ENVS450 Social Survey Analysis. Assignment 2: Examining Predictors for Home Owner-Occupancy

201442927 *University of Liverpool*

The phrase 'housing crisis' has been in widespread use since the 1970s (Google, 2019). What are the predictors of home owner-occupancy in the 2010s? Using data from the 2011 Census, I investigate five other variables to find a minimal predictive model, and then discuss the implications of this result.

Keywords: Social Statistics, Home Ownership, Owner-Occupancy, Regression Models

Introduction

Home ownership is a source of emotional and ontological security (Dupuis and Thorns, 1998), a recurring political target (BBC, 2019), and a key element of contemporary capitalism (Forrest and Hirayama, 2015). In this report I explore data from the 2011 UK Census to find possible predictors for a district's owner-occupancy percentage, as an indicative proxy for home-ownership. I then compare five potential models of cumulative complexity, and suggest a best linear model, demonstrating how well it meets key regression assumptions. In the interests of *reproducible research* (Stodden, 2010), all code has been automatically attached as an appendix to this report.

Reviewing the Literature

According to Forrest et al. (2013), "individual home ownership" is a common goal "across a broad range of developed economies". Badarinza and Ramadorai (2018) note that "[r]esidential real estate is the single largest asset held by most households, and so this question is deeply political".

Schofield (2017) suggests that England has a 'North-South Divide on Home Ownership'. Hoggart (1997), focuses on rural home occupancy and its problems, noting both "under-provision in social housing" and "a lack of alternative rental accommodation". Others focus on the distinctive dynamics of house prices and home ownership in *London*, where the average house price paid in the last twelve months was over £635,000, compared to £277,000 for the UK as a whole (Zoopla, 2019).

This regional disparity in house values is similar to that observed thirty years ago (Giussani and Hadjimatheou, 1991). Studies have investigated the link between London house prices and crowding (Johnston et al. (2016)), gentrification and an increase in private renting (Hamnett (2009)), rent-volatility (Bracke (2015)), and working class 'white flight' (Tim and Chris, 2011). Rather than there being any strict regional divide, Holmes and Grimes (2008) suggest that " all UK regional house prices are driven by a single common stochastic trend… [but that] those regions that are more distant from London exhibit the highest degrees of persistence with respect to deviations in house price differentials".

The likelihood of the *elderly* to be home-owners is assumed and reiterated by the literature discussing care provision for elderly citizens eg. Toussaint and Elsinga (2009). Hamnett (1984) found a reciprocal "link between *fertility* and [housing] tenure". Gurney (1999) suggested that "the normalisation of home ownership" is responsible for "the residualisation of *social rented housing*".

Methodology

Data

This study uses data from the 2011 UK Census.The dataset included aggregated information for 348 districts in England and Wales. Eighteen percentage observations were given for each district – for the full list see Table 1. Some variables concerned the percentage of *individuals* exhibiting a particular characteristic, such as being 65 or over on the night of the census"; other concerned the percentage of *households* exhibiting a particular characteristic, such as not owning a van or car. Also included in the dataset was information about the region and country that each district was in.

The dataset does not include any information about non-response; instead the Office for National Statistics (2013) use a corrective 'edit and imputation' methodology for dealing with "missing data whilst preserving the relationships between census characteristics".

Dependent Variable

Figure 1: Distribution Density of Owner-Occupancy by District

The dependent variable will be the percentage of households that are `Owner_occupied`. This is a derived variable that includes outright ownership, and ownership with a loan or mortgage, as well as shared ownership (Office for National Statistics, 2014, p.93). As it is a percentage it is of course continuous.

Owner-occupancy varies from 24.94% to 83.01%. It has a negative skew of -1.71, making it non-symmetric, as confirmed by the density distribution (Figure 1). This means we must use Spearman's rank correlation coefficient when comparing it with other continuous variables. It also means we use the median as the measure of central tendency, and the interquartile range as the measure of dispersion: its median value is 68.91%, and its interquartile range is 8.81%.

Examining Correlation

Before deciding what predictor variables to investigate, I first explored the correlation between owner-occupancy and the other variables.

The correlation coefficients of all the variables except three (*Professionals*, *No Qualifications*, and *Full-Time Employees*) had p-values of less than 0.01, which means that if there were no correlation between ower-occupancy and the variable in question, the chances of the calculated correlation coefficient being caused merely by the randomness of the sample is less than 1%. The other three

Table 1: Correlation Matrix for Owner Occupancy, sorted by absolute r-value

Variables	Spearman's r	p Value		95% C.I. Min.	95% C.I. Max.
No_Cars	-0.8088430	0.00	**	-0.920	-0.880
Two_plus_Cars	0.8000132	0.00	**	0.825	0.881
Social_Rented	-0.7929545	0.00	**	-0.872	-0.811
Crowded	-0.7464387	0.00	**	-0.791	-0.698
Unemployed	-0.6655372	0.00	**	-0.657	-0.520
Private_Rented	-0.6637339	0.00	**	-0.845	-0.772
Age_65plus	0.6165023	0.00	**	0.606	0.723
Students	-0.5508639	0.00	**	-0.667	-0.533
Flats	-0.5502967	0.00	**	-0.832	-0.755
Lone_persons	-0.5406119	0.00	**	-0.666	-0.532
White_British	0.5082761	0.00	**	0.664	0.766
UK_Born	0.4917068	0.00	**	0.679	0.777
Couple_with_kids	0.4193364	0.00	**	0.477	0.623
Professionals	0.3158935	0.90		-0.112	0.099
No_Quals	-0.1156525	0.10		-0.017	0.192
FT_Employees	0.1068464	0.71		-0.085	0.125
illness	0.0288629	0.00	**	0.070	0.274

Statistical Significance:
** means p < 0.01, * means p < 0.05, no asterisk means p > 0.05

variables had p-values of more than 5%, which is not statistically significant. The correlation figures are all shown in Table 1.

Also shown are the lower (labelled 'Min.') and upper bounds ('Max.') of the 95% Confidence Intervals for the correlation coefficients, which is is the range within which the value will lie in 95 out of every 100 survey samples assuming that any difference between our data and the true population value is free from systematic error.

Examining Regional Effects

I then explored the regional variation of owner-occupancy, by examining a box-plot in the style of McGill et al. (1978). This is shown in Figure 2. "[T]he middle is fixed at the median; the upper and lower *hinges* are fixed at the upper and lower quartiles" (Wickham et al., 2015).

It is clear upon visual inspection that the distribution of owner-occupancy is very different in London than the rest of the country: the mean is much lower and the interquartile range is much wider. This is not surprising when one considers London house prices; cf. Badarinza and Ramadorai (2018), Hamnett (2009)). We therefore added a new `London` dummy variable to the dataset, giving a district the value `100` (as we are working with percentages) if its Region was London, and `0` for anything else.

Choosing Predictor Variables

At this point I knew from the box plot that `London` would be an interesting predictor variable to investigate, and from the correlation matrix that `No_Cars` would be a significant predictor. But the next highest predictor variable in absolute correlation rank, `Two_plus_Cars` would likely not add

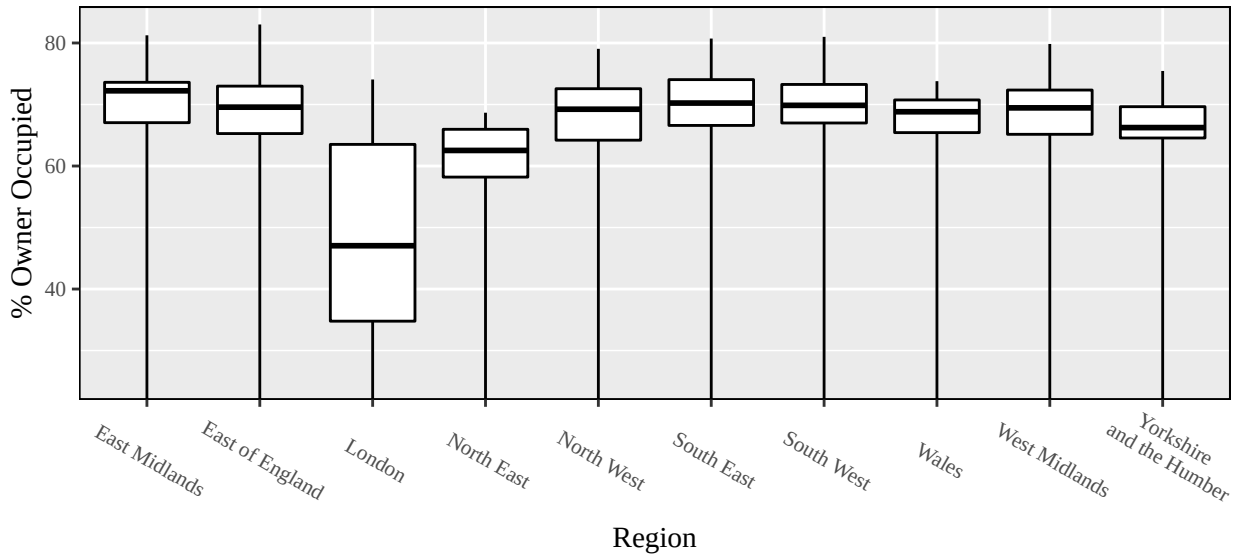

Figure 2: Box Plot Showing Regional Variation in Owner Occupancy

Table 2: Five-Variable Regression Subsets (Automated Analysis)

Variables	Intercept	No_Cars	UK_Born	Age_65plus	Couple_with_kids	London
1	85.564904	-0.8415826	NA	NA	NA	NA
2	58.626660	-0.6818875	0.2621817	NA	NA	NA
3	-19.839004	NA	0.2569634	1.3470769	1.5582157	NA
4	9.268576	-0.2576973	0.2259305	0.9468128	1.0308842	NA
5	5.604206	-0.3298471	0.3413740	0.8066349	0.9138243	0.0520087

Notes:
Based on 2011 Census Data.

that much extra information, since the upper bound on a district's percentage of households with more than two cars is clearly 100% minus the percentage with no cars.

The literature had suggested several other variables that relate to owner-occupancy, but it had also highlighted the politicized nature of this discussion. For an assignment due on the day of election, it therefore seemed appropriate to adopt a data-driven inductive strategy, lest our conclusions be warped by any presuppositional bias.

So I used the `regsubsets()` function from R's `leaps` package (Lumley, 2017) – which is based on the work of Miller (2002) – to run an automated test of all possible combinations of five or less variables, to see which would be worth investigating. The returned results (shown in Table 2) confirmed the significance of `No_Cars` and `London` (though the latter only adds explanatory value in the case of five predictors).

To those two predictor variables, the automated analysis suggested we also consider `UK_Born`, `Age_65plus`, and `Couple_with_kids`. Interestingly, in the case of three predictor variables, those seem to be more significant than either of the other two.

Table 3: Owner-Occupancy Models Compared in Order of Complexity

	1	2	3	4	5
(Intercept)	85.6***	55.3***	-24.7**	-107*	-199*
No_Cars	-0.842***	-0.592***		0.772	1.15
UK_Born		0.301***	0.334*	1.84**	2.85**
Age_65plus			2.36***	2.41	2.9
Couple_with_kids			1.21***	2.69*	5.27**
London					0.951**
No_Cars:UK_Born		-0.00111		-0.0177***	-0.021*
No_Cars:Age_65plus				0.02*	0.00679
UK_Born:Age_65plus			-0.0123*	-0.0299*	-0.0331*
No_Cars:Couple_with_kids				0.00755	0.0129*
Age_65plus:Couple_with_kids			0.00583	0.0309	0.0279
UK_Born:Couple_with_kids			0.0026	-0.0275*	-0.0567**
No_Cars:London					-0.00721*
UK_Born:London					-0.00138
Couple_with_kids:London					-0.0241***
Age_65plus:London					0.00402
Coefficient of Determination	0.813	0.858	0.882	0.896	0.909
Adjusted R squared	0.813	0.857	0.88	0.893	0.905
Akaike's Information Criterion	2010	1910	1860	1820	1780

Notes:
Calculations based on 2011 Census Data. Statistical significance is visualized for regression coefficients (though not for model fit measures): *** means p<0.001, ** means p < 0.01, * means p < 0.05; no asterisk means p > 0.05. Numeric values are rounded to three significant figures.

Results

Having obtained five suggested models of increasing complexity by automated analysis of subsets, I next humanly evaluated which are necessary for a *minimally adequate* predictive model of owner-occupancy, making sure to also consider the significance of two-way interaction effects. I did not however consider three-way interaction effects, deciding they are likely to unnecessarily complicate the analysis and unhelpfully confuse its interpretation.

First, the relevant coefficients for each model were obtained and compared, together with the statistical significance for each value; as well as the Coefficient of Determination (ie. the R-Squared value), the Adjusted R-Squared value, and the Akaike's Information Criterion for each model. These results are shown in Table 3.

Comparing Models of Cumulative Complexity

Inspection of the table shows that each degree of complexity added to model owner-occupancy does make a demonstrable contribution to the model fit, regardless of whether one measures this using the r-squared figure, the 'adjusted r-squared' figured, or Akaike Information Criterion. One might therefore conclude that the best model is the one that includes all five predictor variables.

On the other hand, the loss of simplicity may not be worth the increase in model fit: the first model, predicting owner-occupancy levels just on the basis of the percentage of households not

owning cars (or vans), already accounts for 81.3% of the variation. Adding four other predictor variables to the model may improve that by almost 10%, but is the loss of model parsimony worthwhile?

The other factor to take into consideration is the decreasing statistical significance of the calculated coefficients of the predictor variables.

As we increase the model complexity from one predictor to two, the regression coefficients maintain high degrees of statistical significance, with p-values less than 0.001 for all the independent variables (although this is not the case for the variable showing the interaction effect between having `No_Cars` and being `UK_Born`; but this effect is two orders of magnitude smaller, so will have a comparitively negligible effect).

When we increase the model complexity again to three predictor variables, the statistical significance of the regression coefficients begins to diminish somewhat, but still remains at an acceptable level. The coefficients for `Age_65plus` and `Couple_with_kids` still have p-values less than 0.001, and that for the intercept is less than 0.01; that for `UK_Born` is greater, but still less than 0.05, and its effect is the smallest. The interaction between `UK_Born` and `Age_65plus` is also statistically significant to this degree; and the other interactions are degrees of magnitude smaller, so again their effect will be comparitively negligible.

However, when increasing the model complexity to four predictor variables, the low levels of statistical significance of some of the regression coefficients begin to cause concern about the validity of the model. Two of the four predictor value coefficients now have p-values greater than 0.05, which suggests it is not that unlikely that their values are merely due to random sampling error.

Choosing the 'Best Model'

It therefore seems at this point that we might identify the model with three predictor variables as the best owner-occupancy rate prediction model:

$$\Omega = 0.334\beta + 2.36\epsilon + 1.21f - 0.0123\beta\epsilon + 0.00583\epsilon f + 0.00260\beta f - 24.7$$

Here Ω is the percentage of households in a district that are *owner-occupied*, β is the percentage of residents of a district *born* in the UK, ϵ is the percentage of residents of a district aged 65 or over (ie. *elderly*), and f is the percentage of households in a district made up of an adult couple with dependent children (ie. nuclear *families* living together).

However, examining the equation it is clear that interaction effects actually have a very insignificant impact. So we would in fact do better to leave them out of our model.

First we check the effect this has on the explanatory power of our model. Comparing the fit-values between a model which includes interaction effects and a model which ignores them, I found that the R-Squared, Adjusted R-Squared and A.I.C. values all decreased by less than 0.2%. I then used the `anova()` function to see is including the interaction effects made a statistically significant contribution, and found it did not (p>0.2).

I suggest that a best owner-occupancy rate prediction model is therefore as follows:

$$\Omega = 0.257\beta + 1.35\epsilon + 1.56f - 19.8$$

This model suggests that districts with more residents born in the UK have higher percentages of owner-occupancy; as do districts with more elderly residents; and as do districts with more family households.

This model further suggests that the greatest impact on the percentage of owner-occupancy is made by the percentage of households, with the coefficient of f slightly greater than that of ϵ, and five times higher than that of β. However, this further interpretation contradicts the evidence of the model which included interaction effects, where the coefficient of ϵ was twice that of f, so we must be wary of drawing overly precise conclusions.

We now examine whether the model satisfies key regression assumptions.

Normality of Residuals

An initial calculation suggested that the model's residuals were quite negatively skewed (-0.915), but visual inspection (see the histogram and Q-Q plot in Figure 3) showed that this result was due to a single outlier. (This turned out to be the district of the *Isles of Scilly*, "an archipelago off the southwestern tip of Cornwall", where "[t]ourism is a major part of the local economy" (Wikipedia, 2019), and the percentage of privately rented houses is almost twice the national average.)

When this outlier was removed, the skew calculation was 0.026, which is to say that the distribution of residuals can be considered symmetrical and (since the mean is very close to zero: 0.069) *normal*.

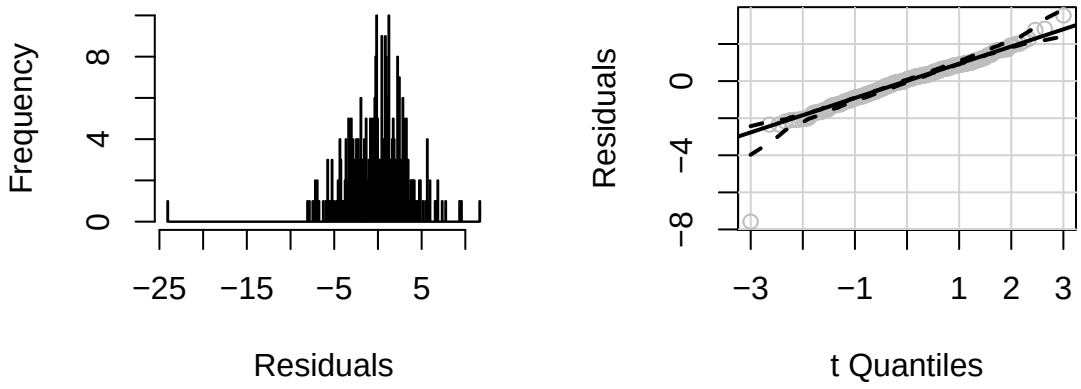

Figure 3: Histogram and Q-Q Plot showing Normality of Residuals.

Homoscedasticity: Constancy of Error Variance

Visual inspection of a Spread-Level Plot (Tukey, 1977) shows that the line of best-fit between the absolute studentized residuals and the fitted values is close to horizontal, so error variance is approximately constant.

This is confirmed by using the `ncvTest()` function, which confirms that this fit is statistically significant ($p<0.01$).

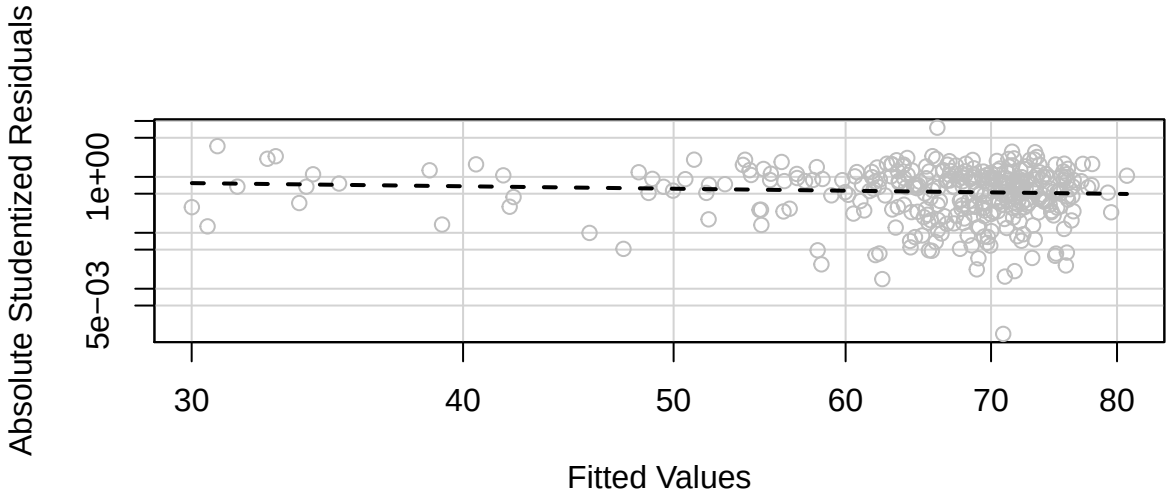

Figure 4: Spread Level Plot to show Homoscedacity

Predictor Variable Independence

Multicollinearity is assessed by calculating Variance-Inflation Factors. These come out (to 3 significant figures) as 2.21 for `UK_Born`, 2.19 for `Age_65plus`, and 1.12 for `Couple_with_kids`. Kabacoff (2011) p200 suggests that "[a]s a general rule, $\sqrt{vif} > 2$ indicates a multicollinearity problem", which is not the case here. Our model therefore shows adequate predictor variable independence.

Linearity of Relationship with Outcome Variable

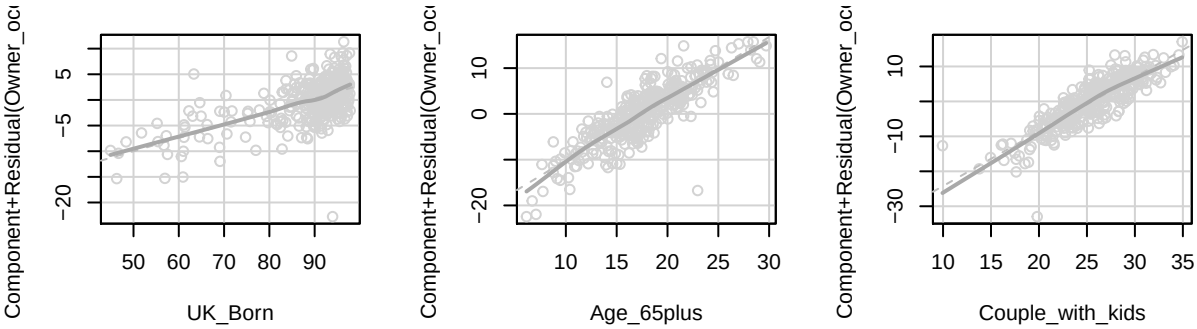

Figure 5: Component + Residual (Partial Residual) Plots

Visualizing the Partial Residuals Plots (see Figure 5) for each predictor variable shows that none of them have a perfectly linear relationship with owner-occupancy. To check more rigorously, we can fit models to a logged version of the variable in question and check if the suggested coefficient for that logged variable is statistically significant. When we do this, it seems that `Age_65plus` has a non-linear relationship with `Owner_occupied` ($p<0.001$); whereas `Couple_with_kids` ($p>0.5$) and `UK_Born` ($p>0.4$) do not.

This suggests an alternative 'best model':

$$\Omega = 19.9 \log \epsilon + 0.263\epsilon + 0.186\beta + 1.47f - 48.7$$

However, this model too has its faults, for the coefficient of ϵ is no longer statistically significant ($p>0.3$). So in fact, it is perhaps not a superior model after all. It is also less clear how to interpret it.

Independence of Model Errors

Regression modelling assumes that the model error for each case should be uncorrelated with the error of all other cases. However, when we sort the census dataset by `Region` and then apply the Durbin-Watson test, we find that there is a statistically significant correlation ($p<0.01$) between errors within the same region.

This is not entirely surprising, as we have already observed a regional difference in owner-occupancy between one region (London) and the rest of the country, but have decided not to include it in our 'best model'. When we exclude London districts from this test, we still find statistically significant correlation, albeit with a p-value one order of magnitude greater ($0.1>p>0.01$). So model errors seem to be geographically autocorrelated outside of London as well as in general.

Conclusion

We conclude by setting our findings in context.

The most significant direct correlation with district owner-occupancy turned out to be with not owning no cars – or if that double negative is confusing, with households owning cars. This is not surprising – if you are able to afford a house, you are doubtless able to afford a car – but it was not something I noticed anyone remarking on. This turned out to be justified, because the impact of lack of car ownership is fully accounted for in the case of three predictor values, by the percentage born in the UK, the percentage of elderly people, and the percentage of couple with dependent children.

The effect of couples with dependent children affirms the conclusion of Gurney (1999), that fertility and tenure have a reciprocal relationship. The implicit recognition we noted in the literature that elderly people are likely to be owner-occupiers was also validated by our analysis.

The 'London effect' was not as significant as might have been expected, only coming into the picture in the fifth and most complex of our examined models. This perhaps justifies the claim of Holmes and Grimes (2008) that there is not a 'divide' between London or 'the South' (Schofield, 2017) and the rest of the country, but rather a convergent continuum.

Lack of social housing was also notable only by its absence as a significant factor in predicting owner-occupancy, which might argue against the claims of Gurney (1999) – though of course the effect might be real and quantifiable, but only visible when one considers the impact of time.

Perhaps we should conclude as well with an admission that our attempt at a best model did not manage to perfectly satisfy the ideal assumptions for a reliable regression model. But this should not dishearten us. If even the movement of air molecules is complex, chaotic, unstable and non-linear, as so memorably suggested by Lorenz (1972) – how much more that of self-aware human beings.

Appendix 1: Code

```
# load essential libraries
library(ENVS450)            #
library(knitr)              # for generating PDF
library(kableExtra)         # for generating tables in PDF
library(dplyr)              # to turn subset regression output to dataframe
load.package('ggplot2')     # for graphing
load.package('ggcorrplot')  # for correlograms
load.package("car")         # for recode
load.package("interplot")   # For marginal effect plots

##################################################

# "(1) Choose a dataset".
# ---- the census dataset

# load data
census_filename <- '../data/2011Census.RData'
load(census_filename)

# explore data
dim(census) # 348 districts x 23 variables
names(census)
summary(census)
head(census)
unique(census$District)

##################################################
# "(2) Identify an outcome variable of interest to you"
# - owner-occupancy.

# "(3) Briefly explore the variation in this variable"

# is it skew?
skew(census$Owner_occupied)

# since it is skew we look at median as the measure of central tendency
summary(census$Owner_occupied)

# since it is skew we look at the interquartile range as the measure of dispersion
summary(census$Owner_occupied)['3rd Qu.']-summary(census$Owner_occupied)['1st Qu.']


##################################################
# confirm skew with visualization
ggplot(data=census) +
```

```
    geom_density( aes(x=Owner_occupied), fill="white" ) +
    theme_bw() +
    ylab("District Density") +
    xlab("% Owner-Occupied") +
    theme(axis.text.x = element_text(
                                    hjust = 0,
                                    vjust = 0,
                                    family="serif",
                                    size=8),
        axis.text.y = element_text(family="serif", size=8),
        axis.title.x = element_text(family="serif"),
        axis.title.y = element_text(family="serif"),
        panel.border = element_rect(colour = "black",
                                     fill = NA))
# "(4) Examine how strongly correlated your outcome variable is with the
#       other variables in the dataset"

# use spearman method because owner-occupancy is skew
correlation.matrix <- cor( census[ , -c(1:5) ] , method = "spearman")
# View results ordered by absolute value of r (i.e. ignoring the sign)
abs.correlations <- cor.results(correlation.matrix, sort.by="abs.r",
                                data=census, var.name="Owner_occupied")

# display correlations as labelled table
column_labels <- c("Variables", "Spearman's r", "p Value",
                   "", "95% C.I. Min.", "95% C.I. Max.")

kable(abs.correlations[,2:7], col.names=column_labels, format = "latex",
      longtable = FALSE, booktabs = TRUE, linesep = "",
      caption = "Correlation Matrix for Owner Occupancy,
                            sorted by absolute r-value")  %>%
      kable_styling(latex_options = c("striped"))  %>%
      footnote(general = "** means p < 0.01, * means p < 0.05, no asterisk means p > 0.05",
               general_title="Statistical Significance:",
               fixed_small_size = TRUE, threeparttable = TRUE)

################################################

# Examine correlation with region

# adjust "Yorkshire and The Humber" for visualisation
ykshire_and_humber <- levels(census$Region)[10]
levels(census$Region)[10] <- "Yorkshire \n and the Humber"

# visualise relationship between owner-occupancy and region
ggplot(data=census) +
  geom_boxplot(color="black", varwidth=FALSE, aes(x=Region, y=Owner_occupied),
```

```
                ymin=16, outlier.alpha = 0) +
    scale_fill_brewer(palette = "Greys") +
    ylab("% Owner Occupied") +
    xlab("Region") +
    theme(axis.text.x = element_text(angle = 330,
                                     hjust = 0.5,
                                     vjust = 0,
                                     family="serif",
                                     size=8),
        axis.text.y = element_text(family="serif", size=8),
        axis.title.x = element_text(family="serif"),
        axis.title.y = element_text(family="serif"),
        panel.border = element_rect(colour = "black",
                                    fill = NA))

# restore full region name for "Yorkshire and The Humber"
levels(census$Region)[10] <- ykshire_and_humber

################################################

# create dummy variable for district based on whether region is London:
# 100 (for easy comparison with percentage values) if so, 0 if not.
census$London <- ifelse(census$Region == "London", 100, 0)

################################################
# continue exploring data to find five key variables

# Use 'All subsets regression'
all_subsets_results <- leaps::regsubsets(Owner_occupied ~
                    No_Cars + Two_plus_Cars + Flats + Crowded + UK_Born + London
                    + White_British + Age_65plus + Students + Lone_persons
                    + Unemployed + Couple_with_kids + illness + No_Quals
                    + FT_Employees + Professionals, nvmax = 5, data=census)
crunch <- coef(all_subsets_results, 1:5)
# put results into dataframe
table<- bind_rows(lapply(crunch, as.data.frame.list))
Vars <- c(1:5)
vartable <- cbind(Vars,table)

# display results
column_labels <- c("Variables", "Intercept", "No_Cars", "UK_Born", "Age_65plus",
                   "Couple_with_kids", "London")

kable(vartable, format = "latex", col.names = column_labels,
      longtable = FALSE, booktabs = TRUE, linesep = "",
      caption = "Five-Variable Regression Subsets (Automated Analysis)") %>%
  kable_styling(latex_options = c("striped"))  %>%
```

```
      footnote(general = "Based on 2011 Census Data.",
          general_title="Notes:",
          fixed_small_size = TRUE, threeparttable = TRUE)

################################################

# "(5) Based on your preliminary data exploration,
#     select FIVE potential predictor variables"
## -- Selected: London + No_Cars + UK_Born + Couple_with_kids + Age_65plus

summary(census$No_Cars)
table(census$London)
summary(census$UK_Born)
summary(census$Age_65plus)
summary(census$Couple_with_kids)

################################################

# "(6) Explicitly adopt one of the regression model fitting strategies
#       outlined in Sessions 8-11."

# -- We have already used 'All Subsets Regression' (Session 8: 6.4)
# in exploring the data and choosing which five variables to focus on.
# The practical notes said "this would...take some time, so it isn't recommended"
# but using the nvmax parameter to restrict searches to subsets of 5 variables
# (ie. the number we want to examine), this didn't take too long.
system.time(
all_subsets_results <- leaps::regsubsets(Owner_occupied ~
                    No_Cars + Two_plus_Cars + Flats + Crowded + UK_Born + London
                    + White_British + Age_65plus + Students + Lone_persons
                    + Unemployed + Couple_with_kids + illness + No_Quals
                    + FT_Employees + Professionals, nvmax = 5, data=census)

)
#   user  system elapsed
#   0.05    0.02    0.08
# --> on my machine this took less than a tenth of a second.

# -- But to actually identify which of these combinations is the *best model*,
# I will adopt a 'narrative approach' (Session 8: 6.1):
# "Create a series of models which cumulatively include additional socio-economic
# and other predictor variables... The goal is to see if these additional
# predictor variables explain away the variation in outcome observed in
# the key predictor variable of interest. Stop adding variables when the
# additional predictor variables turn out to be statistically non-significant;
# or when the model becomes overly complex to readily interpret / report."
```

```r
# "(7) Use your chosen strategy to identify a 'best' model"

# [i] create the series of models
model1 <- lm(Owner_occupied ~ No_Cars, census)

model2 <- lm(Owner_occupied ~
               No_Cars + UK_Born
             + No_Cars:UK_Born,
             census)

model3 <- lm(Owner_occupied ~
              UK_Born + Age_65plus + Couple_with_kids
             + UK_Born:Couple_with_kids + UK_Born:Age_65plus
             + Age_65plus:Couple_with_kids,
             census)

model4 <- lm(Owner_occupied ~
            No_Cars + UK_Born + Age_65plus + Couple_with_kids
            + No_Cars:UK_Born + No_Cars:Couple_with_kids + No_Cars:Age_65plus
            + UK_Born:Couple_with_kids + UK_Born:Age_65plus
            + Age_65plus:Couple_with_kids,
            census)

model5 <- lm(Owner_occupied ~
               No_Cars + UK_Born + No_Cars:UK_Born
             + Age_65plus + No_Cars:Age_65plus + UK_Born:Age_65plus
             + Couple_with_kids + No_Cars:Couple_with_kids
             + Age_65plus:Couple_with_kids + UK_Born:Couple_with_kids
             + London + London:No_Cars + London:UK_Born + London:Couple_with_kids +
               London:Age_65plus,
                  census)


# [ii] "see if these additional predictor variables explain away the variation
# in outcome observed in the key predictor variable of interest"

# to do this I visualise the model data for easy side-by-side comparison

## First extract the model coefficients

coef1 <- coefficients(model1)
coef2 <- coefficients(model2)
coef3 <- coefficients(model3)
coef4 <- coefficients(model4)
coef5 <- coefficients(model5)

## Then extract data about the statistical significance of those coefficients
```

```r
# I created this function to return a column of asterisks
# signifying the p-values of each figure in the model.
# (There must be a simpler way of doing this, but I can't find it...)
stars <- function(modelobject){
  coeff.details <- summary(modelobject)$coefficients[,4]
  star.column <- ifelse(coeff.details<0.05,
                 ifelse(coeff.details<0.01,
                        ifelse(coeff.details<0.001,
                               "***",
                               "**"),
                        "*"),
                 "")

  return(star.column)
}

stars1 <- stars(model1)
stars2 <- stars(model2)
stars3 <- stars(model3)
stars4 <- stars(model4)
stars5 <- stars(model5)

# An inelegant way of getting a five-dimensional matrix
# with labels corresponding to our models.
zero <- coef5 - coef5
table <- cbind(zero,zero,zero,zero,zero)
# replace zeroes with blank space: less cluttered table is simpler to interpret
newtable <- ifelse(table == "0", "", "?")
table <- newtable

# round coefficients to 3sig.figs. and add asterisks indicating p values.
table[names(coef5),5] <- paste0(signif(coef5,3),stars5)
table[names(coef4),4] <- paste0(signif(coef4,3),stars4)
table[names(coef3),3] <- paste0(signif(coef3,3),stars3)
table[names(coef2),2] <- paste0(signif(coef2,3),stars2)
table[names(coef1),1] <- paste0(signif(coef1,3),stars1)


#"(8) For your 'best' model:
#   (a) Assess the model fit"
# --- we will do this for all the models, as it is an essential part of assessing
#   which is the best model.

# compare r-squared
r.squared1 <- summary(model1)$r.squared
r.squared2 <- summary(model2)$r.squared
r.squared3 <- summary(model3)$r.squared
```

```
r.squared4 <- summary(model4)$r.squared
r.squared5 <- summary(model5)$r.squared

r_squared <- signif(c(r.squared1, r.squared2, r.squared3, r.squared4, r.squared5),3)

# compare adjusted r-squared
adj1 <- summary(model1)$adj.r.squared
adj2 <- summary(model2)$adj.r.squared
adj3 <- summary(model3)$adj.r.squared
adj4 <- summary(model4)$adj.r.squared
adj5 <- summary(model5)$adj.r.squared

adjusted_r_squared <- signif(c(adj1, adj2, adj3, adj4, adj5),3)

# compare Akaike's Information Criterion

aic5 <- AIC(model5)
aic4 <- AIC(model4)
aic3 <- AIC(model3)
aic2 <- AIC(model2)
aic1 <- AIC(model1)

A.I.C. <- signif(c(aic1, aic2, aic3, aic4, aic5),3)

# combine data into single table

model_comparison <- rbind(table, r_squared, adjusted_r_squared, A.I.C.)

row_names <- labels(coef5)
row_names <- append(row_names, "Coefficient of Determination")
row_names <- append(row_names, "Adjusted R squared")
row_names <- append(row_names, "Akaike's Information Criterion")

# display table
display <- cbind(row_names, model_comparison)
column_labels <- c("", "1","2","3","4","5")

kable(display, format = "latex", row.names=FALSE, col.names=column_labels,
      longtable = FALSE, booktabs = TRUE, linesep = "",
      caption = "Owner-Occupancy Models Compared in Order of Complexity",
      digits=4) %>% row_spec(c(16,17,18), hline_after=T) %>%
  kable_styling(latex_options = c("striped"))    %>%
      footnote(general = "Calculations based on 2011 Census Data. Statistical significance is vi
          general_title="Notes:",
          fixed_small_size = TRUE, threeparttable = TRUE)

################################################
```

```r
# examine model improvements
anovaALL <- anova(model1, model2, model3, model4, model5)
anova1_2 <- anova(model1, model2)
anova4_5 <- anova(model4, model5)

anovaALL
anova1_2
anova4_5

anova(model3)

# calculate coefficienct for simple_model ignoring interaction effects
simple_model <- lm(Owner_occupied ~
                     UK_Born + Age_65plus + Couple_with_kids, data=census)
# check statistical significance
stars(simple_model) # p-values all <0.001
summary(simple_model) # in fact p-values all <2e-16 (!!)

summary(simple_model)$r.squared # 0.8801621

summary(simple_model)$adj.r.squared # 0.879117

AIC(simple_model) # 1855.463

summary(model3)$r.squared - summary(simple_model)$r.squared # 0.001528662
((summary(model3)$r.squared -
    summary(simple_model)$r.squared)
  /summary(model3)$r.squared * 100) # 0.1733785% decrease

summary(model3)$adj.r.squared - summary(simple_model)$adj.r.squared # 0.0004920726
((summary(model3)$adj.r.squared -
    summary(simple_model)$adj.r.squared)
  /summary(model3)$adj.r.squared * 100) # 0.0559422% decrease


AIC(model3) - AIC(simple_model) # 1.532328
(AIC(model3) - AIC(simple_model))/AIC(model3) * 100 # 0.08251652 % reduction

anova(simple_model, model3)['Pr(>F)']

simple_model$coefficients

#"(8) For your 'best' model...
#   Use model diagnostics to confirm that this 'best' model
#         satisfies key regression assumptions
#
# -- Q: What are these 'key regression assumptions'?
```

```r
# -- A:
## Session 11:5. "The validity of an OLS regression model depends upon
##    a number of assumptions being met: ...
##  - normality of residuals

# we have chosen the model with 3 predictor variables and no interactions
model <- simple_model


# "If the model errors are normally distributed,
# then the model is just as likely to under-estimate as over-estimate,
# meaning that it is not biased. i.e. the average model error should be zero.
# In addition, most model errors should be close to zero,
# with few very large positive or negative model errors."

# confirm average model error is zero
mean(model$residuals)

# calculate skew
skew(model$residuals)  #The skew() function is from the ENVS450 helper file

res <- model$residuals
sorted <- sort(res)
sorted[1:5]
str(sorted)
remove.outlier <- sorted[2:348]
mean(remove.outlier)
skew(remove.outlier)

census[labels(sorted[1]),]
census[149,"Private_Rented"]/mean(census$Private_Rented)

  # plot distribution of residuals to check for normality
g <- ggplot(data=simple_model) +
      geom_histogram( aes(x=.resid) , binwidth=1) +
      theme_bw() +
      ylab("Density") +
      xlab("Residuals") +
      theme(axis.text.x = element_text(#angle = 330,
                                       hjust = 0,
                                       vjust = 0,
                                      # family="serif",
                                       size=8),
          axis.text.y = element_text(#family="serif",
                                       size=8),
          axis.title.x = element_text(#family="serif"
                                       ),
```

```
  #        axis.text.x = element_blank()
          axis.title.y = element_text(#family="serif"
                                      ),
          panel.border = element_rect(colour = "black",
                                      fill = NA))

# ggplot doesn't balance well with qqlot, so use base-package hist() instead
hist(simple_model$residuals, breaks=348, main="", xlab="Residuals")

  # check normality of errors with QQ plot
  # "The studentized residuals (standardized model errors)
  #     should follow a straight line."
qq <-  qqPlot(simple_model, main="", xlab="t Quantiles", ylab="Residuals",
         las=0, col="grey", col.lines="black", id=F, )
  #The qqPlot() function is from the 'car' package



################################################
##  "...constancy of error variance (homoscedasticity)..."

# "The precision of the model (i.e. the variability of the model error)
# should be constant across the range of predicted model values.
# i.e. the model should be just as accurate when predicting low values
# as it is when predicting high values.

slp <- spreadLevelPlot(simple_model, main="", col="grey", col.lines="black")
#The spreadLevelPlot() function is from the 'car' package

# "A horizontal best-fit line, plus no obvious curvilinear shape
#   to the scatterplot indicates constant error variance."

ncvTest(simple_model) #The ncvTest() function is from the 'car' package
# p = 0.0058213

##  [e] predictor variable independence

# Variable independence is easily asseessed with vif() from the 'car' package

vif(simple_model) #      2.208036         2.193485         1.118877

mean(vif(simple_model)) # 1.84

vif(simple_model)^0.5 #   1.485946         1.481042         1.057770

# According to Field... If the largest VIF is greater than 10,
# then there is cause for concern If the average VIF is
```

```
# substantially greater than 1, then the regression may be biased
# --- average VIF = 1.84

# According to Kabacoff... If the square root of the VIF
# for a given variable is > 2 you have a problem. (i.e. VIF > 4 !)
# --- square root of vif is <2 for all

# According to Allison... If the VIF for a given variable
# is > 2.5 you may have a problem
# --- vif for each is <2.5

##  "... linearity of relationship with outcome variable..."

# decided not to include in report
plot_data_column <- function (x.vars, column, outcome) {
  p <- ggplot(data = x.vars, aes_string(x = column, y=outcome) ) +
  geom_point( ) +
  xlab(paste("%",column)) +
  ylab(paste("%", outcome)) +
  theme_bw() +
      theme(axis.text.x = element_text(#angle = 330,
                                       hjust = 0,
                                       vjust = 0,
                                       family="serif",
                                       size=8),
        axis.text.y = element_text(family="serif", size=8),
        axis.title.x = element_text(family="serif", size=8),
        axis.title.y = element_text(family="serif", size=8),
        panel.border = element_rect(colour = "black",
                                    fill = NA))

  if (is.numeric(x.vars[, column])) {
    p <- p + geom_smooth(method=lm, colour="grey", linetype="dashed", se=FALSE)
  }
  return(p)
}

myplots <- lapply(colnames(census), plot_data_column, x.vars = census,
                  outcome="Owner_occupied")



multi <- multiplot( myplots[[13]], myplots[[8]],
                    myplots[[7]], myplots[[21]], cols = 4)
##################################################
```

```
crPlots(simple_model, main="",  col="light grey",
        col.lines=c("grey","dark grey"), lwd=1,
        layout = c(1,3))
#crPlots() is a function of the 'car' package

# could Age_65Plus be non-linear?
#  can be more formally tested using a Box-Tidwell test for linearity:
model.logAge <- lm(Owner_occupied ~ Age_65plus + I(log(Age_65plus)) +
                   UK_Born + Couple_with_kids, data=census)
coeffs(model.logAge)['I(log(Age_65plus))','p.value']
summary(model.logAge)
AIC(model.logAge)

simple.logAge <- lm(Owner_occupied ~ Age_65plus + I(log(Age_65plus)), data=census)
summary(simple.logAge)
#  p = 0.00014 ==> statistically significant

# could Couple_with_kids be non-linear?
#  can be more formally tested using a Box-Tidwell test for linearity:
model.logKids <- lm(Owner_occupied ~ Age_65plus + UK_Born
                    + Couple_with_kids+ I(log(Couple_with_kids)) , data=census)
coeffs(model.logKids)['I(log(Couple_with_kids))','p.value']
#  p = 0.5630743 ==> not statistically significant

# while we're at it, check UK_Born too...
#  can be more formally tested using a Box-Tidwell test for linearity:
model.logUK <- lm(Owner_occupied ~ Age_65plus + UK_Born + Couple_with_kids
                  + I(log(UK_Born)) , data=census)
coeffs(model.logUK)['I(log(UK_Born))','p.value']
# p = 0.4055975 ==> not statistically significant

summary(model.logAge)

(AIC(model.logAge)-AIC(simple_model))/(AIC(simple_model))*100

##  "... independence of model errors (autocorrelation)..."

# Sort census dataset by cluster (region)
census.sorted <- census[ order(census$Region),  ]

# Fit model to sorted data
model.sorted <- lm(Owner_occupied ~
              UK_Born + Age_65plus + Couple_with_kids,
             census.sorted)

# Apply Durbin-Watson test to fitted model
durbinWatsonTest(model.sorted) # p =0.008
```

```r
# what if we separate London from the rest?
census$London[1]
outside_London <- subset(census, London==0)
London <- subset(census, London==100)

# Sort census dataset by cluster (region)
outside_London.sorted <- outside_London[ order(outside_London$Region),  ]

# Fit model to sorted data
model.outside_London.sorted <- lm(Owner_occupied ~
              UK_Born + Age_65plus + Couple_with_kids,
             outside_London.sorted)

# Apply Durbin-Watson test to fitted model
durbinWatsonTest(model.outside_London.sorted) # p =0.026


## Session 11:9 "The 'best' model is ultimately a matter of judgement rather
# than scientific fact, since it involves a trade-off, in priority order, between:
## [a] interpretability (data transformations make models harder to interpret)
## [b] model fit (r-squared; AIC)
## [c] normality of residuals
## [d] linearity of relationship with outcome variable
## [e] homoscedasticity

# "(9) Include adequately documented code in your submission:
# either as an appendix to your report if using Word (or similar);
# or as an integral part of the report if using R Notebook or R Markdown
# (or similar). "

# I'm using Knitr to convert my R Notebook to PDF.
# Knitr author Yihui Xie suggests various possible methods for attaching R code
# to the appendix of a report ( https://yihui.org/en/2018/09/code-appendix/ ),
# including "the cool hacker Lucy's method"
# ( https://twitter.com/LucyStats/status/1039178545715662848 ), adding a final
# codechunk with a reference label pointing to all the other code chunks, and
# the instruction that its code be echoed but not evaluated:
#
# ```{r ref.label=knitr::all_labels(), echo = T, eval = F}
# ```
# This seemed simplest, so I have used it.
```

Appendix 2: References

Badarinza, C. and Ramadorai, T. (2018). *Home away from home? Foreign demand and London house prices*. Journal of Financial Economics. 130(3):pp. 532–555.

BBC (2019). *Labour and Tories push rival housing policies*.
https://www.bbc.com/news/election-2019-50496700

Bracke, P. (2015). *House Prices and Rents: Microevidence from a Matched Data Set in Central London*. Real Estate Economics. 43(2):pp. 403–431. ISSN 1540-6229. doi:10.1111/1540-6229.12062.
https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.12062

Dupuis, A. and Thorns, D.C. (1998). *Home, home ownership and the search for ontological security*. The sociological review. 46(1):pp. 24–47.

Forrest, R. and Hirayama, Y. (2015). *The financialisation of the social project: Embedded liberalism, neoliberalism and home ownership*. Urban Studies. 52(2):pp. 233–244. ISSN 0042-0980, 1360-063X. doi:10.1177/0042098014528394.
http://journals.sagepub.com/doi/10.1177/0042098014528394

Forrest, R., Murie, A., and Murie, A. (2013). *Housing and family wealth in comparative perspective*. doi:10.4324/9780203416273-8.
https://www.taylorfrancis.com/

Giussani, B. and Hadjimatheou, G. (1991). *Modeling regional house prices in the United Kingdom*. Papers in Regional Science. 70(2):pp. 201–219. ISSN 1435-5957. doi:10.1007/BF01434329.
https://doi.org/10.1007/BF01434329

Google (2019). *Google Ngram Viewer: 'Housing Crisis'*.
https://books.google.com/ngrams/graph?content=%22housing+crisis%22&year_-start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_-url=t1%3B%2C%22%20housing%20crisis%20%22%3B%2Cc0

Gurney, C.M. (1999). *Pride and Prejudice: Discourses of Normalisation in Public and Private Accounts of Home Ownership*. Housing Studies. 14(2):pp. 163–183. ISSN 0267-3037, 1466-1810. doi:10.1080/02673039982902.
http://www.tandfonline.com/doi/abs/10.1080/02673039982902

Hamnett, C. (1984). *Housing the two nations: Socio-tenurial polarization in England and Wales, 1961-81*. Urban Studies. 21(4):pp. 389–405.

——— (2009). *Spatially displaced demand and the changing geography of house prices in London, 1995–2006*. Housing Studies. 24(3):pp. 301–320.

Hoggart, K. (1997). *Home Occupancy and Rural Housing Problems in England*. The Town Planning Review. 68(4):pp. 485–515. ISSN 0041-0020.
www.jstor.org/stable/40113472

Holmes, M.J. and Grimes, A. (2008). *Is there long-run convergence among regional house prices in the UK?*. Urban Studies. 45(8):pp. 1531–1544.

Johnston, R., Owen, D., Manley, D., and Harris, R. (2016). *House price increases and higher density housing occupation: The response of non-white households in London, 2001–2011*. International Journal of Housing Policy. 16(3):pp. 357–375. ISSN 1949-1247. doi:10.1080/14616718.2015.1130607. https://doi.org/10.1080/14616718.2015.1130607

Kabacoff, R. (2011). *R in Action : Data Analysis and Graphics with R*. Manning. ISBN 978-1-935182-39-9.

Lorenz, E. (1972). *Predictability: Does the Flap of a Butterfly's Wings in Brazil set off a Tornado in Texas*. http://eaps4.mit.edu/research/Lorenz/Butterfly_1972.pdf

Lumley, T. (2017). *Package 'leaps'*. https://cran.r-project.org/web/packages/leaps/leaps.pdf

McGill, R., Tukey, J.W., and Larsen, W.A. (1978). *Variations of Box Plots*. The American Statistician. 32(1):pp. 12–16. ISSN 0003-1305. doi:10.2307/2683468. https://www.jstor.org/stable/2683468

Miller, A. (2002). *Subset Selection in Regression*. Boca Raton: Chapman and Hall/CRC. 2 edition edn.. ISBN 978-1-58488-171-1.

Office for National Statistics (2013). *2011 Census Statistics for England and Wales*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/r

——— (2014). *2011 Census Variable and Classification Information: Part 4*.

Schofield, C. (2017). *England's North-South Divide on Home Ownership*. Data & Analytics. https://www.slideshare.net/CobainSchofield1/englands-northsouth-divide-on-home-ownership

Stodden, V. (2010). *Reproducible research: Addressing the need for data and code sharing in computational science*. Computing in Science and Engineering. 12:pp. 8–13. doi:10.1109/MCSE.2010.113.

Tim, B. and Chris, H. (2011). *Ethnicity, Class and Aspiration: Understanding London's New East End*. Policy Press. ISBN 978-1-84742-650-5.

Toussaint, J. and Elsinga, M. (2009). *Exploring 'Housing Asset-based Welfare'. Can the UK be Held Up as an Example for Europe?*. Housing Studies. 24(5):pp. 669–692. ISSN 0267-3037. doi:10.1080/02673030903083326. https://doi.org/10.1080/02673030903083326

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. ISBN 978-0-201-07616-5.

Wickham, H., Chang, W., and Henry, L. (2015). *A box and whiskers plot (in the style of Tukey) — geom_boxplot*. https://ggplot2.tidyverse.org/reference/geom_boxplot.html

Wikipedia (2019). *Isles of Scilly*. Wikipedia. page Version ID: 930283887. https://en.wikipedia.org/w/index.php?title=Isles_of_Scilly&oldid=930283887

Zoopla (2019). *House prices in UK*. https://www.zoopla.co.uk/house-prices/