02424 Assignment 1

August Thomas Hjortshøj Schreyer s163716 Mikkel Mathiasen s174344
Nicolai Piet Dittmann s170589

12/03/2021

Contents

Exploratory analysis of the data

This report aim to analyze the clothing level in an office based on different explanatory variables. The data set that are used in the process of analyzing the clothing level consist of 6 variables described in Table 1. For the first part of the modeling, the identifier for subject and day variables will be omitted.

The data set consists of 136 observations with 66 male and 70 female observations. Firstly, the summary statistics are computed for the continuous variables, seen in Table 2

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: Variable description of the data set.

	$\hat{\mu}$	$\hat{\sigma^2}$	Min	Max
`clo`	0.55	0.02	0.25	0.96
`tOut`	21.54	17.27	11.93	33.08
`tInOp`	26.82	1.69	23.11	29.55

Table 2: Summary statistics of the continouos variables.

On Figure 1 below, the variables are plotted against each other separately for male and females. On the first upper plot, it should be noted how the clothing level of females seems to vary to a greater extent compared to the male counterpart which seem more constant. Additionally, it appears that the level of female clothing decreases as the indoor operating and outdoor temperature increases. On the middle plot, it appears that the outdoor temperature and indoor operating temperature are positively correlated which intuitively makes sense. For the lower plot, the indoor operating temperature is plotted against the clothing level. The same applies as for upper plot; the clothing seems more constant for males compared to females.

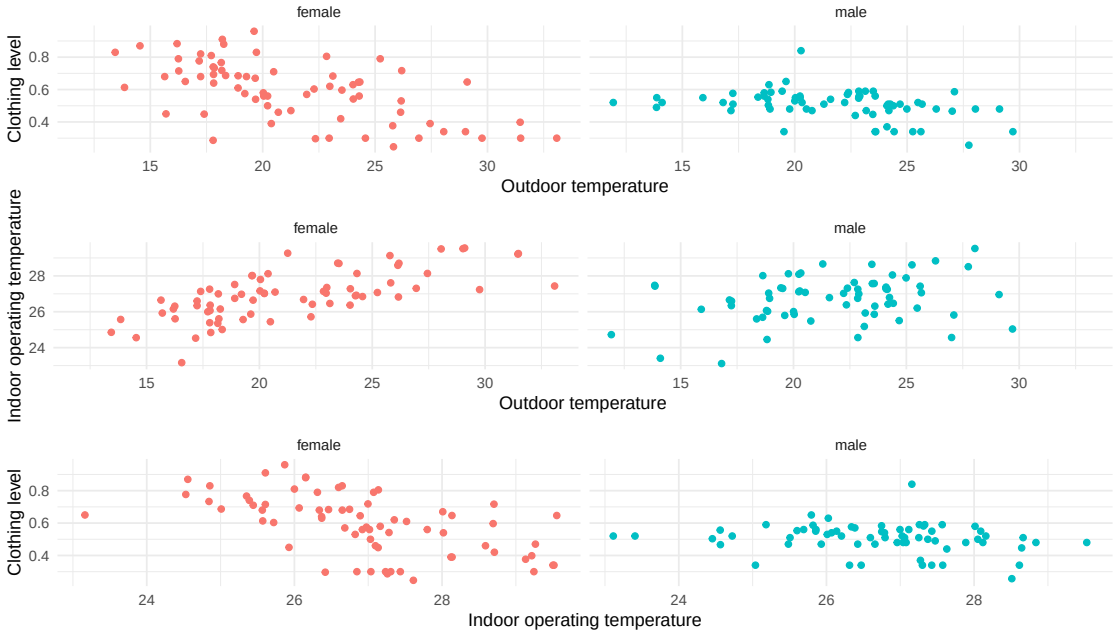


Figure 1: Scatter plot of the continuous variables plotted separately for sex.

On Figure 2 below, histograms have been computed for the continuous variables for male and females separately. The continuous variables appear to be centered and roughly symmetric with no apparent skewness.

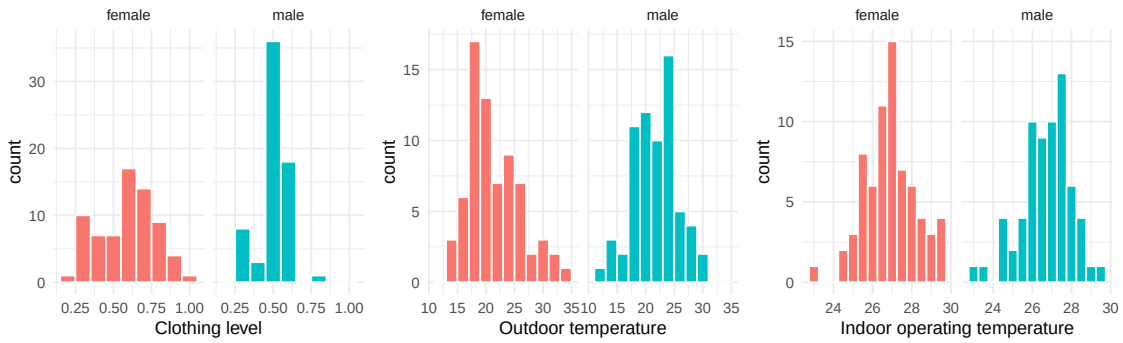


Figure 2: Histograms of the continuous variables plotted separately for sex.

It is assumed for the modelling, that all the data points are independent.

Problem A

Model Selection

A general linear model (GLM) predicting the level of clothing (`clo`) using outdoor temperature (`tOut`), indoor operating temperature (`tInOp`) and gender (`sex`) of the subject as explanatory variables will be fitted. The full GLM model using all variables and interactions are used as a starting point, hereafter the model will be reduced using the backward selection based on Type III partitioning and higher order variables are removed first if these are insignificant. If a higher order variables are significant, but the corresponding first order of the variable is non-significant, we will keep both the first order and higher order terms.

In table 3 the different models fitted through the backward selection are written in R notation (first column) and the statistics calculated for each model in the following columns.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
`clo ~ tOut * tInOp * sex`	128	1.82				
`clo ~ tOut + tInOp + sex +tOut:sex + tInOp:sex + tOut:tInOp`	129	1.85	-1	-0.02	1.71	0.1927
`clo ~ tOut + tInOp + sex + tOut:sex + tInOp:sex`	130	1.85	-2	-0.02	0.86	0.4259
`clo ~ tOut + tInOp + sex + tInOp:sex`	131	1.88	-1	-0.03	2.03	0.1569

Table 3: Analysis of deviance table from backward selection of the optimal model.

Estimated parameters

The final model from the backward selection using Type III partitioning are shown below where the third-order interaction has been removed and as well the second order interaction between the outdoor temperature & gender as well as indoor operating temperature.

$$Clo_i = \beta_0 + \beta_1 t_{Out,i} + \beta_2 t_{InOp,i} + \beta_3 sex_i + \beta_4 t_{InOp,i} \cdot sex_i + \epsilon_i$$

where i denote the observation number, t_{Out} is the outdoor temperature, t_{InOp} is the indoor temperature of the office and sex_i is either 0 or 1 depending on the sex; $sex(female) = 0$ and $sex(male) = 1$. When our categorical variable, `sex`, has been transformed to ones and zeros, we can easily interpret the rest of the parameters.

$\hat{\beta}_0+\hat{\beta}_3 \cdot sex_i$ is the intercept of the model. The intercept for a female individual is $\hat{\beta}_0$ because $sex(female) = 0$. The combination of $\hat{\beta}_0 + \hat{\beta}_3$ is the intercept for a male individual as $sex(male) = 1$.

$\hat{\beta}_1$ determines the impact of the outdoor temperature on the clothing level regardless of sex.

$(\hat{\beta}_2 + \hat{\beta}_4)t_{InOp}$ models the change in clothing level as the indoor temperature changes. And as for each sex grouping for the intercept, $\hat{\beta}_2$ denotes the slope for the female group while $\hat{\beta}_2 + \hat{\beta}_4$ denotes the slope for the male group.

Finally, the noise term is distributed as $\epsilon_i \backsim N(0, \sigma^2)$, meaning that the noise is independent and identically distributed. The standard deviation of the residuals ($\hat{\sigma}$) is estimated as 0.120, and an estimate of the standard deviation and confidence hereof are computed based on theorem 3.5 resulting in the following 95% confidence interval: [0.107, 0.136] (can also be seen in table 4).

It is wanted to test whether the model reduction is appropriate and in order to test this a likelihood-ratio test are performed based on theorem 3.6. The likelihood test are performed by comparing the initial model using all interactions (first row in table 3) and the reduced model (last row in table 3) and by comparing the deviances of the two models, the resulting p-value computed is: $p = 0.29$ and the model reduction is therefore kept.

	2.5%	Estimate	97.5%
$\hat{\beta}_0$	1.51	2.13	2.76
$\hat{\beta}_1$	-0.02	-0.01	-0.01
$\hat{\beta}_2$	-0.07	-0.05	-0.02
$\hat{\beta}_3$	-2.16	-1.28	-0.40
$\hat{\beta}_4$	0.01	0.04	0.08
$\hat{\sigma}$	0.107	0.120	0.136

Table 4: Parameter estimates with 95% confidence intervals and an estimate of σ, the standard deviation of the uncertainty of the model.

The estimated parameters (and the estimate of the standard deviation of the noise) and 95% confidence intervals can be seen in table 4. As the exploratory analysis showed, there is a negative correlation (The estimated slopes are negative) between either of the temperatures and the clothing level.

Model predictions

Now that we have found our model, we have to test its ability to make predictions.

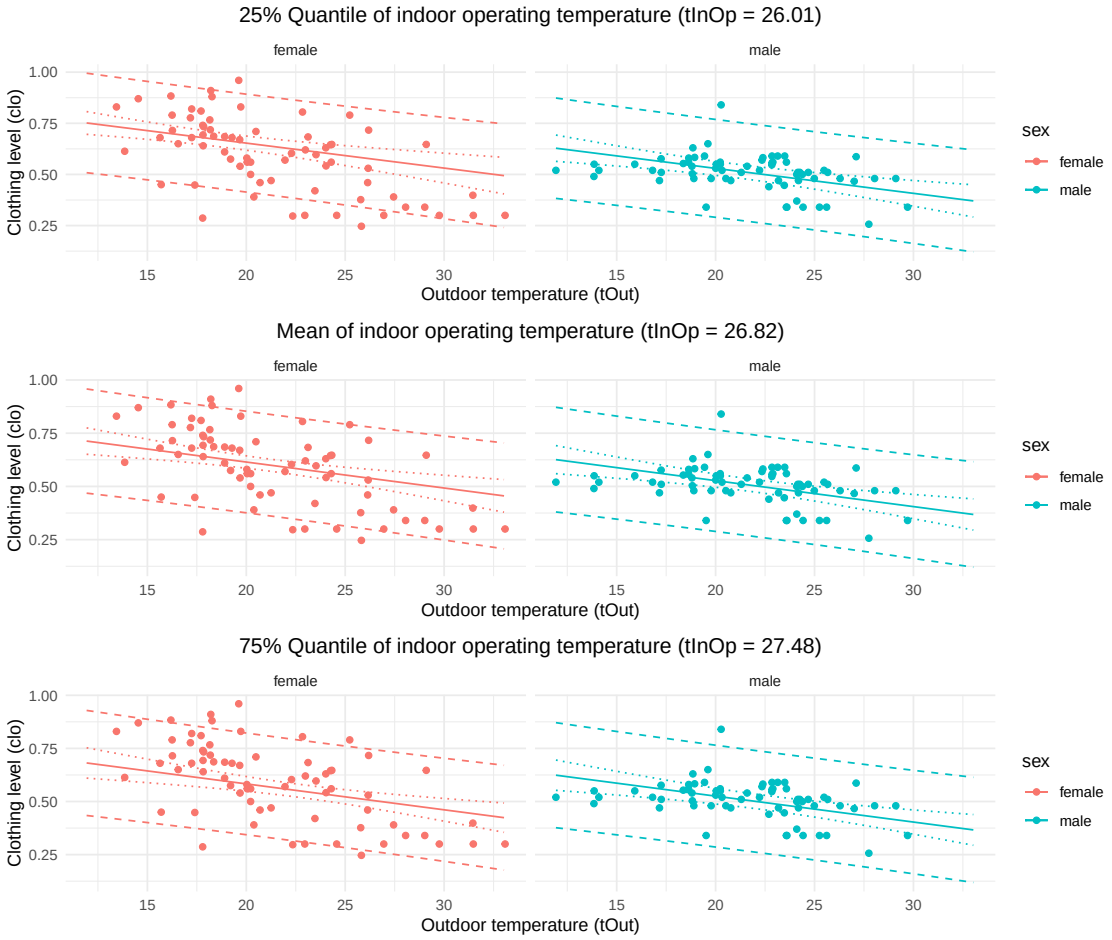


Figure 3: The figure presents the 95% prediction and confidence interval for three fixed values of the indoor operating temperature (tInOp)

In figure 3 95% confidence and prediction intervals are presented for three indoor operating temperature levels being 25% quantile, mean and 75% quantile. From the figure it can be seen that almost all samples for men except one lies within the prediction interval, whereas for females a few samples are out of the prediction interval. As the prediction interval only is a 95% prediction interval it is expected that some points lies without of the interval. From the figure it should also be noted that males appears to be more constant, whereas females appear to contain larger variance. Also from the fitted line this appears to fit the male samples better than the female samples. Also it should be noted that when elevating the indoor operating temperature this results in the prediction and prediction interval for females are shifted to lower clothing levels whereas this appears to be constant for men. This also makes sense when considering the slope $t_{InOp,i}\hat{\beta}_2$ for the indoor operating temperatures effect on clothing level for females which is negative resulting in lower clothing level when the temperature rises. Whereas the effect of elevated indoor operating temperature for males are dependent on the following expression: $t_{InOp,i} \cdot (\hat{\beta}_2 + \hat{\beta}_4 \cdot sex_i)$ where $\hat{\beta}_4$ and $\hat{\beta}_2$ almost cancel out resulting in almost no change in clothing level when the indoor operating temperature rises.

Residual analysis

To ensure that the model lives up to the assumptions that the this type of model requires, we have to analyze the residuals.

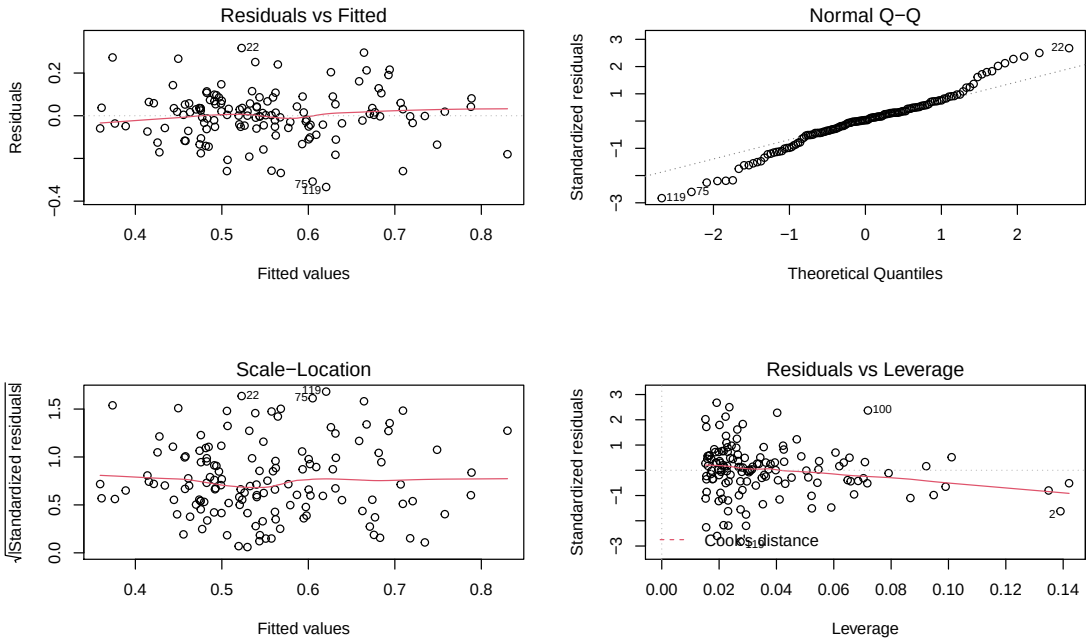


Figure 4: The figure presents the residuals of the model selected from the backward selection with Type III partionioning.

From figure 4 in the Normal-QQ plot it can be seen that the model has tails indicating the model and assumptions of the residuals does not fit the data well. It should be noted that some of the points resulting in the tails (22, 119 and 75) does not have a high leverage (looking at the Residuals vs Leverage plot) meaning that these points does not affect the model a lot compared to eg. point 2 which has a high leverage. This also indicates that outlier investigation of these three points are not necessary. From the Residuals vs Fitted and Scale-Location plots the residuals appears to be randomly distributed.

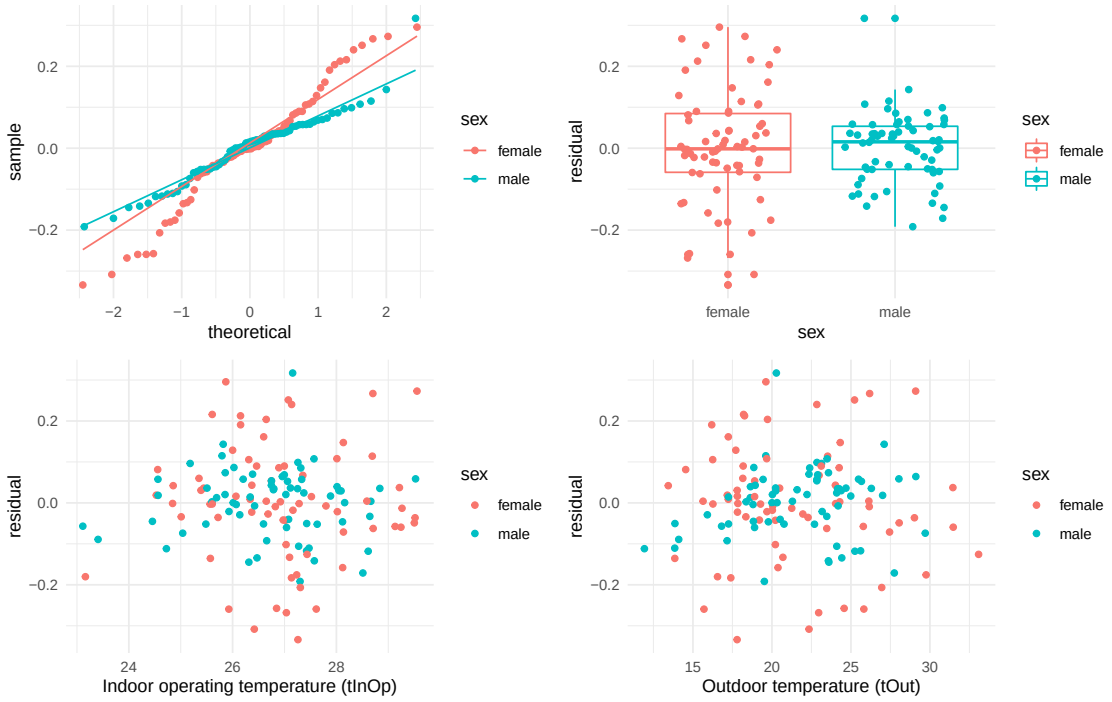


Figure 5: The figure presents the residuals of the model selected from the backward selection with Type III partionioning as a function of the explanatory variables.

From figure 5 it can be seen from the boxplot that the variance of the residuals differ between males and females. When looking at the residuals as a function of indoor operating temperature and outdoor temperature the residuals appears to be randomly distributed and this appears to be the case for both males and females. From the QQ-plot it should be noted that the variance of females and males also appear to be different and this could also give rise to the tails seen in figure 4. This could indicate that the assumption that all residuals should be identical is not correct, and instead this should be switched with an assumption of the residuals being identical within each gender group. This would result in a weighted residual model, where an optimal weight of the residuals for females compared to males needs to be estimated. In order to investigate the effect of having this weighted residual model the residuals of females are divided by the fraction 2.83 (estimated above) and a new QQ-plot using these residuals are presented in figure 6. From figure 6 it can be seen that a weighting residual model appears to be beneficial.

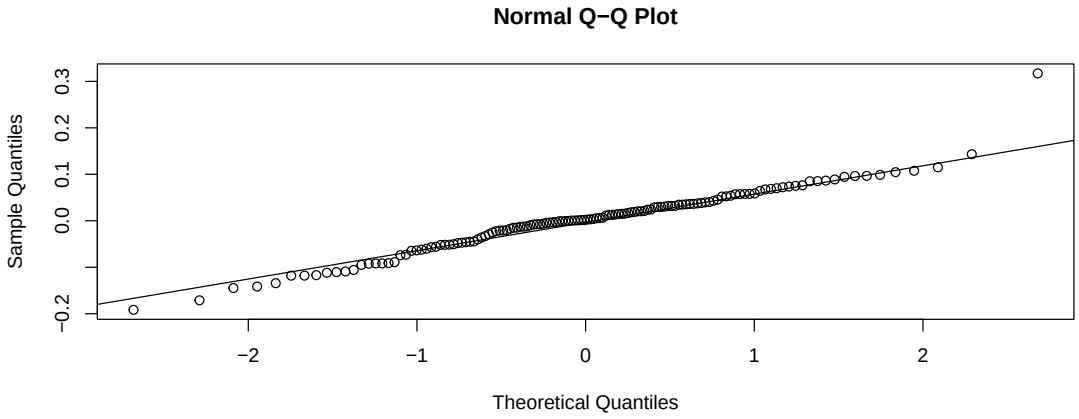


Figure 6: The figure presents a QQ-plot of the effect of normalizing the residuals across females/males

Model with weights on sex

Firstly, the optimal ratio needs to be estimated, this will be done by optimizing the log-likelihood with respect to the weighting of the variance. The optimal weighting of the female residual variance is estimated to: 2.93 which will be used in the weighted residual model. A confidence interval of the weighing of the residuals for females (called: $\hat{w}$) are also computed an can be seen in table 5.

In order to compare the weighted model with the unweighted model information criteria AIC and BIC are used as a measure. The likelihood are not comparable as the models are not nested. The AIC are therefore chosen as criterion and are computed as: -184.59 and -202.64 for the unweighted and weighted model respectively. The BIC are therefore chosen as criterion and are computed as: -167.1172 and -185.16 for the unweighted and weighted model respectively. Both information criteria appears to prefer the weighted residual model. The model parameters and confidence intervals are presented in table 5. It should be noted that $\hat{\beta}_3$ has become smaller (-1.365) compared to the previous model with -1.2834 whereas the other parameters appears to be similar. The residual standard deviation $\hat{\sigma}$ is estimated to 0.085, and an estimate of the standard deviation and confidence hereof are computed based on theorem 3.5 resulting in the following 95% confidence interval: [0.076, 0.097] (can also be seen in table 5). The residual variance is smaller than for the previous model. Afterwards a full model using all interactions and explanatory variables (indoor operating temperature, outdoor temperature and sex) was reduced using the same approach in the beginning (backward selection using type III partitioning where higher order terms are removed before lower order terms). This resulted in the same model as before and therefore this model is used for the proceeding analysis.

	2.5%	Estimate	97.5 %
$\hat{\beta}_0$	1.49	2.22	2.95
$\hat{\beta}_1$	-0.02	-0.01	-0.01
$\hat{\beta}_2$	-0.08	-0.05	-0.02
$\hat{\beta}_3$	-2.22	-1.37	-0.51
$\hat{\beta}_4$	0.02	0.05	0.08
$\hat{\sigma}$	0.076	0.085	0.097
$\hat{w}$	1.51	2.93	4.34

Table 5: Parameter estimate with 95% confidence intervals and an estimate of σ, the standard deviation of the uncertainty of the model.

Weighted model predictions

Once again, we have to test our models ability to make predictions.

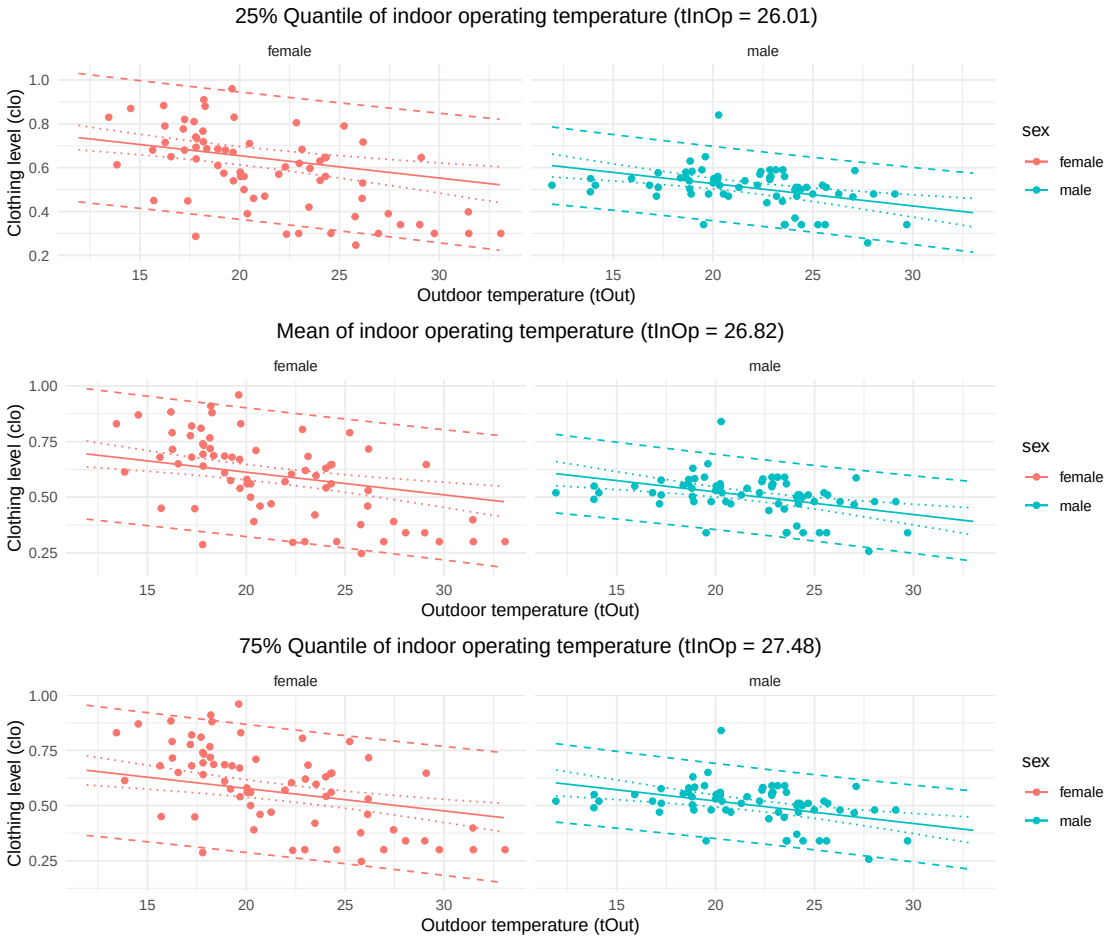


Figure 7: The figure presents the 95% prediction and confidence interval for three fixed values of the indoor operating temperature (tInOp)

From figure 7 it can be seen that the prediction interval for the females are more wide compared to the unweighted model (see figure 3) resulting in more points are within the prediction interval. Clearly, it can be seen that the prediction interval for females are more wide compared to males, which is due to the weighting of the residuals. Also the prediction interval for males appears to have become more narrow compared to the unweighted residual model (see figure 3). As in figure 3 it can be seen that the prediction and prediction interval are shifted down when the indoor operating temperature are elevated, whereas it appears to be rather constant for males. From the figure it is apparent that the slope for females has been changed to fit better to the male samples, driving the slope away from the optimal slope for females. This change in slope might be due to the less weight put on the female residuals which is weighted with 2.93 compared to males. This could indicate that a gender specific slope might be needed in order to model this appropriately.

Weighted model residuals

Once again, we have to ensure that the model lives up to the assumptions that the this type of model requires, we have to analyze the residuals.

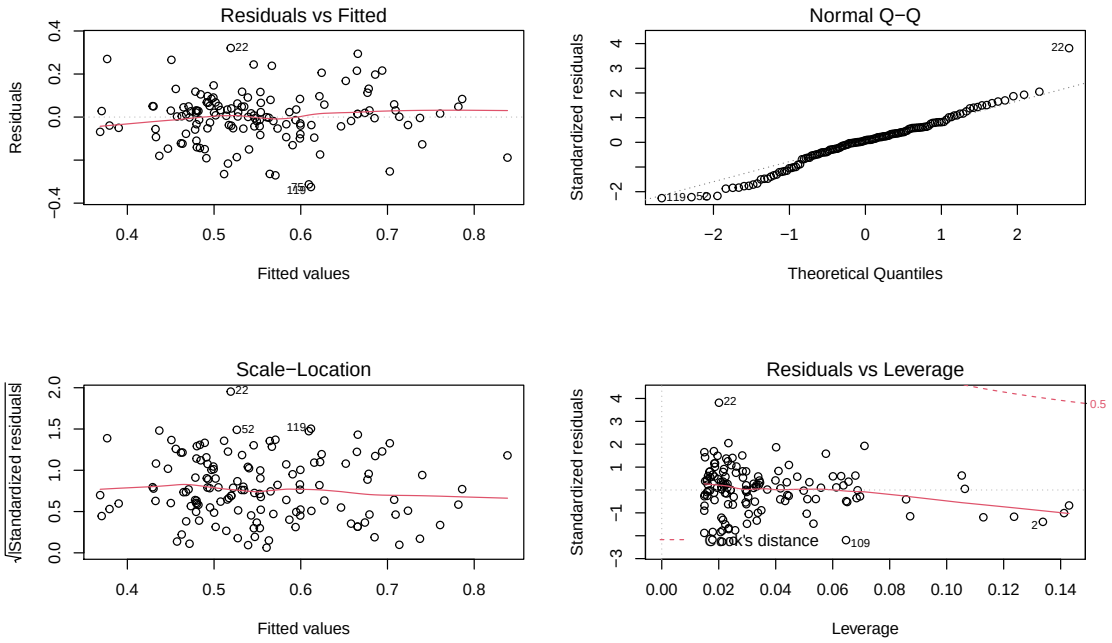


Figure 8: The figure presents the residuals of the weighted model

From figure 8 it can be seen that the Normal-QQ plot clearly has improved compared to the one in figure 4. All points except point 22 appear to follow the QQ-line nicely. From the Residuals vs Leverage plot it can be seen that point 22 has a low leverage and therefore does not affect the model fit substantially and further investigation of this point are omitted. From the Residuals vs Fitted and Scale-location plots the residuals appears to be randomly distributed. Clearly the weighted residual model improved the model fit. As the residuals as a function of indoor operating temperature and outdoor temperature for the unweighted residual model appeared to look randomly distributed these will not be further investigated for this model. Also as the residual variance between gender has been taken into account and the residuals in figure 8 these will neither be further investigated. Figure 9 shows the residuals when looking at the subject ID, and from this it can clearly be seen that the subject ID contains information about the variance in the residuals. The subject ID should therefore be included in the modeling.

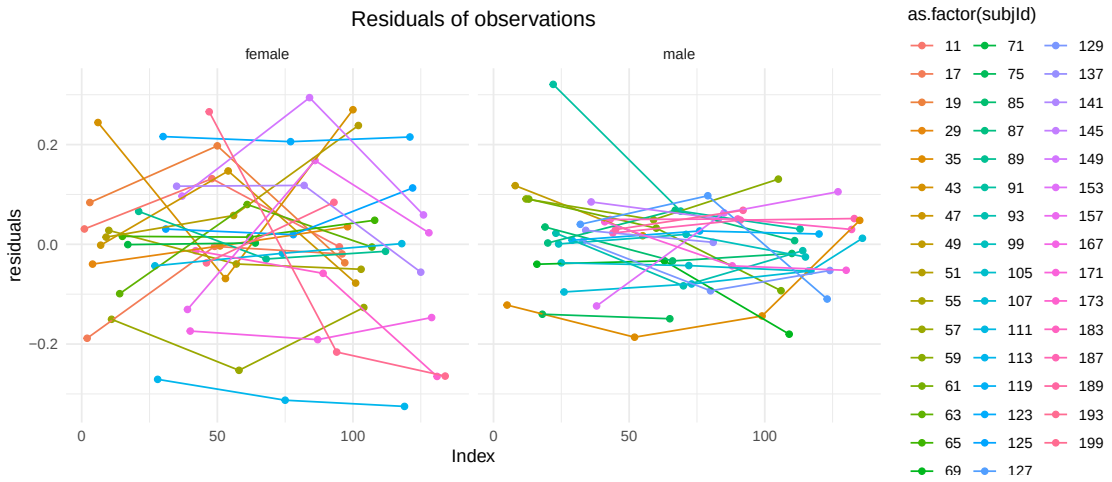


Figure 9: The figure presents the residuals as a function of the subject ID.

Problem B

GLM with subject ID

As we saw from the residuals in Problem A, some of the individuals do produce residuals that are either all positive or all negative. This could indicate that the individuals could have different base preferences of clothing. That is, one individual might be warm most of the time and therefore wear less clothing. Likewise an individual might often be cold and would therefore wear more clothes. This could indicate that our model should have an individual intercept.

Furthermore, dividing the data into individuals rather than just sex, is a more precise grouping. There is also no individual who is both male and female, so the individuals are nested within the sex grouping. Therefore we can discard the groups depending on sex.

We cannot estimate any models with both an interaction between both of the indoor/outdoor temperatures and an individual intercept because of a lack of degrees of freedom. Third order interactions are therefore also not possible. We can however estimate a model with an interaction between either the indoor or outdoor temperature and the individuals. This will lead to either of the following models:

The first model:

$$\begin{aligned} Clo_i =& \beta_0 + \beta_1 \cdot t_{InOp,i} + \beta_2 \cdot t_{Out,i} + \beta_3 \cdot t_{InOp,i} \cdot t_{Out,i} + \\ & a(individual_i) + b(t_{InOp,i}, individual_i) \end{aligned} \tag{1}$$

The second model:

$$\begin{aligned} Clo_i =& \beta_0 + \beta_1 \cdot t_{InOp,i} + \beta_2 \cdot t_{Out,i} + \beta_3 \cdot t_{InOp,i} \cdot t_{Out,i} + \\ & a(individual_i) + b(t_{Out,i}, individual_i) \end{aligned} \tag{2}$$

where β_j is parameters for the overall intercept and continuous variables, while a and b are functions of either just a categorical variable or a combination of categorical and continuous variables.

The following table shows the initial models with either an interaction between the indoor/ temperature and the individuals:

Model		Sum Sq	Df	F value	Pr(>F)
Model 2, eq (1)	(Intercept)	0.022	1	3.386	0.073
	tOut	0.023	1	3.446	0.071
	tInOp	0.017	1	2.531	0.120
	subjId	0.480	46	1.571	0.074
	tOut:tInOp	0.023	1	3.417	0.072
	tOut:subjId	0.442	46	1.449	0.117
	Residuals	0.266	40		
Model 2, eq (2)	(Intercept)	0.002	1	0.270	0.606
	tOut	0.015	1	1.695	0.200
	tInOp	0.004	1	0.417	0.522
	subjId	0.370	46	0.925	0.603
	tOut:tInOp	0.019	1	2.176	0.148
	tInOp:subjId	0.360	46	0.899	0.639
	Residuals	0.348	40		

Table 6: The first rows (until the horizontal line in the middle) significant parameters for a model with an interaction term between the outdoor temperature and the individuals. Below the horizontal line we have a model with interaction term between the indoor temperature and the indivuduals. We can see that none of these interaction terms between the individuals and the temperature are significant. Furthermore we see that those are the first terms to be dropped. We used a Type III partitioning for the ANOVA table.

Table 6 shows that neither of the interaction terms between the temperatures and the individuals are significant. This means that both of our two initial models from equation Equation (1) and Equation (2) will be reduced to the same model.

As in part 1, we remove first higher order terms before we remove lower order terms. We choose which terms to remove by looking at the Type III partitioning in the ANOVA table. We start by removing the terms with the highest p-value, and keep removing terms until all terms are significant. Each removed term can be seen in table 7

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
clo ∼ tOut * tInOp + subjId * tOut	40	0.27				
clo ∼ tOut * tInOp + subjId	86	0.71	-46	-0.44	1.45	0.1171
clo ∼ tOut + tInOp + subjId	87	0.71	-1	-0.01	1.06	0.3092
clo ∼ tOut + subjId	88	0.72	-1	-0.00	0.07	0.7978

Table 7: Deviance table of each of the models. The first column shows the model in R notation. The last column with the P-value shows the P-value when compared to the model from the line above

When all the non-significant terms have been removed we end up with the following model

$$Clo_i = \beta_0 + \beta_1 \cdot t_{out,i} + a(individual_i) \tag{3}$$

When we estimate the model in R, we do however not get an individual estimate of β_0. We estimate $\mu = \beta_0 + a(individual_1)$ which means the intercept and the first individual parameter (subject ID 11) is estimated as one. This also means that each of the parameters for the other individuals are estimated as $g(individual_i) = a(individual_i) - a(individual_1)$.

Instead of fitting the above model we can instead fit a model without an overall intercept as they are equivalent:

$$Clo_i = \beta_1 \cdot t_{out,i} + a(individual_i) \tag{4}$$

This model will ensure an easier interpretation of the subject ID parameters.

Visual presentation of the parameters

By investigating the individual subject ID parameters from Equation (4) we might find an underlying distribution:

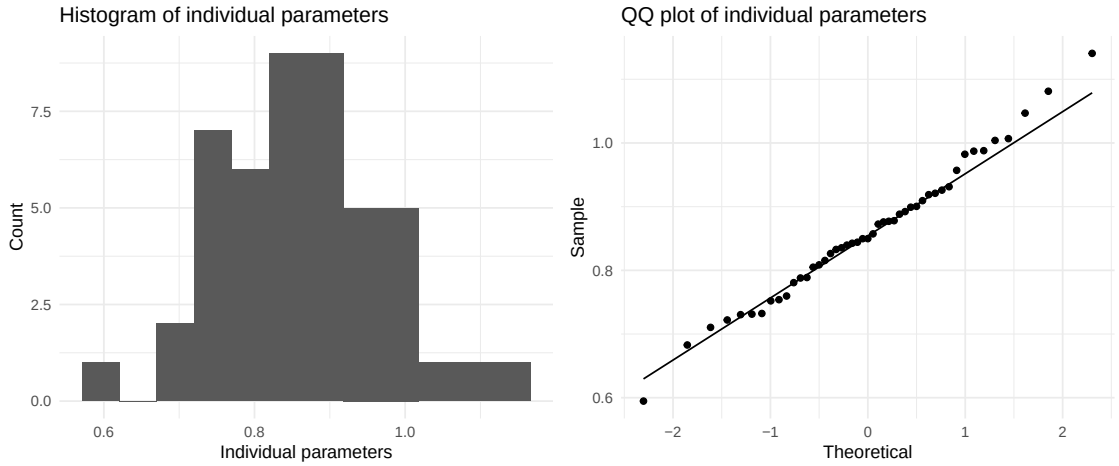


Figure 10: The figure to the left shows a histogram of the individual parameters, $g(individual_i)$ (except for the first individual), while the figure to the right shows a QQ plot. Both suggest that the parameters could be gaussian distributed.

Figure 10 suggests that the parameters of the individuals could be normally distributed. We do see some heavy tails, but it is close enough for the assumption. Ideally, we would have fitted a mixed random and fixed effect model.

Interpretation of the parameters.

The interpretation of the parameters by using Equation (2) instead of Equation (1) become very simple. The subject ID estimates is the individual intercepts, while $\hat{\beta}_1$ is a parameter defines the increase/decrease in clothing level as the indoor operating temperature increases/decreases. The uncertainty of $\hat{\beta}_1$ can be found in table 8.

	Estimate	2.5 %	97.5 %
$\hat{\beta}_1$	-0.0143	-0.0203	-0.0082
$\hat{\sigma}$	0.0902	0.0786	0.1058

Table 8: This table shows the estimate and 95 % confidence interval of the β_1 parameter from (2) and the estimate of σ. The uncertainty is estimated with a Wald type confidence interval. The estimate confidence interval for σ is calculated with the distribution from Theorem 3.5 (page 53) in the book.

We could have added each of the intercepts in the table as well, but the table would become quite large and comparing each of the individual intercepts would become difficult. Instead the individual intercepts can be found in figure 11 along with their confidence intervals. From figure 11 it can be seen that most confidence intervals for subject IDs are within each others range. Though it should be noticed that some does not eg. 113 and 123.

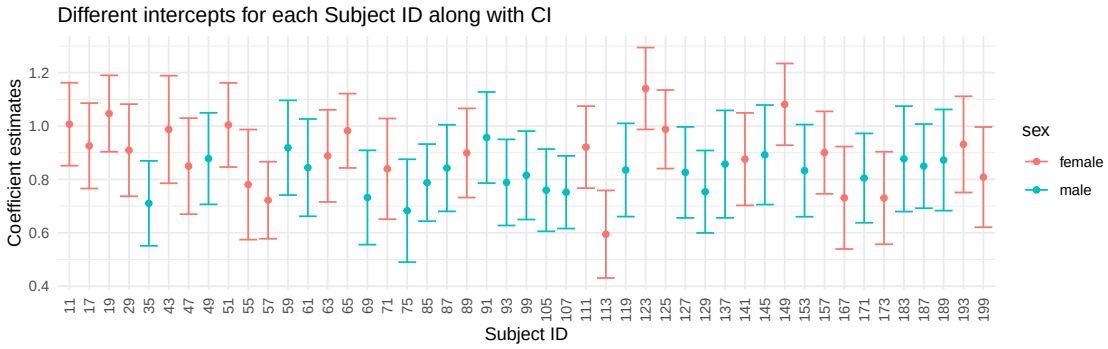


Figure 11: This figure shows each parameter with its 95 % confidence interval. The uncertainty is estimated with a Wald type confidence interval.

Prediction & Residual analysis

For this particular model, we run into issues when we want to start making predictions. We can estimate the predictions for our data points and compute the residuals but we cannot make a mean prediction. At least not directly. We can make a prediction for each subject ID. This means we can pick the subjects with the median, 25% and 75% quantiles intercepts. We can however not predict a new individual, at least not without assuming some kind of probability distribution.

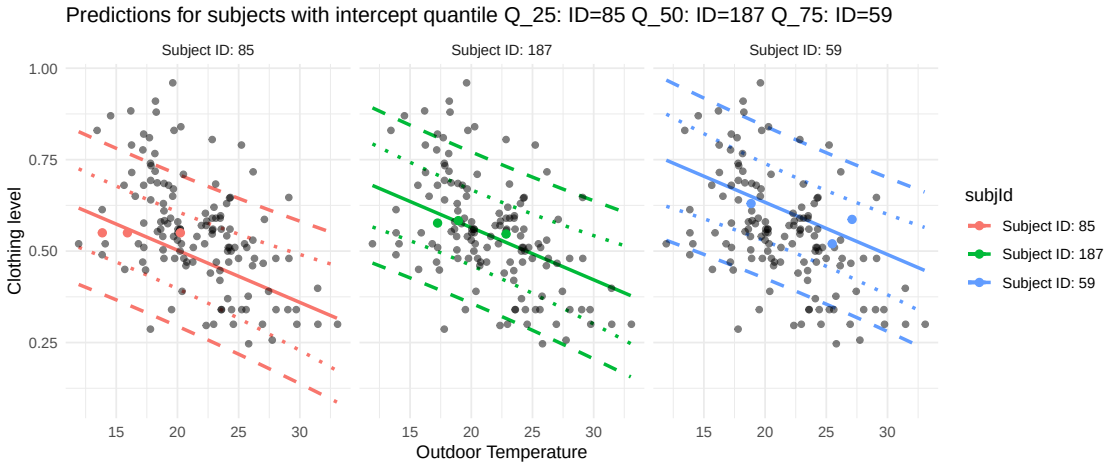


Figure 12: This figure shows the predictions for the individuals who had an intercept corresponding to the 25, 50 and 75% quantiles. All plots contain all the data points.

Figure 12 shows that we can make predictions for each of our individuals. These predictions does not matter for all of the other individuals as each of them have their own intercept. We have to look at the residuals for each data point individually:

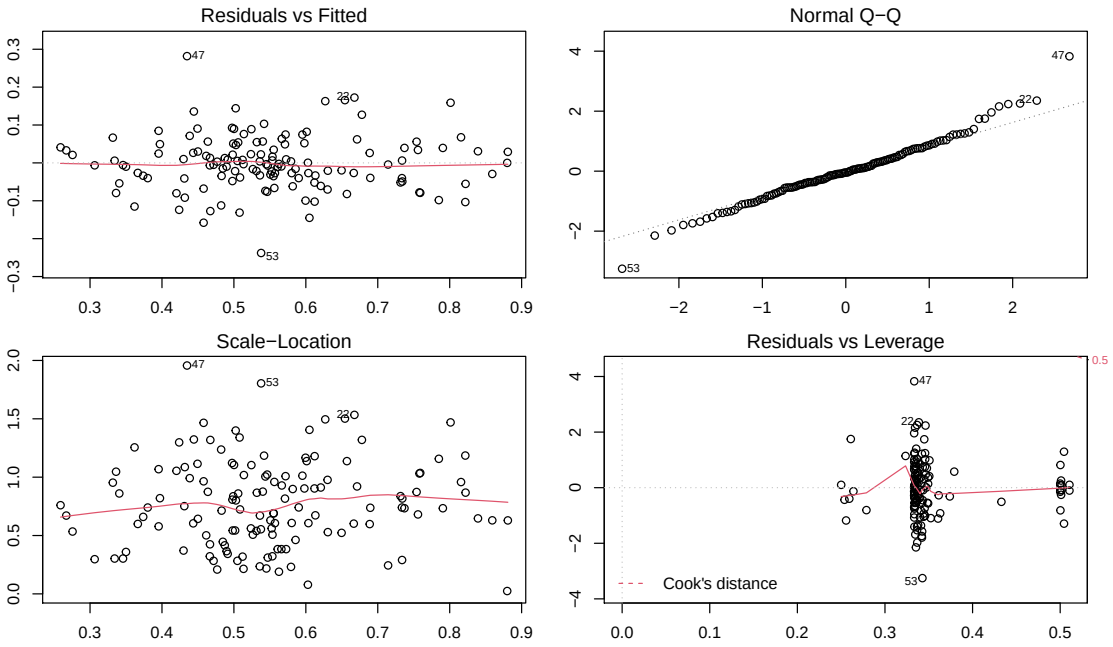


Figure 13: The upper left plot shows the Residuals vs. the fitted values. The plot to the upper right shows a QQ plot. The plot to the lower left is a Scale-Location plot and the final plot to the lower right is a residual vs leverage plot

Figure 13 shows various plots of the residuals. The plot of the residuals vs the fitted values looks like white noise. This is an indication that our model did indeed catch the underlying pattern in the data. The QQ-plot do show some heavy tails, especially to most positive residuals. It is however still close enough to the theoretical quantiles, that it could be normally distributed. The Scale-Location plot also show an almost flat line, meaning that the variance seem to be constant for all observations. The residuals vs the leverage (Cook's Distance) do not indicate that any of the observation have a high influence that does not correspond with the fitted model.

In problem A, we had issues with the female observations would contain more noise. Even though we discarded this grouping we can still investigate if this is the case for this model as well:

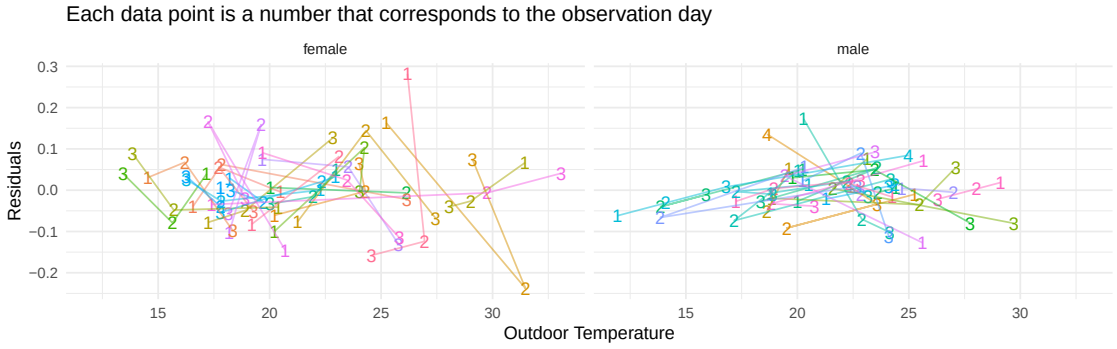


Figure 14: This figure shows the residuals vs. the outdoor temperature, where each point is symolised by the observation day. The different colors are the subject IDs.

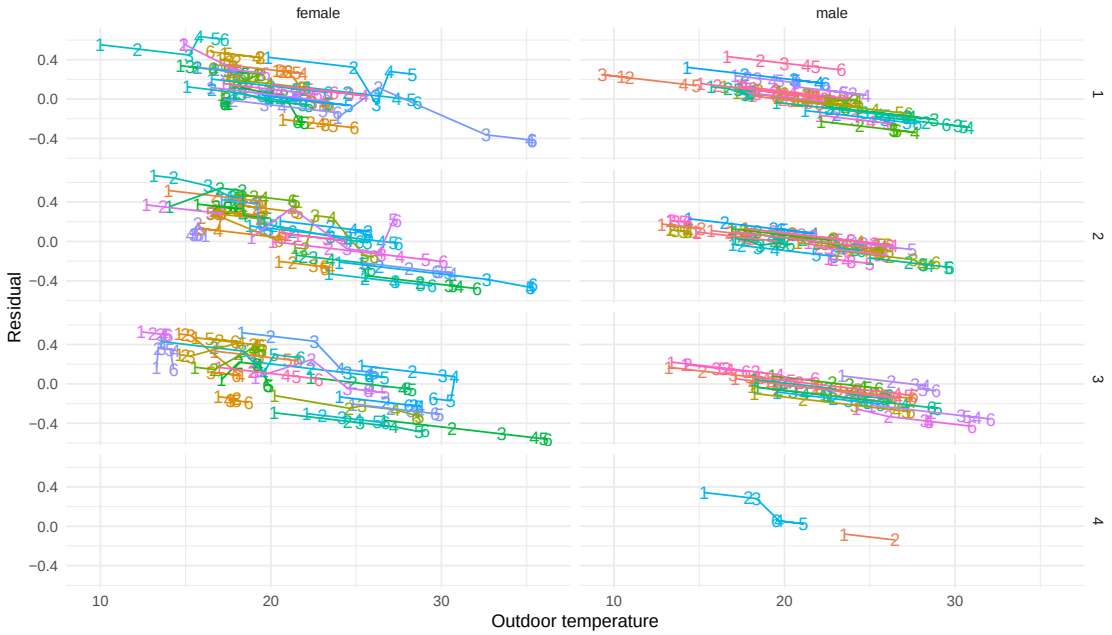
Figure 14 does show that the female group have slightly more variance than the male group. If we fitted a new model with a weight on each group, we would see that the new model would get a lower AIC than

our final (part B) model, but the actual predictions/residuals would hardly change enough for us to see the visual difference. This does however still mean that the model with the weights would theoretically be better because of the lower AIC.

Figure 14 also shows each data point as a number corresponding to the observation day, while the color indicates different subject IDs. It does not seem like, from a visual perspective, that the observation days have any pattern, which means they probably would not have been useful for modelling.

Problem C

For this section, the full data set of the previously used data will be used. This includes the same variables as previously but now includes six observations per day. The model selected in Problem B will be fitted to the full data and the residuals will be investigated.



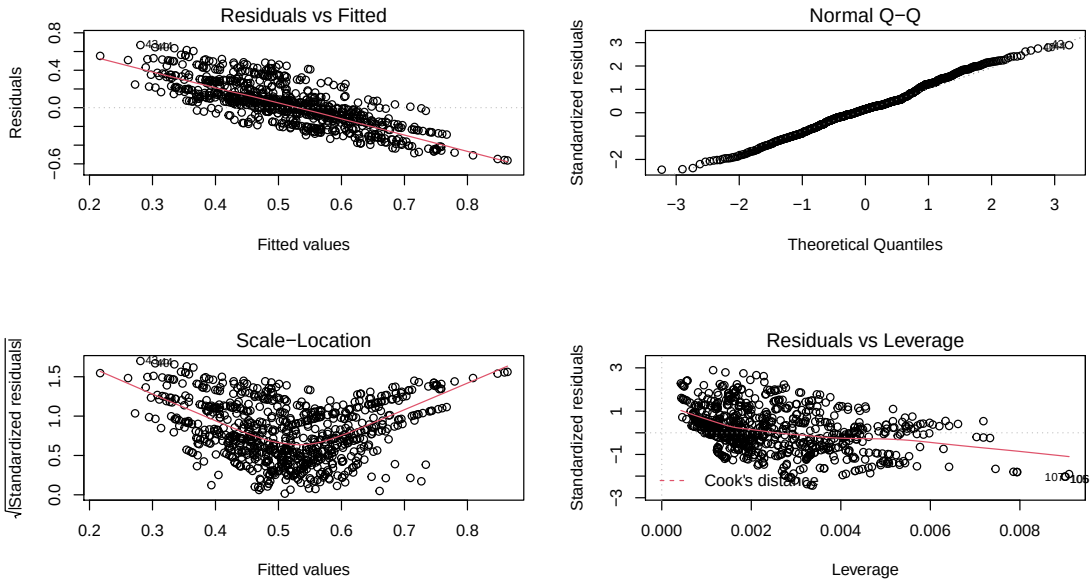


Figure 15: Residuals of model (eq. 4) fitted on the whole data set.

From Figure 15, the residuals appear to be normally distributed with no heavy tails according to the QQ-plot. When looking at the residuals and the standardized residuals, however, these do not appear to be independently distributed. The residuals plotted as a function of the outdoor temperature is also presented below. The residuals are split into to columns according to gender and columns depending on the day.

Figure 16: The residuals of the input variables.

From Figure 16 it can be seen, that a higher variance for the female residuals compared to the male is still observed. A negative correlation between the outdoor tempearature and the model residuals can also be seen.

Finally, the residuals are plotted as a function of the outdoor temperature separately for the 4 different days where data is available. The observations are numbered according to their obsersavation number of the day.

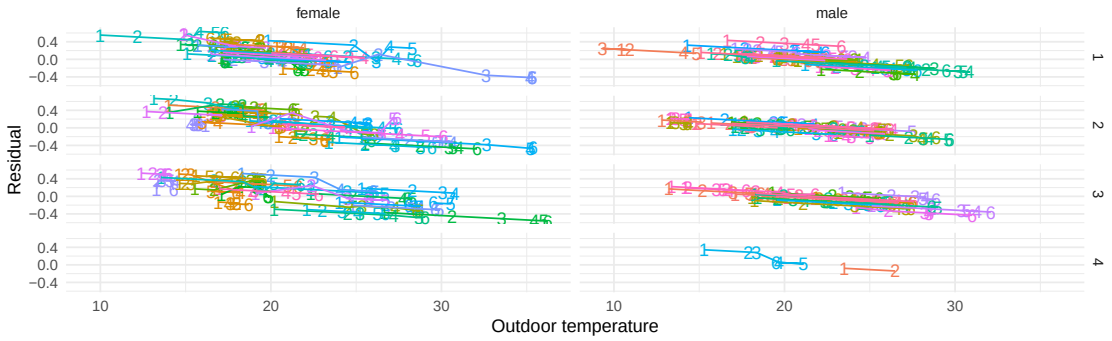


Figure 17: The residuals as a function of outdoor temperature over the 4 observed days. The observations are numbered accordingly to their observation number of the day.

From Figure 17 it is apparent that the outdoor temperature observations are correlated by day. This violates the assumption for the data that it must be independently distributed.

Appendix: R-code

```
source('setup.R')
rm(list=ls())
library(ggplot2)
library(knitr)
library(gridExtra)
library(latex2exp)
library(tidyverse)
library(xtable)
library(directlabels)
source('setup.R')
library(ggplot2)
library(lmtest)
library(car)
library(gridExtra)
df1=read.csv(file="./Data/clothingFull.csv")
df2=read.csv(file="./Data/clothingSum.csv")
df3 = df2[,c(3:6)]


p1 <- ggplot(df3,aes(x=tOut,y=clo, col = as.factor(sex)) )+geom_point()+facet_wrap(~sex)+xlab("Outdoor
p2 <- ggplot(df3,aes(x=tOut,y=tInOp, col = as.factor(sex)))+geom_point()+facet_wrap(~sex) +xlab("Outdoo

p3 <- ggplot(df3,aes(x=tInOp,y=clo, col = as.factor(sex)) )+geom_point()+facet_wrap(~sex) +xlab("Indoor

grid.arrange(p1,p2,p3, nrow = 3)
#histogrammer

hist1 <- ggplot(df3,aes(x=clo,fill = as.factor(sex)))+geom_histogram(color="white", binwidth = .1)+face
hist2 <- ggplot(df3,aes(x=tOut,fill = as.factor(sex)))+geom_histogram(color="white", binwidth = 2)+face
hist3 <- ggplot(df3,aes(x=tInOp,fill = as.factor(sex)))+geom_histogram(color = "white", binwidth = .5)+
```

```
grid.arrange(hist1,hist2,hist3, nrow = 1)
source('setup.R')
rm(list=ls())
library(ggplot2)
library(knitr)
library(RColorBrewer)
library(tidyverse)
library(latex2exp)
library(car)
library(xtable)
library(gridExtra)
df1=read.csv(file="./Data/clothingFull.csv")
df2=read.csv(file="./Data/clothingSum.csv")
names(df2)[5]="tIn"

#######Unweighted model - using backward selection and type III partitioning:
fit1=lm(clo~ tOut*tIn*sex,data=df2)
Anova(fit1, type = 'III')
#Remove third order interaction tOut:tIn:sex

fit2a=lm(clo~ tOut+tIn+tOut*sex+tIn*sex + tOut*tIn,data=df2)
Anova(fit2a, type = 'III')
#Remove tIn*tOut

fit2b=lm(clo~ tOut+tIn+tOut*sex+tIn*sex,data=df2)
Anova(fit2b, type = 'III')
#Remove tOut:sex

fit3=lm(clo~ tOut+tIn+tIn:sex+sex,data=df2)
Anova(fit3, type = 'III')
#Final optimal model:
#Create tabel:
xtable(anova(fit1, fit2a, fit2b, fit3))

#Estimate sigma:

sigma_est = sigma(fit3)
anova(fit1,fit3)

lm_sigma_conf<-function(fit,CI=0.95){
  # Book page 53: sigma_hat^2~sigma^2*chisq_f/f where f=n-k
  # rearanging: sigma_hat^2*f/chisq_f
  sigma=sigma(fit)
  d_f=fit$df.residual
  conf_sigma2=d_f*sigma^2/qchisq(c((1-CI)/2,1-(1-CI)/2), df = d_f, lower.tail = FALSE)
  conf=sqrt(conf_sigma2)
  names(conf)=paste(c((1-CI)/2,1-(1-CI)/2)*100,"%")
  return(conf)
}
lm_sigma_conf(fit3)

#LRT:
```

```r
anova(fit3, fit1)

#CL of parameters:

conf = confint(fit3)
df_param = cbind(conf[,1], fit3$coefficients, conf[,2])
#xtable(df_param)

#########################Plot prediction and confidence intervals of unweighted model using three fixed
#tInOp:
tmp_male=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=quantile(df2$tIn,0.25),sex="ma
tmp_female=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=quantile(df2$tIn,0.25),sex="
new=rbind(tmp_male,tmp_female)

conf = predict(fit3,new,interval = 'confidence')
pred = predict(fit3, new,interval = 'prediction')

out = data.frame(cbind(new, conf, pred))
names(out)
q1 = ggplot(data  = out, aes(x = tOut, y = fit, col = sex)) + geom_line()+
  geom_line(aes(x = tOut, y = lwr), lty = 3)+ geom_line(aes(x = tOut, y = upr), lty = 3)+
  geom_line(aes(x = tOut, y = upr.1), lty = 2) + geom_line(aes(x = tOut, y = lwr.1), lty = 2)+
  geom_point(data = df2, aes(x = tOut, y = clo, col= as.factor(sex))) + facet_wrap(~sex) +
  labs(title="25% Quantile of indoor operating temperature (tInOp = 26.01)",
       x ="Outdoor temperature (tOut)", y = "Clothing level (clo)")+ theme(plot.title = element_text(h


###- Mean:
tmp_male=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=mean(df2$tIn),sex="male")
tmp_female=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=mean(df2$tIn),sex="female")
new=rbind(tmp_male,tmp_female)

conf = predict(fit3,new,interval = 'confidence')
pred = predict(fit3, new,interval = 'prediction')
out = data.frame(cbind(new, conf, pred))
q2 = ggplot(data  = out, aes(x = tOut, y = fit, col = sex)) + geom_line()+
  geom_line(aes(x = tOut, y = lwr), lty = 3)+ geom_line(aes(x = tOut, y = upr), lty = 3)+
  geom_line(aes(x = tOut, y = upr.1), lty = 2) + geom_line(aes(x = tOut, y = lwr.1), lty = 2)+
  geom_point(data = df2, aes(x = tOut, y = clo, col= as.factor(sex))) + facet_wrap(~sex) +
  labs(title="Mean of indoor operating temperature (tInOp = 26.82)",
       x ="Outdoor temperature (tOut)", y = "Clothing level (clo)")+ theme(plot.title = element_text(h

#75 % quantile
tmp_male=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=quantile(df2$tIn,0.75),sex="ma
tmp_female=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=quantile(df2$tIn,0.75),sex="
new=rbind(tmp_male,tmp_female)

conf = predict(fit3,new,interval = 'confidence')
pred = predict(fit3, new,interval = 'prediction')
out = data.frame(cbind(new, conf, pred))

q3 = ggplot(data  = out, aes(x = tOut, y = fit, col = sex)) + geom_line()+
  geom_line(aes(x = tOut, y = lwr), lty = 3)+ geom_line(aes(x = tOut, y = upr), lty = 3)+
```

```
  geom_line(aes(x = tOut, y = upr.1), lty = 2) + geom_line(aes(x = tOut, y = lwr.1), lty = 2)+
  geom_point(data = df2, aes(x = tOut, y = clo, col= as.factor(sex))) + facet_wrap(~sex) +
  labs(title="75% Quantile of indoor operating temperature (tInOp = 27.48)",
       x ="Outdoor temperature (tOut)", y = "Clothing level (clo)")+ theme(plot.title = element_text(h

grid.arrange(q1,q2,q3)
par(mfrow = c(2,2))
plot(fit3)

#Further residual analysis:
df2$residual = resid(fit3)
q1 = ggplot(df2, aes(sample = residual, colour = sex)) +
  stat_qq() +
  stat_qq_line()
q2 = ggplot(data = df2, aes(x = sex, y= residual, colour = sex))+ geom_boxplot() + geom_jitter(width = 

q3 = ggplot(data = df2, aes(x = tIn, y= residual, col = sex))+ geom_point()  + xlab('Indoor operating te

q4 = ggplot(data = df2, aes(x = tOut, y= residual, col = sex))+ geom_point() + xlab('Outdoor temperature

grid.arrange(q1,q2,q3,q4, ncol = 2)
#Normalize residuals and plot QQ-plot:
frac = var(resid(fit3)[df2$sex == 'female'])/var(resid(fit3)[df2$sex == 'male'])
par(mfrow=c(1,1))
res = resid(fit3)
res[df2$sex == 'female'] = res[df2$sex == 'female']/frac
qqnorm(res)
qqline(res)

#Optimize weight of weighted residual model:

ll_partA<-function(a,data=df2){
  weights=rep(1,nrow(data))
  weights[data$sex=="female"]=1/a
  fit=lm(clo~ tOut+tIn+tIn:sex+sex,w=weights,data=data)
  return(logLik(fit))
}
weighted_lm<-function(a,data=df2){
  weights=rep(1,nrow(data))
  weights[data$sex=="female"]=1/a
  fit=lm(clo~ tOut+tIn+tIn:sex+sex,w=weights,data=data)
  return(fit)
}
# Finding optimal weights
opt=optim(1,ll_partA,control=list(fnscale=-1),hessian=T)
# Fitting model with optimal weights
fit=weighted_lm(opt$par)

#Confidence intervals:
m=confint(fit)
m=rbind(m,weights=opt$par+c(-1,1)*sqrt(diag(solve(-opt$hessian)))*qt(0.975,df=fit$df.residual))
m=cbind(c(coef(fit),weigh_par=opt$par),m)
m
```

```
conf = confint(fit)
df_param = cbind(conf[,1], fit$coefficients, conf[,2])

c(lm_sigma_conf(fit), sigma(fit))
c(AIC(fit3), AIC(fit))
c(BIC(fit3), BIC(fit))
tmp_male=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=quantile(df2$tIn,0.25),sex="ma
tmp_female=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=quantile(df2$tIn,0.25),sex="
new=rbind(tmp_male,tmp_female)

w = rep(0, length(new[,1]))
w[new$sex == 'female'] = opt$par
w[new$sex =='male'] = 1


conf = predict(fit,new,weights = 1/w, interval = 'confidence')
pred = predict(fit, new,weights = 1/w, interval = 'prediction')

out = data.frame(cbind(new, conf, pred))
names(out)
q1 =   ggplot(data  = out, aes(x = tOut, y = fit, col = sex)) + geom_line()+
  geom_line(aes(x = tOut, y = lwr), lty = 3)+ geom_line(aes(x = tOut, y = upr), lty = 3)+
  geom_line(aes(x = tOut, y = upr.1), lty = 2) + geom_line(aes(x = tOut, y = lwr.1), lty = 2)+
  geom_point(data = df2, aes(x = tOut, y = clo, col= as.factor(sex))) + facet_wrap(~sex) +
  labs(title="25% Quantile of indoor operating temperature (tInOp = 26.01)",
       x ="Outdoor temperature (tOut)", y = "Clothing level (clo)")+ theme(plot.title = element_text(h


###- Mean:
tmp_male=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=mean(df2$tIn),sex="male")
tmp_female=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=mean(df2$tIn),sex="female")
new=rbind(tmp_male,tmp_female)

w = rep(0, length(new[,1]))
w[new$sex == 'female'] = opt$par
w[new$sex =='male'] = 1


conf = predict(fit,new,weights = 1/w, interval = 'confidence')
pred = predict(fit, new,weights = 1/w, interval = 'prediction')

out = data.frame(cbind(new, conf, pred))
q2 = ggplot(data  = out, aes(x = tOut, y = fit, col = sex)) + geom_line()+
  geom_line(aes(x = tOut, y = lwr), lty = 3)+ geom_line(aes(x = tOut, y = upr), lty = 3)+
  geom_line(aes(x = tOut, y = upr.1), lty = 2) + geom_line(aes(x = tOut, y = lwr.1), lty = 2)+
  geom_point(data = df2, aes(x = tOut, y = clo, col= as.factor(sex))) + facet_wrap(~sex) +
  labs(title="Mean of indoor operating temperature (tInOp = 26.82)",
       x ="Outdoor temperature (tOut)", y = "Clothing level (clo)")+ theme(plot.title = element_text(h

#75 % quantile
tmp_male=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=quantile(df2$tIn,0.75),sex="ma
tmp_female=data.frame(clo=NA,tOut=seq(min(df2$tOut),max(df2$tOut),0.1),tIn=quantile(df2$tIn,0.75),sex="
new=rbind(tmp_male,tmp_female)
```

```
w = rep(0, length(new[,1]))
w[new$sex == 'female'] = opt$par
w[new$sex =='male'] = 1


conf = predict(fit,new,weights = 1/w, interval = 'confidence')
pred = predict(fit, new,weights = 1/w, interval = 'prediction')

out = data.frame(cbind(new, conf, pred))

q3 = ggplot(data  = out, aes(x = tOut, y = fit, col = sex)) + geom_line()+
  geom_line(aes(x = tOut, y = lwr), lty = 3)+ geom_line(aes(x = tOut, y = upr), lty = 3)+
  geom_line(aes(x = tOut, y = upr.1), lty = 2) + geom_line(aes(x = tOut, y = lwr.1), lty = 2)+
  geom_point(data = df2, aes(x = tOut, y = clo, col= as.factor(sex))) + facet_wrap(~sex) +
  labs(title="75% Quantile of indoor operating temperature (tInOp = 27.48)",
        x ="Outdoor temperature (tOut)", y = "Clothing level (clo)")+ theme(plot.title = element_text(h

grid.arrange(q1,q2,q3)
par(mfrow = c(2,2))
plot(fit)
df2$residuals = resid(fit)
df2$subj_number = 1:nrow(df2)
ggplot(data = df2, aes(x = subj_number, y = residuals, col = as.factor(subjId))) + geom_point() + geom_
  xlab('Index') + ggtitle('Residuals of observations')+facet_wrap(~as.factor(sex))+ theme(plot.title = e

source('setup.R')
library(ggplot2)
library(car)
library(xtable)
df2=read.csv(file="./Data/clothingSum.csv")
names(df2)[5]="tIn"
df2$subjId = factor(df2$subjId)
df2$sex=factor(df2$sex)

fitB1 = lm( clo ~ tOut*tIn+subjId*tOut,data = df2)
m1=Anova(fitB1,type=3)
#m1
fitB2 = lm( clo ~ tOut*tIn+subjId*tIn,data = df2)
m2=Anova(fitB2,type=3)
#m2
#xtable(rbind(m1,m2),digits=c(0,3,0,3,3))

# Fitting the reduced model
fitB = lm(clo ~ tOut*tIn+subjId,data = df2)
Anova(fitB,type=3)

# Dropping interaction term and refitting
fit2B = lm(clo ~ tOut+tIn+subjId,data = df2)
Anova(fit2B,type=3)

# Dropping tIn and refitting
fit3B = lm( clo ~ tOut+subjId,data = df2)
Anova(fit3B,type=3)
```

```
#xtable(anova(fitB1,fitB,fit2B,fit3B))

fit3B2 = lm( clo ~ -1+tOut+subjId,data = df2)
#Anova(fit3B,type=3)
#Question 2:
p1=ggplot(data.frame(c()),aes(x=fit3B2$coefficients[-1]))+
  geom_histogram(bins=12)+xlab("Individual parameters")+ylab("Count")+
  ggtitle("Histogram of individual parameters")

p2=ggplot(data.frame(c()),aes(sample=fit3B2$coefficients[-1]))+
  stat_qq()+stat_qq_line()+xlab("Theoretical")+ylab("Sample")+
  ggtitle("QQ plot of individual parameters")

library(gridExtra)
grid.arrange(p1,p2,ncol=2)
# Confidence interval for the fixed effects
m=confint(fit3B2)[1,]
m=cbind(fit3B2$coef[1],t(m))
# CI for sigma function
lm_sigma_conf<-function(fit,CI=0.95){
  # Book page 53: sigma_hat^2~sigma^2*chisq_f/f where f=n-k
  # rearanging: sigma_hat^2*f/chisq_f
  sigma=sigma(fit)
  d_f=fit$df.residual
  conf_sigma2=d_f*sigma^2/qchisq(c((1-CI)/2,1-(1-CI)/2), df = d_f, lower.tail = FALSE)
  conf=sqrt(conf_sigma2)
  names(conf)=paste(c((1-CI)/2,1-(1-CI)/2)*100,"%")
  return(conf)
}
# Combining the results
tmp = lm_sigma_conf(fit3B2)
tmp = cbind(sigma=sigma(fit3B2),t(tmp))
m = rbind(m,tmp)
#xtable(m,digits=4)
l = length(fit3B2$coefficients)
std_coef = sqrt(diag(vcov(fit3B2)))[-1]
coef_est = fit3B2$coefficients[-1]

df_coef = data.frame(cbind(coef_est, confint(fit3B2)[-1,]))
library(stringr)
df_coef$names =gsub("subjId","",rownames(df_coef))
df_coef$sex = NaN
for (i in 1:(l-1)){
  df_coef$sex[i] = as.character(df2$sex[df_coef$names[i]==df2$subjId][1])
}
df_coef = df_coef[order(as.numeric(as.character(df_coef$names))),]
df_coef$names = factor(as.numeric(as.character(df_coef$names)))
df_coef$sex = df_coef$sex
#levels(df_coef$names)

ggplot(data = df_coef, aes(x = names, y = coef_est, color = sex)) +
  geom_point()+
  geom_errorbar(aes(ymin=X2.5.., ymax=X97.5..)) +
```

```
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  xlab("Subject ID")+ylab("Coefficient estimates")+
  ggtitle("Different intercepts for each Subject ID along with CI")
nam=c()
quant=c(0.25,0.5,0.75)
for(i in quant){
  tmp=names(coef(fit3B2)[coef(fit3B2)%in%quantile(coef(fit3B2)[-1],probs = i,type=3)])
  nam=c(nam,tmp)
}

nam=gsub("subjId","",nam)
nam
t_seq=seq(min(df2$tOut),max(df2$tOut),0.1)

new_data=data.frame(clo=NA,tOut=rep(t_seq,3),subjId=rep(nam,each=length(t_seq)))
conf = predict(fit3B2,new=new_data,interval = 'confidence')
pred = predict(fit3B2, new=new_data,interval = 'prediction')

new_data$subjId=paste("Subject ID:",new_data$subjId)
new_data$subjId=factor(new_data$subjId,levels=paste("Subject ID:",nam))


new_data$clo = conf[,"fit"]
new_data$CI_lwr = conf[,"lwr"]
new_data$CI_upr = conf[,"upr"]
new_data$PI_lwr = pred[,"lwr"]
new_data$PI_upr = pred[,"upr"]
tmp = df2[df2$subjId%in%nam,]
tmp$subjId=as.character(tmp$subjId)
tmp$subjId=paste("Subject ID:",tmp$subjId)
tmp$subjId = factor(tmp$subjId,levels=paste("Subject ID:",nam))
ggplot(data  = new_data, aes(x = tOut, y = clo)) + geom_line(aes(col=subjId),lwd=1)+
  geom_line(aes(x = tOut, y = CI_lwr,col=subjId), lty = 3,lwd=1)+
  geom_line(aes(x = tOut, y = CI_upr,col=subjId), lty = 3,lwd=1)+
  geom_line(aes(x = tOut, y = PI_lwr,col=subjId), lty = 2,lwd=1)+
  geom_line(aes(x = tOut, y = PI_upr,col=subjId), lty = 2,lwd=1)+
  facet_wrap(~subjId)+
  geom_point(data = select(df2,-subjId), aes(x = tOut, y = clo),col='black',alpha=0.5)+
  geom_point(data = tmp,aes(x = tOut, y = clo,col=subjId),size=2)+
  xlab("Outdoor Temperature")+ylab("Clothing level")+
  ggtitle(paste("Predictions for subjects with intercept quantile",
                paste0("Q_",quant*100,": ID=",nam,collapse = " ")))

par_opt=par()
par(mfrow=c(2,2),mar=c(2,2,2,2))
plot(fit3B2)
par(mfrow=c(1,1),mar=par_opt$mar)
df2$residual=resid(fit3B2)
ggplot(df2,aes(x=tOut,y=residual,col=subjId))+
  geom_text(aes(label=day))+geom_path(alpha=0.5)+
  facet_wrap(~sex)+theme(legend.position = "none")+
  ggtitle("Each data point is a number that corresponds to the observation day")+
  xlab("Outdoor Temperature")+ylab("Residuals")
```

```
source('setup.R')
df1=read.csv(file="./Data/clothingFull.csv")
names(df1)[4]="tIn"


fit3B = lm( clo ~ -1 +tOut+subjId,data = df1)
df1$obs.no=as.character(df1$obs.no)
df1$subjId=as.character(df1$subjId)
ggplot(df1,aes(x=tOut,y=resid(fit3B),group=subjId,col=subjId))+
  geom_text(aes(label=obs.no))+geom_path()+facet_grid(cols=vars(sex),rows=vars(day))+
  theme(legend.position = "none")+
  xlab("Outdoor temperature")+ylab("Residual")
par(mfrow=c(2,2))
plot(fit3B)
df1$residual = resid(fit3B)
#q1 = ggplot(df1, aes(sample = residual, colour = sex)) +
#  stat_qq() +
#  stat_qq_line()
#q2 =  ggplot(df1, aes(sample = residual)) +
#  stat_qq() +
#  stat_qq_line()

q3=   ggplot(data = df1, aes(x = sex, y= residual, colour = sex))+ geom_boxplot() + geom_jitter(width =


q4 = ggplot(data = df1, aes(x = tOut, y= residual, col = sex))+ geom_point() + xlab('Outdoor temperatur

grid.arrange(q3,q4, ncol = 2)
df1$obs.no=as.character(df1$obs.no)
df1$subjId=as.character(df1$subjId)
ggplot(df1,aes(x=tOut,y=residual,group=subjId,col=subjId))+
  geom_text(aes(label=obs.no))+geom_path()+facet_grid(cols=vars(sex),rows=vars(day))+
  theme(legend.position = "none")+
  xlab("Outdoor temperature")+ylab("Residual")

# ll_partA<-function(a,data=df1){
#   weights=rep(1,nrow(data))
#   weights[data$sex=="female"]=1/a
#   fit=lm(clo~ tOut+tIn+tIn:sex+sex,w=weights,data=data)
#   return(logLik(fit))
# }
# weighted_lm<-function(a,data=df1){
#   weights=rep(1,nrow(data))
#   weights[data$sex=="female"]=1/a
#   fit=lm(clo~ tOut+tIn+tIn:sex+sex,w=weights,data=data)
#   return(fit)
# }
# # Finding optimal weights
# opt=optim(1,ll_partA,control=list(fnscale=-1),hessian=T)
# # Fitting model with optimal weights
# fit=weighted_lm(opt$par)
```