Linear Regression for Energy Demand

Ben Rickard

November 2024

Contents

1 Introduction to Linear Regression

1.1 Model Specification

Assume we have a response variable $\mathbf{Y}$ which we wish to model as a linear combination of some predictors $\mathbf{X} = (x_1, x_2, ..., x_p)$. We notate such model as:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon \\ &= \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon \end{aligned}$$

where ϵ represents the extra noise term. We can also write this in terms of matrices where we now have n response variables. This gives us:

$$\begin{aligned} \mathbf{Y} &= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \\ &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned}$$

We call $\mathbf{X}$ the design matrix, and $\boldsymbol{\beta}$ is a p-column vector of fixed but unknown parameters which we want to learn. Row-wise, this gives us the exact same equation as previous.

1.2 Assumptions

In order for our noise term $\boldsymbol{\epsilon}$ to be meaningful, we make some assumptions. First note that we consider $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ a random vector. We then have the following three assumptions regarding $\boldsymbol{\epsilon}$:

1. Linearity: $\mathbb{E}[\epsilon_i] = 0 \qquad$ i.e. $\mathbb{E}[y_i | \mathbf{x}_i] = \mathbf{x}_i^T \boldsymbol{\beta}$
2. Homoscedasticity & Independence: $\mathbb{V}\text{ar}[\epsilon_i] = \sigma^2$ and $\mathbb{C}\text{ov}[\epsilon_i, \epsilon_j] = 0$ giving $\mathbb{V}\text{ar}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}$
3. Normal: ϵ_i is normally distributed

Joining all of these together gives us $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, and noting that $\mathbf{X}\boldsymbol{\beta}$ is constant, $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$. These assumptions allow our linear models to be meaningful, and also let us perform data analysis on the model parameters, such as hypothesis testing and confidence intervals.

1.3 Estimation of Model Parameters

We calculate the least-squares estimate of $\boldsymbol{\beta}$ called $\hat{\boldsymbol{\beta}}$ by minimising the sum of the residuals squared $\sum_{i=1}^{n} \epsilon_i^2$. Basic differentiation gives us, provided that $(\mathbf{X}^T\mathbf{X})^{-1}$ is invertible, the LS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ which can also be shown to be an unbiased estimator.

The unbiased estimator of σ^2, based on known residuals $\hat{\epsilon_i}$, is:

$$s^2 = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n-p}$$

It can also easily be shown that:

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

which allows us to perform hypothesis tests on the significance of each β_i.

2 Modelling UK Energy Demand

2.1 Motivation

TNEI works with numerous different companies and agencies who generate energy for the UK. In particular, a lot of these companies generate energy from renewable sources. Hence it is vital for these people to have an idea of what the UK's energy demand looks like, in order so that it can consistently be met by renewable sources. This means the government will know when we can keep moving further away from fossil fuels and importing energy internationally.

It seems intuitive that energy demand will be seasonal, with more being used in colder months when more heating will be on. Hence we need data for energy usage for every day through several years. Temperature also seems like a valuable predictor, so we will also want data for daily temperatures going back numerous years.

2.2 Getting the Data

First of all, we need some data of the UK's energy demand over several years. The National Energy System Operator (NESO) have data for daily UK energy transmission going back to June 2021. This is available to be downloaded as a CSV file.

Next we need to find some temperature data. The MetOffice has monthly maximum and minimum temperatures going back over a hundred years. I was

unable to find daily data, so I settled on downloading this as a txt file. It seems clear that there is likely a high correlation between the maximum and minimum temperature, which is something that is considered in the modelling later.

2.3 First Model

First I used a simple linear model, where the predictors were simply date, maximum and minimum temperature for the month. However, calculating the correlation between the maximum and minimum temperature gave a value of 0.975 meaning they are highly correlated. Hence, I decided to only use the maximum temperature as it is more significant. In fact both maximum temperature and date were significant predictors for our model, both having p-values less than 2×10^{-16}.

However, the adjusted-R^2 value was only 0.586. This means that there is still a significant amount of our variability in demand unexplained by our predictors. We also check some common diagnostics to ensure our assumptions hold in Figure 1. In the first plot, it appears that our residuals are indeed random and centered on zero against the fitted values, implying our first two assumptions hold well. However, we can see the residuals don't fit the QQ Normal plot strongly, especially at the tails. Hence, there could be issues when calculating the significance of our predictors, as this assumed Normality.

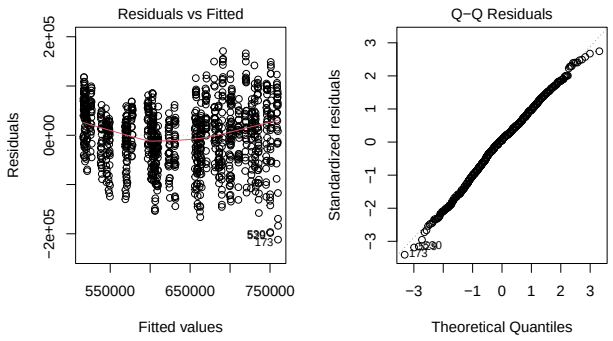

Figure 1: Diagnostics for Model 1

We now showcase the plot of our data, and the best fit line calculated in accordance with Model 1 in Figure 2. We can see the regression line follows the general shape of the data, but is not particularly smooth which we wouldn't expect, and doesn't quite capture the peaks.

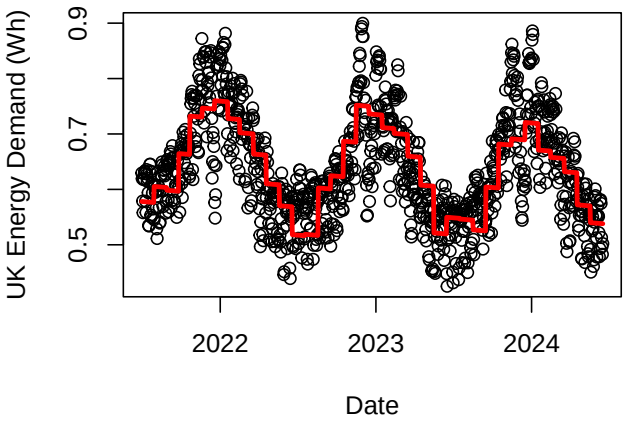

Figure 2: Model 1 Regression

2.4 Second Model

In an attempt to smooth the regression line, and capture more easy the oscillatory nature of the relation, we use $cos(\frac{2\pi \text{time}}{365})$ and $sin(\frac{2\pi \text{time}}{365})$ as our predictors. These are both significant predictors, but our adjusted-R^2 value is still only 0.606. This doesn't affect our linear regression process, and we get very similar residual diagnostics.

However, the regression plot looks different as seen in Figure 3. The regression line is now much smoother which we prefer. Yet we still are not capturing the full range of the demand. This is a limitation of our approach. Perhaps using a linear regression is too simple here, and a generalised linear model may be more appropriate. In addition our relatively low adjusted-R^2 implies that we need further predictors to capture the full variability in energy demand.

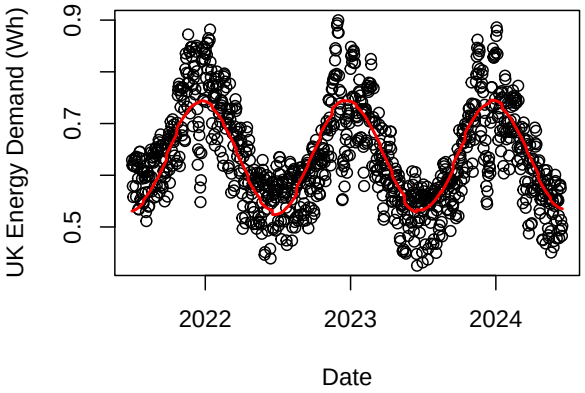

Figure 3: Model 2 Regression