Modelling Unemployment in the United Kingdom with Census Data

201374125[a]

[a]Department of Geography and Planning, University of Liverpool, Liverpool, L69 7ZX

This version was compiled on January 8, 2019

Unemployment in the UK is still above that of many other developed countries, and shows significant regional variation. This paper explores the use of a model to predict the unemployment rate, and assess what causes varying levels of unemployment.

census | unemployment | united kingdom

1. Introduction

Unemployment in the United Kingdom has been steadily decreasing following the 2008 recession (ONS, 2018). However, despite an increase in employment, the rate of increase is slow compared with rising employment following previous recessions in the 1980s and 1990s, and UK unemployment still exceeds that of many other developed countries (Bell and Blanchflower, 2010).

Regional variation in unemployment had remained persistent for many years prior to the recessions following the reduction in industry and mining in the North of England and Wales. This had led to a loss of jobs and higher unemployment, compared with the South East and Midlands, where newer industries were more common (McCormick, 1997). However, following the two recessions in the late 20th century the regional unemployment variation was almost eliminated by 1993 (Gordon, 1995).

Subsequent to the 2008 recession, regional variation in unemployment is once again more prominent, with higher overall unemployment in the North (See Figure 1). This article aims to make explicit links between unemployment in the United Kingdom, and other socio-economic factors that are observed to significantly affect the regional variation in unemployment. Unemployment data is taken from the 2011 census which is a self assessment and may be subject to some subjectivity as to what constitutes unemployment. The true level of unemployment in the UK from 2011 ranges from 7.8% to 8.5% (ONS, 2018), whereas the mean observed in the 2011 Census data is 4%. It is however assumed that this inconsistency is down to definitional differences and will likely not affect the observed regional variation.

This article first assesses the level of regional variation in unemployment in the UK, focusing primarily on the divide in the North and South. The associations between unemployment and other variables in the dataset are then assessed through correlation, and through data exploration and literature, five predictor variables are chosen to use in a regression model. A best model will be selected and used to interpret what the model shows.

2. Literature Review

2.1. Why does unemployment matter?. An increase in unemployment is often associated with an increase in crime (Ward and Carmichael, 2001; Jennings, Farrall and Bevan, 2012), and unemployed individuals are more likely to be victims of crime. Workers who are unemployed are at risk of losing the skills required to return to work, in particular for those who are long term unemployed, finding work becomes more challenging (Machin and Manning, 1999).

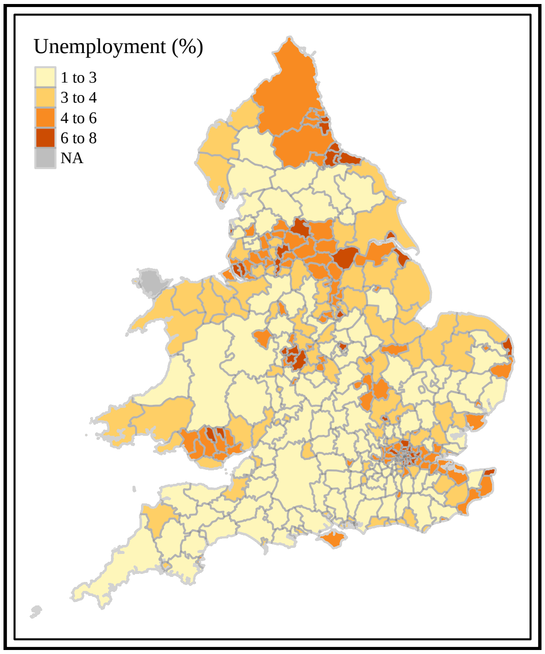

Fig. 1. Unemployment by District with Fisher Jenks Breaks

The unemployed are overall unhappier than those in work (Blanchflower and Oswald, 2004), and often have higher levels of mental stress and depression (Darity and Goldsmith, 1996). In addition Daly et al. (2008) found that suicide rate among the unemployed is significantly higher than those in work, notably those who are physically unable to work had the highest suicide risk overall.

2.2. Evidence for socio-economic factors effecting unemployment. Long-term unemployment (LTU) within most countries varies between both men and women, with men typically more likely to be LTU, older workers less likely to be LTU, and the less educated more likely to be LTU (Machin and Manning, 1999).

Macintyre, Maciver and Sooman (1993) state that when assessing socio-economic relationships, and considering the reasoning behind regional variation, it is important to understand the difference between compositional effects and contextual effects. Compositional effects are due how people vary over space, where people living in similar areas share the same characteristics that make them more likely to be unemployed. Contextual effects occur when unemployment results as an effect of where they live, and not purely due to an individuals own characteristics. For example, the North South divide in the United Kingdom (the divide in deprivation; Green, 1988) should be considered a contextual effect, as people with similar characteristics are more deprived in the North compared with the South. This article aims to primarily focus on factors that are considered contextual effects based on the literature available.

Lack of personal transport is shown to limit access to essential facilities in areas with poor public transport (Lucas, 2012). In particular, job opportunities are limited through a lack of means to reach certain employment locations.

Illness and unemployment both strongly link together in terms of socio-economic deprivation, and areas in which individuals live where there are higher levels of illness tend to be more deprived (Shouls, Congdon and Curtis, 1996). Heffernan and Pilkington (2011) found that mental illness in particular is associated with high levels of unemployment, where individuals may not be fit for normal employment, and that while there is help available for people who are incapable of normal work in the UK, this help is very limited, however the scope of the paper does not fully assess its impact on employment among those with mental illness.

Brown and Sessions (1997) found that unskilled males living in council accommodation are at the highest risk of unemployment, notably in the North of England and Wales. This observed variation in unemployment they attribute to regional variation in living cost, housing cost and industrial job opportunities.

Despite the concern that immigration is a threat to UK job prospects for British born workers, evidence suggests that immigration has very little effect on the employment opportunities (Dustmann, Fabbri and Preston, 2005), and in fact, all immigration groups have a much greater difficulty in gaining employment than White British persons (Frijters, Shields and Price, 2005). It is suggested this results due to a bias against hiring persons not of British descent in some areas of work, or cultural differences making it difficult to hire certain ethnic groups.

Few papers work with UK Census data with regards to modelling unemployment, although some model long term illness, (see Shouls, Congdon and Curtis, 1996; Charlton, Wallace and White, 1994), at present, there are no articles that model unemployment with 2011 census data.

3. Methodology

The dataset analysed in this article comes from the 2011 Census (27th March 2011), collected by the UK government for the Office for National Statistics (ONS, 2018). Data was collected from every household through a questionnaire in the post, which was then completed and returned.

The Census data was aggregated into local authority districts, with 348 in total, results were given as a percentage response per district. At the time of the Census the population of England and Wales was 56,075,912 (ONS, 2018).

3.1. Variation in Unemployment. Spatial variation in unemployment was preliminarily identified by a choropleth map (Figure 1), in addition a Kruskal-Wallis Test was conducted to examine the spatial variation in unemployment by 'SuperRegion' (data is not normal $Shapiro-Wilk$, $p < 0.05$; McCrum-Gardner, 2008). A Pairwise Wilcoxon test was used to determine between which regions there was significant spatial variation (Mehta, Patel and Tsiatis, 1984).

3.2. Correlation with other variables. Correlation between unemployment and all other variables was established through Spearman's rank correlation coefficient as data was not normal ($Shapiro-Wilk$, $p < 0.05$). The highest significant positive or negative correlations, alongside examples in literature were used to choose five predictor variables to use in a regressive model.

3.3. Backwards elimination. With the predictor variables identified, a backwards elimination approach was chosen to fit the most appropriate model. First, a Crawley maximal model was created including all predictor variables, if any of the variables had a statistically insignificant effect on the model ($ANOVA$, $p > 0.05$) it was removed. This was be repeated until the model only contained statistically significant predictor variables, as per Crawley (2015):

> *"Keep removing terms from the model... until the model contains nothing but significant terms. This is the minimal adequate model."*

The model was constructed through multivariate ordinary least squares regression, using the `lm()` function in *R* (linear model). Each model iteration was checked with an analysis of variance to report the total sum of squares for the model, *R* function `aov()`. In addition, an anova was used to check that the removal of a variable in a model iteration was significantly different to the previous model. If the removal was significantly different, but the analysis of variance gave a value that was reasonably similar to the last model, the variable was still removed for simplicity. The r^2 value was checked between each model iteration, and judged independently, considering model simplicity over a small increase in the model r^2 (*R* function: `summary(model)$r.squared`).

3.4. Model Validation. Once the Crawly minimal adequate model was achieved, it was checked for outcome variable normality, model homoscedasticity, multicollinearity and model linearity to determine whether the model accurately represents the relationship between the outcome variable and chosen predictor variables.

3.5. Standardised Regression Coefficients. As the sole purpose of this article was to directly compare the impact predictor variables have on the outcome, Unemployment, standardised (or $beta$) regression coefficients allowed for comparison of the variables through changes in standard deviation, rather than changes in percentages (Nick Horton, 2010). This analysis utilised the `lm.beta()` function from the `QuantPsyc` *R* package.

4. Results

4.1. Variation in Unemployment. Unemployment appears to follow a certain level of positive spatial autocorrelation, with low unemployment towards the South West of England but higher levels towards parts of the South East and across the North (Figure 1). Notably there are high levels of unemployment in major cities, including Hull (8.0%), Middlesbrough (7.6%), Birmingham (7.1%), and surrounding the City of London.

The average unemployment by 'Super Region' ('North West', 'Central' and 'South East') was assessed through A Kruskal-Wallis Test, to examine the spatial variation in unemployment (Unemployed data does not conform with a normal distribution, Shapiro-Wilk normality, $p < 0.05$). There was a significant difference between regions ($Chi-Square = 24.394$, $p < 0.05$, $df = 2$). A Wilcox rank sum test confirmed that the 'North West' region is significantly different to the other regions (Pairwise Wilcoxon rank sum test, North West - South East $p < 0.05$, North West - Central $p < 0.05$).

Table 1. Spearman's rank correlation coefficients for all variables in relation to Unemployment.

Variable	Rho	Lower CI †	Upper CI †
No_Cars	**0.83 ****	**0.64**	**0.75**
Two_plus_Cars	-0.79 **	-0.78	-0.68
Social_Rented	**0.69 ****	**0.64**	**0.74**
Owner_occupied	-0.67 **	-0.66	-0.52
Crowded	0.63 **	0.46	0.61
Professionals	-0.62 **	-0.59	-0.43
No_Quals	**0.57 ****	**0.47**	**0.62**
Age_65plus	-0.51 **	-0.59	-0.43
Lone_persons	0.41 **	0.24	0.42
Students	0.41 **	0.19	0.38
Couple_with_kids	-0.34 **	-0.43	-0.24
illness	**0.34 ****	**0.22**	**0.41**
White_British	**-0.30 ****	**-0.51**	**-0.34**
Flats	0.29 **	0.20	0.39
Private_Rented	0.26 **	0.20	0.39
FT_Employees	-0.20 **	-0.32	-0.12
UK_Born	-0.17 **	-0.41	-0.22

* Significant at the 0.05 level;
** Significant at the 0.01 level;
*** Significant at the 0.001 level;
† 95% Confidence Interval

4.2. Correlation with other variables. As the 'Unemployed' data does not conform with a normal distribution (see above) a Spearman's rank correlation coefficient was ran for all variables in relation to unemployment. Table 1 shows the results of these tests.

The five chosen predictor variables were chosen based on the results of this test and literature from the literature review. They are 'No Cars', significant high positive correlation ($Rho = 0.83$, $p < 0.01$), 'No Quals', significant medium positive correlation ($Rho = 0.57$, $p < 0.01$), 'Social Rented', significant positive correlation ($Rho = 0.69$, $p < 0.01$), 'Illness', significant slight positive correlation ($Rho = 0.34$, $p < 0.01$), and 'White British', significant slight negative correlation ($Rho = -0.30$, $p < 0.01$).

4.3. Construction of the Model. The first model was constructed utilising all five predictor variables, called the Maximal Model (Crawley, 2015). This model was constructed as:

```
Unemployed = No Cars + No Quals + Social
Rented - White British + illness
```

Model t statistics are the estimates ($\hat{\beta}_i$) divided by their standard errors ($\hat{\sigma}_i$), therefore $t_i = \frac{\hat{\beta}_i}{\hat{\sigma}_i}$. These are shown for each predictor variable in this model on Table 2, which indicates the individual significance of each parameter, with the p significance values indicted as asterisks. In this model all terms are significant (at the $p < 0.05$ level) excluding Illness ($p > 0.05$) and by the definition of the Minimal Adequate Model, Illness must be removed (Crawley, 2015). The Akaike's Information Criterion (AIC) was used to measure the fit of a model, the addition of any term to a model will make the AIC go down, but for model simplicity and increased explanatory power, if a term did not change the AIC by a large amount, it was removed (Crawley, 2015). The AIC for the maximal model was: 549.50, after removing 'Illness' this changed to 550.47, a very slight increase. In addition, the degree of fit of the model was assessed through the r^2 value, this was multiplied by 100 to give a percentage (shown in Table 2). The lower r^2, the worse the model fit. The r^2 for the maximal model was 81.98 and fell only slightly to 81.83 with the removal of Illness. The next highest p value in this model was 'Social Rented', while this was significant, it was removed to see whether its removal significantly reduced the effectiveness of the model. The removal of 'Social Rented' significantly altered the model ($ANOVA$, $p < 0.05$), increased the AIC to 581.12 from 550.47, and reduced the r^2 to 80.05 from 81.83 so was considered an integral part of the model, and kept.

Table 2. Multiple Regression Model showing t values

	Model 1	Model 2	Model 3
No Cars	6.40 ***	6.38 ***	11.88 ***
No Quals	15.01 ***	22.60 ***	23.29 ***
White British	-7.34 ***	-10.01 ***	-10.45 ***
Social Rented	5.00 ***	5.81 ***	
Illness	-1.72		
R Squared	81.99	81.83	80.05

* Significant at the 0.05 level;
** Significant at the 0.01 level;
*** Significant at the 0.001 level

4.4. Model Validation. The distribution of the outcome variable 'Unemployed' is not normal (see Figure 2a; skew: 0.744), because of this, a `logit()` transformation was performed which reduced the skew to 0.0427. Logit is known to improve distribution in percentage data, as logit removes the cap of 0 to 100 values and extends the highest 80% and lowest 20% values (Manning, 2002). While the distribution of the model errors showed little skew (skew = 0.25), Figure 2c shows that some studentised residuals (standardised model errors) fall outside the straight line, confirming that either the outcome variable (unemployment) or any of the explanatory variables may have been in need transforming.

However, once a logit transformation was performed, the residual skew was increased dramatically, suggesting this transformation was not appropriate, alternative transformation such as log or square root including transformation of predictor variables did not solve this issue without extreme model complication, therefore in the effort to retain model simplicity, the small skew observed in the model errors suggested that the standard model should be appropriate.

4.5. Homoscedasticity. The studentised residuals show only slight homoscedasticity (Figure 2b), as there is an almost straight line, indicating a constant error variance. A Non Constant Variance test also gave this result (NCV Test, $Chi = 26.07$, $df = 1$, $p < 0.05$). The suggested power transformation from the `spreadLevelPlot()` R function only further increased homoscedasticity, therefore no power transformations were used.

4.6. Multicollinearity. Multicollinearity was measured through the square root of the Variance Inflation Factor (VIF). The mean average VIF was above 2 ($\overline{VIF} = 2.36$) which indicated that there may be a multicollinearity problem (Kabacoff, 2015), VIF values were provided by the `vif()` function from the `car` R package. The VIF for % No Cars was the highest at 3.19. This means that the variance of the 'No Cars' regression coefficient is 119% larger than it would be if 'No Cars' was uncorrelated with the other explanatory variables in the model ('No Quals', 'White British', 'Social Rented'), or that the standard error is 79% larger ($\sqrt[2]{VIF}$), the VIF of Social

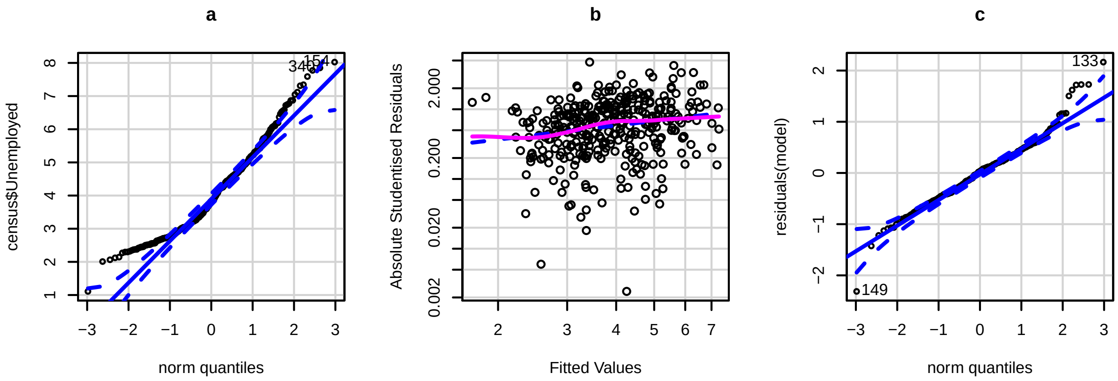

Fig. 2. (a) Distribution of the outcome variable. (b) Studentised Residuals of the model . (c) Homoscedasticity of the model.

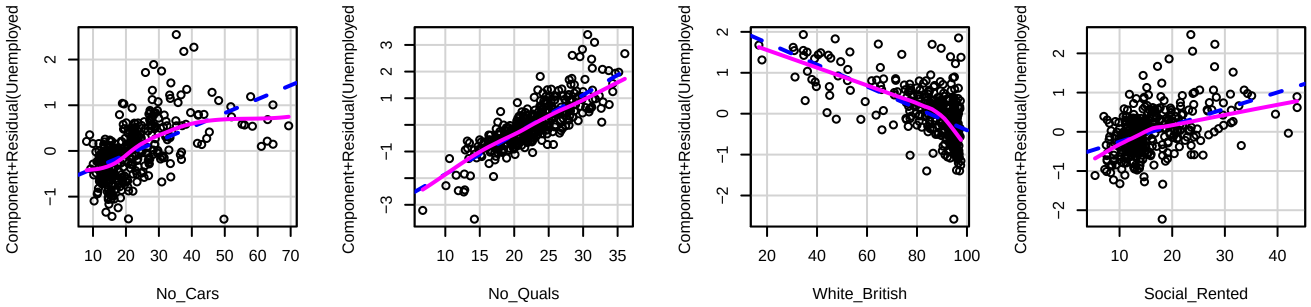

Fig. 3. Partial Residual Plots for Each Predictor Variable.

Rented was also above 2.5 (2.58). Paul Allison (2012) suggests that if predictor variables show a VIF above 2.5, there is some cause for concern.

However, Paul Allison (2012) states that with very large samples, the effects of multicollinearity is reduced greatly, although shouldn't be ignored completely. Due to the removal of 'No Cars' having a severely negative effect on the effectiveness of the model (reduced r^2 and increased AIC), this variable was kept as the data being worked with comes from a very large sample that consists effectively of the entire population of the UK.

4.7. Model Linearity. There was a departure from linearity between the outcome variable, and the predictor variables (Figure 3). Notably 'White British' trends downwards at higher values, as does 'No Cars' and 'Social Rented' at higher values. This issue stems from analysis intended for scale data being used on percentage data. As mentioned above, an attempt to use a `logit()` transformation on the data was unsuccessful, and introduced very large model complexity when attempting to transform outcome variables. Due to the reasonably linear correlation between the outcome variable and all the predictor variables, excluding extreme values, the model was considered to be adequate.

4.8. Standardised Regression Coefficients. All variables in this analysis have differing ranges, for example the mean value for Unemployed is 4% with a range of around 7%, while 'White British' has a mean of 85% with a range of 81%. These vast differences in the scale of the data makes direct comparison difficult without standardisation. Standardisation is commonly utilised when modelling regional demographics (e.g. Eames, Ben-Shlomo and Marmot, 1993).

Table 3. Standardised coefficients

Variable	Standard Deviation
No_Cars	0.26
No_Quals	0.61
White_British	-0.35
Social_Rented	0.21

Table 3 shows the results of this analysis, with the highest influence given by 'No_Quals' by which one standard deviation change results in a 0.61 standard deviation change in unemployment, and lowest, 'Social_Rented' with a 0.21 standard deviation change in unemployment.

5. Discussion of the Model

Unemployment significantly correlates with all socio-economic variables in the census dataset (Table 1), an observation that is noted in previous studies (Shouls, Congdon and Curtis, 1996). As mentioned above however, care has been taken to select predictor variables that represent contextual effects, and not compositional ones (Macintyre, Maciver and Sooman, 1993; Jones and Duncan, 1995). For example, although there is a significant correlation between Flats and Unemployment, this could be considered a compositional effect, as flats are typically cheaper than houses. However, from Figure 1 and analysis of the raw data, the highest levels of unemployment are concentrated within city centres, where there are bound to be fewer houses, and a larger number of flats, therefore it was considered a contextual effect.

Unemployment is considered a major issue in many cities within the UK (Diep, Drabble and Young, 2006), most notably in the North where former industries have been lost, and recent gentrification of city centres has only enhanced the wealth divide between the city and surrounding suburbs.

The standardised coefficients of the final model give insight into how much influence the predictor variables have on Unemployment by district (Table 3).The largest influence given by 'No Quals' suggests that regions with a higher proportion of people with no formal qualifications are more likely to be areas with higher rates of unemployment. This can be attributed to the components of demand that vary between industrial sectors, especially following a recession (Bell and Blanchflower, 2010). Following the 2008 recession, Bell and Blanchflower (2010) suggest that demand for construction and goods industries fell, and as such, areas in which these industries are more prominent have experienced an increased rate of unemployment. These industries typically do not require any formal qualification, and as now they no longer exist, these regions have fewer jobs for those lacking qualifications.

The impact of immigration on less-skilled native workers has been analysed with census data from the 1970s and 1980s (Joseph and Card, 1991), where a small increase in the rate of immigration was shown to slightly decrease the wages for unskilled workers, a suggestion that immigration may negatively impact the job prospects of British persons. The data analysed in this article however is more limited, and cannot assess wage variation, however Table 3 gives a result that indicates a lower rate of employment for regions in which there are fewer White British persons. This result mirrors analysis on more recent census data, where non White British persons are found to have much more difficulty in finding jobs (Frijters, Shields and Price, 2005). The length of time spent in the UK increase job search success, which Borjas (2000) attributes to an increase in language fluency over time, which at first hinders job search success.

An interesting observation with the results is that the predictor variable 'No Cars', while having a fairly significant effect on the standard deviation on Unemployment, the effect is much lower than its correlation would imply (0.83; Table 1). This suggests that the resultant correlation between No Cars and Unemployment may primarily be due resultant contextual effects, for example lack of money through unemployment, or unemployed persons primarily focused towards the city centre. However, social exclusion through lack of suitable and affordable public transport is considered an issue in the UK for persons without person transport (Lucas, 2012), where social exclusion is defined by Levitas et al. (2007) as:

> *"the lack or denial of resources, rights, goods and services, and the inability to participate in the normal relationships and activities, available to the majority of people in a society, whether in economic, social, cultural or political arenas."*

In particular Church, Frost and Sullivan (2000) define seven features in which poor transport results in social exclusion, in relation to unemployment; *economic exclusion:* the higher cost of travel may limit access to facilities and employment, thus impacting income.

Mccormick (1983) found council tenants (Social Rented) between 60% and 70% more likely to be unemployed than the average worker in other tenure when controlling for other variables, while Brown and Sessions (1997) found that during their study one in four council tenants were unemployed and suggest that those in council accommodation often have less financial constraints due to subsidised rents, while persons with a mortgage are subject to more stringent benefits. In addition they note that Council housing is typically concentrated within city areas or outer city areas where job prospects are more limited.

6. Conclusion

Defining what constitutes contextual effects over compositional ones is often open to interpretation and the predictor variables selected in this study likely do not fully reflect context (Macintyre, Maciver and Sooman, 1993). Most notably 'No Cars' was selected due to how this would limit job opportunity due to travel distance limitations. However, fewer people require cars in a city centre, where unemployment is high, and unemployed people are less likely to own a car due to the cost, independent of whether this reduces their employment opportunity.

As this data is percentage data, the techniques used are not fully appropriate, which has led to some complication in determining model validity. A further study may look to utilising methods that are able to handle percentages to further improve model validation techniques. However, as approached, the model validation revealed that overall this model gives a reliable fit.

Acknowledgments. Thanks to Robin Lovelace, Jakub Nowosad and Jannes Muenchow who recently released their fantastic book free and open source 'Geocomputation with R' which allowed me to produce Figure 1 in this article (Lovelace, Nowosad and Muenchow, 2019), with data from GDAL (GDAL, N.d.).

References

Bell, David and David Blanchflower. 2010. "UK Unemployment in the Great Recession." pp. 3–25.

Blanchflower, David G. and Andrew J. Oswald. 2004. "Well-being over time in Britain and the USA." *Journal of Public Economics* 88(7-8):1359–1386.

Borjas, George. 2000. "Introduction to"Issues in the Economics of Immigration'." I(January):1–15.

Brown, S and J G Sessions. 1997. "A profile of UK unemployment: regional versus demographic influences." *Regional studies* 31(4):351–66.

Charlton, John, Merryl Wallace and Ian White. 1994. "Long-term illness: results from the 1991 census." *Population Trends* (75):18–25.

Church, A, M Frost and K Sullivan. 2000. "Transport and social exclusion in London." 7.

Crawley, Michael J. 2015. *Statistics: An introduction using R (Second Edition).*

Daly, Mary C, Daniel J Wilson, Norman J Johnson and U S Census Bureau. 2008. "Relative Status and Well-Being: Evidence from U.S. Suicide Deaths." *Suicide* (May).

Darity, William and Arthur H Goldsmith. 1996. "Social Psychology, Unemployment and Macroeconomics." *Journal of Economic Perspectives* 10(1):121–140.

Diep, Martina, Stephanie Drabble and Craig Young. 2006. "Living with difference? The 'cosmopolitan city' and urban reimaging in Manchester, UK." *Urban Studies* 43(10):1687–1714.

Dustmann, Christian, Francesca Fabbri and Ian Preston. 2005. "The Impact of Immigration on the British Labour Market.".

Eames, M, Y Ben-Shlomo and M G Marmot. 1993. "Social deprivation and premature mortality: regional comparison across England." *Bmj* 307(6912):1097–1102.

Frijters, Paul, Michael A. Shields and Stephen Wheatley Price. 2005. "Job search methods and their success: A comparison of immigrants and natives in the UK." *Economic Journal* 115(507):359–376.

GDAL. N.d. "GDAL - Geospatial Data Abstraction Library.".

Gordon, Ian. 1995. "Migration in a Segmented Labour Market." *Transactions of the Institute of British Geographers* 20(2):139.

Green, A.E. 1988. “The North-South Divide in Great Britain: An Examination of the Evidence.” *Transactions of the Institute of British Geographers* 13(2):179–198.

Heffernan, John and Paul Pilkington. 2011. “Supported employment for persons with mental illness: Systematic review of the effectiveness of individual placement and support in the UK.” *Journal of Mental Health* 20(4):368–380.

Jennings, Will, Stephen Farrall and Shaun Bevan. 2012. “The economy, crime and time: An analysis of recorded property crime in England & Wales 1961-2006.” *International Journal of Law, Crime and Justice* 40(3):192–210.

Jones, Kelvyn and Craig Duncan. 1995. “Individuals and their ecologies: analysing the geography of chronic illness within a multilevel modelling framework.” *Health and Place* 1(1):27–40.

Joseph, Altonji and David Card. 1991. *The Effects of Immigration on the Labor Market Outcomes of Less-skilled Natives*.

Kabacoff, Robert I. 2015. *R in Action, Second Edition*.

Levitas, Ruth, Christina Pantazis, Eldin Fahmy, David Gordon, Eva Lloyd and Demi Patsios. 2007. “The Multi-Dimensional Analysis of Social Exclusion. Department of Sociology and School for Social Policy Townsend Centre for the International Study of Poverty and Bristol Institute of Public Affairs.” (January).

Lovelace, Robin, Jakub Nowosad and Jannes Muenchow. 2019. *Geocomputation with R*. CRC Press.

Lucas, Karen. 2012. “Transport and social exclusion: Where are we now?” *Transport Policy* 20:105–113.

Machin, Stephen and Alan Manning. 1999. “Chapter 47 The causes and consequences of longterm unemployment in Europe.” *Handbook of Labor Economics* 3 PART(3):3085–3139.

Macintyre, Sally, Sheila Maciver and Anne Sooman. 1993. “Area, Class and Health: Should we be Focusing on Places or People?” *Journal of Social Policy* 22(02):213.

Manning, Christopher D. 2002. “Probabilistic Syntax (Draft).” *Probabilistic Linguistics* (2001):1–40.

Mccormick, Barry. 1983. “Housing and unemployment in Great Britain.” *Oxford Economic Papers* 35:283–305.

McCormick, Barry. 1997. “Regional unemployment and labour mobility in the UK.” *European Economic Review* 41(3-5):581–589.

McCrum-Gardner, Evie. 2008. “Which is the correct statistical test to use?” *British Journal of Oral and Maxillofacial Surgery* 46(1):38–41.

Mehta, CR R, NR R Patel and AA a Tsiatis. 1984. “Exact significance testing to establish treatment equivalence with ordered categorical data.” *Biometrics* 40(3):819–25.

Nick Horton. 2010. “Generating standardized regression coefficients.”.

ONS. 2018. “Unemployment - Office for National Statistics.”.

Paul Allison. 2012. “When Can You Safely Ignore Multicollinearity?”.

Shouls, Susanna, Peter Congdon and Sarah Curtis. 1996. “Modelling inequality in reported long term illness in the UK: Combining individual and area characteristics.” *Journal of Epidemiology and Community Health* 50(3):366–376.

Ward, Robert and Fiona Carmichael. 2001. “Male unemployment and crime in England and Wales.” *Economics Letters* 73(1):111–115.

Code Appendix

201374125[a]

[a] Department of Geography and Planning, University of Liverpool, Liverpool, L69 7ZX

This version was compiled on January 8, 2019

Setup.

```r
# Required packages
library(QuantPsyc) # used to standardise coefficients
## note: load first as requres MASS which masks 'select' from dplyr
library(ggplot2) # graphical visualisations
library(ENVS450) # statistical functions
library(car) # more statistical functions
library(sf) # for processing geopackage files/shapefiles
library(tmap) # for plotting maps/choropleths
library(ggcorrplot) # corrolation plots
library(dplyr) # data manipulation
library(extrafont) # use additional fonts for visualisations
library(kableExtra) # addition kable functions
library(broom) # convert summaries into data.frames
library(rowr) # used for cbind.fill to add NA values when missing
```

```r
# Load the census dataset
census <- "./data/2011 Census.RData"
load(census)
```

Figure 1.

Read in Geopackage data.

```r
# read geopackage file for uk polys, choose layer that includes UK district names
# data from GDAL
uk = st_read("./data/gbr.gpkg", layer="gadm36_GBR_3", quiet=TRUE)
```

```r
# select only polygons from England and Wales (same as census data)
uk = subset(uk, NAME_1 == "England" | NAME_1 == "Wales")
# keep only names and polygons
uk = uk[ ,10, 17]
# 'NAME_3' same as 'District' from census
names(uk)[1] = "District"

# Rename certain districts to match census data
uk$District <- recode(uk$District, "Kingston upon Hull"=
                                   "Kingston upon Hull, City of",
                                   "London"="City of London",
                                   "Bristol"="Bristol, City of",
                                   "Durham"="County Durham",
                                   "Herefordshire"="Herefordshire, County of",
                                   "Rhondda, Cynon, Taff"="Rhondda Cynon Taf",
                                   "Suffolk coastal"="Suffolk Coastal",
                                   "Vale of Glamorgan"="The Vale of Glamorgan",
                                   "Saint Albans"="St Albans",
                                   "Saint Edmundsbury"="St Edmundsbury",
                                   "Saint Helens"="St. Helens")
```

```
# Merge polys to dataframe by District name, outer join to keep NAs
census_poly = merge(uk, census, by = "District", all.x=T)
```

Construct Figure 1.

```
# Map: fill by unemployment, change line aesthetics
uk_unemp = tm_shape(census_poly) +
  tm_fill(col = "Unemployed", title = "Unemployment (%)",
          n = 4, # number of bins
          style=’jenks’, # fisher jenks bins
          textNA = "NA",
          colorNA = "gray") + # NAs shown as grey in legend
  tm_borders(lwd = 1, alpha=0.3) + # add borders
  tm_layout(legend.format = list(digits = 0), # alter layout, no decimals
            frame.double.line = TRUE, # double frame
            inner.margin = 0.05, # stop map overlapping frame
            fontfamily = ’Liberation Serif’) # the best font

uk_unemp # plot choropleth map (Figure 1)
```

Correlation

Table 1.

```
# Create a matrix of all correlations between continuous variables
correlation.matrix = cor( census[ , -c(1:5) ], method="spearman")

# Sort by highest correlations in relation to Unemployed
corResults = cor.results(correlation.matrix, sort.by="abs.r", data=census,
                         var.name="Unemployed")
# Show only two numbers after the decimal
corResults = format(corResults, digits = 2)

# Combine significance column with r values (asterisks)
corResults$r = paste(corResults$r,corResults$sig.)

# Drop the columns of the dataframe
corResults = select(corResults,-c(’x’,’p.value’,’sig.’))

# Rename all columns
names(corResults) <- c("Variable", "Rho", "Lower CI †",
                       "Upper CI †")

# Create Table 1
kable(corResults, # align col 1 left, rest centred
      caption = "Spearman’s rank correlation coefficients for all variables
      in relation to Unemployment.", align=c(’l’, ’c’, ’c’, ’c’),
      # longtable stops splitting over pages, booktabs for formatting
      longtable = FALSE, booktabs = TRUE,  linesep = "", format = "latex") %>%
  # add the footnote
  footnote(general_title="",
           general = c("* Significant at the 0.05 level;",
                       "** Significant at the 0.01 level;",
                       "*** Significant at the 0.001 level;",
                       "† 95% Confidence Interval")
         ) %>%
  row_spec(c(1,3,7,12,13), bold = T) # add bold rows showing chosen variables
```

Models

Construct Model 1.

```
# Create maximal linear model
model1 <- lm(Unemployed ~ No_Cars + No_Quals + White_British + Social_Rented +
             illness, data=census)
summary(model1) # summary showing T values etc
AIC(model1) # Akaike’s Information Criterion
summary(model1)$r.squared * 100 # R Squared %
```

Construct Model 2.

```
# Create minimal adequate linear model
model2 <- lm(Unemployed ~ No_Cars + No_Quals + White_British + Social_Rented,
             data=census)
summary(model2) # summary showing T values etc
AIC(model2) # Akaike’s Information Criterion
anova(model1, model2) # compare the two models
summary(model2)$r.squared * 100 # R Squared %
```

Construct Model 3.

```
# Create minimal adequate linear model
model3 <- lm(Unemployed ~ No_Cars + No_Quals + White_British,
             data=census)
summary(model3) # summary showing T values etc
AIC(model3) # Akaike’s Information Criterion
summary(model3)$r.squared * 100 # R Squared %
anova(model2, model3) # compare the two models
```

Create Table 2 showing t values.

```
## Prepare data for Table 2

# tidy from broom package changes summary information into data.frames
m1 = tidy(summary(model1))
# select only the important columns
m1 = select(m1, term, statistic, p.value)
# remove (intercept)
m1 = m1[-1, ]
# repeat all this for other two models (tried a loop but was a bit complex)
m2 = tidy(summary(model2))
m2 = select(m2, term, statistic, p.value)
m2 = m2[-1, ]
m3 = tidy(summary(model3))
m3 = select(m3, term, statistic, p.value)
m3 = m3[-1, ]

# Change p values into vectors for each model
p1 = as.vector(m1$p.value)
p2 = as.vector(m2$p.value)
p3 = as.vector(m3$p.value)

# For each vector of p values, change from numeric to the significance asterisks
p1 <- symnum(p1, corr = FALSE, na = FALSE, cutpoints = c(0,
    0.001, 0.01, 0.05, 1), symbols = c("***", "**", "*", " "))
p2 <- symnum(p2, corr = FALSE, na = FALSE, cutpoints = c(0,
    0.001, 0.01, 0.05, 1), symbols = c("***", "**", "*", " "))
p3 <- symnum(p3, corr = FALSE, na = FALSE, cutpoints = c(0,
    0.001, 0.01, 0.05, 1), symbols = c("***", "**", "*", " "))

# Round statistic values to 2 digits and change to vector as well, keep zeros
t1 = sprintf(’%.2f’, m1$statistic)
t2 = sprintf(’%.2f’, m2$statistic)
t3 = sprintf(’%.2f’, m3$statistic)
```

```
# Join statistic values and p asterisks
mod1 = paste(t1, p1)
mod2 = paste(t2, p2)
mod3 = paste(t3, p3)

# Find R squared % for each model
m1r = summary(model1)$r.squared * 100
m2r = summary(model2)$r.squared * 100
m3r = summary(model3)$r.squared * 100

# Make vector of R squareds
r = c(m1r, m2r, m3r)
# Round them to 2 digits
r = round(r, 2)
# Change to character (so it can be appended to the statistic values)
# Stat value no longer numeric as they have asterisks on them
r = as.data.frame.character(r)
# Transform the dataframe
r = t(r)

# Column bind all statistic values and p values
table = cbind.fill(mod1, mod2, mod3, fill = NA)

# column names
columnn = c("Model 1", "Model 2", "Model 3")
# rename column names
colnames(table) = columnn
# same for r values
colnames(r) = columnn

# join by row and shared column names
table = rbind(table, r)
# rename row names of table
rownames(table) = c('No Cars', 'No Quals', 'White British', 'Social Rented',
                    'Illness', 'R Squared')
```

Plot Table 2.

```
# tell Kable to not show anything for NA values
options(knitr.kable.NA = '')
# Table 2
kable(table, digits = 2, caption = "Multiple Regression Model showing $t$ values",
      longtable = FALSE, booktabs = TRUE,
      linesep = '', format = "latex",
      col.names = c("Model 1",
                    "Model 2",
                    "Model 3")) %>%
  # add line after row 5
  row_spec(5, hline_after = T) %>%
  column_spec(4, color = "gray") %>%
  # footnote for significance
  footnote(general_title="",
           general = c("* Significant at the 0.05 level;",
                       "** Significant at the 0.01 level;",
                       "*** Significant at the 0.001 level")
         )
```

Model Validation

Skew and transformations.

```
# Define the final model again (just in case)
model <- lm(Unemployed ~ No_Cars + No_Quals + White_British + Social_Rented,
            data=census)
skew(model$residuals) ## little skew of residuals (0.25)
skew(census$Unemployed) # Some skew 0.744
skew(logit(census$Unemployed)) # Much lower skew (0.04)
```

Figure 2: Studentised Residuals.

```
# create a matrix for multiple graph plots
layout(matrix(1:4, ncol = 3, nrow=1))
# Show the normality of the outcome, and call it plot a
qqPlot(census$Unemployed, main="a")
# create spread level plot of model showing absolute studentised residuals
spreadLevelPlot(model,
  ylab="Absolute Studentised Residuals", las=par("las"),
  main=paste("b"))
# Also show normality of the residuals of model
qqPlot(residuals(model), main="c")
```

Plot in *LaTeX*.

```
\begin{figure*}
  \begin{center}
    \includegraphics{Assess2PINP_files/figure-latex/fig2-1}
  \end{center}
  \caption{(a) Distribution of the outcome variable.
    (b) Studentised Residuals of the model.
    (c) Homoscedasticity of the model.}\label{fig}
\end{figure*}
```

Note: For figures and tables that cover two columns I needed to convert to a *LaTeX* float as above.

Homoscedasticity.

```
# Non-constant Variance Score Test to show the error variance
ncvTest(model)
```

Multicollinearity.

```
# Variance Inflation Factor
# root of vif of model
vif(model)^0.5

# mean of vif
mean(vif(model))
```

Partial Residual Plots.

```
crPlots(model, main = " ", layout=c(1,4))
```

Plot in *LaTeX*.

```
\begin{figure*}
  \begin{center}
    \includegraphics{Assess2PINP_files/figure-latex/fig3-1}
  \end{center}
  \caption{Partial Residual Plots for Each Predictor Variable.}\label{fig}
\end{figure*}
```

Table 3: Standardised Regression Coefficients.

```
## find outcome mean
mean(census$Unemployed)
## find outcome range
range(census$Unemployed)
## find predictor mean
mean(census$White_British)
## find predictor range
range(census$White_British)
```

```
## Create table of beta coefficients, tidy into a data.frame
cftable = tidy(lm.beta(model)) # beta.lm from QuantPsych

# Table 3
kable(cftable, digits = 2, caption = "Standadised coefficients",
      longtable = FALSE, booktabs = TRUE,
      linesep = ’’, format = "latex",
      col.names = c("Variable",
                    "Standard Deviation"))
```

Markdown

This article was complied in RMarkdown with an adapted version of the PNAS LaTeX style class from the rticles package.

References

Chang, Winston. 2014. *extrafont: Tools for using fonts*.

Fletcher, Thomas D. 2012. *QuantPsyc: Quantitative Psychology Tools*.

Fox, John and Sanford Weisberg. 2011. *An {R} Companion to Applied Regression*. Second ed. Thousand Oaks {CA}: Sage.

Kassambara, Alboukadel. 2018. *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'*.

Pebesma, Edzer. 2018. “Simple Features for R: Standardized Support for Spatial Vector Data.” *The R Journal* .

Tennekes, Martijn. 2018. “{tmap}: Thematic Maps in {R}.” *Journal of Statistical Software* 84(6):1–39.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, Hadley, Romain François, Lionel Henry and Kirill Müller. 2018. *dplyr: A Grammar of Data Manipulation*.

Williamson, Paul. N.d. *ENVS450: Helper functions for ENVS450*.

Zhu, Hao. 2018. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*.