Final Exam

AØKA08084U Advanced Microeconometrics

University of Copenhagen

Character Count Part 1 (incl. spaces): 12,000

Character Count Part 2 (incl. spaces): 11,200

Sebastian Hørlück, Matias Piqueras

December 20, 2021

Contents and distribution of work

Part 1

1.1 Introduction

In economic theory, the assumption of companies showing constant returns to scale, is common. The assumption is related to the Cobb-Douglas representation of production output as a function of labor, L, capital, K, and total factor productivity, A

$$F(K, L) = AK^{\beta_K} L^{\beta_L} \tag{1.1}$$

and states, that if both labor and capital are scaled by the same factor, λ, then production output is scaled by the same factor. This assumption holds if, and only if $\beta_K + \beta_L = 1$.

In this paper we investigate this assumption through a Fixed Effect (FE) and First Difference (FD) estimator. We apply the estimators on a panel dataset of $N = 441$ French firms in the period 1968–1979 ($T = 12$ years), to estimate the parameters β_K and β_L.

We find statistically significant diminishing returns to scale using both estimators, though we state that these results should be interpreted with caution, since essential assumptions regarding the FE and FD estimators could be violated.

1.2 Econometric Theory

We estimate basic model specification given by the log transformation of (1.1)

$$y_{it} = \beta_K k_{it} + \beta_L \ell_{it} + v_{it} \tag{1.2}$$

where the outcome y_{it} is the log of deflated sales, k_{it} is the log of adjusted capital stock and ℓ_{it} the log of employment. We define the error term as $v_{it} := \ln A_{it}$ from the Cobb-Douglas model to capture time-varying and time-invariant factors affecting the outcome y_{it}.

As a starting point will we assume that the error term is the composite of two components $v_{it} := c_i + u_{it}$, where c_i is the captures the unobserved time-invariant effect for i and u_{it} is the idiosyncratic error that varies across both i and t. That is, we assume that deflated sales will depend partly on firm-specific, time-invariant effects, captured in c_i. Such firm-specific effects could be the geographical location of firms, or which industry the firm operates within. Moreover, we will also assume that the expectation of c_i across firms is $\mathrm{E}(c_i) = 0$ to include an intercept to arrive at the *unobserved effects model* (UEM) (Wooldridge, 2010, p.285).

$$y_{it} = \beta_0 + \beta_K k_{it} + \beta_L \ell_{it} + c_i + u_{it} \tag{1.3}$$

For convenience and going forward we will often use matrix notation so $\boldsymbol{\beta} = [\beta_0, \beta_K, \beta_L]$ and $\boldsymbol{x}_{it} = [1, k_{it}, \ell_{it}]$ are row vectors. $\boldsymbol{x}_i$ is a $T \times 3$ matrix of x_{it}'s stacked within individuals, while $\boldsymbol{X}$ is a $N \cdot T \times 3$ matrix, of stacked $\boldsymbol{x}_{it}$'s. $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ are T vectors of stacked y_{it}'s and u_{it}'s within individuals, while $\boldsymbol{y}$, $\boldsymbol{c}$, and $\boldsymbol{u}$ are $N \cdot T$ vectors of stacked y_{it}'s, c_{it}'s, and u_{it}'s respectively.

1.2.1 Consistency

In order to describe our estimation approach we start at the standard Pooled OLS estimator (POLS). We estimate the partial effects $\boldsymbol{\beta}$ in (1.3) by solving $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (y_{it} - \boldsymbol{X\beta})$. This yields the POLS estimator

$$\hat{\boldsymbol{\beta}}_{POLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1} (\boldsymbol{X}'\boldsymbol{y})^{-1} \tag{1.4}$$

$$= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1} (\boldsymbol{X}'\boldsymbol{c} + \boldsymbol{X}'\boldsymbol{u}) \tag{1.5}$$

Taking the probability limit of this then yields

$$\operatorname*{plim}_{N\to\infty} \hat{\boldsymbol{\beta}}_{POLS} = \boldsymbol{\beta} + \underbrace{\mathrm{E}(\boldsymbol{X}'\boldsymbol{X})^{-1} [\mathrm{E}(\boldsymbol{X}'\boldsymbol{c}) + \mathrm{E}(\boldsymbol{X}'\boldsymbol{u})]}_{0} \tag{1.6}$$

$$= \boldsymbol{\beta} \tag{1.7}$$

From (1.6) it is seen, that POLS rely on the assumption that the regressors are uncorrelated with both time-variant and time-invariant omitted variables, that is

$$\mathrm{E}(\boldsymbol{X}(\boldsymbol{c}+\boldsymbol{u}))=0 \tag{POLS.1}$$

However, if $\mathrm{E}(\boldsymbol{X}'\boldsymbol{c}) \neq 0$ we would have an omitted variable bias, and the POLS would not be consistent. This case can be illustrated through the example of industry specific affects, where some industries are more labor-intensive relative to capital than others. In this case c_i would be correlated with ℓ_{it} and thus violate the exogeneity assumption of POLS. This implies that $\boldsymbol{\beta}$ cannot be estimated consistently through POLS. To estimate model parameters even when unobservable individual specific effects are present we introduce the FD and FE-estimators. Moving on, we will not discuss the further assumptions needed for POLS .

We apply the FD estimator by first differencing all variables and then apply POLS to the transformed dataset:

$$\begin{aligned} y_{it}-y_{it-1} &= (\beta_0-\beta_0)+\beta_K(k_{it}-k_{it-1})+\beta_L(\ell_{it}-\ell_{it-1})+ \qquad (1.8)\\ &\quad (c_i-c_i)+(u_{it}-u_{it-1}) \\ \Delta y_{it} &= \beta_K\Delta k_{it}+\beta_L\Delta\ell_{it}+\Delta u_{it} \qquad (1.9) \end{aligned}$$

From (1.9) it shows that by transforming the data, we effectively remove the individual specific time-invariant effects, seen in (1.3), and therefore the endogeneity problems stemming from $\mathrm{E}(\boldsymbol{X}'\boldsymbol{c}) \neq 0$ are no longer present. Similarly, for FE we also transform the data, by taking the difference from the mean of each variable within individuals. This yields:

$$\begin{aligned} y_{it}-\bar{y}_i &= (\beta_0-\beta_0)+\beta_K(k_{it}-\bar{k}_i)+\beta_L(\ell_{it}-\bar{\ell}_i)+(c_i-\bar{c}_i)+(u_{it}-\bar{u}_i) \qquad (1.10)\\ \ddot{y}_{it} &= \beta_K\ddot{k}_{it}+\beta_L\ddot{\ell}_{it}+\ddot{u}_{it} \qquad (1.11) \end{aligned}$$

where a bar above a variable indicates the mean of this variable across all $t \in T$, such that $\bar{x}_i=\frac{1}{T}\sum_{i=1}^{T}x_{it}$. Since c_i is time-invariant, we have that $\frac{1}{T}\sum_{i=1}^{T}c_i=c_i$, which is why, c_i is cancelled out when going from (1.10) to (1.11).

The consistency of the FD and FE estimators rely on two similar assumptions. Firstly, the transformed regressors need to be exogenous to the idiosyncratic error-term. Due to the transformation of the data when estimating $\hat{\boldsymbol{\beta}}$ through FD, we have that the exogeneity assumption for FD not only depends on the explanatory variables from the current period, but also for the two surrounding periods. This can be written as

$$\mathrm{E}(\boldsymbol{\Delta x'_{it}}\Delta u_{it}) = \mathrm{E}(u_{it}|\boldsymbol{x_{it-1}}, \boldsymbol{x_{it}}, \boldsymbol{x_{it+1}}) = 0 \tag{FD.1}$$

Similarly, for the FE estimator we have that the exogeneity assumption relies on the explanatory variables from all time-periods, since the estimator is based upon a mean of the explanatory variables across all time-periods. This can be written as

$$\mathrm{E}(\boldsymbol{\ddot{x}'_{it}}\ddot{u}_{it}) = \mathrm{E}(u_{it}|\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{iT}}) = 0 \tag{FE.1}$$

Further, for both the FD and the FE estimator we need the transformed regressors to show full rank, such that

$$\mathrm{E}(\Delta\boldsymbol{X}'\Delta\boldsymbol{X}) \text{ is invertible} \tag{FD.2}$$

and

$$\mathrm{E}(\boldsymbol{\ddot{X}}'\boldsymbol{\ddot{X}}) \text{ is invertible} \tag{FE.2}$$

For now, we assume both (FD.1) and (FE.1) to hold. This assumption is not a given, and there are several cases under which the assumption might be violated. We test the assumptions further in section 1.3 and discuss the implication of a possible violation in section 1.4.

1.2.2 Inference and Asymptotics

In this section we describe how we can derive the asymptotic variance of both the FD and FE-estimator. By combining (1.5) with (1.9) and (1.11) respectively, as well as a bit of rewriting, we have the following two representations of the differences in estimated effects and true effects

$$(\hat{\boldsymbol{\beta}}_{FD} - \boldsymbol{\beta}_{FD}) = (\Delta\boldsymbol{X}'\Delta\boldsymbol{X})^{-1}(\Delta\boldsymbol{X}'\Delta\boldsymbol{u}) \tag{1.12}$$

$$(\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}_{FE}) = \left(\boldsymbol{\ddot{X}}'\boldsymbol{\ddot{X}}\right)^{-1}\left(\boldsymbol{\ddot{X}}'\boldsymbol{\ddot{u}}\right) \tag{1.13}$$

By multiplying both sides of (1.12) and (1.13) by $\sqrt{N}$ it can be shown that these differences converge in distribution to a normal distribution with mean zero and variance

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{FD} - \boldsymbol{\beta}_{FD}) \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1}) \tag{1.14}$$

with $A = \mathrm{E}(\Delta \boldsymbol{X}' \Delta \boldsymbol{X})$ and $B = \mathrm{E}[\Delta \boldsymbol{X}_{\boldsymbol{i}}' \mathrm{E}(\Delta \boldsymbol{u}_{\boldsymbol{i}} \Delta \boldsymbol{u}_{\boldsymbol{i}}' | \Delta \boldsymbol{X}_{\boldsymbol{i}}) \Delta \boldsymbol{X}_{\boldsymbol{i}}]$, and

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}_{FE}) \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1}) \tag{1.15}$$

with $A = \mathrm{E}(\ddot{\boldsymbol{X}}' \ddot{\boldsymbol{X}})$ and $B = \mathrm{E}[\ddot{\boldsymbol{X}}_{\boldsymbol{i}}' \mathrm{E}(\ddot{\boldsymbol{u}}_{\boldsymbol{i}} \ddot{\boldsymbol{u}}_{\boldsymbol{i}}' | \ddot{\boldsymbol{X}}_{\boldsymbol{i}}) \ddot{\boldsymbol{X}}_{\boldsymbol{i}}]$. Both these convergences in distribution rely on the exogeneity assumption related to the idiosyncratic error-term u_{it}, since it uses the Central Limit Theorem (CLT) of $\frac{1}{\sqrt{N}} \sum_i w_i \xrightarrow{d} \mathcal{N}(0, \mathrm{V}(w_i))$ if $\mathrm{E}(w_i) = 0$, (Riis-Vestergaard Sørensen, 2021, p.41). As explained in section 1.2.1 we assumed both $\mathrm{E}(\boldsymbol{\Delta x_{it}} \Delta u_{it}) = 0$ and $\mathrm{E}(\boldsymbol{\ddot{x}_{it}} \ddot{u}_{it}) = 0$ for consistency. Therefore we can apply the CLT to (1.12) and (1.13) to arrive at (1.14) and (1.15) respectively. Thereby we can estimate the variance of the estimators, $\widehat{\mathrm{Avar}}(\hat{\beta}_{FD})$ and $\widehat{\mathrm{Avar}}(\hat{\beta}_{FE})$ as

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}_{FD}) = (\Delta \boldsymbol{X}' \Delta \boldsymbol{X})^{-1} \left(\sum_{i=1}^{N} \Delta \boldsymbol{X}_{\boldsymbol{i}}' \Delta \hat{\boldsymbol{u}}_{\boldsymbol{i}} \Delta \hat{\boldsymbol{u}}_{\boldsymbol{i}}' \Delta \boldsymbol{X}_{\boldsymbol{i}} \right) (\Delta \boldsymbol{X}' \Delta \boldsymbol{X})^{-1} \tag{1.16}$$

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}_{FE}) = (\ddot{\boldsymbol{X}}' \ddot{\boldsymbol{X}})^{-1} \left(\sum_{i=1}^{N} \ddot{\boldsymbol{X}}_{\boldsymbol{i}}{}' \hat{\ddot{\boldsymbol{u}}}_{\boldsymbol{i}} \hat{\ddot{\boldsymbol{u}}}_{\boldsymbol{i}}' \ddot{\boldsymbol{X}}_{\boldsymbol{i}} \right) (\ddot{\boldsymbol{X}}' \ddot{\boldsymbol{X}})^{-1} \tag{1.17}$$

where $\Delta \hat{u}_i$ and $\hat{\ddot{u}}_i$ are the residuals obtained from the FD and FE estimations respectively. These variance estimators are equivalent to (10.70) and (10.59) from (Wooldridge, 2010) respectively. (1.16) and (1.17) estimate heteroskedastic robust variances of $\boldsymbol{\beta}_{FD}$ and $\boldsymbol{\beta}_{FE}$ relying only on T being relative small compared to N (Wooldridge, 2010, p.311). Thus we allow for serial correlation in the idiosyncratic error-terms.

If we were further willing to assume that Δu_{it} is IID (FD.3), which corresponds to u_{it} being a unit root process, (1.16) collapses to a simpler equation, and FD will be efficient. If instead we are willing to assume that $\ddot{u}_{it}$ is IID (FE.3), (1.17) collapses to a simpler statement and FE will be efficient. We can test for both these cases, using a test for serial correlation, further explained in section 1.3.

We can use (1.16) and (1.17) to test whether our estimates, (1.9) and (1.11) imply constant returns to scale in the Cobb-Douglas production function. To do this we apply a Wald test, to test the null hypotheses of

$$\mathcal{H}_0 : \beta_k + \beta_\ell = 1 \tag{1.18}$$

The Wald test allows for the test of multiple hypotheses relating to the estimated parameters, as well as testing hypotheses building on combinations of the estimated parameters. In our case we want to test whether the estimated Cobb-Douglas production functions shows constant returns to scale. Formally this corresponds to $\beta_k + \beta_\ell$ being equal to one.

The Wald test used for testing, whether the estimated production function shows constant returns to scale, tests the null hypotheses of $R\boldsymbol{\beta} = r$, where $R = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and $r = \begin{pmatrix} 1 \end{pmatrix}$. The Wald test statistic is calculated as

$$W = (R\hat{\boldsymbol{\beta}} - r)'(\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}))^{-1}(R\hat{\boldsymbol{\beta}} - r) \tag{1.19}$$

Following (Wooldridge, 2010, p.42) it can be shown, that under the null, the Wald test statistic will converge in distribution to a χ^2_Q distribution with Q amounting to the number of hypotheses being tested and the degrees of freedom in the distribution. This convergence rely on the asymptotic distribution of $\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\beta}, \text{Avar}(\hat{\boldsymbol{\beta}}))$ which again rely on the consistency of $\hat{\boldsymbol{\beta}}$, and the assumptions required for this.
The results of the Wald test are shown in section 1.3.

1.2.3 Tests of strict exogeneity and serial correlation

Since the assumption of strict exogeneity as per (FE.1) and (FD.1) are key for consistent estimates, we introduce a formal test of these assumptions. As described in (Wooldridge, 2010, p.325) we can do this by estimating

$$\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \ddot{\boldsymbol{x}}_{it+1}\boldsymbol{\delta} + \ddot{u}_{it} \tag{1.20}$$

since strict exogeneity would imply $\delta = 0$. We present these results in table 1.3.

The assumptions (FD.3) and (FE.3) are key for doing valid inference on the estimates thus we test these following the steps laid out by Wooldridge (2010, p.320), using the error term obtained from FD $e_{it} \equiv \Delta u_{it}$ to run the pooled OLS regressions

$$\hat{e}_{it} = \hat{\rho}_1 \hat{e}_{i,t-1} + \text{error}_{it}, \qquad t = 3, 4, \ldots, T; i = 1, 2, \ldots, N. \tag{1.21}$$

The quantity of interest here is $\hat{\rho}_1$, since it can be shown that under FE3 the correlation will be $\text{Corr}(e_{it}, e_{i,t-1}) = -.5$. However, if we instead think that e_{it} follows a random-walk as under FD3, then the correlation should be $\text{Corr}(e_{it}, e_{i,t-1}) = 0$. The results from estimating (1.21) are reported in table 1.1.

1.3 Empirical Results

As a first step in our analysis we evaluate (FD.3) and (FE.3) by testing for serial correlation in the idiosyncratic error term, and present the results in table 1.1. The coefficient $\hat{\rho}_1$ is -0.199 with a standard error of .015 and significant at $p < 0.001$. This provides evidence of negative serial correlation in Δu_{it} warranting the use of heteroscadasticity-robust standard errors derived in (1.16) and (1.17) for the estimation of FE and FD. Moreover, the results are coherent with Wooldridge (2010, p.321)'s observations that the behavior of the error term is likely to lie somewhere in-between (FD.3) and (FE.3) in many empirical settings, in our case marginally suggesting FD being more efficient.

Table 1.1: Test of Serial Correlation

	Dependent variable:
	$\hat{e}_{it}$
$\hat{e}_{i,t-1}$	-0.199***
	(0.015)
Observations	4,410
R^2	0.039
Adjusted R^2	0.039
Note:	*p<0.1; **p<0.05; ***p<0.01

In table 2.4 we report the FE and FD estimates from estimating (1.9) and (1.11) respectively with heteroscedasticity-robust standard errors. We can see from the FE estimates that a one percent increase in ℓ_{it} suggests an increase of about 0.155% increase in y_{it}, with a standard error of 0.03%. The corresponding estimate for FD suggests an 0.063% increase with a standard error of 0.023%. Also the estimates for k_{it} differ significantly between the models, with an estimate of 0.694% and a standard error of 0.042% for FE compared to 0.549% and 0.029% for FD. All estimates are significant at $p < 0.001$. The large difference in the estimated effect size between the models likely renders any interpretation thereof problematic. In fact, the difference between estimates suggests that we should be worried about assumption FE.1 and FD.1 being violated, causing estimates to be inconsistent (Wooldridge, 2010, p.321). We consider attributing such differences to sampling error ill-advised without any insights into the data-generating process. We discuss this further in sections 1.3.1 and 1.4.

Table 1.2: Main results

	Dependent variable:	
	Log deflated sales	
	FE	FD
	(1)	(2)
Log adjusted capital	0.155***	0.063***
	(0.030)	(0.023)
Log employment	0.694***	0.549***
	(0.042)	(0.029)
Observations	5,292	4,851
R^2	0.477	0.165
Adjusted R^2	0.429	0.165
Note:	*p<0.1; **p<0.05; ***p<0.01	

As a final step in this part of the analysis we explicitly evaluate if the production function shows constant returns to scale by testing the *null* hypothesis (1.18) using the Wald test statistic derived in (1.19). For FE we have $\beta_K + \beta_L = 0.849$ with the test statistic $\chi^2 = 19.403$. For FD we have $\beta_K + \beta_L = 0.612$ with the test statistic $\chi^2 = 251.73$. Thus, in both cases significant at $p < 0.001$ leading us to reject the *null*-hypothesis. More specifically, as $\beta_k + \beta_\ell < 1$ for both estimators, we find diminishing returns to scale to be a more accurate characterisation of the relation between employment, capital and deflated sales. We want to stress again, that these results should be interpreted with caution since (FD.1) and (FE.1) are likely violated.

1.3.1 Test of strict exogeneity

We present the results of estimating (1.20) in table 1.3 and see that $\delta \neq 0$ since both β_K and β_L are significant and positive. This implies that strict exogeneity is violated and neither FD or FE produce consistent estimates of $\boldsymbol{\beta}$. Further, since δ is positive, we assume the bias to arise from (FD.1) and (FE.1) being violated is positive, indicating that (1.9) and (1.11) both overestimate $\boldsymbol{\beta}$, thus not challenging the conclusion of the Cobb-Douglas production function showing diminishing returns to scale.

Table 1.3: Fixed effects with leads

	Dependent variable:
	Log deflated sales
Log adjusted capital	0.028
	(0.038)
Log employment	0.541***
	(0.043)
Log adjusted capital$_{t+1}$	0.167***
	(0.046)
Log employment$_{t+1}$	0.142***
	(0.028)
Observations	4,851
R^2	0.478
Adjusted R^2	0.426
Note:	*p<0.1; **p<0.05; ***p<0.01

1.4 Discussion

There are several points that challenge the results we find in this paper, and thus to which extent the results should be interpreted. These challenges are connected to the consistency of the results we yield from our FD and FE estimators.

Firstly, as shown in section 1.3, there are relatively large differences in the estimates of $\boldsymbol{\beta}$ based on whether the FD or the FE estimator is used. This should not be the case, had the consistency assumptions not been violated, thus indicating a such violation.

More precisely, the exogeneity assumptions used to derived the estimators could be violated, which could happen through several channels. One example could be how labour force, ℓ_{it} could be connected to shocks, u_{it}, where labour strikes could happen more frequently in larger companies, with a higher degree of unionization, following Jeppesen and Nielsen (2021). This would cause both estimators to be inconsistent, and have different probability limits, (Wooldridge, 2010, p.321-322). Generally speaking, correlation between x_{it} and u_{is} can appear in either the same period, $t = s$, in the lagged regressor, $t < s$, and/or in the future regressor, $t > s$. Any of these cases causes inconsistency of both FD and FE, and challenges any conclusions based on the results from section 1.3.

As seen in section 1.3.1, the leaded regressors, $t+1$, are significant, thus implying that they should be included in the model to properly estimate target parameters for period t. But as Wooldridge (2010) mentions in p. 322, it only rarely makes sense to include leaded variables. It is difficult to explain how the sales in period t depends on future labour force. Thus, including leaded variables does not add information, but the significance of these variables reveals problems with the model specification.

Lastly, we wish to mention that we purposely chose not to estimate a Random Effects model as we have assumed it to yield inconsistent estimates for the same reasons as OLS would. Given the page limits we therefore chose to focus on FE and FD which we deemed to be more appropriate models. From a pragmatical viewpoint, there is also little to be gained from RE since the data does not include any time-invariant variables who's coefficients could be estimated by RE but not by FE or FD.

Part 2

2.1 Derive the likelihood contribution

The regression model we are considering is

$$y_i = h\left(\mathbf{x}_i\boldsymbol{\beta}\right) + \varepsilon_i, \quad \varepsilon_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right).$$

Since the stochastic component ε_i is conditionally independent from $\boldsymbol{x}_i$ we can express the distribution of y_i as drawn from a normal with mean $h(\boldsymbol{x}_i\boldsymbol{\beta})$ such that $y_i \sim \mathrm{N}\left(h(\boldsymbol{x}_i\boldsymbol{\beta}), \sigma_\varepsilon^2\right)$. Using this we can write conditional density of y_i on the data $\boldsymbol{x_i}$ and as a function of $(\boldsymbol{\beta}, \sigma_\varepsilon, h)$ using the definition of the normal distribution

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \sigma_\varepsilon, h) = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} \exp\left\{-\frac{\left[y_i - h(\boldsymbol{x}_i\boldsymbol{\beta})\right]^2}{2\sigma_\varepsilon^2}\right\}. \tag{2.22}$$

To find the log-likelihood contribution for the model we take logs of (2.22) and simplify

$$\ell_i(\boldsymbol{\beta}, \sigma_\varepsilon, h) = -\ln\sigma_\varepsilon\sqrt{2\pi} - \frac{\left[y_i - h(\boldsymbol{x}_i\boldsymbol{\beta})\right]^2}{2\sigma_\varepsilon^2} \tag{2.23}$$

2.2 Explain why the parameters cannot be identified

If we suppose that $h(z) = \gamma z$ where is the linear transformation $z = \boldsymbol{x}_i\boldsymbol{\beta}$ and $\gamma \neq 0$ identification fails as there is no unique solution to the population problem. If we say that $\boldsymbol{\theta}_0 = \{\boldsymbol{\beta}_0, \gamma_0\}$ are the true parameters and solve the population problem, then we see, that for an alternative set of parameters $\boldsymbol{\theta}_a = \left\{2\boldsymbol{\beta}_0, \frac{\gamma_0}{2}\right\}$ we have

$$h(z) = \boldsymbol{x}_{it}\boldsymbol{\theta}_0 = \boldsymbol{x}_{it}\gamma_0\boldsymbol{\beta}_0 = \boldsymbol{x}_{it}\frac{\gamma_0}{2}2\boldsymbol{\beta}_0 = \boldsymbol{x}_{it}\boldsymbol{\theta}_a$$

and thus $\boldsymbol{\theta}_a$ can recover the same function value as $\boldsymbol{\theta}_0$, which will not be a unique solution to the population problem. In intuitive terms, there are many (in fact infinitely many) solutions since you can always change γ proportionally to β and vice-versa to achieve the same solution.

2.3 Estimate model parameters

To estimate the models parameters we will use the log-likelihood contribution in (2.23) to define the corresponding criterion function over samples $i \in 1, \ldots, N$. We can apply this to a maximum likelihood estimator (ML). For computational reasons we want to turn it to a minimization problem, thus we make a sign change so the criterion becomes

$$\begin{aligned} q(\boldsymbol{w}_i, \boldsymbol{\beta}, \sigma_\varepsilon, h) &= \underset{\boldsymbol{\beta}, \sigma_\varepsilon}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}, \sigma_\varepsilon, h) \\ &= \underset{\boldsymbol{\beta}, \sigma_\varepsilon}{\operatorname{argmin}} \frac{1}{N} \left\{ N \ln \sigma_\varepsilon \sqrt{2\pi} + \frac{\sum_{i=1}^{N} \left[y_i - h(\boldsymbol{x}_i \boldsymbol{\beta})\right]^2}{2\sigma_\varepsilon^2} \right\} \\ &= \underset{\boldsymbol{\beta}, \sigma_\varepsilon}{\operatorname{argmin}} \ln \sigma_\varepsilon \sqrt{2\pi} + \frac{||\boldsymbol{y} - h(\boldsymbol{X}\boldsymbol{\beta})||^2}{2N\sigma_\varepsilon^2}. \end{aligned}$$

Where $\boldsymbol{w}_i$ is the pair $(y_i, \boldsymbol{x}_i)$.

We know that ML will produce consistent estimates under the following conditions:

1. The parameter space, $\boldsymbol{\Theta}$, is compact
2. $q(\boldsymbol{w}_i, .)$ is continuous
3. There is a unique solution to the population problem, $\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}\left[(y - h(\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma_\varepsilon))^2\right]$, i.e. that the true parameters, $\boldsymbol{\theta}_0 = \{\boldsymbol{\beta}_0, \sigma_{\varepsilon 0}\}$ are identified.

We assume condition 1 holds, and we observe that $q(\boldsymbol{w}_i, .)$ is indeed continuous. Lastly we assess whether there are multiple solutions to the population problem, that is if there exists a $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ which also solves the population problem. For both given functions we observe, that the model specifications offers an unique solution to the population problem.

The functional form h that generated the data is for now unknown but one out of two alternatives, so we fit the two separates models where h is substituted by $h(z) = 3\Phi(z)$ and $h(z) = \exp(-z)$. Findings are reported in 2.4 and we find that $\exp(-z)$ is a better fit based on a lower criterion value of 1.434 and a higher proportion of explained variance, as given by $R^2 = 0.324$. An analysis of the residuals does not suggest any immediate problems with heteroskedasticity for any of the two models. Taking these considerations together, we chose to continue with the $h = \exp(-z)$ model.

Table 2.4: Cross-Section Model Results

	Dependent variable: y	
	$3\Phi(z)$	$\exp(-z)$
β_1	0.9192	−1.0229
	(0.0797)	(0.0377)
β_2	−1.1848	1.0371
	(0.0720)	(0.0707)
β_3	−1.1459	0.8234
	(0.1240)	(0.0828)
σ_ε	1.0200	1.0155
	(0.0235)	(0.0237)
$q(\theta)$	1.438	1.434
Observations	1000	1000
R^2	0.318	0.324

Note: Standard errors in parenthesis

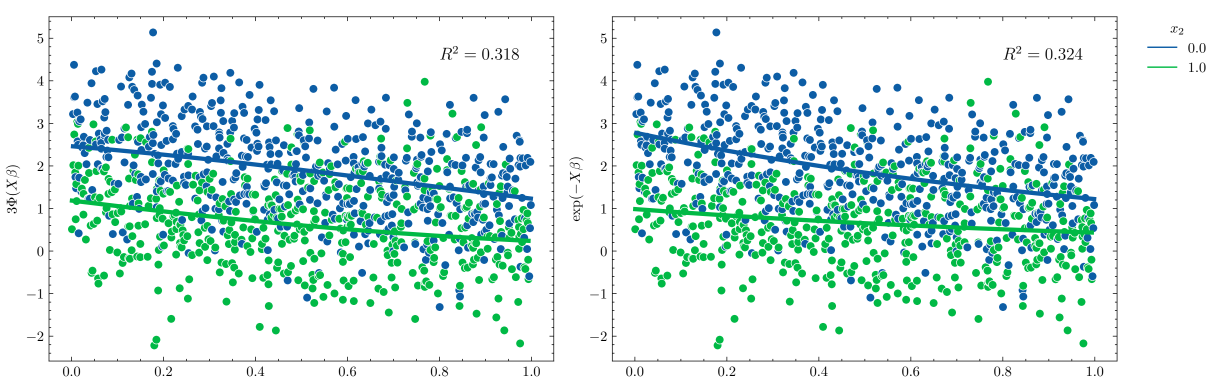

Figure 2.1: Model fit

2.4 Estimate Partial Effects

Continuing with the model

$$y_i = \exp(-\boldsymbol{x}_i\boldsymbol{\beta}) + \varepsilon. \tag{2.24}$$

We now want to find the partial effects (PE) for $\mathrm{E}\left[y_i \mid \boldsymbol{x}_i = \boldsymbol{x}^0\right]$ evaluated at $\boldsymbol{x}^0 = (1, 1, x_3)$. Where x_3 varies over the specific values $x_3 = 0, 0.1, 0.2, \ldots, 1$. The PE can be calculated as the derivative w.r.t. to features x_j which in our case becomes

$$\delta_j(\boldsymbol{x}^0) = \left.\frac{\partial \mathrm{E}(y_i|\boldsymbol{x}_i)}{\partial \boldsymbol{x}_j}\right|_{\boldsymbol{x}=\boldsymbol{x}^0} = -\exp(-\boldsymbol{x}^{\boldsymbol{0}}\boldsymbol{\beta})\beta_j \tag{2.25}$$

To calculate standard errors we use the bootstrap procedure following Wooldridge (2010, p.438 - 442) by taking samples with replacement of from the original sample of size $N = 1000$. We then fit the model to the data and estimate the PE's. This is repeated $2,500$ times and standard errors calculated based on the resulting PE's from each bootstrap repetition. To increase the numerical stability and reduce the computational time we note that σ_ε^2 is not necessarily of interest right now and that the criterion can be simplified to

$$q(\boldsymbol{w}_i, \boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, ||\boldsymbol{y} - \exp(-\boldsymbol{X}\boldsymbol{\beta})||^2$$

since we have independence between the error term and features $\varepsilon_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right)$. The results are displayed in Figure 2.2 (vertical lines represent the standard errors) and Table 2.5. As can be seen the partial the estimated negative partial effect δ_2 and δ_3 become less significant in magnitude as x_3 increases. The same tendency can be seen for δ_3 who's positive partial effect decreases as x_3 increases. This also makes sense on the basis of Figure 2.3 where the two lines gradually come closer to each other. Given the property of the functional form of systematic component where $\exp(-\boldsymbol{x}_i\beta) > 0$, this result can also be viewed as the two lines approaching the bound of 0 where observations who drew $x_2 = 1$ are "ahead" in their trajectory. This would also explain why the standard errors are gradually decreasing for all PE's, since any variation after a certain point have a very small effect on the function value.

Table 2.5: Partial Effects

x_3	δ_1	δ_2	δ_3
0.0	1.0086	-1.0225	-0.8119
	(0.0950)	(0.0382)	(0.1188)
0.1	0.9289	-0.9417	-0.7477
	(0.0829)	(0.0295)	(0.1040)
0.2	0.8554	-0.8673	-0.6886
	(0.0725)	(0.0229)	(0.0909)
0.3	0.7878	-0.7987	-0.6342
	(0.0636)	(0.0186)	(0.0793)
0.4	0.7255	-0.7356	-0.5840
	(0.0562)	(0.0166)	(0.0691)
0.5	0.6682	-0.6774	-0.5379
	(0.0500)	(0.0167)	(0.0602)
0.6	0.6154	-0.6239	-0.4954
	(0.0450)	(0.0180)	(0.0523)
0.7	0.5667	-0.5746	-0.4562
	(0.0409)	(0.0198)	(0.0454)
0.8	0.5219	-0.5291	-0.4201
	(0.0377)	(0.0216)	(0.0394)
0.9	0.4807	-0.4873	-0.3869
	(0.0351)	(0.0232)	(0.0343)
1.0	0.4427	-0.4488	-0.3564
	(0.0331)	(0.0246)	(0.0298)

Note: Standard errors in parenthesis

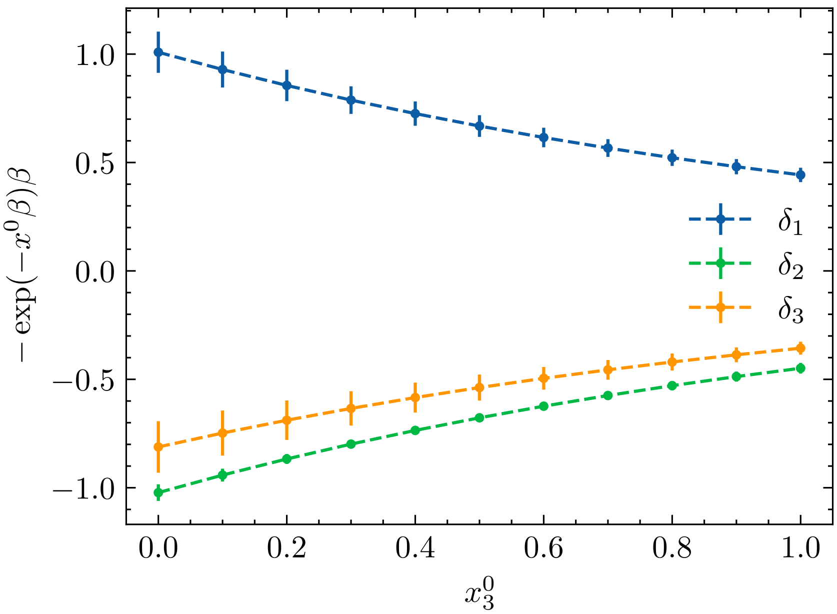

Figure 2.2: Partial Effects

2.5 Derive log-likelihood contribution

We have the equation $y_{it} = h(\boldsymbol{x}_{it}\boldsymbol{\beta}+c_i)+\varepsilon_{it}$, where $c_i|x_i \sim \mathrm{N}(0, \sigma_c)$ and $\{\varepsilon_i\}_{t=1}^{T} \sim \mathrm{N}(0, \sigma_\varepsilon)$. To derive the log-likelihood contribution

$$\ell_i(\boldsymbol{\beta}, \sigma_\varepsilon, \sigma_c, h) = \ln f(y|x)$$

we must first find an expression of $f(y|x)$, i.e. the marginal density of y given x. From probability theory we have that if c depends on x, then for $y \in \mathcal{R}$

$$\begin{aligned} f(y|x) &= \int_{-\infty}^{\infty} f(y, c|x)dc, \\ f(y, c|x) &= f(y|x, c)f(c, x), \\ f(y|x) &= \int_{-\infty}^{\infty} f(y|x, c)f(c|x)dc = \int_{-\infty}^{\infty} f(y|x, c)\phi(c)dc \end{aligned}$$

In the last step we have used, that c, given x is a random normal distribution. Since ε is also a random distribution we have that Thus we can say that $f(y_{it}|\boldsymbol{x}_{it}, c_i, \varepsilon_{it}) = f(y_{it}|\boldsymbol{x}_{it}, c_i)$

is normally distributed with mean $f(y_{it}|\boldsymbol{x}_{it}, c_i)$ and variance σ_ε^2, thus we can write

$$\begin{aligned} f(y_{it}|\boldsymbol{x}_{it}, c_i = \sigma_c \cdot c) &= \frac{1}{\sigma_\varepsilon}\phi\left(\frac{y_{it} - h(\boldsymbol{x}_{it}\boldsymbol{\beta} + \sigma_c \cdot c)}{\sigma_\varepsilon}\right) \\ &= \frac{1}{\sigma_\varepsilon}\left(\frac{\exp\left(-\frac{\left(\frac{y_{it}-h(\boldsymbol{x}_{it}\boldsymbol{\beta}+\sigma_c\cdot c)}{\sigma_\varepsilon}\right)^2}{2}\right)}{\sqrt{2\pi}}\right) \end{aligned}$$

and insert in the log-likelihood contribution

$$\ell_i(\boldsymbol{\beta}, \sigma_\varepsilon, \sigma_c, h) = \ln \int_{-\infty}^{\infty} \prod_{t=1}^{T}\left[\frac{1}{\sigma_\varepsilon}\left(\frac{\exp\left(-\frac{\left(\frac{y_{it}-h(\boldsymbol{x}_{it}\boldsymbol{\beta}+\sigma_c\cdot c)}{\sigma_\varepsilon}\right)^2}{2}\right)}{\sqrt{2\pi}}\right)\right]$$

From here we derive:

$$\begin{aligned} &\ell_i(\boldsymbol{\beta}, \sigma_\varepsilon, \sigma_c, h) \\ &= \ln \int_{-\infty}^{\infty} \prod_{t=1}^{T}\left[\frac{1}{\sigma_\varepsilon}\left(\frac{\exp\left(-\frac{\left(\frac{y_{it}-h(\boldsymbol{x}_{it}\boldsymbol{\beta}+\sigma_c\cdot c)}{\sigma_\varepsilon}\right)^2}{2}\right)}{\sqrt{2\pi}}\right)\right]\phi(c)dc \\ &= \ln \int_{-\infty}^{\infty}\left[\prod_{t=1}^{T}\left[\frac{1}{\sigma_\varepsilon}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(\frac{y_{it}-h(\boldsymbol{x}_{it}\boldsymbol{\beta}+\sigma_c\cdot c)}{\sigma_\varepsilon}\right)^2}{2}\right)\right]\right]\phi(c)dc \\ &= \ln\left(\frac{1}{\sigma_\varepsilon}\right)^T + \ln\left(\frac{1}{\sqrt{2\pi}}\right)^T + \ln \int_{-\infty}^{\infty}\left[\prod_{t=1}^{T}\left[\exp\left(-\frac{\left(\frac{y_{it}-h(\boldsymbol{x}_{it}\boldsymbol{\beta}+\sigma_c\cdot c)}{\sigma_\varepsilon}\right)^2}{2}\right)\right]\right]\phi(c)dc \\ &= \ln\left(\frac{1}{\sigma_\varepsilon}\right)^T + \ln\left(\frac{1}{\sqrt{2\pi}}\right)^T + \ln \int_{-\infty}^{\infty}\left[\exp\left(-\sum_{i=1}^{T}\left[\left(\frac{y_{it}-h(\boldsymbol{x}_{it}\boldsymbol{\beta}+\sigma_c\cdot c)}{\sigma_\varepsilon}\right)^2\frac{1}{2}\right]\right)\right]\phi(c)dc \\ &= \ln\left(\frac{1}{\sigma_\varepsilon}\right)^T + \ln\left(\frac{1}{\sqrt{2\pi}}\right)^T + \ln \int_{-\infty}^{\infty}\left[\exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{T}\left[(y_{it}-h(\boldsymbol{x}_{it}\boldsymbol{\beta}+\sigma_c\cdot c))^2\right]\right)\right]\phi(c)dc \\ &= -T\ln(\sigma_\varepsilon) + \ln \int_{-\infty}^{\infty}\left[\exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{T}\left[(y_{it}-h(\boldsymbol{x}_{it}\boldsymbol{\beta}+\sigma_c\cdot c))^2\right]\right)\right]\phi(c)dc - \frac{T}{2}\ln(2\pi) \end{aligned}$$

thus arriving at the expected log-likelihood contribution.

2.6 Discussion of ε_{it} distribution

For $\varepsilon_{it}|x_i, c_i \sim N(0, \sigma_\varepsilon^2)$ we know that when observing one individual for all time periods, then ε_{it} is normally distributed with mean 0 and variance σ_ε^2, but not necessarily for each time period. If instead we have $\varepsilon_{it}|x_{it}, c_i \sim N(0, \sigma_\varepsilon^2)$, then for each time period within each individual ε_{it} is normally distributed with mean 0 and variance σ_ε^2. Thus if ε_{it} is given by the alternative assumption, then the original assumption will still hold, but the reverse will not necessarily be the case. In this matter, it can be said that $\varepsilon_{it}|x_i, c_i \sim N(0, \sigma_\varepsilon^2)$ yields a more strict assumption on ε_{it} compared to $\varepsilon_{it}|x_{it}, c_i \sim N(0, \sigma_\varepsilon^2)$.
Following this reasoning, the alternative assumption will not change the likelihood contribution.

2.7 Estimate model parameters

To estimate the model parameters, we apply maximum likelihood method (ML), by solving

$$\hat{\boldsymbol{\beta}}, \hat{\sigma}_\varepsilon, \hat{\sigma}_c = \underset{\boldsymbol{\beta}, \sigma_\varepsilon, \sigma_c}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} -\ell_i(\boldsymbol{\beta}, \sigma_\varepsilon, \sigma_c, h)$$

using the $\ell_i(\boldsymbol{\beta}, \sigma_\varepsilon, \sigma_c, h)$ derived in (2.5).
In terms of consistency we follow the same steps as in 2.3 to reach the conclusion of ML yielding consistent estimates.

We can not compute the actual integral over c when doing maximum likelihood, but we can approximate the log-likelihood contribution through simulated maximum likelihood (SML). Through this approach, we draw R values of $c_r \sim N(0, \sigma_c)$, and thus we let the estimation function affect the draws by different inputs of σ_c. We then use these R draws of c_r to compute the inside of the integral, and take the average over these values, which we treat as an approximation of the integral. Thus we say

$$\begin{aligned} &\ell_i(\boldsymbol{\beta}, \sigma_\varepsilon, \sigma_c, h) \cong \\ &- T \ln(\sigma_\varepsilon) + \ln \frac{1}{R} \sum_{r=1}^{R} \left[\exp \left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{T} \left[(y_{it} - h(\boldsymbol{x}_{it}\boldsymbol{\beta} + \sigma_c \cdot c_r))^2 \right] \right) \right] - \frac{T}{2} \ln(2\pi) \end{aligned}$$

and use the right hand side of this as the log likelihood contribution in our estimation approach. We set $R = 100$ and present the results in table 2.6, while figure 2.3 show predictions based on the estimated parameters.

Table 2.6: Model Results, panel data

	Dependent variable: y	
	$3\Phi(z)$	$\exp(-z)$
β_1	0.7877	−0.9154
	(0.0731)	(0.0510)
β_2	−1.0959	0.8848
	(0.0501)	(0.0297)
β_3	−1.1808	0.8907
	(0.0794)	(0.0405)
σ_ε	−0.5850	−0.5149
	(0.0137)	(0.0121)
σ_c	0.5513	0.4603
	(0.0478)	(0.0373)
$q(\theta)$	10.54	8.912
N	100	100
T	10	10
R^2	0.351	0.352
Note:	Standard errors in parenthesis	

From this we see that the β estimates from $h(z) = \exp(-z)$ are similar to the negative β estimates from $h(z) = 3\phi(z)$, though not identical to. Further, the σ estimates from the two functions are also similar, but smaller for the exponential function.
Looking at the average log-likelihood contribution from the estimated parameters, the exponential function yields $\frac{1}{N}\ell_i(.) = 8.91$ while the normal function yields $\frac{1}{N}\ell_i(.) = 10.78$. Thus we determine that the exponential function has a better fit and that the data is generated from $h(z) = \exp(-z)$.

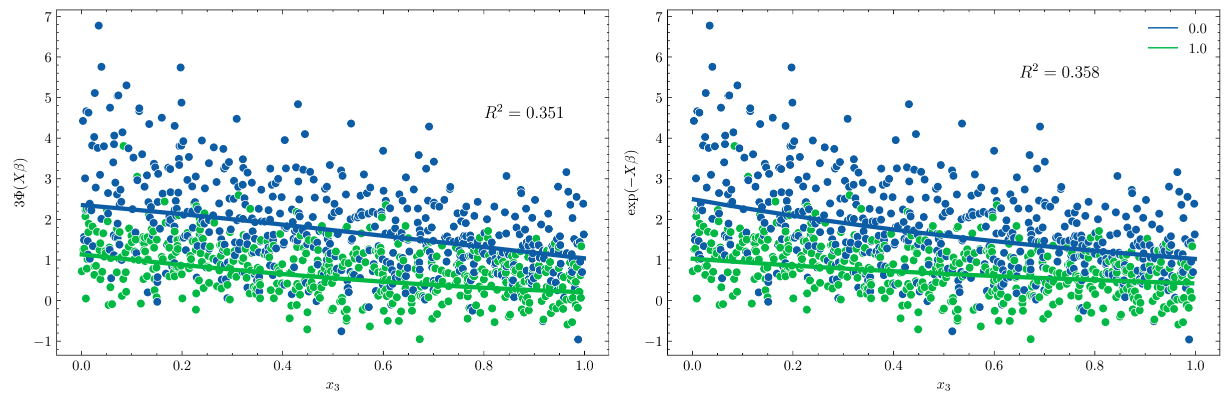

Figure 2.3: Panel Model Fit

2.8 Test null hypotheses

Finally, we test the two null-hypotheses

$$\mathcal{H}_0^1 : \beta_1 = \beta_2 = \beta_3$$
$$\mathcal{H}_0^2 : \beta_2 = \beta_3$$

In order to test test the hypotheses we need the covariance matrix of $\boldsymbol{\beta}$. $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is normally distributed with variance $\boldsymbol{A}_0^{-1}$ where $\boldsymbol{A}_0 := -\mathrm{E}\left[\boldsymbol{H}_i(\boldsymbol{\beta}_0)\right]$, with $\boldsymbol{H}_i(\boldsymbol{\beta}_0)$ being the double derivative of the log-likelihood contribution at $\boldsymbol{\beta}_0$ under two main conditions and some technical conditions. The first main condition is that $\boldsymbol{\beta}_0$ should be identified and compact to the parameter space. Secondly, the log-likelihood contribution should be continuous and twice differentiable on the interior of the parameter space. Thus we can compute the estimated the variance of $\hat{\boldsymbol{\beta}}$

$$\widehat{Avar(\hat{\boldsymbol{\beta}})} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{H}_i(\hat{\boldsymbol{\beta}})$$

Using the obtained covariance matrix we conduct Wald-tests of the two null-hypotheses. We use, that for a hypothesis $\boldsymbol{R\beta} = \boldsymbol{r}$ then, under the null, $W \equiv (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}\widehat{Avar(\hat{\boldsymbol{\beta}})}\boldsymbol{R}')(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ will be χ^2_Q distributed Q degrees of freedom, where Q is the number of hypothesis being tested. For $\mathcal{H}_0^1$ we set

$$\boldsymbol{R}^1 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}, \boldsymbol{r}^1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and for $\mathcal{H}_0^2$ we set

$$\boldsymbol{R}^2 = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}, \boldsymbol{r}^2 = \begin{pmatrix} 0 \end{pmatrix}$$

From this we get

Table 2.7: Wald tests

	W	p-value
$\beta_1 = \beta_2 = \beta_3$	871.9044	0.000
$\beta_2 = \beta_3$	0.0141	0.9054

Thus since $p < 0.05$ for $\boldsymbol{H}_0^1$ then we can reject the null hypothesis of $\beta_1 = \beta_2 = \beta_3$ with a confidence of 95 pct. For $\boldsymbol{H}_0^1$ we have $p > 0.05$, thus we cannot reject the null hypothesis of $\beta_2 = \beta_3$ with a confidence of 95 pct.

One thing to note regarding the Wald tests is the possibility of the standard errors not being valid. Looking at figure 2.3, it seems like here could be heteroskedasticity in the error terms, specifically c_i. We know that given x_i, c_i is distributed with constant variance, but that does not mean that the variance is constant across x_{it}'s Thus valid standard errors should be obtained using a robust variance estimator. For m-estimators in general, robust standard errors can be obtained by estimating the variance using the "sandwich" method, that is $Avar(\hat{\boldsymbol{\beta}}) = \boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}$ where $\boldsymbol{B}_0 = \mathrm{E}\left[\boldsymbol{s}_i(\boldsymbol{\beta}_0)\boldsymbol{s}_i(\boldsymbol{\beta}_0)'\right]$ with $\boldsymbol{s}_i(\boldsymbol{\beta}_0)$ being the first derivative of the log-likelihood contribution at $\boldsymbol{\beta}_0$.

References

Jeppesen, V. B., & Nielsen, J. A. R. (2021). Project 1.

Riis-Vestergaard Sørensen, J. (2021). *Linear model in high dimensions, 2: Implementation and theory (annotated).* University Lecture.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.