Assignment 1

02424 Advanced Dataanalysis and Statistical Modelling

Clara Brimnes Gardner (s153542)
Tobias Engelhardt Rasmussen (s153057)
Christian Dandanell Glissov (s146996)

May 26, 2021

Introduction and Description of Data

This assignment concerns the level of clothing worn at offices. The analysis is based on a data-set which contains 6 variables. These are described in Table 1.

Variable	Type	Description
`clo`	Continuous	Level of clothing
`tOut`	Continuous	Outdoor temperature
`tInOp`	Continuous	Indoor operating temperature
`sex`	Factor	Sex of the subject
`subjId`	Factor	Identifier for subject
`day`	Factor	Day (within the subject)

Table 1: *List of included variables in the data-set.*

A General Linear Model

In this section we model the clothing level, `clo`, by using the explanatory variables `tInOp`, `tOut` and `sex`.

Explorative Analysis

Table 2 shows summary statistics for the three continuous variables, `clo`, `tInOp` and `tOut`.

Variable	Variance	Min	25% Quantile	Median	Mean	75% Quantile	Max
`clo`	0.02295	0.2467	0.4700	0.5483	0.5511	0.6412	0.9600
`tInOp`	1.694	23.11	26.01	26.94	26.82	27.48	29.55
`tOut`	17.28	11.93	18.57	21.01	21.54	24.25	33.08

Table 2: *Summary Statistics of the three continuous variables*

Figure 1 shows boxplots of `clo`, `tInOp` and `tOut`. The figure shows that the distribution of `tInOp` and `tOut` are quite similar for the two sexes. However the distribution of `clo` varies - the variance in `clo` is much larger for females than for males. The median value of `clo` is lower for males.

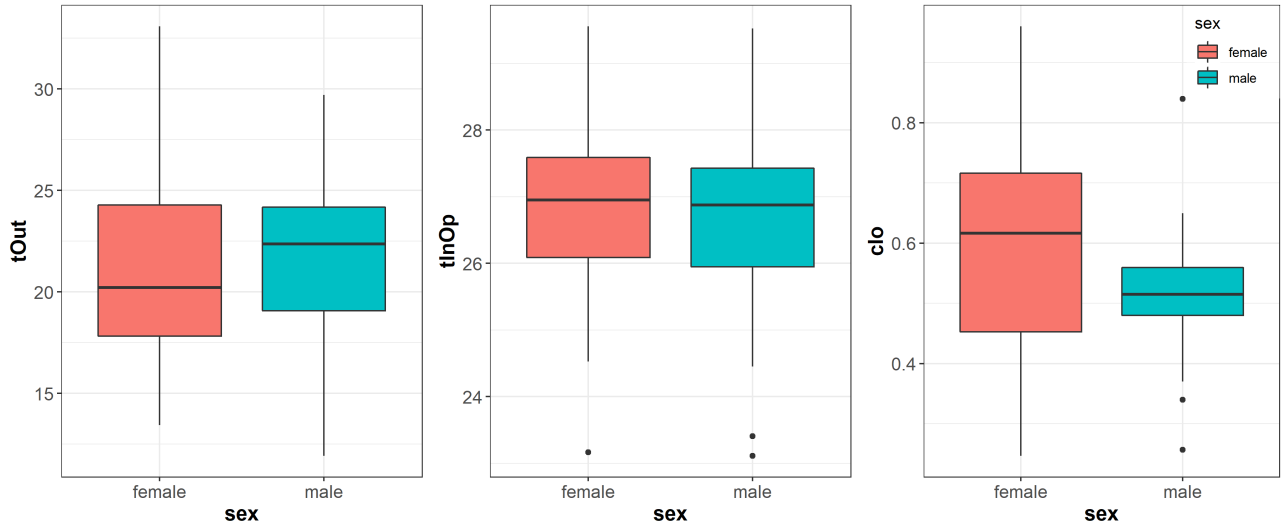


Figure 1: *Boxplots of the three continuous variables for each sex*

Figure 2 shows the values of `clo` as a function of `tInOp` and `tOut` respectively. The level of `clo` does not change much for males as the variance is smaller. For females the value of `clo` decreases when `tInOp` and `tOut` increases. One thing to notice is there seem to be an outlier with clothing level on 0.84 for male.

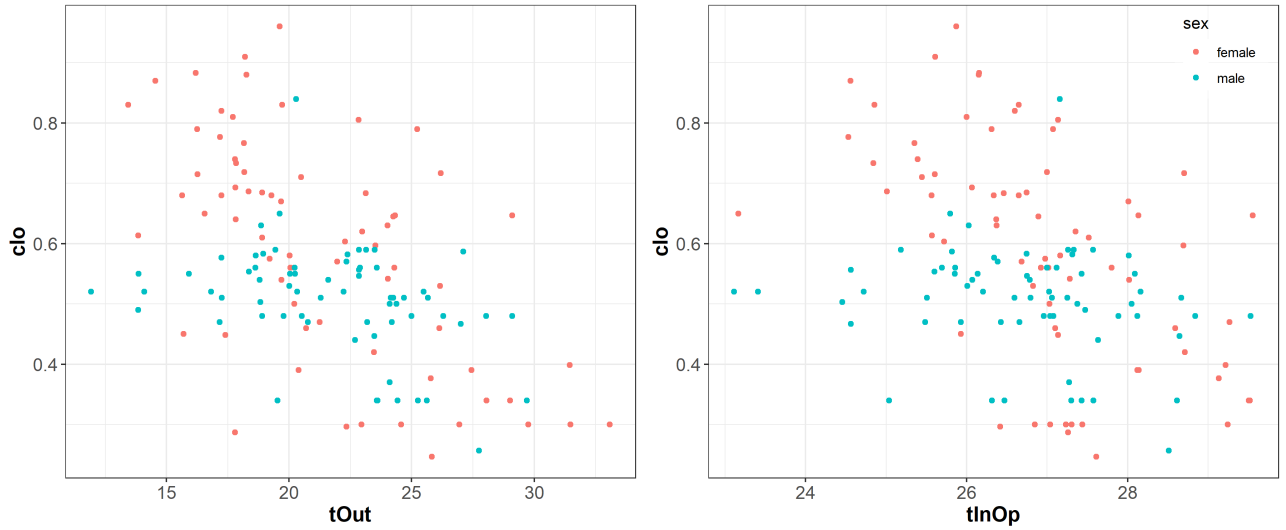


Figure 2: *The values of* `clo` *against the values of* `tInOp` *and* `tOut` *respectively. The colors represents* `sex`

Looking at the density of Figure 3, it is clear that the variance is larger, with most of the density of the data for male based on narrow density peaks. Comparing male and female for the temperatures the data seem to be well distributed, with not much a change in variance, which means that that the two genders have been subject to similar environments.

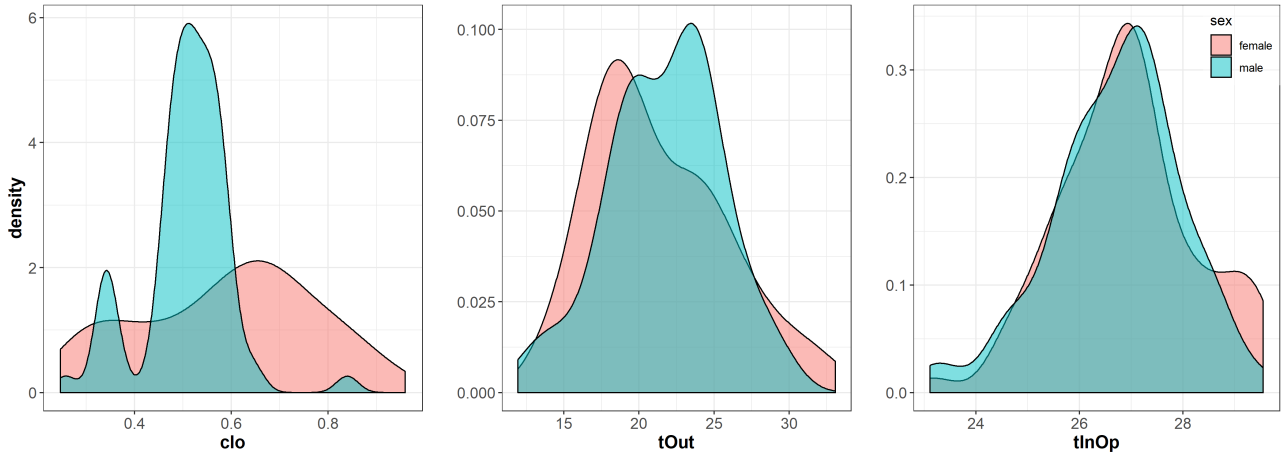


Figure 3: *The density of* `clo`, `tInOp` *and* `tOut` *respectively. The colors represents* `sex`

In Table 3 the correlation can be seen of the variables. High levels of correlation is not observed.

	clo	tOut	tInOp
clo	1.00		
tOut	-0.51	1.00	
tInOp	-0.38	0.52	1.00

Table 3: *Correlation matrix of* `clo`, `tInOp` *and* `tOut`

Fitting a General Linear Model

A general linear model is now fit to the data. We first consider a model with the variables `tInOp`, `tOut` and `sex`. For the numerical variables `tInOp` and `tOut` we consider up to second-order terms. Furthermore we consider interactions between the `sex` and the numerical terms, as well as the three way-interaction between `tInOp`, `tOut` and `sex`. To account for numerical issues and higher order correlations we subtract the mean-value for all numerical variables ($\mu_{\texttt{tInOp}}$ and $\mu_{\texttt{tOut}}$). The initial model is therefore on the form

$$\begin{aligned}\texttt{clo}_i =&\beta_0(\texttt{sex}_i) + \beta_1(\texttt{sex}_i)\cdot(\texttt{tInOp}_i - \mu_{\texttt{tInOp}}) + \beta_2(\texttt{sex}_i)\cdot(\texttt{tOut}_i.\mu_{\texttt{tOut}}) + \beta_4(\texttt{sex}_i)\cdot(\texttt{tInOp} - \mu_{\texttt{tInOp}})\\ &\cdot(\texttt{tOut} - \mu_{\texttt{tOut}}) + \beta_5(\texttt{sex}_i)\cdot(\texttt{tInOp} - \mu_{\texttt{tInOp}})^2 + \beta_6(\texttt{sex}_i)\cdot(\texttt{tOut} - \mu_{\texttt{tOut}})^2 + \epsilon_i\end{aligned} \tag{1}$$

where $\epsilon_i \sim N(0,\sigma)$ and $\texttt{sex}_i \in (male, female)$. The significance of the different effects are tested using Likelihood-ratio-test. Table 4 shows an anova table of the model with the relevant p-values

Parameter	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
`sex`	1	0.25	0.25	18.51	0.0000
$(\texttt{tOut} - \mu_{\texttt{tOut}})$	1	0.78	0.78	56.73	0.0000
$(\texttt{tInOp} - \mu_{\texttt{tInOp}})$	1	0.09	0.09	6.62	0.0112
$I(\texttt{tOut} - \mu_{\texttt{tOut}})^2$	1	0.07	0.07	4.83	0.0299
$(\texttt{tOut} - \mu_{\texttt{tOut}})^2$	1	0.00	0.00	0.21	0.6485
$sex : I(\texttt{tOut} - \mu_{\texttt{tOut}})$	1	0.06	0.06	4.34	0.0393
$sex : I(\texttt{tInOp} - \mu_{\texttt{tInOp}})$	1	0.04	0.04	2.63	0.1072
$I(\texttt{tOut} - \mu_{\texttt{tOut}}) : I(\texttt{tInOp} - \mu_{\texttt{tInOp}})$	1	0.02	0.02	1.70	0.1943
$sex : I(\texttt{tOut} - \mu_{\texttt{tOut}})^2$	1	0.00	0.00	0.27	0.6048
$sex : I(\texttt{tInOp} - \mu_{\texttt{tInOp}})^2$	1	0.01	0.01	0.63	0.4278
$sex : I(\texttt{tOut} - \mu_{\texttt{tOut}}) : I(\texttt{tInOp} - \mu_{\texttt{tInOp}})$	1	0.09	0.09	6.27	0.0135
Residuals	124	1.69	0.01		

Table 4: *Anova table for the initial model given in* (1)

The model is reduced. Interaction-terms are removed before main-effects, and higher order terms are removed before lower order terms. By refitting the model the following effects are removed in the following order:

1. $(\texttt{tOut} - \mu_{\texttt{tOut}})^2 \times \texttt{sex}$
2. $(\texttt{tInOp} - \mu_{\texttt{tInOp}})^2 \times \texttt{sex}$
3. $(\texttt{tInOp} - \mu_{\texttt{tInOp}})^2$
4. $(\texttt{tInOp} - \mu_{\texttt{tInOp}}) \times (\texttt{tOut} - \mu_{\texttt{tOut}}) \times \texttt{sex}$
5. $(\texttt{tInOp} - \mu_{\texttt{tInOp}}) \times (\texttt{tOut} - \mu_{\texttt{tOut}})$
6. $(\texttt{tOut} - \mu_{\texttt{tOut}}) \times \texttt{sex}$
7. $(\texttt{tOut} - \mu_{\texttt{tOut}})^2$

The final model is therefore on the form

$$\boldsymbol{\mu} = \beta_0(\texttt{sex}_i) + \beta_1(\texttt{sex}_i)\cdot(\texttt{tInOp}_i - \mu_{\texttt{tInOp}}) + \beta_2\cdot(\texttt{tOut} - \mu_{\texttt{tOut}})$$

The anova-table for the final model is shown in table

Parameter	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
`sex`	1	0.25	0.25	17.66	0.0000
$\texttt{tOut} - \mu_{\texttt{tOut}}$	1	0.78	0.78	54.13	0.0000
$\texttt{tInOp} - \mu_{\texttt{tInOp}}$	1	0.09	0.09	6.32	0.0132
$(\texttt{tInOp} - \mu_{\texttt{tInOp}}) \times \texttt{sex}$	1	0.10	0.10	7.25	0.0080
Residuals	131	1.88	0.01		

The parameters for the final model are

Variable	Estimate	Standard Error
$\beta_0(\texttt{sex} = female)$	0.595724	0.014396
$\beta_0(\texttt{sex} = male)$	0.508458	0.014806
$\beta_1(\texttt{sex} = female)$	-0.047494	0.012912
$\beta_1(\texttt{sex} = male)$	-0.002893	0.012033
β_2	-0.012204	0.003024

Table 5: *Parameters for the final model along with their standard-error*

Table 5 reveals that the level of clothing is higher for females at the mean-value of `tInOp` and `tOut`, as the intercept is higher for females. Both for males and females, the level of clothing decreases when `tInOp` increases - however the slope is much larger for females. When the `tOut` increases the level of clothing also decreases. Figure 4 shows graphical presentations of the model - predictions, confidence-intervals and prediction intervals as a function of `tInOp` and `tOut` respectively. In both cases the value of the other parameter is taken as the mean. The described tendencies are evident in the plots. It is seen that the slope is nearly flat for men in figure 4.A, while it is much steeper for women. In figure 4.B the slope is the same for men and women, but due to a higher intercept for women the lines are not overlapping.

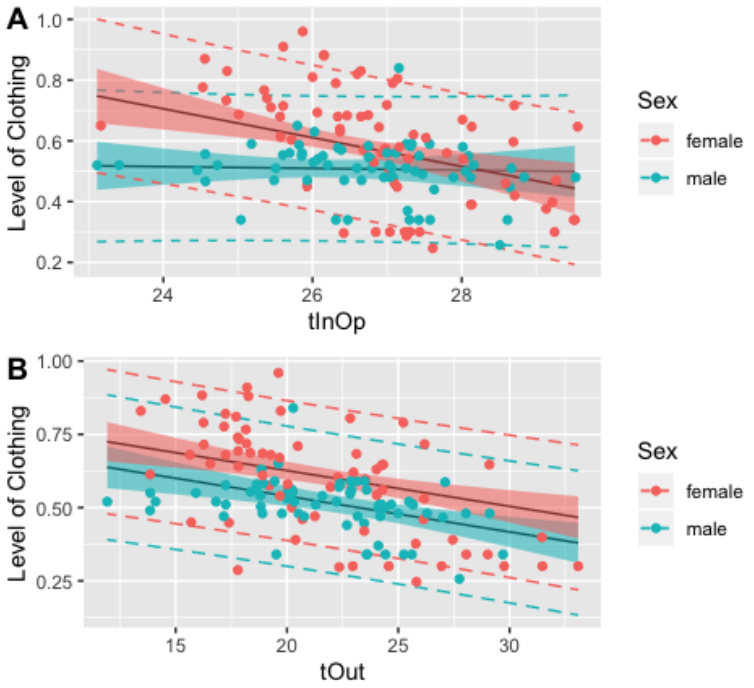


Figure 4: *A: The predictions, the confidence intervals and the prediction intervals of the linear model for different values of* `tInOp`*. The value of* `tOut` *is taken to be the mean,* $\mu_{\texttt{tOut}}$*. B: The predictions, the confidence intervals and the prediction intervals of the linear model for different values of* `tOut`*. The value of* `tInOp` *is taken to be the mean,* $\mu_{\texttt{tInOp}}$*.*

The assumptions for the linear model are now checked. We impose four assumptions:

1. The model structure captures the systematic behaviour of the data
2. The residuals follow a normal-distribution
3. The variance of the residuals is constant
4. The residuals are independent.

To check whether the residuals follow a normal-distribution a qq-plot is used. Figure 5 shows a QQ-plot. Considering the red and blue lines, the tail issues mostly seem to be due to the observations of females.

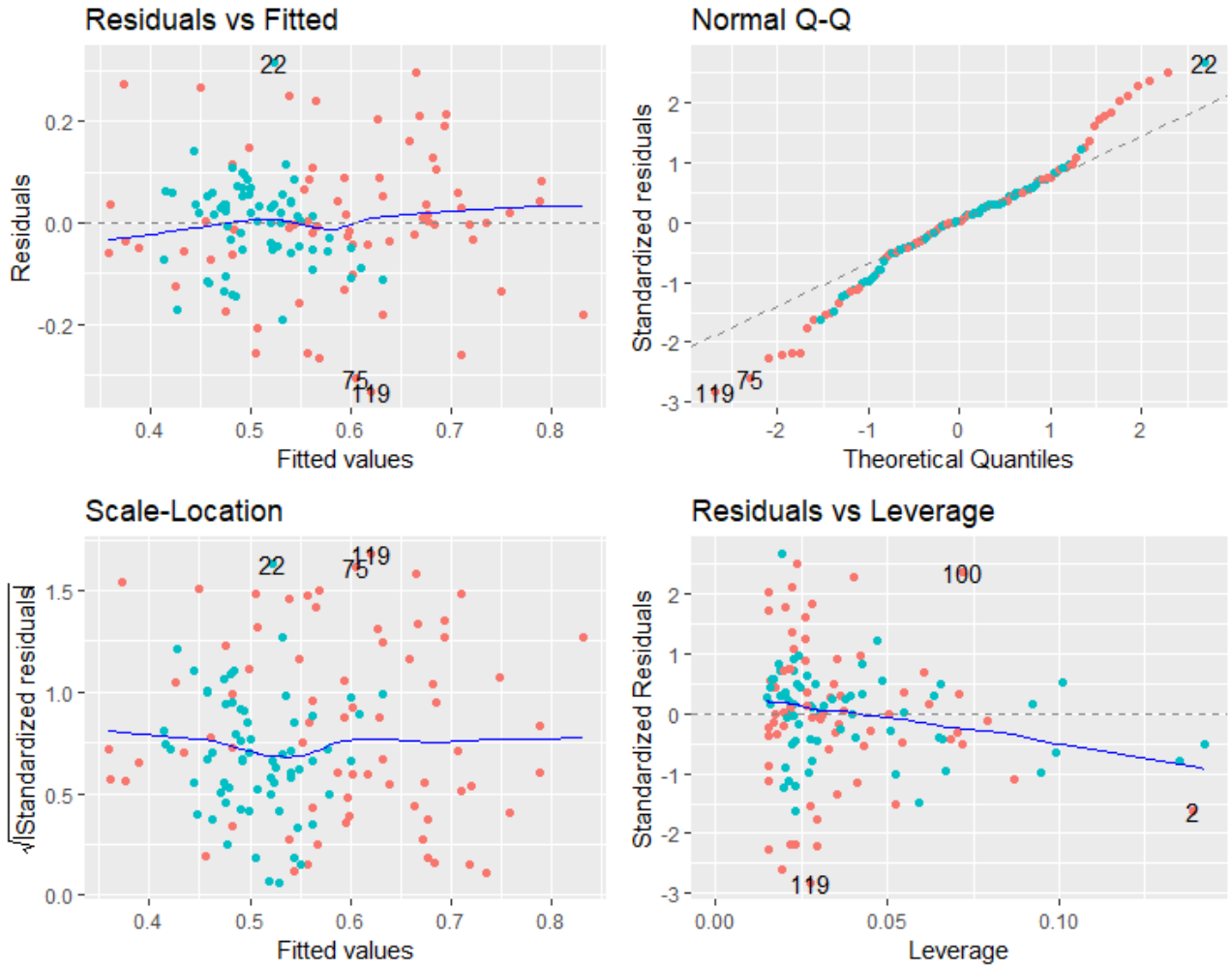


Figure 5: *Diagnostic plots for the unweighted model in item*

To check whether the variance of the residuals are constant, the residuals are plotted against the fitted values. The plot is also shown in figure 5. There seems to be a problem with variance homogeneity for two reasons. Firstly there seems to be a decreasing trend; the residuals decrease when the fitted values increase. Secondly the variance for females and males are very different with the variance for female observations being much larger than the variance for male observations.
At last the residuals are plotted against the three explanatory variables. This can help see if the residuals are indeed independent, and if there are hidden trends which had not been taken into account (for instance higher order terms in the numerical variables). Figure 6.A) shows the residuals against the two values of `sex`. It is seen that the variance of the residuals of female residuals are larger than for the males. For `tInOp` and `tOut` there seem to be no alarming trends.

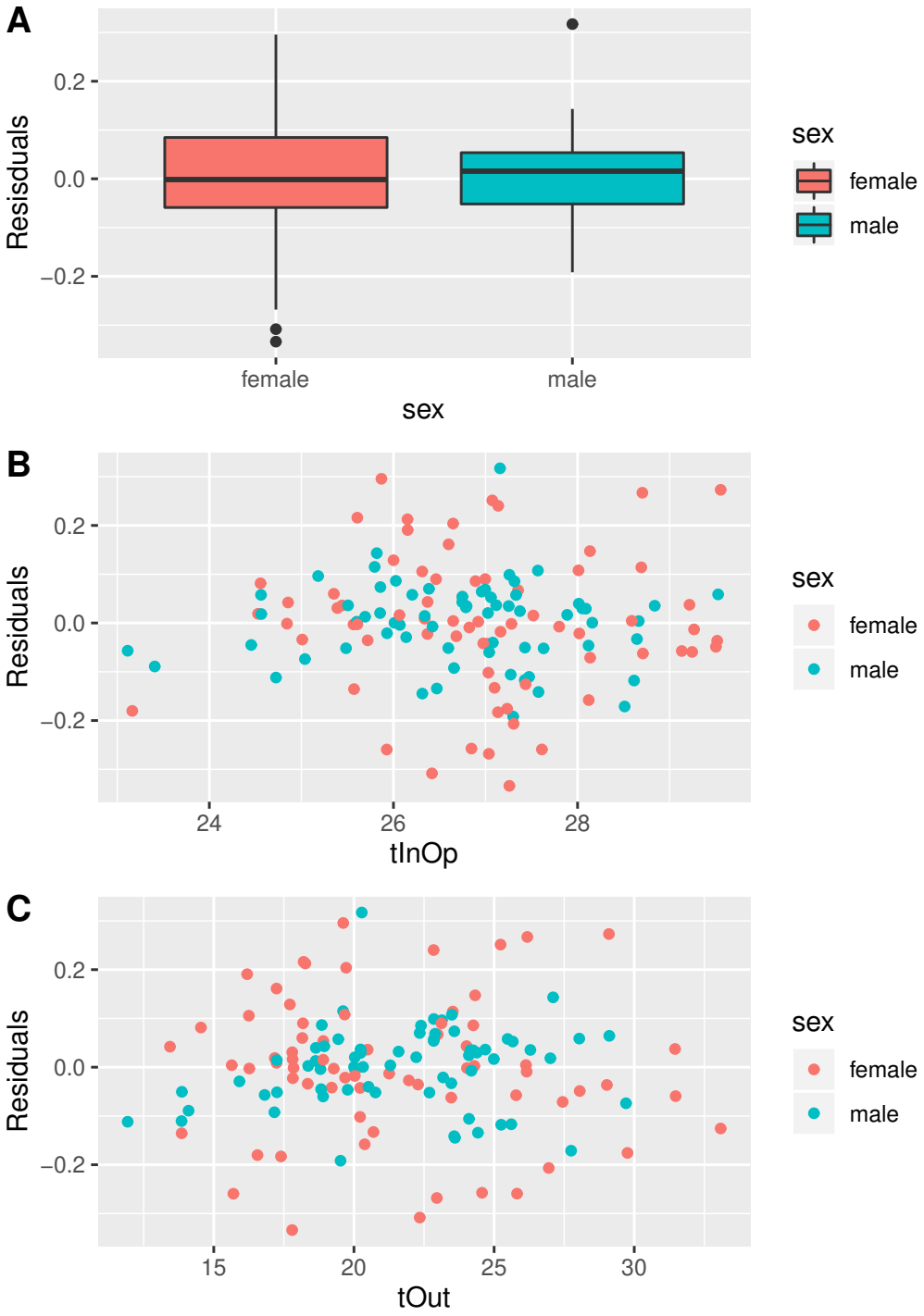


Figure 6: *The residuals against the three explanatory variables*

The residual analysis shows that there is an issue with the sex-differences, as the variance seems to be very different for male and female observations.

A Weighted Approach

One way to account for the different variances for the female and male observations, is to weight the variance of the observations. We consider an approach where male observations are assigned one weight, and female

observations are assigned another weight. The model can then be written as

$$\texttt{clo}_i \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}), \tag{2}$$

where

$$\begin{aligned}\boldsymbol{\mu} &= \beta_0(\texttt{sex}_i) + \beta_1(\texttt{sex}_i)\cdot(\texttt{tInOp}_i - \mu_{\texttt{tInOp}}) + \beta_2(\texttt{sex}_i)\cdot(\texttt{tOut}_i - \mu_{\texttt{tOut}}) + \beta_4(\texttt{sex}_i)\cdot(\texttt{tInOp} - \mu_{\texttt{tInOp}}) \\ &\quad \cdot(\texttt{tOut} - \mu_{\texttt{tOut}}) + \beta_5(\texttt{sex}_i)\cdot(\texttt{tInOp} - \mu_{\texttt{tInOp}})^2 + \beta_6(\texttt{sex}_i)\cdot(\texttt{tOut} - \mu_{\texttt{tOut}})^2 \\ \boldsymbol{\Sigma}_{i,j} &= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \text{ and } \texttt{sex}_i = male \\ c & \text{if } i = j \text{ and } \texttt{sex}_i = female \end{cases}\end{aligned} \tag{3}$$

The value of c is estimated using maximum-likelihood estimation. The log-likelihood function for a multivariate normal distribution is given by the log of the product of individual densities.

$$l(\mu, \boldsymbol{\Sigma}|\texttt{clo}_i) = \log\prod_{i=1}^{n} f_{\texttt{clo}}(\texttt{clo}_i)|\mu, \boldsymbol{\Sigma}) \tag{4}$$

$$= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\texttt{clo}_i - \mu)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\texttt{clo}_i - \mu) \tag{5}$$

Figure 7 shows the likelihood-function for the model given in (2) and (3) as a function of c. By using a numerical optimizer the optimal value for c is found to be $c = 2.94$. The log-likelihood is normalized such that the maximal log-likelihood is zero.

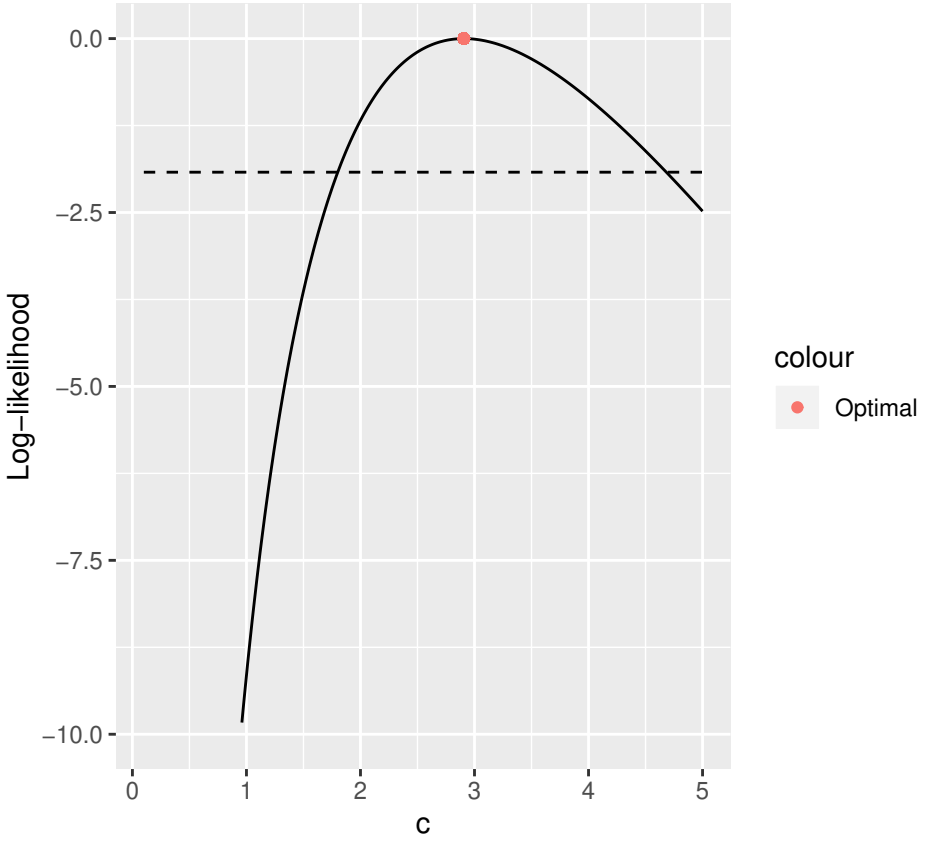


Figure 7: *The likelihood as a function of the weight parameter c. The dotted line indicates* $\dfrac{\chi^2(0.95, df = 1)}{2} = -1.92$

The significance of the weight-parameter can be tested using a likelihood-ratio test. We test the null-hypothesis

$$\mathcal{H}_0 : c = 1$$

against the alternative

$$\mathcal{H}_1 : c \neq 1$$

To carry out this test, the test-statistic given in (6) is calculated, and it is hereafter compared to χ^2-distribution with one degree of freedom. To account for boundary behaviour we divide the p-value with 2.

$$\chi_{obs} = -2(\log(L)(\theta_0) - \log L(\theta_1)) \tag{6}$$

The test-statistic and the p-value becomes

$$\begin{aligned} \chi_{obs} &= -2(104.83 - 113.99) = 18.30 \\ p &= \frac{\Pr\left[\chi_{obs} < \chi^2(0.95, 1)\right]}{2} = 9.39 \cdot 10^{-6}, \end{aligned}$$

which means that the weight parameter is highly significant. The ANOVA-table for the full-model using $c = 2.94$ is shown in table 6

Parameter	**SumSq**	**Df**	**F value**	**Pr(>F)**
Intercept	3.92	1	568.53	0.0000
sex	0.07	1	9.60	0.0024
$\texttt{tOut} - \mu_{\texttt{tOut}}$	0.05	1	7.76	0.0062
$\texttt{tInOp} - \mu_{\texttt{tInOP}}$	0.03	1	3.94	0.0493
$(\texttt{tOut} - \mu_{\texttt{tOut}})^2$	0.03	1	4.60	0.0340
$(\texttt{tInOp} - \mu_{\texttt{tInOP}})^2$	0.02	1	2.75	0.0997
$\texttt{tOut} - \mu_{\texttt{tOut}} \times \texttt{sex}$	0.01	1	1.13	0.2891
$\texttt{tInOp} - \mu_{\texttt{tInOP}} \times \texttt{sex}$	0.01	1	1.46	0.2295
$(\texttt{tOut} - \mu_{\texttt{tOut}}) \times (\texttt{tInOp} - \mu_{\texttt{tInOP}})$	0.03	1	5.07	0.0262
$(\texttt{tOut} - \mu_{\texttt{tOut}})^2 \times \texttt{sex}$	0.01	1	1.21	0.2732
$(\texttt{tInOp} - \mu_{\texttt{tInOP}})^2 \times \texttt{sex}$	0.01	1	1.86	0.1752
$(\texttt{tOut} - \mu_{\texttt{tOut}}) \times (\texttt{tInOp} - \mu_{\texttt{tInOP}}) \times \texttt{sex}$	0.03	1	5.02	0.0268
Residuals	0.86	124		

Table 6: *ANOVA-table for the model given in* (2) *and* (3)

The model can now be reduced as done previously. The procedure differs from the unweighted approach, as a new weight is estimated for every reduction in the model. That is the model reduction procedure is as follows

1. Choose an initial model
2. Estimate the weights using maximum likelihood
3. Test is there are any insignificant effects (including the weights)
4. Remove most insignificant effect. If no effects are removed then the model reduction is done. If an effect is removed return to step 2.

Using this approach the final model becomes

$$\begin{aligned} \boldsymbol{\mu} &= \beta_0(\texttt{sex}_i) + \beta_1(\texttt{sex}_i) \cdot (\texttt{tInOp}_i - \mu_{\texttt{tInOp}}) + \beta_2 \cdot (\texttt{tOut} - \mu_{\texttt{tOut}}) + \beta_3(\texttt{tOut} - \mu_{\texttt{tOut}})^2 \\ \boldsymbol{\Sigma}_{i,j} &= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \text{ and } \texttt{sex}_i = male \\ 3.16 & \text{if } i = j \text{ and } \texttt{sex}_i = female \end{cases} \end{aligned} \tag{7}$$

The estimates of the parameters are given in table 8 Table 8.

Variable	Estimate	Standard Error
$\beta_0(\texttt{sex} = female)$	0.6172887	0.0192558
$\beta_0(\texttt{sex} = male)$	0.5221356	0.0117773
$\beta_1(\texttt{sex} = female)$	-0.0495678	0.0142751
$\beta_1(\texttt{sex} = male)$	-0.0080188	0.0085080
β_2	-0.0098442	0.0025133
β_3	-0.0010435	0.0004282

Table 7: *Parameters for the final weighted linear model along with their standard-error*

The MLE of the weight parameter is found to be $c = 3.16$, and the corresponding p-value is $2.63 \cdot 10^{-6}$.

Figure 8 shows a graphical presentation of the final model. The main difference from the unweighted model presented in figure 4 is that the fit is a second order polynomial for `tOut`.

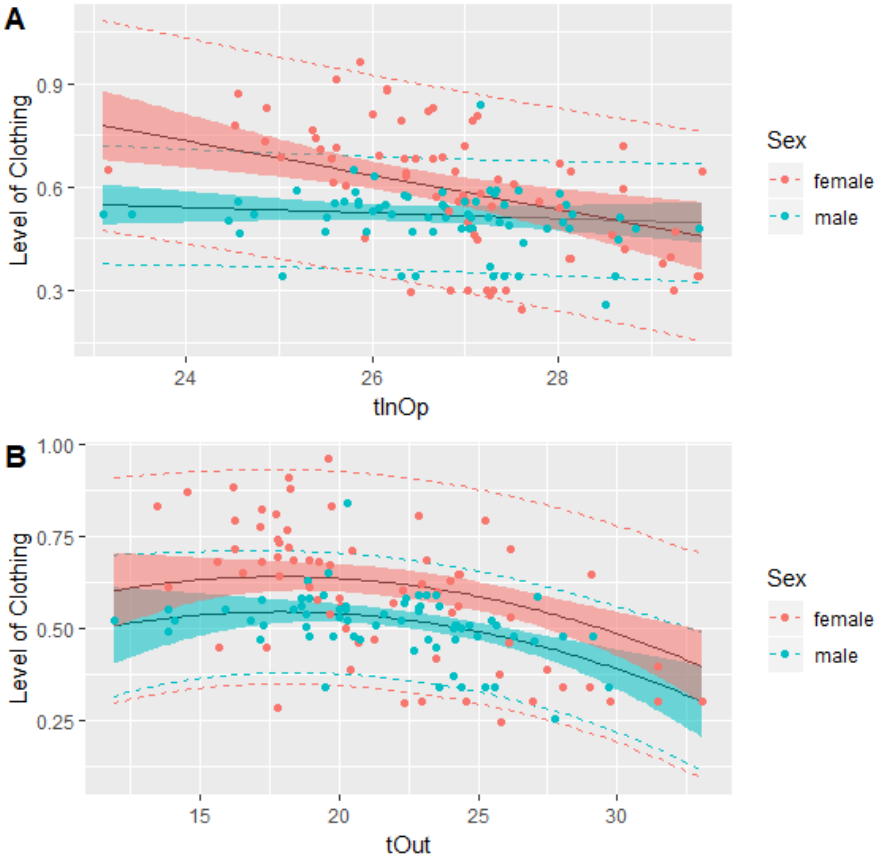


Figure 8: *The confidence (shaded) and prediction (two-dashed) intervals of the linear model for different values of* `tInOp` *and* `tOut`*, respectively. Once again the values* `tOut`*,* `tInOp` *is taken to be the mean,* $\mu_{\texttt{tOut}}$ *and* $\mu_{\texttt{tInOp}}$*.*

Once again as seen from the coefficients the clothing level decreases for both men and female as the temperature both indoor or outdoor increases. It is also seen that the variance is higher for females as expected, while the clothing level for male don't decrease as much in a rise of temperature. Looking at the assumptions of the residuals in Figure 9.

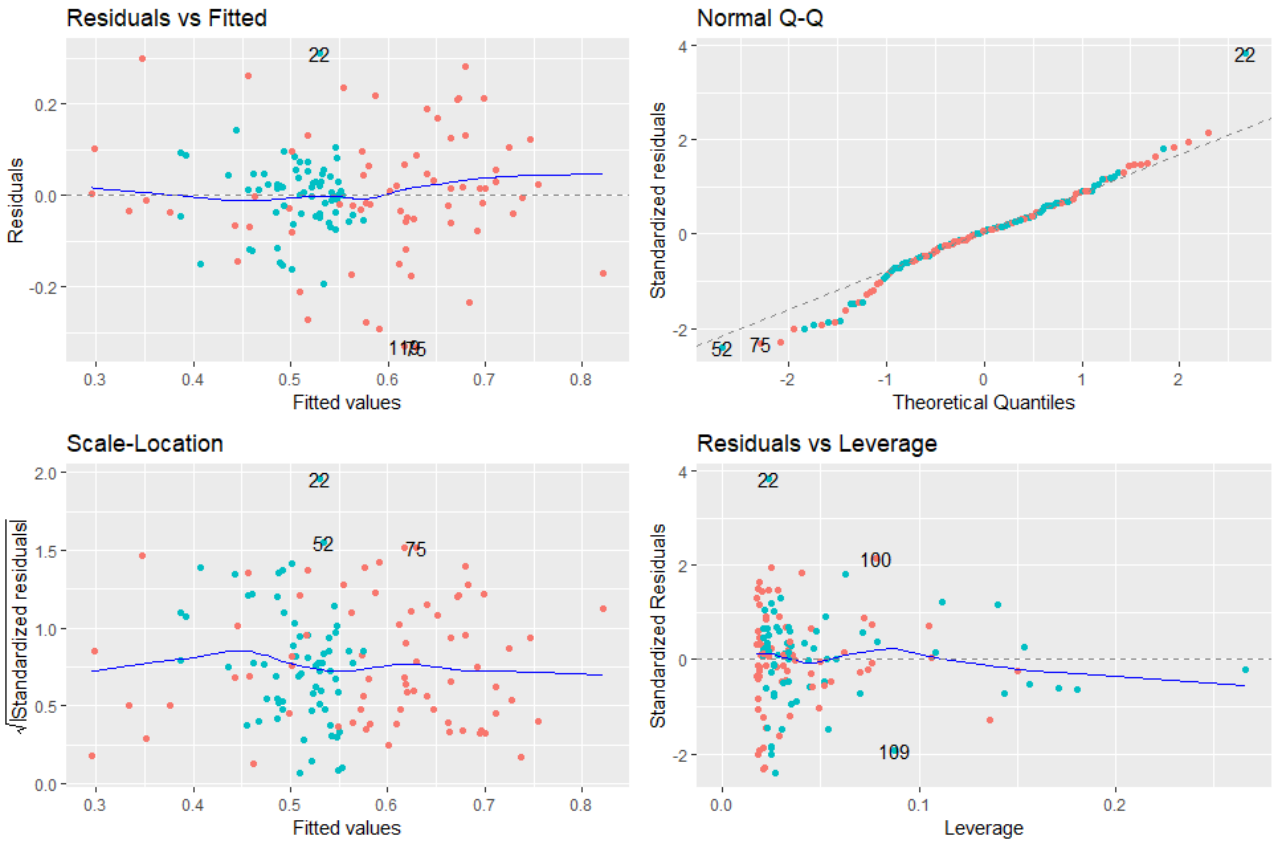


Figure 9: *Diagnostic plots for the weighted model in Equation 7*

Compared to the unweighted Q-Q plot in Figure 5 the residuals of this model seem more normal, hence we are more likely to accept the satisfaction of the assumption on normally distributed residuals for this model. This is due to the variance being smaller in the male group which will therefore be penalized less because of the higher weighting compared to the female group. The decreasing trend for the residuals against the fitted values also seem to have disappeared, there is no clear decreasing trend. Observation 22 is still an outlier for males, this is also seen Figure 3, because the value is still within a valid range (`clo` $\in [0, 1]$) it will not be removed. Figure 10 shows the residuals against the three explanatory variables. Besides that, only small differences are noticed when comparing with figure 6.

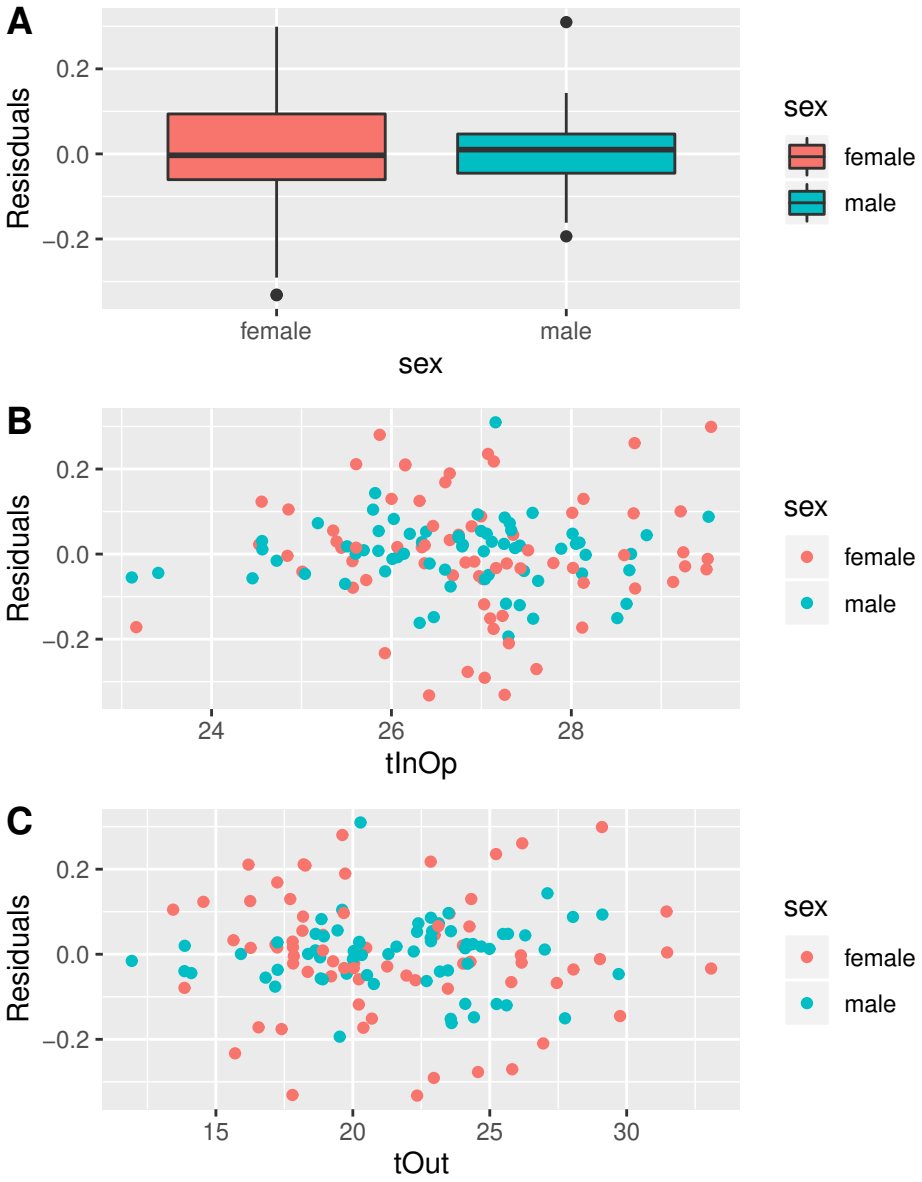


Figure 10: *The residuals against the three explanatory*

Including subject Id

The data set also includes the variable named `subjId`, which is an identifier for the individual subjects. To investigate if it should be included in the model, boxplots of the residuals of the model are plotted against the subject identifiers. Looking at Figure 11 a varying variance within each subject identifier can be observed, some more than others. It seems that the amount of clothing is very much a matter of taste of the individual, hence should be attempted added to the model to add the extra information, which we will do.

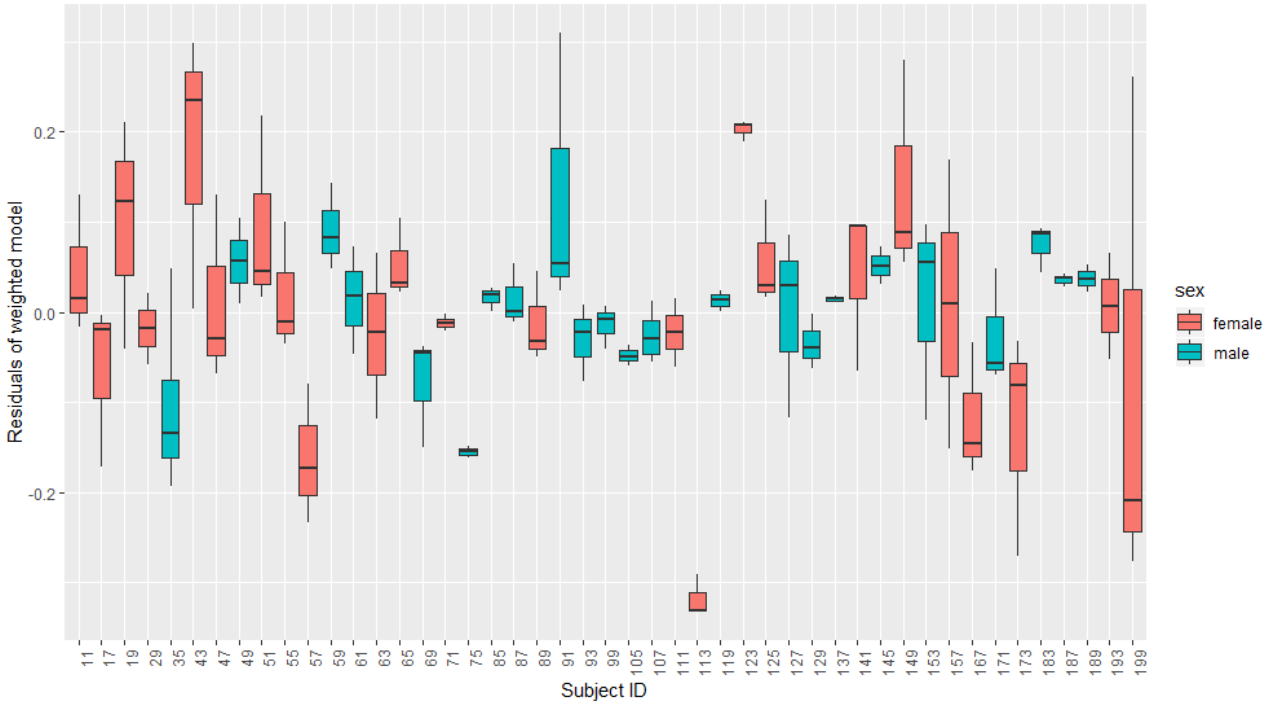


Figure 11: *Subject ID against residuals of the weighted model given in* (7)

Due to the limited data set and the numerous individuals that it consists of, it is not possible to try out complicated models with interactions, hence we will start by trying a simple additive model. Since the `subjId` variable already holds the information about sex differences we will not add this, however the interaction between `sex` and the different temperature variables will be added since these are cheaper in relation to degrees of freedom. The initial model does not contain all the higher order terms and interactions used in the other initial models ((1) and (3)) because of the many degrees of freedom, that `subjId` uses. This model will also include weights that are optimized at each step of the reduction. The initial model is:

$$\begin{aligned}\texttt{clo}_i =& \beta_0(\texttt{subjectID}_i) + \beta_1(\texttt{sex}_i)\cdot(\texttt{tInOp}_i - \mu_{\texttt{tInOp}}) + \beta_2(\texttt{sex}_i)\cdot(\texttt{tOut}_i - \mu_{\texttt{tOut}}) \\ &+ \beta_5)\cdot(\texttt{tInOp} - \mu_{\texttt{tInOp}})^2 + \beta_6\cdot(\texttt{tOut} - \mu_{\texttt{tOut}})^2 + \epsilon_i\end{aligned} \tag{8}$$

Where $\texttt{subjId}_i$ is one of the 47 different subject IDs, and where $\texttt{sex}_i$ is the corresponding sex. The ANOVA table for the initial model looks as follows: The model is reduced using the p-values of the type III partitioning,

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
subjId	46	2.20	0.05	5.88	0.0000
$I(\texttt{tInOp} - \mu_{\texttt{tInOp}})$	1	0.01	0.01	1.83	0.1802
$I(\texttt{tOut} - \mu_{\texttt{tOut}})$	1	0.16	0.16	20.16	0.0000
$I((\texttt{tInOp} - \mu_{\texttt{tInOp}})^2)$	1	0.00	0.00	0.36	0.5483
$I((\texttt{tOut} - \mu_{\texttt{tOut}})^2)$	1	0.02	0.02	2.97	0.0887
$\texttt{sex} : I(\texttt{tInOp} - \mu_{\texttt{tInOp}})$	1	0.01	0.01	1.29	0.2591
$\texttt{sex} : I(\texttt{tOut} - \mu_{\texttt{tOut}})$	1	0.00	0.00	0.04	0.8456
Residuals	83	0.68	0.01		

removing higher order terms before lower order terms. By refitting the model the following effects are removed in the following order:

1. $I((\texttt{tInOp} - \mu_{\texttt{tInOp}})^2)$
2. $\texttt{sex} : I(\texttt{tOut} - \mu_{\texttt{tOut}})$
3. $\texttt{sex} : I(\texttt{tInOp} - \mu_{\texttt{tInOp}})$

4. $I(\texttt{tInOp} - \mu_{\texttt{tInOp}})$

Hence the final model will be of the form and with the following weight matrix:

$$\boldsymbol{\mu} = \beta_0(\texttt{subjId}_i) + \beta_1 \cdot (\texttt{tOut} - \mu_{\texttt{tOut}}) + \beta_2(\texttt{tOut} - \mu_{\texttt{tOut}})^2$$
$$\boldsymbol{\Sigma}_{i,j} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \text{ and } \texttt{sex}_i = male \\ 2.99 & \text{if } i = j \text{ and } \texttt{sex}_i = female \end{cases} \tag{9}$$

Which means that only the individual and the outdoor temperature will have an impact on the amount of clothing people wear, and that the females are weighted one third of the men. The final anova-table becomes:

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
subjId	47	26.01	0.55	140.29	0.0000
$I(\texttt{tOut} - \mu_{\texttt{tOut}})$	1	0.10	0.10	26.58	0.0000
$I((\texttt{tOut} - \mu_{\texttt{tOut}})^2)$	1	0.03	0.03	7.52	0.0074
Residuals	87	0.34	0.00		

Each subject ID will have its own intercept and they will all have a dependency on the outdoor temperature with estimated parameters:

Variable	**Estimate**	**Standard Error**
β_1	-0.0134558	0.0025857
β_2	-0.0012300	0.0004485

Table 8: *The slope and curvature given the outdoor temperature of the model including subject IDs*

The rest of the parameters which are the intercepts of all the individul subjects can be found in appendix in Table 11. To investigate the assumptions on the residuals we look at the residuals for the model, which are shown in figure Figure 12.

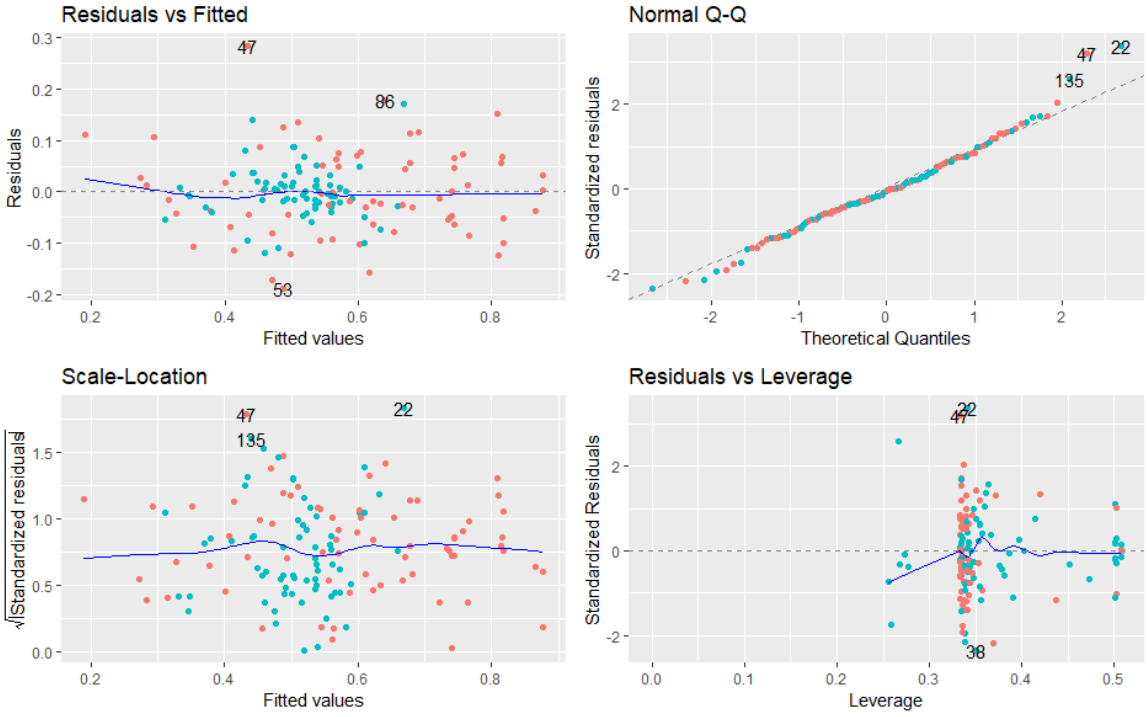
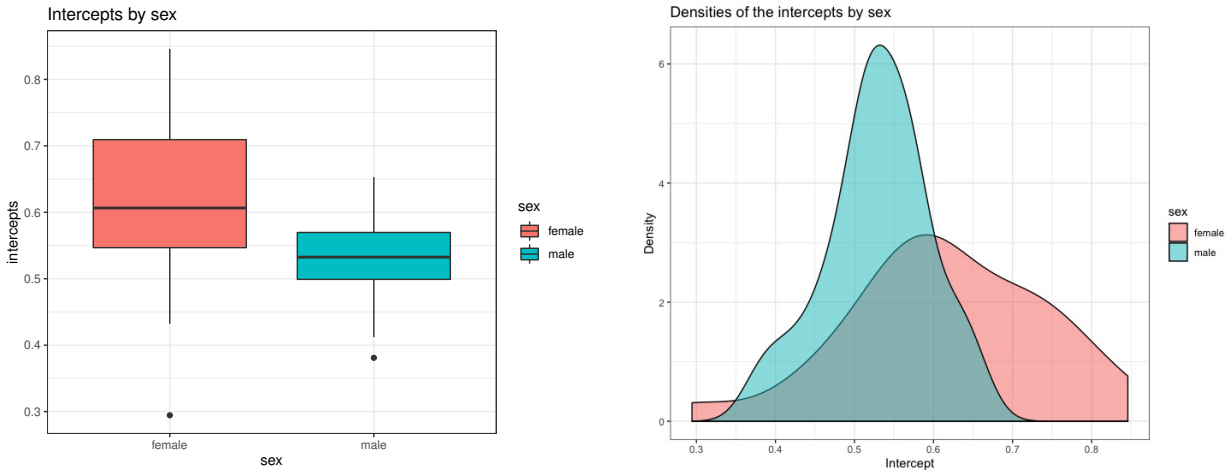


Figure 12: *Diagnostic plots for the weighted* `subjID` *model in Equation 9*

The residuals look good, hence it is assumed that the assumptions are fulfilled. Figure Figure 13 shows boxplots of the intercepts given sex. It is clear that the variance of the intercepts for women is much greater than for men, which corresponds well with the earlier findings, but now we see that we have a couple of outliers, i.e. one man and one woman who wear significantly less than the others. According to this model, the amount of clothing varies greatly from person to person, more for females than for males. The only parameter apart from the individual that is significant is the outdoor temperature.



(a) *Boxplots of the intercepts in the two groups*

(b) *Densities of the intercepts in the two groups*

Figure 13: *It is clear from the plots that the females in general wear more clothes and that they have a greater variance.*

This model will be good at estimating the amount of clothes given the outdoor temperature of the individuals that it already knows, but if a new subject comes along we will need some kind of estimate of the intercept of that individual. We could use the mean of the males or females that we already know, but this estimate would be uncertain, especially for women. What we do have, is an estimate of how much the clothing will change as a function of temperature. Figure Figure 14 shows the predictions of clothing as a function of the outdoor temperature using estimated intercepts as the mean of the males and females respectively.

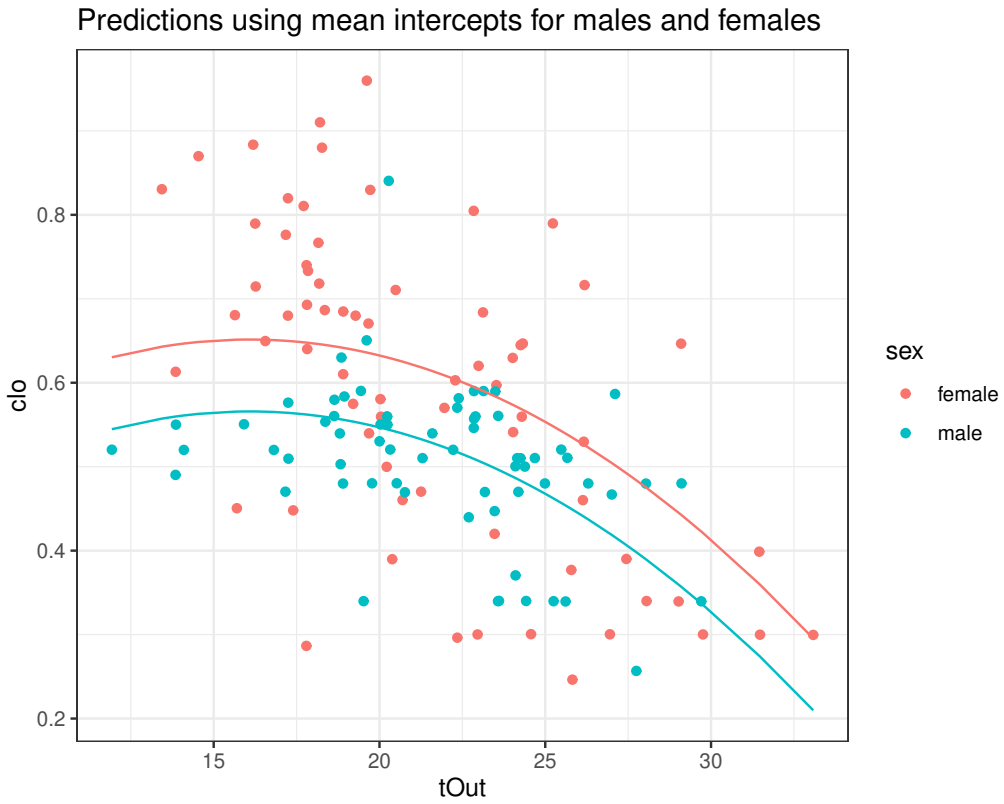


Figure 14: *The predictions of level of clothing given the outdoor temperature using mean intercepts of females and males respectively. Since it is based on the means, there is no confidence- or prediction interval*

This model looks very similar to the one made in part A. Again it is clear that women in general wear more clothes. We will look at the residuals to check of the assumptions hold. It is shown in figure Figure 15.

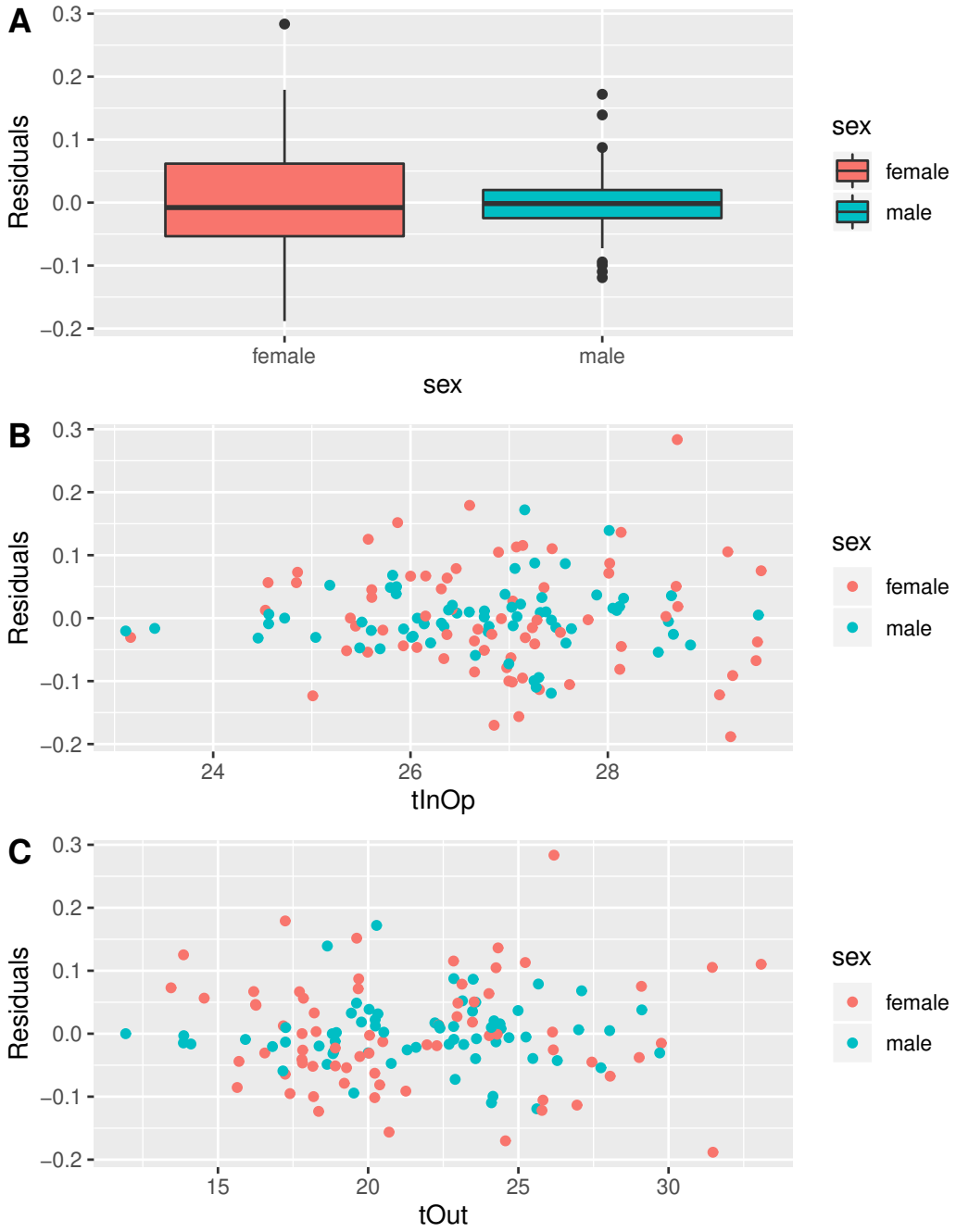


Figure 15: *The residuals against explanatory variables for the model with subject ID.*

It does not look like there is any structure in the data that we did not capture. A more appropriate way to include the `subjId` in the model, would be to include it as a random effect. In that way, the differences between individuals would be part of the variance estimate, and the model would be valid for new individuals as well.

The full data-set

The full data-set is now considered. The data-set differs from the previous data because it contains several measures from the same individual on the same days. Figure 16 shows how the numerical variables in the full data-set against each other, and figure 17 shows boxplots of the numerical variables against `sex`, `day`

and `subjId`.At last figure 18 shows boxplots of the numerical variables against `obs.no` - that is the number of ovservation during the day. It should be noted that both temperature measurements increase with the observation number. The clothing descreases a little for females with `obs.no`, but not for males. This is also illustrated in figure 16, where the first two figure shows "horizontal" lines, which indicates that the clothing level does not change.

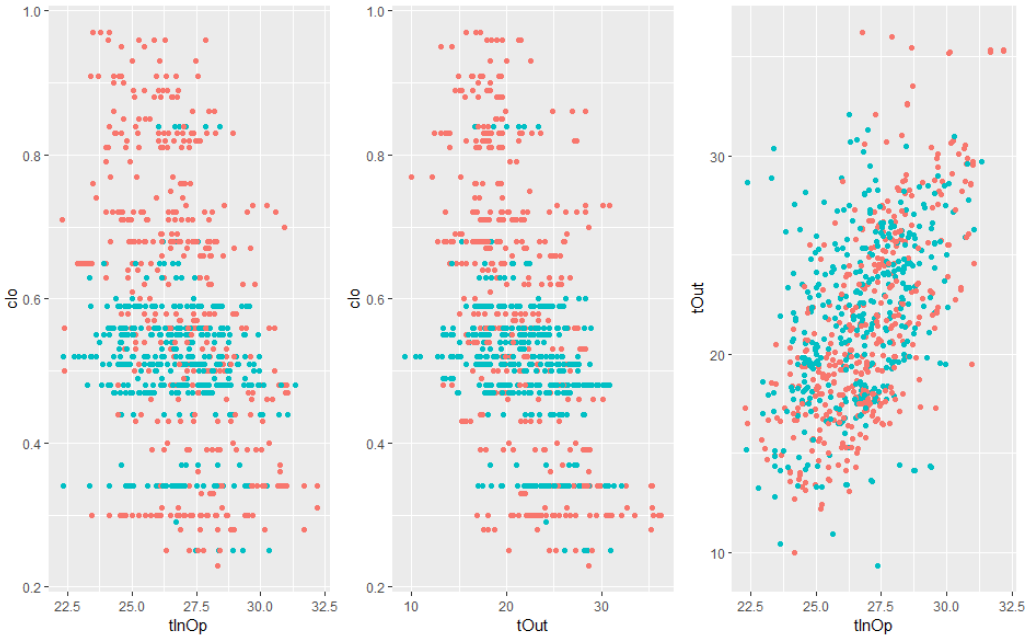


Figure 16: *Plots of the observations for the full model*

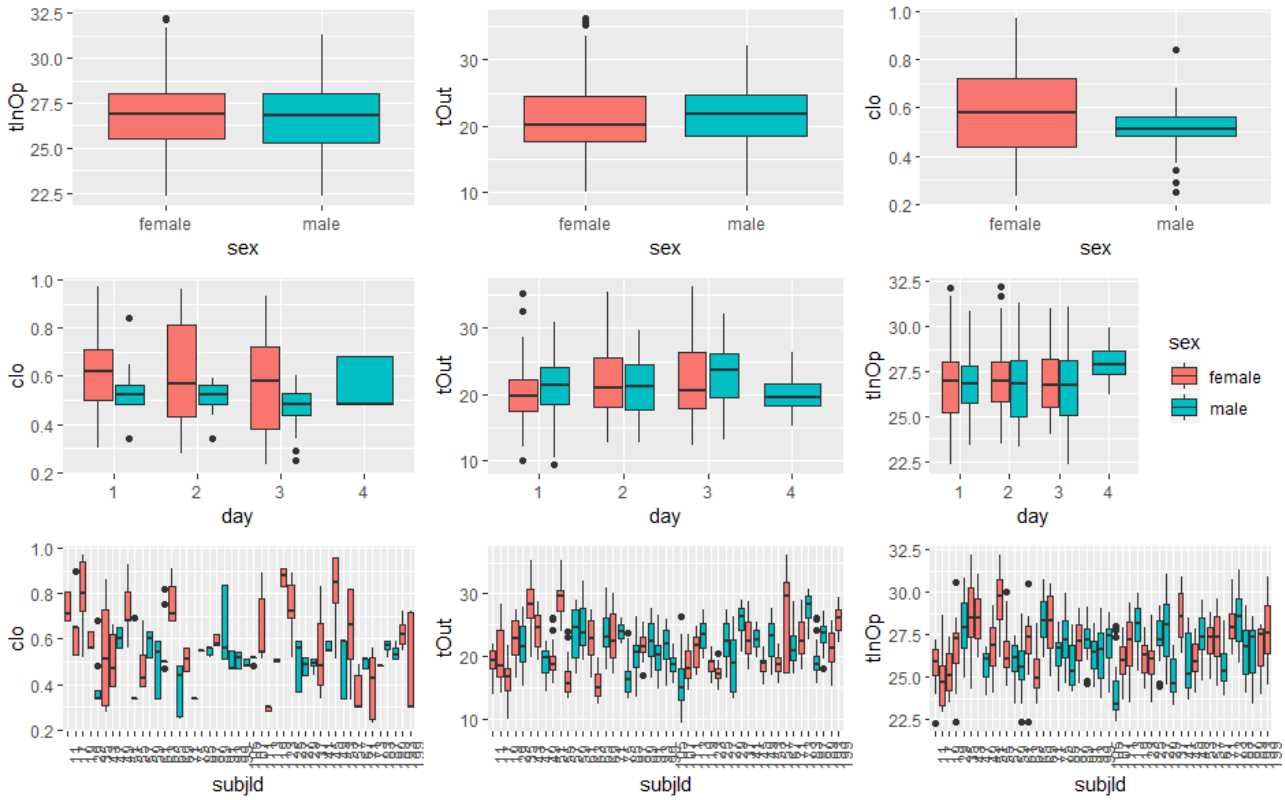


Figure 17: *Plots of the observations for the full model*

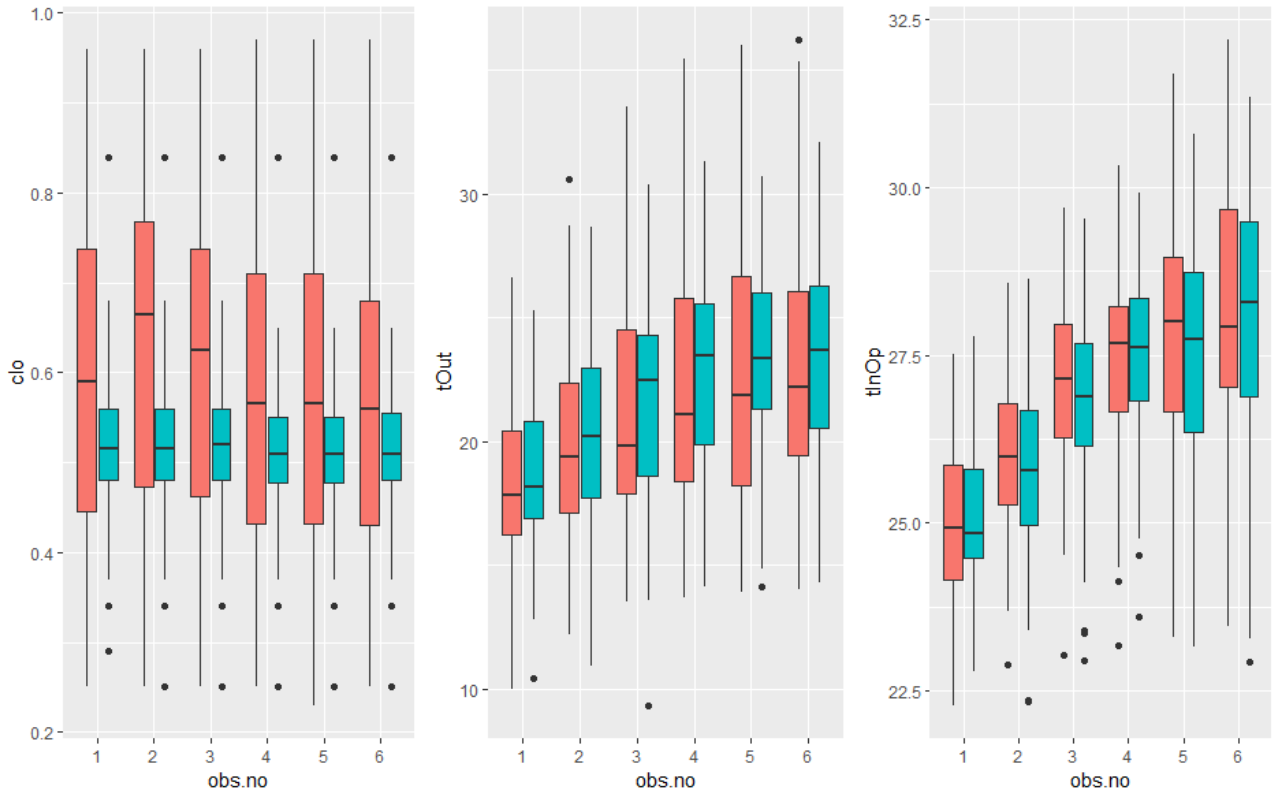


Figure 18: *Plots of the observations for the full model*

The model with the weighted analysis and the model which includes `subjID` are now tried on the full-data set. That is we fit models of the structures given in (7) and (9).

For the weighted analysis the estimates of the terms in the fitted model are given in table 9

Variable	Estimate	Standard Error
$\beta_0(\texttt{sex} = female)$	0.6069607	0.0089718
$\beta_0(\texttt{sex} = male)$	0.5187665	0.0051527
$\beta_1(\texttt{sex} = female)$	-0.0256982	0.0046839
$\beta_1(\texttt{sex} = male)$	0.0003597	0.0026442
β_2	-0.0089647	0.0010126
β_3	-0.0005675	0.0001537

Table 9: *Parameters for the final weighted linear model along with their standard-error*

Figure 19 shows a diagnostics plot for this model. There seem to be some weird tendencies in the plot. The residuals follow a decreasing negative trend when plotted against the fitted values (upper left corner), and in the QQ-plot the female observations produces heavy tails. The standardized residuals in the lower left corner also presents a very uneven distribution for males and females, with the male residuals being much smaller.

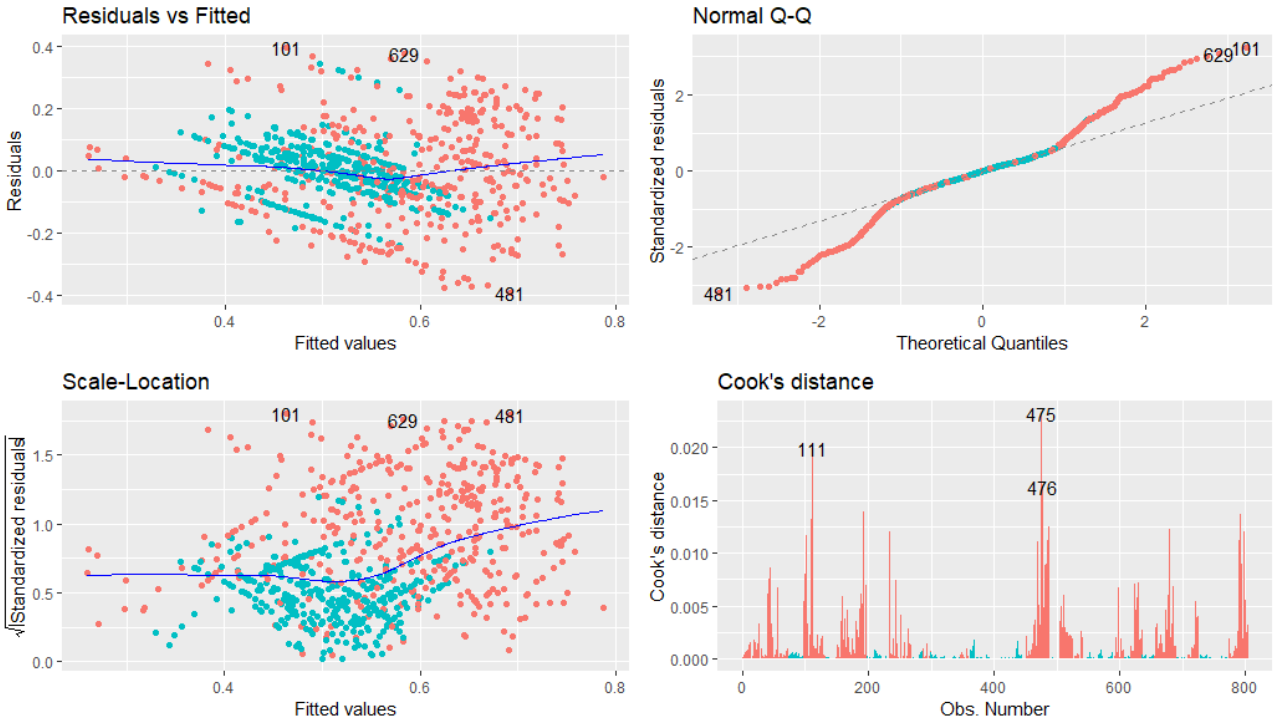


Figure 19: *Diagnostic plots for the weighted* `sex` *model in Equation 7 on the full data set*

For the analysis including the subject ID into the model the estimates of β_1 and β_2 are given in Table 10, all the estimated intercepts can be seen in appendix in Table 12.

Variable	Estimate	Standard Error
β_1	-0.0103	0.0012
β_2	-0.0002	0.0002

Table 10: *Slope and curve parameters for the model including subject ID.*

A diagnostics plot is seen in figure 20. The same weird tendencies as in figure 19 are seen - decreasing linear tendencies when the residuals are plotted against the fitted values, and much lower-valued male residuals.

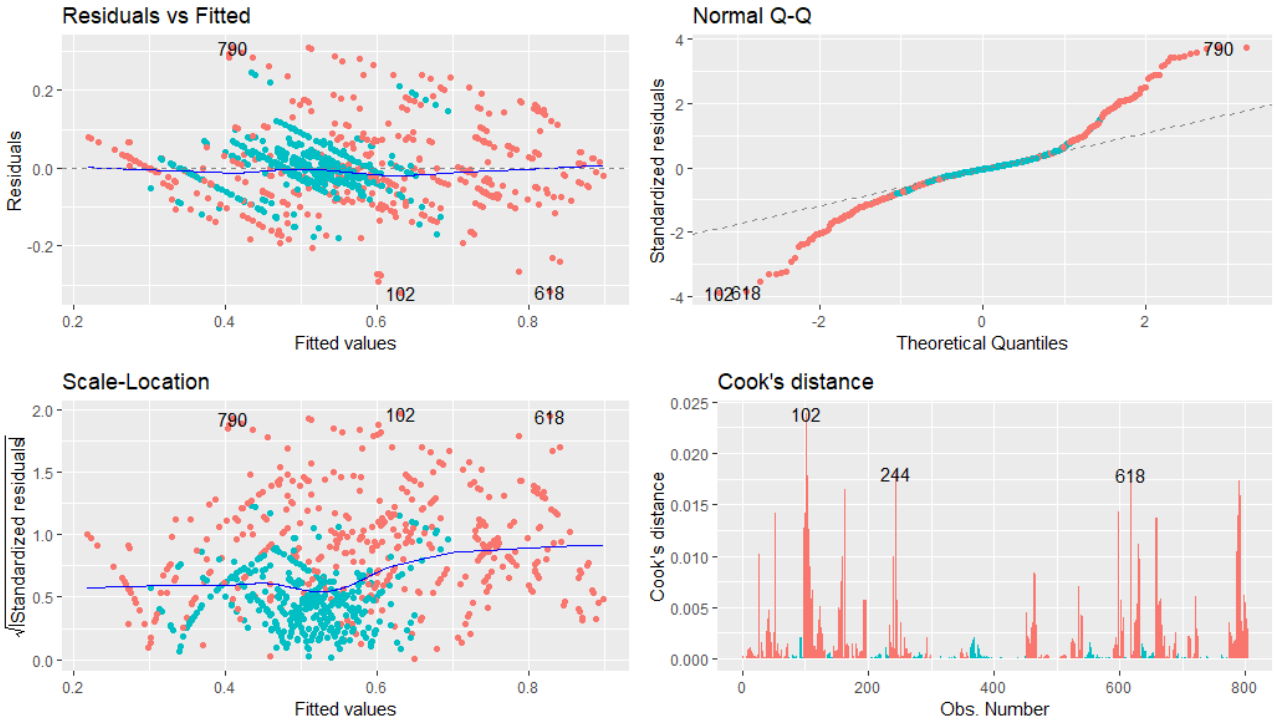


Figure 20: *Diagnostic plots for the weighted* `subjID` *model in Equation 9 on the full data set*

Both figure 19 and 20 showed problems with the residuals. Some of the weird tendencies might be due to a violation of the independent observations assumption. As figure 18 the temperature changes during the day, but the level of clothes does not (due to people not changing clothes during the day). That is the response variable depends much on other observations, which is very problematic.

Conclusion

After having done multiple statistical analyses, we see that the amount of clothing people wear varies significantly both from person to person, but also with the temperature. It is also clear from multiple models that the variance of the amount of clothing is greater for women than for men, and that women generally wear more than men. This led us to making the weighted analysis, optimizing a weight for women relative to men. We did three models:

- **Initial model**
  - We saw that sex plays a role on the relation to the indoor temperature
  - The outdoor temperature also has a relation to clothing
  - The residual analysis was problematic
- **Weighted model**

  - Now also the squared outdoor temperature was significant
  - The residuals looked better

- **Model including subject Id**
  - Clothing varied greatly from person to person
  - The squared outdoor temperature was significant
  - The residuals looked good
  - Poor prediction given an unknown subject

The structures of the weighted model and the model including subject Id was used to estimate a model on the full data set. The residuals for these models were in both cases problematic due to dependence between observation from the same day (people are not expected to change clothing throughout a day), and the residuals in general looked questionable.

Appendix

B) All intercepts by subjects

Subject ID	Estimate	Std. Error	Subject ID	Estimate	Std. Error
subjId 11	0.7099	0.063	subjId 105	0.4652	0.0371
subjId 17	0.6441	0.0634	subjId 107	0.5043	0.0392
subjId 19	0.7796	0.0654	subjId 111	0.6331	0.0634
subjId 29	0.6072	0.0628	subjId 113	0.2945	0.0627
subjId 35	0.4121	0.0316	subjId 119	0.5325	0.0366
subjId 43	0.7433	0.0701	subjId 123	0.8458	0.0632
subjId 47	0.5574	0.0634	subjId 125	0.7068	0.0641
subjId 49	0.5799	0.0449	subjId 127	0.5223	0.0364
subjId 51	0.7092	0.063	subjId 129	0.474	0.038
subjId 55	0.5473	0.0724	subjId 137	0.5707	0.0467
subjId 57	0.4573	0.0654	subjId 141	0.5777	0.0629
subjId 59	0.6315	0.0376	subjId 145	0.5854	0.0445
subjId 61	0.5624	0.0387	subjId 149	0.787	0.0632
subjId 63	0.5859	0.0628	subjId 153	0.5344	0.0366
subjId 65	0.7288	0.067	subjId 157	0.6058	0.0631
subjId 69	0.4416	0.0373	subjId 167	0.5088	0.072
subjId 71	0.5452	0.077	subjId 171	0.5075	0.0365
subjId 75	0.381	0.045	subjId 173	0.4322	0.0629
subjId 85	0.5223	0.0407	subjId 183	0.6145	0.0438
subjId 87	0.5426	0.0364	subjId 187	0.5548	0.0368
subjId 89	0.5934	0.0627	subjId 189	0.5686	0.0447
subjId 91	0.653	0.0364	subjId 193	0.629	0.0768
subjId 93	0.4938	0.0366	subjId 199	0.5222	0.0643
subjId 99	0.5125	0.0363			

Table 11: *The intercepts for all the subject IDs*

C) Full dataset intercepts by subject ID.

Subject ID	Estimate	Std. Error	Subject ID	Estimate	Std. Error
subjId 11	0.7112	0.0205	subjId 105	0.4651	0.0387
subjId 17	0.6294	0.0206	subjId 107	0.4686	0.0426
subjId 19	0.7685	0.0223	subjId 111	0.6284	0.0207
subjId 29	0.5996	0.0203	subjId 113	0.2907	0.0202
subjId 35	0.4049	0.0334	subjId 119	0.5229	0.0385
subjId 43	0.6575	0.0237	subjId 123	0.8469	0.0206
subjId 47	0.535	0.0205	subjId 125	0.7022	0.0213
subjId 49	0.5832	0.0473	subjId 127	0.5182	0.0385
subjId 51	0.707	0.0205	subjId 129	0.4628	0.0389
subjId 55	0.4589	0.0237	subjId 137	0.5381	0.0474
subjId 57	0.4422	0.0221	subjId 141	0.5658	0.0203
subjId 59	0.6094	0.0397	subjId 145	0.5805	0.0493
subjId 61	0.5297	0.0388	subjId 149	0.7878	0.0206
subjId 63	0.5784	0.0203	subjId 153	0.5229	0.0385
subjId 65	0.7094	0.0231	subjId 157	0.6057	0.0205
subjId 69	0.4206	0.0386	subjId 167	0.4213	0.0233
subjId 71	0.5304	0.0249	subjId 171	0.4992	0.0385
subjId 75	0.3676	0.0493	subjId 173	0.4202	0.0203
subjId 85	0.5079	0.0395	subjId 183	0.5548	0.0394
subjId 87	0.5386	0.0397	subjId 187	0.5527	0.0386
subjId 89	0.5922	0.0202	subjId 189	0.5602	0.0472
subjId 91	0.6487	0.0385	subjId 193	0.6279	0.0248
subjId 93	0.4887	0.0386	subjId 199	0.4892	0.0209
subjId 99	0.5111	0.0385			

Table 12: *The intercepts for all the subject IDs using the full data set.*