Analysis of Suicide Rates in England

Zoe Ganess

1 A. Questions

1.1 Introduction

For this analysis, data from 326 local authorities in England with information about suicides, personal well-being, unemployment, income, deprivation and population was considered. This report aims to determine the relationship between these explanatory factors and suicides. The importance of this report is therefore reflected in the fact that studying these relationships will lead to a better understanding of what exactly influences an individual to undergo suicidal thoughts. With better knowledge of this, we can address those influential factors in an attempt to mitigate the occurrence of suicidal incidences in the general population.

NB: where (*) appears, that plot/table/graph etc. was not included due to the page limit but is not central to the main analysis…refer to rmarkdown file for original output

1.2 Q1. To what extent are personal well-being, unemployment, income and deprivation associated with the number of suicides in England?

The overall trend of each factor is shown in figure 1. It can be observed that living score (deprivation) and unemployment rate have a positive relationship with suicide rate whereas median income and worthwhile (wellbeing) have a negative relationship with same. More specifically, increases in deprivation and unemployment and decreases in income and personal wellbeing result in more persons committing suicide. We also recognise the relationships each of these factors have with each other and how they affect suicide rates jointly. For example, deprivation intuitively should be related to high unemployment rates and specifically low income and the combination of these factors, in turn would have a negative effect on ones wellbeing, ultimately leaving room for suicidal thoughts.

1.3 Q2. What are the regional differences in the number of suicides in England?

Looking at the table above, we can infer that Yorkshire and The Humber, North East and East of England have the three highest suicide rates whereas London, South East and East Midlands have the three lowest. The average suicide rate among those living in the region of Yorkshire and The Humber was approximately 1.5 more than those living in London. Interestingly, the three regions with the highest suicide rates are situated in the North of the UK whereas the two lowest suicide rates originate from the South. We note that the difference is not particularly significant (as the pink intervals in the graphs below interlap) but in order to quantify the further quantify the effects we are investigating, we take the instance of high and low unemployment rates as a reference. If all regions were to have the same unemployment rate, the average suicide rate would be fairly similar across each region, as seen in the table (refer to figure 3). However, under a low unemployment rate of 2.0, we observe subtle changes in the suicide rates (most increases), but the relative shape of the graph is unchanged. A higher unemployment rate of 7.0 sees substantial changes with both the average suicide rate and relative shape of the graph. One interesting thing to note is that the regions who initially had the highest suicide rates are now among the lowest, inferring that other factors such as deprivation, especially in the case of regional concerns, also affect the average suicides rates.

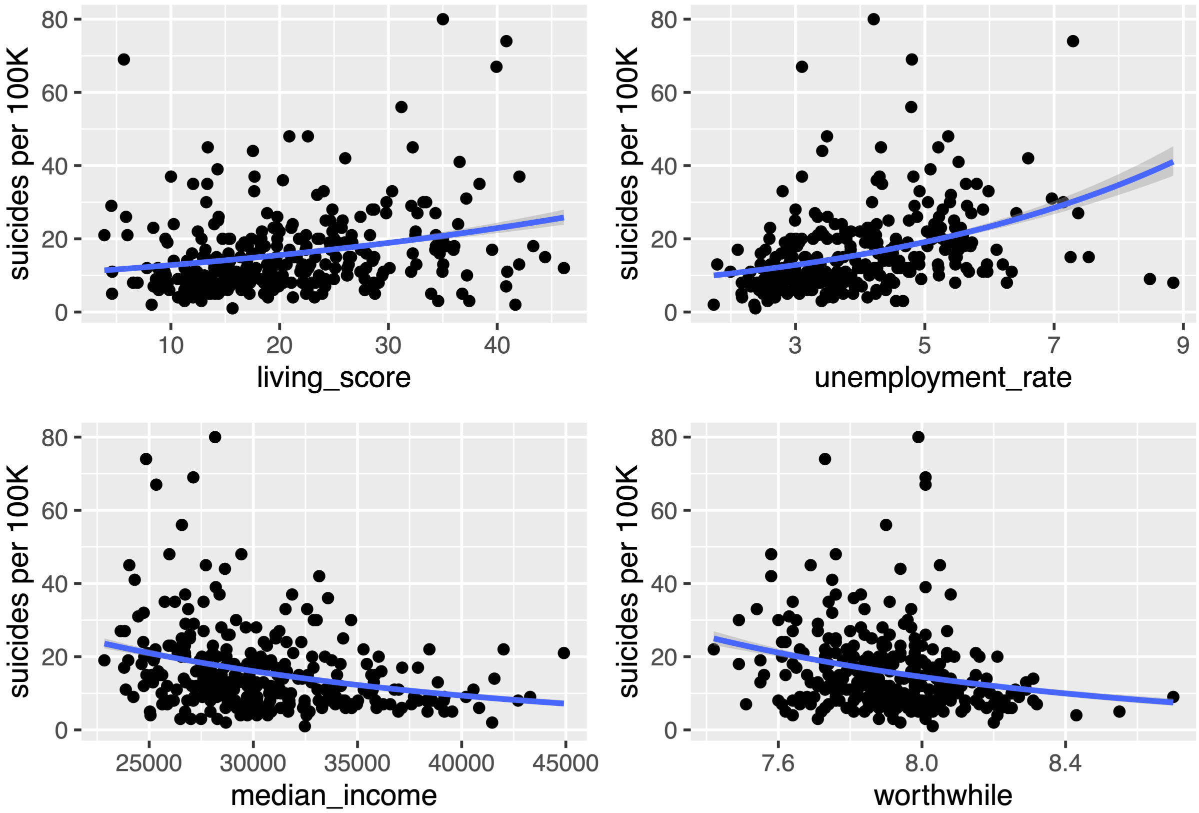


Figure 1: Fitted Curves for Respective Explanatory Variables Against Suicide Rate

Table 1: Regional Suicides Summary Statistics

region	mean_suicides	average_suicide_rate	total_suicides	total_pop
East Midlands	10.03	8.21	331	4030457
East of England	13.42	10.05	577	5742271
London	20.62	7.49	660	8817347
North East	23.92	10.85	287	2644727
North West	17.49	9.40	682	7258627
South East	10.71	7.87	664	8433480
South West	17.65	9.90	459	4634080
West Midlands	17.13	8.77	514	5860706
Yorkshire and The Humber	28.70	10.64	574	5396431

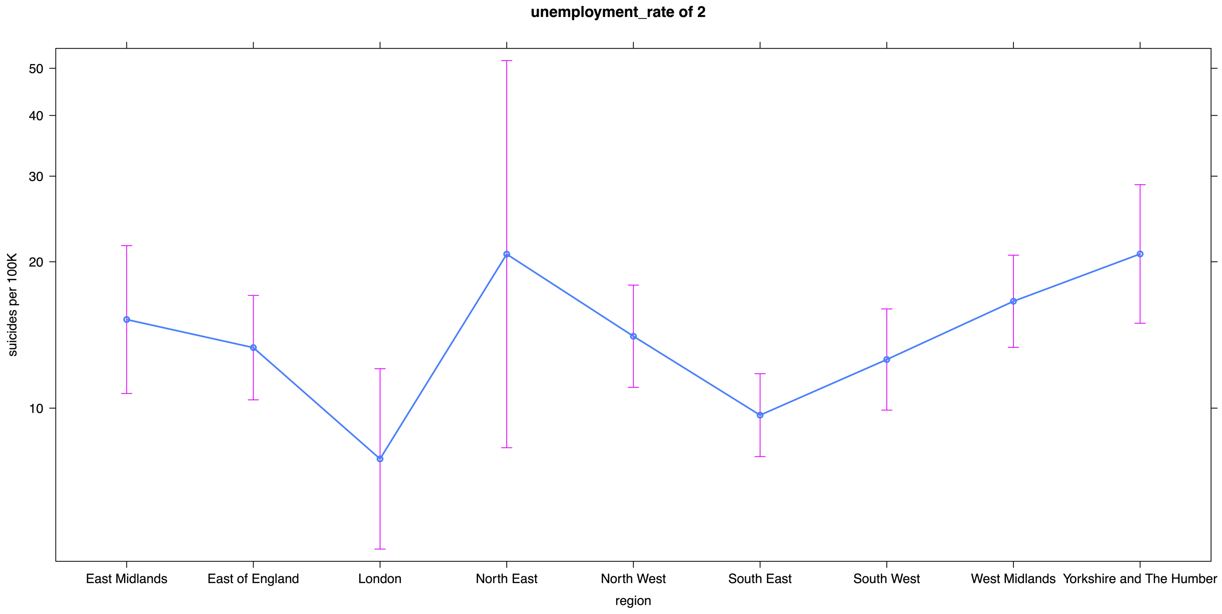


Figure 2: Effect of Low Unemployment Rates on Regional Suicide Rates

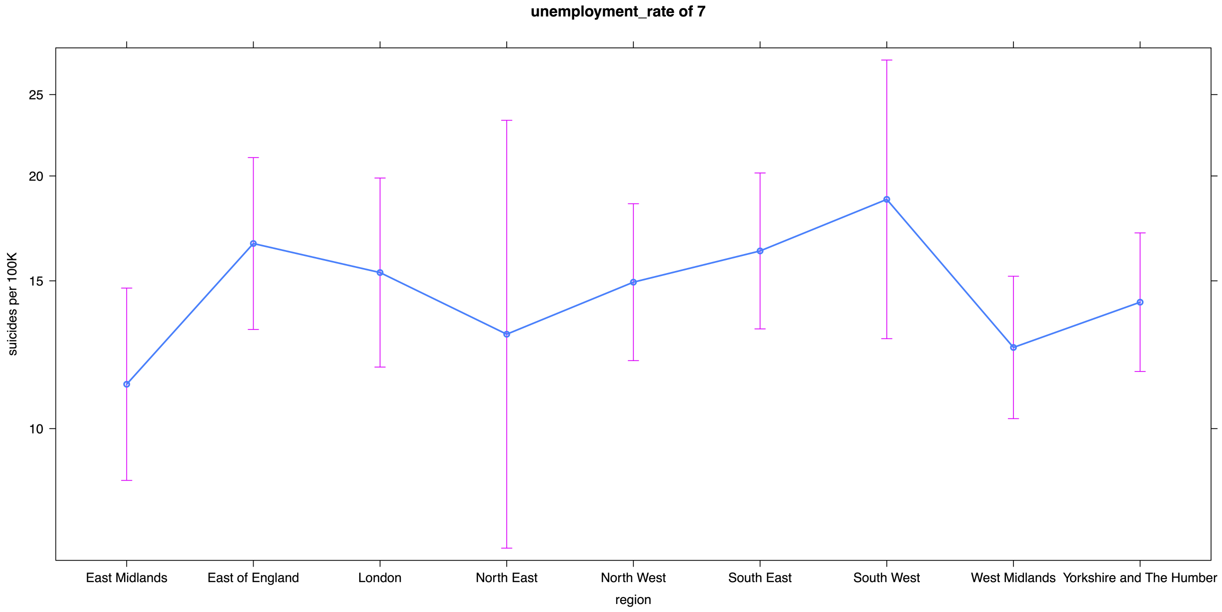


Figure 3: Effect of High Unemployment Rates on Regional Suicide Rates

1.4 Q3. Are there any difference in suicides with regards to the North-South divide?

Marginalising the regions even further into North and South proves that there is a somewhat significant difference overall. At low unemployment rates and living scores the north tends to have more suicides but the difference is debatable as the intervals just barely overlap. At higher unemployment rates both regions see similar suicide rates (approx 0.5 difference between suicide rate), implying that this has significant influence over suicide rates in both regions. Likewise, at higher levels of deprivation, there is approximately 1.0 difference in the suicide rates between both regions. Thus, overall there is a difference in suicides with regards to the North-South divide, mostly owing to differences in unemployment rates and deprivation; in other words, the North sees higher unemployment rates and more deprivation when compared to the South. When we consider the traditional notions surrounding the North-South divide in England, this conclusion supports aspects regarding the socioeconomic makeup of these regions.

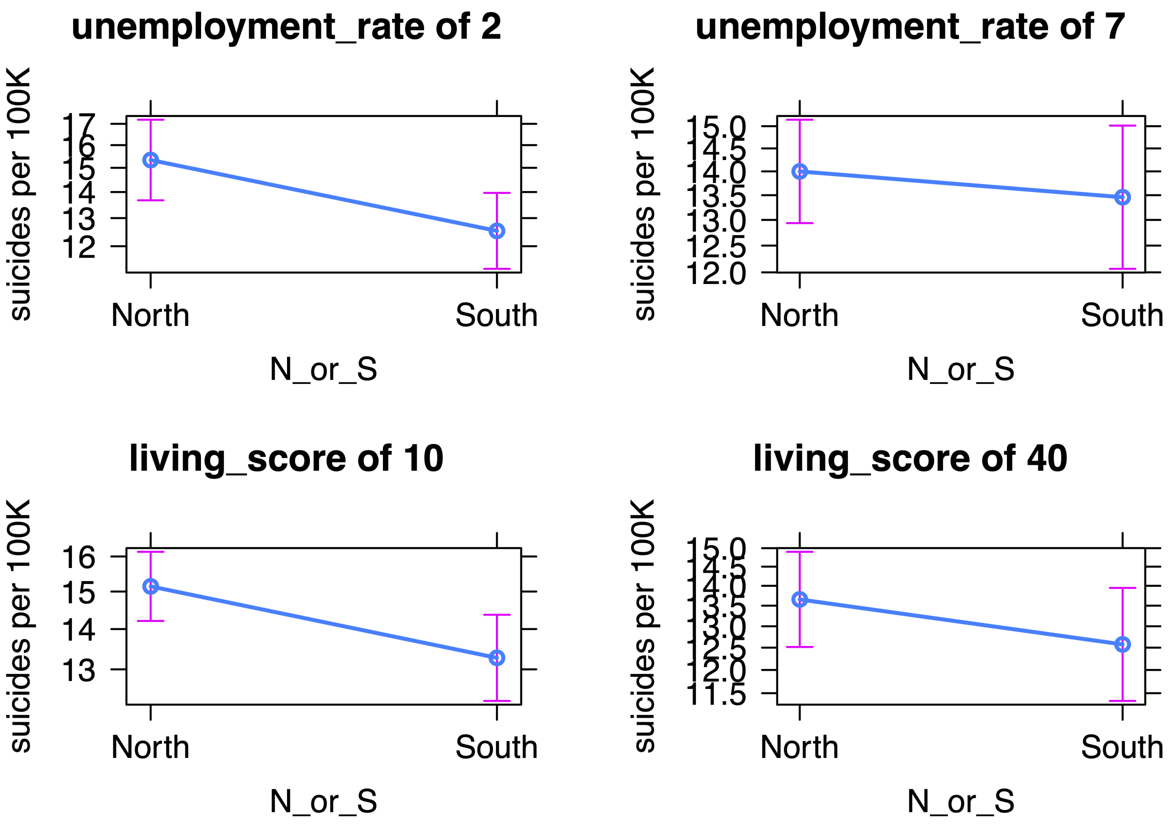


Figure 4: Effect of Unemployment and Deprivation on Regional Suicide Rates

2 B. Materials and Methods

2.1 Data

After creating a dataframe without missing values, we first investigate the relationship between population size and the number of suicides. Intuitively, it is usually assumed that the number of suicides increases with growing population size and this is indeed observed in this case, as there is a strong linear relationship seen in the plot below (*). However, this creates an issue in terms of our analysis, as we are required to determine the association of the other four factors with suicides. More specifically, since a great amount of the suicides can be explained by the population size, we see the need to hold population constant (assume it to be a value of 1) so that it is not estimated by our model. We implement this by creating two offsets: one in terms of log(population/100,000) and just population/100,000; these will be particularly useful when having to try different models further on.

In order to visualise the data, we first plot a histogram of the frequency of the number of suicides, which demonstrates a strong positive skew as seen below (*). Hence, we should anticipate having to either fit a poisson generalised linear model or a transformed linear model in order to best model the data.

In order to implement further analysis, categorical variables are generated in terms of low, medium and high, with regards to income, wellbeing, unemployment and deprivation. We employ visualisation of correlations between the numerical variables (*). We note the high correlation between population and suicides as we also explained earlier. None of our other variables seem to have a particularly high correlation with suicides, however, some variables, such as income and unemployment, seem to display moderate correlation between each other. We note this as they may be an indication of interaction taking place.

2.2 Q1. To what extent are personal well-being, unemployment, income and deprivation associated with the number of suicides in England?

My first thought was to construct a simple linear model using the required variables to be estimated and the non-log offset (as this is a non-transformed model). This model resulted in an R-squared value of 0.334, however the plot of the actual values vs fitted values proved to be very poor. In order to improve on this, I decided it might be best to plot each explanatory variable against suicides and use LOESS smoothing to better visualise their relationship.

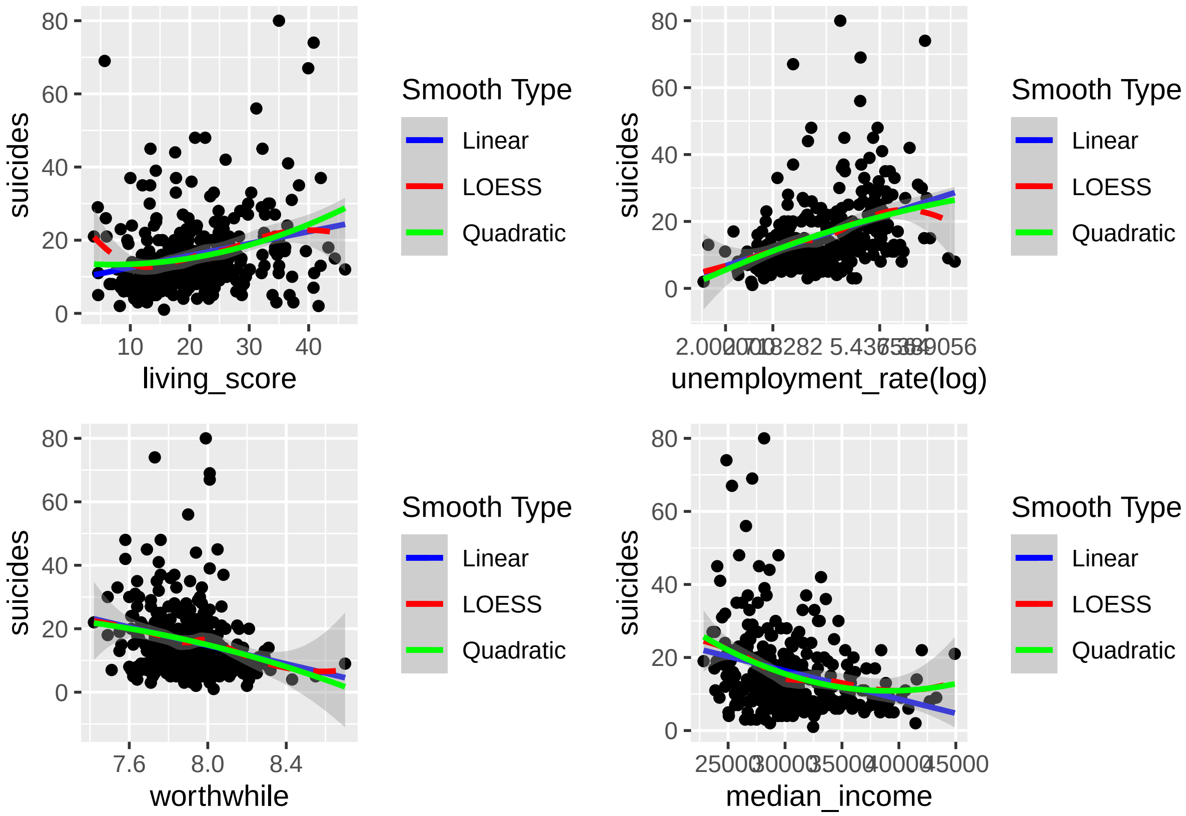


Figure 5: Smoothing

Along with LOESS smoothing, I also included linear and quadratic smoothing for ease of analysis. Firstly, it should be noted that unemployment_rate was log-transformed as it allowed for better linear smoothing; while the linear line slightly overestimates for higher unemployment rates, it still follows the general linear trend well and so we will consider this transformed variable in our model. Living_score can be seen to appear fairly cubic but this may just be possible due to extreme values (or outliers). Instead, we consider the linear trend which follows the LOESS line fairly consistently, only slightly underestimating at lower living scores and overestimating even less at higher scores. Likewise for worthwhile, the linear line follows the LOESS line consistently and so we keep the linear term in our model. Lastly, we consider a quadratic transformation for median_score.

Implementing the transformed explanatory terms only saw little improvement to the R-squared value to 0.340 and thus, not much change in the plot of the actual values vs fitted values. Because of this, I decided it would be best to work backwards with the current full model and use the AIC values as they are a better representation of model fit.

From the drop1 command (*), we can see that p-value for log(unemployment_rate) is greater than 0.05 and so we reject the null hypothesis of this term being insignificant in our model at the 95% confidence interval. Related to its significance is the fact that its deletion would cause the AIC of the current model to increase from 1345.5 to 1356.8. On the other hand, deletion of the quadratic median_income term variable offers the greatest minimisation in our AIC value to 1343.7. In order to further determine whether this decision is justified, we implement the stepwise AIC command.

```
## Start:  AIC=1345.47
## suicides ~ log(unemployment_rate) + worthwhile + median_income +
##     I(median_income^2) + living_score
##
##                      Df Sum of Sq   RSS    AIC
## - I(median_income^2)  1     24.67 26488 1343.7
## - median_income       1     45.45 26509 1344.0
## - worthwhile          1    146.44 26610 1345.1
## <none>                            26463 1345.5
## - living_score        1    338.42 26802 1347.2
##
## Step:  AIC=1343.74
## suicides ~ log(unemployment_rate) + worthwhile + median_income +
##     living_score
##
##                 Df Sum of Sq   RSS    AIC
## - worthwhile     1    146.60 26635 1343.4
## <none>                       26488 1343.7
## - median_income  1    352.30 26840 1345.7
## - living_score   1    358.83 26847 1345.7
##
## Step:  AIC=1343.38
## suicides ~ log(unemployment_rate) + median_income + living_score
##
##                 Df Sum of Sq   RSS    AIC
## <none>                       26635 1343.4
## - median_income  1    286.54 26921 1344.6
## - living_score   1    465.28 27100 1346.5
```

From the above output, it is suggested that we remove worthwhile and the quadratic median_income terms from the full model. In particular, it is highly plausible that the relationship between the worthwhile variable and suicides can be explained by other variables in the model, hence making this removal valid. In fact, additional removal of the quadratic median_income term would possibly imply that the aforementioned relationship could be explained by the linear median_income term. Indeed, when observing the smoothing plots, one can see that worthwhile and median_income appear very similar, and so, only one term, in this case median_income, may be needed in the model. However, when we implement this into a reduced model, we can a slight decrease in our R-squared value 0.337. Conversely, removing only the quadratic median_income term sees the R-squared remain constant at 0.340. Moreover, because of this somewhat negligible difference in the AIC of the model without the worthwhile term (1343.38) and with the worthwhile term (1343.74), I decided to keep the term in the model for ease of interpretation with regards to question 1.

With regards to the residual plots (see Fig.?), we note the tendency for there to be increasing variation in the residuals as the fitted values increase. This phenomenon is known as heteroskedasticity and hence implies that we are dealing with a non-normal distribution due to a non-constant variance. This relates to the point made earlier using the histogram which revealed a strong positive skew in the data. To solve this problem, we can build our way up to implementing a non-normal distribution to ultimately model our data.

Following this discovery, I decided to create a transformed linear model by transforming the response variable and using the reduced model. Here, we implement the log offset as we are using the log of the response.

The transformed linear model shows great improvement as the R-squared has increased significantly to a value of 0.762 so we continue with this model. Considering the correlation plots, we can investigate whether there is any interaction taking place between the explanatory variables. In particular, out of the entire dataset, the most correlation is seen between unemployment_rate and median_income then between unemployment_rate and worthwhile. Though the correlation is only moderate, it may be an indication of some interaction and so we plot their respective interaction plots below.

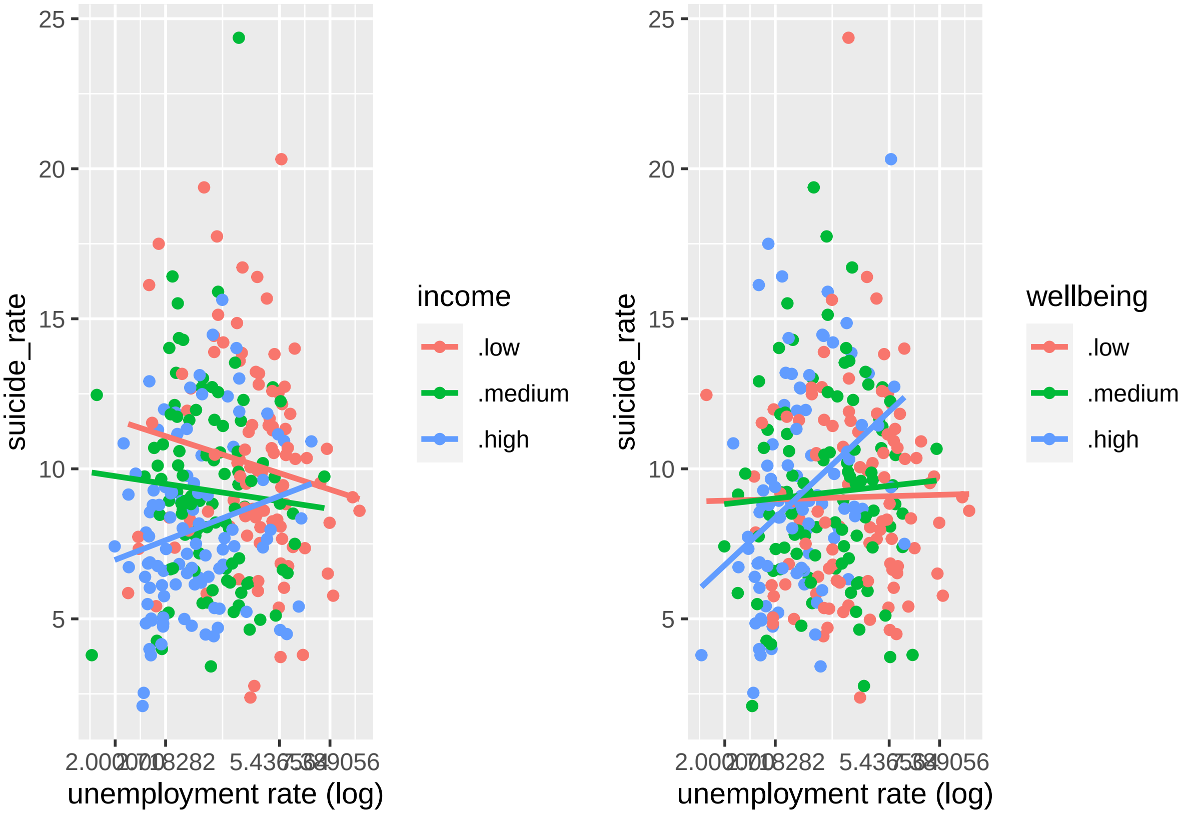


Figure 6: Interaction Plots

With differing levels of income and wellbeing, we can see that the plotted lines cross at differing unemployment rates, thus implying that there is interaction taking place between the respective variables. But how significant is this interaction in terms of improving our model? Adding the interaction of log(unemployment_rate) and median_income sees a considerable improvement in our R-squared value to 0.792. On the other hand, the interaction of log(unemployment_rate) and worthwhile only sees less improvement in our R-squared value to 0.773. In order to further verify the significance of each interaction term, we conduct an ANOVA using the original transformed linear model and the models containing the added interaction terms below.

```
## Analysis of Variance Table
##
## Model 1: log(suicides) ~ log(unemployment_rate) + median_income + living_score +
##     worthwhile
## Model 2: log(suicides) ~ log(unemployment_rate) + median_income + living_score +
##     worthwhile + log(unemployment_rate) * median_income
##   Res.Df    RSS Df Sum of Sq Pr(>Chi)
## 1    292 39.305
## 2    291 38.319  1   0.98546 0.006226 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## Analysis of Variance Table
##
## Model 1: log(suicides) ~ log(unemployment_rate) + median_income + living_score +
##     worthwhile
## Model 2: log(suicides) ~ log(unemployment_rate) + median_income + living_score +
##     worthwhile + log(unemployment_rate) * worthwhile
##   Res.Df    RSS Df Sum of Sq Pr(>Chi)
## 1    292 39.305
## 2    291 38.691  1   0.61422  0.03161 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Both generated p-values for the models the respective interaction terms prove to be less than 0.05, meaning that we can reject the null hypothesis of these interaction terms being insignificant in our model at the 95% confidence interval. In particular, the p-value corresponding to the log(unemployment_rate) * median_income term is seen to be considerably smaller than the p-value for the log(unemployment_rate) * worthwhile term. Considering this, we can infer that the log(unemployment_rate) * median_income term has greater significance with regards to our model, and so we retain this term rather than the log(unemployment_rate) * worthwhile term. In my final search for an improved model, I decided to implement two generalised linear models below: one poisson with a log link and one gaussian with a log link.

```
##          df     AIC
## mod_glm   7 1812.61
## mod_glm2  6 1769.95
```

Fitting with these generalised linear models shows notable improvement to our R-squared value to 0.810 under the gaussian model and 0.806 under the poisson model. In order to determine which one to include in our final model, we consider the AIC of each model which is seen to be 1812.61 under the gaussian model and 1769.95 under the poisson model. The poisson model thus minimises the AIC considerably and so our final and best fitted model can be studied by the summary output below (*). This closely follows the point we made previously regarding the heteroskedasticity and thus non-constant variance of the data. In particular, we know that for a poisson distribution variance increases as mean increases and so the use of this distribution is clearly the best choice. Finally, note how the corresponding p-values for each term, except than worthwhile, is less than 0.05, implying that those terms is significant here. Again, we reiterate the decision to retain the worthwhile term is for ease of interpretation and we will discuss possible reasons why this term ends up being insignificant.

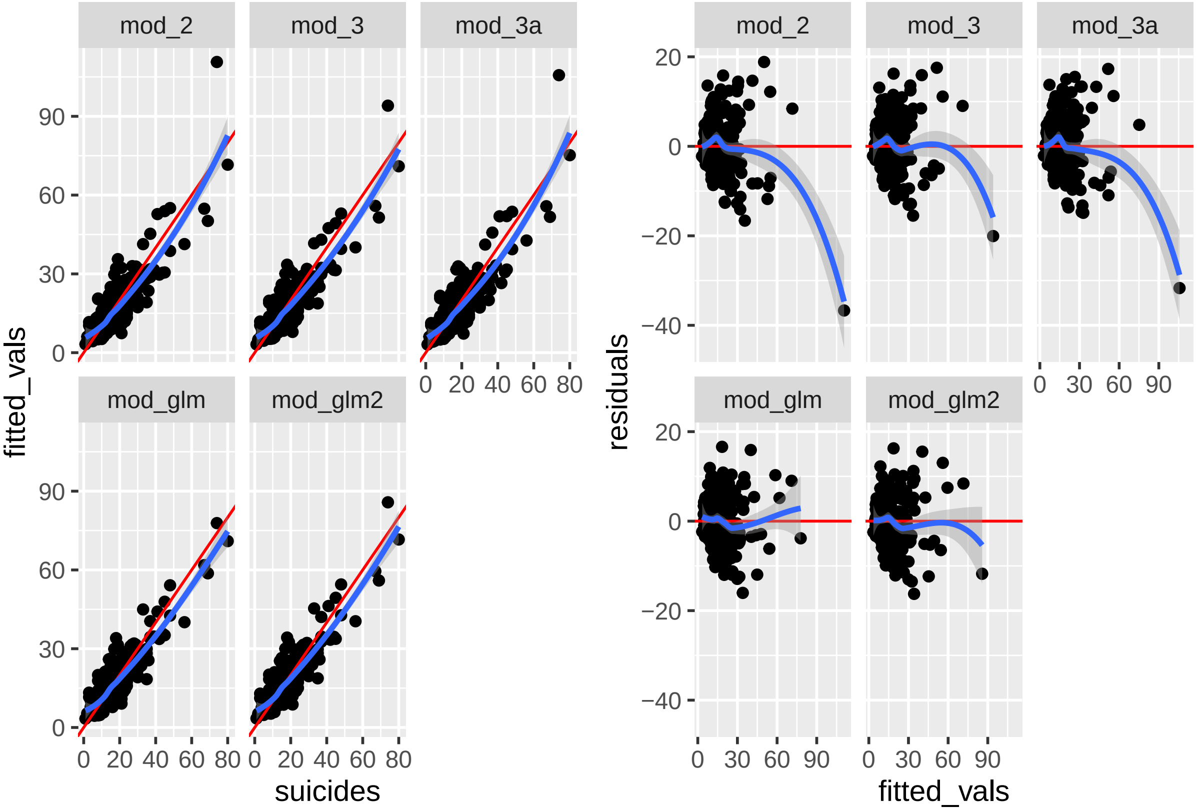


Figure 7: Actual vs Fitted (left) and Fitted vs Residuals (Right) for Transformed and GLM Models

With regards to the residual plots, note how the model building process up to the final poisson model slowly reduces the effect of heteroskedasticity. For easier visualisation we can refer to the Pearsons residuals corresponding to our final model as seen below and observe that the residuals are now evenly distributed about the zero reference line (*). When the fitted values exceed 75 there is overestimation of suicides, this of which is negligible as there is only one point above that range.

From figure 1 from section A we know that suicides increase with increasing levels of unemployment and deprivation and decrease with increasing income and wellbeing. Intuitively this makes sense, however, when looking at the effect plots below we see a different outcome than expected.

Table 2: Final Revised Model: Estimated Parameters with Log Link Poisson GLM

term	estimate	std.error	statistic	p.value
(Intercept)	0.25881	0.79235	0.32663	0.74394
log(unemployment_rate)	0.09958	0.05677	1.75403	0.07943
worthwhile	0.23426	0.09499	2.46621	0.01365
living_score	-0.00216	0.00168	-1.28241	0.19970

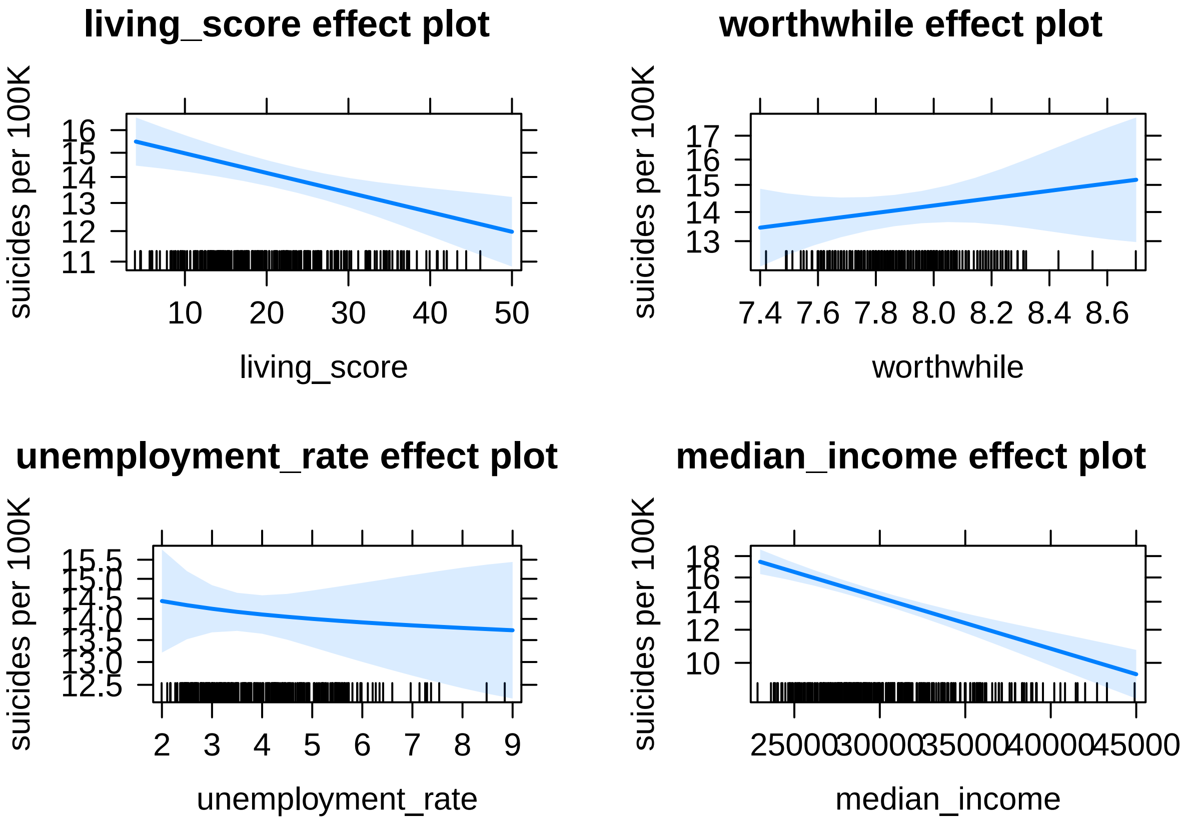


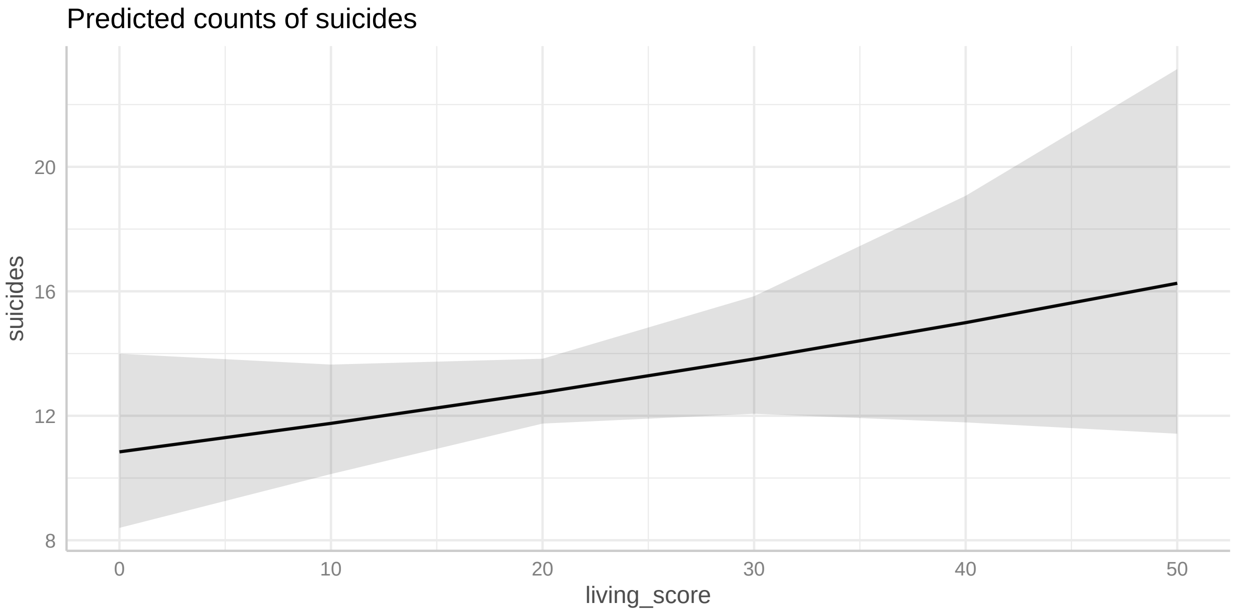
Figure 8: Effect Plots We see that suicide rates tend to decrease with increasing unemployment rates and deprivation while they increase with increasing wellbeing. This is completely opposite to what we saw in the fitted curves and goes against our intuition but we know that the effect plots allow for investigation of individual variable (marginal) effects while holding other explanatory variables constant. Therefore, the reason for this difference is most likely due to the presence of a confounding variable. As we previously mentioned, median_income and unemployment_rate are moderately correlated, possibly to the point where this increase in unemployment_rate seen in figure 1 is actually due to the presence of the variable in our model. In particular, we know that as the unemployment rate declines, persons feel less suicidal (positive relationship), however we also know that as a person's salary/income decreases, unemployment rate will increase and then they become more susceptible to suicidal thoughts. Closely related to this is the fact that less deprivation should result in less suicides, but less income as a result of less employment means an increase in deprivation, resulting in more suicidal thoughts. Likewise, increased personal wellbeing should result in less suicides, but more less as a result of less employment income means a decrease in wellbeing, resulting in more suicidal thoughts.We remove terms associated with this confounding variable and fit a revised final model as follows:

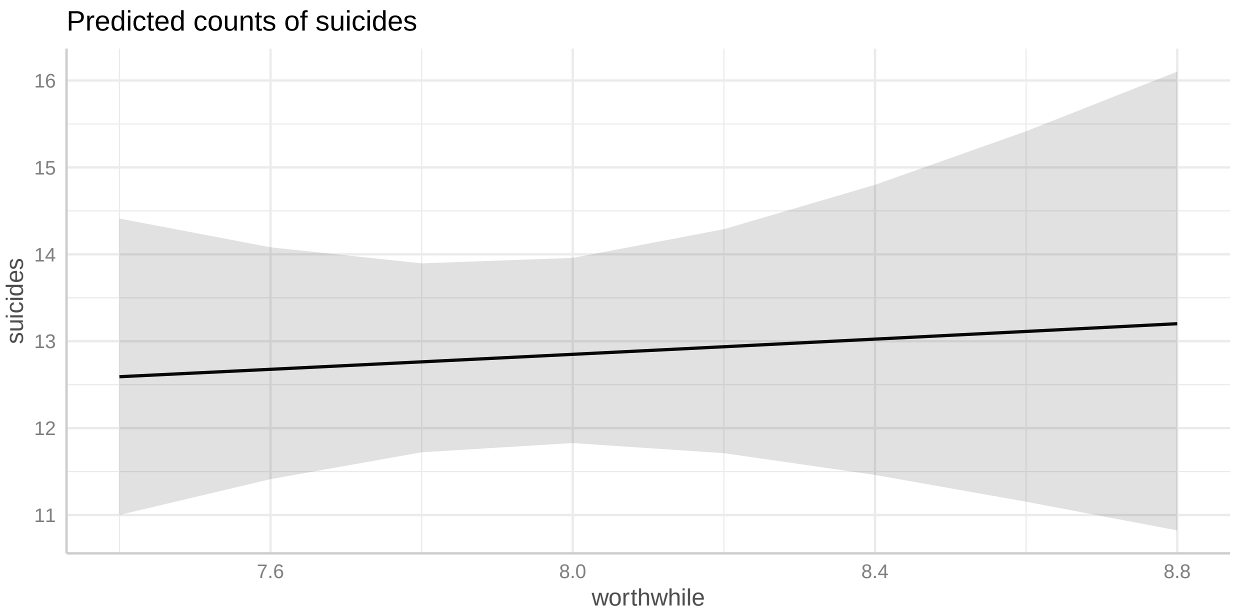
The interpretation for the table above is as follows: for each extra unit that unemployment rate increases by, we expect to see a relative increase of about 10.5% (exp(0.09958)-1 $\approx$ 0.1047068) in the average number of suicides per 100K people. Likewise, the average number of suicides double for every log(2)/(0.09958) ≈ 6.96 unit increase in unemployment rate. From this we are told to expect a relative increase of about 26% in the average number of suicides per 100K people with every increase in worthwhile effects and a relative decrease of about 0.22% in the average number of suicides per 100K people with every increase in living_score; we know these inferences are misleading so we disregard them.

This improves on the unemployment_rate effect but the living_score and worthwhile effects remain the same, possibly indicating that an underlying factor is responsible for such disparity. In particular, we know that living deprivation, personal wellbeing and unemployment usually go hand in hand, as well as the fact that deprivation is strongly influenced by environmental factors. Another way to think about it is that unemployed persons may not be able to afford to live in 'better' environments due to lack of income, this of which affects their wellbeing. With this in mind, we investigate the effects of regional differences in the following section in hopes of improving our model.

2.3 Q2. What are the regional differences in the number of suicides in England?

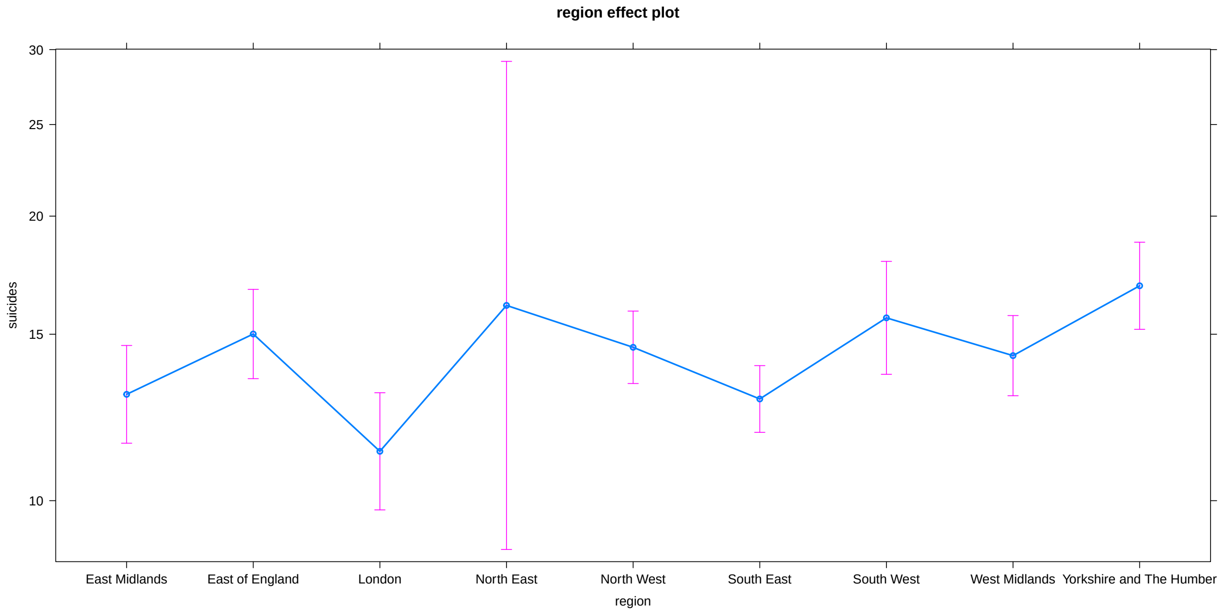
From the above plot (*), we can see that interaction is taking place between the different regions and so we implement this in our revised model below. Conducting an ANOVA confirms that the log(unemployment_rate)* region and living_score * region terms are significant (*).





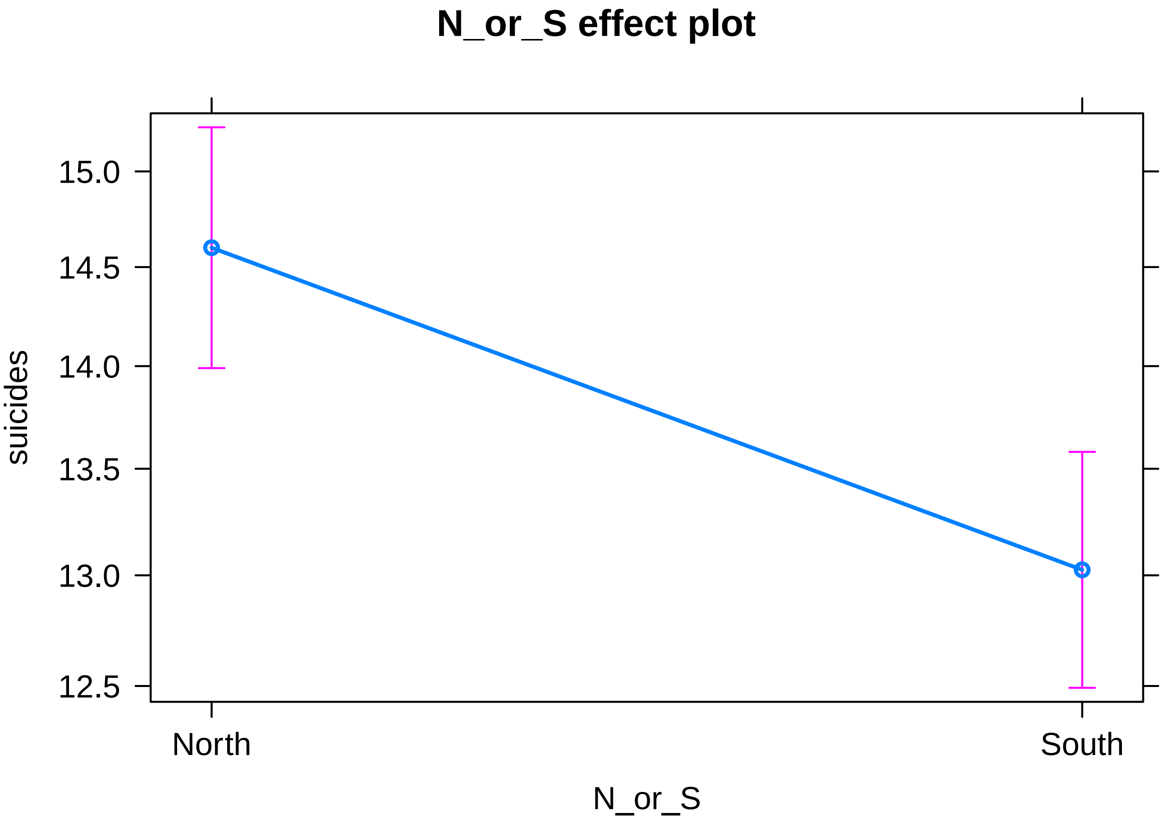
We now see a more sensible result for the marginal effect of living_score appears to be positive in relation to suicide counts. Conversely, we still see no improvement to the worthwhile plot. This may be due to the fact that the range provided for the sampled worthwhile values proves to be very small and thus is not able to confer any significant effects in this case.

From the plot below we can see that a difference between the effect of different regions on suicide rate actually depends on the underlying assumed unemployment rates and living scores. However, note that this difference is not so significant as the intervals for each region overlap with each other.



2.4 Q3. Are there any difference in suicides with regards to the North-South divide?

In order to implement this analysis we must first create a new column in our existing/initial dataframe. To further catergorise the regions into North and South, I referred to this (https://geographical.co.uk/images/2020/North_South_divide/Map.png) map. For this question I opted to remove the worthwhile term from the model due to reasons discussed in the previous question.



3 References

Include your references in this section. The references below are examples of books and a website.

1. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* on-line version
2. Wickham, H. & Grolemund, G. *R for Data Science.* on-line version
3. Department of Maths website