Coursework_2

229040002

2023-01-07

GY7702 Assignment 2

Introduction

The following report aims to explore how energy consumption in the LSOAs of Stockton-on-Tees is related to the number of people living in a household and in how many dimensions are the households deprived according to the indices of multiple deprivation, for each LSOA. The report aims to analyse the given data, build a robust multiple linear regression model for the concerned authorities to use for prioritising support to the key areas, and end with a conclusion discussing the main aspects of the analysis and the model. Two models were analysed for the given data, however, the best out of the two (based on robustness and the ability to predict the energy consumption data) was included in the report.

Data selection and manipulation

For the analysis and finally the model, the gas and the electricity consumption data for 2020 was selected. The LAD code for Stockton-on-Tees is E06000004. However, the LSOA data for the two was available for 2011 from the census. The two tables above were joined with the census data available for 2021 so as to get the appropriate LSOA codes for the electricity and gas consumption data.

The resulting table was joined with the household sizes data and the dimensions of deprivation data. The desired columns were mutated and a resulting normalised dataset was formed and a summary of the essential columns was presented. Based on the summary, only the necessary columns were selected for further exploratory analysis (removing null value columns and such)


```
# loading libraries
library(tidyverse)
library(readr)
library(knitr)
library(ggplot2)
library(pastecs)
library(GGally)
library(magrittr)
library(lmtest)
library(lm.beta)
```

```
# reading household deprivation
# csv for all of UK

uk_household_deprivation_2021 <- read_csv("census2021-ts011-lsoa.csv")
```

```
# reading household size
# csv for all of UK
```

```r
uk_household_size_2021 <- read_csv("census2021-ts017-lsoa.csv")
```

```r
# reading domestic gas
# consumption for all of UK

uk_domestic_gas_consump <- read_csv("LSOA_domestic_gas_2010-20.csv")
```

```r
# reading domestic electricity
# consumption for all of UK

uk_domestic_electricity_consump <- read_csv("LSOA_domestic_elec_2010-20.csv")
```

```r
# reading lookup table
# for LSOA 2011 to 2021

lookup_lsoa11_lsoa21 <-
  read_csv("LSOA_(2011)_to_LSOA_(2021)_to_Local_Authority_District_(2022)_Lookup_for_England_and_Wales.
```

```r
# filtering electricity data
# for Stockton-on-Tees

stockton_on_tees_elec_consp_lsoa11 <- uk_domestic_electricity_consump %>%
  filter(`LAD code` == "E06000004",
         Year == "2020"
) %>%
  select(
    -`MSOA code`, -`MSOA name`,
    -`LAD name`, -`LSOA name`
  )
```

```r
# filtering gas data
# for Stockton-on-Tees

stockton_on_tees_gas_consp_lsoa11 <- uk_domestic_gas_consump %>%
  filter(`LAD code` == "E06000004",
         Year == "2020"
) %>%
  select(
    -`MSOA code`, -`MSOA name`,
    -`LAD name`, -`LSOA name`
  )
```

```r
# joining the 2020 electricity consumption with
# lookup table LSOA 2011 to 2021 to
# obtain the LSOA codes of Stockton-on-Tees for 2021

stockton_on_tees_elec_2020_lsoa21 <-
  stockton_on_tees_elec_consp_lsoa11 %>%
    inner_join(
      lookup_lsoa11_lsoa21,
      by = c("LSOA code" = "F_LSOA11CD")
    ) %>%
# removing unnecessary columns
# from Stockton-on-Tees electricity consumption 2020
# for 2021 LSOAs
```

```
   select(
    -LSOA11NM, -LSOA21NM,
    -CHGIND, -LAD22NM
  )
```

```
# joining the 2020 gas consumption with
# lookup table LSOA 2011 to 2021 to
# obtain the LSOA codes of Stockton-on-Tees for 2021

stockton_on_tees_gas_2020_lsoa21 <-
  stockton_on_tees_gas_consp_lsoa11 %>%
  inner_join(
    lookup_lsoa11_lsoa21,
    by = c("LSOA code" = "F_LSOA11CD")
  ) %>%
# removing unnecessary columns
# from Stockton-on-Tees gas consumption 2020
# for 2021 LSOAs
  select(
    -LSOA11NM, -LSOA21NM,
    -CHGIND, -LAD22NM,
    -LAD22NMW, -ObjectId
  )
```

```
# joining gas consumption and
# electricity consumption data
# to get energy consumption data

energy_consp_stockton_2020 <-
  stockton_on_tees_gas_2020_lsoa21 %>%
  inner_join(
    stockton_on_tees_elec_2020_lsoa21,
    by = c("LSOA21CD")
  )
```

```
# cleaning energy consumption data
# by removing unnecessary columns

energy_consp_final <-
  energy_consp_stockton_2020 %>%
  select(
    -`LAD code.x`, -`LSOA code.x`,
    -Year.x, -LAD22CD.x,
    -`LAD code.y`, -`LSOA code.y`,
    -Year.y, -LAD22CD.y,
    -LAD22NMW, -ObjectId
)
```

```
# reordering the energy consumption data

energy_cons_stockton_final <-
  energy_consp_final %>%
  select(
    LSOA21CD, `Number of gas meters`,
    `Number of non-consuming gas meters`, `Mean gas consumption (kWh per meter)`,
```

```
    `Median gas consumption (kWh per meter)`, `Total gas consumption (kWh)`,
    `Number of electricity meters`, `Mean electricity consumption (kWh per meter)`,
    `Median electricity consumption (kWh per meter)`, `Total electricity consumption (kWh)`
  )
```

```
# joining the energy consumption data
# with the household size data
# for Stockton-on-Tees

energy_cons_stockton_household_size <-
  energy_cons_stockton_final %>%
  inner_join(
    uk_household_size_2021,
    by = c("LSOA21CD" = "geography code")
  )
```

```
# joining the energy consumption and
# household size joined data with deprivation
# data to get combined data for Stockton-on-Tees

energy_cons_stockton_household_deprn <-
  energy_cons_stockton_household_size %>%
  inner_join(
    uk_household_deprivation_2021,
    by = c("LSOA21CD" = "geography code")
  )
```

```
# A. combining more than four people in
# a household into one column,

energy_census_data <-
  energy_cons_stockton_household_deprn %>%
  mutate(
    ppl_hh_more_than_4 = rowSums(
      across(
      `Household size: 5 people in household; measures: Value`:
        `Household size: 8 or more people in household; measures: Value`)
    )
  ) %>%
# B. combining household data
# for 1 to 4 people in a household

  mutate(
    ppl_hh_1_to_4 = rowSums(
      across(
      `Household size: 1 person in household; measures: Value`:
        `Household size: 4 people in household; measures: Value`)
    )
  ) %>%
# C. renaming columns appropriately
  rename(
    tot_hh_spaces =
      `Household size: Total: All household spaces; measures: Value`,
    ppl_hh_0 =
      `Household size: 0 people in household; measures: Value`,
```

```
    ppl_hh_1 =
      `Household size: 1 person in household; measures: Value`,
    ppl_hh_2 =
      `Household size: 2 people in household; measures: Value`,
    ppl_hh_3 =
      `Household size: 3 people in household; measures: Value`,
    ppl_hh_4 =
      `Household size: 4 people in household; measures: Value`,
    tot_hh_depr =
      `Household deprivation: Total: All households; measures: Value`,
    hh_not_depr_any_dim =
      `Household deprivation: Household is not deprived in any dimension; measures: Value`,
    hh_depr_1_dim =
      `Household deprivation: Household is deprived in one dimension; measures: Value`,
    hh_depr_2_dim =
      `Household deprivation: Household is deprived in two dimensions; measures: Value`,
    hh_depr_3_dim =
      `Household deprivation: Household is deprived in three dimensions; measures: Value`,
    hh_depr_4_dim =
      `Household deprivation: Household is deprived in four dimensions; measures: Value`
)
```

```
# reordering energy-census data and
# removing unnecessary columns

energy_census_clean <-
  energy_census_data %>%
  select(
    LSOA21CD, `Number of gas meters`, `Number of non-consuming gas meters`,
    `Mean gas consumption (kWh per meter)`,
    `Median gas consumption (kWh per meter)`,
    `Total gas consumption (kWh)`,
    `Number of electricity meters`,
    `Mean electricity consumption (kWh per meter)`,
    `Median electricity consumption (kWh per meter)`,
    `Total electricity consumption (kWh)`,
    tot_hh_spaces, ppl_hh_0,
    ppl_hh_1, ppl_hh_2,
    ppl_hh_3, ppl_hh_4,
    ppl_hh_1_to_4, ppl_hh_more_than_4,
    tot_hh_depr, hh_not_depr_any_dim,
    hh_depr_1_dim, hh_depr_2_dim,
    hh_depr_3_dim, hh_depr_4_dim
  )
```

```
# Normalising the dataset
# calculating percentages of household size
# and household deprivation;
# calculating total, mean and median
# energy consumption

energy_census_data_final <-
  energy_census_clean %>%
  mutate(
```

```
    ppl_hh_0_perc = (ppl_hh_0/tot_hh_spaces) * 100
  ) %>%
  mutate(
    ppl_hh_1_to_4_perc = (ppl_hh_1_to_4/tot_hh_spaces) * 100
  ) %>%
  mutate(
    ppl_hh_more_than_4_perc = (ppl_hh_more_than_4/tot_hh_spaces) * 100
  ) %>%
  mutate(
    across(
      ppl_hh_1:ppl_hh_4,
      function(x)(x/tot_hh_spaces) * 100,
      .names = "{.col}_perc"
    )
  ) %>%
  mutate(
    across(
      hh_not_depr_any_dim:hh_depr_4_dim,
      function(x)(x/tot_hh_depr) * 100,
      .names = "{.col}_perc"
    )
  ) %>%
  mutate(
  Total_energy_consumption_kWh =
    (`Total gas consumption (kWh)`+ `Total electricity consumption (kWh)`)
  ) %>%
  mutate(
    Mean_energy_consumption_kWh_per_meter =
      (`Mean gas consumption (kWh per meter)`+
      `Mean electricity consumption (kWh per meter)`)
  ) %>%
  mutate(
    Median_energy_consumption_kWh_per_meter =
     ( `Median gas consumption (kWh per meter)`+
      `Median electricity consumption (kWh per meter)`)
  )
```

```
# getting the summary of the required columns

energy_census_data_final %>%
  select(
     Total_energy_consumption_kWh,
     Mean_energy_consumption_kWh_per_meter,
     Median_energy_consumption_kWh_per_meter,
     ppl_hh_0_perc:ppl_hh_4_perc,
     ppl_hh_1_to_4_perc, ppl_hh_more_than_4_perc,
     hh_not_depr_any_dim_perc,
     hh_depr_1_dim_perc:hh_depr_4_dim_perc
  ) %>%
  summary()
```

```
##  Total_energy_consumption_kWh Mean_energy_consumption_kWh_per_meter
##  Min.   : 6989693             Min.   :12095
##  1st Qu.: 9629536             1st Qu.:14883
```

```
##  Median :11158530             Median :16048
##  Mean   :12220335             Mean   :16876
##  3rd Qu.:13417530             3rd Qu.:18666
##  Max.   :25731432             Max.   :27349
##  Median_energy_consumption_kWh_per_meter ppl_hh_0_perc ppl_hh_1_to_4_perc
##  Min.   :10114                           Min.   :0     Min.   :80.00
##  1st Qu.:13479                           1st Qu.:0     1st Qu.:92.62
##  Median :15080                           Median :0     Median :94.49
##  Mean   :15405                           Mean   :0     Mean   :94.08
##  3rd Qu.:17028                           3rd Qu.:0     3rd Qu.:95.99
##  Max.   :24202                           Max.   :0     Max.   :98.19
##  ppl_hh_more_than_4_perc ppl_hh_1_perc    ppl_hh_2_perc   ppl_hh_3_perc
##  Min.   : 1.806          Min.   : 7.979   Min.   :17.14   Min.   : 7.286
##  1st Qu.: 4.005          1st Qu.:25.047   1st Qu.:30.30   1st Qu.:14.096
##  Median : 5.513          Median :29.950   Median :34.72   Median :16.682
##  Mean   : 5.919          Mean   :30.810   Mean   :34.47   Mean   :16.537
##  3rd Qu.: 7.376          3rd Qu.:35.883   3rd Qu.:39.52   3rd Qu.:18.641
##  Max.   :20.000          Max.   :60.423   Max.   :48.46   Max.   :25.053
##  ppl_hh_4_perc    hh_not_depr_any_dim_perc hh_depr_1_dim_perc
##  Min.   : 4.791   Min.   :20.06            Min.   :21.43
##  1st Qu.: 8.747   1st Qu.:37.79            1st Qu.:30.26
##  Median :11.879   Median :47.69            Median :33.01
##  Mean   :12.264   Mean   :48.07            Mean   :32.65
##  3rd Qu.:14.349   3rd Qu.:57.35            3rd Qu.:35.61
##  Max.   :35.006   Max.   :76.25            Max.   :40.57
##  hh_depr_2_dim_perc hh_depr_3_dim_perc hh_depr_4_dim_perc
##  Min.   : 1.827     Min.   : 0.000     Min.   :0.0000
##  1st Qu.: 9.398     1st Qu.: 1.129     1st Qu.:0.0000
##  Median :14.395     Median : 2.793     Median :0.0000
##  Mean   :15.176     Mean   : 3.944     Mean   :0.1636
##  3rd Qu.:20.294     3rd Qu.: 6.197     3rd Qu.:0.2678
##  Max.   :33.236     Max.   :13.031     Max.   :2.0498
```

```
# getting the normalised dataset
# keeping all necessary columns

energy_census_normalised <-
  energy_census_data_final %>%
  select(
    LSOA21CD, ppl_hh_0_perc,
    ppl_hh_1_perc : ppl_hh_4_perc,
    ppl_hh_more_than_4_perc, ppl_hh_1_to_4_perc,
    hh_not_depr_any_dim_perc : hh_depr_4_dim_perc,
    Total_energy_consumption_kWh,
    Mean_energy_consumption_kWh_per_meter,
    Median_energy_consumption_kWh_per_meter
  )

head(energy_census_normalised)
```

```
## # A tibble: 6 x 16
##   LSOA21CD  ppl_hh_0_p~1 ppl_h~2 ppl_h~3 ppl_h~4 ppl_h~5 ppl_h~6 ppl_h~7 hh_no~8
##   <chr>            <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 E01012242            0    26.3    42.5    15.2   11.5     4.54    95.5    50
## 2 E01012243            0    31.1    35.8    17.4   11.7     4       96      53.6
```

```
## 3 E01012244            0    31.5    37.3    16.9   11.1     3.23    96.8    56.1
## 4 E01012246            0    26.5    44.5    13.4   12.3     3.34    96.7    53.9
## 5 E01012296            0    41.4    36.5    10.4    8.71    3.02    97.0    47.7
## 6 E01012297            0    19.8    39.4    19.7   15.5     5.68    94.3    62.7
## # ... with 7 more variables: hh_depr_1_dim_perc <dbl>,
## #   hh_depr_2_dim_perc <dbl>, hh_depr_3_dim_perc <dbl>,
## #   hh_depr_4_dim_perc <dbl>, Total_energy_consumption_kWh <dbl>,
## #   Mean_energy_consumption_kWh_per_meter <dbl>,
## #   Median_energy_consumption_kWh_per_meter <dbl>, and abbreviated variable
## #   names 1: ppl_hh_0_perc, 2: ppl_hh_1_perc, 3: ppl_hh_2_perc,
## #   4: ppl_hh_3_perc, 5: ppl_hh_4_perc, 6: ppl_hh_more_than_4_perc, ...
## # i Use `colnames()` to see all variable names
```

Exploratory Analysis

A variable "energy_census_statdesc" is created from the "energy_census_normalised" data where the preliminary exploratory statistical analysis takes place. The required variables – households with 1 person to household with 4 people (percentages), households with people 1-4 and households with more than 4 people (percentages); household deprivation values 0-4 (percentages); the total energy consumption, the mean and median energy consumption (kWh and kWh/meter resp.) are chosen for further analysis. Variables like households with 0 people (%) were discarded due to presence of null values and that were not appropriate for further consideration.

```
# exploratory analysis

energy_census_statdesc <-
  energy_census_normalised %>%
  select(
    ppl_hh_1_perc : ppl_hh_4_perc,
    ppl_hh_more_than_4_perc, ppl_hh_1_to_4_perc,
    hh_not_depr_any_dim_perc : hh_depr_4_dim_perc,
    Total_energy_consumption_kWh,
    Mean_energy_consumption_kWh_per_meter,
    Median_energy_consumption_kWh_per_meter
  ) %>%
  stat.desc(norm = TRUE)

energy_census_statdesc
```

```
##              ppl_hh_1_perc ppl_hh_2_perc ppl_hh_3_perc ppl_hh_4_perc
## nbr.val        124.0000000   124.0000000  124.00000000  1.240000e+02
## nbr.null         0.0000000     0.0000000    0.00000000  0.000000e+00
## nbr.na           0.0000000     0.0000000    0.00000000  0.000000e+00
## min              7.9794080    17.1428571    7.28571429  4.791345e+00
## max             60.4229607    48.4629295   25.05307856  3.500644e+01
## range           52.4435527    31.3200723   17.76736427  3.021509e+01
## sum           3820.4997746  4274.2890431 2050.59344456  1.520718e+03
## median          29.9500835    34.7161933   16.68247944  1.187907e+01
## mean            30.8104821    34.4700729   16.53704391  1.226386e+01
## SE.mean          0.8603800     0.5842839    0.31890825  4.190362e-01
## CI.mean.0.95     1.7030695     1.1565541    0.63125932  8.294564e-01
## var             91.7914730    42.3320680   12.61110639  2.177332e+01
## std.dev          9.5807867     6.5063099    3.55121196  4.666189e+00
## coef.var         0.3109587     0.1887524    0.21474285  3.804830e-01
## skewness         0.5437278    -0.2897017   -0.08856918  1.512620e+00
```

```
## skew.2SE         1.2506890    -0.6663753   -0.20372783  3.479346e+00
## kurtosis         0.4561129    -0.3147603   -0.15470473  4.614756e+00
## kurt.2SE         0.5285549    -0.3647521   -0.17927566  5.347694e+00
## normtest.W       0.9758641     0.9867868    0.99363212  8.989029e-01
## normtest.p       0.0253814     0.2738194    0.85012235  1.181527e-07
##              ppl_hh_more_than_4_perc ppl_hh_1_to_4_perc
## nbr.val                 1.240000e+02       1.240000e+02
## nbr.null                0.000000e+00       0.000000e+00
## nbr.na                  0.000000e+00       0.000000e+00
## min                     1.806240e+00       8.000000e+01
## max                     2.000000e+01       9.819376e+01
## range                   1.819376e+01       1.819376e+01
## sum                     7.338995e+02       1.166610e+04
## median                  5.512546e+00       9.448745e+01
## mean                    5.918544e+00       9.408146e+01
## SE.mean                 2.523332e-01       2.523332e-01
## CI.mean.0.95            4.994780e-01       4.994780e-01
## var                     7.895331e+00       7.895331e+00
## std.dev                 2.809863e+00       2.809863e+00
## coef.var                4.747558e-01       2.986628e-02
## skewness                1.490012e+00      -1.490012e+00
## skew.2SE                3.427343e+00      -3.427343e+00
## kurtosis                4.247233e+00       4.247233e+00
## kurt.2SE                4.921799e+00       4.921799e+00
## normtest.W              9.031905e-01       9.031905e-01
## normtest.p              1.978830e-07       1.978830e-07
##              hh_not_depr_any_dim_perc hh_depr_1_dim_perc hh_depr_2_dim_perc
## nbr.val                 124.00000000       124.00000000       124.00000000
## nbr.null                  0.00000000         0.00000000         0.00000000
## nbr.na                    0.00000000         0.00000000         0.00000000
## min                      20.05856515        21.42857143         1.82724252
## max                      76.24584718        40.56795132        33.23572474
## range                    56.18728202        19.13937989        31.40848222
## sum                    5960.35156521      4048.48960457      1881.81490527
## median                   47.69245658        33.01451758        14.39486590
## mean                     48.06735133        32.64910971        15.17592666
## SE.mean                   1.16138510         0.36958225         0.65673014
## CI.mean.0.95              2.29889059         0.73156539         1.29995705
## var                     167.25310210        16.93728853        53.48051543
## std.dev                  12.93263709         4.11549372         7.31303736
## coef.var                  0.26905242         0.12605225         0.48188407
## skewness                  0.04324764        -0.49988563         0.36089138
## skew.2SE                  0.09947872        -1.14984265         0.83012649
## kurtosis                 -0.89106377        -0.13800319        -0.75255008
## kurt.2SE                 -1.03258671        -0.15992150        -0.87207362
## normtest.W                0.98309818         0.97514264         0.97050232
## normtest.p                0.12427688         0.02171308         0.00811449
##              hh_depr_3_dim_perc hh_depr_4_dim_perc Total_energy_consumption_kWh
## nbr.val           1.240000e+02       1.240000e+02                 1.240000e+02
## nbr.null          1.000000e+00       7.200000e+01                 0.000000e+00
## nbr.na            0.000000e+00       0.000000e+00                 0.000000e+00
## min               0.000000e+00       0.000000e+00                 6.989693e+06
## max               1.303075e+01       2.049780e+00                 2.573143e+07
## range             1.303075e+01       2.049780e+00                 1.874174e+07
```

```
## sum                4.890578e+02       2.028612e+01                1.515322e+09
## median             2.793080e+00       0.000000e+00                1.115853e+07
## mean               3.944015e+00       1.635977e-01                1.222034e+07
## SE.mean            2.930346e-01       2.616363e-02                3.516854e+05
## CI.mean.0.95       5.800440e-01       5.178930e-02                6.961396e+05
## var                1.064779e+01       8.488240e-02                1.533664e+13
## std.dev            3.263095e+00       2.913458e-01                3.916203e+06
## coef.var           8.273536e-01       1.780868e+00                3.204661e-01
## skewness           8.159140e-01       3.303050e+00                1.732302e+00
## skew.2SE           1.876775e+00       7.597712e+00                3.984661e+00
## kurtosis          -3.036193e-01       1.511938e+01                3.169146e+00
## kurt.2SE          -3.518416e-01       1.752071e+01                3.672484e+00
## normtest.W         8.908048e-01       6.048547e-01                8.282812e-01
## normtest.p         4.617067e-08       9.544140e-17                1.027011e-10
##              Mean_energy_consumption_kWh_per_meter
## nbr.val                                1.240000e+02
## nbr.null                               0.000000e+00
## nbr.na                                 0.000000e+00
## min                                    1.209510e+04
## max                                    2.734889e+04
## range                                  1.525379e+04
## sum                                    2.092641e+06
## median                                 1.604843e+04
## mean                                   1.687613e+04
## SE.mean                                2.667210e+02
## CI.mean.0.95                           5.279578e+02
## var                                    8.821368e+06
## std.dev                                2.970079e+03
## coef.var                               1.759928e-01
## skewness                               9.581275e-01
## skew.2SE                               2.203896e+00
## kurtosis                               9.047165e-01
## kurt.2SE                               1.048408e+00
## normtest.W                             9.386419e-01
## normtest.p                             2.588161e-05
##              Median_energy_consumption_kWh_per_meter
## nbr.val                                  1.240000e+02
## nbr.null                                 0.000000e+00
## nbr.na                                   0.000000e+00
## min                                      1.011404e+04
## max                                      2.420230e+04
## range                                    1.408826e+04
## sum                                      1.910269e+06
## median                                   1.508038e+04
## mean                                     1.540539e+04
## SE.mean                                  2.561895e+02
## CI.mean.0.95                             5.071114e+02
## var                                      8.138500e+06
## std.dev                                  2.852806e+03
## coef.var                                 1.851823e-01
## skewness                                 5.943803e-01
## skew.2SE                                 1.367200e+00
## kurtosis                                 3.326591e-01
## kurt.2SE                                 3.854935e-01
```

```
## normtest.W                                   9.732363e-01
## normtest.p                                   1.442882e-02
```

Stat.desc() Results

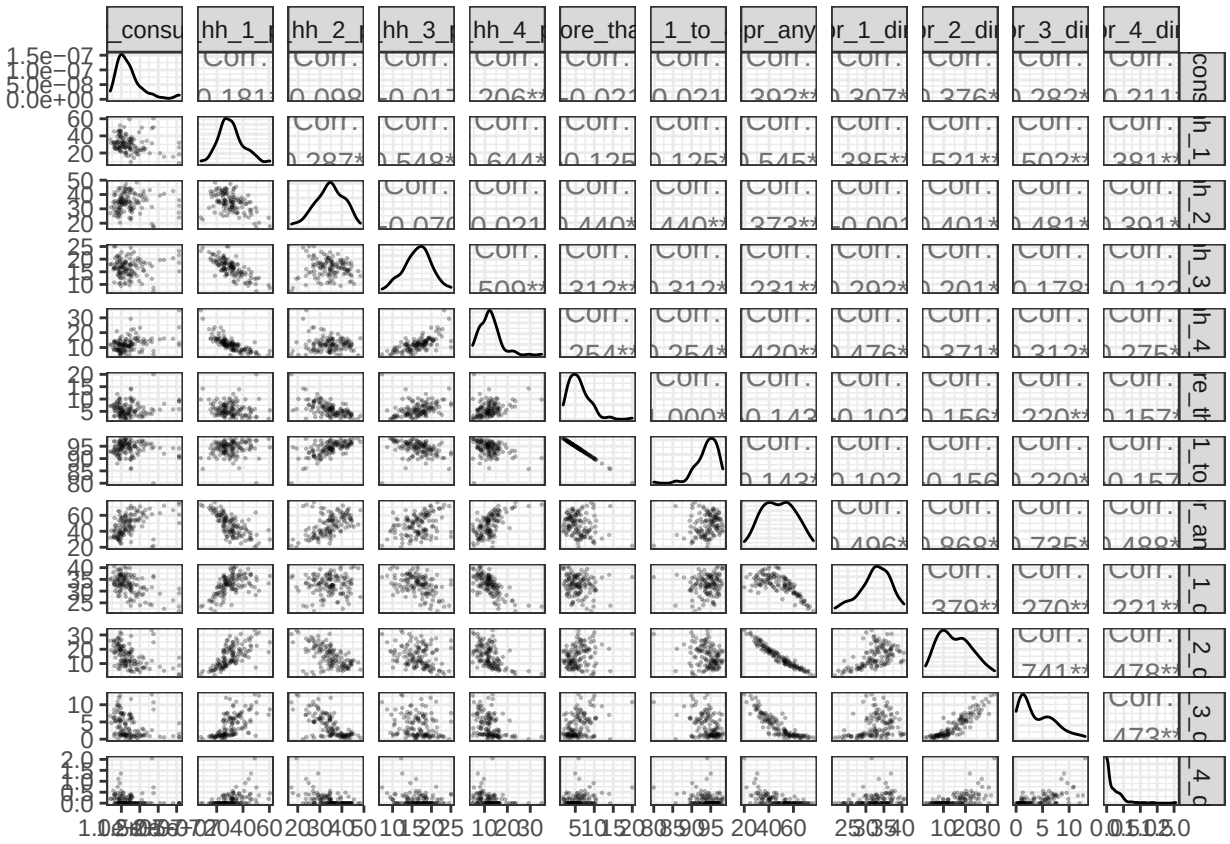
From the resulting table after applying the stat.dec() function, the following variables with a significant p-value (<0.01) were taken for further analysis – households with 4 people(%), households with more than 4 people(%), households with 1-4 people (%); household deprivation values 2-4 (%); the total energy consumption and the mean energy consumption. However, due to all but two variables – household deprivation values 2 and 3 (%), having skewness, the stat.desc() function was later applied to these variables after inverse hyperbolic sine transformations.

Pairs plot

Two pairs plot for Total energy consumption v variables and Mean energy consumption v variables were plotted to get an idea about the correlation amongst the variables, while also showing the skewness of data. (De Sabbata, 2022a)

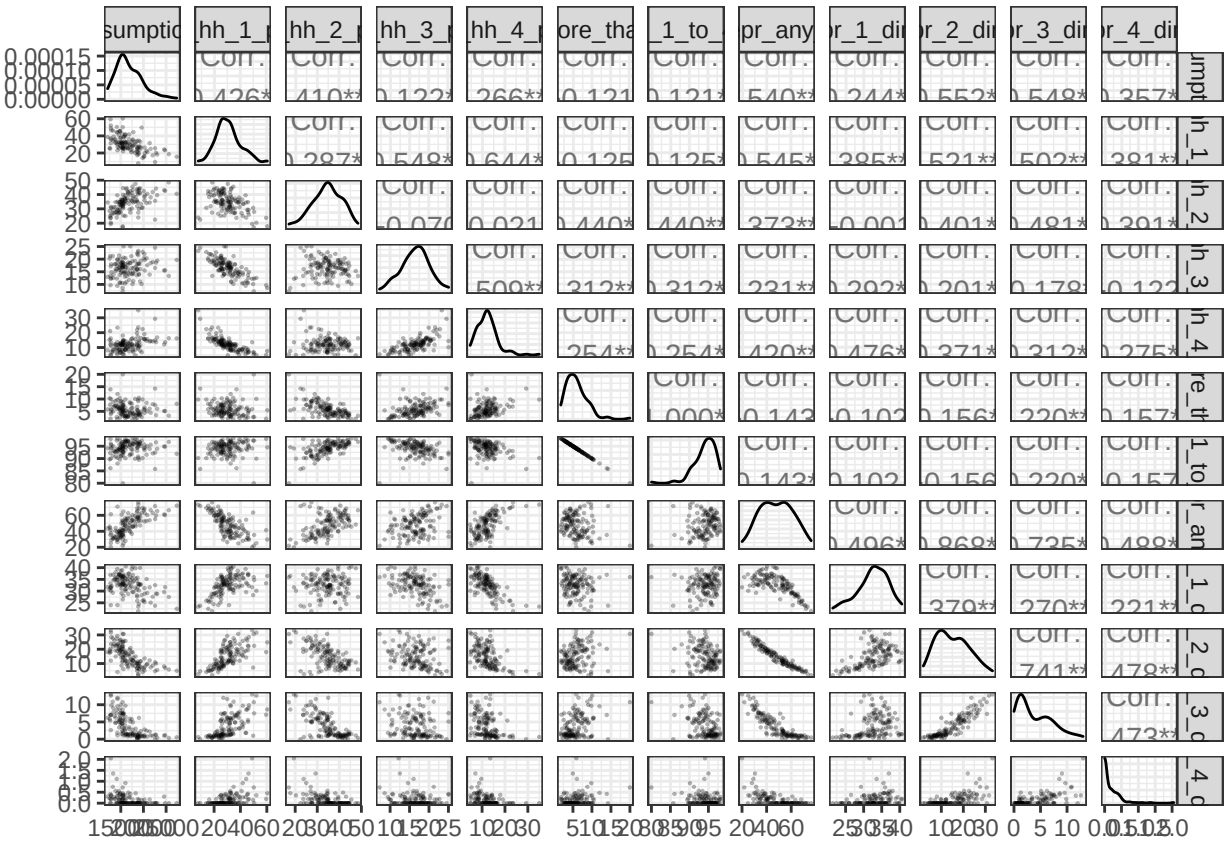
```
# pairs plot of Total_energy_consumption v the variables

energy_census_normalised %>%
  select(
    Total_energy_consumption_kWh,
    ppl_hh_1_perc : ppl_hh_4_perc,
    ppl_hh_more_than_4_perc, ppl_hh_1_to_4_perc,
    hh_not_depr_any_dim_perc : hh_depr_4_dim_perc,
  ) %>%
  ggpairs(
    upper = list(continuous =
        wrap(ggally_cor, method = "kendall")),
    lower = list(continuous =
        wrap("points", alpha = 0.3, size=0.1))
  ) +
  theme_bw()
```



```
# pairs plot of Mean_energy_consumption v the variables

energy_census_normalised %>%
  select(
    Mean_energy_consumption_kWh_per_meter,
    ppl_hh_1_perc : ppl_hh_4_perc,
    ppl_hh_more_than_4_perc, ppl_hh_1_to_4_perc,
    hh_not_depr_any_dim_perc : hh_depr_4_dim_perc,
  ) %>%
   ggpairs(
    upper = list(continuous =
        wrap(ggally_cor, method = "kendall")),
    lower = list(continuous =
        wrap("points", alpha = 0.3, size=0.1))
  ) +
  theme_bw()
```



Data Transformation

A new variable "transformed_energy_census" was created after applying inverse hyperbolic transformation on the variables (due to presence of null values for household deprivation values 3 and 4 (%)) and changing the variable names appropriately.

```
# transforming data to
# unskew desired variables

energy_census_transform <-
  energy_census_normalised %>%
  select(
    Total_energy_consumption_kWh, Mean_energy_consumption_kWh_per_meter,
    ppl_hh_1_to_4_perc, ppl_hh_4_perc, ppl_hh_more_than_4_perc,
    hh_depr_2_dim_perc, hh_depr_3_dim_perc, hh_depr_4_dim_perc
  )
# inverse hyperbolic sine transformation
transformed_energy_census <-
  energy_census_transform %>%
  mutate(
    ihs_tot_energy = asinh(Total_energy_consumption_kWh),
    ihs_mean_energy = asinh(Mean_energy_consumption_kWh_per_meter),
    ihs_hh_1_to_4_perc = asinh(ppl_hh_1_to_4_perc),
    ihs_hh_4_perc = asinh(ppl_hh_4_perc),
    ihs_hh_more_than_4_perc = asinh(ppl_hh_more_than_4_perc),
    ihs_depr_2_dim = asinh(hh_depr_2_dim_perc),
    ihs_depr_3_dim = asinh(hh_depr_3_dim_perc),
```

```
    ihs_depr_4_dim = asinh(hh_depr_4_dim_perc)
  )
```

```
# statdesc for transformed data

transformed_energy_census_statdesc <-
  transformed_energy_census %>%
  select(
    ihs_tot_energy, ihs_mean_energy,
    ihs_hh_1_to_4_perc, ihs_hh_4_perc, ihs_hh_more_than_4_perc,
    ihs_depr_2_dim, ihs_depr_3_dim, ihs_depr_4_dim
  ) %>%
  stat.desc(
    norm = TRUE
  )

transformed_energy_census_statdesc
```

```
##              ihs_tot_energy ihs_mean_energy ihs_hh_1_to_4_perc ihs_hh_4_perc
## nbr.val        1.240000e+02    124.00000000       1.240000e+02  124.00000000
## nbr.null       0.000000e+00      0.00000000       0.000000e+00    0.00000000
## nbr.na         0.000000e+00      0.00000000       0.000000e+00    0.00000000
## min            1.645309e+01     10.09370276       5.075213e+00    2.27067451
## max            1.775637e+01     10.90957834       5.280116e+00    4.24888303
## range          1.303277e+00      0.81587558       2.049029e-01    1.97820852
## sum            2.104319e+03   1291.14353307       6.493730e+02  388.95342964
## median         1.692085e+01     10.37651329       5.241642e+00    3.16968117
## mean           1.697031e+01     10.41244785       5.236879e+00    3.13672121
## SE.mean        2.486690e-02      0.01506523       2.749915e-03    0.03251501
## CI.mean.0.95   4.922251e-02      0.02982070       5.443289e-03    0.06436146
## var            7.667699e-02      0.02814319       9.376923e-04    0.13109598
## std.dev        2.769061e-01      0.16775931       3.062176e-02    0.36207179
## coef.var       1.631709e-02      0.01611142       5.847330e-03    0.11543002
## skewness       9.218298e-01      0.50666977      -1.689042e+00   -0.05161288
## skew.2SE       2.120403e+00      1.16544761      -3.885153e+00   -0.11872055
## kurtosis       7.036480e-01     -0.06288726       5.429483e+00    0.35895576
## kurt.2SE       8.154047e-01     -0.07287531       6.291819e+00    0.41596681
## normtest.W     9.384183e-01      0.97486080       8.852912e-01    0.98131706
## normtest.p     2.499406e-05      0.02043277       2.493991e-08    0.08398908
##              ihs_hh_more_than_4_perc ihs_depr_2_dim ihs_depr_3_dim
## nbr.val                  124.00000000   1.240000e+02   1.240000e+02
## nbr.null                   0.00000000   0.000000e+00   1.000000e+00
## nbr.na                     0.00000000   0.000000e+00   0.000000e+00
## min                        1.35346702   1.363595e+00   0.000000e+00
## max                        3.68950387   4.196999e+00   3.261928e+00
## range                      2.33603685   2.833404e+00   3.261928e+00
## sum                      295.26946466   4.066294e+02   2.177785e+02
## median                     2.40829994   3.361147e+00   1.750859e+00
## mean                       2.38120536   3.279270e+00   1.756278e+00
## SE.mean                    0.04012398   5.000157e-02   7.743119e-02
## CI.mean.0.95               0.07942296   9.897504e-02   1.532703e-01
## var                        0.19963177   3.100194e-01   7.434530e-01
## std.dev                    0.44680171   5.567939e-01   8.622372e-01
## coef.var                   0.18763678   1.697920e-01   4.909456e-01
```

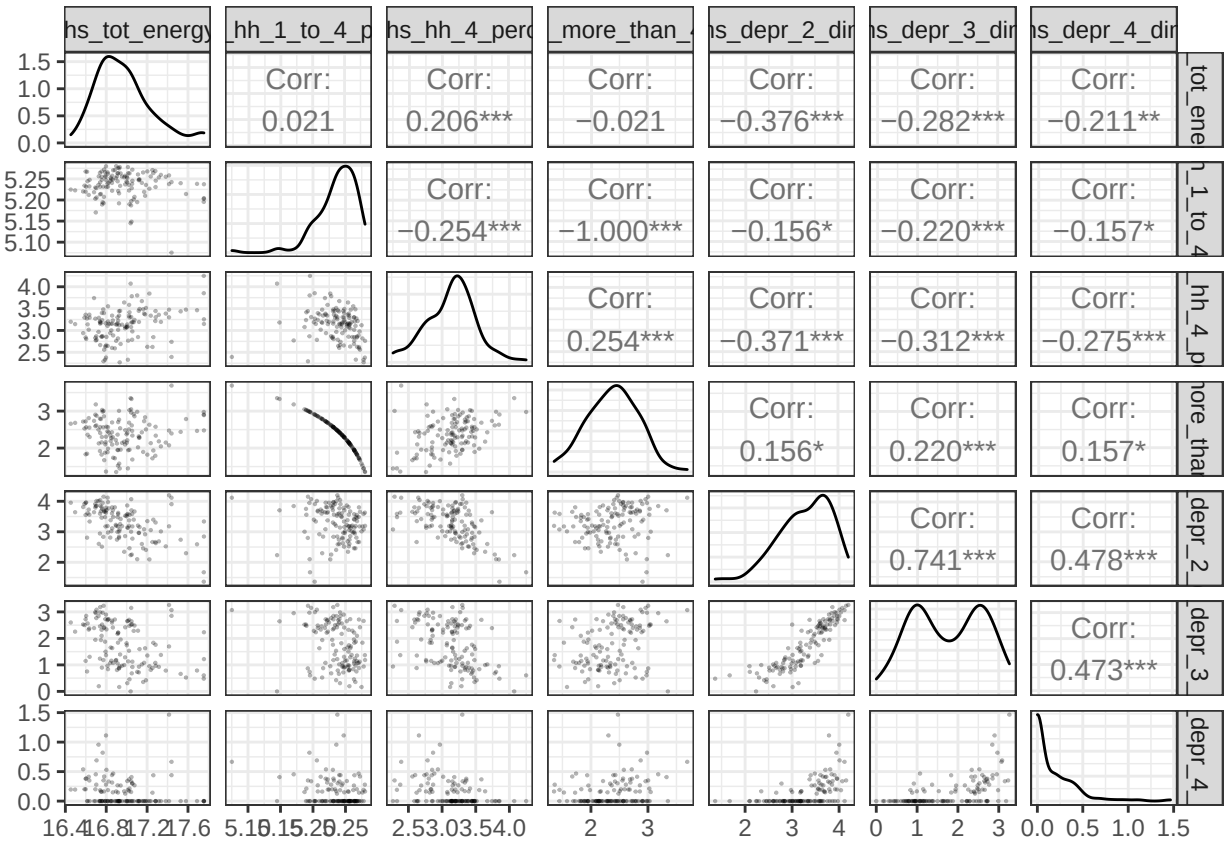
```
## skewness                -0.01727019  -6.928316e-01  -3.977287e-02
## skew.2SE                -0.03972509  -1.593659e+00  -9.148601e-02
## kurtosis                -0.20046599   2.334930e-01  -1.312364e+00
## kurt.2SE                -0.23230495   2.705775e-01  -1.520800e+00
## normtest.W               0.99395771   9.598847e-01   9.440584e-01
## normtest.p               0.87590365   9.882513e-04   6.149258e-05
##              ihs_depr_4_dim
## nbr.val        1.240000e+02
## nbr.null       7.200000e+01
## nbr.na         0.000000e+00
## min            0.000000e+00
## max            1.465679e+00
## range          1.465679e+00
## sum            1.883080e+01
## median         0.000000e+00
## mean           1.518613e-01
## SE.mean        2.200872e-02
## CI.mean.0.95   4.356491e-02
## var            6.006358e-02
## std.dev        2.450787e-01
## coef.var       1.613833e+00
## skewness       2.410256e+00
## skew.2SE       5.544099e+00
## kurtosis       7.633182e+00
## kurt.2SE       8.845520e+00
## normtest.W     6.711355e-01
## normtest.p     2.735198e-15
```

Stat.desc() for skewness

All but one variable – household deprivation value 4 inverse hyperbolic transformed (%) now have skewness < 1.0 and therefore, can be used in the following pair plots to obtain correlation amongst the variables.

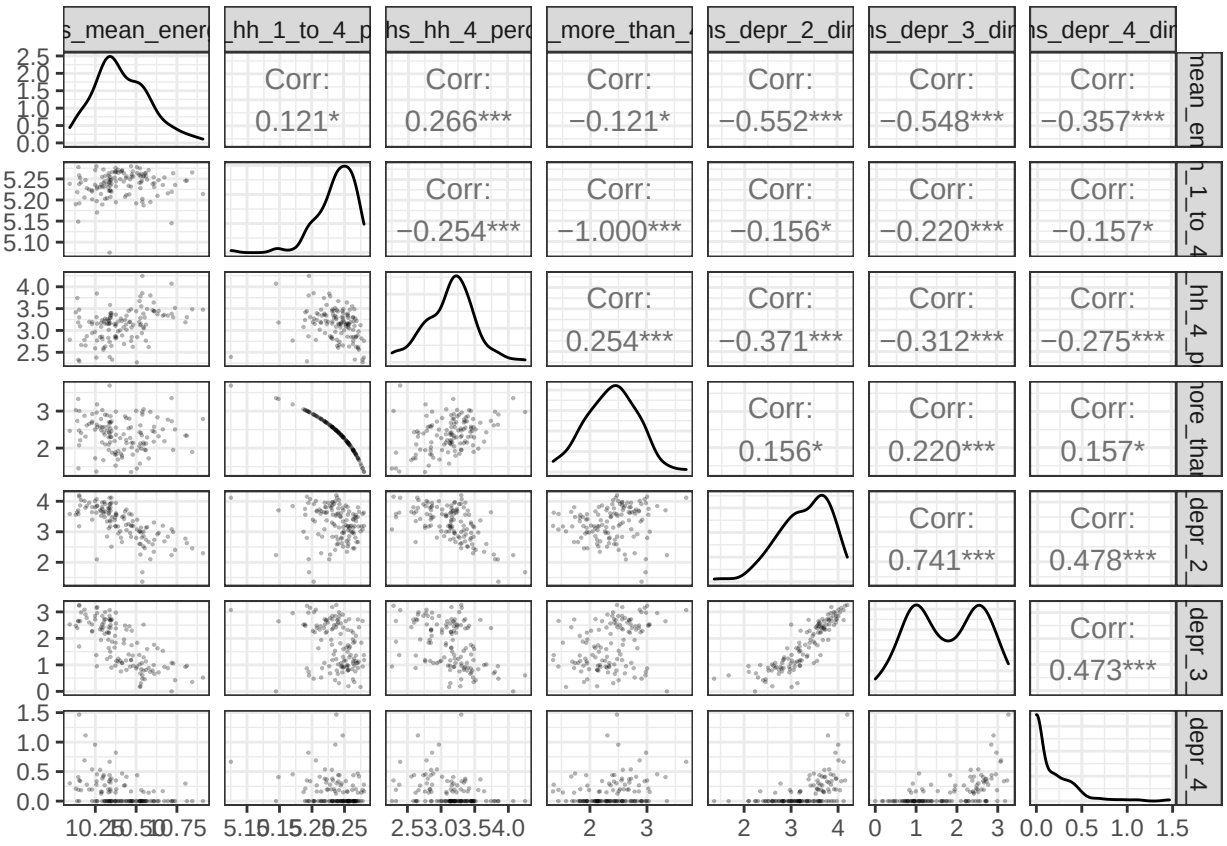
```
# pairs plot of Total energy consumption
# vs the variables

transformed_energy_census %>%
  select(
    ihs_tot_energy,
    ihs_hh_1_to_4_perc, ihs_hh_4_perc, ihs_hh_more_than_4_perc,
    ihs_depr_2_dim, ihs_depr_3_dim, ihs_depr_4_dim
  ) %>%
  ggpairs(
    upper = list(continuous =
        wrap(ggally_cor, method = "kendall")),
    lower = list(continuous =
        wrap("points", alpha = 0.3, size=0.1))
  ) +
  theme_bw()
```




```
# pairs plot of Mean energy consumption
# vs the variables

transformed_energy_census %>%
  select(
    ihs_mean_energy,
    ihs_hh_1_to_4_perc, ihs_hh_4_perc, ihs_hh_more_than_4_perc,
    ihs_depr_2_dim, ihs_depr_3_dim, ihs_depr_4_dim
  ) %>%
  ggpairs(
    upper = list(continuous =
        wrap(ggally_cor, method = "kendall")),
    lower = list(continuous =
        wrap("points", alpha = 0.3, size=0.1))
  ) +
  theme_bw()
```



Pair plots Results

From the pair plots, it is evident that – 1. Total energy consumption is significantly correlated to households with 4 people, and household deprivation values 2-4. All but one variable – households with 4 people, are negatively correlated with the total energy consumption. 2. Mean energy consumption is significantly correlated to households with 4 or more people and deprivation values 2-4. All but one variable – households with 4 people, are negatively correlated with the mean energy consumption.

Model 1 – Mean energy consumption v variables

Model created using Mean energy consumption variable and the households with 4 or more people and deprivation values 2-3 variables. Household deprivation value 4 was removed since it was not significant, reduced the adjusted R2 value for the model and was not linearly correlated with the mean energy consumption variable in the pairs plot. All the remaining variables, however, were linearly correlated with the Mean energy consumption variable. The second model – where the dependent variable was Total energy consumption and the independent variables chosen from above pairs plot results, was discarded since the adjusted R2 value was 0.30, which is less than 0.4. Moreover, it failed the Shapiro-Wilk test for Normality of standard residuals, so it would not have been a robust model in any case.

```
# creating a model with
# dependent variable = Mean energy consumption
# independent variables = household size 4 and above
# and household deprivation 2 and 3

energy_census_model_1 <-
  transformed_energy_census %$%
  lm(
```

```
    ihs_mean_energy ~
      ihs_hh_4_perc +
      ihs_hh_more_than_4_perc +
      ihs_depr_2_dim + ihs_depr_3_dim
  )

energy_census_model_1 %>%
  summary()
```

```
##
## Call:
## lm(formula = ihs_mean_energy ~ ihs_hh_4_perc + ihs_hh_more_than_4_perc +
##     ihs_depr_2_dim + ihs_depr_3_dim)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.312876 -0.073991 -0.005444  0.065118  0.270034
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              10.77806    0.18764  57.440  < 2e-16 ***
## ihs_hh_4_perc            -0.03474    0.04182  -0.831   0.4078
## ihs_hh_more_than_4_perc   0.06273    0.03043   2.061   0.0414 *
## ihs_depr_2_dim           -0.05419    0.04165  -1.301   0.1958
## ihs_depr_3_dim           -0.13000    0.02671  -4.866 3.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1118 on 119 degrees of freedom
## Multiple R-squared:  0.5705, Adjusted R-squared:  0.556
## F-statistic: 39.51 on 4 and 119 DF,  p-value: < 2.2e-16
```

Model 1 – Results

The model computed above is fit ($F(4,119) = 39.51$, $p < 0.01$), indicating that the independent variables chosen are 55.6% responsible for the mean energy consumption for the household.

Calculating the confidence intervals

Small intervals (essential for a good model)

```
# Confident intervals

energy_census_model_1 %>%
  confint()
```

```
##                                2.5 %      97.5 %
## (Intercept)              10.406516593 11.14960623
## ihs_hh_4_perc            -0.117539057  0.04806027
## ihs_hh_more_than_4_perc   0.002476322  0.12298312
## ihs_depr_2_dim           -0.136669428  0.02829108
## ihs_depr_3_dim           -0.182896595 -0.07710457
```

finding outliers and influential cases

energy_output <- transformed_energy_census %>% filter(ihs_mean_energy | ihs_hh_4_perc | ihs_hh_more_than_4_perc | ihs_depr_2_dim | ihs_depr_3_dim) mutate(model_stdres = energy_census_model_1 %>% rstandard(), model_cook_dist = energy_census_model_1 %>% cooks.distance())

energy_output %>% select(ihs_hh_4_perc, ihs_hh_more_than_4_perc, ihs_depr_2_dim, ihs_depr_3_dim, model_stdres, model_cook_dist) %>% filter(abs(model_stdres) > 2.58 | model_cook_dist > 1)

Checking for assumptions

Checking for Normality – Shapiro-Wilk test for standard residuals Checking for Homoscedasticity - Breusch-Pagan test for homoscedasticity of standard residuals Checking for Independence - Durbin-Watson test for the independence of residuals Checking for Multicollinearity - Checking the variance inflation factor (VIF) Standardised Coefficients – Indicate the amount of change in dependent variable per one standard deviation change in the independent variable

```
# testing first assumption

energy_census_model_1 %>%
  rstandard() %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.98974, p-value = 0.4865
```

```
# testing second assumption

 energy_census_model_1 %>%
   bptest()
```

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 16.332, df = 4, p-value = 0.002605
```

```
# testing third assumption

energy_census_model_1 %>%
  dwtest()
```

```
##
##  Durbin-Watson test
##
## data:  .
## DW = 1.6741, p-value = 0.02649
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# testing fourth assumption

library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
energy_census_model_1 %>%
  vif()
```

```
##           ihs_hh_4_perc ihs_hh_more_than_4_perc          ihs_depr_2_dim
##                2.256540                1.819655                5.295236
##          ihs_depr_3_dim
##                5.222709
```

```
# since the units of all variables
# are not same, standardised beta test is performed

lm.beta(energy_census_model_1)
```

```
##
## Call:
## lm(formula = ihs_mean_energy ~ ihs_hh_4_perc + ihs_hh_more_than_4_perc +
##     ihs_depr_2_dim + ihs_depr_3_dim)
##
## Standardized Coefficients::
##             (Intercept)           ihs_hh_4_perc ihs_hh_more_than_4_perc
##              0.00000000              -0.07497738               0.16707118
##          ihs_depr_2_dim          ihs_depr_3_dim
##             -0.17985411              -0.66816759
```

Results of tests for assumptions

Normality – The Shapiro Wilk test is not significant ($p > 0.01$). Therefore, standard residuals are normally distributed as confirmed by the Q-Q plot below.

Independence – Statistic = 1.67, which is close to 2 and between 1 and 3. Moreover, the Durbin- Watson test is not significant ($p > 0.01$)

Multicollinearity – Since the largest VIF is less than 10, there seems to be no multicollinearity

Homoscedasticity – The interpretation of studentized Breusch-Pagan test is disputed. While some sources indicate that a p-value > 0.005 results in heteroscedasticity of the standard residuals, (De Sabbata, 2022b) and p-value > 0.05 indicates heteroscedasticity, ((https://stats.stackexchange.com/users/133928/fornit), 2016) the others state that a p-value < 0.05 results in homoscedasticity. (Zach, 2020) Taking into account the latter assumption, the Breusch-Pagan test for the resulting model is not significant (p-value = 0.0026). Therefore, the standard residuals are homoscedastic.

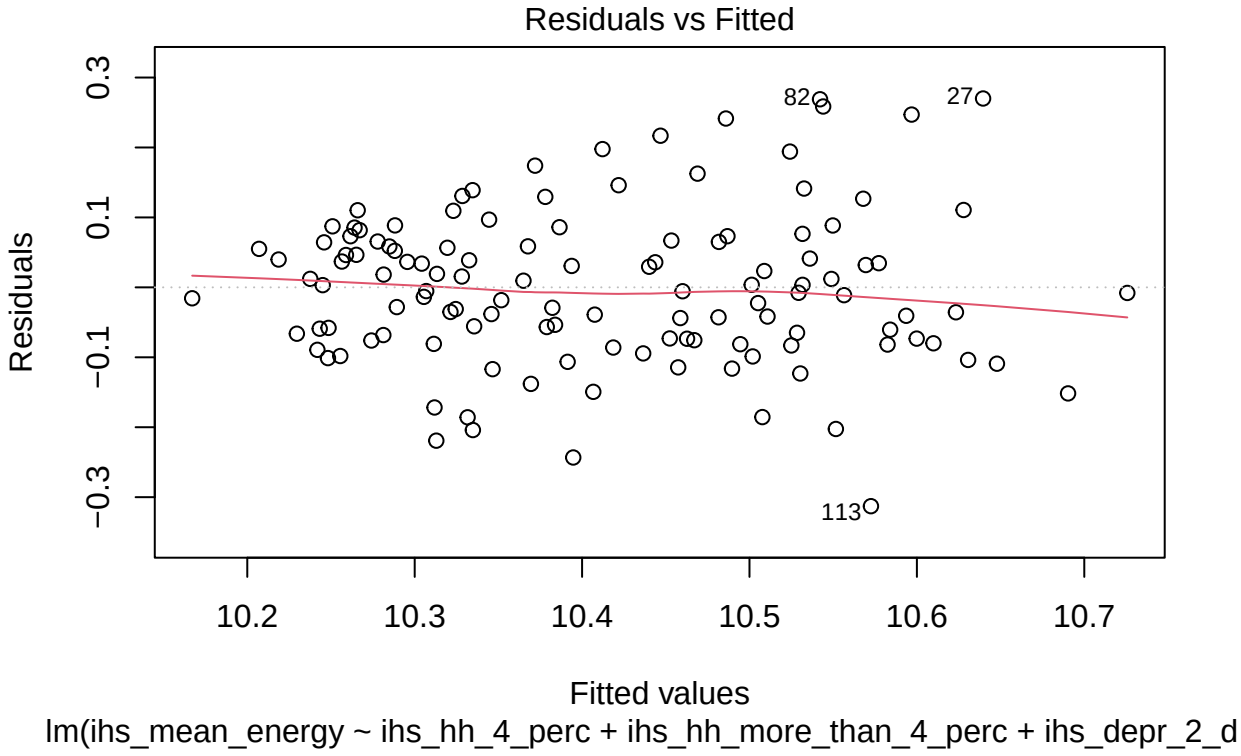
Standardised Coefficients – The results of the lm.beta() function indicate that the household deprivation value 3 was the most significant variable resulting in the change in the mean energy consumption of the household. The household with 4 people was the least significant variable that results in the change of the mean energy consumption variable, when the former changes. The positive and the negative values indicate the relationship of change among the variables. This test was conducted since the units of the variables

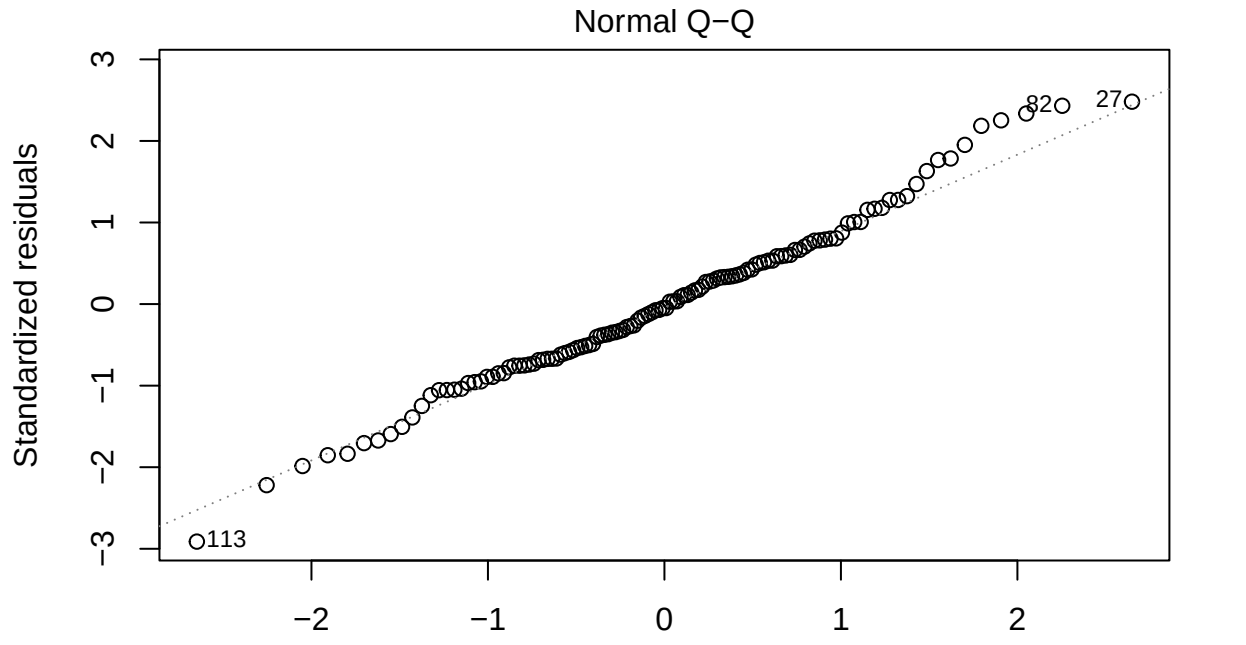
were not the same (instead of VIF being used for the interpretation) (kWh/meter for energy while the other variables are in counts or percentages) (De Sabbata, 2022a)

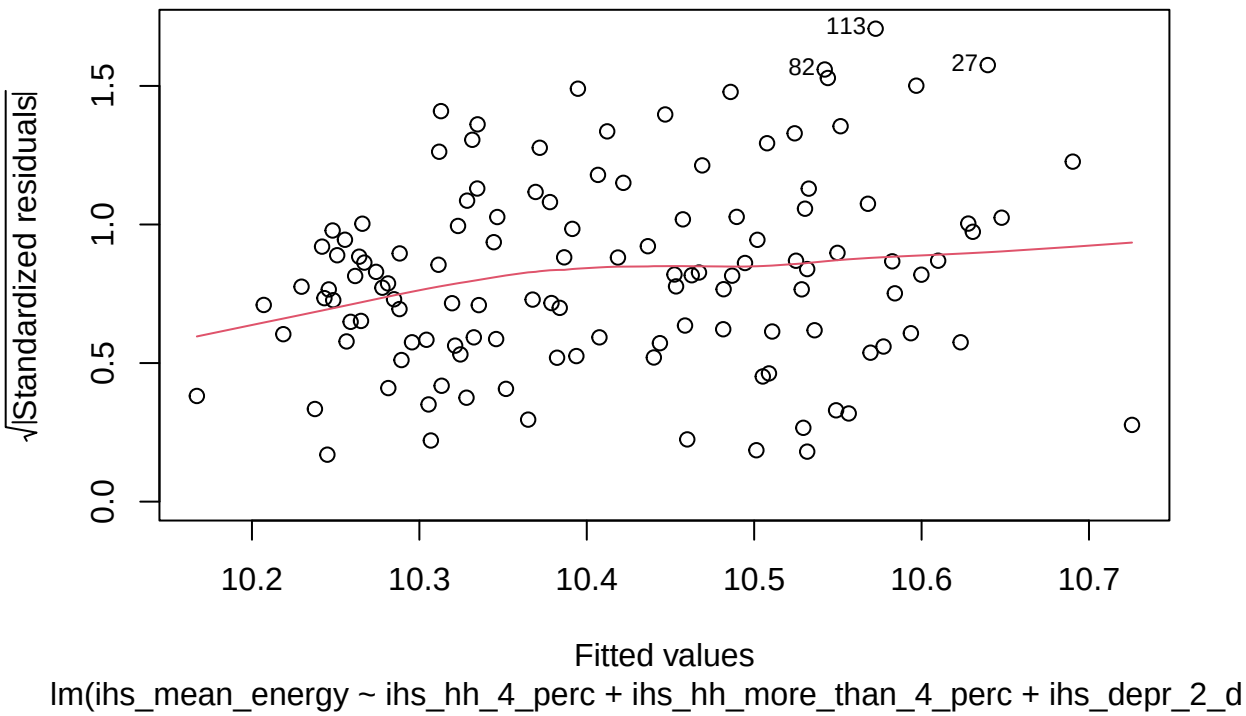
Reporting the model

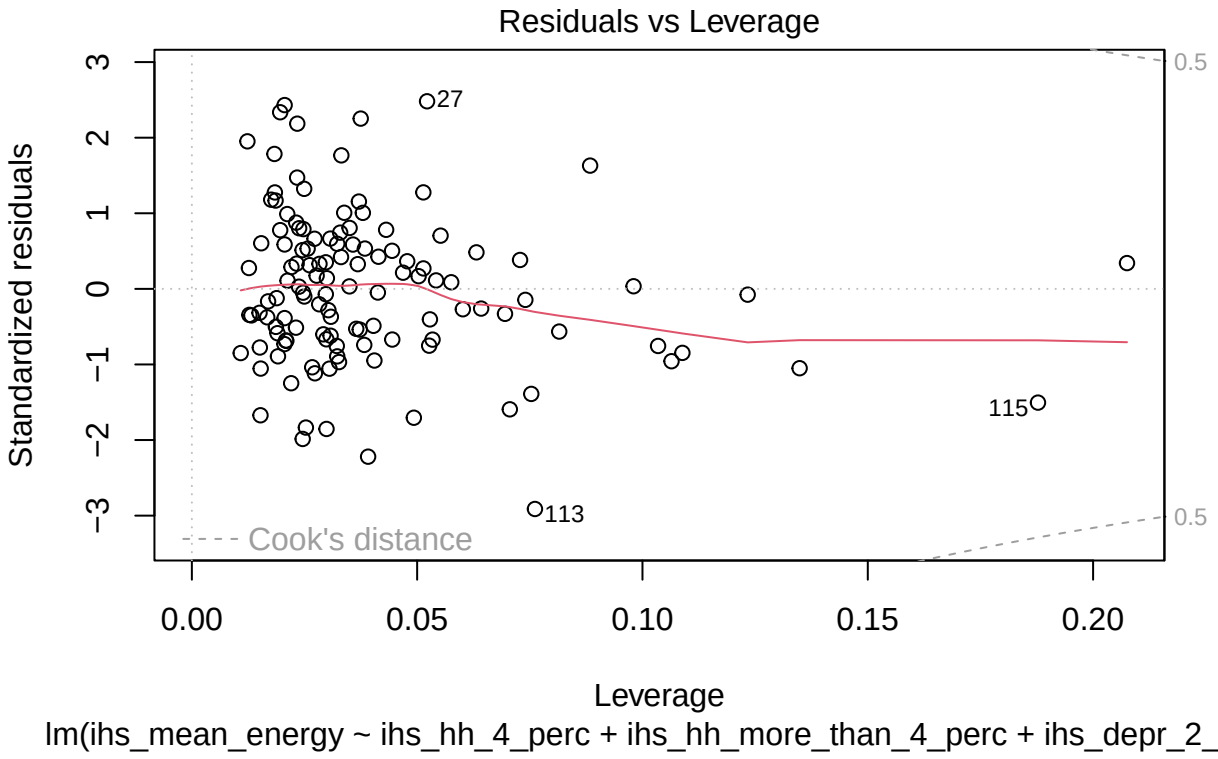
The model computed above is fit ($F(4,119) = 39.51$, $p < 0.01$), indicating that the independent variables chosen are 55.6% responsible for the mean energy consumption for the household. The model is robust. The residuals are normally distributed, there seems to be no multicollinearity, the residuals satisfy the independence assumption and the homoscedasticity assumption. (depends on interpretation)

```
energy_census_model_1 %>%
  plot()
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(ihs_mean_energy ~ ihs_hh_4_perc + ihs_hh_more_than_4_perc + ihs_depr_2_d
113
82
27


Scale−Location
√|Standardized residuals|
Fitted values
lm(ihs_mean_energy ~ ihs_hh_4_perc + ihs_hh_more_than_4_perc + ihs_depr_2_d
113
82
27



Final Model Equation:

Mean energy consumption = {10.78 + (-0.034) * Household with 4 people + (0.062) * Household with more than 4 people + (-0.054) * Household Deprivation value 2 + (-0.130) * Household Deprivation value 3} + error

Conclusion

As seen from the analysis and the model, the mean energy consumption in a household depends largely on the factor that in how many dimensions is the household deprived. If the household is deprived in 3 dimensions, the mean energy consumption goes down. Moreover, the mean energy consumption is also dependent on the number of people living in the household. If the household has 4 people, then the mean energy consumption will be lower when compared with households with more than 4 people, which seems to be an expected result. If the household is deprived in 2 dimensions, the effect on mean energy consumption is not as great with household deprived in 3 dimensions, but the factor is still significant. This also seems to be an expected result. Based on this analysis and the model, the local authorities can act likewise on the issue of increase in the cost of energy, affecting the overall cost of living for the people of Stockton-on-Tees.

References

1. (https://stats.stackexchange.com/users/133928/fornit) (2016) Interpretation of breusch-pagan test bptest() in R. Available at: https://stats.stackexchange.com/questions/239060/interpretation-of-breusch-pagan-test-bptest-in-r (Accessed: January 07, 2023).
2. De Sabbata, S. (2022a) Regression analysis. Available at: https://sdesabbata.github.io/r-for-geographic-data-science/regression-analysis.html (Accessed: January 07, 2023).
3. De Sabbata, S. (2022b) Simple regression . Available at: https://sdesabbata.github.io/r-for-geographic-data-science/slides/204-slides-regression.html#38 (Accessed: January 07, 2023).

4. Zach (2020) The breusch-pagan test: Definition & example. Available at: https://www.statology.org/breusch-pagan-test/ (Accessed: January 07, 2023).