# 04_Modelling

06010567

## Contents

Firstly, we will focus on the most recent year data (2023). I include all potential associated variables in the first model. Model1:

$$
\begin{aligned}
\text{fertility} = \alpha &+ \beta_1 \cdot \text{region} + \beta_2 \cdot \text{income} + \beta_3 \cdot \text{edu}_{\text{upper}} + \beta_4 \cdot \text{urban} + \beta_5 \cdot \text{female}_{\text{labor}} \\
&+ \beta_6 \cdot \text{contraceptive} + \beta_7 \cdot \text{eys}_f + \beta_8 \cdot \text{gii} + \beta_9 \cdot \text{mmr} + \beta_{10} \cdot \text{pr}_f + \beta_{11} \cdot \log(\text{gdp})
\end{aligned}
$$

```
#>
#> Call:
#> lm(formula = fertility ~ region + income + edu_upper + urban +
#>     female_labor + contraceptive + eys_f + gii + mmr + pr_f +
#>     log_gdp, data = df_2023)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -0.91785 -0.26618 -0.01992  0.22582  1.48682
#>
#> Coefficients:
#>                                   Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                      1.657e+00  1.019e+00   1.626  0.10661
#> regionEurope & Central Asia      2.295e-01  1.683e-01   1.364  0.17532
#> regionLatin America & Caribbean -3.632e-01  1.867e-01  -1.946  0.05414 .
#> regionMiddle East & North Africa  1.623e-01  2.438e-01   0.666  0.50692
#> regionNorth America              4.754e-02  5.108e-01   0.093  0.92602
#> regionSouth Asia                -2.027e-01  2.184e-01  -0.928  0.35532
#> regionSub-Saharan Africa         3.487e-01  1.839e-01   1.896  0.06046 .
#> incomeLower middle income       -4.428e-01  1.692e-01  -2.617  0.01007 *
#> incomeUpper middle income       -8.387e-01  2.674e-01  -3.136  0.00218 **
#> incomeHigh income               -6.081e-01  3.867e-01  -1.573  0.11858
#> edu_upper                        3.329e-03  2.736e-03   1.217  0.22617
#> urban                            3.042e-03  3.154e-03   0.964  0.33687
#> female_labor                     4.919e-03  3.738e-03   1.316  0.19074
#> contraceptive                   -1.822e-02  2.826e-03  -6.448 2.83e-09 ***
#> eys_f                           -1.147e-02  2.573e-02  -0.446  0.65652
#> gii                              4.026e+00  7.341e-01   5.484 2.53e-07 ***
#> mmr                              1.945e-05  3.901e-04   0.050  0.96032
#> pr_f                             1.466e-02  4.410e-03   3.325  0.00119 **
#> log_gdp                          1.018e-02  1.175e-01   0.087  0.93112
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.4625 on 114 degrees of freedom
```

```
#>   (84 observations deleted due to missingness)
#> Multiple R-squared:  0.882,  Adjusted R-squared:  0.8634
#> F-statistic: 47.35 on 18 and 114 DF,  p-value: < 2.2e-16
#>                    GVIF Df GVIF^(1/(2*Df))
#> region        28.212289  6        1.320900
#> income        15.180473  3        1.573551
#> edu_upper      3.774472  1        1.942800
#> urban          2.813568  1        1.677369
#> female_labor   2.196157  1        1.481944
#> contraceptive  2.053076  1        1.432856
#> eys_f          4.416561  1        2.101561
#> gii           12.172902  1        3.488969
#> mmr            3.914431  1        1.978492
#> pr_f           1.824716  1        1.350821
#> log_gdp       15.716376  1        3.964388
```

This full model includes all theoretically motivated predictors. However, several predictors show weak or insignificant effects, and VIFs suggest moderate multicollinearity, particularly between log(GDP), education, and GII. `gii` is temporarily kept as it is significantly associated with the response.

We can see that `region` has 7 levels. To reduce the number of region categories while preserving structural variation, we performed hierarchical clustering on average region-level values of key indicators. The resulting three-cluster groups regions with similar socioeconomic and gender characteristics. Figure 1 is the visualisation of clustering.

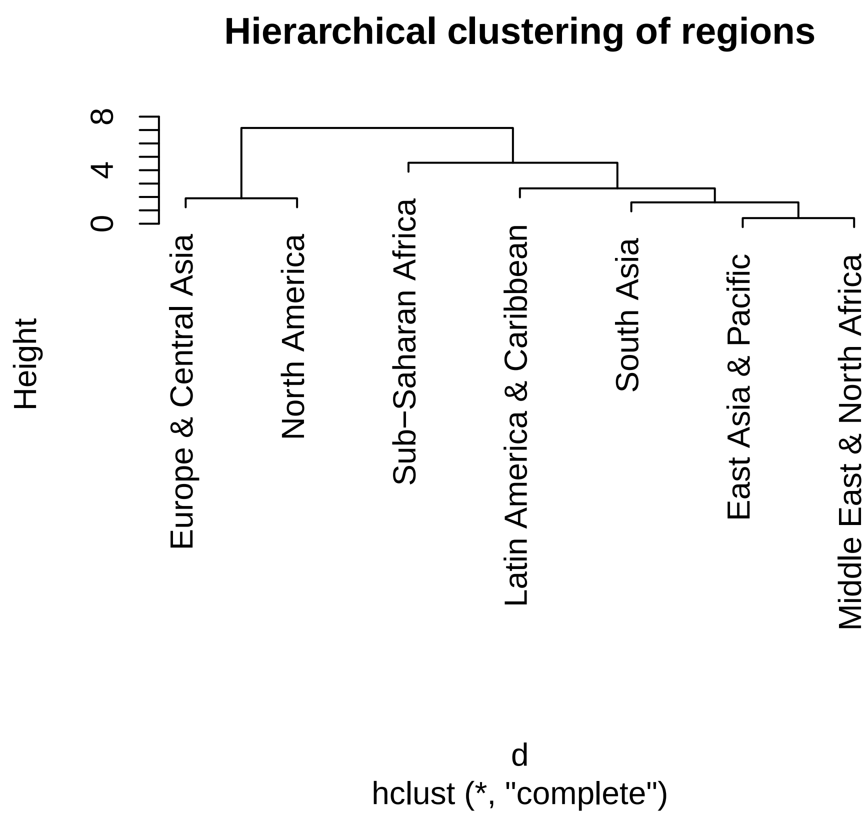


Figure 1: Dendogram of region

We remove variables with persistently weak explanatory power (`urban`, `mmr`, `female_labor`) and replace `region` with the clustered `region_group`. This reduced model retains strong interpretability and maintains high explanatory strength. Model2:

$$fertility = \alpha + \beta_1 region_{group} + \beta_2 income + \beta_3 edu_{upper} + \beta_4 contraceptive + \beta 5 gii + \beta_6 pr_f$$

```
#> 
#> Call:
#> lm(formula = fertility ~ region_group + income + edu_upper + 
#>     contraceptive + gii + pr_f, data = df_2023)
#> 
#> Residuals:
#>      Min       1Q   Median       3Q      Max 
#> -0.92699 -0.27041 -0.01231  0.24492  1.48966 
#> 
#> Coefficients:
#>                            Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)                2.513560   0.359036   7.001 1.45e-10 ***
#> region_groupC2             0.288100   0.197689   1.457  0.14757    
#> region_groupC3            -0.312781   0.144794  -2.160  0.03270 *  
#> incomeLower middle income -0.476013   0.141458  -3.365  0.00102 ** 
#> incomeUpper middle income -0.922601   0.180941  -5.099 1.25e-06 ***
#> incomeHigh income         -0.678166   0.233627  -2.903  0.00438 ** 
#> edu_upper                  0.003817   0.002678   1.425  0.15661    
#> contraceptive             -0.018248   0.002630  -6.940 1.98e-10 ***
#> gii                        3.495102   0.518871   6.736 5.56e-10 ***
#> pr_f                       0.009819   0.003777   2.599  0.01048 *  
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 0.4636 on 123 degrees of freedom
#>   (84 observations deleted due to missingness)
#> Multiple R-squared:  0.8721, Adjusted R-squared:  0.8628 
#> F-statistic:  93.2 on 9 and 123 DF,  p-value: < 2.2e-16
```

I also tried `log_gdp` instead of `income` to see the differences. Model3:

$$fertility = \alpha + \beta_1 region_{group} + \beta_2 log_{gdp} + \beta_3 edu_{upper} + \beta_4 contraceptive + \beta 5 gii + \beta_6 pr_f$$

```
#> 
#> Call:
#> lm(formula = fertility ~ region_group + log_gdp + edu_upper + 
#>     contraceptive + gii + pr_f, data = df_2023)
#> 
#> Residuals:
#>      Min       1Q   Median       3Q      Max 
#> -1.23635 -0.30906 -0.01209  0.34051  1.65416 
#> 
#> Coefficients:
#>                 Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)     3.440656   0.744379   4.622 9.32e-06 ***
#> region_groupC2  0.358067   0.215749   1.660   0.0995 .  
#> region_groupC3 -0.407978   0.157349  -2.593   0.0107 *  
#> log_gdp        -0.142526   0.068660  -2.076   0.0400 *  
#> edu_upper       0.001721   0.002865   0.601   0.5491    
#> contraceptive  -0.018931   0.002876  -6.583 1.15e-09 ***
```

```
#> gii             3.162868   0.612340   5.165 9.21e-07 ***
#> pr_f            0.008849   0.004112   2.152   0.0333 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.5057 on 125 degrees of freedom
#>   (84 observations deleted due to missingness)
#> Multiple R-squared:  0.8453, Adjusted R-squared:  0.8367
#> F-statistic: 97.61 on 7 and 125 DF,  p-value: < 2.2e-16
```

Income shows more consistent significance and interpretability than log(GDP), despite slightly higher degrees of freedom. We choose to retain income in models focused on demographic and social heterogeneity.

Then I compared the two education related covariates through Model2 and Model4: Model4:

$$fertility = \alpha + \beta_1 region_{group} + \beta_2 income + \beta_3 eys_f + \beta_4 contraceptive + \beta 5 gii + \beta_6 pr_f$$

```
#>
#> Call:
#> lm(formula = fertility ~ region_group + income + eys_f + contraceptive +
#>     gii + pr_f, data = df_2023)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -1.20397 -0.26763  0.01496  0.25085  1.39698
#>
#> Coefficients:
#>                            Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                 2.647859   0.439690   6.022 1.63e-08 ***
#> region_groupC2              0.115884   0.188687   0.614 0.540175
#> region_groupC3             -0.427030   0.138812  -3.076 0.002552 **
#> incomeLower middle income -0.518011   0.143395  -3.612 0.000431 ***
#> incomeUpper middle income -0.873345   0.184700  -4.728 5.76e-06 ***
#> incomeHigh income         -0.596490   0.240800  -2.477 0.014518 *
#> eys_f                      0.011520   0.023740   0.485 0.628309
#> contraceptive             -0.018868   0.002685  -7.028 1.04e-10 ***
#> gii                        3.621864   0.548935   6.598 9.39e-10 ***
#> pr_f                       0.007686   0.003829   2.007 0.046802 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.4802 on 131 degrees of freedom
#>   (76 observations deleted due to missingness)
#> Multiple R-squared:  0.8655, Adjusted R-squared:  0.8562
#> F-statistic: 93.64 on 9 and 131 DF,  p-value: < 2.2e-16
```

Including both `edu_upper` and `eys_f` introduces some redundancy. We avoid including both in the same final model to reduce multicollinearity. While both `edu_upper` (attainment-based) and `eys_f` (expectation-based) indicators are associated with fertility outcomes, we choose to retain `edu_upper` in the main specification due to its stronger interpretability and policy relevance. Unlike expected years of schooling, which is a projection subject to assumptions about school-age enrollment patterns, `edu_upper` represents the observed percentage of women aged 25 and above who have completed at least upper secondary education. This makes it a more concrete and retrospective measure of educational attainment. Moreover, `edu_upper` is more directly aligned with education policy targets and is consistently reported and tracked in the World

Bank's World Development Indicators (WDI), enhancing reproducibility and future comparability. It also serves as a clearer proxy for structural shifts in female human capital that can influence long-term fertility behaviour. Therefore, for clarity and policy relevance, we adopt `edu_upper` in the main model.

As I said in the EDA, there may be potential interaction effect of education and income, so I add the interaction term in the following models. Model5:

$$fertility = \alpha + \beta_1 region_{group} + \beta_2 income + \beta_3 edu_{upper} + \beta_4 contraceptive + \beta 5 gii + \beta_6 pr_f + \beta_7 income * edu_{upper}$$

```
#>
#> Call:
#> lm(formula = fertility ~ region_group + income * edu_upper +
#>     contraceptive + gii + pr_f, data = df_2023)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -1.07551 -0.28809 -0.02957  0.22726  1.25738
#>
#> Coefficients:
#>                                       Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                           2.906544   0.367029   7.919 1.35e-12 ***
#> region_groupC2                        0.307893   0.191210   1.610  0.10997
#> region_groupC3                       -0.281665   0.141550  -1.990  0.04888 *
#> incomeLower middle income            -0.979845   0.209526  -4.676 7.71e-06 ***
#> incomeUpper middle income            -1.302385   0.279311  -4.663 8.16e-06 ***
#> incomeHigh income                    -1.620566   0.532022  -3.046  0.00285 **
#> edu_upper                            -0.037428   0.012823  -2.919  0.00420 **
#> contraceptive                        -0.018565   0.002644  -7.023 1.41e-10 ***
#> gii                                   3.556564   0.503666   7.061 1.16e-10 ***
#> pr_f                                  0.009913   0.003701   2.679  0.00843 **
#> incomeLower middle income:edu_upper   0.043589   0.013162   3.312  0.00123 **
#> incomeUpper middle income:edu_upper   0.040540   0.013332   3.041  0.00290 **
#> incomeHigh income:edu_upper           0.048649   0.014421   3.374  0.00100 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.447 on 120 degrees of freedom
#>   (84 observations deleted due to missingness)
#> Multiple R-squared:  0.884,  Adjusted R-squared:  0.8724
#> F-statistic: 76.22 on 12 and 120 DF,  p-value: < 2.2e-16
#> Analysis of Variance Table
#>
#> Model 1: fertility ~ region_group + income + edu_upper + contraceptive +
#>     gii + pr_f
#> Model 2: fertility ~ region_group + income * edu_upper + contraceptive +
#>     gii + pr_f
#>   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
#> 1    123 26.434
#> 2    120 23.975  3    2.4586 4.102 0.00824 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Education appears to reduce fertility more strongly in lower-income countries. The interaction term is significant, suggesting heterogeneous marginal effects. The results of the model and the anova test show that the interaction term is useful.

I also check the other two income-related and education-related variables. Model6:

$$fertility = \alpha + \beta_1 region_{group} + \beta_2 log_{gdp} + \beta_3 eys_f + \beta_4 contraceptive + \beta 5 gii + \beta_6 pr_f + \beta_7 log_{gdp} * eys_f$$

```
#>
#> Call:
#> lm(formula = fertility ~ region_group + log_gdp * eys_f + contraceptive +
#>     gii + pr_f, data = df_2023)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -1.05856 -0.28861  0.01537  0.23821  1.48640
#>
#> Coefficients:
#>                  Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      9.381046   1.259874   7.446 1.11e-11 ***
#> region_groupC2   0.286434   0.184299   1.554   0.1225
#> region_groupC3  -0.327152   0.138701  -2.359   0.0198 *
#> log_gdp         -0.847889   0.146100  -5.803 4.58e-08 ***
#> eys_f           -0.425160   0.081931  -5.189 7.77e-07 ***
#> contraceptive   -0.018146   0.002632  -6.894 2.03e-10 ***
#> gii              3.071672   0.560633   5.479 2.09e-07 ***
#> pr_f             0.004463   0.003782   1.180   0.2401
#> log_gdp:eys_f    0.050302   0.009266   5.429 2.63e-07 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.4706 on 132 degrees of freedom
#>   (76 observations deleted due to missingness)
#> Multiple R-squared:  0.8698, Adjusted R-squared:  0.8619
#> F-statistic: 110.3 on 8 and 132 DF,  p-value: < 2.2e-16
```

The interaction between log(GDP) and expected years of schooling is highly significant. This suggests the educational context modifies the fertility-income relationship, especially in high-growth or highly educated countries.

Across all models, the most consistently significant predictors are contraceptive prevalence, GII, and women's political representation (PR_F). We also find strong support for education × income/GDP interactions. Subsequent forecasting or causal modelling can benefit from stratifying by these interactions.

To account for different aspects of female education, we explored both attainment-based (edu_upper) and expectation-based (eys_f) models. While model5 shows slightly better adjusted R-squared, both models point to a significant interaction between education and economic development. In practice, we adopt model5 for clarity, but validate our results against model6 as a robustness check.

Based on adjusted R-squared, AIC, and BIC, Model5 outperforms all other candidate models (shown is Table 1). It achieves the highest adjusted R-squared (0.872), the lowest AIC (177.56), and maintains competitive BIC compared to simpler models. This indicates that accounting for heterogeneity in the education–fertility relationship across income groups yields superior explanatory power.

Figure 2 shows the visualisation of the assessment of model5. Residual plots for Model 5 show no strong violations of linear regression assumptions. Residuals are approximately centered, with mild deviation in the QQ plot suggesting light right-skewness or a few outliers. The Scale-Location plot reveals a minor increase in variance with fitted values, indicating mild heteroscedasticity. A few high-leverage points are observed but do not appear to overly influence the model. Overall, the model performs well under classical linear regression assumptions.

Table 1: Model Comparison Summary

| Model | Adj_R2 | AIC | BIC |
|---|---|---|---|
| Model 2 (edu_upper) | 0.863 | 184.548 | 216.341 |
| Model 4 (eys_f) | 0.856 | 204.933 | 237.370 |
| Model 5 | 0.872 | 177.563 | 218.028 |
| Model 6 | 0.862 | 198.294 | 227.781 |

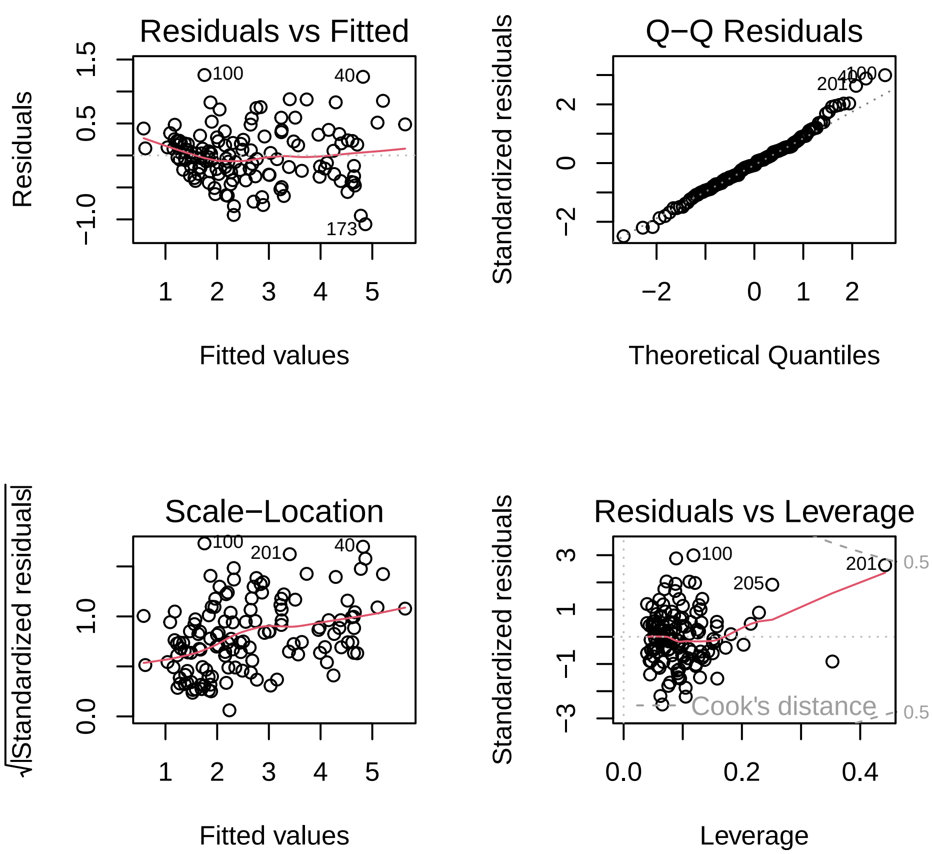


Figure 2: Diagnostic plot of model5

```
#>
#> t test of coefficients:
#>
#>                                      Estimate Std. Error t value  Pr(>|t|)
#> (Intercept)                         2.9065439  0.3922074  7.4107 1.933e-11 ***
#> region_groupC2                      0.3078930  0.1784853  1.7250 0.0870961 .
#> region_groupC3                     -0.2816653  0.1174300 -2.3986 0.0179989 *
#> incomeLower middle income          -0.9798452  0.2997428 -3.2690 0.0014086 **
#> incomeUpper middle income          -1.3023851  0.3496835 -3.7245 0.0002996 ***
#> incomeHigh income                  -1.6205662  0.4489608 -3.6096 0.0004486 ***
#> edu_upper                          -0.0374280  0.0203734 -1.8371 0.0686701 .
#> contraceptive                      -0.0185651  0.0024731 -7.5067 1.175e-11 ***
#> gii                                 3.5565643  0.3805403  9.3461 6.210e-16 ***
#> pr_f                                0.0099128  0.0035937  2.7584 0.0067181 **
#> incomeLower middle income:edu_upper  0.0435892  0.0206004  2.1159 0.0364169 *
#> incomeUpper middle income:edu_upper  0.0405400  0.0208080  1.9483 0.0537151 .
#> incomeHigh income:edu_upper          0.0486495  0.0208463  2.3337 0.0212738 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The differences between the default summary() output and the robust coeftest() results suggest that the original model may exhibit mild heteroskedasticity, and robust inference helps confirm which variables remain reliable under relaxed assumptions. The main interpretation — that education's effect on fertility is stronger in low-income regions — remains robust.

Female education significantly reduces fertility in low-income countries, but this effect diminishes with rising national income. This suggests that education-based interventions may be most effective in high-fertility, low-income contexts, whereas more complex socioeconomic strategies may be required elsewhere.

Figure 3 shows that female education significantly reduces fertility in low-income countries, but this effect diminishes with rising national income. This suggests that education-based interventions may be most effective in high-fertility, low-income contexts, whereas more complex socioeconomic strategies may be required elsewhere.

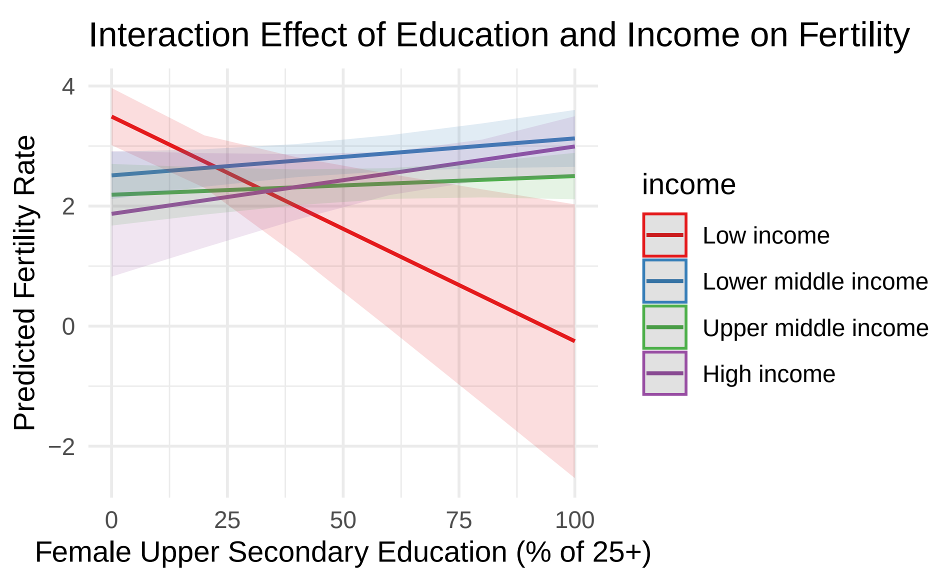


Figure 3: Diagnostic plot of model5

While Model5 performs well under classical linear regression assumptions, some extensions may help address structural and temporal complexities, for example: Panel Modelling (PLM). Since we have panel data, we will next incorporate fixed and random effects to account for unobserved country-specific heterogeneity.

While the cross-sectional models provide important insights, they capture only a static snapshot from 2023. Since the dataset spans over 20 years for most countries, we leverage panel data methods to explore whether the relationships between fertility, education, and economic development persist over time.

For panel modelling, we retain the key covariates from Model5, focusing on female upper secondary education, income group, and their interaction, along with consistent controls (contraceptive prevalence, gender inequality, and women's political representation). These variables demonstrate both theoretical relevance and sufficient variation over time, making them appropriate for longitudinal modelling. Region grouping is excluded in panel models, as country-specific unobserved heterogeneity is captured through fixed or random effects.

I explored the fixed-effect model and random-effect model, and then I used the Hausman test to check the preference. Fixed effects models allow us to control for time-invariant country-level heterogeneity, while random effects models can yield greater efficiency under appropriate assumptions. The Hausman test guides model selection.

```
#> Oneway (individual) effect Within Model
#>
#> Call:
#> plm(formula = fertility ~ income * edu_upper + contraceptive +
#>     gii + pr_f, data = df_panel, model = "within", index = c("iso3c",
#>     "year"))
#>
#> Unbalanced Panel: n = 147, T = 1-24, N = 3206
#>
#> Residuals:
#>      Min.   1st Qu.    Median   3rd Qu.      Max.
#> -1.362140 -0.159483  0.008979  0.178457  1.066944
#>
#> Coefficients:
#>                                       Estimate Std. Error  t-value  Pr(>|t|)
#> edu_upper                            0.0107921  0.0019512   5.5311 3.451e-08 ***
#> contraceptive                       -0.0320630  0.0010517 -30.4859 < 2.2e-16 ***
#> gii                                  1.9658083  0.1555340  12.6391 < 2.2e-16 ***
#> pr_f                                -0.0094592  0.0012971  -7.2925 3.857e-13 ***
#> incomeLow income:edu_upper          -0.0253394  0.0034291  -7.3896 1.892e-13 ***
#> incomeLower middle income:edu_upper -0.0143797  0.0022736  -6.3246 2.910e-10 ***
#> incomeUpper middle income:edu_upper -0.0198341  0.0022394  -8.8568 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    490.99
#> Residual Sum of Squares: 280.42
#> R-Squared:      0.42885
#> Adj. R-Squared: 0.40022
#> F-statistic: 327.378 on 7 and 3052 DF, p-value: < 2.22e-16
#> Oneway (individual) effect Random Effect Model
#>    (Swamy-Arora's transformation)
#>
#> Call:
#> plm(formula = fertility ~ income * edu_upper + contraceptive +
#>     gii + pr_f, data = df_panel, model = "random", index = c("iso3c",
```

```
#>     "year"))
#>
#> Unbalanced Panel: n = 147, T = 1-24, N = 3206
#>
#> Effects:
#>                   var std.dev share
#> idiosyncratic 0.09188 0.30312 0.259
#> individual    0.26281 0.51265 0.741
#> theta:
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>  0.4910  0.8802  0.8802  0.8785  0.8802  0.8802
#>
#> Residuals:
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#> -1.3275 -0.1720  0.0139  0.0027  0.1963  1.1189
#>
#> Coefficients:
#>                                       Estimate  Std. Error  z-value  Pr(>|z|)
#> (Intercept)                          2.84797389  0.17646361  16.1392 < 2.2e-16 ***
#> incomeLow income                     2.30906713  0.18880557  12.2299 < 2.2e-16 ***
#> incomeLower middle income            1.26660133  0.16529497   7.6627 1.821e-14 ***
#> incomeUpper middle income            1.01023170  0.17147019   5.8916 3.825e-09 ***
#> edu_upper                            0.01038728  0.00185266   5.6067 2.062e-08 ***
#> contraceptive                       -0.03154189  0.00099555 -31.6830 < 2.2e-16 ***
#> gii                                  2.11395755  0.14842040  14.2430 < 2.2e-16 ***
#> pr_f                                -0.00770009  0.00123837  -6.2179 5.039e-10 ***
#> incomeLow income:edu_upper          -0.02603949  0.00336269  -7.7436 9.661e-15 ***
#> incomeLower middle income:edu_upper -0.01415069  0.00215288  -6.5729 4.934e-11 ***
#> incomeUpper middle income:edu_upper -0.01950079  0.00213408  -9.1378 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    616.23
#> Residual Sum of Squares: 297.55
#> R-Squared:      0.51724
#> Adj. R-Squared: 0.51573
#> Chisq: 3278.12 on 10 DF, p-value: < 2.22e-16
#>
#>  Hausman Test
#>
#> data:  fertility ~ income * edu_upper + contraceptive + gii + pr_f
#> chisq = 61.67, df = 7, p-value = 6.998e-11
#> alternative hypothesis: one model is inconsistent
```

The p-value is smaller than 0.05, so I can reject the null hypothesis, which means that the fix-effect model is more appropriate to adopt.

While the baseline fixed-effects model focuses on the core explanatory variables drawn from cross-sectional analysis, we also construct an extended model incorporating additional time-varying structural variables—urbanisation, maternal mortality, and female labour force participation. These variables were excluded in the initial cross-sectional specifications due to limited variability, but become informative in a panel setting.

```
#> Oneway (individual) effect Within Model
#>
```

Table 2: Comparison of Baseline and Extended Fixed Effects Panel Models

| Model | R2 | Residual SS | Significant Interaction? | Additional Variables |
|---|---|---|---|---|
| Baseline FE | 0.429 | 280.424 | Yes | None |
| Extended FE | 0.614 | 189.677 | Yes | urban, mmr, female_labor |

```
#> Call:
#> plm(formula = fertility ~ income * edu_upper + contraceptive +
#>     gii + pr_f + urban + mmr + female_labor, data = df_panel,
#>     model = "within", index = c("iso3c", "year"))
#>
#> Unbalanced Panel: n = 147, T = 1-24, N = 3206
#>
#> Residuals:
#>       Min.    1st Qu.     Median    3rd Qu.       Max.
#> -1.0657817 -0.1320361  0.0095425  0.1445733  1.0391492
#>
#> Coefficients:
#>                                       Estimate  Std. Error  t-value  Pr(>|t|)
#> edu_upper                           4.8514e-03  1.6413e-03   2.9559  0.003142 **
#> contraceptive                      -1.8864e-02  9.4318e-04 -20.0008 < 2.2e-16 ***
#> gii                                 9.5734e-01  1.3449e-01   7.1182 1.357e-12 ***
#> pr_f                               -2.2678e-03  1.0840e-03  -2.0920  0.036517 *
#> urban                              -3.6202e-02  1.6783e-03 -21.5702 < 2.2e-16 ***
#> mmr                                 1.5524e-03  7.0135e-05  22.1341 < 2.2e-16 ***
#> female_labor                        8.8645e-03  1.5114e-03   5.8650 4.973e-09 ***
#> incomeLow income:edu_upper         -1.6005e-02  2.8571e-03  -5.6018 2.310e-08 ***
#> incomeLower middle income:edu_upper -6.1442e-03  1.9076e-03  -3.2209  0.001291 **
#> incomeUpper middle income:edu_upper -9.3158e-03  1.8644e-03  -4.9967 6.162e-07 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    490.99
#> Residual Sum of Squares: 189.68
#> R-Squared:      0.61368
#> Adj. R-Squared: 0.59392
#> F-statistic: 484.345 on 10 and 3049 DF, p-value: < 2.22e-16
```

Table 2 indicates that the extended fixed effects model outperforms the baseline specification across all key diagnostic metrics. It achieves a substantially higher adjusted R-squared and a lower residual sum of squares, indicating better overall model fit. Importantly, the inclusion of additional structural controls—urbanisation, maternal mortality (MMR), and female labour force participation—not only enhances explanatory power but also preserves the significance of the core interaction between income and female education. This suggests that the education–fertility relationship is robust to the inclusion of broader development indicators and may be influenced by institutional and demographic factors beyond education and economic status alone.

Our modelling strategy begins with cross-sectional linear models and proceeds to panel models with fixed effects. Across all models, contraceptive prevalence, gender inequality (GII), and female education emerge as the most robust predictors of fertility.

Interaction effects between income and female education are statistically significant and consistent across specifications. Panel modelling confirms the robustness of this relationship over time. Furthermore, additional socioeconomic indicators — such as MMR and urbanisation — improve model fit while preserving core conclusions.