A Assignment 1

A.1 Plot Data

We are given a tabular data set "`BIL54`" describing the total number of registered motor driven vehicles in Denmark. The data given as a time series with monthly observations starting in January 2018.

We divide the data set into a *training set* containing data from January 2018 to December 2023, and a *test set* containing data from January 2024 to December 2024.

A.1.1 Construct time variable

The data set is indexed by an `ISO8601` date-time column, which we transform into a floating point time variable, x as follows:

$$x = \text{Year} + \frac{\text{Month}}{12} \qquad \begin{aligned} \text{Month} &\in \{0, 1, ..., 11\} \\ \text{Year} &\in \{2018, 2019, ..., 2024\} \end{aligned} \tag{1}$$

When referring to a vector of time values, we will follow the vector notation $\underline{\mathbf{x}}$. Other vectors are denoted similarly, while matrices are denoted with a double underline, $\underline{\underline{\mathbf{X}}}$.

It is important to distinguish between the time-like vector-valued variable $\underline{\mathbf{x}}$ and the design matrix $\underline{\underline{\mathbf{X}}}$, which will be introduced later.

A.1.2 Plotting observations

Using only the training data, we plot the total number of registered vehicles in Denmark as a function of time in Figure 1.

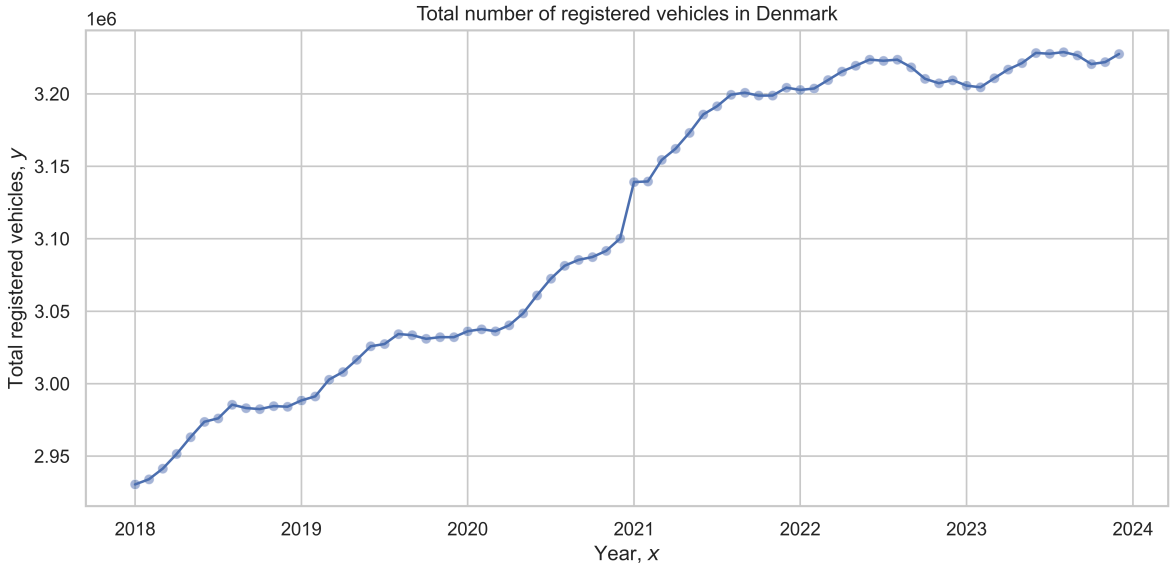

Figure 1: Training data describing the total number of registered vehicles in Denmark as a function of time.

A.2 Linear Trend Model

We are given the General Linear Model (GLM) in sloppy notation:

$$Y_t = \theta_i + \theta_2 \cdot x_t + \epsilon_t \tag{2}$$

A.2.1 Matrix Form

We immediately rewrite Eq. 2 as a matrix, observing the tensor notation outlined previously:

$$\underline{\mathbf{y}} = \underline{\underline{\mathbf{X}}}\underline{\boldsymbol{\theta}} + \underline{\boldsymbol{\epsilon}} \tag{3}$$

Where $\underline{\underline{\mathbf{X}}}$ denotes the *design matrix*, $\underline{\boldsymbol{\theta}}$ the *parameter vector*, and $\underline{\boldsymbol{\epsilon}}$ represents a stochastic noise term, $\underline{\boldsymbol{\epsilon}} \sim \mathcal{N}(0, \sigma^2)$.

In our case, the *design matrix* is constructed with an intercept θ_1 and a single trend parameter θ_2:

$$\underline{\underline{\mathbf{X}}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \tag{4}$$

Using the first 3 data points of the given data, we can construct the following model:

$$\begin{aligned} \underline{\mathbf{y}} &= \underline{\underline{\mathbf{X}}}\underline{\boldsymbol{\theta}} + \underline{\boldsymbol{\epsilon}} \\ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} &= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \\ \begin{bmatrix} 2930483 \\ 2934044 \\ 2941422 \end{bmatrix} &= \begin{bmatrix} 1 & 2018.000 \\ 1 & 2018.083 \\ 1 & 2018.167 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \end{aligned} \tag{5}$$

A.2.2 Parameter Estimation

We can estimate the parameters $\underline{\boldsymbol{\theta}}$ leveraging the *normal equations*.

This is done by minimizing the *residual sum of squares* (RSS) between the observations and the model predictions, that is:

$$\underline{\hat{\boldsymbol{\theta}}} = \underset{\theta}{\operatorname{argmin}} \left\| \underline{\mathbf{y}} - \underline{\underline{\mathbf{X}}}\underline{\boldsymbol{\theta}} \right\|^2 \tag{6}$$

We state without proof that the solution to the above optimization problem is given by the *normal equations*:

$$(6) \quad \Rightarrow \quad \underline{\underline{\mathbf{X}}}^T \underline{\mathbf{y}} = \underline{\underline{\mathbf{X}}}^T \underline{\underline{\mathbf{X}}}\underline{\hat{\boldsymbol{\theta}}} \quad \Leftrightarrow \quad \underline{\hat{\boldsymbol{\theta}}} = \left(\underline{\underline{\mathbf{X}}}^T \underline{\underline{\mathbf{X}}} \right)^{-1} \underline{\underline{\mathbf{X}}}^T \underline{\mathbf{y}} \tag{7}$$

Where $\underline{\underline{\mathbf{X}}}$ is assumed to be invertible.

We additionally consider the standard errors of the parameter estimates, $\underline{\hat{\boldsymbol{\theta}}}$.

Under the assumption that the observations are described by Eq. 2, that is, the data is drawn from a *Simple Linear Model* overlaid with a stochastic (i.i.d) noise term, we find that the residuals of the mdoel are exactly the noise term, $\underline{\boldsymbol{\epsilon}}$.

$$\underline{\boldsymbol{\epsilon}} = \underline{\mathbf{y}} - \underline{\hat{\mathbf{y}}} \tag{8}$$

Where

$$\underline{\hat{\mathbf{y}}} = \underline{\underline{\mathbf{X}}}\underline{\hat{\boldsymbol{\theta}}} \tag{9}$$

For the benefit of the reader, we reproduce the relationship between the residuals and the covariance matrix, $\underline{\underline{\boldsymbol{\Sigma}}}$ [1, p. 36-37]:

From Eq. 7, we have:

$$\begin{aligned}
\mathbb{E}\left[\hat{\underline{\theta}}\right] &= \mathbb{E}\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{y}\right] \\
&= \mathbb{E}\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T(\underline{\underline{X}}\underline{\theta} + \underline{\epsilon})\right] && (3) \\
&= \mathbb{E}\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)\underline{\theta}\right] && \underline{\epsilon} \sim \mathcal{N}(0, \sigma^2) \\
&= \underline{\theta}
\end{aligned} \tag{10}$$

We consider the following:

$$\begin{aligned}
\hat{\underline{\theta}} - \mathbb{E}\left[\hat{\underline{\theta}}\right] &= \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{y} - \underline{\theta} \\
&= \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T(\underline{\underline{X}}\underline{\theta} + \underline{\epsilon}) - \underline{\theta} \\
&= \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\epsilon}
\end{aligned} \tag{11}$$

We can now evaluate the covariance of the predicted parameters, $\hat{\underline{\theta}}$:

$$\begin{aligned}
\mathrm{Cov}\left[\hat{\underline{\theta}}\right] &= \mathbb{E}\left[\left(\hat{\underline{\theta}} - \mathbb{E}\left[\hat{\underline{\theta}}\right]\right)\left(\hat{\underline{\theta}} - \mathbb{E}\left[\hat{\underline{\theta}}\right]\right)^\intercal\right] \\
&= \mathbb{E}\left[\left(\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\epsilon}\right)\left(\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\epsilon}\right)^\intercal\right] \\
&= \mathbb{E}\left[\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{\epsilon}\underline{\epsilon}^T\underline{\underline{X}}\left(\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\right)^\intercal\right] \\
&= \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\mathbb{E}[\underline{\epsilon}\underline{\epsilon}^T]\underline{\underline{X}}\left(\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^\intercal\right)^{-1} \\
&= \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\ \mathrm{Var}[\underline{\epsilon}\underline{\epsilon}^T]\underline{\underline{X}}\left(\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^\intercal\right)^{-1} \\
&= \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\sigma^2\underline{\underline{X}}\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1} && \underline{\epsilon} \sim \mathcal{N}(0, \sigma^2) \\
&= \sigma^2\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}
\end{aligned} \tag{12}$$

Where we understand the variances of the predicted variables to be the diagonal elements of the covariance matrix:

$$\mathrm{Var}\left[\hat{\underline{\theta}}\right] = \mathrm{diag}\left(\mathrm{Cov}\left[\hat{\underline{\theta}}\right]\right) \tag{13}$$

Which gives the following standard errors for the parameter estimates:

$$\sigma_{\hat{\theta}_i} = \sqrt{\mathrm{Var}\left[\hat{\theta}_i\right]} = \sqrt{\sigma^2\left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}_i} \qquad i \in \{1, 2\} \tag{14}$$

Where σ is calculated using the *residual sum of squares* (RSS) with N and p denoting the number of rows and parameters in the design matrix $\underline{\underline{X}}$ respectively:

$$\sigma = \frac{\left\|\underline{y} - \underline{\underline{X}}\hat{\underline{\theta}}\right\|^2}{\sqrt{N - p}} \tag{15}$$

Computing these for the entire training dataset gives:

$$\begin{aligned}\hat{\theta}_1 &= (-110 \pm 4) \times 10^6 \\ \hat{\theta}_2 &= (56.1 \pm 1.8) \times 10^3\end{aligned} \tag{16}$$

Using these predicted parameters, we are able to produce an estimate of the vehicle registrations for the training dataset using the General Linear Model as seen in Figure 2.

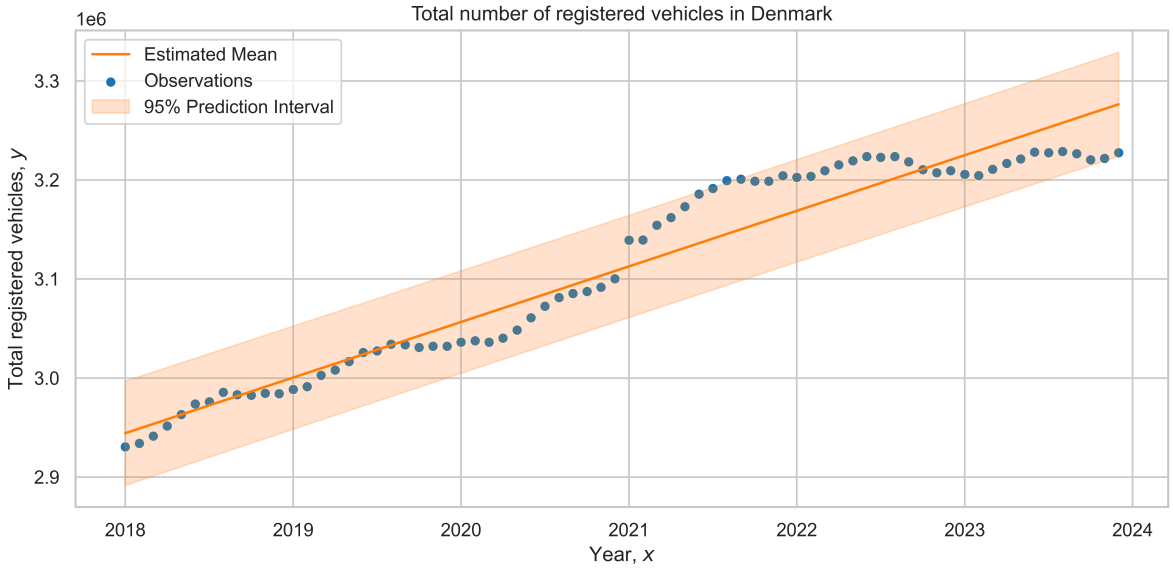

Figure 2: Estimation of vehicle registrations using a Linear Trend Model.

A.2.3 Prediction

We now wish to predict future vehicle registrations using our simple model. From Eq. 9, we see that doing so simply requires the construction of an appropriate design matrix, $\hat{\underline{\underline{\mathbf{X}}}}$. This may be constructed simply as:

$$\hat{\underline{\underline{\mathbf{X}}}}_i = \begin{bmatrix} 1 & 2024 + \frac{i-1}{12} \end{bmatrix} \qquad i \in \{1, 2, ..., 12\} \tag{17}$$

Where i indexes the rows of the design matrix.

Carrying out the forecasting as described by Eq. 9 along with appropriate estimation of the confidence interval of our prediction, we obtain Table 1. It should be noted that the estimation of the confidence interval is valid only in the case where the observations are described by the General Linear Model.

Time	Total Vehicles Registered	95% Confidence Interval, lower bound	95% Confidence Interval, upper bound
2024.000	3281153	3228504	3333803
2024.083	3285832	3233123	3338540
2024.167	3290511	3237741	3343280
2024.250	3295189	3242358	3348021
2024.333	3299868	3246973	3352764
2024.417	3304547	3251586	3357508
2024.500	3309225	3256198	3362253
2024.583	3313904	3260808	3367000
2024.667	3318583	3265417	3371749
2024.750	3323262	3270024	3376499
2024.833	3327940	3274630	3381250
2024.917	3332619	3279235	3386003

Table 1: 12 month forecast of vehicle registrations in Denmark.

A.2.4 Plot of Forecast

We can now present the forecasted data in Table 1 along with the training, test, and prediction data sets

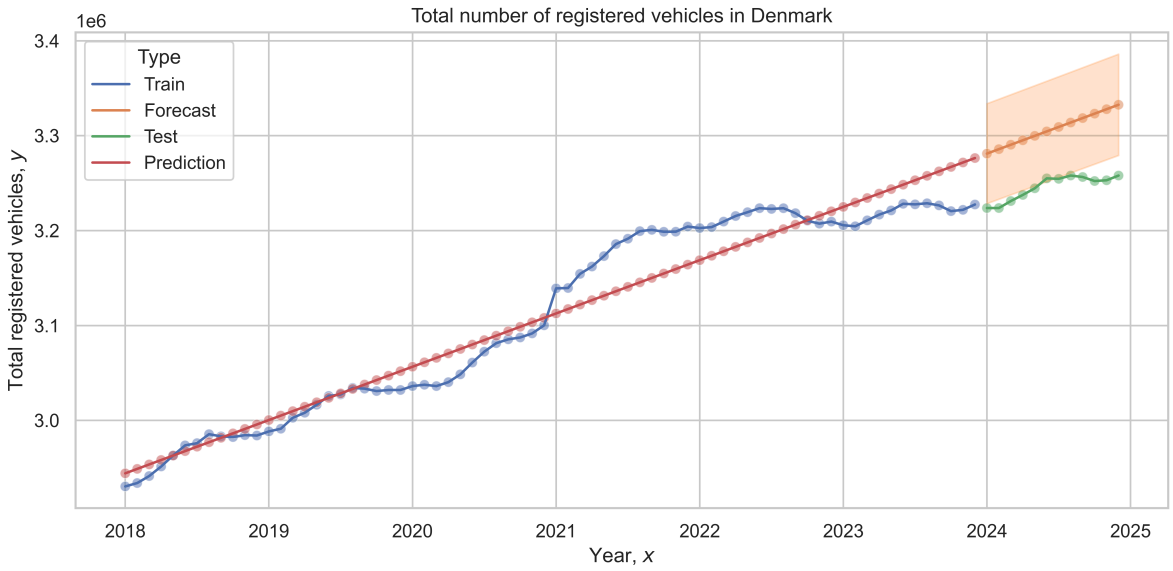

Figure 3: 12 month forecast of vehicle registrations in Denmark.

A.2.5 Commentary on Forecast

We find that the prediction is a relatively poor match against the test data, from which we understand that a Simple Linear Model is likely not an appropriate model for the data.

In particular, we observe significant local deviations from the model, which can be understood by considering the modelling domain. It is reasonable to assume that the number of registered vehicles will be influenced by market conditions, such as government subsidies, registration fees, and taxation schemes. Additionally, we would expect supply chain disruptions to have strongly influence the contemporary pricing and availability of vehicles.

A better model would likely incorporate these factors. A model that incorporates locality without apriori domain knowledge could be a *Weighted Least Squares* model with local weights.

A.2.6 Residual Analysis

In order to substantiate Section A.2.5, we perform a residual analysis on the prediction and forecasting data.

Recalling the assumptions of our model, we expect the residuals to be normally distributed around zero:

$$\underline{\mathbf{r}} = \underline{\mathbf{y}} - \underline{\hat{\mathbf{y}}} \sim \mathcal{N}(0, \sigma^2) \tag{18}$$

It is readily apparent in Figure 4 that the residuals are not normally distributed, nor do they average to zero:

$$\mathbb{E}[\underline{\mathbf{r}}] = -8739 \tag{19}$$

It should be noted, that the sum of the residuals is relatively close to zero, which is by construction as can be seen in Section A.2.2.

We conclude that the model is not appropriate to describe the data.

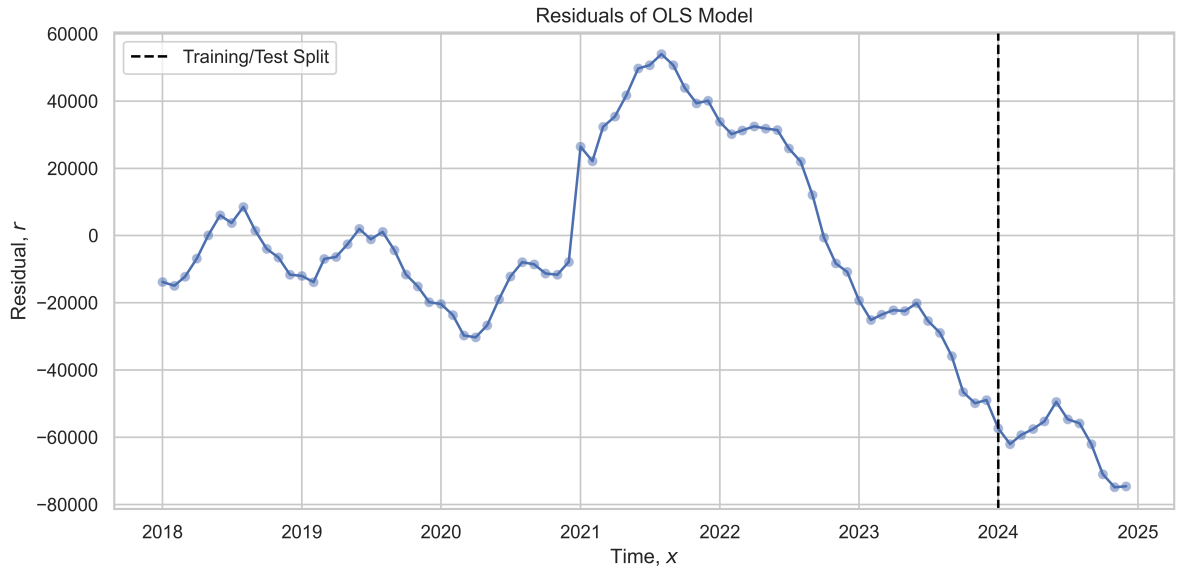

Figure 4: Residual analysis on prediction and forecasting of total vehicle registrations in Denmark. Dashed black line delineates the transition from prediction to forecasting.

A.3 WLS - Local Linear Trend Model

In order to mitigate some of the issues observed in the Linear Trend Model of Section A.2, we propose a *Weighted Least Squares* model with local weights.

That is, more recent observations are weighted higher than older observations.

This is facilitated by repurposing the variance-covariance matrix of the residuals. For Ordinary Least Squares (OLS), we assumed the residuals to be modelled as $\underline{\boldsymbol{\epsilon}} \sim \underline{\mathcal{N}}(0, \sigma^2)$. That is, we have imposed an assumption of *homoscedasticity* on the residuals $\underline{\boldsymbol{\epsilon}}$. In Weighted Least Squares (WLS) modelling, we relax this assumption and instead consider the variances of the residuals to be *heterogeneous* – that is, the residuals are *heteroscedastic*:

$$\underline{\boldsymbol{\epsilon}} \sim \underline{\mathcal{N}}(0, \underline{\boldsymbol{\sigma}}^2) \tag{20}$$

For Global Weighted Least Squares, one would ideally know the variances of the residuals a priori and simply weigh the residuals by the inverse of the variances.

We instead choose to *exploit* the Weighted Least Squares model to introduce *locality* in the estimation by letting the covariances of the residuals be modelled by $\sigma^2 \underline{\underline{\mathbf{W}}}$. We recall that the solution to the OLS problem (Eq. 7) was obtained by optimization of the *residual sum of squares* (RSS) between the observations and the model predictions. If we choose to weigh the residuals by their recency, we able to obtain an estimator that values recent information more highly than older information.

Referring to [1, p. 38-39], we state without proof that the *normal equations* for the WLS model are given by:

$$\left(\underline{\underline{\mathbf{X}}}^T \underline{\underline{\mathbf{W}}} \underline{\underline{\mathbf{X}}}\right) \underline{\hat{\boldsymbol{\theta}}} = \underline{\underline{\mathbf{X}}}^T \underline{\underline{\mathbf{W}}} \underline{\mathbf{y}} \tag{21}$$

Note that we have changed the notation slightly, denoting the weight matrix as $\underline{\underline{\mathbf{W}}}$ rather than perverting the sigma as is done in [1]. That is, we have made the substitution $\underline{\underline{\mathbf{W}}} = \underline{\underline{\boldsymbol{\Sigma}}}^{-1}$.

Assuming invertibility of $\underline{\underline{\mathbf{X}}}^T \underline{\underline{\mathbf{W}}} \underline{\underline{\mathbf{X}}}$, we obtain the parameter estimator:

$$\underline{\hat{\boldsymbol{\theta}}} = \left(\underline{\underline{\mathbf{X}}}^T \underline{\underline{\mathbf{W}}} \underline{\underline{\mathbf{X}}}\right)^{-1} \underline{\underline{\mathbf{X}}}^T \underline{\underline{\mathbf{W}}} \underline{\mathbf{y}} \tag{22}$$

We again follow the proof given in Section A.2.2, though noting that the variances of the residuals are now given by $\mathbb{E}\left[\underline{\boldsymbol{\epsilon}} \underline{\boldsymbol{\epsilon}}^T\right] = \sigma^2 \underline{\underline{\mathbf{W}}}$. As such, Eq. 12 becomes:

$$\mathrm{Cov}\left[\underline{\hat{\boldsymbol{\theta}}}\right] = \sigma^2 \left(\underline{\underline{\mathbf{X}}}^T \underline{\underline{\mathbf{W}}} \underline{\underline{\mathbf{X}}}\right)^{-1} \tag{23}$$

A typical choice of the weight matrix would be to have exponentially decaying weights:

$$\underline{\underline{\mathbf{W}}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} \lambda^{N-i} \qquad\qquad \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{24}$$

Where δ_{ij} is *Kroneckers delta*, λ is the *forgetting factor*, and N denotes the dimension of $\underline{\mathbf{y}}$.

A.3.1 Variance-Covariance and Weight Matrices

Encouraging the reader to grit their teeth we summarise the 'variance-covariance' matrices for the OLS and WLS models as follows:

$$\begin{aligned} \underline{\underline{\boldsymbol{\Sigma}}} &= \sigma^2 \underline{\underline{\mathbb{I}}} && && \text{OLS} \\ \underline{\underline{\mathbf{W}}} &= \sigma^2 \mathrm{diag}(\underline{\boldsymbol{\lambda}}) && \underline{\boldsymbol{\lambda}} = \begin{bmatrix} \lambda^{N-1} & \lambda^{N-2} & \dots & \lambda^0 \end{bmatrix} && \text{WLS} \end{aligned} \tag{25}$$

Stating these more explicitly, they become:

$$\begin{aligned} \underline{\underline{\boldsymbol{\Sigma}}} &= \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix} && && \text{OLS} \\ \underline{\underline{\mathbf{W}}} &= \begin{bmatrix} \sigma^2 \lambda^{N-1} & & & & \\ & \sigma^2 \lambda^{N-2} & & & \\ & & \ddots & & \\ & & & \sigma^2 \lambda^1 & \\ & & & & \sigma^2 \end{bmatrix} && && \text{WLS} \end{aligned} \tag{26}$$

For the rest of the analysis, we will consider the *forgetting factor* $\lambda \equiv 0.9$.

A.3.2 Weighting Regimen

Considering 72 time steps, we visualise the decay of the weights in Figure 5

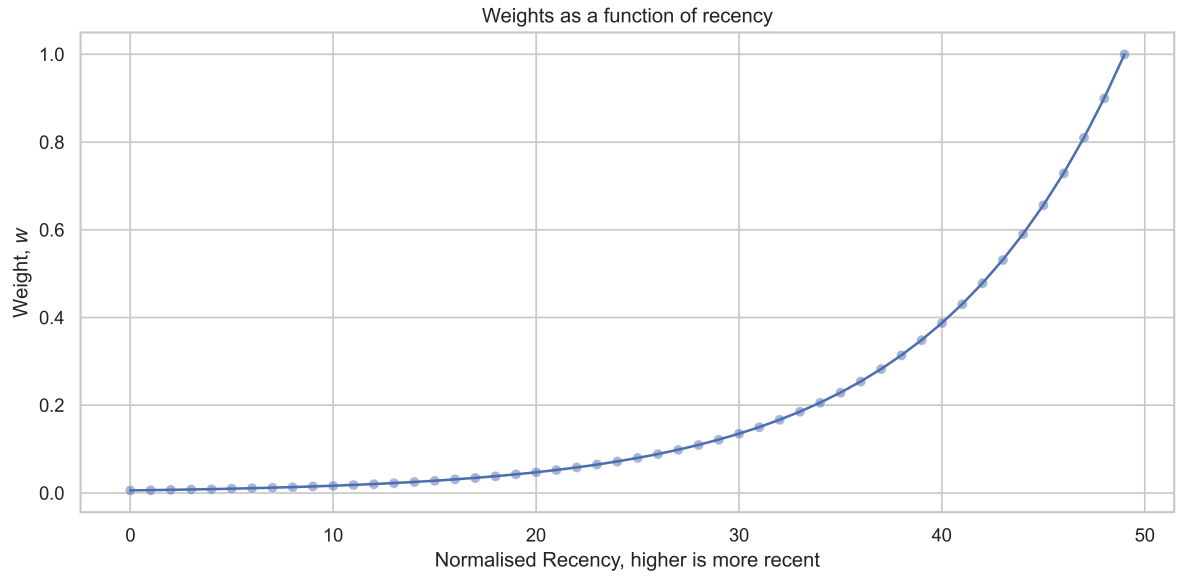

Figure 5: Decay of weights in the Weighted Least Squares model for $\lambda \equiv 0.9$.

A.3.3 Sums of weights

We find that the sum of the weights is given by:

$$\sum_{i=1}^{N} \lambda^{N-i} = \frac{1-\lambda^N}{1-\lambda} \tag{27}$$

For the limit $N \to \infty$ and $\lambda = 0.9$, we find:

$$\lim_{N\to\infty} \sum_{i=1}^{N} \lambda^{N-i} = 10 \tag{28}$$

We consider $N = 72$ and find the truncation gives:

$$\sum_{i=1}^{72} \lambda^{72-i} \approx 9.995 \qquad \text{WLS} \tag{29}$$

The corresponding sum for the OLS model is clearly not bounded and just becomes:

$$\sum_{i=1}^{N} 1 = N \qquad \text{OLS} \tag{30}$$

For $N = 72$, this simply becomes 72.

A.3.4 Parameter Estimation

We can now estimate the parameters of the WLS model using Eq. 22:

$$\begin{aligned} \hat{\theta}_1 &= (-1159200 \pm 600) \times 10^2 \\ \hat{\theta}_2 &= (5173.5 \pm 2.6) \times 10^2 \end{aligned} \tag{31}$$

Where the standard error is computed using Equation (3.43) in [1, p. 39].

A.3.5 Forecasting

We now perform a 12 month forecast using the WLS model in similar to fashion to the one carried out in Section A.2.3, as seen in Figure 6.

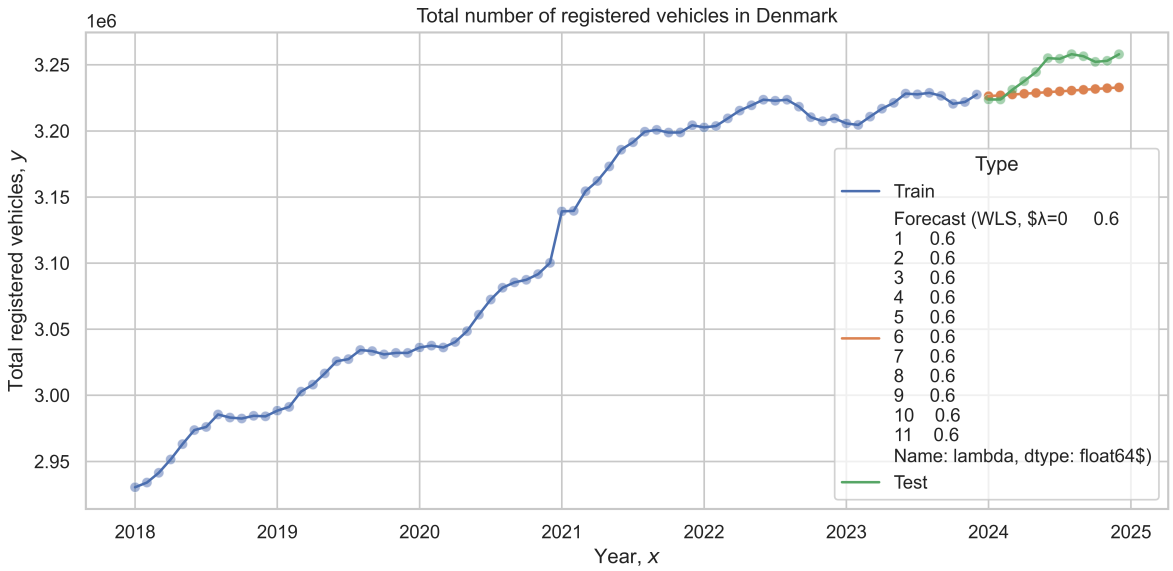

Figure 6: Comparison of the OLS and WLS ($\lambda = 0.9$) forecasts of total number of registered vehicles in Denmark.

It should be noted that the standard error estimate used in the WLS model has been corrected with respect to Eq. 15 for the truncating effect of the weights by calculating the degrees of freedom using the *total memory*, T of the model instead:

$$\sigma = \frac{\left\| \underline{\mathbf{y}} - \underline{\underline{\mathbf{X}}}\underline{\hat{\boldsymbol{\theta}}} \right\|^2}{\sqrt{T - p}} \tag{32}$$

Where the *total memory* is given by the sum of the weights:

$$T = \sum_{i=1}^{N} \lambda^{N-i} = \frac{1 - \lambda^N}{1 - \lambda} \tag{33}$$

For $\lambda < 1$ and $N \in \mathbb{Z}_+$ we observe that $T < N$, which agrees with intuition – a model with less memory *should* be less *confident* in its predictions, all other things being equal.

Even with this correction reducing the model confidence, we find an improved prediction and associated prediction interval as seen in Figure 6. This can be understood as the sum of squared residuals being reduced by the weighting scheme.

While the prediction does exhibit an improved fit to the test data when compared to the prediction carried out with an OLS model, it remains important to note that the model may not generally be expected to forecast well, particularly in the event of drastic changes in legislation or governance of motor driven vehicles.

Given a choice between the two models, short-term forecasts would be better served by the WLS model. It is not possible to confidently state which model would be preferable for long-term forecasting.

A.3.6 Different forgetting factors

Our choice of $\lambda = 0.9$ was somewhat arbitrary and wholly unjustified. In order to gauge which choice of forgetting factor may best fit out test data, we run the WLS analysis repeatedly. It should be noted that the choice of forgetting factor will inevitably be a qualitative matter and there is no guarantee that the λ that minimises the sum of squared residuals is necessarily a good or sensible choice for prediction for data sets that are not the current test set.

The outcome of the analysis can be seen in Figure 7, where we have omitted the confidence interval estimates for clarity.

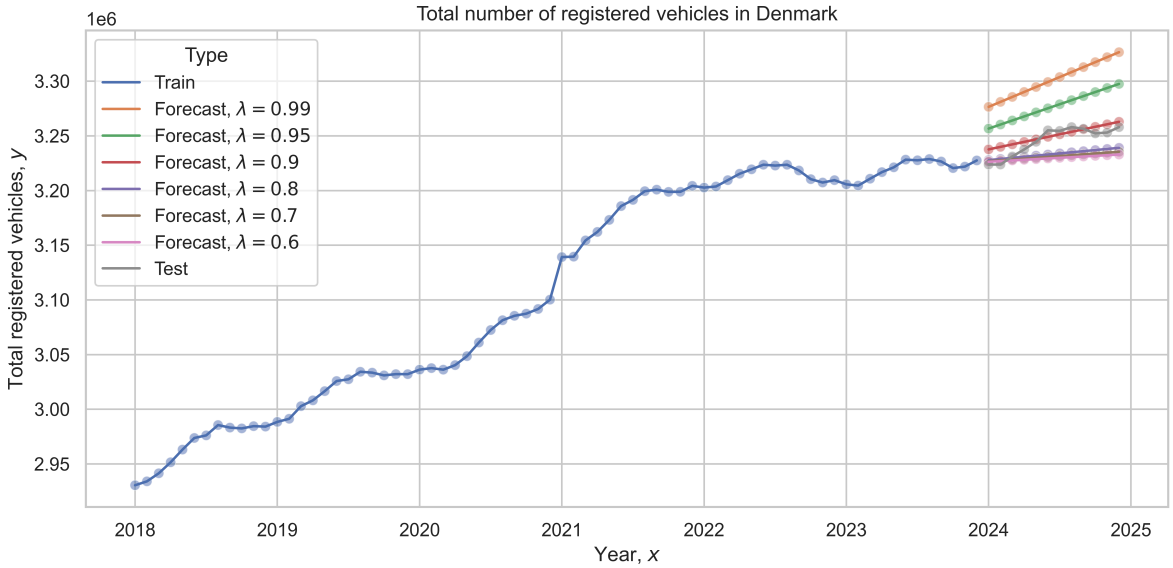

Figure 7: 12 month forecasts of total number of registered vehicles in Denmark using Local Weighted Least Squares with a variety of forgetting factors.

We find that $\lambda = 0.9$ is a reasonable choice for the test data set, though it is not necessarily the best choice for other data sets.

A.4 Recursive Estimation and Optimization of λ

Bibliography

[1] H. Madsen, *Time Series analysis*. Chapman & Hall/CRC, 2008.