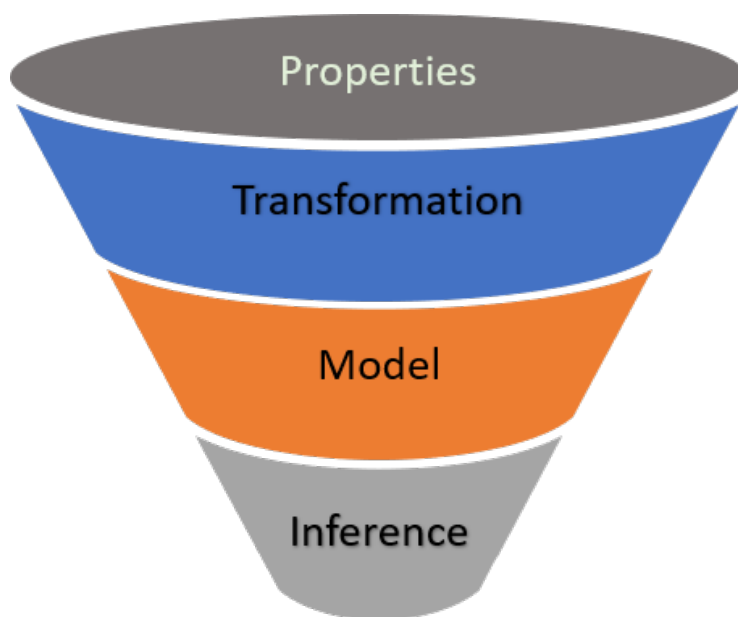


# Is there an Economic Case for Energy-Efficient Dwellings in the UK Private Rental Market?

A Project Report Submitted  
for the course

**Statistics I**  
in  
**Bachelor of Mathematics**



*by*

Aprameya Girish Hebbar  
Atreya Choudhury  
Sanchayan Bhowal  
Shreyan Saha



INDIAN STATISTICAL INSTITUTE  
BANGALORE- 560 059, INDIA  
*December 2021*

## ABSTRACT

This is an attempt to replicate and enhance the analyses presented in the paper ‘Is there an Economic Case for Energy-Efficient Dwellings in the UK Private Rental Market?’ by Franz Fuerst, Michel Haddad, Hassan Adad. This report provides a summary of the methods we used and the sources we used to complete this project.

**Keywords** - Energy efficiency; Hedonic model; Regression; Split incentive problem; Repeat sales index.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Data and descriptive statistics</b>	<b>4</b>
2.1	Repeat sales transactions variables . . . . .	5
2.2	Energy performance certificate (EPC) variables . . . . .	6
2.3	Index of multiple deprivation (IMD) variables . . . . .	7
2.4	Geographical location variables . . . . .	10
<b>3</b>	<b>Hedonic pricing model and results</b>	<b>11</b>
3.1	Summary of important terms in regression . . . . .	12
3.1.1	Multiple Regression Model . . . . .	12
3.1.2	Estimated coefficients . . . . .	13
3.1.3	Standard Error . . . . .	13
3.1.4	Testing if a coefficient is zero . . . . .	13
3.1.5	Residual standard error . . . . .	14
3.1.6	Multiple R-squared . . . . .	14
3.1.7	Adjust R-squared . . . . .	14
3.1.8	Confidence intervals for $\beta_{jt}$ . . . . .	14
3.2	Results of hedonic regression of prices . . . . .	15
3.3	Diagnostic plots . . . . .	18
3.3.1	Residuals vs. Fitted . . . . .	18
3.3.2	Normal Q-Q plots of residuals . . . . .	19
3.3.3	Scale-location plot and assumption of equal variance . . . . .	19
3.3.4	Leverage . . . . .	20
3.3.5	Cook's Distance . . . . .	21
<b>4</b>	<b>Price vs socio-economic variables</b>	<b>23</b>
4.1	Box-cox transformation . . . . .	24
4.2	Results . . . . .	25
4.3	Summary . . . . .	26
4.4	Partial residue plots . . . . .	26
<b>5</b>	<b>Repeat sales index</b>	<b>28</b>
5.1	Preparing the data for repeat sales index . . . . .	28
5.2	Regression . . . . .	29

<b>6</b>	<b>Conclusions</b>	<b>30</b>
6.1	Limitations . . . . .	31
6.2	Extensions . . . . .	31
6.3	Future Scope . . . . .	32
<b>7</b>	<b>References</b>	<b>32</b>

# 1 Introduction

The split incentive problem, or the lack of proper incentives to undertake energy efficiency measures, is one of the most pressing issues in the private rental sector. One party, the landlord, makes capital investments in energy efficiency and does not gain immediately, but the advantages are reaped by another, the tenant, who benefits from lower utility costs and greater thermal comfort. This affects landlords' investment decisions and creates a barrier to achieving improved energy efficiency. The key point of contention is whether the landlord can charge a higher rent for a dwelling that is more energy-efficient. "Does improved house energy efficiency contribute to higher home sales prices?"

We address the topic in section 3. The hedonic regression model is used, in which the price of a property is separated into several components connected with its related attributes, and obtains estimates of the contributory value of each component. Implicit prices are estimated by linear regression.

In section 4, we relate the housing prices to the socio-economic environment of the area. In section 5, we address another important question 'how does the sale prices of houses fluctuate with time in years?' We use the repeat sales method introduced by Bailey, Muth, and Nourse (1963). We conclude in section 6.

## 2 Data and descriptive statistics

The data set contains information on dwellings from a large sample of real estates in the United Kingdom ( $n=4,201$ ). There are 43 variables altogether and can be categorised into four classes: Repeat sales transactions variables, Energy performance certificate (EPC) variables, Index of multiple deprivation (IMD) variables, Geographical location variables. The authors of the original publication also wrote the data article on this dataset. The data was gathered by them from a number of publically available sources and filtered to make it easier to analyse.

- **Step 1:** The information on residences that have been sold at least twice were obtained from Her Majesty's Land Registry's online database, which had residential transaction values from 1995 to 2012.
- **Step 2:** The EPC rating data was acquired from the Domestic Energy Performance Certificate Register online database and then combined with the dataset.
- **Step 3:** The socioeconomic data was then added to the dataset established in the previous stage, and a random draw was performed to collect observations from hundreds of different neighbourhoods across England and Wales to provide a representative sample.

We proceed to explain the variables and highlight the key features of data. In the dataset the 4201 houses are indexed by id running from 1 to 4201.

## 2.1 Repeat sales transactions variables

The variables `price_1` and `price_2` represent the first and second sales prices, respectively which were brought on `date_1` and `date_2` respectively. From these raw variables, two derived variables are formed `ln_price_1`, `ln_price_2` are formed by taking their respective logarithms of `price_1` and `price_2`; `days_between_sale` represents the number of days between `date_1` and `date_2`. Another variable denoted by `perc_change_p2_to_p1`, notes the percentage change of `price_1` and `price_2`.

Variable	Mean	Median	Std.Dev	Skewnsws	Kurtosis	Smallest	Largest	Obs	Normal
<code>price_1</code>	120190.8638	100000	189750.5168	23.9595	689.7001	6000	5660000	4201	1.3e-80
<code>price_2</code>	154575.3033	120000	263755.4659	24.4936	706.5711	25000	7900000	4201	8.317e-82
<code>ln_price_1</code>	11.4558	11.5129	0.6549	-0.0051	1.645	8.6995	15.5489	4201	1.6227e-21
<code>ln_price_2</code>	11.7524	11.6952	0.5217	1.1602	5.0146	10.1266	15.8824	4201	1.7283e-35
<code>perc_change_p2_to_p1</code>	0.4986	0.2	0.8323	3.1254	19.9221	-0.622	10.4167	4201	8.4553e-59
<code>days_between_sale</code>	2400.0838	2196	1236.9192	0.5589	-0.3239	187	6156	4201	2.245e-28

Table 1: Descriptive statistics of the transactional variables

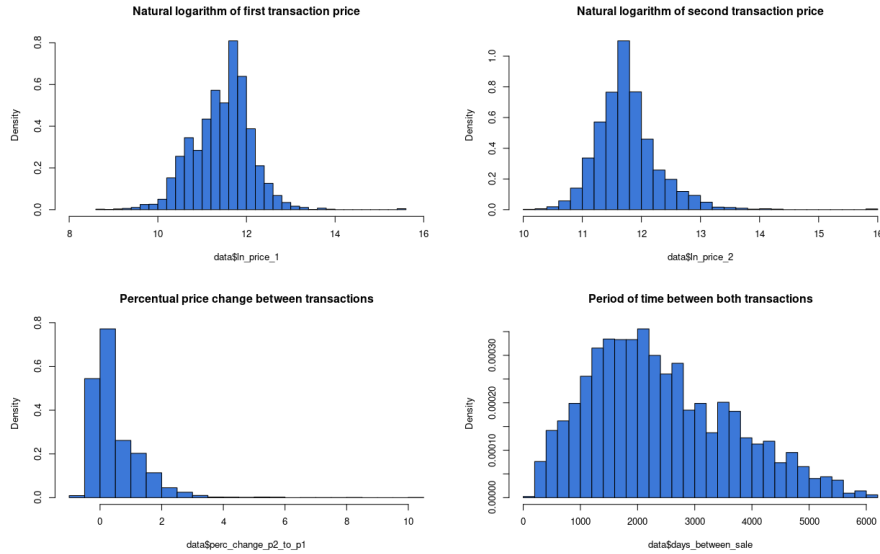


Figure 1: Distributions of the log price paid in the first (upper left hand side) and second (upper right hand side) property sale transactions, price variation between the first and second sale transaction (bottom left hand side), and period of time from the first to the second transaction (bottom right hand side)

## 2.2 Energy performance certificate (EPC) variables

An EPC is a rating system that describes the energy efficiency of real estate properties in the European Union. A score out of hundred denoted by `epc_100` is given to each house, where 1 is the least least efficient, and is based on standard assessment procedure (SAP) points. Each house was given an EPC band according to its EPC rating. A house belongs to a particular band in A,B,C,D,E,F,G if the SAP points is in the range of 92-100, 81-91, 69-80, 55-68, 39-54, 21-38, 1-20 respectively. The variables `epc_rating_a`, `epc_rating_b`, `epc_rating_c`, `epc_rating_d`, `epc_rating_e`, `epc_rating_f`, `epc_rating_g` are boolean variables which takes 1 if and only if the house is assigned to band A, B, C, D, E, F, G respectively. Another variable `ln_epc_100` is obtained by taking logarithm of `epc_100`. The table and histograms show the descriptive statistics of these variables.

EPC band	Frequency	Fraction
EPC A	0	0
EPC B	379	0.0902
EPC C	1442	0.3433
EPC D	1480	0.3523
EPC E	699	0.1664
EPC F	162	0.0386
EPC G	39	0.0093

Table 2: Frequency and fraction of the seven EPC bands

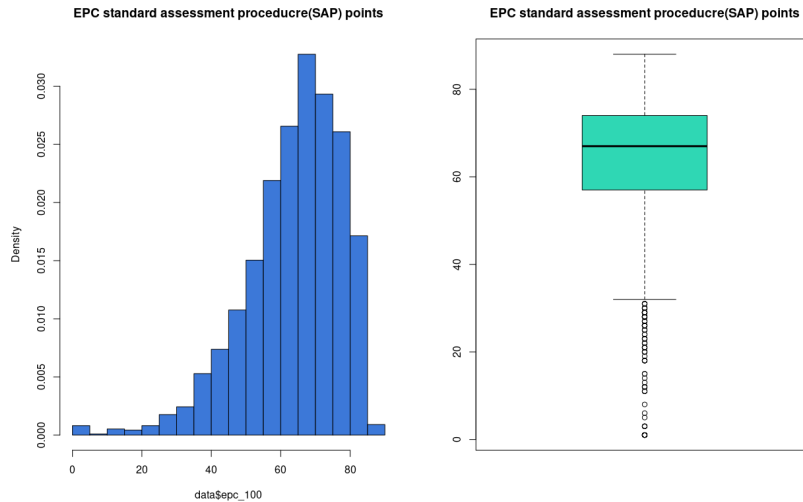


Figure 2: Histogram (left hand side) and box plot (right hand side) of the distribution of the SAP points

In this dataset, no property is assigned to band A. Most houses, around 70%, have the rating

C and D. Also, few data is available on EPC F and G.

## 2.3 Index of multiple deprivation (IMD) variables

A rank is assigned to each of the Lower Layer Super Output Areas (LSOA) in the United Kingdom in terms of crime (crime\_score), education (educ\_score), income (income\_score) and employment (emp\_score), health (health\_score) and disability, income, and living environment (living\_score). A rank indicating the level of deprivation (imd\_score) is also given. In all these a rank of 1 is the most deprived. Each LSOA is given score out of 1 to 10 in each of these criteria, where score 1 is the most deprived 10% of LSOAs. Thus we have the following categorical variables: imd\_level, income\_level, emp\_level, educ\_level, health\_level, crime\_level, barrier\_score, barrier\_level, living\_level.

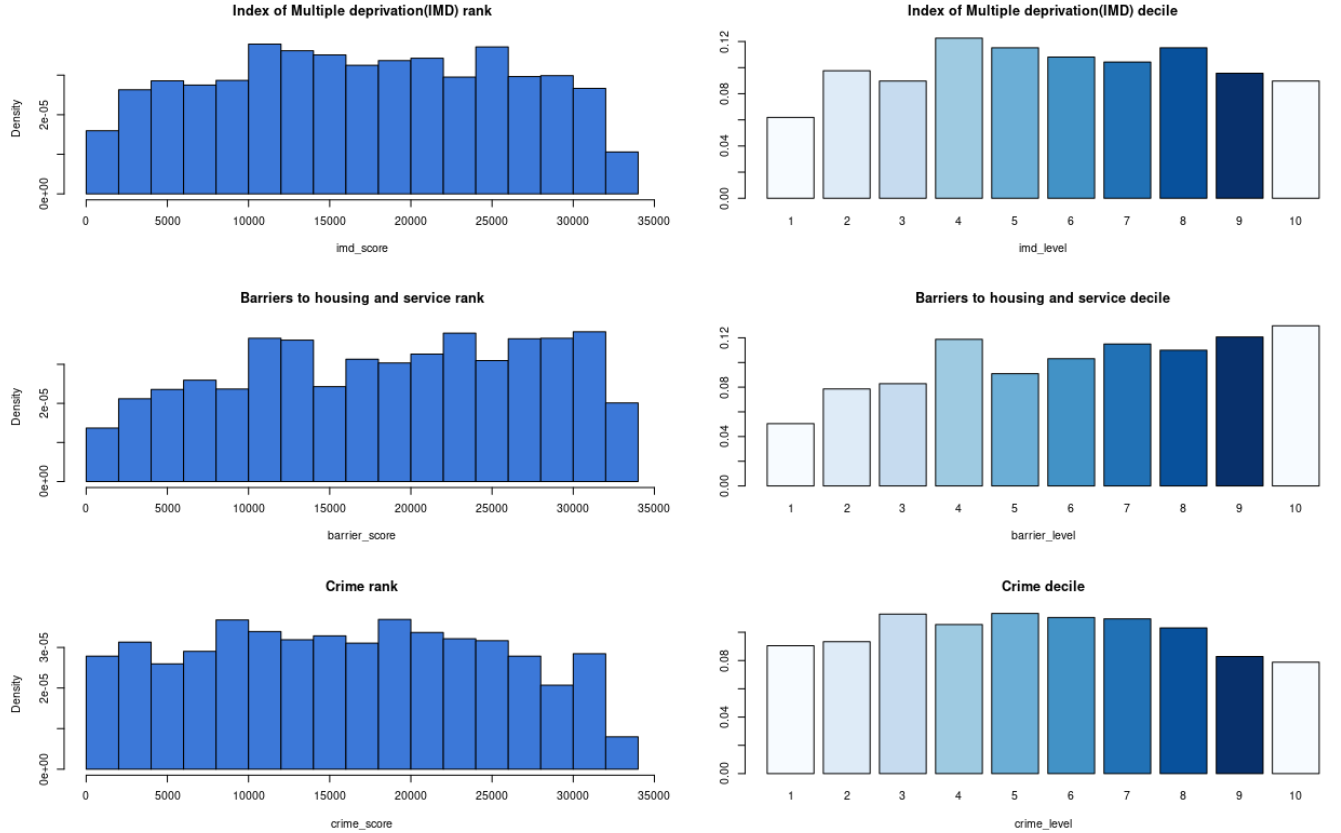


Figure 3: Histograms (left hand side) and bar charts (right hand side) of the IMD and its seven domains, considering their ranks and deciles, respectively



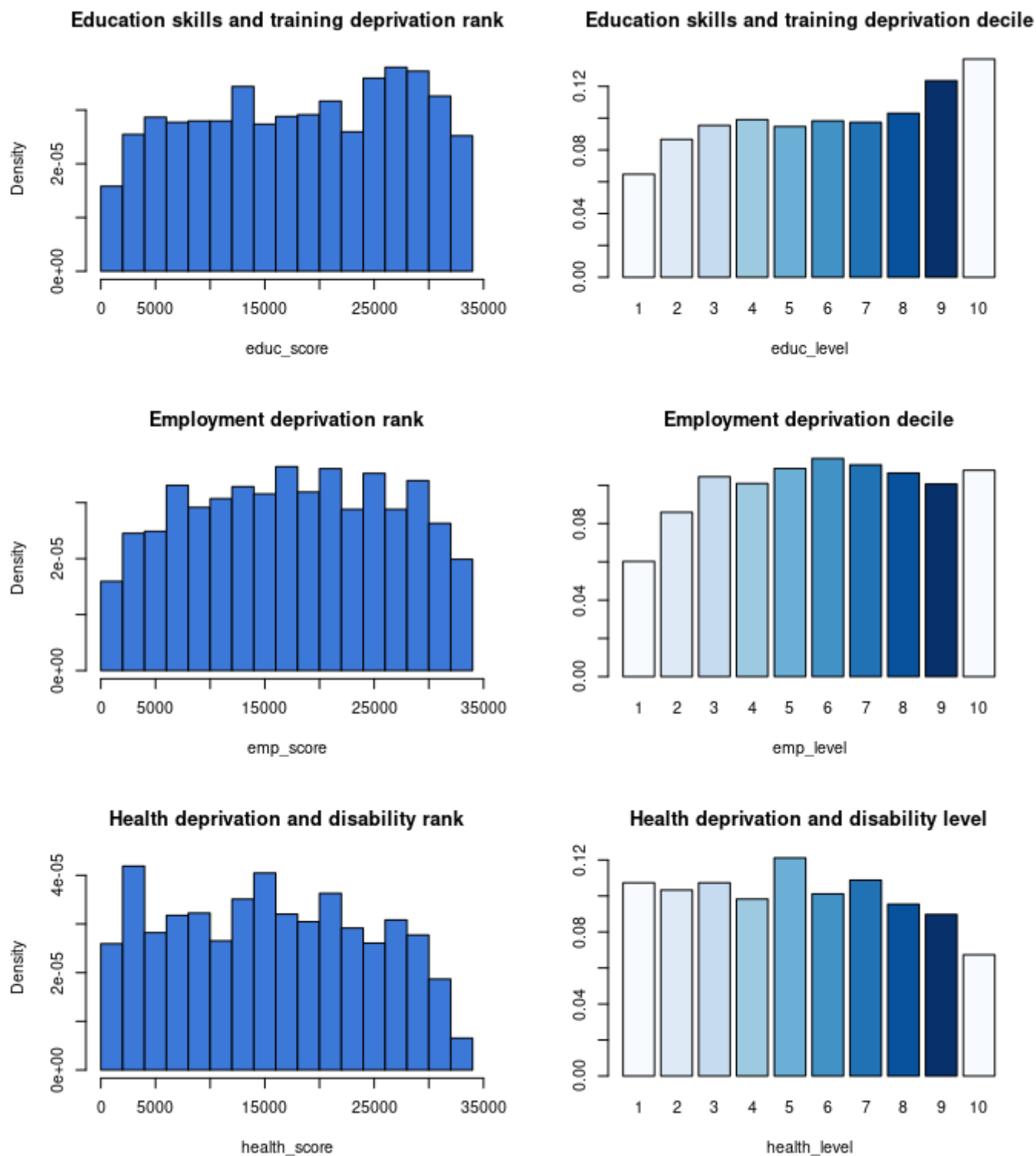


Figure 4: Continued

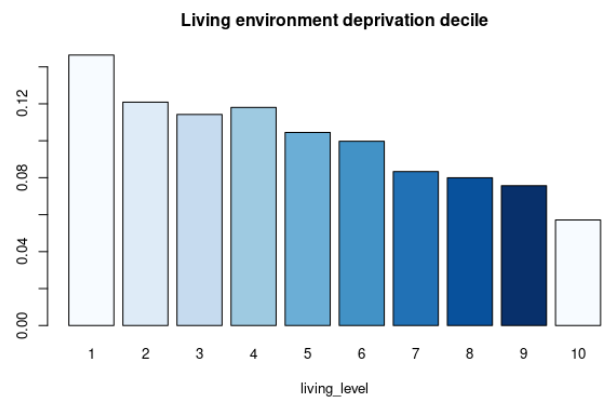
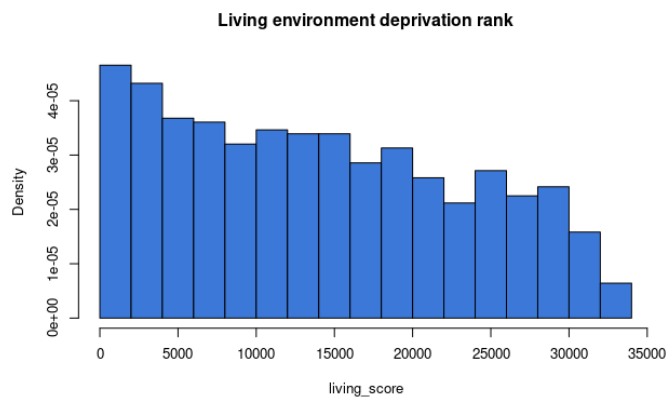
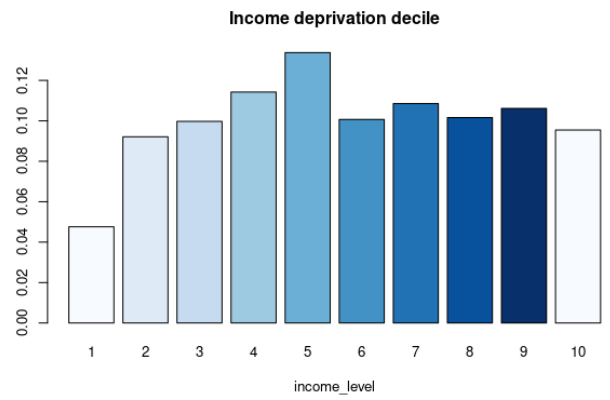
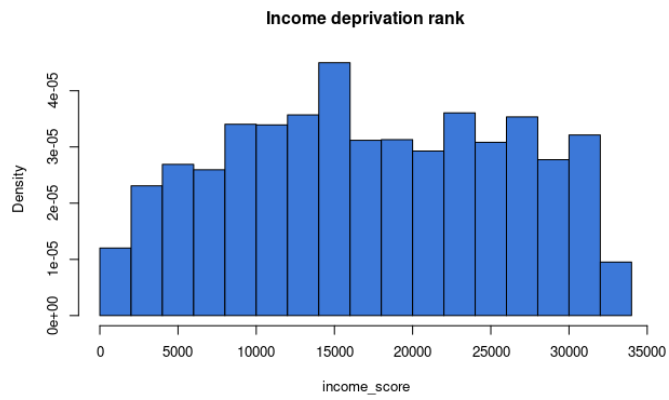


Figure 5: Continued

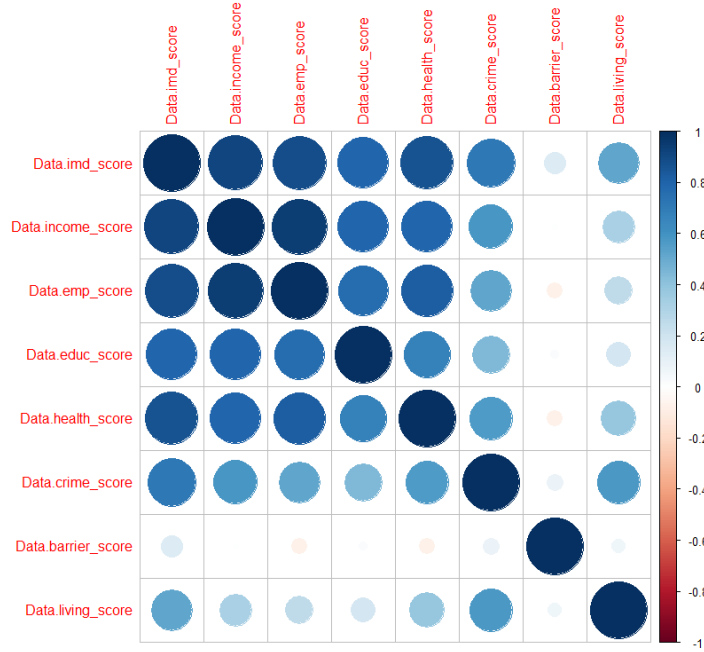


Figure 6: Correlation matrix plot of socio-economic variables labelled according to the colors

## 2.4 Geographical location variables

The properties in dataset were distributed geographically according to the Office for National Statistics (ONS) categorization, which includes nine regions (previously known as “government offices for the regions” or GOR). The variables `reg_north_east`, `reg_north_west`, `reg_yorkshire_and_the_humber`, `reg_east_midlands`, `reg_west_midlands`, `reg_east_of_england`, `reg_london`, `reg_south_east`, `reg_south_west` are the boolean variables indicating the region in which property belongs to.

Geography	Transaction Frequency	Transaction fraction
North West	840	0.2
Yorkshire and the Humber	839	0.1997
West Midlands	599	0.1426
East Midlands	435	0.1035
South East	407	0.0969
London	361	0.0859
South West	287	0.0683
East of England	279	0.0664
North East	154	0.0367

Table 3: Geographical distribution of the transactions included in the dataset

### 3 Hedonic pricing model and results

Let  $x_{i1t}, \dots, x_{iKt}$  denote locational and physical characteristics, including categorical variables related to  $i$ th property. Let  $P_{it}$  denote the price of  $i$ th property. As in the sample data we let  $t = 1, 2$ . We make a probabilistic model as follows. The dependent variable  $\ln(P_{it})$  is written as a linear function of  $x_i$  and an error term  $\varepsilon_i$  is added.

$$\ln(P_{it}) = \beta_{0t} + \sum_{j=1}^K \beta_{jt} x_{ijt} + \varepsilon_i$$

The coefficients  $\beta_{jt}$  are called hedonic price indices and are to be estimated.

The random errors capture additional information on the housing prices. We make four assumptions on the distribution of residuals  $\varepsilon$  in the linear model:

- The mean of  $\varepsilon$  is 0
- The variance of  $\varepsilon$  is  $\sigma^2$
- $\varepsilon \sim N(0, \sigma^2)$
- The random errors are independent of each other.

The method of least square(OLS) is applied to find estimators  $\hat{\beta}_{jt}$  of  $\beta_{jt}$  and the estimated model is

$$\ln(\hat{P}_t) = \hat{\beta}_{0t} + \sum_{j=1}^K \hat{\beta}_{jt} x_{jt}$$

where  $\hat{P}_t$  is the estimated sales price per square meter. The coefficients  $\hat{\beta}_{jt}$  can be interpreted as follows. Consider  $x_{jt}$  for a particular  $j$ . If  $x_{jt}$  is a numerical variable, then the above equation can be differentiated to yield

$$\begin{aligned}\frac{\partial(\ln \hat{P}_t)}{\partial x_{jt}} &= \hat{\beta}_{jt} \\ \Rightarrow \frac{\partial \hat{P}_t}{\partial x_{jt}} &= \hat{\beta}_{jt} \hat{P}_t\end{aligned}$$

On the other hand if  $x_{jt}$  is a categorical variable, hold all other variables constant and let  $x_{jt}$  change to  $x_{jt} + \Delta x_{ijt}$  and suppose  $\hat{P}_t$  changes to  $\hat{P}'_t$ . Then,

$$\begin{aligned}\ln(\hat{P}'_t) - \ln(\hat{P}_t) &= \hat{\beta}_{jt} \Delta x_{jt} \\ \Rightarrow \hat{P}'_t &= \hat{P}_t e^{\hat{\beta}_{jt} \Delta x_{jt}}\end{aligned}$$

Thus, for categorical variables,  $e^{\hat{\beta}_{jt}}$  measures the proportion of change brought in the price for unit change in  $x_{jt}$ .

## 3.1 Summary of important terms in regression

### 3.1.1 Multiple Regression Model

Suppose we would like to predict variable  $y$  using  $x_1, \dots, x_K$ . We have the data:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{iK} + \varepsilon_i = \mathbf{x}_i^T \vec{\beta} + \varepsilon_i$$

for  $i = 1, 2, \dots, n$ . Write

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nK} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{e} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

A model of the form  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_n x_K$  is to be fit to data, i.e.  $\hat{Y} = X\hat{\beta}$ . We need to find  $\hat{\beta}$  that minimizes the standard square error (SSE):  $S(\hat{\beta}) = \|Y - \hat{Y}\|^2$ .

- Equate  $\partial S / (\partial \hat{\beta}_i) = 0$  and we obtain:  $(X^T X) \hat{\beta} = X^T Y$ .
- A formal solution is  $\hat{\beta} = (X^T X)^{-1} X^T Y$ .

- The fitted values are  $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1}XY = HY$  where  $H$  is called the projection matrix.
- We make usual four assumptions on  $\mathbf{e}$  :  $E[\mathbf{e}] = 0, \mathbf{e} \sim N_n(\vec{0}, \sigma^2 I_K)$
- The covariance matrix of random vector  $\hat{\beta}$  is  $\Sigma_{\hat{\beta}\hat{\beta}} = \sigma^2(X^T X)^{-1}$

The following is a summary of the important terms in the hedonic regression that will be used.

### 3.1.2 Estimated coefficients

The *Estimate* is the estimated coefficient by OLS regression. If the estimated coefficient is positive, then there is positive associativity between the corresponding variable and the price.

### 3.1.3 Standard Error

*Std.Error*: The standard error for each coefficient is given by

$$\sigma_{\hat{\beta}_{jt}} = \sigma^2[(X^T X)^{-1}]_{jj}$$

### 3.1.4 Testing if a coefficient is zero

The testing of hypothesis whether a coefficient might be zero, using  $t$ -distribution is as follows. We take the level of significance to be 0.01.

$$\begin{aligned} \text{Null Hypothesis: } & \beta_{jt} = 0 \\ \text{Alternate Hypothesis: } & \beta_{jt} \neq 0 \\ \text{test statistic: } & t_c = \frac{\hat{\beta}_{jt}}{s_{\hat{\beta}_{jt}}} \\ \text{Rejection region: } & |t_c| > t_{\frac{0.01}{2}} \\ \text{p-value: } & 2P(t > t_c) \text{ if } t_c \text{ is positive} \\ & \text{and } 2P(t < t_c) \text{ if } t_c \text{ is negative.} \end{aligned}$$

where the  $t$ -distribution has  $n - (K + 1)$  degrees of freedom.

If we are 99% confident that  $\beta_{jt} \neq 0$ , then we consider  $x_{jt}$  to be a statistically significant in our model. The p-value for this test is given in the last column. The significance code is given to each coefficient: \*, \*\*, \*\*\* indicate that the  $p$ -value is less than 0.001, 0.01, 0.05 respectively. The variables whose coefficients have \*, \*\*, \*\*\* are statistically significant.

### 3.1.5 Residual standard error

If  $y$  is the dependent variable estimated by  $\hat{y}$ , the residual standard error  $s$  is given by

$$s = \sqrt{\frac{SSE}{n - (K + 1)}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - (K + 1)}}$$

$s$  is an unbiased estimator of  $\sigma$ .

### 3.1.6 Multiple R-squared

If  $y$  is the dependent variable estimated by  $\hat{y}$ , then the Multiple coefficient of determination  $R$  is given by

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$$

represents the proportion of the total sample variation in  $y$  that can be “explained” by the multiple-regression model.

### 3.1.7 Adjust R-squared

Adjust R-squared is defined as

$$\bar{R}^2 = 1 - \frac{n - 1}{n - (K + 1)}(1 - R^2)$$

The value of  $R$  may be large due to the excess number of regressors, which may not add to the regression’s explanatory power. This is penalised by adjusted R-squared.

### 3.1.8 Confidence intervals for $\beta_{jt}$

To each parameter  $\beta_{jt}$ , the 99% ( $\alpha = 0.01$ ) confidence interval using  $t$ -distribution is given by

$$(\hat{\beta}_{jt} - t_{\alpha/2} s_{\hat{\beta}_{jt}}, \hat{\beta}_{jt} + t_{\alpha/2} s_{\hat{\beta}_{jt}})$$

where  $t_{\alpha/2}$  is the value such that, if  $t$  has  $t$ -distribution of  $n - (K + 1)$  degrees of freedom,

$$P(t > t_{\alpha/2}) = \frac{\alpha}{2}$$

### 3.2 Results of hedonic regression of prices

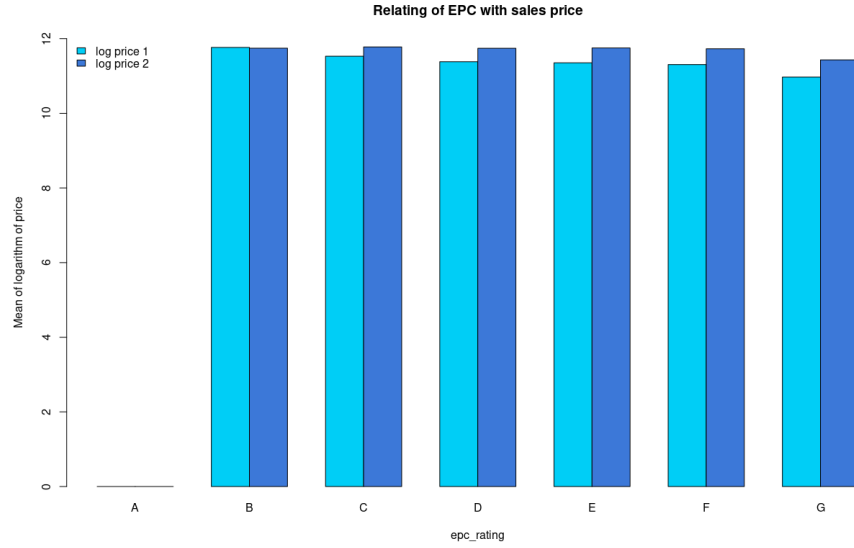


Figure 7: Plot of mean of  $\ln\_price\_1$  and  $\ln\_price\_2$  across different EPCs

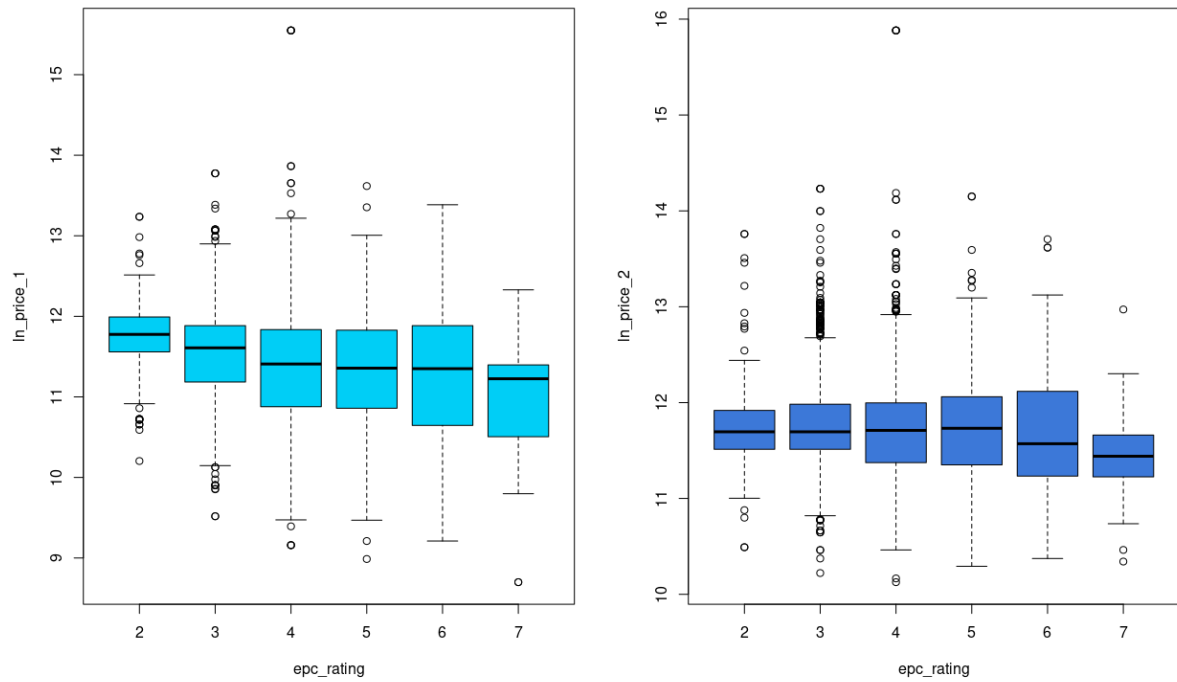


Figure 8: Box Plot for  $\ln\_price\_1$  and  $\ln\_price\_2$  respectively categorised by EPCs where 2:B,3:C,4:D,5:E,6:F,7:G



We have used R to estimate the hedonic pricing indices. The logarithm of prices was regressed on categorical variables `epc_rating_b`, `epc_rating_c`, `epc_rating_e`, `epc_rating_f`, `epc_rating_g`, and various regional indicators. As most houses have EPC rating D, `epc_rating_d` was taken as the baseline and is held out of the regression. `epc_rating_a` is removed as no houses in the sample has EPC rating A. `reg_north_west` is also taken as reference, as most houses in the sample are from this region. [Table 4](#) and [Table 5](#) shows the summary of this linear model.

Variable	Estimate	Std Error	t-value	Pr(>  t )	
Intercept	11.201346	0.024276	461.414	0(approx)	***
<code>epc_rating_b</code>	0.412648	0.033911	12.168	0(approx)	***
<code>epc_rating_c</code>	0.139870	0.021929	6.378	$2 \cdot 10^{-9}$	***
<code>epc_rating_e</code>	-0.021471	0.027018	-0.795	0.42684	
<code>epc_rating_f</code>	-0.103260	0.048702	-2.120	0.03405	*
<code>epc_rating_g</code>	-0.371917	0.095488	-3.895	0.00001	***
<code>reg_north_east</code>	-0.002641	0.051658	-0.051	0.95923	
<code>reg_yorkshire_and_the_humber</code>	0.061126	0.028788	2.123	0.03378	*
<code>reg_east_midlands</code>	-0.002814	0.035063	-0.080	0.93605	
<code>reg_west_midlands</code>	0.101695	0.031686	3.209	0.00134	**
<code>reg_east_of_england</code>	0.203688	0.040687	5.006	0.0000006	***
<code>reg_london</code>	0.882822	0.037107	23.791	0(approx)	***
<code>reg_south_east</code>	0.496382	0.035536	13.968	0(approx)	***
<code>reg_south_west</code>	0.239821	0.040296	5.951	$3 \cdot 10^{-8}$	***

Table 4: Hedonic regression of `price_1`. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Variable	Estimate	Std Error	t-value	Pr(>  t )	
(Intercept)	11.56700	0.01797	643.775	0(approx)	***
epc_rating_b	0.03577	0.02510	1.425	0.15415	
epc_rating_c	0.02668	0.01623	1.644	0.10026	
epc_rating_e	0.02064	0.02000	1.032	0.30208	
epc_rating_f	-0.04066	0.03605	-1.128	0.25942	
epc_rating_g	-0.27357	0.07067	-3.871	0.00011	***
reg_north_east	-0.06989	0.03823	-1.828	0.06763	.
reg_yorkshire_and_the_humber	0.05109	0.02131	2.398	0.01653	*
reg_east_midlands	-0.02497	0.02595	-0.962	0.33599	
reg_west_midlands	0.05897	0.02345	2.515	0.01195	*
reg_east_of_england	0.19067	0.03011	6.332	0(approx)	***
reg_london	0.98121	0.02746	35.727	0(approx)	***
reg_south_east	0.47361	0.02630	18.007	0(approx)	***
reg_south_west	0.25482	0.02982	8.544	0(approx)	***

Table 5: Hedonic regression of price\_2. Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

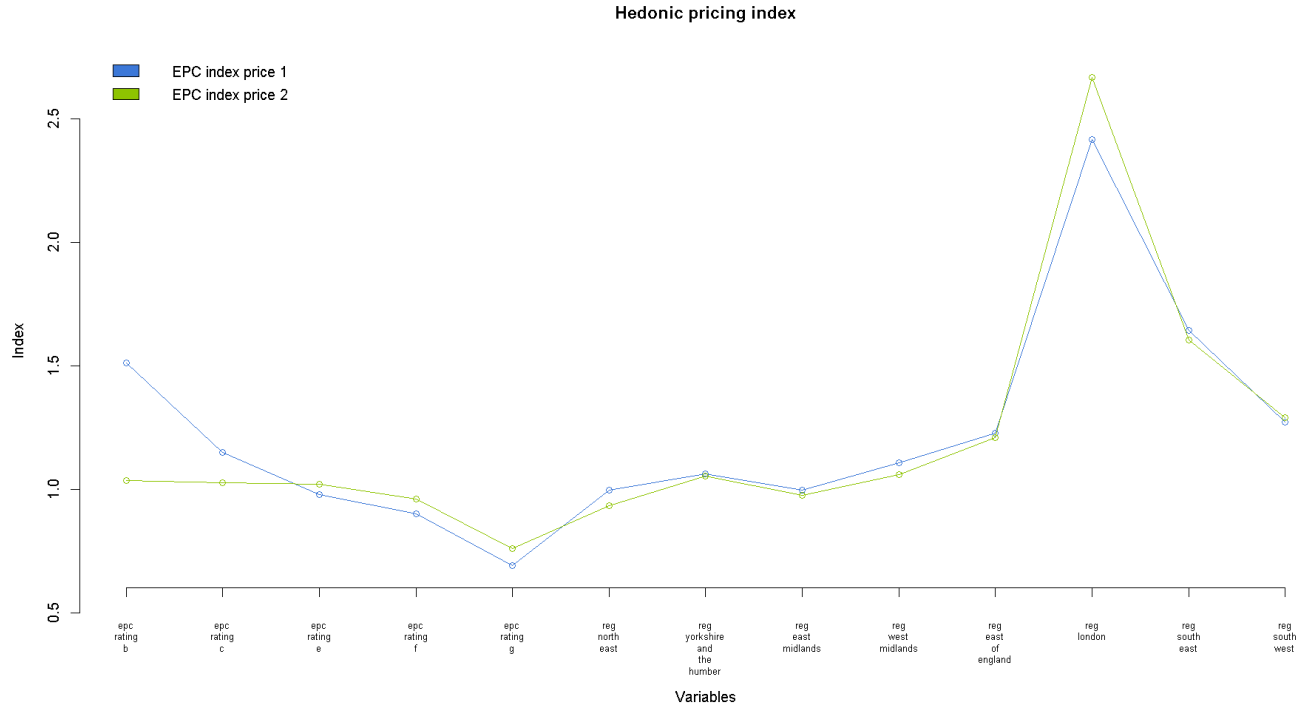


Figure 9: Hedonic Index of variables indicating the exponents of the regression coefficients

The exponent of coefficients can be plotted on a line graph and is shown in above Figure 9. This shows the relation between EPC rating and prices. From the Table 4, we see that the epc\_rating\_b, epc\_rating\_c are statistically significant variables in this model. For price\_1, epc\_rating\_b has

hedonic index around 1.5, indicating that houses with EPC rating B experience a price premium. The hedonic index of `epc_rating_g` is less than 1, indicating that lower energy efficient homes tend to have lower prices. Region London has the highest hedonic index indicating that, with respect to houses in northwest region, houses in London are expensive.

### 3.3 Diagnostic plots

Residual is nothing but the difference between fitted values and true values.

$$\varepsilon_i = y_i - \hat{y}_i$$

We made four assumptions on the distribution of these residuals in the beginning and now we diagnose if these assumptions hold, and assess if the model explains the relationship well.

#### 3.3.1 Residuals vs. Fitted

This is a scatter plot of fitted values on x-axis and residual=true value–fitted values. These graphs indicate if the price and energy performance have a non-linear connection. If the model does not capture the non-linear connection between predictor factors and outcome variables, the pattern may appear non-linear

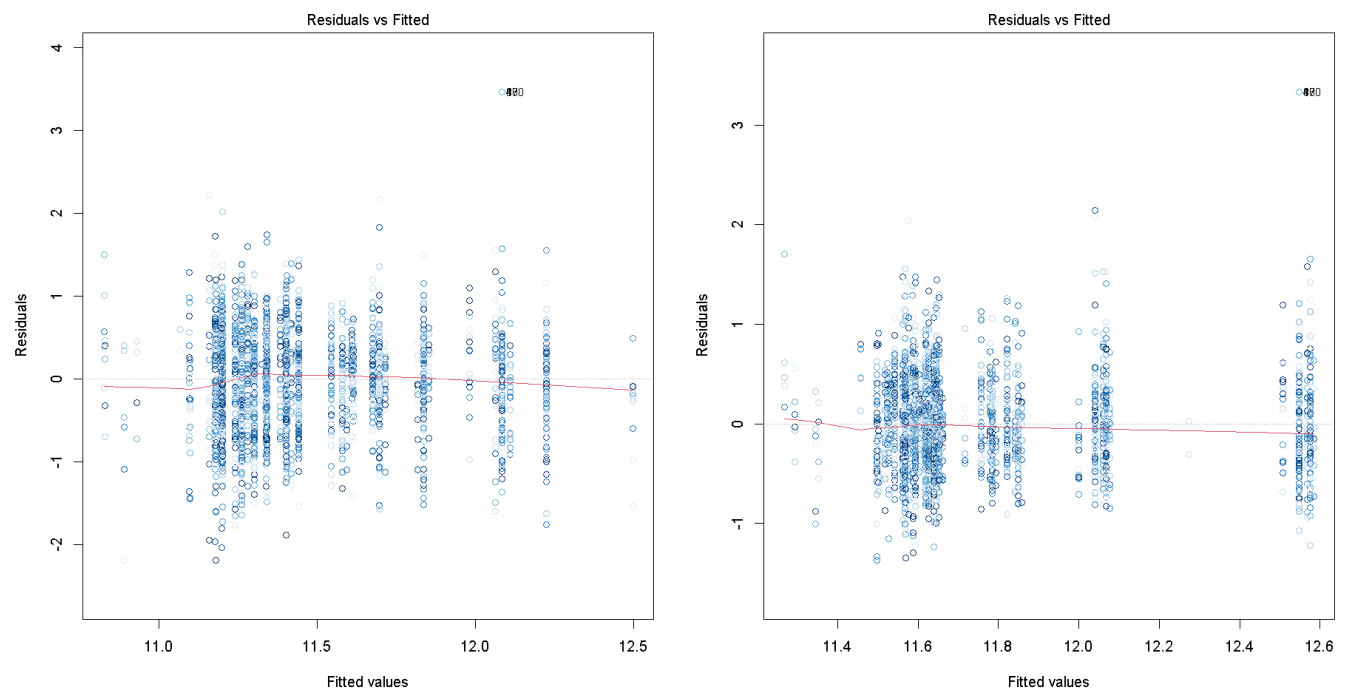
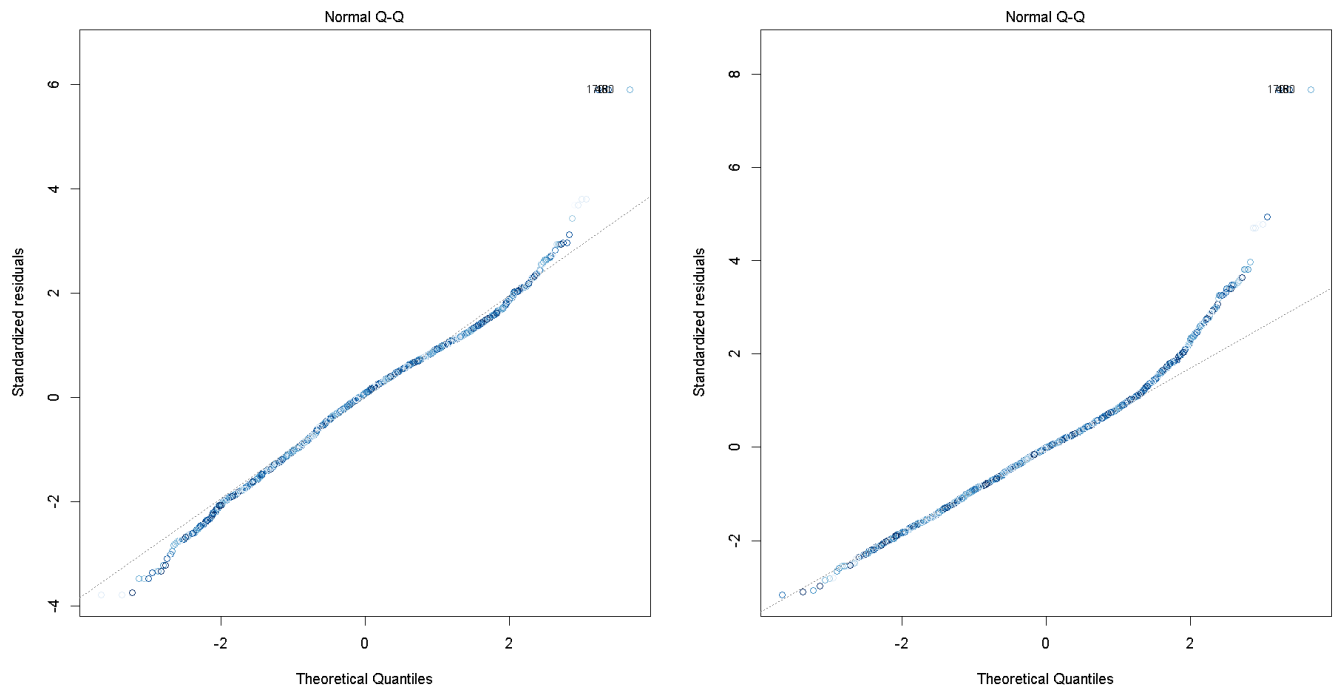


Fig shows the residuals versus fitted plot for the two models. Both the graphs show that the relationship between prices and locational factors, energy performance is nearly linear and our model explains this relationship quite well.

### 3.3.2 Normal Q-Q plots of residuals

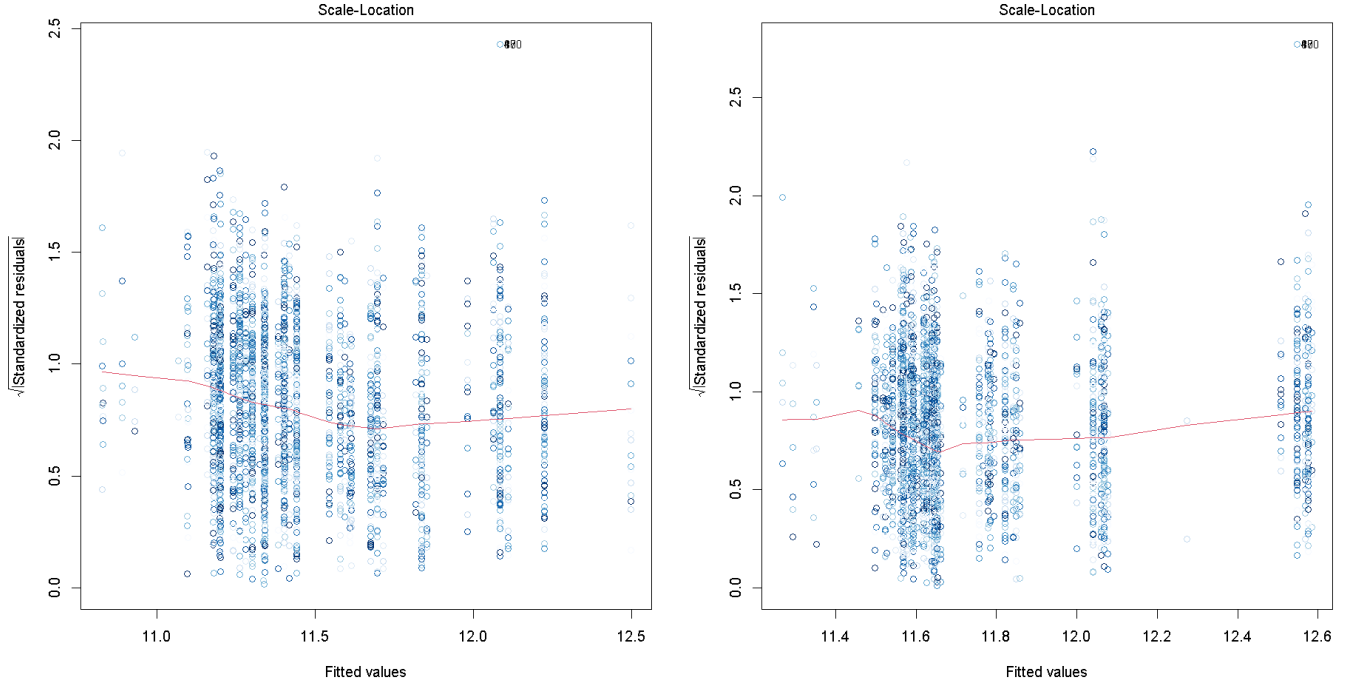
This is a scatter plot with theoretical normal quantiles on x-axis and quantiles of residues in y-axis. This plot diagnoses if the residuals are normally distributed. The figures below show that the residuals in this model are almost normally distributed, in accordance with our assumptions.



### 3.3.3 Scale-location plot and assumption of equal variance

The standardized residual is given by  $\frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$ . The standardized residual is a measure of the strength of the difference between observed and expected values.

A scatter plot showing the square root of the absolute value of standardised residuals against fitted values is known as a scale-location plot. This graph indicates if residuals are distributed evenly across predictor ranges. This is how we may graphically test the equal variance assumption. If the graph is spread equally accross the horizontal line, the assumption is meaningful and it is so in our linear model:



All the plots above show that the hedonic pricing model with OLS is indeed a good way to model the dependence of prices and locational, energy factors.

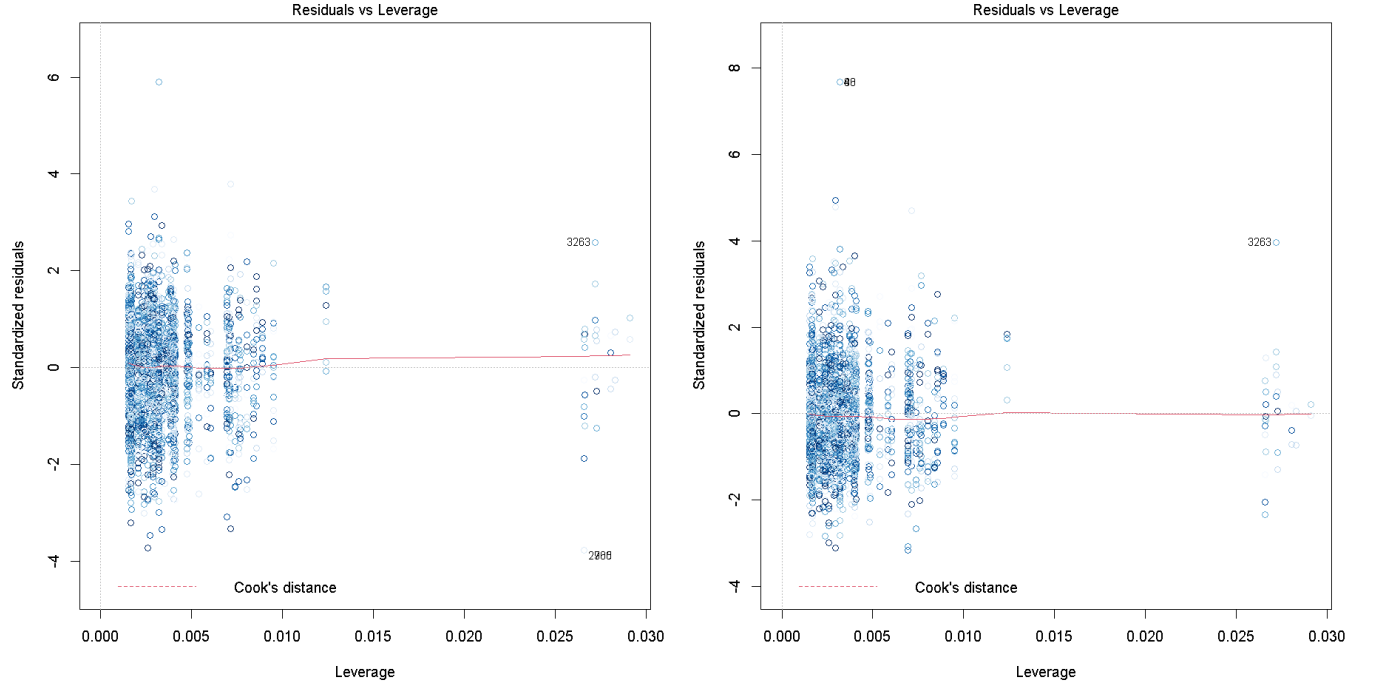
### 3.3.4 Leverage

The leverage score for the  $i^{th}$  independent observation  $x_i$  is given as:

$$h_{ii} = [H]_{ii} = \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i = h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$$

It is the degree by which  $i$ th measured value influences the  $i$ th fitted value (i.e.,  $\hat{y}_i$ ). Points with extreme values of variables  $x_i$  are said to have high leverage. High leverage points have a stronger capacity to shift the regression line and can be influential, causing the outcome and accuracy to be distorted.

The figure below shows the standardized residuals plotted against the leverage.



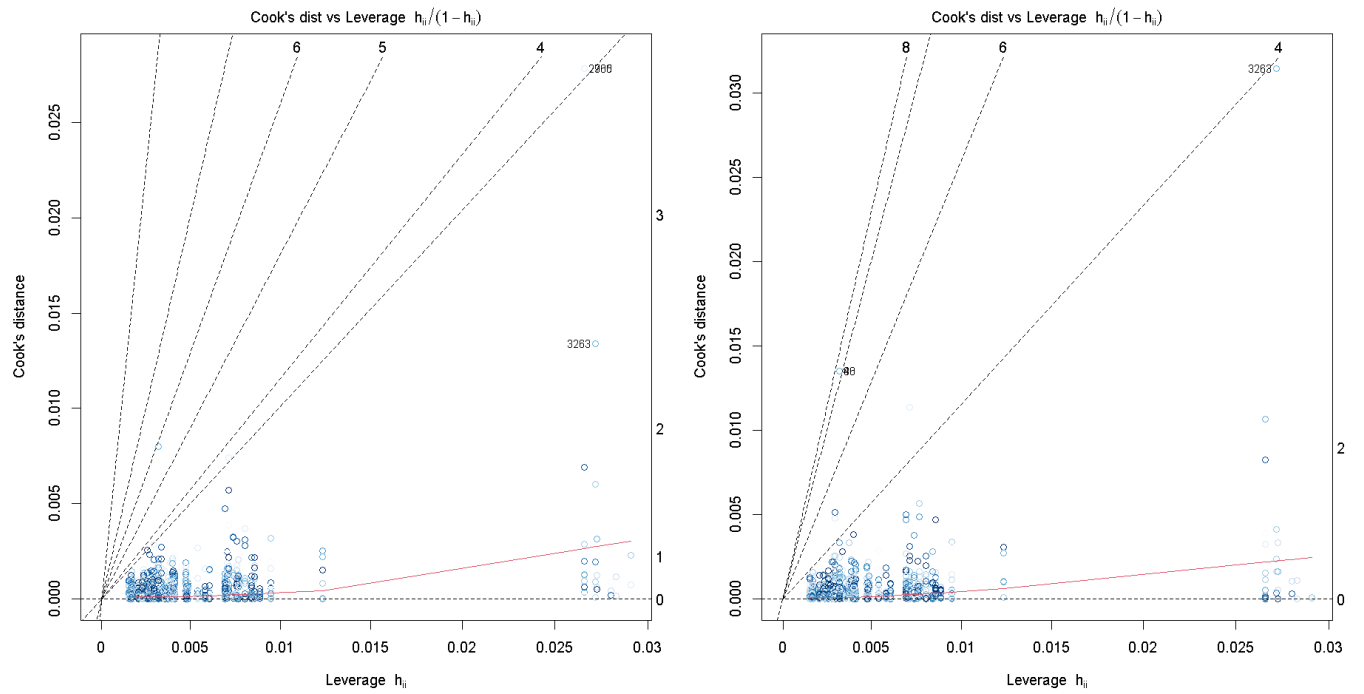
The contour lines of Cook's Distance are not visible in the graph. Hence, it is evident that the number of outliers are too less. So, we do not need to rerun the model by removing outliers.

### 3.3.5 Cook's Distance

Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. The Cook's distance statistic for every observation measures the extent of change in model estimates when that particular observation is omitted. Cook's distance  $D_i$  of observation  $i$  is defined as the sum of all the changes in the regression model when observation  $i$  is removed from it

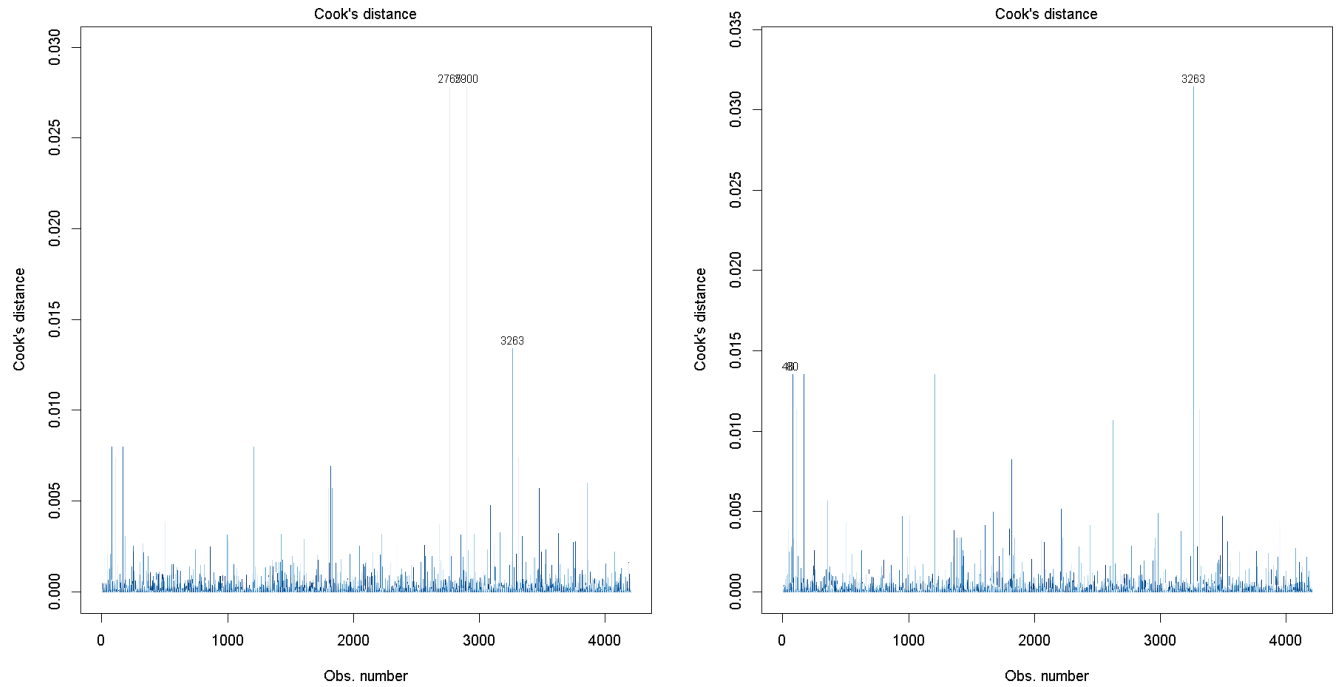
$$D_i = \frac{1}{(K+1)s^2} \sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2 = \frac{e_i^2}{(K+1)s^2} \left[ \frac{h_{ii}}{(1-h_{ii})^2} \right].$$

where  $\hat{y}_{j(i)}$  is the fitted response value obtained when excluding  $i$ , and  $s$  is the standard error. The figure below shows the Cook's Distance of each response value.



The graph shows that the Cook's Distance is mostly small for each value. Hence, our fitted model is not highly distorted.

The figure below shows the plot of leverage and cooks distance:



## 4 Price vs socio-economic variables

We would like to relate price\_1 and price\_2 with the socio-economic variables imd\_score, income\_score, emp\_score, educ\_score, health\_score, crime\_score, barrier\_score, living\_score. This would reveal the dependency of housing prices on socio-economic environment. In many statistical techniques, we assume that the errors are normally distributed and regression is one of them. The plots given below clearly show that the price\_1 and price\_2 variables are not normal.



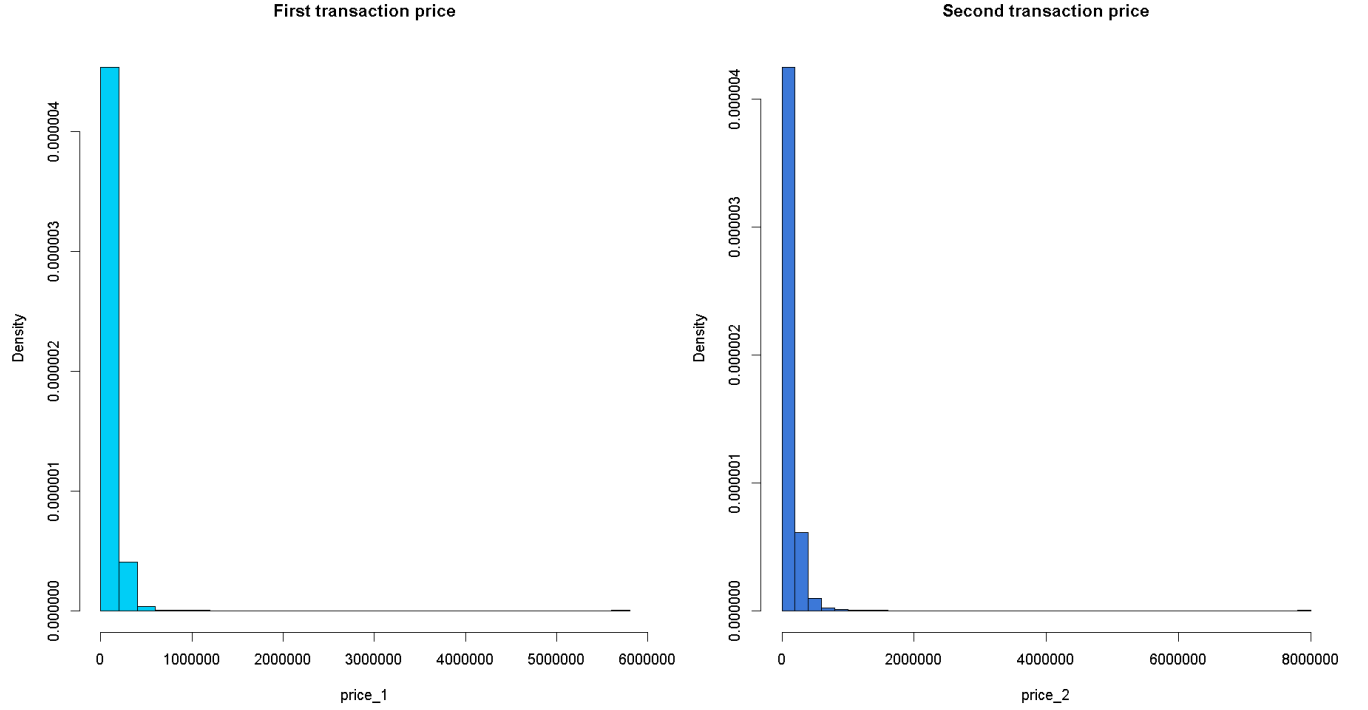


Figure 10: Distributions of the price paid in the first (left hand side) and second (right hand side) property sale transactions

Instead of directly applying linear regression, we start by applying box-cox transformation on the dependent variable, developed by Box and Cox(1964).

## 4.1 Box-cox transformation

The Box-Cox transformation transforms our data so that it closely resembles a normal distribution. Suppose variable  $y$  is regressed against variable  $x$ . The linear model is a very good fit if  $y$  is normally distributed. One reason is that confidence intervals and hypothesis tests are all based on this assumption. The objective of box-cox transformation is to transform  $y$  to  $y^{(\lambda)}$  so that transformed  $y$  is approximately normal. The family of transformations considered are:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}$$

By using maximum likelihood estimators, a value of  $\lambda$  is found which makes  $y^{(\lambda)}$  approximately normal.

## 4.2 Results

R already has a function in its MASS library, `boxcox(object, ...)` where `object` refers to the linear model we want to fit. After running the code, it is noticed that,

$$\lambda_{\text{price}_1} = 0.02$$

$$\lambda_{\text{price}_2} = -0.3$$

We take  $0.02 \approx 0$ , and we transform `price_1` variable to  $\ln(\text{price}_1)$  and `price_2` variable to  $\frac{(\text{price}_2)^{-0.03}-1}{-0.03}$ .

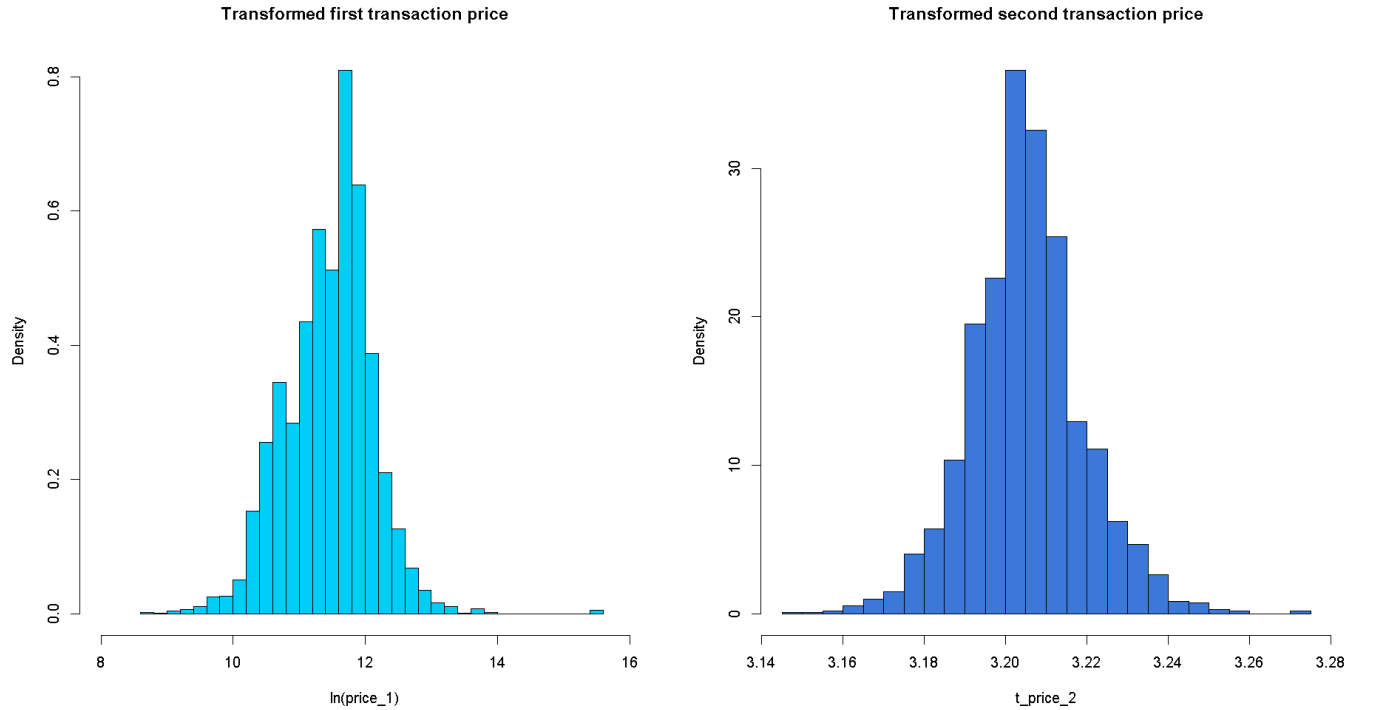


Figure 11: Distributions of the transformed price paid in the first (left hand side) and second (right hand side) property sale transactions

The plots above are symmetric and look like normal distribution. We can now move to regression analysis.

### 4.3 Summary

Variable	Estimate	Std Error	t-value	Pr(>  t )	
(Intercept)	10.82817670995	0.04523532930	239.374	0(approx)	***
imd_score	-0.00006253022	0.00000696566	-8.977	0(approx)	***
income_score	-0.00001487051	0.00000390421	-3.809	0.000142	***
emp_score	0.00003609509	0.00000368207	9.803	0(approx)	***
educ_score	0.00003438949	0.00000196613	17.491	0(approx)	***
health_score	0.00003242077	0.00000250168	12.960	0(approx)	***
crime_score	0.00000174045	0.00000167774	1.037	0.299622	
barrier_score	-0.00000001551	0.00000146222	-0.011	0.991539	
living_score	0.00001157480	0.00000167902	6.894	0(approx)	***

Table 6: Trasformed price\_1 vs socio-econoic variables.  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Variable	Estimate	Std Error	t-value	Pr(>  t )	
(Intercept)	3.18665988102	0.00087845134	3627.588	0(approx)	***
imd_score	-0.00000193203	0.00000013527	-14.283	0(approx)	***
income_score	-0.00000017160	0.00000007582	-2.263	0.0237	*
emp_score	0.00000103479	0.00000007150	14.472	0(approx)	***
educ_score	0.00000088244	0.00000003818	23.112	0(approx)	***
health_score	0.00000085907	0.00000004858	17.683	0(approx)	***
crime_score	0.00000014584	0.00000003258	4.476	0.000008	***
barrier_score	0.00000004633	0.00000002840	1.632	0.1028	
living_score	0.00000026311	0.00000003261	8.069	0(approx)	***

Table 7: Trasformed price\_2 vs socio-econoic variables.  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

### 4.4 Partial residue plots

A partial residual plot is a scatterplot to show the relationship between a given independent variable( $x_i$ ) and the response variable( $y$ ) given that other independent variables are also in the model. The partial residual plot is a useful tool to verify if there exists any linear relationship of

a given variable on the response variable. Partial residuals are computed as:

$$\text{Residuals} + \hat{\beta}_i x_i$$

Note that the Residuals are the residuals from the full model.

In partial residual plot, these partial residuals are plotted against  $x_i$ . The `termplot` function of R helps us to create such plots. This function also takes the projection of the regression hyperplane to the  $x_i$  vs  $y$  plane. The slope of the line is the coefficient of  $x_i$  in the original linear model. Moreover, it further adds a smooth curve to the plot. This helps us analyse whether there is any non-linear relationship with the  $x_i$ .

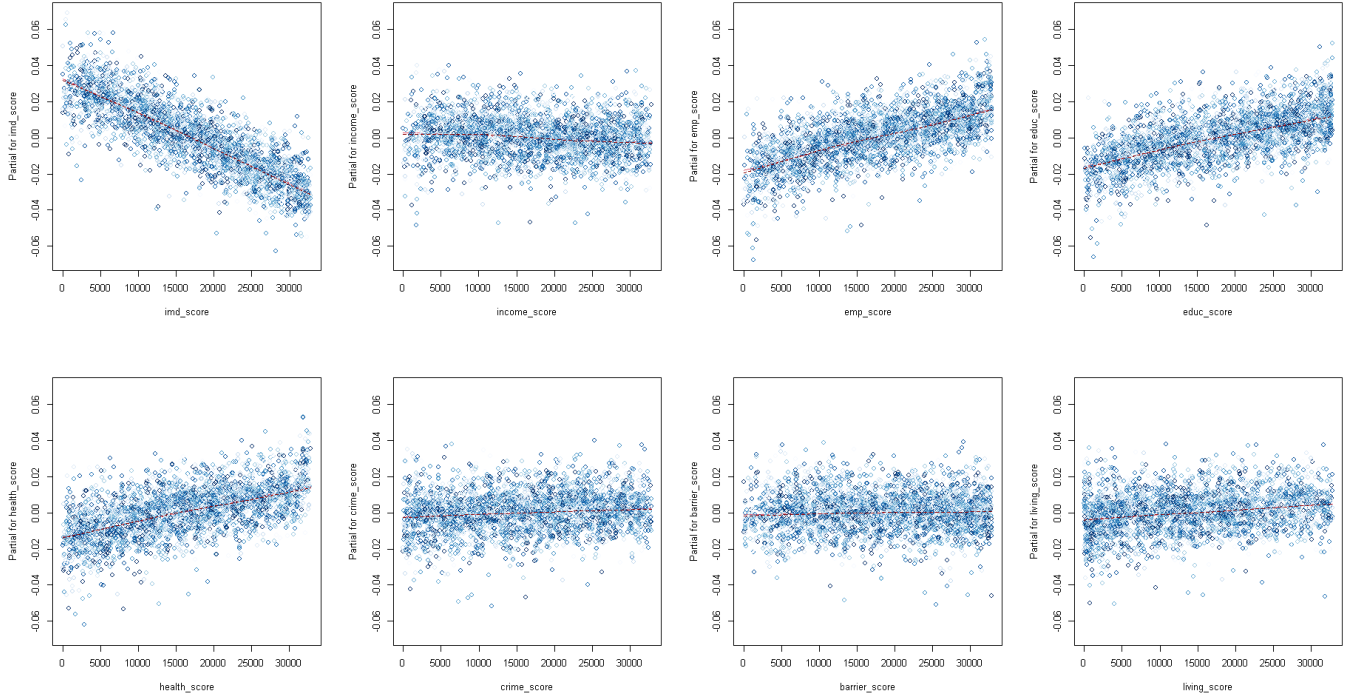


Figure 12: Plot of partial residuals and regression terms of transformed price\_1 across the predictor variables along with smooth fitted curve

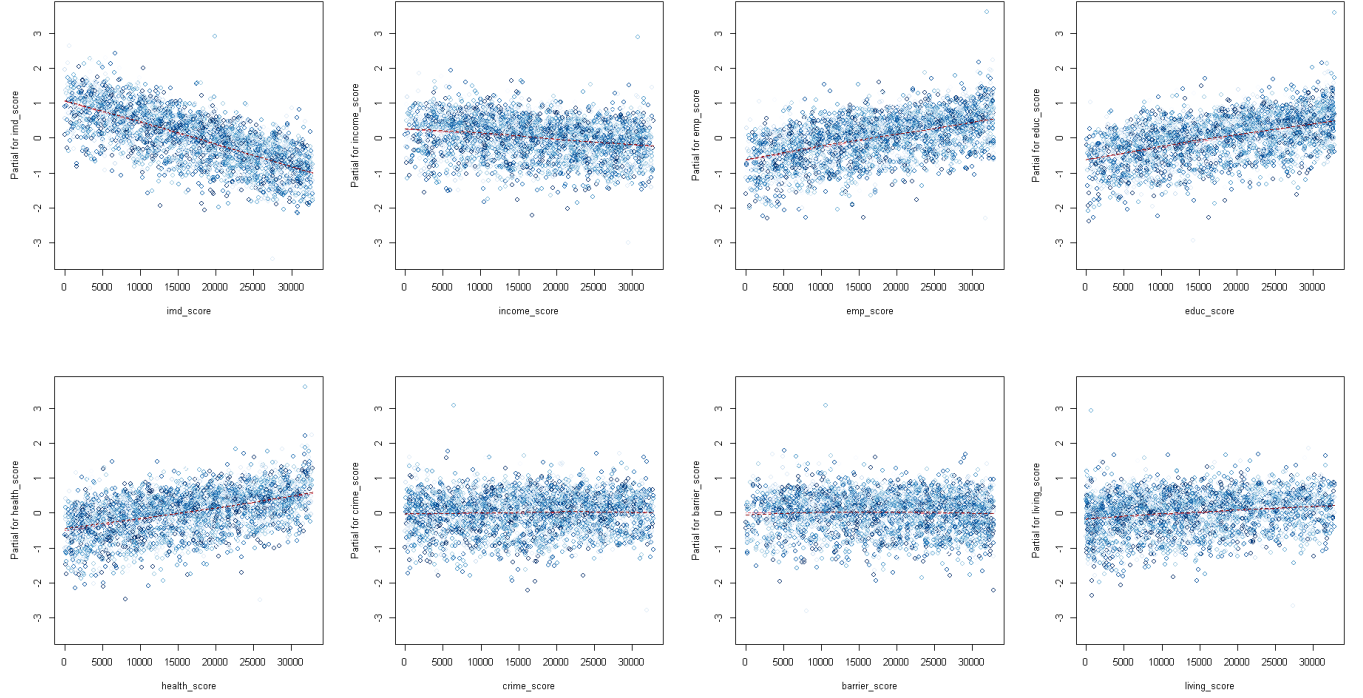


Figure 13: Plot of partial residuals and regression terms of transformed price\_2 across the predictor variables along with smooth fitted curve

The above plots suggests that there is no non-linear relationship and our model quite efficiently measures the relationship.

## 5 Repeat sales index

In this section we use the repeat-sales method to examine how house values fluctuate over time by comparing the sale prices of the same piece of property. Using this method, assuming a house undergoes no changes, to assess how prices change over time, one need only look at the difference in sale prices of the same house.

### 5.1 Preparing the data for repeat sales index

The original dataset had the repeat sales prices of houses. The repeat sales transactions took place from 1995 to 2012, for this data set. We made a new dataset that consisted of variables Y1995, Y1996, ... and so on upto Y2012. For any house, these variables take value -1 if the first sales transaction took place in that year, 1 if the second sales transaction took place in that year and 0 otherwise. The variable log\_change each derived as:

$$\text{log\_change} = \ln\_price\_2 - \ln\_price\_1$$

## 5.2 Regression

Let there be  $T$  time periods where sales can occur from  $y_1, \dots, y_T$ . Each  $y_t$  is a categorical variable:  $y_t = -1$  ( $y_t = 1$  resp.) if first (second resp.) sale takes place in year  $y_t$  and  $y_t = 0$  otherwise. Let  $B'_t$  denote the (true, but unknown) repeat sales index for year  $y_t$ . Let  $P_{it}$  and  $P_{it'}$  denote the sales prices of  $i$ th house at the  $t$ th period. A probabilistic model is made as follows:

$$\frac{P_{it'}}{P_{it}} = \frac{B_{t'}}{B_t} U_{itt'}$$

Taking logarithm,

$$\ln(P_{it'}) - \ln(P_{it}) = \ln(B_{t'}) - \ln(B_t) + u_{itt'}$$

where  $u_{itt'} = \ln(U_{itt'})$  is the residual. Again, four assumptions on the distribution of the residuals: They have mean 0, constant variance,  $u_{itt'} \sim N(0, \sigma^2)$ , uncorrelated with each other. The above equation can be written as:

$$\ln(P_{it'}) - \ln(P_{it}) = \sum_{j=1}^T y_j \ln(B_j) + u_{itt'}$$

The model in the above equation is fit using linear regression and the estimated log indices are transformed to repeat sales indices with the exponential function. The estimated linear equation will be:

$$\ln(\hat{P}_{t'}) - \ln(\hat{P}_t) = \hat{B}_0 + \sum_{j=1}^l \ln(\hat{B}_j) y_j$$

We run OLS regression using R where the dependent variable is log\_change and the regressors are Y1995, Y1996, ..., Y2012 (called  $y_1, \dots, y_T$ ,  $T = 18$ ). The Y1995 is taken as base year and is held out of the regression.

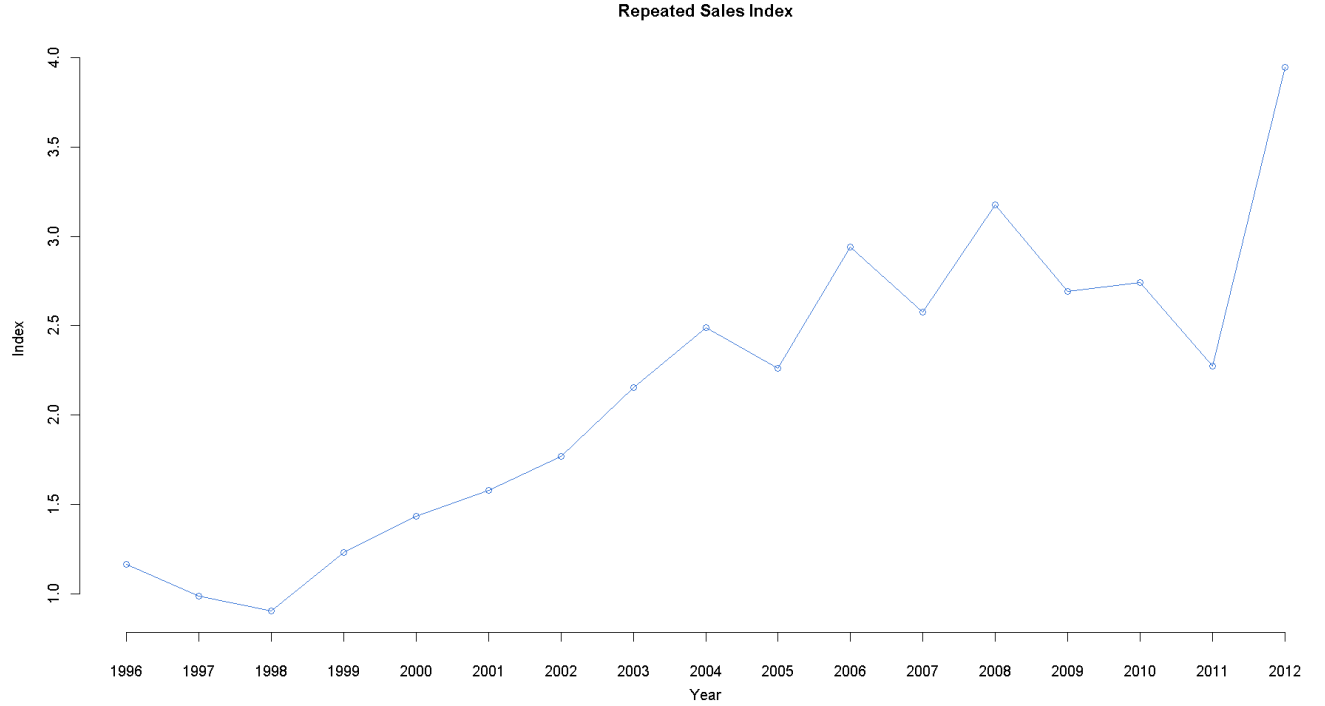


Figure 14: Repeat Sales Index of each year with base year being 1995

The Figure 14 shows the line graph of the estimated coefficients. There is increase of change in prices over years. The inflation in prices is evident in the figure. There are major retractions in 1998, 2011. The index reaches its maximum at 2012.

## 6 Conclusions

Landlords can make informed investments when there is a clear link between price and energy efficiency. In addition, we have taken regional aspects into account. We linked the pricing to socio-economic parameters in the second model. The empirical findings confirm that energy efficiency characteristics have a minor but considerable impact on both transaction prices. Furthermore, the current study was unable to control the correlation between factors. The graph clearly demonstrates that several factors that we expected to be independent have a high degree of association. In some circumstances, better transformations or the deletion of associated columns may be beneficial. Furthermore, our model was only focused on pricing, and we were unable to establish a link between EPCs and Regions. In order to produce a comprehensive idea on this divided incentive dilemma, the data can also be enlarged to a much larger database.

## 6.1 Limitations

- For hedonic regression, a large data must be collected and processed with. However, the data that was utilised only had 4201 values. The Hedonic Price Index calculates how much individuals are prepared to pay for alleged differences in environmental quality and its implications. However, if individuals are ignorant of the link between environmental attributes and property worth, the value will not be reflected in the price.
- One major issue is that repeat sales approach remove residences that sell only once during the data gathering period. Another issue is that all houses age, and as a result, the houses alter, which might impact the index, which is not taken into account here.
- While using box-cox transformation, the lambda obtained for price\_2 was not zero, and is therefore not possible to interpret. Also box-cox transformation does not always guarantee normality.

## 6.2 Extensions

The chief findings support existing literature on “Is there an Economic Case for Energy-Efficient Dwellings in the UK Private Rental Market?” in the housing market by adding to empirical evidence on the relationship between energy efficiency ratings and pricing decisions and demonstrating that the features captured by EPCs are widely significant to housing price.

Here is the list of extension we have attempted, which were not there in the original paper.

1. Plots for visualization viz.
  - [Figure 6](#)
  - [Figure 7](#)
  - [Figure 8](#)
  - [Figure 9](#)
2. Diagnostic plots in [subsection 3.3](#)
3. Box-Cox transformation in [subsection 4.1](#)
4. Partial Residue Plots, [subsection 4.4](#)
5. Repeat sales index of the the price variables across years, [section 5](#)



### 6.3 Future Scope

More complex regressions can be applied in terms of statistical approaches. Case and Quigley's (1991) hybrid index can also be used. The Case-Shiller(1987,1989) regression method can be used to make a development on the repeat sales index. A recent method called autoregressive index has been developed by Nagaraja, Brown, and Zhao (2011). On a bigger dataset, the analysis may be run with additional variables, taking into account, housing attributes.etc.

## 7 References

1. Franz Fuerst, Michel Ferreira Cardia Haddad(2020). [“Real estate data to analyse the relationship between property prices, sustainability levels and socio-economic indicators.”](#) Data in Brief. 33 106359.
2. F. Fuerst, M.F.C. Haddad, H. Adan(2015). [“Is there an economic case for energy-efficient dwellings in the UK private rental market?”](#) Journal for Cleaner Prod. 245 (2020) 118642.
3. F. Fuerst , P. McAllister , A. Nanda , P. Wyatt(2015). [“Does energy efficiency matter to home-buyers? an investigation of EPC ratings and transaction prices in England?”](#) Energy Economy. 48 145–156 .
4. Bailey, M.J., Muth, R.F., Nourse, H.O. (1963). “A regression method for real estate price index construction. Journal of the American Statistical Association.” 58 933-942
5. Ryan, Thomas P (2015). Modern Regression Methods. Wiley-Interscience.
6. G. E. P. Box and D. R. Cox(1964). [“An Analysis of Transformations.”](#) Journal of the Royal Statistical Society. Series B Vol. 26. No. 2.
7. Case, K.E., Shiller, R.J. (1987). Prices of single-family homes since 1970: new indexes for four cities. New England Economic Review. Sept./Oct. 45-56.
8. Case, K.E., Shiller, R.J. (1989). The efficiency of the market for single family homes. The American Economic Review. 79 125-137
9. Case, B., Quigley, J.M. (1991). The dynamics of real estate prices. The Review of Economics and Statistics. 73 50-58
10. Chaitra H. Nagaraja. Lawrence D. Brown. Linda H. Zhao(2011). [“An autoregressive approach to house price modeling.”](#) Ann. Appl. Stat. 5 (1) 124 - 149, March 2011.

11. [Repeat Sales House Price Index Methodology](#). Journal of Real Estate Literature. Vol. 22, No. 1 (2014), pp. 23-46.