

1 Introduction

We consider data on the gas consumption of dwellings of various types throughout North-East England. We have data on the gas consumption, age of the property, type of property, floor area of property, depth of loft insulation, whether cavity wall insulation is fitted, and whether a new boiler has been installed.

All but the gas consumption data is categorical, meaning that for example we do not know the precise age of the property but rather that it lies in a range of ages. This will impact our analysis and make it somewhat less accurate, but is a common trade-off, as it is easier to collect and store data of this sort. Obviously the type of property must be categorical, for example. We must note that the gas consumption has been somewhat severely rounded, leading to it, although truly continuous, having some properties (and difficulties) of a categorical variable. We will handle it as continuous, although we must keep the rounding in mind.

We will note that the gas consumption has been pre-corrected for weather, so this will not be a factor in our analysis. The data has also been selected to be representative of the housing stock, which will allow for us to draw more general conclusions. All of the properties use gas as the main mode of heating.

2 Statistical Analysis

2.1 Handling the Missing Data

First we must check the data for missing values and decide how to continue. An aggregate plot of the data is found in Figure 1. Obviously one must think about the ramifications of this, but we can clearly see that a large proportion of the data for boilers is missing, making it somewhat less useful for our analysis. We assume a completely random missing data mechanism, as this is the most likely given the nature of our data. This means that using predictive mean matching should give us results that are unbiased. We will use this going forward.

Another way of dealing with missingness in categorical variables is to make NA a category. This is not what we are going to use, but it is a simple way of getting the analysis to work.

We use the R package `mice` to perform multiple imputation on our data to obtain a full dataset. We will carefully consider the differences between our complete case analysis and our analysis of the imputed data and attempt to explain any discrepancies.

2.2 Initial Analysis

There are a few factors that we will account for. The first is that houses are not lines, but (nearly) cubes. We must consider that a house loses heat primarily through its four enclosing walls and its roof. One could think that a four fold increase in gas consumption would be needed to heat a house that is twice larger in length and depth. This means that it would be appropriate to use a square-root transformation on the gas consumption. We must be careful to acknowledge that the age and floor area, although inputted as numerical, are treated as factors. To do otherwise would invalidate the analysis.

The decision to use the square root is validated by looking at the quantile-quantile plots in Figure 2. We see that the de-trended plot for the transformed response variable gives us far superior results. We will continue with our analysis using the transformed response. We also see clear evidence of a quadratic relationship in the non-transformed plot, further reinforcing our decision.

2.3 Model Comparison and Selection

To generate the full model we must first look at the independence of terms in the data. We do this using the χ^2 test, the results of which are given in Table 1. We see that many of the variables are interdependent, which will influence our results. We see that the age, type, and floor area of a property are highly related, which is what we realistically expected. We regress the square root of the gas consumption on all of the predictor variables to obtain our full model. The diagnostic plots are given in Figure 3. These plots are perfectly sufficient, the leverages are low so our model is quite stable.

We will now compare the full model to two submodels. The submodels regress gas consumption on the base property (age, type, floor area) and on modifications to the property (loft insulation, cavity wall insulation, new boiler) respectively. The diagnostic plots are given in Figure 4. The diagnostic plots show that both are valid models which give similar results to the initial model. We compare them to the initial model using an ANOVA test. We see that the modifications model is significantly different from the initial model, so the base properties of the property have significant effects. However the property model is different, but only at the five percent level. We compare the residual sum of squares and see that the full model has an RSS of 1254843, the property model has an RSS of 1262560, and that the modification model has an RSS of 1902782. We see that the modifications to the property have a significantly lesser effect on the consumption than the property properties.

We choose to use the full model for our analysis as both of the sub models do not explain the data sufficiently. The regression coefficients are given in Table 2.

2.4 Discussion of Results

We are discussing the effect of the different variables on the square root of the gas consumption. The ‘base’ results are for a Bungalow in age band 1, in floor area band 1, with loft insulation greater than 150mm, without cavity walls, and without a new boiler. The predicted square root of consumption for this property is 102.3.

2.4.1 Effect of Age

We see that age has less of an effect than might be expected. The first four age bands have minimal effect. The fifth and the sixth age bands have more significant differences, with a decrease of 7.94 and 15.6 respectively. This means that new builds have decreased gas consumption. This may be due to the types of property built in new builds, as in 1 we see that they are not independent.

2.4.2 Effect of Type

We see that the type of property is very significant in determining the gas consumption. Detached houses have an increased root gas consumption compared to bungalows, consuming 12.2 root units more. Semi detached consume only very slightly more, to the point that it is not significant (it is less than the standard error of the estimate). Terraced and end of terrace houses consume less. Flats consume significantly less, with 21.5 less root units, corresponding to 461 units less.

2.4.3 Effect of Floor Area

We see that floor area has a significant effect on the gas consumption. All of them are significantly different from each other. The largest floor area consumes 52.43 root units, or 2749 units more than the smallest floor area. This is also not independent of the type of property, so we must know that changing one should have an effect on the other (ie a property with a large floor area is less likely to be a flat than detached).

2.4.4 Effects of Modifications to The Property

3 Conclusions

A Plots and Figures

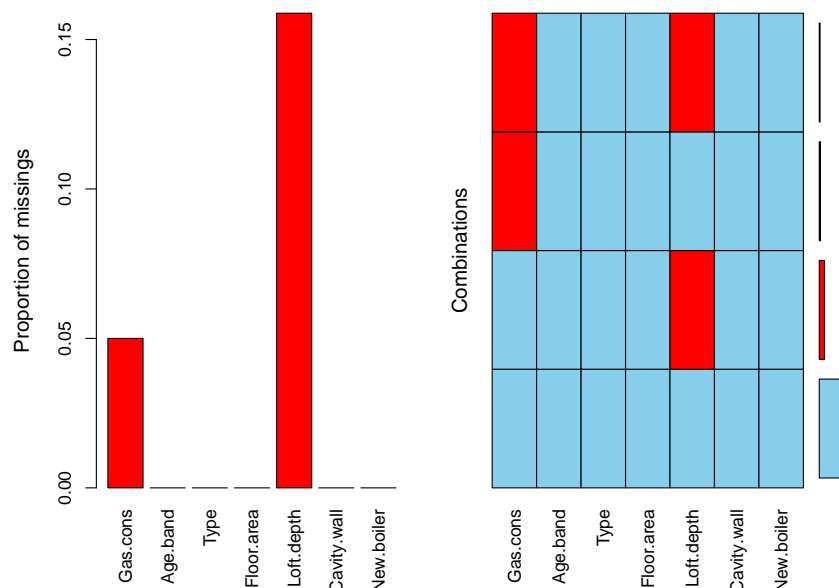


Figure 1: Aggregate plot of missingness in our data

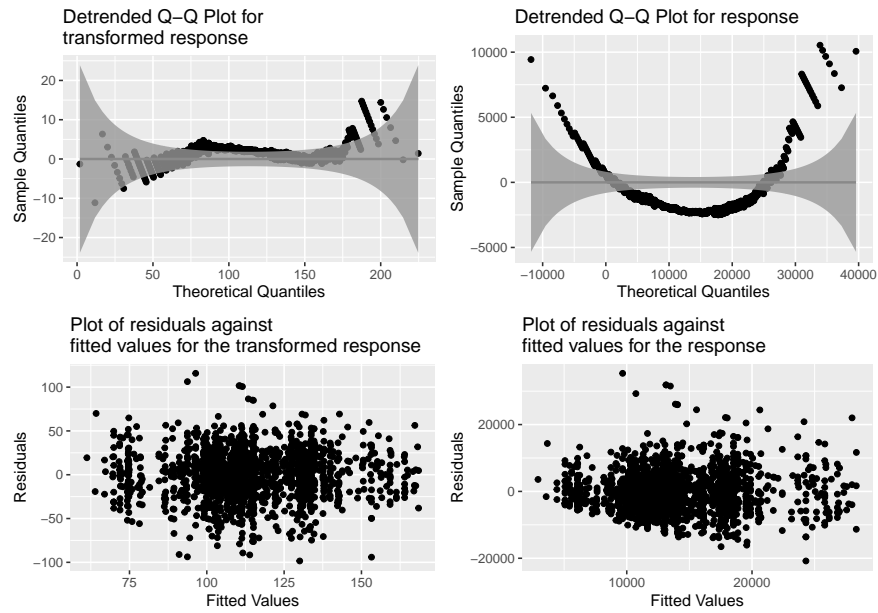


Figure 2: Model diagnostic plots for the transformed and non-transformed models

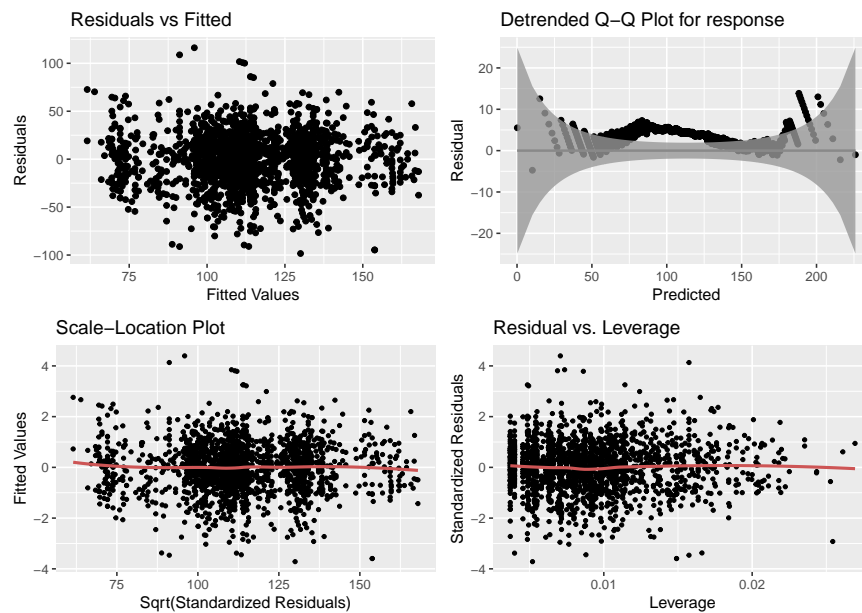


Figure 3: Model diagnostic plots for our full model

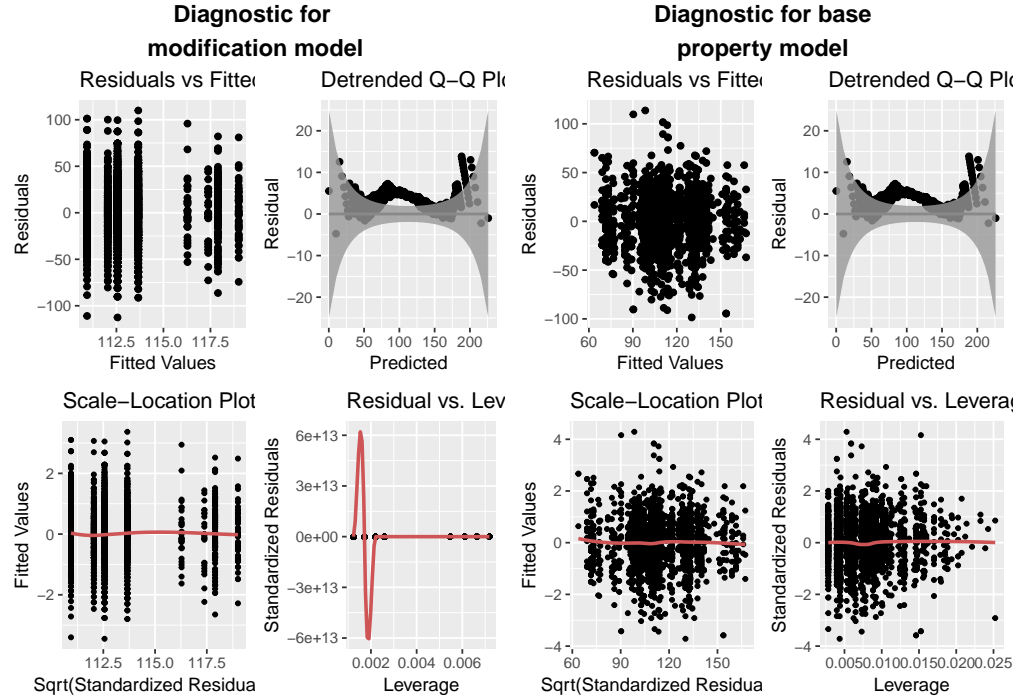


Figure 4: Model diagnostic plots for our sub models

B Tables

	Row	Column	Chi.Square	df	p.value
1	Gas.cons	Age.band	704.58	635	0.03
2	Gas.cons	Type	1114.13	635	0.00
3	Gas.cons	Floor.area	1126.06	381	0.00
4	Gas.cons	Loft.depth	118.46	127	0.69
5	Gas.cons	Cavity.wall	141.97	127	0.17
6	Gas.cons	New.boiler	129.45	127	0.42
7	Age.band	Type	649.32	25	0.00
8	Age.band	Floor.area	154.38	15	0.00
9	Age.band	Loft.depth	16.01	5	0.01
10	Age.band	Cavity.wall	91.75	5	0.00
11	Age.band	New.boiler	6.81	5	0.23
12	Type	Floor.area	1338.01	15	0.00
13	Type	Loft.depth	4.63	5	0.46
14	Type	Cavity.wall	71.07	5	0.00
15	Type	New.boiler	7.94	5	0.16
16	Floor.area	Loft.depth	6.46	3	0.09
17	Floor.area	Cavity.wall	36.19	3	0.00
18	Floor.area	New.boiler	3.29	3	0.35
19	Loft.depth	Cavity.wall	1.57	1	0.21
20	Loft.depth	New.boiler	1.96	1	0.16
21	Cavity.wall	New.boiler	0.47	1	0.49

Table 1: Table of results from Chi Squared test for independence.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	102.2815	4.0866	25.03	0.0000
Age.band2	-0.9813	2.3843	-0.41	0.6807
Age.band3	-4.2592	2.2810	-1.87	0.0620
Age.band4	-4.6301	2.3540	-1.97	0.0493
Age.band5	-7.9362	2.8934	-2.74	0.0062
Age.band6	-15.5629	2.9704	-5.24	0.0000
TypeDetached	12.1527	2.8993	4.19	0.0000
TypeEnd-terrace	-7.0452	2.7567	-2.56	0.0107
TypeFlat	-21.4717	3.1333	-6.85	0.0000
TypeMid-terrace	-11.7961	2.3468	-5.03	0.0000
TypeSemi-detached	0.4266	2.0914	0.20	0.8384
Floor.area2	13.0348	3.3290	3.92	0.0001
Floor.area3	32.3176	3.6443	8.87	0.0000
Floor.area4	52.4287	4.5895	11.42	0.0000
Loft.depthLe	1.4559	1.9928	0.73	0.4651
Cavity.wallY	-0.2868	1.3193	-0.22	0.8279
New.boilerY	-2.2588	1.3441	-1.68	0.0930

Table 2: Table of regression coefficients for the full model.