Electricity Demand Regression Analysis in R

Oguzhan Gurbuz

#This is a multivariate regression analysis examining whether there is a relationship electricity demand (response variable) and GDP per capita in USD, urban population percentage of total population and precipitation in a country (predictor variables).

#I collected the data used in the model from different databases. The key points about the data are explained below:

#1) GDP per capita: Gross Domestic Product (GDP) per capita in USD. Data Source: Our World in Data

#2) Precipitation: Average precipitation of the country in depth (mm per year). Data Source: World Bank Climate Portal

#3) Electricity Gross Demand: The amount of the electricity demand in Terawatt Hours including the quantity of electricity that gets lost on the way and never makes it to the end user. It does not only include electricity, but also other areas of consumption including transport, heating and cooking. Data Source: United Nations

#4) Population: The population of the country in 2017.Data Source: Our World in Data

#Each value in the dataset represents the values from the year 2017.

```
#Let's load the necessary packages
library(ggplot2) #to create plots
library(dplyr) #for data manipulation
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(broom)
library(ggpubr) #to create correlation plots
library(readr)
library(psych) #to create correlation plots
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

#Let's load the dataset.

```
regression_data <- read_csv("Regression Analysis Data.csv")
```

```
## Rows: 139 Columns: 5
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (5): gdp_pct, Precipitation, Urban Population Percentage, elect_gross_de...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Checking the class of the data.
class(regression_data) #It's a data frame
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
#Let's look at the quality of the data and if there is any null value in the data frame
is.null(regression_data) # There's not a single "null" value which makes the data set more
explanatory.
```

```
## [1] FALSE
```

```
#Let's see a few rows and columns of the data.
head(regression_data)
```

gdp_... <dbl>	Precipitation <dbl>	Urban Population Percentage <dbl>	elect_gross_demand <dbl>	populatio <dbl
530	298.38	25	5976.30	3564342
4531	1113.15	59	7412.44	287936
4135	78.24	72	68202.00	4113654
2283	1057.08	65	10454.00	3020863
14613	632.53	92	151981.00	4405461
4042	453.27	63	6291.58	285192

6 rows

```
#Now, it's time to load the columns in the dataset as variables.
precipitation <- regression_data$Precipitation
gdp_pct <- regression_data$gdp_pct
gross_demand <- regression_data$elect_gross_demand
population <- regression_data$population
urban <- regression_data$`Urban Population Percentage`
```

#DATA MUTATION #To find the per capita electricity demand per capita, we need to use the 'mutate' function of the 'dplyr' package. Because the data 'gross electricity demand' shows the values in million KwHs or Gigawatt Hours, we can start by first mutating it to the values in KwH. Then, we need to get divided by the country's population. Finally, we're going create a new variable as 'electricity demand per capita or simply "demand_pct", showing the KWh per capita values.

```
regression_data <-
  regression_data %>% mutate(demand_pct = (gross_demand * 1000000) / regression_data$popul
ation)
demand_pct <- regression_data$demand_pct #Now, the new variable for our model is ready.
```

#Let's check the first values again. So we can see the newly created variable as "low_elec_per_capita".

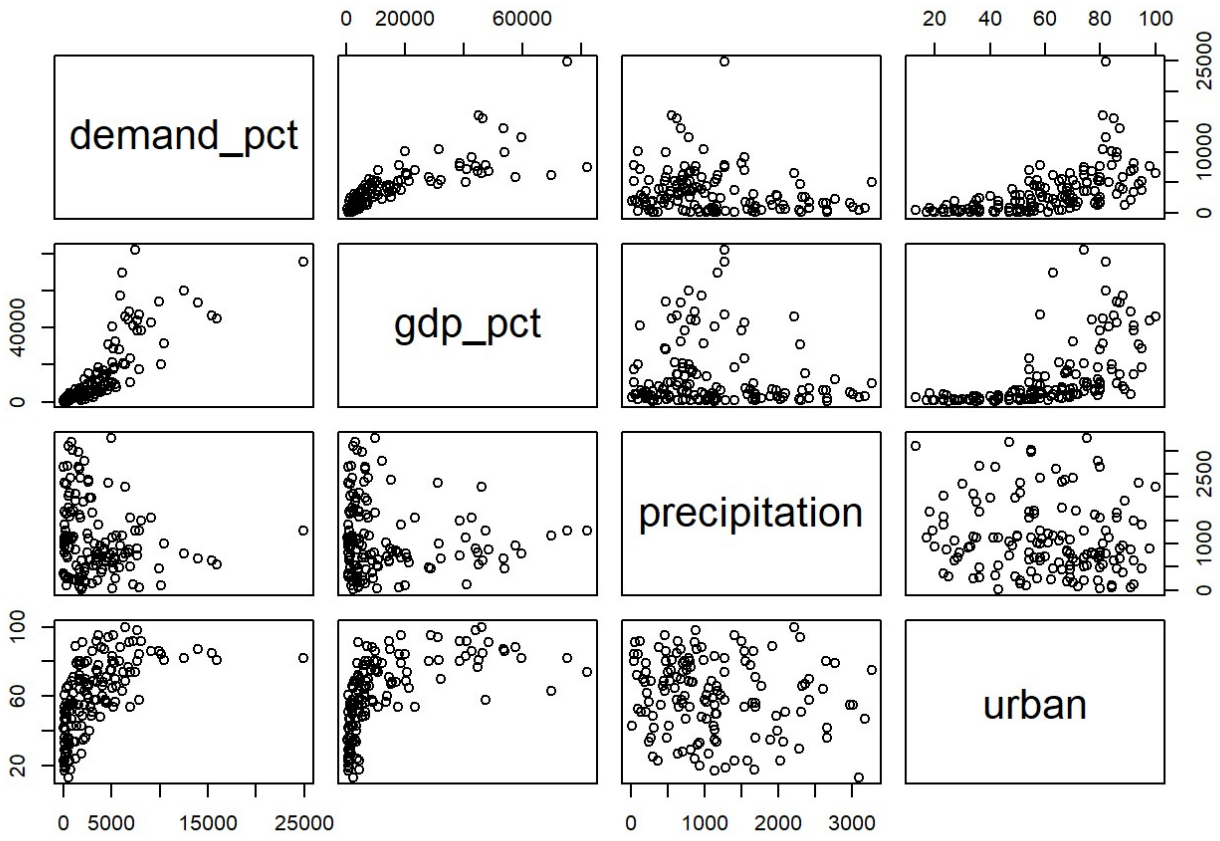
```
head(regression_data)
```

gdp_… <dbl>	Precipitation <dbl>	Urban Population Percentage <dbl>	elect_gross_demand <dbl>	populatio <db
530	298.38	25	5976.30	3564342
4531	1113.15	59	7412.44	287936
4135	78.24	72	68202.00	4113654
2283	1057.08	65	10454.00	3020863
14613	632.53	92	151981.00	4405461
4042	453.27	63	6291.58	285192

6 rows | 1-5 of 6 columns

#CHECKING THE CORRELATION AND MULTICOLLINEARITY

```
#Pair plots
pairs(demand_pct ~ gdp_pct + precipitation + urban, data = regression_data)
```




```
# Pair plots show a strong linearity between GDP per capita and electricity demand per cap
ita. We need further investigation to see the link between other variables.
```

#Let's check for the multicollineraity which can be problematic for our regression model. In other words, let's check the correlation level between the independent variables.

```
cor(gdp_pct, precipitation)
```

```
## [1] -0.08599934
```

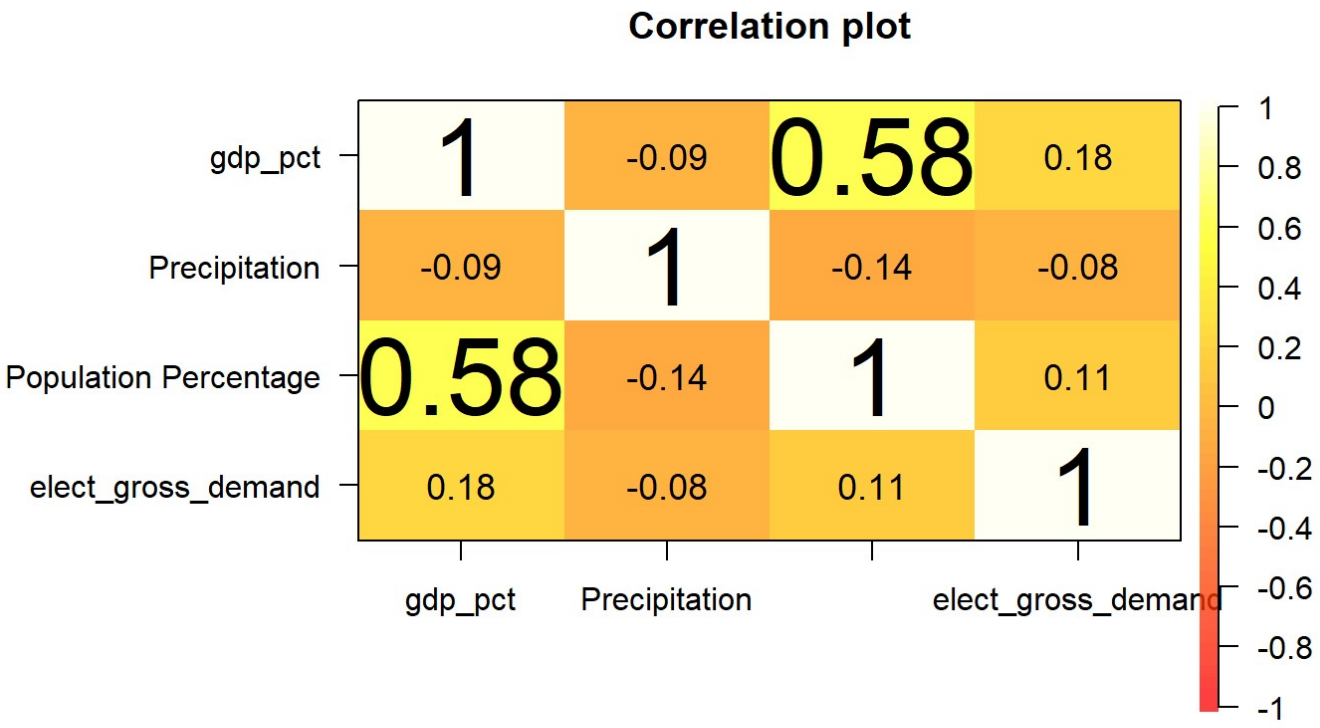
```
cor(precipitation, urban)
```

```
## [1] -0.1383724
```

```
cor(gdp_pct, urban)
```

```
## [1] 0.576139
```

```
#We can also see the multicollinearity and create another correlation plot
corPlot(regression_data[1:4],  gr = colorRampPalette(heat.colors(30)))
```




```
summary(demand_pct)
```

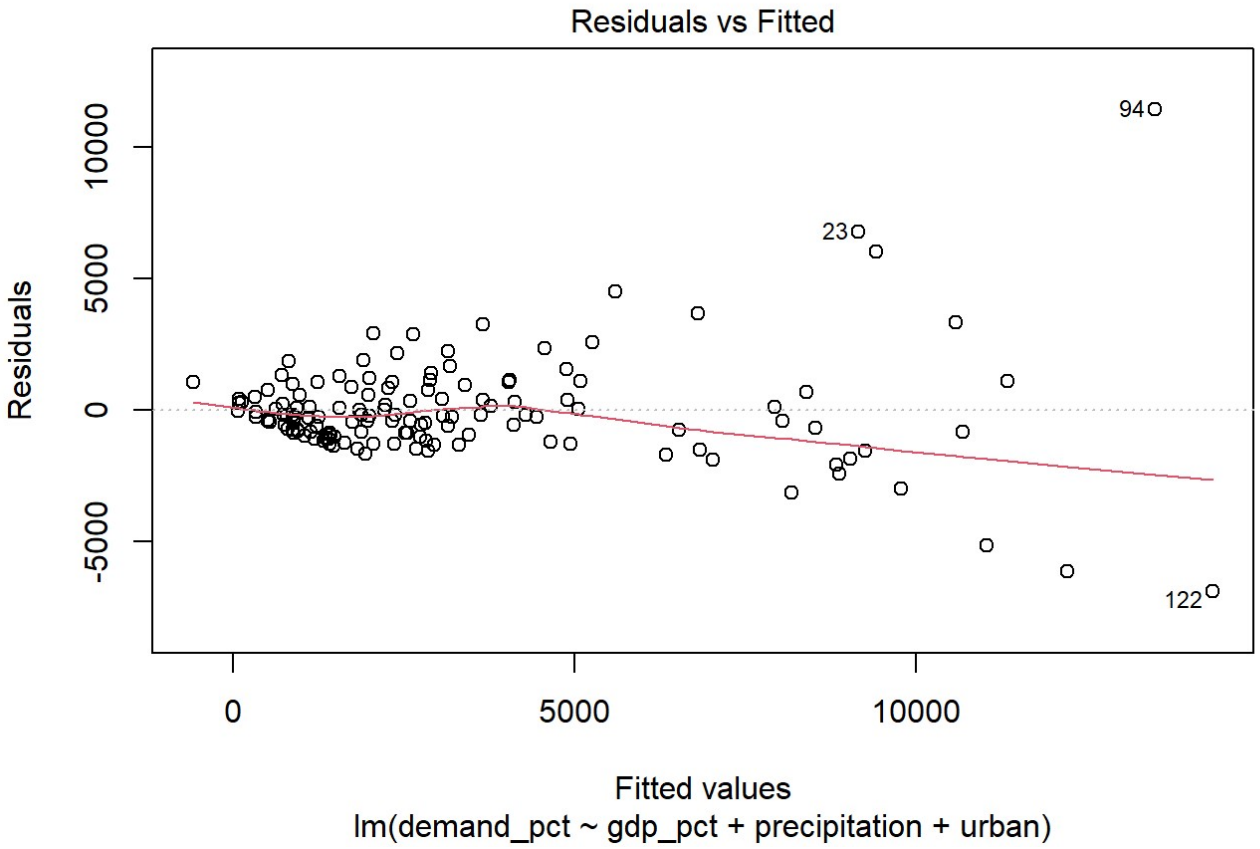
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    19.29   640.31  2212.48  3360.20  5008.19 24918.16
```

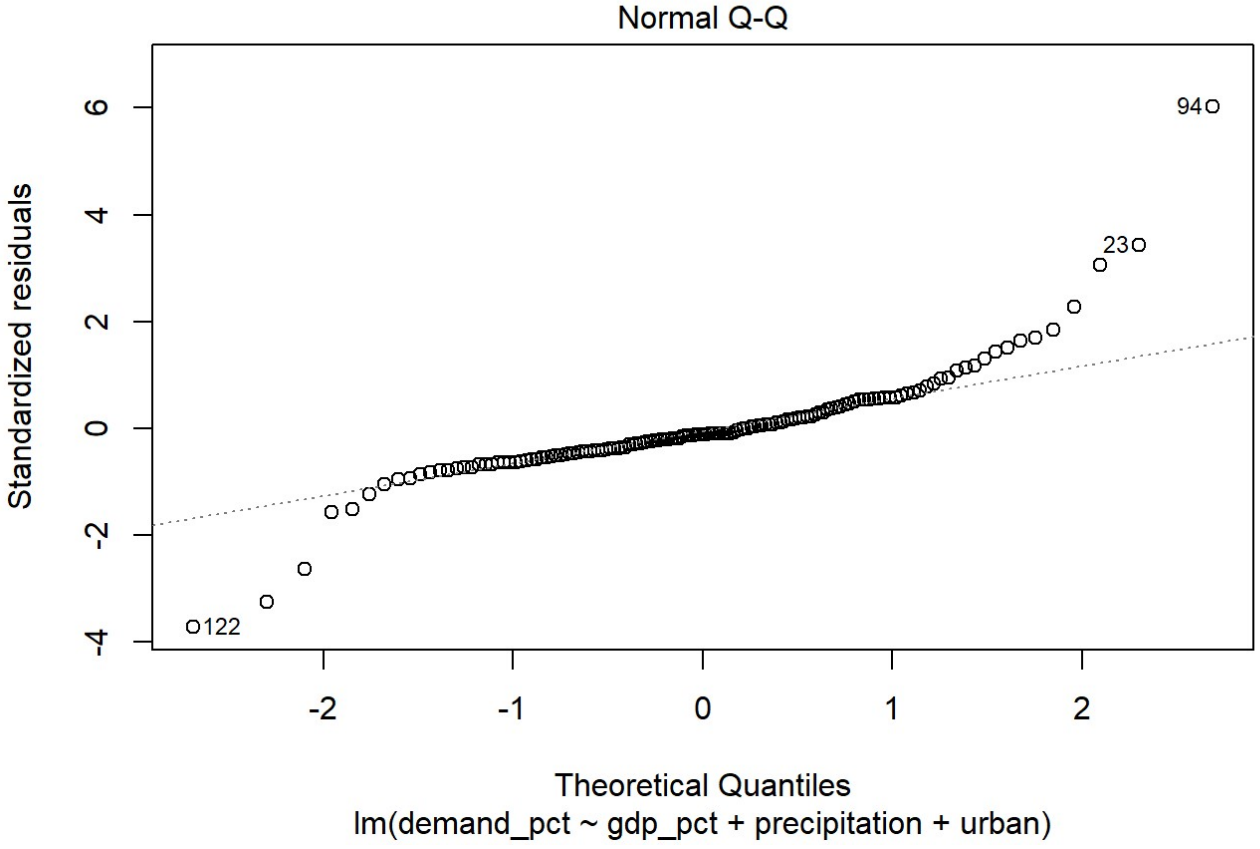
#Multivariate Regression Analysis

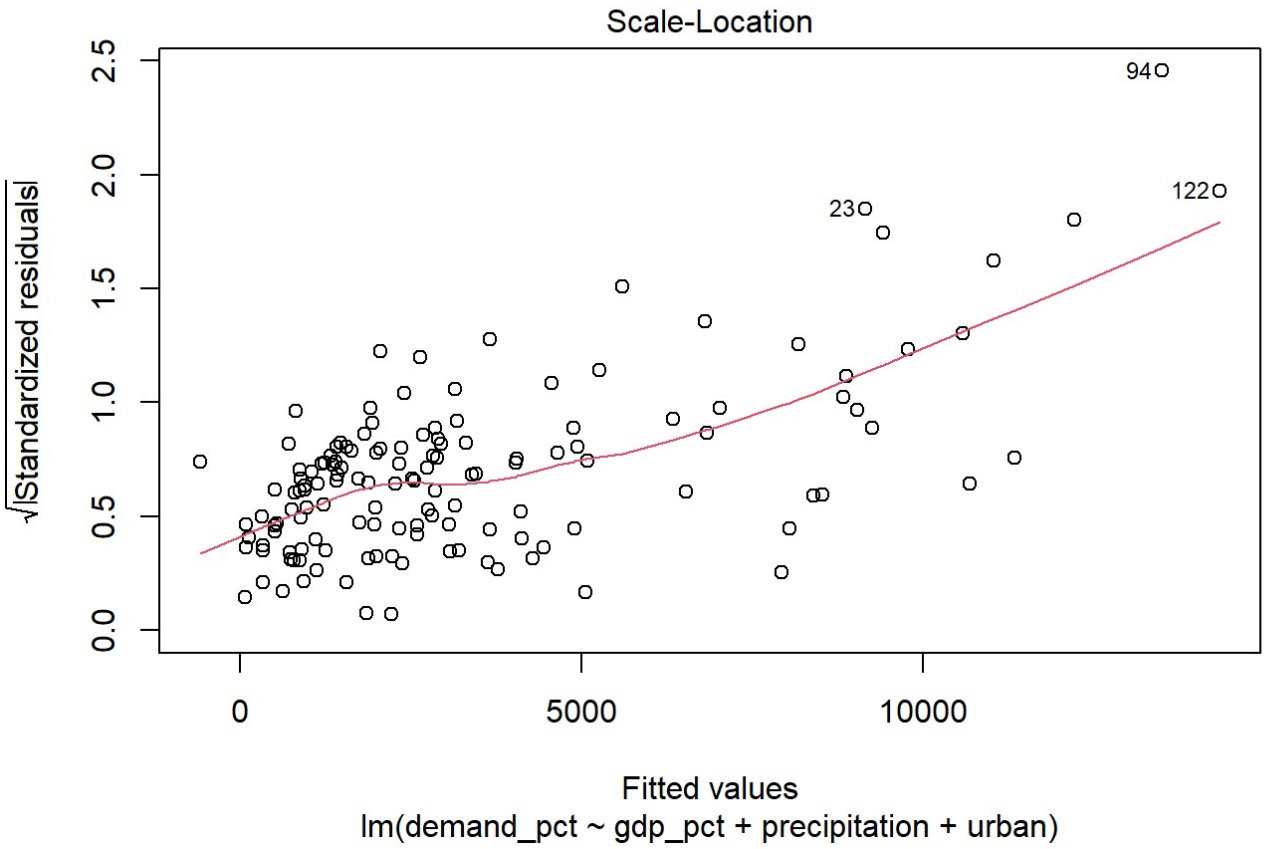
```
#This analysis examines whether there is a relationship between the electricity demand per
capita (as the response variable) and precipitation, GDP per capita and urbanization perce
ntage (as explanatory variables).

lin.model <- lm(demand_pct ~ gdp_pct + precipitation + urban, data = regression_data)

#Let's see the ANOVA table and some key statistics of the regression model.

summary(lin.model)
```

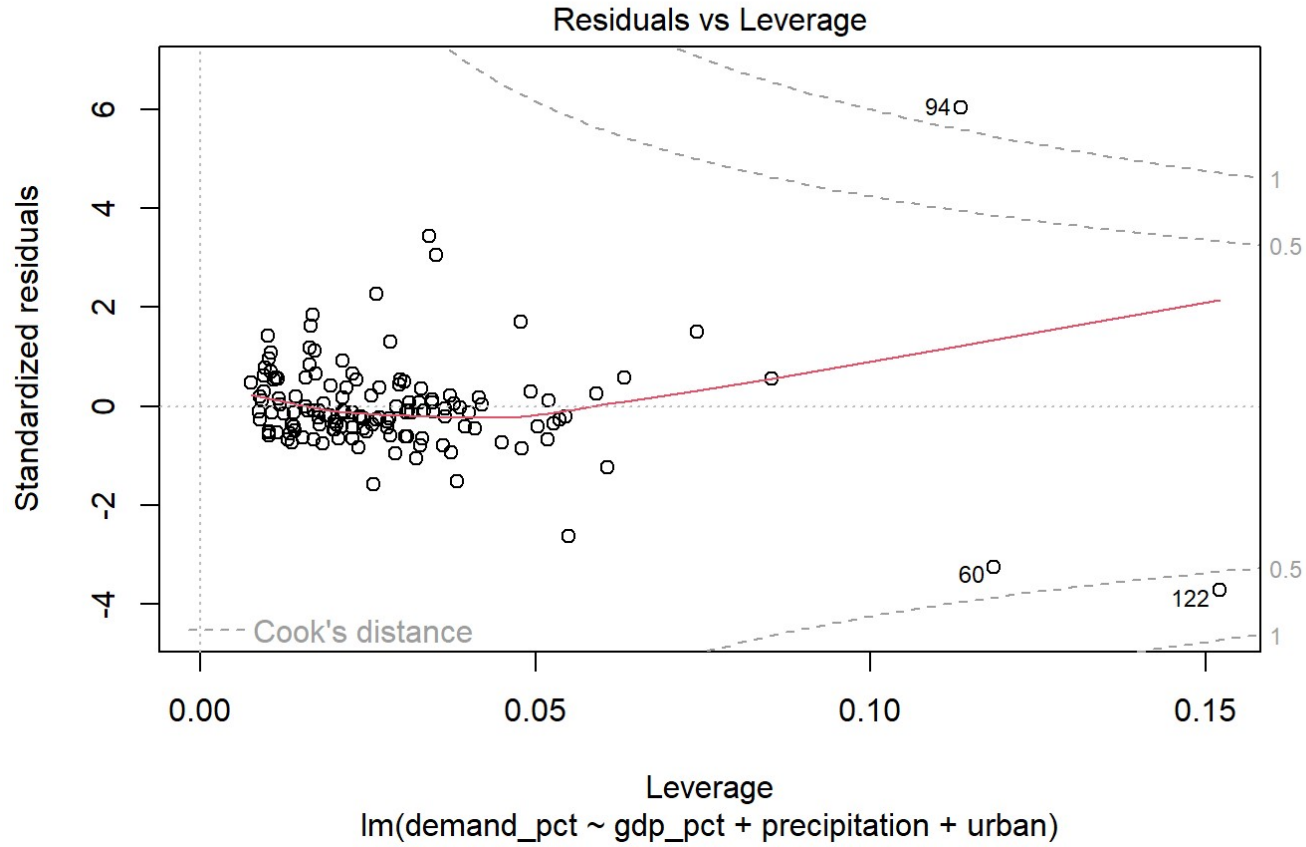
```
##
## Call:
## lm(formula = demand_pct ~ gdp_pct + precipitation + urban, data = regression_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6889.7  -904.0  -241.9   716.7 11416.7
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   408.41308  628.92199   0.649   0.5172
## gdp_pct         0.15522    0.01204  12.896   <2e-16 ***
## precipitation  -0.55065    0.22102  -2.491   0.0139 *
## urban          25.26479    9.89793   2.553   0.0118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2010 on 135 degrees of freedom
## Multiple R-squared:  0.7105, Adjusted R-squared:  0.704
## F-statistic: 110.4 on 3 and 135 DF,  p-value: < 2.2e-16
```

```
#Comments:
#i. Our model explains more than 70% of the data based on various R-squared and modified R
-squared values. This is not a great percentage, but it explains the majority of the data.
#ii. According to the p-value of GDP per capita, it is statistically significant at the 1%
level. In other words, the null hypothesis can be rejected with a 1% probability for the G
DP per capita in our model. On the other hand, the p-values for the average precipitation
and urbanization show that they are statistically significant at the 5% level. Lastly, the
p-value for the regression model is lower than 0.01 (2.2e-16).
#iii. The standard errors of the GDP per capita and precipitation do not seem to be high.
However, the error for the intercept appears to be slightly higher. Because the minimum va
lue for electricity demand per person is 19.29 and the 1st quantile is 640, our model may
not explain the pattern for the countries with the lowest electricity demand per-capita va
lues.
#iv. We can show our model as Electricity demand per person = 408.41 + -0.55 x precipitati
on + 0.15 x GDP per capita + 25.26 x urbanization rate + ϵ

#It would also be beneficial to build and examine the plots for our linear regression mode
l.

plot(lin.model)
```


Residuals vs Fitted
Residuals
10000
5000
0
-5000
94
23
122
0
5000
10000
Fitted values
lm(demand_pct ~ gdp_pct + precipitation + urban)


Normal Q-Q
Standardized residuals
6
4
2
0
-2
-4
94
23
122
-2
-1
0
1
2
Theoretical Quantiles
lm(demand_pct ~ gdp_pct + precipitation + urban)

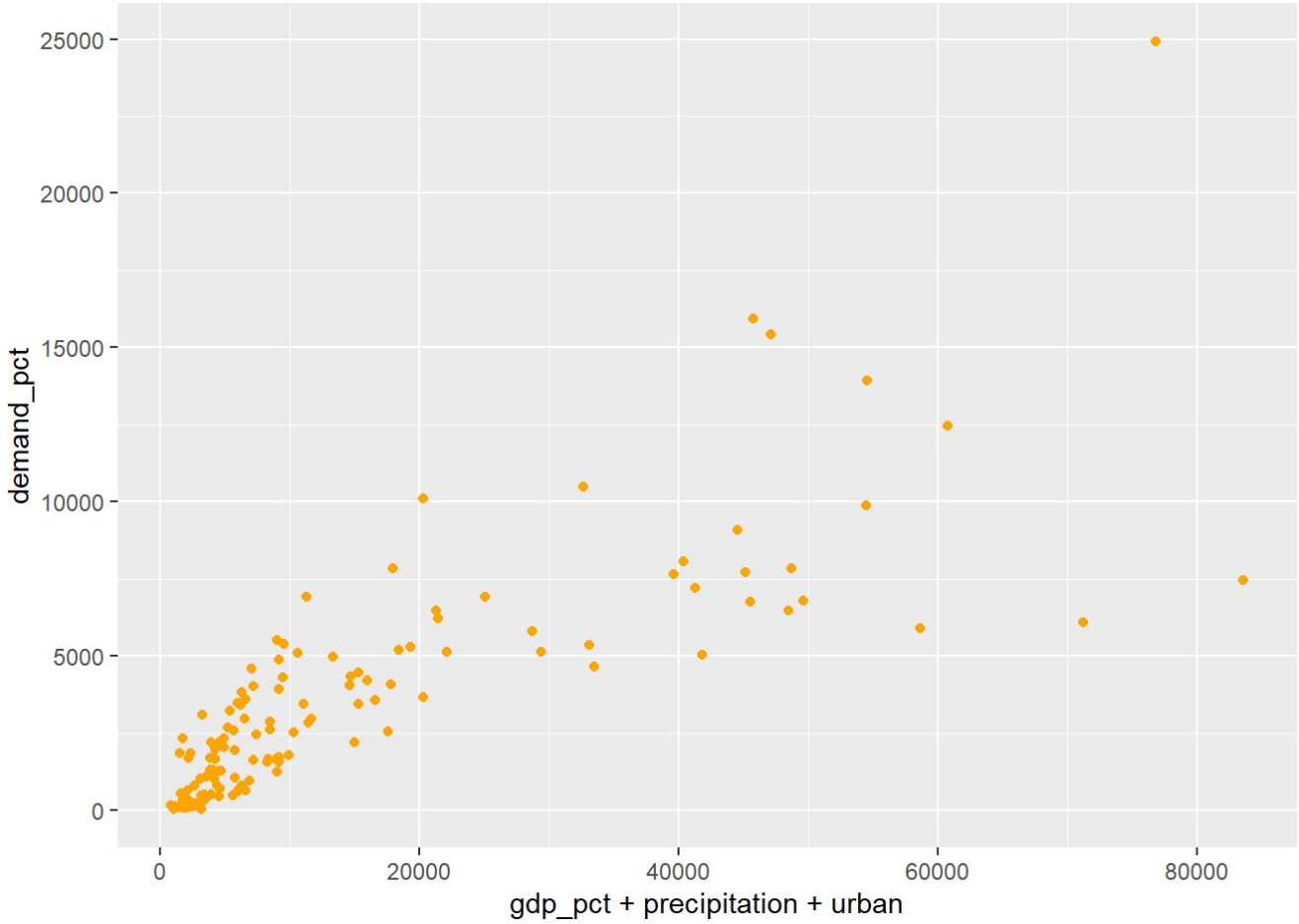

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
5000
10000
94
23
122
Fitted values
lm(demand_pct ~ gdp_pct + precipitation + urban)


Residuals vs Leverage
Standardized residuals
-4
-2
0
2
4
6
0.00
0.05
0.10
0.15
94
60
122
1
0.5
0.5
1
Cook's distance
Leverage
lm(demand_pct ~ gdp_pct + precipitation + urban)

```
#Based on the "residuals vs fitted" graph, we see that linearity seems to hold in the firs
t part of the red line, as the red line is close to the dashed line. We can also see the h
eteroskedasticity: when we move to the right on the fitted values line (x-axis), the sprea
d of the residuals seems to be increasing. The observations 23, 94, and 122 may be outlier
s, having large residual values.
#To conclude from all graphs, we can say the points 23, 94 and 122 may be the outliers in
our model.
```

#Creating a scatter plot

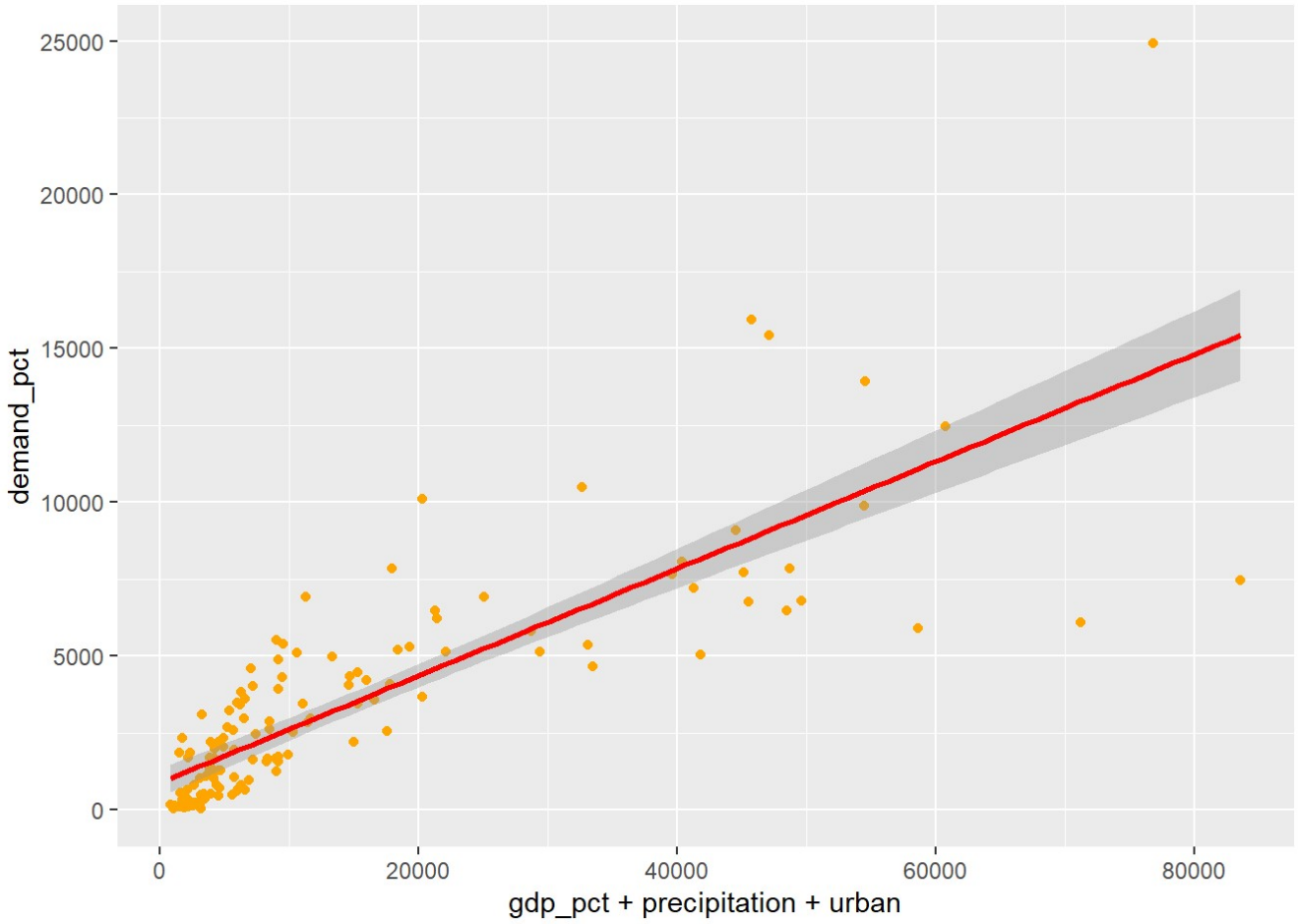
```
ggplot(regression_data, aes(x = gdp_pct + precipitation + urban, y= demand_pct)) + geom_po
int(color = "orange")
```



#Adding the linear line

```
ggplot(regression_data, aes(x = gdp_pct + precipitation + urban, y= demand_pct)) + geom_po
int(color = "orange") + geom_smooth(method="lm", col="red")
```

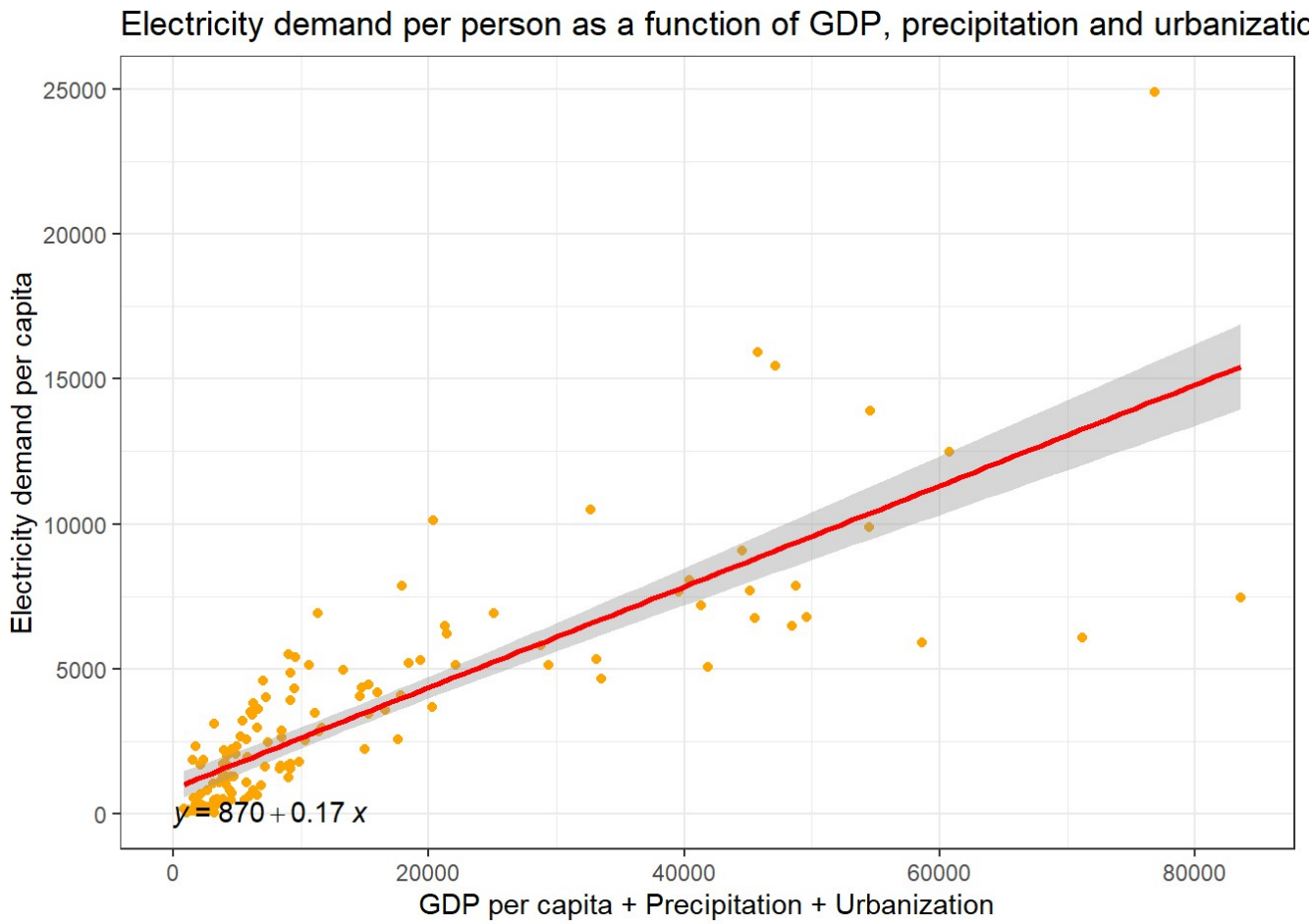
```
## `geom_smooth()` using formula = 'y ~ x'
```



#Adding the title and function for the linear regression graph.

```
ggplot(regression_data, aes(x = gdp_pct + precipitation + urban, y= demand_pct)) + geom_po
int(color = "orange") + geom_smooth(method="lm", col="red") + stat_regline_equation(label.
x = 5, label.y = 9) + theme_bw() +
  labs(title = "Electricity demand per person as a function of GDP, precipitation and urba
nization rate",
       x = "GDP per capita + Precipitation + Urbanization",
       y = "Electricity demand per capita")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



#FINAL COMMENTS #Our model is not great, however, it helps us understand the link between the electricity demand per person and some variables. To conclude we can say, a 100 USD increase in GDP per capita can increase the electricity demand per capita by 1 KwH. Besides, an increase in urbanization can also increase electricity demand. However, we can see the precipitation can decrease the level of electricity of demand.

#We can also conclude that the model has also some weaknesses.

#Firstly, it only has one year to observe. The model can benefit from if we can include more than one year and turn the model into a time series/panel data analysis. Lastly, there are also some statistically significant variables at the 10% level which is below the level of 1%.

#Secondly, the standard error for the urbanization is high. A standard error of 9.9% for a country is a bit high which may lead to incorrect insights.

#Last but not least, the R-squared value of the model is around 70 %. A higher percentage may better explain the pattern.