Assignment 2

Assignment 2

Task A. Plotting monthly averages of temparature, volume, price and wind production.

To discuss the relationships between temperature, volume, price and wind power production we chose The variables in the data set have vastly different scales. To surpass this issue when comparing the specified variables and in discussing their relationship, we chose to scale the values of the monthly averages. We chose to only included the last year of the sample, to make the plot more readable.

```
# Create a data frame of monthly averages of the variables volume, price,
# temperature and wind_production in long format.
df_avg_long <-
  df %>%

    # Select the specified variables
    select(date, volume, temperature, price, wind_production) %>%

    # Calculate monthly averages for each variable
    mutate(year_month = floor_date(date, "month")) %>%
    group_by(year_month) %>%
    summarise(
      avg_vol = mean(volume, na.rm = TRUE),
      avg_temp = mean(temperature, na.rm = TRUE),
      avg_price = mean(price, na.rm = TRUE),
      avg_wind = mean(wind_production, na.rm = TRUE)
    ) %>%

    # Scale the values
    mutate(
```

```
      avg_vol = scale(avg_vol),
      avg_wind = scale(avg_wind),
      avg_temp = scale(avg_temp),
      avg_price = scale(avg_price)
    ) %>%

    # Pivot to long format
    pivot_longer(
      cols = -year_month,
      names_to = "variable",
      values_to = "scaled values"
    )
```

```
# Plot the specified variables
df_avg_long %>%
  arrange(year_month) %>%
  tail(60) %>%
  ggplot(
    aes(
      x = year_month,
      y = `scaled values`,
      col = variable
    )
  ) +
  geom_line() +
  labs(
    title = "Relationship between variables",
    x = "Months",
    y = "Scaled values"
  ) +
  theme_classic()
```

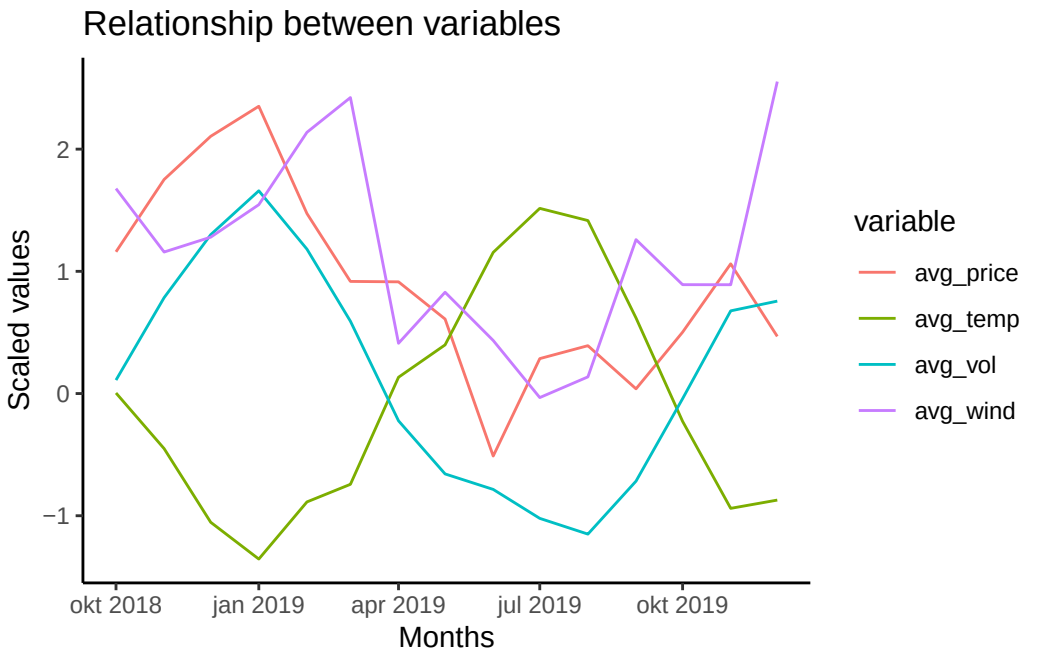

Figure 1: Monthly averages of temperature, volume, price and wind power production. The values have been scaled to fit into a single plot.

We notice large seasonal variations in all variables. Especially in average temperatures and volume there is little variation between each year. The monthly average price had a dip in the summer of 2015, but increased steadily up until the Covid19-pandemic in the spring of 2019. The wind power production have strong seasonal variations, but we notice that the average wind production has increased over the period by a large margin.

When analyzing the relationship between the variables, we noticed what is likely a negative correlation between temperature and volume. This seems likely as the temperature decreases the demand for energy increases. Also we notice the negative correlation between wind production and temperature. This is likely due to the fact that the colder periods like winter and autumn is more windy. Lastly, we notice the positive correlation between volume and price. This is likely because in periods of high demand, the supply of energy is strained resulting in a higher price for electric energy.

There is likely more relationships between the variables that can be identified and discussed using the plot we have provided. We chose to only comment on the most apparent of the relationships at this point.

Task B. Why will a OLS regression on quantity not provide an estimate of demand?

An OLS regression of quantity on price alone, even with controls, is not fitted to estimate demand due to many reasons:

1. Endogeneity exists as price and quantity in markets are simultaneously determined
2. Omitted variables like for example consumer income, advertising and seasonality. These can affect and lead to omitted variable bias.
3. Measurement errors in price and quantity can contribute to bias for the estimates for the parameters
4. OLS (or Gauss-Markow theorem) assumes a linear relationship. This may not hold if the demand curves are nonlinear.
5. Heteroskedasticity which is variability in the error term given any explanatory variable. In this case the variability in quantity varies across price levels which in return can affect standard erros.
6. The demand's dynamic behavior and seasonality may not be sufficiently captured. To estimate demand accurately, more advanced econometric methods are often required. These advanced methods usually contain instrumental variable regression and structural equation modelling. Controlling for relevant factors and improve our data quality can also make the estimation precision better.

Task C. Considering choice of instrument for price when estimating demand

When choosing a instrument for price when estimating demand, it's important to consider the validity of the instrument in terms of the requirements for a valid instrument. These requirements can be relevance, exogeneity, and exclusion restrictions:

1. **Why is temperature not a valid instrument for price when estimating demand?** The key requirements for exogeneity may not be satisfied if one were to use temperature as an instrument for price when estimating demand. (Exogeneity refers to the condition in which an independent variable is unrelated to the error term, indicating that it is not influenced by unobserved factors in the model.) Temperature is usually affected by factors that also may affect demand, like seasonality and consumber behavior. The instrument would not be valid due to endogeneity concerns if temperature is correlated with unobservable demand shocks.

2. **Why can magazine levels (or deviations) and wind power production potentially be good instruments for price when estimating demand?** Magazine levels and wind power production could potentially be sufficient instruments for price as they can fulfil the requirements for a valid instrument. They should be relevant in other terms they should be affecting the price. In addition, they should be exogenous, meaning that they should not be correlated with unobservable factors. But establishing their exogeneity requires thoughtful consideration and testing.

3. **Why is it necessary to control for seasonality (say, calendar month) and temperature?** It is necessary to control for seasonality and temperature to avoid omitted variable bias. (Omitted variable bias appear if a relevant variable that affects the dependent variable is excluded in a regression model. This leads to biased and estimates of the coefficients of the variables that are included.) Seasonality captures variation in the systematic demand throughout the year unrelated to price. Meanwhile temperature affects demand independently. When we control for these factors, we ensure precise demand estimates.

4. **How could controlling for weekday and year be useful?** Incorporating controls for weekday and year is beneficial to justify for demand patterns variation. Weekdays and weekends usually give different demand profiles given the consumers consumption. Meanwhile years would capture trends that are long-term and macroeconomic affects on demand. These controls strengthen the accuracy of demand estimates, while taking into account sources of variation unrelated to price.

Task D. Performing OLS regressions

With basis in the supplied equations, we sat out to perform a set of regressions. First we had to rescale some of the data to make the results more readable. We chose to divide the variables for volume and wind power production by 1,000. We also added variables to the data containing the year, month and weekday for each observation (or row).

```
# Create new variables in data frame containing the dummies for
# year, month and day in which the observation is made and scale variables.
df_new <-
  df %>%
  mutate(
    volume = volume / 1000,
    wind_production = wind_production / 1000,
    year = as.factor(year(date)),
    month = as.factor(month(date)),
    weekday = as.factor(wday(date,
                  label = FALSE,
```

```
                    week_start = 1))
  )
```

With our new data set, we were able to perform four regression, the first stage, reduced form, IV/2SLS and OLS. In the first stage regression we regressed P_t on all the exogenous variables from both the basis equations. For the reduced form regression, we regressed Q_t^D on all the exogenous variables from both equations as well. In the OLS regression, we estimated the main equation without considering endogeneity. Lastly, in the IV/2SLS regression, we used a two-stage procedure to predict the values of an endogenous explanatory variable using instrumental variables, and then used those predicted values to estimate the impact on the dependent variable.

```
# Perform specified regressions

# First stage
first_stage <- lm(
  price ~ magazine_level + temperature + wind_production + year + month + weekday,
  data = df_new
)

# Reduced form
reduced_form <- lm(
  volume ~ magazine_level + temperature + wind_production + year + month + weekday,
  data = df_new
)

# IV/2SLS
iv_model <- ivreg(
  volume ~ price + temperature + year + month + weekday
  | magazine_level + temperature + wind_production + year + month + weekday,
  data = df_new
)

# OLS
ols_model <- lm(
  volume ~ price + temperature + year + month + weekday,
  data = df_new
)
```

With the regressions at hand, we were able to create one table showing the results of the first stage, reduced form, IV/2SLS and OLS regressions. To make the table more readable, we omitted all the dummies and regression statistics from the table.

```r
# Regression table of all regression.
# Omitting statistics to make the table more readable.
# Note! Scaling of variables 'volume' and 'wind_production' of 10^-3, and
# that all dummies is omitted from the regression table.
stargazer(
  title = "Table including all regressions",
  first_stage,
  reduced_form,
  iv_model,
  ols_model,
  type = "text",
  omit = c("year", "month", "weekday"),
  omit.table.layout = "sn",
  dep.var.caption = "First stage (1) | Reduced form (2) | IV/2SLS (3) |OLS (4)"
)
```

```
Table including all regressions
================================================================================
                 First stage (1) | Reduced form (2) | IV/2SLS (3) |OLS (4)
                ------------------------------------------------------------
                     price                          volume
                      OLS             OLS        instrumental       OLS
                                                   variable
                      (1)             (2)            (3)             (4)
--------------------------------------------------------------------------------
magazine_level     -0.380***       1.282***
                    (0.022)        (0.162)

price                                              1.831***        1.173***
                                                   (0.364)         (0.146)

temperature        -0.321***      -12.755***      -12.036***      -12.341***
                    (0.036)        (0.270)         (0.331)         (0.292)

wind_production    -0.035***      -0.356***
                    (0.002)        (0.019)

Constant           63.158***     1,062.206***    1,037.639***    1,064.506***
                    (1.287)        (9.653)         (15.479)        (7.300)

================================================================================
```

```
================================================================================
```

Firstly, all included variables had statistically significant coefficients ($p < 0.01$). In the first stage regression, we notice that the standard errors of the coefficients are small, which can be interpreted that the included variables(especially magazine level and temperature due to the size of the coefficients) might be robust predictors of the endogenous variable, price. However, all unit increases of the variables lowers the price in the model.

The results of the reduced form regression indicated that in our model, all instruments and other exogenous variables together predict demanded quantity (volume). We take particular notice of the relatively large coefficient of the temperature variable, suggesting the large negative demand impact of increased temperature.

Comparing the IV/2SLS and the OLS regression, we notice that the respective size of the coefficients is similar but with substantial differences in the corresponding standard errors. This might be due to the fact that when addressing potential endogenity with IV estimation, price has a larger effect on quantity demanded (volume) than what is captured through OLS. This is highlighting the potential bias in the OLS regression.

Task E. Commenting on the first stage

In the previous task, we already made some short comments on the first stage regression. However, to comment further we can include the regression statistics to enable an assessment of instrument strength.

```r
# Regression table for first stage, including stats
stargazer(
  title = "First stage",
  first_stage,
  type = "text",
  omit = c("year", "month", "weekday")
)
```

```
First stage
================================================
                         Dependent variable:
                     ---------------------------
                                price
------------------------------------------------
magazine_level                -0.380***
                               (0.022)
```

```
temperature                   -0.321***
                               (0.036)

wind_production               -0.035***
                               (0.002)

Constant                      63.158***
                               (1.287)

-----------------------------------------------
Observations                    2,556
R2                              0.717
Adjusted R2                     0.714
Residual Std. Error       5.285 (df = 2529)
F Statistic          246.224*** (df = 26; 2529)
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

To assess the strength of instruments, one typically analyse the F-statistic of the regression. A general rule of thumb is that an F-statistic above 10 in the first stage regression indicates sufficiently strong instrument. The F-statistic of 246.224 is well above this threshold, indicating that magazine levels (M_t) is a good instrument - meaning that it explains a lot of the variation in price (P_t). Moreover, a Goodness of Fit (R^2) of 71.7% implies that the model explains a lot of the variability in price. On should in general not use the R^2 as an indicator as a validation of good instruments, but seen in combination with the F-statistic it at the minimum suggests that M_t is a strong instrument.

Task F. Commenting on the reduced form

Previously, we commented on the first stage regression. Before comparing the first stage and reduced form regressing, we include to regression table of the latter to make supplement the comments made in task D.

```r
# Regression table for reduced form, including stats
stargazer(
  title = "Reduced form",
  reduced_form,
  type = "text",
  omit = c("year", "month", "weekday")
)
```

Reduced form

	Dependent variable:
	volume
magazine_level	1.282***
	(0.162)
temperature	-12.755***
	(0.270)
wind_production	-0.356***
	(0.019)
Constant	1,062.206***
	(9.653)
Observations	2,556
R2	0.944
Adjusted R2	0.943
Residual Std. Error	39.648 (df = 2529)
F Statistic	1,627.381*** (df = 26; 2529)
Note:	*p<0.1; **p<0.05; ***p<0.01

Most notably, the F-static is statistically high ($p < 0.01$) and equals 1,627.381. This result suggests that the variables included in the regression jointly have a significant effect on the quantify demanded. Furthermore, the Goodness of Fit is exceptionally high at 94.4%. If we were to trust this result, it suggests that the model explains almost all the variation in quantity demanded. However, such a result might be an indicator of potential data leaking or overfitting, possibilities that we decided just to comment and not investigate further.

When comparing the first stage and the reduced form regression we systematically compared the coefficients. Regarding magazine level, in the first stage regression an increase is associated with a decrease in price. In the reduced form regression however, the association is the other way around. The objective of performing these two separate regressions is to consider magazine levels as an instrument, because it is believed related to price but not quantity demanded directly. Such a relationship might be plausible, as dropping magazine levels might cause the prices to rise, resulting in a drop in the quantity demanded. This relationships is also plausible the other way around, as rising magazine levels might reduce the price, causing the demand

to increase. This suggested relationship should be analysed further with the help of statistical tests (which we will not do here).

We notice that the coefficients of temperature and wind production suggests that increases is associated with decreases in the dependent variables. By using the correlations identified in Figure 1, these relationships might be plausible. However, as with magazine levels, further discussions of the relationships should be backed up by statistical tests.

Task G. Application of the IV/2SLS estimates