Assignment 2
Generalized Linear Models

Students s212460@dtu.dk, s212676@dtu.dk

Part A: Clothing insulation

Find suitable generalized linear models for clothing insulation level and compare the results. Also illustrate the differences between some of the models.

As with any data analysis task, the initial step is to take a look at the data set and perform some fundamental descriptive statistics, in order to get an understanding for the data. The dependent variable (clo) seems to be multimodal, indicating that a simple linear regression might not be sufficient to explain the data.

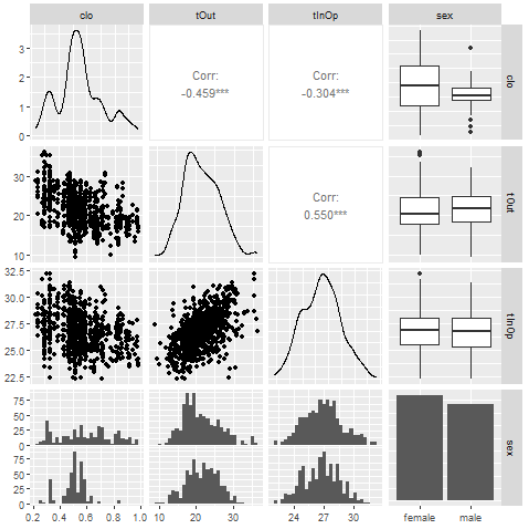
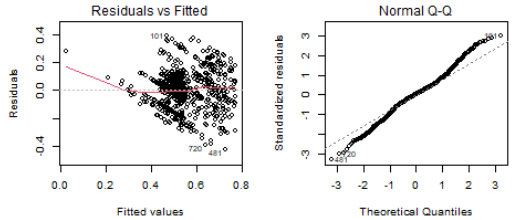


Figure 1: Pair plot between clo, tOut, tInOp and sex

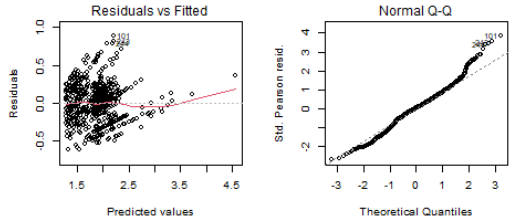
Our initial models are gaussian models with the identity link. However, even a quite extensive model formulation does not lead to a satisfactory fit. It is easy to see that the Gaussian distribution might not be the most appropriate one for describing the dependent variable. For instance, the insulation levels are non-negative, violating one basic assumption of normal data. Therefore, exploring other distributions seems beneficial to the analysis. One distribution that satisfies the aforementioned constraint is the Gamma. This family of distributions has also the characteristic that its variance increases as the mean increases, which might be beneficial particularly when modelling heteroscedastic data. The gamma density is parametrized in two terms, the mean μ and the shape parameter k. As seen in table 1, the gamma distribution is a more appropriate fit as indicated by the model performance.

Distribution	Formula	AIC
Gaussian	clo∼tOut+tInOp+sex	-917.9
Gaussian	clo∼(tOut+tInOp+sex)^2	-955.3
Gaussian	clo∼(tOut+tInOp+sex)^3	-958.2
Gaussian	clo∼poly(tOut,2)*sex*poly(tInOp,2)	-971.6
Gamma	clo∼tOut+tInOp+sex	-948.6
Gamma	clo∼(tOut+tInOp+sex)^2	-971.5
Gamma	clo∼poly(tOut,2)*sex*poly(tInOp,2)	-1007.4

Table 1: Model evaluation between Gaussian and Gamma distribution



(a) residual in gaussian model with second order polynomial



(b) residual in gamma model with second order polynomial

Figure 2: residuals of models

	Estimate	Std. Error	Pr(>t)
(Intercept)	1.97	0.05	0.00
poly(tOut, 2)1	11.69	1.52	0.00
poly(tOut, 2)2	7.63	1.16	0.00
sexmale	0.01	0.05	0.79
poly(tInOp, 2)1	0.01	1.39	1.00
poly(tInOp, 2)2	5.58	1.13	0.00
poly(tOut, 2)1:sexmale	-7.31	1.77	0.00
poly(tOut, 2)2:sexmale	-5.22	1.51	0.00
poly(tOut, 2)1:poly(tInOp, 2)1	-276.63	50.38	0.00
poly(tOut, 2)2:poly(tInOp, 2)1	-45.34	30.62	0.14
poly(tOut, 2)1:poly(tInOp, 2)2	25.96	29.62	0.38
poly(tOut, 2)2:poly(tInOp, 2)2	27.60	15.46	0.07
sexmale:poly(tInOp, 2)1	0.51	1.66	0.76
sexmale:poly(tInOp, 2)2	-4.57	1.48	0.00
poly(tOut, 2)1:sexmale:poly(tInOp, 2)1	284.06	57.19	0.00
poly(tOut, 2)2:sexmale:poly(tInOp, 2)1	66.69	42.58	0.12
poly(tOut, 2)1:sexmale:poly(tInOp, 2)2	-20.02	38.00	0.60
poly(tOut, 2)2:sexmale:poly(tInOp, 2)2	-18.91	31.41	0.55

Table 2: Parameter estimates for Gamma Model with second order polynomials

	LR Chisq	Df	Pr(>Chisq)
poly(tOut, 2)	129.16	2	0.00
sex	72.13	1	0.00
poly(tInOp, 2)	6.41	2	0.04
poly(tOut, 2):sex	9.82	2	0.01
poly(tOut, 2):poly(tInOp, 2)	11.22	4	0.02
sex:poly(tInOp, 2)	1.69	2	0.43
poly(tOut, 2):sex:poly(tInOp, 2)	26.48	4	0.00

Table 3: Type II Anova for Gamma model with second order polynomials

$$f(y;k,\mu) = \frac{(k/\mu)^k}{\Gamma(k)} e^{(-ky/\mu)} y^{(k-1)}, y \geq 0 \tag{1}$$

For which $E(y) = \mu$ and $var(y) = \mu^2/k$. The dispersion parameter $\phi = 1/k$ is treated as unknown, thus being estimated alongside $\boldsymbol{\beta}$. This leaves us with a analogous method of variable elimination to the ordinary linear model, by using F-statistics comparing deviances D of nested models M_0 and M_1 in the form.

$$\frac{[D(M_0) - D(M_1)]/(p_1 - p_0)}{D(M_1)/(n - p_1)} \; F_{p_1-p_0, n-p_1} \tag{2}$$

By fitting the Gamma GLM to the data and doing variable selection we end up with a much more parsimonious model when comparing to the Gaussian model. We notice the interactions are not statistically significant anymore considering the 95% confidence level. The model also yields a lower AIC, indicating a better goodness of fit as illustrated in table 1.

Make a residual analysis, as part of this you should investigate if the variation/dispersion in clothing level differs between men and women.

When looking at the residual plots of the gamma model shown in figure 2 (b), there are some noticeable comments to make. Most of the residuals concentrate towards the lower predicted values, but there seems to be neither a violation of homoscedasticity nor any overly influential observations. There is a slight tail towards the right in the Normal Q-Q plot, but otherwise the residuals seem fine.

We can check whether there is a difference in variation between men and women by comparing the residuals of both genders. By plotting the residuals of the last model split by sex, we can see that they indeed seem to show

different dispersion.

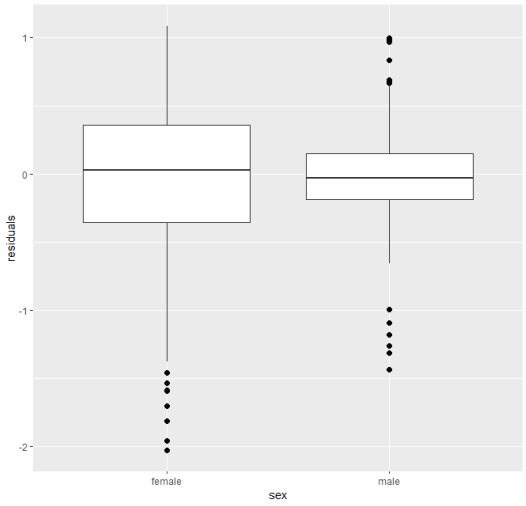


Figure 3: Box plots of residuals by gender

We could also use the F-test for equality of two variances. This test would be formulated as:

$$H_0 : \sigma_M^2 = \sigma_F^2$$
$$H_a : \sigma_M^2 \neq \sigma_F^2$$
$$\text{Test statistic} : s_M^2/s_F^2 \sim F_{N_m-1,N_F-1}$$

Ratio of vars	lower CI	Upper CI
0.431	0.354	0.525

Thus we reject the null hypothesis with confidence.

Give an interpretation of your model, including some graphical presentation.

Due to the higher order terms in the model formula, parameter interpretation becomes more challenging. As seen in table 3, practically all parameters are significant at $\alpha = 5\%$. There is a positive and significant intercept, indicating that the overall mean is relevant (people dress to some sort of insulation level).

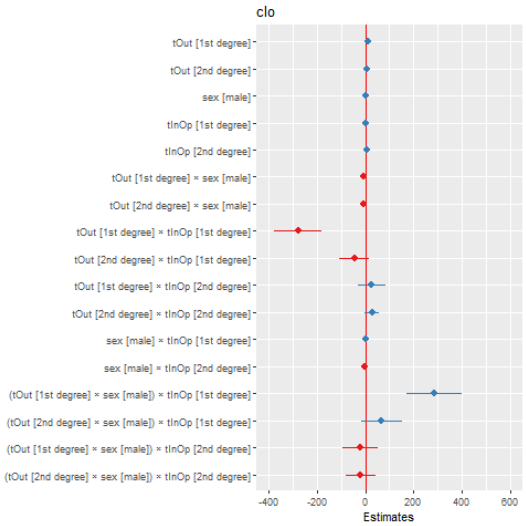


Figure 4: Model parameter plot

Fit the model but including subjId rather than sex and conclude.

Using the gamma model identified in the previous section, we replace the sex variable with with subjectID. The AIC is significantly lower, with -1859.60 instead of -1007.4 previously. This indicates that there is also a significant variation within the two groups, meaning individuals do not necessarily dress the same, irregardless if they belong to the same sex. However, the model now also has significantly more parameters, and instead of abstracting to the two groups, does now predict on the individual instead. Additionally, the Normal Q-Q plot indicates that the residuals are now a lot less evenly distributed, and the Residuals vs Fitted plot shows a decrease in variance.

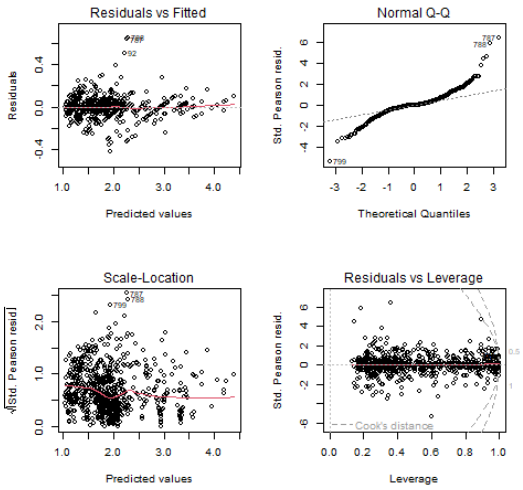


Figure 5: Residuals of gamma model with second order polynomials and interactions

Make a residual analysis that include analysis of within day auto-correlation.

The within day auto-correlation analysis can be conducted by grouping the residuals by subjectId and day of measurement. Now, we can compare the auto-correlation for the measurements of each subject, for each day. This yields an array of correlation values of lag(1). In order to assess if those are statistically different than zero we could perform a t-test in the following manner:

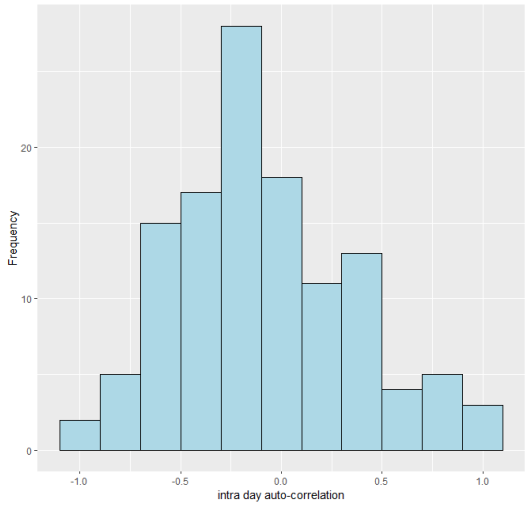


Figure 6: Intra-day auto-correlation on the residuals

$$H_0 : \mu = 0$$
$$H_a : \mu \neq 0$$
$$\text{Test statistic} : \frac{\bar{x} - 0}{s/\sqrt{n}} \sim t_{n-1}$$

mean of x	lower CI	Upper CI	p-value
-0.9	-0.17	-0.011	0.0245

thus we can reject the null hypothesis with the given α of 95%. This result hints towards the inadequacy of the model for the data. Models that account for the covariance structure of the measurements could be more appropriate for the analysis.

Set up a model that estimate the optimal weight/dispersion parameter for men and women, also make a profile likelihood plot for this parameter.

This can be done in a similar way as the last part of Assingment 1. Now, instead of optimizing the fit of a linear combination of the variables $\boldsymbol{X\beta}$ as the mean μ of a normal density, and the standard deviation σ_i for each lab i, now we use the Gamma density. This time, also, the link between the linear combination of variables and the location parameters is not the identity. We decide to use the canonical link $1/x$ as while fitting all the gamma GLMs in the analysis. Following those lines, optimizing the $\boldsymbol{\beta}$ vector will be done for each possible combination of ϕ_M and ϕ_F on a pre-defined grid of dispersion parameters. The optimization of $\boldsymbol{\beta}$ will be done by optimizing the data fit by minimizing the negative log-lokelihood of those parameters, under different dispersion parameters for each gender. The optimx function in R is used in each iteration. For the dgamma function in R, we can assume the following parametrization:

$$f(y) = \frac{1}{\sigma^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-y/\sigma} \tag{3}$$

Thus, for different dispersion parameters $\phi_i = 1/\sigma_i$ for each gender we have the log-likelihood as:

$$\ell(\hat{\boldsymbol{\beta}}, \sigma_M, \sigma_L | \boldsymbol{y}) = \sum_{i \in A} n_i(\alpha_i - 1) log\bar{g}y_i - n_i log\Gamma(\alpha_i) - n_i\alpha_i log\sigma_i - n_i\bar{y}_i/\sigma_i, A = \{M, F\} \tag{4}$$

$$\text{where, } \eta_i = 1/(\boldsymbol{X_i\beta}) \text{ (canonical link)}$$
$$\text{and } \alpha_i = \eta_i/\sigma_i \text{ (parametrization)}$$

Where n_i is the number of observations in the dataset for each gender i, X_i and y_i are the design matrix and observed response variables for gender i, respectively.

Algorithm 1 Profile likelihood

Output: (log likelihoods for each combination of σs)

$results \leftarrow$ new array of size NxN
$i \leftarrow 0$
for $\sigma_M \leftarrow 1$ to N **do**
    for $\sigma_F \leftarrow 1$ to N **do**
        $i \leftarrow i + 1$
        $fit \leftarrow Optimize\ \ell(\hat{\boldsymbol{\beta}}, \sigma_M, \sigma_L | \boldsymbol{y})$
        $results[i] \leftarrow (\sigma_M, \sigma_F, fit.objective)$

        end for
        return $results$

The final results array contains all combinations of σ_M and σ_F and the respective negative log likelihood. The results are from the simple gamma model with no interactions. The values are consistent with figure 3 and the results from the F-test to compare the residual variance of the different genders. The dispersion from females is again shown to be larger than the male one. A profile likelihood contour plot is displayed showing the minimum value and the respective parameter combination that yields it.

ϕ_M	ϕ_F	neg. Log Likelihood
0.0159	0.052	-534

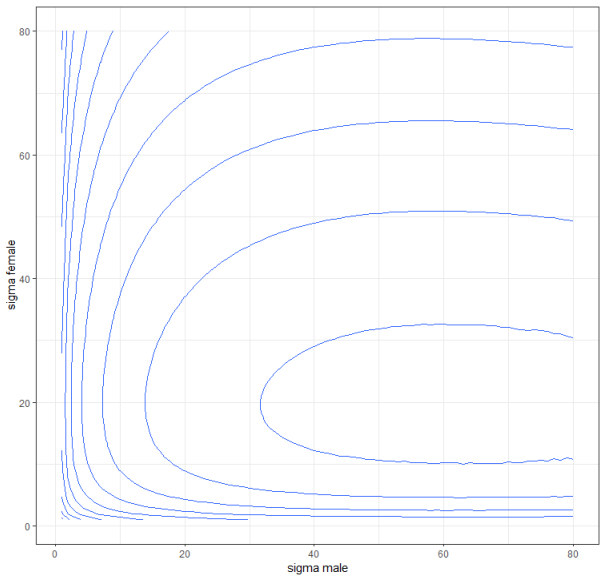


Figure 7: Profile likelihood contour plot

Write a small conclusion of your findings.

Due to the nature of more complex models, it gets more challenging in interpretation. Nonetheless, one can look at the fitted model for to generate understanding. We conclude that Men and Women have different variances in insulation level. The variation between individual seems more important than the variation between men and women (better model performance for subjID than sex). Additionally, there is a intra-day correlation. Intuitively this makes sense, as subjects are unlikely to change clothes while out during the day, but the temperature can change.

There is a significant interaction between outside and inside temperature in regards to clo, as well as a significant three way interaction when including sex.

Part B: Ear infections

Describe the content of this dataset in words.

The data set contains 24 observations of different groups of swimmers. These are differentiated between the *location* of swimming, the *age* within the group, and the corresponding *sex*. The dependent variable is the number of reported infections within a given group. Notably, the group size varies in the data set. In total 287 swimmers have been observed, with a total of 136 self reported ear-infections.

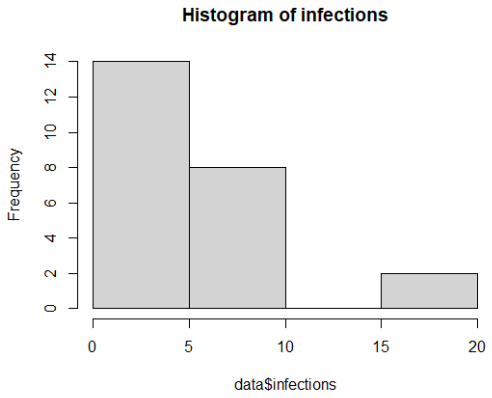


Figure 8: Histogram of total number of ear infections

Test	pval
$G^2(H_0\|H_1)$	0.621
$G^2(H_0)$	0.084
$G^2(H_1)$	0.020

Table 4: Analysis of Goodness of Fit between H0 and H1

Explain why a linear model would be inappropriate.

First, the distribution of the response variable y_i (infections) is not continuous, hinting to the inadequacy of the normal assumption. In some cases however, if we are dealing with high counts, and values substantially above zero, the normal distribution might be a decent approximation. In our case however, none of these requirements are satisfied, as shown is the histogram plot. The data thus hints towards a specific model for count data, such as the Poisson loglinear model. This has been chosen in favor of the binomial distribution, as one swimmer can have more than 1 infection in a year, thus the response is not bounded by the number of people in the group, even though the ratio of infections is between 0 and 1 in this particular instance.

Explain why a model with an offset might be appropriate.

The number of people in each group varies greatly, with $\mu = 12, \sigma^2 = 58.46$, the smallest group consisting of 4 and the largest group of 32 people. Even if all groups were to have the same likelihood of getting infected ($H_0 = true$), one would expect more infections in a larger group. This means, that the response count variable y_i (*infections*) is proportional to the population size t_i (*persons*). This dependency is referred to as an offset, and as such a model with an offset might be appropriate in our case. In this case, the sample rate y_i/t_i, has expected value μ_i/t_i. The respective loglinear model with explanatory variables then assumes the form.

$$log(\mu_i/t_i) = \sum_{j=1}^{p} \beta_j x_{ij} \tag{5}$$

Due to log properties the $log(\mu_i/t_i) = log(\mu_i) - log(t_i)$ thus we can see the offset is an adjustment of $-log(t_i)$ to the log of the mean.

Fit a full model. What can you say about its goodness of fit (explain).

Our initial full model encompasses an offset as reasoned in the previous section plus two-way interactions. To assess model sufficiency, we use residual deviance as our test-statistic:

$$\frac{D(\boldsymbol{\mu}(\hat{\beta}), \hat{\boldsymbol{\mu}}_{\boldsymbol{null}})}{k-1} \sim \chi^2(k-1) \tag{6}$$

We reject the hypothesis that we can reduce the model H_1 to the null-model H_0 for large values of the test statistic. Using the equation above, we calculate the Goodness of Fit as displayed in table 4. When looking at the unconditional goodness of fit for H_0, we reject the hypothesis that the model is sufficient at the 5%-level. The full model would be sufficient, however, the distance between the two models is not deemed significant. This indicates that there might be a model between the H_0 and H_1, which is a more appropriate fit.

Try to reduce this model by successive likelihood ratio tests. Explain how you proceed to compare two models.

For reducing the model we can use Likelihood-ratio model comparison using deviance differences. For the standard poisson model, where $\phi = 1$ is assumed, the deviance D reduces to:

$$D(y; \hat{\boldsymbol{\mu}}) = -2[L(\hat{\boldsymbol{\mu}}; y) - L(y; y)] = 2\sum_i y_i log(y_i/\hat{\mu}_i) \tag{7}$$

For deviance differences from two nested models we then have:

$$D(y; \hat{\boldsymbol{\mu}}_0) - D(y; \hat{\boldsymbol{\mu}}_1) = \sum_i y_i log(\hat{\mu}_{1i}/\hat{\mu}_{0i}) \tag{8}$$

	Estimate	Std. Error	Pr(>z)
(Intercept)	-0.9078149	0.2128480	0.0000200
locationNonBeach	0.0177664	0.3074703	0.9539219
sexMale	-0.2783041	0.2649529	0.2935385
swimmerOccas	0.2082634	0.1724761	0.2272429
locationNonBeach:sexMale	0.5542410	0.3809972	0.1457492

Table 5: Parameter estimates of final model

	Estimate	Std. Error	Pr(>t)
(Intercept)	-0.9078149	0.1177785	0.0000003
locationNonBeach	0.0177664	0.1701374	0.9179275
sexMale	-0.2783041	0.1466105	0.0729602
swimmerOccas	0.2082634	0.0954389	0.0418573
locationNonBeach:sexMale	0.5542410	0.2108232	0.0165305

Table 6: Parameter estimates using Quasi-Poisson

With an approximate chi-squared null distribution with $df = p_1 - p_0$ where p_i is the number of explanatory variables in each model. This way we are able to compare the nested models performing model reduction until we check the 'distance' between two models are not significant to the chosen confidence level.

In the standard Poisson model, the response variable y is assumed to have expected value $E(y) = \mu$ and variance $var(y) = \mu$. In many cases, this seems like an unreasonable assumption to make. This happens due to the fact that under the standard assumption the dispersion parameter $\phi = 1$, thus the mean variance relation $v(\mu_i) = \phi\mu_i$ is reduced to the equality. There are some ways of dealing with that such as using the Negative-binomial GLM, or, as we will do, assume ϕ is unknown. This can be done using the Quasi-Likelihood approach estimating the dispersion parameter as:

$$\hat{\phi} = \frac{1}{n-p}\sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = X^2/(n-p) \tag{9}$$

Where n is the number of observations, p is the number of explanatory variables in the model and X^2 is the Pearson statistic. However, now that the ϕ parameter is also estimated, the test statistic doesn't follow a chi-squared distribution anymore, now following the F-distribution, similarly to other models in which the dispersion parameter is estimated by default, such as the Gaussian and the Gamma models. This situation is called under-dispersion, when $\phi < 1$ and over-dispersion when $\phi > 1$.

Report your best model (formula, goodness of fit).

When looking at the estimated coefficients of the model, we have assumed the dispersion parameter to equal to one. However, The likelihood of getting an infection varies between people (even within the same group). When looking at the dispersion parameter ϕ, the GLM function estimates it to be 0.3061914, and thus being under-dispersed. The goodness of fit for our final model is calculated using equation . The unconditional test has a p-Value of *0.002131307*, indicating that it is a better fit than the previous model H_1. Testing the conditional goodness of fit $G^2(H_0|H_b)$ the test still rejects at a significance level of 5%, however would fail to reject at the 10%-level (p-Value = 0.08013). This is a substantial improvement in comparison to $G^2(H_0|H_1)$

Looking at the parameter estimates for the model, we need to be aware that the estimates are in the log-domain (canonical link for poisson). The mean infection rate (independent of the offset) $\gamma - log(t_i) = exp(\beta_1)$ given by the parameter estimate of the intercept is approximately 0.40. Moreover, one can conclude that men have fewer ear infections than women. The location of swimming does not have any impact by itself. However men swimming on Non-beach locations are more likely to get an infection. Surprisingly, occasional swimmers are at a higher risk of getting an infection compared to regular swimmers. In order to understand the mechanics of this in detail, one would need to consult domain experts. However, there either might be an actual biological phenomenon at play (such as frequent swimmers develop an immunity), or the data has a certain bias (over-representation of occasional swimmers).