MATH-493 : Individual report

Théo Imler

May 2019

This report has for objective the statistical analysis of the data collected by JR Caplehor and J. Bel for their study published in 1991 in the *Medical Journal of Australia*. This study objective was to compare two methadone treatment clinics for heroin addicts who entered maintenance programmes (between February 1986 and August 1987) to assess patient **time remaining** under methadone treatment, an opioid used for maintenance therapy in opioid dependence.

The dataset is composed of 238 individuals and the following variables :

- **Survival** : The time (in **days**) until the patient dropped out of the clinic or was censored
- **Status** : If the patient droppred out of the clinic (code : **1**) or was censored (code : **0**)
- **Clinic** : I which methadone clinic the patient was (**1** for the first, **2** for the second)
- **Prison** : If the patient has a prison record (Yes : **1**, No : **0**)
- **Dose** : patient maximum methadone dose (**mg/day**)

Note that the two clinics differ from their live-in policies.

Since the variable of interest is here a time to an event the dataset is survival data. In this case, the event is the departure from clinic (drop out). If the patient drops out, the therapy is considered as a failure (negative event). In case where the study ends before the patient drops out we only know that the waiting time exceeds the observation time (censoring) and we can't model this situation with logistic regression. For the survival analysis methods to be valid, the censoring mechanism must be independent of the survival mechanism. In this study, the total duration was fixed thus it is right censoring (fixed type I), a non-informative mechanism that leads to the same likelihood function.

In this report, the analysis focus on how single parameter affects the survival. Lastly, the hazard function was modelled using Cox regression. Every outputs and results were obtained from **R** software (v.3.5.2).

Data Investigation

The dataset is consisted of two categorical variables : **Clinic** and **Status**. In order to be able to represent survival curve for the **Methodone** variable, an ordonnored factor *Dosage* was created as follow :

- Methodone < 55 : Dosage = **Low**
- $55 \geq$ Methodone ≤ 70 : Dosage = **Medium**
- Methodone > 70 : Dosage = **High**

As observable on **Table 1** below, the proportion of censored events is high since it represent more than 60% of the sample. It is then not adequate to use regression analysis.

Variable	**Levels**	**n**	%
Clinic	Clinic 1	163	68.5
	Clinic 2	75	31.5
	all	238	100.0
Status	left	88	37.0
	censored	150	63.0
	all	238	100.0
Dosage	Low	72	30.2
	Medium	119	50.0
	High	47	19.8
	all	238	100.0

Table 1: Patient characteristics: categorical variables.

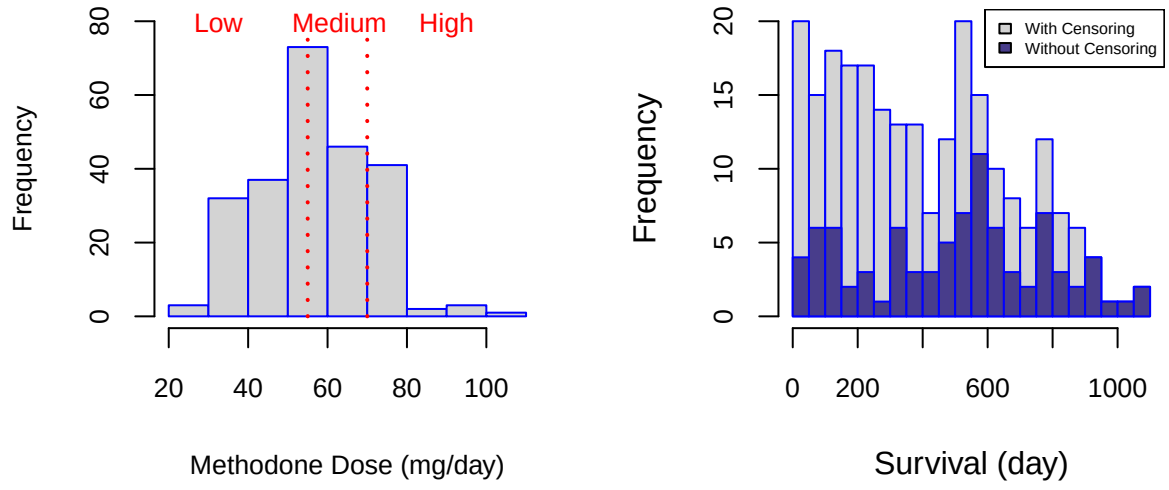


Figure 1: Dosage and Survival distribution

We can observe than the repartion of patients into clinic and administered Methodone dose are not uniformly distributed. Indeed, there is two times more patients in Clinic 1 than in Clinic 2 and half of the patients are under a medium dose of Methodone, the two other regimes are less frequent.

This can be easely visualized on Figure **1**(left panel) and resumed in Table **2**, we can see that the medium dosage is the most used among patient for an average dose of 60.4 g by day.

Variable	**n**	**Min**	$\mathbf{q_1}$	$\widetilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\mathbf{q_3}$	**Max**
Methodone	238	20	50.0	60.0	60.4	70.0	110
Survival	238	2	171.2	367.5	402.6	585.5	1076

Table 2: Patient characteristics: continuous variables

Logically, the survival decreases with time, but it is not possible to know if the decrease is due to an event or to censoring. In order, to have a better visualization, the survival distribution of patients who experienced an event during the study observation time was plot on the same histogram (blue, no censoring). As observed, the subset distribution is more uniform than the population one, which indicates that a majority of observed events in the population is in fact censoring.

It is now interesting to see how these different factors impact the survivall of patient.

Survival Function

The **survival probability** is one of the two related probabilities used to decribe survival data with the **hazard probability**. The survival probability, also known as the survivor function $S(t)$, is the probability that an individual survives from the time origin (of the study) to a specific future time t. Thus, this function focuses on not having an event. It can be estimated using a non-parametric method from observed survival time : the **Kaplan-Meier** method. It calculated the survival probability at time t_i, $S(t_i)$ as follow :

$$S(t_i) = S(t_i - 1)(1 - d_i/n_i)$$

,where : n_i = the number of patients alive just before t_i and d_i = the number of events at t_i.

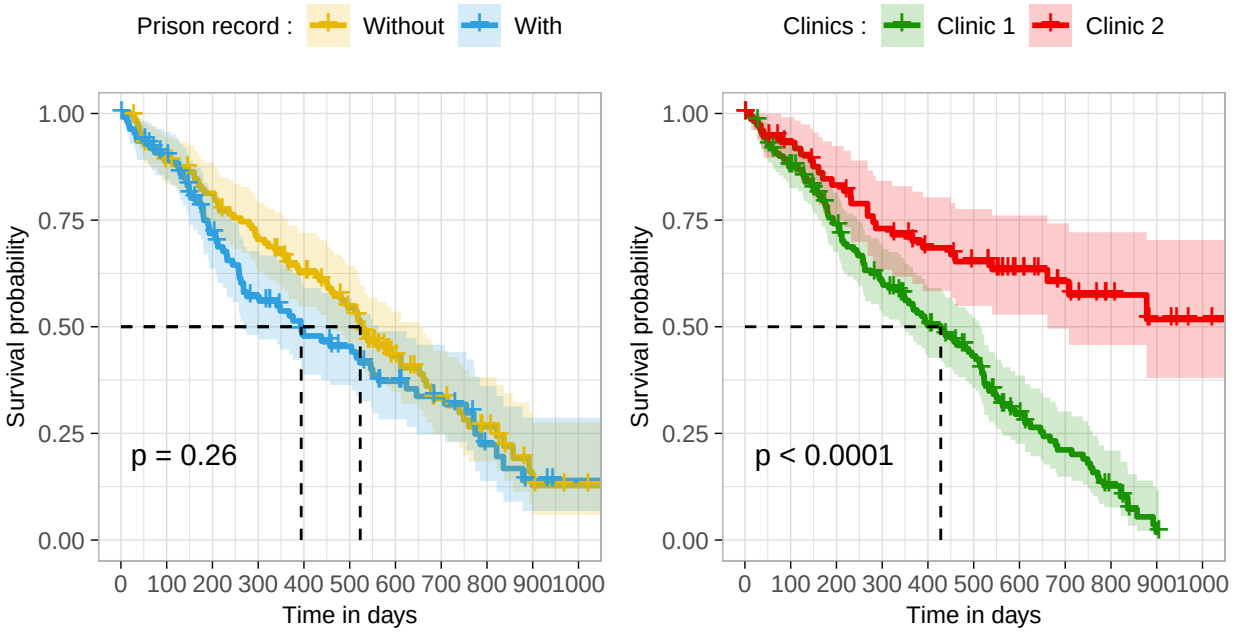


Figure 2: Survival curves stratidied by Prison record and Clinic variables

The discrete calculation method used results in a step function. The survival probability can be visualized considering one factor in order to compare survival curves between two groups.

In this section, the survival function was visualized for the Clinic, Prison and Dosage variable. Overall, we want to know which factor have a statistical significant effect on the survival curves. For that we need to test for a significant difference between the curves.

Since the data are censored, the survival distribution comparaison was made using a non-parametric test : the **logrank test**. It compares estimates of the hazard functions of the two groups (three in case of Dosage variable) at each observed time, assuming that the chance of dying for subjects at risk is the same for each group.

Then the test compares the observed number of deaths in each group to the expected number using a χ^2 test. The p-value was indicated on each graph above as the confidence intervall for the survival curve.

As we can observe on Figure **2** the clinics have a clear impact on the survival rates of Methodone patients (p-value < 0.0001). The patients in Clinic 1 drop out faster than the patients in the clinic 2. After 428 days, 50 % of the patients of Clinic 1 drop out while Clinic 2 will never reach 50% of drop out on the study duration.

Therefore, the live-in policy of Clinic 2 have a real benefit for patients. On the other hand, the possession of a prison record does not impact the drop out rate of patients. As seen on the table **3**, with a p-value of 0.262 the difference is not statistically significative even if the patients with a prison record have a median drop out time shorter than the patients without.

The results above show that the remission times differ in function of the clinic. It is not demonstrated that the difference between clinics is not due to other explanatory variables.

As an example, to test for the clincis variable significancy without effect of the Prison variable, it is good to compare the two clinics accounting for prison records by stratifying on the variable prison.

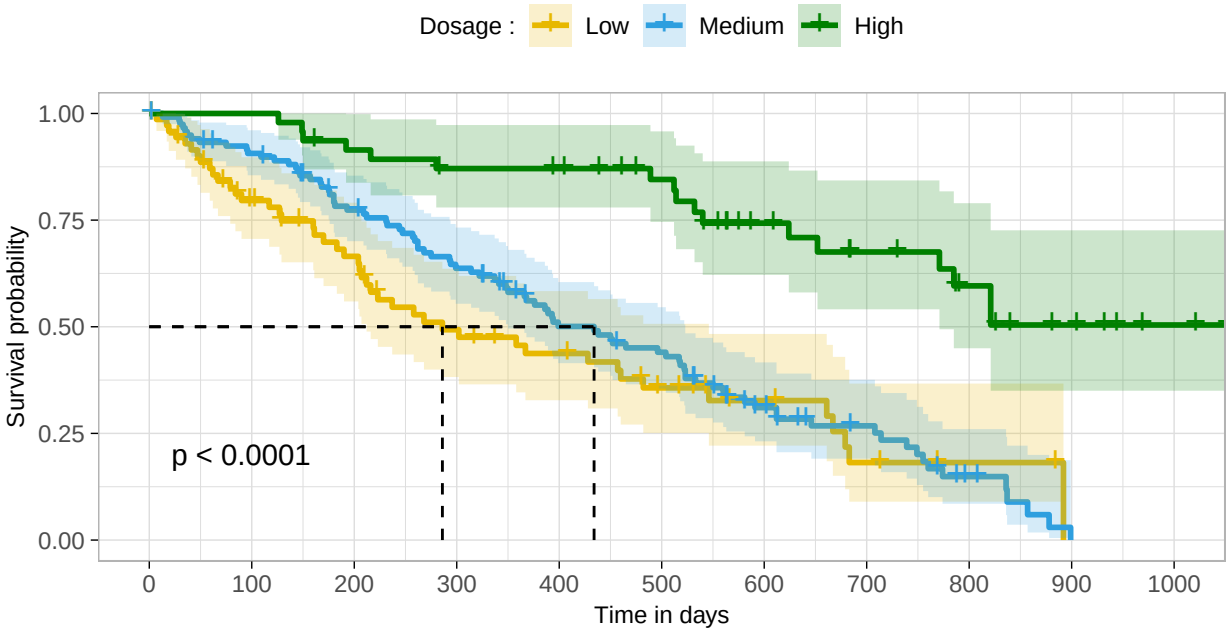


Figure 3: Survival curves stratidied by Methodone dosage

Prison		1 (N=163)	2 (N=75)	Total (N=238)	p value
0	**Surv(Survival, Status)**				< 0.001
	Events	66	15	81	
	Median Survival	496.000	NA	523.000	
1	**Surv(Survival, Status)**				< 0.001
	Events	56	13	69	
	Median Survival	341.000	878.000	394.000	

Table 3 : Logrank test for Clinic variable stratified by Prison

As we can see on Table **3** above, in both case the remission time is significantly shorter in clinic 1. The patients in clinic 2 have a consistently better prognosis for remaining under treatment that do parients in clinic 1. It is interesting to note that this difference appears to be really small before one year of treatment but diverges after.

Even if the Methodone is a continuous variable it was interessant to visualize its effect with the corresponding survival curve using the Dosage variable. As observed on Figure **3** and Table **4** below, the test statistic is highly significant, indicating that these three curves are not equivalent. Patients with high dose of Methadone have a consistently better survival prognosis than patients with medium or low doses.

The maximum daily dose of methadone dispensed during the study period seems to be a highly significant predictor of retention. Indeed, less than 50% of the patients dropped out clinics after the end of the study (which explains the **NA** value in table **4** against none of the patients under lower dosage.

	Low (N=72)	Medium (N=119)	High (N=47)	Total (N=238)	p value
Surv(Survival, Status)					< 0.001
Events	45	88	17	150	
Median Survival	286.000	434.000	NA	504.000	

Table 4 : Logrank test for Dosage variable

Now that we have compared survival curves in specific treatments, it will be interesting to analyze the effect of several risk factors on survival.

Cox Proportional Hazards Models

In survival analysis we often want to assess which time periods have high or low chance of failure among those still at risk at the time. We can characterize these risks using the instantaneous failure rate or hazard function $h(t)$ which describes the probability of an event (its hazard) if the subject survived up to a specific time point t.

Therefore, it measures the instantaneous risk of death (here drop out of the clinic). This hazard function is needed to consider covariates when comparing survival of patients groups.The **hazard function** and **survival function** can be related together : $S(t) = exp(-H(t))$, where $H(t)$ is the hazard function integrated on the time intervall.

As the survival probility, it is not a parametric method, *i.e* it does not assume an underlying probability distribution. Although, it assumes that the hazards of the patient groups compared are constant over time. Thus, it is called a **semi-parametric** method.

One way to model hazard function is **Cox regression**, it is carried out similarly to regression but with lineraity assumed on the log hazard scale.

Cox proportional hazards models can be interpreted by looking to the **hazard ratios** (HR) which are derived from the model for all covariates included in the model forumla.

As a reminder, the hazard function is modeled as : $h(t) = h_0(t)exp(\beta(x))$, where $h_0(t)$ is the baseline hazard and $exp(\beta(x))$ is the hazard ratio (HR) between two individuals whose values of x differ by one unit when all other covariates are held constant.

An HR represents a relative risk of death that compares one instance of a binary feature to the other instance. > 1 indicates an increased risk of death whereas an HR < 1 indicates a decreased risk.

The Cox regression model used in this report takes every variable in account: **Survival** ~ **Prison** + **Clinic** + **Methodone**. The results are visualized in the table **5** below :

	Variable	HR	95 CI for HR	P-value
1	Clinic	0.36	(0.24, 0.56)	<0.001
2	Prison	1.39	(1.00, 1.92)	0.0508
3	Methodone	0.97	(0.95, 0.98)	<0.001

Table 4: Cox regression results summary

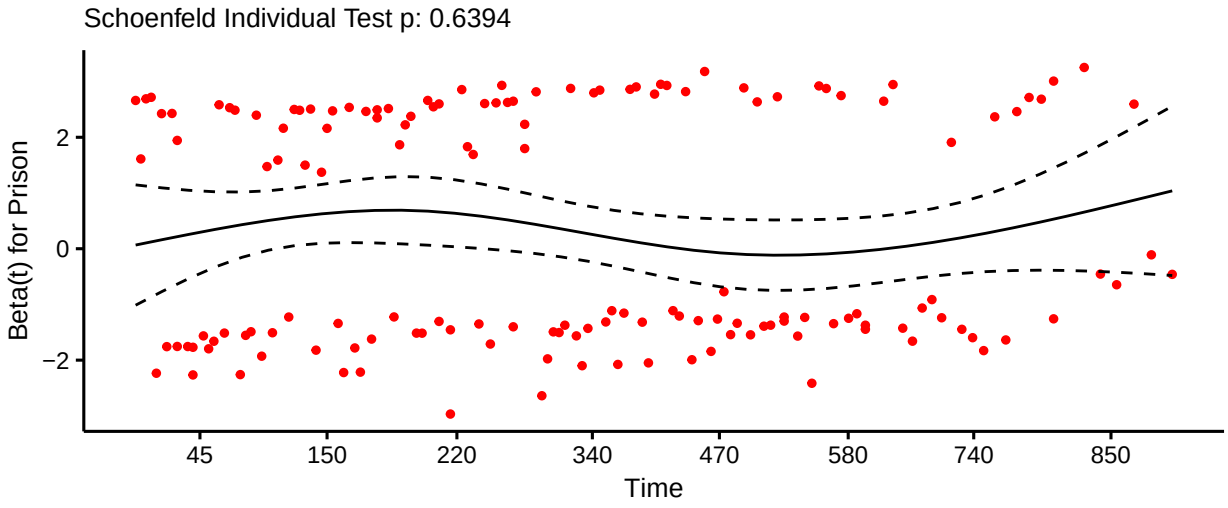
From the outputs of this model we can see that the Methodone dosage and the clinic policy seem to affects positively the patients treatment since the associated HR are < 1 and their p-value significative.

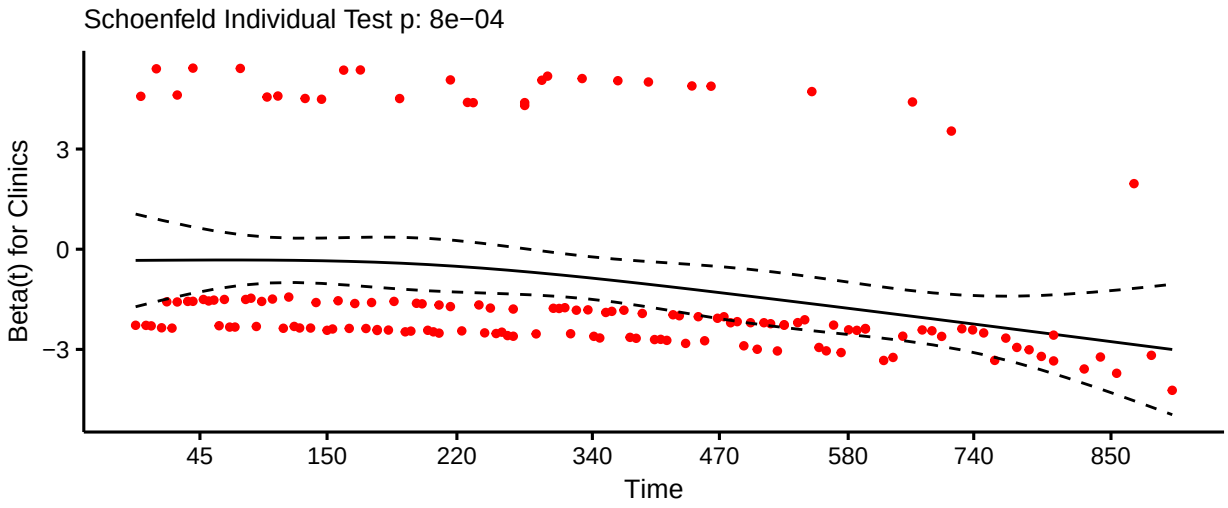
On the other hand, the fact to have a prison record seems to affect treatment efficacity but note that the p-value is not significative. As a first conclusion, we can observe that the model confirms the first assumptions made from the study of the survival curve: the non-significance of prison variable with the efficacity of methadone dosage and clinic live-in-policies. Note that the Clinic appear to be the most important predictor of survival since from the HR we can see that the risk of death is reduced by 64 %..

But in order to return significative results the **proportionnal hazard assumption** must hold. It means that the ratio of the hazards for an option two individuals is constant over time. Indeed, if the coefficient vary in function of time, the model is irrelevant.

One possible solution to check this assumption is the Schoenfeld residuals test which test independence between residuals and time. It correlates the corresponding set of scaled Schoenfeld residuals with time.

For each covariates, we can plot the scale Schoenfeld residuals against time and the coefficient estimation. The proportionnal hazards assumption is then supported by a non-significant relationship between residuals and time, and refuted by a significant relationship. These residual





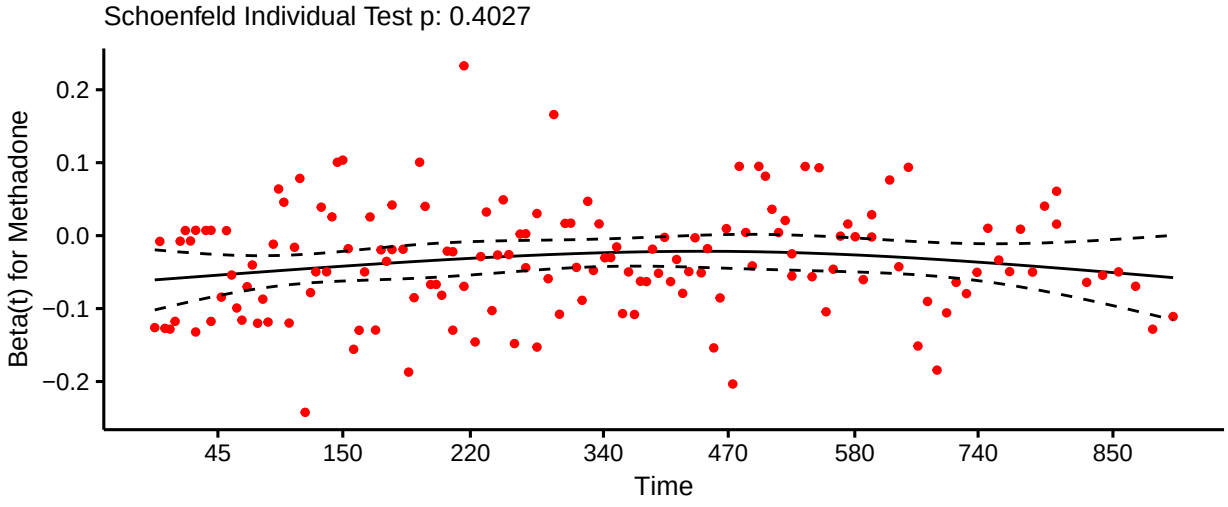


Figure 4 : Schoenfeld residuals for Prison, Clinics and Methadone variable

First, we can notice the clear "two-band" patterns of Prison and Clinics residuals. This is easely explained by the fact that both variable are two-level factor.

Even if we observe coefficient fluctuations for Methodone and Prison, there is no clear pattern with time, as confirmed by the non-significant p-value. Thus, the proportional hazards appears to be supported for the covariates Methdone and Prison.

On the other hand, we can observe a clear patern for the clinics variable since the coefficient strongly decreases with time. This observation is clearly confirmed by the Schoenfeld residuals test with a significant p-value. These results are summarized in table **5** below.

Therefore, we have one covariate breaking the assumptions. In order to fix it for future models, we can create an interaction term with clinics and time or we can use stratification. Note that the stratification option is not the best in our case since we would not be able to examine the effect of the stratified variable anymore.

	rho	chisq	p
Clinics	-0.26	11.19	0.00
Prison	-0.04	0.22	0.64
Methodone	0.07	0.70	0.40

Table 6 : PH assumption testing

How can we assess the fit of our model ? Martingale residuals can be used to check the model fit. When evaluated at the true coefficient value, the expected martingale residual is zero. We can therefore check for systematic deviations from the assumed model by inspecting scatterplots of the martingale residuals. In order to have a good model, the martingale residuals should be scattered fairly evenly above and below 0, and in addition they should not seem to show any particular pattern.

For clear visualization, we can use the deviance residual which is a normalized transform of the martingale residual. These residuals should be roughtly symmetrically distributed about zero with a standard deviation of 1 :

- Positive values correspond to individuals that "died too soon" compared to expected survival times.
- Negative values correspond to individual that "lived too long".
- Very large or small values are outliers, which are poorly predicted by the model.

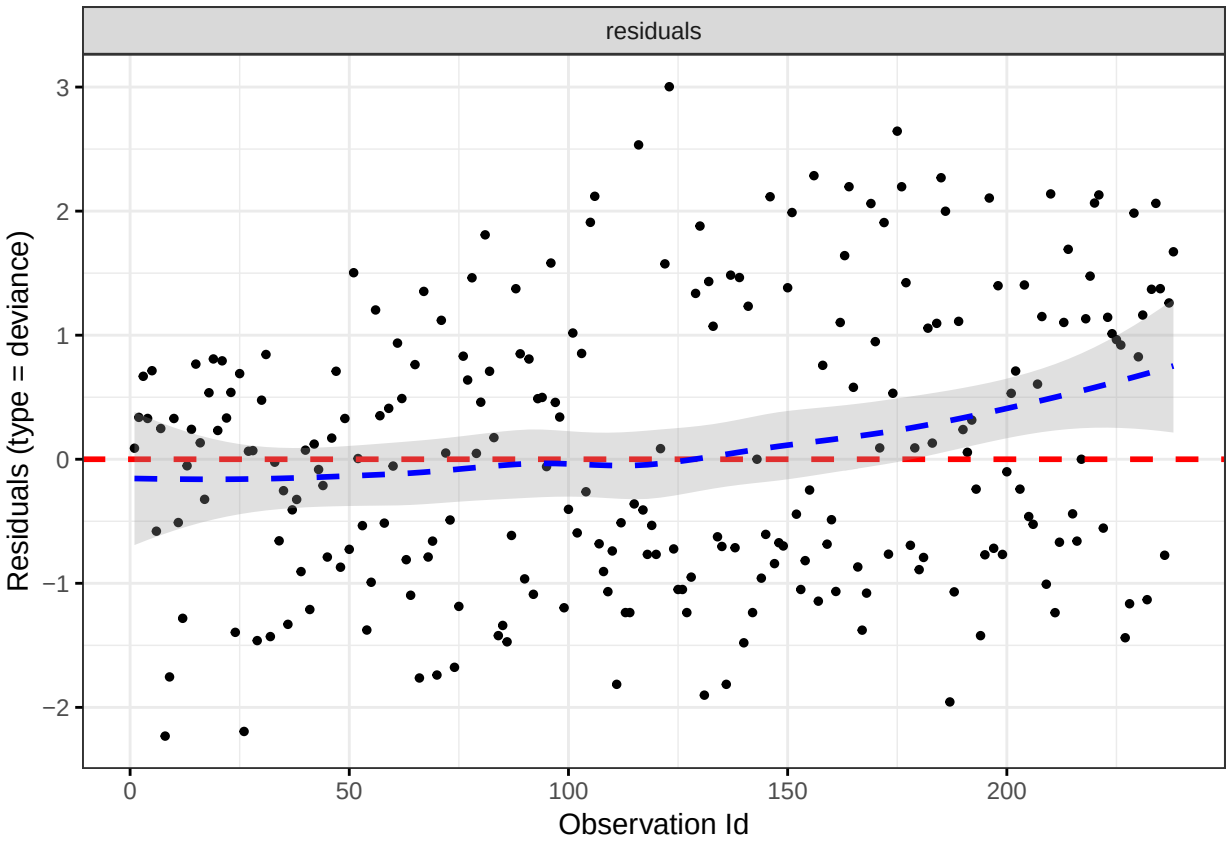


Figure 5 : Variance residuals of the model

From the diagnostic plot above on Figure **5**, we can see that the residuals distribution is roughly symetric as wanted but the distribution deviate from the theoretical line (red) for the last individuals. The model seems to fit well the data for the first observation as already observed previously but it is not robust enough to be choosed.

Conclusion

The results of this study suggests that the clinic live-in-policies and the opioid agonist (Methadone) dose seems to be the best predictors for a successful treatment.

The individual variable study have confirm these two variables as significant but the regression model was not enough robust and adequate to be chosen.

Indeed, a modified model which would take in account the time dependance of the clinics variable with an interaction term seems to be the best option with our dataset and should be investigated.