Linear Regression, Robustness and Smoothing Practical

Name: Dominic Lee
Student ID: 201691989

All R code can be found in the order I used it in the appendix at the end.
All propositions/theorems/lemmas mentioned refer to the online notes.
'LE' means life.expectancy', 'GPC' means GDP.per.capita, 'HS' means health.spending, 'PS' means population.size, 'PD' means population.density

1 Task 1

Since there are no NA values for LE or GPC, I didn't have to remove anything. I first plotted a scatterplot of LE vs GPC (fig. 1) to get an idea for how the data looks. I also included the command 'stat_smooth()' to help me spot any patterns in the data.

To me, it looked like there was a clear curve in the data, especially at low GDP.per.capita. To me it looked like a log curve, so I used a log transformation of GPC and plotted LE vs logGPC (fig. 2). This looked much better immediately, but there might still be room for improvement. I thought it would make sense if the variance were positively correlated to life.expectancy in some way, although from the scatterplot it looks like the variance is higher for **both** low and high life.expectancy.

Despite this, I tried using the transformations $y' = \sqrt{y}$ and $y' = \log(y)$. However the scatterplots (fig. 3) and (fig.4) don't look any better than without any transformation, and the R^2 values are 0.6193 and 0.5941 resp. with the transformations and 0.64 without. I also decided to plot the residuals vs the fitted values (fig.5, fig.6 and fig.7) and weirdly it actually seemed like the variance was slightly higher for smaller fitted values, so I decided these transformations don't actually improve the model.

The R^2 value of 0.64 isn't particularly high, but since LE probably depends on many of the other factors, I'm satisfied with it in this case, and in my opinion the most obvious improvement (variance stabilisation) didn't work. The final model was:

$$\text{LE} = \beta_0 + \beta_1 \log(\text{GPC})$$

with $\hat{\beta_0} = 32.844, \hat{\beta_1} = 4.509$. To get a 95% CI for the life expectancy of a country with 5000 GDP per capita, I used the 'predict' function in R. It gave a 95% CI of $[70.87039, 71.63363]$.

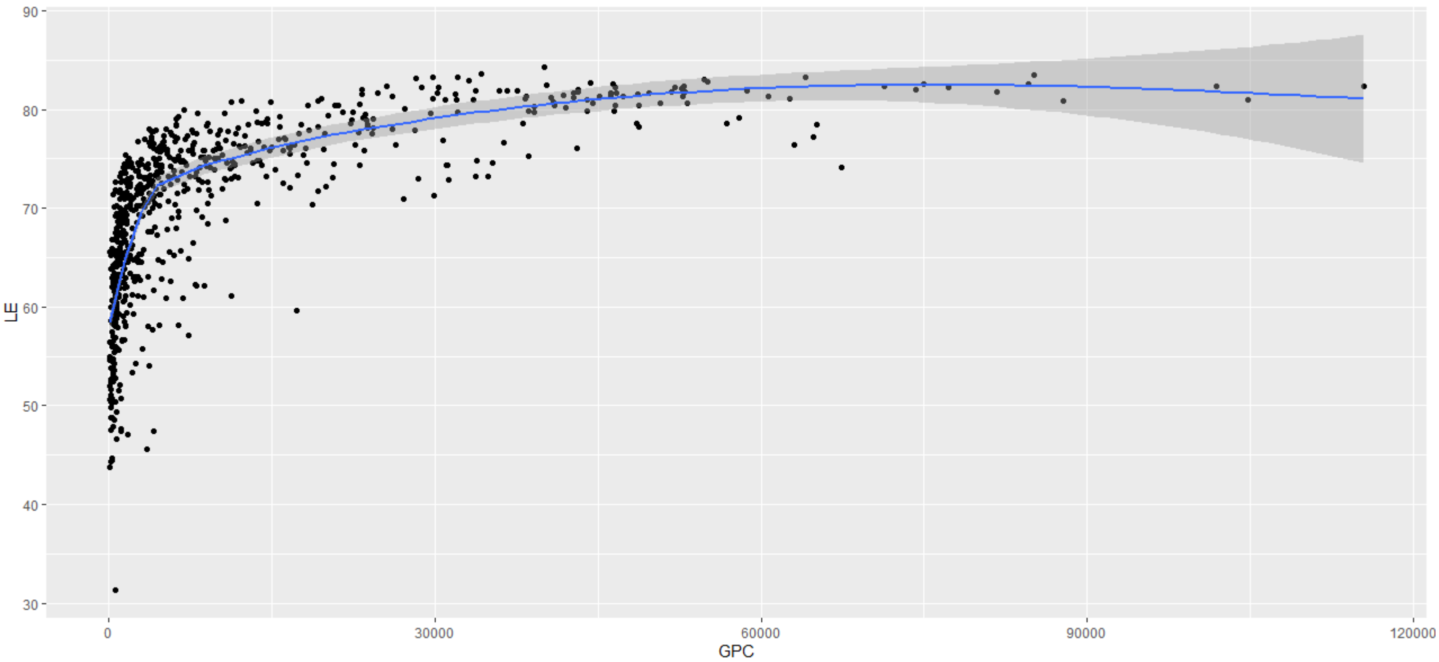

Figure 1: Scatterplot of life.expectancy vs GDP.per.capita

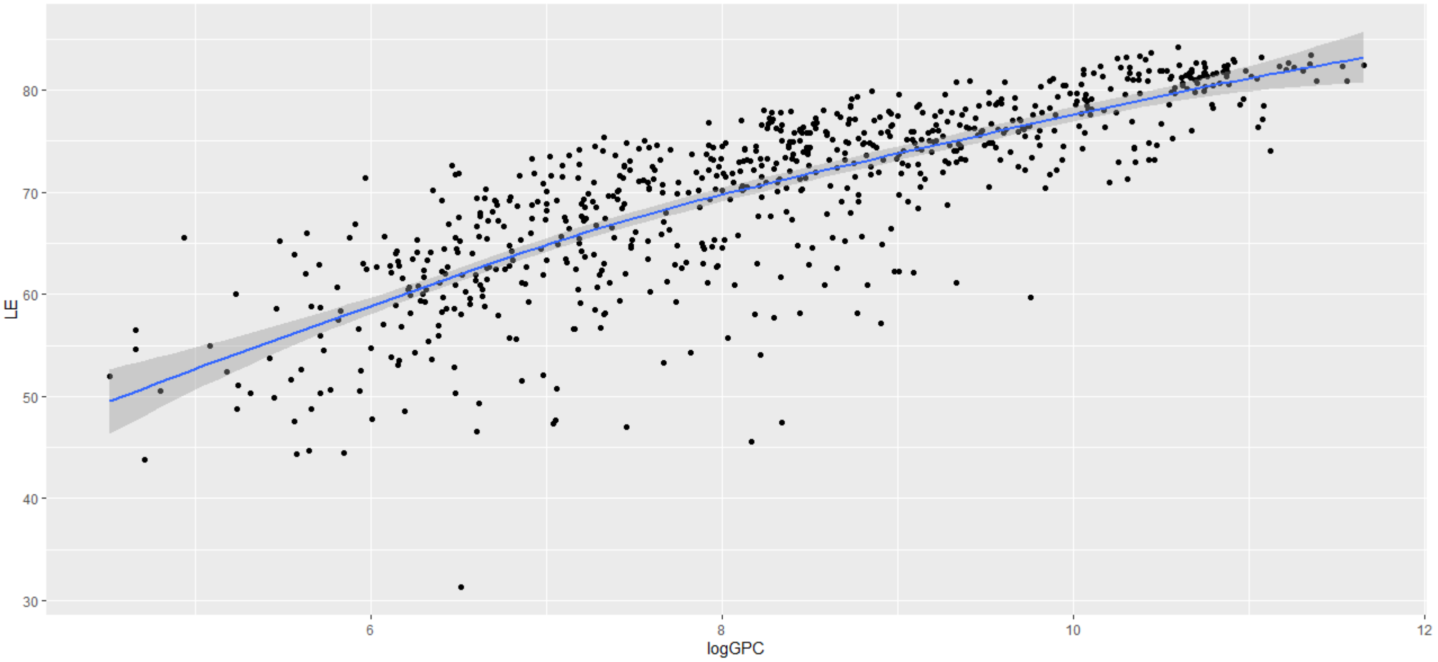

Figure 2: Scatterplot of life.expectancy vs log(GDP.per.capita)

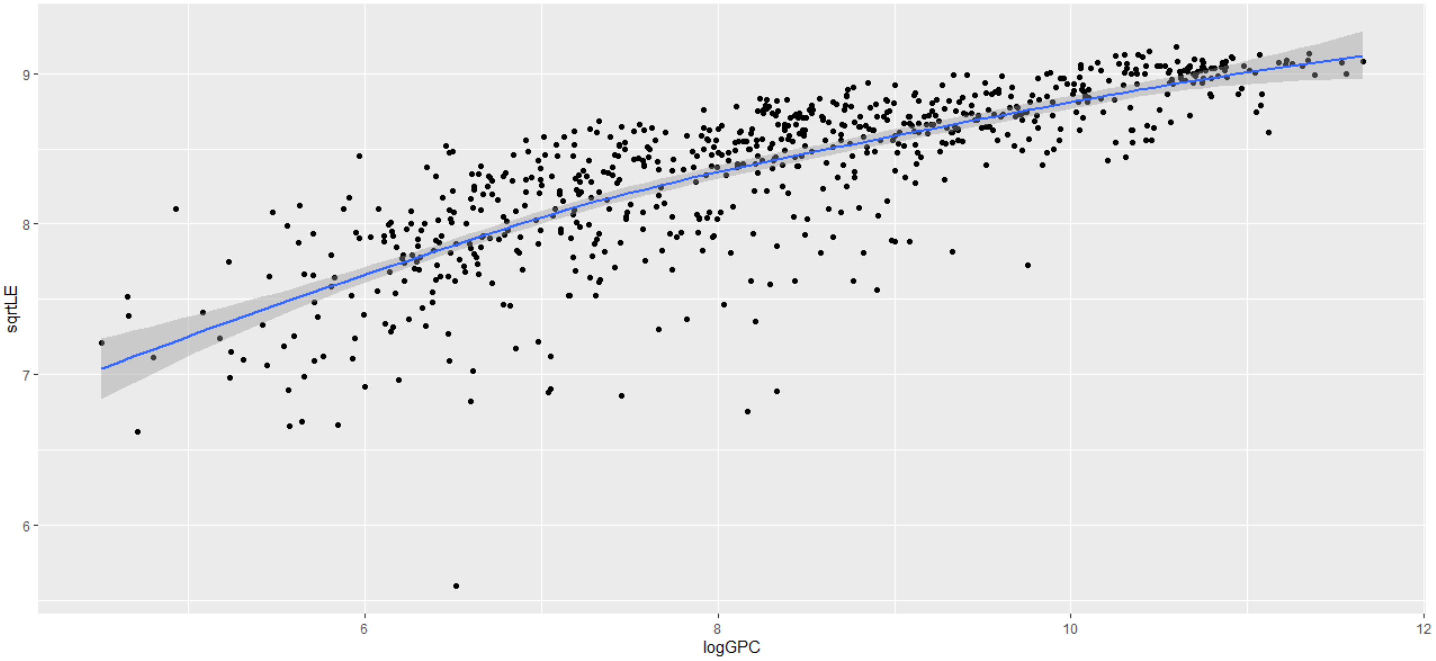

Figure 3: Scatterplot of sqrt(life.expectancy) vs log(GDP.per.capita)

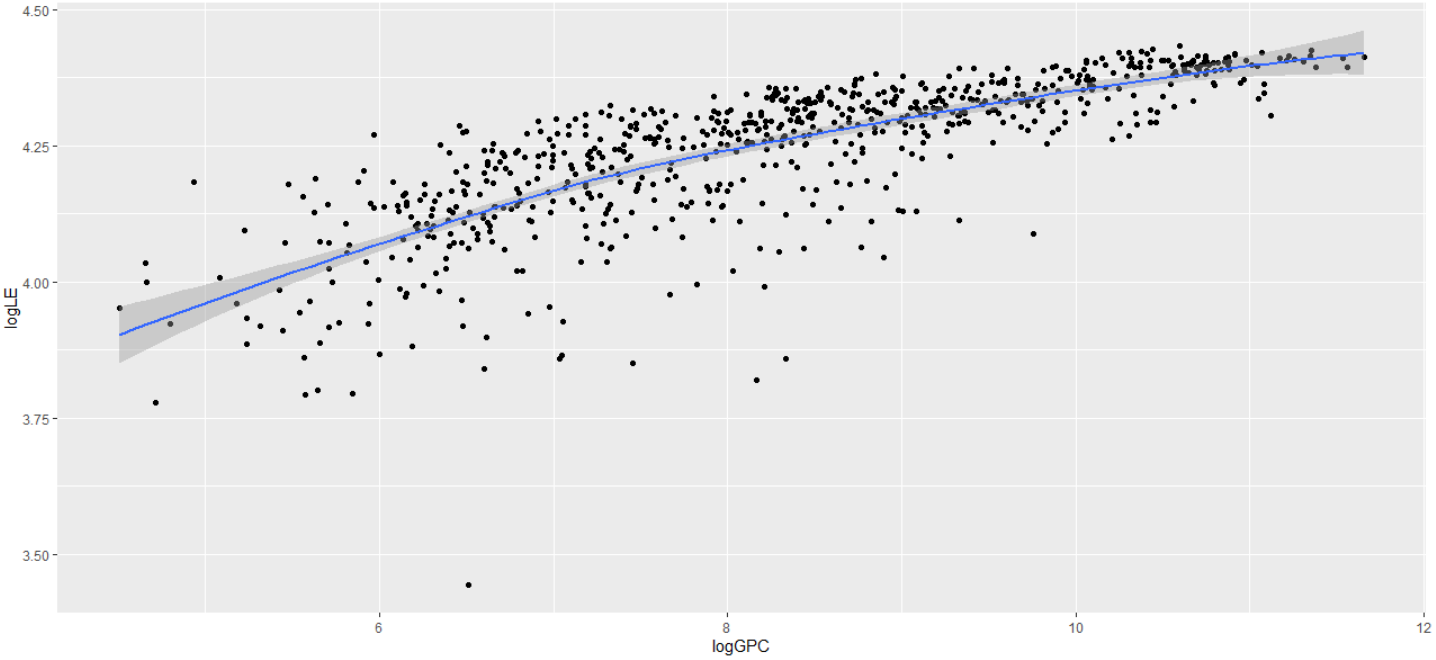

Figure 4: Scatterplot of log(life.expectancy) vs log(GDP.per.capita)

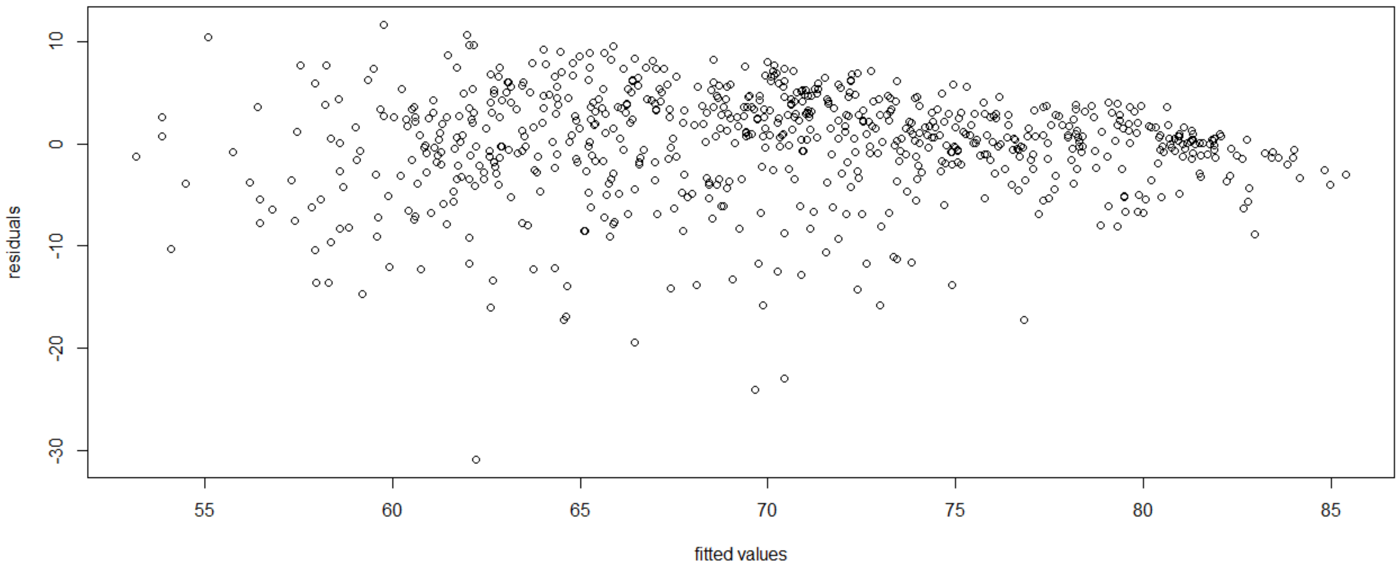

Figure 5: Scatterplot of residuals of life.expectancy vs log(GDP.per.capita)

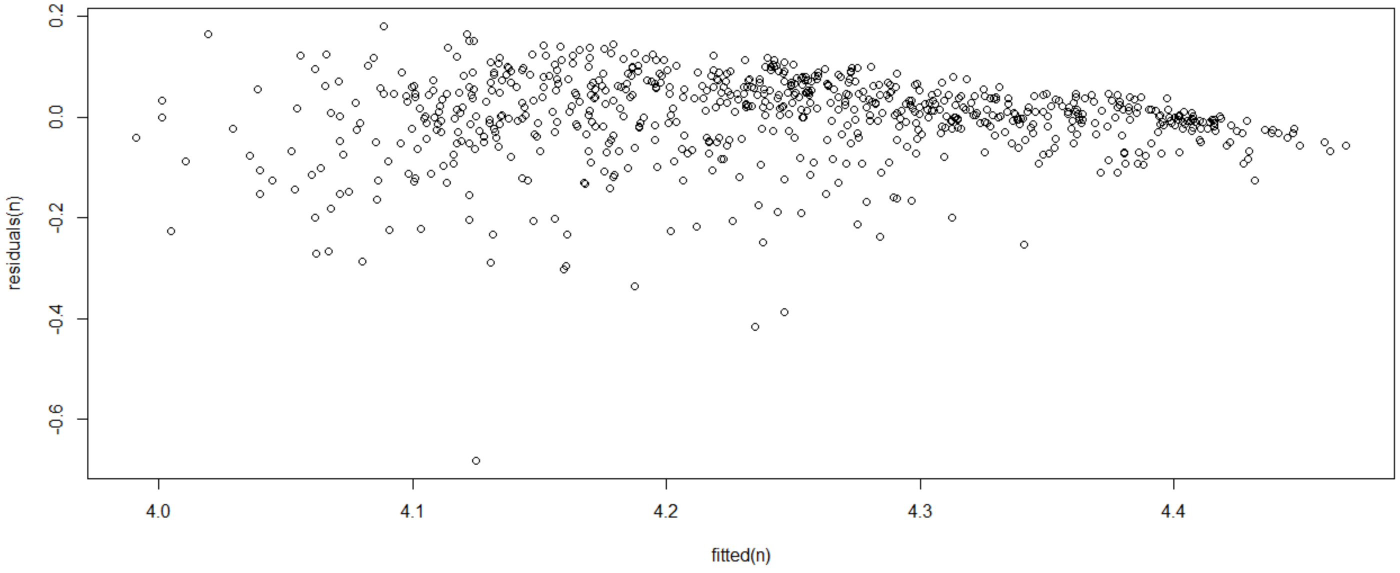

Figure 6: Scatterplot of residuals of sqrt(life.expectancy) vs log(GDP.per.capita)

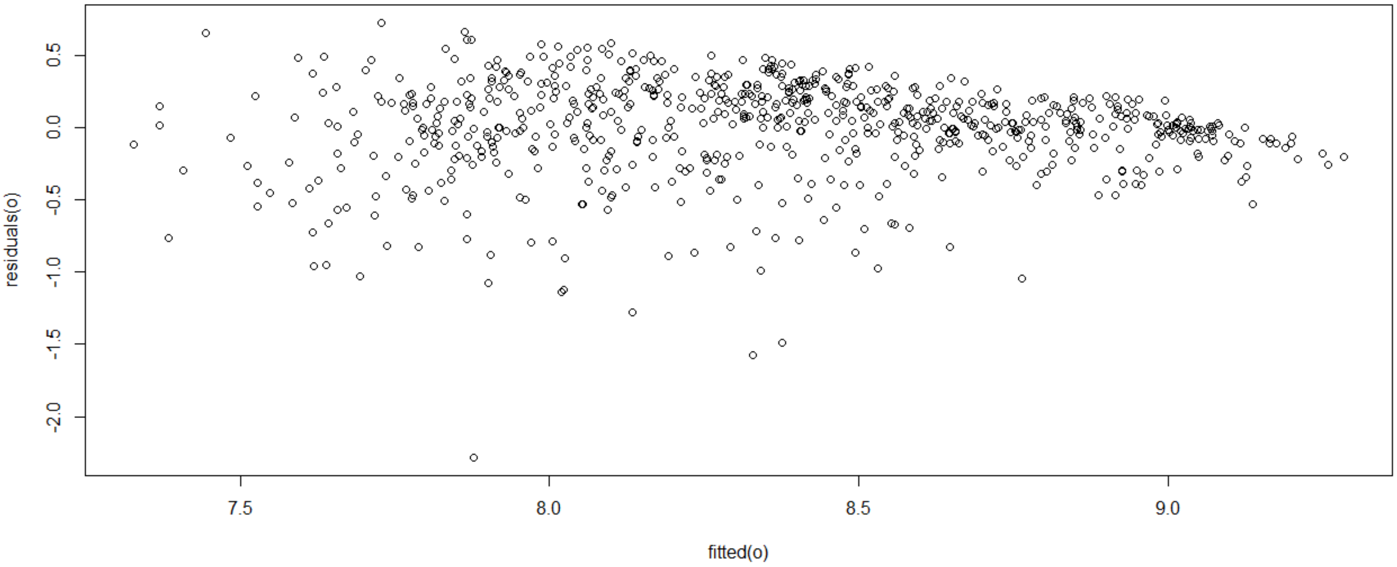

Figure 7: Scatterplot of log(life.expectancy) vs log(GDP.per.capita)

2 Task 2

There are lots of tradeoffs to make when choosing a model, the main things I looked for were: High accuracy, low complexity, and easy explainability.

The rough outline of the process I planned to go through was:

1) Identify which transformations are needed for the explanatory variables to make LE depend on them linearly (if there exist any), and choose some to include in the model to balance accuracy with simplicity.
2) Include some combinations of explanatory variables and their transformations (not all because there are lots of combinations and also the model should be fairly easy to interpret)
3) Check for any multicolinearity.
4) Check if there are any variance stabilising transformations for y to improve the model.
5) Check the model is adequate

2.1 1)

There are 12 columns in the dataframe d. One is the name of the country which we will ignore, one is LE, and one is GPC for which we have already found the log transformation.

Region and subregion will be very linked, so in my opinion we don't need to include both. There are far more subregions, so to not risk overfitting, to simplify things, and make the model more explainable I will ignore the subregion column. In fact, I will ignore the region column too. This is mainly an intuitive decision, since I think that so many of the other variables (GPC, HIV, population.density etc.) will depend on region that there will be lots of multicolinearity and the effect of factors that are related to the region but aren't in the data (climate, soil fertility etc.) will still be correlated to factors like GPC and the effect of including them will be minimal, and their effect would be hard to explain. It also makes the model less complex, since there are 6 regions and later we'll include 4 years, so I didn't want to have 24 different models.

All data is either from the year 2000, 2010, 2015 or 2019, so I will treat this as a factor rather than a continuous variable (even though it could be argued that it is since it's really just time) because I don't think having only 4 different inputs is enough to draw a good conclusion as a continuous variable. The mean LEs of data from 2000, 2010, 2015, 2019 are 66.81824, 70.00448, 71.50033, 72.54049 resp, so there does seem to be a definite positive correlation with year, which makes sense as healthcare gets better over time LE should go up.

I plotted a scatterplot of LE vs HS (fig. 8), and again it looked like a log curve. (Note there is a lot of missing data here, I decided to simply not include it. I could be introducing a bias here, for example if less developed countries tend to spend less on health and also tend to not have data available for how much they spent, so I would only be taking data from one end of the spectrum, however I don't want to make any assumptions like this off nothing other than speculation, and if I did try to deal with it the model would become much more complex so I decided to just omit it.) I then plotted a scatterplot of LE vs logHS (fig. 9) and it looked very interesting, there seemed to be not much correlation except for roughly logHS≥ 6 when all the life expectancies are very high. It also had 4 outliers (the 3 leftmost points and the bottommost point), I decided to include them since they are still valid data points. I plotted a graph of the residuals (fig. 10) and it seemed like if anything the variance was lower for the higher fitted values, so no variance stabilising transformation was necessary.

I plotted a scatterplot of LE vs HIV (fig. 11). Again there was missing data and for the same reason I decided to exclude it. This time it looked like an exponential curve (roughly LE $= \beta_0 + \beta_1 e^{-\mathrm{HIV}}$). So again I used the transformation HIV $\rightarrow$ log(HIV). Now the scatterplot (fig. 12) looks better, but still a bit like a curve as can be seen from the figure, maybe a cubic in log(HIV). Since this is complicated to improve via a transformation and such a transformation would complicate the model, and also because a plot of the residuals (fig. 13) has no major problems, I decided to stick with this simple transformation.

I then plotted a scatterplot of LE vs alcohol (fig. 14). There seemed to be very little correlation at all, and counterintuively a positive correlation if any (for which my instincts - and google - suggests that a higher GPC means people can afford more alcohol but also means a high LE). Because of this, I decided to leave it out completely, as I think the vast majority of the effect it has on LE is through being correlated with GPC, so including it wouldn't add much information anyway.

For tobacco the exact same thing is true (other than I couldn't find anything on google to corroborate my theory), so I also didn't include it. (The scatterplot is fig. 15.)

I plotted a scatterplot of LE vs PS (fig. 16), and it was hard to see a pattern because of the few countries with a massive population. To try make the relationship easier to interpret, I plotted a scatterplot of LE vs log(PS) (fig. 17). Another reason why this transformation is 'natural' to me is because it's well known that often population sizes of various things approximately follow a power law distribution, so by taking logs we get values which are more evenly distributed. This scatterplot seemed to have almost no correlation whatsoever, (the linear model has an R^2 value of only 0.0006887!) so I can't see how any transformation or anything else could help the model take into account PS in any meaningful way, so I didn't include it.

I plotted a scatterplot of LE vs PD (fig. 18), and again it was hard to see a pattern, so again I plotted a scatterplot of LE vs logPD (fig. 19). There was slightly more correlation but still very little, (an R^2 value of 0.05977), so again I feel it would just be overcomplicating things for no reason to include this factor.

In summary, the only response variables we will use in our model are logGPC, logHS, logHIV and year.

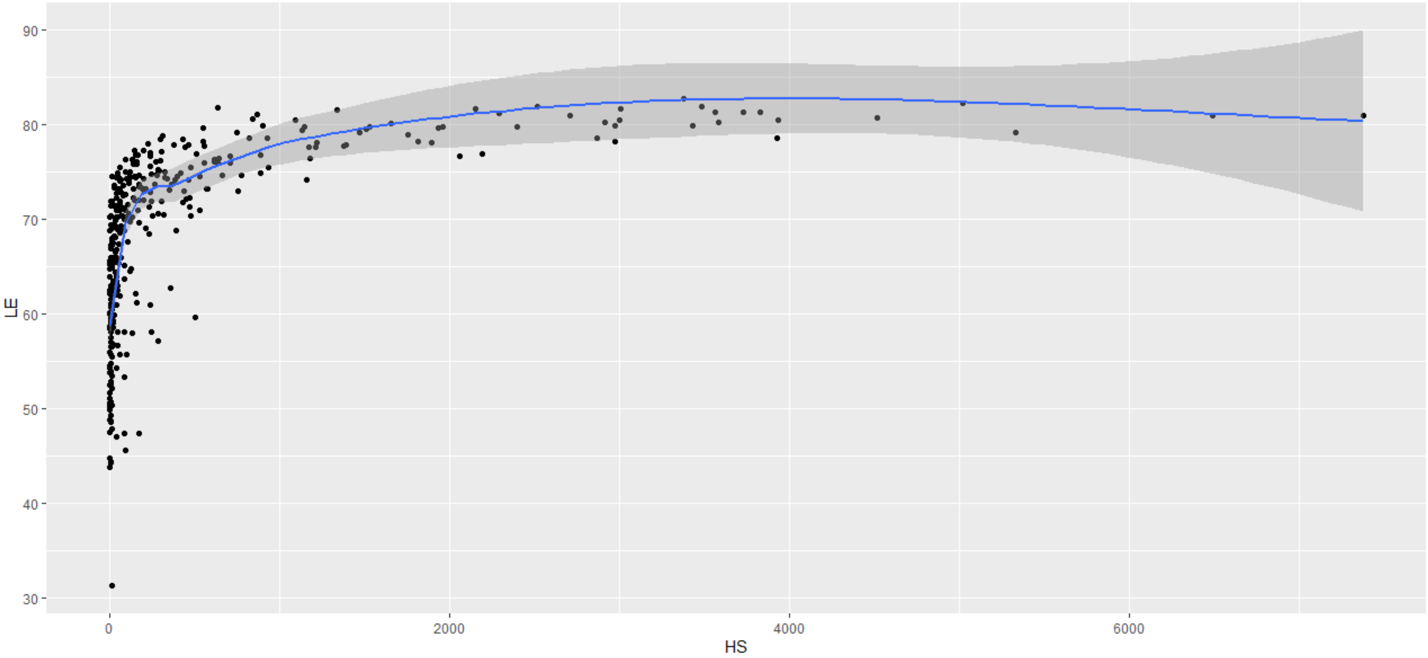

Figure 8: Scatterplot of LE vs HS

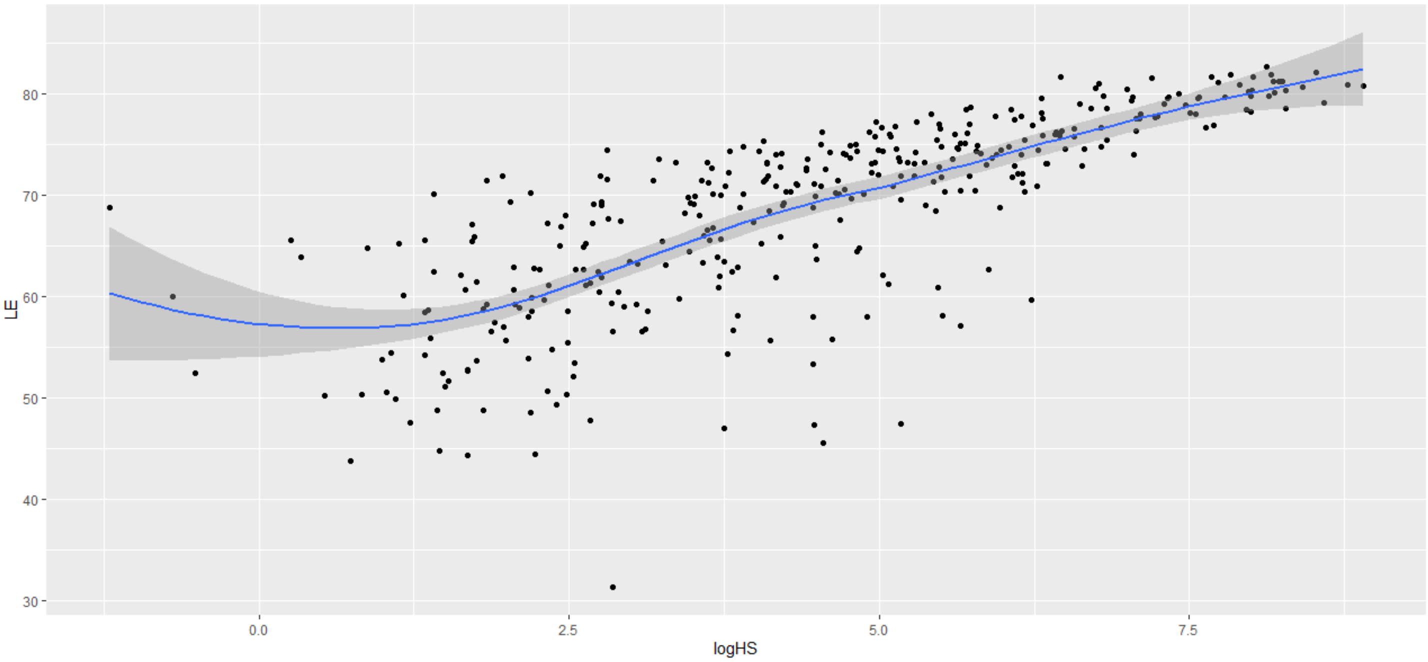

Figure 9: Scatterplot of LE vs logHS

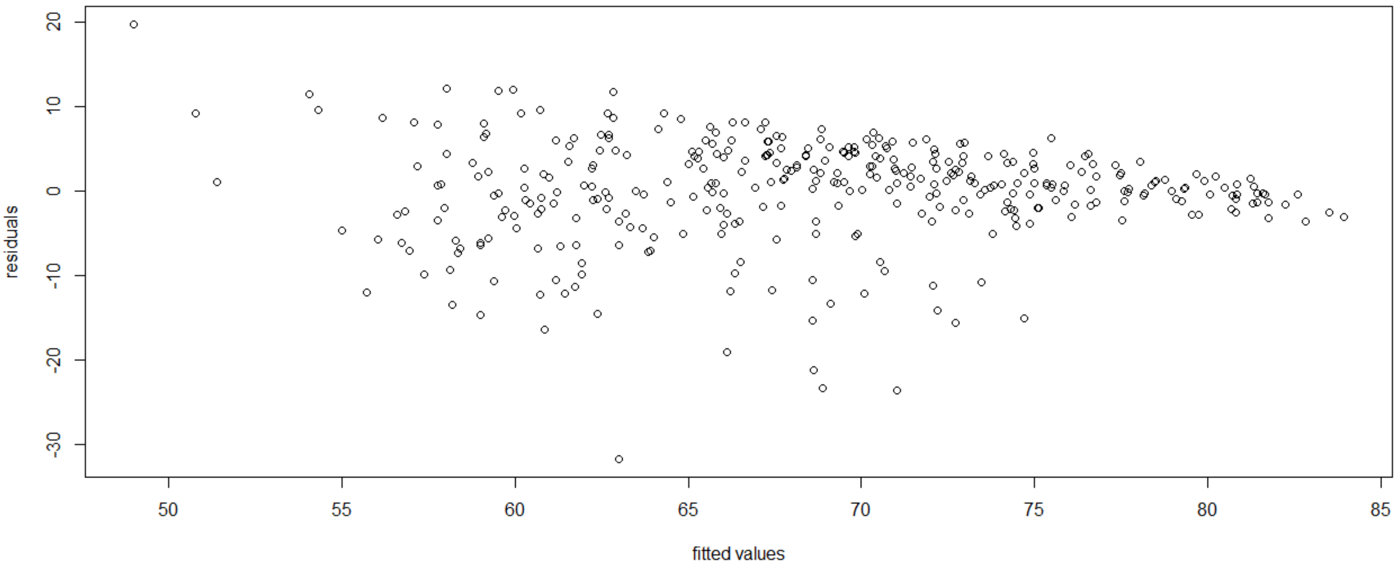

Figure 10: Scatterplot of residuals in the LE vs logHS model

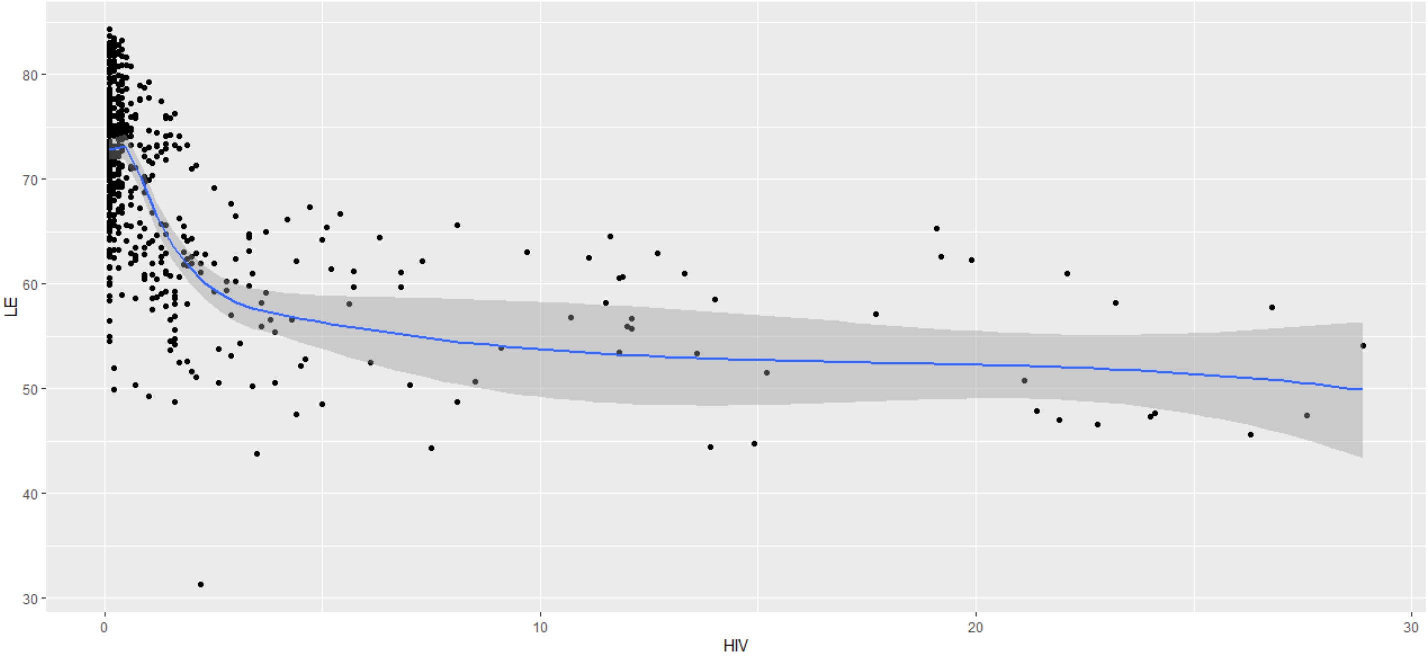

Figure 11: Scatterplot of LE vs HIV

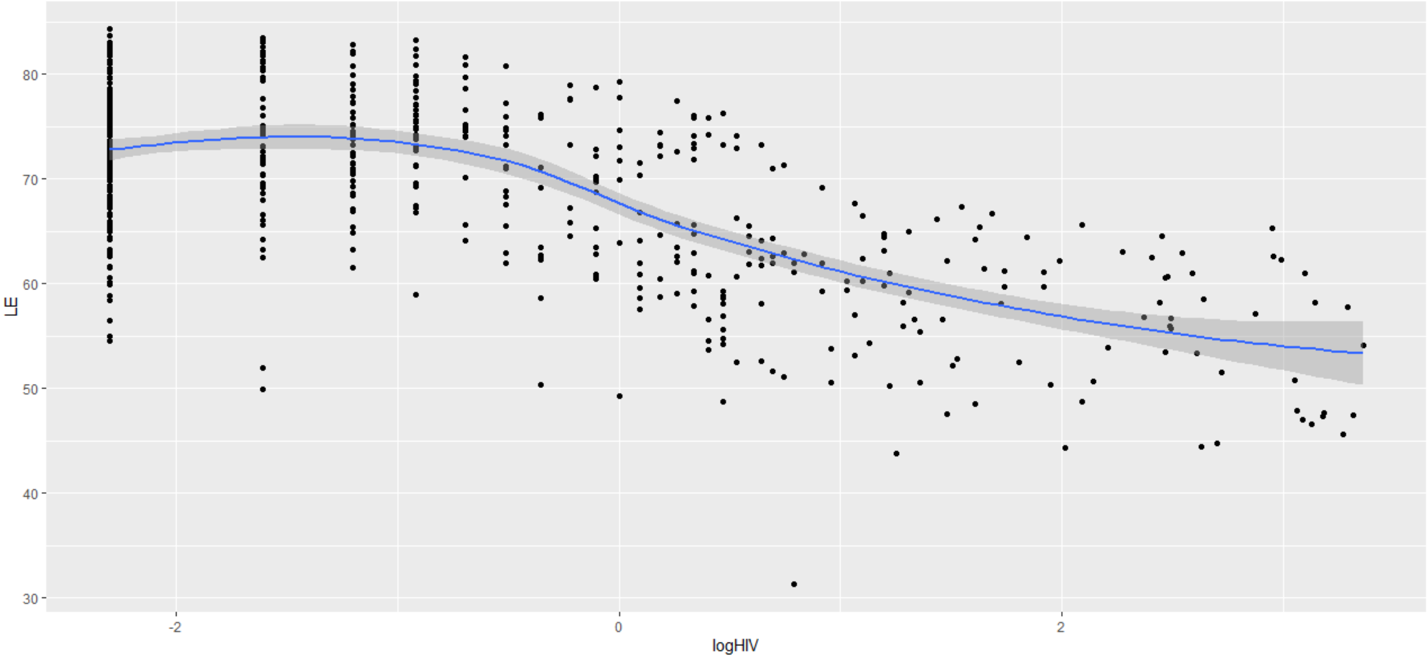

Figure 12: Scatterplot of LE vs logHIV

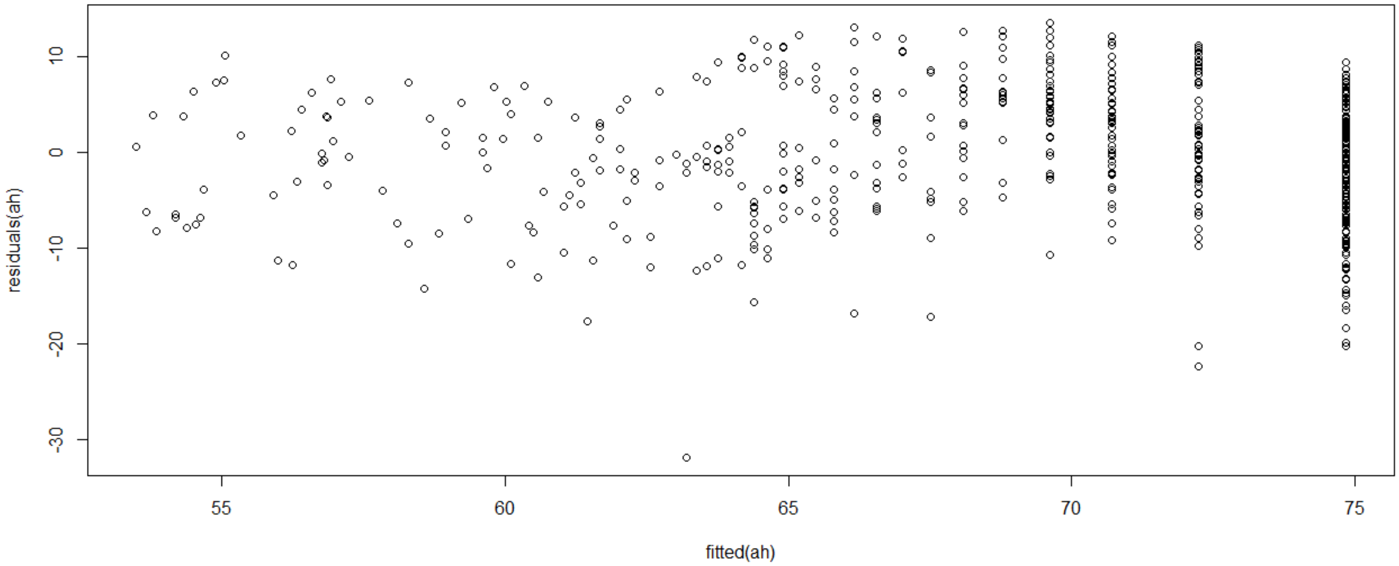

Figure 13: Scatterplot of the residuals in the LE vs logHIV model

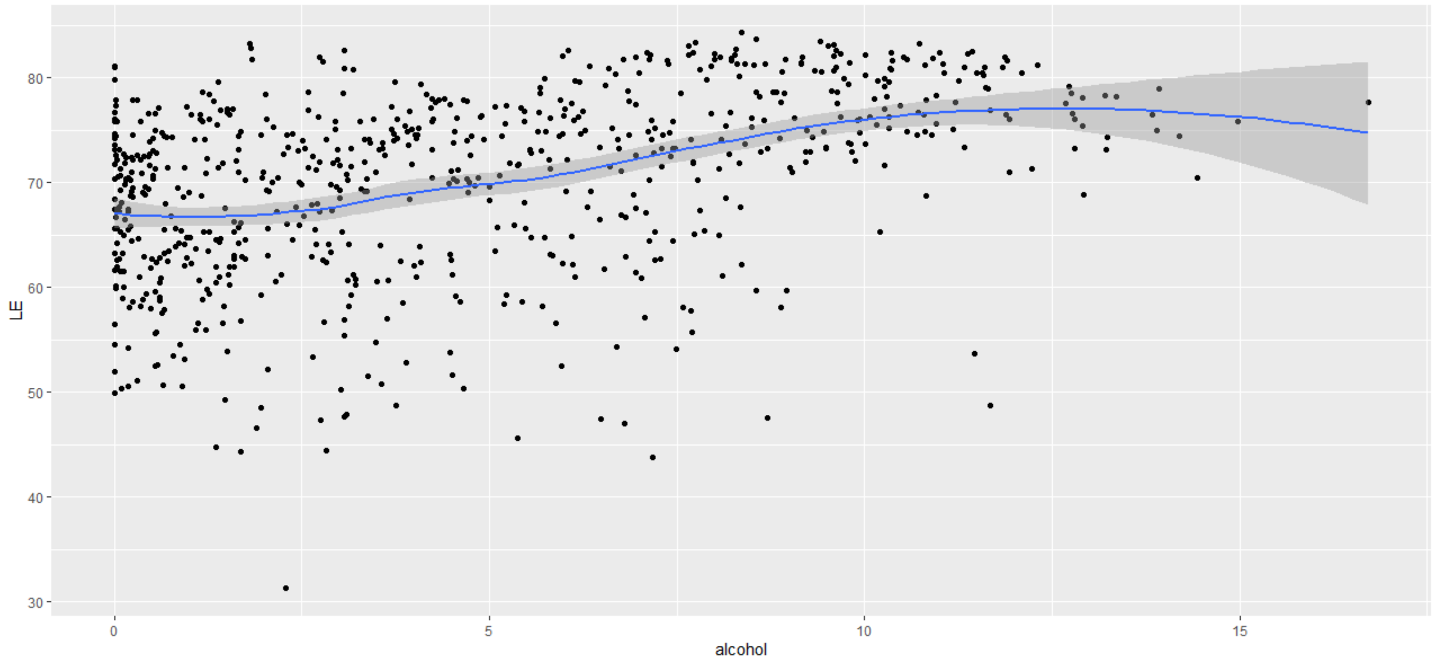

Figure 14: Scatterplot of LE vs alcohol

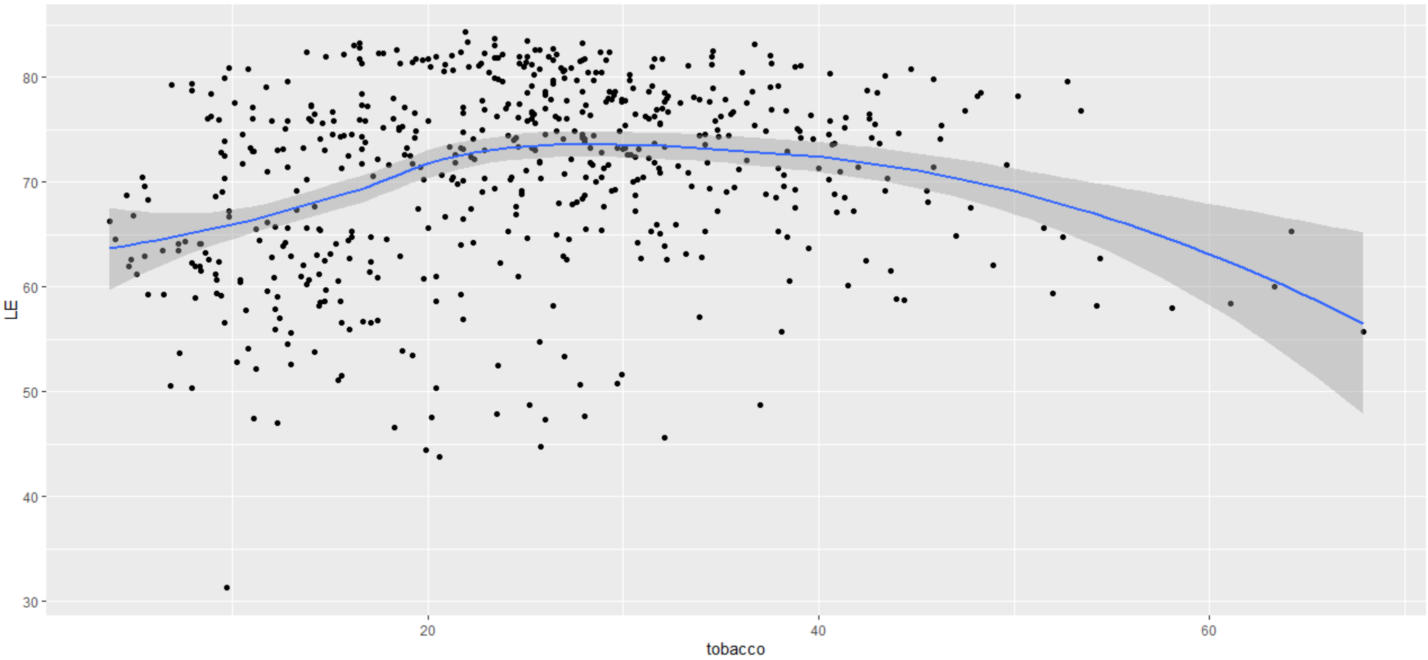

Figure 15: Scatterplot of LE vs tobacoo

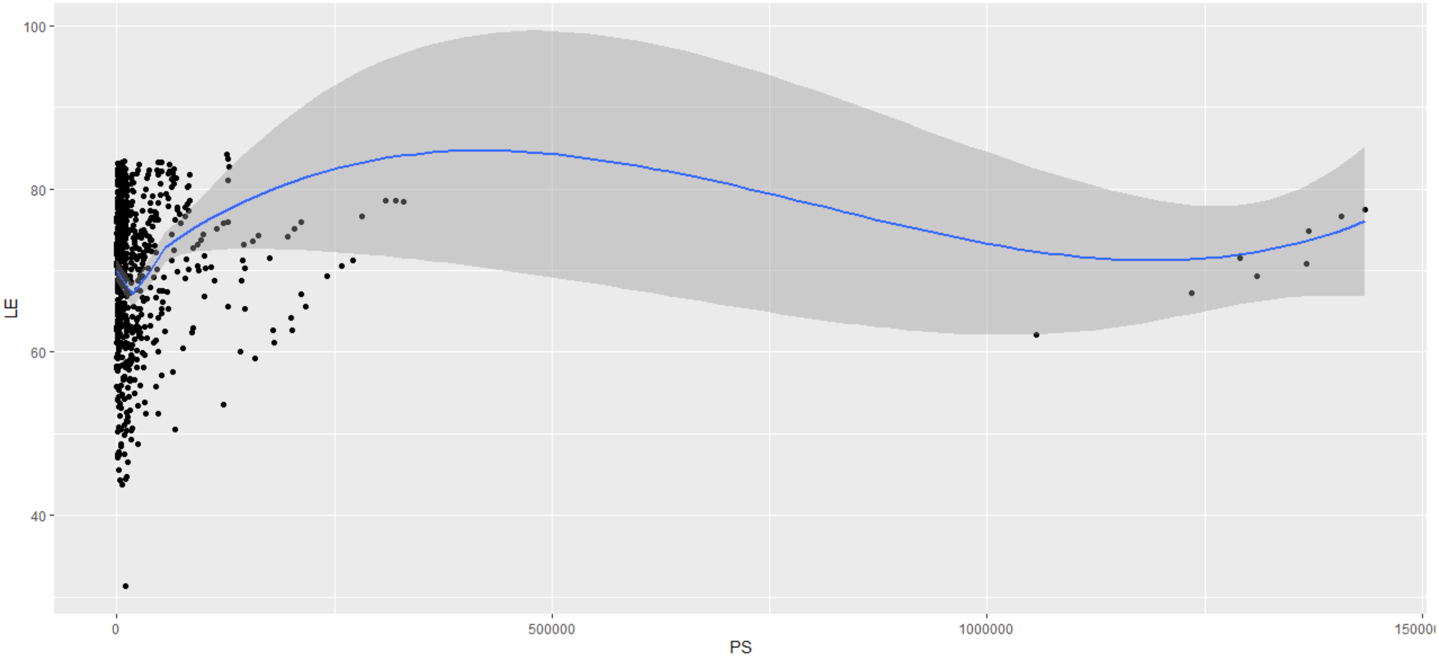

Figure 16: Scatterplot of LE vs PS

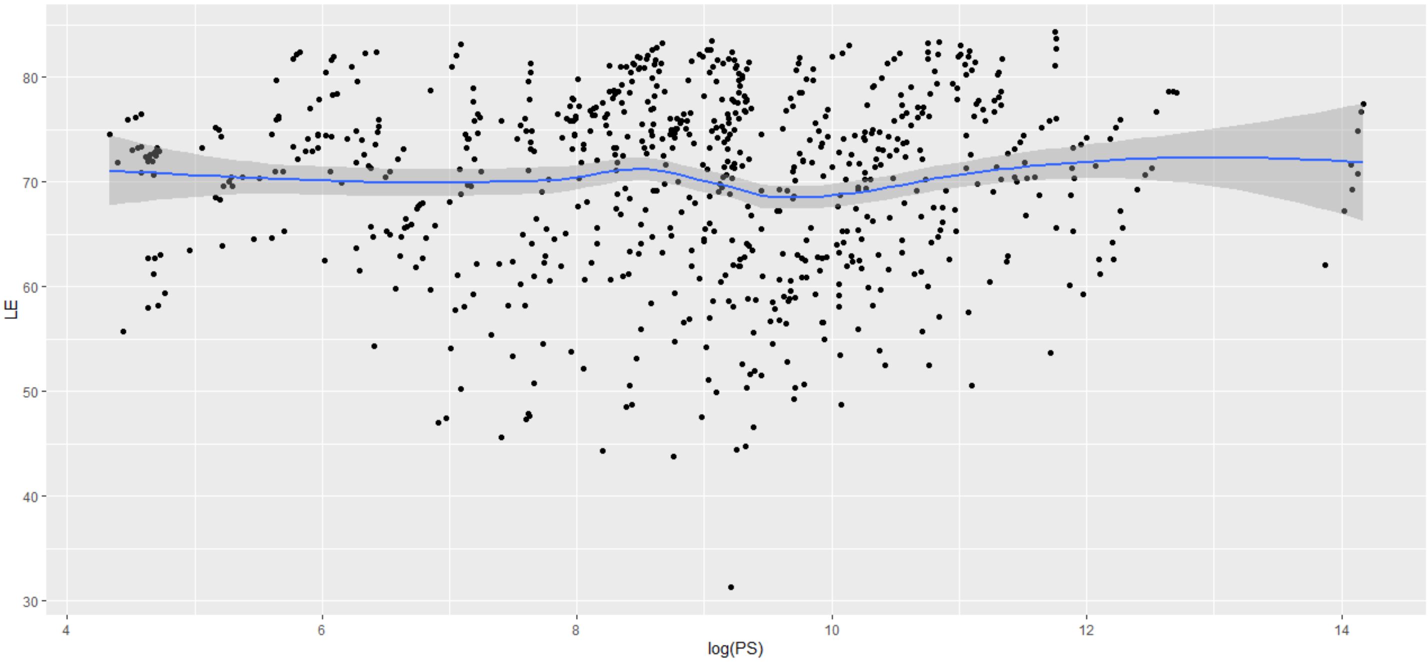

Figure 17: Scatterplot of LE vs logPS

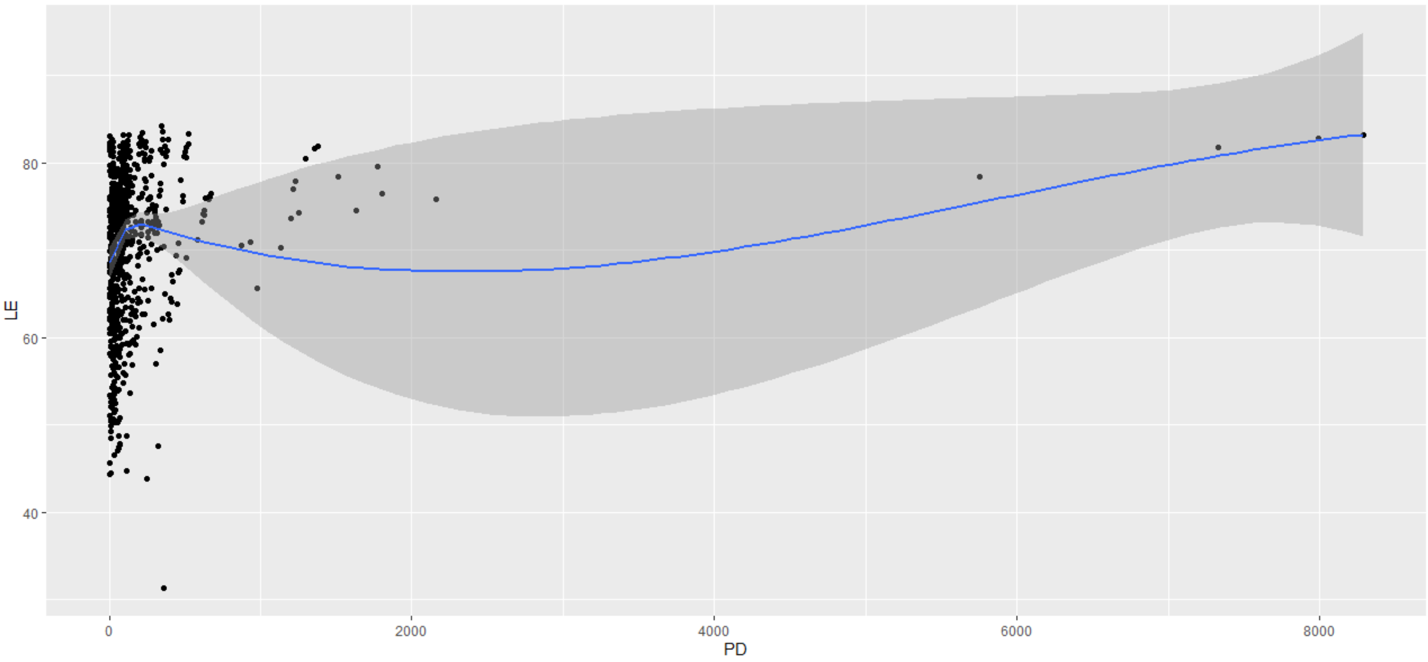

Figure 18: Scatterplot of LE vs PD

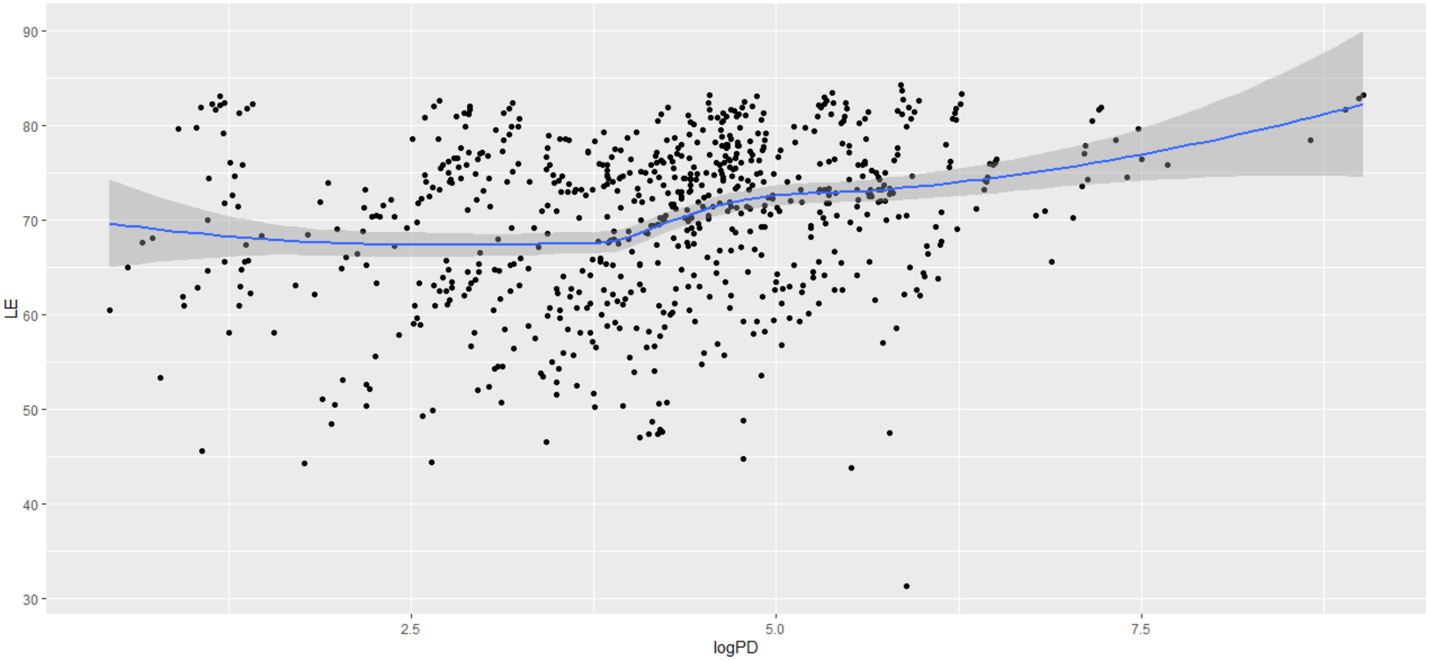

Figure 19: Scatterplot of LE vs logPD

2.2 2)

Now I would use the 'lm' command to make a linear model using these 4 variables, unfortunately there's a problem. There is a lot of missing data - out of 731 total rows, 371 are missing logHS and 194 are missing logHIV, luckily there is no clear pattern I could find to which rows are missing logHIV, however the only data available for HS is from 2000 and 2010, so this suggests that the data isn't missing at random, rather in a systematic way. If I simply deleted the missing data, I would only have 264 rows left. There are three main ways to tackle this problem that I can see:
1) Delete the rows with missing values - If I did this I would lose a significant amount of information.
2) Remove one or more of the explanatory variables. If I removed one of the variables, I would have more data points, but I would also have less input to go on, meaning my model might be worse.
3) Fill in the values with some sort of average. If I did this, I would keep all the information, however this doesn't account for the correlation between the explanatory variables and it will affect the amount of variance in the variables.
4) Run a regression on the explanatory variables and estimate the missing variables from the others. This would likely introduce bias and multicolinearity.

None of these are perfect, but since I have a lot of data, and a rule of thumb sometimes used is that 10 data points per independent variable are needed for a linear regression to be useful, 264 should still be plenty, and every other method would introduce bias or information in different ways, so I will go with a version of 1). Since there are only 5 missing values of logHS in the years 2000 and 2010, I will delete these rows, then set the NA values to 0 for 2015 and 2019 since the year is a factor, this will give a coefficient error for those 2 years since any coefficient works, so I will just interpret them as 0 which is the best we can do anyway. I will also delete the rows which have an NA in logHIV. I now still have 533 data points, which I think is a good compromise to deal with this problem.

I now won't use '(lm(LE . , data=f))' since I need separate coefficients for logHS for different years, so I did (lm(LE .$\hat{2}$, data=f)). The adjusted R^2 value is now 0.801, already much better than any previous model! Here are the full results:

```
Residuals:
     Min       1Q   Median       3Q      Max
-27.4112  -2.2942   0.1852   2.5730  10.1506

Coefficients: (2 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     33.73012    5.22322   6.458 2.46e-10 ***
logGPC           3.26719    1.01082   3.232  0.00131 **
logHS            4.26846    0.92976   4.591 5.55e-06 ***
logHIV          -3.73379    0.72658  -5.139 3.93e-07 ***
year2010         4.02901    5.75559   0.700  0.48423
year2015         4.67221    5.64919   0.827  0.40859
year2019         7.28458    5.66484   1.286  0.19904
logGPC:logHS    -0.34810    0.10553  -3.299  0.00104 **
logGPC:logHIV    0.06668    0.12684   0.526  0.59936
logGPC:year2010 -1.11322    1.20641  -0.923  0.35656
logGPC:year2015  0.43483    1.04086   0.418  0.67629
logGPC:year2019  0.22884    1.04132   0.220  0.82614
logHS:logHIV    -0.01166    0.15310  -0.076  0.93933
logHS:year2010   1.23547    0.96959   1.274  0.20316
logHS:year2015        NA         NA      NA       NA
logHS:year2019        NA         NA      NA       NA
logHIV:year2010  0.38459    0.37027   1.039  0.29945
logHIV:year2015  1.04749    0.64602   1.621  0.10553
logHIV:year2019  1.18286    0.65446   1.807  0.07129 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.977 on 516 degrees of freedom
Multiple R-squared:  0.807,     Adjusted R-squared:  0.801
F-statistic: 134.8 on 16 and 516 DF,  p-value: < 2.2e-16
```

This model is much more complex and harder to explain, so I decided to remove a lot of terms. I decided to just stick with logGPC, logHIV, logHS:year, year, because this model had an R^2 value of 0.7857 and the coefficients are easily explainable. I almost opted for logGPC:year, logHIV:year, logHS:year, year to have more accuracy, however this made too many coefficients, compromising explainability and simplicity, and only improved the adjusted R^2 value by 0.0097, which I thought at that point meant I was just overfitting. I could have used an algorithm to look for the best model with ≤ 7 explanatory variables, however it can't have done better than an R^2 value of 0.801 which isn't much better anyway, and having a weird choice of 7 variables would be much less natural and explainable, and would also probably just be overfitting.

2.3 3)

Unfortunately, we can't use the simple kappa function to check for colinearity, since the values for logHS in 2015 and 2019 are 0, meaning they are perfectly colinear. I plotted graphs of the scatterplots against each other: (fig. 20)

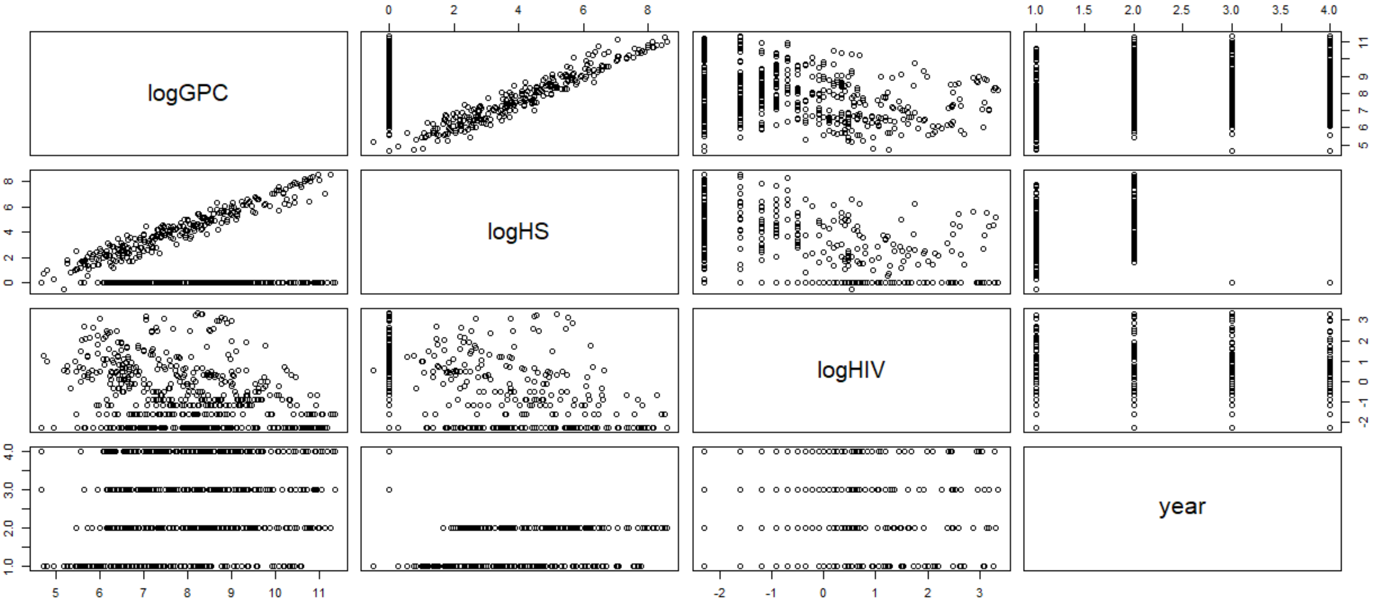

Figure 20: Scatterplot of the pairs of explanatory variables

Unfortunately, as makes sense, logGPC is highly correlated with logHS. I don't think logHS contains all the information that logGPC does, since people in richer countries can probably afford healthier food and generally have more time to exercise, etc. My hypothesis was shown to be correct, as when the regression is run with all variables other than logGPC, the adjusted R^2 is only 0.6547. However to my surprise, when I ran the regression lm(LE logGPC+logHIV+year) the adjusted R^2 value was 0.7846! Barely less than when we included logHS. So the clear choice was to remove logHS as a variable, this meant we could add in a few more data points again. If this hadn't been the case, and both inputs were important, then it seems believable to me that LE depends on GPC and the the proportion of GPC spent on healthcare, independently, so the I would've tried the transformation logHS → logHS-logGPC (=log(HS/GPC)). Now there are no obvious correlations in the graphs of the pairs (fig. 21) that I can see (other than maybe a slight negative correlation between logGPC and logHIV:

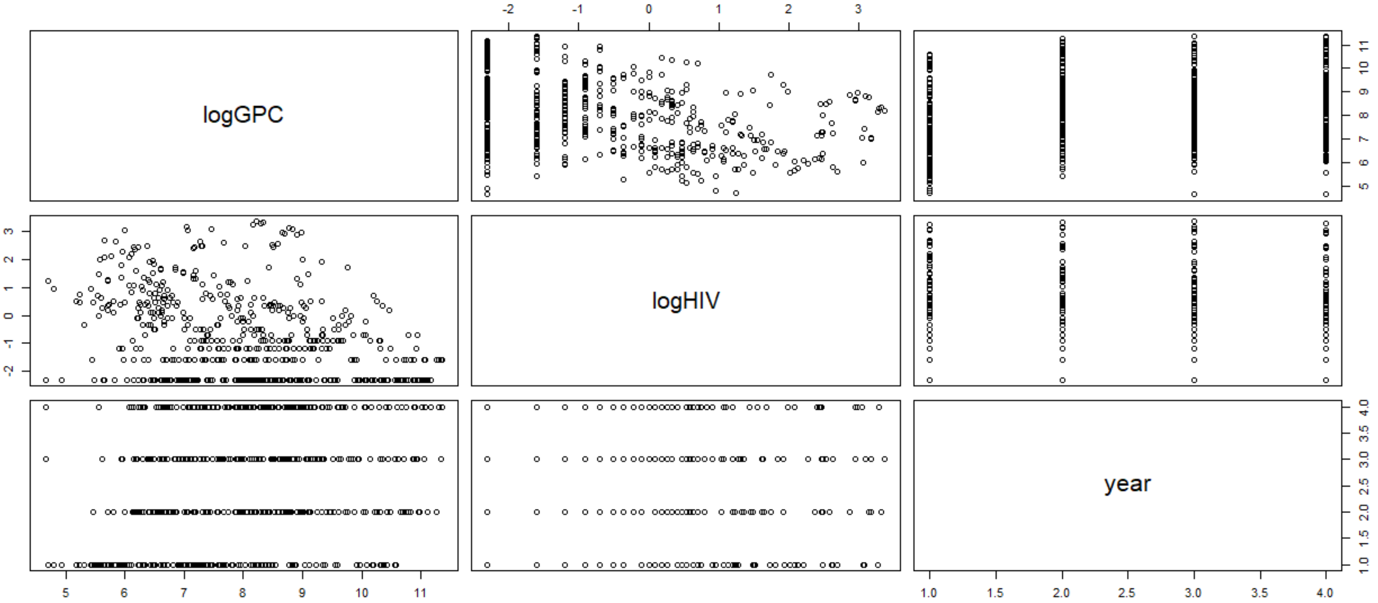

Figure 21: Scatterplot of the pairs of new explanatory variables

2.4 4)

Now I plotted the residuals of the model against the fitted values (fig. 22)

Actually it seems like the variance decreases as LE increases. This seems to be counterintuitive, so I didn't want to use a transformation unless there was more strong evidence. I plotted a Q-Q plot of the residuals (fig. 23) and they (roughly) formed a straight line, so I was satisfied, and decided against using any transformation.

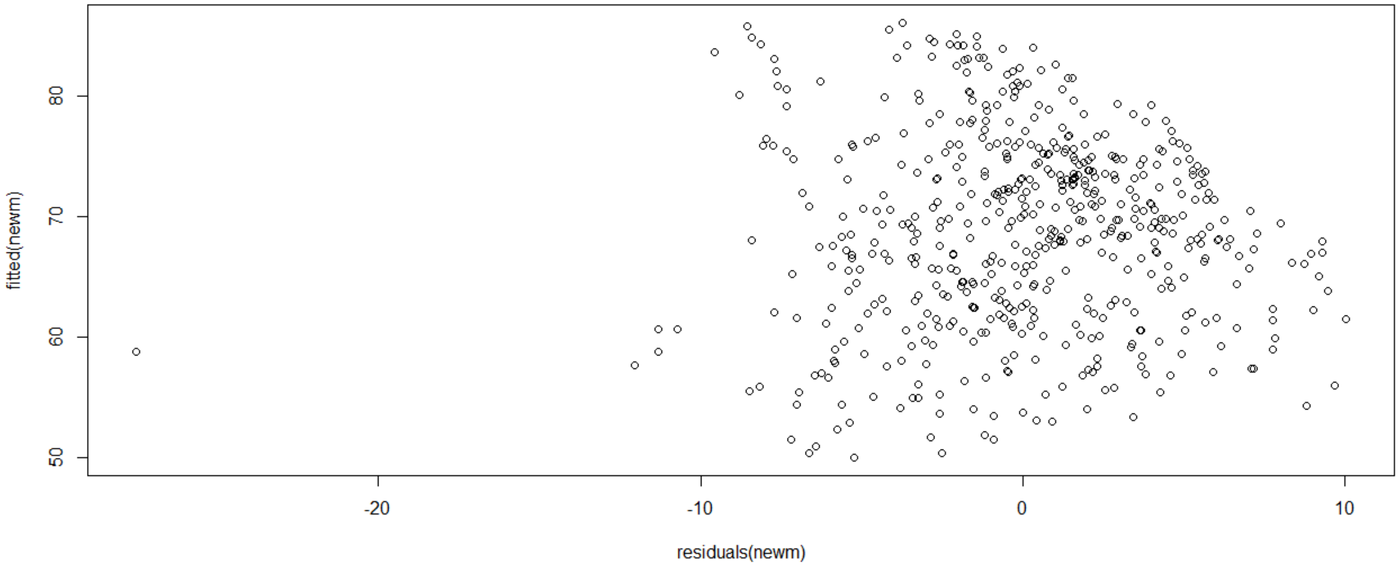

Figure 22: Residuals vs fitted values in the new model

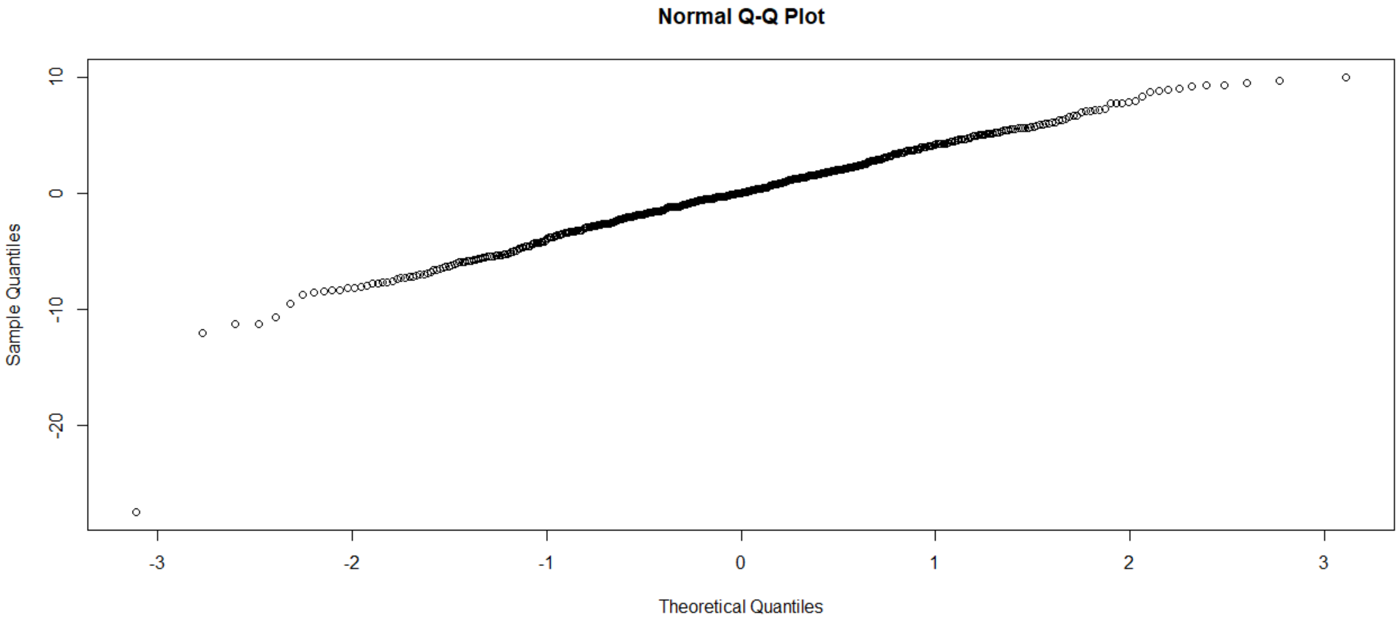

Figure 23: QQ plot of residuals in new model

2.5 5)

There are many ways to check a model is adequate, we've already seen the adjusted R^2 value is 0.7846. Of course this is subjective as to whether it's 'high', but I think considering the simplicity of the model and the number of complex factors that truly go into LE it's good enough. The Q-Q plot of residuals looks roughly straight and from the scatterplots there isn't too much colinearity, so I'm satisfied. In this model, we have:

```
Residuals:
     Min       1Q   Median       3Q      Max 
-27.4995  -2.5815   0.0351   2.7795  10.0250 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.5907     1.0131  36.117  < 2e-16 ***
logGPC        3.6310     0.1337  27.158  < 2e-16 ***
logHIV       -2.6335     0.1249 -21.092  < 2e-16 ***
year2010      0.6148     0.5247   1.172 0.241801    
year2015      2.0306     0.5274   3.851 0.000132 ***
year2019      2.8254     0.5296   5.335 1.42e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.183 on 531 degrees of freedom
  (194 observations deleted due to missingness)
Multiple R-squared:  0.7866,    Adjusted R-squared:  0.7846 
F-statistic: 391.4 on 5 and 531 DF,  p-value: < 2.2e-16
```

i.e.

$$\text{LE} = \begin{cases} 36.5907 + 3.631\log(\text{GPC}) - 2.6335\log(\text{HIV}) \text{ if the data is from 2000} \\ 37.2055 + 3.631\log(\text{GPC}) - 2.6335\log(\text{HIV}) \text{ if the data is from 2010} \\ 38.6213 + 3.631\log(\text{GPC}) - 2.6335\log(\text{HIV}) \text{ if the data is from 2015} \\ 39.4161 + 3.631\log(\text{GPC}) - 2.6335\log(\text{HIV}) \text{ if the data is from 2019} \end{cases}$$

The interpretation of the constants is that 3.631 is the amount that the life expectancy goes up by when the log of the GDP per capita of a country goes up by 1, and 2.6335 is how much the life expectancy goes down by when the log of the prevalence of HIV in adults aged 15 to 49 goes up by 1, and the four constants 36.5907, 37.2055, 38.6213, 39.4161 are the respective shifts to the life expectancy needed, i.e the life expectancy for a country with a GDP per capita of 1 USD and a HIV prevalence of 1% for the years 2000, 2010, 2015, 2019 respectively.

However, after removing logHS I tried the linear model lm(LE logGPC:year + logHIV:year+year, data=d). The coefficients of logGPC in the years 2000, 2010, 2015, 2019 are 3.8323, 3.6152, 3.6380, 3.4242 so there isn't a clear trend and it feels like I'm just overfitting to the specific data, and I couldn't explain them otherwise. However, the coefficients of logHIV in the 4 years are: -3.3415, -2.7662, -2.1604, -2.0210, which is a clear upward trend, and could be explained by the fact that HIV treatment has got better over time since 2000. Because of this, I tried the model lm(LE logGPC + logHIV:year+year, data=d) and the adjusted R^2 value was 0.7932, I thought this was a significant enough increase to justify adding 3 more coefficients, so I decided on this as my final model.

It's very similar to the previous model (lm(LE logGPC+logHIV+year,data=d)), it for the same reasons I think this model has good accuracy, they are very similar for model selection criteria and it sacrifices some simplicity for even better accuracy, however it is still very easily interpreted, so I decided this was the better one. The final model now has:

```
Residuals:
     Min       1Q   Median       3Q      Max 
-27.3153  -2.2644  -0.0434   2.7134  10.5229 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      36.0259     1.0016  35.967  < 2e-16 ***
logGPC            3.6309     0.1310  27.715  < 2e-16 ***
year2010          1.0962     0.5572   1.967   0.0497 *  
year2015          2.8992     0.5609   5.168 3.35e-07 ***
year2019          3.8588     0.5677   6.798 2.88e-11 ***
logHIV:year2000  -3.4051     0.2166 -15.722  < 2e-16 ***
```

```
logHIV:year2010  -2.7610     0.2316 -11.921  < 2e-16 ***
logHIV:year2015  -2.1626     0.2408  -8.980  < 2e-16 ***
logHIV:year2019  -1.9544     0.2501  -7.815 2.99e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.099 on 528 degrees of freedom
  (194 observations deleted due to missingness)
Multiple R-squared:  0.7962,	Adjusted R-squared:  0.7932 
F-statistic: 257.9 on 8 and 528 DF,  p-value: < 2.2e-16
```

and it is:

$$\text{LE} = \begin{cases} 36.0259 + 3.6309\log(\text{GPC}) - 3.4051\log(\text{HIV}) \text{ if the data is from 2000} \\ 37.1221 + 3.6309\log(\text{GPC}) - 2.761\log(\text{HIV}) \text{ if the data is from 2010} \\ 38.9251 + 3.6309\log(\text{GPC}) - 2.1626\log(\text{HIV}) \text{ if the data is from 2015} \\ 39.8847 + 3.6309\log(\text{GPC}) - 1.9544\log(\text{HIV}) \text{ if the data is from 2019} \end{cases}$$

The interpretation of the constants is that 3.6309 is the amount that the life expectancy goes up by when the log of the GDP per capita of a country goes up by 1, and in 2000, 2010, 2015, 2019 the life expectancy goes down by 3.4051, 2.761, 2.1626, 1.9544 years resp. as the log of the prevalence of HIV in adults aged 15 to 49 goes up by 1 (decreasing over time as healthcare improves), and the four constants 36.0259, 37.1221, 38.9251, 39.8847 are the respective shifts to the life expectancy needed (increasing over time as healthcare improves), i.e the life expectancy for a country with a GDP per capita of 1 USD and a HIV prevalence of 1% for the years 2000, 2010, 2015, 2019 respectively.