Modeling Lateral Position of Driving

Mark Krysan

August 4, 2023

Introduction

In the field of driving simulator studies, there is a need for accruate and informative models for infrence. Many reserchers use the Standard Devation of Lane Position (SDLP) as a primary statistic for classifying drives. However, this statistic drastically reduces the information from driving simulators, resulting in a large loss of statisitcal power. There has not been much reserch into more complex and powerful models to handing simulator driving, but in 2010, Dawson et al. proposed a third order autoregressive time series model with a signed error as a new way to model time series [2]. This model was sucesfully able to distinguish bewtween drivers with and without Alzhimers. There are two part of this model, a projection component, which predicts the next position of a drive based off a linear combination of the past three positions, the signed error component, which models the probability of the projection component over or under predicting the true path is model using a logsitic regression.

Including the origonal paper [2], there have been two other papers written about this autoregressive model. In 2018, O'Shea et al. wrote a paper describing a new estimation technique, called the Modificed-Single Pass, replacing the orginal single-pass technique [4]. The Modfied Single Pass method was meant to account for bias in the estimation of parameters in the model, and was sucuessfully able to reduce some of the mean percent bias through a simulation study. Also, in 2019, Dawson et al. wrote another paper, investigating other estimation techniques aside from the Modified Single Pass [1]. These included ad hoc and liklihood based techniques, such as a Grid Search and Newton-Raphson algorithim. However, neither of these technqiues were able to out preform the Modified Single Pass in reducing the bias in parameter estimation.

In this paper, we investiate the model as decribed in [4], specfically the residuals of the projection component and their impact on bias innate in the model. To account for serial autocorrelation found in the signed error component, we add another predicitor to the logisitc model for the sign of the projection error. However, we also found that the distribution of the projection component errors was misspecified, which we belive adds to unnaccounted variablilty in simulation studies. We therefore deceided that for simulations, we should use bootstrapped projection residuals, rather than misspecifed errors. To test the impact of our two modifications, we ran a simulation study, selecting 1000 sets of parameters from a sample multivariate normal distribution defined by real data and generating 1000 drives each. We generated data as described by the Dawson paper, as well as generating data by also bootstrap the magnitude of the projection residuals in order to have more realsitic drives

Our simulation study was centered on determining if our two changes to the original model had any impact on the mean percent bais (MPB) in λ_1. We found that for the original data generation method, there was almost no change to the MPB model with and without lateral velocity. However, generating data with bootstrapped projection residuals, there was a noticeable decrease of about 1-2% in the MPB for larger values of λ.

Methods

Proposed Model

We now turn to a discription of the model proposed by Dawson et al. [2]. For a single simulated drive, let Y_t be the lane position at time $t = 0, 1, 2, \ldots, T$, where $Y_t > 0$ represents the left side of the lane, $Y_t < 0$ represents the right side of the lane, and $Y_t = 0$ represents the middle of the lane. In this model, the vector $[Y_{t-3}, Y_{t-2}, Y_{t-1}]$ is reparameterized to $[W_{1t}, W_{2t}, W_{3t}]$, where

$$\begin{aligned} W_{1t} &= Y_{t-1} \\ W_{2t} &= Y_{t-1} + [Y_{t-1} - Y_{t-3}]/2 \\ W_{3t} &= 3Y_{t-1} - 3Y_{t-2} + Y_{t-3}. \end{aligned}$$

W_{1t} represents a flat projection, W_{2t} a linear projection, and W_{3t} to a quadradic projection. These projections predict the next point in the time series as if the lateral position, veleocity, or acceleration is mantained, as estimated by the past three points. The vectors, $\mathbf{W}_1 = [W_{14}, W_{15}, \ldots, W_{1T}]^T$, and similar for $\mathbf{W}_2$ and $\mathbf{W}_3$, can then be calcualted based on all time positions. With this reparameterization defined, we now look at the thrid-order autogregressive time seris model proposed by Dawson [2], defined below:

$$Y_t = \beta_1 W_{1t} + \beta_2 W_{2t} + \beta_3 W_{3t} + |e_t| I_t, \text{ for } t > 3.$$

In this autoregressie model, the β_i's are constrained such that $\beta_1 + \beta_2 + \beta = 1$ and $0 \leq \beta_1, \beta_2, \beta_3 \leq 1$, so that each predicited position is a weighted average of the flat, linear, and quadradic projection, plus an error term.

This error term is the product of two componenets, a magnitude component and a sign compopnent. The magnitude componenet is assumed to be normally distribted with mean of zero and variance of σ_e^2, which is estimated from the model. The sign component is an indicator variable which represents the residual of the projection model: when $Y_t < \hat{Y}_t$, where $\hat{Y}_t$ is the predicited position at time t, $I_t = -1$ with probabtiliy p_t, and when $Y_t > \hat{Y}_t$, $I_t = 1$ with probability of $1 - p_t$. In [2], the functional form of p_t was characterized by the logisitc regtression model:

$$\log\left[\frac{p_t}{(1 - p_t)}\right] = \lambda_0 + \lambda_1 Y_{t-1},$$

where λ_0 is the intercept term and λ_1 is the slope term. In [2], λ_1 is intepreted as the 're-centering' parameter, where higher values indicate a higher tendancy to return to the center of the lane. Although in [2], there was some interpratation of how different values for β_1, β_2, and β_3 relate to a drivers preformance, [4] and this paper, foucs on λ_1 to determine how well a driver preforms.

Estimation

[4] decribed a new method for estimating the parameters of the model above, called the Modififed Single Pass, which was a improvment on the original Single Pass method proposed in [2] in regard to decreasing mean percent bias. The algorthim for MSP (Modified Single Pass) begins with calculating $\mathbf{W}_1, \mathbf{W}_2$, and $\mathbf{W}_3$ based on the first, second, and third lags of the position of a drivers path. The values of the $\hat{\beta}_i$'s are then calculated using OLS, without an intercept term. However, in order to ensure that $\hat{\beta}_1+\hat{\beta}_2+\hat{\beta} = 1$ and $0 \leq \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \leq 1$, the values of $\hat{\beta}_2$ and $\hat{\beta}_3$ are remapped to fit to these constrains, and $\hat{\beta}_1$ is set equal to $1-\hat{\beta}_3-\hat{\beta}_2$. An estimated path is calcualted then using the projection vectors $\mathbf{W}_1, \mathbf{W}_2$, and $\mathbf{W}_3$, along with their respective values of β. In symobols, we get,

$$\hat{\mathbf{Y}} = \hat{\beta}_1\mathbf{W}_1 + \hat{\beta}_2\mathbf{W}_2 + \hat{\beta}_3\mathbf{W}_3.$$

The reisual vector is then calcualted, with $\mathbf{Y} - \hat{\mathbf{Y}} = \hat{\mathbf{e}}$ and an indicator vector is then calcualted based on the sign of the reisudal, with $I_t = 1$ when $Y_t < \hat{Y}_t$ and $I_t = 0$ when $Y_t > \hat{Y}_t$. The probability of the resiual being negative is then estimated using logisitc regression, with

$$\log\left[\frac{\hat{p}_t}{(1-\hat{p}_t)}\right] = \hat{\lambda}_0 + \hat{\lambda}_1 Y_{t-1}.$$

[2] paper uses Firths ias reduction logisitc regression to overcome either quasi or complete seperation. We follow suit and use this method instead of the GLM method in R. The value of the varaince of the model, $\hat{\sigma}_e^2$, is calculated by traditional means, using the sum of squared differences. The equation is:

$$\hat{\sigma}^2 = \sum_{t=4}^{T} \frac{(Y_t - \hat{Y}_t)^2}{T^* - 3}, \text{ where } T^* = T - 3.$$

A detailed account of the psuedo-code for the estimation and data generation process that we used can be found in [4].

Analysis of Dawson Model

In our investiagtion of the model proposed by Dawson *et al.*, we found two impoortant inconsitiancies: a lack of normallity of the residuals of the linar model of the projections and correlation within the sign of those residuals. We belive that some of the bias that the Dawson group has incountered may be due to these two actributes. The distribution of the resiudals is important for simulation and infrence, such as confidence intervals. The correlation of the sign of the residuals is very importnat, since they are the response variables for the logisitic regression, and correlation in the observations lead to inflated standard errors, and, as we will show, an increase in bias.

Correlation of Residuals

In the Dawson model, the sign of the residual of the projection model is vitial in the estimation of λ_0 and λ_1. The sign of the residual indicates whether the model is over or under prediciting the path, and is associated with changes the drivier makes to the cars position in the lane. λ_1 is therefore given a significant amount of importantce for determining differences in dirvers. However, when plotting the drive path with color representing the sign of the residual, we found bands of the same color, with very little random/indpendnet points of differnt color (See figrue 1). These bands indicated that lane position is not the only predictor in the sign of the residual. It also points out that the sign of the reisuals of the projection model are not indepndent of each other. This correlation in the sign of the residuals has a considerable impact on the estimation of λ_0 and λ_1, as well as the interpretation of the model as a whole.

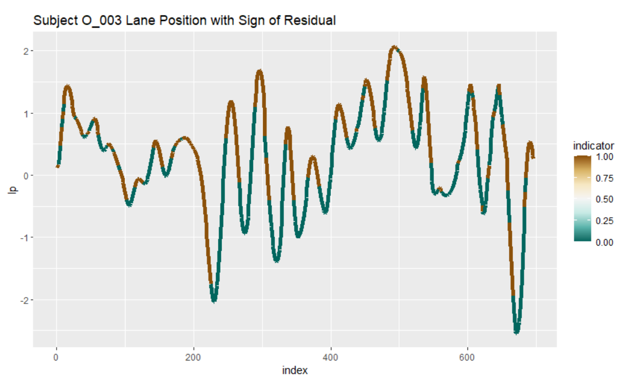

Figure 1

The correlation found in the sign of the residuals of the projection model results in corrolated residuals in the logsitic model, inflating standard error estimates for each coefficent. We also hypotheisized that the autocorrelation in the logisitic model residuals could be impacting the bias in the estimates. In maximum liklihood estimation of linear models, the bias of the ML estimate grows at the rate of $O(n^{-1})$ [3]. However, if the errors are correlated in a model, then the effective n, the number of independent errors, is dramaticially reduced. This provides a theoretical reason for why as correlation between resiudals increases, the bias also increases. We decided to emperically test this hypothesis for λ_1, since in was unclear what our "true" n would be for the logistic model. We therefore created multiple synthetic datasets of indicators, meant to simulate the residual indicator from the OLS model. For simplicity's sake, we modeled the drive as a sine curve with an amplitude of three and calulated the sign of the resiudal based on the same log-odds as the Dawson model. However, we added in a correlation parameter, which, for the specificed probability, set the ith value of the indicator equal to the $i-1$th value. We were therefore able to match the bands of the same value of the indicator seen in the true data (See Apendix, Figure 1). Using this parameter, we were then able to change the total amount of correltion in the sign of the residuals for a simulated drive. We found that as the correlation increased, from 0 correaltion to .9 correlation, the bias of the λ_0 and λ_1 estimates increased when the "correlation" probability reached .6, and was significantly higher at .9 (Figure 2). This emperically showed that autocorrealtion in the logsitic residuals leads to biased estimates of λ_1.

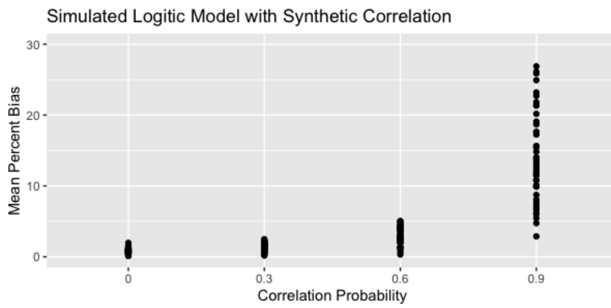

Figure 2

Distribution of Residuals

While studying the diagonsitic plots of our application of the Dawson model, we found that the residuals of the projection model were not normally distributed. We compared the histogram of residuals of the projection model to a normal distribution with a mean of 0 and a variance equal to that estimated from the model. There was a noticable difference between the normal residual histogram and true residual histogram. To be specific, the tails of the distribution of the true residuals were much heavier than the normally distributed residuals (See the histogram and QQ plot in Figure 3 and 4). The main effect of mispecification of the projection residuals is that the simulated drives will have a different type of variability in the driving path. In the case of our simulation study, by assuming that the projection residuals are normal, when they are not, there will be a larger discrepancy in the simulated drives for a given set of estimated parameters. We hypotheiszed that the incorrect distribution of the projection residuals may be such changes in the driving paths that parameter estimates are being biased. We hope that by boostrapping the residuals, the simulated driving paths are more consitiant and could produce less biased estiamtes.

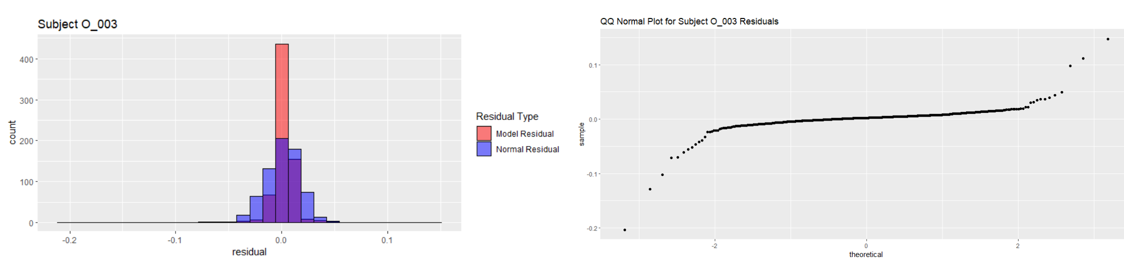

Figure: 3 and 4

Our Proposed Changes

For the two issues we found in the application of the Dawson model, we propose one change to the model, and one change to simulation technqiue. For the model, we propose adding in the lateral velocity to the logisitc model. In our analysis of the sign's of the projection component residuals, we found that there was a relationship bewteen the shape of the drive path and the sign of the reisual. For instance, when a path "takes an excursion" from the center of the lane, going to the left/positive values of Y_t, the sign of the residual is negative, until just before the local maximum of the parobola that the path makes, where the sign of the residuals becomes positive, until just before the next local minimum of the path (See Figure 2). This pattern, which appears is most driving path, indicates that we should include some measure of the shape of the driving path into the logistic model. We belive that by adding lateral velocity of the privous point to our logisitc model, the shape of the path will be incorperated into prediciton, leading to more accurate estimation. This model, called the LV model, has a functional form for the probability of a negative residual, p_t, as,

$$\log\left[\frac{p_t}{(1-p_t)}\right] = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 V_{t-1},$$

where V_{t-1} is the lateral velocity at time $t-1$. However, the parameter of interest will still be λ_1. Estimating the same real drive with the original model and the LV model, the parameter estimates for $\beta_1, \beta_2, \beta_3$ and σ are all these same, but λ_0 and λ_1 do have different values (Figure 3).

Real Drive Parameter Estimation

Original Model vs LV Model

Type	beta1	beta2	beta3	lambda0	lambda1	sigma
Original	0.05590029	0.2595608	0.6845389	-0.5679039	0.4567025	0.00929247
LV	0.05590029	0.2595608	0.6845389	-0.6112553	0.4991876	0.00929247

Figure 3

Our proposed change to the distribution of the residuals is to not assume any distributional form and to use non-parameteric boostrapping of real drives residuals for the projection residual magnitudes. Since the distribution of the errors has no impact on the model fitting, but only for simulation, we decided to turn to bootstrapping the residuals from real drives for simulation. We only bootstrap the magnitudes of the projection residuals, so as not to impact the sign of the residuals. Therefore, the only impact that this change will have is on data generation during our simulation study. Our intention is to remove any bias in the estimation of λ_1 that may have come from variation in simulated drives.

Simulation Study

Introduce Data

For our simulation study, we used driving simulator data from the University of Colorado, where about 100 participants took multiple drives in mutliple different scenarios. The dataset we worked with had 96 drivers each driving for about 3 minutes. The raw data was camptured in about 60 hz, but we averaged every 5 frames to get data in about 6 hz. After first estimating all 6 parameters for each driver using the orginal model, we recongized that the joint distirbution of all 6 parameters was approximatly multivaraite normal. Therefore, to increase our number of subjects, we created a MVN distribution based on a mean vector and variance-covariance matrix calcualted from the parameter estiamtes from our driving dataset.

Simulation Technique

We ran two simulations, both comapring the model proposed by Dawson and the model we have proposed in this paper, which adds in lateral velocity to the logsitic model. In the first simulation, we generate data according to the Dawson model (see [4] for an overview of generation technique), and in the second simulation, we instead bootstrap the magnitude of the residual from a true drive. In the simulation which bootstrapped the magnitudes of the residuals from true drives, we first found all true drives which had a variance within $\pm.005$ of the sampled variance. From these sets of parameters, we then found the true drive with the euclidian distiance of all 6 parameters closest to that of the sampled parameters. Then, instead of sampling the magnitudes of the error from a normal distribution with a set variance, we instead sampled from the reisudals of the true drive. For both simulations, we first sampled 1000 observations from the MVN distribution we described above. These sets of 6 parameters can each be though of as a specific drive. We then generate 1000 drives based on the respective data generation model and estimate the parameters with the Dawson model and our model. For each set of 6 parameters, we then calcualte the mean, vairance, bias, mean squared error, mean percent bias, and the approximate coverage rate for the Dawson model and our model.

Results

The results of our two simulation studies can be found in plots (whatever number plots). Looking at plot (showing difference between models), where we have the differnce in mean percent bias between the LV and Dawson model, with normal data generation, for values of λ_1 over 0, we see that the values hover around 0, getting closer and closer and λ_1 continues. However, when we generated data by bootstrapping true residuals, we see that for values of λ_1 greater than .25, the difference between the LV and Dawson model is negative, indicating that the MPB of the LV model was less than the Dawson model.

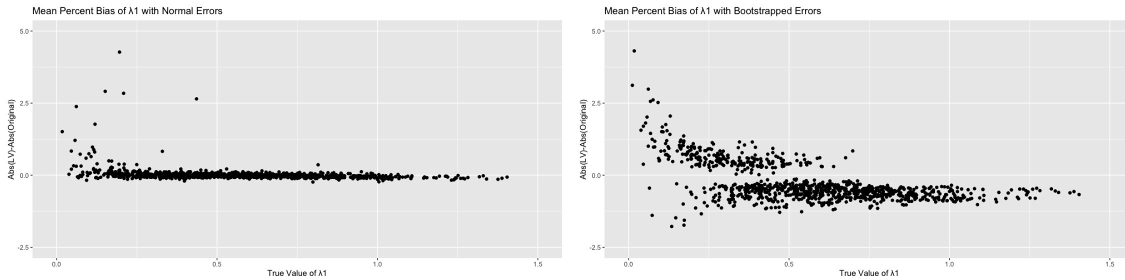

Figure: 4 and 5

Discussion

The time series model proposed by Dawson has clear problems in estimation of parameters of high relevance (λ_1). Our investigation into the effect of correlated errors in the logistic model points out that not only are the standard errors impacted by correlation, but as the correlation increases, the bias of the estimation also increases. This bias has been unaccounted for in previous work [2, 4, 1]. In our simulation study, we found that with the data generation method described in [4], including lateral velocity in the logsitic model had minimal impact to the mean percent bias. As λ_1 increaesed, the difference between the models with and without lateral velocity hovered around 0. However, when generating data with bootstrapped error magnitudes from true drives, there was a minimual, but noticiable decrease in mean percent bias when including lateral velocity in the logistic model. Since the data for this simulation was generated with bootstrapped residuals, the simulated drives had errors which more accuratly represented reality. Therefore, by making the simulated drives more similar to real drives and by including the lateral velocity into the logistic model, there is reduction in the bias of the estimation of λ_1. However, looking at the characteristics of the simulated drives with and without boostrapped residuals, it was very difficult to decern a differnece in statsitics such as range and SDLP. But, based on the histograms of normal residuals vs real residuals, we see that the real residuals have much heavier tails. It follows then that there may be high leaveraged points in the real residuals which are unaccounted for when generating data with normal residuals.

Although we were able to find a small decrease in bias for estimates of λ_1, there are still issues in the model which are unnacoutned for. First off, our inclustion of lateral velocity in the logistic model was based on the visual impact that the shape of a driving path has on the sign of residuals. However, in our investigations, we found that there was no decrase in the autocorrealtion of the residuals of the logsitic model when inclduign lateral velocity. We also noted that there is still a large varaition in drive paths for the same set of parameters, even when bootstrapping residuals from a real drive. Although we tried to account for this variablity in our simulation by running 1000 drives per set of parameters, it is difficult to understand how this variation impacts our simulation results.

Taking a step back, there is still autocorrealtion in the residuals of the logisitc model which are influening both the bias and standard error of the estimates, which impact infrence using this model. It is important that studies which use this model account for these two large impacts to hypothesis tests. There is also still bias in the estiamtes of the β parameters which is not accounted for in our LV model.

References

[1] Jeffrey D Dawson, Amy MJ O'Shea, and Joyee Ghosh. "Reducing High-Frequency Time Series Data in Driving Studies". In: ().

[2] Jeffrey D. Dawson et al. "Modeling lateral control in driving studies". In: *Accident Analysis & Prevention* 42.3 (2010). Assessing Safety with Driving Simulators, pp. 891–897. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2009.04.022. URL: https://www.sciencedirect.com/science/article/pii/S000145750900102X.

[3] Peter McCullagh and J Nelder. *Generalized linear models*. Routledge, 1989.

[4] Amy M. J. O'Shea and Jeffrey D. Dawson. "Modeling time series data with semi-reflective boundaries". In: *Journal of Applied Statistics* 46.9 (2019), pp. 1636–1648. DOI: 10.1080/02664763.2018.1561834. eprint: https://doi.org/10.1080/02664763.2018.1561834. URL: https://doi.org/10.1080/02664763.2018.1561834.

Appedix

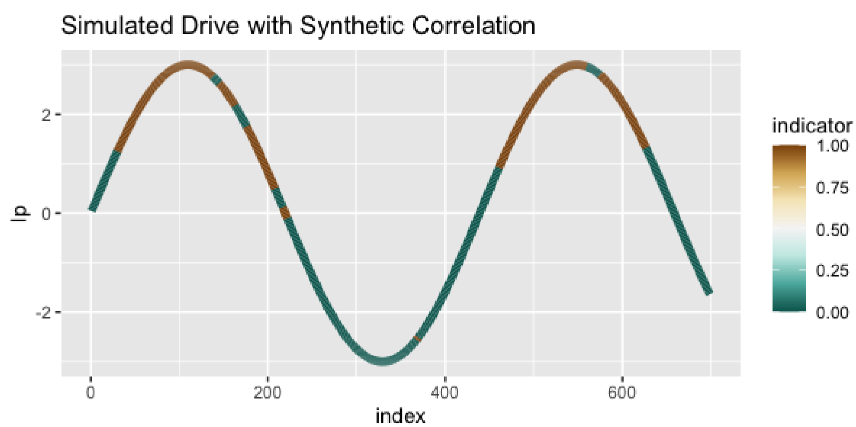

Figure: