Modeling BBT to Predict Ovulation

Daniel Huang, SMU UbiComp Lab, June - August 2016

# Introduction

Predicting the period of highest fertility is the chief concern for couples who are trying to conceive. This window of time comprises the days on and directly following ovulation. Our goal is to predict the day on which ovulation occurs, given previous data.

# Background

## The Menstrual Cycle

A woman’s menstrual cycle comprises alternating phases of menstruation and ovulation.

During menstruation, which typically lasts from three to five days, the lining of the uterus is shed and estrogen levels in the bloodstream increase, prompting the production of a new lining. The secretion of menstrual blood is a way to detect the occurrence of menstruation visually.

Afterwards, ovulation, when a matured egg is released from the ovary, occurs and is accompanied by a release of progesterone into the bloodstream, which serves to facilitate the reproductive process. The resulting presence of elevated levels of progesterone in the bloodstream causes the body to warm considerably. Accordingly, a sudden increase in basal body temperature (BBT) after menstruation suggests that ovulation has occurred (a rule of thumb is given later). **The probability of fertilization is greatest on the day of ovulation and decreases each passing day directly thereafter.**

Thus, to determine the days with the highest chance of fertilization, it is necessary to find the day of ovulation.

## The Data

The data, collected by the University of Bologna in the 1960s, comprise various reproductive indicators and miscellaneous information about the female subjects’ menstrual cycles.

The data is split into three files: London, Alletaev, and Anomali. The London file contains the records of all cycles that are 100 days or fewer in length. The Anomali file contains the records of all cycles that are over 100 days in length. The Alletaev file contains information about the same cycles recorded in the London file, but with additional information about the woman’s background.

Each file shares the following data fields:

* DONNA - unique identifier (number) for each woman
* P\_SPEZZ - number of the group of consecutive cycles that the current cycle belongs to (1, 2, ...)
* P\_CICLO - number of that cycle within the group
* ANNO\_NAS - year of birth (last two digits), incomplete data
* DATA - date of the beginning of the current cycle (DD-MM-YY)
* T\_SPEZZ - total number of (groups of) consecutive cycles
* T\_CICLI - total number of cycles in the (group of) consecutive cycles that the current cycle belongs to
* QUALIFI - qualification (description) of the current cycle  
  0: Chart missing (information only available on total length)  
  1: Complete information  
  2: Presence on illness  
  3: Chart illegible (information only available on total length)  
  4: Cycle monophasic  
  5: Missing critical temperatures so that is impossible to identify BBT rise
* TIPOTEMP - temperature measured in Fahrenheit (1) or Celsius (2)
* L\_CICLO - length of the cycle in days
* L\_PREOV - number of days before BBT rise
* L\_PERIOD - length of period/menstruation

The London file contains the following additional data fields:

* TEMP1...TEMPn - temperature for day n of the cycle (n<=99)

The Alletaev file contains data about the subjects from both the London and Anomali data sets. The Alletaev data set purportedly contains additional data fields (listed below), but the data therein is scarce and mostly incomplete.

Additional data fields in the Alletaev file:

* ETA - age of subject (date of first cycle - birth date)
* GG\_MATR, MM\_MATR, AA\_MATR - day, month, year (last two digits) of marriage respectively
* GG\_UL\_EV, MM\_UL\_EV, AA\_UL\_EV - day, month, year of previous event (it is not known what an “event” is)
* FIGLI - number of children
* GG\_n\_EV, MM\_n\_EV, AA\_n\_EV - day, month, year of nth event (n<=4)

Note that the data field names are, by default, all in Italian. They have been renamed their English equivalents.

Of particular interest is how the data was recorded for the L\_PREOV column, since the only outward physical symptom of ovulation is the increase in BBT temperature. The widely accepted rule of thumb for calculating the day of ovulation based on a set of consecutive daily BBT measurements is the “three over six” rule.

The three over six rule states that the minimum of the three daily BBT measurements directly on and following the day of ovulation is greater than the maximum of the six daily BBT measurements directly before the day of ovulation.

In order to determine whether the data collectors had used this procedure to fill in the data for the length of pre-ovulatory period field, the three over six rule was applied to all the cycles in the London file, the only file to contain daily BBT measurements, and the calculated result for each cycle was compared to the recorded day.

|  |  |
| --- | --- |
| Different calculated value | 3445 |
| No pre-ovulatory period calculated | 3023 |
| Total differences | 6468 |
| Total number of cycles | 28357 |

The majority of the recorded days of ovulation matched the calculated day of ovulation. We tentatively conclude that the experimenters used the three over six rule and in the case that that rule did not yield a day of ovulation, a secondary rule was used.

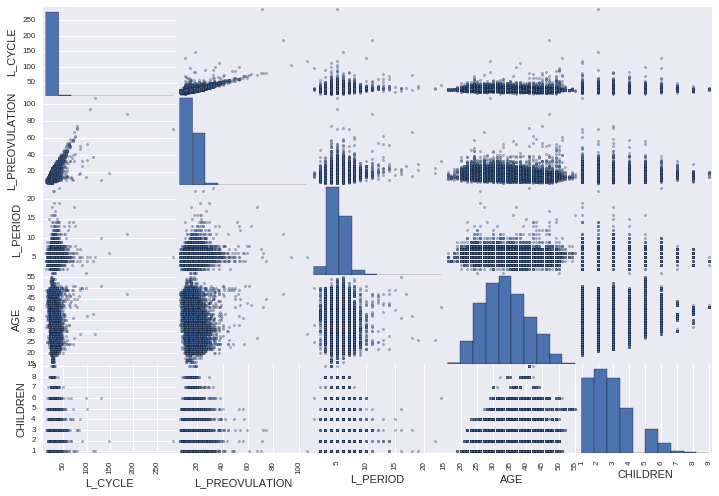
# Data Exploration

## Tools Used

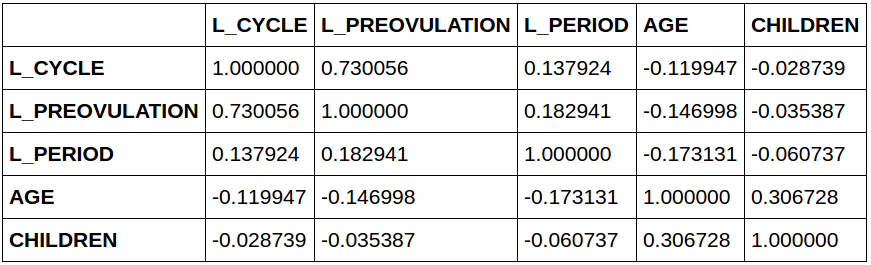
Pandas was used in conjunction with Matplotlib and Seaborn to process and visualize the data. Scikit-Learn was used to create models and perform machine learning tasks.

## Scatter and Correlation Matrices

A scatter matrix was first used:

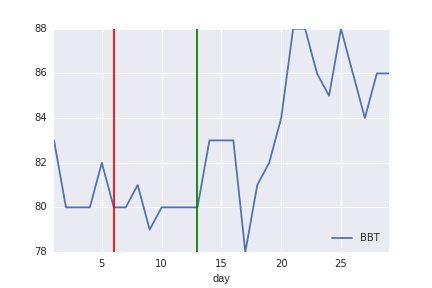


An apparent correlation between the length of the preovulatory period (L\_PREOVULATION) and length of the menstrual period (L\_PERIOD) was observed. The correlation matrix, shown below, indicates a correlation coefficient of 0.730056.



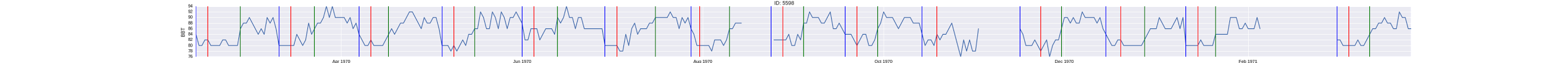
## BBT Graph

Since the London file gave daily BBT measurements for each cycle, it was possible to create BBT vs time graphs of each cycle. The red vertical line marks the end of the period and the green vertical line marks the end of the pre-ovulatory period (i.e. the day of ovulation). Many of these graphs showed an increase in BBT measurements on and after the day of ovulation.



Some BBT measurements were recorded as zero. This suggests that a measurement was not taken that day, so these fields were replaced with NaN values.

## Continuous BBT Graphs (Consecutive Cycles)

Since the daily BBT measurements for multiple consecutive cycles were recorded in the London file, it was possible to create a continuous BBT graph displaying the movement of the BBT measurements over several consecutive menstrual cycles for a specific subject.

# Predicting the Day of Ovulation

Machine learning was employed to model the BBT movements, with the ultimate goal of predicting the day of ovulation for a cycle. The set of features comprises the measurements of BBT for each day in the cycle and the label is the day of ovulation, represented in the data by the length of pre-ovulation.

## Imputation

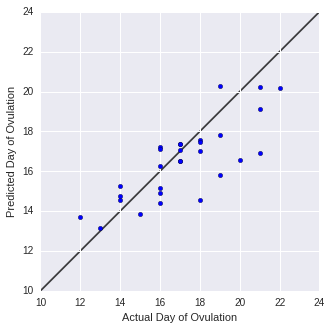
Since some features are missing in the data set, imputation was used to fill in the gaps in the BBT data. Originally, a missing BBT measurement was written in with the placeholder zero, but models built with imputed values seemed to be more robust. Sklearn’s Imputer class was used to replace missing values with the calculated median along the column.

## Linear Models

### One Participant

We first consider building a model specific to one single participant by using only data from the cycles recorded for that specific participant. This approach did not yield desirable results because most participants did not have a sufficient number of cycles recorded for this approach to be practical.

Participant 5080 had the greatest number of cycles, 112 cycles, of all the participants. A linear model (ordinary least squares) was built with the test set size being 20% of the entire data set. The variance score of the model on test data was 0.55815154860205007.

Clearly, this approach does not yield a useable model.

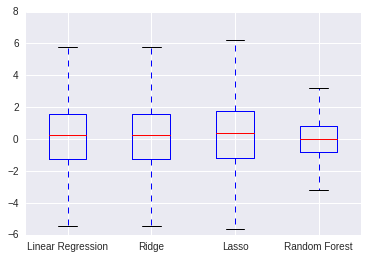
### Over many participants

In order to build a better model, the whole data set was used, leading the resulting model to be not specific to any particular participant. K-Fold cross-validation was used. Initially, the cycles of each participant was assigned to a unique fold, but this method proved to be too time consuming. Instead, each cycle was randomly assigned to one of 10 folds, which yielded similar results.

The data set allots space for 99 days of BBT measurements per cycle, but each cycle typically does not last for that long. All BBT columns past the number of days of the longest cycle in the group were dropped.

#### Analysis of the Residuals’ Statistics

Three linear models, ordinary least squares, ridge, and lasso, were built. Since they performed similarly, a more numeric method was used to measure performance: a boxplot of the residuals of each model was plotted and the statistics for the residuals were calculated. Additionally, a Random Forest model was also included.

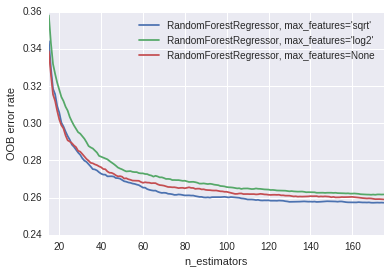


Linear Regression STD: 3.41036164681 Median: 0.26318359375  
Ridge STD: 3.36738670657 Median: 0.263861146368  
Lasso STD: 3.44618615727 Median: 0.369928157639  
Random Forest STD: 2.33115008185 Median: 0.0

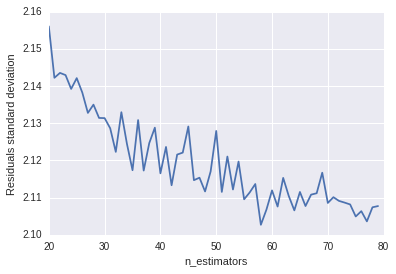
A smaller standard deviation is desired. This box plot suggests that a Random Forest model would be the most appropriate.

## Random Forest

The main concern with building a Random Forest model was the number of estimators to use. We wanted to know using which number of estimators would produce a model that has the smallest standard deviation for the residuals. Since the optimal number of estimators could potentially be very high and creating Random Forest models with 10-fold cross validation is very time consuming, we first found an approximate range for the optimal number of estimators by examining the out of bag (OOB) error rate. Random Forest models were built for a large range of n\_estimators, but they were trained over the entire data set, which is less resource intensive. The OOB error rate for each model was collected and graphed. It plateaued between 20 and 80 estimators, so the optimal number of estimators would likely fall in that range.



After narrowing in on that specific range, Random Forest models with 10-fold cross-validation were built.



From this graph, it appears that the optimal n\_estimators is between 50 and 60.

# Next Steps

* We did not examine the data from the Anomali file. This file is made up of outliers (cycle length greater than 100 days), so it may be useful in fine-tuning the model later.
* The method of imputation (of BBT measurements) used was not particularly sophisticated. A better way would be the following:

For missing BBT values in between two known values:

Impute the missing value by taking the mean of the two nearest known values on either sides.

For missing BBT values at the beginning of the data set:

Replace with the first non-missing value

For missing BBT values at the end of the data set:

Replace with the last non-missing value

* In this data set, it is not known whether pregnancy occurred for any cycle. If, after ovulation, the BBT values stay at a sustained elevated level for more than 18 days, pregnancy likely occurred. However, this deals more with predicting fertilization than with ovulation.
* Since the length of the period has a weak positive correlation with the day of ovulation, one could sort cycles into folds based on the length of the period.