National College of Ireland

Statistics for Data Analytics

on

Statistical regression

Martin Mohan
x18191339

MSc/PGDip Data Analytics – 2019

Submitted to: Tony Delaney

(Individual Project - 30% of marks for the module)

National College of Ireland
Project Submission Sheet – 2019
School of Computing

National College of Ireland

Student Name:	Martin Mohan
Student ID:	x18191339
Programme:	PGDip Data Analytics - Part Time
Year:	2019
Module:	Statistics
Lecturer:	Tony Delaney
Submission Due Date:	07/04/2019 at 23:59
Project Title:	Statistical regression

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

All materials are referenced in the bibliography section. The Harvard Referencing Standard in accordance with project guidelines. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. .

Signature:	
Date:	August 13, 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Statistical regression

Martin Mohan
x18191339

August 13, 2019

1 Introduction

This is the report for the course statistics for Data analysis Delaney (2019). A research report is represented incorporating two pieces of statistical analysis a multiple regression and a binary regression. The choice of 3 data sets was provided to use with the regressions. Data was selecte from the UN and WHO databases.

http://www.who.int/gho/en/	World Health Organisation Data
http://www.who.int/gho/en/	United Nations data
https://data.govt.nz	The New Zealand Government data depository

Table 1: Regression dependent variable **WHS2_138** and 3 independent variables

2 Multiple Regression

Due to antretroviral therapy HIV is no longer a life sentence but expectation of death by AIDS is still a threat in many countries especially poor African countries. Adult HIV prevalence exceeds 20% in Eswatini (Swaziland), Botswana, and Lesotho, , while an additional six countries report adult HIV prevalence of at least 10% Wikipedia (2017*b*) . According to Mansori et al. (2017) The prevalence and incidence rates of HIV/AIDS each had an inverse correlation with HDI and its four indicators (life expectancy at birth, mean years of schooling, expected years of schooling, and GNI per capita).
In this report data was taken from the WHO database to see how the death rate from AIDS around the world related to "life expectancy at birth", GNI per capita in US dollars and Antiretroviral coverage. According to Burrage et al. (n.d.) "the percentage of children with HIV infection receiving coverage increased from 24% in 2012 to 44% in 2016. This report indicated more antiretroviral coverage is available in countries with higher incidences of HIV/AIDS.
Data was taken from the WHO database and the UN database. The variable being investigated (dependent) was Death due to HIV/AIDS (per 100000 population) **WHS2_138**
2

WHS2_138	**Deaths due to HIV/AIDS (per 100 000 population)**
WHOSIS_000001	Life expectancy at birth (years)
MDG_0000000033	Antiretroviral therapy coverage among people with HIV infection eligible for ART according to 2010 guidelines (percent)
Schooling	School life expectancy Wikipedia (2018*b*)
GNI	GNI per capita in US dollars Wikipedia (2017*a*)

Table 2: Regression dependent variable **WHS2_138** and 3 independent variables

2.1 The objectives of the analysis and the context of the data being analysed.

The objective is to create a multiple regression matrix showing correlation between the dependent variable WHS2_138 (Deaths due to HIV/AIDS) and several independent variables chosen from 2 and 1.

2.2 Data cleaning tranformation of all independent / dependent variables in the analysis.

The data was viewed and downloaded from the WHO using the R package called "WHO". The UN data was downloaded as csv files and loaded into R using read.csv() function
The R package dplyr was used to clean and merge the data using the functions select(),filter(),arrange(),mutate(),summarise() ,group_by().
Some data could not be cleaned programmatically so was downloaded to a csv file and cleaned using tools like excel and grep. One rule for the minimum number of cases required for a multiple linear regression is "50+8k" see Field (2009) [pp 274] which means 66 (50+8(2)) cases in the case of 2 predictors. After trial and error it decided to filter years after 2009.The further back in time filtered the more data was available but the data then becomes more out of date. A csv file with 158 different values was extracted (although some values were NA). See below...

```
> HIV_view
# A tibble: 158 x 8
X1 country          WHS2_138   GNI WHOSIS_000001 MDG_0000000033 school
<dbl> <chr>            <dbl> <dbl>         <dbl>          <dbl>  <dbl>
1     1 Afghanistan         NA    616.          63.2              8   6.62
2     2 Bhutan              NA   2383.          70.2             12   7.31
3     3 Botswana           282   6424.          65.3             95   7.63
4     4 Cuba                 2.6 7511.          78.8             95   5.84
5     5 Democratic Repu~    48    487.          60.1             31   6.30
6     6 Ghana               46   1752.          63.1             58   6.67
7     7 Guinea              44    737.          58.6             50   5.44
8     8 Lesotho            755   1283.          52.1             54   7.50
9     9 Madagascar          28    446.          65.7              1   7.39
10    10 Niger              20    357.          59.2             46   4.21
# ... with 148 more rows
>
```

2.3 Report on the results of preliminary tests to check that the assumptions of the technique being used are not violated.

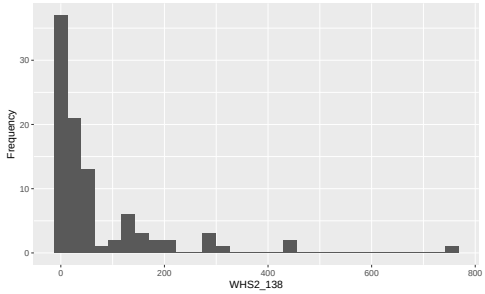
Histograms were used to check if the data was parametric and scatter plots were used to view the predictors against the dependent variable (WHS_238). The data 2 was inspected using histograms and qq plots 1.
Heavy positive skew was seen in histogram WHS_138 because some African countries have an unusually high death rate due to AIDS (e.g Lesotho=755) as mentioned in the introduction in 2.GNI is also positively skewed due to unequal distribution of wealth in the world. The other histograms had normal distribution. To reduce skew richer countries with GNI<10000 dollar were filtered and countries with very high deaths/per capita (287>) were filtered.
After filtering the skew of WHS2_198 and GNI was reduced but it was still worryingly high and this error was reflected in the final model4.

2.4 Analysis in R and discussion of the output from the model.

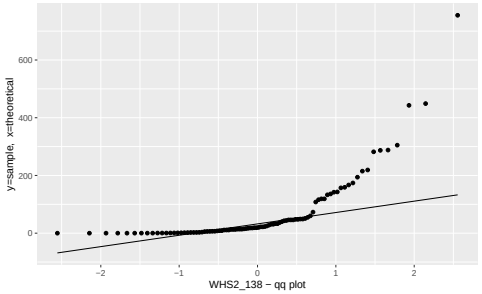
A hierachical regression was done.i.e A single regression using independent variable WHOSIS_00001 2.4.1 followed by a multiple regression.2 in 2.4.2

2.4.1 Hierarchical regression one predictor

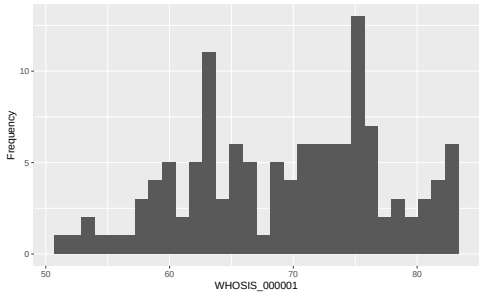
```
Call:
lm(formula = WHS2_138 ~ WHOSIS_000001, data = HIV_view)

Residuals:
Min     1Q Median     3Q    Max
-82.83 -35.90 -10.04  14.97 219.35

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    366.777     68.037   5.391 1.20e-06 ***
WHOSIS_000001   -4.657      1.021  -4.560 2.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.99 on 61 degrees of freedom
Multiple R-squared:  0.2542,Adjusted R-squared:  0.242
F-statistic:  20.8 on 1 and 61 DF,  p-value: 2.516e-05
```



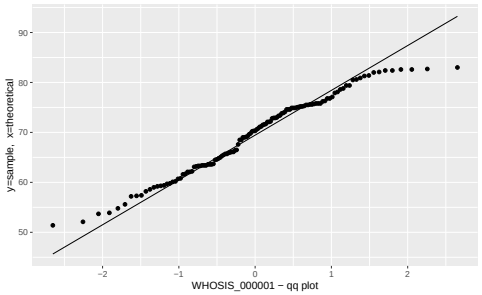
(a)



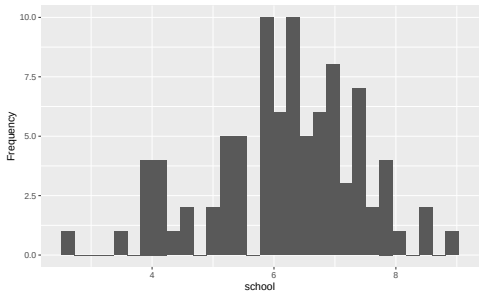
(b)



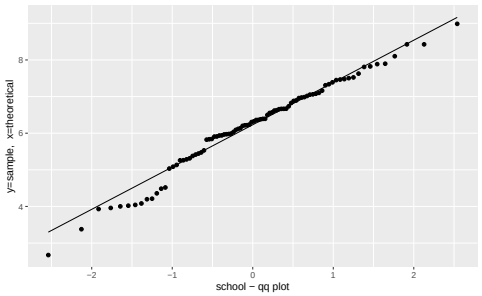
(c)



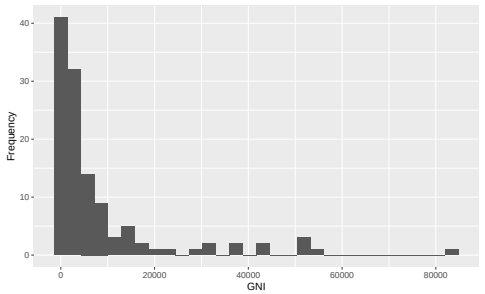
(d)



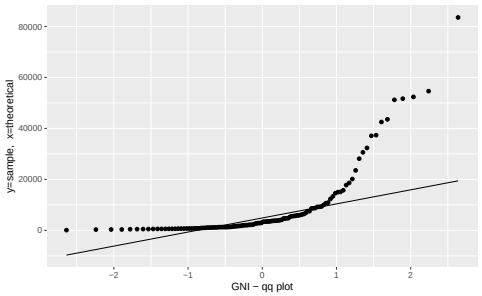
(e)



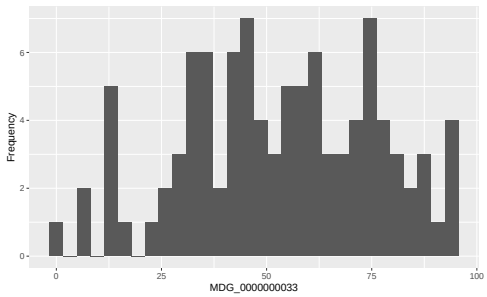
(f)



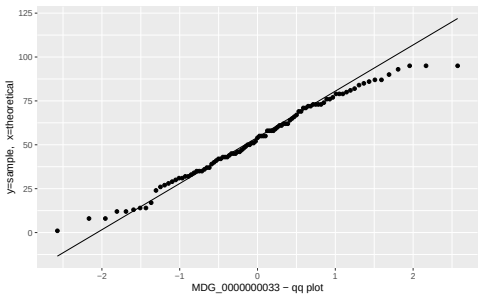
(g)



(h)

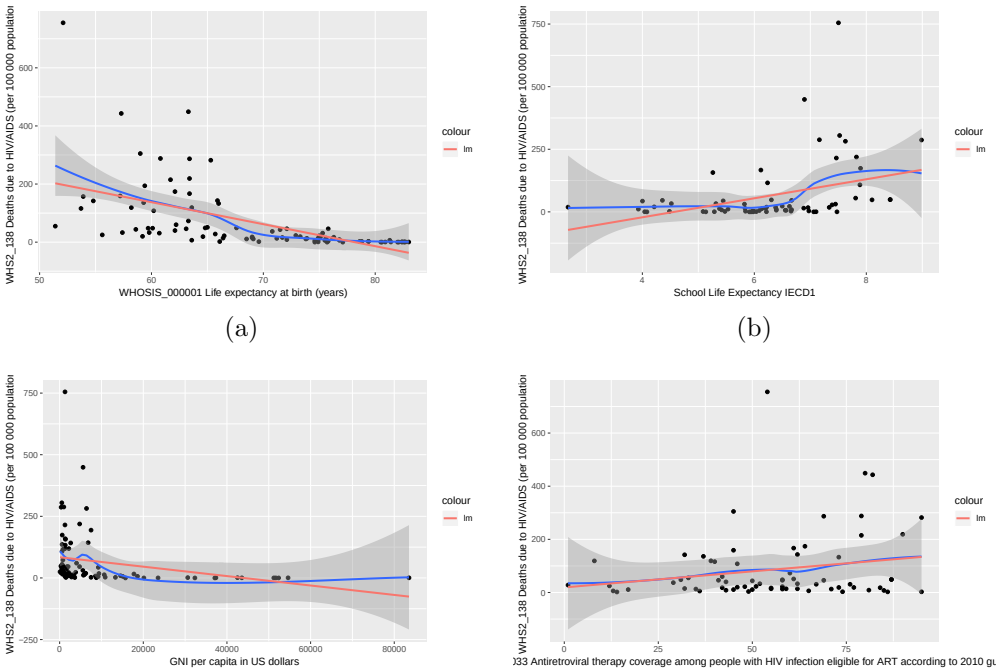


(i)



(j)

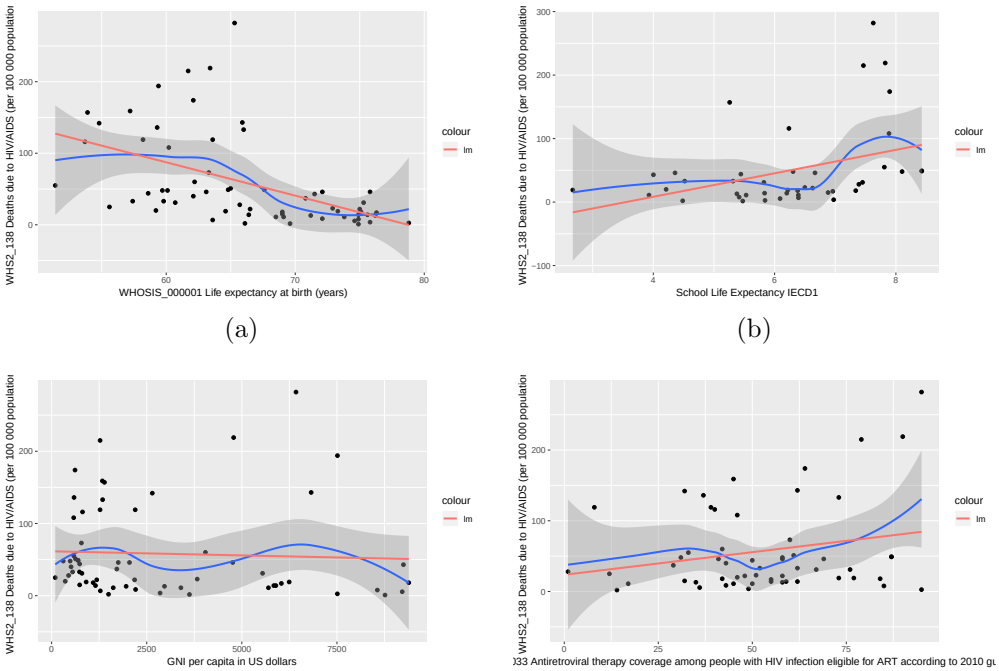
Figure 1: Histograms / qqplots of HIV data unfiltered



(a) (b)

(c) (d)

Figure 2: Scatter plot of HIV data unfiltered



(a) (b)

(c) (d)

Figure 3: Scatter HIV data filtered (GNI<10000)

2.4.2 Hiearchical regression two predictors

```
Call:
lm(formula = WHS2_138 ~ WHOSIS_000001 + MDG_0000000033, data = HIV_view)

Residuals:
Min      1Q  Median      3Q     Max
-70.569 -36.922  -2.927  23.278 173.624

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    338.7289    66.3120   5.108 4.37e-06 ***
WHOSIS_000001   -5.1003     1.0330  -4.937 7.99e-06 ***
MDG_0000000033   1.0810     0.3287   3.289  0.00177 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.51 on 54 degrees of freedom
(6 observations deleted due to missingness)
Multiple R-squared:  0.3452,Adjusted R-squared:  0.321
F-statistic: 14.24 on 2 and 54 DF,  p-value: 1.083e-05
```

2.5 Testing the accuracy of the regression model

Various tests were carried out on the final regression model using R to test the model

In an ordinary sample we expect 95% of cases to have standardized residuals within about +/-2. Three cases were found with values outside this range and the **cooks distance** for the values was calulated as **0.435, 0.146 and 0.101** which is well below 1 (note: several large values had already been filtered previously).

Durbin-watson was 2.07 (close to 2) confirming the assumption of independence. The bf**p-value of 0.8** which is much greater than 0.05 confirmed this assumption.

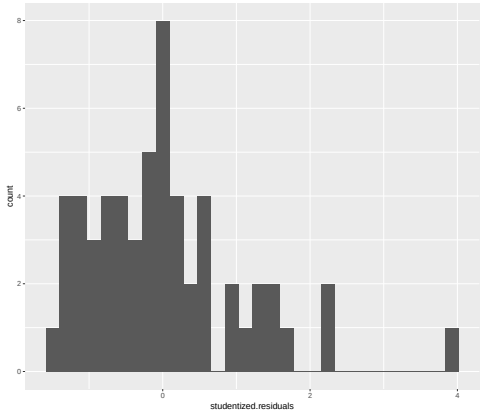
The Vif is nowhere near the 10 which would cause alarm about multicollinearity. The **mean of VIF is 1.08** and this causes no concern as a value much greater than 1 is needed to indicate regression bias. The tolerance of vif (=1/vif) is above 0.9 (a tolerance below 0.2 indicates a problem) so there is no tolerance problem.

2.5.1 Checking assumptions about residuals

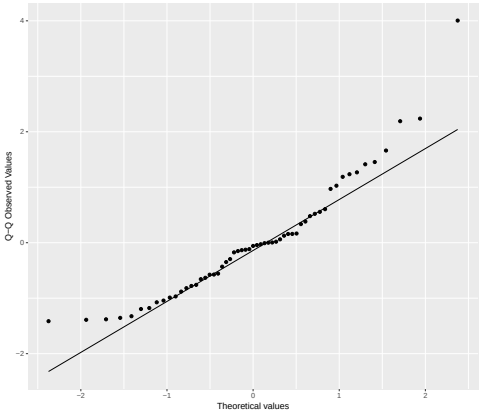
Residual plots of the final model 4 show clearly that the problem with skew was not solved which threw into doubt all other tests although the graph of the fitted model seemed acceptable.

2.5.2 Cross validation of the model using data splitting

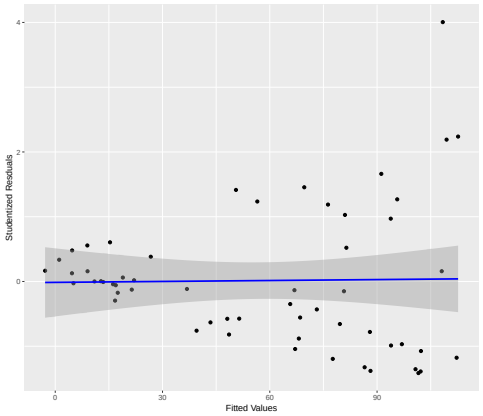
The data was randomly split in half and regression test run to see how the output compared with 2.4.2. The B values were similar to the original, as were R^2 values.



(a)



(b)



(c)

Figure 4: Multiple regression plots for the final model

The F-statistic differed from 14.o4 to 10.05 but there was less DF in the test model.

```
Call:
lm(formula = WHS2_138 ~ WHOSIS_000001 + MDG_0000000033, data = HIV_view2)

Residuals:
Min      1Q  Median      3Q     Max
-72.007 -33.527  -9.213  25.389  96.255

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    313.7326    85.3775   3.675  0.00104 **
WHOSIS_000001   -4.7881     1.3086  -3.659  0.00108 **
MDG_0000000033   1.2714     0.3852   3.300  0.00272 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.98 on 27 degrees of freedom
(3 observations deleted due to missingness)
Multiple R-squared:  0.4267,Adjusted R-squared:  0.3843
F-statistic: 10.05 on 2 and 27 DF,  p-value: 0.0005466
```

2.5.3 R^2 and $Adjusted - R^2$ for the hierarchical model

R^2 measures how much of the variability in the outcome is accounted for by the predictors. In the first model it's values is .2542 meaning life expectancy at birth accounts for 25.4% of the variation in HIV/Aids deaths. When the other predictor MDG_0000000033 is added R^2 increases to .3452 (34.5%).
The adjusted R^2 give us an idea of how well our model generalizes Field (2009) [page 273] ,and ideally we would like its value to be the same, or very close to R^2. In this case the final difference .345-.321= (about 2.4%)). This shrinkage means that if the model were derived from the population rather than a sample it would it would account for approximately 2.4% less variance in the outcome.

2.5.4 Discussion of the multiple regression model

The first column in 2.4.2 gives estimates of the β-values to be used in the multiple regression equation. Therefore we can define the model as follows.

$$\widehat{Y} = \beta_0 + \beta_1 WHOSIS\_000001 + \beta_2 MDG\_0000000033 \quad (1)$$

$$= 338.73 - 5.1 WHOSIS\_000001 + 1.08 MDG\_0000000033 \quad (2)$$

Each of these beta values has an associated standard error. If the t-test associated with the beta value is significant (the value of the column labelled Pr(¿t) is less than .05) then the predictor is making a significant contribution to the model.The smaller the value of Pr(¡t) the bigger the contribution to the model. For this model the value WHOSIS_000001, t(54)=-4.94, p¡.001 is a significant predictor of HIV death and MDG_0000000033 t(54)=3.29 p=.002 is also significant but by less.

Standardised beta-values tell us the number of standard deviations by which the outcome will change as a result of one standard deviation in the predictor and are used on the final report (bigger absolute value = more important). QuantPsyc::lm.beta() function in the QuantPsyc package was used to find these.
WHOSIS_000001 = -0.5649774 MDG_0000000033 = 0.3763352
The F-ratio is used in to find the change when new predictors are added. The first model 2.4.1 causes R^2 to change from 0 to .208. The F-ratio which showed a value of F(1,61)=20.8, p<.001, which is significant
The 1 additional predictor changed F-ration to 14.24 2.4.2 F(2,54)=14.24 p<.001 which is significant.

2.6 Results of the findings

Due to a small number of measurements only 2 predictors were used in the multiple regression instead of the 4 originally envisaged.
Preliminary test showed the distribution of the dependent variable WHS2_198 was skewed and an attempt was made to address this but the skew still had a significant effect on the final model throwing results into doubt. .2.5.2
As described in the discussion 2.4 the life expectancy at birth is a (negative) significant predictor of HIV deaths. MDG_0000000033 is a (positve) significant predictor so antiretroviral therapy coverage increases were AIDS deaths are higher. The final multiple regression model is summarized in the table 3 in accordance with American Psychological guidelines Field (2009) [page. 301] (Step 2).

	δR^2	B	SE B	β	P
Step 1	.24				
Constant		386.78	68.03		<.001
WHOSIS_000001 - Life expectancy at birth (years)		-4.66	1.02	-.05	<.001
Step 2	.10				
Constant		338.72	66.31		<.001
WHOSIS_000001 - Life expectancy at birth (years)		-5.10	1.03	-.57	<.001
MDG_0000000033 - Antiretroviral therapy coverage among people with HIV infection		1.08	.39	.38	.002

Table 3: Multiple regression report

3 Binary Logistic

3.1 The objectives of the analysis and the context of the data being analysed.

A binary regression was done to see if it is possible to predict whether a country can be regarded as a low-income economy based on life expectancy and schooling level 2. The world bank describes a low-income economy as an economy which has a GNI less than 1005 dollars per year Wikipedia (2018*a*). The data collected for the multiple regression 2 was used in this binary regression. The GNI data for each country was split into 2 parts low-income economy (<1005 dollars)and above.

3.2 The data used

The full data 2.2 (158 entries) referenced in 2.2 was used. A new column 'income' represented the GNI data split into **0** = 'Low income countries (<1005 dollars)' and **1** = 'low-middle-income and above'.

3.3 Analysis in R and discussion of the output from the model.

The R command glm function with option 'family=binomial' was used to analyse the binary logistic data.

```
Call:
glm(formula = income ~ WHOSIS_000001, family = binomial, data = df)

Deviance Residuals:
Min        1Q    Median        3Q       Max
-2.04677   0.09069   0.24001   0.46772   2.23896

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)   -15.93857    3.41153  -4.672 2.98e-06 ***
WHOSIS_000001   0.25945    0.05301   4.894 9.88e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 127.959  on 119  degrees of freedom
Residual deviance:  81.984  on 118  degrees of freedom
(5 observations deleted due to missingness)
AIC: 85.984

Number of Fisher Scoring iterations: 6
```

Binary regression was initially done with one predictor 'WHOSIS_000001'. The residual deviance 81.98 is much less than the null deviance 127.96 indicating the model is predicting the outcome variable more accurately. The chisq probability was calculated using pchisq()

as 1.197509e-11 which is much less than 0.05.
i.e. $\tilde{\chi}^2(1)$=46.00,p<.001
The z values are significantly different from zero which means the b value is making signficant contribution (the significance of the z-statistic is less than .05).
B(WHOSIS_000001)=0.26,z=4.90,p<.001 Note(Wald=Z^2) 4
Binomial regression was also attempted with a second predictor 'school' but analysis showed that 'school' added very little. The summary was.
B(WHOSIS_000001)=0.30,z=4.26,p<.001
B(school)=-0.26,z=-1.046,p=0.30 (*0.30>0.05*)
In the model with 2 predictors there were also many more 'missing' values . There were only 85 degreess of freedom as opposed to 118 degrees.

It was decided to only use 1 predictor WHOSIS_000001 in the rest of this report.

The first column in 3.3 gives estimates of the β-values to be used in the binary regression equation. Therefore we can define the model as follows.

$$\widehat{Y} = \frac{\exp(\beta + \beta_1 WHOSIS_000001}{1 + \exp(\beta_0 + \beta_1 WHOSIS_000001} \quad (3)$$

$$= \frac{\exp(-17.26 + 0.30 WHOSIS_000001}{1 + \exp(-17.26 + 0.30 WHOSIS_000001} \quad (4)$$

Some statistical data was collected for the model such as the binary equivalent values or R^2
Pseudo R^2 for logistic regression
Homer and Lemeshow R^2 0.359
Nagelkerke R^2 0.485
The confidence interval did not cross 1.

Residuals were examined. Three studentiszed residuals were found to be above 2 but all dfbeta's were below 1 and Leverage values all close to 0.018 so it was decided to make no modifications.

```
# A tibble: 3 x 3
leverage studentized.residuals dfbeta[,"(Intercept)"] [,"WHOSIS_000001"]
<dbl>                  <dbl>                    <dbl>              <dbl>
1   0.0404                  2.11                     1.35            -0.0202
2   0.0363                  2.33                     1.43            -0.0215
3   0.0155                 -2.07                     0.546           -0.0093
```

3.4 Results of the findings

The binary regression with a single predictor was created. It showed with good statistical significance that living in low income countries means you have a lower life expectancy. A summary of the binary regression model is in 4.

		B	S.E	Wald	df	Sig	Exp(P)
Step 1	Constant	-15.94	3.41	21.87	1	.000	.000
	WHOSIS_000001 - Life expectancy	.259	.053	23.95	1	.000	1.30

Table 4: Binary regression report

References

Burrage, A., Patel, M., Mirkovic, K., Dziuban, E., Teferi, W., Broyles, L. & Rivadeneira, E. (n.d.).

Delaney, T. (2019), 'Statistics for data analytics', `https://moodle.ncirl.ie/course/view.php?id=1633`. Accessed: 2019-03-20.

Field, A. (2009), *Discovering statistics using R*, Sage publications.

Mansori, K., Ayubi, E., Shadmani, F. K., Hanis, S. M., Khazaei, S., Sani, M., Moradi, Y., Khazaei, S. & Mohammadbeigi, A. (2017), 'Estimates of global hiv/aids mortality, prevalence and incidence rates, and their association with the human development index', `www.bmrat.org/index.php/BMRAT/article/view/181`. Accessed: 2019-04-01.

Wikipedia (2017*a*), 'Gross national income'. [Online; accessed 02/04/2019].
URL: *https://en.wikipedia.org/wiki/Gross_national_income*

Wikipedia (2017*b*), 'List of countries by hiv/aids adult prevalence rate'. [Online; accessed 05/04/2019].
URL: *https://en.wikipedia.org/wiki/School_life_expectancy*

Wikipedia (2018*a*), 'New country classifications by income level: 2017-2018'. [Online; accessed 05/04/2019].
URL: *https://blogs.worldbank.org/opendata/new-country-classifications-income-level-2017-2018*

Wikipedia (2018*b*), 'School life expectancy'. [Online; accessed 05/04/2019].
URL: *https://en.wikipedia.org/wiki/School_life_expectancy*