02424 - Advanced Dataanalysis and Statistical Modelling
Assingment 2

Tymoteusz Barcinski - s221937
Soren Skjernaa - s223316

April 10, 2023

Contents

1 Introduction

All tests in the following two sections are carried out using a significance level of $\alpha = 0.05$

2 Part A: Clothing Insulation level

The purpose of the following section is to establish a model for the variation in clothing insulation level, measured at an experiment conducted at the Laboratory of Occupant Behavior, Satisfaction Thermal comfort and Environmental Research (LOBSTER). We will split the model building into two parts.

First we will investigate four different generalized linear models for the clothing insulation level (clo) using the independent variables sex, indoor operating temperature (tInOp) and outdoor operating temperature (tOut).

Next we will investigate the effect of changing the independent variable sex, with a independent variable modelling the subject level.

2.1 Description of experiment and an initial look at the data

The experiment was carried out at the LOBSTER, where the clo was measured for 47 different subjects over multiple days. Each subject participated for 1-4 days (with most subjects participating for 3 days), and had 2-6 measurements recorded each day (with most having 6 recordings per day). All in all the data set consists of 803 observations with no missing data, and the following variables were recorded for each observation:

Variable	**Type**	**Explanation**
clo	Clothing insulation level	Positive variable, with higher values implying higher insulation.
tOut	Outdoors air temperature	Measured in C^o
tInOp	Indoor operating temperature	Measured in C^o
sex	Sex	Female/male.
subjId	Subject ID	Uniqiue ID for each subject.
time	Time	Time difference since last observation for the subject.
day	Day	Number of experimentation day for the subject.

Table 1: Overview of the variables recorded.

The main variables of interest are the outcome clo, and how it is affected by the temperature variables, and the sex and subject variables.

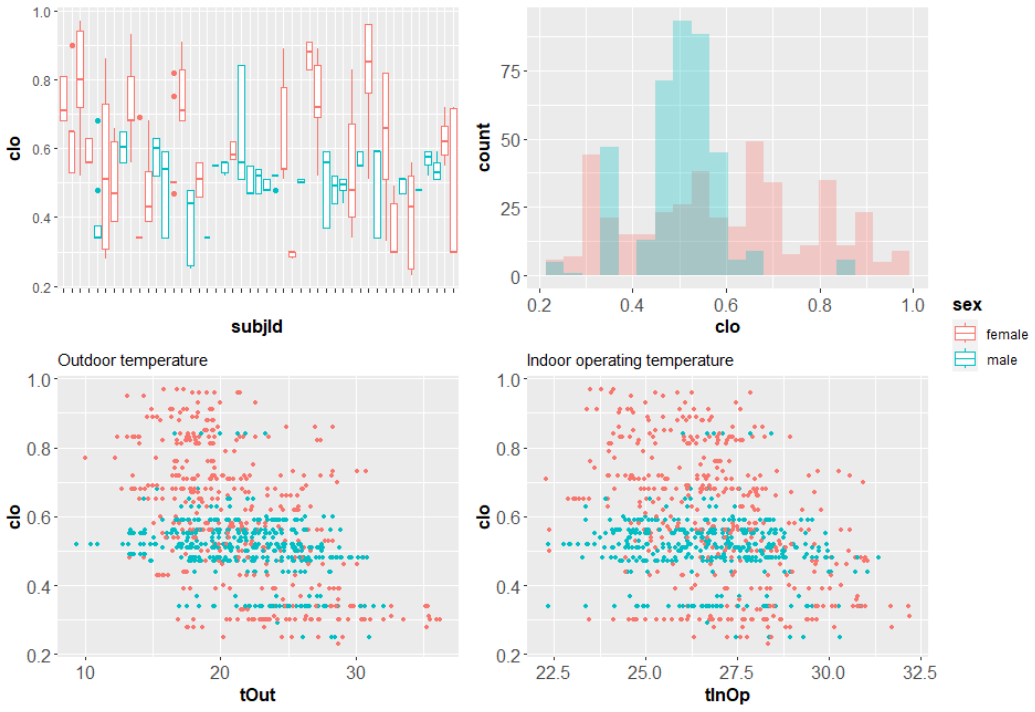

Figure 1: Plots of clo against the different independent variables.

Figure 1 shows how the outcome, clo is distributed for different levels of the independent variables. The top-left plot shows that there is a large subject-to-subject variation in the distribution of clo. We also see that for some subjects, there is no within-subject variation, while for others there is a large within-subject variation. From all four plots, we clearly see that there is a larger variation for the clo measurements for female subjects than for male subjects. Furthermore, it seems like the clo level is lower for the male subjects compared to the female subjects. From the two bottom plots, it is not clear if clo depends on the two temperature measurements, although there might be a tendency of clo decreasing when the temperatures increase.

2.2 Experimentation with 4 models based on sex and temperatures

In this section, we will focus on modelling clo based on the two temperature variables and the sex variable.

2.2.1 Model descriptions

We will investigate four different generalized linear models, in order to determine how best to model the variation in the data. Common for all four models is that we will use the following linear dependency for the linear predictor

$$\begin{aligned}\eta_i = \beta_0 &+ \beta_1(\text{sex}_i) \\ &+ \beta_2(\text{sex}_i)\cdot \text{tOut}_i + \beta_3(\text{sex}_i)\cdot \text{tOut}_i^2 \\ &+ \beta_4(\text{sex}_i)\cdot \text{tInOp}_i + \beta_5(\text{sex}_i)\cdot \text{tInOp}_i^2 \\ &+ \beta_6(\text{sex}_i)\cdot \text{tOut}_i\cdot \text{tInOp}_i \\ &+ \beta_7(\text{sex}_i)\cdot \text{tOut}_i\cdot \text{tInOp}_i^2 \\ &+ \beta_8(\text{sex}_i)\cdot \text{tOut}_i^2\cdot \text{tInOp}_i \\ &+ \beta_9(\text{sex}_i)\cdot \text{tOut}_i^2\cdot \text{tInOp}_i^2\end{aligned} \tag{1}$$

with $i = 1, 2, \ldots, 803$. Thus we model the linear predictor using second-order polynomials in the two temperature variables (and their interactions), and with different coefficients depending on the sex. As the outcome is positive, we choose to investigate the following four generalized linear models:

1. The general linear model with no transformation of the outcome clo. The assumptions are that
$$Y_i = \eta_i + \epsilon_i$$
where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. error terms.

2. The general linear model with a square root transformation of the outcome clo. The assumptions are that
$$\sqrt{Y_i} = \eta_i + \epsilon_i$$
where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. error terms.

3. The generalized linear model using a Gamma distribution with an inverse link function (the canonical link). The assumptions are that $Y_i \sim \text{Gamma}\left(\phi, \frac{\mu_i}{\phi}\right)$, where $Y_1, Y_2, \ldots, Y_{803}$ are mutually independent, the shape parameter $\phi > 0$ is a constant precision parameter and
$$\frac{1}{\mu_i} = \eta_i$$

4. The generalized linear model using a Gamma distribution with a logarithmic link function. The assumptions are that $Y_i \sim \text{Gamma}\left(\phi, \frac{\mu_i}{\phi}\right)$, where $Y_1, Y_2, \ldots, Y_{803}$ are mutually independent, the shape parameter $\phi > 0$ is a constant precision parameter and
$$\log(\mu_i) = \eta_i$$

Note that the independence assumption of the error terms for model 1. and 2. and the independence assumption for the observations for model 3. and 4. are most likely violated, as we are dealing with repeated measurements for several subjects. Furthermore, note that models 1-2 assume that the dependent variable takes values on the entire real line, where as the clo variable is strictly positive. Thus we would expect clo values for the same subject to be correlated. A more correct modelling choice would probably be to treat the subjects as random effects, with a time dependent correlation between observations from the same subject, such that measurements taken closer together in time are more correlated. This is however outside the scope of this project, and we shall not investigate this further, except examine if there is an autocorrelation of the observations from the same subject on the same day.

2.2.2 Model diagnostics

Fitting the four models, we first perform model diagnostics for all the models, to note if any of the models clearly do not fit the data. For all the models, we extract the standardized deviance residuals and look at the following 6 plots:

1. Standardized deviance residuals against the fitted values $\hat{\mu}_i$.
2. QQ-plot of the standardized deviance residuals against the theoretical $\mathcal{N}(0,1)$-quantiles.
3. Standardized deviance residuals against sex.
4. Standardized deviance residuals against tInOp.
5. Standardized deviance residuals against tOut.
6. Studentized deviance residuals against the leverage of the observations. A cut off for possible outliers of ± 2 is used for the residuals, where as a cut off of $2 \cdot$ "number off parameters" / "number of observations" is used for the leverage for detecting influential observations as recommended in (Conradsen et al., 2019).

For all six plots, the observations belonging to males and female are clearly marked, with black observations being female and red observations being male.

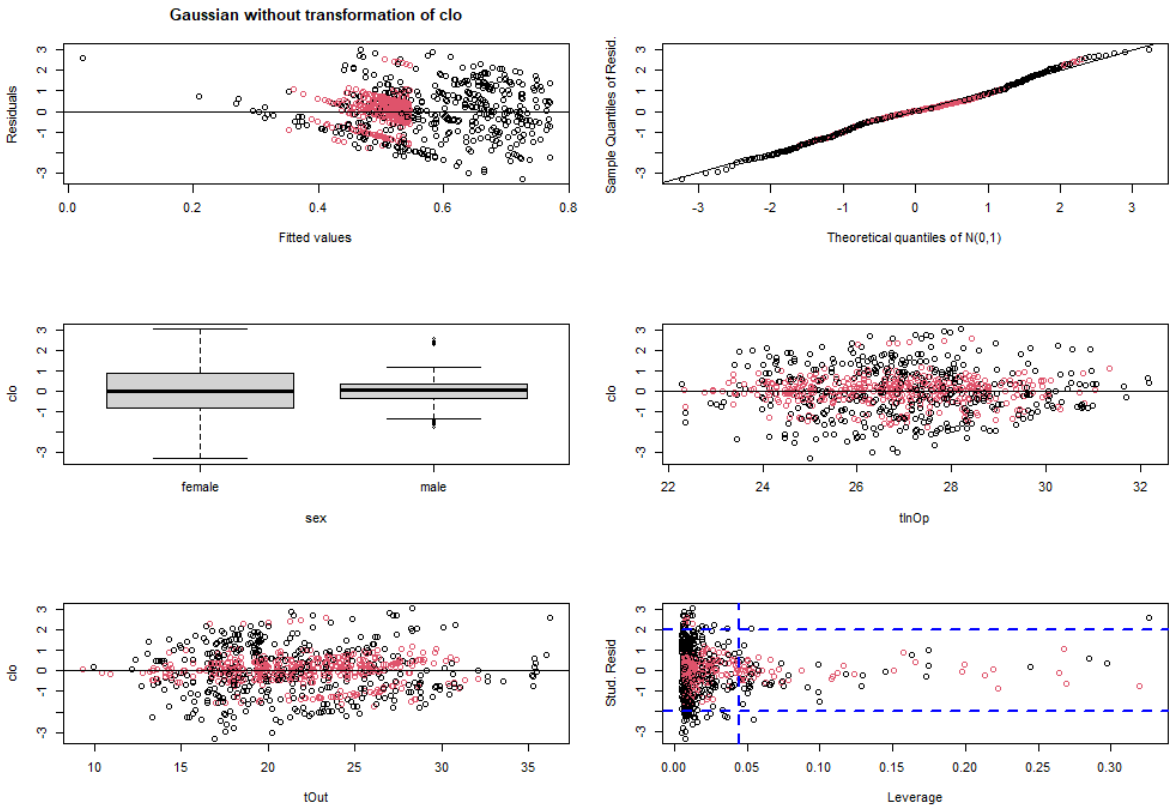

Figure 2: Model diagnostic plots for model 1.

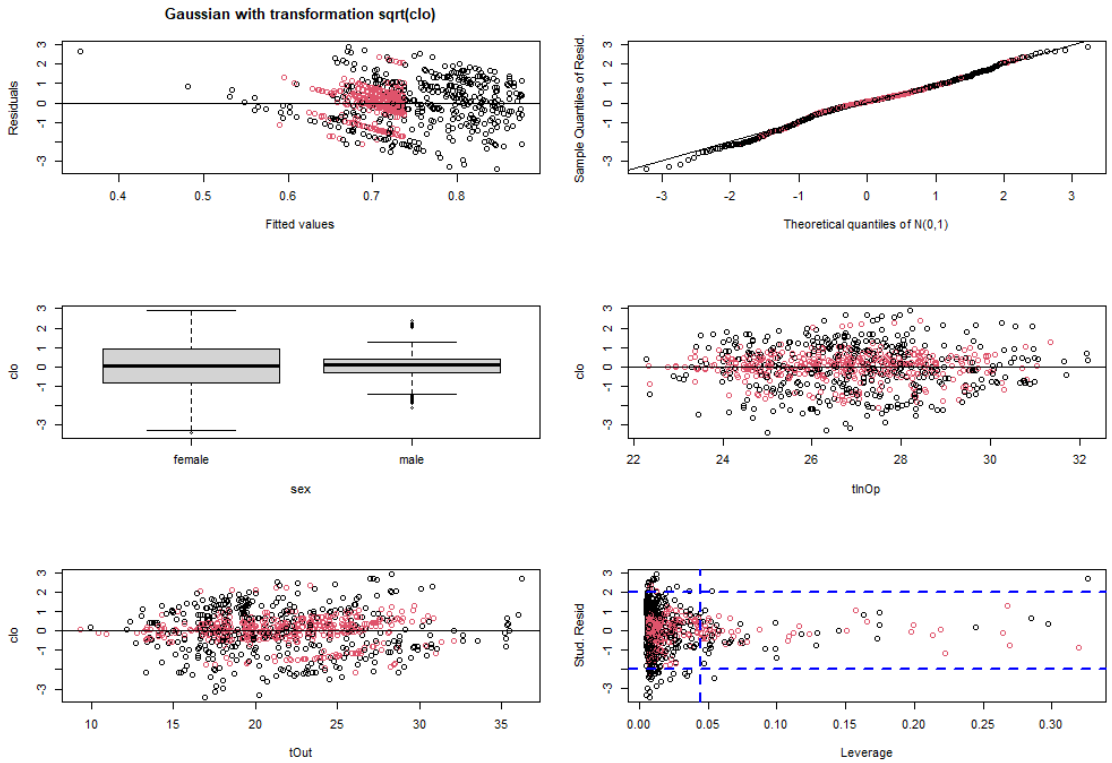

Figure 3: Model diagnostic plots for model 2.

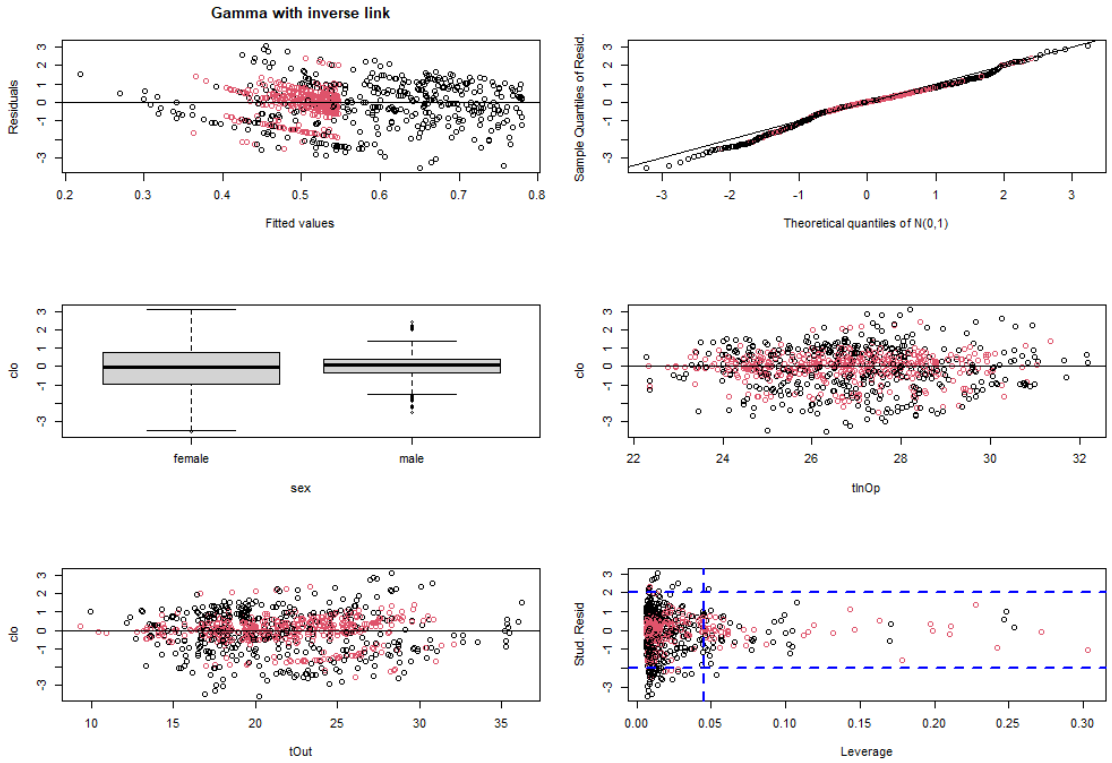

Figure 4: Model diagnostic plots for model 3.

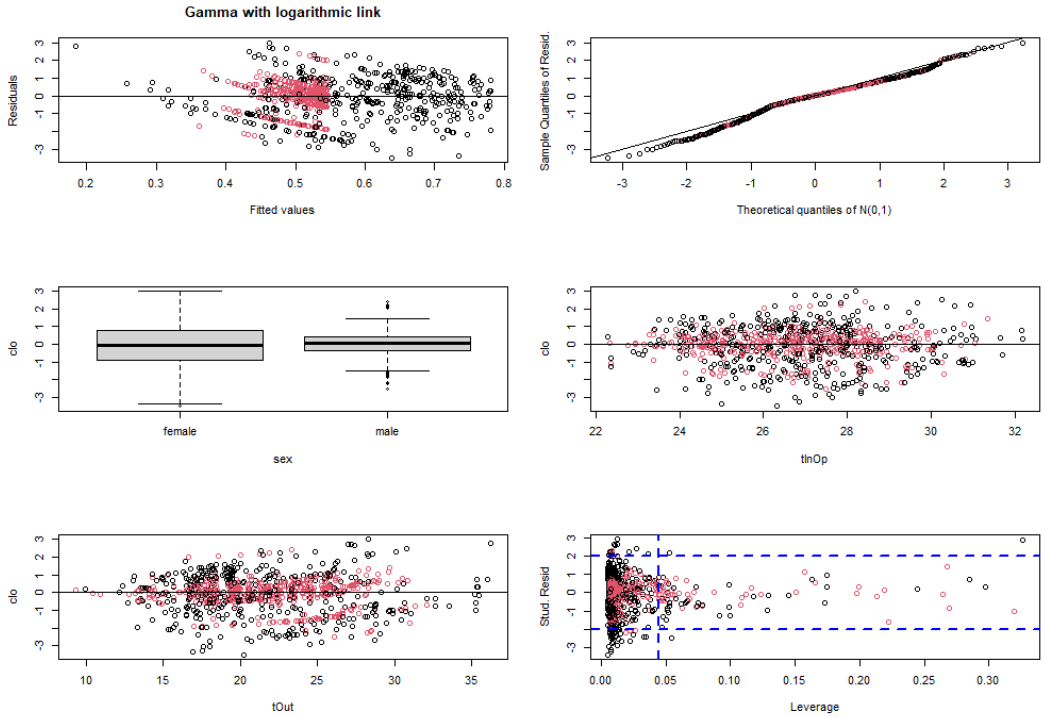

Figure 5: Model diagnostic plots for model 4.

Looking at the six plots we draw the following conclussions:

1. Looking at plot 1 for the four different models, we see that for the two gaussian models there is a clear shape to the distribution of the residuals against the fitted values, with an increase in the fitted value, leading to a decrease in the residuals. This shape is also notable for the two gamma models, although it is much less pronounced. Furthermore, we note that the range of the fitted values for males is much lower than the range for females, with the male subjects generally having a lower fitted value.

2. For the QQ-plots we see that there is no problem with assuming a $\mathcal{N}(0,1)$ distribution of the standardized deviance residuals. However, we do note that the QQ-plot for the gaussian models better follow a straight line.

3. From plot 3, it is clear that there is a difference in size of the residual variance for males and females. This trend is pronounced throughout all the plots. This indicates that we should look at a model, with differing dispersion parameters for males and females. We will estimate such a model, in the end of this part.

4. The model used for clo dependence on tInOp, seems to contain no systematic errors, as the residuals are evenly distributed for all four models.

5. The same conclusion holds for clo dependence on tOut.

6. For the residuals against leverage plot, we generally see the same picture for all four models. With 803 observations, we would expect 37 observations with a residual numerically greater than 2. If we count the number of residuals numerically greater than 2 we find between 51-59. Thus the models have more outliers than expected. Looking at the leverages of the observations we see that according to the chosen cutoff a lot of the observations are considered influential.

Overall we draw the conclusion that it may be more correct to use one of the two gamma models, as there is less systematic pattern to the deviance residuals against the fitted values. However, the systematic shape to the plot of the residuals against the fitted values, and the large amount of outliers indicate that it may not be reasonable to model clo using only sex and the two temperature variables. As mentioned earlier, this may be due to the repeated measurements violating the independence assumptions underlying the four models and/or a possible difference in the dispersion between the two sexes, which we investigate later on.

2.2.3 Model reduction

We first compare the initial models, in order to find the best candidate for a final model. As none of the four models are nested within each other, we compare the models using their AIC. As model 2 is a model for $\sqrt{\text{clo}_i}$ we need to take this into account when comparing the four AIC. As AIC is defined as

$$\text{AIC} = 2k - 2\log(\hat{L})$$

with k being the number of parameters in the model and $\hat{L}$ being the maximized likelihood, we need to adjust the AIC for model two by subtracting $-2\sum_{i=1}^{803}\log\left(\frac{d}{d\,\text{clo}_i}\sqrt{\text{clo}_i}\right)$, as transforming the observations clo_i correspond to multiplying the likelihood by the jacobian of the transformation. Taking this transformation into account we find the following AIC

	Model	**AIC**
1	Gaussian with no transformation	-971.6
2	Gaussian with square root transformation	- 995.6
3	Gamma with inverse link	-1007.4
4	Gamma with logarithmic link	-1001.1

Table 2: AIC for the four models.

Based on this table we choose to proceed with our model building using model 3, i.e. the Gamma model with inverse link (the canonical link).

Guided by a t-test for setting the individual parameters $\beta_j = 0$ we suspect that a suitable sub-model might be given by using the following linear predictor

$$\begin{aligned}\eta_i = \beta_0 &+ \beta_1(\text{sex}_i) \\ &+ \beta_2(\text{sex}_i)\cdot\text{tOut}_i + \beta_3(\text{sex}_i)\cdot\text{tOut}_i^2 \\ &+ \beta_4(\text{sex}_i)\cdot\text{tInOp}_i + \beta_5(\text{sex}_i)\cdot\text{tInOp}_i^2 \\ &+ \beta_6(\text{sex}_i)\cdot\text{tOut}_i\cdot\text{tInOp}_i\end{aligned} \tag{2}$$

i.e. we reduce model (1) by only keeping the first order interactions of the two temperatures (but still keep a sex specific interaction between them). Performing the likelihood ratio test of the two models, we find a test statistic of 0.268 and comparing this to the $\chi^2(6)$-distribution we find a p-value of 0.549. Thus, the data do not reject the sub-model. A type II partitioning of the model deviance for further model reduction using the likelihood ratio test, gives us the following ANOVA table

	Df	Deviance	scaled dev.	Pr(>Chi)
Model (2)		45.68		
sex:poly(tInOp, 2)	2	46.11	7.89	0.0193
sex:poly(tOut, 2)	2	46.84	21.49	0.0000
sex:poly(tInOp, 1):poly(tOut, 1)	1	46.92	22.89	0.0000

Table 3: ANOVA table for further reductions

so no further reduction of the model can be done.

2.2.4 Model interpretation

The final model is given by the linear predictor (2). Table 4 gives the model parameters, and their 95% profile likelihood intervals.

	Estimate	Std. Error	2.5 %	97.5 %
(Intercept)	1.94	0.04	1.86	2.02
sexmale	0.05	0.05	-0.05	0.15
poly(tInOp, 2)1	1.25	0.80	-0.33	2.82
poly(tInOp, 2)2	4.96	1.03	2.97	7.01
poly(tOut, 2)1	10.08	0.93	8.27	11.93
poly(tOut, 2)2	7.13	1.12	4.97	9.37
sexmale:poly(tInOp, 2)1	-0.71	1.14	-2.94	1.52
sexmale:poly(tInOp, 2)2	-3.89	1.39	-6.62	-1.17
sexmale:poly(tOut, 2)1	-5.61	1.29	-8.15	-3.08
sexmale:poly(tOut, 2)2	-4.87	1.46	-7.75	-2.01
poly(tInOp, 1):poly(tOut, 1)	-222.47	40.56	-303.06	-144.10
sexmale:poly(tInOp, 1):poly(tOut, 1)	223.15	47.33	130.96	316.50

Table 4: Model parameters and their uncertainties for model (2).

Note that we have used orthogonal polynomials in order to fit the model. Furthermore, note that the parameters: "sexmale", "poly(tInOp, 2)1" and "sexmale:poly(tInOp, 2)1" does not seem to be significant based on a t-test for setting the parameter equal to zero (as the parameter confidence intervals includes zero).

The estimated dispersion for the model is 0.0539.

As the model contains interactions, interpreting the parameters directly is challenging. However, figures 6, 7 and 8 gives us a clearer picture of the model.

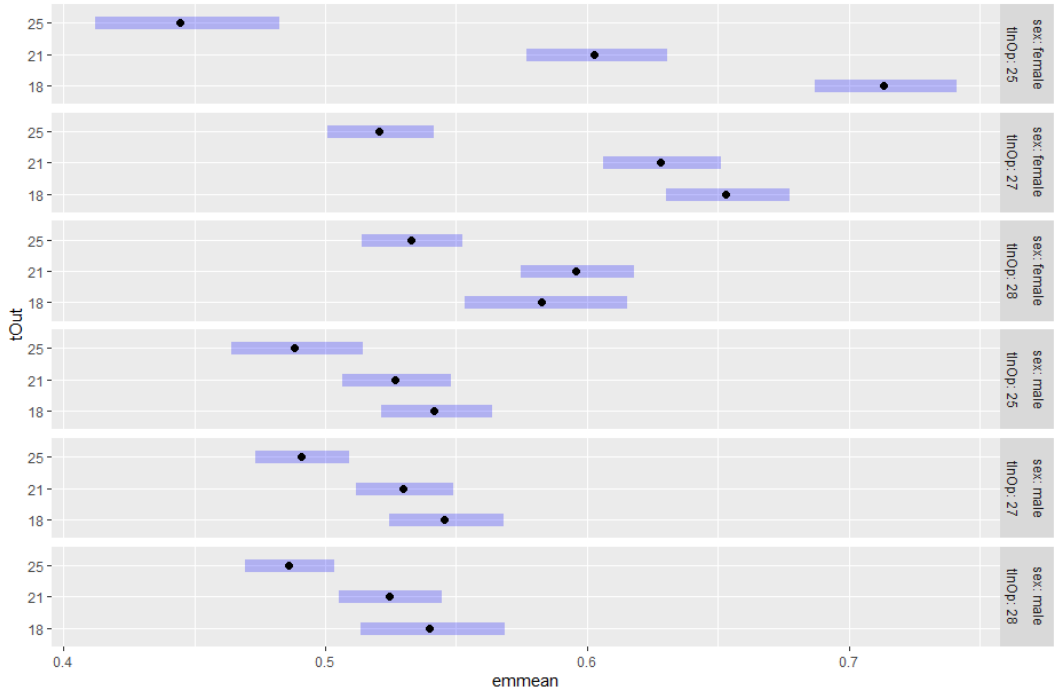

Figure 6: Estimated marginal models for levels of the temperatures and sex. The 1., 2. and 3. quantiles of the temperatures is used for the levels.

Looking at the estimated marginal means, we see a similar pattern for the confidence intervals for the male subjects for the three different indoor temperatures. Thus the indoor temperature has little effect for men. This is also seen by the three regression lines in figure 8, being almost identical for the men. For the outdoor temperatures, we see the emmean for tOut = 25 is lower, than the others, while the emmeans for tOut = 21 and tOut = 18 have a large overlap in their confidence intervals, implying no significant difference.

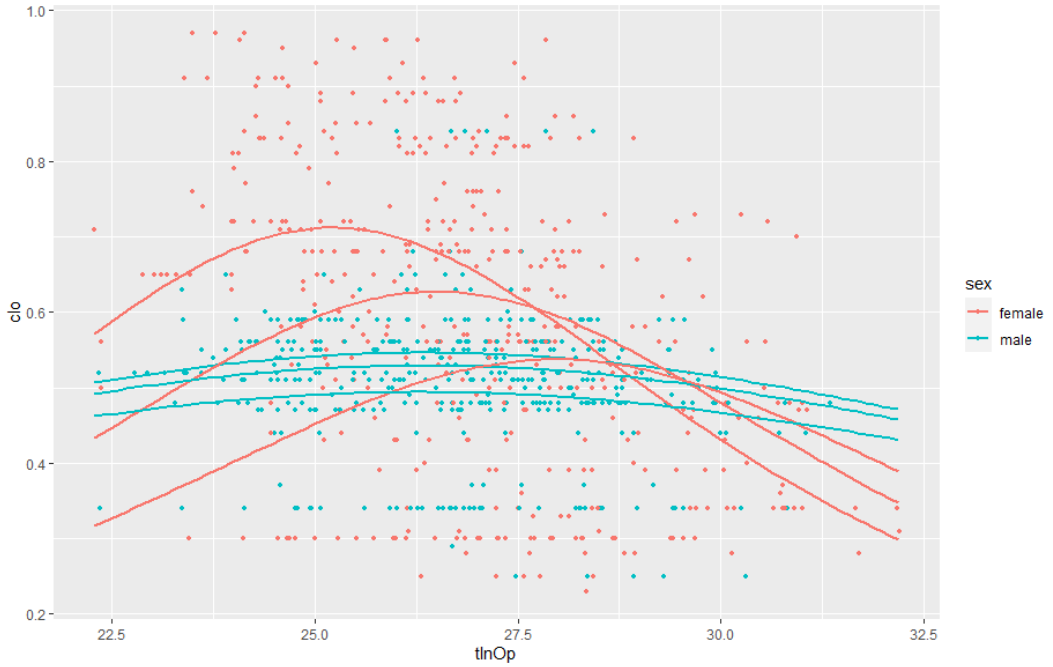

Figure 7: Scatter plot of clo against the indoor operating temperature. The lines represent model predictions for different sex and the 1., 2. and 3. quantiles for tOut.

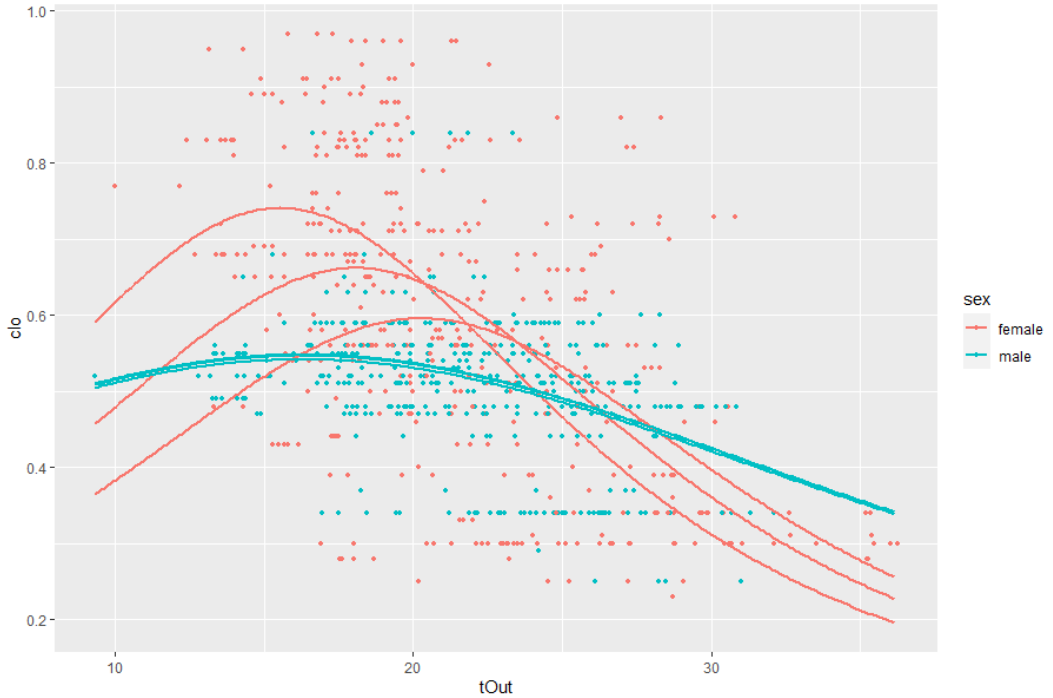

Figure 8: Scatter plot of clo against the outdoor temperature. The lines represent model predictions for different sex and the 1., 2. and 3. quantiles for tInOp.

For the female subjects, we clearly see an interaction between the two temperature variables. For an indoor temperature of tInOp = 25 we see a clear distinction between the emmeans, with a lower outdoor temperature leading to a higher clo. For an indoor temperature of tInOp = 27 we see the same pattern however the low outdoor temperatures of tOut = 18 and tOut = 21 are not distinguishable. For the highest indoor temperatures tInOp = 28, the effect of the outdoor temperature is barely significant, with the confidence interval for the emmeans of tOut = 25 being close to overlapping with the two lower outdoor temperatures.

Overall we conclude that the indoor temperature has almost no effect for the male subjects, with the clo decreasing as outdoor temperatures increase past 17 degrees. For the females, the pattern of decreasing clo as tOut increases is the same, however the indoor temperatures have an effect on when the maximum of clo is reached for different outdoor temperatures. Furthermore increasing the indoors temperature decreases the effect of the outdoor temperature.

2.3 Examining a subject based model

As mentioned, the four models using sex and temperatures for the linear predictor all showed problems, with a high amount of outliers and a pattern in the deviance residuals against the fitted values. As we have also mentioned, a reason for this might be that we have repeated measurements for the same subject. An obvious approach to modelling this, would be to interpret the subjects as randomly drawn from a larger population and model the between subject variation, with a time-dependent correlation within subjects. However, in this section we will treat the subjects as fixed effects and investigate the consequences of changing the sex variable with a specific subject variable for the modelling.

2.3.1 Model description

We choose to restrict our focus to the gamma model with an inverse link function. Thus the new model can be described by the linear predictor

$$\begin{aligned}\eta_i = \beta_0 &+ \beta_1(\text{subjId}_i) \\ &+ \beta_2(\text{subjId}_i)\cdot \text{tOut}_i + \beta_3(\text{subjId}_i)\cdot \text{tOut}_i^2 \\ &+ \beta_4(\text{subjId}_i)\cdot \text{tInOp}_i + \beta_5(\text{subjId}_i)\cdot \text{tInOp}_i^2 \\ &+ \beta_6(\text{subjId}_i)\cdot \text{tOut}_i\cdot \text{tInOp}_i \\ &+ \beta_7(\text{subjId}_i)\cdot \text{tOut}_i\cdot \text{tInOp}_i^2 \\ &+ \beta_8(\text{subjId}_i)\cdot \text{tOut}_i^2\cdot \text{tInOp}_i \\ &+ \beta_9(\text{subjId}_i)\cdot \text{tOut}_i^2\cdot \text{tInOp}_i^2\end{aligned} \tag{3}$$

with $i = 1, 2, \ldots, 803$. Thus we model the linear predictor using second-order polynomials in the two temperature variables (and their interactions), and with different coefficients depending on the subject. The model assumptions are that $Y_i \sim \text{Gamma}\left(\phi, \frac{\mu_i}{\phi}\right)$, where $Y_1, Y_2, \ldots, Y_{803}$ are mutually independent, the shape parameter $\phi > 0$ is a constant precision parameter and

$$\frac{1}{\mu_i} = \eta_i$$

2.3.2 Model diagnostics

Fitting model (3), we perform model diagnostics by plotting the same six plots as mentioned in the previous model diagnostics section. We see that the plot of the standardized deviance residuals against the fitted values, is much better for model (3) than for model the models based on the linear predictor (1), as there is no systematic pattern this time. Looking at the QQ-plot we however see heavier tails than before, implying that the residuals are not necessarily normal distributed. For the plot of sex, we see that there is still a difference in the size of the residual variance for the two sexes, with the variance being smaller for the male subjects. It is however, less distinct than for the models based on (1). The residual plots also shows no systematic dependency on the temperature variables. For the outlier plot, we see that the number of outliers is reduced with only 44 observations having a studentized deviance residuals numerically greater than 2. However, as the model contains 423 parameters we cannot use the previously defined cutoff for investigating influential observations. We note however, that the size of the leverages is more evenly distributed among the observations.

Overall we find that the model better fits the data. However, it is concerning that we have 423 parameters to 803 observations. This large amount of parameters may cause overfitting, which could explain the improved residual plots.

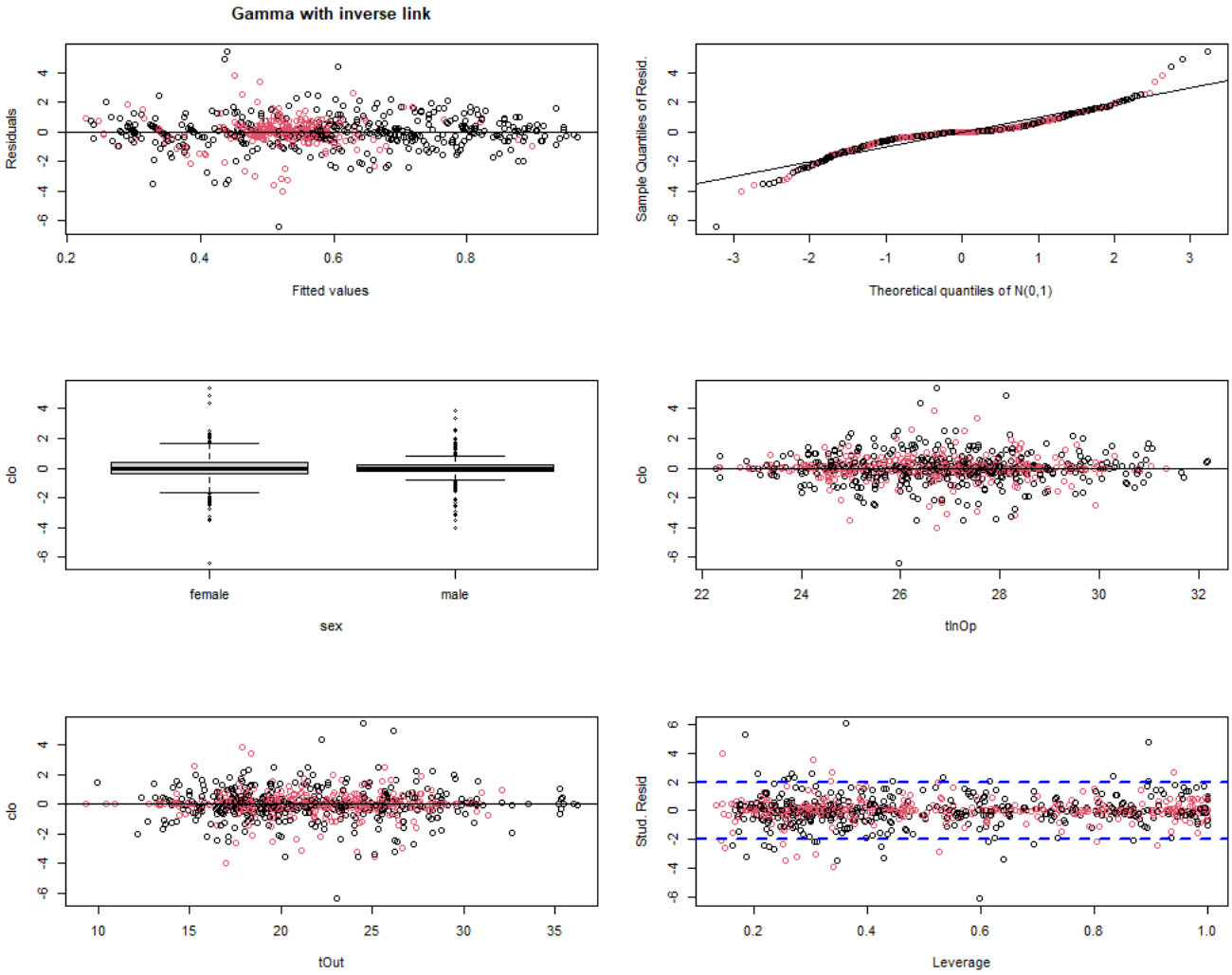

Figure 9: Model diagnostic plots for model (3).

2.3.3 Examining within day autocorrelation

As mentioned, we are dealing with repeated measurements on the subjects. The previous models all assume mutual independence of the observations. Figure 10 looks into the lag 1 autocorrelation of the standardized deviance residuals within each subject and day combination. There are 136 six combinations of subjects and test days. Out of these, 86 combinations had a constant clo value for all observations. This is reflected in figure 10 by the large amount of indexes with an autocorrelation of 1.

The conclusion is that any model assuming independence of observations or error terms is making a wrong assumption. As mentioned a repeated measurements model, using the subjects as random effects and a within subject correlation, would be better suited for analyzing the experiment.

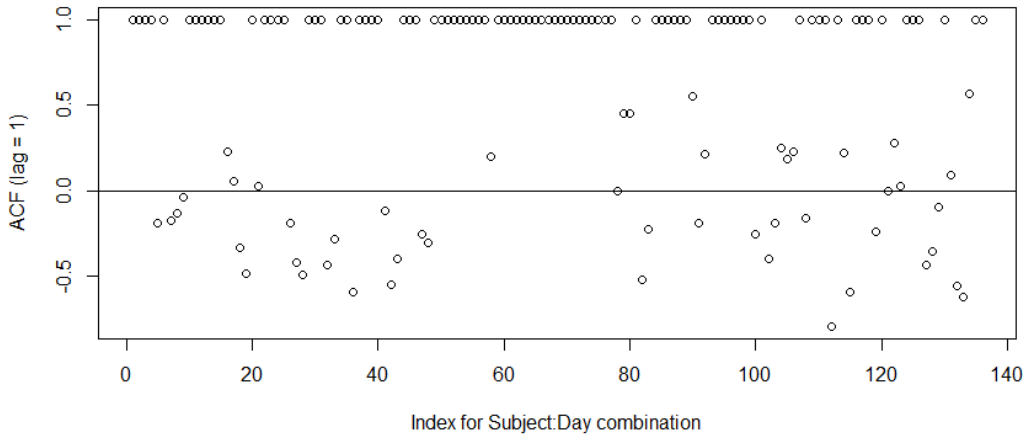

Figure 10: Plot of lag 1 autocorrelations for each combination of subject and day.

2.3.4 Model reduction

Performing a likelihood ratio test of model (3) against a reduced sub-model based on the linear predictor (2) (but with subjId instead of sex), we get a test statistic of 3.1183 (204.08 when scaling with the estimated dispersion parameter) and comparing this to the $\chi^2(141)$-distribution we find a p-value of 0.0004. Thus, using subjId instead of sex, we reject the model reduction. Performing a type II partitioning of the model deviance for further model reduction using the likelihood ratio test, gives us the following ANOVA table

	Df	Deviance	scaled dev.	Pr(>Chi)
Model (3)		5.78		
subjId:poly(tInOp, 2):poly(tOut, 2)	184	10.11	283.51	0.0001

Table 5: ANOVA table for further reductions.

As such, the data does not support further model reduction.

2.3.5 Conclusion

As model (3) contains 423 parameters we choose not to include a table of the parameters for this model. This high amount of parameters is due to the subject interaction with the second order polynomials of the temperature variables. In practice this means that the model fits a second order multivariate polynomial in the two temperature variables for each subject. As such, the parameters will only allow us to infer about the temperatures effect on the clo values for the specific subjects.

As noted, we saw in the model diagnostics that model (3) fits the data better than model (2). This is also reflected in the AIC, with model (3) having an AIC of -1859.6 compared to an AIC of -1014.6 for model (2). An explanation for the improved model fit, is the large auto correlation we noted for each subject and day combination. As most of the variation in the data seems to be explained by the subject and day combination, this variation is better captured in a subject dependent model. However using model (3) we give up on an easier interpretation of the sex and temperature effect on clo.

2.4 Estimating the optimal weight/dispersion parameter for sex

As was noted in previous sections, the dispersion of residuals from the models between men and women differs significantly. We want to take this into account to improve the performance of the model. We choose to consider the model (2) instead of the model (3) since the latter has artificially many parameters by including the subject Id as a fixed effect. Moreover, the optimization procedure for the latter model would be infeasible. We extend the model (2) by defining a weight parameter τ as the ratio of the variances of residuals between females and

men. Equivalently, we may consider the model (2) with the regression on the dispersion parameter. Formally our extended model has the form:

$$Y_i \sim Gamma(\mu_i, \phi_i), \quad \frac{1}{\mu_i} = \eta_i = \boldsymbol{X}_\mu \boldsymbol{\beta}_\mu, \quad \log(\phi_i) = \boldsymbol{X}_\phi \boldsymbol{\beta}_\phi \quad \forall i \in \{1, \ldots, n\} \tag{4}$$

The regression on the dispersion parameter can be rewritten

$$\log(\phi_i) = \boldsymbol{X}_\phi \boldsymbol{\beta}_\phi \iff \phi_i = \exp(\boldsymbol{X}_\phi \boldsymbol{\beta}_\phi) = \exp(\beta_{\phi,0} + \beta_{\phi,1}(\text{sex}_i)) = \exp(\beta_{\phi,0}) \exp(\beta_{\phi,1}(\text{sex}_i)) = \phi\tau \tag{5}$$

where ϕ is the overall dispersion but under our model, it can be interpreted as the dispersion associated with men. To obtain the dispersion for females one should multiply it by τ. We note that the optimization in this section will be based on parameters $(\beta_{\phi,0}, \beta_{\phi,1})$ since they are in the linear domain and do not have any constraints contrary to τ and ϕ which have to be positive numbers. Hence the model is parameterized by $\boldsymbol{\theta} = (\boldsymbol{\beta}_\mu, \beta_{\phi,0}, \beta_{\phi,1})$. Table 6 presents the estimated parameters. Note that the regression parameters associated with μ ($\boldsymbol{\beta}_\mu$) have not changed significantly compared with the model (2). The overall dispersion parameter was estimated to be $\exp(\beta_{\phi,0}) = 0.0303$.

	Estimate	Std. Error	2.5 %	97.5 %
(Intercept)	1.94	0.05	1.84	2.04
sexmale	0.05	0.05	-0.06	0.16
poly(tInOp, 2)1	1.25	0.98	-0.67	3.16
poly(tInOp, 2)2	4.96	1.26	2.49	7.42
poly(tOut, 2)1	10.08	1.14	7.85	12.31
poly(tOut, 2)2	7.13	1.37	4.45	9.81
sexmale:poly(tInOp, 2)1	-0.71	1.15	-2.97	1.54
sexmale:poly(tInOp, 2)2	-3.89	1.44	-6.71	-1.07
sexmale:poly(tOut, 2)1	-5.61	1.32	-8.20	-3.02
sexmale:poly(tOut, 2)2	-4.87	1.54	-7.89	-1.86
poly(tInOp, 1):poly(tOut, 1)	-222.47	49.40	-319.29	-125.65
sexmale:poly(tInOp, 1):poly(tOut, 1)	223.15	52.68	119.91	326.40
$\beta_{\phi,0}$	-3.50	0.07	-3.64	-3.35
$\beta_{\phi,1}$	0.97	0.10	0.78	1.16

Table 6: Model parameters and their uncertainties for model 4.

We are interested in an inference only on the $\beta_{\phi,1}$ parameter and hence we consider the profile likelihood of it and treat the rest of the parameters as nuisance parameters. The profile likelihood is defined as:

$$L_P(\beta_{\phi,1}; \boldsymbol{y}) = \sup_{\boldsymbol{\beta}_\mu, \beta_{\phi,0}} L((\beta_{\phi,1}, \boldsymbol{\beta}_\mu, \beta_{\phi,0}); \boldsymbol{y}) \tag{6}$$

where the maximization is performed at a fixed value of $\beta_{\phi,1}$ and the vector $\boldsymbol{y}$ is the vector of the dependent variable clo. Therefore we solve the following optimization problem to estimate $\beta_{\phi,1}$:

$$\hat{\beta}_{\phi,1} = \sup_{\beta_{\phi,1}} L_P(\beta_{\phi,1}; \boldsymbol{y}) \tag{7}$$

The uncertainty associated with the estimation is captured by the profile likelihood function. We consider two types of confidence intervals: Wald's and likelihood-based. The following table presents the results

	value	Wald Lower CI	Wald Upper CI	Profile Lower CI	Profile Upper CI
$\hat{\beta}_{\phi,1}$	0.97	0.78	1.16	0.78	1.16
$\tau = \exp(\hat{\beta}_{\phi,1})$	2.64	2.17	3.2	2.18	3.19

Table 7: Estimate of the precision parameter with uncertainties.

Figure 11 presents the profile likelihood and the quadratic approximation at the optimum. We note that the profile likelihood is symmetric and the quadratic approximation is almost exact. It is a result of the fact that the maximum likelihood estimator is asymptotically normally distributed (Theorem 2.4) under mild assumptions. In our model, there are more than 800 observations so the convergence in distribution granted by the Central Limit Theorem takes place. Moreover, it is known that for the normal distribution, the quadratic approximation is exact. We

further note that another result of the above-mentioned theorem is that the Wald and Likelihood-based confidence intervals are almost exact.

We are interested in testing whether the precision parameter τ is different than 1, investigating whether the dispersion parameter is different between the two groups. Formally, we test a null hypothesis $H_0 : \tau = 1$ against the alternative $H_1 : \tau \neq 1$. We use the likelihood ratio test which is invariant to the parameter transformations and hence can rewrite the hypothesis in the following way

$$H_0 : \tau = 1 \iff exp(\beta_{\phi,1}) = 1 \iff \beta_{\phi,1} = 0; \quad H_1 : \tau \neq 1 \iff exp(\beta_{\phi,1}) \neq 1 \iff \beta_{\phi,1} \neq 0 \tag{8}$$

The test is equivalent to checking whether the parameter from the null hypothesis ($\beta_{\phi,1} = 0$) belongs to the profile likelihood confidence interval. We see that it is indeed the case, and hence we reject the null hypothesis that $\tau = 1$. Therefore, there is evidence to conclude that the dispersion in the two groups is different and should be modelled.

We note, that the maximum likelihood estimator is known to be biased for the dispersion parameters and one might consider using the restricted maximum likelihood to account for that.

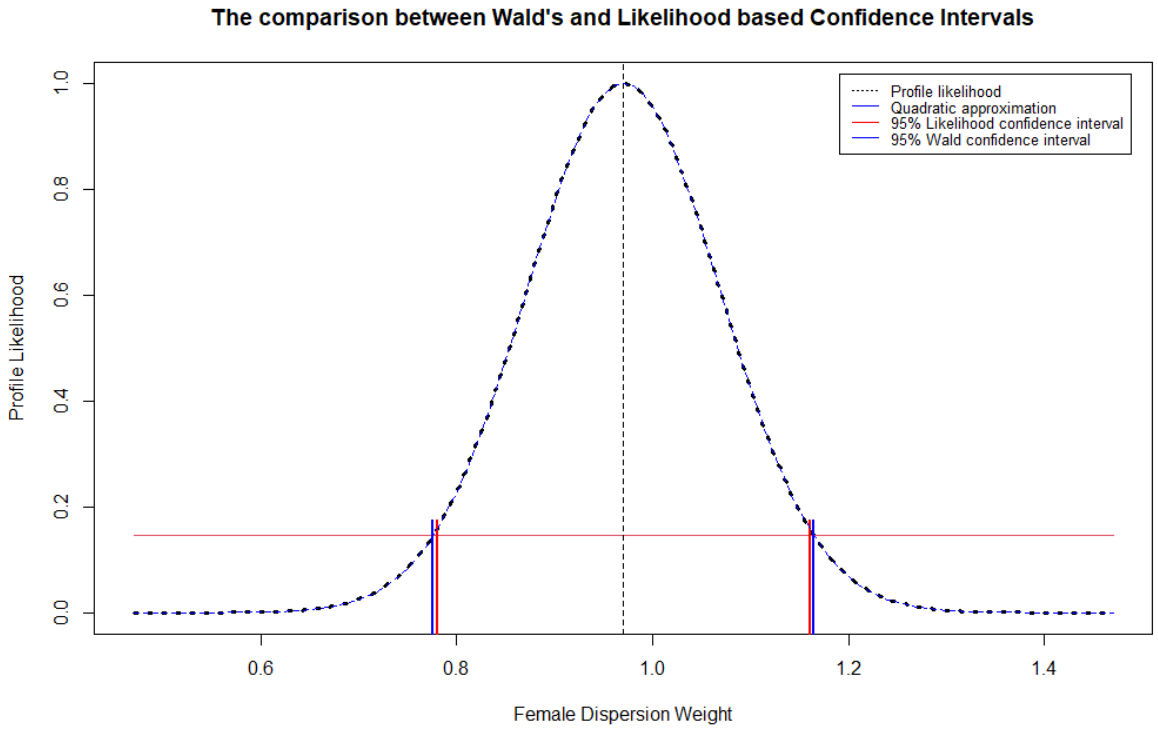

Figure 11: Plot of the profile likelihood for the parameter weight $\beta_{\phi,1}$. The vertical line indicates the value that maximizes the likelihood. Blue rugs correspond to Wald's confidence interval, whereas red rugs correspond to the Likelihood-based confidence intervals.

2.5 Conclusions

We investigated different models within the general linear model and generalized linear models family for the variation in the clothing insulation level. Guided by the information criteria AIC and model diagnostic plots we found that the Gamma regression model with a canonical inverse link performed the best, with the variables sex, tOut, tInOp and their interactions as independent variables. We found that the variation in residuals is significantly greater for female than men. We found a reduced model (2). Based on this model we concluded, that the temperatures effect is dependent on the sex. Furthermore increasing the indoors temperature decreases the effect of the outdoor temperature. We found a model with the subject Id included as the fixed effect instead of the sex to have a lower AIC. However, using subject ID as a fixed effect results in a model only relevant for the particular subjects in the experiment, and does not necessarily generalize to a larger population. We found a significant within-day autocorrelation explained by the repeated measurements set up of the study. We estimated the model (ref) with different dispersion parameter for men and women and based on a statistical test we concluded

that the two dispersion parameters should be included in the model. We concluded that the assumptions of a generalized linear models were violated by the autocorrelation. We proposed to investigate the subject ID as a random effect and incorporate the within-day autocorrelation in the variance-covariance structure of the mixed effect model.

3 Part B: Ear infection in swimmers

The purpose of the following section is to establish a model for the variation in the number of ear infections in swimmers. It is further of interest to investigate whether location, age, whether the swimmer is a frequent or occasional swimmer or interactions between these have any effects on the number of ear infections.

3.1 Description of the experiment and an initial look at the data

The dataset comes from an observational study, where for each group comprised of a different number of people the overall number of ear infections was reported by the individuals in the year 1990. There are 4 categorical variables that differentiate the groups:

1. swimmer - Indicates if the swimmer is a frequent or an occasional ocean swimmer (2 levels)
2. location - Indicates the usually chosen swimming location: beach or non-beach (2 levels)
3. age - The age of the swimmer: 15-19, 20-24, 25-29 (3 levels)
4. sex - The gender of the swimmer male or female (2 levels)

Therefore the number of groups is 24 and this is the size of the dataset. We note that there are no continuous independent variables. It was verified that the dataset is balanced. Since the number of people in each group is different is it sensible to consider the rate defined as the number of infections in a group divided by the number of people in the group. Figure 12 presents the dependent variable rate as a function of independent variables.

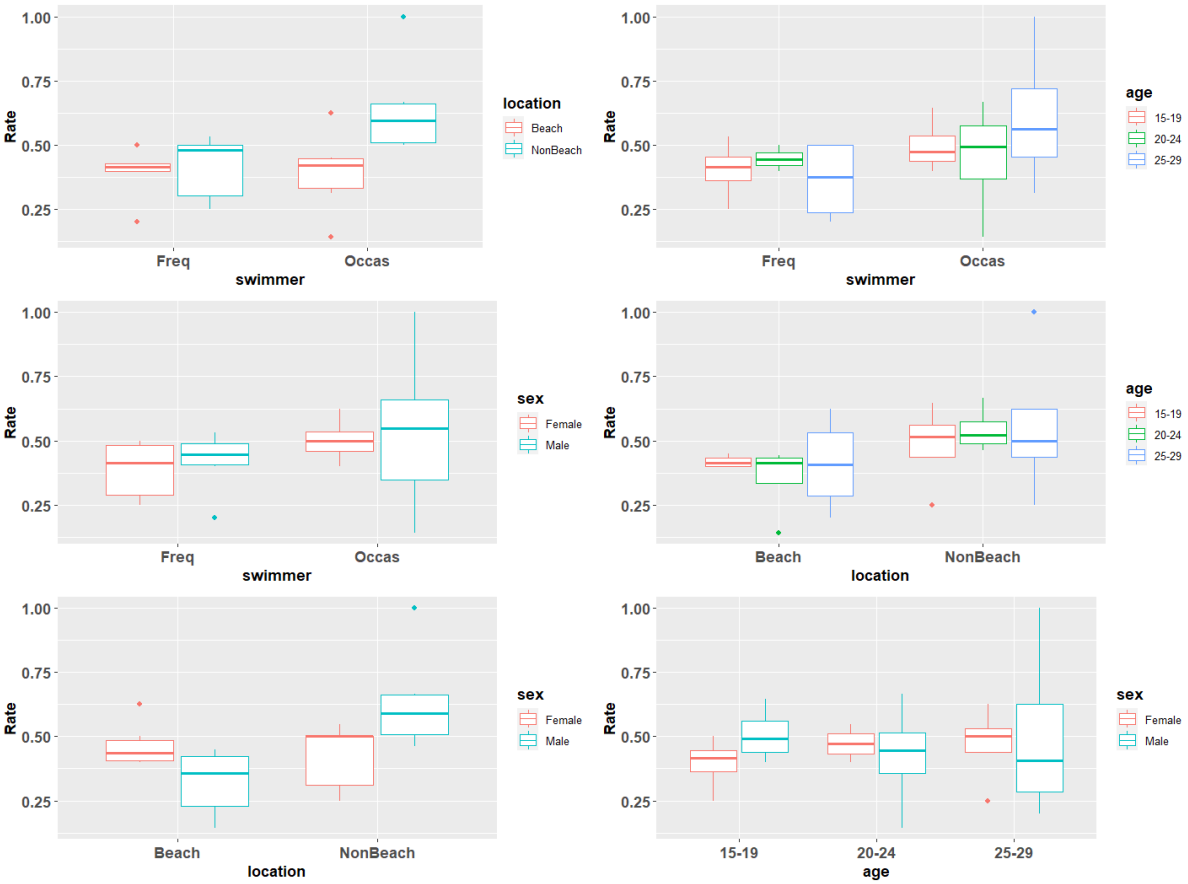

Figure 12: Plots of rate against the different independent variables.

We see that there is some variation in the rate based on different independent variables. However, variables sex, age and swimmer, seem to contribute little to the variation of rate, but based on the plots there might be an interaction between location and other variables. Based on the left top plot variable location seems to influence the variation in rate the most. It is further supported by the right middle plot where we see that the overall mean between all ages is clearly different based on the location.

3.2 Considerations on the general linear model

A general linear model assumes that the variation in the dependent variable can be well-modelled by the normal distribution, namely

$$Y_i \sim N(X_i^T \beta, \sigma^2), \quad \forall i \in \{1, \dots, n\}. \tag{9}$$

This distribution assumption would be inappropriate to model the rate of ear infections since the dependent variable rate takes positive values but the support of the normal distribution is the entire real line. Hence, inevitably, some probability mass would be placed on the dependent variable being negative which is physically impossible. Therefore it is more appropriate to search between models that do not violate the physical constraints.

3.3 Considerations on the offset

The data of ear infections are count data and it is natural to consider a Poisson model for it. We write the following model with the canonical link function:

$$\frac{Y_i}{n_i} \sim Poiss\left(\mu_i, n_i\right) / n_i, \quad \log\left(\frac{\mu_i}{n_i}\right) = \eta_i = \boldsymbol{X}\boldsymbol{\beta} \iff \log(\mu_i) = \boldsymbol{X}\boldsymbol{\beta} + \log(n_i) \tag{10}$$

we see that different group sizes can be accounted for as the known value - the variable with the coefficients 0 which depends on the experimental conditions. This offset is needed (either expressed as above or included in the calculation of the rate) to account for different group sizes. Hence in the following sections, we will consider a Poisson model for the number of ear infections with the offset indicating the group size.

3.4 Model descriptions

We fit the full (saturated) model to the data, which has as many parameters as observations. The formula is the following:

$$\eta_i = \beta(\text{swimmer}_i, \text{sex}_i, \text{location}_i, \text{age}_i) \tag{11}$$

The Poisson distribution is a member of the exponential dispersion family with the dispersion parameter $\phi = 1$. Since we did not estimate the dispersion parameter we can perform the test for model sufficiency - the goodness of fit test for our initial full model (Remark 4.22). We test the null hypothesis that the model is a sufficient fit to the data against the alternative hypothesis that states otherwise. The test is a special case of the Likelihood Ratio Test which compares the calculated test statistics with the quantiles of the chi-squared distribution. There are as many parameters as observations and the full model fits the data perfectly. Hence, the residual deviance of the model is 0 and the goodness of fit test fails to reject the null hypothesis. Therefore we conclude that the initial model has a sufficient fit to the data, which is a trivial statement since we fit a full model.

3.5 Model reduction

We perform the successive Likelihood Ratio Tests (Theorem 4.3) on the saturated model, which are a type II partitioning of the model deviance for further model reduction. The final ANOVA table is presented below, where no further reductions can be performed:

	Df	Deviance	LRT statistic	Pr(>Chi)
Model after reductions		9.629		
location	1	14.357	4.729	0.03

Table 8: ANOVA table for further reductions

The reduced model has the following form:

$$Y_i \sim Poiss(\mu_i), \quad \log(\mu_i) = \eta_i + \log(n_i),$$
$$\eta_i = \beta_0 + \beta_1(\text{location}_i) = \tilde{\beta}_0(\text{location beach}_i) + \tilde{\beta}_1(\text{location nonbeach}_i)$$

where we choose to parametrize the model by $(\tilde{\beta}_0, \tilde{\beta}_1)$ which will ease the interpretation. The estimated coeficients along with the measuere of uncertinty are presented below:

We note that both coefficients are significant since the value 0 does not belong to the confidence intervals. Moreover, the profile likelihood confidence intervals are slightly wider.

	Estimate	Std. Error	Wald 2.5 %	Wald 97.5 %	Likelihood 2.5 %	Likelihood 97.5 %
location beach	-0.947	0.133	-1.207	-0.688	-1.219	-0.699
location nonbeach	-0.572	0.113	-0.793	-0.352	-0.801	-0.360

Table 9: Model parameters for the final model.

3.6 Model diagnostic

We perform the model diagnostics for the reduced model. We look at the following 4 plots to check the fit:

1. Standardized Pearson residuals against the predicted values $\hat{\mu}_i$.
2. QQ-plot of the standardized Pearson residuals against the theoretical $\mathcal{N}(0,1)$-quantiles.
3. Square root of the standardized Pearson residuals against predicted values to check for homoscedasticity of the variance of residuals.
4. Standardized Pearson residuals against the leverage to detect influential observations.

Figure 13 presents the results. Based on the QQ-plot We see that the Pearson residuals are normally distributed, potentially with heavy tails. The heavy tails may be due to the few number of observations in the experiment, as the Pearson residuals are only asymptotically normally distributed. The variance of the residuals appears to be constant which satisfies model assumptions. The residuals vs leverage plot doesn't indicate any influential observations. The p-value associated with the goodness of fit test for this model was calculated to be 0.989. Hence, the test fails to reject the null hypothesis, indicating that the final model has a sufficient fit to the data.

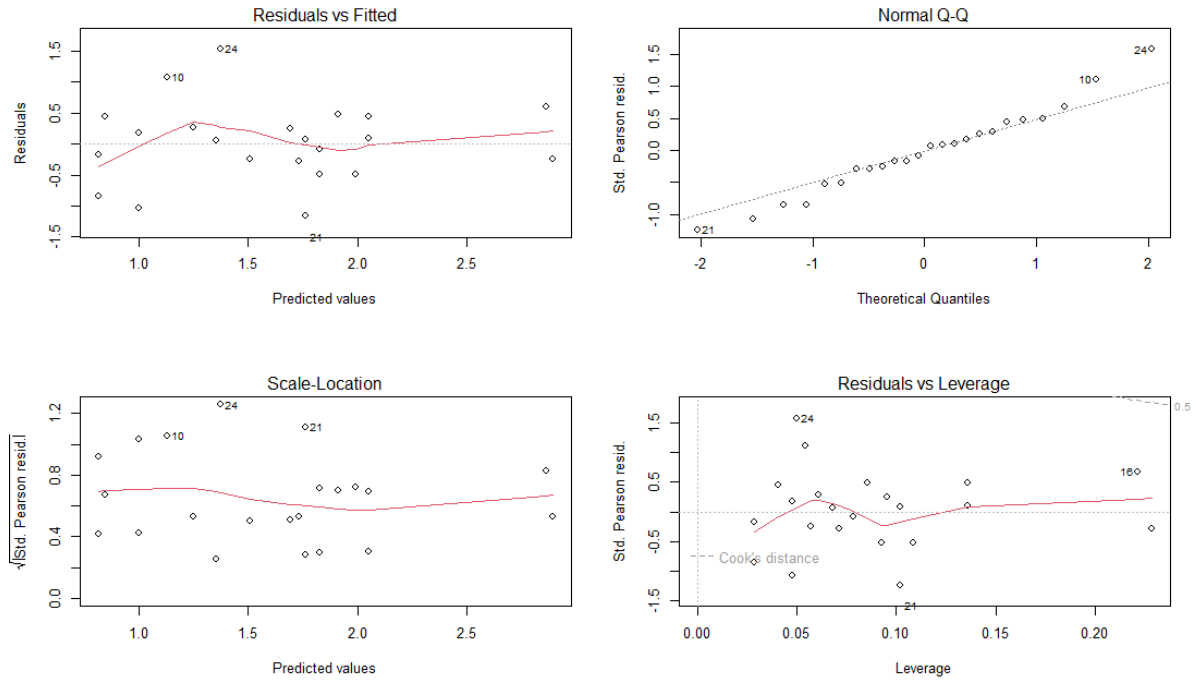

Figure 13: Plots of the number of ear infections against the different dependent variables.

Figure 14 presents the number of ear infections against different groups with fitted values and observations. We note that 6 out of 24 observations do not belong to the confidence intervals which is concerning. More investigation should be conducted to understand this.

Figure 15 presents the deviance residuals against different independent variables. For the variable swimmer, we see a difference in the distribution for the two levels. Also, for the the variable age, we see that the variation in the deviance residuals is larger for the 25-29 age group. However, these differences are to be expected with the low number of observations.

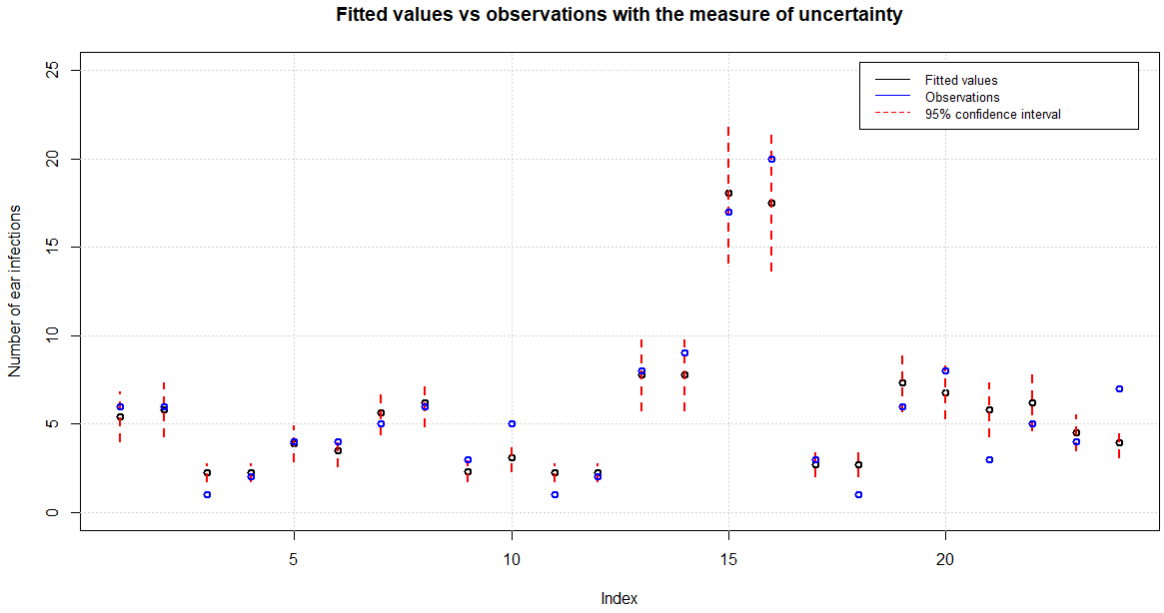

Figure 14: Plots of the number of ear infections against different groups with fitted values and observations

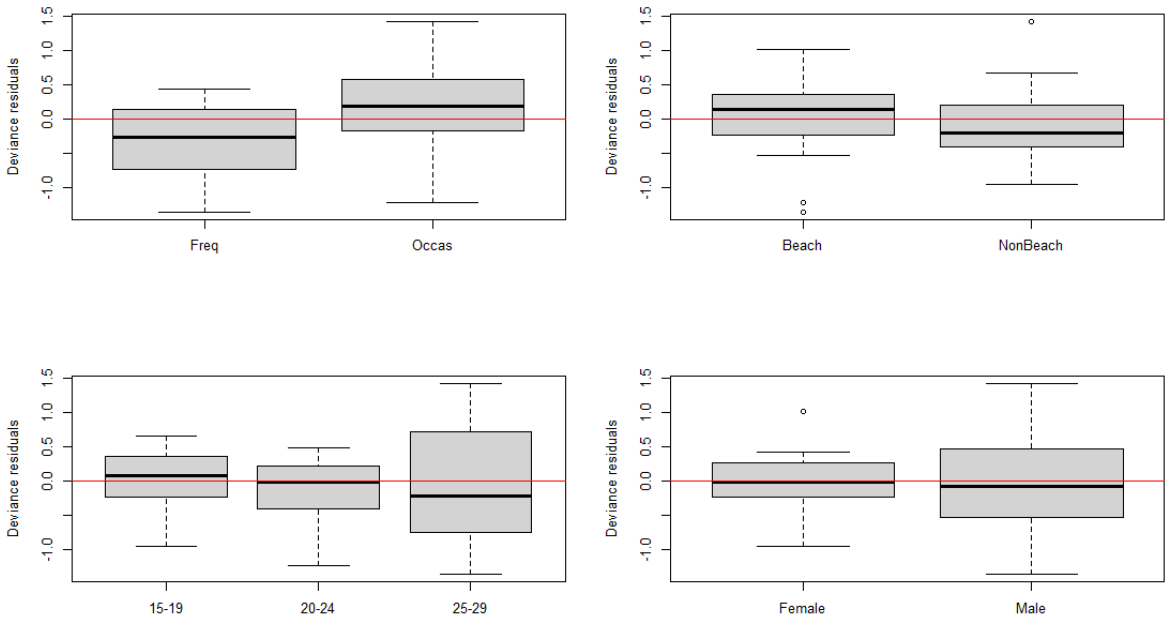

Figure 15: Plots of the deviance residuals against different dependent variables.

3.7 Model interpretation

We write the final model in the following form

$$Y_i \sim Poiss(\mu_i), \quad \mu_i = \exp(\tilde{\beta}_0(\text{location beach}_i)) \exp(\tilde{\beta}_1(\text{location nonbeach}_i))n_i \tag{12}$$

We conclude that only the location variable has an effect on the number of ear infections since it's the only variable present in the final model. Given the number of people in a group n_i the expected number of ear infections is the

following

$$E[Y_i] = \mu_i = \begin{cases} \exp(-0.947) n_i = 0.388 \cdot n_i, & \text{for swimmer in beachs} \\ \exp(-0.572) n_i = 0.564 \cdot n_i, & \text{for swimmer in non beachs} \end{cases} \tag{13}$$

Equivalently, for a given group with size n_i the expected number of ear infections in non-beaches will be $\frac{0.564}{0.388} = 1.455$ times greater than the expected number of ear infections for swimmers in beaches.

4 References

1. Conradsen, K., Christensen, A.M., Nielsen, A.A., Ersbøll, B.K. (2019, v.0.94). *Multivariate Statistics: For the Technical Sciences*. DTU Compute Lyngby.

2. Madsen H., Thyregod P. (2011). *Introduction to General and Generalized Linear Models*. CRC Press.

5 Appendix

```
# Preamble ####################################################################

### File Description ##########################################################
##
##   Soren Skjernaa - s223316
##   Tymotuesz Barcinski - s221937
##   10/04-2023
##
##   Advanced Dataanalysis and Statistical Modelling
##   Assignment 2
##
##   Note: The data file for the analysis is assumed to be found at the relative
##         paths "data/clothing.csv" and "data/earinfect.txt".
##
##
###############################################################################

### Clean up ##################################################################
rm(list = ls())
if(!is.null(dev.list())) dev.off()

### Libraries #################################################################

# Plotting
library(ggplot2)      # Nice plots
library(patchwork)    # Layout of plots
library(gridExtra)

# Handling data frames
library(tidyr)        # Reshape data frames
library(dplyr)        # Rename column in data frames

# Statistics
library(nlme)
library(betareg)
library(Gammareg)

# Post-hoc analysis
library(emmeans)
library(multcomp)

# Model diagnostics
library(boot)
library(MASS)
library(MESS)
library(nortest)
library(influence.ME)

# Other
library(xtable)       # Latex tables
library(numDeriv)     # Finding Hessians


# - ###########################################################################
# Part A: Clothing level ######################################################


### Load the data and get initial overview ####################################

# Load data
df_data <- read.csv("data/clothing.csv", header = TRUE)
str(df_data)
df_data$subjDay <- 2^df_data$subjId * 3^df_data$day
df_data <- mutate_at(df_data, c("sex", "subjId", "day", "subjDay"), as.factor)
summary(df_data)
attach(df_data)

# Tabulate data
```

```
table(sex)
table(day, sex)

# Summary statistics of data
tapply(clo, sex, mean)
tapply(clo, sex, sd)

# Check the average standard deviation for each combination subjId and Day
summary(clo)                        # 0.23 - 0.97
sd(clo)                             #      0.16008
mean(tapply(clo, subjDay, sd))      # Only 0.02635
mean(tapply(tInOp, subjDay, sd))
mean(tapply(tOut, subjDay, sd))

### Plot of the data ###########################################################

# Boxplot of subjId
p1 <- ggplot(df_data, aes(x=subjId, y=clo, col=sex)) +
  geom_boxplot() +
  labs() +
  theme(plot.title = element_text(hjust=0.5, size=14, face="bold"),
        axis.title = element_text(size=12, face="bold"),
        legend.title = element_text(size=12, face="bold",),
        axis.text.x=element_blank())

# Histograms of clothing insulation level
p2 <- ggplot(df_data, aes(x = clo, fill = sex)) +
  geom_histogram(alpha=0.3, position="identity", bins=20) +
  labs() +
  theme(plot.title = element_text(hjust=0.5, size=14, face="bold"),
        axis.text = element_text(size=12),
        axis.title = element_text(size=12, face="bold"),
        legend.title = element_text(size=12, face="bold"),
        legend.position = "none",
        legend.text = element_text(size=11))


# Plot of clothing insulation levels against temperatures
# df_data = df_data[df_data$subjId %in% c(11,17,35,49),]
p3 <- ggplot(df_data, aes(x=tOut, y=clo, col=sex)) +
  geom_point(size = 1) +
  labs(subtitle = "Outdoor temperature") +
  theme(plot.title = element_text(hjust=0.5, size=14, face="bold"),
        axis.title = element_text(size=12, face="bold"),
        legend.title = element_text(size=12, face="bold",),
        legend.position = "none",
        axis.text = element_text(size=12))

p4 <- ggplot(df_data, aes(x=tInOp, y=clo, col=sex)) +
  geom_point(size = 1) +
  labs(subtitle = "Indoor operating temperature") +
  theme(plot.title = element_text(hjust=0.5, size=14, face="bold"),
        axis.title = element_text(size=12, face="bold"),
        legend.title = element_text(size=12, face="bold",),
        legend.position = "none",
        axis.text = element_text(size=12))

p1 + p2 + p3 + p4 + plot_layout(guides = "collect", ncol = 2) +
  plot_annotation(title = "",
                  theme = theme(plot.title = element_text(size = 16,
                                                          face = "bold",
                                                          hjust = 0.5)))



### Sex models #################################################################


#### Fitting the four models ###################################################

# Gaussian without transformation of clo
lmId <- lm(clo ~ sex * poly(tInOp, 2) * poly(tOut, 2))
summary(lmId)
#boxcox(lmId)


# Gaussian with sqrt transformation
lmSqrt <- lm(sqrt(clo) ~ sex * poly(tInOp, 2) * poly(tOut, 2))
summary(lmSqrt)


# Gamma with inverse link
glmGammaInv <- glm(clo ~ sex * poly(tInOp, 2) * poly(tOut, 2),
```

```
                    family = Gamma(link = "inverse"))
summary(glmGammaInv)


# Gamma with logarithmic link
glmGammaLog <- glm(clo ~ sex * poly(tInOp, 2) * poly(tOut, 2),
                    family = Gamma(link = "log"))
summary(glmGammaLog)



#### Model diagnostics ##########################################################

# Gaussian without transformation of clo
par(mfrow = c(3,2))
stdDevResid <- rstandard(lmId)
studDevResid <- studres(lmId)
infl <- influence.measures(lmId)$infmat
leverage <- infl[, ncol(infl)]
cutoff <- 2 * length(coef(lmId)) / length(studDevResid)

p1 <- plot(fitted(lmId), stdDevResid, col = sex,
     main = "Gaussian without transformation of clo",
     xlab = "Fitted values", ylab = "Residuals")
p1 <- p1 + abline(0,0)

qqnorm(stdDevResid, col = sex,
       main = "", xlab = "Theoretical quantiles of N(0,1)",
       ylab = "Sample Quantiles of Resid.")
abline(0,1)

plot(sex, stdDevResid, xlab = "sex", ylab = "clo")

plot(tInOp, stdDevResid, col = sex, xlab = "tInOp", ylab = "clo")
abline(0,0)

plot(tOut, stdDevResid, col = sex, xlab = "tOut", ylab = "clo")
abline(0,0)

plot(leverage, studDevResid, col = sex,
     xlab = "Leverage", ylab = "Stud. Resid")
abline(h = -2, col = "blue", lwd=2, lty=2)
abline(h =  2, col = "blue", lwd=2, lty=2)
abline(v = cutoff, col = "blue", lwd=2, lty=2)
sum(studDevResid > 2 | studDevResid < -2)


# Gaussian with sqrt transformation
par(mfrow = c(3,2))
stdDevResid <- rstandard(lmSqrt)
studDevResid <- studres(lmSqrt)
infl <- influence.measures(lmSqrt)$infmat
leverage <- infl[, ncol(infl)]
cutoff <- 2 * length(coef(lmSqrt)) / length(studDevResid)

p1 <- plot(fitted(lmSqrt), stdDevResid, col = sex,
           main = "Gaussian with transformation sqrt(clo)",
           xlab = "Fitted values", ylab = "Residuals")
p1 <- p1 + abline(0,0)

qqnorm(stdDevResid, col = sex,
       main = "", xlab = "Theoretical quantiles of N(0,1)",
       ylab = "Sample Quantiles of Resid.")
abline(0,1)

plot(sex, stdDevResid, xlab = "sex", ylab = "clo")

plot(tInOp, stdDevResid, col = sex, xlab = "tInOp", ylab = "clo")
abline(0,0)

plot(tOut, stdDevResid, col = sex, xlab = "tOut", ylab = "clo")
abline(0,0)

plot(leverage, studDevResid, col = sex,
     xlab = "Leverage", ylab = "Stud. Resid")
abline(h = -2, col = "blue", lwd=2, lty=2)
abline(h =  2, col = "blue", lwd=2, lty=2)
abline(v = cutoff, col = "blue", lwd=2, lty=2)
sum(studDevResid > 2 | studDevResid < -2)



# Gamma with inverse link
par(mfrow = c(3,2))
```

```
stdDevResid <- rstandard(glmGammaInv, type = "deviance")
studDevResid <- rstudent(glmGammaInv, type = "deviance")
infl <- influence.measures(glmGammaInv)$infmat
leverage <- infl[, ncol(infl)]
cutoff <- 2 * length(coef(glmGammaInv)) / length(studDevResid)

p1 <- plot(fitted(glmGammaInv), stdDevResid, col = sex,
           main = "Gamma with inverse link",
           xlab = "Fitted values", ylab = "Residuals")
p1 <- p1 + abline(0,0)

qqnorm(stdDevResid, col = sex,
       main = "", xlab = "Theoretical quantiles of N(0,1)",
       ylab = "Sample Quantiles of Resid.")
abline(0,1)

plot(sex, stdDevResid, xlab = "sex", ylab = "clo")

plot(tInOp, stdDevResid, col = sex, xlab = "tInOp", ylab = "clo")
abline(0,0)

plot(tOut, stdDevResid, col = sex, xlab = "tOut", ylab = "clo")
abline(0,0)

plot(leverage, studDevResid, col = sex,
     xlab = "Leverage", ylab = "Stud. Resid")
abline(h = -2, col = "blue", lwd=2, lty=2)
abline(h =  2, col = "blue", lwd=2, lty=2)
abline(v = cutoff, col = "blue", lwd=2, lty=2)
sum(studDevResid > 2 | studDevResid < -2)


# Gamma with logarithmic link
par(mfrow = c(3,2))
stdDevResid <- rstandard(glmGammaLog, type = "deviance")
studDevResid <- rstudent(glmGammaLog, type = "deviance")
infl <- influence.measures(glmGammaLog)$infmat
leverage <- infl[, ncol(infl)]
cutoff <- 2 * length(coef(glmGammaLog)) / length(studDevResid)

p1 <- plot(fitted(glmGammaLog), stdDevResid, col = sex,
           main = "Gamma with logarithmic link",
           xlab = "Fitted values", ylab = "Residuals")
p1 <- p1 + abline(0,0)

qqnorm(stdDevResid, col = sex,
       main = "", xlab = "Theoretical quantiles of N(0,1)",
       ylab = "Sample Quantiles of Resid.")
abline(0,1)

plot(sex, stdDevResid, xlab = "sex", ylab = "clo")

plot(tInOp, stdDevResid, col = sex, xlab = "tInOp", ylab = "clo")
abline(0,0)

plot(tOut, stdDevResid, col = sex, xlab = "tOut", ylab = "clo")
abline(0,0)

plot(leverage, studDevResid, col = sex,
     xlab = "Leverage", ylab = "Stud. Resid")
abline(h = -2, col = "blue", lwd=2, lty=2)
abline(h =  2, col = "blue", lwd=2, lty=2)
abline(v = cutoff, col = "blue", lwd=2, lty=2)
sum(studDevResid > 2 | studDevResid < -2)


#### Model reduction ##########################################################

# Comparison of the models
AIC(lmId)
AIC(lmSqrt)
AIC(glmGammaInv)
AIC(glmGammaLog)

# Adjust AIC for square root transformation
sqrt_derivative <- function(y_input){
  return(1/(2*sqrt(y_input)))
}
to_subtract <- sum(log(sqrt_derivative(clo)))
AIC(lmSqrt) - 2*to_subtract


# Model reduction
summary(glmGammaInv)
```

```
glmGammaInv1 <- glmGammaInv
glmGammaInv2 <- glm(clo ~ sex + poly(tInOp, 2) + poly(tOut, 2) +
                          sex:poly(tInOp, 2) + sex:poly(tOut, 2) +
                          poly(tInOp, 1):poly(tOut, 1) +
                          sex:poly(tInOp, 1):poly(tOut, 1),
                    family = Gamma(link = "inverse"))

anova(glmGammaInv2, glmGammaInv1, test = "LRT")
drop1(glmGammaInv2, test = "LRT")
xtable(drop1(glmGammaInv2, test = "LRT")[, c(1,2,3,4)])



#### Post-Hoc Analysis ##########################################################

# Model parameters
parameters <- cbind(summary(glmGammaInv2)$coefficients[, 1:2],
                    confint(glmGammaInv2))
xtable(round(parameters, 2))
summary(glmGammaInv2)$dispersion

# Quantiles of temperatures
tOutLevels <- as.numeric(summary(tOut)[c(2, 3, 5)])
tInOpLevels <- as.numeric(summary(tInOp)[c(2, 3, 5)])

# Model contrasts
observedEmmeans <- emmeans(glmGammaInv2, "tOut", by = c("sex", "tInOp"),
                           at = list(tOut = round(tOutLevels,0),
                                     tInOp = round(tInOpLevels,0)),
                           type = "response")
plot(observedEmmeans)


# Graphical presentation - tInOp
tInOpSeq <- seq(min(tInOp), max(tInOp), 0.1)

df_tInOp <-
  data.frame(sex = c(rep("female", length(tInOpSeq) * length(tOutLevels)),
                     rep("male", length(tInOpSeq) * length(tOutLevels))),
             tInOp = rep(tInOpSeq, 2 * length(tOutLevels)),
             tOut = rep(c(rep(tOutLevels[1], length(tInOpSeq)),
                          rep(tOutLevels[2], length(tInOpSeq)),
                          rep(tOutLevels[3], length(tInOpSeq))), 2))
df_tInOp <- cbind(df_tInOp,
                  "clo" = predict(glmGammaInv2, df_tInOp, type = "response"),
                  "group" = 2^(df_tInOp$tOut) *
                    3^as.numeric(as.factor(df_tInOp$sex)))

ggplot(df_data, aes(x=tInOp, y=clo, col=sex)) +
  geom_point(size = 1) +
  geom_line(data = df_tInOp, aes(x=tInOp, y=clo, group=group), linewidth = 1)
labs(subtitle = "Outdoor temperature") +
  theme(plot.title = element_text(hjust=0.5, size=14, face="bold"),
        axis.title = element_text(size=12, face="bold"),
        legend.title = element_text(size=12, face="bold",),
        axis.text = element_text(size=12, face="bold"))


# Graphical presentation - tOut
tOutSeq <- seq(min(tOut), max(tOut), 0.1)

df_tOut <-
  data.frame(sex = c(rep("female", length(tOutSeq) * length(tInOpLevels)),
                     rep("male", length(tOutSeq) * length(tInOpLevels))),
             tOut = rep(tOutSeq, 2 * length(tInOpLevels)),
             tInOp = rep(c(rep(tInOpLevels[1], length(tOutSeq)),
                          rep(tInOpLevels[2], length(tOutSeq)),
                          rep(tInOpLevels[3], length(tOutSeq))), 2))
df_tOut <- cbind(df_tOut,
                 "clo" = predict(glmGammaInv2, df_tOut, type = "response"),
                 "group" = 2^(df_tOut$tInOp) *
                   3^as.numeric(as.factor(df_tOut$sex)))

ggplot(df_data, aes(x=tOut, y=clo, col=sex)) +
  geom_point(size = 1) +
  geom_line(data = df_tOut, aes(x=tOut, y=clo, group = group), linewidth = 1)
  labs(subtitle = "Outdoor temperature") +
  theme(plot.title = element_text(hjust=0.5, size=14, face="bold"),
        axis.title = element_text(size=12, face="bold"),
        legend.title = element_text(size=12, face="bold",),
        axis.text = element_text(size=12, face="bold"))
```

```
### Subject model ##################################################################

#### Fit model #####################################################################

glmGammaSubj <- glm(clo ~ subjId * poly(tInOp, 2) * poly(tOut, 2),
                    family = Gamma(link = "inverse"))
summary(glmGammaSubj)


#### Model Diagnostics #############################################################

par(mfrow = c(3,2))
stdDevResid <- rstandard(glmGammaSubj, type = "deviance")
studDevResid <- rstudent(glmGammaSubj, type = "deviance")
infl <- influence.measures(glmGammaSubj)$infmat
leverage <- infl[, ncol(infl)]
cutoff <- 2 * length(coef(glmGammaSubj)) / length(studDevResid)

p1 <- plot(fitted(glmGammaSubj), stdDevResid, col = sex,
           main = "Gamma with inverse link",
           xlab = "Fitted values", ylab = "Residuals")
p1 <- p1 + abline(0,0)

qqnorm(stdDevResid, col = sex,
       main = "", xlab = "Theoretical quantiles of N(0,1)",
       ylab = "Sample Quantiles of Resid.")
abline(0,1)

plot(sex, stdDevResid, xlab = "sex", ylab = "clo")

plot(tInOp, stdDevResid, col = sex, xlab = "tInOp", ylab = "clo")
abline(0,0)

plot(tOut, stdDevResid, col = sex, xlab = "tOut", ylab = "clo")
abline(0,0)

plot(leverage, studDevResid, col = sex,
     xlab = "Leverage", ylab = "Stud. Resid")
abline(h = -2, col = "blue", lwd=2, lty=2)
abline(h =  2, col = "blue", lwd=2, lty=2)
abline(v = cutoff, col = "blue", lwd=2, lty=2)
sum(studDevResid > 2 | studDevResid < -2)
nrow(summary(glmGammaSubj)$coefficients)


#### Model reduction ###############################################################

# Model reduction
summary(glmGammaSubj)

glmGammaSubj1 <- glmGammaSubj
glmGammaSubj2 <- glm(clo ~ subjId + poly(tInOp, 2) + poly(tOut, 2) +
                       subjId:poly(tInOp, 2) + subjId:poly(tOut, 2) +
                       poly(tInOp, 1):poly(tOut, 1) +
                       subjId:poly(tInOp, 1):poly(tOut, 1),
                     family = Gamma(link = "inverse"))
anova(glmGammaSubj2, glmGammaSubj1, test = "LRT")

drop1(glmGammaSubj, test = "LRT")
xtable(round(drop1(glmGammaSubj, test = "LRT"), 2))

# stepAIC(glmGammaSubj, direction = "backward")


#### Final model ###################################################################

summary(lmSqrt1)
AIC(lmSqrt1)


### Within subj:day autocorrelation ################################################

acf1SubjDay <- numeric(length(unique(subjDay)))
j <- 1
tempDf1 <- cbind(df_data, stdDevResid)

for (i in unique(subjDay)){
  tempDf2 <- filter(tempDf1, subjDay == i)
  if (length(unique(tempDf2$clo)) == 1){
    acf1SubjDay[j] <- 1
  } else{
    acf1SubjDay[j] <- acf(tempDf2$stdDevResid, lag.max = 1, plot = FALSE)$acf[2]
  }
```

```
  j <- j + 1
}

par(mfrow = c(1,1))
plot(acf1SubjDay,
     xlab = "Index for Subject:Day combination",
     ylab = "ACF (lag = 1)")
abline(0,0)
mean(acf1SubjDay)


### Different dispersions ###################################################

#### Optimizing dispersion parameters ########################################

# Model
glmGammaInv2 <- glm(clo ~ sex + poly(tInOp, 2) + poly(tOut, 2) +
                      sex:poly(tInOp, 2) + sex:poly(tOut, 2) +
                      poly(tInOp, 1):poly(tOut, 1) +
                      sex:poly(tInOp, 1):poly(tOut, 1),
                    family = Gamma(link = "inverse"))


# Negative log likelihood
objective <- function(beta, formula){

  Xmu <- model.matrix(formula)
  Xdispersion <- cbind(rep(1, 803), as.integer(df_data$sex == "female"))

  eta <- Xmu %*% beta[1:ncol(Xmu)]
  mu <- 1 / eta
  dispersion <- exp(Xdispersion %*% beta[c(ncol(Xmu) + 1, ncol(Xmu) + 2)])

  alpha <- 1 / dispersion
  beta <- mu * dispersion

  return(-sum(dgamma(clo, shape = alpha, scale = beta, log = TRUE)))
}

# Model formula
linModel <- clo ~ sex + poly(tInOp, 2) + poly(tOut, 2) +
  sex:poly(tInOp, 2) + sex:poly(tOut, 2) +
  poly(tInOp, 1):poly(tOut, 1) +
  sex:poly(tInOp, 1):poly(tOut, 1)
# linModel <- clo ~ tInOp + tOut

# Initial parameters
coef <- coefficients(glmGammaInv2)
initialBeta <- c(coef, "Overall_dispersion" = 1, "female_weight" = 1)
# initialBeta <- c(rep(1, 5))

# Optimization
opt <- nlminb(initialBeta, objective, formula = linModel,
              control = list(eval.max = 1000, iter.max = 1000))
opt


# Control of results
stdDevResid <- rstandard(glmGammaInv2, type = "deviance")
sd(stdDevResid[sex == "female"])^2 / sd(stdDevResid[sex == "male"])^2
exp(opt$par[length(opt$par)])

### Gammareg ###
# formula.mean = linModel
# Z1 <- as.integer(df_data$sex == "female")
# formula.shape = ~ Z1
# a=Gammareg(formula.mean, formula.shape, meanlink="log")
# summary(a)
# opt$par


#### Profile likelihood ######################################################

# Optimize zeta, for each value fixed tau (see page 34)
profile_objective <- function(tau, formula){

  inner_objective <- function(zeta, tau_input){
    objective(c(zeta, tau), formula)
  }

  zeta_length <- ncol(model.matrix(formula)) + 1
  initial_zeta <- c(rep(1, zeta_length))
  nlminb(initial_zeta, inner_objective, tau_input = tau)$objective
}
```

```
# Calculate profile log likelihoods
tau <- seq(opt$par[14] - 0.5, opt$par[14] + 0.5, 0.01)
profile_logLikelihood <- numeric(length(tau))
for (i in 1:length(tau)){
  profile_logLikelihood[i] <- profile_objective(tau[i], formula = linModel)
  print(c(i, tau[i], profile_logLikelihood[i]))
}

# Finding the profile likelihood confidence interval (see page 36)
profile_likelihood <- exp(-profile_logLikelihood)
profile_likelihood <- profile_likelihood / max(profile_likelihood)
# profile_likelihood <- exp(profile_likelihood)
L_CI_lower <- min(tau[profile_likelihood > exp(-(1 / 2) *  qchisq(0.95, df=1))])
L_CI_upper <- max(tau[profile_likelihood > exp(-(1 / 2) *  qchisq(0.95, df=1))])


# Finding the quadratic approximation (Wald Confidence intervals) (see page 23)
# observed_hessian <- hessian(objective, opt$par, formula = linModel)
# observed_hessian_tau <- - observed_hessian[14, 14]
observed_hessian_tau <- hessian(profile_objective, opt$par[14], formula = linModel)
quadratic_approx_logLik <- - opt$objective +
  (1 / 2) * -observed_hessian_tau * (tau - opt$par[14])^2
quadratic_approx_Lik <- exp(quadratic_approx_logLik)
quadratic_approx_Lik <- quadratic_approx_Lik / max(quadratic_approx_Lik)
# plot(tau, quadratic_approx_Lik)
# standard_error = sqrt(diag(solve(observed_hessian)))[14]
standard_error = sqrt(diag(solve(observed_hessian)))[1]


# Plot of profile likelihood and quadratic approximation.
par(mfrow = c(1,1))
plot(tau, profile_likelihood, type = "l",
     xlab="Female Dispersion Weight", ylab="Profile Likelihood",
     main="The comparison between Wald's and Likelihood based Confidence Intervals")
lines(tau, rep(exp(-(1 / 2) *  qchisq(0.95, df=1)), length(tau)), col = 2)
rug(L_CI_lower, ticksize = 0.2, lwd = 2, col = "red")
rug(L_CI_upper, ticksize = 0.2, lwd = 2, col = "red")
c(L_CI_lower, L_CI_upper)
abline(v = opt$par["female_weight"], lty = 2)
lines(tau, quadratic_approx_Lik, col = "blue")
rug(opt$par["female_weight"] - qnorm(0.975) * standard_error,
    ticksize = 0.2, lwd = 2, col = "blue")
rug(opt$par["female_weight"] + qnorm(0.975) * standard_error,
    ticksize = 0.2, lwd = 2, col = "blue")
opt$par["female_weight"] - qnorm(0.975) * standard_error * c(1, -1)
legend("topright", 95,
       legend=c("Profile likelihood", "Quadratic approximation",
                "95% confidence interval"),
       col=c("black", "blue", "red"), lty = 1:1, cex=0.8,
       inset = 0.02)




# - ##########################################################################
# Part B: Ear infections ######################################################


### Load the data and get initial overview ####################################

# Load data
df_data <- read.csv("data/earinfect.txt", sep = " ", header = TRUE)
str(df_data)
df_data <- mutate_at(df_data, c("swimmer", "location", "age", "sex"),
                     as.factor)
summary(df_data)
attach(df_data)


# Table data to get picture if balanced
table(swimmer, location)
table(swimmer, age)
table(swimmer, sex)
table(location, age)
table(location, sex)
table(age, sex)


### Plot of the data ##########################################################

# We plot the rates of infections for each group, to have a comparable Y-value
df_data$rate <- df_data$infections / df_data$persons
comb <- t(combn(colnames(df_data)[1:4], 2))
```

```
plot_list <- list()
for (i in 1:nrow(comb)){
  plot_list[[i]] <- ggplot(df_data,
                           aes_string(x=comb[i,1], y="rate", col=comb[i,2])) +
    geom_boxplot() +
    labs(x = comb[i,1],
         y = "Rate",
         col = comb[i,2]) +
    theme(plot.title = element_text(hjust=0.5, size=14, face="bold"),
          axis.title = element_text(size=12, face="bold"),
          legend.title = element_text(size=12, face="bold",),
          axis.text = element_text(size=12, face="bold"))


}
grid.arrange(grobs = plot_list, nrow = 3, ncol = 2)

df_data = subset(df_data, select = -c(rate))
rm(list=setdiff(ls(), "df_data"))


### Fit Poisson model ##########################################################

# Second order interaction model
m1 <- glm(infections ~ (swimmer + location + age + sex)^2,
          family = poisson(link = "log"), offset = log(persons),
          data = df_data)
summary(m1)

# Full model
mf <- glm(infections ~ swimmer*age*sex*location,
          family = poisson(link = "log"), offset = log(persons),
          data = df_data)
summary(mf)


### Model diagnostics ##########################################################

# Initial Goodnes of fit test
dev <- summary(m1)$deviance
df <- summary(m1)$df.resid
p <- 1 - pchisq(dev, df)
p # p = 0.98 so the data do not reject the initial model.


## Model diagnostic plots ##
par(mfrow=c(2,2))
plot(m1, which = 1:4)
par(mfrow=c(1,1))

# Plot of deviance residuals against the different factors
stddevresid <- rstandard(m1, type = "deviance")
par(mfrow=c(2,2))
plot(swimmer, stddevresid)
abline(h=0, col="red")
plot(location, stddevresid)
abline(h=0, col="red")
plot(age, stddevresid)
abline(h=0, col="red")
plot(sex, stddevresid)
abline(h=0, col="red")
par(mfrow=c(1,1))


### Model reduction ############################################################

# Reduction of second order interaction model
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - swimmer:sex)
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - swimmer:age)
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - location:age)
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - age:sex)
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - age)
drop1(m1, test = "Chisq")
```

```
m1 <- update(m1, . ~ . - swimmer:location)
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - swimmer:location)
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - swimmer)
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - location:sex)
drop1(m1, test = "Chisq")

m1 <- update(m1, . ~ . - sex)



# # Control of result
# test <- glm(infections ~ swimmer + age + sex + location +
#                          swimmer:age + swimmer:sex + swimmer:location +
#                          age:sex + age:location,
#                        family = poisson(link = "log"), offset = log(persons),
#                        data = df_data)
# tdev <- summary(test)$deviance
# tdf <- summary(test)$df.resid
# d <- tdev - dev
# 1 - pchisq(d, 1)


### Post-Hoc analysis ##########################################################

summary(m1)
m1 <- update(m1, . ~ . - 1)
summary(m1)


# Confidence intervals
confint(m1)
```