École polytechnique fédérale de Lausanne

Project 2023

Regression Methods

Regression Methods Project 2023

Authors:
Jean-Sébastien Delineau
Martin Leclercq

Instructor:
Anthony Davison
Assistant:
Paul Freulon

Introduction

A cigarette butt contains more than 150 substances toxic to the environment [5]. The contemporary challenges around environmental issues are becoming increasingly more urging. Therefore, providing accurate and powerful models to guide our decisions is necessary. The data frame of interest is **SLR-Data.csv**. It contains information collected by trained volunteers of litters on the shores of Swiss lakes and rivers. As a proxy for the overall littering on Swiss beaches, we will consider the feature "Cigarette butts and filters" as our response variable.

This project is composed of three sections. The first will be about data exploration. The second will be on models construction to predict the number of cigarette butts on the beaches of Switzerland. The third will be on intervals of prediction on any given beach of Switzerland on any days of August in 2023 for the number of cigarette butts.

Since we are modeling the counts of random positive occurrences, the models presented will be of Poisson, as Poisson distribution counts the number of times an event occurs in a **fixed interval**. Therefore it begs the question, what is the fixed interval considered? As we can see in Table 1, the time interval in between measure is not fixed over the months, which could disrupt the results obtained and the truthfulness of the Poisson distribution. However, since other people might clean the beach, this random effect might counterbalance this flaw. Hence we will not take it into account and take the fixed interval to be the length of the beach in meter. We also get that this flaw of ordered measurement propagate on the third part. We will assume that the measurement predicted of a random beach and day in August can only be precise for one measurement per month. All our analysis has been done with the `R` software.

1 Data exploration

1.1 Dataset

SLR-Data.csv has 1052 observations of 134 variables. The observations start in April 2017 and finish in March 2018. We assume that there should not be growth of the variable of interest over the years since we only have a one-year span of observations and hence we can difficulty study the yearly effect on the growth. In other words, we predict the number of cigarette butts in August 2023 just like we would predict in in August 2017!

After disregarding the NA values and the blatant outliers (see 1.2 for more details), we reduced the data to **891** observations. Assuming the data has been collected uniformly over the months and the days, we should have around **2.5** observations per days over all the beaches of Switzerland. Since there is **109** beaches it makes less than 12 observations per beach over the year, therefore some beaches are not observed certain months. This lack of data forced us to make the assumption that a given month, the distribution of the number of cigarette butts is the same for all the weeks. Table 1 reports the number of measures by month.

Table 1: Number of measures of cigarette butts from January to December

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Number of measures	57	55	68	90	95	94	89	93	75	77	55	55

This assumption has one weakness, it does not encapsulate the random events intrinsic to certain days, i.e.

social events, parties, festivals... Which we could complement with an extra feature "event" considered as a factor obtained by another data set listing the major events in Switzerland.

1.2 Features

The initial dataset is composed of **134** covariates, but we only consider {*BeachName*, *BeachLength_*m, *EventDate*, *Gebiet*}. After renaming and arranging, we obtain {*Beach*, *Length*, *Month*, *Season.custom Day*, *Settlement*}, as presented in Table 2, where we extracted *Day* and *Month* features from the *EventDate* covariate. We also create four customized seasons { *J/F/M (January/Febuary/March); A/M/J (April/May/June); J/A/S (July/August/September); O/N/D (October/November/December)*} by grouping months three by three. The categorical variables {*Month*, *Day*, *Settlement*} are considered as factors, while *Length* is considered as an offset term. There are 3 area types {City, Agglomeration, Land}, 7 days of the week, 12 months of the year. Due to the distribution of y, we propose two transformations of y:

$$y_1 = 2\sqrt{(\frac{y}{4}+1)} \quad y_2 = \log(y+1). \tag{1}$$

Table 2 reports the first row of the dataframe obtained.

Table 2: First row of our data set

Beach	**Length**	**Month**	**Season.custom**	**Day**	**Settlement**	**y**	**y1**	**y2**
Aare_Bern_CaveltiN	28	4	A/M/J	Sunday	Stadt	12	4	2.565

We dropped **146** rows of NA values, which consists in **13.89%** of the original data. We then drop the two rows associated to *untersee_steckborn_siedlerm*. The median beach length is 36 meters, the second beach is 375 meters wide while *untersee_steckborn_siedlerm* spans **1335049** meters! Every fit provides poor residuals for the beach *Rhein_Beach near Tinguely Museum_Bolger. O_Sigrist F.* This beach has a particular behavior since it has the four highest values of butts measured **2630, 1532, 1500, 803** in the entire data set, when the mean is at **33.09** and the median at **5**. After doing such adjustment our distribution looks like this Figure 1. We can see from the distribution of the number of butts that Poisson distribution will have difficulties to encapsulate all the zeros. There is **142** zeros in this new data set, which corresponds to **13.51 %** of the data. Which hints that a simple Poisson model will fit poorly.

1.2.1 Cigarette Butts

As mentioned previously, the exceeding number of zeros will not yield a proper fit. We therefore use different transformations of y (cf. (1)) as presented in Figure 1. As depicted in the top plot, the original density is closer to an exponential than a Poisson, which has extreme value ranging up to 600. The transformation y_1 has a distribution quite close to a Poisson, the only major problem is its increased variance or outliers, which a Poisson does not display. The bottom plot obtained with y_2 has no outliers or increases variance but the distribution only sort of resembles to a Poisson.

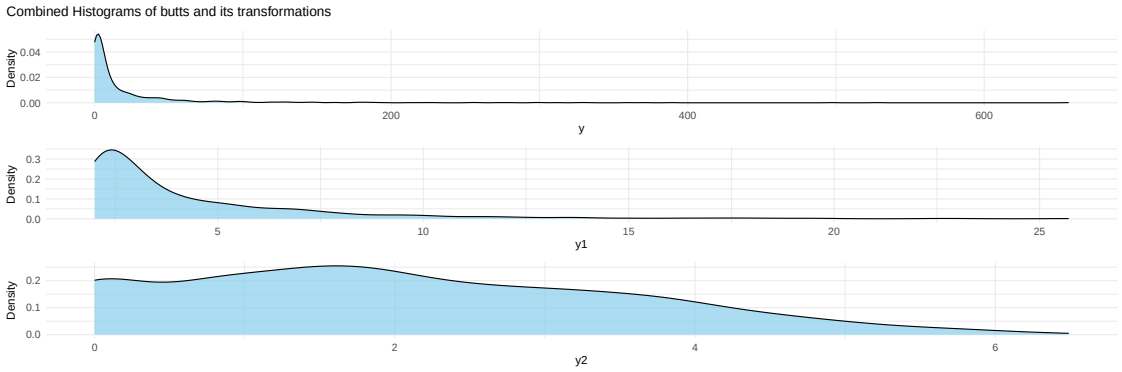


Figure 1: Distribution of cigarette butts and its transformations

1.2.2 Length

As our models will be of Poisson, we have decided to define the covariate *Length* has an offset. The models will then predict the number of cigarette butts divided by the unit of exposure, which is per length meters of beach. Since we consider it as an offset we have to take the logarithm, which yield the following fit Figure 2.

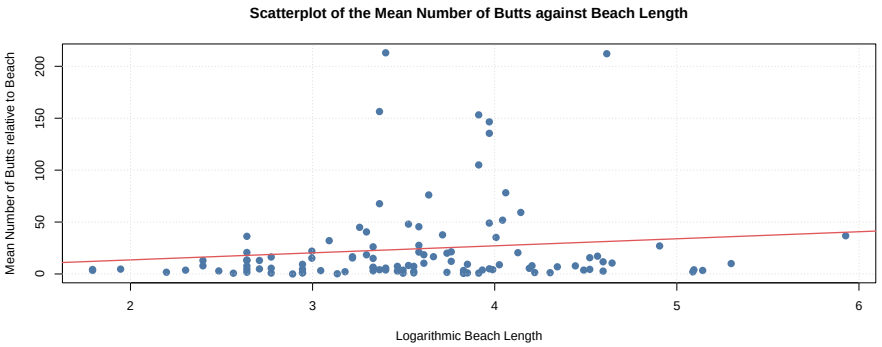


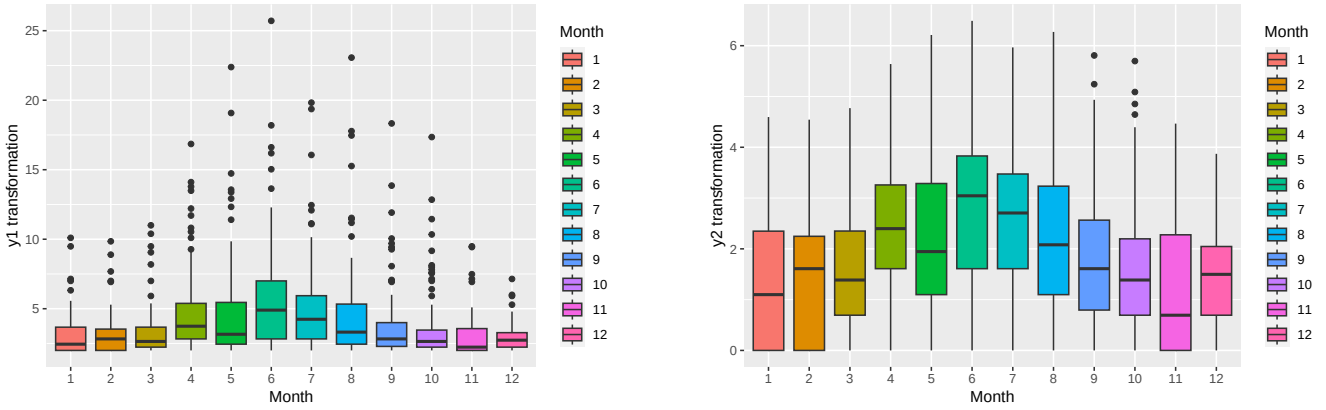
Figure 2: Relation between number of cigarette butts and beach length

We fitted a linear model onto the mean number of cigarette butts per beach. The resulting prediction is the red curve, of slope **6.67**. The fit is poor, but as this is the simplest model to fit, it is not abnormal to see outliers, which could be explained from other covariates. However, as the slope suggests, there is indeed a small positive correlation in between the number of cigarette butts collected and the length of the beach.

1.2.3 Seasons, Months and Days

As stated in the introduction, we assume independence from year to year, although this assumption would probably be false (e.g environment sensitisation campaigns, littering fines, number of trash cans increased to counter littering

etc.). In this sense, as 2017 and 2018 do not overlap in our dataset and represent the whole 12 months, we consider the data to be over a full one year span.
We are interested in knowing whether seasons, months and days are correlated with the number of cigarette butts we can find on Swiss beaches and lakes. Figure 3 gives insights into how y_1, y_2 and months are related. It seems that there are disparities between months, and that months 4 to 8 (i.e late Spring to late Summer) are strongly correlated with high number of cigarette butts collected. It makes sense as people tend to go to the beach when temperatures rise. Hence more people would potentially litter the beaches and lake sides.



(a) y_1 transformation

(b) y_2 transformation

Figure 3: Boxplot of number of cigarette butts by month

Notice the outliers in Figure 3b. They may correspond to the last sunny days to organise parties at the beach before winter. Additionally, notice all the outliers in Figure 3a. We may expect some overdispersion of the data as there are many outliers far from their empirical mean. Similarly to months, we look at the interaction between season and littering. We estimate the density of cigarette butts by season and obtain Figure 4.

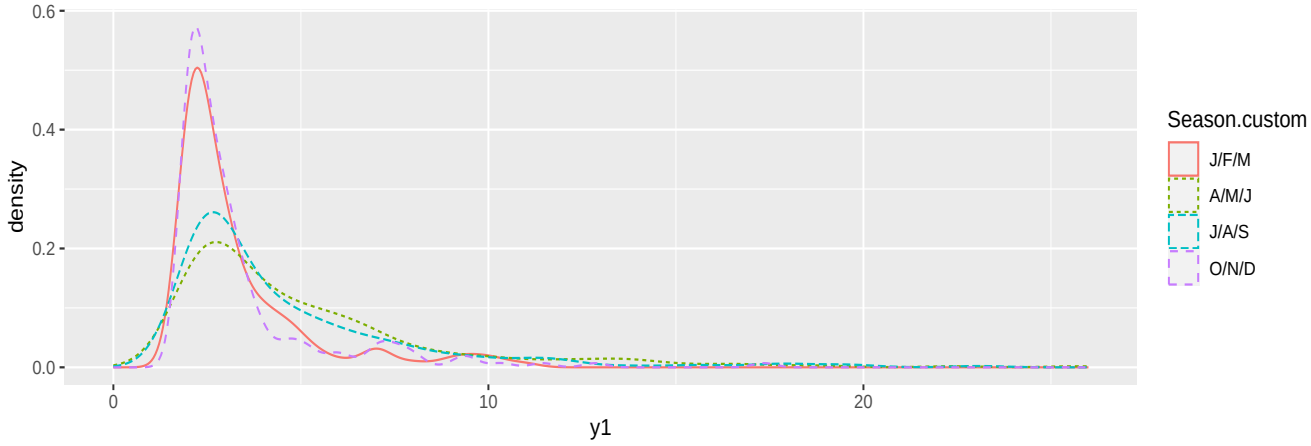


Figure 4: Distribution of cigarette butts by season, transformation y_1

It seems that categories A/M/J and J/A/S have higher densities towards higher values of cigarette butts. Additionally, in J/F/M and O/N/D have high densities towards small values of y_1. The findings corroborate findings of Figure 3.

Grouping by seasons, we can see in Figure 5 that in a given season, there are differences between the number of cigarette butts collected each day.

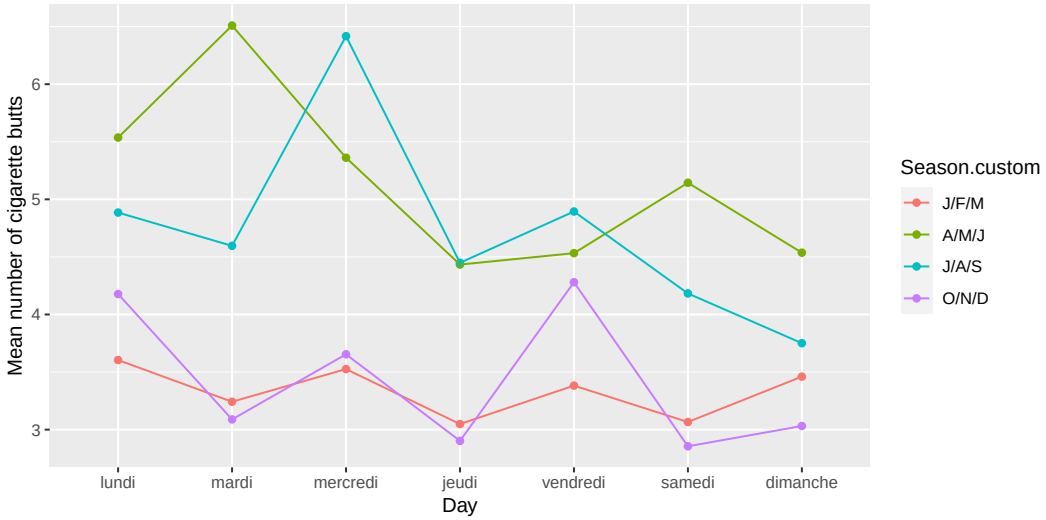


Figure 5: Mean number of butts by day by season, transformation y_1

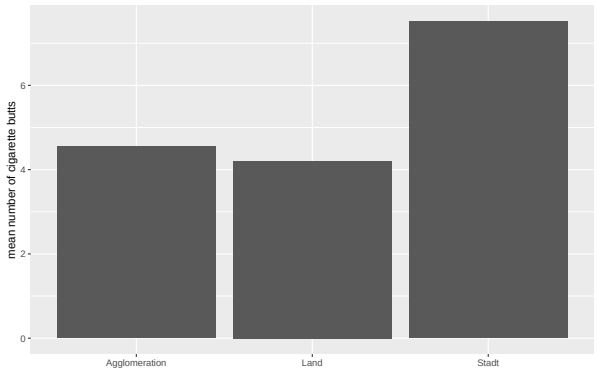
As expected, A/M/J and J/A/S have the highest values in average across all days. On the contrary, J/F/M values are on quite low across the week as during winter period, people rarely go to the beach. However, high values on Tuesday and Wednesday are quite surprising. Indeed, we would expect to have higher values rather towards the end of the week as people would traditionally unwind during the weekend. These unexpected results are probably due to the day the trash is collected. Perhaps the volunteers collect during the week and spend time elsewhere during the weekends.

1.2.4 Location and seasonality effect

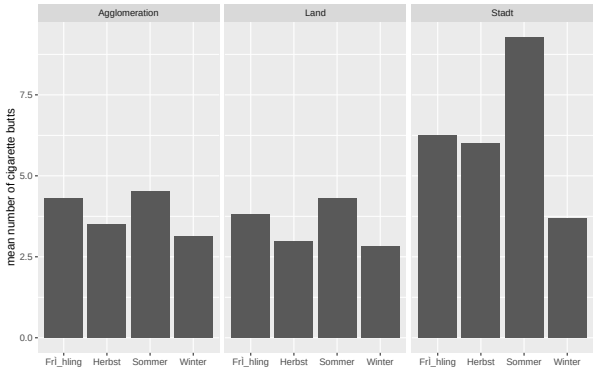
Intuitively, location seems like an important factor. Leman's lake side around Vidy (right in the city, nearby EPFL and Unil) is probably different from Leman's lake side near Lutry, where it is more residential and less frequented.

The 3 settlements Agglomeration, Land and City (Stadt) are roughly partitioned as 32.9%, 42.1%, 24.9% respectively. Figure 6a reports the mean of the number of cigarette butts collected in each area type. Along side that, we plot the interaction between customed seasons and settlements in Figure 6b. Mountain lakes and beaches are probably more frequented during summer when people go hiking whereas in winter people ski instead.

We can see that a lot more trash is collected in cities in average, which again is sensible as cities are often more frequented. The seasonality effect in Figure 6b reveals the difference between settlements across seasons. Again,



(a) Mean number of cigarette butts by settlement



(b) Mean cigarette butts by season for different settlements

Figure 6: Cigarette butts by settlement, transformation y_1

cities are more frequented and hence subject to higher littering rate.

2 Models

As mentioned previously, since we are considering count values, we would like to fit a Poisson model for the variable of interest y. Given the distribution of y presented in Figure 1, the mean of our observations must not depend linearly on the covariates, which is why we start with a generalized linear model (GLM) with a non linear log function. The data exploration in part 1 leans us towards the following relation for the mean μ with g the log link function:

$$E[y] = \mu, \quad g(\mu) = X\beta. \tag{2}$$

$$\mu_{mb(sd)(st)} = \exp(\alpha_m + \beta_b + \gamma_{sd} + \lambda_{st}) \cdot \text{offset}(\text{Length}).$$

With $m = 1, \ldots, 12$ for the months, $b = 1, \ldots, 109$ for the beaches, $s = 1, \ldots, 4$ for the seasons, $d = 1, \ldots, 7$ for the days and $t = 1, \ldots, 3$ for the settlements, i.e. a cross relation in between the seasons/days and season/settlement.

In the previous equations, we could have considered the cross relation in between Month/Day and Month/Settlement, but we chose to take seasons, since as explained in Figure 1, August demonstrates a similar behavior to July and September. Therefore it will allocate less degrees of freedom and will less likely over-fit.

2.1 GLM

Using the relationship for covariates presented in equation (2), we fitted a GLM model on y_1. Taking a Poisson family yields poor results as the empirical mean (≈ 4.28) and empirical variance (≈ 10.01) are different while in a Poisson distribution they coincide. Which is why we implemented a Quasi-Poisson and a Negative binomial to account for the over-dispersion [1].

[1] Similar idea presented in [3]

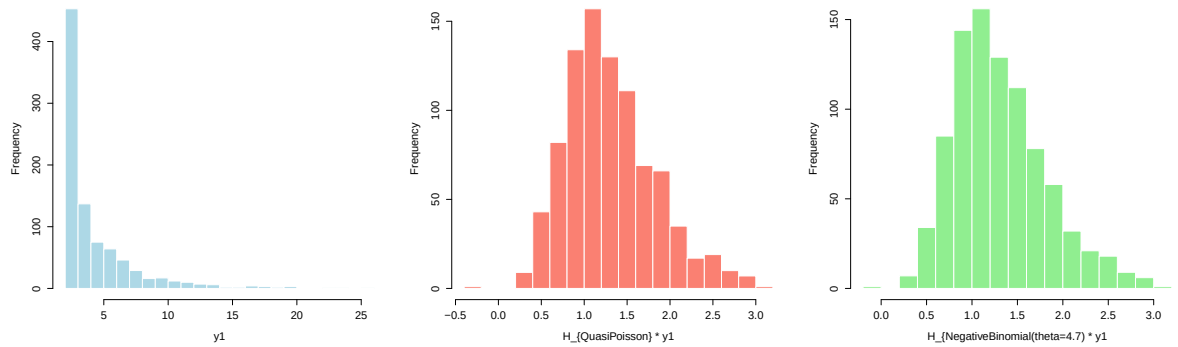


Figure 7: Distribution: y_1 | y_1 predicted with Quasi-Poisson | y_1 predicted with Negative Binomial

As depicted in Figure 7, neither the model with a quasi-Poisson family, nor with a negative binomial produce a good fit. The fitted values are in the range **[-0.2, 3]** while y_1 is in the range of **[2,25.7]**.

To define the negative binomial, we have to plug in a shape parameter Θ. We came across an abnormality with our fit. The smaller the Θ the smaller the deviance and the bigger the Θ, the smaller the BIC. Which means that the fit is not properly working and we decided to split the apple in two and take arbitrarily $\Theta = 5$.

As the Quasi-Poisson is based on a Quasi-likelihood approach, the concept of likelihood is not well-defined, which is why we get **NA** Values for AIC and BIC, therefore our only metric is the deviance, which is defined for GLM as:

$$\text{Deviance} = -2 \times (\text{log-likelihood of the model} - \text{log-likelihood of the saturated model}).$$

The Quasi-Poisson and negative binomial showed similar performance. For the negative binomial with $\Theta = 5$, we obtain an AIC of **3793.62**, a BIC of **4507.68** and a deviance of **144.94**. This result should be taken with care as we had an abnormality with Θ. For the Quasi-Poisson we obtain a deviance of **301.90**, which seems more reasonable.

We did not present the results for the logarithmic transformation even-though its BIC is in the range of the **3000** as a non negligible part of its density is negative.

Model Term	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Cp**	**F value**	**Pr(>F)**
NULL			891	59834	59834.26		
Month	12	57046	879	2788	2798.29	11465.55	< 2.2e-16 ***
Beach	107	2456	772	332	431.03	55.36	< 2.2e-16 ***
Season.custom	0	0	772	332	431.03		
Day	6	3	766	329	432.64	1.35	0.2311
Settlement	0	0	766	329	432.64		
Season.custom:Day	18	15	748	314	432.81	1.98	0.0091 **
Season.custom:Settlement	6	12	742	302	425.46	4.96	5.503e-05 ***

Table 3: ANOVA Table for Quasi Poisson

The covariates explaining most of the deviance are the *Beach* and *Month*. Since *Season.custom* is constructed

with the covariate *Month*, its null deviance reduction is expected. *Day* and *Settlement* are complex covariates, which is why they do not account individually for a lot of deviance. Mallow's Cp hints that the model is a bit too complex and we are bit over-fitting for the covariates Beach. The F-statistic confirms our analysis on the covariates *Season.custom*, *Day* and *Settlement*.

Using a cook distance with threshold $2C = 16/(n-p)$, we obtain that the influential points are, **121** (0.035), **235** (0.02), **422** (0.056), **726** (0.026), **746** (0.032). Looking at the dataset, they all deserve to be dropped.

2.2 GAM

To improve the fit obtained by the GLM methods, we add a random effect on the *Beach* covariate. This is done by the Generalized Additive Models[2]. In the same logic as for the GLM, we consider a Quasi-Poisson family with a log link function and an offset term for the *Length*. Therefore, we obtain a model predicting the number of cigarette butts per meter of beach length. Even if in this instance y does not have to depend linearly on the covariates, the model yields better results when we consider the transform y_1.
As all the covariates except *Length* are factors, to include random effect with splines, we have to resort to splines with a random basis, i.e. include random intercepts for each level of a factor. Most of the variance and randomness is due to the *Beach* covariate, which is why we consider it as the only random effect. Not combined with another random effect to not over-fit. The smoothing parameter estimation method used is REML.

The model is of the following form:

$$E[y_1] = \mu, \quad g(\mu) = \eta = B\Theta = X\beta + Zb. \tag{3}$$

$$\mu_{mb(sd)(st)} = \exp(f_b + \alpha_m + \beta_b + \gamma_{sd} + \lambda_{st}) \cdot \text{offset(Length)}.$$

For f_b the random effect smooth function for the variable Beach, with $m = 1, \ldots, 12$ for the months, $b = 1, \ldots, 109$ for the beaches, $s = 1, \ldots, 4$ for the seasons, $d = 1, \ldots, 7$ for the days and $t = 1, \ldots, 3$ for the settlements.

In the following table we assess different models with different grouping for the random effect on *Beach*, which allows it to vary randomly across different values of z, for $z \in \{Months, Settlements, Season.custom\}$ depending on the grouping chosen. Which yields the following model:

$$\mu_{mb(sd)(st)} = \exp(f_{b|z} + \alpha_m + \beta_b) \cdot \text{offset(Length)}. \tag{4}$$

Grouping	**Residual Deviance**	**R-squared**	**Residual Df**
Beach +cross term (season, days) and (season, settlements)	301.90	0.75	742
Beach	332.35	0.72	772
Beach by Settlements	332.35	0.72	772
Beach by Season.custom	203.52	0.81	666.27
Beach by Months	40.69	0.95	331.83

Table 4: Summary of fits with main criteria

The model explaining most of the variance is the one where *Beach* has been grouped by *Month*. The independent variables explain 95% of the deviance, as the residual deviance is **40**, this model highly over-fit. Therefore to not

[2]We constructed this part with the help of [6]

over-fit the grouping by season is preferable. There is no added benefit with grouping by settlements so we disregard this one. When *Beach* is not grouped, adding the cross terms yields a similar performance. We have decided to take the model with the cross terms, but we could have consider the one without it. The following table does suggest to consider the model without the cross terms. Considering the other way around, first without the cross terms and then with it, the Chi-square statistic also suggest to take the other model as its value is 1.3e-05.

Grouping	**Resid. Df**	**Resid. Dev**	**Df**	**Deviance**	**Pr(>Chi)**
Beach + cross term (season, days) and (season, settlement)	742.00	301.90			
Beach	772.00	332.35	-30.00	-30.45	< 2.2e-16 ***
Beach by Settlements	772.00	332.35	0.00	0.00	
Beach by Season.custom	616.59	203.52	155.41	128.83	< 2.2e-16 ***
Beach by Months	151.70	40.69	464.89	162.82	< 2.2e-16 ***

Table 5: Comparison of fit with Anova and Chi-squared statistic

2.2.1 GAM assessment

We assess the first model presented in Table 4 which we will refer to as model 1. The deviance reduces from 1540.78 for the null model (with 890 df) to 303.86 [3] (with 745.77). We estimate the over-dispersion parameter $\kappa \approx 0.415$ for the GAM and use it to obtain the Anova in Table 6.

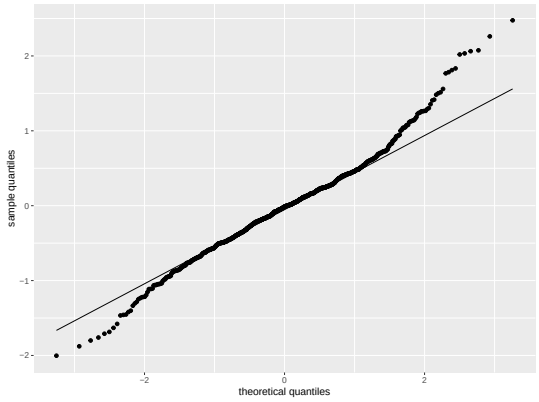
Term	**df**	**Chi.sq**	***p*-value**
Month	9	296.04	$< 2 \times 10^{-16}$
Season.custom	3	242.43	$< 2 \times 10^{-16}$
Day	6	10.48	0.106
Settlement	2	1.49	0.475
Season.custom:Day	18	33.35	0.015
Season.custom:Settlement	6	29.81	4.26×10^{-5}
Approximate significance of smooth terms			
s(Beach)	101.2	4469	$< 2 \times 10^{-16}$

Table 6: ANOVA table for the GAM model. Family: quasipoisson, Link function: log.

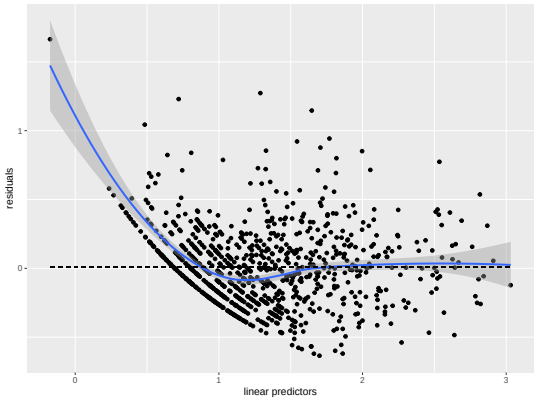
As we can see from the $p-$values of the χ^2 tests, the covariates *Month, Season.custom, s(Beach)* and the cross terms *Season.custom:Day, Season.custom:Settlement* significantly improve the goodness of fit. Additionally $\kappa < 1$, therefore we still have over dispersion despite the transformation, the random effect and the quasi Poisson fit. This dispersion translates in the assessment of the fit in Figure 8.

The residuals in Figure 8d are roughly normally distributed around 0 with a very slight long-right tail (probably some outliers). In other words, we sometimes make "very" wrong predictions. It is confirmed by Figure 8a where we can see that the fit is correct in the middle but we tend to underestimate values (lower quantile values of the samples than expected by the theoretical quantiles) and overestimate values (higher sample quantile values than expected).

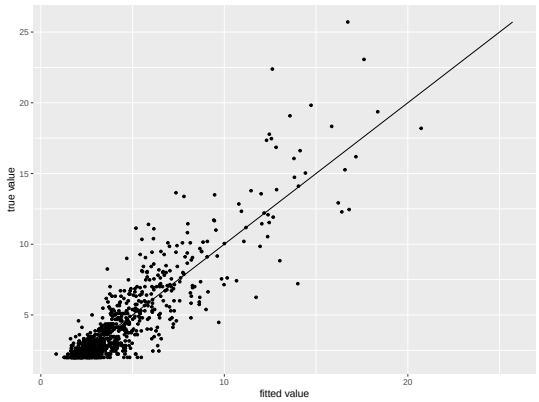
[3] The optimization method or parameters might differ from version to version. The one in Tables 4 and 5 were obtained with mgcv 1.9.1 package while in section 2.2.1 with version 1.8.42.



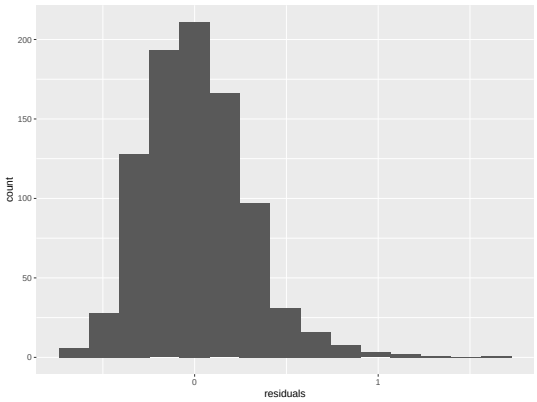
(a) QQ-plot



(b) Linear predictors against residuals



(c) Fitted values against response values



(d) Histogram of residuals

Figure 8: GAM assessment, using y_1

From Figure 8c, we can see the model "stays close" to the true responses, while still making quite large errors for large number of cigarette butts. Figure 8b reveals some heteroscedastic behavior for low linear predictors between 0 and 1. There is a significant outlier towards 0 that corresponds to the beach *sihl_leimbach_kruegerm* during Winter. However, for higher values of the linear predictors, the residuals are roughly uniformly distributed around 0 as the smooth line corroborates this observation. We can see that for very high values of the linear predictors, we have more variance in the residuals (see confidence interval). This may either be because of the overdispersion or the fact that we have few high values of y_1 and hence low confidence in the results.

In Figure 9, we represent the "importance" of the coefficients. Recall that we have taken the log as a link function and hence the coefficients should be interpreted by taking the exponential of the values given in the Figure 9. We represent the coefficients this way to better visualize their importance. In other words, the lower the coefficient, the lower it contributes to the number of cigarette butts. As we can see, colder months and seasons like O/N/D and January, February, March barely increase the predicted number of cigarette butts while in a city in A/M/J significantly increases the littering.

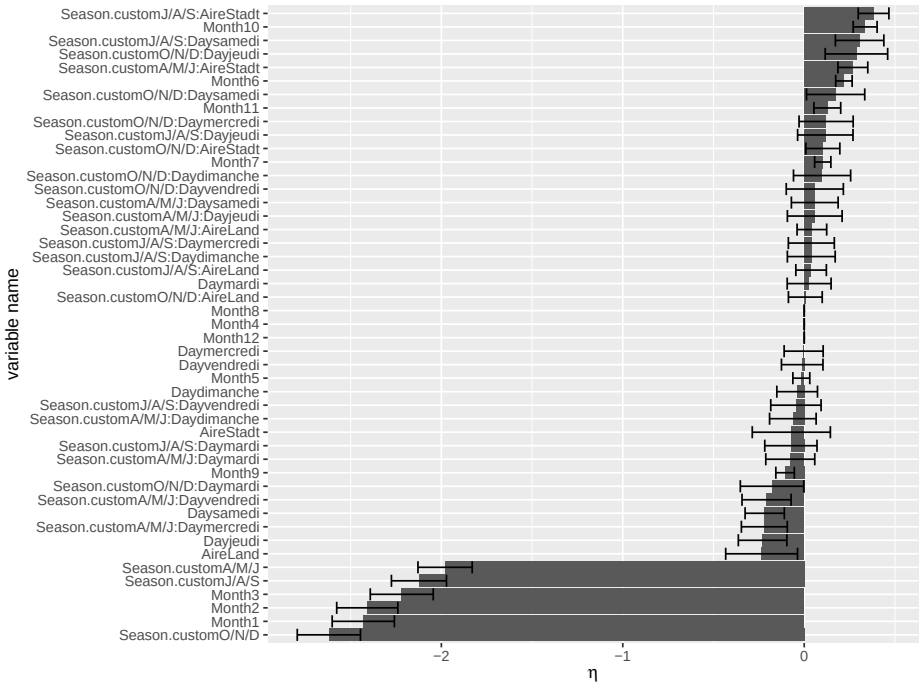


Figure 9: Multiplicative coefficients with standard errors of model 1, GAM

3 Predictions

For the reasons explained in the introduction , we assume that the weeks within a month are identically distributed. Hence, we only make predictions for a week in August. We make predictions using model 1 from Table 4, ie GAM with transformation y_1 on the response variable, random effects on the beaches, an offset on the length, crossed terms and quasi Poisson family with log link. Additionally, we want to compute a prediction interval PI on the predicted response variable y_+ for the beaches. For a given tolerance α, we want to find $\mathcal{U}, \mathcal{L}$ such that:

$$\mathbb{P}_{\mathcal{D}}(\mathcal{U}(X) \leq Y_1 \leq \mathcal{L}(X) \mid \lambda) \geq 1 - \alpha \tag{5}$$

where $Y_1 \sim \mathcal{D}(\lambda)$ for some $\lambda > 0$. We use a $(1-\alpha)\%$ basic prediction interval at $\alpha = 0.05$, i.e $y_+ \pm z_{\alpha/2}\sqrt{\Sigma_x}$ where Σ is the covariance matrix, x a point, $\Sigma_x = (\Sigma)_{xx}$ its variance and $z_{\alpha/2}$ quantile for normal distribution. In `R`, we estimate the standard error provided by the `mgcv::predict.gam` function [1]. We arbitrarily focus on the beach *Aare_bern_gerberm* in Bern and obtain Figure 10.

Notice that the PI computed is very basic and does not account for the over-dispersion of the data. To account for the dispersion, we can compute calibrated PI using the package `predint` [4] as presented by Taeho Kim et al. [2], section 5.3 combined with equation (10).

The final step is reverse back to the original scale. The transformation $\phi(y) := y_{1y} = 2\sqrt{1 + 0.25y}$ being bijective from $\mathbb{R}_+ \to \mathbb{R}_+$, we can simply apply its inverse $\phi^{-1}(y_1) = y_1^2 - 4$ to the predictions. Computing the PI requires a bit more effort as it requires to take into account the transformation.

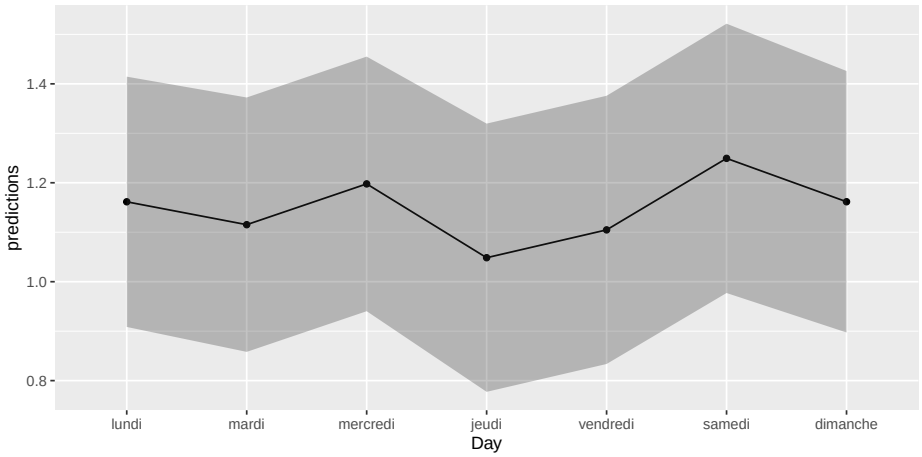


Figure 10: Prediction model 1 for days in August for beach *Aare_bern_gerberm*, 95% prediction interval, transformation y_1

4 Conclusion

After cleaning the data set and highlighting the relationship in between the covariates we have been able to construct two models with their respective strengths and flaws. The GLM one is more basic and interpretable. It predicts efficiently small values of cigarette butts but fails to account for values bigger than 3 for the transformation y_1 of y. However the GAM can account for these bigger values, it improved the fit and its flexibility, but it suffers in interpretability. The values obtain in Table 5 shows that our GAM is slightly over-fitting. To counter this, we could have implemented a resampling methods or a k-fold cross validation. The results were not presented but the fit is greatly improved when the values bigger than the threshold for the cook distance in the GLM are dropped.
Overall we have made a robust analysis and are able to present the predicted number of cigarette butts on any beach and day of August in a concise and clear manner as presented in Figure 10.

References

[1] Trevor Hastie. Generalized Additive Models. https://cran.r-project.org/web//packages/gam/gam.pdf. Accessed: 16.01.2024, version 1.22-3.

[2] Taeho Kim et al. "Prediction intervals for Poisson-based regression models". In: WIREs Computational Statistics 14.5 (2022). DOI: 10.1002/wics.1568.

[3] Richard Berk Æ John M. MacDonald. "Overdispersion and Poisson Regression". In: Journal of Quantitative Criminology 24 (2008), pp. 269–284. DOI: https://doi.org/10.1007/s10940-008-9048-4.

[4] Max Menssen. Predint R package. https://cran.r-project.org/web/packages/predint/index.html. Accessed 16.01.2024, version 2.2.0. Sept. 29, 2023.

[5] Summit Foundation. https://www.summit-foundation.org/. Accessed: 31.12.2023.

[6] Wood. Generalized Additive Models: An Introduction with r. CRC Press, 2006.