Analysis of the Variance of the Principal Component Regression Coefficients and the Estimated Outcome

Does Knowing the True Variance Covariance Matrix Decrease the Variance?

Manuel Huth*

August 27, 2020

Computational Statistics
University of Bonn
Summer Term 2020
Submitted to Prof. Dr. Lena Janys

*Address: Siemensstrasse 240, 53121 Bonn Germany, e-mail: *s6mahuth@uni-bonn.de*

Contents

1 Introduction

In economics, one is often interested in structural modelling which abstracts real world dependencies in a realistic theoretical set-up. For example Blundell et al. (2016) examine how female labour supply and human capital accumulation are affected by tax credit reforms. Fan et al. (2015) study, inter alia, the dependencies of schooling and wages. In the latter paper, the authors estimate a key reduced form equation that is similar to the original Mincer Equation (Mincer, 1974), which has been examined extensively in the economic literature, see Heckman et al. (2006) for a more recent review. Depending on the chosen covariates, models estimating wages have to deal with highly multicollinear data, and thus high variances of OLS coefficients, leading to high variances of the outcome estimates. For the outcome estimate it is desirable to have a low variance yielding a reliable estimate of the counterfactual outcome. Reducing the variance can be addressed applying principal components, that builds up on computing orthogonal projections of the original data such that the variance of the original data is inherited in the projections (James et al., 2013). Subsequent, an OLS regression is performed using the chosen projections. Since the projections are uncorrelated the issue of high multicollinearity is solved. The projection matrix consists of eigenvectors of the Variance covariance matrix and is therefore unique up to a sign flip of the projections (Jolliffe, 1986). In practice this variance covariance matrix is unknown and must be estimated using the given sample and thus the projection matrix is also an estimate.

In this paper I examine if the estimation of the projection matrix increases the variance of the regression coefficients and thus the variance of the estimated outcome relative to the variances obtained by using the true projection matrix. I do so by simulating data for an equation in style of the Mincer Equation, that sets wages in dependence to highly correlated covariates and should approximate realistic data. Subsequent I derive that simulating the distribution of the principal component regression coefficients is, due to the non-uniqueness of the projection matrix in the case of the estimated projection, not suitable to derive the unconditional variance of the estimated coefficients. I present another simulation strategy, to obtain proper variances of the estimates, that builds up on some algebra, and justify this strategy by comparing both cases using the true projection matrix. Finally, I compare the variances of the coefficients of the true and of the estimated projection matrices and find that there is no substantial difference. Furthermore, I validate that this leads to no difference in the variance of the estimated outcome.

The structure of the paper is the following. In section 2 I introduce how the data is build to yield the desirable features of the analysis. Subsequent, I show in section 3 how the principal components can be derived, how the principal component regression is applied and discuss two techniques to derive the variance of the estimators using a simulation study. In section 4, the simulation study, I simulate the data using the data generating process proposed in section 2,

compare the two techniques to derive the variance of the estimates and finally compare the variances of the coefficient estimates and the outcome estimates using the proper technique. Section 5 concludes and proposes extensions for further analysis. The appendix consists of calculations A and an exact description of how the data generating process is build B.

2 Data Generating process

To model the real hourly wages of an individual i I follow the well-known Mincer Equation (Mincer, 1974), that models the logarithmic wage as linear function of schooling and work experience as a second order polynomial. Additionally, I add the number of siblings and the years of parent's education and interpret the individual intercept coefficient as a persons ability. I follow Björklund and Kjellström (2002) and divide the squared work experience by 100 to have more numerical stability. The notation is the following $Y_i :=$ log-hourly wage of individual i, $\alpha :=$ constant term for all individuals, $a_i :=$ individual's ability, $s_i :=$ individual's years of schooling, $w_i :=$ individual's work experience, $n_i :=$ number of Siblings, $e_i :=$ parent's years of education, $\varepsilon_i :=$ normally distributed error and uncorrelated with the other variables and $T_{i,g,c} :=$ individual's test score in grade g of subject c.

$$\ln(Y_i) = \alpha + a_i + \beta_1 s_i + \beta_2 w_i + \beta_3 \frac{w_i^2}{100} + \beta_4 n_i + \beta_5 e_i + \varepsilon_i \tag{2.1}$$

Since ability is not observable in the real world, I use math and reading test scores from the seventh and eleventh grade as a proxy variable in the analysis. I have chosen this type of model since it is well-known and well-analyzed (see inter alia: Heckman et al. (2006); Lemieux (2006)) and all the covariates are (highly) correlated. Thus it provides a well-fitting set-up to examine the question whether the variance of the parameter estimates and the outcome estimates increases, if the true variance covariance matrix of the regressors is unknown in principal component regression. For the sake of this principal component analysis, I have omitted regional or gender specific dummy variables. However, it would be interesting to incorporate them in further analysis.

Setting up the data generating Process, I faced three main issues

1. there should be a reasonable correlation structure between the variables
2. there should not be impossible values, e.g. negative count variables
3. benchmark population moments from Blundell et al. (2005) should be met

My strategy was to first determine the dependencies (given in Figure 1), determining the respective scaling of all variables and finally determine equations to impose a structure that meets the given population moments and the dependence structure. The dependence structure

can be best illustrated using a causal graph 1.
Parent's with high education usually have fewer children (Cygan-Rehm and Maeder, 2013), ability influences test scores (Hansen et al., 2004) and parents with education tend to have children with more schooling years (Davis-Kean, 2005). A key assumption I make is that ability is random and not affected by any other variable, in particular parent's education. This assumption facilitates the derivation of the parameterization of the model, especially of the number of years of schooling of an individual, and seemed reasonable for the purposes of my data generating process. I have set up an extensive system of equations that describes

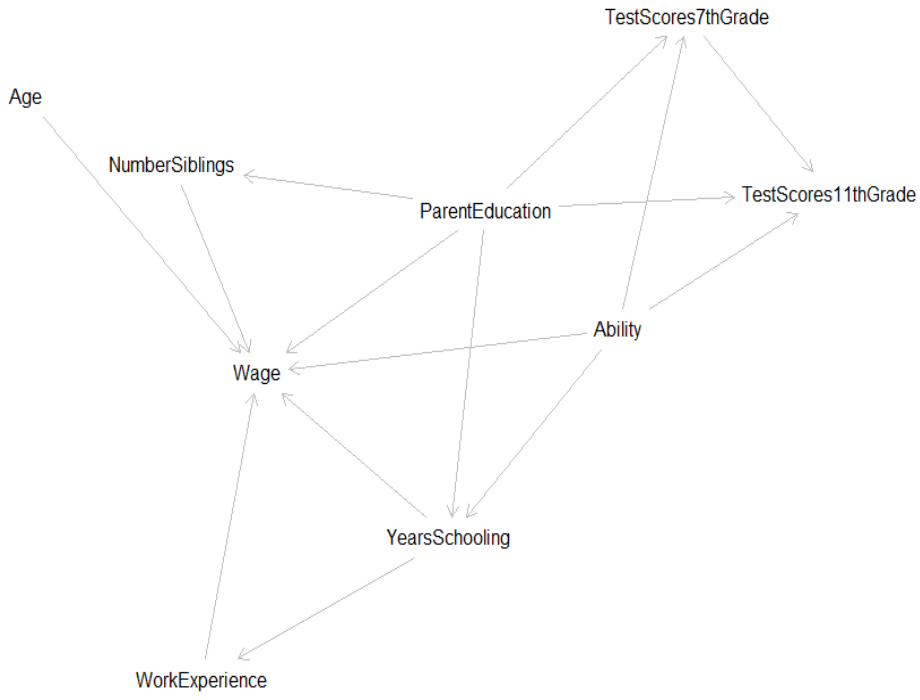

Figure 1: Causal Graph of Dependence Structure

these dependencies, which can be found in the appendix B.

3 Principal Component Methodology

In this section I show how the principal components can be computed and how the variance of the parameter estimate in the principal component regression can be derived in different ways using a simulation study. Let $I = 1, \dots p$ be an index set, $x_i \in \mathbb{R}^n \ \forall i \in I$ be a random vector, such that its entries are independent and follow the same distribution X_i that has finite first and second moments. For notation purposes let $X = \begin{pmatrix} X_1 & X_2 & \dots & X_p \end{pmatrix}'$. Note that it would be sufficient to assume that the expected values and variances are equal across entries

in x_i. However, for ease of notation I decided to stick to the case of equal distributions. Note that this definition yields the same results as defining a matrix containing random vectors from $\mathbb{R}^N, N > n$, that represent every observation from a finite sample population. To stay more general and allow the theoretical results to hold for infinite samples, I stick to the case of general random variables for the theory part. The variance covariance matrix of this random variables is denoted by $\mathbf{\Sigma}$. The collection of the x_i's vectors is defined by the sample matrix

$$\boldsymbol{X} = \begin{pmatrix} x_1 & x_2 & \dots & x_p \end{pmatrix} \in \mathbb{M}_{n \times p}. \tag{3.1}$$

It is assumed that the random vector x_i is already demeaned. For ease of notation it is therefore assumed that $\mathrm{E}(X_i) = 0, \ \forall i \in I$. Note that in some applications it might be desirable to use standardized variables with unit variances. However, since the scaling of the variables used in the regression mostly consists of count numbers and thus the same scales, I only normalize the data and thus stick in the methodology section to the case with normalized data.

3.1 Principal Component Analysis

The aim of the principal component analysis is to build M new vectors $z_1, z_2, \dots, z_M \in \mathbb{R}^n$ as orthogonal linear combinations of $x_1, x_2, \dots, x_p$. Note that $M \leq p$ since otherwise the z_m's cannot be orthogonal to each other. Denoting the scalars that are used to build z_m by $\phi_m \in \mathbb{R}^p$, one can express the m-th principal component as $z_m = \boldsymbol{X} \cdot \phi_m$. By defining $\boldsymbol{\phi} = \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_M \end{pmatrix}$ it is possible to shorten the above notation. This is useful to compute the values for all z_m in one equation.

$$\boldsymbol{Z} = \begin{pmatrix} z_1 & z_2 & \dots & z_M \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} \cdot \phi_1 & \boldsymbol{X} \cdot \phi_2 & \dots & \boldsymbol{X} \cdot \phi_M \end{pmatrix} = \boldsymbol{X}\boldsymbol{\phi}. \tag{3.2}$$

It will turn out that $\boldsymbol{\phi}$ is the matrix containing all eigenvectors with length one of $\mathbf{\Sigma}$ as columns. However, since an eigenvector with length one multiplied by minus one has the the same direction and length, this new vector can also be used as suitable eigenvector. Hence, there are 2^M different $\boldsymbol{\phi}$ and thus 2^M different $\boldsymbol{Z}$ matrices (proven in 2). From equation (3.2) it can be observed that flipping the sign of ϕ_m changes the sign of z_m, which will be relevant to derive that the coefficient estimates are variant to the choice of the sign (Theorem (3)).

I follow many textbooks and take $\boldsymbol{\phi}$ as deterministic (Jolliffe, 1986; Shlens, 2014). However, in practice it is a matrix to estimate and therefore adds additional randomness turning the matrix $\boldsymbol{Z}$ stochastic. Subsequent, I use hat notation for the quantities derived in an empirical set-up to distinguish the theoretical and the empirical cases. E.g $\hat{\phi}_m$ are the solutions to the empirical counterpart of (3.3), $\hat{\boldsymbol{\phi}}$ their collection in a matrix, $\hat{z}_m$ the estimated principal components and $\hat{\boldsymbol{Z}}$ their collection in a matrix. This two different approaches build the

foundations to obtain the estimators of interest in the principal components regression.

Since the derivation of $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ is very similar I only show the derivation of the former and briefly state the differences of the latter subsequent. For the derivation, I assume that the eigenvalues are distinct and thus Σ is ensured to be diagonalizable, which is the case in most applications. I derive all p possible principal components and therefore use p as highest index. The desired number of them can subsequent be adjusted by choosing the first M vectors with highest corresponding eigenvalues. The problem that has to be solved for all $m \in I$ is found to be

$$\begin{aligned} \phi_m = \quad & \arg\max_{w \in \mathbb{R}^p} \mathrm{Var}(Z_m) \\ \text{s.t.} \quad & Z_m = Xw \ \forall m \in I \\ & \mathrm{Cov}(Z_m, Z_{m'}) = 0 \ \text{ if } m \neq m' \ \ \forall m, m' \in I \\ & \mathrm{Var}(Z_m) \geq \mathrm{Var}(Z_{m+1}) \ \forall m \in I_{(-p)} \\ & ||w|| = 1 \end{aligned} \tag{3.3}$$

Theorem 1. *Let $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0, \lambda_i \in \mathbb{R}$ be eigenvalues of $\boldsymbol{\Sigma}$ and $v_1, v_2, \ldots, v_p \neq 0$ the corresponding eigenvectors with length one. Then $\phi_m = v_m \ \forall m \in I$ is a solution to the problem given in* (3.3).

Proof. Due to the length, the proof, a modification of Jolliffe (1986) that suits the set-up and notation presented in the paper, is given in A.2. □

Theorem 2. *There are 2^p solutions of problem* (3.3) *and 2^M if M principal components are used.*

Proof. From A.2 it is known that the solutions of (3.3) must be eigenvectors v_i of $\boldsymbol{\Sigma}$ with length one and corresponding eigenvalues $\lambda_i > 0$. Thus, $\boldsymbol{\Sigma} v_i = \lambda_i v_i$. Multiplying both sides with -1 yields that $-v_i$ is also an eigenvector of λ_i. Moreover, $1 = ||v_i|| = ||-v_i||$ by the properties of the norm. Since $v_i \neq 0$, v_i and $-v_i$ are distinct and thus there are two choices of an eigenvector corresponding to any of the p eigenvalues. Applying some combinatorics, this yields 2^p distinguished solutions, which reduces to 2^M, if M principal components are used. □

There is no clear theory which eigenvectors should be chosen. In subsection 3.2 I show that the choice affects the regression coefficients but not the estimated outcome. Moreover, I show in section 4 that the non-uniqueness of $\boldsymbol{\phi}$ restricts the way how the variance of the principal component regression coefficients can be obtained and that if these restrictions are taken into account, the variances of the parameter estimates are invariant to the choice of $\boldsymbol{\phi}$. For notation purposes I will denote the whole class of solutions to the specific principal component analysis problem (3.3) as $[\phi]$ and any representative of this class by $\boldsymbol{\phi}$.
In an empirical set-up one would replace $\mathrm{Var}(Z_m)$ by its empirical estimate $\frac{1}{N-1}\hat{z}_m'\hat{z}_m$ and the

vector of random variables X by the matrix of empirical outcomes $\boldsymbol{X}$. The solution is found to be the eigenvectors of the estimated variance covariance matrix $\hat{\boldsymbol{\Sigma}}$.

3.2 Principal Component Regression

In this subsection I will show how the estimator of the principal component regression can be derived using principal component analysis. Moreover, I show the differences in the estimates that occur due to using either $\boldsymbol{\phi}$ or $\hat{\boldsymbol{\phi}}$. First I assume a linear true relationship of the expected value of Y and γ

$$Y = \boldsymbol{X}\gamma + \varepsilon, \tag{3.4}$$

whereby $\varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma_x^2 \boldsymbol{I}_{n\times n}\right)$ is a vector of uncorrelated error terms with conditional mean zero and $\gamma \in \mathbb{R}^p$.

The idea of principal component regression is that Y can be estimated in a linear model by using the constructed M principal components. The advantage compared to an ordinary OLS model is that often a small number of principal components is sufficient to explain the variability in X and its correlation with Y (Jolliffe, 1986). In this paper I focus on the variance of the principal component regression coefficient estimates and the variance of the predicted outcome and therefore always use all principal components. In equation (3.6) I show that the respective betas are invariant to the choice of M and thus using all principle components yields more information for the purposes of my study.

The line of argumentation on the usefulness of principal component regression in the literature, for example in Friedman et al. (2001) and James et al. (2013), is reflected by: Imagine a case where the number of regressors p is close to the number of observations n (a) or a case of very high multicollinearity (b). In such a case the OLS estimate tends to overfit the data (a) or yields high variance of the coefficients and therefore of the estimated outcome (b). The former feature (a) is not part of this paper but combining both issues (a) and (b) in one study would be a desirable extension. Since all principal component are uncorrelated, see A.1, principal component regression circumvents the problem of high multicollinearity (b).

This is true if the true variance covariance matrix $\boldsymbol{\Sigma}$ is known. However, as stated in the previous subsection, in practice $\boldsymbol{\Sigma}$ is unknown and $\boldsymbol{\phi}$ is estimated by the eigenvectors of $\hat{\boldsymbol{\Sigma}}$. So instead of p parameters to estimate, there are M parameters to estimate in the regression and p vectors to estimate before the regression. The latter adds additional randomness to the M estimated parameters and thus also to the estimate of the outcome. Subsequent I derive the necessary tools to set up a simulation study that addresses this issue in section 4.

Quantities denoted with a superscript t are used in the theoretical framework and quantities denoted with a superscript s are computed using the stochastic $\hat{\boldsymbol{\phi}}$.

The equations to be estimated are

$$\begin{aligned} Y &= \boldsymbol{Z}\beta^t + \varepsilon_Z^t = \boldsymbol{X}\boldsymbol{\phi}\beta^t + \varepsilon_Z^t, \\ Y &= \hat{\boldsymbol{Z}}\beta^s + \varepsilon_Z^s = \boldsymbol{X}\hat{\boldsymbol{\phi}}\beta^s + \varepsilon_Z^s, \end{aligned} \tag{3.5}$$

whereby ε_Z^t and ε_Z^t are again vectors of independent homoscedastic error terms with conditional expectation of zero and $\beta^t, \beta^s \in \mathbb{R}^M$ are the parameter vectors of interest.
Since in a simulation study drawing different samples at each iteration $\hat{\boldsymbol{\phi}}$ is estimated at each iteration, the variance estimate of $\hat{\beta}^s$ simulating the distribution yields a higher variance induced by different $\hat{\beta}^s$ that are obtained using different $\hat{\boldsymbol{\phi}}$. Since $\boldsymbol{\phi}$ is chosen prior to the simulation, this problem does not appear for $\hat{\beta}^t$.

Theorem 3. *$\hat{\beta}^s$ is variant to the choice of the signs of the eigenvectors in $\hat{\boldsymbol{\phi}}$.*

Proof. $\hat{\beta}^s$ can be computed as an OLS estimate of (3.5). The solution is therefore given by

$$\begin{aligned} \hat{\beta}^s =& \arg\min_{b\in\mathbb{R}^M} ||Y - \hat{\boldsymbol{Z}}b||_2^2 = \left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}\hat{\boldsymbol{Z}}'Y = \left(\hat{\boldsymbol{\phi}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\phi}}\right)^{-1}\hat{\boldsymbol{\phi}}'\boldsymbol{X}'Y \\ =& \operatorname{diag}\begin{pmatrix}\lambda_1^{-1} & \dots & \lambda_M^{-1}\end{pmatrix}\hat{\boldsymbol{Z}}Y = \begin{pmatrix}\frac{\hat{z}_1'Y}{\hat{\lambda}_1} & \dots & \frac{\hat{z}_M'Y}{\hat{\lambda}_M}\end{pmatrix}'. \end{aligned} \tag{3.6}$$

From the last term it follows that flipping the sign of $\hat{z}_m$ flips the sign of $\hat{\beta_m^s}$. An alternative way of proving is given in the appendix. □

Hence, an alternative way of computing the variance of $\hat{\beta}^s$ in a simulation study must be found. I use some algebra to find a representation that is only dependent on the eigenvalues, which are unique, and not dependent on the eigenvectors.

Theorem 4. *The unconditional variance of $\hat{\beta}^s$ is given by*

$$Var(\hat{\beta}^s) = E\left[\hat{\sigma}_z^2 diag\begin{pmatrix}\lambda_1^{-1} & \dots & \lambda_M^{-1}\end{pmatrix}\right] \tag{3.7}$$

Proof. First, I derive that the estimate $\hat{\beta}^s$ is conditioned on $\hat{\boldsymbol{Z}}$ unbiased (ii).

$$\begin{aligned} \mathrm{E}\left(\hat{\beta^s} \mid \hat{\boldsymbol{Z}}\right) =& \mathrm{E}\left[\left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}\hat{\boldsymbol{Z}}'Y \mid \hat{\boldsymbol{Z}}\right] = \mathrm{E}\left[\left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}\hat{\boldsymbol{Z}}'(\hat{\boldsymbol{Z}}\beta^s + \varepsilon_Z) \mid \hat{\boldsymbol{Z}}\right] \\ =& \beta_s + \mathrm{E}\left[\left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}\hat{\boldsymbol{Z}}'\varepsilon_Z \mid \hat{\boldsymbol{Z}}\right] = \beta^s + \left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}\hat{\boldsymbol{Z}}'\underbrace{\mathrm{E}\left[\varepsilon_Z \mid \hat{\boldsymbol{Z}}\right]}_{=0} = \beta^s \end{aligned} \tag{3.8}$$

Hence, the conditional variance of $\hat{\boldsymbol{Z}}$ can be computed as in the OLS case yielding $\mathrm{Var}(\hat{\beta}^s|\boldsymbol{Z}) = \sigma_z^2\left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}$ (iii). The latter three findings are subsequent used to compute the variance of $\hat{\boldsymbol{Z}}$,

making use of the law of total variance (A.3).

$$\begin{aligned}\mathrm{Var}(\hat{\beta}^s) \overset{\text{i}}{=} &\mathrm{E}\left[\mathrm{Var}(\hat{\beta}^s|\hat{\boldsymbol{Z}})\right] + \mathrm{Var}\left[\mathrm{E}(\hat{\beta}^s|\hat{\boldsymbol{Z}})\right] \overset{\text{ii}}{=} \mathrm{E}\left[\mathrm{Var}(\hat{\beta}^s|\hat{\boldsymbol{Z}})\right] \\ \overset{\text{iii}}{=} &\mathrm{E}\left[\hat{\sigma}_z^2\left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}\right] \mathrm{E}\left[\hat{\sigma}_z^2 \mathrm{diag}\begin{pmatrix}\lambda_1^{-1} & \dots & \lambda_M^{-1}\end{pmatrix}\right] \end{aligned} \tag{3.9}$$

□

To validate that using the formula (3.7) is equal as using the simulated distribution, I derive a similar formula in the same way for $\hat{\beta}^t$ and show in the simulation study 4 that its outcome is equal to the the variance obtained by the simulated distribution of $\hat{\beta}^t$.

Theorem 5. *The unconditional variance of $\hat{\beta}^t$ is given by*

$$Var(\hat{\beta}^t) = E\left[\hat{\sigma}_z^2\left(\boldsymbol{\phi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\phi}\right)^{-1}\right] \tag{3.10}$$

Proof. I make use of the well-known fact that the OLS estimate is unbiased conditional on Z, thus the conditional expectation is constant (ii), and that the conditional variance is given by $\mathrm{Var}(\hat{\beta}^t|\boldsymbol{Z}) = \sigma_z^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$ (iii). Furthermore, I apply the law of total variance (i) (A.3). If the variance of ε_Z must be estimated the variance of $\mathrm{Var}(\hat{\beta}^t)$ is derived by

$$\begin{aligned}\mathrm{Var}(\hat{\beta}^t) \overset{\text{i}}{=} &\mathrm{E}\left[\mathrm{Var}(\hat{\beta}^t|\boldsymbol{Z})\right] + \mathrm{Var}\left[\mathrm{E}(\hat{\beta}^t|\boldsymbol{Z})\right] \overset{\text{ii}}{=} \mathrm{E}\left[\mathrm{Var}(\hat{\beta}^t|\boldsymbol{Z})\right] \overset{\text{iii}}{=} \mathrm{E}\left[\hat{\sigma}_z^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right] \\ = &\mathrm{E}\left[\hat{\sigma}_z^2\left(\boldsymbol{\phi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\phi}\right)^{-1}\right] \end{aligned} \tag{3.11}$$

□

Even though $\hat{\beta}^s$ is variant to the choice of the eigenvectors, $\hat{Y}^s$ is invariant and thus the variance of it can be computed simulating the distribution. That $\hat{Y}^t$ is invariant to the choice of the eigenvectors follows trivially since the eigenvectors are chosen prior to the simulation study.

Theorem 6. *$\hat{Y}^s$ is invariant to the choice of the signs of the eigenvectors in $\hat{\boldsymbol{\phi}}$.*

Proof. Using (3.6) it follows

$$\hat{Y^s} = \hat{Z}\hat{\beta}^s = \hat{Z}\begin{pmatrix}\frac{\hat{z}_1'Y}{\hat{\lambda}_1} & \dots & \frac{\hat{z}_M'Y}{\hat{\lambda}_M}\end{pmatrix}' = \sum_{m=1}^{M}\frac{\hat{z}_m\hat{z}_m'Y}{\hat{\lambda}_m} \tag{3.12}$$

From (3.2) it follows that a sign flip of $\hat{\phi}_m$ only flips the sign of $\hat{z}_m$. As derived above, $\hat{Y^s}$ is invariant to a sign flip of $\hat{z}_m$ and thus invariant to a sign flip of $\hat{\phi}_m$. □

3.3 Estimate the Variance of the Coefficients in a Simulation Study

If the notation $\hat{\beta}$ is used in this section, this indicates that the given statement holds for for $\hat{\beta}^s$ and $\hat{\beta}^t$. Since $\text{Var}(\hat{Y}^s)$ and $\text{Var}(\hat{Y}^t)$ are invariant to the choice of the matrix of eigenvectors, they can be computed using the simulated distribution as presented for beta subsequently.

3.3.1 Framework to Use Simulated Distributions of the Coefficients to Estimate the Coefficients' Variances in a Simulation Study

To compute the simulated distributions, I fix a specific sample size n and simulate I different samples yielding I results for $\hat{\beta}$ denoted by $\hat{\beta^{(i)}} \quad i = 1, \ldots, I$ respectively. This beta sample serves as an estimate of the simulated distribution of $\hat{\beta}$. Denoting the simulation sample mean of $\hat{\beta}^{(i)}$ by $\overline{\hat{\beta}}$, an estimate of the variance $\text{Var}(\hat{\beta})$ can therefore be obtained by

$$\widehat{\text{Var}(\hat{\beta})}_n^E = \frac{1}{I-1} \sum_{i=1}^{I} \left(\hat{\beta}^{(i)} - \overline{\hat{\beta}} \right)^2 \tag{3.13}$$

$\widehat{\text{Var}(\hat{\beta})}_n^E$ denotes the estimate of the variance of $\hat{\beta}$ using the simulated empirical distributions (E) and sample sizes of n. As shown in Theorem 3 $\hat{\beta}^s_{(i)}$ is dependent on the choice of $\hat{\boldsymbol{\phi}}$ and therefore the coefficients vary more than they would do if the representation of $[\hat{\phi}]$ would be unique and thus the variance of the coefficients increase. I examine the size of this effect by comparing the empirically simulated variance of $\hat{\beta}^t$ and the one using simulations of Theorem 5 in the simulation study in section 4 and find that using Theorem 5 serves a lower bound of the empirically simulated variance.

3.3.2 Framework to Use Formula the Algebraically Derived Formulas to Estimate the Coefficients' in a Simulation Study

As stated above, the non-uniqueness of the matrix of eigenvectors in the practical case raises a need to estimate the variance of the coefficient differently. Let I again denote the number of iterations, n a fixed sample size, $\hat{\lambda}_k^{(i)}$ the k-th eigenvalue of the i-th sample's variance covariance matrix, $X^{(i)}$ the matrix of regressors at iteration i and $(\hat{\sigma}_z^2)_{(i)}$ the estimated error variance at iteration i. According to the stated formulas the variance of $\hat{\beta}$ can be estimated in a simulation study by computing the simulation sample averages

$$\widehat{\text{Var}(\hat{\beta}^s)}_n^F = \frac{1}{I} \sum_{i=1}^{I} (\hat{\sigma}_z^2)_{(i)} \begin{pmatrix} \frac{1}{\hat{\lambda}_1^{(i)}} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{\hat{\lambda}_M^{(i)}} \end{pmatrix} \tag{3.14}$$

$$\widehat{\text{Var}(\hat{\beta}^t)}_n^F = \frac{1}{I}\sum_{i=1}^{I}(\hat{\sigma}_z^2)_{(i)}\left(\boldsymbol{\phi}'\boldsymbol{X}'^{(i)}\boldsymbol{X}^{(i)}\boldsymbol{\phi}\right)^{-1} \tag{3.15}$$

In the simulation study in section 4 I show that for $\hat{\beta}^t$ the procedure in (3.13) and the one presented in (3.15) yield the same results but for $\hat{\beta}^s$ the procedure in (3.13) yields a higher variance than the one presented in (3.14), due to the non-uniqueness of $\hat{\boldsymbol{\phi}}$.

4 Simulation

Armed with the theory of the data generating process and the variance of the principal component regression coefficients, I examine how $\text{Var}(\hat{\beta}^s)$, $\text{Var}(\hat{\beta}^t)$ and $\text{Var}(\hat{Y}_i)$ behave for different sample sizes. The results show that if the strategy from equation (3.14) is used to compute the $\text{Var}(\hat{\beta}^s)$, the results for $\text{Var}(\hat{\beta}^s)$ and $\text{Var}(\hat{\beta}^t)$ do not differ.

4.1 Simulate the whole Population

First I simulate the whole population. To check if the simulated data meets the requirements imposed in B, I plot histograms of any variable that was generated and observe that the requirements seem to be met for all variables.

To increase the validity of the simulated data, I compared the first two simulated population moments with the imposed moments reported in Blundell et al. (2005). All variables have means that are close to the imposed means. The variances deviate of around 13% to 14% for schooling, number of siblings and the logarithmic wage. However, since these numbers are small in magnitude, I see this as reasonable population moments for the upcoming principal component analysis.

4.2 Variance of the Principal Component Regression Coefficients $\hat{\beta}$ in a Simulation Study

Subsequent I derive a 3-step strategy to validate that $\widehat{\text{Var}(\hat{\beta}^s)}_n^F$ and $\widehat{\text{Var}(\hat{\beta}^t)}_n^F$ build reasonable quantities on how to decide if the variance of the principal component regression coefficients is larger if the true eigenvectors are unknown.

1. I compare $\widehat{\text{Var}(\hat{\beta}^t)}_n^E$ and $\widehat{\text{Var}(\hat{\beta}^t)}_n^F$ and derive that the simulated distribution of $\hat{\beta}^t$ yields the same variance estimates as using formula (3.15) to estimate the variance.
2. I compare $\widehat{\text{Var}(\hat{\beta}^s)}_n^E$ and $\widehat{\text{Var}(\hat{\beta}^s)}_n^F$ and observe that the variance estimates using the

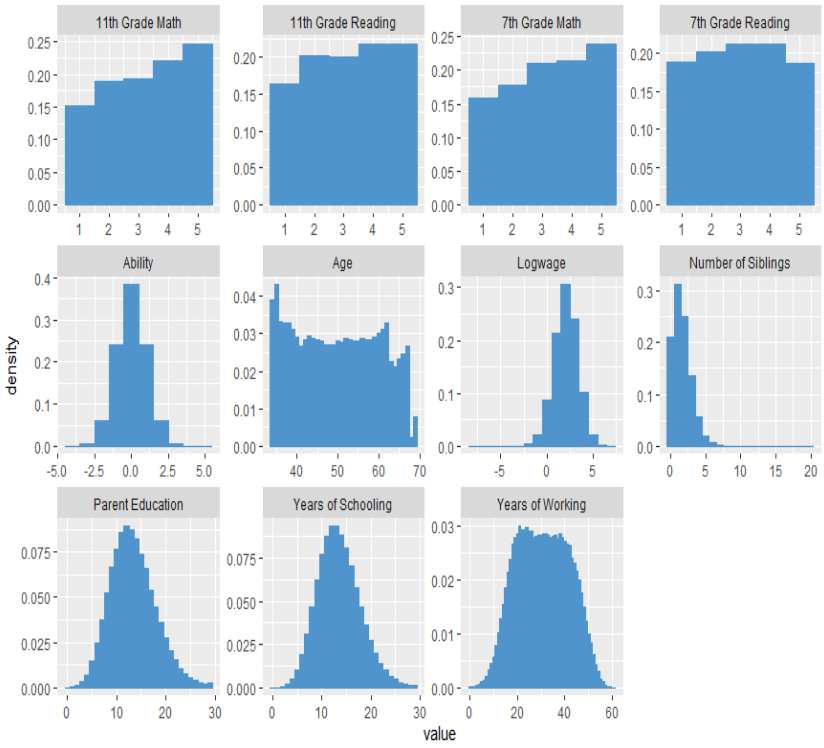

Figure 2: Histograms of the Simulated Population Variables

simulated distribution are for most parameters significantly higher and for some parameters equal than the estimates obtained by using formula 3.14.

3. I compare $\widehat{\text{Var}(\hat{\beta}^s)}_n^F$ and $\widehat{\text{Var}(\hat{\beta}^t)}_n^F$ and find the there is no remarkable difference in the variances.

I have chosen sample sizes of $n = 50, 80, 100, 150, 200, 400, 600, 800, \ldots, 5000$ to check the validity of the results in small and large samples. I chose I, the number of iterations (see formulas (3.13), (3.14) and (3.15)), to be 400, which seemed large enough to build the simulated distributions of the moments given in the stated formulas. To ensure robustness of my results I have simulated the variances 400 times for each sample size.

4.2.1 Compare the Variances of the Theoretical Case using Simulated Distributions and the Simulated Algebraically Derived Formula

I plot the mean of the simulated variances per sample size and the simulated quantiles of the variances (transparent areas), even though the latter cannot be seen since they are too small, as a function of the corresponding sample sizes. As theory suggests, the variances increase in an ascending order of the beta coefficients. The means of the variances and the quantiles nearly

coincide and thus I conclude that there is no remarkable difference in using(3.13) or (3.15). Hence, the formulas derived in (3.14) and (3.15) are reasonable estimates of the true variance and therefore it seems reasonable to use $\widehat{\text{Var}(\hat{\beta}^t)}_n^F$ and $\widehat{\text{Var}(\hat{\beta}^s)}_n^F$ rather than the variances obtained by simulated distributions, if the variance derived from the simulated distribution $\widehat{\text{Var}(\hat{\beta}^s)}_n^E$ is greater than $\widehat{\text{Var}(\hat{\beta}^s)}_n^F$. I show that this is the case in the next subsection.

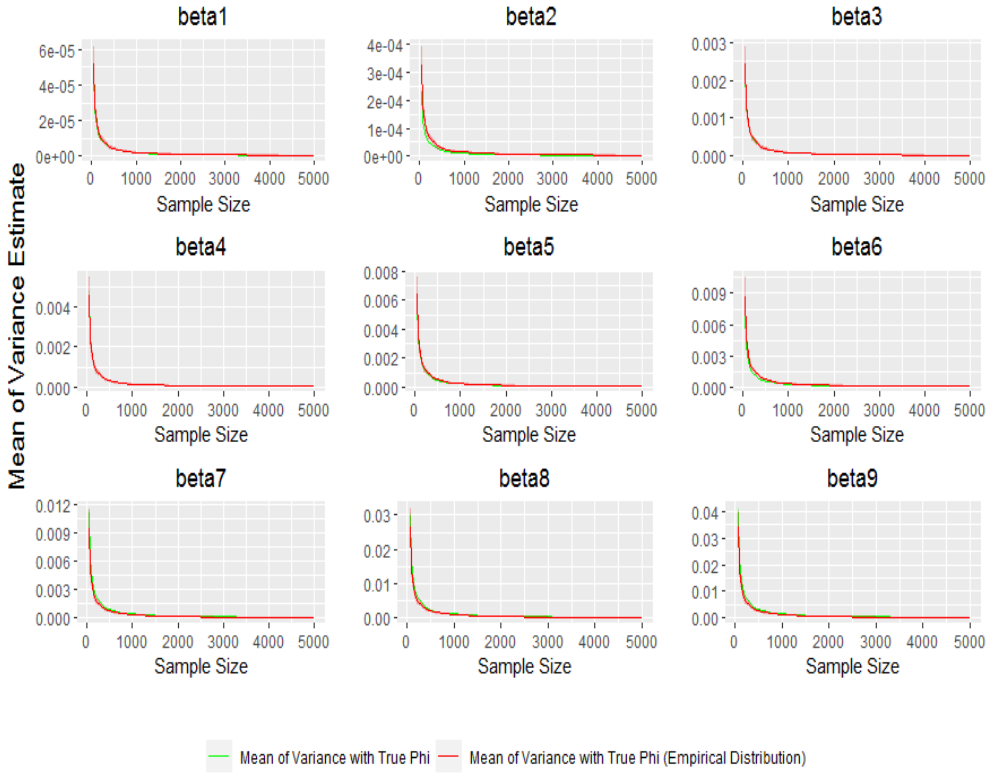

Figure 3: Variances of Regression Coefficients Using the Theoretical Matrix of Eigenvectors Computed by Analytical Formula and Simulated Distribution

4.2.2 Compare the Variances of the Empirical Case using Simulated Distributions and the Simulated Algebraically Derived Formula

Again, I plot the mean of the simulated variances per sample size and the simulated quantiles of the variances (transparent areas), as a function of the corresponding sample sizes. Contrarily to the non-stochastic case, the variances computed with formula (3.14) are all lower or at least equal to the variances computed with (3.13) and thus can serve as an lower bound. From the findings I conclude that using the results of the simulated distribution can lead to misleading results when the variance of the coefficients should be compared. Summed up I conclude that

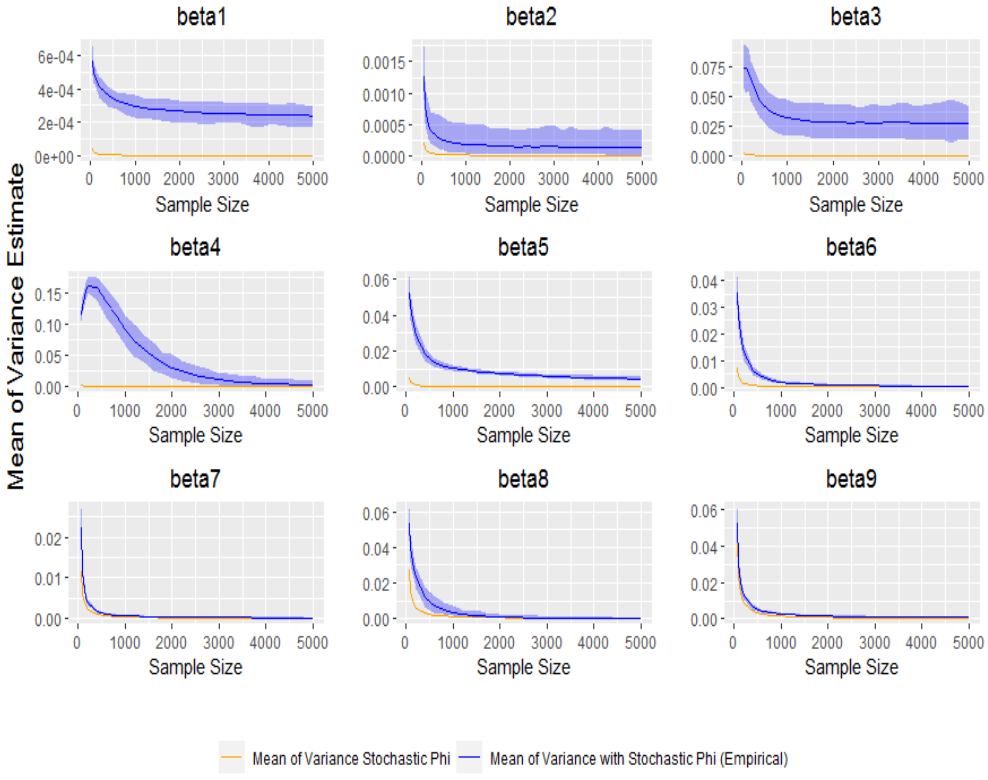

Figure 4: Variances of Regression Coefficients Using the Stochastic Matrix of Eigenvectors Computed by Analytical Formula and Simulated Distribution

$\widehat{\text{Var}(\hat{\beta}^s)}_n^F$ and $\widehat{\text{Var}(\hat{\beta}^t)}_n^F$ should be used to compare the variances, which is done subsequent.

4.2.3 Compare the Variances of the Empirical and the Theoretical Coefficient Estimates

The means and the quantiles of the variances $\widehat{\text{Var}(\hat{\beta}^t)}_n^F$ and $\widehat{\text{Var}(\hat{\beta}^s)}_n^F$ are basically the same for nearly any sample size. For some parameters in small samples ($n = 50, 80$) it is even the case that $\hat{\beta}^t$ has a greater mean of the variances and its quantiles are higher. From here I conclude that there is no increase of the variances of the principal component regression coefficients when either the true matrix of eigenvectors $\boldsymbol{\phi}$ or the estimated $\hat{\boldsymbol{\phi}}$ matrix is used. This finding should imply that the variances of $\hat{Y}^s$ and $\hat{Y}^t$ do not differ as well. This is examined and confirmed subsequent.

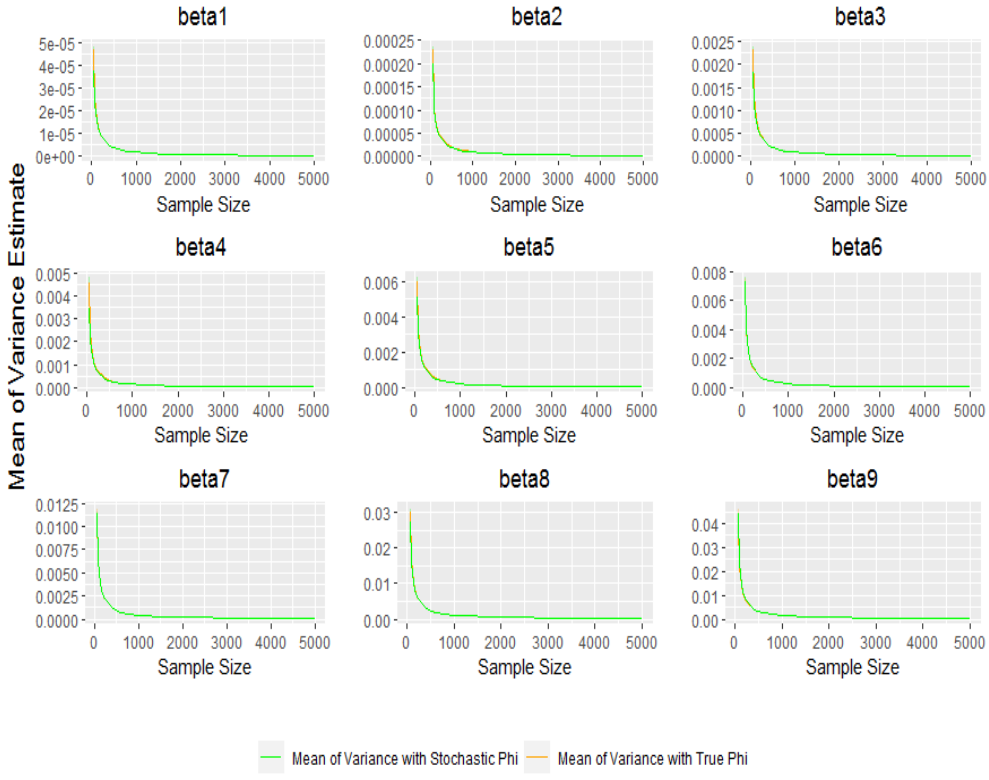

Figure 5: Variances of Regression Coefficients Using the Stochastic Matrix of Eigenvectors vs. the Theoretical Matrix of Eigenvectors

4.3 Variance of Estimated Outcomes

Since $\hat{Y^s}$ is invariant to the choice of the eigenvectors (3.12) and $\hat{Y^t}$ is invariant since its representative is chosen prior to the regressions, I use the simulated distributions of the random variables to derive the unconditional variances. As expected from 4.2.3 there is no difference in the variance of $\hat{Y^s}$ and $\hat{Y^t}$.

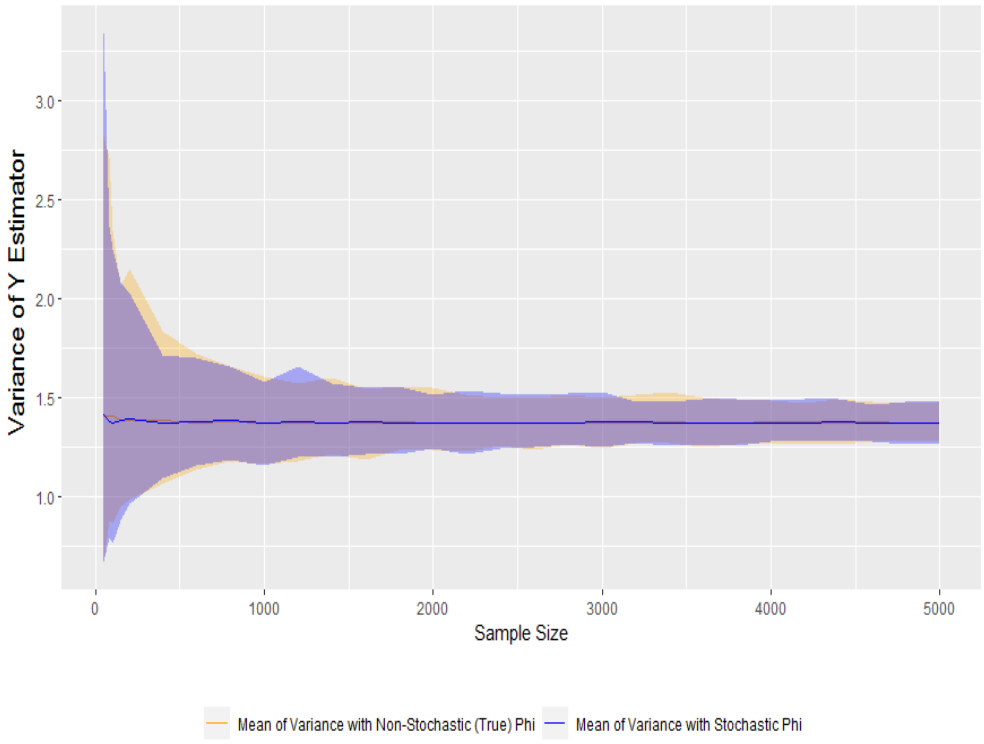

Figure 6: Variances of Wage Estimates Using the Stochastic Matrix of Eigenvectors vs. the Theoretical Matrix of Eigenvectors

5 Conclusion

I have used a data generating process, that builds on various assumptions, is able to reflect given first and second moments, to build a data set that sets wages in dependence of other socioeconomic factors and to yield realistic dependence structures. I used this data to examine how the variance of the principal component regression coefficients and the variance of the estimated wages differ, if the principal components are unknown and must be estimated with the empirical observations. I did so by algebraically deriving necessary conditions on how the estimates of the variances can be obtained in a simulation study. In particular I find that simulating the distribution is feasible for the estimated outcome but infeasible for the coefficients, which must be simulated by using an algebraic representation of the variance. I validate that this algebraically derived procedures can be applied in the simulation study. The final results yield that the variances of all estimated coefficients and of the estimated wages, computed using the estimated principal components, are not substantial different to the variances obtained using the theoretical principal components. Thus, I find that there is no considerable loss of accuracy using the estimated variance covariance matrix estimating

counter factual outcomes.
For further analysis it is desirable to study whether the results change, if only a subset of all principal components is used. Even though this would not change the variances using the estimated principal components, the variances using the theoretical principal components could change. Moreover, it would be interesting to see in a simulation study how fast the eigenvalues of the estimated variance covariance matrix converge to the true eigenvalues and if there is convergence of the absolute values of the eigenvectors.
Additionally, it might be desirable to increase the number of regressors to see how the variance behaves in high-dimensional cases and if the derived theory can help to perform dimension reduction.

Appendix A Formulas and Further Derivations

A.1 Distribution of Z_m Random Variables

The exact distributions are dependent on the distributions of the X_i variables, their convolution properties and the eigenvectors ϕ_m. Since all n random variables in z_m come from the same distribution Z_m, they are all the same linear combinations of random vectors x_i with random variables X_i given by

$$Z_m = \begin{pmatrix} X_1 & X_2 & \dots & X_p \end{pmatrix} \phi_m \iff \mathcal{L}(Z_m) = \mathcal{L}\left(\begin{pmatrix} X_1 & X_2 & \dots & X_p \end{pmatrix} \phi_m\right) \tag{A.1}$$

Theorem 7. *The expected value of each principle component is zero and thus invariant to the choice of the eigenvectors.*

Proof.

$$\mathrm{E}(Z_m) = \mathrm{E}(\begin{pmatrix} X_1 & X_2 & \dots & X_p \end{pmatrix} \cdot \phi_m) = \mathrm{E}\left(\sum_{j=1}^{p} X_j \phi_{jm}\right) = \sum_{j=1}^{p} \underbrace{E(X_j)}_{=0} \phi_{jm} = 0. \tag{A.2}$$

□

Theorem 8. *The variances of the principal components are equal to the respective eigenvalues and the covariances are zero.*

Proof. Since $\mathrm{E}(Z_m) = 0 \ \forall m$

$$\begin{aligned}
\boldsymbol{\Sigma}_Z =& \mathrm{E}\left(\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_M \end{pmatrix} \begin{pmatrix} Z_1 & Z_2 & \dots & Z_M \end{pmatrix}\right) = \boldsymbol{\phi}' \mathrm{E}\left(\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \begin{pmatrix} X_1 & X_2 & \dots & X_p \end{pmatrix}\right) \boldsymbol{\phi} \\
=& \boldsymbol{\phi}' \boldsymbol{\Sigma} \boldsymbol{\phi} \overset{\text{eigendecomposition}}{=} \boldsymbol{\phi}' \begin{pmatrix} v_1 & \dots & v_M \end{pmatrix} \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & \lambda_M \end{pmatrix} \begin{pmatrix} v_1' \\ \vdots \\ v_M' \end{pmatrix} \boldsymbol{\phi} \\
=& \underbrace{\boldsymbol{\phi}' \boldsymbol{\phi}}_{=I_{M\times M}} \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & \lambda_M \end{pmatrix} \underbrace{\boldsymbol{\phi}' \boldsymbol{\phi}}_{=I_{M\times M}} = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & \lambda_M \end{pmatrix}
\end{aligned} \tag{A.3}$$

□

Note that since $\boldsymbol{\Sigma}$ is a positive definite matrix, it is ensured that the variance of Z_m is always positive. Moreover, since the eigenvalues are equal for each representative in $[\phi]$, the

result does not change for different choices of the representative.

A.2 Derivation of Principal Components

Proof of Theorem 1.

Proof. First, note that since $\mathbf{\Sigma}$ is a positive definite $p \times p$ matrix, it has p real-valued eigenvalues greater than zero and therefore all λ_i and v_i exist.
The problem is first divided into M subproblems for each ϕ_m. The first subproblem is given by

$$\phi_1 = \arg \max_{||w||=1} \operatorname{Var}(Z_1) = \arg \max_{||w||=1} w' \mathbf{\Sigma} w \tag{A.4}$$

Since it is the This problem can be solved using the Lagrangian with the constraint $w'w = 1$, which is equivalent to $||w||^2 = 1$ that follows from $||w|| = 1$.

$$\begin{aligned} L(w, \lambda) =& w' \mathbf{\Sigma} w - \lambda \left(w'w - 1 \right) \\ \frac{\partial L}{\partial \lambda} =& w'w - 1 = 0 \\ \frac{\partial L}{\partial w} =& 2\mathbf{\Sigma} w - 2\lambda w = 0 \end{aligned} \tag{A.5}$$

From equation (A.5) it follows that $(\mathbf{\Sigma})\, w = \lambda w$. Hence, the solution vector w is an eigenvector of $\mathbf{\Sigma}$ with eigenvalue λ. Armed with the eigenvectors and eigenvalues, the question arises which eigenvalue to use to maximize the variance.

$$\begin{aligned} &\arg \max_{\lambda \in \{\lambda_1, \ldots, \lambda_p\}} w' \mathbf{\Sigma} w \overset{(A.5)}{=} \arg \max_{\lambda \in \{\lambda_1, \ldots, \lambda_p\}} w' \lambda w \\ =& \arg \max_{\lambda \in \{\lambda_1, \ldots, \lambda_p\}} \underbrace{w'w}_{1} \lambda = \arg \max_{\lambda \in \{\lambda_1, \ldots, \lambda_p\}} \lambda = \lambda_1 \end{aligned} \tag{A.6}$$

Hence, the maximum variance is captured by choosing the largest eigenvalue, which is by definition, λ_1. The corresponding eigenvector v_1 ischosen to be ϕ_1. Thus, it is enough to compute the largest eigenvalue of the variance covariance matrix of all the X_i to compute ϕ_1 and with that z_1. Armed with ϕ_1, it is possible to compute $\phi_2, \phi_3, \ldots, \phi_M$ iteratively by defining new random variables

$$\begin{pmatrix} X_1^{(m)} \\ X_2^{(m)} \\ \vdots \\ X_p^{(m)} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} - \sum_{j=1}^{m-1} Z_j \phi_j \tag{A.7}$$

and solving the former problem for $\mathrm{Var}(Z_m)$ using the respective random variables of the m-th iteration $\begin{pmatrix} X_1^{(m)} & X_2^{(m)} & \dots & X_p^{(m)} \end{pmatrix}'$ to get ϕ_m. It turns out that $\boldsymbol{\phi}$ equals the matrix $\boldsymbol{V} = \begin{pmatrix} v_1 & v_2 & \dots & v_M \end{pmatrix}$, whereby v_i is the eigenvector with length one of $\boldsymbol{\Sigma}$ to the corresponding i-th highest eigenvalue λ_i (Jolliffe, 1986). Hence, $\boldsymbol{\phi}$ can computed by only computing the eigenvectors of $\boldsymbol{\Sigma}$ and sort them in descending order by the corresponding eigenvalues.

□

A.3 Law of Total Variance

Theorem 9. *Let $y \in \mathbb{M}_{p\times 1}$ be a random vector and x be a random matrix. The variance of y can be decomposed by*

$$Var(y) = E\left[Var(y|x)\right] + Var\left[E(y|x)\right] \tag{A.8}$$

Proof.

$$\begin{aligned}
\mathrm{Var}(y) =& \mathrm{E}(yy') - \mathrm{E}(y)\mathrm{E}(y') \\
=& \mathrm{E}\left(\mathrm{E}(yy'|x)\right) - \mathrm{E}\left[\mathrm{E}(y|x)\right]\mathrm{E}\left[\mathrm{E}(y'|x)\right] \\
=& \mathrm{E}\left(\mathrm{Var}(y|x) + \mathrm{E}(y|x)\mathrm{E}(y'|x)\right) - \mathrm{E}\left(\mathrm{E}(y|x)\right)\mathrm{E}\left(\mathrm{E}(y'|x)\right) \\
=& \mathrm{E}\left[\mathrm{Var}(y|x)\right] + \mathrm{E}\left[\mathrm{E}(y|x)\mathrm{E}(y'|x)\right] - \mathrm{E}\left[\mathrm{E}(y|x)\right]\mathrm{E}\left[\mathrm{E}(y'|x)\right] \\
=& \mathrm{E}\left[\mathrm{Var}(y|x)\right] + \mathrm{Var}\left[\mathrm{E}(y|x)\right],
\end{aligned} \tag{A.9}$$

whereby it is made us of the law of iterated expectations in the second row. The other reshapings are basically applications of the same law given in the first line. □

A.4 Alternative proof of Theorem 3

Proof. We know that by construction that the covariances of the principal components are zero. Thus, each coefficient $\hat{\beta_m^s}$, $m = 1, \dots, M$ equals the solution of the OLS problem regressing only $\hat{z}_m$ on Y. This solution is well-known to be $\frac{\mathrm{Cov}(Z_m, Y)}{\mathrm{Var}(Z_m)}$. Since $\mathrm{E}(Z_m) = 0$ the estimated covariance is given by $\hat{z}_m' Y$. From A.1 it follows that the estimated variance of Z_m is λ_m, which yields the given formula in (3.6). □

Appendix B Data Generating Process

B.1 Scaling of the Variables

To have a realistic data generating process set up, it is necessary that the variables have the right scaling, e.g. negative values for the count variables schooling, work experience and number of siblings and for the wage must be avoided. The latter is already incorporated in equation (2.1) by using the logarithmic value of the wages as dependent variable. Moreover, test Scores should be distributed in an interval that is bounded below and above or even following an ordinal scale. Since Blundell et al. (2005) report the share of individuals in each of the quintiles, I decided to allow for grades between 5 (best) and 1 (worst). Since there is no natural scaling for ability, I assume that ability can possibly take any real value.

I use the derived dependency structure and scaling to build the fundamentals for the data generating process in the next subsection.

B.2 Equations of the Data Generating Process

The causal graph given above is the foundation of my strategy to create the data. Every variable is dependent on each variable that has an arrow pointing at it. Summing up this structure in general equations yields

$$\begin{aligned} Y_i &= f_0(a_i, s_i, w_i, n_i, e_i) \\ T_{i,7,c} &= f_1(a_i, e_i) \\ T_{i,11,c} &= f_2(a_i, e_i, T_{i,7,c}) \\ s_i &= f_3(a_i, e_i) \\ w_i &= f_4(s_i) \\ n_i &= f_5(e_i) \end{aligned} \tag{B.1}$$

whereby f_j is a function that models the relationship between the dependent and independent variable. Note that f_0 is already determined in equation (2.1). My aim is now to find reasonable forms for the f_j such that parameterizations meet the requirements for a realistic data generating process derived in the previous subsections. Thus, I first determine parent's education and ability, since they are not caused by other variables. Subsequently, I use the imposed assumptions to build the data step by step.

The structure I use is very similar for every step. I use the population moments, set them equal to the theoretical moments from the assumed distributions and solve for the characteristic parameters of the respective distribution. In the simulation I use this computed parameterizations to approximate the population moments given in Blundell et al. (2005).

Parents' Education

Since the years of parent's schooling is not caused by any other variable and is positive integer valued, I decided to draw N independent times from a Generalized Poisson distribution. The distribution yields only non-negative integers and allows, other than the Poisson distribution, for overdispersion in the data. Following Jung and Tremayne (2011), I denote the generalized Poisson distribution for parent's education with parameters $\lambda_e \in \mathbb{R}_+$ and $\eta_e \in [0, 1)$ by $GP(\lambda_e, \eta_e)$. The probability mass function is given by

$$Prob\,(E = e) = \frac{\lambda(\lambda + e\eta)^{e-1} \cdot \exp\left[-(\lambda + e\eta)\right]}{e!}, \tag{B.2}$$

The first and second moment can be expressed by the introduced parameters as $\mathrm{E}(e_i) = \frac{\lambda_e}{1-\eta_e}$ and $\mathrm{Var}(e_i) = \frac{\lambda_e}{(1-\eta_e)^3}$. Setting $\eta_e = 0$ yields the ordinary Poisson distribution and therefore the setting allows to use the nested Poisson distribution by adjusting η. For a full summary of the distribution I reference to Consul and Jain (1973). To find the parameterization that suits the first and second population moments of Blundell et al. (2005), I set the expected outcome of one individual equal to the population mean μ_e and the variance equal to the population variance σ_e^2. Subsequent, I solve this system of two equations for λ_e and η_e.

$$\sigma_e^2 = \mathrm{Var}(e_i) = \frac{\lambda_e}{(1-\eta_e)^3} = \frac{\mathrm{E}(e_i)}{(1-\eta_e)^2} = \frac{\mu_e}{(1-\eta_e)^2} \iff \eta_e = 1 - \frac{\sqrt{\mu_e}}{\sigma_e} \implies \lambda_e = \frac{\mu_e^{1.5}}{\sigma_e} \tag{B.3}$$

A note of notation, for subsequent derivations I denote the distribution function of e_i by D_e for all $i = 1, \ldots, N$.

Ability

Since there is no clear scaled measure for ability and hence I cannot match the distribution to population means, I decided to use a normal distribution with mean zero, such that $a_i \sim \mathcal{N}\left(0, \sigma^2_{ability}\right)$. It seemed reasonable to me that that most individuals have similar abilities but there are some individuals with very low or very high abilities. As parent's education, ability is not caused by any other variable and can therefore be drawn N independent times from the stated normal distribution.

Since I will make use of this when I define the process for the years of schooling, I denote the cumulative distribution function of a_i by ϕ_a for all $i = 1, \ldots, N$.

Age

I assume that individuals are aged between 33 and 68 years. The former is the age of the individuals in Blundell et al. (2005) and the latter is the UKs state pension age (British Government, 2020a). The age distribution is derived from the Office for National Statistics (2020) data on how many people lived in the United Kingdom per each year of age in 1991. I use the total number of individuals aged between 33 and 68 and compute the relative frequency per year to create a sample that uses the computed values as true probabilities and draw N independent times from this discrete distribution. For this purposes I used my user written

functions that allow generating discrete random variables given a specific probability mass.

Test Scores

As the individuals in Blundell et al. (2005). Individuals in my data take two tests in reading and math in the seventh and eleventh grade. To build the test scores, I follow Hansen et al. (2004) and determine the test scores as a function of ability. Moreover, I found it reasonable to add parent's education to account for parental influence on the children's test scores. Since in Blundell et al. (2005) the relative frequency of the quintiles is given, I decided that individuals can score from 5 (best) to 1 (worst). The problem that test scores are integer valued is solved in Hansen et al. (2004) by rounding the test scores (on a scale up to 100) to meet the integer requirement. However, since my scaling of test scores only consists of five numbers, I chose a different approach that builds on the property that the cumulative distribution function of a random variable is uniform distributed to prevent that rounding changes the population frequencies of the grades.

Theorem 10. *The Cumulative Distribution Function (cdf) of a Random Variable is Uniform Distributed*

Proof. Let $Y = F_X(x)$ be the cdf of a random variable X with density $f_X(x)$. Since $F_X(x)$ is continuous and strictly increasing, it follows that $F_X(x)$ is surjective and injective and hence invertible. The inverse function is denoted by $F_x^{-1}(x)$. $F_Y(y)$, the cdf of the cdf, is defined as $P(Y \leq y)$. Hence $F_Y(y) = P(Y \leq y) = P(F_X(x) \leq y)$. Since $F_X(x) \in [0,1]$, it follows that $P(F_X(x) \leq 1) = 1$ and $P(F_X(X) < 0) = 0$. Thus, the case of interest is when $y \in [0,1]$. Using the inverse, one gets $P(F_X(X) \leq y) = P(X \leq F_x^{-1}(y)) = F_X(F_x^{-1}(y)) = y$. Thus, the cdf of Y is given by

$$F_Y(y) = \begin{cases} 0 & , y < 0 \\ y & , 0 \leq y \leq 1 \\ 1 & , 1 < y \end{cases}, \tag{B.4}$$

which is the uniform distribution. □

From equations (B.1) it follows that the test scores are a function of ability and parent's education. First, I model a continuous latent process $T^*_{i,g,c}$ for every test, whereby $g \in \{7, 11\}$, $c \in \{\text{math}, \text{reading}\}$, $i = 1, \ldots, N$. The latent test scores for the 7th grade tests are defined as a linear function of ability and to a normal distribution transformed parent's education

$$T^*_{i,7,c} = \gamma_{ability} \cdot a_i + \gamma_{parent} \cdot \underbrace{\phi_a^{-1}\left(\underbrace{D_e(e_i)}_{\sim unif(0,1)}\right)}_{\sim \mathcal{N}(0, \sigma^2_{ability})} + \varepsilon_{i,7,c}, \tag{B.5}$$

whereby I assume that $\varepsilon_{i,7,c} \sim \mathcal{N}(0, \sigma_7^2)$ and uncorrelated with ability and parent's education. To derive that $D_e(e_i)$ is uniform and $\phi_a^{-1}(D_e(e_i))$ is normal, I make use of that the distribution function of a random variable is uniform distributed (B.4). Thus, I scale the variable to a uniform distribution and interpret the N values of of this as N independent draws from a uniform distribution. By applying the the inverse of the normal distribution with the population moments from parent's education as distribution moments, I generate a normally distributed variable that meets the requirements of the population moments. Hence, this newly created variable inherits the structure of individual's parent's education, thus the general correlation structure, and also meets the parent's education population moments. It would be desirable for further analysis to examine how the correlation structure is inherited by this transformation and how this could be applied to other approaches of modelling. The advantage is that with this transformation the distribution of the latent test scores exist in closed form, $T^*_{i,7,c} \sim \mathcal{N}\left(0, (\gamma^2_{ability} + \gamma^2_{parent})\sigma^2_{ability} + \sigma_7^2\right)$ with distribution function ϕ_7, which is needed subsequent to compute the grade frequencies. Moreover, it turns out that in the application the specification allows to model the desired correlation structure.

Since the latent test score's distribution is known in closed form, the same trick can be applied again to transform the latent test score to a uniform distributed variable. In notation that means $\phi_7\left(T^*_{i,7,c}\right)$ is uniform distributed. From here on, it is possible to give the test scores any desired distribution, since any variable can be simulated from a uniform distribution using appendix (B.4) vice versa. Defining the k-th quintiles of test scores of the test in c in seventh grade by $q^k_{7,c}$, I obtain the test scores by allocating the individuals according to their transformed latent test score

$$T_{i,7,c} = \begin{cases} 1 & , 0 \leq \phi_7\left(T^*_{i,7,c}\right) \leq q^1_{7,c} \\ 2 & , q^1_{7,c} < \phi_7\left(T^*_{i,7,c}\right) \leq \sum_{k=1}^{2} q^k_{7,c} \\ 3 & , \sum_{k=1}^{2} q^k_{7,c} < \phi_7\left(T^*_{i,7,c}\right) \leq \sum_{k=1}^{3} q^k_{7,c} \\ 4 & , \sum_{k=1}^{3} q^k_{7,c} < \phi_7\left(T^*_{i,7,c}\right) \leq \sum_{k=1}^{4} q^k_{7,c} \\ 5 & , \sum_{k=1}^{4} q^k_{7,c} < \phi_7\left(T^*_{i,7,c}\right) \leq 1 \end{cases} \tag{B.6}$$

Moving to further stages of their education, the individuals in my model take a new test and score the same latent value as in the 7th grade plus/minus a stochastic error $\varepsilon_{i,11,c} \sim \mathcal{N}(0, \sigma^2_{11})$. The correlation between the test scores can be determined by increasing (lower correlation) or decreasing (higher correlation) the variance of the error term. In the application I chose a rather high variance of 11 for the error to set the correlation between the two test scores to around 0.9. A higher correlation lead to perfect multicollinearity of the test scores in some

small samples.

$$T^*_{i,11,c} = T^*_{i,7,c} + \varepsilon_{i,11,c} \tag{B.7}$$

Hence, $T^*_{i,11,c} \sim \mathcal{N}\left(0, (\gamma^2_{ability} + \gamma^2_{parent})\sigma^2_{ability} + \sigma^2_7 + \sigma^2_{11}\right)$. To obtain the test scores, I apply the same transformation as in equation (B.6). Note that this transforming ensures flexibility to allow basically any distribution of grades for the test scores in seventh or eleventh grade and both subjects.

Number of Siblings

To determine the structure of f_5, I have used a model that is similar to the ones used for integer valued count data models, see Jung and Tremayne (2011). Economic literature argues that higher educated individuals have fewer children because the demand of children decreases with higher education, due to increasing opportunity costs (see inter alia: Willis (1973); Bailey (2010)). Cygan-Rehm and Maeder (2013) found that one additional year of education reduces the expected number of children of around 0.1. I have built an equation that can account for this by preserving the count structure of the data and takes into account that higher educated parents are expected to have fewer children. For every individual i I draw the maximum possible years of education minus the years of their parent's education independent Bernoulli experiments with success probability p and add an independent error term $W^{sib}_i \sim GP(\lambda^{sib}, \eta^{sib})$.

$$n_i = \sum_{j=1}^{m-e_i} B_{ij} + W^{sib}_i, \tag{B.8}$$

m is the maximum years of parent's education, $B_{ij} \sim \mathcal{B}(p)$ is Bernoulli distributed. The sum of B_{ij} and W^{sib}_i are independent. Since the error is assumed to be independent of all the Bernoulli trials $\mathrm{E}(W^{sib}_i) = \mathrm{E}(W^{sib}_i|e_i)$. Moreover, the sum of the independent Bernoulli trials is, conditioned on parent's education, binomial distributed $\sum_{j=1}^{m-e_i} B_{ij}|e_i \sim Bin(m - e_i, p)$. Thus, the expected difference in a one year increase in parent's education can be computed by $= \mathrm{E}(n_i|e_i+1) - \mathrm{E}(n_i|e_i) = (m-e_i-1)p + \mathrm{E}(W^{sib}_i) - (m-e_i)p - \mathrm{E}(W^{sib}_i) = -p$, such that one additional year of parent's education decreases the expected number of siblings by p. The imposed structure has the benefit that, similar as done before, it can be parameterized such that the first two moments of the random variable n_i meet the population moments reported in Blundell et al. (2005) by computing λ^{sib} and η^{sib} according to the desired moments.

Theorem 11. *The first two population moments of the number of Siblings are met if the*

error term is parameterized in the following way

$$\begin{aligned}\eta^{sib} =& 1 - \left(\frac{\mu_n - (m-\mu_e)\cdot p}{\sigma_n^2 - (m-\mu_e)\cdot p\cdot(1-p) + p^2\cdot\sigma_s^2}\right)^{0.5} \\ \lambda^{sib} =& [\mu_n - (m-\mu_e)\cdot p]\cdot(1-\eta^{sib}),\end{aligned} \tag{B.9}$$

whereby μ_n and σ_n^2 are the first and second moment of the number of siblings.

Proof.

$$\begin{aligned}\mu_n =& \mathrm{E}(n_i) = \mathrm{E}\left(\sum_{j=1}^{m-e_i} B_{i,j} + W_i^{sib}\right) = \mathrm{E}\left[\mathrm{E}\left(\sum_{j=1}^{m-e_i} B_{ij}|e_i\right)\right] + \frac{\lambda^{sib}}{1-\eta^{sib}} \\ =& \mathrm{E}\left[(m-e_i)\cdot p\right] + \frac{\lambda^{sib}}{1-\eta^{sib}} = (m-\mu_e)\cdot p + \frac{\lambda^{sib}}{1-\eta^{sib}}\end{aligned} \tag{B.10}$$

using the law of total variance of formula (A.3) for the second equality sign

$$\begin{aligned}\sigma_n^2 =& \mathrm{Var}(n_i) = \mathrm{Var}\left(\sum_{j=1}^{m-e_i} B_{i,j}\right) + \mathrm{Var}(W_i^{sib}) \\ =& \mathrm{E}\left[\mathrm{Var}\left(\sum_{j=1}^{m-e_i} B_{i,j}|e_i\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\sum_{j=1}^{m-e_i} B_{i,j}|e_i)\right)\right] + \frac{\lambda^{sib}}{(1-\eta^{sib})^3} \\ =& \mathrm{E}[(m-e_i)\cdot p\cdot(1-p)] + \mathrm{Var}[(m-e_i)\cdot p] + \frac{\lambda^{sib}}{(1-\eta^{sib})^3} \\ =& (m-\mu_e)\cdot p\cdot(1-p) + p^2\cdot\sigma_s^2 + \frac{\lambda^{sib}}{(1-\eta^{sib})^3}\end{aligned} \tag{B.11}$$

from the equations it follows that

$$\begin{aligned}\frac{\lambda^{sib}}{1-\eta^{sib}} =& \mu_n - (m-\mu_e)\cdot p \\ \frac{\lambda^{sib}}{(1-\eta^{sib})^3} =& \sigma_n^2 - (m-\mu_e)\cdot p\cdot(1-p) + p^2\cdot\sigma_e^2\end{aligned} \tag{B.12}$$

$$\begin{aligned}\Longleftrightarrow& \frac{\mu_n - (m-\mu_e)\cdot p}{(1-\eta^{sib})^2} = \sigma_n^2 - (m-\mu_e)\cdot p\cdot(1-p) + p^2\cdot\sigma_s^2 \\ \Longleftrightarrow& \frac{\mu_n - (m-\mu_e)\cdot p}{\sigma_n^2 - (m-\mu_e)\cdot p\cdot(1-p) + p^2\cdot\sigma_e^2} = (1-\eta^{sib})^2 \\ \Longleftrightarrow& \eta^{sib} = 1 - \left(\frac{\mu_n - (m-\mu_e)\cdot p}{\sigma_N^2 - (m-\mu_e)\cdot p\cdot(1-p) + p^2\cdot\sigma_e^2}\right)^{0.5}.\end{aligned} \tag{B.13}$$

Note that I excluded the $-$ case after applying the square root since this would imply η^{sib} to be greater than one, for which the generalized Poisson distribution is not defined for. From the above calculations it is derived that

$$\lambda^{sib} = [\mu_n - (m - \mu_e) \cdot p] \cdot (1 - \eta^{sib}) \tag{B.14}$$

□

Years of Schooling

The structure I impose is similar to the one for the number of siblings including one Bernoulli experiment for every year of parent's education and an additive error. To account for ability in the years of schooling, the error in my model is dependent on an individuals ability, such that $W_i^{sch}|a_i \sim GP(\lambda_i^{\text{sch}}, \eta^{sch})$ and $\lambda_i = k \cdot \phi_a(a_i)$, whereby $k \in \mathbb{R}$. Since ability and parent's education are assumed to be independent, the error is independent from the sum of the bernoulli experiments. Using (B.4), a cumulative distribution function is uniform distributed, it follows that $\lambda_i^{sch} \sim unif[0, k]$ and thus $\text{E}(\lambda_i^{sch}) = \frac{k}{2}$ and $\text{Var}(\lambda_i^{sch}) = \frac{k^2}{12}$. The equation is similar to equation (B.8).

$$s_i = \sum_{j=1}^{e_i} B_{i,j}^{sch} + W_i^{sch}, \tag{B.15}$$

with $B_{i,j}^{sch} \sim \mathcal{B}(q)$, such that q is the probability that one year of parent's education is inherited to the child. To meet the population moments μ_s and σ_s^2, it is necessary to find a suitable parameterization for k and η^{sch}.

Theorem 12. *The population moments for the years of schooling are met if the error term is parameterized such that*

$$\begin{aligned} k =& \pm \left(\frac{2(\mu_s - \mu_e \cdot q)^3}{\left[\sigma_s^2 - \mu_e \cdot q \cdot (1-q) - \sigma_e^2 \cdot q - \frac{1}{3}(\mu_s - \mu_e \cdot q)^2\right]} \right)^{0.5} \\ \eta =& 1 - \frac{k}{2(\mu_s - \mu_e \cdot q)}, \end{aligned} \tag{B.16}$$

whereby μ_s and σ_s^2 are the population mean and variance of the years of schooling.

Proof. Using the law of iterated expectations and the law of total variance formula (A.3) it

follows that

$$\begin{aligned}
\mathrm{E}(W_i^{sch}) =& \mathrm{E}\left[\mathrm{E}\left(W_i^{sch}|a_i\right)\right] = \mathrm{E}\left(\frac{\lambda_i}{1-\eta}\right) = \frac{1}{1-\eta}\mathrm{E}\left[k\cdot\phi(a_i)\right] = \frac{k}{2(1-\eta)} && \text{(B.17)}\\
\mathrm{Var}(W_i^{sch}) =& \mathrm{E}\left[\mathrm{Var}\left(W_i^{sch}|a_i\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(W_i^{sch}|a_i\right)\right]\\
=& \mathrm{E}\left(\frac{\lambda_i}{(1-\eta)^3}\right) + \mathrm{Var}\left(\frac{\lambda_i}{1-\eta}\right) = \frac{k}{2(1-\eta)^3} + \frac{k^2}{12(1-\eta)^2} && \text{(B.18)}
\end{aligned}$$

with the above formulas and what was derived for the number of siblings. Second step of variance equality sign I make use that ability and parent's years of education are uncorrelated.

$$\begin{aligned}
\mu_s =& \mathrm{E}(s_i) = \mathrm{E}\left(\sum_{j=1}^{e_i} B_{i,j}^{sch} + W_i^{sch}\right) = \mathrm{E}\left(\sum_{j=1}^{e_i} B_{i,j}^{sch}\right) + \mathrm{E}\left(W_i^{sch}\right) = \mu_e\cdot q + \frac{k}{2(1-\eta)} && \text{(B.19)}\\
\sigma_s^2 =& \mathrm{Var}(s_i) = \mathrm{Var}\left(\sum_{j=1}^{e_i} B_{i,j}^{sch} + W_i^{sch}\right) = \mathrm{Var}\left(\sum_{j=1}^{e_i} B_{i,j}^{sch}\right) + \mathrm{Var}\left(W_i^{sch}\right)\\
=& \mu_e\cdot q\cdot(1-q) + \sigma_e^2\cdot q + \frac{k}{2(1-\eta)^3} + \frac{k^2}{12(1-\eta)^2} && \text{(B.20)}
\end{aligned}$$

from the first equation it follows

$$\frac{1}{1-\eta} = \frac{2(\mu_s - \mu_e\cdot q)}{k} \tag{B.21}$$

yielding for the second equation that

$$\begin{aligned}
\sigma_s^2 =& \mu_e\cdot q\cdot(1-q) + \sigma_e^2\cdot q + \frac{k}{2}\cdot\left(\frac{2(\mu_s-\mu_e\cdot q)}{k}\right)^3 + \frac{k^2}{12}\cdot\left(\frac{2(\mu_s-\mu_e\cdot q)}{k}\right)^2\\
\Longleftrightarrow& \sigma_s^2 - \mu_e\cdot q\cdot(1-q) - \sigma_e^2\cdot q - \frac{1}{3}(\mu_s-\mu_e\cdot q)^2 = 4\frac{(\mu_s-\mu_e\cdot q)^3}{k^2}\\
\Longleftrightarrow& k^2 = \frac{4(\mu_s-\mu_e\cdot q)^3}{\sigma_s^2 - \mu_e\cdot q\cdot(1-q) - \sigma_e^2\cdot q - \frac{1}{3}(\mu_s-\mu_e\cdot q)^2}\\
\Longrightarrow& k = \pm\left(\frac{2(\mu_s-\mu_e\cdot q)^3}{\left[\sigma_s^2 - \mu_e\cdot q\cdot(1-q) - \sigma_e^2\cdot q - \frac{1}{3}(\mu_s-\mu_e\cdot q)^2\right]}\right)^{0.5} && \text{(B.22)}
\end{aligned}$$

□

k is chosen such that $k > 0$.

Theorem 13. *If $\sigma_s^2 = \sigma_e^2$, $\mu_s = \mu_e$ and $\mu_s > 3$, it follows that*

$$\frac{(\sigma_s^2 - \frac{1}{3}\mu_s^2)}{(\mu_s - \frac{1}{3}\mu_s^2)} < q \tag{B.23}$$

and thus q can be chosen from the interval $\left[\frac{(\sigma_s^2-\frac{1}{3}\mu_s^2)}{(\mu_s-\frac{1}{3}\mu_s^2)}, 1\right]$.

Proof. Note that the first equation in (B.22) implies that either $\mu_s - \mu_e \cdot q > 0$ and $\sigma_s^2 - \mu_e \cdot q \cdot (1-q) - \sigma_e^2 \cdot q - \frac{1}{3}(\mu_s - \mu_e \cdot q)^2 > 0$ or both terms are negative. In the application section I will assume that $\mu_s = \mu_e$ and $\sigma_s^2 = \sigma_e^2$. This directly yields that $\mu_s - \mu_s \cdot q = \mu_s(1 - \cdot q) > 0$ and thus it must be that

$$\sigma_s^2 - \mu_s \cdot q \cdot (1-q) - \sigma_s^2 \cdot q - \frac{1}{3}(\mu_s - \mu_s \cdot q)^2 > 0. \tag{B.24}$$

Assuming that $\sigma_s^2 = \sigma_e^2$ and that $\mu_s = \mu_e$ it follows

$$\begin{aligned}
&\sigma_s^2 - \mu_s \cdot q \cdot (1-q) - \sigma_s^2 \cdot q - \frac{1}{3}(\mu_s - \mu_s \cdot q)^2 > 0 \\
\iff \quad &\sigma_s^2(1-q) - \mu_s \cdot q \cdot (1-q) - \frac{1}{3}(\mu_s(1-q))^2 > 0 \\
\iff \quad &(\sigma_s^2 - \mu_s \cdot q)(1-q) > \frac{1}{3}(\mu_s(1-q))^2 \\
\iff \quad &(\sigma_s^2 - \mu_s \cdot q) > \frac{1}{3}\mu_s^2(1-q) \\
\iff \quad &(\sigma_s^2 - \mu_s \cdot q) > \frac{1}{3}\mu_s^2 - \frac{1}{3}\mu_s^2 q \\
\iff \quad &(\sigma_s^2 - \frac{1}{3}\mu_s^2) > \mu_s \cdot q - \frac{1}{3}\mu_s^2 q \\
\iff \quad &(\sigma_s^2 - \frac{1}{3}\mu_s^2) > (\mu_s - \frac{1}{3}\mu_s^2) \cdot q
\end{aligned}$$

if $\mu_s - \frac{1}{3}\mu_s^2 = \mu_s(1 - \frac{1}{3}\mu_s) < 0$, which is satisfied if $\mu_s > 3$

$$\frac{(\sigma_s^2 - \frac{1}{3}\mu_s^2)}{(\mu_s - \frac{1}{3}\mu_s^2)} < q \tag{B.25}$$

and if $\mu_s - \frac{1}{3}\mu_s^2 = \mu_s(1 - \frac{1}{3}\mu_s) > 0$

$$\frac{(\sigma_s^2 - \frac{1}{3}\mu_s^2)}{(\mu_s - \frac{1}{3}\mu_s^2)} > q \tag{B.26}$$

□

Work Experience

Following Heckman et al. (2006) work experience w_i of an individual i is determined by imposing a linear structure in the age age_i of an individual. Additionally, I subtract the years before school enrollment (British Government, 2020b) and allow individuals to take gap years

gap_i.

$$w_i = age_i - age\_schoolenrollment - s_i - gap_i. \tag{B.27}$$

gap_i, a discrete random variable with probability mass F_{gap}. In the application I allow for zero up to four gap years. The probabilities are the relative frequencies reported in Holmlund et al. (2008).

Wages

As given in equation (2.1), log-wages are a function of all the mentioned covariates but the test scores. To ease readability, I restate the equation and write it in compact vector notation.

$$\ln(Y_i) = \alpha + a_i + \beta_1 s_i + \beta_2 w_i + \beta_3 \frac{w_i^2}{100} + \beta_4 n_i + \beta_5 e_i + \varepsilon_i = \alpha + X_i^T \beta + \varepsilon_i, \tag{B.28}$$

whereby $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is an error term uncorrelated the other covariates. I denote the expected value of the logarithmic hourly wage by μ_Y, its variance by σ_Y and by μ_x the vector of the means the variables in X. Σ_x is the respective variance covariance matrix. The first and second moments of w_i imposed by the model can be computed by

$$\begin{aligned} \mathrm{E}(Y_i) =& \alpha + \mu_x \beta \\ \mathrm{Var}(Y_i) =& \mathrm{Var}(\alpha + X_i^T \beta + \varepsilon_i) = \mathrm{Var}(X_i^T \beta) + \mathrm{Var}(\varepsilon_i) = \beta' \Sigma_x \beta + \sigma_\varepsilon^2 \end{aligned} \tag{B.29}$$

The challenge is now to set the β parameters such that the population moments reported in Blundell et al. (2005) are met. Additionally, it might be desirable to set upper and lower bounds $\beta_j^{max}, \beta_j^{min}$ for each β_j. Using the weighted squared deviation from the desired population mean and standard deviation as objective function, the optimization problem at hand is

$$\begin{aligned} \beta = \arg \min_{b \in \mathbb{R}^{|\beta|}} \quad & \tau \cdot \left[\mu_Y - \alpha - \mu_x' b\right]^2 + (1 - \tau) \left[\sigma_Y - \sqrt{b' \Sigma_x b - \sigma_\varepsilon^2}\right]^2 \\ \text{s.t.} \quad & \beta^{min} \leq b \leq \beta^{max}, \end{aligned} \tag{B.30}$$

with $\tau \in [0, 1]$. I chose to use the standard deviation rather than the variance, since the variance is already squared. This minimization problem can then be solved numerically to find approximations for β.

References

Bailey, M.J., 2010. " momma's got the pill": how anthony comstock and griswold v. connecticut shaped us childbearing. American economic review 100, 98–129.

Björklund, A., Kjellström, C., 2002. Estimating the return to investments in education: how useful is the standard mincer equation? Economics of Education Review 21, 195–210.

Blundell, R., Costa Dias, M., Meghir, C., Shaw, J., 2016. Female labor supply, human capital, and welfare reform. Econometrica 84, 1705–1753.

Blundell, R., Dearden, L., Sianesi, B., 2005. Evaluating the effect of education on earnings: models, methods and results from the national child development survey. Journal of the Royal Statistical Society: Series A (Statistics in Society) 168, 473–512.

British Government, 2020a. Pension age. https://www.gov.uk/state-pension-age, Last accessed on 2020-08-11.

British Government, 2020b. School starting age. https://www.gov.uk/schools-admissions/school-starting-age, Last accessed on 2020-08-11.

Consul, P.C., Jain, G.C., 1973. A generalization of the poisson distribution. Technometrics 15, 791–799.

Cygan-Rehm, K., Maeder, M., 2013. The effect of education on fertility: Evidence from a compulsory schooling reform. Labour Economics 25, 35–48.

Davis-Kean, P.E., 2005. The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. Journal of family psychology 19, 294.

Fan, X., Seshadri, A., Taber, C., 2015. Estimation of a life-cycle model with human capital, labor supply and retirement. Australia. University of New South Wales .

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. volume 1. Springer series in statistics New York.

Hansen, K.T., Heckman, J.J., Mullen, K.J., 2004. The effect of schooling and ability on achievement test scores. Journal of econometrics 121, 39–98.

Heckman, J.J., Lochner, L.J., Todd, P.E., 2006. Earnings functions, rates of return and treatment effects: The mincer equation and beyond. Handbook of the Economics of Education 1, 307–458.

Holmlund, B., Liu, Q., Nordström Skans, O., 2008. Mind the gap? estimating the effects of postponing higher education. Oxford Economic Papers 60, 683–710.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. volume 112. Springer.

Jolliffe, I.T., 1986. Principal components in regression analysis, in: Principal component analysis. Springer, pp. 129–155.

Jung, R.C., Tremayne, A.R., 2011. Convolution-closed models for count time series with applications. Journal of Time Series Analysis 32, 268–280.

Lemieux, T., 2006. The mincer equation thirty years after schooling, experience, and earnings, in: Jacob Mincer a pioneer of modern labor economics. Springer, pp. 127–145.

Mincer, J., 1974. Schooling, experience, and earnings. human behavior & social institutions no. 2. .

Office for National Statistics, 2020. Population estimates. https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland/mid2015/ukandregionalpopulationestimates18382015.zip, Last accessed on 2020-08-11.

Shlens, J., 2014. A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100 .

Willis, R.J., 1973. A new approach to the economic theory of fertility behavior. Journal of political Economy 81, S14–S64.