Modeling temperature in chickens transport with R

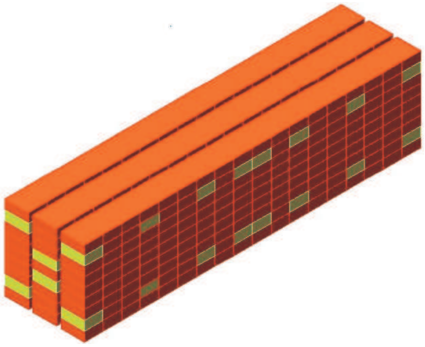
Bruno. H.F. Fonseca[*]
ESALQ-USP

Description

This project is to be conducted during the discipline Statistical modeling in R and will be developed using the R environment to fixed effects and mixed-effects linear model.
Keywords: Linear Model;Mixed-Effect ;Variable Selection;Residual Analysis.

1 Introduction

The NUPEA is a laboratory that study animal behavior, a project about transport of chickens is the source of the data for this work. The chickens are carried in trucks that have 486 boxes, which support of six to nine animals each. The boxes have a position in the truck, this position has three coordinates and the first coordinate is between 1 and 18 and correspond to backward direction of the truck, the second coordinate is between 1 and 3 and correspond to block of the boxes in the truck and the last coordinate is between 1 and 9 and correspond to height of the box in the truck. Therefore, there are 18x3x9 = 486 positions in the trucks. Follow illustrative picture:



In the picture above each yellow box is a example of the position that had the temperature monitored in someone trip.

The researchers have information about temperature by minute of sixteen trips in many positions of the trucks, but by data missing motive, will be used only 47 positions. The goal is to model the temperature in each position which are responses and can be explicated by some explanatory variables, for example, duration of the trips, season of the year, journey of the work and number of the chickens by boxes. Some different approaches are going to study, for example, linear model and linear mixed-effect model.

The second goal is to compare predictions between models of the different positions, the researchers think that places in the central block, more near of the driver and more low have higher temperature than others positions.

But the data were in different files and in different formats, then before of to model we needed to do manipulation of the data, this process was very complicated and needed of very time for to prepare the correct variables. The program is annexed.

2 Data

The research has eight trips by season, follow the data characteristics:

[*] Endereço: ESALQ, Universdade de São Paulo, Piracicaba, Brasil. E-mail: bhffonse@esalq.usp.br

Season	Date	Loggers	Duration in min	Journey	No of Chickens
$Winter$	2006/04/27	46	44	$12hto18h$	6
$Winter$	2006/05/13	45	25	$12hto18h$	7
$Winter$	2006/06/10	45	23	$06hto12h$	9
$Winter$	2006/06/25	46	124	$18hto24h$	7
$Winter$	2006/07/13	46	56	$06hto12h$	8
$Winter$	2006/07/13	46	46	$12hto18h$	7
$Winter$	2006/07/13	46	146	$18hto24h$	6
$Winter$	2006/08/31	46	36	$00hto06h$	8
$Summer$	2006/11/11	46	111	$06hto12h$	7
$Summer$	2006/11/30	41	71	$12hto18h$	8
$Summer$	2006/12/02	39	51	$06hto12h$	7
$Summer$	2006/12/08	40	26	$18hto24h$	8
$Summer$	2007/03/22	45	54	$18hto24h$	7
$Summer$	2007/03/23	44	111	$00hto06h$	7
$Summer$	2007/03/31	44	36	$12hto18h$	6
$Summer$	2007/04/01	45	146	$00hto06h$	7

Table 1: Brief Data Description

The number of the loggers is different in each trip because there are data missing. Other important characteristic is that there are trips with measurements of temperature in different positions of the general behavior, therefore there are positions with few observations, we will model only positions with more that 500 observations. O number of positions modeled is equal to 47, which has some data missing that are of the trips with out standard position, for example the first trip was the with date 2006/04/27 and this trip was considered like experimental, therefore only two positions this trip will be used. The duration of the trips is a random variable and corresponds to number of measurements of the temperature in each trip, for example in some trip the first measure of the temperature will have 0 for Duration, the second measure will have 1 for Duration... . This variable will be included in the random part of the models with mixed-effect. The researchers think that the temperature has a behavior that depends of the time in that were done the measure. We will try to model this behavior with inclusion of splines in the duration variable throughout of the trips.

Other covariate that could be considered is journey of work that has four levels and, for example, the researchers think that trips in hours between 12h and 18h have higher temperature than trips with hours between 0h and 6h.

The number de chickens by box is a covariate that has 4 levels and can to increase the temperature, trips with more chickens by boxes could have higher temperature. Again, the manipulation of the data for to begin the modeling were very hard and expend very time, the program is in annex.

3 Methodology

Like said before the objective is to model the temperature in each position of the trucks which will be modeled with two approaches linear model with fixed-effects and mixed-effects.

However, the temperature has a not constant behavior to length of the trips. The researchers find that it could begin high than it could be less in the middle of the trip and it could be higher in the final of the trips again. Therefore in all approaches of linear model we can adjust some polynomials of order 3 or 4 for to capture the behavior described. But we choose to smooth the fixed part of the model with a spline in the Duration of the trips.

The first approach is fixed-effects linear model that will be used only as initial analysis, however will be conducted many analysis of the models with respect to temperature behavior throughout the trips.

The other approach is linear mixed-effect model, were the random effect has an intercept and the Duration for each Trip.

Begging the modeling with the full models, therefore the start models always will have all possible covariate, will be done a variable selection, the analysis residuals and comparisons between positions.

For variable selection will be analyzed two estimators of the parameters the Maximum Likehood and Restrict Maximum Likehood, the first is used to test nested models with same fixed-effect, therefore will test the random effect, the second estimator is used to test the fixed-effects of nested models.

In the residuals analysis will be done residuals plots to see the behavior of the residuals with respect to normality and variability. Hypothesi tests about the assumptions will be done too.

With the objective of to compare the predictions of the different positions, will be showed plots with real temperature values and predictions throughout the trips.

All analyses will be made in the R ambient, and will be used the packages stats, MASS, nlme, splines, lattice, ASOR and Hmisc.

4 Results

This topic brings the results of the data modeling. Were choused positions to show some situations, the others positions have same or similar analysis.

4.1 Linear Model with fixed-effects

A initial data analysis can be done this approach. Follow the results for three positions:

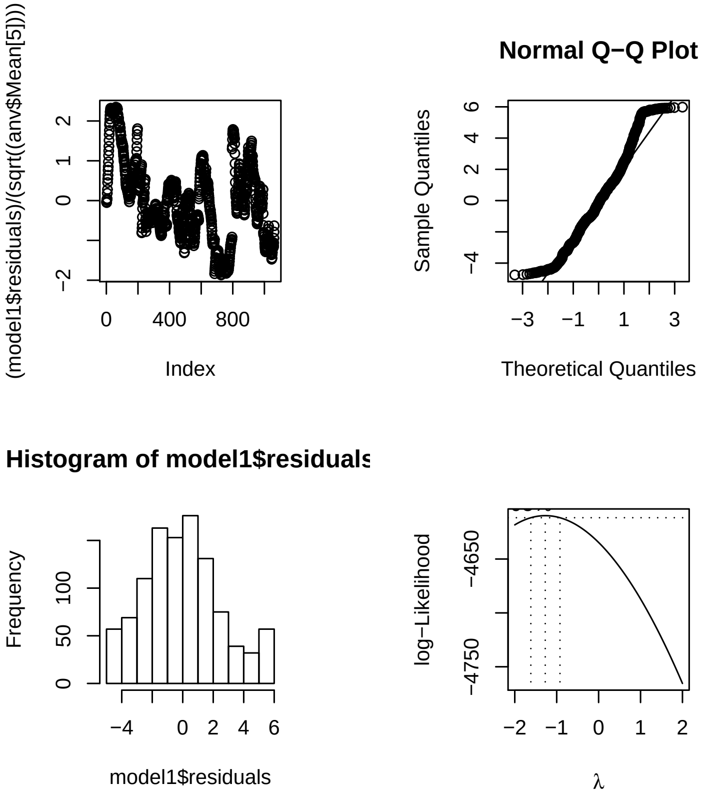
First Position 1 1 2:

The initial model is with all covariate, follow the results:

anova1.tex	Df	Sum Sq	Mean Sq	F value	Pr(¿F)
$ns(ETime, 4)$	4	1102.73	275.68	42.48	< 0.0001
$Season$	1	4866.37	4866.37	749.80	< 0.0001
$QDay$	3	1744.96	581.65	89.62	< 0.0001
No_per_cage	3	196.90	65.63	10.11	< 0.0001
$Residuals$	1050	6814.79	6.49		

Table 2: Analysis of Variance of the Temperature in the Position 1 1 2

The results above show that the model of the temperature in the position 1 1 2 have all parameter are significants, the calculate r-square statistic is equal to 0.53, therefore the variability explicated by model is not very high. The next step is to do residual analysis, follow the results:



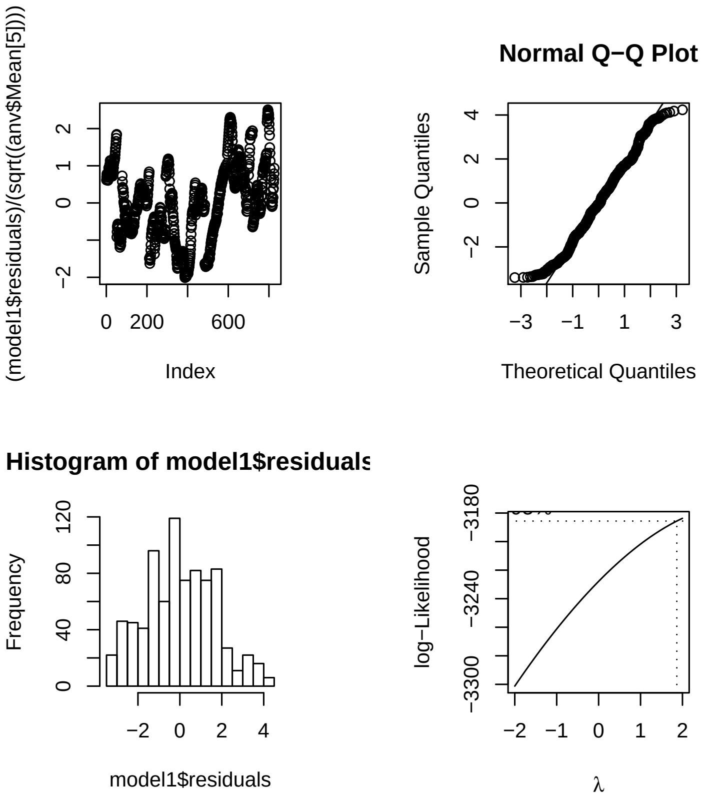
The first plot show that there are not standard residuals out of the interval of -3 to 3, but the behavior is not homogeny, there are trend because of each trip were there are trends throughout the time of the trips. The QQ-plot show that in the middle of the residuals there is a good approximation with the normal, but in the tails there are discrepant values. The histogram show a behavior asymptotic of the residuals. Therefore, there are strong evidences against the assumptions. The Shapiro-Wilk normality test show that reject the null hypothesi, therefore the residuals are not normality distributed. The Box-Cox transformation give a lambda equal to -1 which indicate a transformation of the type 1/temperature. But before of to transform the data, will be showed results of others positions because there are evidences of that in others positions will occur the same assumptions problems.

Second Position 9 2 6:

The same approach above will be used:

lm1.tex	Df	Sum Sq	Mean Sq	F value	Pr(¿F)
$ns(ETime, 4)$	4	1314.48	328.62	114.94	< 0.0001
$Season$	1	333.50	333.50	116.65	< 0.0001
$QDay$	3	2943.67	981.22	343.21	< 0.0001
No_per_cage	3	1584.48	528.16	184.74	< 0.0001
$Residuals$	814	2327.19	2.86		

Table 3: Analysis of Variance of the Temperature in the Position 9 2 6

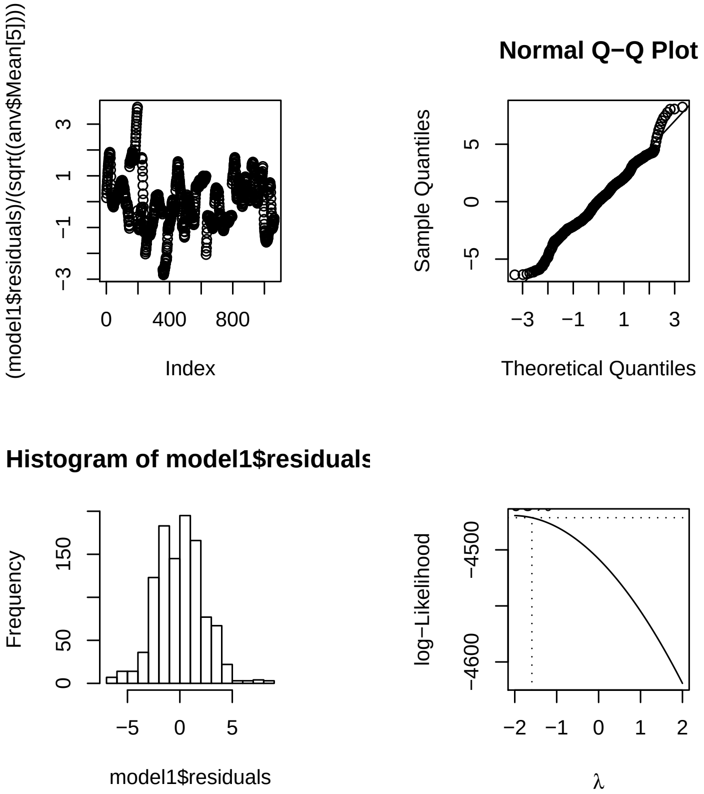


Again, all parameters are significants and the same assumptions problems occur, but now the r-square is larger. The Box-Cox transformation indicate a many values to lambda.

Third Position 10 3 8:

anova3.tex	Df	Sum Sq	Mean Sq	F value	Pr(¿F)
$ns(ETime, 4)$	4	3334.42	833.61	164.80	< 0.0001
$Season$	1	1213.91	1213.91	239.98	< 0.0001
$QDay$	3	927.47	309.16	61.12	< 0.0001
No_per_cage	3	52.38	17.46	3.45	0.02
$Residuals$	1050	5311.35	5.06		

Table 4: Analysis of Variance of the Temperature in the Position 10 3 8



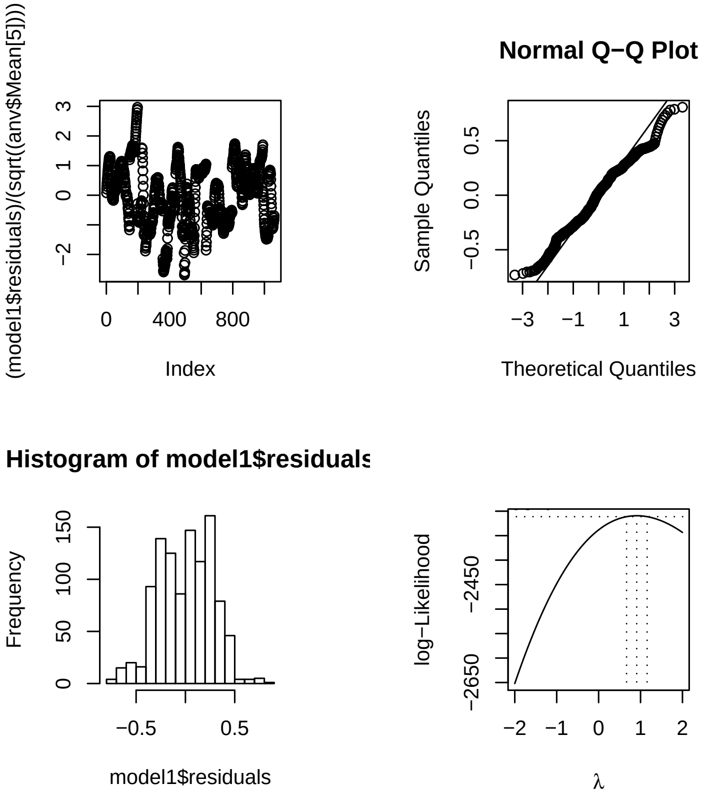
Occurred significant parameters again, but the assumptions problems were larger than above positions, the plots show a strange behavior were there are standard residuals out of the interval of -3 to 3, in the QQ-plot the tails are far of the normal, and Box-Cox transformation give many lambda again.

A observation is that for each position there is different problem with the assumptions, a propose solution is to transform the temperatures of all positions with the same transformation ln(temperature - 16.5), the constant included is next, but shorter, than the minimum of all temperature of all positions. Therefore, this transformation will be used in all next models.

Now will be showed again the modeling of the temperature of the position 10 3 8, but will be used a transformation above, follow the results:

anova4.tex	Df	Sum Sq	Mean Sq	F value	Pr(¿F)
$ns(ETime, 4)$	4	56.90	14.22	192.86	< 0.0001
$Season$	1	21.70	21.70	294.17	< 0.0001
$QDay$	3	11.43	3.81	51.67	< 0.0001
No_per_cage	3	0.54	0.18	2.46	0.06
$Residuals$	1050	77.44	0.07		

Table 5: Analysis of Variance of the Temperature Transformed in the Position 10 3 8



The results above show that with the transformation occurred a better approximation of the assumptions, were there are a little number of the standard residuals out of the interval, and the lambda indicated is equal to 1 which indicate nothing transformation.

However, using the transformation or not, this approach is not good for modeling the problem, therefore, follow in the next subsection mixed-effect approach.

4.2 Linear Model with mixed-effects

As the fixed-effects approach didn't adjust well the problem, now will be tried the mixed-effects approach were the start model will have response equal to log(temperature - 16.5) and the same fixed-effects than the last approach, however will be included the random effect that will have the intercept and the Duration (ETIME) of the trip by Trip.

For all positions will be showed the test about inclusion or not of the random effect, the variable selection, the graphical analysis and the residual analysis.

First Position 10 3 8:

anova5.tex	numDF	denDF	F-value	p-value
$(Intercept)$	1	1043	573.81	< 0.0001
$ns(ETime, 4)$	4	1043	241.17	< 0.0001
$Season$	1	7	1.73	0.23
$QDay$	3	7	0.55	0.66
$No_{p}er_{c}age$	3	7	0.67	0.60

Table 6: Analysis of Variance of the Temperature Transformed in the Position 10 3 8 with mixed-effects

The first test will be about the ETime random effects which will be extracted of the model and its significance will be tested with the Likehood Ratio, a important observation is that the estimator of the parameters is Restrict Maximum Likehood that test the random effect only, follow the results:

The results above show that the random effect is significant to model and it must to continue in the all analysis. Therefore the inclusion this term is helping to explicate the variability of the temperature of this position.

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
1	16	-1022.47	-943.17	527.24			
2	14	-607.69	-538.30	317.85	$1vs2$	418.78	< 0.0001

Table 7: Likehood Ratio Test For Random Effect in the Position 10 3 8

Now will be showed the first test about the fixed-effect. Will be used again the Likehood Ratio to test the significance of the Number of the chickens by box (No_per_cage), a observation is that now will be used Maximum Likehood Estimator, only thus we can to test the fixed-effect, follow the results:

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
1	16	-1049.69	-970.20	540.84			
2	13	-1054.48	-989.90	540.24	$1vs2$	1.21	0.75

Table 8: Likehood Ratio Test For No_per_cage Effect in the Position 10 3 8

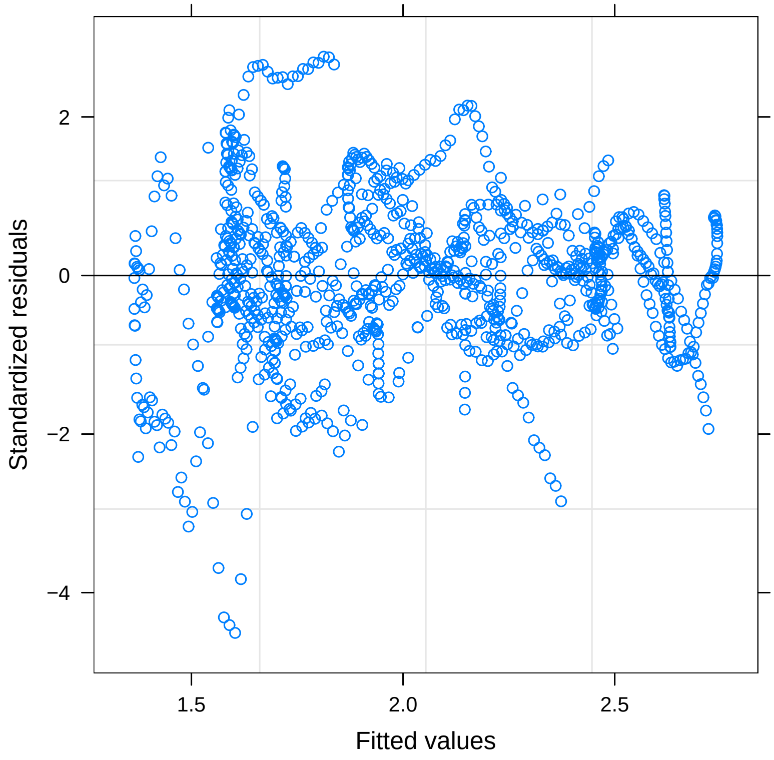
The result above show that the No_per_cage effect is not significant to model, therefore it will be extracted of the next analysis.

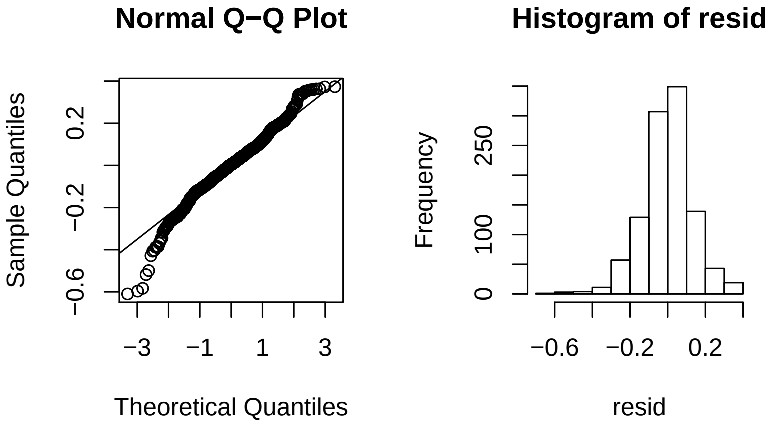
For all others covariate of the fixed effect of the model were applied the same methodology used in the test about No_per_cage. Follow the fixed-effect of the final model:

parameter	numDF	denDF	F-value	p-value
$Intercept$	1	1043	876.18	< 0.0001
$ns(ETime, 4)$	4	1043	240.36	< 0.0001
$Season$	1	13	4.80	0.047

Table 9: Significance of the final model

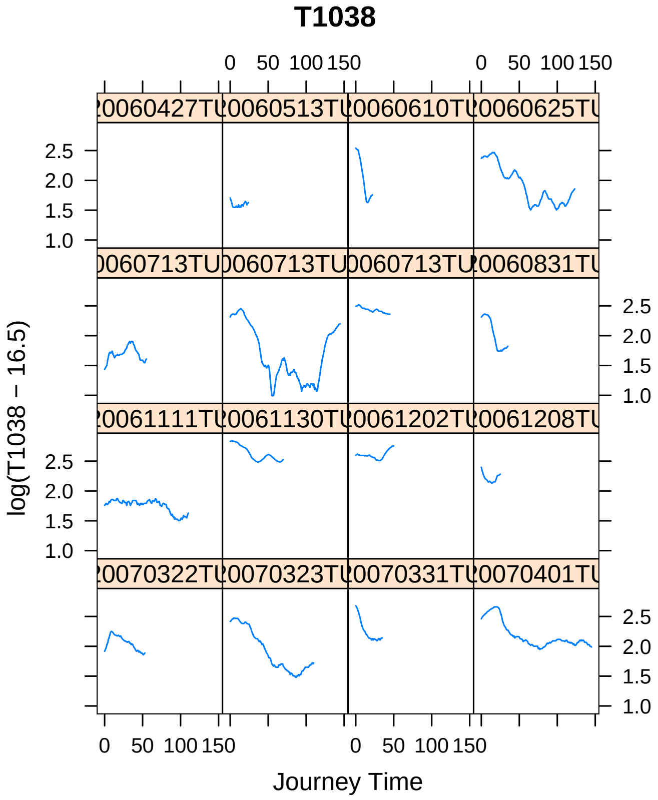
The three variables above are significant to model, and using the function summary of the R, program in annex, can see that there aren't high correlation between the covariate. before of the to study the estimative of the parameters, will be done the residual analysis for to verify the assumptions of this modeling with mixed-effects. Follow the graphics:

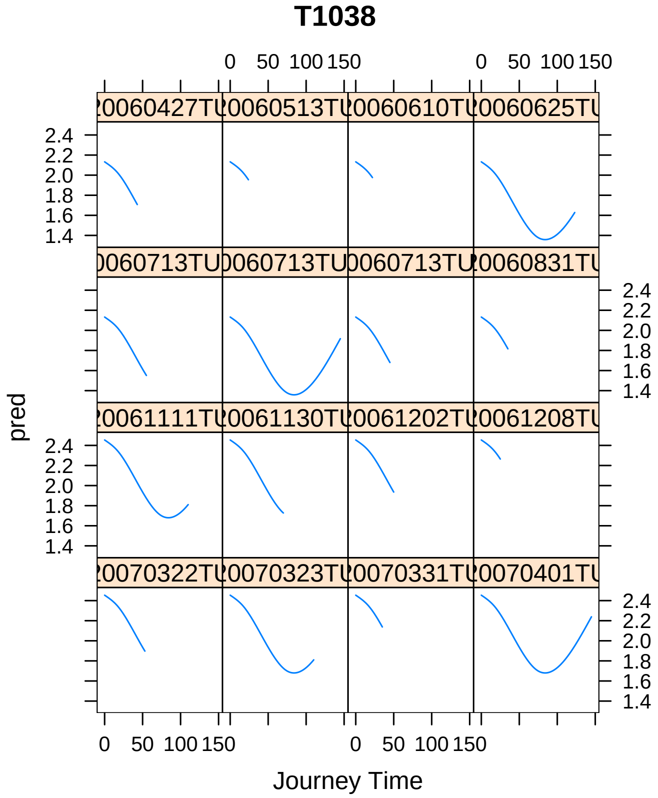




The plots above show that this approach have problems with the assumptions and even with the random effect inclusion, what helped to explicate better the variability, there are problems with the residuals distribution, in the first plot there are standard residuals out of the interval of the -3 to 3 and there are trends explicated by sequence of the measure in each trip. The QQ-plot show there are problems with the tails of the distribution of the residuals. And the histogram show that there are a little asymmetric behavior of the residuals.

As the assumptions are violated the analysis of the estimative will not be done, but will be showed plots with the real values and predict values. Follow the graphics:





The plots above show that for some trips the predictions are seemed with the real values, but there are temperature behavior very different of the expected by researchers, therefore the predictions are not good. This characteristic confirm that there are variability that are not controlled, therefore can be that there is some covariate that wasn't included in the study. Other motive that can be influencing are the assumptions, the variability is not homogeneous and we can try other distribution for the data too.

The same techniques above can be used for others positions, follow the modeling of more a position.

Second Position 18 1 2:

Again the start model is with all covariate. Follow the results:

anova9.tex	numDF	denDF	F-value	p-value
$(Intercept)$	1	933	1313.91	< 0.0001
$ns(ETime, 4)$	4	933	168.34	< 0.0001
$Season$	1	6	13.14	0.011
$QDay$	3	6	2.28	0.179
No_per_cage	3	6	2.54	0.152

Table 10: Analysis of Variance of the Temperature Transformed in the Position 18 1 2 with mixed-effects

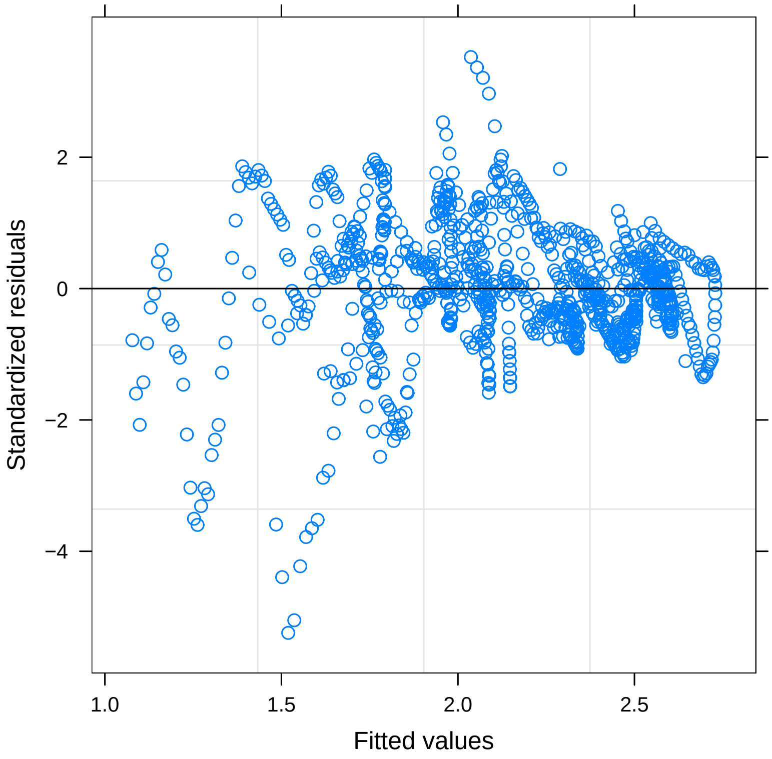
After of the variable selection the final model has the results follow:

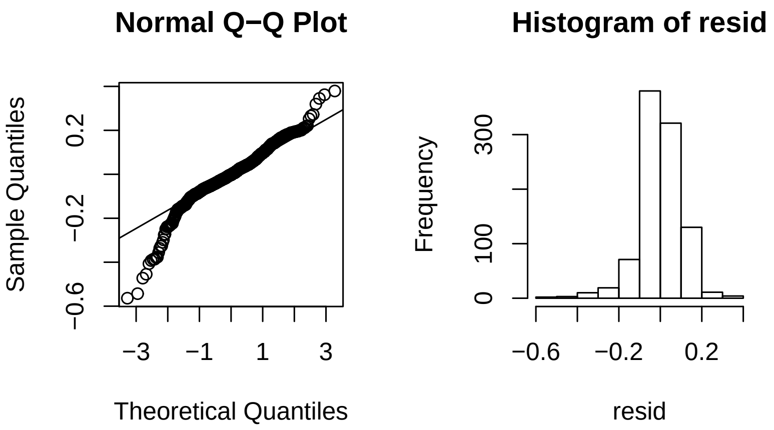
anova10.tex	numDF	denDF	F-value	p-value
$(Intercept)$	1	933	2351.12	< 0.0001
$ns(ETime, 4)$	4	933	166.36	< 0.0001
$Season$	1	9	26.57	0.0005
No_per_cage	3	9	8.64	0.005

Table 11: Significance of the final model in the position 18 1 2

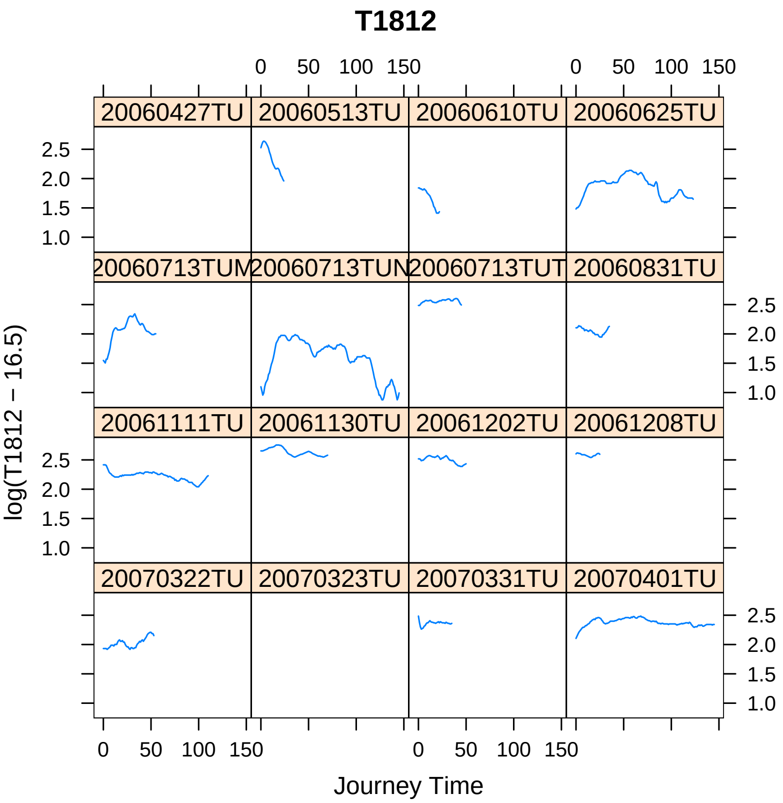
A important observation is that the final model this position has more covariate than the last position, now the No_per_cage covariate is included in the final model.

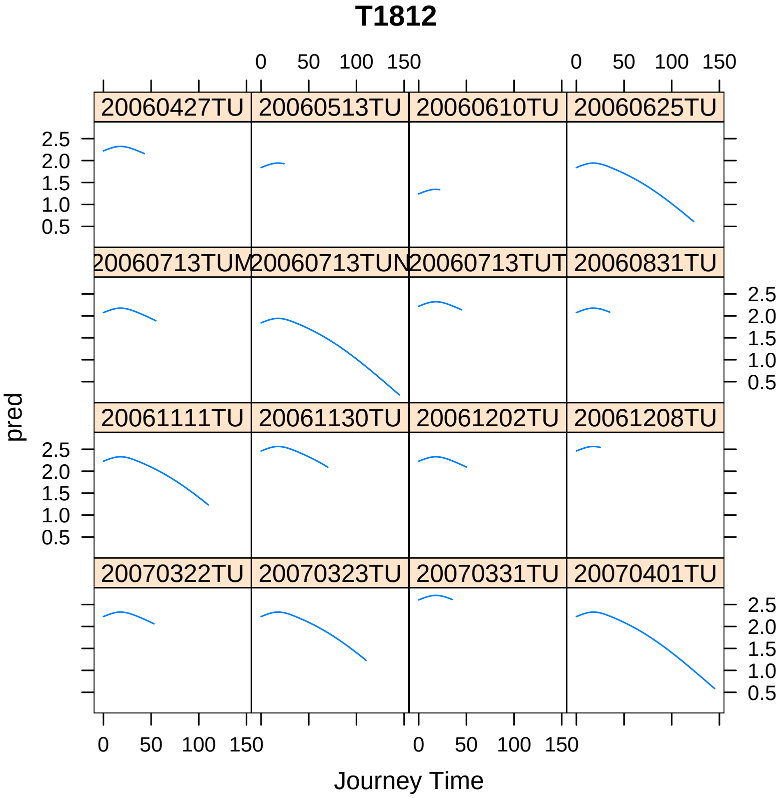
Now will be showed the residual analysis, the same problems of the last position are occurring in the position 18 1 2:





Follow the plots of the real and predict values:





The graphics above show that the temperature of this position has a different behavior than expected by researchers, now the temperature up first and down in the final of the trips.

5 Conclusion

This data was studied with others approaches before, for example geostatics, principal components and analysis of the variance three way and all analysis based in inference theory have some problems. We propose the linear model approach with fixed an mixed-effects to model the temperature throughout the time of the trips. The final models presented significant parameters, but were evidenced problems in the assumptions with the residual analyses. This problems can be explicated by many causes, in a brainstorm session with researchers were identified some possible causes of variability didn't control, for example, in some trips the driver watered the boxes and the chickens and this practice between drivers is a usual, therefore the researchers not to control this variability because the objective is to model the reality with observational data and not to model experimental data. Other possible cause is the variability between the Seasons and the QDay, this covariate don't consider the climatic variation, for example, trips with date in the winter and QDay equal to Q1 can have higher temperature than trips in the summer in the QDay equal to Q3 because in the trip of the summer can to have very rain, for example. A suggestion for this problem is to choose the Season and QDay by the ambient temperature in each date of the trip. Beyond of temperature data the researchers have humidity data with the same structure than temperature, other suggest is to use a index that use the two measure, maybe with the index a variability of the data is shorter. Beyond of the suggest for to use the same approach of this study with other response, other idea is to use some distribution of the probability that has domain of the zero to infinite, for example, to use mixed-effect generalized linear model with gamma family. In the session brainstorm other goal was thought, the researchers want to model the temperature in the trips with the position been modeled as covariate, therefore can or not there are correlations between the observations of the neighbor positions. Therefore, with this study were propose many things about the data, many possible solutions were thought to model the data, beyond of this study was very good for to develop knowledge in data manipulation with R.

Reference

Pinheiro, J. C., Bates, D. M (2000). Mixed-Effects Models in S and S-PLUS, Hoboken: Statistics and Computing Series, Springer-Verlag, New York, NY.

Marx, B., Eilers, P. (2005). Splines, knot and penalties. 9th School on Rgression Models, Sao Paulo.