Relating tCO2 to Energy Consumption in Steel Industry Data – Executive Summary

William Stults

Western Governors University

May 10, 2023

As governments worldwide give greater attention to environmentally friendly legislation, the focus on reducing CO_2 emissions has been increasing. In the United States, regulations are already in place at the federal level limiting CO_2 emissions from motor vehicles. It is estimated that "30 percent of U.S. greenhouse gas emissions come from industry, including indirect emissions from the sector's electricity consumption" (Center for Climate and Energy Solutions, 2021). The industrial sector will be a logical area of focus for reduction of greenhouse gasses in the coming years. My research project utilized one year's worth of energy consumption and tCO2 (total carbon dioxide) data taken from the DAEWOO Steel Co. Ltd in Gwangyang, South Korea, provided courtesy of the UCI Machine Learning Repository (Dua and Graff, 2021). I hypothesized that a statistically significant relationship exists between energy consumption kWh and tCO2, which I would consider proven via p-value ≤ 0.05 and r-squared $> .90$.

The data set consists of 35040 rows across 11 variables. Using Python via Jupyter Lab the data was imported, cleaned, and reduced to the variables relevant to the hypothesis, energy_kWh and tCO2_CO2. A standard set of libraries was employed during data cleaning and analysis, including Pandas, NumPy, Seaborn and Statsmodels. Beyond cleaning and reduction, the following data analysis tasks were completed:

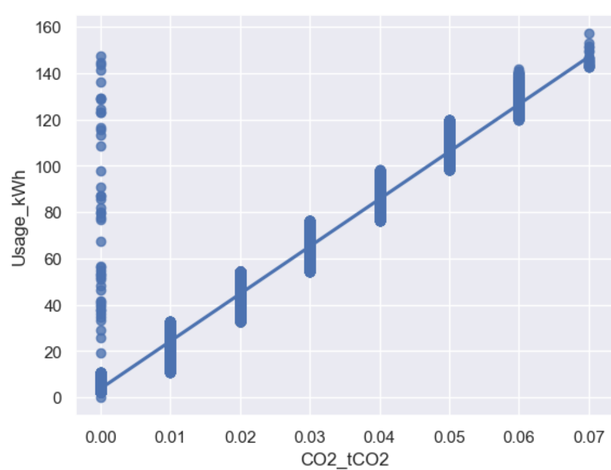
1) A correlation matrix was generated to determine the degree to which the relevant features of the data set might be related.
2) A linear regression model was fitted and evaluated for further statistical corroboration of correlation.
3) The resulting model was evaluated for MSE, RSE, and R-Squared metrics.

Per the results of the correlation matrix, a high degree (~0.988) of positive correlation was identified between the two variables:

```
df.corr()
```

	Usage_kWh	CO2_tCO2
Usage_kWh	1.00000	0.98818
CO2_tCO2	0.98818	1.00000

The linearity of this correlation was visualized via a regplot using Seaborn:



The results of the linear regression model can be seen here, indicating a p-value of 0.00 and an r-squared value of 0.976:

```
# Create model and display summary
mdl_Usage_kWh_vs_CO2_tCO2 = ols("Usage_kWh ~ CO2_tCO2", data=df).fit()
print(mdl_Usage_kWh_vs_CO2_tCO2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Usage_kWh   R-squared:                       0.976
Model:                            OLS   Adj. R-squared:                  0.976
Method:                 Least Squares   F-statistic:                 1.456e+06
Date:                Sun, 07 May 2023   Prob (F-statistic):               0.00
Time:                        13:50:07   Log-Likelihood:            -1.0699e+05
No. Observations:               35040   AIC:                         2.140e+05
Df Residuals:                   35038   BIC:                         2.140e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.8051      0.034    113.086      0.000       3.739       3.871
CO2_tCO2    2046.2773      1.696   1206.604      0.000    2042.953    2049.601
==============================================================================
Omnibus:                    49038.339   Durbin-Watson:                   1.136
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         46356205.792
Skew:                           7.768   Prob(JB):                         0.00
Kurtosis:                     180.509   Cond. No.                         61.9
==============================================================================
```

The strong relationship between the variables was further confirmed via the MSE and RSE of the model:

```
# Display MSE, RSE, Rsquared and Adjusted Rsquared for model
mse_all = mdl_Usage_kWh_vs_CO2_tCO2.mse_resid
print('MSE of model: ', mse_all)
rse_all = np.sqrt(mse_all)
print('RSE of model: ', rse_all)
print('Rsquared of model: ', mdl_Usage_kWh_vs_CO2_tCO2.rsquared)
print('Rsquared Adjusted of model: ', mdl_Usage_kWh_vs_CO2_tCO2.rsquared_adj)
```

```
MSE of model:  26.286950187221958
RSE of model:  5.127080083948559
Rsquared of model:  0.9764992611554846
Rsquared Adjusted of model:  0.9764985904340152
```

The data analysis was limited in that only 2 of the 11 variables of the data set were utilized for the study. In addition, the data utilized came from a single manufacturing plant and cannot be assumed to reflect the conditions of manufacturing plants worldwide. Additionally, while one year's worth of data was sufficient to establish the suspected relationship, a greater number of observations over a longer period of time would offer more conclusive results.

Within the context of this study, a recommendation can be made to businesses in the manufacturing sector wishing to minimize CO2 emissions that they monitor their energy consumption and make efforts to mitigate against large energy consumption spikes occurring over short periods.

My research also revealed some additional approaches for future study of this data set. With the simple regression model showing significant strength, a logical next step would be a “train” and “test” split of the data to determine more conclusively how accurate its predictions are.

The second potential approach would be to expand the analysis to include all the variables in the data set. Upon reviewing a correlation matrix of the continuous variables, there are already additional variable pairs that indicate strong relationships which may also be useful in predictive modeling (code and output for the matrix is shown below). Expanding the model to include additional variables may result in a multiple linear regression model with greater predictive accuracy.

To conclude, according to the results my research yielded, manufacturing companies will be able to predict the tCO_2 presence in ppm based on a manufacturing plants energy consumption in kWh to a 97.6% degree of accuracy. This will be of great benefit when utilized for logistical planning aimed at reducing CO_2 emissions and adhering to any future regulations imposed by the federal government.

Citations

Dua, D. and Graff, C. (2021, March 30). *Steel Industry Energy Consumption Dataset Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Steel+Industry+Energy+Consumption+Dataset#

Center for Climate and Energy Solutions. (2021). *Controlling Industrial Greenhouse Gas Emissions.* https://www.c2es.org/content/regulating-industrial-sector-carbon-emissions/