Danmarks Tekniske Universitet

DTU

Assignment 1

Author

Emil Johannesen Haugstvedt - s211818
In collaberation with Louis Raskin and Olgeir Ingi Árnason.

October 6, 2021

Contents

Question 1.1: Plotting

I imported the data and plotted for driver 9. The plot can be seen in Figure 1.

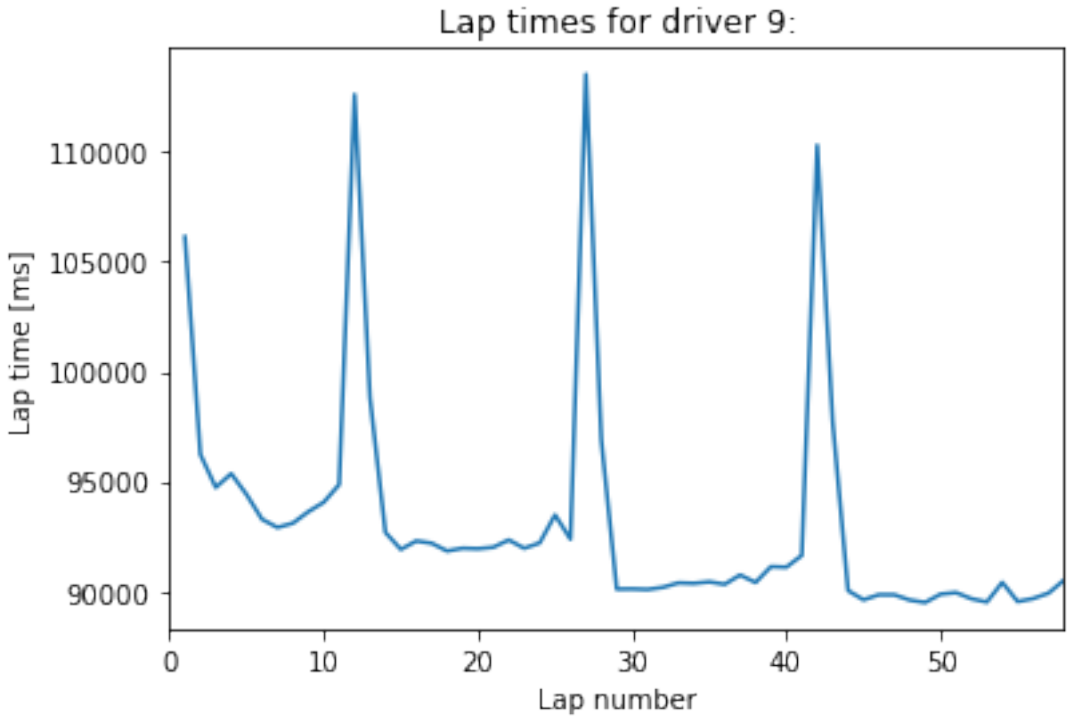

Figure 1: Plot showing the lap times for driver 9.

As seen from the plot there are four spikes in the plot. The data set provided the following information for the laps: lap time, lap number, driver ID, if the lap is a pitstop lap and if the lap is lagging a pitstop lap. After some investigation of the data I found that the first spike were related to the first lap taking longer than the other. And the rest of the spikes were related to pitstop laps and the laps lagging pitstop laps.

To illustrate this I plotted the data, now with dots representing the first lap, the pitstop laps and the laps lagging pit stops. This can be seen in Figure 2.

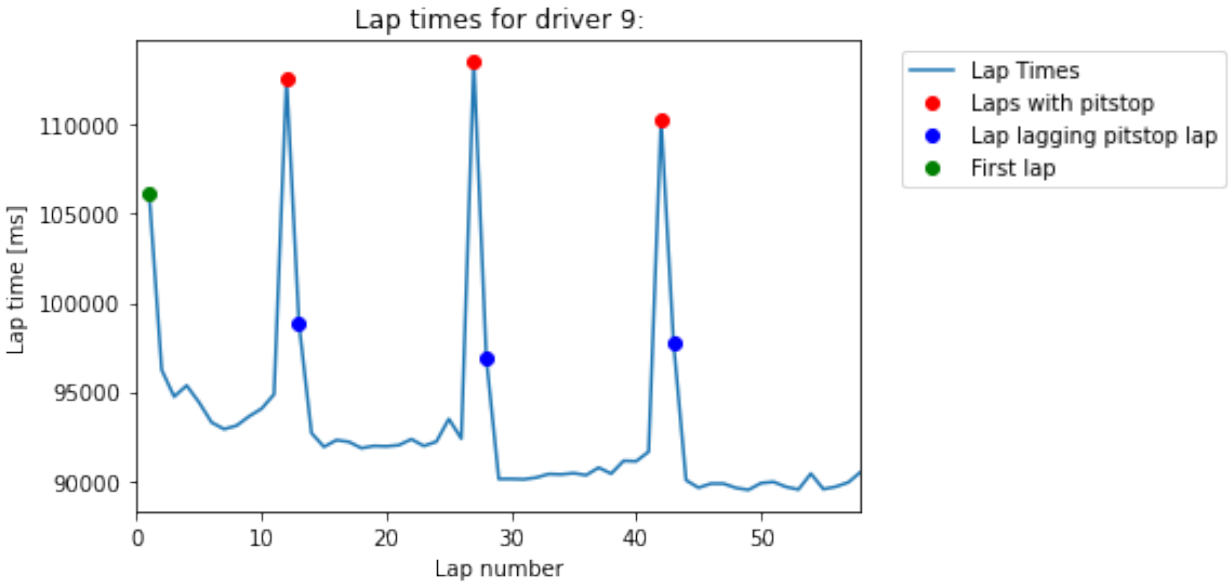

Figure 2: Plot of the lap times for driver 9 with first lap, pitstop laps and laps lagging pitstop laps marked.

From the above figure it is clear that the spikes in the lap times are due to the first lap, the laps with pitstops and the laps lagging pitstop laps taking longer time than a normal lap.

Question 1.2: Ordinary least squares

In this part we are trying to use ordinary least squares to make a general linear model for the lap times of the drivers.

1.2.1

The model created in this task should contain intercept for each individual driver, a linear realtionship to the lap number, and indicators to correct for the larges spikes in lap times due to that the first lap, pitstop lap and laps lagging pit stop laps has a longer lap time than normal laps. A model taking the individual intercepts and linear realtionship with the lap number into account are easy to formulate and will be on the following form (also given in the example).

$$lapTime_k = \boldsymbol{\alpha}_{driverId_k} + \boldsymbol{\beta} * lapNumber_k + \boldsymbol{\epsilon}_k \tag{1}$$

To include indicators to correct for the big spikes in the lap times I added a new term to the model, namely the correction term. The correction term are again made up of three terms, each with one parameter. The first parameter, $\boldsymbol{\delta}$, is related to the laps with pitstop. The second parameter, $\boldsymbol{\mu}$, is related to the laps lagging a pit stop lap. And the third parameter, $\boldsymbol{\phi}$, are related to the first lap of the drivers. The correction terms are given as:

$$correctionTerms = \boldsymbol{\delta} * isPitstop + \boldsymbol{\mu} * isLaggingPitstop + \boldsymbol{\phi} * isFirstLap. \tag{2}$$

As you can see the term also includes three vectors, $isPitstop$, $isLaggingPitstop$ and $isFirstLap$. All of these vectors are made up of ones and zeros, where the ones are indicating that a lap is respectively a pitstop, lagging a pitstop lap and a first lap of a driver.

The inclusion of this correction term gives a final general linear model on the following form:

$$lapTime_k = \boldsymbol{\alpha}_{driverId_k} + \boldsymbol{\beta} * lapNumber_k + correctionTerms_k + \boldsymbol{\epsilon}_k \tag{3}$$

1.2.2

For making the model we are to separate the data into a training set of the first 51 laps for each driver and a test set of the last seven laps. Some drivers had less laps completed than other, in that case I just used the laps they completed for training and ignored them when testing. See code for implementation.

1.2.3

The model are to be trained using ordinary least squares. By introducing θ as a vector containing all the parameters for the model, the parameters are given directly from the following equation.

$$\theta = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}Y \tag{4}$$

Here $\boldsymbol{X}$ is a matrix where each column represents one parameter. Column 1 to 21 represent the intercept for each driver, the last four columns represents, respectively, the lap number coefficient, pitstop coefficient, lap lagging pitstop coefficient and first lap coefficient. Y is a vector consisting of all the lap times for all drivers.

All the columns in $\boldsymbol{X}$ representing the driver specific intercept, $\boldsymbol{\alpha}$, are built up such that each row represent a lap, and that row contains a one if it is a lap performed by that driver and zero otherwise. The rest of the columns of $\boldsymbol{X}$ are not driver specific, thus the column for $\boldsymbol{\beta}$ just contains the lap numbers corresponding to the lap times. The last three columns has ones in the rows corresponding to a pitstop lap, a lap lagging a pitstop lap and a first lap, respectively.

We were also asked to provide an estimate for the standard deviation, $\hat{\sigma}$, for the model. The standard deviation can be estimated by as the square root of the variance of the model given by the following formula:

$$\hat{\sigma}^2 = \frac{(Y - \boldsymbol{X}\boldsymbol{\theta})^T(Y - \boldsymbol{X}\boldsymbol{\theta})}{N - p} \tag{5}$$

The Y term is the actual lap times and the $\boldsymbol{X}\boldsymbol{\theta}$ is the lap times computed by the model. N is number of samples and p the number of parameters. In words, this equation is the squared error divided by the number of samples, minus the number of parameters.

As a measure of uncertainty for the parameters, the standard deviation of each parameter can be calculated via the following formula:

$$Var(\theta) = \hat{\sigma}^2(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1} \tag{6}$$

By taking the square root of the variance of each parameter I obtain the standard deviation of the parameter. The above equation results in a pxp matrix, where p is the number of parameters. The variance of each parameter is found along the diagonal of the row and column corresponding to that parameter.

1.2.4

By calculating the above equations the following parameters are given for driver 9:

Parameter:	Value:	Standard deviation:
$\boldsymbol{\alpha}$ (intercept)	93527.72	369.56
$\boldsymbol{\beta}$ (lap number)	-76.77	5.69
$\boldsymbol{\delta}$ (pitstop)	22000.77	384.40
$\boldsymbol{\mu}$ (pitstop lag)	7475.00	384.42
$\boldsymbol{\phi}$ (first lap)	16310.25	546.53
$\hat{\boldsymbol{\sigma}}$	2398.56	-

Table 1: Table showing the parameters of the model found using ordinary least squares.

1.2.5

$\boldsymbol{\alpha}$ is the intercept of driver 9. This is the constant term for each driver and is the value where the estimation crosses the y axis in order to get the smallest error. This parameter contains information about the average lap time of the driver, and thus by comparing the $\boldsymbol{\alpha}$ value for different drivers one can find which driver performing best on average. Since they all have the same β, the driver with the lowest $\boldsymbol{\alpha}$ will have the best lap times.

The term, $\boldsymbol{\beta}$, is a estimation of how the lap times are changing throughout the race as a function of the lap number. This term is how the trend is for all drivers on average. Thus one can conclude that all the drivers on average see a decrease in their lap time throughout the race.

The pitstop coefficient, $\boldsymbol{\delta}$, is an estimate of much extra time a lap with pitstop takes, compared to a normal lap. Since this parameter are trained on all the drivers this will be the average extra time spent on a lap with pitstop. The same is for the two last parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$, but for the lap lagging the pitstop and the first lap respectively.

From Figure 2 it can be seen that a lap with pitstop takes the longest amount of time followed by the first lap, laps lagging pitstop laps and normal laps. This relationship can also be seen in the values of the last three parameters in Table 1. $\boldsymbol{\delta}$, the pitstop parameter, has the highest value and thus a pitstop lap has the highest extra lap time. The first lap has the second highest lap time and thus also the second highest parameter value, $\boldsymbol{\mu}$. Lastly the lap lagging a pitstop lap has the fastest lap time compared to a normal lap time, and thus also the smallest parameter values, $\boldsymbol{\phi}$.

1.2.6

The below table shows the a prediction of the last 7 laps with a 95% prediction interval as a measure of uncertainty. The 95% prediction interval is calculated by the following formula:

$$PredictionInterval = \hat{Y} \pm t_{\alpha/2,(N-p)}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{t+l}^T(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}_{t+l}} \tag{7}$$

Here $t_{\alpha/2,N-p}$ is the $\alpha/2$ quantile of the a student-t distribution with $N-p$ degrees of freedom. In this case α is 0.05, N is 933 and p is 25. The fact that the $\boldsymbol{x}_{t+l}$ vector becomes longer and longer when predicting further and further forward in time results in the prediction interval increasing.

Lap number:	Prediction:	5% prediction interval:
52	89535.86	± 6755.35
53	89459.09	± 6855.80
54	89382.33	± 6958.54
55	89305.56	± 7063.55
56	89228.79	± 7170.78
57	89152.03	± 7280.22
58	89075.26	± 7391.82

Table 2: Table showing the lap times predicted by the OLS model and a 5% prediction interval as a measure of uncertainty.

Figure 3: Plot showing the predicted lap times for the test set of driver 9 with upper and lower prediction interval. The same data as shown in Table 2

1.2.7

In Figure 4 the true lap times and the estimated lap times are shown. The plot also includes a 95% prediction interval.

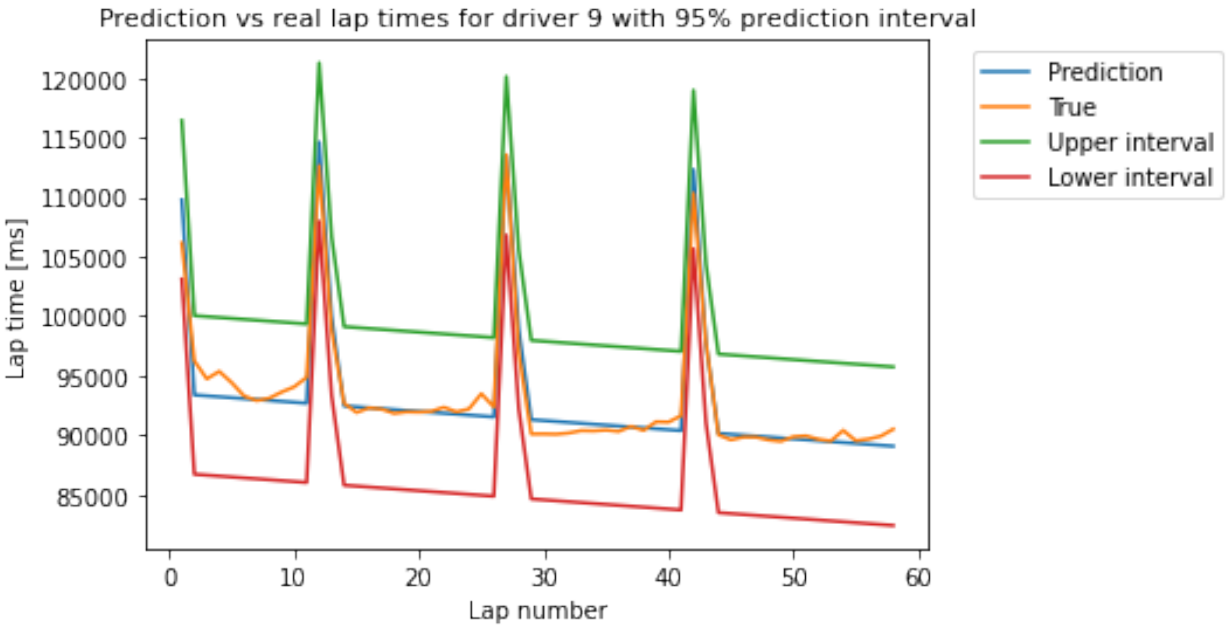

Figure 4: Plot showing predicted and real lap times for all laps for driver 9 together with a 95% prediction interval.

Question 1.3: Weighted least squares

In this part of the assignment we are trying to improve our predictions by adding a covariance matrix and thus use weighted least squares instead of ordinary least squares to make the model.

1.3.1

$\boldsymbol{\Sigma}_{i,j}$ will be all zeros for $i \neq j$. The assumption that all lap times for different drivers are independent yields that there are noe covariance between a lap of driver 9 and driver 10. Thus will all rows and columns in the covariance matrix where different laps of different intersects be equal to zero. This will result in a block diagonal covariance matrix with the each drivers covariance matrix, $\boldsymbol{\Sigma}_{i,i}$, as the inputs along the diagonal.

1.3.2

Now we are going to estimate the model parameters using weighted least squares. This is done in almost the same way as for ordinary least squares, the different is that the calculation of the parameters and variance includes some extra factors:

$$\theta = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}Y \tag{8}$$

$$\hat{\sigma}^2 = \frac{(Y - \boldsymbol{X}\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(Y - \boldsymbol{X}\boldsymbol{\theta})}{N - p} \tag{9}$$

In both the above equations the $\boldsymbol{\Sigma}$ is the covariance matrix.

1.3.3

Table 3 shows the parameters for the weighted least square models calculated using the above equations.

Parameter:	Value:	Standard deviation:
$\boldsymbol{\alpha}$	93489.16	417.1
$\boldsymbol{\beta}$	-75.73	6.42
$\boldsymbol{\delta}$	21638.08	433.85
$\boldsymbol{\mu}$	7384.06	433.88
$\boldsymbol{\phi}$	15654.99	616.83
$\hat{\boldsymbol{\sigma}}$	2707.12	-

Table 3: Table showing the parameter values for the model made with weighted least squares.

1.3.4

The variance of a lap is given as σ^2, where σ is the standard deviation. Thus increasing the standard deviation by a factor of γ will result in increasing the variance of that lap by a factor of γ^2. The covariance is given as $Cov(x, y) = Corr(x, y)\sigma_x\sigma_y$. Thus multiplying the standard deviation of either x or y by a factor of γ will result in increasing the covariance between x and y by a factor of γ.

Put mathematically, $V(lap\_time_k) = \gamma^2 * \sigma^2$, and $Cov(lap\_time_k, lap\_time_i) = \gamma * sigma_{i,j}$.

Thus the variance of laps containing a pitstop would be multiplied by a factor of γ^2. And the covariance of two laps, one which contains a pitstop, will increase by a factor of γ.

1.3.5

I multiplied both the columns and the rows containing a lap with a pitstop, a lap lagging a pitstop and a first lap by 1.5. This resulting in the covariance of these laps increasing by 1.5 and the variance of these laps increasing by $1.5^2 = 2.25$. Then I recalculated the parameters using weighted least squares.

1.3.6

The parameter values with the new covariance matrix including the introduction of γ are presented in Table 4.

Parameter:	Value:	Standard deviation:
$\boldsymbol{\alpha}$	93960.58	305.16
$\boldsymbol{\beta}$	-82.77	4.70
$\boldsymbol{\delta}$	21487.09	317.41
$\boldsymbol{\mu}$	7364.41	317.43
$\boldsymbol{\phi}$	15349.59	451.29
$\hat{\boldsymbol{\sigma}}$	1980.58	-

Table 4: Table showing the estimated parameter values for the model after introducing γ to update the covariance matrix.

1.4: Optimal covariance matrix

In this part we are trying to optimize the covariance matrix by using the relaxation algorithm to estimate the covariance parameters presented in 1.3.

1.4.1

We are using the relaxation algorithm to estimate the values of ρ, γ_{first_lap}, $\gamma_{pitstop}$ and $\gamma_{pitstop\_lag}$. The relaxation algorithm consists of first selecting an initial $\mathbf{\Sigma}$, than find parameter estimates with that covariance matrix, than consider the residuals and calculate a new $\mathbf{\Sigma}$ that reflects the structures of the residuals. This is done until the covariance matrix converges. In our case we will do this five times.

We assume a correlation matrix where the diagonal entries are matrices and all the non-diagonal elements are equal to zero. This is due to the assumption that the lap times of the different drivers are independent We also assume that the rows and columns of the diagonal entry matrices with laps related to laps with pitstops, laps lagging pitstop laps and first laps are multiplied by $\gamma_{pitstop}$, $\gamma_{pitstop\_lag}$ and γ_{first_lap} respectively. This is because we assume that special laps both has a higher variance and a bigger covariance with the other laps compared to normal laps. The covariance matrix will then take the following form:

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{1,1} & 0 & \dots & 0 \\ 0 & \Sigma_{2,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma_{I,I} \end{bmatrix}$$

Where each diagonal matrix, $\Sigma_{i,i}$, are on this form:

$$\Sigma_{i,i} = \begin{bmatrix} \gamma^2_{first\_lap}\rho_i^0 & \gamma_{first\_lap}\rho_i^1 & \gamma_{first\_lap}\rho_i^1 & \dots & \gamma_{first\_lap}\rho_i^{N_i} \\ \gamma_{first\_lap}\rho_i^1 & \rho_i^0 & \rho_i^1 & \dots & \rho_i^{N_i-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{first\_lap}\rho_i^{N_i} & \rho_i^{N_i-1} & \rho_i^{N_i-2} & \dots & \rho_i^0 \end{bmatrix}$$

Here the multiplication by γ_{first_lap} are shown for the first lap for one driver. For the other γ parameters the rows and corresponding to pitstop laps and laps after pitstop laps will by multiplied in the same way.

We know that the covariance of the residuals are given as $Cov(\varepsilon) = \sigma^2\Sigma = \varepsilon\varepsilon^T$. An approach could be to just use this matrix as our new covariance matrix, but this will introduce many more parameters. Thus the best approach is to exploit the structure of Σ and $\varepsilon\varepsilon^T$ to estimate the values for ρ, γ_{first_lap}, $\gamma_{pitstop}$ and $\gamma_{pitstop\_lag}$ that best reflects the covariance structure.

Some observations of the structure of $\Sigma_{i,i}$:

- The diagonal are all $\rho_i^0 = 1$. Thus the gamma values can be calculated by extracting the right diagonal elements of the $\varepsilon\varepsilon^T$ matrix.

- The diagonal above the main diagonal are all ρ^1 which is exactly the ρ value we are searching.

Some obervations on the structure of $\varepsilon\varepsilon^T$ matrix:

- The diagonal elements will all be $\varepsilon_i\varepsilon_i$, or $Var(\varepsilon_i)$.
- The elements in the diagonal above the main diagonal will all be $\varepsilon_i\varepsilon_{i+1}$, or the $Cov(\varepsilon_i, \varepsilon_{i+1})$

By combining these observations ρ can be estimated by finding the residuals for the normal laps, ε_{normal} and calculating $\frac{Cov(\varepsilon_{normal,i},\varepsilon_{normal,i+1})}{\sigma_{\varepsilon_{normal,i}}\sigma_{\varepsilon_{normal,i+1}}}$. Observing that the residuals are assumed to have the same variance I get the following expression for the estimation of ρ:

$$\rho = \frac{Cov(\varepsilon_{normal,t}, \varepsilon_{normal,t+1})}{\sigma^2} = Corr(\varepsilon_{normal,t}, \varepsilon_{normal,t+1}) \tag{10}$$

The other observations can be used for estimating the values of gamma. Noting that all the i diagonal elements of $\varepsilon\varepsilon^T$ are $Var(\varepsilon_i)$ an estimate of $\gamma^2_{pitstop}$ can be made by finding the ratio between the variance of the pitstop laps and the normal laps. And thus an estimate of $\gamma_{pitstop}$ are found by taking the square root of this ratio. The same approach can be done for the other γ values as well, resulting in the following equations:

$$\gamma_{pitstop} = \sqrt{\frac{Var(\varepsilon_{pitstop})}{Var(\varepsilon_{normal})}} \tag{11}$$

$$\gamma_{lag\_pitstop} = \sqrt{\frac{Var(\varepsilon_{lag\_pitstop})}{Var(\varepsilon_{normal})}} \tag{12}$$

$$\gamma_{first\_lap} = \sqrt{\frac{Var(\varepsilon_{first\_lap})}{Var(\varepsilon_{normal})}} \tag{13}$$

For each iteration of the relaxation algorithm the above equations are used to estimate values for ρ, γ_{first_lap}, $\gamma_{pitstop}$ and $\gamma_{pitstop\_lag}$. The new values are then used to remake the covariance matrix at each iteration.

1.4.2

Below I present the two tables, one includes the estimated values of ρ, γ_{first_lap}, $\gamma_{pitstop}$ and $\gamma_{pitstop\_lag}$ and the other presents the parameters for the model calculated with the newly estimated covariance matrix.

Parameter	Value:
ρ	0.36
$\gamma_{pitstop}$	4.57
γ_{pistop_lag}	1.51
γ_{first_lap}	12.24

Table 5: Table showing the estimated values for calculating the optimal covariance matrix

Parameter:	Value:	Standard deviation:
$\boldsymbol{\alpha}$	97852.29	169.09
$\boldsymbol{\beta}$	-83.44	2.6
$\boldsymbol{\delta}$	20843.22	175.88
$\boldsymbol{\mu}$	7420.16	175.89
$\boldsymbol{\phi}$	11315.31	250.05
$\hat{\boldsymbol{\sigma}}$	1097.42	-

Table 6: Table showing the estimated parameter values for the model after using the relaxation algorithm to estimate the covariance matrix

1.4.3

Lap number:	Prediction:	5% prediction interval:
52	89546.56	± 3811.04
53	89466.04	± 3867.71
54	89385.53	± 3925.67
55	89305.01	± 3984.91
56	89224.50	± 4045.41
57	89143.98	± 4107.15
58	89063.47	± 4170.10

Table 7: Table showing the predicted lap times for the last seven laps for driver 9 using weighted least squares with a covariance matrix estimated by the relaxation algorithm. It also shows a 95% prediction interval as a measure of uncertainty.

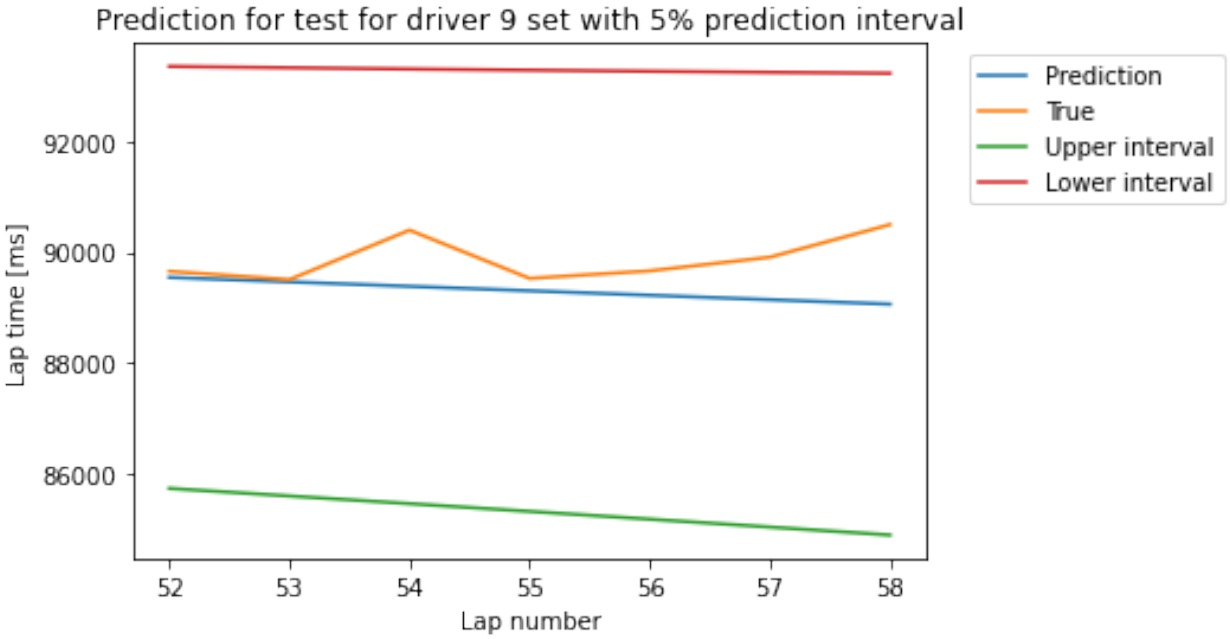

Figure 5: Plot showing the prediction for the last 7 laps for driver 9, including the true lap times and a 95% prediction interval. This is the same as shown in Table 7

1.4.4

In Figure 6 the prediction using the covariance matrix calculated by the relaxation algorithm are shown with the true lap times and a 95% prediction interval.

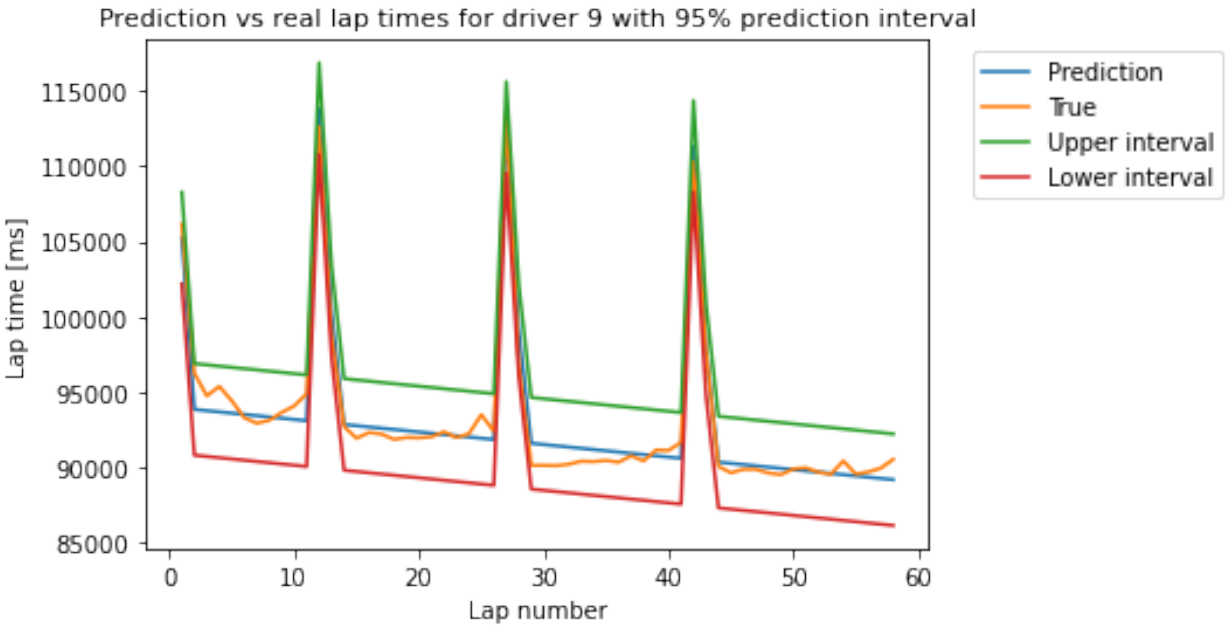

Figure 6: Plot showing the estimated lap times using the covariance matrix calculated with the relaxation algorithm, the true lap times and a prediction interval of 95% are also included.

1.4.5

The main difference between this model and the first OLS model is that this model takes into account the covariance between the different laps of each driver. The OLS model assumes the covariance matrix to be the identity matrix and thus that the different lap times has no connection with each other.

By introducing the covariance matrix we introduce the ability for some laps to co-vary. This means that the lap time may be affected by the lap time of prior laps. The reason why this works for this model is that some of the laps clearly are connected to what happens before. For example for driver 9 one can clearly see that the lap timed decreases after a pitstop, and thus there is a connection between the different laps.

1.5

I prefer the model with the covariance matrix trained by the relaxation algorithm. This is du to the fact that this model includes to covariance matrix which reflect the actual covariance in the model. This results in a much lower variance and thus a more accurate model. The standard deviation is actually more then halved from 2398 using OLS to 1079 with WLS and the optimized covariance matrix.

To improve the model I would have tried to estimate the covariance structure even more. This could be done by exploring the data even more and use more sophisticated algorithms to approximate the covariance matrix.

Also, in Formula 1, it is clear that there are some covariance between the lap times of the different drivers. Many drivers driving together may driver slower than drivers driving alone. Thus doing some approximation on the correlation between some of the laps for some of the drivers could be a good approach.

List of Figures

References

[1] Madsen H.(2007) *Time Series Analysis*, Chapman & Hall/CRC.