02424 - Assignment 1

Spring 22, 02424
Assignment 1
March 11, 2022

Nicolaj Hans Nielsen, s184335
Anton Ruby Larsen, s174356

Danmarks
Tekniske Universitet

Contents

A | A General Linear Model

A.1 Data Exploration

In this section, we take a close look at our dataset.

In the table below, you will find a table of the available data for this project.

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 A.1 – Description of data available

In the initial part of this project, we will limit the scope to the variables *clo*, *tOut*, *tInOp*, and *sex*.

A.1.1 Covariance

To have an overall idea of the variance and covariance of our data, we depict them in pairs below. To include *sex*, we have colored the points such that males are blue and females are orange. This distinction between males and females by use of color, is the same thorough out all figures in this exercise.

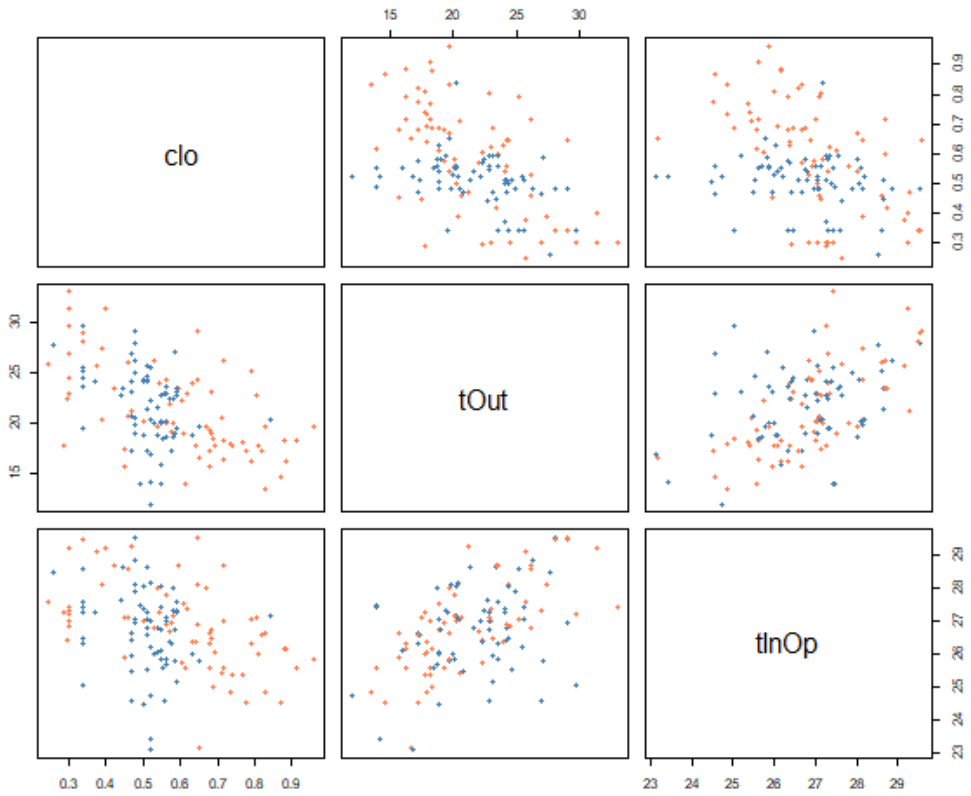

Figure A.1 – Pairs of figures to understand in the data. The factor sex is plotted using a coloring-scheme where blue is male and orange is female.

In figure A.1, we see that in general there seems to be more clear linear patterns for the females than for the male. Consider e.g. *clo-tOut* plot; as the outer temperature increases it seems quite evident that the females tends to have a lower level of clothing. For the males, this trend is not as evident. The same patterns goes for the *clo-tInOp*. If we disregard the gender, it is quite evident that when it gets warmer, the level of clothing decreases. There also seems to be a clear relation between the temperatures; when it is warmer outside, it tends to be warmer in the office as well. Notice again that this trend seems more obvious for the females than for the males. To support this claim, consider the correlations:

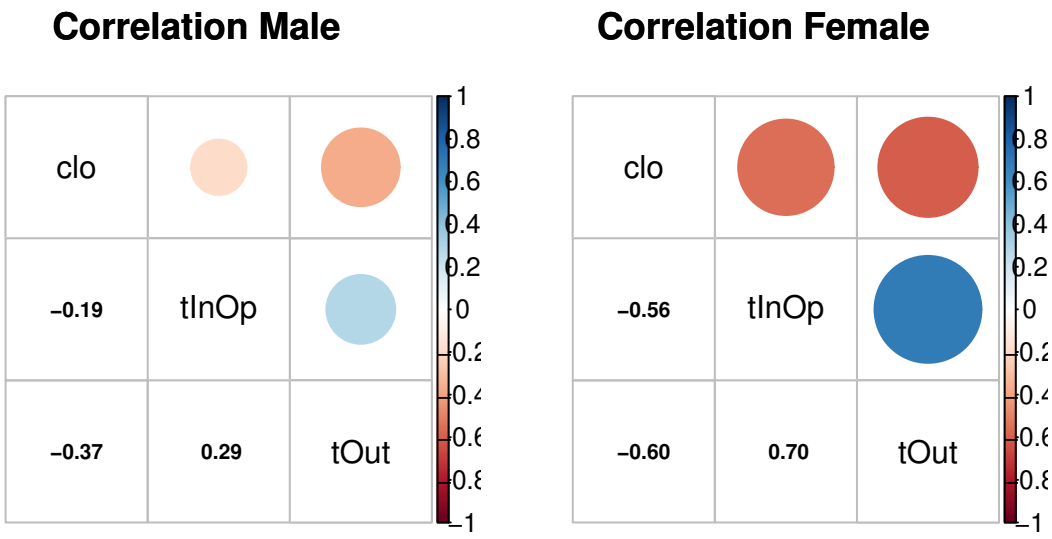

Figure A.2 – The correlation matrix for numerical varibels for each sex. Note that the colorbar is the same for each sex.

In figure A.2, we see that the magnitude of the correlation between variables in the data is much higher for the females than for the males. Even for the temperatures have a much larger correlation.

A.2 Model Building

To find a suitable model, we will first specify a rather comprehensive model which we will refer to as a *sufficient* model. This model must satisfy a number of assumptions. From there we do backward selection using type II testing to reduce the model until only necessary terms are left. The stopping criteria is when all terms are significant with level of 0.95.

A.2.1 Specifying a sufficient model

For now, consider a model of the form:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\Sigma) \tag{A.2.1}$$

Here $\mathbf{Y}$ is the response variable we want to predict, $\mathbf{X}$ is the design matrix, β is the linear relation between $\mathbf{X}$ and $\mathbf{Y}$ we want to find and Σ is the covariance structure which we here assume to be given by the identity matrix, $\mathbf{I}$.
To estimate β we will use the maximum likelihood estimate (MLE) as in theorem 3.2 in [1] and here stated in equation A.2.2.

$$\hat{\beta} = \left(\mathbf{X}^T\Sigma^{-1}\mathbf{X}\right)\mathbf{X}^T\Sigma^{-1}Y \tag{A.2.2}$$

Again we here assume Σ to be given as the identity matrix and the design matrix to be full rank, for $\hat{\beta}$ to be unique.
With an estimate of $\hat{\beta}$, the residuals of A.2.1 must meet the following three assumptions for the found model to be sufficient.

- *Normality*: The standardized residuals must follow a standard normal distribution.
- *Independence*: The standardized residuals must be independent
- *Homoscedasticity*: The standardized residuals must be homoscedastic.

We will first consider the most advanced model, we can build with interactions between all variables. Using the notation explained in section 3.12 in [1] which is adapted from **R**, we start with the model:

$$\text{clo} \sim \text{tOut} * \text{tInOp} * \text{factor(sex)} \tag{A.2.3}$$

First we will check the normality assumption for the residuals of A.2.3 by inspecting the qq-plot in figure A.3.

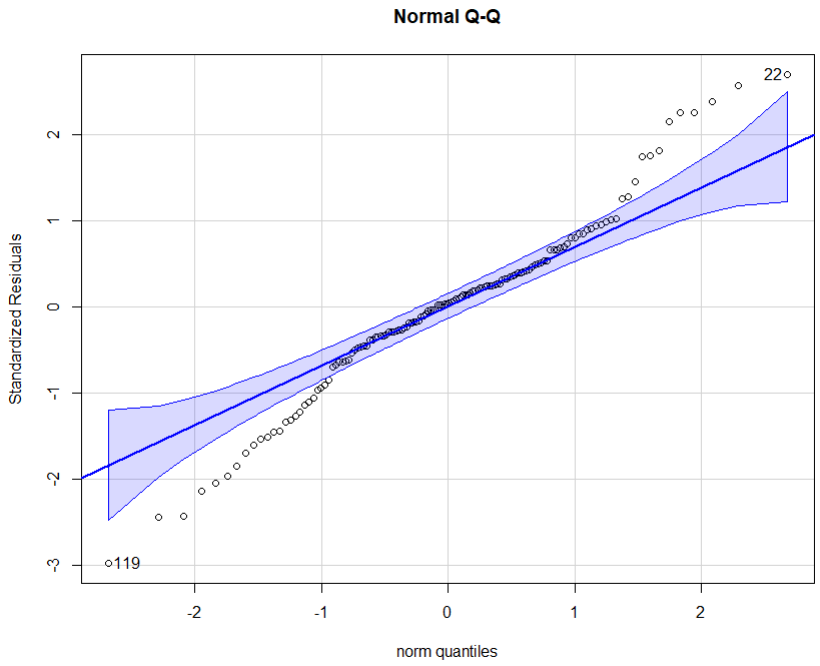

Figure A.3 – QQ-plot of the residuals of A.2.3.

We see the tails are outside the confidence interval hence the normality assumption is violated. We continue checking if independence and homoscedasticity is met by investigating the standardized residuals given in definition 3.12 in [1]. We plot them against fitted and explanatory values in figure A.4 as is advised in section 3.10 in [1].

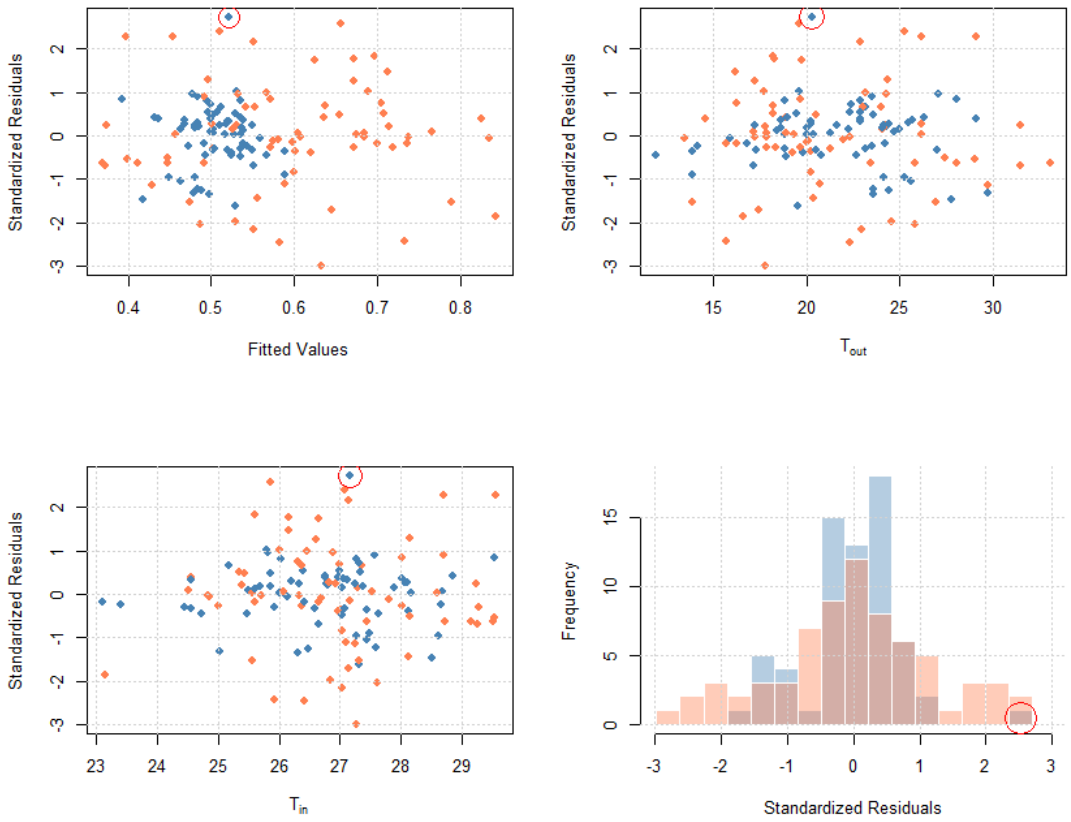

Figure A.4 – The blue color indicates when the explanatory variable *sex* equals *'male'* and the red color indicates *'female'*. The encircled point could be a potential outlier.

If we were to investigate the residuals without taking the factor variable *sex* into consideration we would see no patterns and on the basis of section 3.10 in [1] conclude dependence and homoscedasticity. If we though look at the variance of *males* and *females*, it seems two different levels of variance are present. To check this we will do a 1-way between-groups ANOVA on the factor *sex*.

Df	F value	p value
1	32.488	$7.30 \cdot 10^{-08}$

Table A.2 – 1-way between-groups ANOVA on the factor *sex* testing for homoscedasticity

From the ANOVA table given in A.2 we can conclude on the basis of the very small p value that the residuals also fail on homoscedasticity.
Even though the model only meets the independence assumption, we will continue with the model for now.

Outlier analysis

Before moving to the backward selection of our model, A.2.3, we notice the extreme *'male'* observation incircled in red in figure A.4. To investigate this outlier further we will use the residuals for which the factor variable *sex* equals *'male'*. If a particular observation

gives rise to a very large residual as the one encircled in figure A.4, then that observation would also inflate the estimate $\hat{\sigma}$ and hence masking the effect of the contamination if the standardized residuals are used. To circumvent this problem, we will make use of the studentized residuals which are defined in definition 3.13 in [1]. Here we scale the residuals with an estimate of σ^2 that does not include the i'th observation as shown in A.2.4.

$$\hat{\varepsilon}_i^{stud} = \frac{\hat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2_{(i)}\,(1-h_{ii})}} \tag{A.2.4}$$

We know from section 3.9 in [1] that the studentized residuals follow a t-distribution with $Df = n - m_0 - 1$ degrees of freedom. Hence we have in figure A.5 plotted the studentized residuals in a qq-plot and against the fitted values.

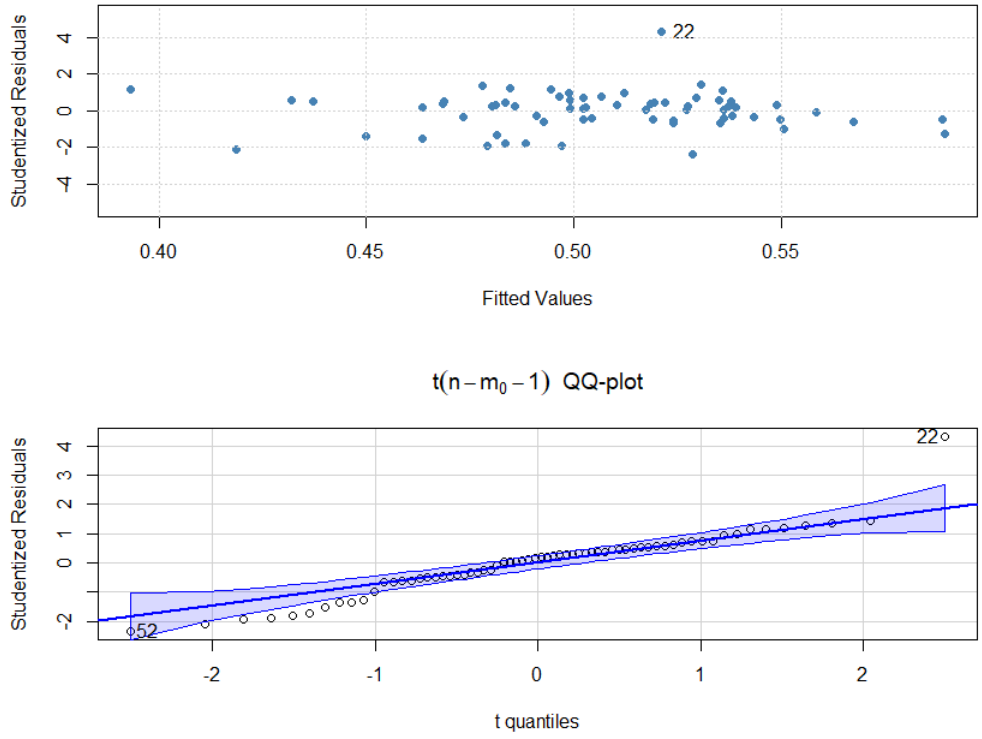

Figure A.5 – The studentized residuals of A.2.3 where the variable *sex* equals *'male'* are plotted against the fitted values and the theoretical quantiles

We see from figure A.5 that the observation 22 is very extreme and we must pay special attention to this point for the further analysis.

Backward selection

With the assumed sufficient model, we can now reduce the model until only necessary terms are left. From the experimental setting, we have no natural chain of hypothesis and hence it would not be natural to choose a type I partition. We have on the basis of lecture 4 chosen to remove the most complex terms before simpler terms. Therefore, it would be natural to choose a type II partitioning using a significance level of 0.05. Here we will display the interactions we remove for each iteration explicitly:

Iteration 1: tOut:tInOp:factor(sex)

Iteration 2: tOut:tInOp

Iteration 3: tOut:factor(sex)

All terms a now significant in our type II test, and the model is:

$$\text{clo} \sim \text{tOut} + \text{tInOp} + \text{factor(sex)} + \text{tInOp:factor(sex)} \tag{A.2.5}$$

Note: in equation A.3.2 we write out the formula explicitly. Above we simple state the syntax used in ***R****.*
We first consult the qq-plot to asses the assumption of normality of the residuals.

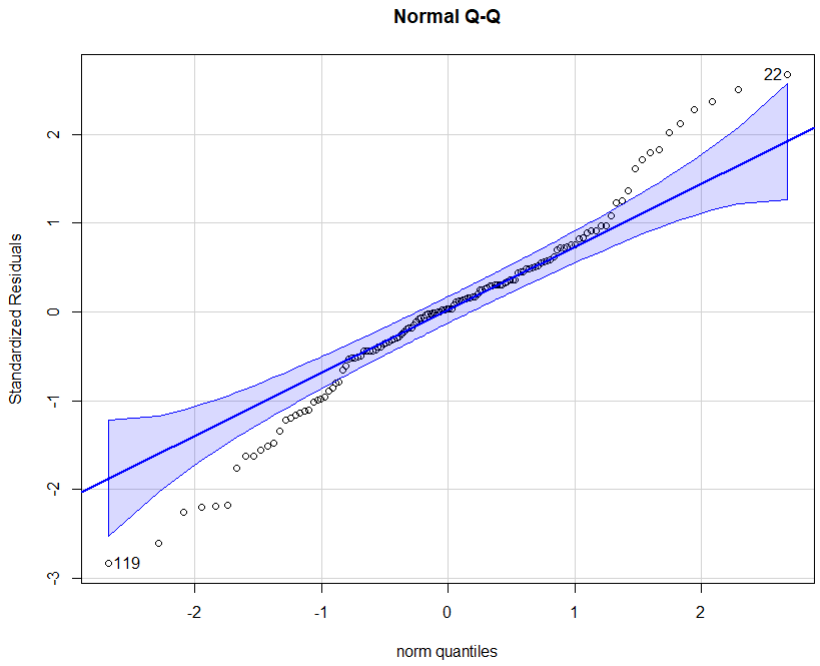

Figure A.6 – QQ-plot of the residuals of A.2.5.

In figure A.6, we see that the normal qq plot has not improved significantly compared to the qq-plot for more advanced model given in figure A.3. We now check for model deficiencies by analyzing the residuals against the fitted and explanatory variables:

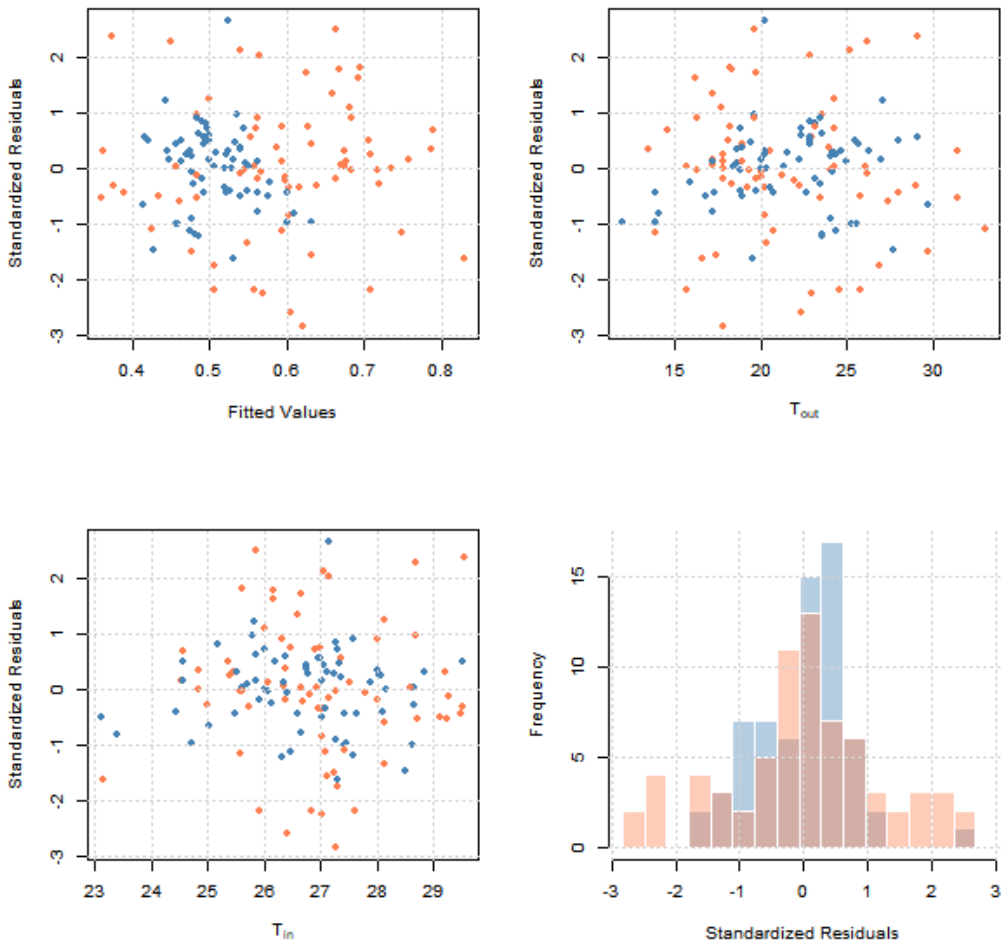

Figure A.7 – Model deficiency figures to asses if the how variance is explained by independent variables.

In figure A.7, it is evident that there is some additional variance for the females that the model is unable describe as we also concluded in section A.2.1. This is very evident in the histogram of the standardized residuals where the females have greater variance. Likewise we see in the plots with fitted values and T_{out}, that the variance seems greater for the females. The outlier pointed out in section A.2.1 is still apparent in all figures.

A.3 Interpretation of the model

The model in equation A.2.5 is as mentioned written in **R** syntax. We will therefore here write out the model in a more classic notation. Let $y_{i,j}$ be the level of clothing for observation i with gender j where $j \in \{0, 1\}$ with 0 coding for female and 1 for male. Introduce the indicator:

$$\mathbb{1}_{\{j=g\}} = \begin{cases} \textit{'female'} & \text{for } g = 0 \\ \textit{'male'} & \text{for } g = 1 \end{cases} \tag{A.3.1}$$

we can write our model as:

$$y_{i,j} = \mu + \alpha\, \mathbb{1}_{\{j=1\}} + \beta_1\, x_{i,j}^{tOut} + \beta_2\, x_{i,j}^{tInOp} + \beta_3\, \mathbb{1}_{j=1} x_{i,j}^{tInOp} + \varepsilon_{i,j}, \quad \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2) \tag{A.3.2}$$

If we sort the observations such that:

$$\mathbf{y} = [\underbrace{y_{1,j=0}, y_{2,j=0}, \ldots, y_{m,j=0}}_{female}, \underbrace{y_{m+1,j=1}, \ldots, y_{N,j=1}}_{male}]^{\intercal} \tag{A.3.3}$$

We can write A.3.2 in matrix notation:

$$\begin{bmatrix} y_{\{1,j=0\}} \\ \vdots \\ y_{\{m,j=0\}} \\ y_{\{m+1,j=0\}} \\ \vdots \\ y_{\{N,j=0\}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_{\{1,j=0\}}^{tOut} & x_{\{1,j=0\}}^{tInOp} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{\{m,j=0\}}^{tOut} & x_{\{m,j=0\}}^{tInOp} & 0 \\ 1 & 1 & x_{\{m+1,j=1\}}^{tOut} & x_{\{m+1,j=1\}}^{tInOp} & x_{\{m+1,j=1\}}^{tInOp} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & x_{\{N,j=1\}}^{tOut} & x_{\{m+1,j=1\}}^{tInOp} & x_{\{m+1,j=1\}}^{tInOp} \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{\{1,j=0\}} \\ \vdots \\ \varepsilon_{\{m,j=0\}} \\ \varepsilon_{\{m+1,j=0\}} \\ \vdots \\ \varepsilon_{\{N,j=0\}} \end{bmatrix}$$

We now see directly how we can apply the normal equation A.2.2 to calculate the parameters. The confidence interval of the mean value estimate is given as equation (3.59) in [1] and here stated in A.3.4.

$$\widehat{\beta}_j \pm t_{1-\alpha/2}\,(n - m_0)\,\widehat{\sigma}\sqrt{\left(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}_{jj}} \tag{A.3.4}$$

where the covariance structure, Σ is still given as $\boldsymbol{I}$. When we do this, we obtain the intervals.

	2.5%	$\hat{\beta}$	97.5%
μ	1.5055	2.1324	2.7593
α	−0.0182	−0.0122	−0.0062
β_1	−0.0730	−0.0475	−0.0220
β_2	−2.1639	−1.2834	−0.4030
β_3	0.0118	0.0446	0.0774

(A.3.5)

The interpretations of the parameters are:

- μ: the average level of clothing for females
- α: the difference between the average level of clothing for females and males. As $\alpha < 0$, males tend to wear less than females in the experiment.
- β_1: how much one temperature increase outside affects the the level of clothing for both sexes. $\beta_1 < 0$ hence when it gets warmer, less cloth is worn.
- β_2: the amount an increase in the indoor temperature affects the level of clothing. As $\beta_2 < 0$ and $|\beta_1| < |\beta_2|$ we see that an increase in the temperature inside tends to mean more for the level of clothing.

- β_3: adjusts the reaction of the males to an increase in indoor temperatures compared to females. As $\beta_2 + \beta_3 = -0.0475 + 0.0446 = -0.00289$, we see that the indoor temperature should affect the females more than the males.

To visualize the interpretation of the slope parameteres above, consider figure A.8.

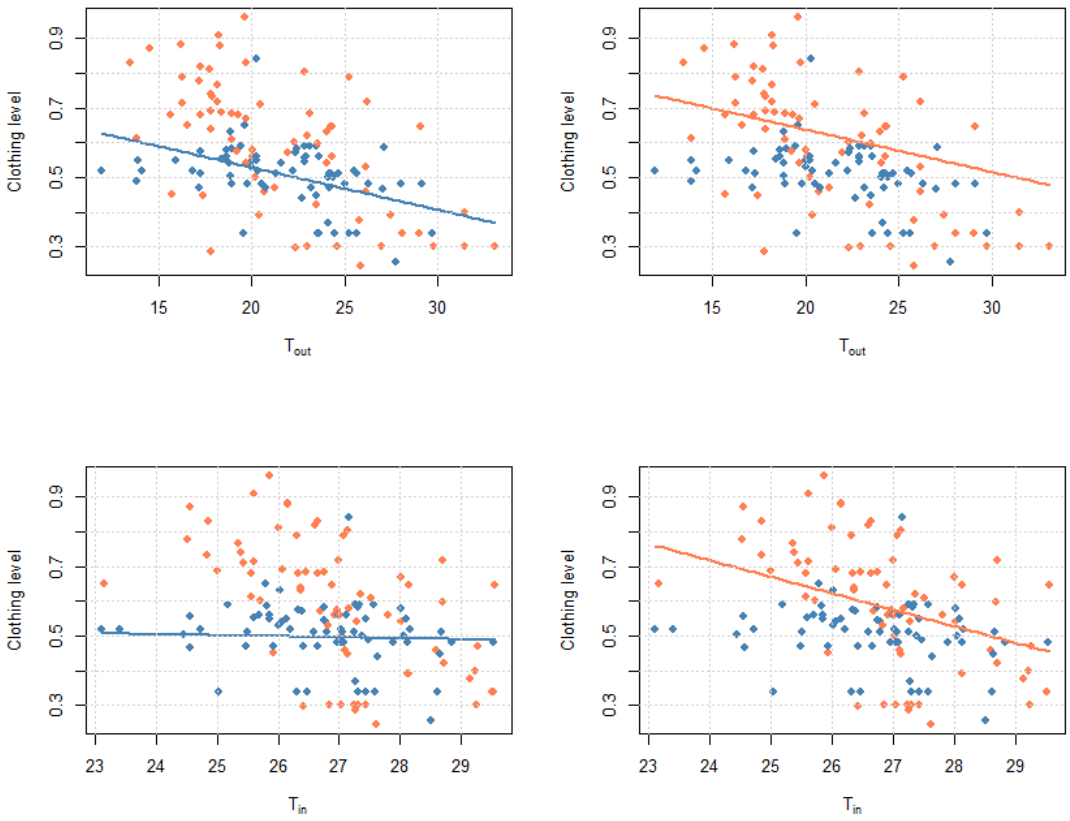

Figure A.8 – Figures to graphical interpretation of parameters. We fix one temperature and consider the isolated effect of the other i.e. the upper left slope is constructed using the average of the observed T_{in} temperatures.

In the upper plots of figure A.8, we see the effect of different outdoor temperatures for males and females. For the outdoor temperature, the sexes have a common slope, however, the sexes have a different mean values hence the vertical shift. In the lower panel we see how much the indoor temperature affects the level of clothing. As hinted above, the indoor temperature affects the females substantially while only minor effects are seen on the males.

A.4 Weighted Analysis

In our model building, section A.2, we noticed in figure A.4 that the factor variable *sex* seemed to have different variances dependent on level. This we also tested in a 1-way between-groups ANOVA and proved that the difference in variance was significant. Therefore, we will perform a weighted analysis where we assume different variances for each sex.

As in section A.3, we will assume the data is sorted. We can then present the chosen covariance structure as a diagonal matrix as the one given in A.4.1.

$$\Sigma = \begin{bmatrix} \sigma_{'female'} & 0 & \cdots & 0 \\ 0 & \ddots & & \\ & & \sigma_{'female'} & \vdots \\ \vdots & & & \\ & & \sigma_{'male'} & \\ & & & \ddots \quad 0 \\ 0 & \cdots & 0 & \sigma_{'male'} \end{bmatrix} \tag{A.4.1}$$

To estimate A.4.1 we will use the **R** library `nlme` which contains the function `gls`, which stands for generalized least squares. This function allows us to specify the covariance structure A.4.1 by passing the argument `weights = varIndent(form=~|factor(sex))`. One could also have used the MLE for the variance given in theorem 3.5 in [1]. If we compare the standard deviation estimated from `gls` and the MLE we get a small difference as shown in table A.3.

	`gls`	MLE
$\frac{\sigma_{'male'}}{\sigma_{'female'}}$	0.5879	0.5948

Table A.3 – Caption

The reason for this small difference is that the MLE for the variance is a biased estimate as shown in example 2.10 in [1]. The estimate from `gls` is obtained by use of restricted MLE which takes care of this bias.
The new parameters are stated in A.4.2.

	2.5%	$\hat{\beta}$	97.5%
μ	1.4889	2.2166	2.9442
α	−0.0153	−0.0102	−0.0051
β_1	−0.0807	−0.0522	−0.0238
β_2	−2.2148	−1.3649	−0.5150
β_3	0.0160	0.0476	0.0792

(A.4.2)

A.4.1 Analyzing the result of a weighted estimate

We now refit the model with the same structure as in section A.2. We then try to redo the 1-way between-groups ANOVA on the factor variable *sex*. We see from table A.4 that the residuals are actually not homoscedastic. We hence try to remove observation 22 which we detected in the outlier analysis in section A.2. From table A.4 we see that now the residuals of the factor variable *sex* are homoscedastic.
We also see from figure A.9 that if one ignores observation 22 the two sexes semes to have a equal variance.

	Without	With
p-value	0.0511	0.0223

Table A.4 – 1-way between-groups ANOVA on the factor variable *sex* with and without observation 22.

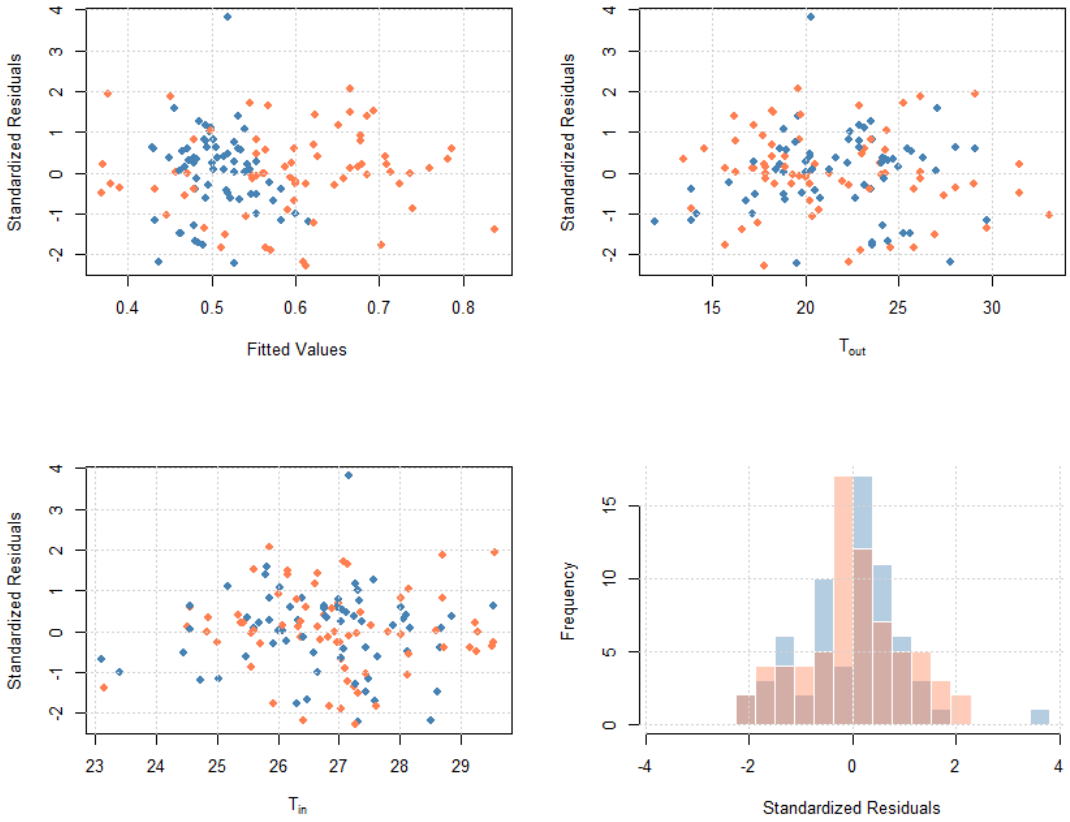

Figure A.9 – Residuals for the weighted model plotted against fitted and explanatory values with observation 22.

From figure A.10 we also see that the qq-plot for the weighted model without observation 22 is a bit better. We would inform the persons who collected the data that there could be a potential outlier but we keep it when we fit this final model.

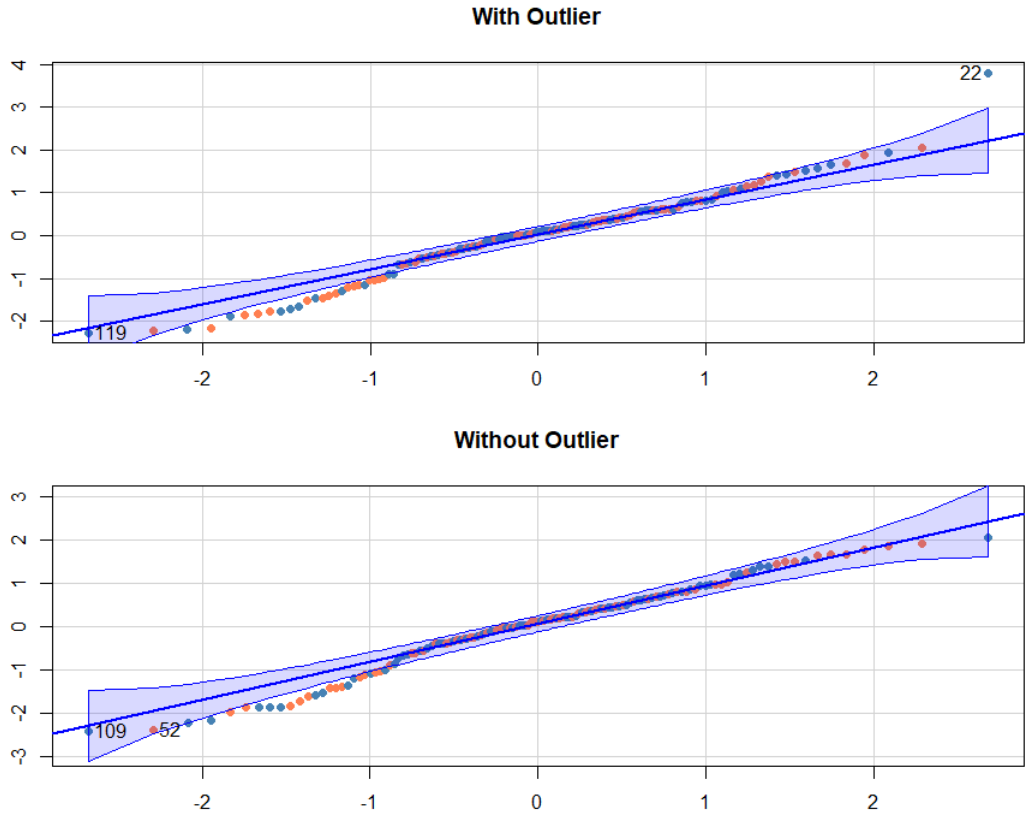

Figure A.10 – QQ-plots for the weighted model with and without observation 22.

A.5 Final Model

On the basis of the previous sections we have concluded on the following model.

$$y_{i,j} = \mu + \alpha\, \mathbb{1}_{\{j=1\}} + \beta_1\, x_{i,j}^{tOut} + \beta_2\, x_{i,j}^{tInOp} + \beta_3\, \mathbb{1}_{j=1} x_{i,j}^{tInOp} + \varepsilon_{i,j}, \quad \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2) \tag{A.5.1}$$

Our model has two continuous explanatory values and hence one total plot would result in a 3D-plot with confidence and prediction 'tubes' around the slopes for males and females. This would be vary hard to take readings off and hence we have decided to split up the regression lines as in figure A.8. The justification for this split is in the linearity of the model. Because it is a linear model shifts in either x^{tOut} or x^{tInOp} would not change the model qualitatively and hence no information is lost in splitting up the full 3D plot.
We have fixed x^{tOut} to be the sample mean when we plot x^{tInOp} and vice versa. The formulas used to calculate the confidence and prediction intervals in figure A.11, can be found in section 3.8 in [1].

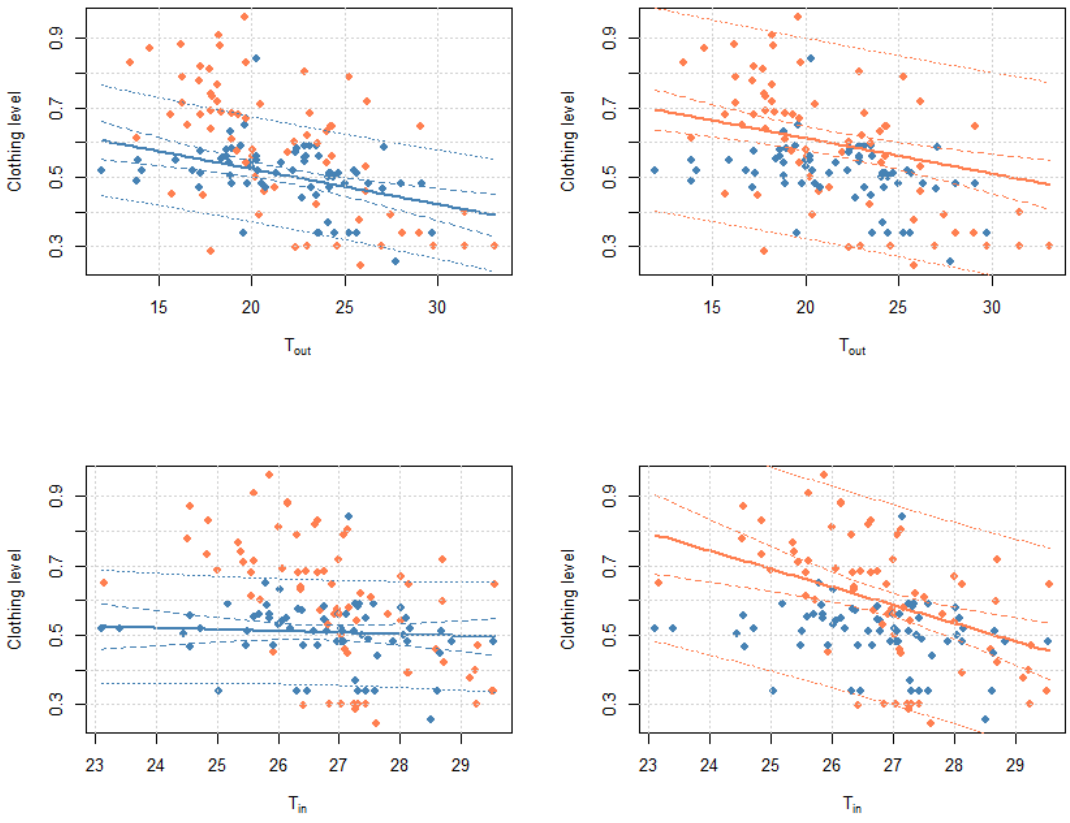

Figure A.11 – Confidence and prediction intervals for the final model A.5.1

In figure A.11 we see a very similar story as in section A.3. In figure A.11 we see a very similar story as the one told in section A.3. We see approximately 5% of observations lies outside the prediction intervals which are as it should be.

A.6 Inclusion of subject ID

To check if we *subject ID* can be ignored, we include the intercept for each subject id such that now the model is

$$\text{clo} \sim \text{tOut} + \text{tInOp} + \text{factor(sex)} + \text{factor(subjId)} + \text{tInOp:factor(sex)} \tag{A.6.1}$$

The model in A.6.1 has 51 parameters. We can test if this model can be simplified to be that of equation A.2.5 using type I partitioning. The model in equation A.2.5 has 5 parameters and using the residuals for each model, we can create the ANOVA table:

Source	f	Deviance	F-Test	p-value
Model versus hypothesis	45	1.1776	3.2212	1.576e-06
Residual under model	86	0.69864		

The above F statistics has a p-value of $1.576e-06$ hence with a significance level of 0.05, we would reject $\mathcal{H}_1$, meaning we cannot ignore the effect of *subject ID*.
A interesting thing regarding A.6.1 is the one of the parameters for *subject ID* is set to NA by **R**. The reason for this is that we also include the variable *sex* for which the effect can be fully described by *subject ID*.

B | Including subject Id

B.1 Include subject Id as explanatory variable

In section A.6, we included *subject ID* in the most simple way with a mean value for each *subject ID*. We have 47 *subject IDs* and 136 data points. As the number of parameters must be less than the number of data points, we can only consider a limited set of model structures. We start with the model:

$$\text{clo} \sim \text{tOut} * \text{tInOp} * \text{sex} + \text{subjId} \tag{B.1.1}$$

In figure E.1, we depict 4 plots to asses model assumptions, however, most noticeably is the qq-plot which can be seen below in figure B.1. It is seen that the qq-plot overall looks fine but with one outlier at point 47.

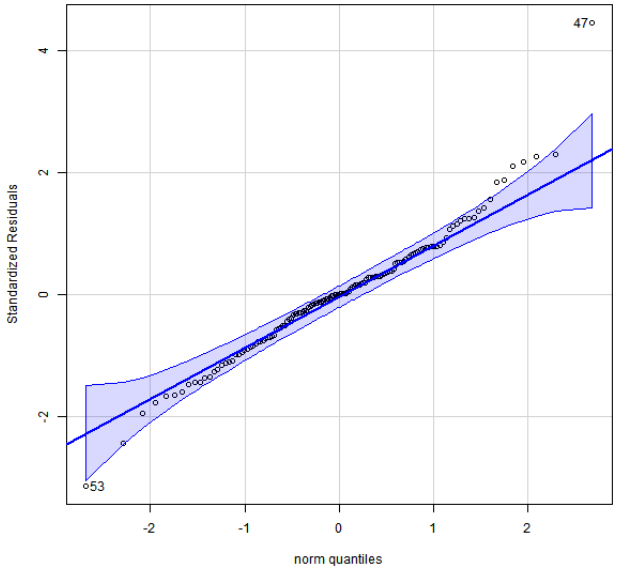

Figure B.1 – Qq-plot to asses the assumption of normality for the model in equation B.1.1.

Point 47 is also evident when we plot standardized residuals against explanatory and fitted values in figure B.2.

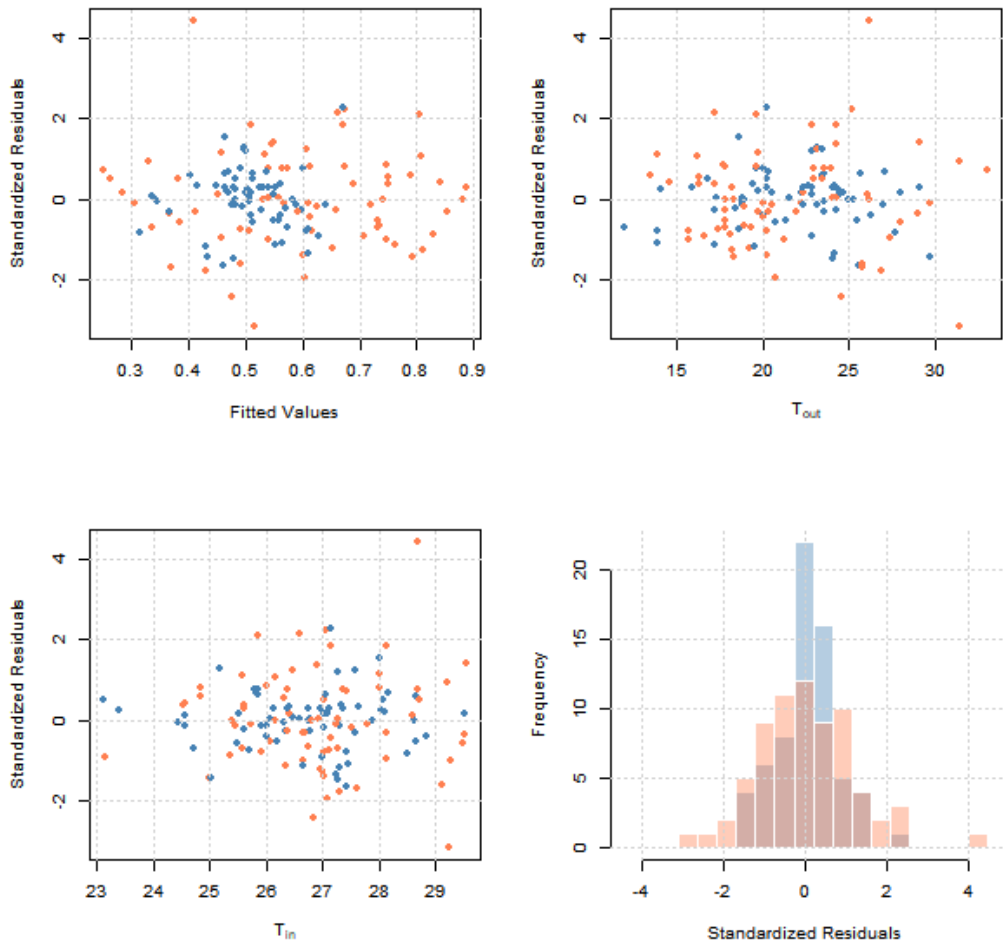

Figure B.2 – Model deficiencies for the model in equation B.1.1.

We also see from figure B.2, that the variance for each sex appears to be the same and the residuals seems to be reasonable. This makes sense because the variable *sex* does not really add any value anymore because we can describe every person's individual traits through *subject ID*. Hence we can conclude that the model B.1.1 fulfills the requirements listed in section A.2.1 and we can proceed with the model reduction.

B.1.1 Model Reduction

Using type II partitioning, we can reduce this model to:

$$\text{clo} \sim \text{tOut} + \text{subjId} \tag{B.1.2}$$

Intuitively, it makes sense that `subjId` is a great predictor. After all, if we know mean value on an individual basis then sex becomes redundant and insignificant. We decide to explicitly remove the intercept as we find it more intuitive to have an intercept for each *subject ID* instead of the offset from some arbitrary *subject ID*. Therefore, we fit the model:

$$\text{clo} \sim \text{tOut} + \text{subjId} - 1 \tag{B.1.3}$$

For this final model, we have the following residuals

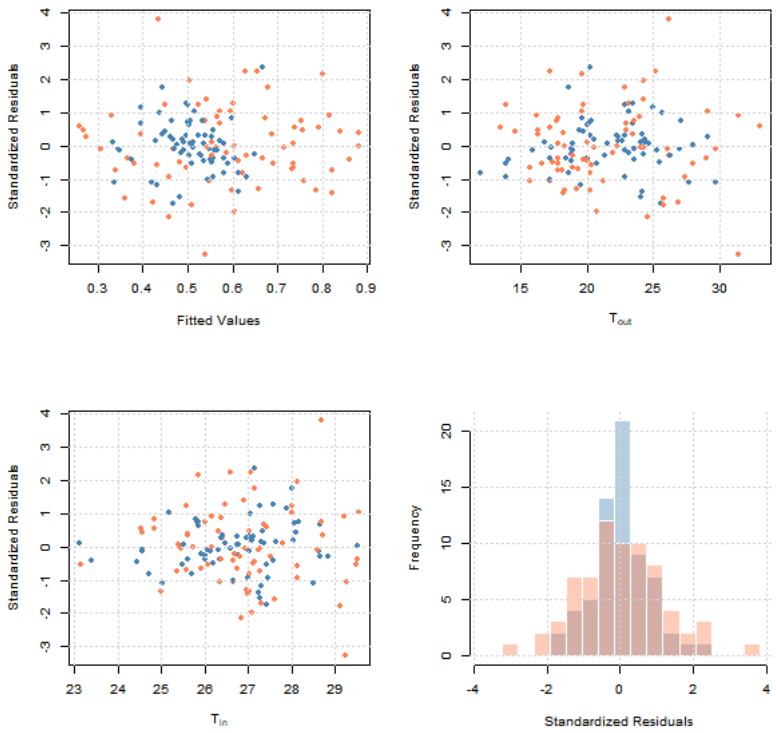

Figure B.3 – Residuals for the final model with subject ID. The model is specified in B.1.3.

In figure B.3, we see that the residuals look reasonable. The females seems to have slightly larger variance but nothing the appears troublesome.

B.2 Visual Representation of Parameters

With the found model in equation B.1.3, we plot the distribution of the mean value for each subject Id:

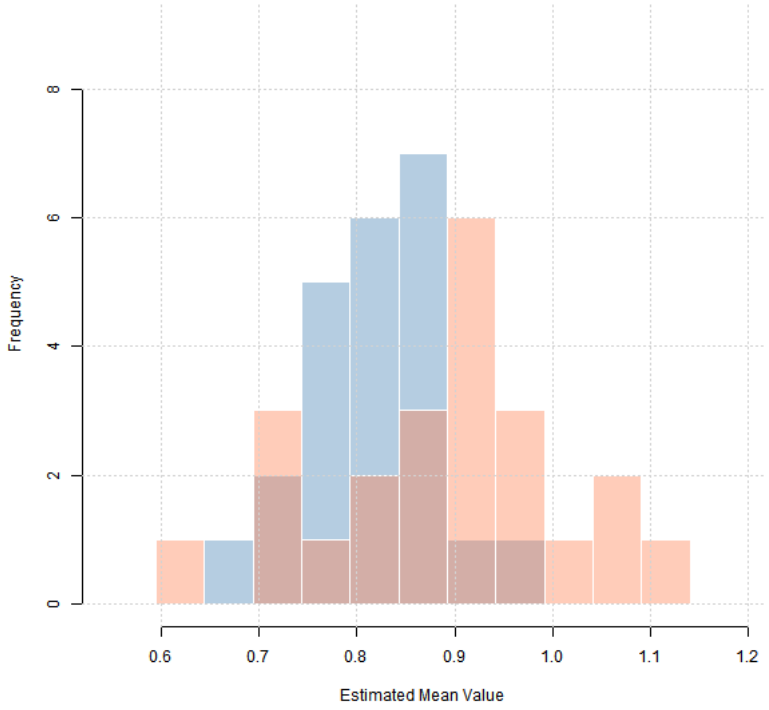

Figure B.4 – Plot for each mean value of each subject Id.

If we neglect the sex, the distribution seems quite normal. If we consider the sex, we see two things. First we notice that that females tends to have a greater level of clothing, i.e. the female parameters are a bit skewed to the right compared with the males. Second we notice that the spread of the female parameters is greater than the spread of the male parameters. From table B.1 we indeed see that the spread of the female parameters is much larger than the spread of the male parameters. This explains why we needed a weighted analysis without *subject ID* in chapter A. This is not needed here because *subject ID* instead takes care of the difference in variance for each sex.

	$\hat{\beta}^{subjId}_{male}$	$\hat{\beta}^{subjId}_{female}$
$\hat{\sigma}$ with observation 47	0.068	0.128
$\hat{\sigma}$ without observation 47	0.068	0.138

Table B.1 – Standard deviation of the parameters for *subject ID* sorted into *males* and *females*.

B.3 Interpretation of Parameters

As noted and displayed in section B.2, we see that each subject Id has it own mean value. This corresponds the level of clothing that each subject wears on average. This should however be considered in relation to the slope parameter $\beta_{tOut} = -0.0143$. This tells us that for each increase in temperature, the subjects will take of 0.0143 level of clothing on average. This model can describe the variation in the data well, however, this is only when we know the *subject ID*. If we where to predict a new *subject ID*, our model would be

flawed. We have found a good model if we have specific information regarding each person but it is not very likely that one would be in the possession of such information. Hence the model is quite useless for prediction of new subjects.

C | The full data-set

C.1 A Model for All Data

We now extend the data set with the variables *day* and *no.obs*. This also extends the number of observations in the data set to 803. We first fit a model to the new data set parameterized the same way as A.2.5 where the observations are weighted by sex. The residuals are shown in figure C.1.

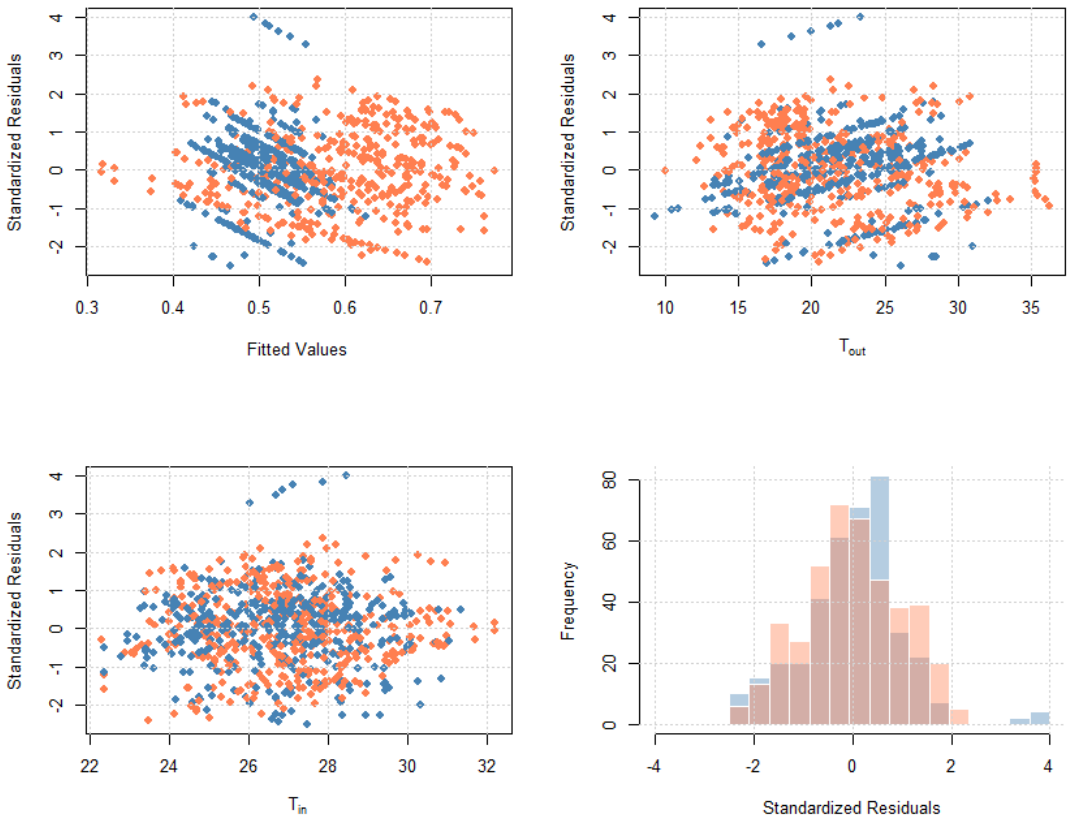

Figure C.1 – Plot for the residuals of a model with the structure found in A.2.5.

Secondly we include *subject ID* and fit a model with the structure found in B.1.3. The residuals are shown in figure C.2

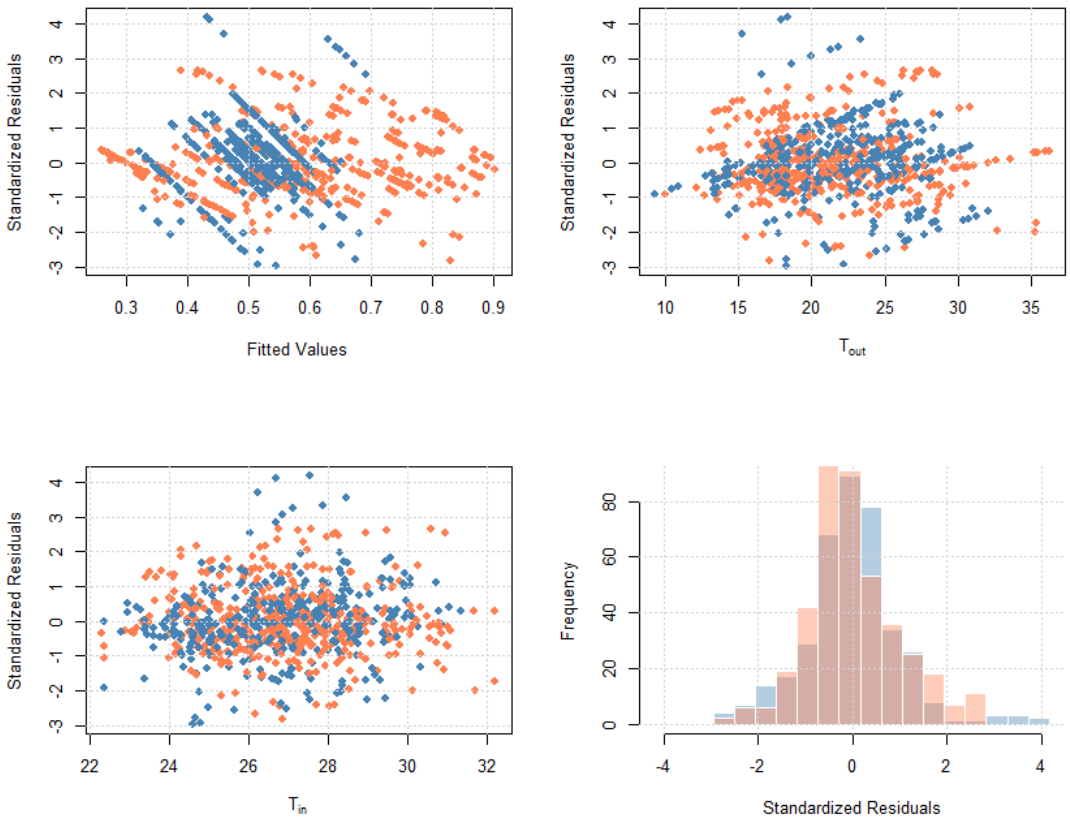

Figure C.2 – Plot for the residuals of a model with the structure found in B.1.3.

We see from both figure C.1 and C.2 that there is a very clear autocorrelation between the days and the independence assumption from section A.2.1 is heavily violated. To backup this conclusion we have plotted the time series for four different *subject IDs* in figure C.3

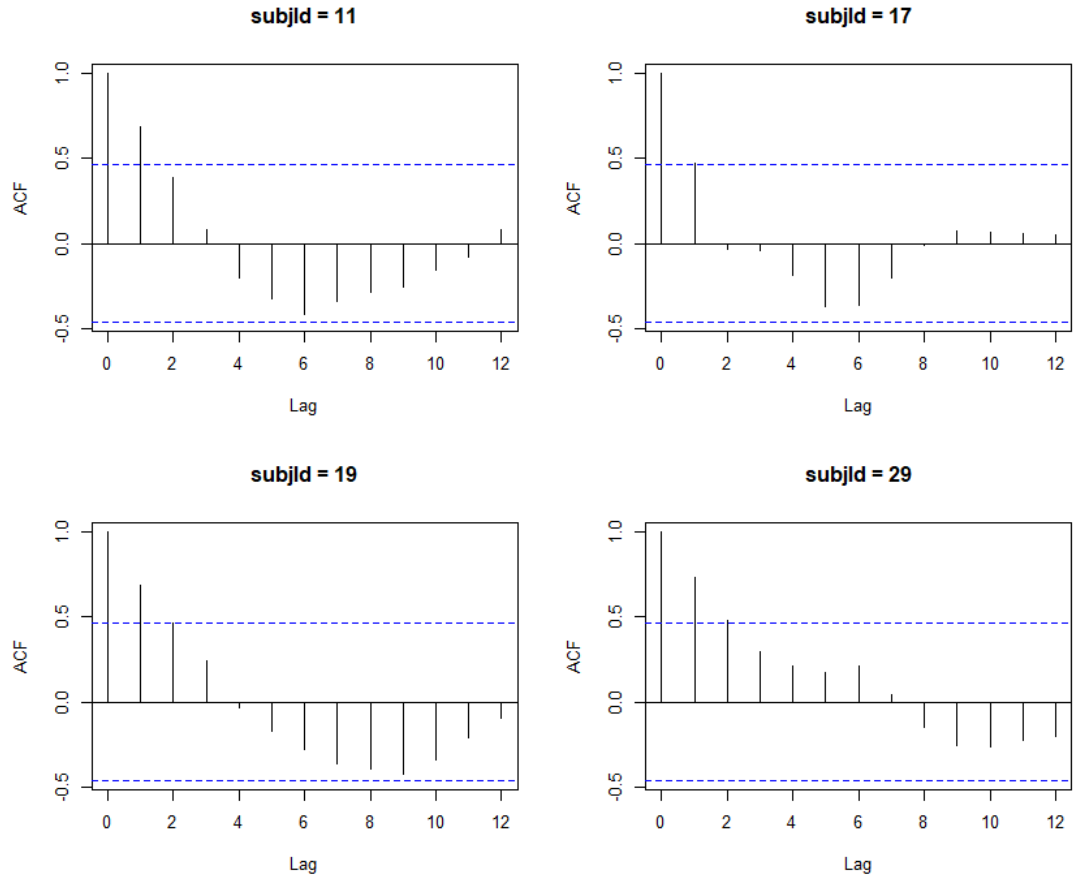

Figure C.3 – ACF plots of the time series for 4 different *subject IDs*

To get some summary statistics regarding the autocorrelation for all the *subject IDs*, we calculate that percentage of all first lags that are significant. For model A the percentage is 53% and for model B is is 61%. The average number of sign shifts in the autocorrelation arrays are 3.84 for model A and 3.31 for model B where the average array length is 17.09 for both. Hence it is very clear that there are a lot of patterns in the residuals that are not described by the model. Therefore to include *day* and *no.obs* one needs to apply some tools from time series analysis to handle the autocorrelation introduced by these variables.

D | Conclusion

We found a reasonable model in exercise A but it had some problems meeting the normality assumption of the residuals.
In exercise B, we added *subject ID* which helped on the normality assumption but introduced a new problem. Now the model is dependent on information which is highly unlikely to be in possession of if we want to use the model for prediction of new subjects.
In C we then extended the data with a *day* and *no.obs* variable. Here it was very clear that it made the residuals highly correlated and hence strongly violated the independence assumption of the residuals.

One thing to notice throughout all the exercises is that using the classic GLM framework to model the clothing level allows us to extrapolate into areas where the clothing level is outside the interval $[0, 1]$. Especially clothing levels below 0 do not make any sense and hence a classic GLM is maybe a poor model choice in the first place and one should probably look towards generalized linear models instead.

E | Appendix

Appendix for B

All figures to asses the model assumption

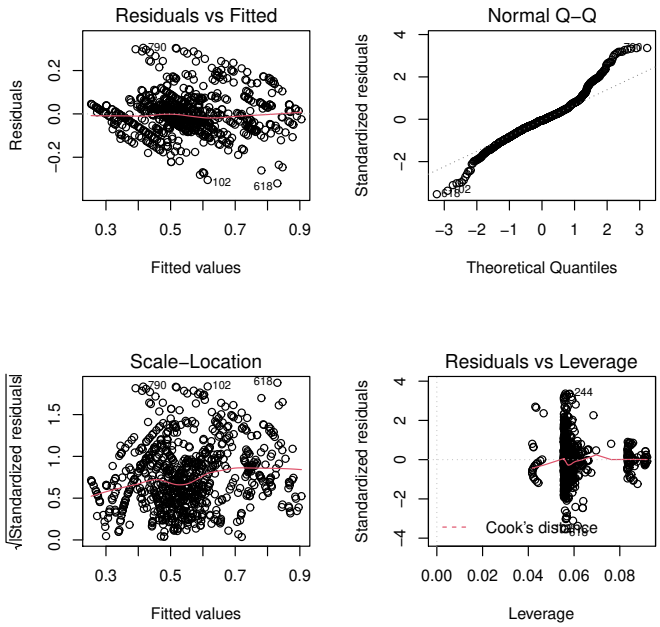

Figure E.1 – Figures to asses the model assumptions of the intial model with subject Id.

Bibliography

[1] H. Madsen and P. Thyregod, *Introduction to general and generalized linear models.* CRC Press, 2010.