Statistical Modelling of Indicators of Social Health Across England

Yung-Wei Ko

Content:

Executive Summary

To provide reliable suggestions to government officials in County Durham detailing how residents' overall satisfaction of living in their local area can be improved, our group set out to analyze the dataset provided through a structured process: data standardization, multivariate linear regression, outlier identification and model robustness assessment. In the model created to investigate this task, multivariate linear regression was applied as the main mathematical method to investigate the relationship between four different variables and residents' overall satisfaction of living in their local area. These variables included; ability to influence decisions in the local area, community cohesion, belonging, and problems with drug/use selling. The findings from our model were used to scientifically identify the primary drivers impacting resident satisfaction, and thus develop targeted policy recommendations for improving living satisfaction scores. Three main suggestions we propose are as follows:

- **Focus directly on combating drug use and selling:** It is highly suggested that County Durham officials should focus on drug related problems first. Our model reveals that this factor has the greatest negative impact on residents' overall satisfaction compared with the other three factors analysed, demonstrating that tackling issues with drug use and selling could lead to a marked improvement in satisfaction levels, as this is a key area that residents consider to be problematic.
- **Attend to wider societal issues related to drug use and selling:** Drug related problems usually are exacerbated by and contribute towards and wider societal problems, such as increasing crime rates and antisocial behaviour, which have knock-on effects to overall resident satisfaction. Therefore, addressing drug related issues within the context of other social problems in County Durham could help improve overall satisfaction score effectively. Providing addiction treatment and prevention programs could be a possible measure to lower the percentage of drug use and selling in the community.
- **Improve community cohesion and foster greater sense of community harmony:** While we argue that addressing drug related issues should be the priority, our model suggests that officials in County Durham should also seek to improve community cohesion, as this is the factor that has the most significant positive impact on overall satisfaction scores. We recommended the local government work towards improving community harmony through organizing more community events and activities and encouraging interaction between residents from different backgrounds. Besides, supporting inclusive policies is another practical approach to foster understanding and strengthen trust among a community of diversity.

Aside from the insights above, we believe **the strategies taken by the officials should be tailored by the region's differences.** Our geographic dummy variable gave results for London which reported a relatively lower satisfaction level compared with non-London regions. This result highlights the fact that London is facing unique challenges rooted from the nature of urban living. That is, besides the four factors, community harmony, drug issue, residents' sense of belonging (Belong) and the power of to affect decisions (Influence_Decisions), there are some other key factors that are fundamental to affect living satisfaction especially in urban regions, however the dataset does not give insight to what these may be or how these may tie into satisfaction. Thus, policies targeting to improve the satisfaction score in County Durham should optimally be taken into consideration with other data in order to meet the unique needs of the area that this dataset cannot highlight.

Findings

The multivariate regression model we produced has allowed us to analyse the key factors influencing overall satisfaction of residents in County Durham. The model utilises social health indicator data from the UK Government website, whereby adult residents in 353 local areas within England were asked 5 questions regarding their opinions on different aspects of their local area, responding with either ‘yes’ or ‘no’ to each. Our model is based around the results of these 5 questions, and seeks to assess what factors have the greatest impact on how residents answered question 5 (“Overall, are you satisfied living in your local area?”). As a result, our model provides insight into which areas should be the highest priority for improvement, in order to increase overall resident satisfaction. The four factors we assessed and used as variables within are model are as follows:

- **Get_On_Well**: Represents the percentage of respondents who believe people from diverse backgrounds in their local area interact positively and harmoniously. This variable reflects the confidence residents have in the quality of social relationships within their community.
- **Influence_Decisions**: Indicates the percentage of respondents who feel they can influence decisions made in their local area. This variable serves as a measure of how empowered residents feel in shaping their community’s future.
- **Belong**: Captures the percentage of respondents who feel a strong sense of belonging to their neighborhood, symbolising the depth of connection and community integration among residents.
- **Drug_Use_And_Selling**: Reflects the percentage of respondents who perceive drug use or drug-related activities as a significant problem in their area. This variable highlights the extent of concerns related to public safety and crime.
- **London**: A binary variable representing whether the area is located in London (1) or outside London (0).

The equation produced by the model is:

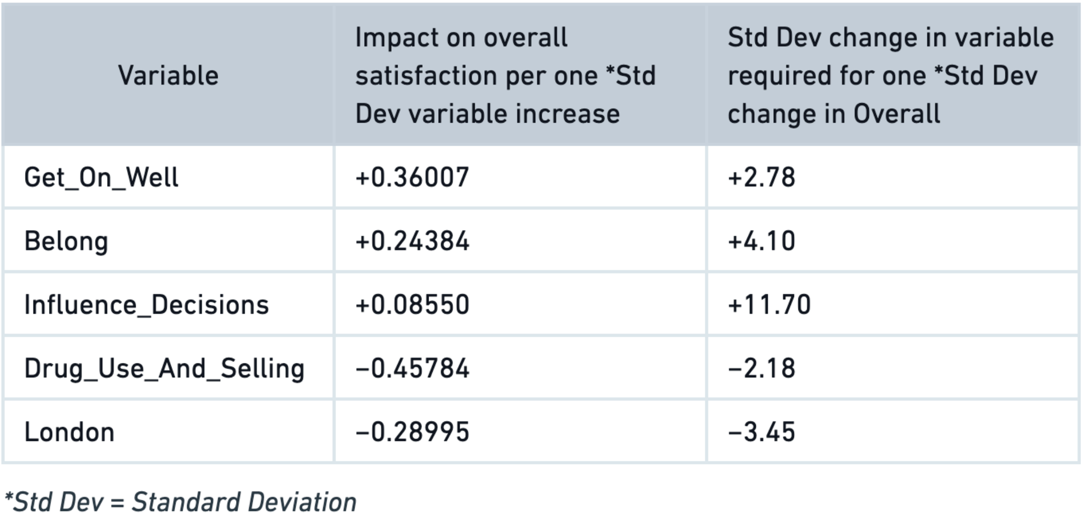
$$\textbf{Overall} = \textbf{0.02718} + \underline{\textbf{0.08550}} \times \textbf{Influence\_Decisions} + \underline{\textbf{0.36007}} \times \textbf{Get\_On\_Well} + \underline{\textbf{0.24384}} \times \textbf{Belong} - \underline{\textbf{0.45784}} \times \textbf{Drug\_Use\_And\_Selling} - \underline{\textbf{0.28995}} \times \textbf{London}$$

(where London = 1 if the area is in London, and 0 otherwise)

The underlined values in the equation above are the model’s coefficients. These coefficients represent the relative importance of each variable within the model, indicating the influence they have on *Overall,* or residents’ overall satisfaction. We say that these coefficients represent the change in *Overall* if the corresponding variable increases by one standard deviation. In other words, if more respondents answered ‘yes’ to a survey question, and the result value for the corresponding variable increases by one, how would *Overall* change? Would it increase, or decrease? And is it greatly impacted or just slightly?

For instance, *Get_On_Well* has the greatest positive influence on overall satisfaction, and thus it is the largest positive coefficient within our model. If the value of *Get_On_Well* increases by one, *Overall* will increase by 0.36007, and residents’ overall satisfaction increases. *Drug_Use_And_Selling* has the greatest negative influence on overall satisfaction, and thus is the largest negative coefficient within our model. If the value of *Drug_Use_And_Selling* increases by one, *Overall* decreases by −0.45784.

The influence of each variable on satisfaction is summarised below, alongside how much the value of each variable would need to increase or decrease by in order for *Overall* to change by one standard deviation. This analysis offers insight into which factors should be prioritised for the greatest impact.



Variable	Impact on overall satisfaction per one *Std Dev variable increase	Std Dev change in variable required for one *Std Dev change in Overall
Get_On_Well	+0.36007	+2.78
Belong	+0.24384	+4.10
Influence_Decisions	+0.08550	+11.70
Drug_Use_And_Selling	−0.45784	−2.18
London	−0.28995	−3.45

**Std Dev = Standard Deviation*

Figure 1: Table of variable coefficients

From the table in Figure 1, *Get_On_Well* emerges as the most effective variable for improving satisfaction, as only a relatively small increase of 2.78 in this factor is needed to achieve a significant boost in overall satisfaction. This demonstrates that not only is the creation of a more inclusive and harmonious community seen as greatly important for residents and key to how they feel about living in their local area, but improving this area is also very meaningful to them, and will successfully make substantial improvements to overall satisfaction. Additionally, although *Belong* also plays a meaningful role within contributions to *Overall,* it requires slightly more effort to produce the same improvement in satisfaction compared to *Get_On_Well.* As such, it will likely take more time and diligent enforcement for equivalent improvements to be seen. It is also evident that *Influence_Decisions* not only has a very small impact on *Overall*, but would also require substantial improvement (about 4.3x more than that required with *Get_On_Well)* in order for *Overall* to increase by one. Resultantly, at this point in time, tackling issues relating to this factor is a low priority, as doing so will not yield the greatest improvements in overall satisfaction, or be the most productive way to spend resources.

Meanwhile, addressing *Drug_Use_And_Selling* is crucial, as at an impact on *Overall* of -0.45784, it is this variable that has the greatest total impact on satisfaction. Unlike the rest of the variables, it has a negative influence on *Overall*, causing satisfaction levels to decrease as *Drug_Use_And_Selling* values increase. With the change required in this variable for a one standard deviation change in *Overall* being -2.18, which is even smaller than that required of *Get_On_Well,* it is clear that its negative impact on overall satisfaction is substantial. Not only would mitigating the impact of *Drug_Use_And_Selling* create a noticeable improvement on overall satisfaction, but failing to attend to issues related to this factor would swiftly reduce residents' satisfaction with their local area too.

We therefore resolve that focussing upon issues relating to drug use and selling should be the priority area for County Durham officials to improve, as this factor contributes the most towards residents' overall satisfaction and would be the most effective use of the county's resources, as a result of how residents place importance upon this issue. Our results also heed a warning to officials that drug related issues should not be ignored, nor neglected in favour of attending to other issues that may work to directly improve overall satisfaction, as only a relatively small change in this variable causes a direct decrease in overall satisfaction. Failing to act is only likely to perpetuate this. If officials in County Durham were to look into another area in order to improve overall satisfaction, we propose that making improvements towards community cohesion would be the second priority, as the variable *Get_On_Well* has the second highest influence on *Overall.* Attending to these recommendations will offer the greatest potential for impactful and efficient improvements, creating a more satisfied community.

Based on these findings, we propose the following recommendations to enhance overall satisfaction among residents:

1. **Directly address existing drug-related issues (Drug_Use_And_Selling):** Public safety programs aimed at reducing drug use and related crimes should be a key focus. This could involve harm-reduction programs, for example introducing mobile support units that provide overdose prevention services, drug dependence treatments, needle exchange programs, etc. (World Health Organisation, 2020). Such steps work to help keep existing drug users safe, minimising the negative consequences of drug use and promoting recovery, thus reducing impacts on communities by improving the cleanliness and safety of neighbourhoods (Drugwise, 2016)
2. **Expand drug education and prevention programs (Drug_Use_And_Selling):** Expanding programs dedicated to prevention and early intervention would also be beneficial. This could be done by increasing youth engagement initiatives in high risk areas, in order to prevent young people from drug involvement, education campaigns particularly in schools or colleges, to promote the risks of drug use, or community outreach sessions. At present, County Durham's Drug and Alcohol Services focuses more on recovery than prevention (CDDARS, 2024) - while this is indeed a vital part of tackling drug related issues, bolstering prevention and education programs could improve this service, reducing problems with drug use/selling and thus improving overall satisfaction.
3. **Foster Community Harmony (Get_On_Well):** Initiatives that promote intergroup understanding and collaboration should be prioritised. Programs that encourage dialogue between diverse groups, community-building events, and inclusive policies are likely to yield the most significant benefits in satisfaction levels.

Despite the valuable insights above, there are two main limitations of our approach and data set used. First, the limited scope of variables makes it difficult for our model to fully capture all the determinants of overall satisfaction. This is because regional and indicators diversity were overlooked, leading to omitted variable bias. Urban areas might have its unique indicators affecting overall satisfaction like housing prices and cost of living. Second, the geographic variable was simplified into a binary dummy variable (1 for London, 0 for non-London), neglecting nuances between other non-London regions. This limits the generosity of our findings in explaining satisfaction across other areas.

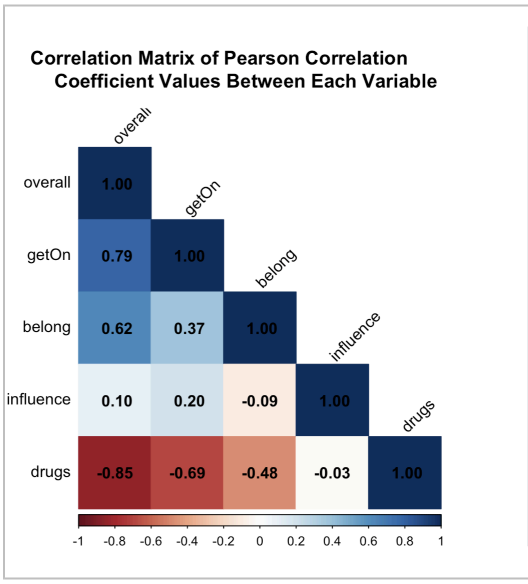
Statistical Methodology

As described in the previous sections, our findings and recommendations were formed as the result of our multivariate linear regression model. We used a series of different approaches in order to construct the model, which were largely driven by steps taken in order to appropriately select our predictor variables. After adjusting the data in order to standardise variables, we started to test the model with different combinations of predictor variables. During testing, we also incorporated a dummy variable to assess region-specific trends in overall satisfaction, providing us with further insight into how to decide upon our final model. The final process of model selection took place after a series of residual checks, including outlier checks which ensure the reliability of our model. The process of model construction and assessment could be divided into the following five steps:

① Standardisation of the dataset:

As our predictors were initially measured on different scales and displayed varying distributions, we standardised all variables to z-scores before fitting the regression model. Converting each variable to have a mean of zero and a standard deviation of one ensures that the resulting coefficients are directly comparable across all predictors. Instead of reflecting changes on original and often non-uniform scales, each coefficient now indicates how the dependent variable shifts with a standard deviation increase of one in the corresponding predictor. This approach enhances interpretability, allowing us to more intuitively understand the relative importance of each predictor. Histograms to demonstrate this standardisation can be seen in appendix A.

② Checking linear relationship between numerical predictors and response variable:



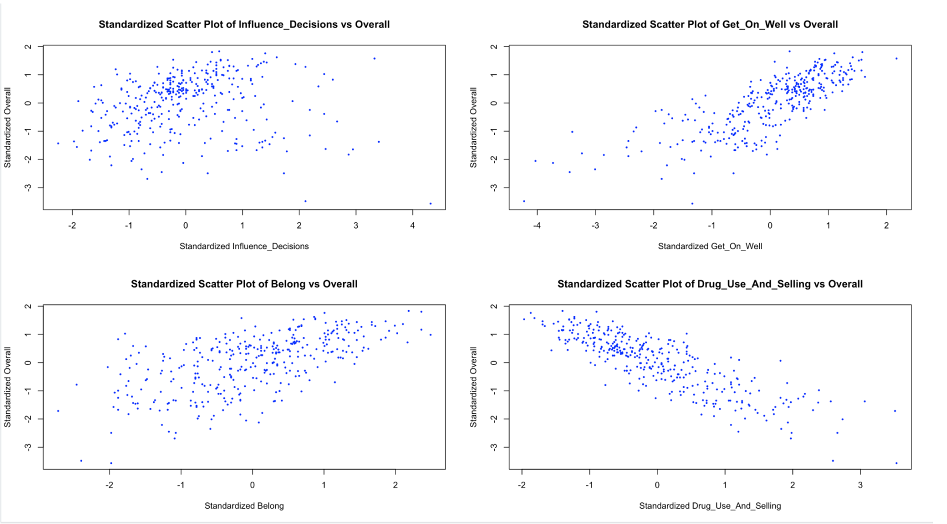


Figure 2 (Left): Correlation matrix of correlation values between each variable and Overall
Figure 3 (Right): Standardised scatter plots of relationship between Overall each predictor

When exploring the linear relationships between response variables and each of the numerical predictors individually, it is helpful to create simple regression models that consider each predictor in isolation, in order to gain an insight into the strength of relationship between each predictor and the response variable. Within our own modelling, this provided insight into direction and magnitude of the influence of predictors before considering the complexity of the full multiple regression model. For instance, when we examined *Influence_Decisions* as a sole predictor, the resulting model shows only modest explanatory power on Overall Satisfaction, and a p-value that hovers around the threshold for statistical significance of 0.05. This suggests that taken alone, *Influence_Decisions* does not strongly explain variations in the response variable, and thus may not have a pronounced linear relationship without the context of other predictors. In contrast, evaluating *Get_On_Well* reveals a high R-squared value, signifying that a substantial portion of the variability in the response can be captured by this predictor. Moreover, the highly significant p-value underscores that the relationship between *Get_On_Well* and *Overall_Satisfaction* is not likely due to random chance. A similar result can be found when examining Belong, which also shows a statistically significant and meaningful linear relationship, although its explanatory power is somewhat more moderate than that of *Get_On_Well*. Perhaps most notably, *Drug_Use_And_Selling* stands out with an even higher R-squared, indicating that changes in this single predictor align strongly with variations in the response variable, and the extremely low p-value provides confidence in the robustness of this linear association.

Taken together, these individual regressions highlight distinct patterns: some predictors, like *Get_On_Well* and *Drug_Use_And_Selling*, exhibit a strong and significant linear relationship on their own, while others, such as *Influence_Decisions*, show less evidence of a standalone influence. These trends can be seen in figures 2 and 3 above. Although these single-predictor models do not account for interactions or shared variance that might emerge when considering all predictors simultaneously, their ability to reveal basic linear connections provide a foundational understanding of variable influence, shaping our subsequent approach to building and refining the comprehensive multiple regression model.

③ **Testing the model with different combinations of predictor variables:**

Determining which predictors to keep in a multiple linear regression often involves weighing both their statistical importance and the practical improvement they offer to the model's explanatory power. During the formulation of our model, we specifically questioned *Influence_Decisions* and whether it should be removed, due to its seemingly limited impact on Overall Satisfaction. As such, we closely examined this predictor within the combined context of other predictors to assess if such interactions improved its explanatory power, creating initial test models both with and without *Influence_Decisions*. Our first test model (Model 1) included *Influence_Decisions* as well as the other available numeric predictors.

The results from Model 1 looked strong, as can be seen in Appendix C. The multiple R-squared value of 0.8475 indicates that the model explains a large share of the outcome's variability, and the adjusted R-squared is similarly impressive. The model's overall significance and the absence of multicollinearity (with VIF values all comfortably below 5) mean the model is both stable and broadly effective. While Influence_Decisions itself may not hold the most influence alone, it performed well within the context of other variables. Following this, we constructed Model 2 without Influence_Decisions to see if anything would be lost if it was to be omitted. Interestingly, this model still performed well, with an R-squared just slightly lower at 0.8474. The other predictors remained highly significant, and the model fit remained robust, thus demonstrating that removing Influence_Decisions didn't cause immediate significant issues with the model.

However, the subtle differences did matter - Model 1's slightly better R-squared and a marginally lower residual standard error (2.764 versus 2.777) suggests that *Influence_Decisions* increases precision of the model by contributing a small but measurable improvement to the model's explanatory and predictive qualities when in tandem with other predictors. Given this, keeping *Influence_Decisions* seems reasonable. All things considered, Model 1, which includes *Influence_Decisions*, feels like the stronger choice, offering a slightly richer and more precise picture of the factors at play.

④ **Incorporating a Geographic Dummy Variable:**

In addition to the core numerical predictors related to community influence, social cohesion, sense of belonging, and perceptions of drug-related issues, we introduced a geographic dummy variable to capture region-specific effects on overall satisfaction. Our dataset encompasses 353 local areas within England, each belonging to a broader geographic region (e.g., North East, London, South West). Despite the fact that we are advising policymakers in County Durham, and geographic location itself is not modifiable, we need to control for latent cultural or structural factors associated with geography to ensure other predictors are not confounded. To do this, we employed a binary geographic dummy. This allowed us to assess whether certain regions, specifically London, consistently deviate from broader trends in satisfaction.

Creation and Rationale for the London Dummy Variable:

To incorporate geography, we initially considered (1) multiple categories representing distinct regions (North East, North West, London, etc.), and (2) three categories representing the North, Midlands, and South. However, exploratory checks indicated that London, in particular, displayed distinct characteristics relative to other regions. It also had a sufficient number of datapoints to be able to reliably generate a coefficient associated with it, unlike the approach in (1). We therefore created a single dummy variable coded as 1 if the observation was from London and 0 otherwise, simplifying interpretation and mitigating the issues of multicollinearity, and the complexity of including multiple dummy variables at once.

The decision to focus on London as the reference point for our geographic dummy was grounded in both data-driven and conceptual reasoning. From a data perspective, London's observations consistently showed different satisfaction levels compared to other regions. From a conceptual standpoint, London's unique socioeconomic environment, population density, housing costs, and diversity, may shape residents' lived experiences distinctly.

Statistical Significance and Influence of the Geographic Dummy

Upon including the geographic dummy in our multivariate linear regression model, we found that being from London was associated with a significantly lower predicted overall satisfaction score. Specifically, the estimated coefficient for the London dummy variable was approximately -0.28995 ($p < 0.002$). This indicates that, ceteris paribus, respondents from London scored about 0.29 points lower on overall satisfaction than those from other regions. While a single-unit shift in a standardised numerical predictor corresponds to a one-standard-deviation change, this dummy variable represents a categorical difference. Simply being from London rather than another region is associated with a marked decrease in satisfaction.

Evaluation of Interactions Between Predictors

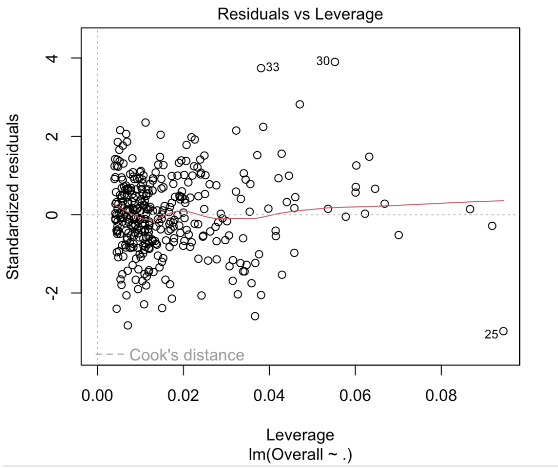
During the model-building process, we explored whether interactions between the geographic dummy variable (*London_vs_Rest*) and other predictors significantly improved model performance. To do this, we considered all possible permutations of interactions, noting the top ten by adjusted R^2, from those with acceptable multicollinearity (indicated by VIF<5). While these interactions offered some insight, the results showed their inclusion did not substantially enhance the model's explanatory power compared to the simpler model without interactions.

Exploration of Interactions

We tested interactions between *London_vs_Rest* and key predictors (*Get_On_Well, Belong, Influence_Decisions, and Drug_Use_And_Selling*) to assess whether the effects of these predictors on satisfaction varied across regions. Some interaction terms, such as *Influence_Decisions: London_vs_Rest and Belong:London_vs_Rest,* were statistically significant, indicating that the relationships between these predictors and satisfaction differed between London and other regions. However, these differences were modest and added complexity without substantially improving the model's fit.

While the interaction model yielded a slightly higher adjusted R2 (Adjusted R2=0.8517), the simpler model – including only the *London_vs_Rest* dummy, and the main effects – was chosen for its clarity and practicality. This model retains robust explanatory power (Adjusted R2=0.8513) and ensures that the relationships between predictors and satisfaction are generalisable across regions. By excluding interaction terms, we provide policymakers with a straightforward framework for interpreting results and developing actionable strategies.

Dealing with outliers:



Checking the influence and leverage diagnostics for our multivariate regression model revealed three notable data points - rows 25, 30, and 33 - corresponding to observations from London (twice) and the northeast region, respectively. These points stood out as potential outliers in the model's residual-versus-leverage plot, as can be seen in figure 4. The initial inclination might be to consider their removal if they were exerting undue influence on the model's predictive accuracy or if their presence suggested errors in data collection. However, there was no compelling reason to believe that the recorded data for these observations were incorrect, nor was there clear evidence that they represented unique anomalies fundamentally at odds with the broader trends in the dataset. To assess their impact, we refitted the model excluding these three outliers and compared the results. The revised model demonstrated only a marginal change in the coefficient of determination (Multiple R-squared: 0.8534, Adjusted R-squared: 0.8513), a difference that did not materially alter our interpretation or conclusions. In other words, the exclusion of these outlying points did not lead to a notably stronger or more robust model. Given this, and without any indication of data mismeasurement or other methodological flaws, we decided not to remove these points from the final model. By retaining these observations, we respect the natural variability in the data and maintain a more comprehensive view of the underlying relationships, ensuring that the resulting model is both honest in its representation and reliable for subsequent inference.

Equation of final model:

$$\text{Overall} = 0.02718 + 0.08550 \times \text{Influence\_Decisions} + 0.36007 \times \text{Get\_On\_Well} + 0.24384 \times \text{Belong} - 0.45784 \times \text{Drug\_Use\_And\_Selling} - 0.28995 \times \text{London}$$

(where London = 1 if the area is in London, and 0 otherwise)

⑤ **Model diagnostics and validation:**

To ensure the reliability and validity of the multivariate linear regression model, we conducted a series of diagnostic checks to assess whether the model meets the standard assumptions of linearity, homoscedasticity, independence, and normality of residuals. Previously, we also evaluated multicollinearity to confirm the stability of coefficient estimates.

Linearity:

Linearity between the predictors and the response variable was assessed through a scatter plot of residuals versus fitted values. The absence of systematic patterns and even distribution suggests that the assumption of linearity is satisfied, confirming that the predictors have an additive, linear effect on satisfaction scores.

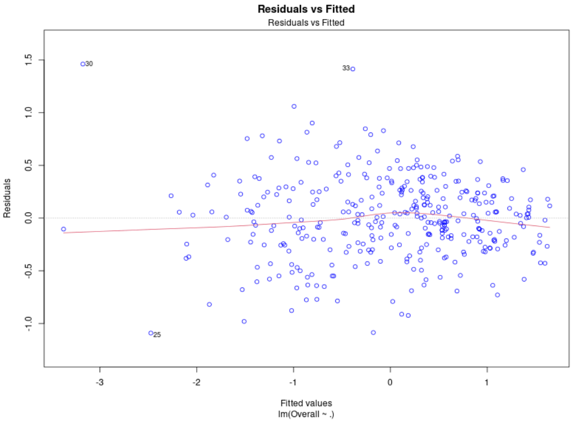


Figure 5 (Above): Scatter plot of residuals vs. fitted value

Homoscedasticity:

While the scale-location plot in figure 6 did not, the Breusch-Pagan test and White's test revealed evidence of heteroskedasticity, returning low p-values and forcing us to reject the null hypothesis of constant error variance. To address this, we used heteroskedasticity robust standard errors (HC(1)) in all hypothesis tests and standard error estimations. This adjustment ensures that the standard errors remain valid, maintaining the reliability of our inference.

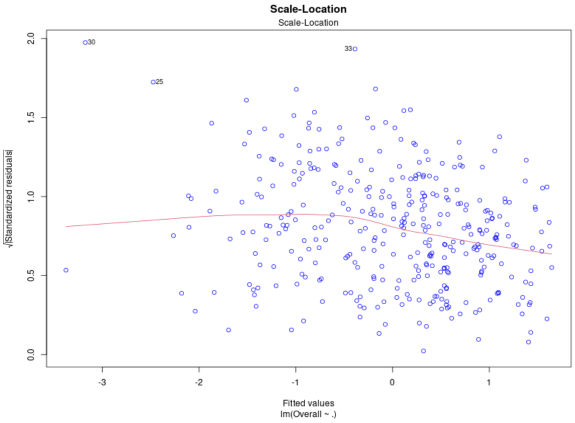
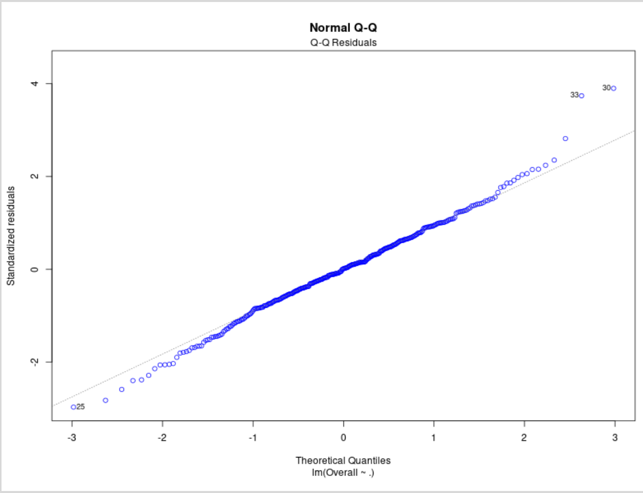


Figure 6 (Below): Scale-location plot

Normality of Residuals:

The normality of residuals was evaluated using a Q-Q plot, as seen in figure 7. The residuals closely align with the theoretical quantile line, indicating that the normality is reasonably satisfied.



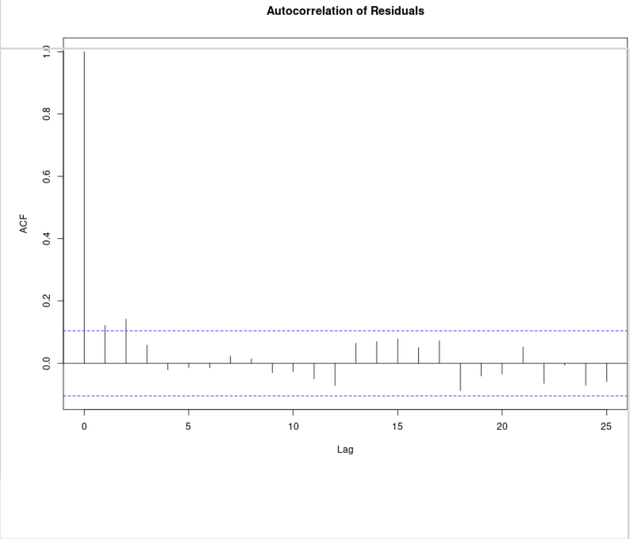


Figure 7: Q-Q plot for normality of residuals

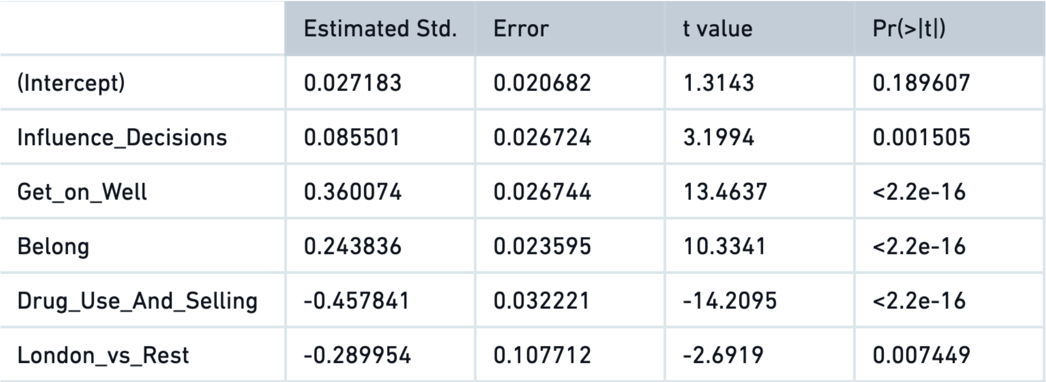
Figure 8: Autocorrelation of residuals plot

Independence of Residuals:

Independence of residuals was assessed using the Durbin-Watson test, which yielded a value close to 2 (DW = 1.7577) despite a p-value of 0.008848. This result, while significant, as well as the ACF plot, suggests very little autocorrelation in the residuals, sufficiently satisfying the assumption of independence.

Final Model with Robust Errors:

In figure 9, the p-value for Get_on_Well, Belong and Drug_Use_And Sellimg are extremely low (<2.2e-16), while Influence_Decisions and London_vs_Rest are statically significant as well (<0.01), confirming that these key factors have statically supported effects on Overall.



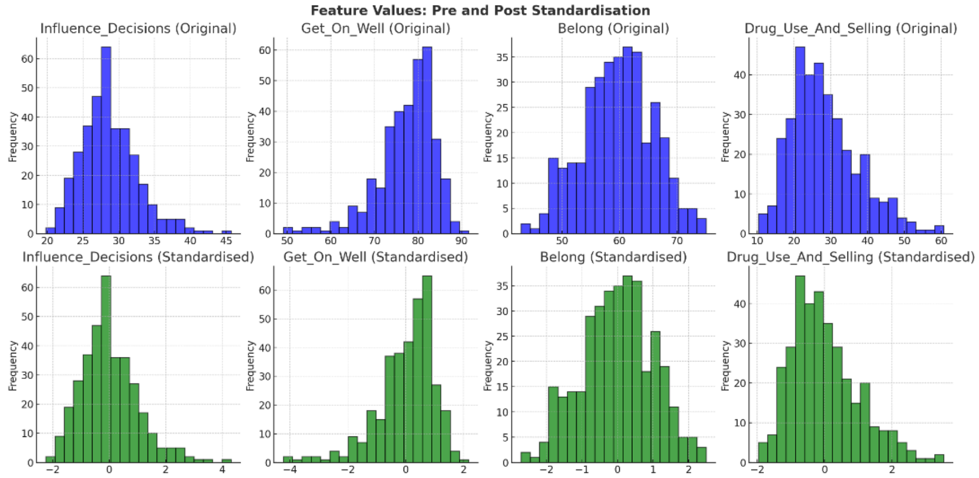
	Estimated Std.	Error	t value	Pr(>|t|)
(Intercept)	0.027183	0.020682	1.3143	0.189607
Influence_Decisions	0.085501	0.026724	3.1994	0.001505
Get_on_Well	0.360074	0.026744	13.4637	<2.2e-16
Belong	0.243836	0.023595	10.3341	<2.2e-16
Drug_Use_And_Selling	-0.457841	0.032221	-14.2095	<2.2e-16
London_vs_Rest	-0.289954	0.107712	-2.6919	0.007449

Figure 9: Table of the t test of coefficients

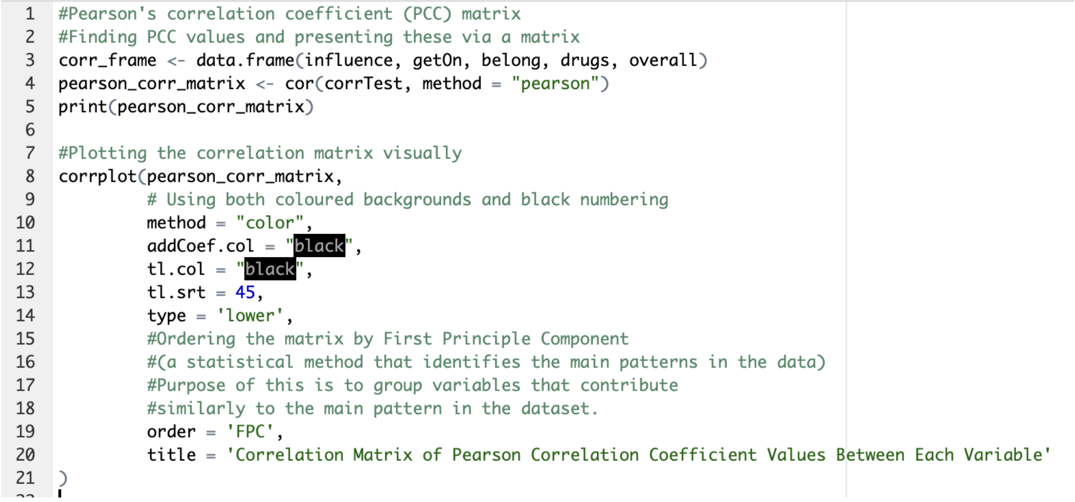
The Wald test is applied to examine the overall significance of the model (Appendix F). With a F-value as 352.8 and p-value < 2.2e-16, showing the model has strong robustness.

Appendix:

Appendix A - Histograms depicting changes after standardisation to z-scores

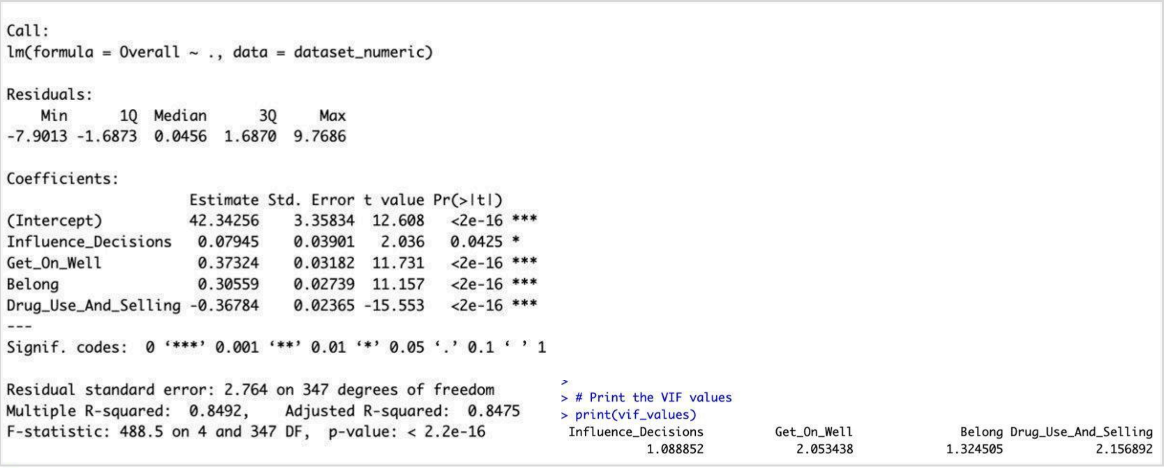


Appendix B - Code for correlation matrix



```
#Pearson's correlation coefficient (PCC) matrix
#Finding PCC values and presenting these via a matrix
corr_frame <- data.frame(influence, getOn, belong, drugs, overall)
pearson_corr_matrix <- cor(corrTest, method = "pearson")
print(pearson_corr_matrix)

#Plotting the correlation matrix visually
corrplot(pearson_corr_matrix,
         # Using both coloured backgrounds and black numbering
         method = "color",
         addCoef.col = "black",
         tl.col = "black",
         tl.srt = 45,
         type = 'lower',
         #Ordering the matrix by First Principle Component
         #(a statistical method that identifies the main patterns in the data)
         #Purpose of this is to group variables that contribute
         #similarly to the main pattern in the dataset.
         order = 'FPC',
         title = 'Correlation Matrix of Pearson Correlation Coefficient Values Between Each Variable'
)
```

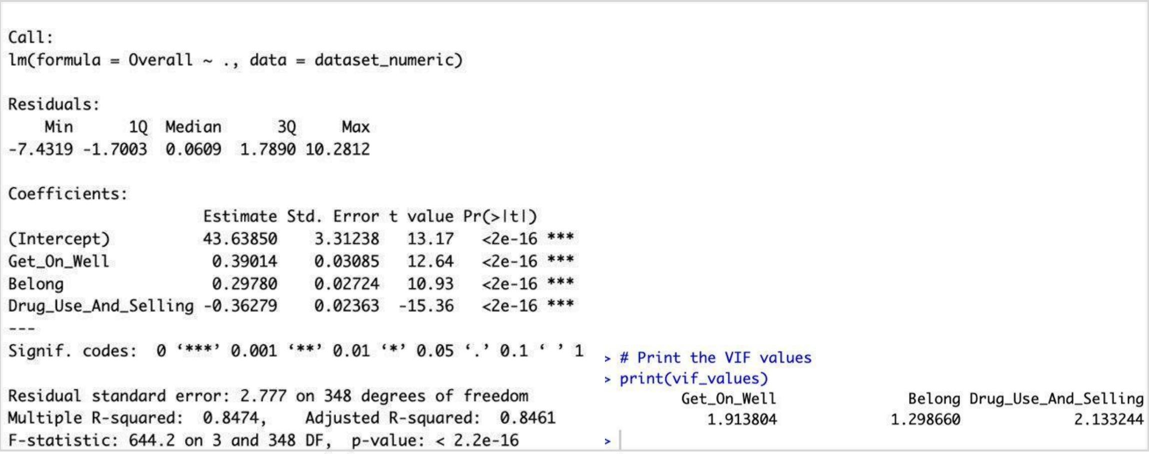
Appendix C - Test model 1 results



```
Call:
lm(formula = Overall ~ ., data = dataset_numeric)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9013 -1.6873  0.0456  1.6870  9.7686 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          42.34256    3.35834  12.608   <2e-16 ***
Influence_Decisions   0.07945    0.03901   2.036   0.0425 *  
Get_On_Well           0.37324    0.03182  11.731   <2e-16 ***
Belong                0.30559    0.02739  11.157   <2e-16 ***
Drug_Use_And_Selling -0.36784    0.02365 -15.553   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.764 on 347 degrees of freedom
Multiple R-squared:  0.8492,	Adjusted R-squared:  0.8475 
F-statistic: 488.5 on 4 and 347 DF,  p-value: < 2.2e-16
```

```
> 
> # Print the VIF values
> print(vif_values)
 Influence_Decisions          Get_On_Well               Belong Drug_Use_And_Selling 
            1.088852             2.053438             1.324505             2.156892 
```

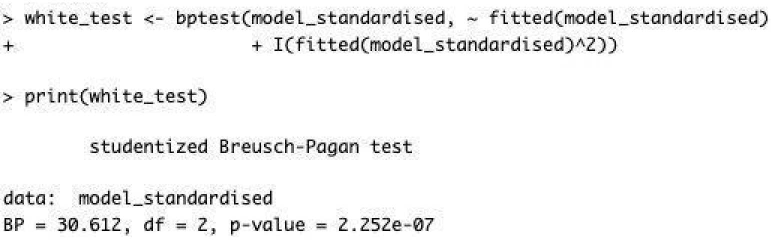
Appendix D - Test model 2 results



```
Call:
lm(formula = Overall ~ ., data = dataset_numeric)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.4319 -1.7003  0.0609  1.7890 10.2812 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          43.63850    3.31238   13.17   <2e-16 ***
Get_On_Well           0.39014    0.03085   12.64   <2e-16 ***
Belong                0.29780    0.02724   10.93   <2e-16 ***
Drug_Use_And_Selling -0.36279    0.02363  -15.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.777 on 348 degrees of freedom
Multiple R-squared:  0.8474,	Adjusted R-squared:  0.8461 
F-statistic: 644.2 on 3 and 348 DF,  p-value: < 2.2e-16
```

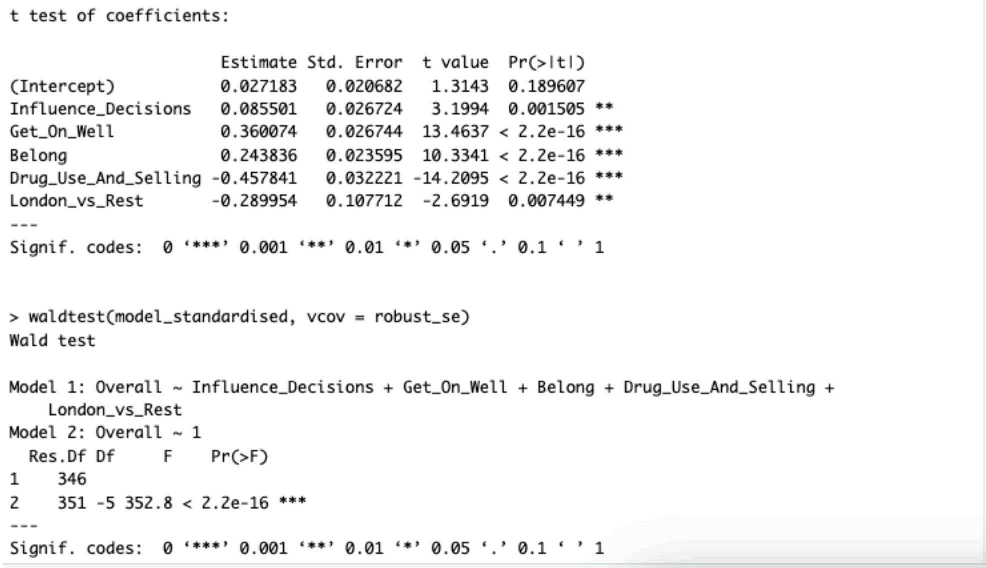
```
> # Print the VIF values
> print(vif_values)
         Get_On_Well               Belong Drug_Use_And_Selling 
            1.913804             1.298660             2.133244 
> 
```

Appendix E - R code for White’s test and Breusch Pagan test



```
> white_test <- bptest(model_standardised, ~ fitted(model_standardised)
+                      + I(fitted(model_standardised)^2))

> print(white_test)

	studentized Breusch-Pagan test

data:  model_standardised
BP = 30.612, df = 2, p-value = 2.252e-07
```

Appendix F - R code for t test of coefficients and Wald test



```
t test of coefficients:

                      Estimate Std. Error  t value  Pr(>|t|)    
(Intercept)           0.027183   0.020682   1.3143  0.189607    
Influence_Decisions   0.085501   0.026724   3.1994  0.001505 ** 
Get_On_Well           0.360074   0.026744  13.4637 < 2.2e-16 ***
Belong                0.243836   0.023595  10.3341 < 2.2e-16 ***
Drug_Use_And_Selling -0.457841   0.032221 -14.2095 < 2.2e-16 ***
London_vs_Rest       -0.289954   0.107712  -2.6919  0.007449 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> waldtest(model_standardised, vcov = robust_se)
Wald test

Model 1: Overall ~ Influence_Decisions + Get_On_Well + Belong + Drug_Use_And_Selling + 
    London_vs_Rest
Model 2: Overall ~ 1
  Res.Df Df     F    Pr(>F)    
1    346                       
2    351 -5 352.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```