Modelling_Assignment_3 (4)

December 11, 2024

1 Modelling Assignment 3: Charging Cost of EVs in New York

Yirui Wang, Mateo Domínguez De La Llera, Yeyang (David) Ou, Xinyi Wang

1.1 Problem Statement

Construct a linear regression model to predict charging cost of Electric Vehicles (EVs) in New York State in U.S.

1.2 Variables and Parameters

Description	Symbol	Unit	Type
Total Battery capacity	X_1	kWh	covariates
Total energy consumed during a charge	X_2	kWh	covariates
Total charging duration	X_3	h	covariates
The average charging rate	X_4	kW	covariates
Total distance driven	X_5	km	covariates
The charging cost	Y	USD	target
Coefficient for intercept	β_0	USD	parameter
Coefficient for battery capacity	β_1	1/kWh	parameter
Coefficient for total energy consumed during a charge	β_2	1/kWh	parameter
Coefficient for total charging duration	β_3	1/h	parameter
Coefficient for average charging Rate	β_4	1/kW	parameter
Coefficient for total distance driven	β_5	1/km	parameter

1.3 Assumptions and Constraints

- The data satisfies the assumptions needed to perform linear regression (i.e. the 4 assumptions on the error term ϵ).
- Trends in the historical data are expected to remain constant and continue into the future.
- All the data provided are accurate.

- The price of electricity remains constant across the city.
- There are no external factors that influence the performance of the EVs.
- The battery capacity of the EVs remains the same.
- The batteries have no leaks, so all the energy in the batteries is used for driving.
- There is no energy loss in the circuit of the EVs.
- The energy needed to power the car's control system, air conditioning, and other systems is the same for all cars.

1.4 Building Solutions

We first make a multiple linear regression using all data and no transformations: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$.

[ ]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.api as sm
from scipy.stats import poisson
from scipy.stats import norm
%matplotlib inline
```

[ ]:
```
# Import data
ev = pd.read_csv('ev_charging_NY.csv')
ev = pd.DataFrame(ev)
# Define and run model
X = ev[['Battery_Capacity', 'Energy_Consumed', 'Charging_Duration',␣
 ↪'Charging_Rate', 'Distance_Driven']]
Y = ev['Charging_cost']
X = sm.add_constant(X)
reg = sm.OLS(Y,X).fit()
summary_table=reg.summary()

coefficients = reg.params; std_errors = reg.bse; p_values = reg.pvalues;
R_squared = reg.rsquared; R_adj = reg.rsquared_adj

summary_df = pd.DataFrame({'Coefficient': coefficients,'Std. Error': std_errors,
    'P>|t|': p_values,'R-squared': R_squared,'R. adj': R_adj}).round(3)

print(summary_df)
```

```
                   Coefficient  Std. Error  P>|t|  R-squared  R. adj
const                   -0.554       1.043  0.596      0.904   0.901
Battery_Capacity        -0.006       0.009  0.559      0.904   0.901
Energy_Consumed          0.212       0.008  0.000      0.904   0.901
Charging_Duration        4.264       0.169  0.000      0.904   0.901
Charging_Rate            0.224       0.012  0.000      0.904   0.901
Distance_Driven         -0.000       0.002  0.998      0.904   0.901
```

1.5 Analyze and Assess

The adjusted R^2 is high at 0.901, but there is a strong multicollinearity between the covariates. We now analyze each of the defining assumptions for linear regression.

Check Assumption 1: Average value of the error is zero ($\mathbb{E}$ = 0).

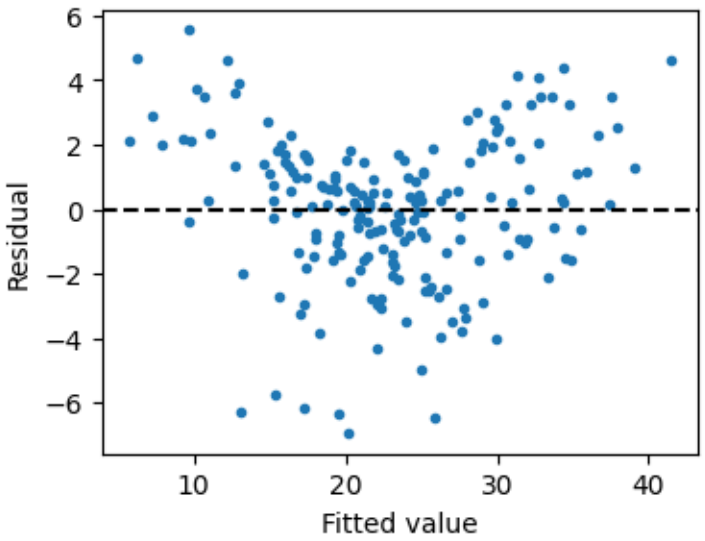
Looking at the residuals vs. covariates plots, we see several problems: 1. Battery capacity clearly shows discrete data, with most values at 50, 62, 75, 85, and 100, but the mean still appears to be 0. 2. Energy consumed and charging duration both shows data in a sort of double-cone, with variance first decreasing and then increasing. This indicates a violation of the regression assumptions. 3. Charging rate displays a weak downward parabolic behavior. 4. Distance driven is great, as all points are randomly scattered.

For the partial regression plots, battery capacity and distance driven both are randomly scattered, so there is no linear relationship with the beta coefficients, again violating the assumption of zero mean error. The plots for energy consumed, charging duration, and charging rate are good.

Check Assumption 2: Variance of Errors is 0 (var( ) = 0)

The residuals vs fitted value graph isn't randomly scattered and has a slight parabolic characteristics. This violates the assumption that variance of the errors equal to 0.

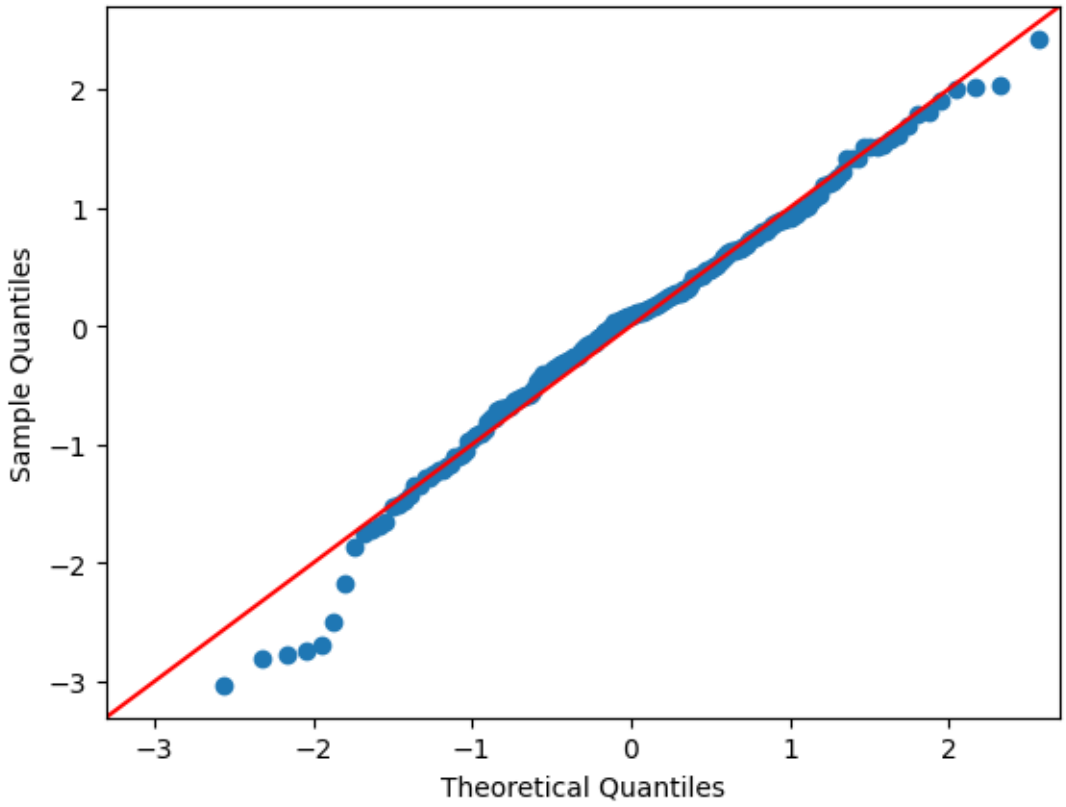
[ ]:
```
fitted_y = reg.fittedvalues
residuals = reg.resid; plt.figure(figsize=(4, 3))
plt.scatter(fitted_y,residuals, marker='.')
plt.axhline(y = 0, color = 'k', linestyle = '--')
plt.xlabel('Fitted value'); plt.ylabel('Residual')
plt.show()
```



Check Assumption 3: Errors are normally distributed.

We don't see major problems with the QQ plot of the models as the datasets display a quite clear linear relationship.

[ ]:
```
#residuals = reg.resid
plt.figure(figsize=(2, 2)); sm.qqplot(residuals,fit=True, line="45"); plt.show()
```



Check Assumption 4: (Errors are independent) Since the data does not involve time series, there aren't dependencies on the error. Hence, we don't (and can't) run the residual vs. index plot.

1.6 Transformed Model

Our preliminary model exhibited violations of assumptions 1 and 2. Hence, we propose a new model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_2 \cdot X_3) + \beta_4 \log(X_4) + \beta_5 \sqrt{X_5}$.

[ ]:
```
ev['Total_energy'] = ev['Energy_Consumed']*ev['Charging_Duration']
ev['Charging_Ratenew'] = np.log(ev['Charging_Rate'])
ev['Distance_Driven_new'] = (ev['Distance_Driven'])**(1/2)

X2 =␣
 ↪ev[['Battery_Capacity','Total_energy','Charging_Ratenew','Distance_Driven_new']]
Y2 = ev['Charging_cost']
```

```
X2 = sm.add_constant(X2)
reg_2 = sm.OLS(Y2,X2).fit()
summary_table=reg_2.summary()

coefficients = reg_2.params; std_errors = reg_2.bse; p_values = reg_2.pvalues
R_squared = reg_2.rsquared; R_adj = reg_2.rsquared_adj

summary_df = pd.DataFrame({'Coefficient': coefficients, 'Std. Error':␣
 ↪std_errors,
    'P>|t|': p_values, 'R-squared': R_squared, 'R. adj': R_adj}).round(3)
print(summary_df)
```

```
                    Coefficient  Std. Error  P>|t|  R-squared  R. adj
const                     0.666       0.225  0.003      0.997   0.997
Battery_Capacity          0.002       0.002  0.342      0.997   0.997
Total_energy              0.100       0.000  0.000      0.997   0.997
Charging_Ratenew          4.317       0.042  0.000      0.997   0.997
Distance_Driven_new      -0.007       0.009  0.398      0.997   0.997
```

- The adjusted R-squared is high at 0.997.
- The multicollinearity reported in the former model is no longer present in the new model.
- From the p-value, we can see that `Total_energy` and `Charging_Ratenew` are highly significant ($p < 0.001$). While `Battery_Capacity` and `Distance_Driven` have p-values greater than 0.05, indicating that they are not statistically significant predictors in this model.
- The skewness (0.887) implies that residuals are not symmetrically distributed. The qq-plot that follows further supports this.

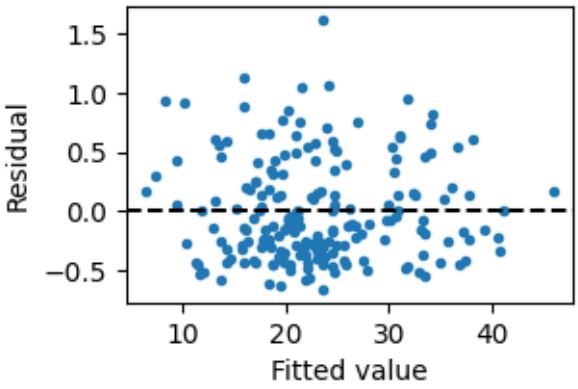
Check Assumption 1: Average value of the error is zero ($\mathbb{E}$ $= 0$).

After the transformation, the residual vs battery capacity graph sthas a discrete patterns and isn't randomly scattered. The partial regression plot of the battery capacity covariates doesn't display a linear relationship.

Check Assumption 2: Variance of Errors is 0 (var( ) $= 0$)

We see an improvement in the residual vs fitted values graph. The graph looks randomly scattered, albeit still with outliers at around (25, 1.6) and (45, 0.2). Hence the assumption that the variance of the error is 0 is satisfied.

[ ]:
```
fitted_y_1 = reg_2.fittedvalues
residuals_1 = reg_2.resid
plt.figure(figsize=(3, 2))
plt.scatter(fitted_y_1,residuals_1, marker='.')
plt.axhline(y = 0, color = 'k', linestyle = '--')
plt.xlabel('Fitted value'); plt.ylabel('Residual')
plt.show()
```

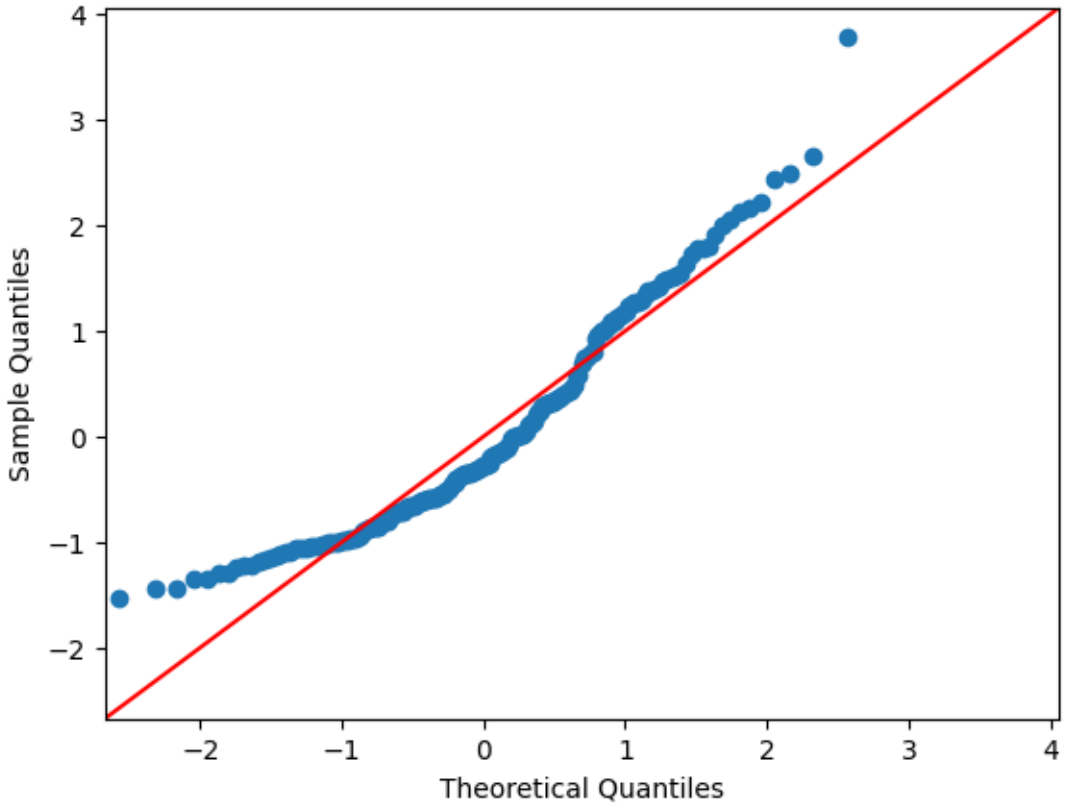


Check Assumption 3: Errors are normally distributed.

The QQ plot of the transformed model displays a slight parabolic characteristics. When we remove the outlier at (2.6, 3,9), the datasets has a slightly better linear relationship.

[ ]:
```
plt.figure(figsize=(2, 2)); sm.qqplot(residuals_1,fit=True, line="45"); plt.
↪show()
```

```
<Figure size 200x200 with 0 Axes>
```



1.7 Conclusions

Our preliminary model with no transformations exhibited several violations to the assumptions of zero mean error and constant variance. Therefore, we proposed the following multiple linear regression model the charging cost: $Y = \beta_0 + \beta_1 X_1 + \beta^*(X_2 \cdot X_3) + \beta_4 \log(X_4) + \beta_5 \sqrt{X_5}$.

The values of each of the beta coefficients are as follow:

- The constant (β_0) : The baseline `Charging_cost` is 0.666 when all predictors are zero.
- Battery_Capacity (β_1): For each unit increase, `Charging_cost` increases by 0.0017 units, but this effect is statistically insignificant.
- Total_energy (β^*): This is a new beta values we calculated. For every unit increase, `Charging_cost` rises by 0.0996 units, a highly significant and strong predictor.
- Charging_Ratenew (β_4): For every unit increase, `Charging_cost` increases by 4.3171 units, confirming it as another strong predictor.
- Distance_Driven (β_5): The coefficient (-0.0073) suggests a slight decrease in `Charging_cost` per unit increase, but this effect is not statistically significant.

Below are the models to predict the charging costs:

$$Y = 0.666 + 0.0017X_1 + 0.0996(X_2 \cdot X_3) + 4.3171 \log(X_4) - 0.0073\sqrt{X_5}$$

The QQ plot of the new model has a parabolic shape, proving a violation of the third assumption for linear regressions. This would usually be enough to discard the model. However, we decided this is the best one to predict the charging cost, as models with different transformations performed worse in predicting the cost or had even more serious violations to the abovementioned assumptions. The source of this problem is very probably in the nature of the data utilized: consumers or producers might have a non-gaussian or non-random preference for the type of batteries or driving behavior. Ergo, it is unavoidable to have problems in the data.

Future Improvements: For future improvements of modelling EV charging costs in New York, we could remove the outliers and continue with the simple linear regression model. Beyond this, more sophicated techniques can be introduced to better capture the relationship between total energy consumed, total charging duration, and other covariates. Also seeing the discrete pattern of battery capacity, we could incorporate a more robust technique into the model.

[ ]: