ucc

Coláiste na hOllscoile Corcaigh, Éire
University College Cork, Ireland

Handling Unusual Cases: OLS Linear Regression versus Robust Regression

Student name (ID): Maxim Chopivskyy (118364841)

Supervisor: Kathleen O'Sullivan

Module Code: ST4092

Module Name: Data Analytics Project

Abstract

After fitting an ordinary least-squares (OLS) linear regression model, the investigation of unusual cases such as outlying and influential cases is an important step (Chatterjee and Hadi). This is because some unusual cases may affect the OLS estimators for a linear model.

How to identify outlying and influential cases is well described in the literature (e.g. *Aguinis et al., 2013*). Nonetheless, there is much debate in the literature as to how to deal with this type of data (e.g. *Dhakal., 2017; Osborne et al., 2004*). One remedy commonly used is to remove these cases. *Judd et al. (1989)* make several strong points for removal but this is not a view held by others *(e.g. Orr et al., 1991).* Alternative robust regression techniques (e.g. using M estimation and MM estimation) could be applied which are less sensitive to these cases *(e.g. Glasser, 2007; Ortiz et al., 2006; Farcomeni et al., 2012)*. This project will examine robust regression techniques as an alternative to least squares linear regression when data are contaminated with unusual cases. Additionally, it will also assess robust regression for the purpose of detecting unusual cases.

The objective of this project is (1) identify unusual cases, (2) assess how robust regression techniques handle these cases and (3) compare results with OLS linear regression.

Table of Contents

1 Introduction

The presence of unusual cases is one of the biggest challenges faced when fitting an ordinary least-squares (OLS) linear regression model to data *(Aguinis et al., 2013).* In statistics, OLS is a type of linear regression model to estimate an unknown variable (*dependent variable*). It works by fitting a function to all the other variables (*independent variables)* of the cases. The function minimises the sum of squares of difference between the observed dependent variable and the predicted dependent variable by the OLS model.

The purpose of this study is to compare robust regression techniques with OLS techniques when fitting a model to a dataset that contains unusual cases. The first step is to identify the independent variables in the dataset, we use to predict the dependent variable, Satisfaction With University Life (SWUL). The second step is to identify all unusual cases from the dataset, using standard and less known unusual case detection techniques. The third and final step is to fit OLS models and robust regression models to the data. We will compare the efficiency of these techniques.

Unusual cases are cases in the dataset which contains dependent and/or independent variables which are greatly deviated from the rest of the data group *(Osborne, 2004).* There are three types of unusual cases: *outliers*, *leverage cases*, *influential cases*. Before defining unusual cases further, residuals should be explained. Residuals of a case are used to determine if it is an outlier. A *residual* is defined as the difference between the predicted dependent value of a case (based on the regression equation) and the actual, observed dependent value.

An *outlier is* a case whose dependant-variable value greatly deviates from the rest of the data group, given its value on the independent variables. In other words, it has a very high residual. A case with high *leverage* has an extreme predictor variable. Cases with unusually high leverage are referred to as *leverage cases.* Leverage is a measure of how far an independent variable deviates from its mean. Leverage cases greatly affect the estimation of regression coefficients in OLS models. An *influential case* is one that has a high combination of residual and leverage. The removal of this case substantially changes the estimation of regression coefficients. *(https://stats.oarc.ucla.edu/r/dae/robust-regression/)*

Unusual cases occurs commonly in modelling. Outliers and leverage cases can occur due to the variability in the dependent variable and independent variables respectively, or due to experimental error *(Grubbs, 1969).* However, unusual cases are especially prevalent in survey data, because of *careless respondents*. Careless responding refers to survey respondents providing random, and inconsistent answers to questions due to lack of effort in completing the survey. Undetected carelessly given responses in survey data diminish the credibility of study findings *(Goldammer et al., 2020).* In 2020, Goldammer et al. examined the effect of careless responding in survey data. They concluded that careless responding inflated variances of variables, and increases in residual and leverage values.

The presence of unusual cases has great effect on the estimators and parameters calculated in an OLS model *(Glasser, 2007).* Because of this, there is a lot of extensive research being done to determine the best way to deal with unusual cases. They can affect the normal distribution of residuals in an OLS model, altering the odds of making both Type I and Type II errors in unusual case diagnosis *(Rasmussen et al., 1988).* Outliers, by virtue of being different from other cases, usually exert disproportionate influence on substantive conclusions regarding relationships among variables. Accordingly, the issue of outliers is of concern to organizational science research spanning all levels of analysis. Any organisation dealing with some form of multivariable regression is affected by unusual cases. These organisations range from organizational behaviour and human resource management *(Orr, Sackett, & DuBois, 1991)* to strategy *(e.g., Hitt, Harrison, Ireland, & Best, 1998).*

There are several suggested methods to deal with unusual cases. If the unusual cases are incorrectly inputted in the dataset, the value from this case should be modified to its correct value. This is normally done when the data is faulty (inaccurate information), and requires the analyst to know how the data was inputted in the first place. An example of faulty data is when the temperature of a patient is recorded as 100 degrees Celsius. This is obviously not possible for someone to have that high temperature, and so it is faulty data.

However, most of the time, analysts need to deal with non-faulty unusual cases. Winsorization of data is one method to deal with non-faulty unusual cases. Winsorization limits the extremity of outliers to within the specified percentile of the data. Charles P. Winsorto created this technique, to deal with outliers in univariate data (Jerome, 2017).
Consider the data set consisting of:

{92, 19, **101**, 58, **1053**, 91, 26, 78, 10, 13, **−40**, 101, 86, 85, 15, 89, 89, 28, **−5**, 41} (N = 20, mean = 101.5)

The data below the 5th percentile lies between −40 and −5, while the data above the 95th percentile lies between 101 and 1053 (pertinent values shown in bold); accordingly, a 90% winsorization would result in the following:

{92, 19, **101**, 58, **101**, 91, 26, 78, 10, 13, **−5**, 101, 86, 85, 15, 89, 89, 28, **−5**, 41} (N = 20, mean = 55.65)

Sweet and Martin (2012) indicated that outliers should be deleted if they a) are faulty, or b) they are some special cases isolated from a common phenomenon in the analysis. Otherwise, model the data with and without the outliers. If the outliers have little influence on the regression results, the models should coincide. This method is an easy and quick way of dealing with unusual cases, and is especially necessary if the unusual cases are faulty data.

However, if the unusual cases are valid, removing them can be a waste of good data, thus reducing the information gained from the dataset. Removal of outliers can reduce important information about the variability of data. Data variability refers to how spread out a dataset is.

Removing outliers greatly reduces data variability, as the outliers which lie at the extremes of the dataset, are essentially being pushed in to the rest of the data.

In 1991, John M. Orr et al. conducted a survey to identify what methods senior authors believed were the best way to deal with outliers. Their respondents were the senior authors of all published papers using correlation or regression in the *Journal of Applied Psychology and Personnel Psychology* from 1984 to 1987. Only 4% of respondents believed it is best to remove the outliers from the data, regardless of why they are outliers. Sixty-seven percent of the respondents believed that outliers should only be removed if there is an identifiable reason to believe they are invalid. Twenty-nine percent believed outliers should be left in the data, regardless of why they are outliers. The results from this survey indicate it is best to not remove valid outlier data, and it is advisable to identify if the outliers are faulty or valid data. However, there is not a clear consensus in the author community (e.g. 67% vs. 29%) on whether non-faulty outliers should be removed or not.

Robust regression is a relevant alternative to deal with this data. Robust regression finds a good compromise between excluding these cases entirely from the analysis and including all the data cases and treating all of them equally in OLS regression *(Bruin, 2006).* Robust regression weighs the importance of each case based on how unusual the case is. Therefore, unusual cases will have less effect on the model than the rest of the cases. This reduces the effect unusual cases have on the parameter estimates of the regression model. This is a huge advantage over OLS, where one unusual case can drastically influence the parameter estimates *(Yu, Yao, & Bai, 2014).*

Another advantage is robust regression follows less restrictive assumptions than OLS regression (*Fox & Weisburg 2013*). When the data does not follow an assumption, results from the OLS model cannot be relied on. One OLS assumption is that the residuals are normally distributed. Outliers can disobey this assumption. They tend to distort the regression coefficients by having more influence than they deserve. *(https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Robust_Regression.pdf)*

OLS regression assumes *homoskedasticity* in a dataset. Homoskedasticity refers to the phenomenon that different cases in the data set have the same residual variances. Different values have the same variances in the dataset. An example where this is broken: lower income people would have lower variance in expenditure than higher income people. The breaking of the *homoskedasticity* assumption affects the OLS model. Although the OLS estimator remains unbiased, the estimated standard error (SE) becomes wrong. Because of this, statistical tests using an OLS model could not be relied on, such as hypothesis tests. Hypothesis tests depend on an accurate SE.
(http://www3.wabash.edu/econometrics/EconometricsBook/chap19.htm) (Yellowlees, Bursa and Fleetwood)

Robust regression does not assume the above assumptions, and so is unaffected in those situations. The less restrictive assumptions of robust regression make it more reliable in estimating coefficients.

The robust methods examined in this project consist of LAD estimation, M-estimation, MM-estimation, and GM-estimation. It will be noticed that most of the robust methods appear to only downweigh outliers, while GM-estimation downweighs all unusual cases. In 2015, Ayinde et al. performed a diagnostics of influential cases in linear regression models. They identified both outliers and influential cases in the models, and noted that not all influential cases are typically outliers. Ayinde et al. suggested that multicollinearity may be the cause of those influential cases. One shortcoming of this paper, is they did not identify leverage cases. Leverage cases are very important in diagnostics and can be the reason a case is influential.

The robust methods examined in this project consist of LAD estimation, M-estimation, MM-estimation, and GM-estimation. It will be noticed that most of the robust methods appear to only downweigh outliers, while GM-estimation downweighs all unusual cases. In 2016, Yellowlees et al. performed an experiment to analyse the appropriateness of robust regression in addressing outliers in an anthrax vaccine potency test. OLS regression, M-estimation and MM-estimation were all performed on a dataset and results were compared with each other. The dataset consisted of 32 immunopotency assays (IPA), which measures the relative potency of a final drug product. The results showed that robust estimation is more effective and less biased when there are outliers present. Additionally, the authors found that robust regression provides similar results to OLS when there are no outliers present but is less efficient. One shortcoming of this experiment is that the effect of leverage cases on regression methods was not analysed. Leverage cases can be detrimental to regression, as well as outliers.

2 Data

This section describes the survey data set used for conducting the analysis. The data set consists of UCC students' survey responses. The data set of survey responses was provided by Student Experience Technical Working Group (TWG). An email was sent to 13,846 UCC undergraduate students with an attached link to questionnaire in January 2013, was used as the primary data to conduct the analysis. The questionnaire consisted of 44 questions containing 123 items on demographic details (9 questions, 9 items), academic life (11 questions, 66 items), social life (8 questions, 8 items), financial aspects (6 questions, 7 items), student support services (6 questions, 29 items), and general comments (4 questions, 4 items). The survey remained open for four weeks.

Over this period a total of 3,462 undergraduate student participated in the survey. The cases which contain any missing (NA) answers are filtered out of the dataset, leaving 1485 survey responses to analyse. One subject was removed from the study, for answering 'Not Applicable' to V16A. One-hundred-ninety-six subjects were removed for answering 'Not Applicable' to

both V33A and V33B. Overall, there are **1288** survey responses to analyse (n=1288). There were 20 of the possible 44 variables (questions), in the data set received. These variables are given in Table 2.1, organised by variable group. Two pairs of questions were combined into single individual variables, leaving **18** independent variables altogether (k=18+1=19).

Each variable group is explained in more detail in sections 2.1 - 2.5. The method of obtaining these variables and the categories for each variable are also explained. The chosen variables are as shown in Table 2.1.

Variable group	Variable name	Variable Type	Labels / Range
Dependent Variable	Satisfaction with University Life (SWUL)	Numerical	[5, 35]
Demographic Measures	Gender (V1)	Categorical	(Male, Female)
	What college department is the student in (V4)	Categorical	4 categories
	Present Year of study (V7)	Categorical	(1, 2, 3, 4, 5)
	Through what process did you enter UCC (V8)	Categorical	4 categories
Academic Measures	Commitment level to study (V12)	Ordinal	5 point Likert scale
	How accommodating the Lecturer was score (V16)	numerical	[1, 5]
	Motivation Scale (MS)	numerical	[-100, 100]
	Appropriate Workload Scale (AWS)	numerical	[-100, 100]
	Good Teaching Scale (GTS)	numerical	[-100, 100]
Social Measures	How socially integrated do you feel in UCC (V22)	Ordinal	5 point Likert scale
	Do you take part in Club activities in UCC (V23)	Ordinal	4 point Likert scale
	Do you take part in Societies activities in UCC (V25)	Ordinal	4 point Likert scale
	Do you take part in Club activities outside of UCC (V27)	Ordinal	4 point Likert scale

***Table 2.1 (a)**: Variables included in the model*

Variable group	Variable name	Variable Type	Labels / Range
Social Measures	Do you take part in Community-based Organisations outside of UCC (V28)	Ordinal	4 point Likert scale
Financial Measures	I have enough money to meet my needs (V30)	Ordinal	5 point Likert scale
	Lecture Availability (V33)	numerical	[1, 5]
	How much paid work you do per week during term time on weekdays (V34A)	Ordinal	5 point Likert scale
	How much paid work you do per week during term time on weekends (V34B)	Ordinal	5 point Likert scale

***Table 2.1 (b)**: Variables included in the model*

2.1 Satisfaction with University Life (SWUL)

This scale measures a student's overall satisfaction with university life, how much the student enjoyed their university life. This scale is determined by the response (Likert scale) they give to five different statements, regarding their university life. Each statement was measured on a 7 point Likert scale. Each point on the scale are defined: (1: strongly disagree, 2: disagree, 3: slightly disagree, 4: neither agree nor disagree, 5: slightly agree, 6: agree, 7: strongly agree).

Statement
In most ways my life in university is close to my ideal
The conditions of my life in university are excellent
I am satisfied with my life in the university
So far I have gotten the important things I want in my life in university
If I had my time in university again I would change almost nothing

***Table 2.2**: SWUL statements*

For a student, their responses for each statement are added and this is the SWUL of that student with range [5, 35]. A SWUL score of 5 means a student is completely unsatisfied with university life, while 35 is completely satisfied. A higher score corresponds to a higher satisfaction and vice versa. SWUL scores are categorised as shown in Table 2.3.

SWUL Score range	Category name
5-9	Extremely dissatisfied
10-14	Dissatisfied

***Table 2.3 (a)**: Categories of SWUL scores*

SWUL Score range	Category name
15-19	Slightly dissatisfied
20	Neutral
21-25	Slightly satisfied
26-30	Satisfied
31-35	Extremely Satisfied

***Table 2.3 (b)**: Categories of SWUL scores*

2.2 Demographic Measures

The student's gender, nationality, whether English is first language, college of study, study status, year of study and process of entry make up the demographic measures. The categories for each variable in the dataset are given in Table 2.3:

Variable	Categories
Gender (V1)	Male Female
What college department is the student in (V4)	College of Arts, Celtic Studies & Social Sciences College of Business & Law College of Medicine & Health College of Science, Engineering & Food Science
Present Year of study (V7)	First Second Third Fourth Fifth
Through what process did you enter UCC (V8)	CAO School Leaving qualifications Medical student process (combination of 2 options) CAO Mature years Other Routes (combination of 5 options)

***Table 2.4**: Demographic variables and their categories*

In the survey, there were seven options to choose from for V4. These were (*Junior Year Abroad (JYA), Erasmus, College of Arts, Celtic Studies & Social Sciences, College of Business & Law, College of Medicine & Health, College of Science, Engineering & Food Science, Centre for Adult and Continuing Education*). For this project, only students from any of the four colleges, shown in *Table 2.3*, were included in the final dataset.

In the survey, V8 had nine options to choose from. The category '*Medical student process'* is a combination of two options: (*CAO School Leaving qualifications and HPAT (Medicine),*

UCC Diploma in Dental Hygiene). Students that chose either of these options were set in the '*Medical student process'* category.

The category '*Other Routes'* is a combination of five options: (*CAO HEAR offer, CAO DARE offer, CAO FETAC route, CAO Degree and GAMSAT (Graduate Entry to Medicine), UCC Non-EU Degree*).

2.3 Academic Measures

The academic measures contain all measures related to the perceptions of a student of the educational aspects of university. The Motivation Scale (MS), Appropriate Workload Scale (AWS) and Good Teaching Scale (GTS) are all numerical values in the range [-100, 100], and the rest of the variables are measured in a 5 point Likert scale.

MS, AWS and GTS are scales which measure the student's perceptions of motivation, academic workload and good teaching respectively. These scales were formulated by combining the responses from the items in the survey.

The MS measures the extent to which students feels motivated by their programme of study. This scale is determined by a 5 point Likert score given to nine different statements, regarding their university life. The Likert scale is defined: (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). The different statements are given in Table 2.4.

Statement
My programme of study is intellectually stimulating
I find my programme of study motivating
My programme of study has stimulated my enthusiasm for further learning
My programme of study has stimulated my interest in the field
I feel that I benefit from being in contact with active researchers/scholars
The academic expectations of me on my programme of study are too high
Intellectual standards at UCC are set too high
I feel part of a community of scholars who are committed to learning
Being selected to study at UCC is a source of motivation to me

***Table 2.5**: MS statements*

Each statement is recalibrated to a -100 to +100 scale with -100 = strongly disagree, -50 = disagree, 0 = uncertain, 50 = agree, 100 = strongly agree. A score above 0 indicates a positive response. Negatively phrased items are reversed scored so that items which infer a negative attitude are reversed scored so that disagreement, which infers a positive attitude, receives a positive score. A respondent's scale score was found by averaging over the items. This was computed for the respondents who answered all items.

The AWS measures the extent to which students perceive that the workload given is manageable. This scale is determined by a 5 point Likert score they give to four different statements, regarding their university life. The Likert scale is defined: (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). The different statements are given in Table 2.5.

Statement
There is a lot of unwanted academic pressure on me as a student
My academic workload is too heavy
I am generally given enough time to understand things I have learned.
The volume of work necessary for my programme of study means that it cannot all be thoroughly comprehended

***Table 2.6**: AWS statements*

After reverse scoring any negatively phrased items, the four item response scores were rescaled to a -100 to +100 range and averaged similar to the GTS score. This was computed for all respondents who answered all items.

The GTS refers to how the students perceive the lecturers' ability to contribute to student learning. This scale is determined by the Likert score they give to six different statements. The Likert scale is defined: (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). The different statements are given in Table 2.6.

Statement
My lecturers normally give me helpful feedback on my progress
My lecturers in this programme motivate me to do my best work
My lecturers make a real effort to understand any difficulties I have
My lecturers are extremely good at explaining things
My lecturers work hard to make their subjects interesting
My lecturers put a lot of time into comments (orally and/or in writing) on my work

***Table 2.7**: GTS statements*

V16 is a variable composed by combining two questions from the survey, shown in Table 2.7:

Variable name	Variable Type	Labels / Range
To what extent do you feel lecturers are available to students? (V16A)	Ordinal	5 point Likert scale or 'Not Applicable'

***Table 2.8 (a)**: V16A and V16B variables. V16 is composed from these two variables.*

Variable name	Variable Type	Labels / Range
To what extent do you feel the schedule of deadlines for assignments are well balanced? (V16B)	Ordinal	5 point Likert scale or 'Not Applicable'

***Table 2.8 (b)**: V16A and V16B variables. V16 is composed from these two variables.*

One student answered 'Not Applicable' to V16A, and was removed from the study. Students that answered 'Not Applicable' to V16B were left in the study.

The answers from V16A and V16B were converted to numerical form, with *Very much so = 1 … Not at all = 5, Not Applicable = NA.* The mean of the respondents answers to V16A and V16B, was set as their response to V16. NA's were stripped before computing the mean (mean of 3 and NA = 3).

2.4 Social Measures

Social measures quantify the satisfaction of student with social aspects of life in university. These measure their involvement with clubs and societies in college. The options of V22 are a 5 point Likert scale. The Likert scale is defined as: (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). The options of V23, V25, V27, V28 are a 4 point Likert scale. The Likert scale is defined as: (1 = No, 2 = Rarely, 3 = Often, 4 = Very Often).

2.5 Financial Measures

Financial measures consist of whether students have enough money, whether students worry about debt, how commitments impact on studies and paid employment. These questions were measured on Likert scales of differing lengths.

Students were asked if they missed scheduled academic classes due to wok and caring commitments. These were captured in V33. V33 is composed by combining two questions from the survey, shown in Table 2.8:

Variable name	Variable Type	Labels / Range
How often do you miss lectures or other scheduled study obligations because of work commitments? (V33A)	Ordinal	5 point Likert scale or 'Not Applicable'
How often do you miss lectures or other scheduled study obligations because of caring for dependents? (V33B)	Ordinal	5 point Likert scale or 'Not Applicable'

***Table 2.9**: V33A and V33B variables*

The 5 point Likert scale in Table 2.8 is defined as: (1 = Always, 2 = Usually, 3 = Sometimes, 4 = Rarely, and 5 = Never).

Students that answered 'Not Applicable' to both V33A and V33B were removed from the study. The answers from V33A and V33B were converted to numerical form, with *Always = 1 … Never= 5, Not Applicable = NA.* The mean of the respondents answers to V33A and V33B, was set as their answer to V33. NA's were stripped before computing the mean. For example, if a student (Osborne and Overbay) answered 'Never' to V33A and 'Not Applicable' to V33B, their answers were converted to 5 and NA respectively. The mean is calculated as 5 (mean of 5 and NA = 5), and set as their answer to V33.

3 Methods

This sections describes the statistical methods applied in this project. First, methods to identify outliers, high leverage cases and influential cases will be described. We will also discuss the robust methods to fit a regression model to the data.

3.1 OLS regression

Linear regression is an approach to model the relationship between a dependent variable *Y* and one or more independent variables denoted *X (Susani et al., 2014).* In linear regression, predictor functions are applied to the data and model coefficients are estimated. Numerical variables are encoded as is, while categorical variables are transformed into dummy variables for to compute the OLS regression model (Hutcheson, 2011). For the Gender variable, a new dummy variable, β_{Male}, is created that takes the value:

- $1\ if\ case\ is\ male$
- $0\ if\ case\ is\ female$

For a multi-level variable with *m* categories, *m-1* dummy variables are created. Either one or none of the dummy variables are set to one, while the rest are set to zero. For the "What college department is the student in (V4)" variable in the survey dataset, there are four categories. Three dummy variable are created as a result. These variables are defined as:

- $\beta_{Arts} = 1\ if\ case\ is\ in$ College of Arts, Celtic Studies & Social Sciences,
  $0\ otherwise$
- $\beta_{Busines} = 1\ if\ case\ is\ in$ College of Business & Law,
  $0\ otherwise$
- $\beta_{Medicine} = 1\ if\ case\ is\ in$ College of Medicine & Health,
  $0\ otherwise$

Note that if the case is in the College of Science, Engineering & Food Science, then all three variables above are set with the value 0.

A regression model with p variables is expressed as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \mathrm{e}_i, \qquad i = 1,2,\cdots,n.$$

Y_i is the dependent variable on the i^{th} case, $\beta_0, \beta_1, \cdots, \beta_p$ are coefficients, X_i is the independent variable value or dummy variable on the i^{th} case, and ϵ_i is a normally distributed random variable. The error $\epsilon_i \sim N(0, \sigma^2)$ is not mutually correlated.

Ordinary Least Squares (OLS) regression is the most common type of linear regression, because it is one of the simplest forms of regression. It is one of the earliest regression models, proposed by Legendre in 1805. The β_p parameters in the OLS model are chosen that minimise the *objective function,* ρ, which is the squared residuals in this case:

$$\min \sum_{i=1}^{n} \rho(\mathrm{e_i}) = \min \sum_{i=1}^{n} \mathrm{e}_i^2$$

If we were to create an OLS model on our dataset using only GTS_SCORE and Gender as the independent variables and SWUL as the dependent variable, the model is expressed as:

$$SWUL = \beta_0 + \beta_1 GTS\_SCORE_i + \beta_2 Gender_i + \mathrm{e}_i, i = 1,2,\cdots,n.$$

For male student with a GTS_SCORE of 40, their SWUL is calculate as:

$$\mathrm{SWUL} = \widehat{\beta_0} + \widehat{\beta_1} * 40 + \widehat{\beta_2} * 1$$

Where $\widehat{\beta_p}$ is the estimate of variable p in the OLS model.

Assessment of model

The accuracy and fit of the model can be assessed using the errors obtained from $\mathrm{e}_i = Y_i - \widehat{Y_\iota}$, where $\widehat{Y_\iota}$ is the fitted dependent variable of the i^{th} case. Two assessment methods will be used to assess the fit and accuracy of the models, the residual standard error (RSE) and mean squared error (MSE). The time taken by the model to fit the full dataset will also be taken into account.

RSE is the standard deviations of the residuals. It is represented as:

$$RSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{df}}$$

Where df is the degrees of freedom. In this case $df = n - k - 1$, where n is the total number of cases, and k is the number of variables in the model.

Smaller RSE means the predictions are better, and so the model is a better fit to the data. The RSE is particularly useful for comparing models, as the RSE is standardised. This means the RSE of two different models are on the same scale, and thus can be compared with each other.

For the MSE, the dataset will be split into a training dataset, and a test dataset. Seventy percent of the data will be the training data, and 30% will be the test data. All the unusual cases are part of the training dataset, and all the test data are non-unusual cases. This is done so we can observe the effect that the unusual cases have on the fit of the model. If the model is robust to the unusual cases in the training dataset, it should predict non-unusual cases in the test dataset very well, and vice-versa if not robust (Yellowlees et al., 2016). The model will be fitted to the training data. The fitted model will then be used to predict the SWUL variable in the test data. The MSE, obtained from the predictions, will be used to assess the model. It is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Smaller values of MSE indicate better prediction accuracy.

3.2 Unusual case identification

Here, methods to identify outliers, high leverage cases and influential cases will be discussed. We will assess the OLS with and without the unusual cases, and compare the assessments. We will examine how the robust regression methods deal with these unusual cases.

3.2.1 Outlier Detection

An outlier is a case whose dependant-variable value greatly deviates from the rest of the data group, given its value on the predictor values. In other words, it has a very high residual. *(https://stats.oarc.ucla.edu/r/dae/robust-regression/)*

There are difficulties with outlier detection. One difficulty is *swamping* (Type I errors), the mislabelling of non-outliers as outliers. (*Farne & Vouldis, 2018*). Another difficulty is *masking* (Type II errors), the mislabelling of outliers as non-outliers. This happens when multiple outliers in a data set conceal the presence of additional outliers. Swamping and masking make a dataset more complicated to diagnose and may cause an analyst to reject non-unusual case, or leave unusual cases in the dataset. For example, in the case of masking, let us assume case *i*, an outlier, masks case *j*, another outlier. Case *i* is observed as an outlier, but case *j* is not. Case *i* would be dealt with, but case *j* goes unnoticed and continues to negatively affect model fitting.

Several authors have suggested methods to overcome masking, including Atkinson (1986) and Hadi (1992), but these methods typically involve removing the 'masking' outliers from the data set before the 'masked' outliers can be identified.

Standardised residuals

Chatterjee and Hadi (1986) discuss that unmodified residuals are not appropriate for diagnosing outliers. The average residual value in one dataset may be completely different to the average residual value in another. For example, the residual on the height of an ant would be a few micrometres, while the residual of the height of a human would be a few centimetres. One would need to have different cut-off values in every dataset to identify their respective outliers. Scaling the residuals to a standardised form makes it easier to identify outliers, as the same cut-off value can be used across different datasets. Standardised and studentised residuals are different forms of scaling the residuals in a model.

The standardised residual is simply the residuals divided by their estimated standard deviation.

$$r_i = \frac{\mathrm{e_i}}{\sqrt{\mathrm{SD_{(RES)}}}} \quad i = 1, \dots, n$$

Where $\mathrm{SD_{(RES)}}$ is the standard deviation of the residuals. The standardised residuals are approximately standard normally distributed with a mean of 0 and a variance of 1.

The studentised residual is defined as:

$$r_i = \frac{\mathrm{e_i}}{\sigma\sqrt{1 - \mathrm{h_{ii}}}} \quad \mathrm{i} = 1, \dots, \mathrm{n}$$

Where h_{ii} is the hat diagonal value (explained further in section 3.1.2). The studentised residual approximately follows a t distribution with $n - p - 1$. They have a mean of zero and variance slightly above one.

For this project, we will use standardised residuals and ignore the studentised residual. The studentised residual give similar results for large datasets *(Vittinghoff et al., 2005)*.

Standardised residuals are more useful than unmodified residuals, as unmodified residuals may have non-constant variance among different cases. The standardised residual makes all cases have constant variance, and thus simplifies the process of finding outliers.

The general consensus is to use 3 as the threshold for the absolute standardised residual, for a case to be considered an outlier, as cases above this cut-off are quite clearly outliers *(https://online.stat.psu.edu/stat462/node/172/)*. Assuming the residuals follow a normal distribution, around 1% cases will be identified as an outlier.

Grubbs' test

Grubbs' test is a hypothesis test used to determine if the case with the most extreme residual should be considered an outlier in the dataset.

Grubbs' test is used to check if an outlier exists in a dataset that approximately follows a normal distribution. We can assume that the residuals of our dataset approximately follow a normal distribution.

Grubbs' test is defined as a two-sided hypothesis test as follows:

H_0: There are no outliers in the data set
H_a: The maximum absolute residual case is an outlier

Grubbs' test statistic is defined as:

$$G = \frac{max\left|\mathrm{Y_i} - \overline{\mathrm{Y}}\right|}{\sigma}$$

with $\overline{Y}$ and σ denoting the sample mean and standard deviation of the dependent variable, respectively. Grubbs' test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation.
The hypothesis of no outliers is rejected if:

$$G > \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{\left(t_{\alpha/(2n),n-2}\right)^2}{n-2+\left(t_{\alpha/(2n),n-2}\right)^2}}$$

With α denoting the significance level; α will be set to 0.05 for the experiment. $t_{\alpha/(2n),n-2}$ denotes the critical value of the t distribution with $(n-2)$ degrees of freedom and a significance level of $\alpha/(2n)$. If the hypothesis is rejected, it means there is at least one outlier that does not follow the data distribution *(Grubbs, 1969)*. In other words, the most extreme residual is an outlier. *(https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h1.htm)*

The disadvantage of Grubbs' test is that it only finds one outlier, and ignores the rest of the outliers in the data.

To overcome this issue, we will recursively apply the Grubbs' test. The diagnosis will involve applying the Grubbs' test and performing one of the below steps, depending on the result of the Grubbs' test:

1. If the null hypothesis is rejected, record the case with the highest absolute residual as an outlier. Remove that case from the data and apply Grubbs' test again.
2. If the null hypothesis cannot be rejected, end the diagnosis.

This method is a good way to prevent masking of cases, as the largest outlier is always removed, thus removing the masking from other cases. Swamping does not affect the outliers, because if the most extreme residual is not an outlier, the less extreme residuals cannot be outliers *(Grubbs, 1969)*.

3.2.2 Leverage case Detection

A case with high leverage has extreme independent variables. Leverage is a measure of how far the independent variable values of a case are from the other cases. High leverage cases greatly affect the estimation of regression coefficients in OLS models.

Hat Diagonal

Also known as leverage values, hat diagonal values measure the extent to which cases are outliers in the space of independent variables *(Aguinis, 2013)*. This is independent of the dependent variable values.

The hat matrix is defined as:

$$H = X(X^T X)^{-1} X^T$$

Where X is the matrix of the independent variables in the regression model.

The i^{th} diagonal entry, h_{ii}, gives the sum of squared entries in its i^{th} row or column. This is the hat diagonal value. h_{ii} describes the contribution that case i has to the fitted regression coefficients. Therefore, it serves as a suitable measure of leverage *(John Fox, 2015)*.

The average value of h_{ii} is p/n, where p is the number of variables, and n is the number of rows. Belsley, Kuh, and Welsch (1980) propose a hat diagonal cut-off value of

$$h_{ii} > 2p/n$$

Every case i with hat diagonal larger than that cut-off value is considered a leverage case.

Mahalanobis Distance

Mahalanobis distance (MD) is defined as the distance between a case and the centroid of all cases altogether, where the centroid is the point created at the intersection of the means of all the predictor variables *(Aguinis, 2013)*.

$$MD_i^2 = (x_i - \bar{x})^T C^{-1} (x_i - \bar{x})$$

Where x_i is the independent values for the i^{th} case, $\bar{x}$ is the mean independent values, and C is the covariance matrix.

MD is closely related to the hat diagonal statistic, but is scaled differently according to the formula:

$$MD_i^2 = (n-1)\left(h_{ii} - \frac{1}{n}\right)$$

A large MD indicates a case has high leverage. Becker & Gather (1999) recommend the MD cut-off for large sample datasets to be $\chi^2_{df=p;alpha=\alpha/n}$, where p is the number of variables, χ^2 = critical value in a chi square distribution, and α/n is the significance level. This is based on the asymptotic distribution of the MD. α is most commonly set to 0.05.

3.2.3 High Influence Detection

In simple terms, an influential case is a case with a high combination of residual and leverage *(https://stats.oarc.ucla.edu/r/dae/robust -regression)*. The removal of this case substantially changes the estimation of regression coefficients.

Cook's Distance (CD)

CD is the scaled change in fitted coefficients, when a case is removed. This shows the influence each case has on the coefficients. CD is the most well-known influence measure. A combination of the residual and leverage measures are used to determine the CD *(Dhakal, 2017)*. CD for case i is defined as:

$$D_i = \frac{\sum_{j=1}^{n}\left(\widehat{Y_j} - \widehat{Y_{j(i)}}\right)^2}{p\widehat{\sigma^2}}$$

Where p is the number of coefficients, $\widehat{Y_j}$ is the j^{th} fitted coefficient value, $Y_{j(i)}$ is the j^{th} fitted coefficient value when case i is removed, σ is the standard deviation.

The higher the leverage and residuals, the higher the influence is, and therefore, the higher the CD *(Andale, 2016)*. The most common cut-off value to use for CD is $4/n$. An alternative cut-off value is the F distribution with $df = (p, n-p)$ and α=0.50, where p is the number of parameters in the model, and n is the number of cases. Cases with CD above the cut-off are considered influential. The F distribution cut-off will be used for this project.

Difference in Fits (DFFITS)

DFFITS is a measure of how the predicted value at the i^{th} case changes when the i^{th} case is deleted. DFFITS is the standardised version of DFFIT. It is defined as:

$$DFFITS_i = \frac{\widehat{Y_i} - \widehat{Y_{i(i)}}}{s_{(i)}\sqrt{h_{ii}}}$$

Where $\widehat{Y_i}$ is the predicted dependent value for case i, $\widehat{Y_{i(i)}}$ is the predicted dependent value for case i with case i excluded from the fitted model, $s_{(i)}$ is the estimated standard deviation without case i, and h_{ii} is the leverage for case i.

The numerator measures the difference in the predicted dependent value obtained when the i^{th} case is included and excluded from the fitted model. The denominator is the estimated standard deviation of the difference in the predicted responses.

The DFFITS formula quantifies the number of standard deviations the fitted dependent value changes when the i^{th} case is removed.

The regular recommended cut-off for DFFITS is 2. The size-adjusted version of the cut-off is $2\sqrt{\frac{p}{n}}$. A case *i* is deemed influential if $|DFFITS| > 2\sqrt{\frac{p}{n}}$ *(Ayinde & Arowolo, 2015).*

Standardized DFBETA (DFBETAS)

DFBETAS is different to the other measures of influence discussed so far, because the ones discussed so far are global measures of influence, they indicate the influence each case has on the regression as a whole. DFBETAS, on the other hand, measures how much the inclusion of a case increases or decreases each individual regression coefficient. Because the other measures of influence only measure influence as a whole, they may neglect cases that are influential on one or two regression coefficients *(Aguinis, 2013).* Therefore it is important to always include DFBETAS in an analysis in addition to global measures of influence.

DFBETAS for the j^{th} coefficient of the i^{th} case is defined as:

$$DFBETAS_{i,j} = \frac{b_j - b_{(i)j}}{s_{(i)}\sqrt{(X'X)^{-1}{}_{jj}}}$$

Where b_j is the j^{th} coefficient, $b_{(i)j}$ is the j^{th} coefficient without the i^{th} case, $s_{(i)}$ is the estimated standard deviation when case i is excluded from the fitted model, X is the matrix of the independent variables in the fitted model, $(X'X)^{-1}{}_{jj}$ is the $(j,j)^{th}$ element of $(X'X)^{-1}$.

The numerator measures the difference in the predicted coefficients obtained when the i^{th} case is included and excluded from the fitted model. The denominator is the estimated standard deviation of the difference in the predicted responses.

The DFBETAS formula quantifies the number of standard deviations the fitted coefficients changes when the i^{th} case is removed.

In general, large values of $DFBETAS_{i,j}$ indicate the i^{th} case is influential in estimating the j^{th} coefficient. *Belsley, Kuh, and Welsch (1980)* recommend 2 as a general cut-off value to indicate influential cases and $\frac{2}{\sqrt{n}}$ as a size-adjusted cut-off.

COVRATIO

Effects on the precision of regression coefficients is measured by the COVRATIO statistic (*Belsley, Kuh, & Welsch, 1980).* The precision of an estimated regression coefficient is how close it is to its true value. The lower the SE of a coefficient, the higher its precision.

COVRATIO measures the change in the determinant of the covariance matrix of the estimated coefficients after deleting the i^{th} case.

The COVRATIO is defined as:

$$COVRATIO_i = \frac{det\left(s_{(i)}^2\left(X_{(i)}^T X_{(i)}\right)^{-1}\right)}{det(s^2(X^T X)^{-1})}$$

Where s is the estimated standard deviation, $s_{(i)}$ is the estimated standard deviation when the i^{th} case is excluded, X is the matrix of the independent variables in the fitted model, $X_{(i)}$ is the matrix of the independent variables in the fitted model without the i^{th} case.

Leverage cases have large COVRATIO (they improve precision), while outliers have small COVRATIO (they degrade precision). Removing outliers increase precision of a regression model, as the SE of coefficients become lower. Because there is lower overall residuals without outliers, the model is more certain its coefficients are correct.

Belsley, Kuh, and Welsch (1980) suggest that cases with

$$|COVRATIO - 1| \geq \frac{3p}{n}$$

are influential cases and are thus worth investigating further.

This cut-off interval is too small when used in our dataset for this project, as too many cases are identified as influential. We will increase the cut-off point to $\frac{7p}{n}$ for this project to solve this.

3.3 Robust regression

Robust regression is a type of regression created to overcome the limitations of OLS regression. Unlike OLS, robust regression is insensitive to outliers and is designed to not be overly affected

by violations of assumptions in the underlying data. Robust regression is an iterative process that seeks to identify outliers and minimise their impact on coefficient estimates.

The amount of weight assigned to each case is determined by a special curve called the *influence function (IF)*. *(NCSS Statistical Software, Chapter 308).* The IF describes the effect the unusual cases have on the regression estimator. In a robust model, the IF must be bounded and its maximum describes the sensitivity of the estimator. Robust methods are generally grouped by whether their IF is bounded on the x-axis (independent variables), the y-axis (dependent variable), or both. If a method is bounded on the y-axis, it means it is resistant to unusual y-axis values (outliers).

There exists a few statistical techniques to measure the robustness of data. The *breakdown point (BP)* of an estimator is the largest fraction of the data that can be given arbitrary values without perturbing the estimator to the boundary of the parameter space. Thus, the higher the BP, the more robust the estimator is against extreme outliers *(Farcomeni et al., 2010).*

For example, the median of a sample has a BP of 0.5. Fifty percent or more of the sample would need to be outliers to significantly affect the median. If 49% of the values are outliers, the median remains unaffected. Meanwhile, the mean has a BP of 0, since only one outlier can throw the mean out of place. The median is the better estimator in this case.

There are many types of robust regression models, i.e., LAD, M-Estimation, MM-Estimation, S-Estimation, GM-Estimation. They work in different ways, but they all give less weight to outliers, that would otherwise influence the regression coefficients.

Least Absolute Deviation (LAD) regression

LAD is one of the earliest regression methods developed. It was introduced in 1757 by Roger Joseph Boscovich, around 50 years before the OLS method. LAD was created to be less sensitive to outliers than OLS regression. While OLS minimises $\sum_{i=1}^{n} e_i^2$, LAD instead minimises $\sum_{i=1}^{n} |e_i|$. It is somewhat more resistant to outliers since the residuals are not squared. This means the outliers have less weight on the model.

LAD is resistant to outliers, but will still be affected by them if they are extreme enough, as the outliers still have weight; they just have less weight than they would have in an OLS model. Therefore, this method has an IF bounded to the x-axis. LAD typically fares even worse with leverage cases, than OLS regression. Estimated coefficients of LAD fit the leverage cases more than the rest of the data. One high leverage case can throw off the estimate, so it has a BP of $\frac{1}{n} \rightarrow 0\%$, where n is the number of cases.

LAD is more complicated and time consuming to compute than OLS regression, but provides no advantage over OLS regression when there are no outliers present. Therefore, it is preferred to use OLS regression in these situations, over LAD.

Maximum Likelihood (M)-Estimation

In 1964, Peter J. Huber proposed the generalising Maximum Likelihood Estimation (MLE) to minimising the objective function, ρ, of a model, thus introducing a class of robust estimation called M-estimation *(Almetwally & Almongy, 2018).* This is the most common general form of robust regression. The M in M-estimation stands for 'Maximum Likelihood type', named after the fact it is regarded as a generalisation of MLE. M-estimation is said to be more efficient and robust to outliers than LAD. Like LAD, this method's IF is bounded to the x-axis only and has a BP of $\frac{1}{n} \to 0$.

M-estimation regression differs depending on the *estimation method* used. There are several types of estimation methods, defined by their *objective function*, ρ. The *influence curve,* $\psi = \rho'$*,* is the derivative of the objective function. The weight of each case in a robust regression is defined by the *weight function,* $w(e_i) = \psi(e_i)/e_i$.

Four different estimation methods will be examined in this project: *Huber estimator, Tukey bisquare (or biweight) estimator, Hampel estimator,* and *Andrew's Sine estimator.* These are the most well estimation methods.

Huber Estimator

Huber estimator is the first to be created and most commonly used estimation method. Its objective function is defined:

$$\rho(e) = \begin{cases} \frac{1}{2}e^2 & for\ |e| \le k \\ k|e| - \frac{1}{2}k^2 & for\ |e| > k \end{cases}$$

Where k is the *tuning constant.* This can be any value. Smaller values of k produce more resistance to outliers, but at the expense of lower efficiency when the errors are normally distributed. The tuning constant is generally picked to give reasonably high efficiency in the normal case *(Fox et al., 2013).* Usually for Huber estimation, $k = 1.345\sigma$ (where σ is the standard deviation of the residuals). This value produces 95% efficiency when the residuals are normal, and still provides protection against outliers. Normally a robust measure of residual deviation is used in place of σ for robust regression; The two most common measures are median absolute deviation (MAD) and Huber Proposal 2. We will not go into depth on these measures, as they are not needed to understand the robust regression techniques. Farcomeni and Ventura (2010) provide descriptions of MAD and Huber Proposal 2.

The weight function is defined as:

$$w(e) = \begin{cases} 1 & for\ |e| \le k \\ k/|e| & for\ |e| > k \end{cases}$$

Andrew's Sine Estimator

D.F. Andrews first proposed the Andrew's Sine estimator in his 1972 paper. He defines the objective function as:

$$\rho(e) = \begin{cases} k[1 - cos(e/k)], for|e| \leq k \\ 2k \qquad\qquad for\ |e| > k \end{cases}$$

and weight function

$$w(e) = \begin{cases} \dfrac{sin(e/k)}{e/k} & for\ |e| \leq k \\ 0 & for\ |e| > k \end{cases}$$

Young (2017) recommends to have $k = 1.339\sigma$ as this value gives an M-estimator 95% efficiency at the Normal Distribution.

Tukey Bisquare Estimator

The objective function is defined as:

$$\rho(e) = \begin{cases} \dfrac{k^2}{3}[1 - (1 - (e/k)^2)^3], for|e| \leq k \\ 2k \qquad\qquad for\ |e| > k \end{cases}$$

and weight function

$$w(e) = \begin{cases} [1 - (e/k)^2]^2 & for\ |e| \leq k \\ 0 & for\ |e| > k \end{cases}$$

Where $k = 4.685$ *(Ripley, 1992)*.

Hampel Estimator

The objective function is defined as:

$$\rho(e) = \begin{cases} \dfrac{1}{2}e^2 & for\ |e| < a \\ a|e| - \dfrac{1}{2}a^2 & for\ a \leq |e| < b \\ a\dfrac{c|e| - \frac{1}{2}e^2}{c - b} - \dfrac{7a^2}{6} & for\ b < |e| < c \end{cases}$$

and weight function

$$w(e) = \begin{cases} 1 & for\ |e| < a \\ \dfrac{a}{|e|} & for\ a \leq |e| < b \\ a\dfrac{\frac{c}{|e|} - 1}{c - b} & for\ b < |e| < c \end{cases}$$

Where a, b, c are chosen by the user *(Almetwally & Almongy, 2018)*. It can be observed from the objective function, that Hampel Estimator is an updated version of the Huber estimation.

S-Estimation

Yohai created the S-estimate regression in 1984. The goal of S-estimation was to have a high BP, while sharing the nice asymptotic properties of M-estimation. S-estimators are a generalisation of LAD and least trimmed squares (LTS) regressions *(Almetwally, 2018).*

S-estimates are defined by

$$\hat{\beta}_s = \min_{\beta} \hat{\sigma}_s(e_1, e_2, \ldots, e_n)$$

Where $\widehat{\sigma_s}$ is the robust estimator of residual deviation (usually MAD or Huber Proposal 2).

S-estimators have several advantages over M-estimators; S-estimators have a BP of 0.5, while M-estimators have a BP of 0. S-estimators have smaller bias and variance when unusual cases are present *(Rousseeuw & Leroy, 2005).* However, they are very inefficient. They have efficiency of 28%, as opposed to MM-estimation which has 85% efficiency.

Modified Maximum Likelihood (MM)-Estimation

MM-estimation is a special type of M-estimation, created by Yohai in 1987. MM-estimation combines the high resistance of S-estimators to outliers and the high efficiency of M-estimators. MM-estimation is defined by a three stage procedure:

1. Find the initial weights of the cases using S-estimation. Calculate an S-estimate with the objective function:

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2k^2} + \frac{x^6}{6k^4}, & for |x| \le k \\ \frac{x^2}{6} & for\ |x| > k \end{cases}$$

   Where k = 1.548.

2. Calculate the MM parameters that provide the minimum value of $\sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}_0}\right)$, where $\rho(x)$ is the objective function from Stage 1 with k = 4.687, and $\widehat{\sigma_0}$ is the estimate of scale from Stage 1. k = 4.687 is the general consensus to use in S-estimation.
3. Compute the MM estimate of scale as the solution to

$$\frac{1}{n-p} \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = 0.5$$

MM-estimate is one of the best robust regression methods, as it has the advantages of both S-estimation and M-estimation. MM-estimation has a BP of $\frac{n}{2} = 0.5$, and has very high efficiency (around 85%). However, like all the other robust regression methods discussed so far, MM-estimation only penalises outliers but not leverage cases *(Özlem Gürünlü Alma, 2011).*

Generalised Maximum Likelihood (GM)-Estimation

To make M-estimation resistant to outliers, Mallows (1975) proposed Mallows GM-estimate. This estimation method penalises both outliers and leverage cases. However, it would also penalise 'good' leverage cases. These are leverage cases that had low residuals. Handschin et al. (1975) proposed the Schweppe GM-estimate which adjusts the leverage weights according to the size of the residual *(Chun Yu et al., 2014).* However, Carroll and Welsh (1988) proved that the Schweppe estimator is not consistent when the residuals are asymmetric. Also, the BP of Mallows and Schweppe estimators are no more than $\frac{1}{p}$. Coakley and Hettmansperger (1993) proposed Schweppe one-step (S1S) estimate, extending from the Schweppe estimator. S1S is different from Mallows and Schweppe in that it computes the M-estimate in one step rather than iteratively. The S1S estimate uses LTS estimation for the initial estimates of the residuals and least median squares (LMS) regression for initial estimate of scale. In summary, LMS is a regression model which minimises the median of squared residuals. Ortiz et al. (2006) describe LMS in more detail. S1S has a BP of 0.5, much higher than Mallows and Schweppe. It also has an efficiency of 0.95. The Schweppe GM-estimator will be used for this project. The S1S is overall the best option for GM-estimation, for its high efficiency.

4 Results

This section will summarise the findings of the survey data, i.e. ratio of males to females. This section will also describe the findings in this analysis, after using the methods described in Section 3. The outliers, high leverage, and influential cases will be identified, and explained why they are so. We will deem these all as unusual cases. An OLS model will be fitted with and without the unusual cases. Different robust regression models will be fitted with the unusual cases. We will compare the regression coefficients of these models and observe how the robust models deal with the unusual cases. We should expect that the unusual cases will be given less importance in the robust models, and the robust models should have more accurate results.

4.1 Data Summary

To understand the data being dealt with, a summary of the variables must be extracted. In this section, the mean (SD), and quantiles will be formulated for the numerical variables. The percentage of cases in each category will be identified for the categorical variables. As stated in Section 2, there are 1288 complete cases in this dataset. Every case has completed all questions in the survey.

Satisfaction with University Life (SWUL)

The summary statistics of SWUL are shown in Table 4.1. As defined in Section 2, the range of SWUL is [5, 35]. From the summary, the SWUL is generally high, overall. A high proportion of students at UCC are satisfied with their university life. Most students (74.9%) are satisfied.

Variable group	Variable name	Summary Statistics	Quantiles or Percentages (%)
Dependent Variable	Satisfaction with University Life (SWUL)	(1^{st} Qu, Median, 3^{rd} Qu)	(20.0, 25.0, 29.0)
		Percentage of students with SWUL $\geq$ 21 (at least satisfied) (%)	74.9

***Table 4.1**: Summary data in SWUL*

Demographic Measures

The summary statistics of demographic measures are shown in Table 4.2. The demographic measures only consist of categorical data, so we can only see the percentage summaries in each variable. The majority of students are female (34.6% vs. 65.4%). There are less students in the higher years of college. This is expected, because of students dropping out at each year.

Variable group	Variable name	Categories	Percentages (%)
Demographic Measures	Gender (V1)	Male	34.6
		Female	65.4
	What college department is the student in (V4)	College of Arts, Celtic Studies & Social Sciences	31.6
		College of Business & Law	18.9
		College of Medicine & Health	18.6
		College of Science, Engineering & Food Science	30.9
	Present Year of study (V7)	First	32.3
		Second	29.9
		Third	22.0
		Fourth	15.4
		Fifth	0.4
	Through what process did you enter UCC (V8)	CAO School Leaving qualifications	80.3
		Medical student process	4.1
		CAO Mature years	8.9
		Other Routes	6.7

***Table 4.2**: Summary data in Demographic measures*

Academic Measures

The summary statistics of Academic measures are shown in Table 4.3. All the variables here are above average, showing that students are performing well academically. However, the MS

is considerably higher than the AWS and GTS (50 vs. 9.8 and 17). It could be surmised that students feel the workload amount and teaching quality could be improved. Students are perhaps mostly pushed by their own motivation.

Variable group	Variable name	Categories or Summary Statistics	Percentages (%) unless otherwise stated
Academic Measures	Commitment level to study (V12)	Very low	1.1
		Low	5.1
		Average	31.8
		High	43.3
		Very High	18.7
	Lecturer Accommodation (V16)	Mean (SD)	3.4 (0.94)
	Motivation Scale (MS)	(1st Qu, Median, 3rd Qu)	(30.0, 50.0, 70.0)
	Appropriate Workload Scale (AWS)	Mean (SD)	9.8 (37.3)
	Good Teaching Scale (GTS)	Mean (SD)	17.0 (37.9)

***Table 4.3**: Summary data in Academic measures*

Social Measures

The summary statistics of Social measures are shown in Table 4.4. From this, students generally feel socially integrated. However, the number of students not taking part in social or club activities is quite high (~46%). This seems contradictory to each other. Students may feel socially integrated for reasons outside of club and social activities.

Variable group	Variable name	Categories	Percentages (%)
Social Measures	How socially integrated do you feel in UCC (V22)	Very Socially Integrated	15.6
		Socially Integrated	43.9
		Neither Isolated nor Socially Integrated	28.6
		Isolated	9.3
		Very Isolated	2.6
	Do you take part in Club activities in UCC (V23)	No	46.4
		Rarely	29.0
		Often	16.2
		Very Often	8.3

***Table 4.4 (a)**: Summary data in Social measures*

Variable group	Variable name	Categories	Percentages (%)
Social Measures	Do you take part in Societies activities in UCC (V25)	No	40.2
		Rarely	32.5
		Often	16.6
		Very Often	10.6
	Do you take part in Club activities outside of UCC (V27)	No	46.9
		Rarely	16.5
		Often	20.0
		Very Often	16.6
	Do you take part in Community-based Organisations outside of UCC (V28)	No	47.1
		Rarely	22.3
		Often	19.2
		Very Often	11.4

***Table 4.4 (b)**: Summary data in Social measures*

Financial Measures

The summary statistics of Social measures are shown in Table 4.5. Students are more likely to work during the weekends. This makes sense, with university work taking place on weekdays.

Variable group	Variable name	Categories / Summary Statistics	Percentages (%) unless otherwise stated
Financial Measures	I have enough money to meet my needs (V30)	Always	14.0
		Usually	41.1
		Sometimes	23.6
		Rarely	15.3
		Never	6.0
	Lecture Availability (V33)	Mean (SD)	4.33 (0.78)
	How much paid work you do per week during term time on weekdays (V34A)	No Hours	60.5
		Less Than 5 Hours	14.4
		5 to 10 Hours	16.1
		10 to 15 Hours	5.3
		Over 15 Hours	3.7
	How much paid work you do per week during term time on weekends (V34B)	No Hours	39.4
		Less Than 5 Hours	9.7
		5 to 10 Hours	24.5
		10 to 15 Hours	16.3
		Over 15 Hours	10.1

***Table 4.5:** Summary data in Financial measures*

4.2 Unusual Case identification

Outliers, leverage cases, and influential cases will be identified in this section using methods described in Section 3. These will all be labelled as unusual cases. Cases that are a combination of outlier, leverage and influential will be extracted from these unusual cases.

Outliers

Outliers will be identified using standardised residuals and Grubbs' test. Cases that are identified by either method will be classed as an outlier for this project. This lets us be certain that all outliers are found. The plot of standardised residuals is shown in Figure 4.1.

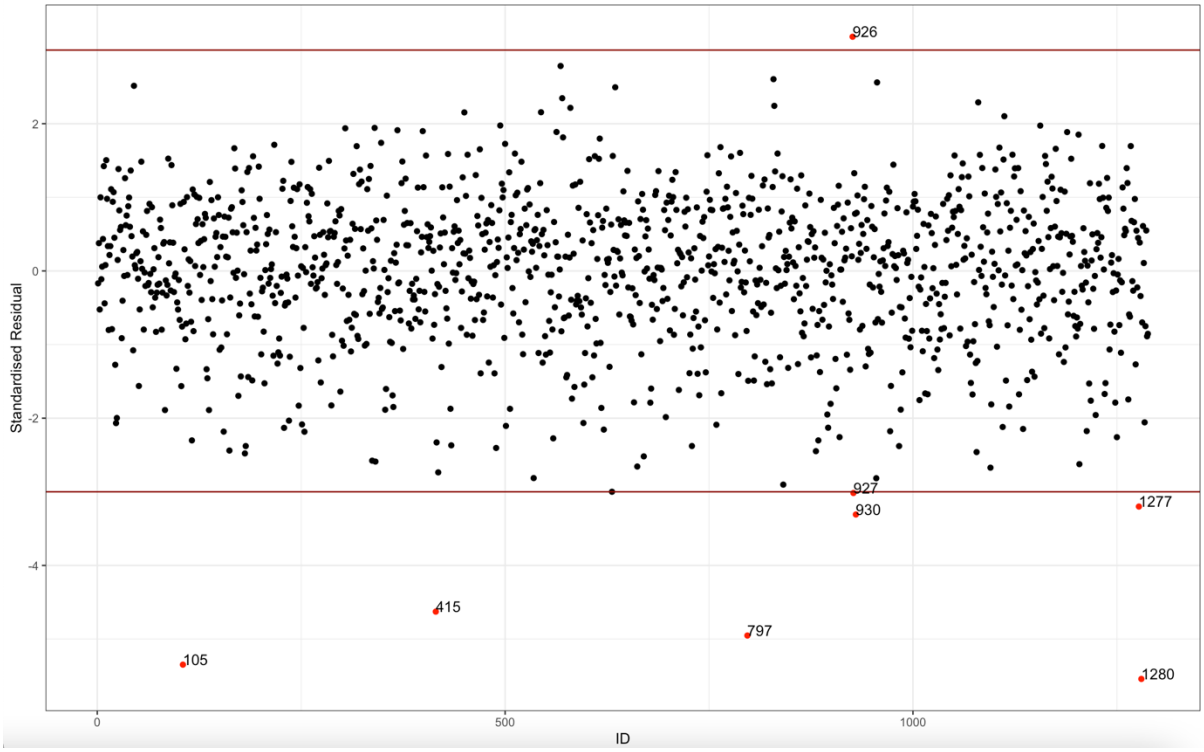

***Figure 4.1**: Plot of cases and their standardised residuals*

The red lines are the cut-offs (-3, 3), used to identify the outliers. The position of each case on the y-axis is their standardised residual. Any case outside the red lines is an outlier. It can be seen that eight cases were identified as outliers. Grubbs' test identified four outliers. Grubbs' test was performed recursively as defined in Section 3. The outliers detected from Grubbs' test are outlined in Table 4.6. The first four loops of the Grubbs' test identify the most extreme residuals as outliers. These are determined because the Grubbs' test p-values in these loops are < 0.05. The case in the fifth loop is not identified as an outlier as the p-value is $>= 0.05$. The test ends because no outliers are found.

Loop #	Most extreme case in terms of residual	Grubbs' test p-value	Case conclusion
1	1280	1.95e-05	Outlier
2	105	2.62e-05	Outlier
3	797	0.0002	Outlier
4	415	0.0004	Outlier
5	930	0.28	Not Outlier

***Table 4.6**: Summary results of recursive Grubbs' test*

Leverage cases

Hat diagonal and MD will be used to identify the leverage cases. The plot of hat diagonals vs. ID is shown in Figure 4.2. The plot of MD vs. ID is shown in Figure 4.3

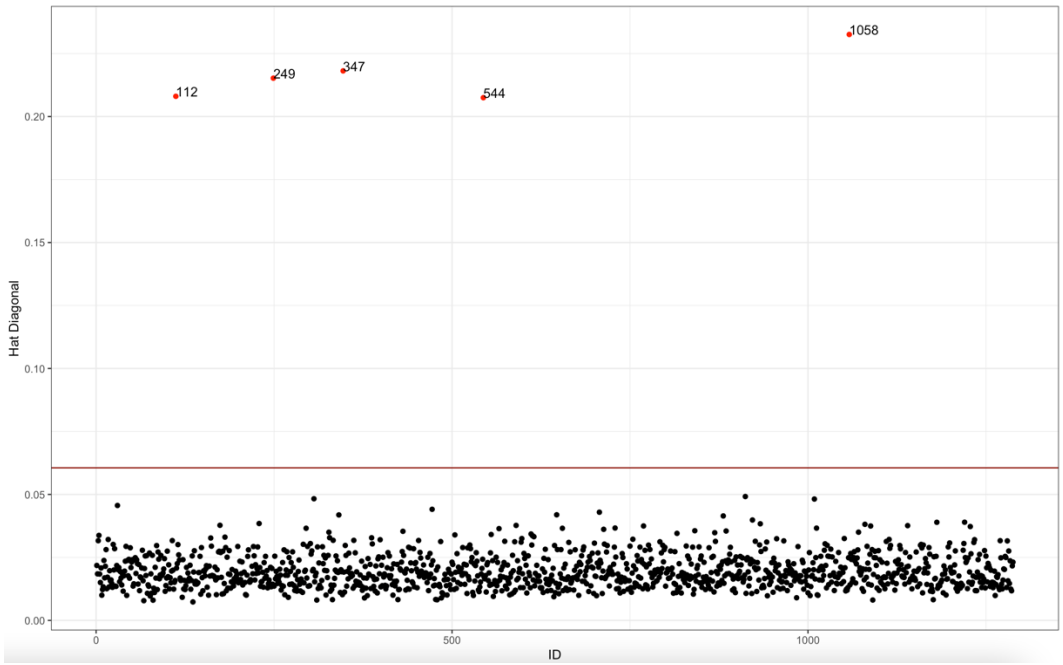

***Figure 4.2**: Plot of cases and their hat diagonals*

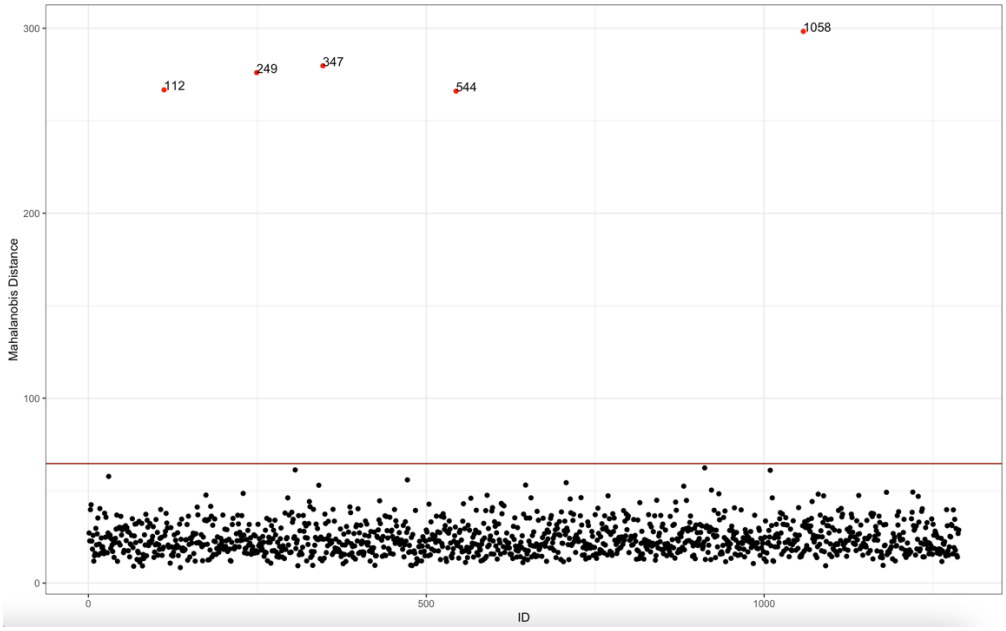

***Figure 4.3**: Plot of cases and their MD*

In Figure 4.2, the red line is the hat diagonal cut-off (0.06). Cases above the red line are considered leverage cases. Five leverage cases are identified. In Figure 4.3, the red line is the MD cut-off (64.6). The MD identifies the same five leverage cases as the hat diagonal. The plots of hat diagonal and MD are very similar. As explained in Section 3, hat diagonal and MD are closely related with each other but scaled differently.

Influential cases

CD is the global measure of influence, used to identify influential cases. In addition, DFBETAS will be used to find cases that are influential on one or more regression coefficients. The plot of CD versus ID is shown in Figure 4.4. The plot of DFBETAS on the V34A variable is shown in Figure 4.5. V34A is the "how much paid work you do per week during term time on weekdays" variable. There were 26 plots created altogether for DFBETAS, one for each OLS coefficient β_p. These coefficients correspond to all independent variables and dummy variables of the OLS model. One plot corresponds to the (Intercept) coefficient. Fourteen plots correspond to numeric variables and 11 plots correspond to dummy variables. The regression coefficients of the OLS model can be seen in Table 4.14. However, the plot for V34A is the only one that contains noteworthy insights, which the global measures of influence did not identify. The plots for the other 25 variable is shown in Appendix A - DFBETAS plots produced from influential case analysis.

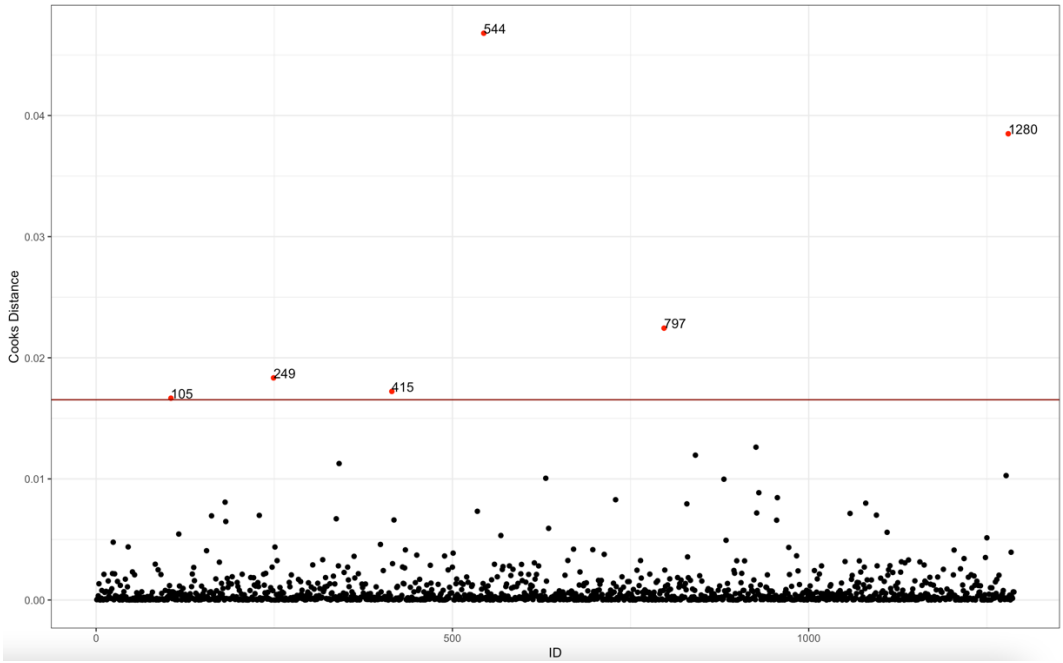

***Figure 4.4**: Plot of cases and their CD*

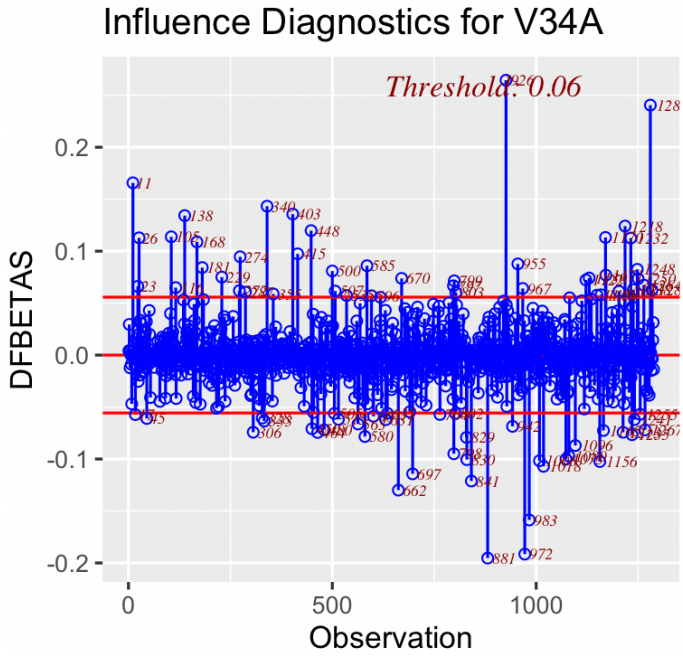

***Figure 4.5**: DFBETAS influence diagnostics for V34A*

In Figure 4.4, the red line is the CD cut-off (0.0165). Six cases were identified by CD. One extra case, 926, will be considered an influential case because of DFBETAS. We can see, from Figure 4.5, case 926 has high influence on "how much paid work you do per week during term time on weekdays (V34A)" variable. The "how much paid work you do per week during term time on weekdays (V34A)" DFBETAS coefficient for case 926 is considerably high compared to the other cases.

Unusual cases summary

All unusual cases found are shown in Table 4.6. The unusual cases identified by multiple methods (outlier, leverage, or influential) are at the top of the table, and the cases identified by fewer methods are near the bottom. Out of 1288 cases, there are 13 unusual cases (1% of cases).

Case ID	Outlier	Leverage case	Influential case
105	Yes	No	Yes
415	Yes	No	Yes
797	Yes	No	Yes
1280	Yes	No	Yes
926	Yes	No	Yes
249	No	Yes	Yes
544	No	Yes	Yes
927	Yes	No	No
930	Yes	No	No
1277	Yes	No	No
112	No	Yes	No
347	No	Yes	No
1058	No	Yes	No

***Table 4.7**: All unusual cases identified in survey dataset*

A plot of residual vs. leverage vs. influence of all cases is shown in Figure 4.6. The standardised residual is on the y-axis, the hat diagonal value is measured on the x-axis, and the CD is the size of each case on the plot. The two red horizontal dotted lines are the standardised residual cut-offs. The single red vertical dotted line is the hat diagonal cut-off. Cases that are red, mean they exceeded the CD cut-off (are influential). All unusual cases are labelled with their case ID. The effect of the residual and leverage on influence can be seen. It can be observed that high combinations of residual and leverage in a case lead to high influence. However, high residuals of a case seem to contribute to high influence more than leverage does. For example, cases 112, 347, and 1058 all have extremely high leverage, but low residual. They also have very low influence. Whereas, cases 105, 415, 797, and 1280 have high absolute residual, and low leverage values. They have very high influence as a result.

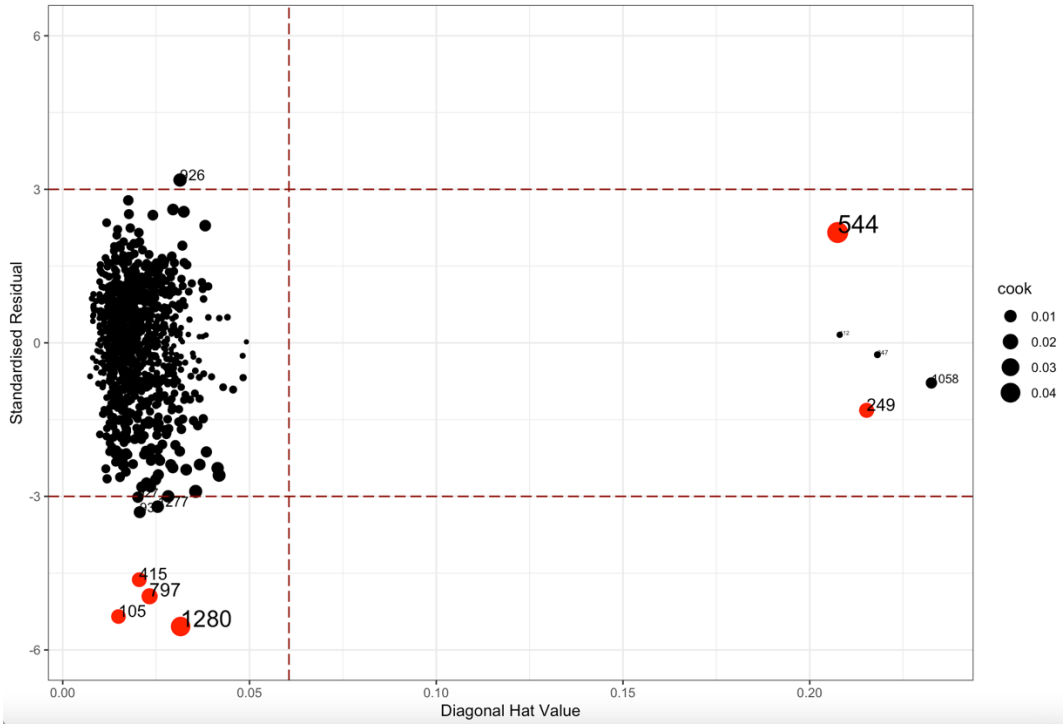

***Figure 4.6**: Standardised residual vs. hat diagonal vs. cook's distance of all cases*

4.3 Model fitting comparison

This section will deal with the efficiency and accuracy of the models. From here, the full dataset with all 1288 cases will be referred to as Dataset A (n=1288). The dataset with the 13 unusual cases removed will be referred to as Dataset B (n=1275). The OLS model will be fitted on Dataset A and Dataset B. The robust models, discussed in Section 3, will be fitted on Dataset A. The results from all these models will be compared. The MSE of all models will be compared, as well as the time taken to fit the model on all 1288 cases.

The summary results of the models are shown in Table 4.8 and Table 4.9. The MSE is calculated by fitting the model on the training dataset (70% cases) (which includes all unusual cases). The fitted model predicts SWUL of the test dataset and the MSE is calculated from the errors. The time taken to fit model is an average. The model is fitted 100 times on all 1288 cases, with time taken each time. The average of these times is taken as the time taken to fit model. The Huber estimator will be the estimator function used in all robust models besides M-estimation, for simplicity and time-saving. All estimators discussed in Section 3 are tested with the M-estimation model.

The findings of the comparisons when fitting the models on Dataset A and Dataset B are shown in Table 4.8 and Table 4.9 respectively. OLS is the fastest and most efficient model, but evidently does not fit well when the data contains unusual cases. The MSE of OLS on Dataset A is significantly higher than most of the other models. The RSE and MSE decrease

significantly when OLS is fitted on Dataset B. The RSE and MSE of all models are lower when fit on Dataset B instead of Dataset A. However, they do not decrease as much as OLS. For example, the MSE of OLS decreases by 1.44 from Dataset A to Dataset B. None of the MSEs of the robust models decrease by more than 0.4 from Dataset A to Dataset B. This demonstrates that the robust regression models are much less affected by unusual cases than OLS.

All of the robust regression models have lower RSE values than OLS on Dataset A except for LAD. LAD is evidently the least accurate model to use and highly inefficient. It is the second slowest model to fit, after S-estimation. The rest of the robust regression models work quite well with unusual cases. The RSE and MSE values of them are lower than that of OLS on Dataset A.

For Dataset B, OLS may be an optimal model to use, as it has a very low MSE and is very quick. However, some robust regression models are very close to its performance in MSE. In fact, Schweppe GM-estimation has a marginally lower MSE than OLS in Dataset B. This may be not completely accurate as these are test results, but it does show that robust regression can be quite reliable for datasets with no unusual cases. However, there is not any reason you would need to choose them over OLS in datasets with no unusual cases.

Overall, Schweppe GM-estimator may be the best model to fit on our survey dataset. It has the lowest RSE value, meaning it fits the dataset most well. The MSE is relatively low, meaning it predicts new cases very well also.

Regression Model	RSE	MSE	Time taken to fit model (milliseconds)
OLS	4.922	24.44	5.895
LAD	5.21	23.756	345.489
M-estimation (Huber estimator)	4.4	23.25	15.581
M-estimation (Andrew's Sine estimator)	4.375	23.295	12.027
M-estimation (Tukey Bisquare estimator)	4.382	23.17	20.317
M-estimation (Hampel estimator)	4.38	23.312	15.275
MM-estimation	4.384	23.115	171.465
S-estimation	4.383	27.169	823.795
Schweppe GM-estimation	4.373	23.239	16.221

***Table 4.8**: Assessment summary of all regression models on Dataset A*

Regression Model	RSE	MSE	Time taken to fit model (milliseconds)
OLS	4.631	23	5.895
LAD	5.056	23.607	345.489
M-estimation (Huber estimator)	4.349	23.15	15.581
M-estimation (Andrew's Sine estimator)	4.355	23.056	12.027
M-estimation (Tukey Bisquare estimator)	4.357	23.17	20.317
M-estimation (Hampel estimator)	4.314	23.017	15.275
MM-estimation	4.346	23.048	171.465
S-estimation	4.345	26.94	823.795
Schweppe GM-estimation	4.353	22.984	16.221

***Table 4.9**: Assessment summary of all regression models on Dataset B*

4.4 Visualisation of robust regression

In Section 3, it was discussed how the robust regression techniques deal with unusual cases. Methods such as M-estimation and MM-estimation give less weight, or importance to cases based on how unusual they are, as judged by the regression model. They judge cases based on their residual/leverage/influence. A plot of weights for each case is given for the M-estimation (Huber estimator) (Figure 4.7), MM-estimation (Figure 4.8), Schweppe GM-estimation (Figure 4.9) methods. The plots show standardised residual vs. hat diagonal vs. weight. These are similar plots to Figure 4.6, except the CD is replaced by the weights. The two red horizontal dotted lines are the standardised residual cut-offs. The single red vertical dotted line is the hat diagonal cut-off. Cases that are red, mean they exceeded the CD cut-off (are influential). It should be noted that weights from robust regression models can be used to identify unusual cases.

It can be observed that M-estimation and MM-estimation appears to not penalise cases based on their leverage at all, judging by the weights given to cases 112, 347 and 1058 in Figure 4.7 and Figure 4.8 respectively. These have extremely high leverage, and low residuals. M-estimation (Huber estimator) and MM-estimation still gives high weightings to these cases (does not penalise them). Outliers are heavily penalised in M-estimation and MM-estimation however. Cases 105, 415 797, 1280 seen in the bottom left of Figure 4.7 and Figure 4.8, have weights near 0. MM-estimation gives these cases weights of 0 and completely disregards these cases when fitting the data.

Schweppe GM-estimation is slightly more balanced in weight distribution between outliers and leverage cases, compared to MM-estimation. Cases 112, 347, and 1058 are given lower weights

compared to MM-estimation. Cases 105, 415 797, 1280 have very low weights, but are not 0. LAD, M-estimation, and S-estimation have similar weight results to MM-estimation, in that they give lower weights to outliers but do not give lower weights to leverage cases.

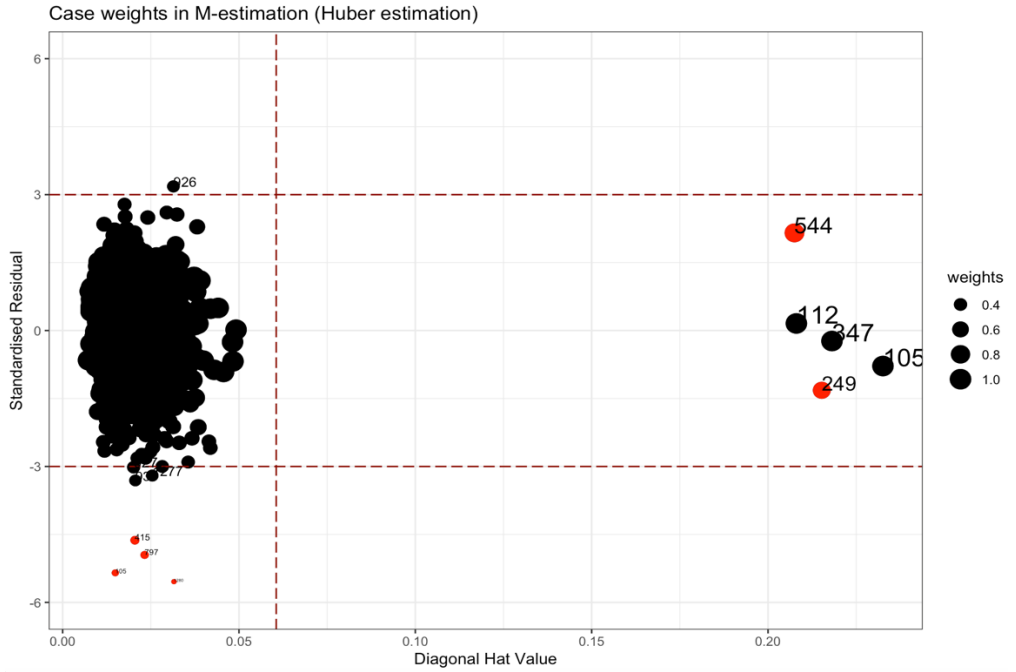

***Figure 4.7**: Standardised residual vs. hat diagonal vs. weight of Dataset A for M-estimation (Huber estimator)*

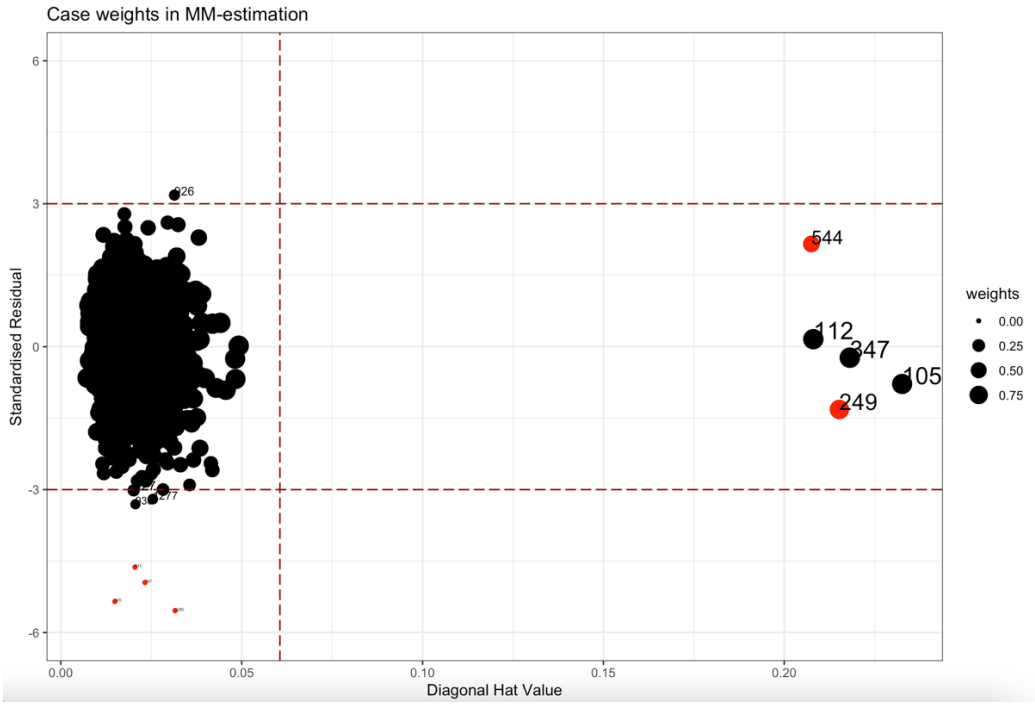

***Figure 4.8**: Standardised residual vs. hat diagonal vs. weight of Dataset A for MM-estimation*

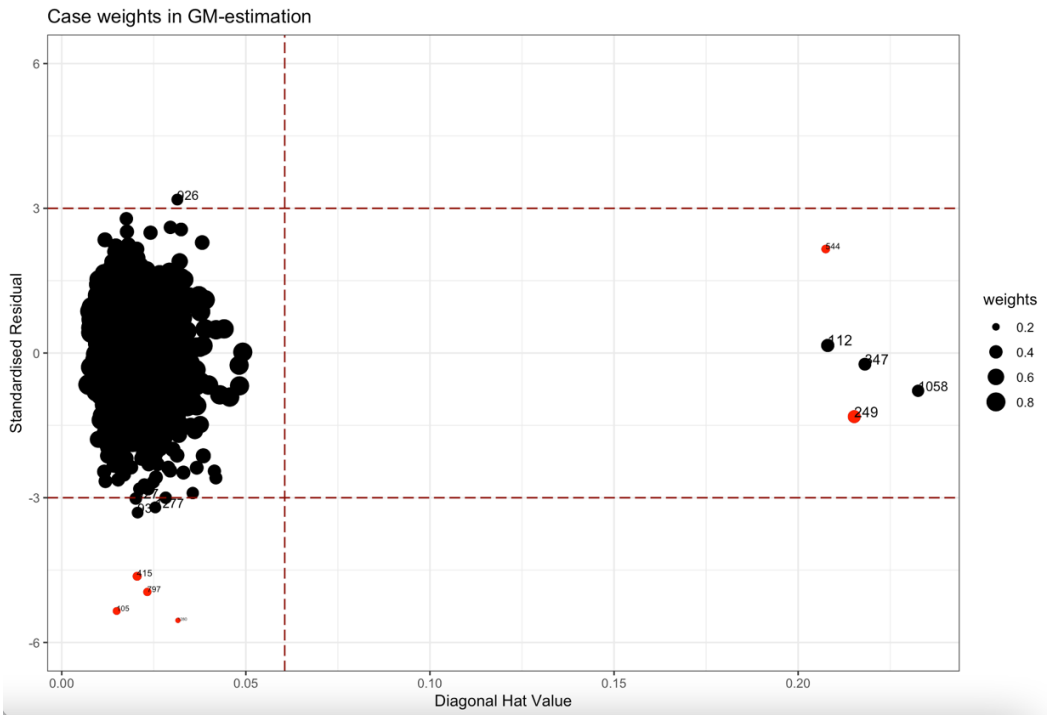

***Figure 4.9**: Standardised residual vs. hat diagonal vs. weight of Dataset B for GM-estimation*

Robust regression can be used to detect unusual cases. The weights they give to the cases give a good indication of how unusual the cases are for the model. The 10 lowest weights given by M-estimation (Huber estimator), MM-estimation and Schweppe GM-estimation are shown in Table 4.10, Table 4.11, and Table 4.12 respectively. The standardised residual, hat diagonal and CD are given for these cases also. M-estimation (Huber estimator) and MM-estimation detect high residual cases as unusual cases only. The higher the absolute standardised residual, the lower the weight of the case. Cases 105, 415, 797 and 1280 are given weights of 0 by MM-estimation model for their extremely high standardised residuals. Therefore, these four cases have no influence on the estimated coefficients of the MM-estimation model. Somewhat similar results are found for most robust regression models, in that they give less weight to cases that have higher standardised residuals.

On the other hand, Schweppe GM-estimator gives more emphasis on the hat diagonal when giving weights to the cases. Seven of the outliers are in the bottom 10 weighted cases by GM-estimator. Three leverage cases are also in the bottom 10 weighted cases. These leverage cases are cases 249, 544, and 1058.

Case ID	Standardised Residual	Hat Diagonal	Cook's Distance	Weight
1280	-5.55	0.032	0.038	0.209
105	-5.35	0.015	0.017	0.218
797	-4.95	0.023	0.022	0.235

***Table 4.10 (a)**: Standardised residual, hat diagonal, CD and weights for the ten lowest weighted cases given by M-estimation (Huber estimator).*

Case ID	Standardised Residual	Hat Diagonal	Cook's Distance	Weight
415	-4.63	0.02	0.017	0.249
930	-3.31	0.021	0.009	0.354
1277	-3.2	0.025	0.01	0.364
926	3.18	0.031	0.013	0.371
927	-3.02	0.02	0.007	0.392
631	-3.0	0.028	0.01	0.404
841	-2.9	0.036	0.012	0.41

***Table 4.10 (b)**: Standardised residual, hat diagonal, CD and weights for the ten lowest weighted cases given by M-estimation (Huber estimator).*

Case ID	Standardised Residual	Hat Diagonal	Cook's Distance	Weight
105	-5.35	0.015	0.017	0
415	-4.63	0.02	0.017	0
797	-4.95	0.023	0.022	0
1280	-5.55	0.032	0.038	0
930	-3.31	0.021	0.009	0.109
1277	-3.2	0.025	0.01	0.128
926	3.18	0.031	0.013	0.152
927	-3.02	0.02	0.007	0.213
631	-3.0	0.028	0.01	0.245
841	-2.9	0.036	0.012	0.249

***Table 4.11**: Standardised residual, hat diagonal, CD and weights for the ten lowest weighted cases given by MM-estimation.*

Case ID	Standardised Residual	Hat Diagonal	Cook's Distance	Weight
1058	-0.78	0.232	0.233	0.007
930	-3.31	0.021	0.021	0.009
1277	-3.2	0.025	0.025	0.01
926	3.18	0.031	0.031	0.013
105	-5.35	0.015	0.015	0.017
415	-4.63	0.02	0.02	0.017
249	-1.32	0.022	0.215	0.018
797	-4.95	0.023	0.023	0.022
1280	-5.55	0.032	0.032	0.038
544	2.16	0.207	0.207	0.047

***Table 4.12**: Standardised residual, hat diagonal, CD and weights for the ten lowest weighted cases given by Schweppe GM-estimation.*

4.5 Regression Coefficients of OLS and robust regression models

The regression coefficients results of OLS fitted on Dataset A, OLS fitted on Dataset B, and Schweppe GM-estimation fitted on Dataset A are shown in Table 4.14, Table 4.15, and Table 4.16 respectively. The coefficients in the 'Coefficient' column are labelled as shown in Table 4.13.

Judging from Table 4.14 and Table 4.15, it can be seen how the OLS model performs better on Dataset A than on Dataset B. The SE of all the coefficients are lower when the OLS model is fit on Dataset B, than on Dataset A. In an intuitive sense, the regression coefficients of the OLS model are more precisely estimated when unusual cases are not present, compared to when they are present. As a result, OLS is more certain of which coefficients are significant when there are no unusual cases present.

	Label format
Numerical coefficients	**Variable name (V#)**
Dummy coefficients	**Variable name (V#_#)** *Category vs. reference category*

***Table 4.13**: Label formats for numerical and dummy coefficients in Table 4.14, Table 4.15, and Table 4.16*

OLS Coefficients on Dataset A

Coefficient	Estimate	Std.Error	P-value	
(Intercept)	20.8	1.57	< 0.001	***
Commitment level to study (V12)	0.627	0.191	0.001	**
How socially integrated do you feel in UCC (V22)	-2.078	0.16	< 0.001	***
Do you take part in Club activities in UCC (V23)	0.277	0.161	0.085	NS
Do you take part in Societies activities in UCC (V25)	0.293	0.154	0.056	NS
Do you take part in Club activities outside of UCC (V27)	-0.04	0.148	0.786	NS
Do you take part in Community-based Organisations outside of UCC (V28)	-0.248	0.151	0.1	NS
I have enough money to meet my needs (V30)	-0.512	0.136	< 0.001	***

***Table 4.14 (a)**: Regression coefficients summary of OLS fitted on Dataset A*

Coefficient	Estimate	Std.Error	P-value	
How much paid work you do per week during term time on weekdays (V34A)	-0.255	0.148	0.086	NS
How much paid work you do per week during term time on weekends (V34B)	0.263	0.112	0.019	*
GTS_SCORE	0.029	0.005	< 0.001	***
MS_SCORE	0.045	0.005	< 0.001	***
AWS_SCORE	0.016	0.004	< 0.001	***
How accommodating the Lecturer was score (V16)	0.354	0.176	0.044	*
Lecture Availability (V33)	0.634	0.214	0.003	**
Gender (V1_2) *Female vs. Male*	0.363	0.311	0.243	NS
What college department is the student in (V4_4) *College of Business & Law vs. College of Arts, Celtic Studies & Social Sciences*	0.86	0.424	0.043	*
What college department is the student in (V4_5) *College of Medicine & Health vs. College of Arts, Celtic Studies & Social Sciences*	-0.629	0.459	0.17	NS
What college department is the student in (V4_6) *College of Science, Engineering & Food Science vs. College of Arts, Celtic Studies & Social Sciences*	-0.109	0.383	0.777	NS
Present Year of study (V7_2) *Year 2 vs. Year 1*	0.155	0.354	0.662	NS
Present Year of study (V7_3) *Year 3 vs. Year 1*	-0.413	0.387	0.287	NS

***Table 4.14 (b)**: Regression coefficients summary of OLS fitted on Dataset A*

Coefficient	Estimate	Std.Error	P-value	
Present Year of study (V7_4) *Year 4 vs. Year 1*	-0.489	0.442	0.289	NS
Present Year of study (V7_5) *Year 5 vs. Year 1*	-2.477	2.26	0.274	NS
Through what process did you enter UCC (V8_2) *Medical student process vs. CAO School Leaving qualifications*	0.033	0.803	0.967	NS
Through what process did you enter UCC (V8_3) *CAO Mature years vs. CAO School Leaving qualifications*	-0.747	0.531	0.159	NS
Through what process did you enter UCC (V8_4) *Other Routes vs. CAO School Leaving qualifications*	-0.258	0.569	0.65	NS

Signif. codes: *** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05, **NS** > 0.05				

Table 4.14 (c): *Regression coefficients summary of OLS fitted on Dataset A*

OLS Coefficients on Dataset B

Coefficient	Estimate	Std.Error	P-value	
(Intercept)	22.397	1.494	< 0.001	***
Commitment level to study (V12)	0.697	0.181	< 0.001	***
How socially integrated do you feel in UCC (V22)	-2.283	0.152	< 0.001	***
Do you take part in Club activities in UCC (V23)	0.193	0.152	0.204	NS
Do you take part in Societies activities in UCC (V25)	0.161	0.145	0.266	NS

Table 4.15 (a): *Regression coefficients summary of OLS fitted on Dataset B*

Coefficient	Estimate	Std.Error	P-value	
Do you take part in Club activities outside of UCC (V27)	-0.046	0.14	0.742	NS
Do you take part in Community-based Organisations outside of UCC (V28)	-0.131	0.143	0.359	NS
I have enough money to meet my needs (V30)	-0.541	0.128	< 0.001	***
How much paid work you do per week during term time on weekdays (V34A)	-0.38	0.14	0.007	**
How much paid work you do per week during term time on weekends (V34B)	0.254	0.106	0.016	*
GTS_SCORE	0.029	0.005	< 0.001	***
MS_SCORE	0.046	0.004	< 0.001	***
AWS_SCORE	0.019	0.004	< 0.001	***
How accommodating the Lecturer was score (V16)	0.353	0.166	0.034	*
Lecture Availability (V33)	0.479	0.203	0.019	*
Gender (V1_2) *Female vs. Male*	0.464	0.294	0.114	NS
What college department is the student in (V4_4) *College of Business & Law vs. College of Arts, Celtic Studies & Social Sciences*	0.618	0.4	0.122	NS
What college department is the student in (V4_5) *College of Medicine & Health vs. College of Arts, Celtic Studies & Social Sciences*	-1.037	0.435	0.017	*

***Table 4.15 (b)**: Regression coefficients summary of OLS fitted on Dataset B*

Coefficient	Estimate	Std.Error		P-value
What college department is the student in (V4_6) *College of Science, Engineering & Food Science vs. College of Arts, Celtic Studies & Social Sciences*	-0.339	0.361	0.347	NS
Present Year of study (V7_2) *Year 2 vs. Year 1*	0.125	0.334	0.708	NS
Present Year of study (V7_3) *Year 3 vs. Year 1*	-0.447	0.366	0.222	NS
Present Year of study (V7_4) *Year 4 vs. Year 1*	-0.667	0.417	0.110	NS
Through what process did you enter UCC (V8_2) *Medical student process vs. CAO School Leaving qualifications*	0.304	0.768	0.693	NS
Through what process did you enter UCC (V8_3) *CAO Mature years vs. CAO School Leaving qualifications*	-0.463	0.502	0.357	NS
Through what process did you enter UCC (V8_4) *Other Routes vs. CAO School Leaving qualifications*	0.182	0.542	0.737	NS
--- **Signif. codes: *** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05, NS > 0.05**				

***Table 4.15 (c)**: Regression coefficients summary of OLS fitted on Dataset B*
*** No cases were in Year 5 for Dataset B, so V7_5 coefficient is removed from Table 4.15*

Schweppe GM-estimator Coefficients on Dataset A

Coefficient	Estimate	Std.Error	P-value	
(Intercept)	21.968	1.571	< 0.001	***
Commitment level to study (V12)	0.643	0.191	0.001	***
How socially integrated do you feel in UCC (V22)	-2.219	0.160	< 0.001	***
Do you take part in Club activities in UCC (V23)	0.271	0.161	0.085	NS
Do you take part in Societies activities in UCC (V25)	0.161	0.154	0.056	NS
Do you take part in Club activities outside of UCC (V27)	-0.043	0.148	0.786	NS
Do you take part in Community-based Organisations outside of UCC (V28)	-0.177	0.151	0.101	NS
I have enough money to meet my needs (V30)	-0.508	0.136	< 0.001	***
How much paid work you do per week during term time on weekdays (V34A)	-0.338	0.148	0.086	*
How much paid work you do per week during term time on weekends (V34B)	0.248	0.112	0.019	*
GTS_SCORE	0.027	0.005	< 0.001	***
MS_SCORE	0.048	0.005	< 0.001	***
AWS_SCORE	0.018	0.004	< 0.001	***
How accommodating the Lecturer was score (V16)	0.368	0.176	0.044	*
Lecture Availability (V33)	0.537	0.214	0.003	*
Gender (V1_2) *Female vs. Male*	0.431	0.311	0.243	NS
What college department is the student in (V4_4) *College of Business & Law vs. College of Arts, Celtic Studies & Social Sciences*	0.802	0.424	0.043	NS

***Table 4.16 (a)**: Regression coefficients summary of Schweppe GM-estimation fitted on Dataset A*

Coefficient	Estimate	Std.Error	P-value	
What college department is the student in (V4_5) *College of Medicine & Health vs. College of Arts, Celtic Studies & Social Sciences*	-0.925	0.459	0.17	*
What college department is the student in (V4_6) *College of Science, Engineering & Food Science vs. College of Arts, Celtic Studies & Social Sciences*	-0.294	0.383	0.777	NS
Present Year of study (V7_2) *Year 2 vs. Year 1*	0.164	0.354	0.662	NS
Present Year of study (V7_3) *Year 3 vs. Year 1*	-0.464	0.387	0.287	NS
Present Year of study (V7_4) *Year 4 vs. Year 1*	-0.63	0.442	0.269	NS
Present Year of study (V7_5) *Year 5 vs. Year 1*	-3.471	2.261	0.274	NS
Through what process did you enter UCC (V8_2) *Medical student process vs. CAO School Leaving qualifications*	0.309	0.803	0.967	NS
Through what process did you enter UCC (V8_3) *CAO Mature years vs. CAO School Leaving qualifications*	-0.533	0.531	0.159	NS

***Table 4.16 (b)**: Regression coefficients summary of Schweppe GM-estimation fitted on Dataset A*

Coefficient	Estimate	Std.Error	P-value	
Through what process did you enter UCC (V8_4) *Other Routes vs. CAO School Leaving qualifications*	0.084	0.569	0.650	***

Signif. codes: *** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05, **NS** > 0.05				

***Table 4.16 (c)**: Regression coefficients summary of Schweppe GM-estimation fitted on Dataset A*

5 Discussion

5.1 Findings

The purpose of this study was to identify unusual cases, assess how robust regression techniques handle these cases and compare results with OLS linear regression. A dataset of UCC student' survey responses, provided by TWG, was used to perform the analysis. OLS regression was fit on this dataset using 18 independent variables and SWUL as the dependent variable. One variable was created as a combination of two variables from the Financial measures in the original dataset. Outliers, leverage cases, and influential cases were identified using the OLS model and diagnostic measures. These are all classed as unusual cases. Finally, multiple robust regression methods were fit on the dataset and compared with OLS regression.

OLS, multiple robust models as well as four estimation functions were described in Section 3. Multiple diagnostic measures of residual, leverage, and influence were also described. Not all diagnostic measures were used in the analysis, as some of them identify the same cases as unusual cases.

Out of 3462, 1288 complete cases were used in this analysis. OLS regression was fitted on all 1288 cases and analysed. Standardised residuals and Grubbs' test were used to find outliers. Hat diagonals and MD were used to find leverage cases. CD and DFBETAS were used to detect influential cases. Eight outliers, five leverage cases, and seven influential cases were identified. Zero cases were both outliers and leverage cases. Five cases were both outliers and influential cases. Two cases were both leverage cases and influential cases. Zero cases was an outlier, leverage case, and influential case all at the same time. Overall, 13 unusual cases were found. OLS regression was then fit on the data set with unusual cases removed (Dataset B). Eight different robust models were then fit on the complete dataset (Dataset A). The RSE, MSE and time taken to fit each model were all measured. The same eight robust models were then fit on Dataset B. The RSE, MSE and time taken to fit each model were all measured.

As expected, OLS regression fit better on Dataset B than on Dataset A. The robust regression models also fit better on Dataset B than on Dataset A, but OLS had the most drastic

improvement in fit and prediction accuracy from Dataset A to Dataset B. Also, the robust models performed better than OLS regression on Dataset A. LAD was the worst model overall and was the only robust model to have a higher RSE than OLS on Dataset A. It was also was extremely slow compared to the rest except for S-estimation. The model chosen as the best model was the Schweppe GM-estimation model. The reason this was chosen is it has the lowest RSE and has very low MSE overall. GM-estimation is also one of the very few robust models to penalise leverage cases. Therefore, this model may be the most well-balanced model. For GM-estimation, weights decrease as residual and/or leverage increase. For the rest of the robust models, weights decrease as residual increases, and is not significantly affected by leverage. This was demonstrated in M-estimation (Huber estimator) and MM-estimation.

OLS had one of the lowest MSE values for the dataset with unusual cases removed (Dataset B). However, most of the robust regression models perform almost as well to OLS in terms of MSE for Dataset B. In fact, GM-estimation had a marginally lower MSE than OLS for Dataset B. Robust regression methods can be completely reliable to use on a dataset with no unusual cases, but there is no reason to use them over OLS in a dataset with no unusual cases.

5.2. Limitations and recommendations

The major limitation of the study was the omission of missing values (62.8%) from the dataset. The easiest way to deal with missing data is to drop all cases that have one or more values missing in any of the variables required for analysis. This is called complete case analysis and is how the data for this analysis was dealt with. Although under missing completely at random (MCAR) *(Papageorgioua et al., 2018)* does not lead to bias of the results, it may result in considerable loss of data and associated loss of power (e.g. wider confidence intervals) because the sample size is reduced. The extent of the loss of power depends on the quantity of missing data. Imputation is an alternative method to deal with missing data, which does not result in loss of data *(Soley-Bori, 2013).* Imputation is the process of substituting each missing value for a reasonable guess, and then carrying out the analysis as if there were no missing values *(Soley-Bori, 2013).* Median imputation or regression imputation may be applicable to use on our incomplete data set. For median imputation, the median of all non-missing values in variable X is determined and that value is used to impute the missing values of X *(Zhang, 2016).* Regression imputation uses the complete cases in all other independent variables to predict the missing values of X, using some form of regression. In general, the main limitation of regression imputation is that it leads to an underestimation of SEs. This happens because the imputed values are determined by a model applied to the observed data *(Allison, 2001).*

One way to enhance the analysis, would be to test the regression models with only outliers, only leverage cases, and only influential cases removed from the data. Three additional Datasets could have been fit on the regression models, labelled Dataset 1, Dataset 2, and Dataset 3. Dataset 1 would have only outliers removed from the survey dataset. Dataset 2 would have only leverage cases removed from the survey dataset. Dataset 3 would have only influential cases removed from the survey dataset. The OLS, and robust regression models

fitted on Dataset 1 would be compared. The same would be done with Dataset 2. The same is done Dataset 3. This would allow us to get a better sense of how each robust model handles outliers, leverage cases, and influential cases specifically. For example, we know that GM-estimation gives lower weights to both outliers and leverage cases, while the rest of the robust models only give lower weights to outliers only. However, the RSE and MSE on Dataset 1, Dataset 2, and Dataset 3 could give a better sense of how effective the robust models are with handling each type of unusual case.

6 Conclusions

The objective of this project is (1) identify unusual cases, (2) assess how robust regression techniques handle these cases and (3) compare regression coefficients with OLS linear regression. After fitting an OLS regression model on all 1288 cases, 13 unusual cases were found. Fitting an OLS on the data set with these unusual cases removed resulted in a better fit (RSE) and more accurate prediction accuracy (MSE) compared to when OLS is fit on a dataset with unusual cases. Unusual cases negatively affect the parameter estimates and increase coefficient SEs.

This analysis was conducted using only complete cases. This results in the omission of 62.8% of cases from the original survey dataset. A simple test to check if the omission of cases from the dataset caused bias in the dataset would be to perform a Chi-Square test. Let us define Dataset 1 which contains all 1288 cases that were retained for the analysis, and define Dataset 2 that contains all 2178 cases removed from the original survey dataset. A Chi-square test was used to check if the distribution of the Gender variable is different between Dataset 1 and Dataset 2. The p-value returned was 0.137, meaning that the distributions of the Gender variable between Dataset 1 and Dataset 2 are the same. Fortunately, there is not any significant bias in the Gender variable created from the omission of cases. However, omission of cases leads to higher SEs and an associated loss of power.

Robust regression fitted on the same data set reduced the effect of unusual cases with the use of weight functions. Multiple robust models were tested. Schweppe GM-estimation reduced the weight of all unusual cases, while the other robust models reduced weights of outliers only. Out of all robust models, Schweppe GM-estimation had the lowest RSE, while MM-estimation had the lowest MSE. The differences in RSE and MSE between most models were slight however. Schweppe GM-estimation was chosen to be the best robust regression model overall, for its low RSE, MSE and for the fact it gives low weights to all unusual cases, and not only outliers.

This report draws the conclusion that robust regression is better than OLS for the analysis of university data containing unusual cases. When no unusual cases are present, robust regression poses no clear advantage over OLS. However, robust regression performs almost as well or just as well as OLS when no unusual cases are present. Therefore, robust regression would be reliable to use on a dataset with no unusual cases, but there would be no reason to choose it over OLS.

Sixty-two percent of cases were deleted from the original dataset for this analysis. The next step for this analysis would be to minimise the number of cases deleted from the original dataset. This could be done using median imputation, described in Section 5.2. With this, the associated loss of power is minimised.

Works Cited

Aguinis, Herman, Ryan K. Gottfredson and Harry Joo. *Best-Practice Recommendations for Defining, Identifying, and Handling Outliers*. Sage, 2013.

Alma, Özlem Gürünlü. *Comparison of Robust Regression Methods in Linear Regression*. International Journal of Contemporary Mathematical Sciences, 2011.

Almetwally, Ehab M. and Hisham M. Almongy. *Comparison Between M-estimation, S-estimation, And MM Estimation Methods of Robust Estimation with Application and Simulation*. International Journal of Mathematical Archive-9(11), 2018, 55-63 Available online through www.ijma.info ISSN 2229 – 5046 International Journal of Mathematical Archive- 9(11), 2018.

Andale. *Cook's Distance/Cook's D: Definition, Interpretation*. 2016. Statistics how to. <http://www.statisticshowto.com/cooks-distance/>.

Atkinson, A. C. *Masking unmasked*. Biometrika, Volume 73, Issue 3, 1986.

Ayinde, Kayode, Adewale Lukman and Olatunji Arowolo. *Robust Regression Diagnostics of Influential Observations in Linear Regression Model*. Open Journal of Statistics, 2015.

Becker, Claudia and Ursula Gather. *The Masking Breakdown Point of Multivariate Outlier Identification Rules*. Taylor & Francis, Ltd. on behalf of the American Statistical Association, 1999.

Belsley, David A., Edwin Kuh and Roy E. Welsch. *REGRESSION DIAGNOSTICS: IDENTIFYING INFLUENTIAL DATA AND SOURCES OF COLLINEARITY*. New York: John Wiley & Sons, Inc., 1980.

Bruin, J. *newtest: command to compute new test. UCLA: Statistical Consulting Group*. 2006. <https://stats.idre.ucla.edu/stata/ado/analysis/>.

Buchanan, E.M. and J.E. Scofield. *Methods to detect low quality data and its implication for psychological research*. Behavior research methods, 50(6):2586– 2596. , 2018.

Capel, R. *L'Evaluation des Personnes: Théories et Techniques*. Geneva: Slatkine, 2009.

Chatterjee, Samprit and Ali S. Hadi. *Influential Observation, High Leverage Points, and Outliers in Linear Regression*. Statistical Science, 1986.

Dhakal, Chuda Prasad. *DEALING WITH OUTLIERS AND INFLUENTIAL POINTS WHILE FITTING REGRESSION*. Kathmandu: Institute of Science and Technology, 2017.

Farcomeni, Alessio and Laura Ventura. *An overview of robust methods in medical research*. Rome, 2012.

Farnè, Matteo and Angelos T. Vouldis. *A methodology for automatised outlier detection in high-dimensional datasets: an application to euro area banks' supervisory data*. ECB Working Paper Series No 2171, 2018.

Fox, John and Sanford Weisberg. *Robust Regression*. 2013.

Fox, John. *Encyclopedia of Social Measurement*. Kimberly Kempf-Leonard, 2015.

Glasser, Leslie. *Dealing with Outliers: Robust, Resistant Regression*. Curtin: Curtin University of Technology, 2007.

Goldammer, Philippe, et al. *Careless responding in questionnaire measures: Detection, impact, and remedies*. The Leadership Quarterly, 2020.

Grubbs, Frank E. *Procedures for Detecting Outlying Observations in Samples*. American Statistical Association, 1969.

Hitt, M. A., et al. *Attributes of successful and unsuccessful acquisitions of US firms*. British Journal of Management, 1998.

Hutcheson, G. D. *Ordinary Least-Squares Regression*. The SAGE Dictionary of Quantitative Management Research, 2011.

Jerome, Frieman. *Principles & methods of statistical analysis*. Thousand Oaks, California. p. 130. ISBN 9781483358598. OCLC 967133901., 2017.

Jr, P.T. Costa and R.R. McCrae. *The Revised NEO Personality Inventory (NEO-PI-R)*. Sage Publications, Inc., 2008.
Judd, Charles M and Gary H MacClelland. *Data analysis: a model-comparison approach*. 1989.
Kaufman, L. and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, Hoboken. http://dx.doi.org/10.1002/9780470316801.ch1, 2005.
Orr, John M, Paul R Sackett and Cathy L.Z. Dubois. *OUTLIER DETECTION AND TREATMENT IN 1/0 PSYCHOLOGY: A SURVEY OF RESEARCHER BELIEFS AND AN EMPIRICAL ILLUSTRATION*. Minnesota, 1991.
Ortiz, M. Cruz, Luis A. Sarabia and Ana Herrero. *Robust regression techniques A useful alternative for the detection of outlier data in chemical analysis*. Burgos, 2006.
Osborne, Jason W and Amy Overbay. *The power of outliers (and why r The power of outliers (and why researchers should AL chers should ALWAYS check for them)* . ScholarWorks, 2004.
O'Sullivan, K, et al. *Report on the Quantitative Aspects of the UCC Student Survey 2009*. Cork: UCC, 2010.
Rasmussen, J. L. *Evaluating outlier identification tests: Mahalanobis DSquared and Comrey D. Multivariate Behavioral Research*. 1988.
Ripley, B. D. *Robust Statistics*. 1992.
Sevcenco, Sabina, et al. *Quantitative Apparent Diffusion Coefficient Measurements Obtained by 3-Tesla MRI Are Correlated with Biomarkers of Bladder Cancer Proliferative Activity*. https://doi.org/10.1371/journal.pone.0106866, 2014.
Sweet, Stephen A. and Karen Grace Martin. *Data Analysis with SPSS: A First Course in Applied Statistics*. Pearson, 2012.
Vittinghoff, Eric, et al. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models (Statistics for Biology and Health) 2nd ed. 2012 Edition* . Springer Publishing Co., 2012.
Yellowlees, Ann, et al. *The Appropriateness of Robust Regression in Addressing Outliers in an Anthrax Vaccine Potency Test*. Oxford University Press BioScience 66: 63–72, 2016.
Yu, Chun, Weixin Yao and Xue Bai. *Robust Linear Regression: A Review and Comparison*. Manhattan: Department of Statistics,, 2014.
Zhang, Zhongheng. *Missing data imputation: focusing on single imputation*. 2016.

Appendix A

DFBETAS plots produced from influential case analysis

DFBETAS was used to identify any influential cases that were not identified by the global measures of influence. Only the plot for the "how much paid work you do per week during term time on weekdays (V34A)" variable was shown in Figure 4.5. All the plots from the DFBETAS output are shown in Figure A.1 (a)-(g).

It can be noted that the influential cases (cases 105, 249, 415, 544, 797, 926, 1280) identified by the global measures of influence appear to have high influence on specific variables. For example, case 415 particularly has high influence on the "Do you take part in Community-based Organisations outside of UCC (V28)", "Gender (Female dummy) (V1_2)", "Present Year of study (Year 3 dummy) (V7_3)"and "Through what process did you enter UCC (CAO Mature Years dummy) (V8_3)" variables. This likely means the student for case 415 answered unusually for these variables in the survey. It can be seen that case 1280 has extremely high influence on a large majority of the variables. This likely means this student made little to no attempt in aswering the questions honestly, and thus their answers do not make sense, given their SWUL score. Thus, they are the most extreme outlier out of all cases, as seen in Figure 4.6.

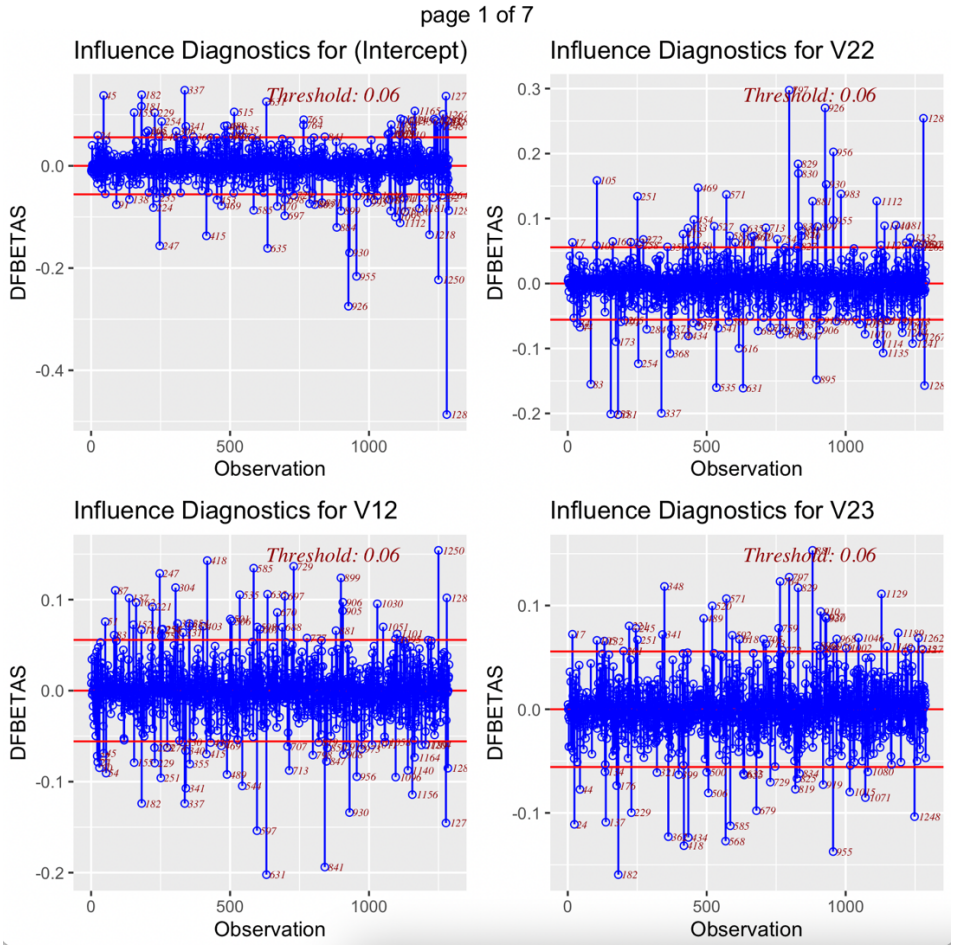

***Figure A.1 (a)**: DFBETAS output for variables in OLS regression model*

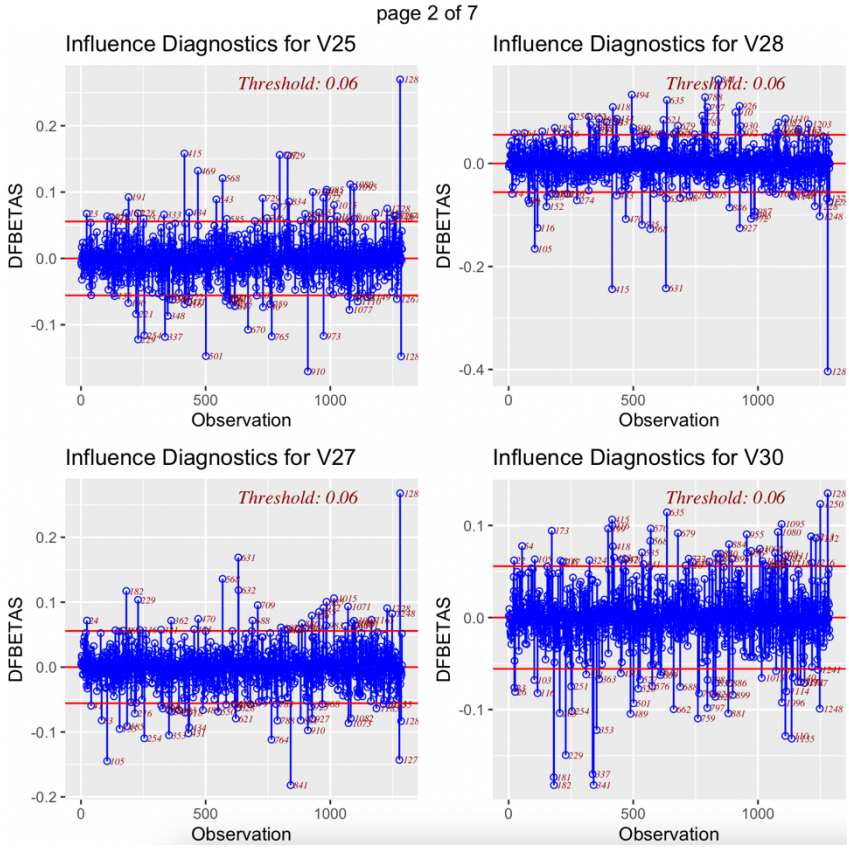

***Figure A.1 (b)**: DFBETAS output for variables in OLS regression model*

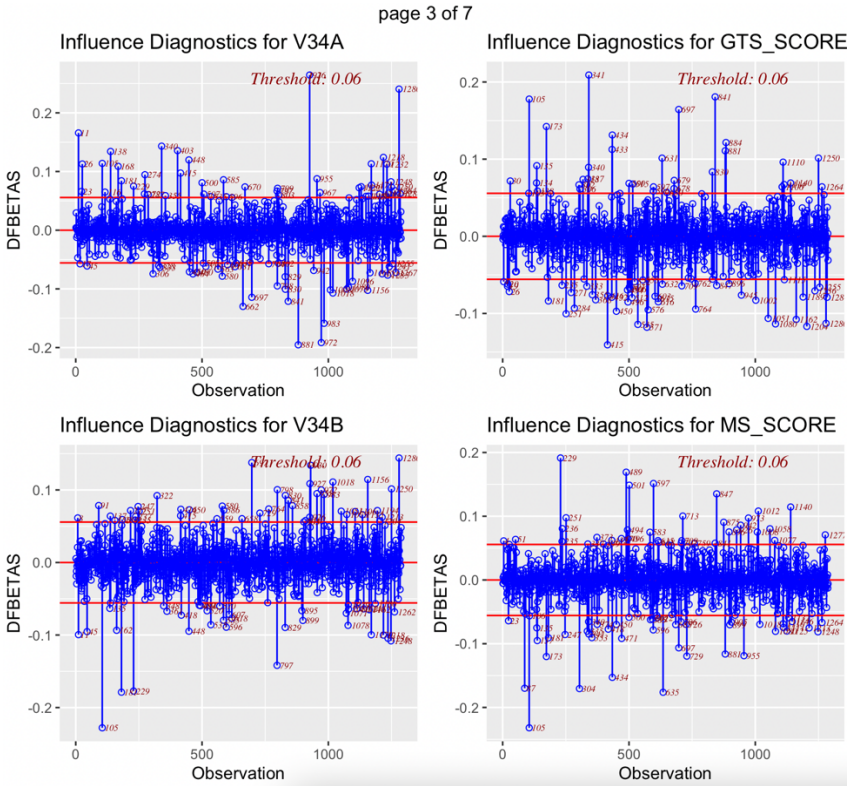

***Figure A.1 (c)**: DFBETAS output for variables in OLS regression model*

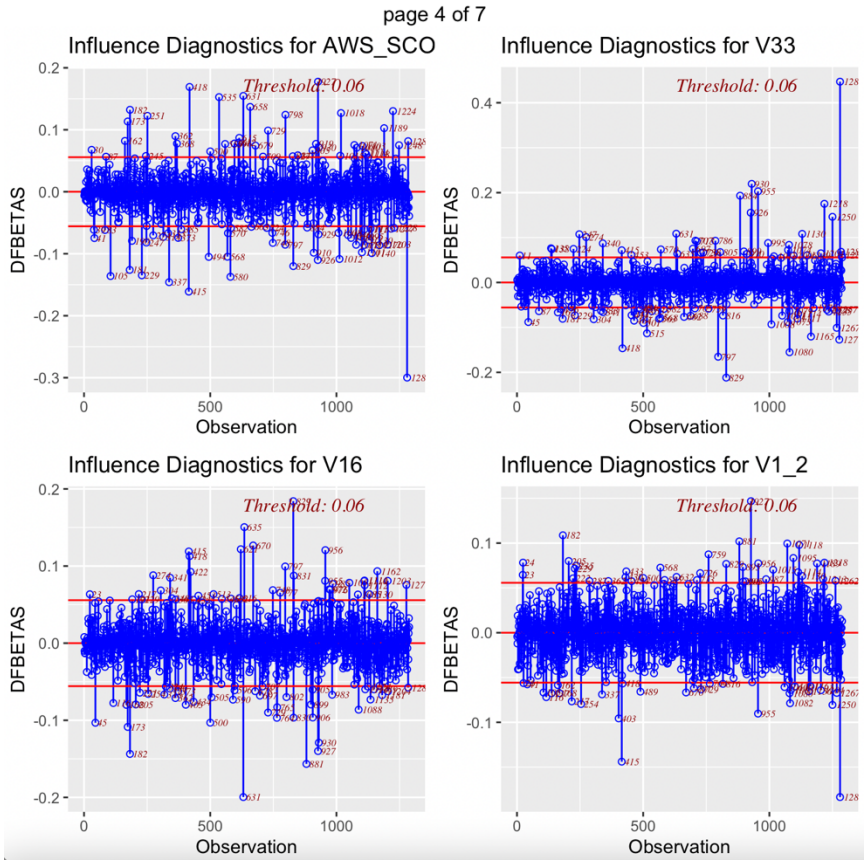

***Figure A.1 (d)**: DFBETAS output for variables in OLS regression model*

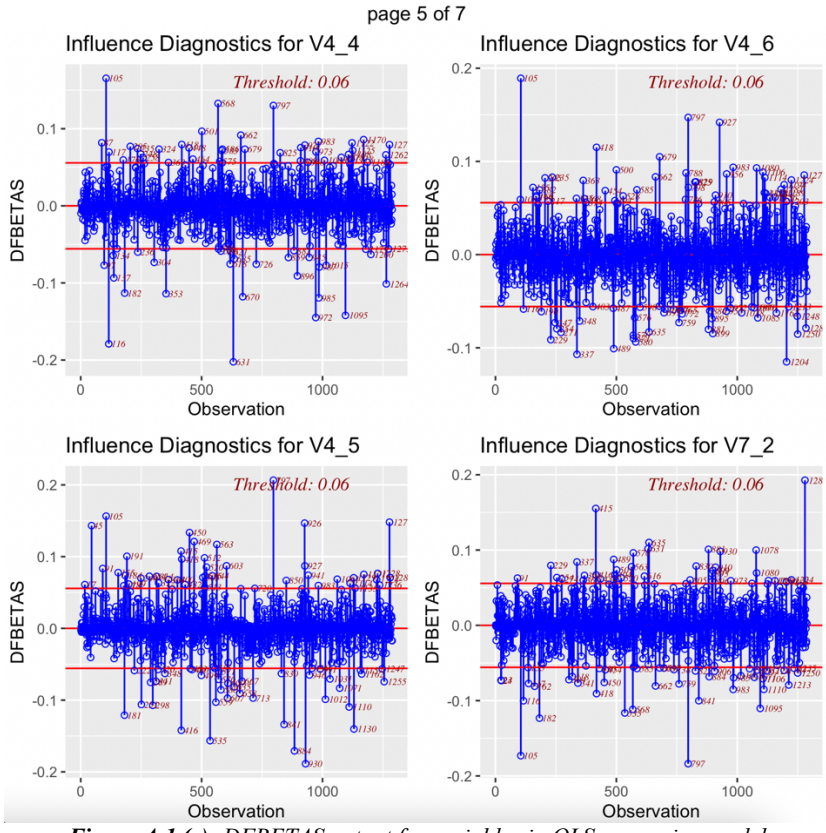

***Figure A.1 (e)**: DFBETAS output for variables in OLS regression model*

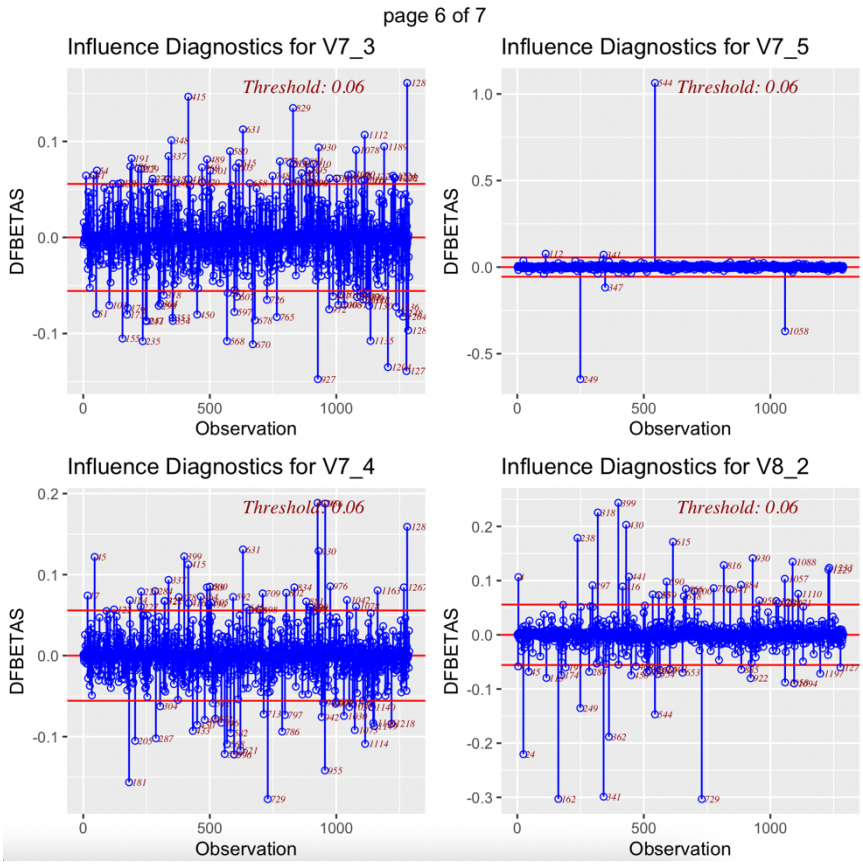

***Figure A.1 (f)**: DFBETAS output for variables in OLS regression model*

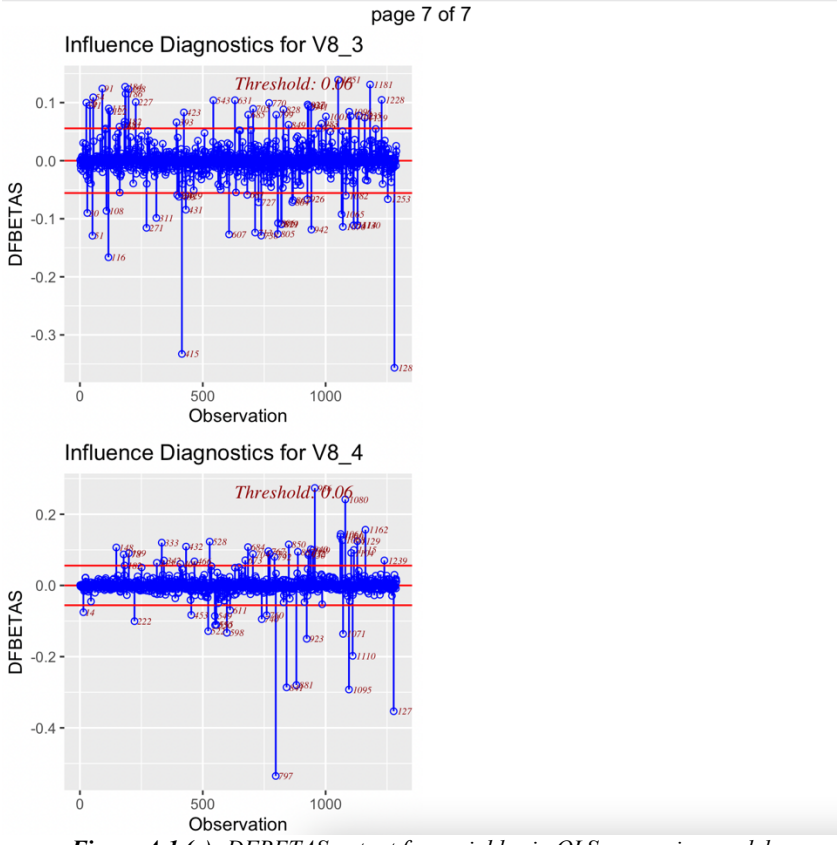

***Figure A.1 (g)**: DFBETAS output for variables in OLS regression model*