Econometrics 1 – Assignment 1

Floris Holstege and Markus Mueller

September, 2020

Question 1

A

A general regression of y on p independent variables and an intercept is given in the form of

$$\mathbf{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The associated fitted values are then given by:

$$\hat{\mathbf{y}} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p,$$

Where the b_i are the estimated coefficients. When we use the ordinary least squares (OLS) method, we aim to minimize the sum of squared errors (SSE).

$$\begin{aligned} SSE &= \sum_i^n e_i^2 \\ &= \sum_i^n (\mathbf{y} - \hat{\mathbf{y}})^2 \\ &= \sum_i^n (\mathbf{y} - (b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p))^2 \end{aligned}$$

If we try to minimize SSE with respect to b_0, we need to take the first order derivative with respect to b_0, and set it equal to zero.

$$\begin{aligned} \frac{\partial SSE}{\partial b_0} &= -2 \sum_i^n (\mathbf{y} - (b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p)) \\ 0 &= -2 \sum_i^n (\mathbf{y} - (b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p)) \\ 0 &= -2 \sum_i^n (\mathbf{y} - \hat{\mathbf{y}}) \\ 0 &= -2 \sum_i^n e_i \end{aligned}$$

The last statement holds if and only if the residuals $e_i = \mathbf{y} - \hat{\mathbf{y}}$ sum to zero.

B

To show that the slope coefficient of the jth regressor for $i = 1, \dots, n$ observations and $j = 1, \dots, k$ independent variables is given as:

$$b_j = \frac{Cov(y_i, \tilde{u}_{ji})}{Var(\tilde{u}_{ji})},$$

Where $\tilde{u}_{ji}$ is the residual from a regression of x_{ji} on all other covariates, we use the approach of a partial regression.

First, we regress y_i on all k covariates excluding x_{ji}. Secondly, we regress x_{ji} on all of the k covariates as well. Lastly, we regress the residuals of the first regression on the residuals of the second. The Frisch-Waugh Theorem suggests that we therewith recover the sought after coefficient b_j.

Let $\mathbf{X}_{-j}$ be an $n \times (k-1)$ dimensional matrix, i.e. let it include all of the n observations of the $k-1$ covariates such that X_j is excluded. Further, let the residual maker for a regression of a variable on these $(k-1)$ covariates be $\mathbf{M}_{-j} = \mathbf{I}_n - \mathbf{X}_{-j}(\mathbf{X}'_{-j}\mathbf{X}_{-j})^{-1}\mathbf{X}'_{-j}$. From the first described regression, we then retrieve the residuals $\mathbf{M}_{-j}\mathbf{y}$, from the second one the residuals $\mathbf{M}_{-j}\mathbf{x}_j$. Notice that $\mathbf{M}_{-j}\mathbf{x}_j = \tilde{\mathbf{u}}_j$.

From the Frisch-Waugh Theorem and the known OLS estimator in matrix form, we know that the following estimator when regressing $\mathbf{M}_{-j}\mathbf{y}$ on $\mathbf{M}_{-j}\mathbf{x}_j$ indeed recovers the b_j.

$$\begin{aligned}
b_j &= ((\mathbf{M}_{-j}\mathbf{x}_j)'\mathbf{M}_{-j}\mathbf{x}_j)^{-1}(\mathbf{M}_{-j}\mathbf{x}_j)'\mathbf{M}_{-j}\mathbf{y} \\
&= (\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j)^{-1}\mathbf{x}'_j\mathbf{M}'_{-j}\mathbf{M}_{-j}\mathbf{y} && \text{(as } \mathbf{M}_{-j}\mathbf{x}_j = \tilde{\mathbf{u}}_j\text{)} \\
&= (\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j)^{-1}\mathbf{x}'_j\mathbf{M}_{-j}\mathbf{y} && \text{(as } \mathbf{M}_{-j} \text{ idempotent and symmetric)} \\
&= (\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j)^{-1}(\mathbf{M}'_{-j}\mathbf{x}_j)'\mathbf{y} \\
&= (\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j)^{-1}(\mathbf{M}_{-j}\mathbf{x}_j)'\mathbf{y} && \text{(as } \mathbf{M}_{-j} \text{ is symmetric)} \\
&= (\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j)^{-1}\tilde{\mathbf{u}}'_j\mathbf{y} && \text{(use } \mathbf{M}_{-j}\mathbf{x}_j = \tilde{\mathbf{u}}_j \text{ again)} \\
&= (\frac{1}{n}\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j)^{-1}\frac{1}{n}\tilde{\mathbf{u}}'_j\mathbf{y} && \text{(multiply and divide by } n\text{)}
\end{aligned}$$

To proceed, we notice that $Var(\tilde{\mathbf{u}}_j) = \mathbb{E}(\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j) - \mathbb{E}(\tilde{\mathbf{u}}_j)^2 = \mathbb{E}(\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j)$ when $\mathbb{E}(\tilde{\mathbf{u}}_j) = 0$. Similarly, $Cov(\mathbf{y}, \tilde{\mathbf{u}}_j) = \mathbb{E}(\tilde{\mathbf{u}}'_j\mathbf{y}) - \mathbb{E}(\tilde{\mathbf{u}}_j)\mathbb{E}(\mathbf{y}) = \mathbb{E}(\tilde{\mathbf{u}}'_j\mathbf{y})$ when $\mathbb{E}(\tilde{\mathbf{u}}_j) = 0$. In terms of sample moments, this suggests that we require $\frac{1}{n}\sum_{i=1}^n \tilde{u}_{ji} = 0$. Since these are residuals, we know from a) that their sum is indeed equal to zero. Thus, we have that $\frac{1}{n}\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j$ and $\frac{1}{n}\tilde{\mathbf{u}}'_j\mathbf{y}$ is the relevant sample variance and the sample covariance, respectively. In terms of individual components, we therefore have

$$\begin{aligned}
b_j &= (\frac{1}{n}\tilde{\mathbf{u}}'_j\tilde{\mathbf{u}}_j)^{-1}\frac{1}{n}\tilde{\mathbf{u}}'_j\mathbf{y} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n \tilde{u}_{ji}y_i}{\frac{1}{n}\sum_{i=1}^n \tilde{u}_{ji}^2} \\
&= \frac{\hat{Cov}(y_i, \tilde{u}_{ji})}{\hat{Var}(\tilde{u}_{ji})}.
\end{aligned}$$

C

The expression in b) suggests that b_j can be derived by dividing how much the residuals $\tilde{u}_{ji}$ and y vary together by the variation in the residuals themselves. Intuitively, by taking the residuals first, we take

away the influences of other variables on x_j and y. Thus, $\tilde{u}_{ji}$ is like a cleaned version of x_{ji} in which the correlations between x_{ji} and the other variables are no longer included. When regressing $\mathbf{M}_{-j}\mathbf{y}$ on $\mathbf{M}_{-j}\mathbf{x}_j$ (see explanation in task b)), we have isolated the correlation between y_i and x_{ji} by removing any potential influences of other variables due to correlations.

D

$$
\begin{aligned}
\frac{\hat{Cov}(\tilde{y}_i, x_{ji})}{\hat{Var}(x_{ji})} = \frac{\frac{1}{n}\sum_{i=1}^{n}\tilde{y}_{ji}x_{ji}}{\frac{1}{n}\sum_{i=1}^{n}x_{ji}^2} &= \frac{\sum_{i=1}^{n}\tilde{y}_{ji}x_{ji}}{\sum_{i=1}^{n}x_{ji}^2} \\
&= (\mathbf{x}_j'\mathbf{x}_j)^{-1}\mathbf{x}_j'\tilde{\mathbf{y}}_j \\
&= (\mathbf{x}_j'\mathbf{x}_j)^{-1}\mathbf{x}_j'(\mathbf{M}_{-j}\mathbf{y}) \\
&= (\mathbf{x}_j'\mathbf{x}_j)^{-1}(\mathbf{M}_{-j}'\mathbf{x}_j)'\mathbf{y} \\
&= (\mathbf{x}_j'\mathbf{x}_j)^{-1}(\mathbf{M}_{-j}\mathbf{x}_j)'\mathbf{y} && \text{(as } \mathbf{M}_{-j} \text{ is symmetric)} \\
&= (\mathbf{x}_j'\mathbf{x}_j)^{-1}\tilde{\mathbf{u}}_j'\mathbf{y} && (\mathbf{M}_{-j}\mathbf{y} = \tilde{\mathbf{u}}_j \text{ from b)}) \\
&= \frac{\sum_{i=1}^{n}\tilde{u}_{ji}y_i}{\sum_{i=1}^{n}x_{ji}^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_{ji}y_i}{\frac{1}{n}\sum_{i=1}^{n}x_{ji}^2} \\
&= \frac{\hat{Cov}(y_i, \tilde{u}_{ji})}{\hat{Var}(x_{ji})} \neq \frac{\hat{Cov}(y_i, \tilde{u}_{ji})}{\hat{Var}(\tilde{u}_{ji})} = b_j && \text{(see subtask b)}
\end{aligned}
$$

We also know that the OLS estimator for b_j can be written as follows:

$$b_j = (\mathbf{x}_j'\mathbf{x}_j)^{-1}\mathbf{x}_j'\mathbf{y}$$

Thus, the OLS estimator for b_j and $\frac{\hat{Cov}(\tilde{y}_i, x_{ji})}{\hat{Var}(x_{ji})}$ are only the same when $(\mathbf{M}_{-j}\mathbf{y}) = \mathbf{y}$, which requires that $\mathbf{M}_{-j} = (I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = I_n$. Intuitively, the outcome in this case could not be projected onto the space of the independent variables.

E

The given situation illustrates the omitted variable bias. Estimating $\mathbf{y} = \mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{e}$ gives us the OLS estimator $\mathbf{b}_r = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{y}$. The true DGP is given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{bmatrix}\mathbf{X}_r & X_k\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_r \\ \beta_k\end{bmatrix} + \boldsymbol{\epsilon} = \mathbf{X}_r\boldsymbol{\beta}_r + X_k\beta_k + \boldsymbol{\epsilon}$, where X_k represents the omitted variable and β_k the related coefficient. Using OLS we arrive at the following estimator for $\mathbf{b}_r$, in which the true DGP can be plugged in.

$$
\begin{aligned}
\mathbf{b}_r &= (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{y} \\
&= (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'(\mathbf{X}_r\boldsymbol{\beta}_r + X_k\beta_k + \boldsymbol{\epsilon}) \\
&= (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{X}_r\boldsymbol{\beta}_r + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'X_k\beta_k + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta}_r + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'X_k\beta_k + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\boldsymbol{\epsilon}
\end{aligned}
$$

The bias of the restricted OLS estimator is given by $\mathbb{E}(\mathbf{b}_r - \boldsymbol{\beta}_r)$, so that we have

$$
\begin{aligned}
\mathbb{E}(\mathbf{b}_r - \boldsymbol{\beta}_r) &= \mathbb{E}(\boldsymbol{\beta}_r + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r' X_k\beta_k + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\boldsymbol{\epsilon} - \boldsymbol{\beta}_r) \\
&= \mathbb{E}((\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r' X_k\beta_k + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\boldsymbol{\epsilon}) \\
&= (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r' X_k\beta_k + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbb{E}(\boldsymbol{\epsilon}) \\
&= (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r' X_k\beta_k && \text{(assuming } \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}\text{)}
\end{aligned}
$$

2

A

We now that the OLS estimator $\mathbf{b}$ is given by $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Rewriting this expression and expressing it including the sample size n sheds light on the asymptotic properties of the estimator.

$$
\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + (\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon} && \text{(multiply and divide by } n\text{)}
\end{aligned}
$$

To continue, we assume that $plim(\frac{1}{n}\mathbf{X}'\mathbf{X}) = \mathbf{Q}$, i.e. that the expression converges in probability to a finite matrix $\mathbf{Q}$, and that this matrix is invertible, i.e. full-rank. Moreover, from the given task we know that $plim(\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}) = (0, \dots, 0, \rho)'$. We call this vector $\mathbf{z}$. To determine the consistency of the estimator $\mathbf{b}$, we have to determine its probability limit.

$$
\begin{aligned}
plim(\mathbf{b}) &= plim(\boldsymbol{\beta}) + (plim(\frac{1}{n}\mathbf{X}'\mathbf{X}))^{-1}\, plim(\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}) \\
&= \boldsymbol{\beta} + \mathbf{Q}^{-1}\mathbf{z} && \text{(using the previous assumptions)}
\end{aligned}
$$

We now let $\mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{Q}_1^{-1} \\ \vdots \\ \mathbf{Q}_k^{-1} \end{bmatrix}$, so that $\mathbf{Q}_i^{-1}$ is the ith row of the matrix $\mathbf{Q}^{-1}$. Additionally, let $\mathbf{Q}_{ik}^{-1}$ be the last component of the ith row of $\mathbf{Q}^{-1}$. Using the previous expressions and the properties of matrix-vector multiplication, we get

$$
\begin{aligned}
plim(\mathbf{b}) &= \boldsymbol{\beta} + \mathbf{Q}^{-1}\mathbf{z} \\
&= \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Q}_1^{-1} \\ \vdots \\ \mathbf{Q}_k^{-1} \end{bmatrix}\mathbf{z} \\
&= \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Q}_{1k}^{-1}\rho \\ \vdots \\ \mathbf{Q}_{kk}^{-1}\rho \end{bmatrix}
\end{aligned}
$$

From this we see that if $\rho \neq \mathbf{0}$, then $plim(\mathbf{b}) \neq \boldsymbol{\beta}$, as each component in $\boldsymbol{\beta}$ gets a non-zero part added on. This makes the estimator $\boldsymbol{b}$ inconsistent for all coefficients β_i.

B

Throughout the analyses we take the $\mathbf{x}_i$ as given. Alternatively, one could interpret the utilised operators as being conditional on the $\mathbf{x}_i$. Assuming unbiasedness of $\mathbf{b}$ for $\boldsymbol{\beta}$ we have

$$\begin{aligned}
\sqrt{n}(\mathbf{b}-\boldsymbol{\beta}) &= \sqrt{n}(\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}-\boldsymbol{\beta}) && (\text{as } \mathbf{b}=\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon})\\
&= \sqrt{n}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\\
&= \frac{n}{\sqrt{n}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\\
&= (\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\epsilon}
\end{aligned}$$

By assumption, $\frac{1}{n}\mathbf{X}'\mathbf{X}$ converges in probability to the finite and invertible matrix $\mathbf{Q}$. Further, we can rewrite $\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\epsilon}$ as follows. Thus, to determine the asymptotic variance, we have to determine the variance of the term $\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\epsilon}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{w}_i=\sqrt{n}\bar{w}$.

$$\begin{aligned}
Var(\sqrt{n}\bar{w}) &= Var(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i)\\
&= \frac{1}{n}Var(\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}Var(\mathbf{x}_i\epsilon_i) && (\text{assuming both } \mathbf{x}_i \text{ and } \epsilon_i \text{ are jointly i.i.d.})\\
&= \frac{1}{n}nVar(\mathbf{x}_i\epsilon_i)\\
&= \mathbb{E}((\mathbf{x}_i\epsilon_i)'\mathbf{x}_i\epsilon_i)-\mathbb{E}(\mathbf{x}_i\epsilon_i)'\mathbb{E}(\mathbf{x}_i\epsilon_i)\\
&= \mathbb{E}(\mathbf{x}_i'\mathbf{x}_i\epsilon_i^2)-\mathbb{E}(\mathbf{x}_i\epsilon_i)'\mathbb{E}(\mathbf{x}_i\epsilon_i) && (\text{as } \epsilon_i \text{ is a scalar})\\
&= \mathbb{E}(\mathbf{x}_i'\mathbf{x}_i\epsilon_i^2)-\mathbf{z}'\mathbf{z} && (\text{where } \mathbf{z}=(0,\ldots,0,\rho)' \text{ as before})\\
&= \mathbb{E}(\mathbf{x}_i'\mathbf{x}_i\epsilon_i^2)-\rho^2\\
&= \boldsymbol{\Omega}_{x\epsilon}
\end{aligned}$$

Using the expressions given, the asymptotic variance can be derived as follows.

$$\begin{aligned}
AVar(\sqrt{n}(\mathbf{b}-\boldsymbol{\beta})) &= AVar(\mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\epsilon})\\
&= \mathbf{Q}^{-1}Var(\frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\epsilon})\mathbf{Q}^{-1}\\
&= \mathbf{Q}^{-1}\boldsymbol{\Omega}_{x\epsilon}\mathbf{Q}^{-1}
\end{aligned}$$

So that $AVar(\mathbf{b})=n^{-1}\mathbf{Q}^{-1}\boldsymbol{\Omega}_{x\epsilon}\mathbf{Q}^{-1}$ with $\boldsymbol{\Omega}_{x\epsilon}=\mathbb{E}(\mathbf{x}_i'\mathbf{x}_i\epsilon_i^2)-\rho^2$ represents the asymptotic variance of the estimator $\mathbf{b}$.

C

Since we know that

$$plim(\mathbf{b})=\boldsymbol{\beta}+\begin{bmatrix}\mathbf{Q}_{1k}^{-1}\rho\\ \vdots\\ \mathbf{Q}_{kk}^{-1}\rho\end{bmatrix}$$

Only the last coefficient, b_k would be inconsistent if:

$$\begin{bmatrix} \mathbf{Q}_{1k}^{-1} \\ \vdots \\ \mathbf{Q}_{kk}^{-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \eta \end{bmatrix}$$

Where η is some real number $\neq 0$. This is because then the bias becomes:

$$\begin{bmatrix} \mathbf{Q}_{1k}^{-1}\rho \\ \vdots \\ \mathbf{Q}_{kk}^{-1}\rho \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \eta\rho \end{bmatrix}$$

and the *plim* of $\boldsymbol{\beta}$ becomes

$$plim(\mathbf{b}) = \boldsymbol{\beta} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \eta\rho \end{bmatrix}$$

$$plim(\mathbf{b}) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k + \eta\rho \end{bmatrix}$$

Then, only the last coefficient would be inconsistent. Under what circumstances might this happen? Let's give one example. Remember that we have defined $\mathbf{Q} = plim(\frac{1}{n}\mathbf{X}'\mathbf{X})$, which is invertible. This occurs when $plim((\mathbf{X}'\mathbf{X})^{-1}) = Q = I$. This occurs when all the independent variables ($\mathbf{X}$) are uncorrelated with each other, and the variance of each $\mathrm{X}_i is 1$.

3

A

From $E(e_i) = 0$ and $Var(e_i) = \sigma^2$ we can derive the expected value and variance of y_i.

$$\begin{aligned} \mathbb{E}(y_i) &= \mathbb{E}(\alpha_0 + \alpha_1 x_i + e_i) \\ &= \alpha_0 + \alpha_1 x_i + \mathbb{E}(e_i) && \text{(since } x_i \text{ assumed non-random)} \\ &= \alpha_0 + \alpha_1 x_i && \text{(since } \mathbb{E}(e_i) = 0 \text{ given)} \end{aligned}$$

$$\begin{aligned} \mathbb{V}\mathrm{ar}(y_i) &= \mathbb{V}\mathrm{ar}(\alpha_0 + \alpha_1 x_i + e_i) \\ &= \mathbb{V}\mathrm{ar}(e_i) && \text{(since } x_i \text{ assumed non-random)} \\ &= \sigma^2 && \text{(given in task)} \end{aligned}$$

Since e_i are said to follow a normal distribution, the y_i will also be normally distributed, with mean $\mathbb{E}(y_i)$ and variance $\mathbb{V}\mathrm{ar}(y_i)$. We have $y_i \sim N(\alpha_0 + \alpha_1 x_i, \sigma^2)$. Substituting the mean and variance into the normal density, we get

$$
\begin{aligned}
p(y_i) &= \frac{1}{\sqrt{2\pi \mathbb{V}\text{ar}(y_i)}} exp\{-\frac{1}{2\mathbb{V}\text{ar}(y_i)}(y_i - \mathbb{E}(y_i))^2\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(y_i - (\alpha_0 + \alpha_1 x_i))^2\} && \text{(plugging in)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(y_i - \alpha_0 - \alpha_1 x_i)^2\}
\end{aligned}
$$

Assuming that the e_i are independently and identically distributed (i.i.d.), implies that the y_i are also i.i.d. and we can derive the likelihood as a simple product over all observations.

$$
\begin{aligned}
L(\alpha_0, \alpha_1 | x_i) &= \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(y_i - \alpha_0 - \alpha_1 x_i)^2\} \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i)^2\}
\end{aligned}
$$

Taking logs on both sides of the equation leads us to the log-likelihood function for y_i, $l(\alpha_0, \alpha_1 | x_i)$.

$$
\begin{aligned}
l(\alpha_0, \alpha_1 | x_i) &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i)^2 \\
&= -\frac{n}{2} log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i)^2
\end{aligned}
$$

B

Deriving the maximum likelihood estimators (MLEs) for α_0 and α_1 involves taking the first order derivatives of the log-likelihood function $l(\alpha_0, \alpha_1 | x_i)$ in A with respect to each parameter and setting these equal to zero. For α_0 we have

$$
\begin{aligned}
\frac{\partial l(\alpha_0, \alpha_1 | x_i)}{\partial \alpha_0} &= -\frac{1}{2\sigma^2} 2 \sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i)(-1) \\
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i) \\
0 &= \sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i) && \text{(multiply by } \sigma^2\text{)} \\
0 &= \sum_{i=1}^{n} y_i - n\alpha_0 - \alpha_1 \sum_{i=1}^{n} x_i \\
n\alpha_0 &= \sum_{i=1}^{n} y_i - \alpha_1 \sum_{i=1}^{n} x_i \\
\hat{\alpha_0} &= \frac{1}{n}\sum_{i=1}^{n} y_i - \alpha_1 \frac{1}{n}\sum_{i=1}^{n} x_i && \text{(rewrite in terms of averages)} \\
&= \bar{y} - \alpha_1 \bar{x} \\
\hat{\alpha_0} &= \bar{y} - \hat{\alpha_1}\bar{x} && \text{(use MLE of } \alpha_1\text{)}
\end{aligned}
$$

Thus, to be able to compute $\hat{\alpha_0}$ we require the MLE of α_1, which is derived as follows.

$$
\begin{aligned}
\frac{\partial l(\alpha_0,\alpha_1|x_i)}{\partial \alpha_1} &= -\frac{1}{2\sigma^2}2\sum_{i=1}^{n}(y_i-\alpha_0-\alpha_1 x_i)(-x_i) \\
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\alpha_0-\alpha_1 x_i)x_i \\
0 &= \sum_{i=1}^{n}(y_i x_i-\alpha_0 x_i-\alpha_1 x_i^2) && \text{(multiply by } \sigma^2 \text{ and distribute } x_i\text{)} \\
0 &= \sum_{i=1}^{n}(y_i x_i)-\alpha_0\sum_{i=1}^{n}x_i-\alpha_1\sum_{i=1}^{n}x_i^2 \\
\alpha_1\sum_{i=1}^{n}x_i^2 &= \sum_{i=1}^{n}y_i x_i-\hat{\alpha_0}\sum_{i=1}^{n}x_i && \text{(use MLE of } \alpha_0\text{)} \\
\alpha_1\sum_{i=1}^{n}x_i^2 &= \sum_{i=1}^{n}y_i x_i-(\bar{y}-\hat{\alpha_1}\bar{x})\sum_{i=1}^{n}x_i && \text{(plug in for } \hat{\alpha_0}\text{)} \\
\hat{\alpha_1}\sum_{i=1}^{n}x_i^2 &= \sum_{i=1}^{n}y_i x_i-\bar{y}\sum_{i=1}^{n}x_i+\hat{\alpha_1}\bar{x}\sum_{i=1}^{n}x_i \\
\hat{\alpha_1}\sum_{i=1}^{n}x_i^2-\hat{\alpha_1}\bar{x}\sum_{i=1}^{n}x_i &= \sum_{i=1}^{n}y_i x_i-\bar{y}\sum_{i=1}^{n}x_i \\
\hat{\alpha_1}(\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i) &= \sum_{i=1}^{n}y_i x_i-\bar{y}\sum_{i=1}^{n}x_i \\
\hat{\alpha_1}(\sum_{i=1}^{n}x_i^2-n\bar{x}^2) &= \sum_{i=1}^{n}y_i x_i-\bar{y}n\bar{x} && \text{(use that } \textstyle\sum_{i=1}^{n}x_i=n\bar{x}\text{)} \\
\hat{\alpha_1} &= \frac{\sum_{i=1}^{n}y_i x_i-n\bar{y}\bar{x}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2} \\
\hat{\alpha_1} &= \frac{\frac{1}{n}\sum_{i=1}^{n}y_i x_i-\bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2} && \text{(multiply numerator and denominator by } \tfrac{1}{n}\text{)} \\
\hat{\alpha_1} &= \frac{\hat{Cov}(y_i,x_i)}{\hat{Var}(x_i)}
\end{aligned}
$$

C

With $\mathbb{E}(e_i)=cz_i$,

$$
\begin{aligned}
\mathbb{E}(y_i) &= \mathbb{E}(\alpha_0+\alpha_1 x_i+e_i) \\
&= \alpha_0+\alpha_1 x_i+\mathbb{E}(e_i) && \text{(since } x_i \text{ assumed non-random)} \\
&= \alpha_0+\alpha_1 x_i+cz_i && \text{(since } \mathbb{E}(e_i)=cz_i \text{ given)}
\end{aligned}
$$

Since the variance remains $Var(e_i)=\sigma^2$, we have $Var(y_i)=\sigma^2$. The likelihood then becomes

$$
\begin{aligned}
L(\alpha_0,\alpha_1,c|x_i) &= \prod_{i=1}^{n}p(y_i)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}(y_i-\alpha_0-\alpha_1 x_i-cz_i)^2\} \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha_0-\alpha_1 x_i-cz_i)^2\}
\end{aligned}
$$

Taking logs, we arrive at the log-likelihood function

$$
\begin{aligned}
l(\alpha_0, \alpha_1, c|x_i) &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i - cz_i)^2 \\
&= -\frac{n}{2} log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i - cz_i)^2
\end{aligned}
$$

D

To derive the MLEs for α_0, α_1 and c, we take the partial derivatives of the log-likelihood function with respect to these parameters and set them equal to zero. In this way, for α_0 we get

$$
\begin{aligned}
\frac{\partial l(\alpha_0, \alpha_1, c|x_i)}{\partial \alpha_0} &= -\frac{1}{2\sigma^2} 2 \sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i - cz_i)(-1) \\
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i - cz_i) \\
0 &= \sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i - cz_i) && \text{(multiply by } \sigma^2\text{)} \\
0 &= \sum_{i=1}^{n} y_i - n\alpha_0 - \alpha_1 \sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} z_i \\
n\alpha_0 &= \sum_{i=1}^{n} y_i - \alpha_1 \sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} z_i \\
\hat{\alpha_0} &= \frac{1}{n}\sum_{i=1}^{n} y_i - \alpha_1 \frac{1}{n}\sum_{i=1}^{n} x_i - c\frac{1}{n}\sum_{i=1}^{n} z_i && \text{(rewrite in terms of averages)} \\
&= \bar{y} - \alpha_1 \bar{x} - c\bar{z} \\
\hat{\alpha_0} &= \bar{y} - \hat{\alpha_1}\bar{x} - \hat{c}\bar{z} && \text{(use MLEs of } \alpha_1 \text{ and } c\text{)}
\end{aligned}
$$

The MLE for α_1 can be derived as follows.

$$
\begin{aligned}
\frac{\partial l(\alpha_0, \alpha_1, c|x_i)}{\partial \alpha_1} &= -\frac{1}{2\sigma^2} 2 \sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i - cz_i)(-x_i) \\
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i - cz_i)x_i \\
0 &= \sum_{i=1}^{n} x_i y_i - x_i\alpha_0 - \alpha_1 x_i^2 - cx_i z_i \\
&= \sum_{i=1}^{n} x_i y_i - \alpha_0 \sum_{i=1}^{n} x_i - \alpha_1 \sum_{i=1}^{n} x_i^2 - c\sum_{i=1}^{n} x_i z_i \\
\alpha_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i - \alpha_0 \sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} x_i z_i \\
\hat{\alpha_1} &= \frac{\sum_{i=1}^{n} x_i y_i - \alpha_0 \sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} x_i z_i}{\sum_{i=1}^{n} x_i^2} \\
\hat{\alpha_1} &= \frac{\sum_{i=1}^{n} x_i y_i - \hat{\alpha_0} \sum_{i=1}^{n} x_i - \hat{c}\sum_{i=1}^{n} x_i z_i}{\sum_{i=1}^{n} x_i^2} && \text{(plug in MLEs for } \alpha_0 \text{ and } c\text{)} \\
\hat{\alpha_1} &= \frac{\bar{xy} - \hat{\alpha_0}\bar{x} - \hat{c}\bar{xz}}{\bar{x^2}}
\end{aligned}
$$

Where the last step expresses all involved quantities in terms of averages, i.e. $\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \bar{xy}$, $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$, $\frac{1}{n}\sum_{i=1}^{n} x_i z_i = \bar{xz}$ and $\sum_{i=1}^{n} x_i^2 = \bar{x^2}$.

Lastly, the MLE for parameter c can be derived by again setting the related first derivative of the log-likelihood function equal to zero.

$$
\begin{aligned}
\frac{\partial l(\alpha_0, \alpha_1, c|x_i)}{\partial c} &= -\frac{1}{2\sigma^2} 2 \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 x_i - c z_i)(-z_i) \\
0 &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 x_i - c z_i) z_i \\
0 &= \sum_{i=1}^{n} z_i y_i - z_i \alpha_0 - \alpha_1 z_i x_i - c z_i^2 \\
&= \sum_{i=1}^{n} z_i y_i - \alpha_0 \sum_{i=1}^{n} z_i - \alpha_1 \sum_{i=1}^{n} z_i x_i - c \sum_{i=1}^{n} z_i^2 \\
c \sum_{i=1}^{n} z_i^2 &= \sum_{i=1}^{n} z_i y_i - \alpha_0 \sum_{i=1}^{n} z_i - \alpha_1 \sum_{i=1}^{n} z_i x_i \\
\hat{c} &= \frac{\sum_{i=1}^{n} z_i y_i - \alpha_0 \sum_{i=1}^{n} z_i - \alpha_1 \sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i^2} \\
\hat{c} &= \frac{\sum_{i=1}^{n} z_i y_i - \hat{\alpha_0} \sum_{i=1}^{n} z_i - \hat{\alpha_1} \sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i^2} && \text{(plug in the MLEs } \hat{\alpha_0}, \hat{\alpha_1}\text{)} \\
\hat{c} &= \frac{\bar{zy} - \hat{\alpha_0}\bar{z} - \hat{\alpha_1}\bar{zx}}{\bar{z^2}} && \text{(express quantities in terms of averages)}
\end{aligned}
$$

Where the last step expresses all involved quantities in terms of averages, i.e. $\frac{1}{n}\sum_{i=1}^{n} z_i y_i = \bar{zy}$, $\frac{1}{n}\sum_{i=1}^{n} z_i = \bar{z}$, $\frac{1}{n}\sum_{i=1}^{n} z_i x_i = \bar{zx}$ and $\sum_{i=1}^{n} z_i^2 = \bar{z^2}$. Lastly, it should be noted that each of the derived MLEs depends on the other MLEs, i.e. they are simultaneously determined.

4

A

We create two models for the dependent variable of *birthweight*:

$$birthweight = b_0 + x_1 b_{age} + x_2 b_{unmarried} + x_3 b_{education} \quad (1)$$
$$birthweight = b_0 + x_1 b_{age} + x_2 b_{unmarried} + x_3 b_{education} + x_4 b_{smoker} + x_5 b_{alcohol} + x_6 b_{drinks} \quad (2)$$

The estimated coefficients of both models, including standard errors (between brackets) and their statistical significance can be found in table 1 and table 2.

Table 1 – Results for model 1

	Dependent variable:
	birthweight
age	−2.493
	(2.319)
unmarried	−277.272***
	(28.288)
educ	13.213**
	(5.560)
Constant	3,342.281***
	(80.209)
Observations	3,000
R^2	0.043
Adjusted R^2	0.042
Residual Std. Error	579.543 (df = 2996)
F Statistic	45.011*** (df = 3; 2996)
Note:	*p<0.1; **p<0.05; ***p<0.01

B

To test the following hypothesis:

$$b_{smoker} = b_{alcohol} = b_{drinks} = 0$$

We conduct an F-test of the following form:

$$\frac{(e^T_{model1} e_{model1} - e^T_{model2} e_{model2})/g}{(e^T_{model2} e_{model2})/n - k} \sim F(g, n-k)$$

This yields a value of 15.28, which has a p-value of 7.187555e-10. This is well below the 5% significance level, and thus we can reject the null hypothesis that these three coefficients are jointly equal to zero.

Table 2 – Results for model 2

	Dependent variable:
	birthweight
age	−2.273
	(2.308)
unmarried	−244.106***
	(28.539)
educ	7.276
	(5.593)
alcohol	−17.815
	(96.208)
smoker	−185.032***
	(27.936)
drinks	−7.230
	(19.187)
Constant	3,442.130***
	(81.213)
Observations	3,000
R^2	0.058
Adjusted R^2	0.056
Residual Std. Error	575.448 (df = 2993)
F Statistic	30.459*** (df = 6; 2993)
Note:	*p<0.1; **p<0.05; ***p<0.01

C

To test if there are non-linear deviations in model 2, we perform a Regression Specification Error Test (RESET test). Here, we add two non-linear transformations of $\hat{birthweight} = \hat{y}$: $\hat{y}^2, \hat{y}^3$ to model 2, creating model 3:

$$birthweight = b_0 + x_1 b_{age} + x_2 b_{unmarried} + x_3 b_{education} + x_4 b_{smoker} + x_5 b_{alcohol} + x_6 b_{drinks} + \hat{y}^2 b_{\hat{y}^2} + \hat{y}^3 b_{\hat{y}^3} \tag{3}$$

We use an F-test to see if $H0 : b_{\hat{y}^2} = b_{\hat{y}^3} = 0$ holds true. If this is the case, then there is no non-linear specification error in model 2. The F-test is defined as follows:

$$\frac{(e^T_{model2} e_{model2} - e^T_{model3} e_{model3})/g}{(e^T_{model3} e_{model3})/n-k} \sim F(g, n-k)$$

Where e notes the errors of a specific model, n the number of observations (3000), k the number of variables in model 3 plus one (9), g the number of restrictions, or the number of variables that are different between model 2 and 3 (2). Performing this test yields a value of 2.35, which according to the F-distribution has a p-value of 0.095. This result tells us that we have a 9.5% chance that $H0 : b_{\hat{y}^2} = b_{\hat{y}^3} = 0$ holds true, which would mean a non-linear deviation in our model. We fail to reject the $H0$ at a 5% significance level.

D

We log the dependent variable *birthweight*, and add the regressors from model 2, creating model 4. We also create another model (5) with two non-linear transformations of $Log(\hat{birthweight}) = \hat{y}$: $\hat{y}^2, \hat{y}^3$ to test for linear deviations.

$$Log(birthweight) = b_0 + x_1 b_{age} + x_2 b_{unmarried} + x_3 b_{education} + x_4 b_{smoker} + x_5 b_{alcohol} + x_6 b_{drinks} \tag{4}$$

$$Log(birthweight) = b_0 + x_1 b_{age} + x_2 b_{unmarried} + x_3 b_{education} + x_4 b_{smoker} + x_5 b_{alcohol} + x_6 b_{drinks} + \hat{y}^2 b_{\hat{y}^2} + \hat{y}^3 b_{\hat{y}^3} \tag{5}$$

The results of model 4 can be found below in table 3.

Table 3 – Results for model 4

	Dependent variable:
	log(birthweight)
age	−0.001
	(0.001)
unmarried	−0.088***
	(0.010)
educ	0.003
	(0.002)
alcohol	−0.013
	(0.035)
smoker	−0.056***
	(0.010)
drinks	−0.002
	(0.007)
Constant	8.134***
	(0.030)
Observations	3,000
R^2	0.050
Adjusted R^2	0.048
Residual Std. Error	0.211 (df = 2993)
F Statistic	26.160*** (df = 6; 2993)
Note:	*p<0.1; **p<0.05; ***p<0.01

We perform the following F-test to check for linear deviations:

$$\frac{(e^T_{model4} e_{model4} - e^T_{model5} e_{model5})/g}{(e^T_{model5} e_{model5})/n - k} \sim F(g, n-k)$$

Where e notes the errors of a specific model, n the number of observations (3000), k the number of variables in model 3 plus one (9), g the number of restrictions, or the number of variables that are different between model 4 and 5 (2). Performing this test yields a value of 1.28, which according to the F-distribution has a p-value of 0.566. This result tells us that we have a 56.6% chance that $H_0 : b_{\hat{y}^2} = b_{\hat{y}^3} = 0$ holds true, which would mean a non-linear deviation in our model. We fail to reject the H_0 at a 5% significance level.

In both models, we fail to reject the H_0, but the p-value was substantially higher once we logged the birthweight variable. This potentially indicates that logging birthweight better captures some of the non-linear dynamics in the model, but we cannot tell for sure given both tests do not provide evidence for a non-linear deviation.

E

The R^2 and adjusted R^2 of model 2 are respectively 0.058 and 0.056 (see table 2). For model 4, The R^2 and adjusted R^2 of model 2 are respectively 0.05 and 0.048. The relevant statistic for comparing the two

models is the adjusted R^2. R^2 indicates the % in variance of the dependent variable explained by the model. But the downside of R^2 is that it monotonically increases when we add variables to the model. To account for this, the adjusted R^2 penalizes the R^2 for the number of variables used. Thus, for comparisons across model specifications the adjusted R^2 should be used.

Model 4 has a (slightly) lower score for the adjusted R^2, which indicates that model 2 is slightly better in terms of explaining the variance of the dependent variable. Since the independent variables in both models are the same (only the dependent variable is different), we can infer that this set of variables better explains the *birthweight* variable when its not logged than when its logged.

F

In our maximum likelihood estimate, we assume that the errors are normally distributed by creating the log-likelihood function on the basis of the normal distribution.

$$l(e, \sigma^2) = \frac{1}{2n} log(2\pi) - \frac{1}{2n} log(\sigma^2) - \frac{e^T e}{2\sigma^2}$$

We use the Broyden - Fletcher - Goldfarb - Shanno (BFGS) algorithm to find the $\hat{\beta}$ that minimizes $e^T e$. We then calculated the corresponding σ^2-estimate with the following formula:

$$\hat{\sigma^2} = \frac{e^T e}{n}$$

The results can be found below in table 4. These coefficients are very similar to the coefficients estimated by OLS in model 1. We do not see much difference between the OLS and the ML estimates because the errors appear to be approximately normally distributed. When taking the normal distribution to create the log-likelihood function, this is the same as assuming the errors are normally distributed. Would the errors in reality differ a lot from this normality assumption, then the ML estimates in turn would differ more from the OLS estimates as well. Under normality of errors, OLS and ML estimators coincide.

Table 4 – ML estimation

Constant	3342.281
Age	−2.493
Unmarried	−277.272
Education	13.213
σ^2	335422.3

G

Comparing the OLS and GMM estimates (see table 5) of the coefficients and σ^2, we see that we get again very similar results. This is due to the moment conditions we impose in the GMM estimation process. Since we impose that the population moments hold exactly in the sample, i.e. $\mathbb{E}(\mathbf{X}'\boldsymbol{\epsilon}) = 0$ implies $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}'\boldsymbol{\epsilon}) = 0$.

However, the GMM estimates are strongly depending on the initial values chosen for the $\mathbf{b}$ vector. Since GMM uses an iterative procedure, a range of starting values apparently leads to a convergence to a local optimum instead of the global solution as found by OLS, leading to vastly different results. Initialising $\mathbf{b}$ using the OLS estimates from before, however, gives the reported results. Initialising them this way also ensures a faster convergence time as it represents a very good guess in the direction of the global minimum in

Table 5 – GMM estimation

Constant	3342.281
Age	−2.493
Unmarried	−277.272
Education	13.213
σ^2	335422.3

this case. Of course, depending on the situation at hand, such an initialisation may not be possible. In that case, many different starting values should be checked to confirm that indeed a global minimum is reached.