MAS360 project 1

160175125

Influencing factors of GDP and Economic Growth.

Abstract

The report analyses various national statistics of 193 countries to understand their influence on the GDP per capita. The aim was to provide information intended to aid the design of economic policies which are tailored towards fostering economic growth. The analysis consisted of a preliminary data analysis to understand the general structure of the data, followed by a data linearization process which allowed for a less complex formal model to describe the data. Outliers and interactions were considered; all outliers found were attributed to influential observations and not discarded, and no interactions were found necessary to include in the formal model. The formal model implied that only the infant mortality data was needed to accurately predict the GDP per capita of a country. The conclusion suggested that possible good choices for economic policies should be based around reducing the fertility rate which could indirectly decrease the cost of healthcare and improve the populace' ability to work.

Introduction

The report investigates various national statistics (referred to as factors) that may influence the GDP per capita of a country. The aim was to establish relationships between the GDP per capita and those influencing factors, which can then be used to predict GDP per capita by only knowing those factors. This information can then be used to help create effective economic policies which foster economic growth. The data used is national statistics of 193 countries (averaged mostly from 2009-2011) taken from the United Nations [1]. The catagories of the data are as follows;

- `country` Country.
- `region` Region of the world: Africa, Asia, Carribean, Europe, America, Oceania.
- `fertility` Total fertility rate, number of children per woman.
- `ppgdp` Per capita gross domestic product (GDP) in US dollars.
- `lifeExpF` Female life expectancy in years.
- `pctUrban` Pecent of poulation in urban areas.
- `infantMortality` Infant deaths by age 1 year per 1000 live births.

Table 1 shows an extract from the data.

The report first looks at the raw data in the form of both box plots and scatter plots. This was done to help understand the general structure of the data and also to determine any necessary data processing that can be done to help simplify the analysis. The boxplots uncovered a positive relationship between the fertility and infant mortality rate, and an inverse relationship between the female life expectancy and the fertility/infant mortality rate. This implied that if one of these factors could predict GDP per capita, then any of them could, meaning that the formal model would likely contain only one of these factors. The scatter plots of the raw data showed that the GDP per capita was exponentially related to the other factors, which led to the plots being unreadable. To solve the readability issue, the GDP per capita data was log transformed. This has the effect of squashing the data and makes it readable, while preserving relationships in the data.

The next part of the report focuses on finding the most simple component models possible for constructing the formal model. A component model here is defined as a model that accurately describes the relationship between the GDP per capita and one of the other factors; fertility rate, infant mortality rate, female life

Table 1: Extract from the dataset

Country	region	fertility	ppgdp	lifeExpF	pctUrban	infantMortality	colour
Algeria	Africa	2.142	4473.0	75.00	67	21.458	black
Angola	Africa	5.135	4321.9	53.17	59	96.191	black
Benin	Africa	5.078	741.1	58.66	42	76.674	black
Botswana	Africa	2.617	7402.9	51.34	62	35.117	black
Burkina Faso	Africa	5.750	519.7	57.02	27	70.958	black
Burundi	Africa	4.051	176.6	52.58	11	94.083	black

expectancy, or urban population %. The aim was to transform the data (in the same way as for the GDP per capita) whenever necesssary and possible in order to linearize the relationships so that straight line component models could be fit to the data. This was achieved by comparing quadratic models against linear models in analysis of variance tests, after transforming the data appropriately. Once the most simple component models were found, this implied the transformed data was in a form that could be described by a simpler formal model.

Once the data was in it's desired state, the data was then checked for interactions between region, fertility rate, infant mortality rate, female life expectancy, and urban population %. The Possible interactions found between region and the other factors were understood to be artifacts of the data, and it was deemed that including them would overfitting the data, therefore they were not included in the formal model analysis. The model diagnostics of the formal model also showed no evidence of absent interactions. The data was also checked for outliers by inspecting the model diagnostics for the component models. Upon examination, no outliers were found to be the result of errors, and therefore were considered influencial observations and included in the analysis.

The last part of the report covers the model building. The formal model was constructed by concatenating the component models found in the previous section. The formal model was then tested using the method of analysis of variance to see whether it could be simplified by excluding some terms. It was found that the fertility rate and the female life exectancy could be excluded from the model, leaving only the infant mortality rate and urban population % needed to predict the GDP per capita of a country. It was also found than the model could be closely approximated by only knowing infant mortality.

Preliminary Analysis

The preliminary analysis looks at the raw data to understand the general structure of the data. Table 2 shows that the sample sizes for the regions "America", "Carribean", and "Oceania" are quite small, so care was taken about any conclusions draw directly from these samples. The variation between the min, the max, and the 1st & 3rd quantiles is large for all the factors, which indicates large differences between development and population health between countries.

Figure 3 shows box plots of the original data for each catagory, split by region. The plots W1 and W5 are strikingly similar, indicating a positive correlation between them. The correlation coefficient for fertility rate and infant mortality is `0.86`, which is a strong positive correlation. Also similar to the plots of W1 and W5 is the W3 plot, except W3 is an inverse of W1 and W5, which implies a negative correlation between them. The correlation coefficient for female life expectancy and infant mortality is `-0.93`, which is a very strong negative correlation. The relationship between GDP per capita and the others factors can also be seen here. As the GDP per capita increases, the fertility rate and infant mortality decrease, and the female life expectancy increases. The strong correlations between fertility rate, infant mortality rate, and female life expectancy indicate that if one of these factors can predict GDP per capita, then any of them can. This implies the formal model likely needs to include only 1 of these factors.

Appendix 1 shows the original data plotted as a series of scatter plots. In each plot the GDP per capita is

Table 2: Summary of the dataset

	Country	region	fertility	ppgdp	lifeExpF	pctUrban	infantMortality	colour
	Afghanistan: 1	Africa :52	Min. :1.134	Min. : 114.8	Min. :48.11	Min. : 11.0	Min. : 1.916	black :52
	Albania : 1	America :22	1st Qu.:1.750	1st Qu.: 1239.8	1st Qu.:65.10	1st Qu.: 39.0	1st Qu.: 7.243	blue :39
	Algeria : 1	Asia :50	Median :2.264	Median : 4495.8	Median :75.57	Median : 59.0	Median : 19.637	green :13
	Angola : 1	Caribbean:13	Mean :2.780	Mean : 12291.1	Mean :72.08	Mean : 57.1	Mean : 30.739	grey :17
	Argentina : 1	Europe :39	3rd Qu.:3.700	3rd Qu.: 14497.3	3rd Qu.:79.07	3rd Qu.: 75.0	3rd Qu.: 45.892	red :22
	Armenia : 1	Oceania :17	Max. :6.925	Max. :105095.4	Max. :87.12	Max. :100.0	Max. :124.535	yellow:50
	(Other) :187	NA	NA	NA	NA	NA	NA	NA

plotted against one of the other factors in the dataset. All plots show an exponential looking relationship between the data, and all plots are unreadable. The log transformed GDP per capita was taken to solve this issue. Appendix 2 shows the same plots as Appendix 1 except the GDP per capita has been rescaled with a log transformation. The data is now readable. Only plot B41 looks to be linear, the rest look to be quadratic relationships. The outliers were not checked at this point for reasons explained later.

The preliminary analysis found that the factors of the dataset are generally strongly correlated to each other and there are some very clear relationships that appear. Most of the relationships however are quadratic which is not ideal for model building and needed to be linearized where possible to simplify the formal model. This is done in the model components analysis section.

Model components analysis

In this section, for each of the factors, fertility rate, infant mortality, female life expectancy, and urban population %, the most simple model explaining the relationship between the factor and the GDP per capita was found. To find the simplest models, the data seen in Appendix 2 was linearized as much as possible using log transformations, which allowed more straight line models to be fitted to the data. The models found in this part of the analysis are the component models that were concatenated to create the most simple, full version of the formal model.

To linearize the realtionships seen in Appendix 2, a series of ANOVA tests were used to compare a linear model of the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

against a quadratic model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

The goal was to transform the data in such a way that the linear model became a better fit than the quadratic model, meaning the relationship between the data could then be assumed linear and not quadratic. The series of ANOVA tests and the list of models used in the ANOVA tests can be found in Appendix 3.

Figure 2 shows the data after the linearization process in Appendix 3. Only 1 plot (C1) is now assumed to be a quadratic relationship as oppose to the 3 plots in Appendix 2, the rest are assumed linear. This transformed form of the data is used for the formal model analysis. The component models used for the analysis are as follows;

$$\begin{aligned} N2 : \log(y_i) &= \beta_0 + \beta_1 l_i + \beta_2 l_i^2, \\ N7 : \log(y_i) &= \beta_0 + \beta_1 \log(m_i), \\ N11 : \log(y_i) &= \beta_0 + \beta_1 \log(f_i), \\ N13 : \log(y_i) &= \beta_0 + \beta_1 u_i, \end{aligned}$$

where,

- y_i: GDP per capita of country i,

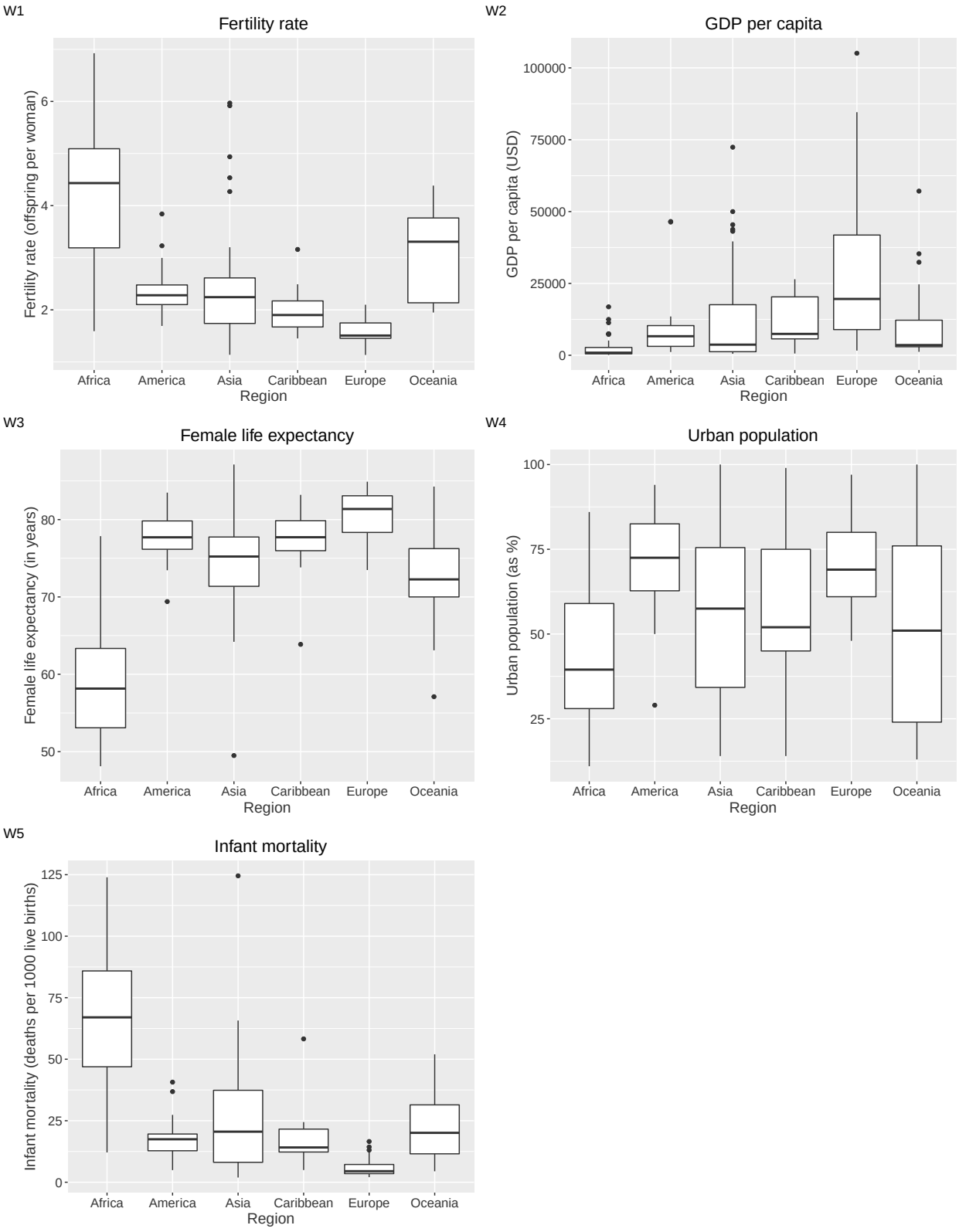

Figure 1: Box plots of the data, split by region. Plots W2 and W3 are positively correlated. Plots W1 and W5 are positively correlated. Plots W2 and W3 are negatively correlated to plots W1 and W5. As female life expectancy and GDP per capita go up, fertility rate and infant mortality rate go down.

Table 3: Outliers from the model diagnostics

	Country	region	fertility	ppgdp	lifeExpF	pctUrban	infantMortality	colour
2	Angola	Africa	5.135	4321.9	53.17	59	96.191	black
4	Botswana	Africa	2.617	7402.9	51.34	62	35.117	black
17	Equatorial Guinea	Africa	4.980	16852.4	52.91	40	93.315	black
20	Gabon	Africa	3.195	12468.8	64.32	86	43.770	black
27	Liberia	Africa	5.038	218.6	58.59	48	76.853	black
43	Somalia	Africa	6.283	114.8	53.38	38	100.017	black
44	South Africa	Africa	2.383	7254.8	54.09	62	45.892	black
84	North Korea	Asia	1.988	504.0	72.12	60	25.053	yellow
89	Qatar	Asia	2.204	72397.9	78.24	96	8.195	yellow
115	Trinidad and Tobago	Caribbean	1.632	15205.1	73.82	14	24.458	green
120	Bosnia and Herzegovina	Europe	1.134	4477.7	78.40	49	12.695	blue
134	Latvia	Europe	1.506	10663.0	78.51	68	6.700	blue
138	Moldova	Europe	1.450	1625.8	73.48	48	14.344	blue

- f_i: the female life expectancy of country i,
- l_i: the infant mortality rate of country i,
- m_i: the fertility rate of country i,
- u_i: is the % of country i's population living in urban areas.

N2 describes the data in plot C1, N7 describes plot C2, N11 describes C3, and N13 describes C4.

Outliers

Table 3 shows all the outliers found by the component model diagnostics (see Appendix 4). The outliers from the model diagnostics were checked rather than the boxplots because region is an arbitrary way to split up the countries, and has no influence itself on whether a country has high or low measurements of a particular statistic. For example just because a region on the whole has a low fertility rate for instance, does not imply a country in that region can not have a high fertlity rate; the same argument applies to the other factors. Therefore the outliers on the boxplots may not be outliers in the dataset as a whole.

After examining the numbers, there is no real reason to believe that any of the outlying observations are caused by errors in the data rather than being real influential observations. There are no obvious mistakes in the data, and since there are quite a high number of outliers, and the variation in the dataset as a whole is large, it is very plausible that the values are real aspects of the dataset. Therefore the outlying observations were accepted as part of the variation in the data and not omitted from the analysis.

Interactions

Figure 3 shows the same plots as figure 2 except the best fit lines through the data for each region are shown. Plots D2 and D4 show no signs of an interaction between infant mortality rate and region, or urban population % and region since the best fit lines are quite similar. The difference in gradient is most likely due to the small sample sizes for some of the regions. Plot D3 has similar best fit lines except for the best fit line for Europe. The orientation of the best fit line for Europe is likely an artifact of the data because the data for Europe is bunched in the top left corner and not spread across the plot. Therefore the best fit line for Europe is not assumed to be the consequence of a real interaction between fertility rate and region, and fitting for this interaction would therefore be overfitting the data. Plot D1 was established as a quadratic relationship between the data. Although the best fit lines do differ, this is due to where the data for each region is found on the plot. If the quadratic best fit line and various tangent lines to the quadratic model were plotted, the same effect would be produced with the tangent lines. So the difference between the lines is a consequence of fitting straight lines to quadratic data, rather than an interaction between fertility rate and region.

No evidence of interactions between region and the other factors has been found in the plots. Interactions between the factors themselves was not considered until after the model diagnostics of the formal model was checked. It wouldn't make sense to think about further interactions if the formal model fit the data well without them.

Formal Model Analysis.

The formal model analysis shows the process of contructing the formal model from the established component models found in the previous section. The analysis starts with the most complex model which is the concatenation of all the component models and is shown below,

$$\log(y_i) = \beta_0 + \beta_1 f_i + \beta_2 f_i^2 + \beta_3 \log(l_i) + \beta_4 \log(m_i) + \beta_5 u_i.$$

- y_i: GDP per capita of country i.
- f_i: the female life expectancy of country i.
- l_i: the infant mortality rate of country i.
- m_i: the fertility rate of country i.
- u_i: is the % of country i's population living in urban areas.

ANOVA tests were then conducted to find out whether some parts of the full were redundent and could be removed. The tests compared the current full model with it's corresponding nested models to find the most simple model that sufficiently describes the transformed data. If a nested model was found to be sufficient in explaining the data, then the nested model took the current full models' place to become the new full model. The new full model was then compared with it's nested models and so on. The list of models which were used in the ANOVA tests is below.

$$\begin{aligned}
M0 &: \log(y_i) = \beta_0 + \beta_1 f_i + \beta_2 f_i^2 + \beta_3 \log(l_i) + \beta_4 \log(m_i) + \beta_5 u_i \\
M1 &: \log(y_i) = \beta_0 + \beta_1 f_i + \beta_2 f_i^2 + \beta_3 \log(l_i) + \beta_4 \log(m_i) \\
M2 &: \log(y_i) = \beta_0 + \beta_1 f_i + \beta_2 f_i^2 + \beta_3 \log(l_i) + \beta_5 u_i \\
M3 &: \log(y_i) = \beta_0 + \beta_1 f_i + \beta_2 f_i^2 + \beta_4 \log(m_i) + \beta_5 u_i \\
M4 &: \log(y_i) = \beta_0 + \beta_3 \log(l_i) + \beta_4 \log(m_i) + \beta_5 u_i \\
M5 &: \log(y_i) = \beta_0 + \beta_1 f_i + \beta_2 f_i^2 + \beta_4 \log(m_i) \\
M6 &: \log(y_i) = \beta_0 + \beta_1 f_i + \beta_2 f_i^2 + \beta_5 u_i \\
M7 &: \log(y_i) = \beta_0 + \beta_4 \log(m_i) + \beta_5 u_i \\
M8 &: \log(y_i) = \beta_0 + \beta_4 \log(m_i) \\
M9 &: \log(y_i) = \beta_0 + \beta_5 u_i
\end{aligned}$$

Tables 4 to 12 show the results of each ANOVA test. The full models were rejected in favour of the nested models if the P-value of the test was greater than 0.05. Tables 4 to 7 show the ANOVA tests for the full model M0 and it's nested models, M1, M2, M3, and M4; there was no evidence to reject the M2 model in favour of M0. Models M1, M3, and M4 were all rejected in favour of the model M0. Since M2 was not rejected in favour of M0, M2 becomes the new full model. Tables 8 to 10 show the ANOVA tests for the full model M2 and it's nested models, M5, M6, and M7; there was no evidence to reject the M7 model in favour of the M2 model. Models M5 and M6 were rejected in favour of M2. Since M7 was not rejected in favour of M2, M7 becomes the new full model. Tables 11 and 12 show the ANOVA tests for the full model M7 and it's nested models, M8, and M9; there was no evidence to reject the M7 model in favour of M8 or M9. Therefore M7 is taken as the best model to describe the data.

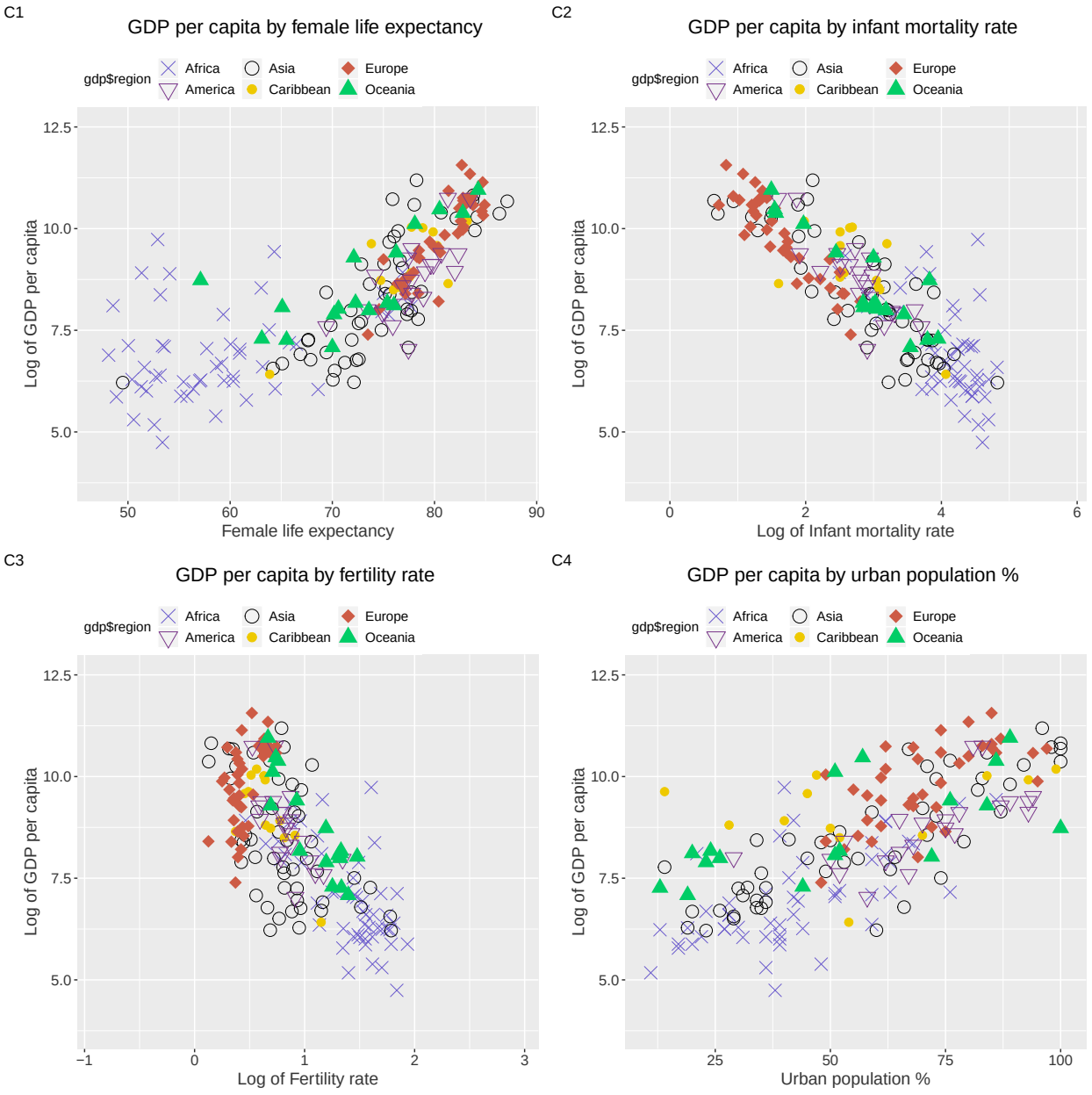

Figure 2: Plots of the log of the GDP per capita and the other trandformed factors. Only plot C1 now shows a quadratic relationship. Plots C2, C3, and C4 are all linear relationships. The model N2 describes the data in plot C1, N7 describes plot C2, N11 describes C3, and N13 describes C4.

Table 4: ANOVA test for M1 versus M0. Reject M1

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
188	102.22266	NA	NA	NA	NA
187	83.93816	1	18.2845	40.73478	0

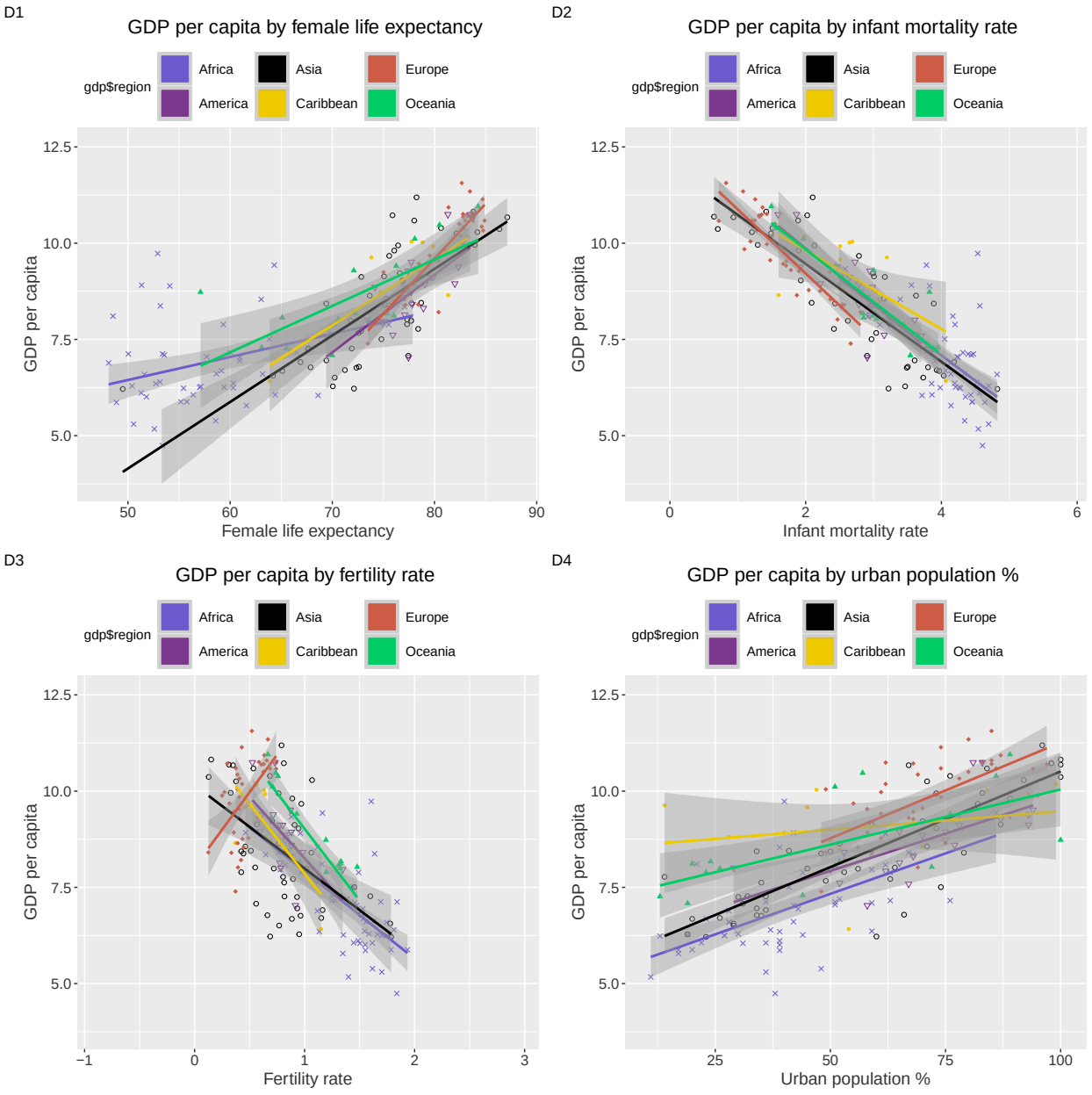

Figure 3: Best fit lines for each region for the transformed data. Differences in the gradients of the best fit lines are likely due to small sample sizes in some of the regions. Plot S3 shows a big difference in the best fit line for Europe, which was deemed an artifact of the data being bunched up in the top left corner. Plot D1 shows differences in the best fit lines as a consequence of fitting straight lines to quadratic data.

Table 5: ANOVA test for M2 versus M0. Reject M0

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
188	85.58787	NA	NA	NA	NA
187	83.93816	1	1.649711	3.675277	0.0567494

Table 6: ANOVA test for M3 versus M0. Reject M3

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
188	124.56128	NA	NA	NA	NA
187	83.93816	1	40.62312	90.50144	0

Table 7: ANOVA test for M4 versus M0. Reject M4

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
189	86.69919	NA	NA	NA	NA
187	83.93816	2	2.761039	3.075563	0.0485063

Table 8: ANOVA test for M5 versus M2. Reject M5

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
189	104.46672	NA	NA	NA	NA
188	85.58787	1	18.87885	41.46877	0

Table 9: ANOVA test for M6 versus M2. Reject M6

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
189	135.51436	NA	NA	NA	NA
188	85.58787	1	49.9265	109.6672	0

Table 10: ANOVA test for M7 versus M2. Reject M2

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
190	87.01760	NA	NA	NA	NA
188	85.58787	2	1.429735	1.570258	0.2107069

Table 11: ANOVA test for M8 versus M7. Reject M8

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	107.2867	NA	NA	NA	NA
190	87.0176	1	20.26911	44.25692	0

Table 12: ANOVA test for M9 versus M7. Reject M9

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	207.9349	NA	NA	NA	NA
190	87.0176	1	120.9173	264.0188	0

Model diagnostics

Model diagnostics were used to determine how well the model M7 fits the data, and whether any interactions were missing from the model. Figure 4 shows the diagnostics for the formal model M7. The QQ plot shows that the residuals for the model are close to normally distributed with slightly heavy tails. The residual plot shows no obvious pattern which indicates the model has no missing interactions that need to be considered. The R-squared statisic for M7 is 0.81 (see Appendix 5 for model summary) which says the model explains over 80% of the variation in the data. The diagnostics imply that the model M7 is complete and describes the data well, and so M7 was taken as the formal model.

Conclusion

The model chosen to describe the data is M7,

$$\log(y_i) = 10.182 - 0.976 * \log(m_i) + 0.019 * u_i$$

The result of the analysis implies that the GDP per capita of a country can be predicted accurately by only knowing the infant mortality rate and proportion of the population that lives in urban areas. Inspecting the co-efficients on the model suggests that only knowing the infant mortality rate of a country will give a very close approximation to the model too. It seems more likely however that infant mortality is dependent on the GDP per capita of a country rather than the other way around, and that high infant mortality is a symptom of a country being too poor to afford effective healthcare. Since fertility rate is strongly positively correlated to infant mortality rate, fertility should also be a good predictor of GDP per capita. Economic policies designed at reducing infant mortality would likely not be effective in increasing the GDP of a country. However ecomomic policies designed to lower the fertility rate in a country may actually be effective in fostering economic growth as reducing the population, or the growth in population, puts less of a burden on the healthcare system which would indirectly increase the quality of the healthcare system. This would decrease the cost of care for people and increase their ability to work.

References

[1] UNSD. *Statistics Division*. https://unstats.un.org/home/. [Online] Accessed: Novermber 2018.

Appendix

Appendix.1

See figure 5.

Appendix.2

See figure 6.

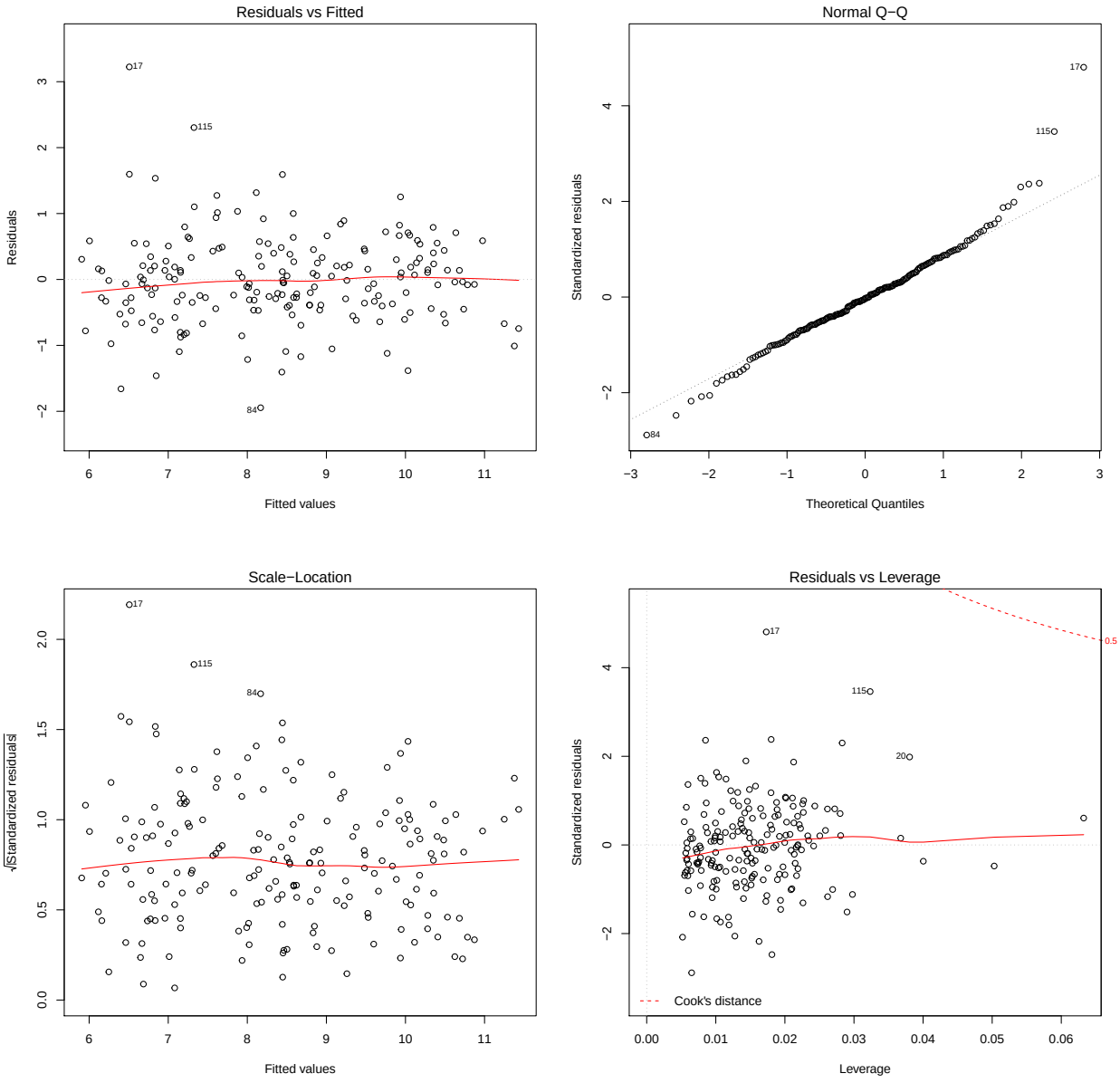

Figure 4: Model Diagnostics for M7. The residual plot shows no signs of an obvious pattern, The QQ-plot looks normally distributed with some slightly heavy tails. There was no evidence to suggest M7 is an incomplete model.

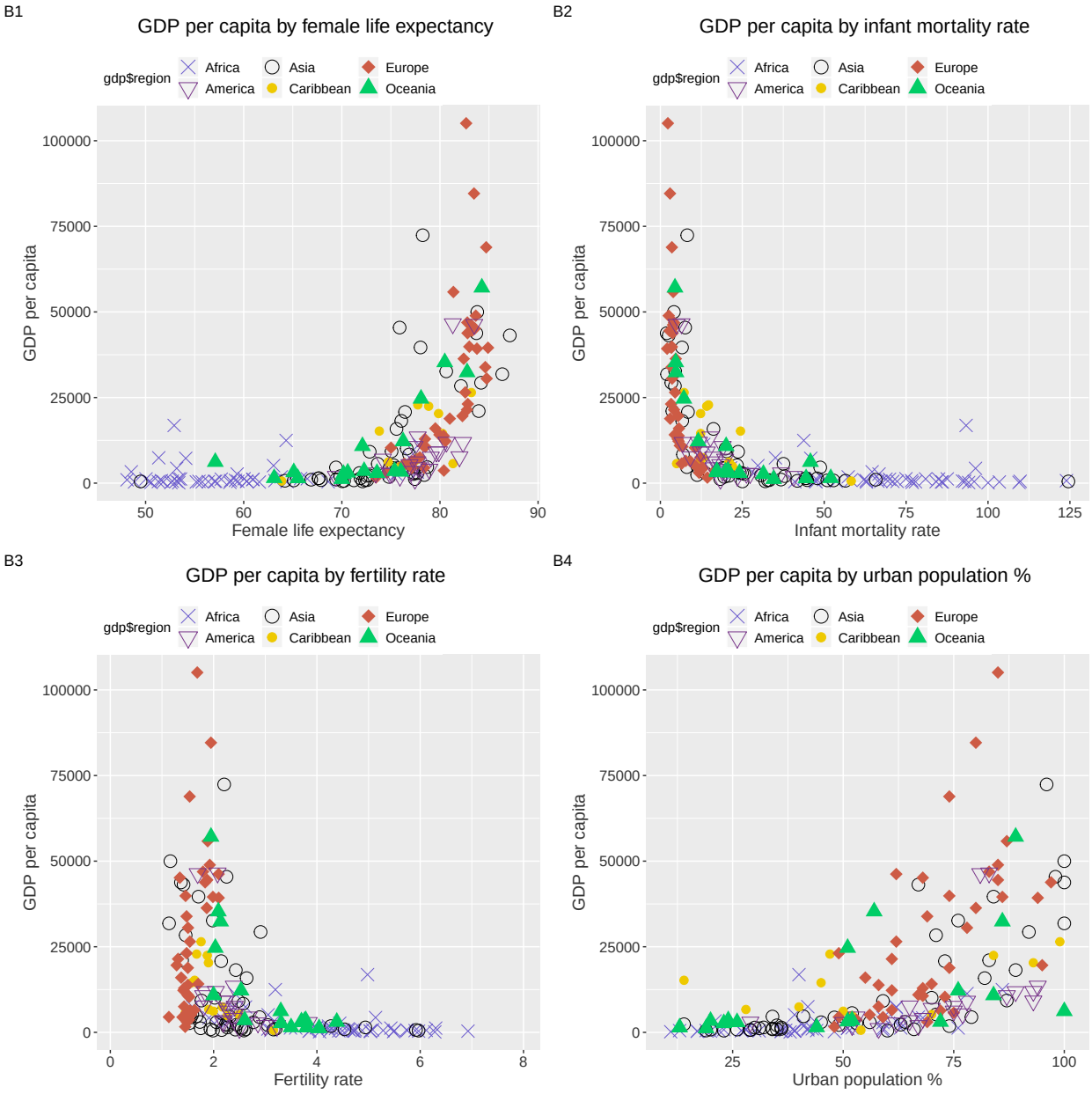

Figure 5: Plots of GDP per capita by female life expectancy, infant mortality rate, fertility rate, and urban population %. Each plot shows and exponential relationship between GDP per capita and the factor. All plots are unreadable and not useful. The data must therefore be transformed appropriately.

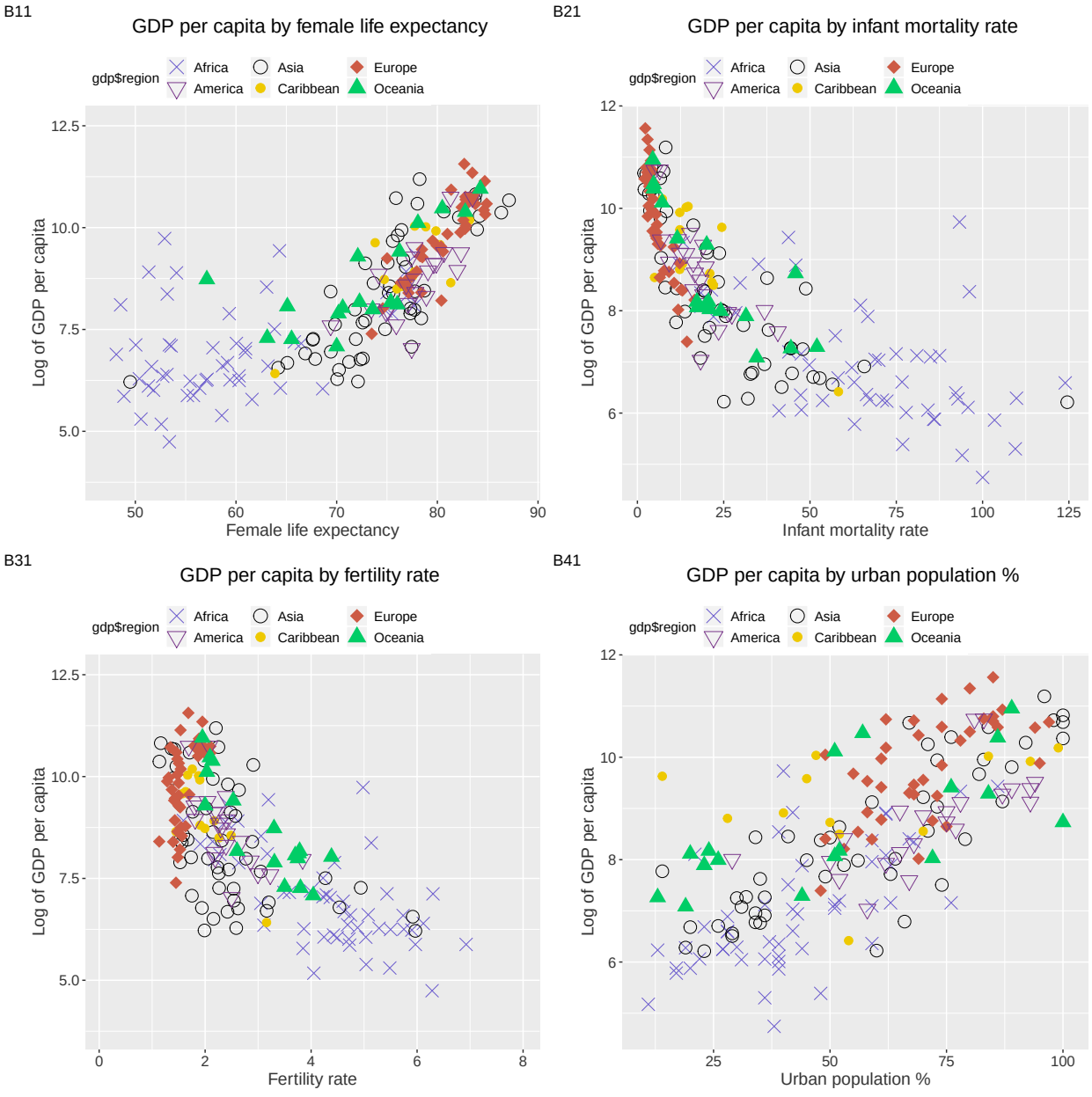

Figure 6: Plots of the log of the GDP per capita by female life expectancy, infant mortality rate, fertility rate, and urban population %. Each plot is now readable. Only B41 looks to be linear. Plots B11, B21,and B31 look to be quadratic relationships and should be linearized for the formal model analysis.

Table 13: ANOVA Linearity test for N1 versus N2. Reject N1.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	184.8164	NA	NA	NA	NA
190	135.3677	1	49.4487	69.40543	0

Table 14: ANOVA Linearity test for N3 versus N4. Reject N3.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	202.0546	NA	NA	NA	NA
190	137.8614	1	64.19324	88.47088	0

Appendix.3 (Component models)

In each of the tables for the ANOVA tests, if the P-value was greater than 0.05, then this was deemed sufficient evidence to reject the quadratic model in favour of the linear model. The goal was to transform the data to achieve P-values greater than 0.05 to linearize the data.

List of linear and quadratic models used in the ANOVA tests

$$
\begin{aligned}
N1 &: \log(y_i) = \beta_0 + \beta_1 l_i, \\
N2 &: \log(y_i) = \beta_0 + \beta_1 l_i + \beta_2 l_i^2, \\
N3 &: \log(y_i) = \beta_0 + \beta_1 \log(l_i), \\
N4 &: \log(y_i) = \beta_0 + \beta_1 \log(l_i) + \beta_1 \log(l_i)^2, \\
N5 &: \log(y_i) = \beta_0 + \beta_1 m_i, \\
N6 &: \log(y_i) = \beta_0 + \beta_1 m_i + \beta_2 m_i^2, \\
N7 &: \log(y_i) = \beta_0 + \beta_1 \log(m_i), \\
N8 &: \log(y_i) = \beta_0 + \beta_1 \log(m_i) + \beta_1 \log(m_i)^2, \\
N9 &: \log(y_i) = \beta_0 + \beta_1 f_i, \\
N10 &: \log(y_i) = \beta_0 + \beta_1 f_i + \beta_2 f_i^2, \\
N11 &: \log(y_i) = \beta_0 + \beta_1 \log(f_i), \\
N12 &: \log(y_i) = \beta_0 + \beta_1 \log(f_i) + \beta_1 \log(f_i)^2, \\
N13 &: \log(y_i) = \beta_0 + \beta_1 u_i, \\
N14 &: \log(y_i) = \beta_0 + \beta_1 u_i + \beta_2 u_i^2,
\end{aligned}
$$

where,

- y_i: GDP per capita of country i,
- f_i: the female life expectancy of country i,
- l_i: the infant mortality rate of country i,
- m_i: the fertility rate of country i,
- u_i: is the % of country i's population living in urban areas.

ANOVA tests for GDP per capita vs female life expectancy.

The relationship between the log of the GDP per capita and female life expectancy is assumed quadratic on the basis of the P-value in Table 13. The relationship between the log of the GDP per capita and the log of the female life expectancy is assumed quadratic on the basis of the P-value in Table 14. Since the P-values in

Table 15: ANOVA Linearity test for N5 versus N6. Reject N5.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	175.6392	NA	NA	NA	NA
190	124.8617	1	50.77753	77.26733	0

Table 16: ANOVA Linearity test for N7 versus N8. Reject N8.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	107.2867	NA	NA	NA	NA
190	107.1533	1	0.1334481	0.236625	0.627215

both tests strongly imply a quadratic relationship, it was not possible to linearize ths relationship with log transformations. Therefore the original female life expectancy data was used for the formal model analysis and the N2 model was used as the component model for the formal model construction.

ANOVA tests for GDP per capita vs infant mortality.

The relationship between the log of the GDP per capita and the infant mortality rate is assumed quadratic on the basis of the P-value in Table 15. The relationship between the log of the GDP per capita and the log of the infant mortality rate is assumed linear on the basis of the P-value in Table 16. It was possible to linearize this relationship with log transformations. Therefore the log transformed infant mortality rate data was used for the formal model analysis and the N7 model was used as the component model for the formal model construction.

ANOVA tests for GDP per capita vs fertility.

The relationship between the log of the GDP per capita and the the fertility rate is assumed quadratic on the basis of the P-value in Table 17. The relationship between the log of the GDP per capita and the log of the fertility rate is assumed linear on the basis of the P-value in Table 18. It was possible to linearize ths relationship with log transformations. Therefore the log of the fertility rate data was used for the formal model analysis and the N11 model was used as the component model for the formal model construction.

ANOVA tests for GDP per capita vs urban population %.

The relationship between the log of the GDP per capita and the urban population % is assumed linear on the basis of the P-value in Table 19. There was no need to trasnform the urban population %. Therefore the original urban poppulation % data was used for the formal model analysis and the N13 model was used as the component model for the formal model construction.

The model diagnostics for the component models N2, N7, N11, and N13 can be found in Appendix 4.

Table 17: ANOVA Linearity test for N9 versus N10. Reject N9.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	220.2738	NA	NA	NA	NA
190	213.3616	1	6.912181	6.155345	0.0139705

Table 18: ANOVA Linearity test for N11 versus N12. Reject N12.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	216.7101	NA	NA	NA	NA
190	214.7795	1	1.930575	1.707841	0.1928452

Table 19: ANOVA Linearity test for N13 versus N14. Reject N14.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
191	207.9349	NA	NA	NA	NA
190	206.4563	1	1.478576	1.360721	0.2448746

Appendix.4 (Model diagnostics for the component models)

Model diagnostics for N2 (female life expectancy)

Figure 8 shows the diagnostics for the model N2. The QQ plot shows a bit of a skewed distribution for the model, most likely coming from the large variation in the Africa data. The residual plot shows no obvious patterns. Therefore the component model looks to describe the data well.

Model diagnostics for N7 (infant mortality)

Figure 9 shows the diagnostics for the model N7. The QQ plot shows the data is very close to normally distributed around the model. The residual plot shows no obvious patterns. Therefore the component model looks to describe the data well.

Model diagnostics for N11 (fertility)

Figure 10 shows the diagnostics for the model N11. The QQ plot shows the data is very close to normally distributed around the model. The residual plot shows no obvious patterns. Therefore the component model looks to describe the data well.

Model diagnostics for N13 (urban population %)

Figure 11 shows the diagnostics for the model N13. The QQ plot shows the data is very close to normally distributed around the model. The residual plot shows no obvious patterns. Therefore the component model looks to describe the data well.

Appendix.5 (Summary of the M7 model)

```
##
## Call:
## lm(formula = log(ppgdp) ~ log(infantMortality) + pctUrban)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9460 -0.3898 -0.0163  0.3823  3.2243
##
## Coefficients:
```

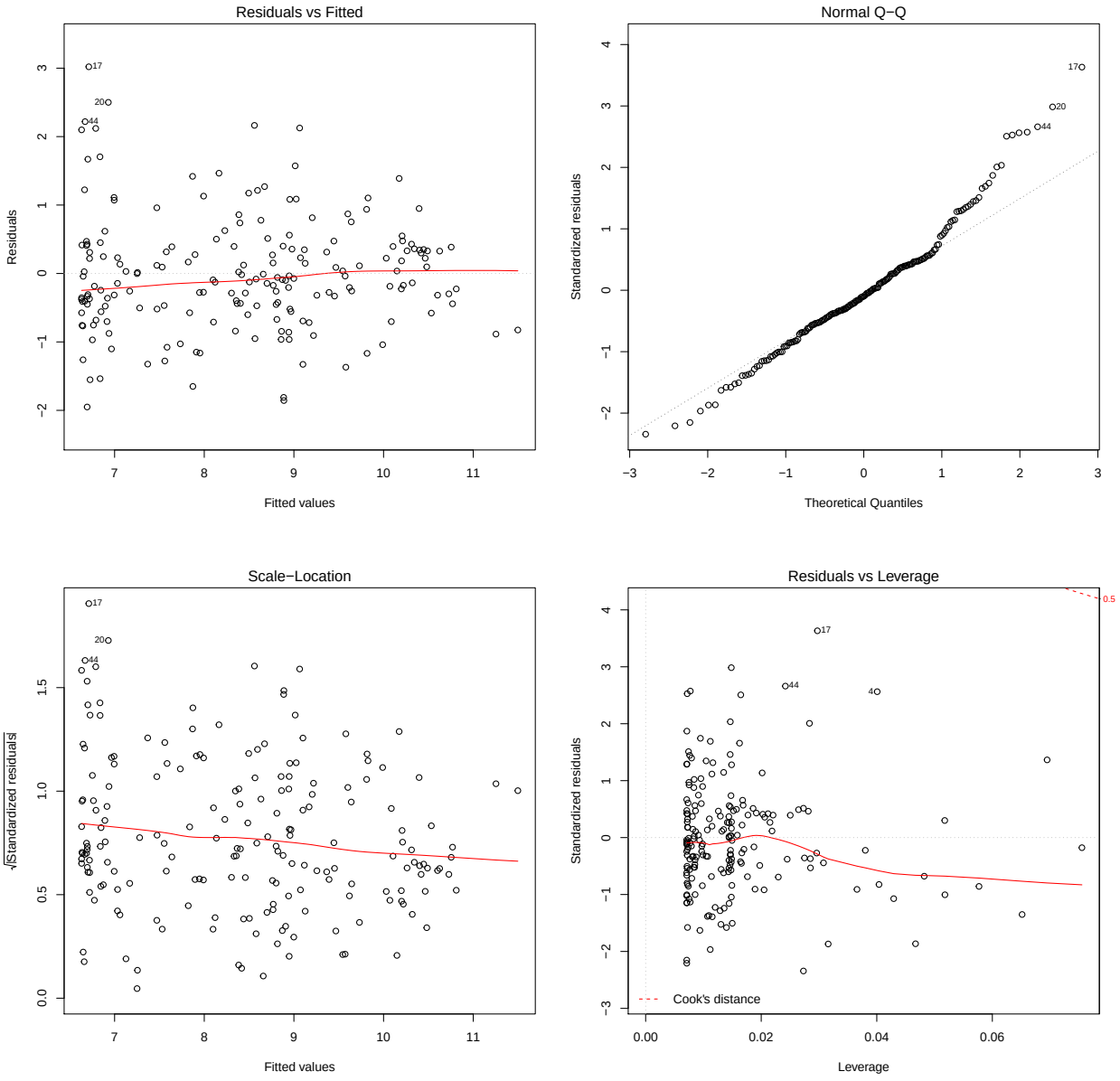

Figure 7: Model diagnostics for N2. The residual plot shows no signs of undyling patterns. The QQ-plot shows a skewed distribution, likely because of the high variation in the africa data. The numbered outliers were checked for validity in the outliers section.

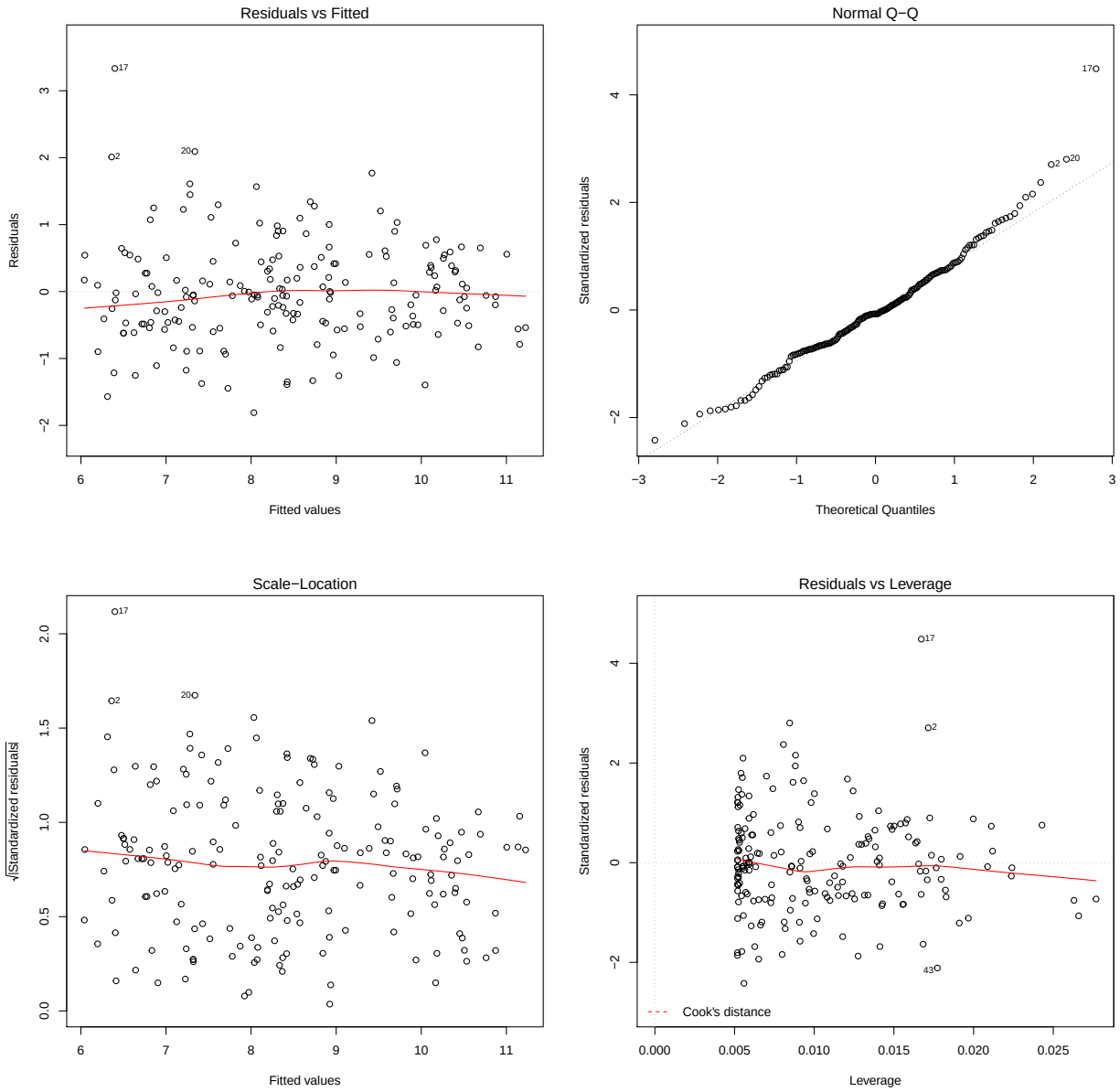

Figure 8: Model diagnostics for N7. The residual plot shows no signs of undyling patterns. The QQ-plot shows a normal looking distribution. The numbered outliers were checked for validity in the outliers section.

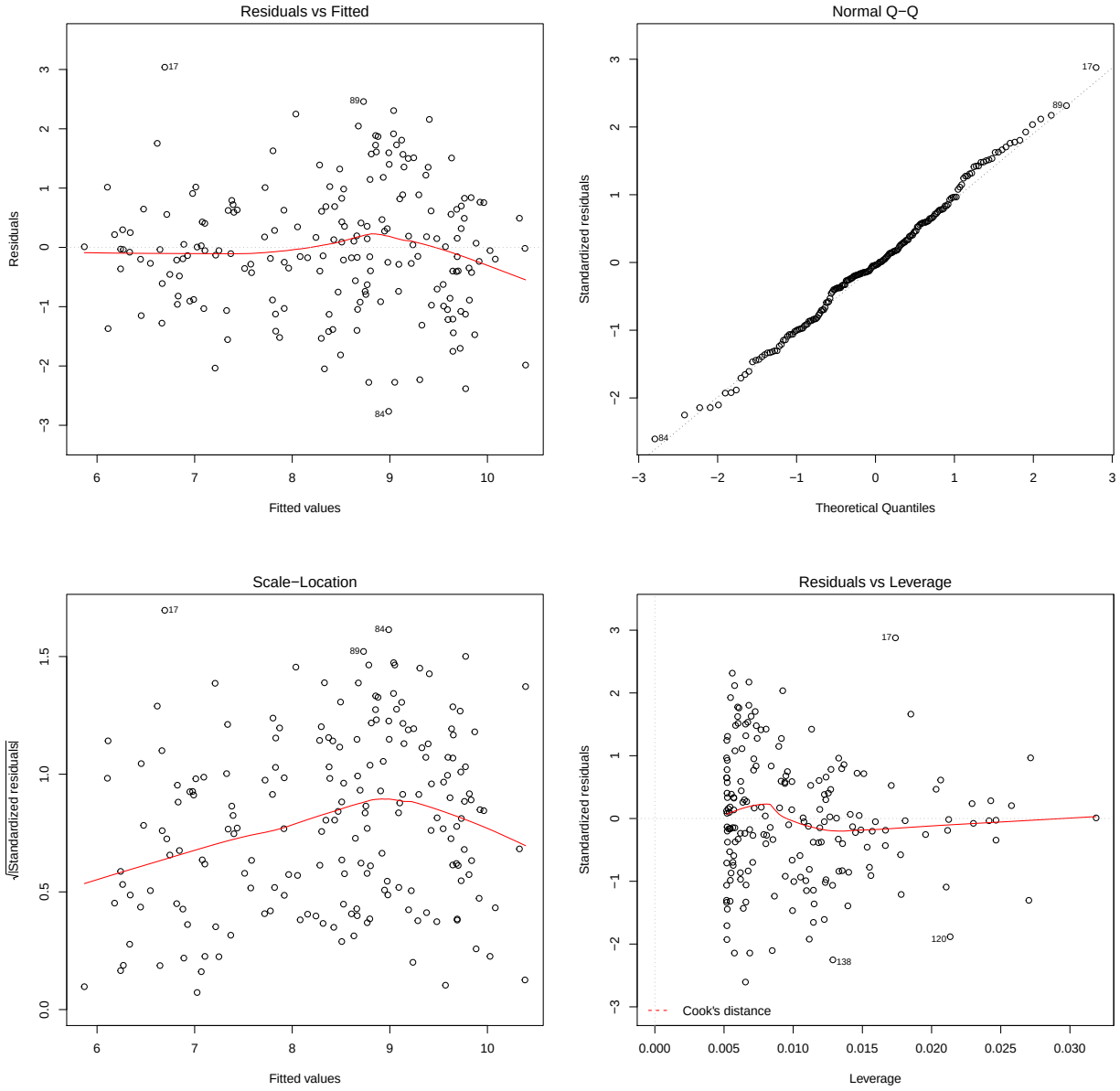

Figure 9: Model diagnostics for N11. The residual plot shows no signs of undyling patterns. The QQ-plot shows a very good looking distribution. The numbered outliers were checked for validity in the outliers section.

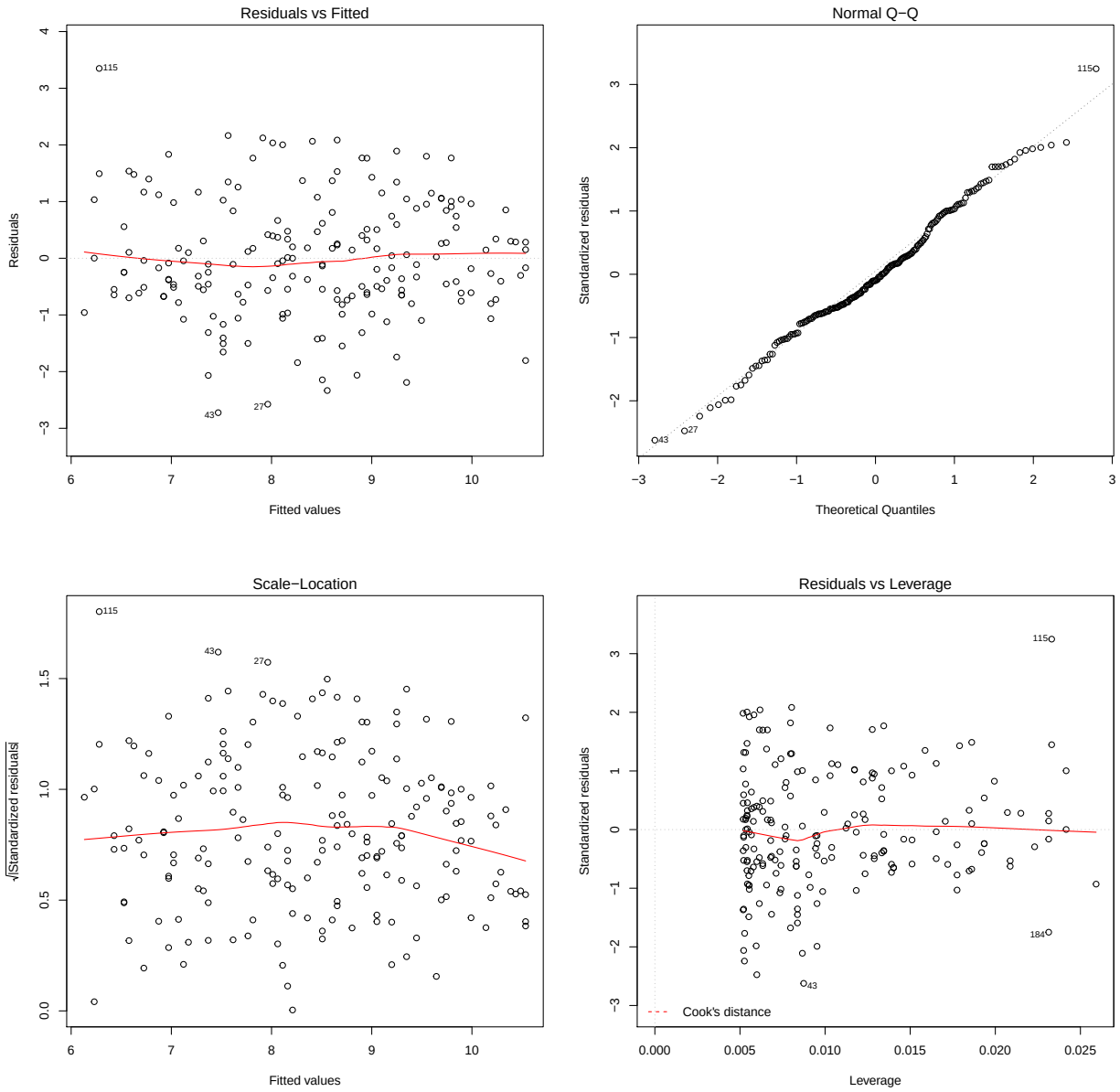

Figure 10: Model diagnostics for N13. The residual plot shows no signs of undyling patterns. The QQ-plot shows a normal looking distribution. The numbered outliers were checked for validity in the outliers section.

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          10.182110   0.311297  32.709  < 2e-16 ***
## log(infantMortality) -0.976172   0.060077 -16.249  < 2e-16 ***
## pctUrban              0.018844   0.002833   6.653 2.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6767 on 190 degrees of freedom
## Multiple R-squared:  0.8104, Adjusted R-squared:  0.8084
## F-statistic: 405.9 on 2 and 190 DF,  p-value: < 2.2e-16
```