Computational Social Science Masters

Introduction to Forecasting

Graham Elliott

September 8, 2022

Approaches to Forecasting

Consider two rather different approaches to constructing a forecast.

1. Carefully construct a model based on theoretical knowledge of the area, simulate it to provide a forecast of what is to come.

2. Construct a dataset and use some statistical models of the relationship between the data to provide a forecast of what is to come.

Approaches to Forecasting

How would you do # 1? We need

(a) Need to know a lot about the fundamentals of the area, construct a model that is close to reality.

(b) Computational aspects still exist, but they are less statistics and more model simulation. Often parameters of the model are estimated using statistical methods and slotted into the model.

This is basically how weather forecasting works.

Approaches to Forecasting

How would you do # 2? We need

(a) To know what data to collect.

(b) Some idea of what types of models to consider.

(c) Some way of evaluating how well the model might work going forward.

All of these require some theoretical understanding of what you are trying to do (CS people call this domain knowledge).

Approaches to Forecasting

Most approaches mix an understanding of the area with a statistical approach using data to actually computing models.

Some criticisms of this approach

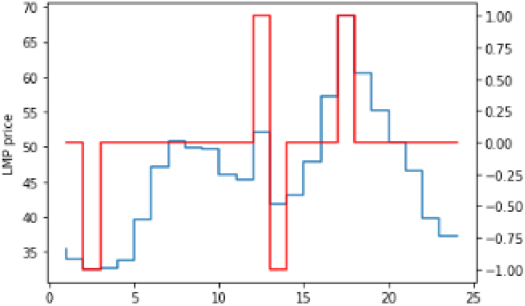
(a) Only works when we have data on the problem.

(b) Maybe a really simple model is better than a complicated statistical analysis.

(c) Can only predict things that have happened before with any confidence.

(d) Often need not just a forecast but a story to go with it.

A Real Forecasting Problem



What am I trying to do?

There are a number of elements to think about here.

1. What is it that good forecasts do better than poor forecasts?
2. What data is relevant to build a model?
3. What types of models can I build?

Basic Setup

A forecaster observes $z = \{z_t\}_{t=1}^{T}$
(for example we may have $z_t = [y_t, x_t']$ for x_t being $kx1$)

We want to forecast an unknown future variable y
(e.g. $y = y_{T+1}$)

The forecasters problem is to use the observation of z as well as possible to predict y

Basic Setup

We consider observed data z as an outcome of a random variable Z.

Object to be forecast is an outcome y of a random variable Y.

A point forecast $f(z)$ is an outcome of a random variable $f(Z)$
Often will write as $f(z, \beta)$ to be clear about the notion of a model, but then we need to estimate β which will be a function of z

There will be a true distribution of the data $p(y, z, \theta)$ that is of course unknown. We refer to this as the DGP.

Basic Setup - Loss Functions

We need a method of choosing between different forecast methods

Define the loss function as $L(f(z), y)$.

The mapping is to the real number line (usually $\mathbb{R}^+$ or a subset of the real number line.
How we measure loss when $f(z) \neq y$

Loss can be considered a random variable $L(f(Z), Y)$

Basic Setup - Loss Functions

Examples of loss functions

Mean Square Error

Mean Absolute Error

But not Mean Absolute Percentage Error

Basic Setup - Loss Functions

Loss functions only need to take values on possible errors

Consider the classification problem.

Basic Setup - Simplest Example

Forecast y_{T+1} having observed i.i.d. normal data $\{y_t\}_{t=1}^{T}$

Suppose we have squared (MSE) loss, and we know the mean of y_{T+1} is μ.

Then

$$min_f E_Y[y_{T+1} - f]^2$$

Basic Setup - Simplest Example

$$E_Y[y_{T+1} - f]^2 = E_Y[(y_{T+1} - \mu) - (f - \mu)]^2$$

Basic Results

Minimizing MSE results in an optimal forecast that is the conditional mean $E[Y|Z]$

Minimizing MAE results in an optimal forecast that is the conditional median of Y given Z.

Other loss functions result in optimal forecasts that are other functions of the conditional distribution of Y given Z.

Basic Results

None of this tells us how to estimate the conditional mean or summary statistic we are looking for.

1. What is the best model for the conditional mean?

2. Given a model, what is the best estimator using observations z?

Basic Results - The General Setup

To summarize this a bit more formally.

The forecasting problem can be considered to be the problem of choosing $f(z)$ to solve

$$f^*(z) = argmin_{f \in \mathcal{F}} \int L(f(z), y) p_Y(y|z, \theta) dy$$

Basic Results - Risk

Once we have selected a forecasting model $f(z)$, we can consider the risk function $R(z, \theta)$

$$R(z, \theta) = \int L(f(z), y) p_Y(y|z, \theta) dy$$

Basic Setup - Simplest Example

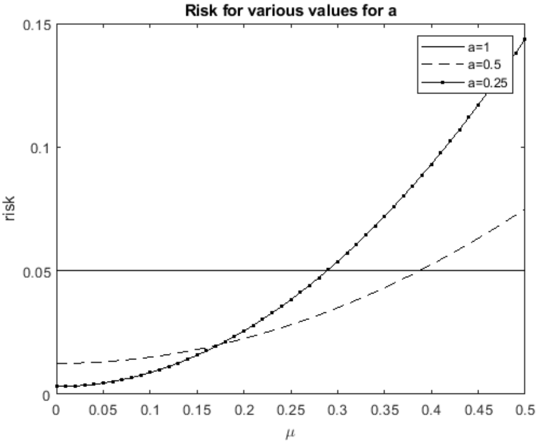
Forecast y_{T+1} having observed i.i.d. normal data $\{y_t\}_{t=1}^{T}$, we want to estimate the mean.

Consider estimators of the form $f = a\bar{y}_T$ where $0 < a < 1$.

What is the risk?

Risk - Example 1

$$E_{Y,Z}[y_{T+1} - a\bar{y}_T]^2 = (1 + T^{-1}a^2) + (1 - a)^2\theta^2$$



The Data and Problem

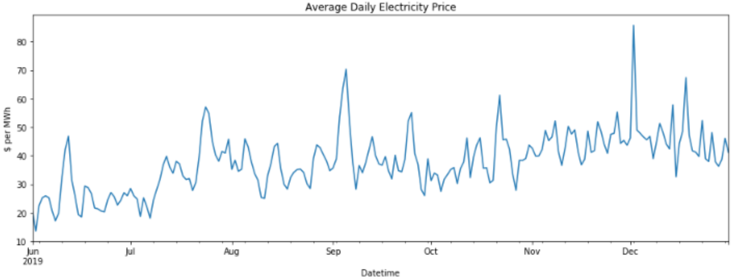
We want to forecast daily average electricity prices, we will abstract from why for now.

We will assume that we want to minimize MSE, this suggests we want an estimator for the conditional mean of the price given any data available.

As a first step we can consider using the dynamic structure of prices to forecast.

The Data and Problem

First we should look at the data.



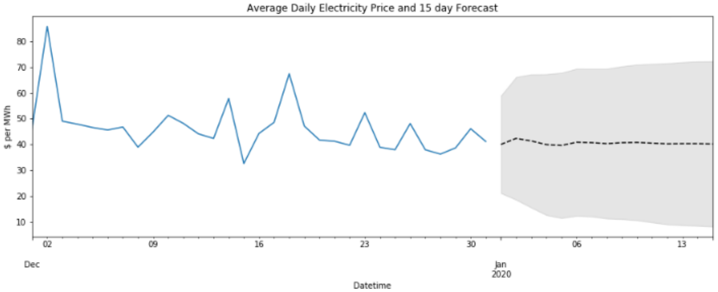
Try an autoregressive model with 7 lags

```
                               Statespace Model Results
==============================================================================
Dep. Variable:                    LMP   No. Observations:                  365
Model:               SARIMAX(7, 0, 0)   Log Likelihood               -1346.063
Date:                Fri, 15 Jul 2022   AIC                           2710.125
Time:                        09:10:24   BIC                           2745.224
Sample:                    01-01-2019   HQIC                          2724.074
                         - 12-31-2019
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      4.0811      1.498      2.723      0.006       1.144       7.018
ar.L1          0.7628      0.032     23.610      0.000       0.699       0.826
ar.L2         -0.0491      0.056     -0.873      0.383      -0.159       0.061
ar.L3          0.0993      0.051      1.934      0.053      -0.001       0.200
ar.L4         -0.0423      0.055     -0.769      0.442      -0.150       0.066
ar.L5         -0.0373      0.059     -0.627      0.531      -0.154       0.079
ar.L6         -0.0281      0.072     -0.392      0.695      -0.169       0.112
ar.L7          0.1919      0.055      3.521      0.000       0.085       0.299
sigma2        93.1003      3.898     23.882      0.000      85.460     100.741
===================================================================================
Ljung-Box (Q):                       60.48   Jarque-Bera (JB):               1365.13
Prob(Q):                              0.02   Prob(JB):                          0.00
Heteroskedasticity (H):               0.42   Skew:                              1.58
Prob(H) (two-sided):                  0.00   Kurtosis:                         11.93
===================================================================================
```

Autoregressive Model

Forecasts and the inbuilt confidence intervals

```
In [8]: fig, ax = plt.subplots(figsize=(15, 5))

        # Plot the data (here we are subsetting it to get a better look at the forecasts)
        bd.LMP.loc['2019-12-01':'2019-12-31'].plot(ax=ax)

        # Construct the forecasts
        fcast = res.get_forecast('2020-01-15').summary_frame()
        fcast['mean'].plot(ax=ax, style='k--')
        ax.fill_between(fcast.index, fcast['mean_ci_lower'], fcast['mean_ci_upper'], color='k', alpha=0.1);
        plt.title('Average Daily Electricity Price and 15 day Forecast')
        plt.ylabel('$ per MWh')

Out[8]: Text(0, 0.5, '$ per MWh')
```

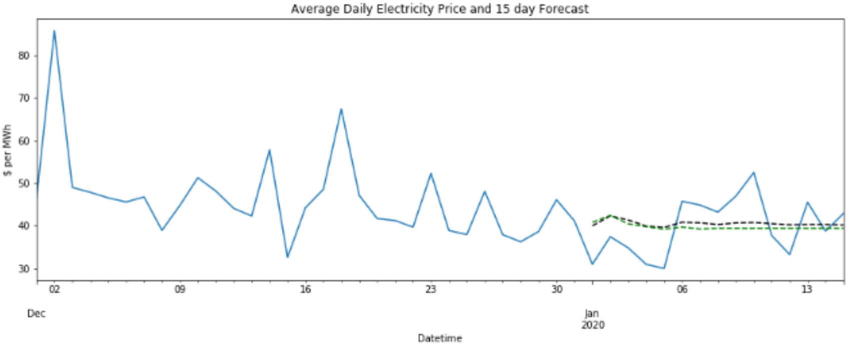


Moving Average Model

Try an MA model with 7 lags

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                    LMP   No. Observations:                  365
Model:               SARIMAX(0, 0, 7)   Log Likelihood               -1362.231
Date:                Wed, 20 Jul 2022   AIC                           2742.461
Time:                        10:00:06   BIC                           2777.560
Sample:                    01-01-2019   HQIC                          2756.410
                         - 12-31-2019
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     39.4413      2.750     14.340      0.000      34.051      44.832
ma.L1          0.8393      0.033     25.627      0.000       0.775       0.904
ma.L2          0.6685      0.047     14.258      0.000       0.577       0.760
ma.L3          0.5924      0.060      9.820      0.000       0.474       0.711
ma.L4          0.4439      0.051      8.776      0.000       0.345       0.543
ma.L5          0.2237      0.049      4.546      0.000       0.127       0.320
ma.L6          0.0406      0.052      0.774      0.439      -0.062       0.143
ma.L7          0.0633      0.042      1.515      0.130      -0.019       0.145
sigma2       101.8142      4.438     22.941      0.000      93.116     110.513
===================================================================================
Ljung-Box (Q):                      120.03   Jarque-Bera (JB):               855.67
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.44   Skew:                             1.38
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.98
===================================================================================
```

Moving Average Model



Seasonal ARMA Model

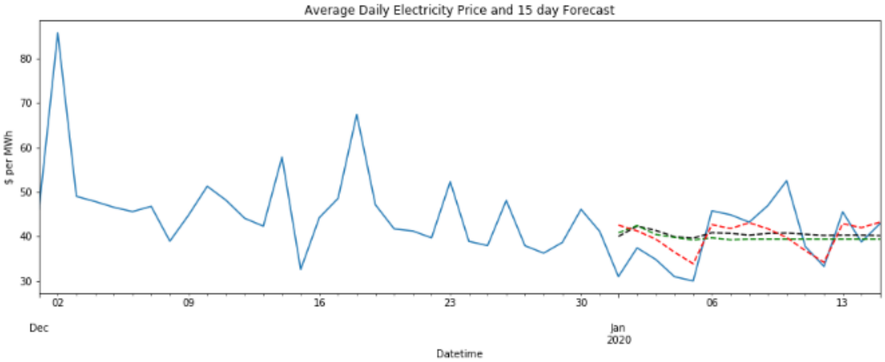
Try an AR with one lag with week seasonals in AR and MA

```
# Construct the model with daily 'seasonal'
mod = sm.tsa.SARIMAX(bd.LMP['2019'], order=(1, 0, 0), trend='c',seasonal_order=(1,0,1,7),freq="D")

# Estimate the parameters
ress = mod.fit()

print(ress.summary())
```

```
D:\Python\lib\site-packages\statsmodels\tsa\base\tsa_model.py:162: ValueWarning: No frequency information was provided, so infe
rred frequency D will be used.
  % freq, ValueWarning)
```

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                LMP   No. Observations:                  365
Model:             SARIMAX(1, 0, 0)x(1, 0, [1], 7)   Log Likelihood               -1332.904
Date:                            Wed, 20 Jul 2022   AIC                           2675.808
Time:                                    10:19:41   BIC                           2695.308
Sample:                                01-01-2019   HQIC                          2683.557
                                     - 12-31-2019
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0418      0.070      0.597      0.551      -0.095       0.179
ar.L1          0.8397      0.016     53.885      0.000       0.809       0.870
ar.S.L7        0.9935      0.010    101.126      0.000       0.974       1.013
ma.S.L7       -0.9458      0.042    -22.574      0.000      -1.028      -0.864
sigma2        85.1655      3.975     21.423      0.000      77.374      92.957
```

Seasonal ARMA Model



Basic Idea

Above we compared the forecast to the price we are forecasting in a casual way.

But we have a loss function, and it is through the measure of the loss function that we should evaluate the forecasts.

But what do we compute?

Basic Idea

Some reminders of the basics.

If we have random variables X_i with means μ, and data x_i from each of these random variables, then the sample mean can be used to estimate μ.

So if we have realizations of our loss function $L(f_i, y_i)$ we could take the sample average of these and this would be an ESTIMATE of the loss generated by the forecasting method.

For example consider

$$n^{-1}\sum_{i=1}^{n} L(f_i, y_i)$$

Basic Idea - Simplest Example

Forecast y_{T+1} having observed i.i.d. normal data $\{y_t\}_{t=1}^{T}$

Suppose we have squared (MSE) loss, and we might use
$\bar{y}_T = T^{-1} \sum_{t=1}^{T} y_t$ for μ.

Then
$E[y_{T+1} - \bar{y}_T]^2 =$

This is the actual out of sample loss.

Basic Idea - Simplest Example

Suppose we did this in sample as an evaluation

Then
$E[T^{-1}\sum_{t=1}^{T}(y_t - \bar{y}_T)^2] =$

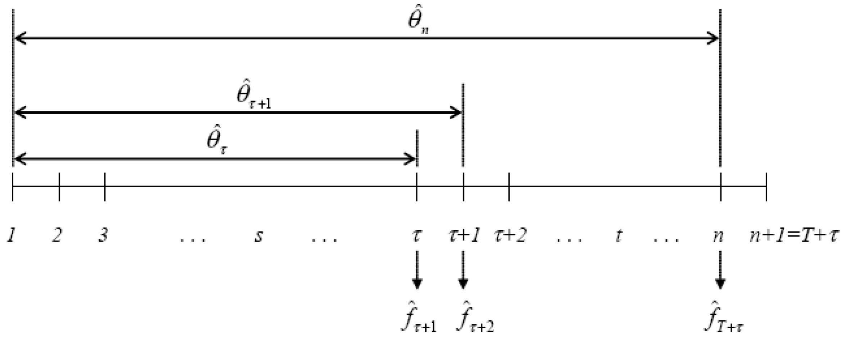
Where do we get the forecasts from?

There are a number of (very related) approaches

(1) Split the sample into two parts, one for estimation and another for evaluation.

(2) Make your forecast for each period over an evaluation component updating the model each time period.

Recursive Forecasts



Key Issues

(a) You need to realize here is that either approach results in an estimate of the loss from the procedure. You should be computing standard errors as you always do when computing a sample mean.

(b) Computing these standard errors is not necessarily straightforward - there are estimates inside the forecasting model and often this needs to be taken into account.

(c) Choosing the best forecast in this approach is really just an approach to model selection.