Source: https://nacchocommunique.com/tag/social-determinants-of-health/

Summary and Assignment Support

Zi Ye

ENVS225

Exploring the Social World

UNIVERSITY OF LIVERPOOL

Quantitative Block

- Week 7: Intro to R for statistics
- Week 8: Correlation & Multiple Linear Regression for numeric variables
- Week 9: Correlation & Multiple Linear Regression for qualitative variables
- Week 10: Logistic Regression
- Week 11: Data visualisation
- Week 12: Wrap up

Regression Model

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

Scale/Continuous variables

Scale/Continuous variables

Dummy variables

Y: What is average % of people with long-term illness in the district?

X1: % of male
X2: % of no qualification
X3: % of higher professional

X4: Region

Logistic Regression

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Log odds

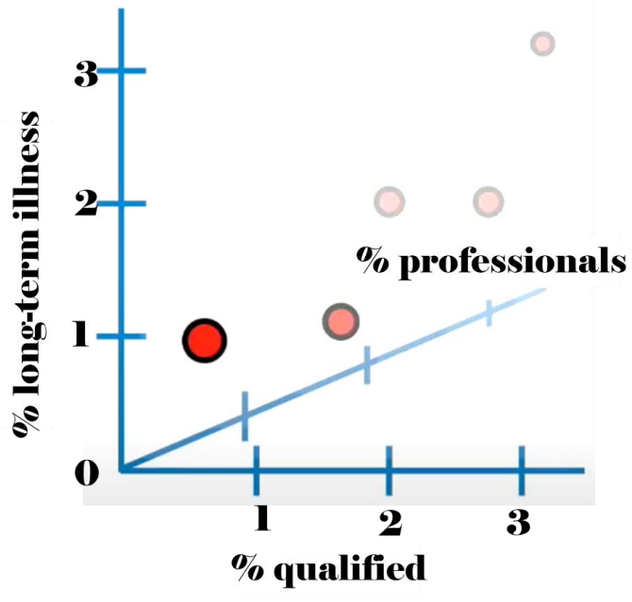
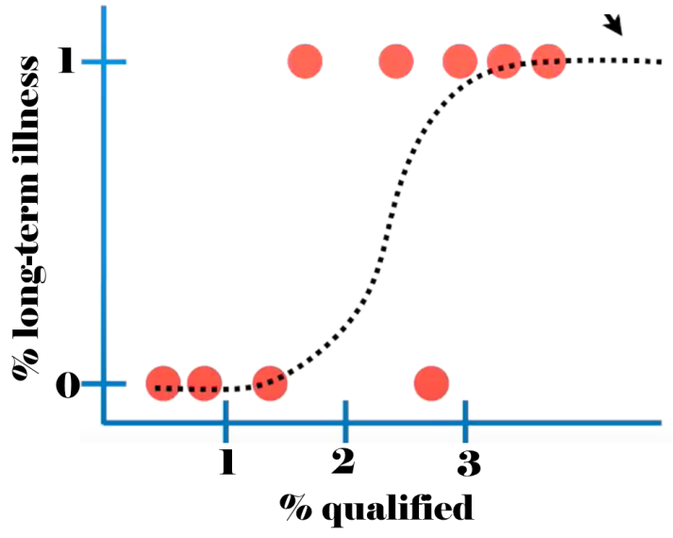
Categorical Variables
*binary

Scale/Continuous variables

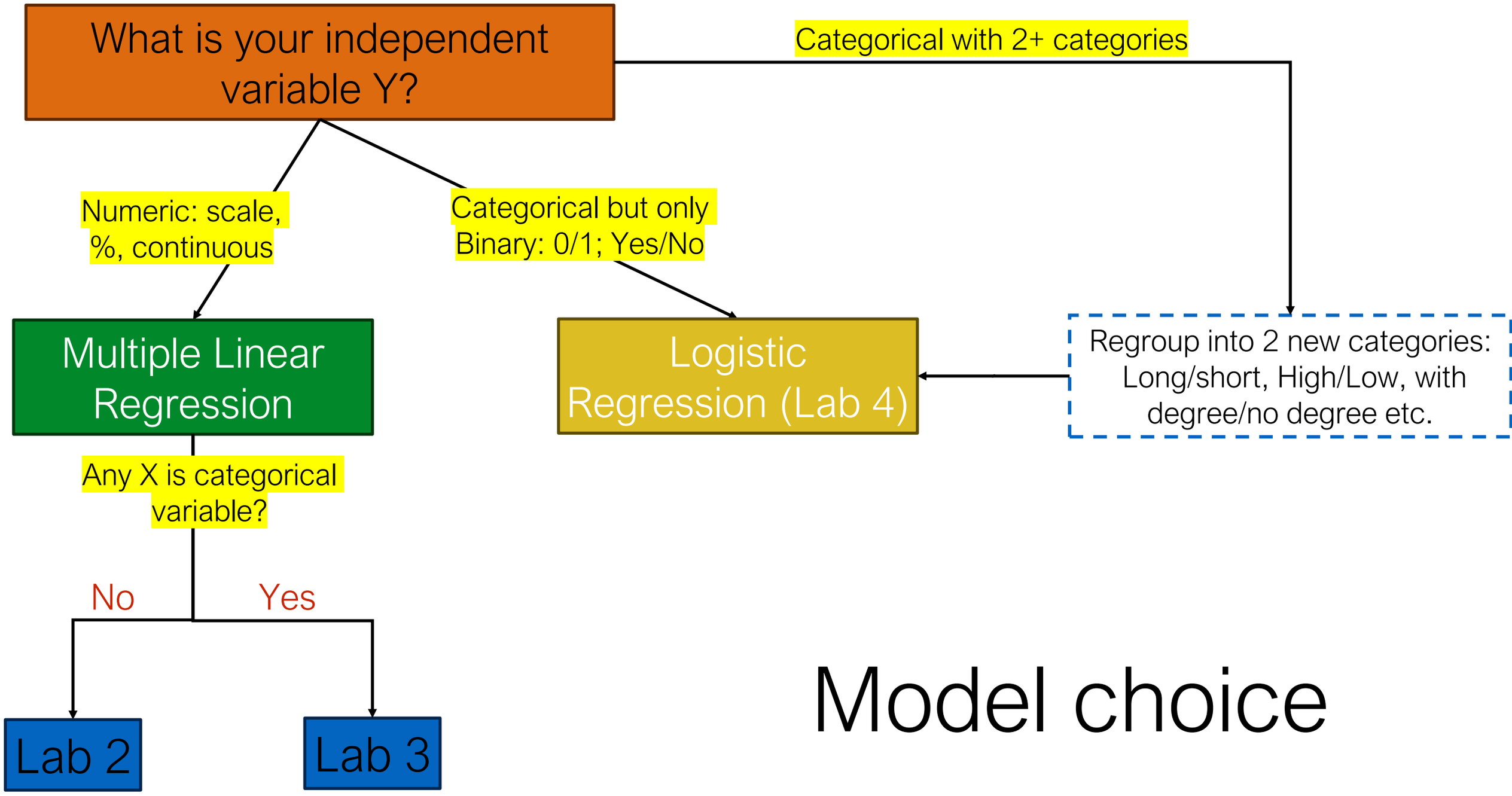
Dummy variables

p: Whether the person is willing to commute long distance?

X1: Sex
X2: NSSEC (higher managers, higher professional, routine occupation

	Multiple Linear Regression	Logistic Regression
Output variables (dependent/response)	Continuous/Scales (e.g. Rate, Age, Distance, Height…)	Categorical (e.g. Yes/No, Male/Female, Win/Not win)
Output to predicted…	Y: Mean of the target variable at the given values of the input variable	Log Odds: The probability of the particular levels of the given values of the input variable
Solve problems	Regression	Classification
Practical	What is the average long-term illness rate (%) in Liverpool?	Do you willing to commute long distance?
		

Model choice



Interpretation

Intercept

Slope coefficient 1

Slope coefficient 2

Slope coefficient 3

Error

Dependent variable

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

$logit(p)$

Independent variable 1

Independent variable 2

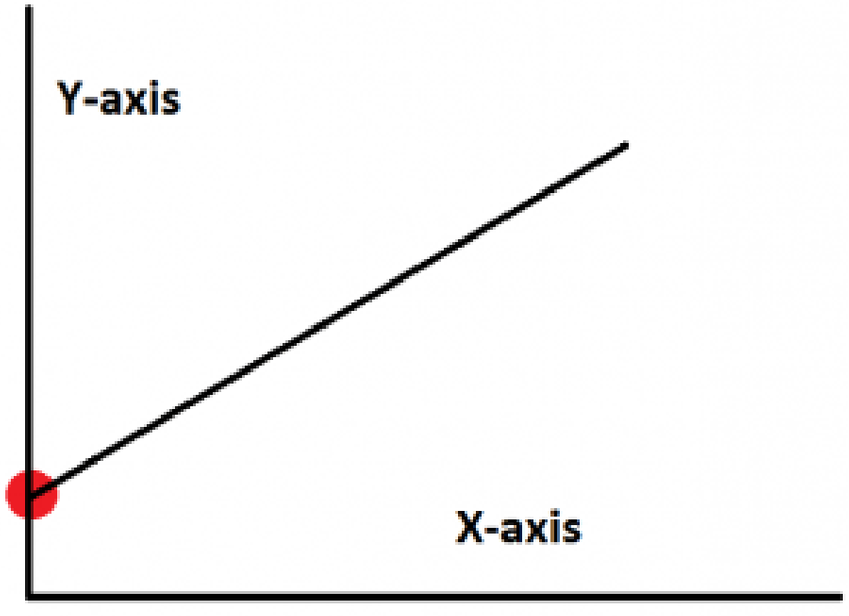
Independent variable 3

Overall Model Fit (R^2)

R-Square / Adjusted R-square: the proportion of variance in the dependent variable (science) which can be predicted from the independent variables

Intercept (Constant)

- The predicted value of Y/Log-odds when all other variables are 0.



P-value (Sig.)

- help to determine whether the relationships that you observe in your sample also exist in the larger population.

- If the p-value of a coefficient is smaller than 0.05, the coefficient is statistically significant. *You can say that the relationship between this independent variable and the outcome variable is statistically significant.*

- If the p-value of a coefficient is larger than 0.05, the coefficient is not statistically significant. *You can say or conclude that there is no evidence of an association or relationship between this independent variable and the outcome variable.*

Coefficient βs

- The estimated change in the Y/Log-odds for one unit change in X_i, holding all other variables constant.

```
Call:
lm(formula = pct_Long_term_ill ~ pct_Males + pct_No_qualifications +
    pct_Higher_manager_prof + Region_label, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2963 -0.9090 -0.1266  0.8168  5.2821

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      41.54134    5.22181   7.955 1.95e-14 ***
pct_Males                        -0.75756    0.10094  -7.505 4.18e-13 ***
pct_No_qualifications             0.50573    0.03062  16.515  < 2e-16 ***
pct_Higher_manager_prof           0.08910    0.03674   2.426  0.01574 *
Region_labelEast Midlands         1.14167    0.35015   3.260  0.00121 **
Region_labelEast of England      -0.01113    0.33140  -0.034  0.97322
Region_labelNorth East            2.70447    0.49879   5.422 1.03e-07 ***
Region_labelNorth West            2.64240    0.35468   7.450 6.03e-13 ***
Region_labelSouth East            0.48327    0.30181   1.601  0.11013
Region_labelSouth West            2.62729    0.34572   7.600 2.22e-13 ***
Region_labelWest Midlands         0.91064    0.37958   2.399  0.01690 *
Region_labelYorkshire and the Humber  1.03930    0.41050   2.532  0.01174 *
Region_labelWales                 4.63424    0.41368  11.202  < 2e-16 ***
Region_labelScotland              0.46291    0.38916   1.189  0.23497
Region_labelNorthern Ireland      0.55722    0.42215   1.320  0.18762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.394 on 391 degrees of freedom
Multiple R-squared:  0.8298,    Adjusted R-squared:  0.8237
F-statistic: 136.2 on 14 and 391 DF,  p-value: < 2.2e-16
```

Assignment

Assessment

Deadline: Tuesday 7th January 2025. **Word count**: 2000 words - including tables, excluding references.

The assignment **Data Exploration and Analysis** consists of writing a research report using one of the regression techniques learned during the module. The basic idea is to put in practice the methods learned during the quantitative block of the module. You are required to apply a linear or logistic regression model to the data provided for the module. The report needs to include the following sections (in brackets, % of the whole lenght):

- Introduction (5%).
- Literature Review (20%).
- Methods and data (30%).
- Results and discussion (40%).
- Conclusion (5%).
- Reference List.

Methodology

For sure

Regression model

Multiple Linear Regression

Or

Logistic Regression

Please make sure you use one regression model, and not just one X (independent variable)

Structure

- 1. Introduction
- 2. Literature Review
- 3. Methodology
- 4. Results and Discussion
- 5. Conclusion
- References

Research Question

- Make sense
- Knowledge gap
- Location: national, regional, local …

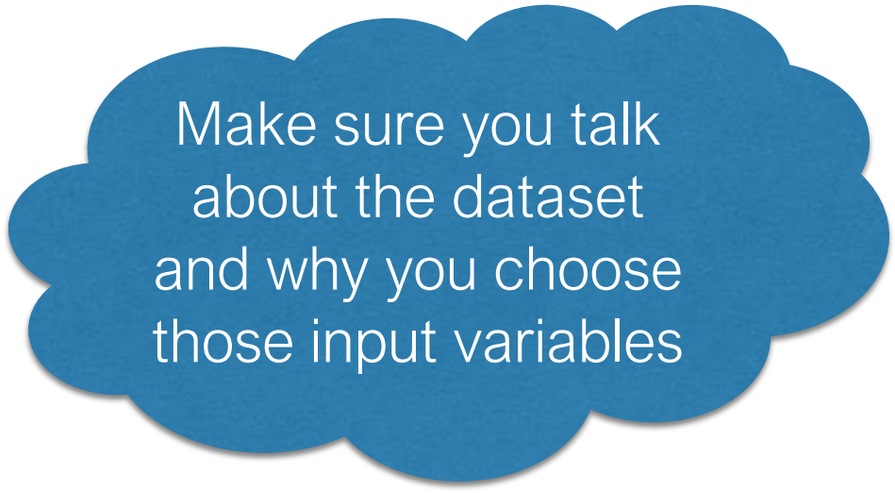
Examples from practical

- **How do local factors affect residents' health in England and Wales?**
- **What is the average long-term illness rate in Liverpool?**
- **How does health vary across regions in the UK?**
- **Who is willing to commute long distances?**

Methodology: dataset

Employing a novel dataset, i.e. not employed during the practical sessions, for the assignment will be awarded with a higher grade. For example, the quantitative dataset from Secondary datasets for Human Geography and Planning Students: 202425-ENVS203.

- 1. 2021 UK Census Data
- 2. 2021 Annual Population Survey
- 3. Family Resource Survey 2016-17
- 4. 2011 Census Sample of Anonymised Records (SAR.sav)



Methodology: descriptive statistic

For continuous variables

Name of variable	Description of variable	Minimum value	Maximum value	Mean	Standard deviation

For categorical data

Name of variable	Description of variable	Number of unique values	*Frequency of each unique value

Wisely use Practical Lab 5 Data Visualisation to help you describe the dataset you used.

Results and Discussion

- *Overall Model Fit (R^2)
- *P-value (Sig.)
- *coefficient βs
- Intercept (Constant)

```
Call:
lm(formula = pct_Long_term_ill ~ pct_Males + pct_No_qualifications +
    pct_Higher_manager_prof + Region_label, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2963 -0.9090 -0.1266  0.8168  5.2821

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       41.54134    5.22181   7.955 1.95e-14 ***
pct_Males                         -0.75756    0.10094  -7.505 4.18e-13 ***
pct_No_qualifications              0.50573    0.03062  16.515  < 2e-16 ***
pct_Higher_manager_prof            0.08910    0.03674   2.426  0.01574 *
Region_labelEast Midlands          1.14167    0.35015   3.260  0.00121 **
Region_labelEast of England       -0.01113    0.33140  -0.034  0.97322
Region_labelNorth East             2.70447    0.49879   5.422 1.03e-07 ***
Region_labelNorth West             2.64240    0.35468   7.450 6.03e-13 ***
Region_labelSouth East             0.48327    0.30181   1.601  0.11013
Region_labelSouth West             2.62729    0.34572   7.600 2.22e-13 ***
Region_labelWest Midlands          0.91064    0.37958   2.399  0.01690 *
Region_labelYorkshire and the Humber  1.03930    0.41050   2.532  0.01174 *
Region_labelWales                  4.63424    0.41368  11.202  < 2e-16 ***
Region_labelScotland               0.46291    0.38916   1.189  0.23497
Region_labelNorthern Ireland       0.55722    0.42215   1.320  0.18762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.394 on 391 degrees of freedom
Multiple R-squared:  0.8298,    Adjusted R-squared:  0.8237
F-statistic: 136.2 on 14 and 391 DF,  p-value: < 2.2e-16
```

Talk according to your table.

Others

- More than 5 references, uniform style (Chicago, APA, Harvard), ENVS203 literature management

- Use Figure 1, Figure 2, Table 1 … with your Graph/Table. Mention them in the text.

- Interpret your results and discuss with relevance to your literature review – use citations!

- Earn points for illustrations (graphs/maps/charts) in discussion part!

Have a nice Christmas break!

Zi Ye zi.ye@liverpool.ac.uk
Gabriele Filomena gabriele.filomena@liverpool.ac.uk

ENVS225
Exploring the Social World

UNIVERSITY OF LIVERPOOL