Predicting Patients – A Smith

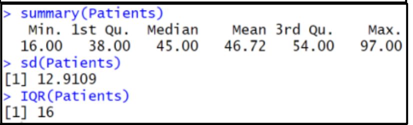
1.Introduction

We were given a time series dataset containing the daily number of patients who visited a specialist surgery from April 2015 to March 2019. We were tasked with fitting various models to explain the data, and asked for a prediction of at least 7 days for the surgery's management team to estimate the number of patients who will visit the surgery in the future.

For this problem, I plotted the time series and performed some rudimentary analysis using *R*, and decomposed the data to examine the underlying trend, seasonality and error. I then experimented on the data by fitting different models to explain the series and analysed the effectiveness of each one. These models include a baseline (*naïve*) model, extrapolation models (*Single Exponential Smoothing (SES), Holt Linear and Holt-Winters*), simple and multiple linear regression, and *ARIMAs*. After describing each model, I fitted it to our data, extracted the required forecast, and calculated the error statistics.

2. Numerical Summary

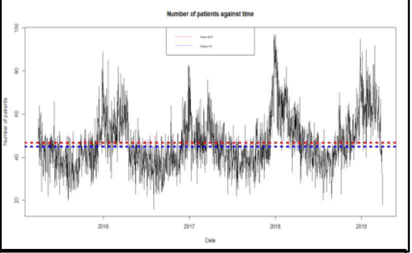
We find the maximum and minimum values, the interquartile range, mean, median and standard deviation:



```
> summary(Patients)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   38.00   45.00   46.72   54.00   97.00
> sd(Patients)
[1] 12.9109
> IQR(Patients)
[1] 16
```

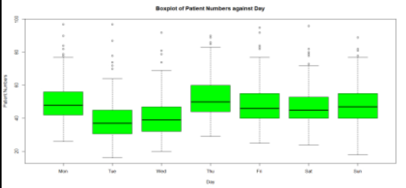
This gives us an indication of the spread of the data and the range of values it includes.

3. Graphical Summary

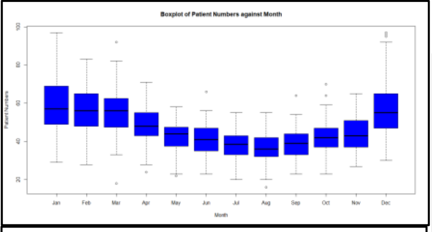


Plot of our dataset with red line representing overall mean (46.72) and blue line representing overall median value (45).

We begin to see that our data contains seasonality. Because we have data over a long time period, we break it down into smaller weeks and months and create boxplots:



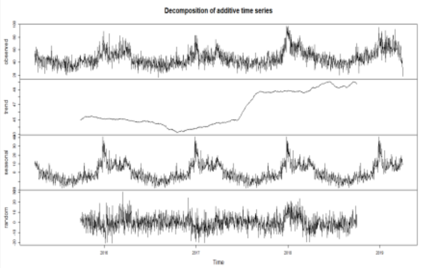
We notice that some days are busier than others



The surgery seems busier in the winter and quieter in late spring and summer, corresponding with flu season.

4. Decomposition

We must assess whether to use an *additive* or *multiplicative* decomposition. If the cycles of the decomposition appear additive, we use *additive decomposition*. If, however, seasonality increases as the mean increases, we favour *multiplicative decomposition*. I could not see any drastic increase in seasonality throughout the data I so opted for *additive decomposition*.

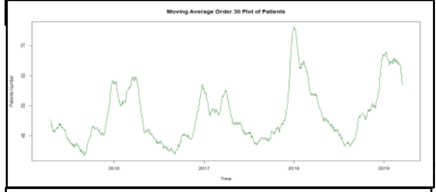


We see a general downward trend from mid-2015 until late 2016. There is a general upward trend from late 2016 until late 2018. The yearly seasonality appears consistent.

5. Moving Averages

Moving averages take the mean of a fixed subset and then shift forward to continue taking the mean of the next subset of the series. This smooths out fluctuations and extreme values and helps to smooth the data. For time series $Y_1, Y_2, \ldots, Y_n$ we define the moving average of period k (MA(k)):

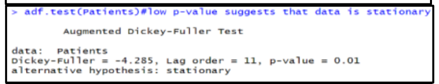
$$\frac{(Y_1+Y_2+\ldots+Y_k)}{k}, \frac{(Y_2+\ldots+Y_{k+1})}{k}, \frac{(Y_3+\ldots+Y_{k+2})}{k}, \ldots$$



MA(30) plot, with within-month variation smoothed.

6. Stationarity

A stationary time series has constant mean, variance and autocorrelation. If we can establish that our time series is stationary, it simplifies the prediction of future values as predicted values will reflect previous observations. We perform an *Augmented Dickey-Fuller Test* for stationarity.



```
> adf.test(Patients)#low p-value suggests that data is stationary

	Augmented Dickey-Fuller Test

data:  Patients
Dickey-Fuller = -4.285, Lag order = 11, p-value = 0.01
alternative hypothesis: stationary
```

The extremely low p-value in our Augmented Dickey-Fuller Test provides strong statistical evidence to reject our null hypothesis and consider that our data is stationary. This is somewhat surprising given that our decomposition appeared to show an upward trend.

7. Error Statistics

Mean squared error (MSE) value given by:

$$MSE = \frac{1}{n}\sum(y-\hat{y})^2$$

where n is the number of datapoints in our series, and $(y-\hat{y})^2$ is the square of the difference between our observed value y and our predicted value $\hat{y}$.

Mean absolute percentage error (MAPE) value:

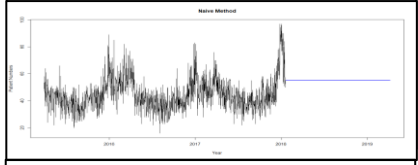
$$MAPE = \frac{100\%}{n}\sum\left|\frac{y-\hat{y}}{y}\right|$$

8. Naive Models

Before performing any modelling, I partitioned the entire *patient numbers* dataset into a training set (70%) and a test set (30%) against which the effectiveness of the model would be measured.

$$\hat{Y}_{t+1} = Y_t$$

The naïve model is the simplest form of model and is often used as a baseline for comparison. It assumes that the forecast for time $t+1$ is equal to the previous observed value.



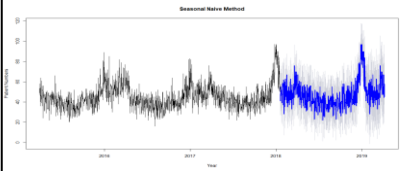
Our baseline (naïve) model predicts a constant number, 55.

One weakness of the naïve model is that it's predicted values are dependent on the last observation which may vary, giving extreme values.

Seasonal Naïve:

$$\hat{Y}_{t+1} = Y_{t-k} \text{ for seasonal lag } k$$

The seasonal naïve model predicts based on the previous observation of the same season.



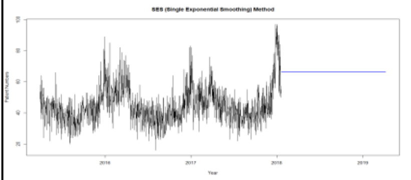
Seasonal naïve model predicted values. The blue section representing the predicted values, light blue represents the bounds for the 90% confidence intervals.

9. Extrapolation Models

We examine the single exponential smoothing (SES) model, this method was first proposed by Robert Goodell Brown:

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)$$

That is, our predicted value $\hat{Y}_{t+1}$ represents the previous period's predicted value plus a smoothing parameter for the level (α), based on the error in the previous prediction, where $0 \le \alpha \le 1$.



SES predicted values. This method predicts a constant number of patients for the future.

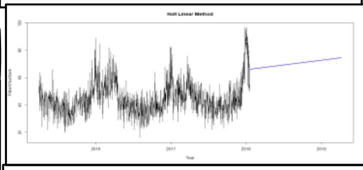
Holt linear method proposed by Charles C. Holt (1957), this method is most useful for data which displays a linear trend. The method uses the value of the time series at time t to estimate the base level of the time series (E_t) and the trend per time period (T_t). The function for Holt's linear method is given by:

$$\hat{Y}_{t+n} = E_t + nT_t$$

Where

$$E_t = \alpha Y_t + (1-\alpha)(E_{t-1} + T_{t-1})$$
$$T_t = \beta(E_t - E_{t-1}) + (1-\beta)T_{t-1}$$

For smoothing parameters α, β



Holt linear predicted values. This method predicts a linear increase in the number of patients without accounting for seasonality.

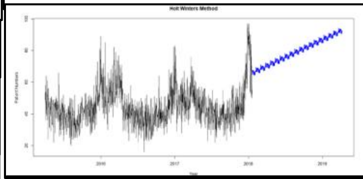
The Holt-Winters method is an extension of Holt's linear method but altered to allow for seasonality. The forecasting function is:

$$\hat{Y}_{t+n} = E_t + nT_t + S_{t+n-p}$$

Seasonality:

$$S_t = \gamma(Y_t - E_t) + (1-\gamma)S_{t-p}$$

Where γ represents the value of the smoothing parameter for the seasonal component. This is estimated by *R*.



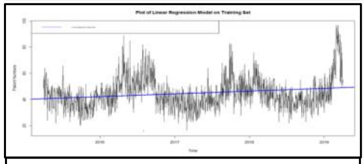
The prediction of future values resulting from the Holt-Winters method. Unfortunately, Holt-Winters is not very effective for time series with a high frequency. Hence the prediction shows a frequency of 24 and not 365. It still accounts for some seasonality.

10. Linear Regression

Linear regression functions have form:

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

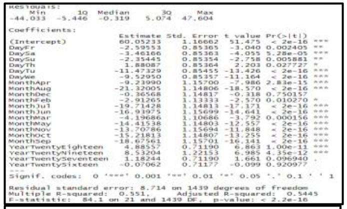
That is, y_t is equal to some constant β_0 plus some coefficient β_1 multiplied by a corresponding x_t value and added with a random error component ϵ_t. We assume that the error terms ϵ_t are i.i.d with a normal distribution and have mean 0.



We see that the linear regression model follows a linearly increasing trend. The values given for β_0 and β_1 can be extracted from R:

For multiple variables, we may use a more sophisticated method of regression, called multiple linear regression. The model takes general form:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_k x_{k,t} + \epsilon_t$$



```
Residuals:
    Min      1Q  Median      3Q     Max 
-44.033  -5.446  -0.319   5.074  47.604 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           60.05233    1.16662  51.475  < 2e-16 ***
DayFr                 -2.59553    0.85365  -3.040 0.002405 ** 
DaySa                 -3.46166    0.85363  -4.055 5.28e-05 ***
DaySu                 -2.35445    0.85354  -2.758 0.005881 ** 
DayTh                  1.88087    0.85364   2.203 0.027727 *  
DayTu                -11.47329    0.85455 -13.426  < 2e-16 ***
DayWe                 -9.52950    0.85357 -11.164  < 2e-16 ***
MonthApr              -9.23990    1.15700  -7.986 2.83e-15 ***
MonthAug             -21.32005    1.14806 -18.570  < 2e-16 ***
MonthDec              -0.36568    1.14817  -0.318 0.750157    
MonthFeb              -2.91265    1.13333  -2.570 0.010270 *  
MonthJul             -19.71428    1.14813 -17.171  < 2e-16 ***
MonthJun             -16.93975    1.15699 -14.641  < 2e-16 ***
MonthMar              -4.19686    1.10686  -3.792 0.000156 ***
MonthMay             -14.41538    1.14803 -12.557  < 2e-16 ***
MonthNov             -13.70786    1.15694 -11.848  < 2e-16 ***
MonthOct             -15.21813    1.14807 -13.255  < 2e-16 ***
MonthSep             -18.67561    1.15701 -16.141  < 2e-16 ***
YearTwentyEighteen     4.88557    0.71190   6.863 1.00e-11 ***
YearTwentyNineteen     8.53204    1.22153   6.985 4.35e-12 ***
YearTwentySeventeen    1.18244    0.71190   1.661 0.096940 .  
YearTwentySixteen     -0.07062    0.71177  -0.099 0.920977    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.714 on 1439 degrees of freedom
Multiple R-squared:  0.551,	Adjusted R-squared:  0.5445 
F-statistic:  84.1 on 21 and 1439 DF,  p-value: < 2.2e-16
```

Coefficients for each of the variables

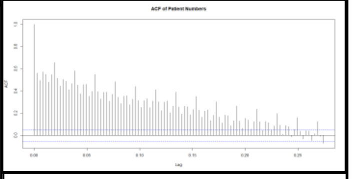
11. ARIMAs

An ARIMA model explains the time series based on its previous observations. The ARIMA model consists of three terms:

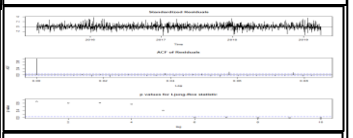
p represents the order of the Auto Regressive (AR) term

q represents the order of the Moving Average (MA) term

d represents the number of differencing required to achieve stationarity in the data.



Plot of ACF of our time series



Plot of diagnostics from recommended ARIMA(5,1,3) model. I experimented with ARIMA(1,1,1), ARIMA(5,1,3) and ARIMA(12,1,3).

12. Summaries

Model	**Mean Squared Error**	**Mean Absolute Error Percentage**	**Prediction for next 10 days**	**Ranking of effectiveness within model type**	**Overall ranking of effectiveness**
Baseline models:					
Naïve	145.11	22.51	Constant 18	1	8
Seasonal Naïve	160.99	22.68	64, 43, 52, 53, 58, 65, 47, 54, 58, 43	2	10
Extrapolation models:					
SES	84.92	17.43	Constant 43	2	5
Holt Linear	84.97	17.47	Constant 43	3	6
Holt-Winters	84.21	17.30	43, 42, 42, 40, 43, 41, 43, 41, 42, 43	1	4
Regression models:					
Simple Linear	154.62	22.32	Constant 53	2	9
Multiple Linear	74.79	14.67	Constant 61	1	**2**
ARIMAs					
ARIMA(1,1,1)	86.55	17.01	Constant 53	3	7
ARIMA(5,1,3)	77.80	15.81	Constant 53	2	**3**
ARIMA(12,1,3)	67.15	14.54	Constant 54	1	**1**

13. Conclusion

We examined 10 different types of time series models and applied each of these models to our *patient numbers* dataset. We also calculated certain measures of error associated with each model and used these values to judge the suitability of the model. Finally, for each model we have predicted the next 10 observations.

We started with a baseline (naïve) model, the level against which we should measure. Any model that outperforms the baseline should be considered. A number of extrapolation, regression and ARIMA models outperformed the naïve model. Simple linear regression models performed badly due to the data's high seasonality. ARIMA and multiple linear regression models fitted the data best, and yielded the lowest error statistics when the test set was compared against the model's predictions.

The best-ranking model was the ARIMA(12,1,3) model as it had the lowest MSE (mean square error). Its prediction was that the surgery would be visited by a consistent 54 people per day. This value did not reflect the weekly variation in the time series, so I propose, as our final prediction, subtracting the day coefficient from the ARIMA(12,1,3) prediction, combining our two best performing models to create a hybrid method:

01/04/2019 Monday = 54−0= 54
02/04/2019 Tuesday = 54−11= 43
03/04/2019 Wednesday = 54−9= 45
04/04/2019 Thursday 54+2= 56
05/04/2019 Friday 54−3= 51
06/04/2019 Saturday 54−3= 51
07/04/2019 Sunday 54−2= 52
08/04/2019 Monday = 54−0= 54
09/04/2019 Tuesday = 54−11= 43
10/04/2019 Wednesday = 54−9= 45

We would recommend to the surgery management team using an ARIMA(12,1,3) model but expect there to be weekly variation in the numbers of patients, consistent with variation seen in other weeks.