4 Random effects model

In this chapter the random effects model is tested on the data set (function „plm“ in R). Again, the model frame is set by company name and observation year. As such, one expects the random effects model also to construct company specific random intercepts, additionally to observation specific intercepts (please also see 1.3). Firstly, we let model to indicate random effects variables, using function ("random"):

```
Effects:
                  var std.dev share
idiosyncratic 1122.3    33.5     1
individual       0.0     0.0     0
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
      0       0       0       0       0       0

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-107.319  -14.598   -2.662    0.808   12.640  129.784 

Coefficients:
                         Estimate Std. Error z-value  Pr(>|z|)    
(Intercept)              -29.2702    15.6668 -1.8683  0.061721 .  
factor(Management)2       14.2478     7.6906  1.8526  0.063936 .  
factor(Management)3       27.1693     8.3425  3.2567  0.001127 ** 
P_B                       -7.7162     3.6283 -2.1267  0.033446 *  
factor(Industry_1)Other -13.5013     8.0549 -1.6762  0.093709 .  
factor(Industry_1)Trade -11.6314    10.3878 -1.1197  0.262835    
factor(Market_cycle)1    19.9012     9.2371  2.1545  0.031201 *  
factor(Market_cycle)2    36.3545     9.6420  3.7704  0.000163 ***
factor(Market_cycle)3    49.6211     9.8314  5.0472 4.484e-07 ***
factor(Rel_return_2y)2   17.4708     9.4826  1.8424  0.065417 .  
factor(Rel_return_2y)3   29.0190    13.6326  2.1286  0.033283 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    245910
Residual Sum of Squares: 138290
R-Squared:      0.27897
Adj. R-Squared: 0.23275
Chisq: 66.0455 on 10 DF, p-value: 2.5563e-10
```

Variables included into model are similar to ordinary regression model. Also, the adjusted R-squared is similar (0.23275). However, the company specific variance is negative (limited to zero), indicating correlation between random coefficients u(i) and regressors x(i), and as a result no need for random effects.

Just in case, let’s also test with the „Hausman test“, if the fixed effects prevail over random ones. As expected, also the Hausmann test indicates with p-value = 0.005394, that the fixed effects model shall be preferred to random effects one.

```
Hausman Test

data:  formula_w
chisq = 12.675, df = 3, p-value = 0.005394
alternative hypothesis: one model is inconsistent
```

Secondly, let’s run “Lagrange multiplier test“ to test if panel effects exist. The test concludes that panel effect is insignificant and as such, again, ordinary linear regression is preferred.

```
Lagrange Multiplier Test - (Breusch-Pagan)

data:  formula_w
chisq = 2.6879, df = 1, p-value = 0.1011
alternative hypothesis: significant effects
```

One can conclude that random effects model is not a viable solution for chosen modelling paradigm. Such can be also considered intuitive. While the existence of fixed effects could be easily linked to limited number of regressors in the model and general market perception about the specific company, then it would be difficult to intuitively explain the existence of random effects across the different companies.

2.5 Mixed effects model

In this chapter the linear mixed effects model is tested on the data set (function „lmer“ in R). The most important feature of the mixed model is, that both fixed and random variables can be handled within the same model. The fixed and the random variables were selected based on the statistical testing carried out in chapter 2.3. and 2.4. One should not conclude purely from the fact that random effects model could not be parametrized on given dataset, that it is also not possible to estimate the mixed effects model. As a result of additional random slope variables or otherwise different set of regressors, there could occur a situation for specific data set, that it is possible to parametrize mixed effects model but not the random effects one.

Finally, the model as below was concluded to fit the data the best:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Name     (Intercept)   0       0.00
 Residual             843      29.03
Number of obs: 167, groups:  Name, 32

Fixed effects:
                       Estimate Std. Error t value
(Intercept)             -44.136     12.499  -3.531
P_B                      -7.074      3.483  -2.031
factor(Market_cycle)1    20.744      8.977   2.311
factor(Market_cycle)2    36.944      9.384   3.937
factor(Market_cycle)3    50.383      9.575   5.262
factor(Rel_return_2y)2   18.219      9.236   1.973
factor(Rel_return_2y)3   31.489     13.188   2.388
factor(Management)2      14.706      7.265   2.024
factor(Management)3      28.686      7.910   3.627

Correlation of Fixed Effects:
            (Intr) P_B    f(M_)1 f(M_)2 f(M_)3 f(R__2)2 f(R__2)3 fc(M)2
P_B         -0.402
fctr(Mrk_)1 -0.564  0.171
fctr(Mrk_)2 -0.555  0.162  0.801
fctr(Mrk_)3 -0.496  0.128  0.763  0.744
fctr(R__2)2 -0.419 -0.042 -0.192 -0.201 -0.177
fctr(R__2)3 -0.357 -0.093  0.033  0.034  0.023  0.640
fctr(Mngm)2 -0.344 -0.035  0.054  0.068  0.001 -0.191   -0.140
fctr(Mngm)3 -0.274 -0.147  0.086  0.132  0.066 -0.254   -0.268    0.745
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

Similarly to random effects model in chapter 2.4, there is clear evidence of missing random effects. Even the inclusion of additional random slope variables did not support evidence of random effects.

2.6 Mixed effects model with autoregressive term

In current chapter the linear mixed effects model together with autoregressive parameter is tested on the data set (function „lme“ in R). One important feature of such a model is that each reoccurring observation (company in our case), is assigned with individual intercept to accompany the autoregressive term. Such intercept enables company specific parametrization of the autoregressive parameter. The simplest, i.e., first order autoregressive function is tested. Following the final conclusions in chapters 2.4 and 2.5 on random effects, all regressive input variables were treated as fixed variables (except for the autoregressive term), in order to increase the accuracy of the model. The model as below was fitted on the data:

```
Linear mixed-effects model fit by REML
  Data: Magistri_andmed_xls
       AIC      BIC    logLik
```

```
  1593.813 1636.511 -782.9067

Random effects:
 Formula: ~1 | Name
        (Intercept) Residual
StdDev:  0.00229179 29.88671

Correlation Structure: ARMA(1,0)
 Formula: ~Year | Name
 Parameter estimate(s):
     Phi1
0.0913845
Variance function:
 Structure: fixed weights
 Formula: ~1/Weight
Fixed effects:  Return_2 ~ P_B + factor(Market_cycle) + factor(Rel_return_2y) +
factor(Management) + factor(Industry_1)
                            Value Std.Error  DF   t-value p-value
(Intercept)             -27.95842 16.301330 129 -1.715101  0.0887
P_B                      -8.80907  3.867336 129 -2.277813  0.0244
factor(Market_cycle)1    20.45793  9.087315 129  2.251263  0.0261
factor(Market_cycle)2    37.11017  9.580292 129  3.873595  0.0002
factor(Market_cycle)3    50.24835  9.534333 129  5.270253  0.0000
factor(Rel_return_2y)2   16.74339  9.562427 129  1.750956  0.0823
factor(Rel_return_2y)3   26.78109 13.732509 129  1.950197  0.0533
factor(Management)2      14.74295  8.173022  27  1.803855  0.0824
factor(Management)3      28.65886  8.861983  27  3.233910  0.0032
factor(Industry_1)Other -14.37876  8.720267  27 -1.648890  0.1108
factor(Industry_1)Trade -12.38049 11.209967  27 -1.104418  0.2792
 Correlation:
                        (Intr)  P_B      f(M_)1  f(M_)2  f(M_)3  f(R__2)2  f(R__2)3
fc(M)2 fc(M)3 f(I_1)O
P_B                     -0.375
factor(Market_cycle)1   -0.465  0.195
factor(Market_cycle)2   -0.452  0.185  0.799
factor(Market_cycle)3   -0.413  0.155  0.757  0.740
factor(Rel_return_2y)2  -0.362 -0.040 -0.182 -0.193 -0.171
factor(Rel_return_2y)3  -0.336 -0.092  0.021  0.019  0.009  0.639
factor(Management)2     -0.334 -0.068  0.036  0.053 -0.003 -0.170   -0.104
factor(Management)3     -0.318 -0.160  0.069  0.115  0.060 -0.226   -0.217    0.751
factor(Industry_1)Other -0.583  0.100  0.049  0.031  0.044  0.045    0.099    0.041
0.119
factor(Industry_1)Trade -0.475 -0.037 -0.014 -0.018  0.011  0.026    0.112    0.218
0.242  0.691

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.52157239 -0.47464811 -0.07923229  0.42659093  4.37933664

Number of Observations: 167
Number of Groups: 32
```

One can recognize in final model similar variables as were applicable also to final linear regressive model (please see chapter 2.2 for further details). Just that model intercept and the weights of the variables are slightly different as a result of autoregressive term added. Similarly to linear regression model, the most influential differentiation between the companies, is driven by variables of market cycle, management quality, P/B and relative return over market average. The industry of main activity plays somewhat a smaller role within the final model.

Similar is concluded by the Anova test on variable level, which suggest excluding as minimum the activity sector from final model:

```
Analysis of Deviance Table (Type II tests)

Response: Return_2
                       Chisq Df Pr(>Chisq)
P_B                   5.1884  1    0.02274 *
factor(Market_cycle) 37.6378  3  3.372e-08 ***
```

```
factor(Rel_return_2y)  4.2351  2    0.12033
factor(Management)    11.3506  2    0.00343 **
factor(Industry_1)     2.7213  2    0.25650
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As a result, the final model is reparametrized without company activity as one of the regressors:

```
Linear mixed-effects model fit by REML
  Data: Magistri_andmed_xls
       AIC      BIC    logLik
  1604.704 1641.455 -790.3518

Random effects:
 Formula: ~1 | Name
        (Intercept) Residual
StdDev: 0.002330659 29.91303

Correlation Structure: ARMA(1,0)
 Formula: ~Year | Name
 Parameter estimate(s):
      Phi1
0.07449832
Variance function:
 Structure: fixed weights
 Formula: ~1/Weight
Fixed effects:  Return_2 ~ P_B + factor(Market_cycle) + factor(Rel_return_2y) +
factor(Management)
                           Value Std.Error  DF   t-value p-value
(Intercept)            -43.71727 13.106684 129 -3.335494  0.0011
P_B                     -7.92447  3.768753 129 -2.102677  0.0374
factor(Market_cycle)1   21.14900  9.106254 129  2.322470  0.0218
factor(Market_cycle)2   37.51822  9.595310 129  3.910058  0.0001
factor(Market_cycle)3   50.86403  9.598004 129  5.299438  0.0000
factor(Rel_return_2y)2  17.60861  9.558724 129  1.842150  0.0677
factor(Rel_return_2y)3  29.49357 13.651572 129  2.160452  0.0326
factor(Management)2     15.10543  7.812310  29  1.933542  0.0630
factor(Management)3     30.00149  8.504205  29  3.527842  0.0014
 Correlation:
                       (Intr) P_B    f(M_)1 f(M_)2 f(M_)3 f(R__2)2 f(R__2)3 fc(M)2
P_B                    -0.414
factor(Market_cycle)1  -0.554  0.181
factor(Market_cycle)2  -0.548  0.174  0.799
factor(Market_cycle)3  -0.487  0.145  0.758  0.741
factor(Rel_return_2y)2 -0.418 -0.045 -0.186 -0.196 -0.174
factor(Rel_return_2y)3 -0.343 -0.094  0.023  0.022  0.010  0.639
factor(Management)2    -0.362 -0.035  0.054  0.069  0.002 -0.179   -0.132
factor(Management)3    -0.288 -0.146  0.082  0.129  0.063 -0.241   -0.255    0.739

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.69298131 -0.50526276 -0.06172051  0.50912525  4.23069833

Number of Observations: 167
Number of Groups: 32
```

All remaining regressors of the model are found significant also by the Anova test:

```
Analysis of Deviance Table (Type II tests)

Response: Return_2
                        Chisq Df Pr(>Chisq)
P_B                    4.4213  1   0.035494 *
factor(Market_cycle)  37.5201  3  3.571e-08 ***
factor(Rel_return_2y)  5.0266  2   0.081002 .
factor(Management)    13.4384  2   0.001208 **
```

The autoregressive parameter (Phi1) of the model is only 0.074, which indicates very low positive correlation and leads as a result to very low autoregressive impact (see respective ACF plot on Figure

5). Such does raise concern if additional complexity of the model compared to, e.g., OLS model is justified.

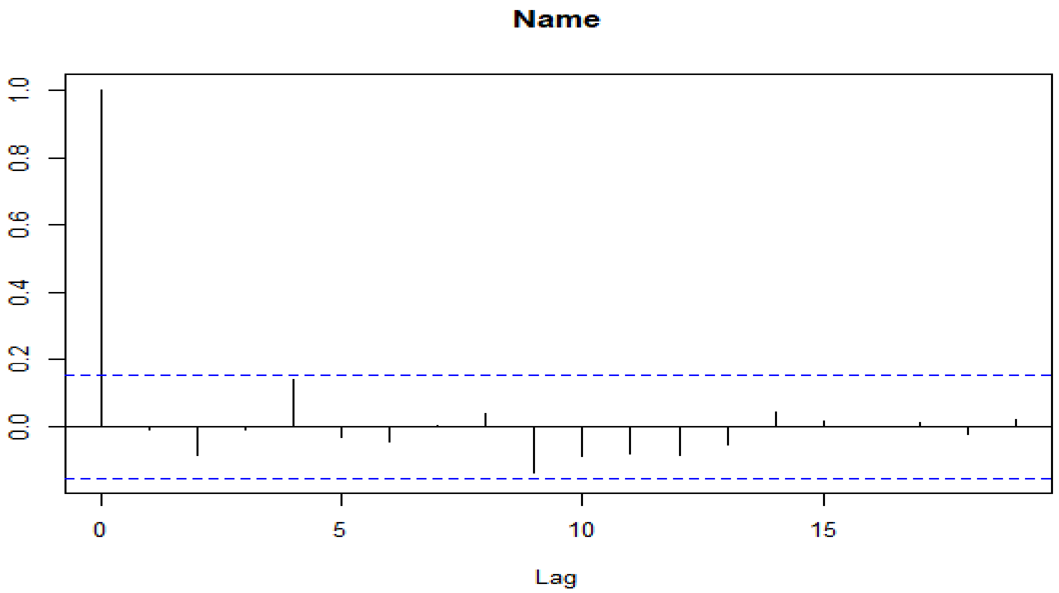

Figure 5. Autocorrelation Function (ACF)

The AIC and the BIC criteria were found to be 1605 and 1641 respectively, which are comparable to linear regression model. As next, let's analyze the residuals of the model predictions. As R software produces standardized residuals already as weighted, then output for standardized residuals was analyzed.

When analyzing the model residuals, it can be seen, that three observations stand out with high positive residuals. Those are similar as for the linear regression model, i.e., Coop Pank and SAF Tehnika returns. As very small weight assigned to autoregressive part of the model, the individual levels of residuals are very similar to regression model.

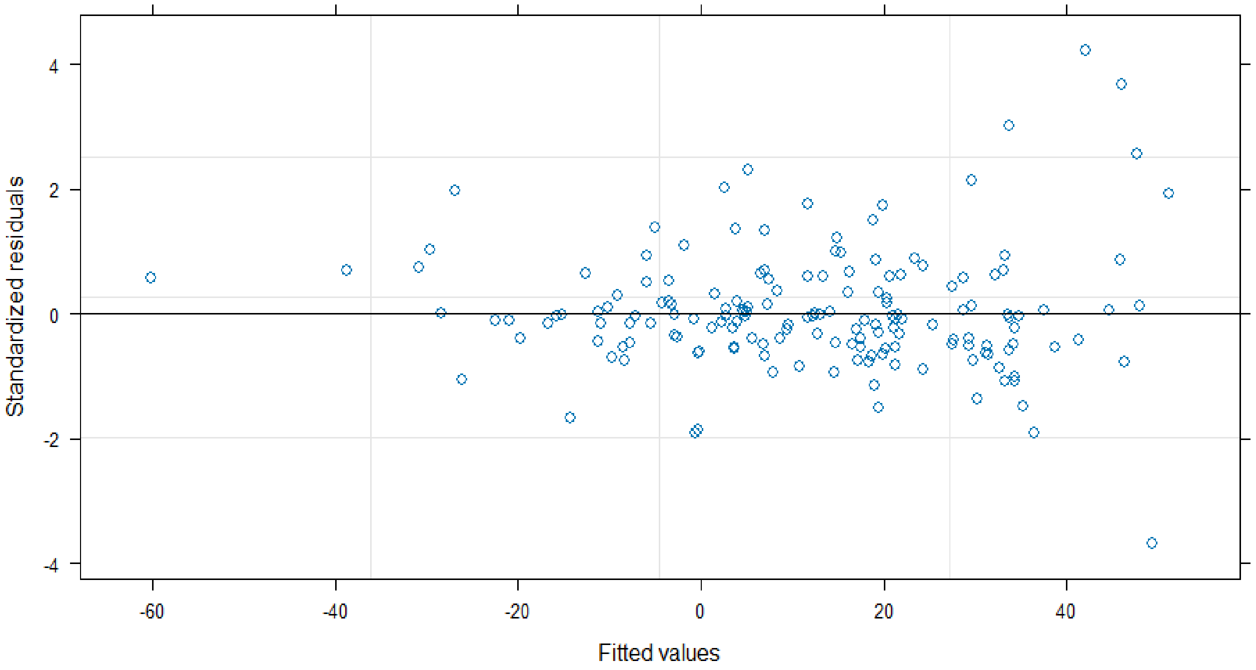

Figure 6. Standardized residuals for linear mixed effects model with autoregressive term

As next, let's analyze ranking of the companies by the model. Table 3, below, lists for last four outcome period the eight companies, for which the highest returns were predicted by the model for the given year. It seems that at least for the highest returns, the results are identical with the regression model. As such, the same limitations, e.g. the limited ability to capture downward trends, also exist in case of additional autoregressive variable. Given the small autocorrelation, however, such was also expected.

Model Rank	**For 2025**	**Return Rank**	**For 2024**	**Return Rank**	**For 2023**	**Return Rank**	**For 2022**	**Return Rank**
1	Ekspress Grupp	11	Vilkyškių pieninė	1	SAF Tehnika	23	Vilkyškių pieninė	2
2	Eften Real Estate	14	Ignitis grupė	8	Ekspress Grupp	17	Rokiškio sūris	6
3	Harju Elekter	21	Rokiškio sūris	6	Vilkyškių pieninė	6	Linas Agro Group	1
4	Coop Pank	13	Linas Agro Group	10	Coop Pank	15	Ignitis grupė	10
5	Vilkyškių pieninė	6	Šiaulių bankas	4	Eften Real Estate	12	Šiaulių bankas	11
6	Merko Ehitus	1	Telia Lietuva	16	Merko Ehitus	5	Grigeo	17
7	TKM Grupp	12	Ekspress Grupp	24	TKM Grupp	7	Apranga	3
8	LHV Group	7	Eften Real Estate	9	Pro Kapital Grupp	1	Ekspress Grupp	8
Sum		85		78		86		58

Table 3. Ranking of the companies by the linear mixed effects model with autoregressive term

Again, model proposes to invest into relatively fixed set of companies. Names of one company can be found in each year's list (Vilkyškių pieninė). Additionally, twelve companies are found at least in two year lists out of four (Ekspress Group, Eften Real Estate, Coop Pank, Merko Ehitus, TKM Group, LHV Group, Ignitis grupė, Rokiškio sūris, Linas Agro Group, Telia Lietuva, Šiaulių bankas and Linas Agro Group). Remianing four companies end up in lists only once (Harju Elekter, SAF Tehnika, Grigeo and Apranga).

In general, the model outperforms linear regression model slightly. The positive sign of the autoregressive parameter seems logical as typically markets react to good news with lag (so called herd instinct of investors). However, it seems that either because of the limited length of the time series (only seven years) or just as a result of low actual autocorrelation, the importance of autoregressive variable in the final model is negligible for impacting the model use.

2.7 Final model chosen

In previous chapters three different alternatives for applicable model were constructed: linear regression model, fixed effects model and mixed effects model with autoregressive term. Each of them has certain limitations (e.g., as a result of limited observations), but all of them add an important insight to analyzed paradigm. To begin with, let's compare predicted returns for last period between the models.

Model Rank	**Regression, for 2025**	**Return Rank**	**Fixed effects, for 2025**	**Return Rank**	**Auto-regressive, for 2025**	**Return Rank**
1	Coop Pank	13	Merko Ehitus	1	Ekspress Grupp	11
2	LHV Group	7	Nordecon	2	Eften Real Estate	14
3	Ekspress Grupp	11	Enefit Green	3	Harju Elekter	21
4	Šiaulių bankas	5	Klaipėdos nafta	4	Coop Pank	13
5	Eften Real Estate	14	Vilkyškių pieninė	6	Vilkyškių pieninė	6
6	Harju Elekter	21	LHV Group	7	Merko Ehitus	1
7	Vilkyškių pieninė	6	Šiaulių bankas	5	TKM Grupp	12
8	TKM Grupp	12	Ignitis grupė	8	LHV Group	7
Sum		89		36		85

Table 4. Ranking of the companies between the models

One can find two companies among predicted highest returns of all four three models (Vilkyškių pieninė and LHV Group). Additionally, eight companies are found at least in two lists out of three (Coop Pank, Ekspress Grupp, Šiaulių bankas, Eften Real Estate, Harju Elekter, TKM Grupp, Merko Ehitus and Enefit Green). Remaining three companies end up in lists only once (Nordecon, Klaipedos nafta and Ignitis grupė).

The fixed effects model seems to outperform other models in good market conditions, that is most of the time, while linear regression in stands out market correction period. Slight edge of mixed effects model with autoregressive term over the linear regression model, is considered too small to be practical for real life application.

As result, even though, all of the three models enable useful insight in forecasting the annual stock returns, author would rather opt for running in parallel both the fixed effects and the linear regression model.

Summary

The majority of the stock market analyses are relying on either the stochastic models, or even more commonly, standalone ratios indicating the relative price of individual shares (like the price-to-equity ratio). As an alternative, the current master's thesis analyzes possibilities to rely on regression models when predicting annual returns of individual stocks. The analysis is carried out on all listed companies in the main markets of the Tallinn, Riga and Vilnius Stock Exchange. All-together, 16 companies from Estonia, 3 of Latvia and 13 from Lithuanian were analyzed.

In order to increase the number of observations, each company is included up to 7 times into analyses: between outcome years of 2019 to 2025. The relatively long outcome prediction horizon, of one year, should enable usability of traditional prediction techniques, like regression models. At the same time the inclusion of one company more than once, which is typical to panel survey, raises the need analyze possible random effects. Such is done using the fixed and mixed effects models.

Based on the author's experience and the conclusions from data analyses, the decision was taken to exclude from the final model construction dataset such observations, where company made loss during the given calendar year. Often, companies in financial difficulties go through substantial re-organizations, which together with volatility in share prices, introduce major obstacles in using traditional statistical modeling techniques. Also, as current thesis carries the practical aim of building a model, which could help an average investor while making long term investment decisions, then focus on profitable companies seems reasonable.

Three viable models were constructed: linear regression model, fixed effects model and mixed effects model with autoregressive term. The random effects model was found unsuitable for analyzing stock returns as random effects were not found to be significant. While the existence of fixed effects could be easily linked to limited number of regressors in the model and general market perception about the specific company, then it would be difficult to intuitively explain the existence of random effects across the different companies.

The linear regression model includes variables of P/B ratio, industry of main activity, market cycle, management experience and relative return during previous 2 years. It can be seen, that most influential differentiation between the companies, is driven by variables of market cycle, management quality, P/B and relative return over market average. The industry of main activity plays somewhat a smaller role within the final model. The fixed effects model excluded time invariant variables of management and main activity of the company, while relying substantially on fixed effects between the companies, instead. The mixed effects model with autoregressive term excluded weakest regressor, i.e., the main activity of the company and utilized AR(1) correlation factor.

The fixed effects model seems to outperform other models in good market conditions, that is most of the time, while linear regression in stands out market correction period. Slight edge of mixed effects model with autoregressive term over the linear regression model, is considered too small to be practical for real life application.

As result, even though, all of the three models enable useful insight in forecasting the annual stock returns, author would rather opt for running in parallel both the fixed effects and the linear regression model. Constructed models suggest investing into rather limited set of companies on stock market. As such, both the fixed effects model and the linear regression model would have clear benefits in real life application.

References

Diggle, P.J. (2002). Analyses of Longitudinal Data

McCulloch, C.E., Searle, R.S. (2001). Generalized, Linear and Mixed Models.

Möls, M. (2019). Dispersioonanalüüsi segamudelid

Pinheiro, J.P, Bates, D.M., (2000). Mixed-Effects Models in S and S-PLUS.

Robinson, G.K. (1991). That BLUP Is a Good Thing: The Estimation of Randon Effects. Statistical Science 6, pages 15-51.

Searle, R.S., Casella, G., McCulloch, C.E. (2006). Variance Components.

Torres-Reyna, O. (2007). Panel Data Analyses Fixed and Random Effects using Strata

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Indrek Künnapas ,
(author´s name)

1. grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

   Predicting Annual Returns of Individual Stocks on the Baltic Stock Exchange ,
   (title of thesis)

   supervised by Raul Kangro ;
   (supervisor´s name)

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in points 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Indrek Künnapas
21/05/2025