Econometrics
Supervision 2

Samuel Lee

Question 1

(a)

On average, for females aged 28-38 across the UK, an IQ score that is higher by 1 point is associated with a natural log of years of education that is 0.005 higher, or approximately 0.5% more years of education relative to the expected years of education without the increase in IQ. This is because of how $\ln educ$ is estimated to change in response to IQ:

$$\begin{aligned}
leduc = \ln educ &= \hat{\beta}_0 + \hat{\beta}_{IQ} IQ + \hat{\beta}_{age} age + \hat{\beta}_{agesq} agesq + \hat{\beta}_{black} black + \hat{\beta}_{meduc} meduc \\
\frac{\partial \ln educ}{\partial IQ} &= \frac{\partial \ln educ}{\partial educ} \cdot \frac{\partial educ}{\partial IQ} \\
\hat{\beta}_{IQ} &= \frac{1}{educ} \cdot \frac{\partial educ}{\partial IQ} \\
\frac{\partial educ}{\partial IQ} &= \hat{\beta}_{IQ} \cdot educ = 0.005 \cdot educ
\end{aligned}$$

meaning the increase in years of schooling associated with an infinitesimal increase in IQ is 0.005 multiplied by the level of education associated with the baseline IQ level.

(b)

An F-test would be done, with the "restricted model" being the same linear regression but without age and $agesq$ as independent variables. The F-statistic would then be $\frac{SSR_r - SSR_{ur}/q}{SSR_{ur}/(n-k-1)}$ where SSR is the residual sum of squares, q is the number of restrictions or $df_r - df_{ur}$ (2 in this case), and $n-k-1$ is df_{ur} (which is $857-5-1=851$ in this case). The F-statistic is distributed according to the F-distribution with q and $n-k-1$ degrees of freedom. The rejection region would be $R = \{F : f(F) \geq 1-\alpha\}$ where $f(F)$ is the cumulative distribution function for the relevant F-distribution (or otherwise defined depending on what the statistical table provides) and α is the significance level. If $F \in R$, then the hypothesis that educational attainment is not affected by age is rejected.

(c)

The intercept would be shifted up by $\ln 12$. This is because the old intercept is $\hat{\beta}_{0old} = \ln(educ_{old})$ and the new intercept is $\hat{\beta}_{0new} = \ln(12 \cdot educ_{old}) = \ln 12 + \ln(educ_{old})$. The coefficients would not change; this is because the dependent variable is in logarithms and the coefficients are essentially semi-elasticities and are unit-free with respect to the dependent variable. To show this,

$$\ln educ = \sum_{i=0}^{k} \hat{\beta}_i x_i \qquad (x_0 = 1)$$

$$\ln(12 \cdot educ) = \ln 12 + \ln educ = \sum_{i=0}^{k} \hat{\beta}_i x_i + \ln 12$$

$$\frac{\partial \ln(12 \cdot educ)}{\partial x_i} = \hat{\beta}_i$$

Which shows that expressing *educ* in months doesn't change the numerical significance of any x_i.

Knowing that the coefficients should not change, the only way for the t-statistics to change is for the standard errors of the coefficients to change. But the standard error of the coefficient is estimated by

$$s_{\hat{\beta}_j} = \sqrt{\frac{\sum u_i/(n-k-1)}{\sum (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)}} = \sqrt{\frac{\sum (y_i - \sum \hat{\beta}_m)/(n-k-1)}{\sum (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)}}$$

where R_j^2 is the R^2 from regressing x_j on all other independent variables. Since there is no change to the x's, the only things that will change in this expression are y_i and $\hat{\beta}_0$, and they both increase by $\ln 12$. The changes negate each other, $\sum (y_i - \sum \hat{\beta}_m)$ does not change, and thus the standard error of any coefficient is the same as before. Therefore there is no change to any t-statistic.

(d)

Such a test is rather meaningless, but can be done if one is so inclined. H_0 would be $3\beta_{meduc} = \beta_{black}$, and H_1 would be $3\beta_{meduc} \neq \beta_{black}$. The t-statistic would presumably be

$$t = \frac{3\hat{\beta}_{meduc} - \hat{\beta}_{black}}{s_{3\hat{\beta}_{meduc} - \hat{\beta}_{black}}}$$

Since $s_{3\hat{\beta}_{meduc} - \hat{\beta}_{black}} = \sqrt{\mathrm{Var}(3\hat{\beta}_{meduc} - \hat{\beta}_{black})}$,

$$\begin{aligned} s_{3\hat{\beta}_{meduc} - \hat{\beta}_{black}} &= \sqrt{\mathrm{Var}(3\hat{\beta}_{meduc} - \hat{\beta}_{black})} \\ &= \sqrt{9\mathrm{Var}(\hat{\beta}_{meduc}) + \mathrm{Var}(\hat{\beta}_{black}) - 6\mathrm{Cov}(\hat{\beta}_{meduc}, \hat{\beta}_{black})} \end{aligned}$$

and whoever volunteers to calculate this will be able to get the t-statistic required to perform the t-test, with H_0 being rejected if $|t| \geq t_{\frac{\alpha}{2}, n-k-1}$ for a significance level α. With $H_0 : 3\beta_{meduc} = \beta_{black}$ or $H_0 : \theta_1 = \beta_{black} - 3\beta_{meduc} = 0$, $s_{3\hat{\beta}_{meduc} - \hat{\beta}_{black}}$ can also be calculated by running the regression

$$\begin{aligned} leduc &= \beta_0 + \beta_{IQ} IQ + \beta_{age} age + \beta_{agesq} agesq + \beta_{black} black + \beta_{meduc} meduc \\ &= \beta_0 + \beta_{IQ} IQ + \beta_{age} age + \beta_{agesq} agesq + \beta_{black} black + \beta_{meduc} meduc - 3\beta_{meduc} black \\ &\quad + 3\beta_{meduc} black \\ &= \beta_0 + \beta_{IQ} IQ + \beta_{age} age + \beta_{agesq} agesq + (\beta_{black} - 3\beta_{meduc}) black + \beta_{meduc}(meduc + 3black) \\ &= \beta_0 + \beta_{IQ} IQ + \beta_{age} age + \beta_{agesq} agesq + \theta_1 black + \beta_{meduc}(meduc + 3black) \end{aligned}$$

and getting the standard error for θ_1 the usual way.

Question 2

The law of iterated expectations is $\mathrm{E}(\mathrm{E}(X|Y)) = \mathrm{E}(X)$ and for discrete X, Y, $\sum_y P(Y = y) \cdot \mathrm{E}(X|Y = y) = \mathrm{E}(X)$.

Question 3

(a)

An estimator is unbiased if its expected value is the value of the parameter it is estimating.

For the regression model $Y_i = \beta_0 + \beta_1 X_i + U_i$,

$$\begin{aligned}
\mathrm{E}\left(\frac{Y_i}{\sum a_j X_j}\right) &= \mathrm{E}\left(\frac{\beta_0 + \beta_1 X_i + U_i}{\sum a_j X_j}\right) \\
&= \beta_0 \mathrm{E}\left(\frac{1}{\sum a_j X_j}\right) + \beta_1 \mathrm{E}\left(\frac{X_i}{\sum a_j X_j}\right) + \mathrm{E}\left(\frac{U_i}{\sum a_j X_j}\right) \\
&= \beta_0 \mathrm{E}\left(\frac{1}{\sum a_j X_j}\right) + \beta_1 \mathrm{E}\left(\frac{X_i}{\sum a_j X_j}\right) + \mathrm{E}(U_i)\mathrm{E}\left(\frac{1}{\sum a_j X_j}\right) \\
&\text{(since } U \text{ and } X \text{ are independent)} \\
&= \beta_0 \mathrm{E}\left(\frac{1}{\sum a_j X_j}\right) + \beta_1 \mathrm{E}\left(\frac{X_i}{\sum a_j X_j}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= \mathrm{E}\left(\frac{Y_3 + Y_1 - 2Y_2}{X_3 + X_1 - 2X_2}\right) \\
&= \mathrm{E}\left(\frac{Y_3}{X_3 + X_1 - 2X_2}\right) + \mathrm{E}\left(\frac{Y_1}{X_3 + X_1 - 2X_2}\right) - 2\mathrm{E}\left(\frac{Y_2}{X_3 + X_1 - 2X_2}\right) \\
&= \beta_0 \mathrm{E}\left(\frac{1}{\sum a_j X_j}\right) + \beta_1 \mathrm{E}\left(\frac{X_3}{\sum a_j X_j}\right) + \beta_0 \mathrm{E}\left(\frac{1}{\sum a_j X_j}\right) + \beta_1 \mathrm{E}\left(\frac{X_1}{\sum a_j X_j}\right) \\
&- 2\left[\beta_0 \mathrm{E}\left(\frac{1}{\sum a_j X_j}\right) + \beta_1 \mathrm{E}\left(\frac{X_2}{\sum a_j X_j}\right)\right] \\
&= \beta_1 \left(\mathrm{E}\left(\frac{X_3}{\sum a_j X_j}\right) + \mathrm{E}\left(\frac{X_1}{\sum a_j X_j}\right) - 2\mathrm{E}\left(\frac{X_2}{\sum a_j X_j}\right)\right)
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\beta_1 \mathrm{E}\left(\frac{\sum a_j X_j}{\sum a_j X_j}\right) \\
&= \beta_1
\end{aligned}$$

and the estimator is unbiased.

(b)

No. When the Gauss-Markov assumptions are fulfilled (which in this case they are), the OLS estimator is the best (meaning lowest variance or most efficient) linear unbiased estimator for this model. It also makes intuitive sense that any estimator which only uses 3 data points out of 1000 observations should not be a better estimator than one which uses all 1000 observations.

(c)

The general result was used in (a).