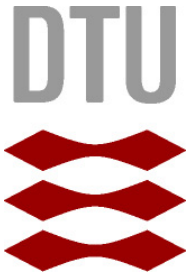
Technical University of Denmark

02424 - Advanced Data Analysis and Statistical Modeling

Project 1

Anders Olsen (s154043)
Patrick Klarskov Jensen (s136392)
Michael Aagaard-Hansen (s961654)



March 13th 2020

1 Introduction

This report summarizes the work done in the first assignment of the course 02424 Advanced Data Analysis and Statistical Modeling, concerning the level of clothing worn in an office as a function of the potential variables outdoor temperature (tOut), indoor operating temperature (tInOp), sex of the subject, subject Id, and day. In the first part of the assignment, modeling is carried out using a small data set with 136 observations and excluding subject Id and day. In the second part, subject ID is included, and in the third part, a larger data set is used. Throughout the report the confidence interval will be the 95% confidence interval unless anything else is stated.

2 Problem A: General Linear Model

This section models clothing insulation based on the indoor temperature, the outdoor temperature, and the sex of the subjects. The subject ID and the day of the observation is therefore not considered in this model.

2.1 Exploratory Analysis

The data is loaded into R and it is found that there are no missing data points. Exploratory analysis is carried out using the *ggpairs* function in R which shows the density plot for each variable, the correlation between continuous variables, and boxplot for factors. Since sex is a factor, either male or female, this function also shows the density plot for each variable for both factors. The data can be seen in Figure 1.

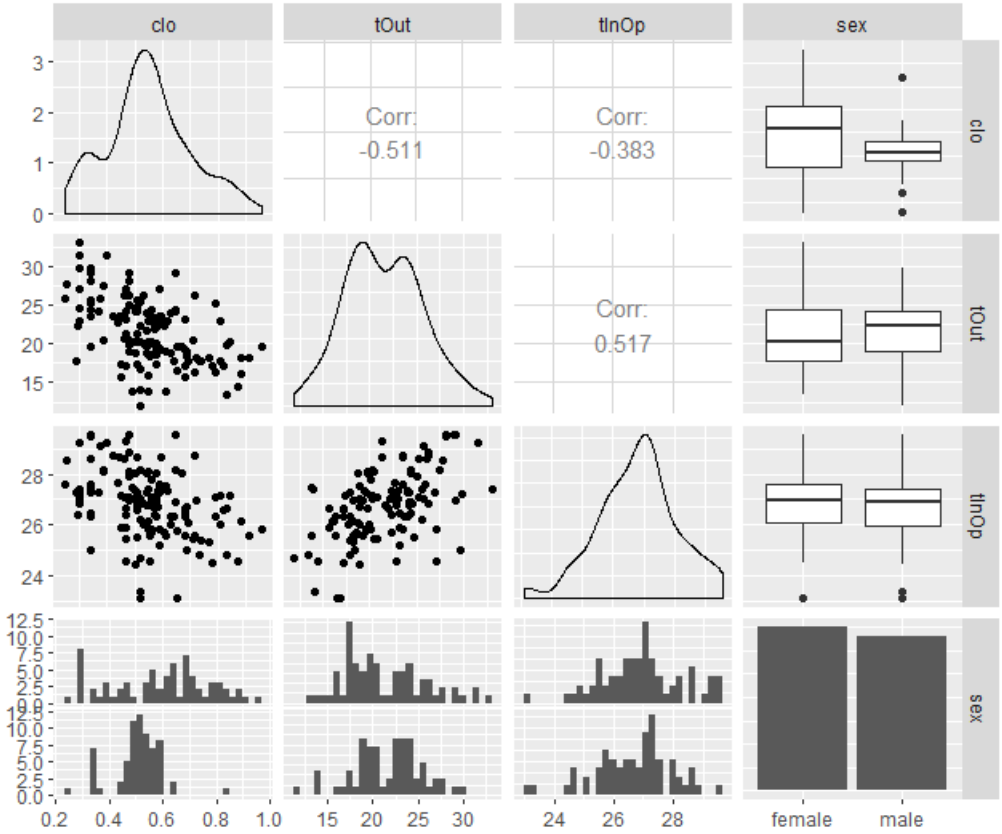


Figure 1: Exploratory analysis of the data

Figure 1 shows that the density functions are close to being normally distributed. However, tOut, which is the outdoor temperature, has two peaks indicating that this distribution might not come from a normal distribution but rather a mix of two normal distribution. The correlation and distribution plot between the variables show no high systematic correlation and are either positive or negative with a correlation value between 0.38 and 0.5. The boxplots show no evidence of significant differences between the two sexes since the boxes overlap. The boxplots show that both males and females were observed at fairly similar indoor and outdoor temperatures, as expected. For the clothing level, however, it is seen that females have a larger variance in the level of clothing insulation and a higher mean, whereas males were in general wearing the same level of less clothing. This is supported by the density distributions of the genders. Furthermore, the investigation shows that there are slightly more females than males represented in the data set.

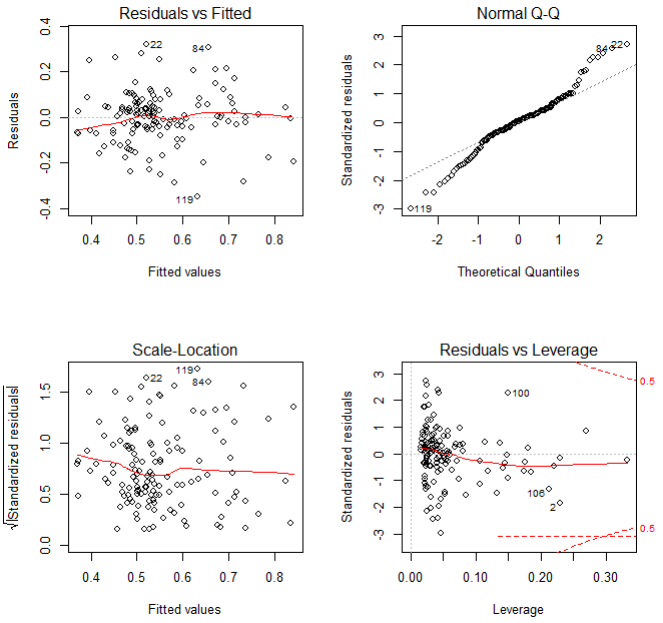
2.2 Find the best model of the data

The response variable *clo* is fitted to a linear model of the variables *tOut*, *tInOp*, and *sex* using the R-function *lm*. The exploratory analysis of the data gave no reason to investigate any of the variables to the second order. However, interactions between the variables were investigated. The first model is the full model where all interactions are included. Backward selection is then used to reduce the model to a more simple form while still retaining accuracy. This is done by removing the least significant variables or interactions and retesting the model until all the variables are significant. A non-significant variable is not removed if the variable is significant in interaction with another variable, and non-significant interactions are removed before non-significant single variables.

The summary and fit of the full model can be seen in Figure 2. It can be seen that sex is only shown as the male in the summary. This is because the female sex is interpreted into the y-intercept and the model checks for significant difference between the intercept(females) and the males. The summary shows that not all of the variables are significant. The investigation of the model is done by looking at the plot. The residuals vs fitted values shows that there is no systematic appearance in the data set which means that there is no hidden trend of the data. The normal Q-Q plot shows that the model is a good fit to some of the data however both tails are fit poorly.

```
Call:
lm(formula = clo ~ tOut * tInOp * sex, data = CS1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34654 -0.05299  0.00547  0.05586  0.31881

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.196897   1.380072   2.316   0.0221 *
tOut               -0.077734   0.064707  -1.201   0.2318
tInOp              -0.082849   0.051361  -1.613   0.1092
sexmale            -3.632267   2.002033  -1.814   0.0720 .
tOut:tInOp          0.002222   0.002357   0.943   0.3475
tOut:sexmale        0.129495   0.092442   1.401   0.1637
tInOp:sexmale       0.125578   0.075416   1.665   0.0983 .
tOut:tInOp:sexmale -0.004498   0.003435  -1.309   0.1927
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1193 on 128 degrees of freedom
Multiple R-squared:  0.4117,    Adjusted R-squared:  0.3796
F-statistic:  12.8 on 7 and 128 DF,  p-value: 2.17e-12
```

(a) Summary of the first model



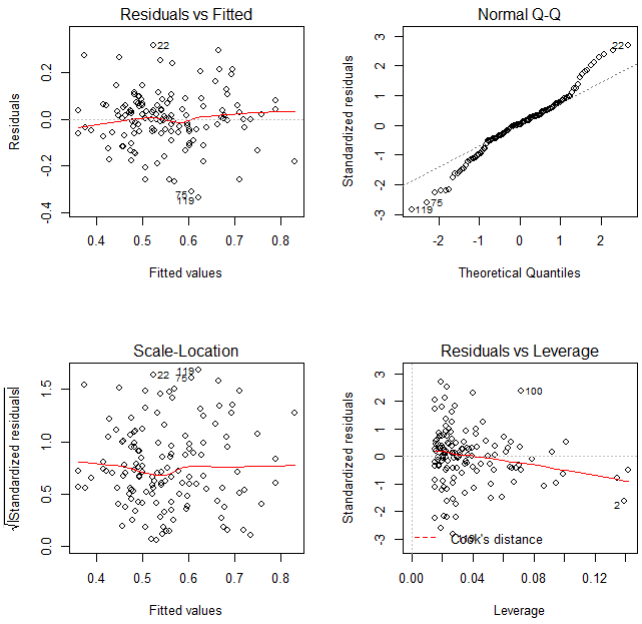
(b) Plot of the final model

Figure 2: Plots of the first model

After backward selection the final model can be seen in Figure 3 where both the summary and the plots can be seen. The confidence interval of each parameter can be seen in the summary of the final model.

```
Call:
lm(formula = clo ~ tOut + tInOp + sex + tInOp:sex, data = CS1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33392 -0.05325  0.00402  0.05806  0.31712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.132395   0.316897   6.729 4.83e-10 ***
tOut        -0.012204   0.003024  -4.036 9.19e-05 ***
tInOp       -0.047494   0.012912  -3.678 0.000342 ***
sexmale     -1.283448   0.445072  -2.884 0.004596 **
tInOp:sexmale  0.044600   0.016563   2.693 0.008013 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1197 on 131 degrees of freedom
Multiple R-squared:  0.3945,    Adjusted R-squared:  0.376
F-statistic: 21.34 on 4 and 131 DF,  p-value: 1.434e-13

> confint(fit4)
                 2.5 %       97.5 %
(Intercept)   1.50549665  2.759294151
tOut         -0.01818621 -0.006222327
tInOp        -0.07303649 -0.021950841
sexmale      -2.16390725 -0.402989584
tInOp:sexmale  0.01183436  0.077366309
```

(a) Summary of the final model



(b) Plot of the final model

Figure 3: Plots of the final model

2.3 Interpretation of the model

The final model includes the three original variables and an interaction term between the temperature outside and sex. The model has an explained R-squared value of roughly 40%. Figure 3a shows that the model will have the following equation where the estimates have been inserted.

$$clo = 2.13 - 0.01tOut - 0.05tInOp - 1.3sexmale + 0.04tInOp : sexmale + \epsilon_i \quad (1)$$

where $\epsilon_i \sim N(0, 0.12^2)$ is independently and identically distributed. Equation 1 shows that if the temperature outside increases by 1, the level of clothing will decrease by 0.01, and if the temperature indoor is increased by 1 the level of clothing will decrease by 0.05. The level of clothing is thus more dependent on the indoor temperature than the outside temperature. The model also shows that male sex will have a lower level of clothing compared to the female. Furthermore, the interaction means that the level of clothing of men will increase by 0.04 when the indoor temperature increases by 1 degree. The indoor temperature is more important for the level of clothing of females whereas the level of clothing for men is equally dependant on the indoor and outdoor temperature due to the interaction between the indoor temperature and male. The model analysis plots can be seen in Figure 3b. The plots show that the data points are evenly spread and the assumption of normal distribution holds. The scale-location and leverage plot shows that there are no extreme observations. The Normal Q-Q plot shows a good fit. However it can be seen that the fit is not perfect as both tails are not fitted on the linear curve. It seems as if there are two distributions on the Q-Q plot.

2.4 Weighted Analysis

In the model, it is assumed that both genders have the same variation for each parameter. However, when looking at the boxplot for clothing level versus gender in Figure 1, it is clear that the variance in the

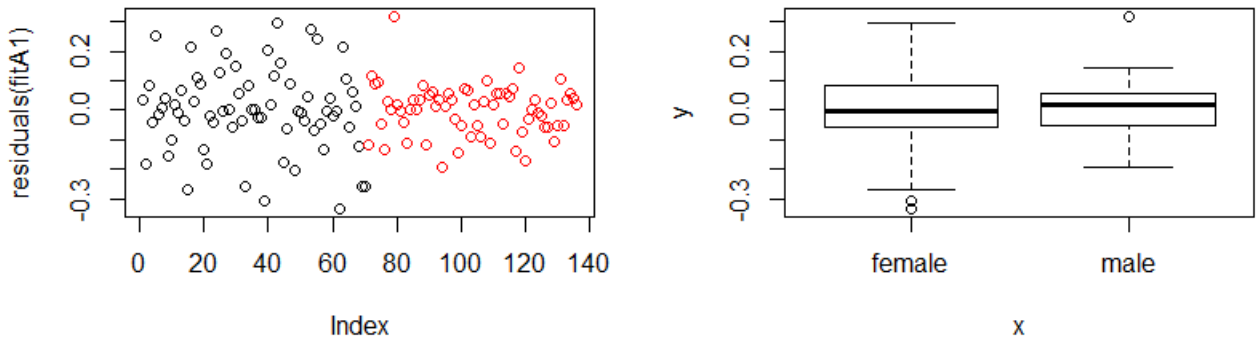


Figure 4: Left: Residuals from the model color-coded according to female (black) and male (red) sex. Right: The same residuals in a boxplot according to sex.

response variable is highly different between men and women. Furthermore, the residuals from the model above, as seen in Figure 4, show a clear difference in variance according to sex. The sample residual variance according to sex is $S^2_{women} = 0.0204$ and $S^2_{men} = 0.00722$, indicating a ratio $\frac{S^2_{women}}{S^2_{men}} = 2.83$.

It is therefore decided to do a weighted analysis of the form

$$f_Y(\mathbf{y}) = \frac{1}{(2\pi)^{k/2}\sigma^k\sqrt{\det \mathbf{\Sigma}}} \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\Big], \tag{2}$$

where $\mathbf{\Sigma}$ is a diagonal matrix of ones for female observations and some number w for the male observations. In the previous model, $\mathbf{\Sigma}$ was the identity matrix. The weight w is defined as the variance ratio between women and men: $w = \frac{\sigma^2_{women}}{\sigma^2_{men}}$, of which the sample variance ratio was calculated to be 2.83. To do this, the same general linear model from before, i.e. using variables *tOut*, *tInOp*, *sex* and an interaction term between *sex* and *tInOp*, was estimated using a weighting w in a range from 0 to 10, and the log-likelihood of the model was estimated for each w. The plot in Figure 5 was achieved. The maximum of this curve is found to be at a weighting $w = 2.93$, i.e. very close to the previously estimated 2.83.

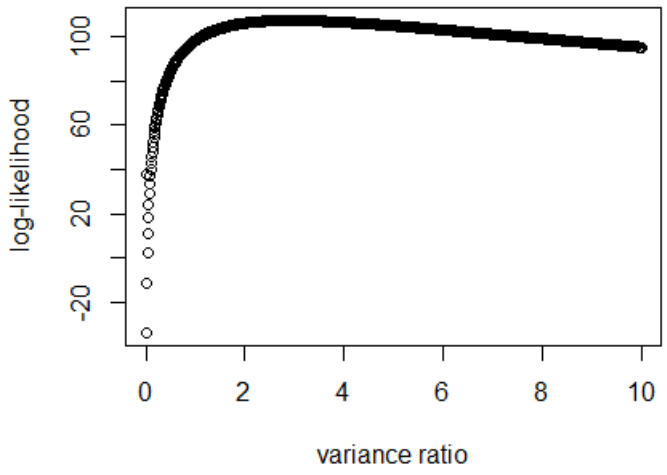


Figure 5: Log-likelihood in a weighted analysis scheme

2.5 The final model

A new model is constructed using the weights obtained in the previous section. In general, p-values go up from the previous model, but upon studying the Q-Q plot, we see a much better fit (Figure 6). One observation stands out, namely observation 79. Upon a closer look, this observation is a male, whose level of clothing level increased with 50% along with a small decrease in outdoor temperature and no difference in indoor temperature. However, this observation is only an outlier in terms of other male observations but does not have an exceptional high leverage compared to the pool of all other observations. Removing this observation would increase the accuracy of the linear fit when looking at the Q-Q plot, but evidence for removal is not deemed large enough.

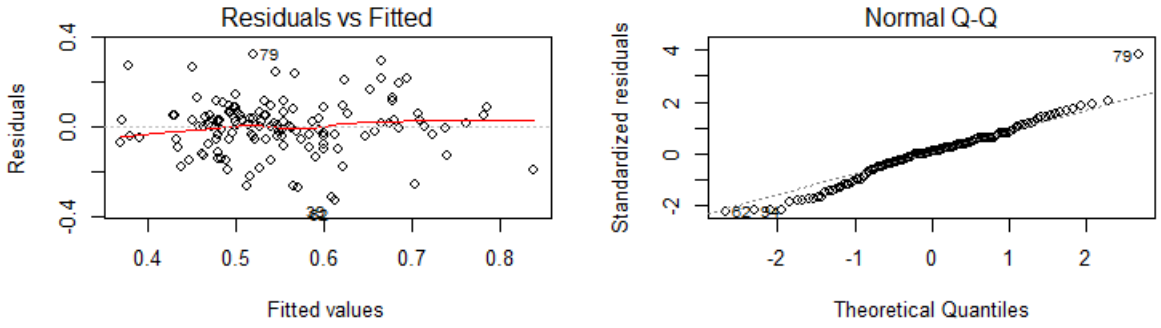


Figure 6: Left: Residuals versus fitted values for the final model. The residuals are not standardized or studentized, so they are exactly the same as in the previous model. Right: Q-Q plot.

The confidence intervals for model parameters are provided in Figure 7. We see that none of the confidence intervals include zero.

$$clo = 2.22 - 0.01tOut - 0.05tInOp - 1.4sexmale + 0.05tInOp : sexmale + \epsilon_i \tag{3}$$

Moreover, we see that the final linear model is the following:

	2.5 %	97.5 %
(Intercept)	1.48779164	2.946913612
tOut	-0.01526629	-0.005099016
tInOp	-0.08077050	-0.023739010
sexmale	-2.21650234	-0.514900768
tInOp:sexmale	0.01600570	0.079227986

Figure 7: Parameter confidence intervals

where $\epsilon_i \sim N(0, 15^2)$ are identical and independent. That is, the model is similar to the one described in Equation 1, albeit with higher variance. A drawback of this model is the fact that the clothing variable is modeled on the real axis. No limits have been stated regarding the response variable, but since all observations are positive, a more reasonable model would utilize generalized linear models, e.g. non-negative transformations.

Finally, a plot of prediction and confidence intervals is included for the outdoor temperature, which was determined to have the highest effect. In Figure 8, all female observations and corresponding intervals are shown to the left, and the male observations to the right. The indoor temperature was fixed to its mean value in both cases. It is clearly seen that the confidence interval show poor prediction of the mean, whereas the prediction interval manages to keep almost all observations within their limits. Again we also see lower variance in the prediction for male observations. The weighting scheme is included in both intervals, which explains that predictions are more precise for males.

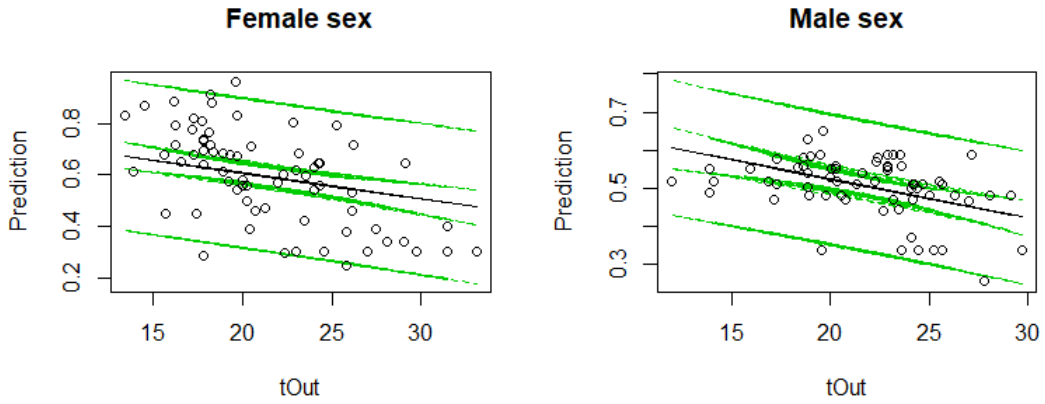


Figure 8: Confidence and prediction interval for female (left) and male observations (right). The outermost green lines show prediction intervals, the innermost green lines show confidence intervals and the black line shows prediction.

2.6 Should subject ID be included?

To determine whether it makes sense to exclude subject ID from the linear model, we include a boxplot of the residuals from the final model above versus subject ID in Figure 9. As seen, both the mean and variance of the residuals vary highly from subject to subject. Therefore, including subject ID would potentially enhance the model to a high degree. In the data so far there are approximately three observations per subject. Of course, including a factor that groups every three observations must exclude other factors. In particular, *sex* is now redundant.

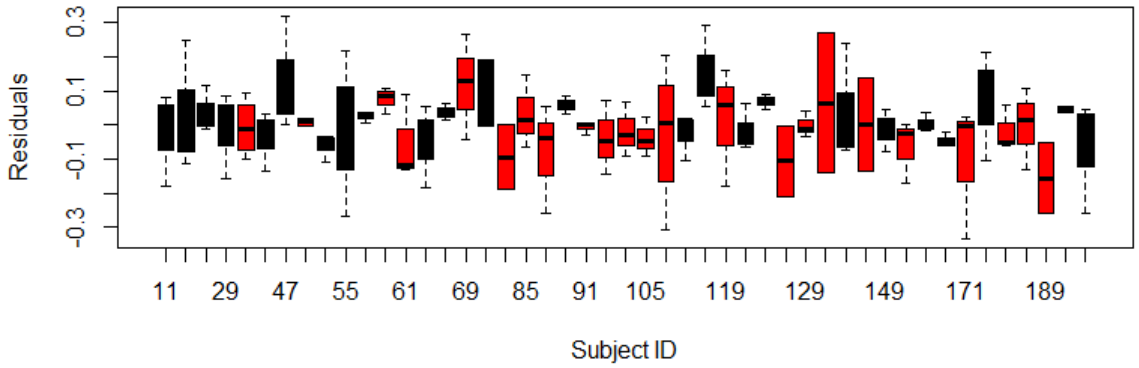


Figure 9: A boxplot of residuals versus subject ID. Black are female, red are male

3 Problem B - Including Subject ID

In the last section, we excluded knowledge of the specific individuals in the general linear model. In this section we include *subjId*, an identifier for the subjects.

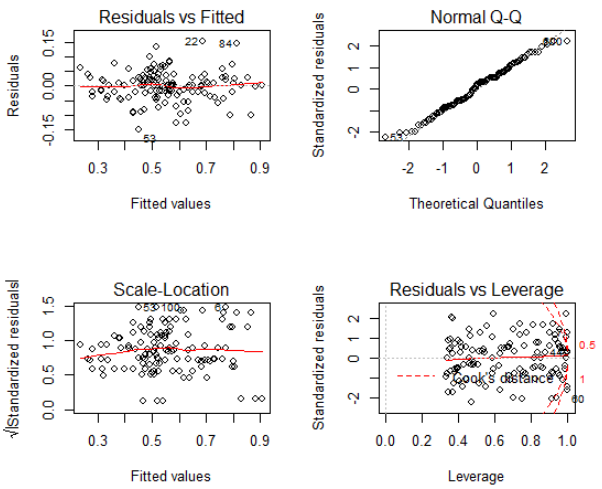
3.1 Develop a GLM including *subjId*

The variable *sex* is a good way to reduce the 47 levels of the variable *subjId* to two, but with *subjId* in the model we remove *sex*. We have already done an exploratory analysis in Problem A, and with 47 levels with two to four observations on each level trying to make sense of *subjId* visually is hopeless. Given that we only have 136 observations we do not have enough data to estimate the parameters in a model with a three-way interaction, so we will start with a model containing the three variables and all two-way interactions using the R-function *lm*.

We get a model with more parameters than observations, which has the problem that some parameters are not estimated at all, and all the parameters estimated have an extremely large standard deviation (all p-values are above 0.4, most significantly higher). Actually, the model itself has a p-value of 0.4288, meaning that the model itself is not significant. Since we have no basis for choosing meaningful contrasts, we will have to choose a new model. In Section 2.3 we found that the interaction between sex and indoor temperature was important, whereas the interaction between sex and outdoor temperature was not. Since the sex variable is a simplification of the subject variable it seems likely that the result from sex carries over to the individual. We therefore remove the interaction between subject Id and outdoor temperature.

```
Call:
lm(formula = clo ~ tOut + tInOp + subjId + tOut:tInOp +
tInOp:subjId, data = CS4)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15047 -0.02274  0.00000  0.02797  0.15261

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -5.997651   6.446740  -0.930    0.358
tOut              0.153182   0.117667   1.302    0.200
tInOp             0.271803   0.249213   1.091    0.282
tOut:tInOp       -0.006471   0.004387  -1.475    0.148
And a lot of factor levels that can not be removed
individually.

Anova Type III

Response: clo
              Sum Sq Df F value Pr(>F)
subjId       0.37019 46  0.9247 0.6033
tInOp:subjId 0.35977 46  0.8987 0.6386


Residual standard error: 0.09329 on 40 degrees of freedom
Multiple R-squared:  0.8877,    Adjusted R-squared:
0.6208
F-statistic: 3.327 on 95 and 40 DF,  p-value: 2.689e-05
```

(a) Statistical summary



(b) Plot of residuals

Figure 10: The initial model utilizing *subjId*

In Figure 10a we see the statistical characteristics of our chosen initial model. The R-function *summary* gives the estimates for the $(47-1)$ level effects and their interactions which is not as interesting and useful as an overall evaluation of the variables since we can exclude variables from the model but not individual observations. The R-function *Anova* can provide a p-value for the variables themselves, which is sufficient for our current purpose.

In Figure 10b we see that the residuals seem to be normally, identically, and independently distributed, and we therefore conclude that the model is sufficient.

We use backward selection exactly as in Problem A. R's *drop1(model)* and *Anova(model,type="III")* both work by comparing the model with and without each individual variable, but *drop1* only displays

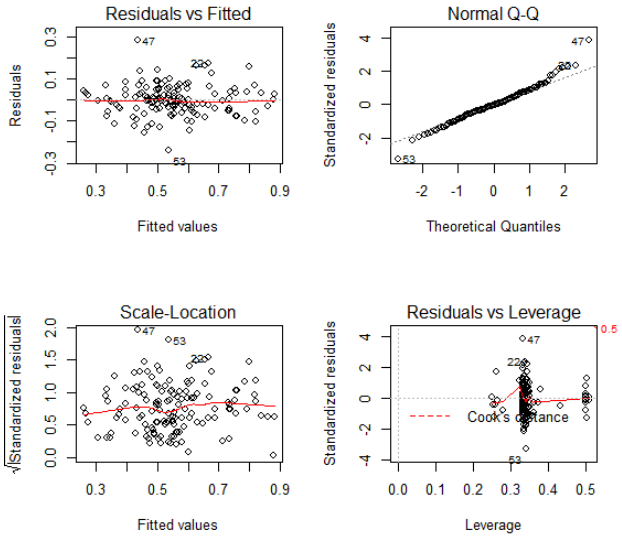
the highest order of interaction, which are the terms we want to remove first in order to be able to interpret the parameters. Using *drop1* we reduce the variables in the order {*tOut*:*subjId*, *tInOp*:*subjId*, *tOut*:*tInOp*, *tInOp*} and end up with the model

$$Y = 1.01 - 0.014 \cdot \text{tOut} + \boldsymbol{\alpha} \cdot \text{subjId} + \epsilon \tag{4}$$

where $\epsilon \sim \text{N}(\mathbf{0}, 0.090^2\mathbf{I})$. In Figure 11a we see the statistical characteristics of our chosen model. As in Figure 10 we have replaced the estimates of the factor intercepts and their confidence intervals with an ANOVA for the terms themselves. In appendix A the full list of parameters with associated confidence intervals can be found.

```
Call:
lm(formula = clo ~ tOut + subjId, data = CS4)

Residuals:
     Min       1Q   Median       3Q      Max
-0.237842 -0.040145 -0.003432  0.039902  0.281815

Coefficients from summary:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  1.006833   0.078165  12.881  < 2e-16 ***
tOut        -0.014275   0.003043  -4.690 9.92e-06 ***
46 levels that can not be excluded individually

ANOVA Type III
Response: clo
          Sum Sq Df  F value    Pr(>F)
subjId    1.57288 46   4.2063 3.688e-09 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09016 on 88 degrees of freedom
Multiple R-squared:  0.7691, Adjusted R-squared:  0.6458
F-statistic: 6.238 on 47 and 88 DF,  p-value: 9.696e-14
```

(a) Summary of the final linear model utilizing *subjId*.



(b) Plot of the final model

Figure 11: The final model utilizing *subjId*

We only have a limited span of outdoor temperatures, so we cannot assess the temperature range in which the linear model is valid. It is therefore assumed that all non-negative predictions of *clo* are true and negative predictions of *clo* are set to 0 (naked, very bad in an office environment :)). This deals with the potential problem of predicted negative clothing levels in a way that is just as valid as using a transformation of *clo*. The advantage is a simpler model that is more intuitive to interpret.

3.2 Examination of the parameters

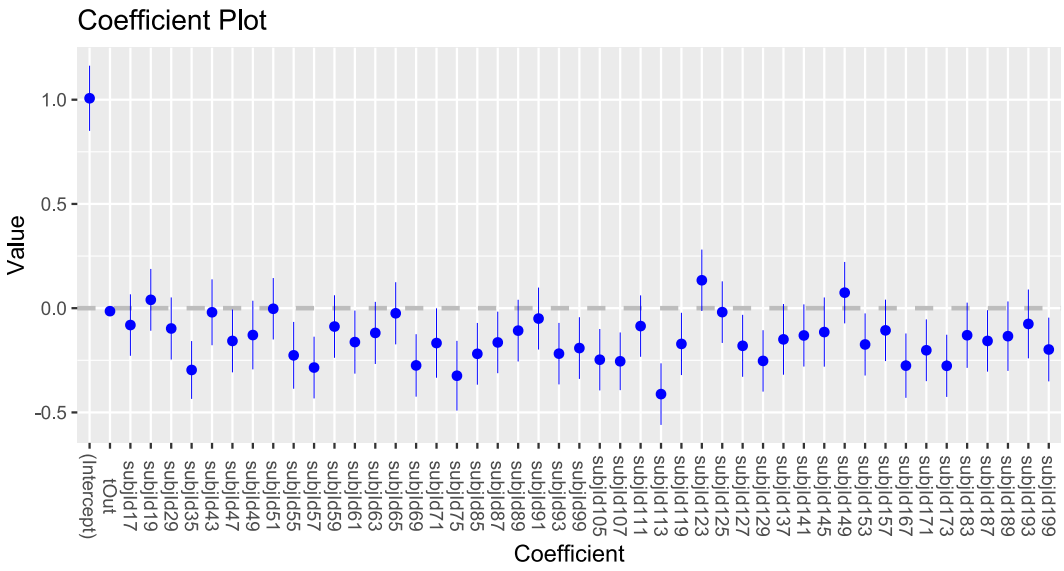


Figure 12: The confidence intervals of the parameters

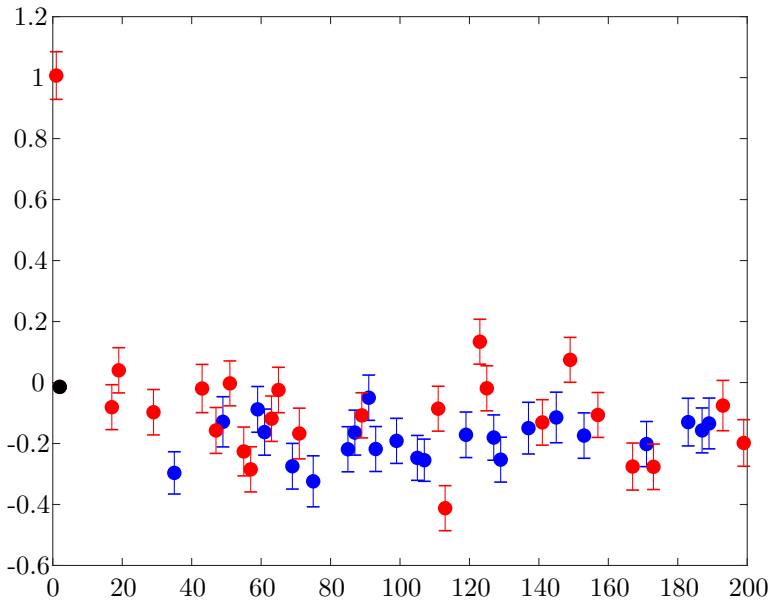


Figure 13: The confidence intervals of the factor coefficients. Men are blue and women are red.

In Figure 12 we see estimates for all parameters and the 95% confidence intervals. The parameter for outdoor temperature is locked down tight, but all the others (intercept includes the factor coefficient for subject 11) suffer from a lack of observations, having only two to four each. From Figure 13 it is obvious that though the estimates vary strongly from individual to individual, the coefficients of the men (blue)

are in general lower than the women (red), and there is a greater variation of the estimates amongst women than men.

Figure 14 predicts the clothing level of each subject. Almost all observations fall well within both confidence- and prediction intervals, which are a bit wide do to having few observations of each subject.

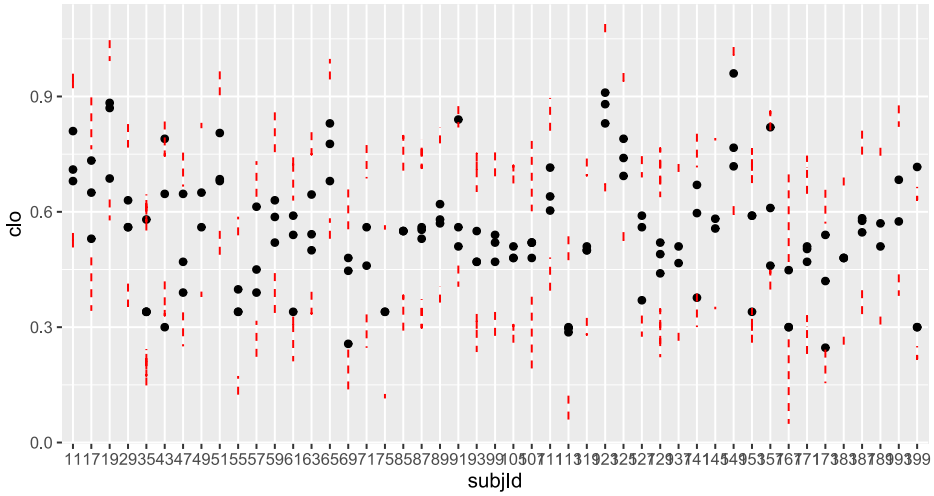


Figure 14: The 95% confidence and 95% prediction intervals for each subject.The dashed lines are outside confidence- but inside prediction interval.

The statistical analysis in this subsection relies heavily on the fact that we have i.i.d. residuals.

3.3 Interpretation

In Problem A we needed five terms in the linear model, one of which was a tedious interaction. By substituting individuals for genders we have managed to need only two variables, while at the same time reducing the standard deviation of the errors from 0.12 to 0.09.

What this model tells us is that each person will dress based on the outdoor temperature and their own personal dress style, with a strong consensus across all individuals on how to take outdoor temperature into account (very small standard error on the slope of $tOut$). In other words, we individually dress for looks with no regard to indoor temperature and then we collectively adjust clothing level according to the outdoor temperature.

From Figure 13 it can be seen that men have a more uniform dress style than women (less variation in factor coefficients) and that they wear less (the mean of the coefficients is lower) than women, which we also saw in Problem A.

This model is great for predicting how a specific individual in the data-set will dress, needing only the outdoor temperature, though the prediction intervals would benefit a good deal from more observations. On the other hand, the model is not good for predicting the clothing level for a new subject, because we do not know their individual preferences. With a bit of effort, we could find an average dress sense across the entire population or the new subject's gender, but in that case, the model from Problem A is more appropriate.

4 Problem C - The Full Data-set

As this is a new data set, the first step is to visualize the data. Figure 15 shows the analysis of the full data set. It was decided that the Subjid should not be tested as this variable was not helpful in prediction of clothing of a random person. Compared to the data set used in problems A and B it can be seen that the distribution of *tOut* is looking more like a normal distribution than it did in the reduced data set where two peaks were found. Just as it was seen in Figure 1, the female population has a much larger variance in the level of clothing they wear compared to males. The correlation and density plots show no systematic behavior which suggests that the variables should not be transformed.

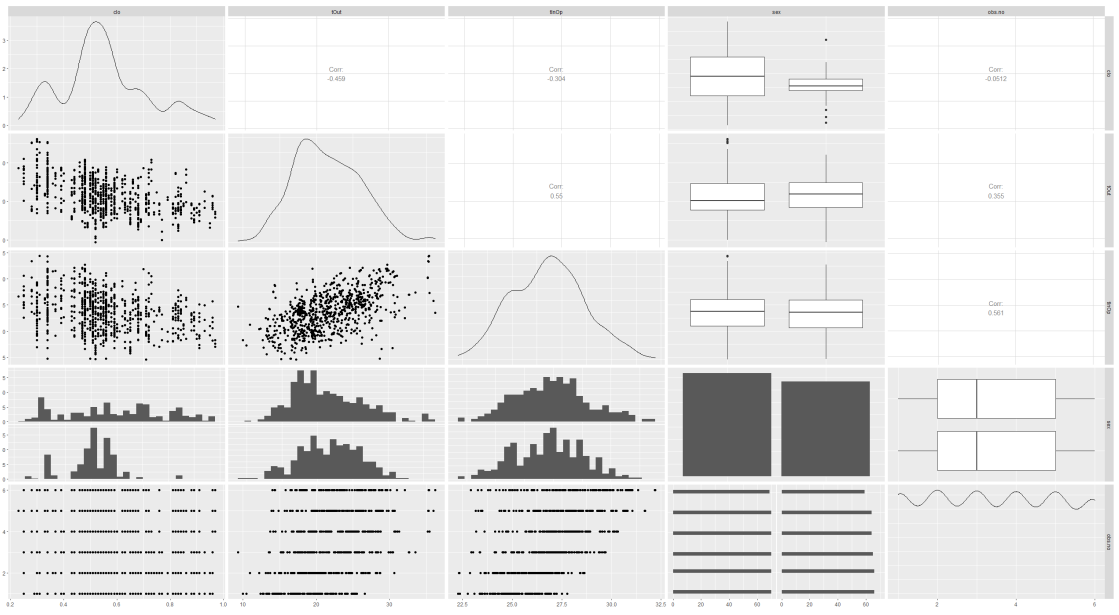


Figure 15: Exploratory analysis of the entire data set

Figure 15 shows that there is a large difference in the variance of the clothing between males and females. This was also seen in analysis of the data used in part A/B. In order to counter this difference, variance weights are calculated by using Equation 2. The relationship between log-likelihood and variance ratio can be seen in Figure 16. The maximum for this curve was found to be $w = 3.67$, which is slightly higher than the weighting found in the smaller data set.

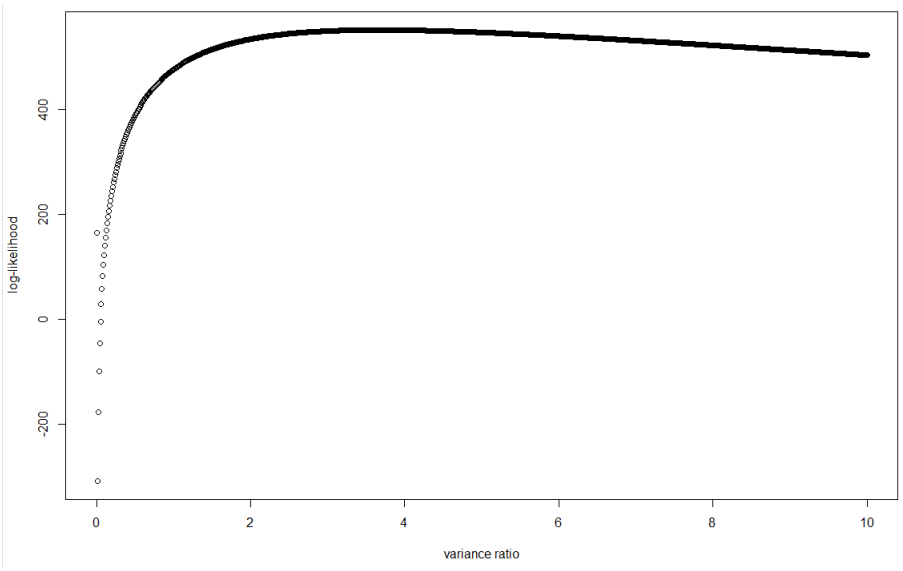
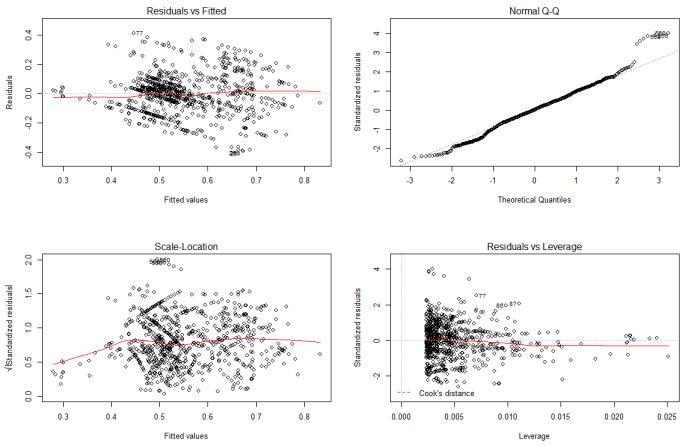


Figure 16: Relationship between the variance ratio and the log-likeihood

With the weights calculated a linear model fitted to the data. Backward selection is used with the highest level of allowed interaction between variables is set to two. Compared to the model found in problem A, the temperature inside is not significant in the model of the entire data set, as was also found in problem B. The summary and analysis of the model of the entire data set can be seen in Figure 17. The model analysis shows that the residuals are evenly distributed and there is no systematic behavior. The distribution plot show that the model is a good fit to the data. However, there are 6 observation which does not fit the linear curve, among these observation 559 and 560. The leverage plot shows that there are no clear outliers, the 559 and 560 observation should therefore not be removed. These two observations belong to the same subject that, compared to the full data set, does not have a large enough leverage to be removed.

```
Call:
lm(formula = clo ~ tOut + sex + tOut:sex, data = CSfull, weights = W)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.43073 -0.09688  0.00187  0.10185  0.66057

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.043280   0.035888  29.071  < 2e-16 ***
tOut         -0.021098   0.001636 -12.892  < 2e-16 ***
sexmale      -0.376856   0.042687  -8.828  < 2e-16 ***
tOut:sexmale  0.013759   0.001941   7.089 2.97e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1653 on 799 degrees of freedom
Multiple R-squared:  0.2745,    Adjusted R-squared:  0.2718
F-statistic: 100.8 on 3 and 799 DF,  p-value: < 2.2e-16
```

(a) Summary of the final model with the full data set



(b) Plots of the final model with the full data set

Figure 17: Summary and plots of the model with the full data set

$$clo = 1.04 - 0.02tOut - 0.38sexmale + 0.01tOut : sexmale + \epsilon_i \tag{5}$$

where $\epsilon_i \sim N(0, 0.17^2)$. From the summary in Figure 17 the parameters of the linear model can be obtained. The estimates can be seen in Equation 5 while the estimate and the confidence interval of the estimates can be seen in Figure 18. This figure shows that the parameters tOut and the interaction between tOut and sex are very significant. However because of the interaction parameter it can be seen that males are less affected by tOut. The model also shows that males are wearing a lower level of clothing than females.

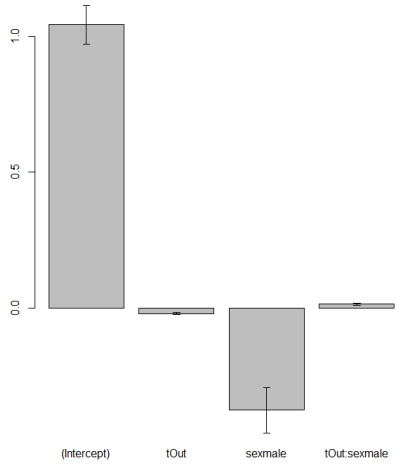


Figure 18: Parameter estimates and their confidence interval

To verify that the model is a good fit to the data the model and the confidence interval are plotted for the observations. The confidence interval model can be seen in Figure 19.

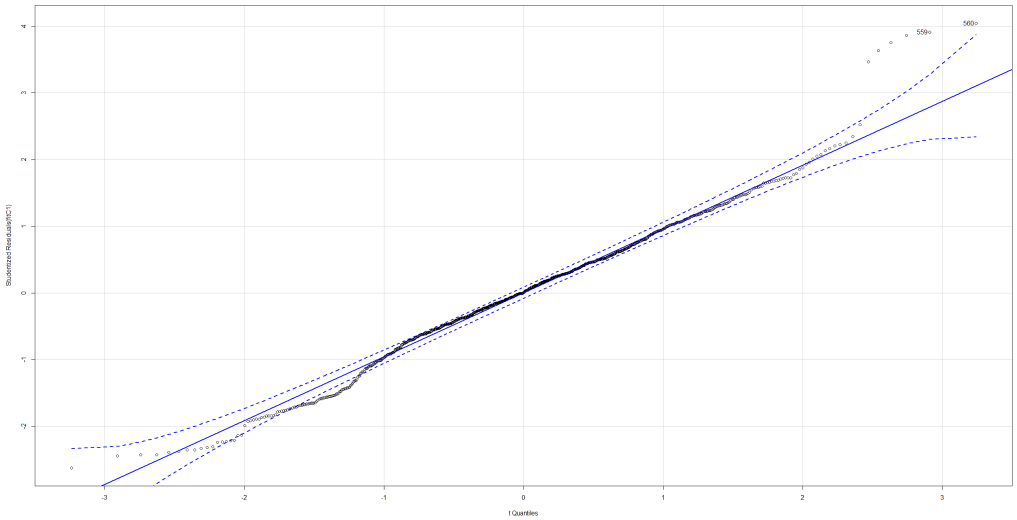


Figure 19: Confidence interval of the model

In Figure 19 it can be seen that the bottom tail has a slightly poor fit, however, there are 6 points which are clear outliers in the top. These are investigated in Figure 21 where it can be seen that the observations are 559 and 560. If these are investigated in the data set it can be seen that these belong to subject 91, which had a significantly higher level of clothing on the first day of observation compared to the rest of the male subjects. The subject did not have a significantly different level of clothing the following days.

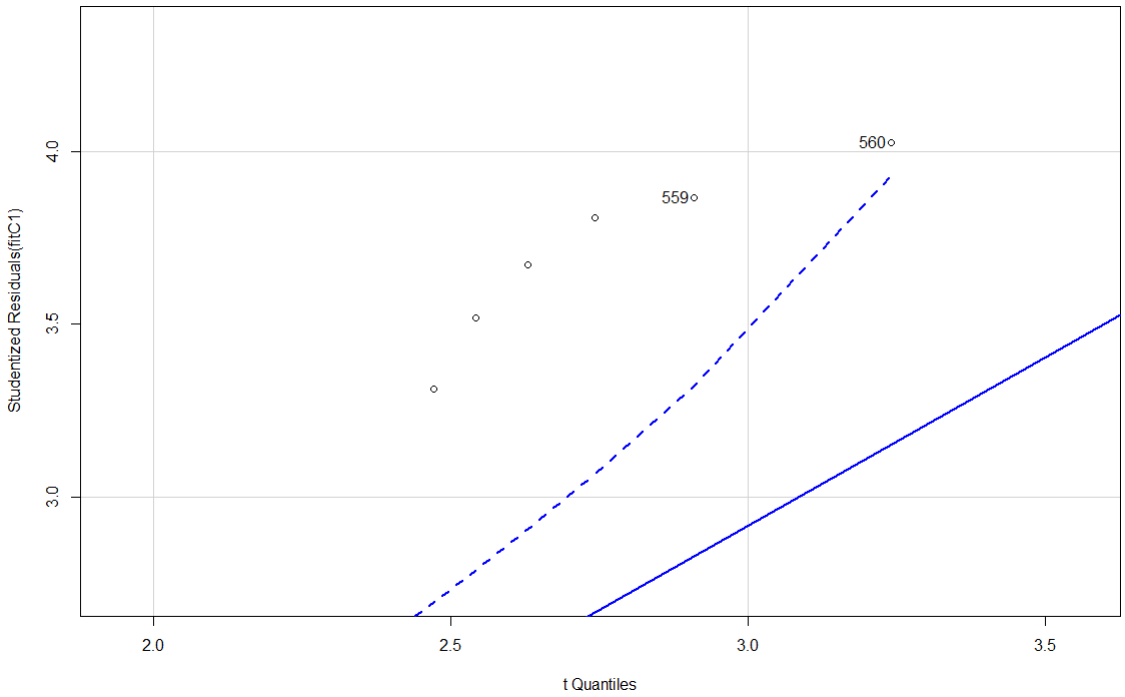


Figure 20: Focus on the far outliers in the model

The model found for the full data set is tested against the model found in problem A by doing a ANOVA test.

```
> Anova(fitA1,fitC1, type = "III")
Anova Table (Type III tests)

Response: clo
             Sum Sq  Df F value    Pr(>F)
(Intercept)  0.6485   1 23.7441 1.327e-06 ***
tOut         0.2333   1  8.5419  0.003569 **
tInOp        0.1938   1  7.0949  0.007886 **
sex          0.1191   1  4.3607  0.037094 *
tInOp:sex    0.1038   1  3.8023  0.051532 .
Residuals   21.8219 799
```

Figure 21: Anova (III) between the model for problem A and problem C

5 Conclusion

This assignment has seen the development of general linear models on an example data set. In problem A, a subset of explanatory variables was selected and analyzed for potential fits to the response variable. A weighted analysis was performed to estimate the difference in observation and residual variance between men and women. The developed models in this section provided reasonable fits. However, the fact that the response variable is non-negative was not taken into account. Consequently, a very high outside temperature would predict a clothing level of less than zero. Likewise, there are no observations for outdoor temperature less than 12°, potentially making the model have poor fit for lower temperatures.

In problem B we included subject ID, making the sex of subjects redundant. The inclusion of subject ID made the model subject specific, thus modeling the clothing variable as person-specific. The analysis found that we collectively change our clothing level according to outdoor temperature, but that indoor temperature has no significant effect, and is determined by our individual style. This model has lowest residual error but it is unfortunately very difficult to generalize to new individuals. It also solved the problem of negative predictions.

Compared to the model in problem A, the model for the full data set is more simple as the model with the full data set is not dependent on the indoor temperature. Anova (type III) test between the model in problem C and problem A showed that the two models were significantly different. The model from problem A would therefore not be as good a fit as the model found for the entire data set. The model for the entire data set estimates that men will wear a lower level of clothing than females, and when the outside temperature increases, the level of insulation will decrease slightly, while a decrease in temperature will increase the level of clothing, which intuitively makes sense. The final model is independent on the indoor temperature. In our group discussion of the results it was backed as we agreed that the temperature within the building would be nearly constant throughout the year, while we were more inclined to wear a hoodie or long t-shirt in the winter months.

6 Appendix

6.1 The Full Final Model of Problem B

```
Call:
lm(formula = clo ~ tOut + subjId, data = CS4)

Residuals:
      Min        1Q    Median        3Q       Max
-0.237842 -0.040145 -0.003432  0.039902  0.281815

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.006833   0.078165  12.881  < 2e-16 ***
tOut        -0.014275   0.003043  -4.690 9.92e-06 ***
subjId17    -0.080873   0.073683  -1.098 0.275380
subjId19     0.040047   0.074108   0.540 0.590294
subjId29    -0.097409   0.074465  -1.308 0.194241
subjId35    -0.296394   0.069311  -4.276 4.81e-05 ***
subjId43    -0.019671   0.079025  -0.249 0.803998
subjId47    -0.157143   0.075287  -2.087 0.039756 *
subjId49    -0.128867   0.082306  -1.566 0.121007
subjId51    -0.002821   0.073632  -0.038 0.969527
subjId55    -0.226138   0.080073  -2.824 0.005863 **
subjId57    -0.284781   0.074013  -3.848 0.000225 ***
subjId59    -0.088014   0.074966  -1.174 0.243546
subjId61    -0.162645   0.075600  -2.151 0.034184 *
subjId63    -0.118644   0.074461  -1.593 0.114664
subjId65    -0.024524   0.074492  -0.329 0.742774
subjId69    -0.274575   0.074883  -3.667 0.000420 ***
subjId71    -0.167212   0.083171  -2.010 0.047444 *
subjId75    -0.324046   0.083621  -3.875 0.000205 ***
subjId85    -0.218847   0.074005  -2.957 0.003985 **
subjId87    -0.164286   0.073756  -2.227 0.028470 *
subjId89    -0.107678   0.074008  -1.455 0.149240
subjId91    -0.049794   0.074292  -0.670 0.504450
subjId93    -0.218089   0.073723  -2.958 0.003973 **
subjId99    -0.191325   0.073932  -2.588 0.011297 *
subjId105   -0.247163   0.073621  -3.357 0.001164 **
subjId107   -0.254759   0.069187  -3.682 0.000398 ***
subjId111   -0.085824   0.073625  -1.166 0.246888
subjId113   -0.412140   0.073837  -5.582 2.60e-07 ***
subjId119   -0.171454   0.074667  -2.296 0.024038 *
subjId123    0.133885   0.073628   1.818 0.072406 .
subjId125   -0.018940   0.073837  -0.257 0.798151
subjId127   -0.180472   0.074272  -2.430 0.017136 *
subjId129   -0.252909   0.073619  -3.435 0.000905 ***
subjId137   -0.149551   0.084784  -1.764 0.081218 .
subjId141   -0.130840   0.074535  -1.755 0.082667 .
subjId145   -0.114582   0.082977  -1.381 0.170812
subjId149    0.074419   0.073633   1.011 0.314940
subjId153   -0.174004   0.074468  -2.337 0.021729 *
subjId157   -0.106379   0.073619  -1.445 0.152013
```

```
subjId167   -0.275601   0.077152  -3.572 0.000577 ***
subjId171   -0.201860   0.074039  -2.726 0.007726 **
subjId173   -0.276415   0.074534  -3.709 0.000364 ***
subjId183   -0.129795   0.078186  -1.660 0.100460
subjId187   -0.157057   0.073633  -2.133 0.035712 *
subjId189   -0.134239   0.083264  -1.612 0.110500
subjId193   -0.075559   0.082531  -0.916 0.362421
subjId199   -0.198203   0.076423  -2.594 0.011126 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09016 on 88 degrees of freedom
Multiple R-squared:  0.7691,Adjusted R-squared:  0.6458
F-statistic: 6.238 on 47 and 88 DF,  p-value: 9.696e-14
```

6.2 R Code for Assignment 1

```
### Clear variables and load data
# clothingSum3 is sorted female first and male second
rm(list=ls())
#set WD
CS <- read.csv(file = "clothingSum3.csv") #Sorted according to sex

library("ggplot2")
library("dplyr")
library("GGally")
library("car")

## Remove subjId and day, and do exploratory analysis
CS1 <- subset(CS, select = -c(subjId,day) )
ggpairs(CS1)
corrplot(cor(CS1))

### Do backwards model selection
fitA1 <- lm(clo ~ tOut*tInOp * sex, data = CS1)
summary(fitA1)

drop1(fitA1,test="F")
fitA1 <- update(fitA1,.~. -tOut:tInOp:sex)
summary(fitA1)

drop1(fitA1,test="F")
fitA1 <- update(fitA1,.~. -tOut:tInOp)
summary(fitA1)

drop1(fitA1,test="F")
fitA1 <- update(fitA1,.~. -tOut:sex)
summary(fitA1)
####### Done with backwards model selection.

############### Do some residual plots
par(mfrow=c(3,2))
plot(residuals(fitA1)) #Residuals versus obs number (not subj id)
```

```
plot(fitted(fitA1),residuals(fitA1)) #residuals versus fitted values
plot(CS1$clo,residuals(fitA1)) #residuals versus response, this will always be linear!
plot(CS1$tOut,residuals(fitA1)) #residuals versus tOut
plot(CS1$tInOp,residuals(fitA1)) #residuals versus tInOp
plot(CS1$sex,residuals(fitA1)) #residuals versus sex
##### The residuals plotted against sex suggests a small difference in variance.
##### The residuals plotted against obs.nr. suggests the same (need color coding)
##### The boxplot in the initial analysis suggested the same.

####### Check for outliers using qqplot
par(mfrow=c(1,1))
qqPlot(fitA1,simulate=FALSE)
range(rstudent(fitA1))
range(rstandard(fitA1))
## which one is the outlier, if any?
which(abs(rstudent(fitA1))==max(abs(rstudent(fitA1))))
par(mfrow=c(2,2))
plot(fitA1)
#### We conclude that there are no outliers here

################# Weighted analysis

ll = 1
vseq = seq(0,10,by=0.01) #Possible weights
for (v in 1:length(vseq)) {
w = c(rep(1,70),rep(vseq[v],66))
fitw <- lm(clo ~ tOut + tInOp + sex + tInOp:sex, weights = w, data = CS1)
#summary(fitw)
ll[v] <- logLik(fitw) #collect the log-likelihood for each weight
}

par(mfrow=c(1,1)) #plot the weights and log-likelihood
plot(vseq,ll,xlab = "variance ratio",ylab="log-likelihood")

# Find the best weight
index = which.max(ll)
optimalweight = vseq[index]
w = c(rep(1,70),rep(optimalweight,66))

# Do a new fit with this weight
fitA2 <- lm(clo ~ tOut + tInOp + sex + tInOp:sex, weights = w, data = CS1)
summary(fitA2)
par(mfrow=c(2,2))
plot(fitA2)
############ p-values get higher, but:

par(mfrow=c(3,2))
plot(residuals(fitA2)) #Residuals versus obs number (not subj id)
plot(fitted(fitA2),residuals(fitA2)) #residuals versus fitted values
plot(CS1$clo,residuals(fitA2)) #residuals versus response, this will always be linear!
plot(CS1$tOut,residuals(fitA2)) #residuals versus tOut
plot(CS1$tInOp,residuals(fitA2)) #residuals versus tInOp
plot(CS1$sex,residuals(fitA2)) #residuals versus sex
```

```r
# Jan: find pearson residuals, maybe option in residuals(fit)
# Otherwise these plots don't make sense


####### Check for outliers
par(mfrow=c(1,1))
qqPlot(fitA2,simulate=FALSE)
range(rstudent(fitA2))
range(rstandard(fitA2))
## which one is the outlier, if any?
which(abs(rstudent(fitA2))==max(abs(rstudent(fitA2))))
par(mfrow=c(2,2))
plot(fitA2)
#### We don't conclude that 79 is an outlier (do we?)

#OPtIONAL  Do a new fit without obs 79
CS2 = CS1[-79,]
w2 = w[-79]
fitA3 <- lm(clo ~ tOut + tInOp + sex + tInOp:sex, weights = w2, data = CS2)
summary(fitA3)
par(mfrow=c(1,1))
qqPlot(fitA3,simulate=FALSE)
par(mfrow=c(2,2))
plot(fitA3)
# Now looks dramatically different (and beter)

# Plot prediction interval
dev.off()
par(mfrow=c(1,2))
tinop <- mean(CS1$tInOp[1:70]) # For given tInOp!
#Sex <- 1 # For given Height!
newdat=data.frame(tOut=CS1$tOut[1:70],tInOp=tinop,sex=CS1$sex[1:70])
pred <- predict(fitA2,se=TRUE, newdata=newdat,interval="prediction")
conf <- predict(fitA2,se=TRUE, newdata=newdat,interval="confidence")

#plot(CS1£tOut[1:69],CS1£clo[1:69])
matplot(CS1$tOut[1:70],pred$fit,type="l",col=c(1,3,3),
        lty=c(1,2,2),main="Female sex",xlab="tOut",ylab="Prediction")


matlines(CS1$tOut[1:70],conf$fit,type="l",col=c(1,3,3),
         lty=c(1,2,2))
points(CS1$tOut[1:70],CS1$clo[1:70])
############## Nr 2
tinop <- mean(CS1$tInOp[71:136]) # For given tInOp!
newdat2=data.frame(tOut=CS1$tOut[71:136],tInOp=tinop,sex=CS1$sex[71:136])
pred2 <- predict(fitA2,se=TRUE, newdata=newdat2,interval="prediction",weights = rep(2.93,136,1))
conf2 <- predict(fitA2,se=TRUE, newdata=newdat2,interval="confidence")

matplot(CS1$tOut[71:136],pred2$fit,type="l",col=c(1,3,3),
        lty=c(1,2,2),main="Male sex",xlab="tOut",ylab="Prediction")
```

```
matlines(CS1$tOut[71:136],conf2$fit,type="l",col=c(1,3,3),
         lty=c(1,2,2))
points(CS1$tOut[71:136],CS1$clo[71:136])


############### See if subject ID can be excluded.
# Load a new subset including subjId
CS3 <- subset(CS, select = -c(day) )
CS3 = CS3[-79,]
CS3$subjId=factor(CS3$subjId)
# Then we look at residuals versus subject ID
par(mfrow=c(1,1))
plot(sort(CS3$subjId),residuals(fitA3),xlab="Subject ID",ylab="Residuals") #residuals versus subjectId
### We can see that the within-subject variance is small, but between-subject is high
# Could be nice to include subject ID
################################### PART B ##########################

########################### Do model selection
### Clear variables and load data
# clothingSum3 is sorted female first and male second
rm(list=ls())
setwd("C:/Users/ander/Documents/10. Semester/Advanced data analysis/Assignment 1")
CS <- read.csv(file = "clothingSum.csv") #Sorted according to sex

library("ggplot2")
library("dplyr")
library("GGally")
library("car")
library("corrplot")

## Remove subjId and day, and do exploratory analysis
CS1 <- subset(CS, select = -c(subjId,day) )
ggpairs(CS1)
corrplot(cor(CS1))

### Do backwards model selection
fitA1 <- lm(clo ~ tOut*tInOp * sex, data = CS1)
summary(fitA1)

drop1(fitA1,test="F")
fitA1 <- update(fitA1,.~. -tOut:tInOp:sex)
summary(fitA1)

drop1(fitA1,test="F")
fitA1 <- update(fitA1,.~. -tOut:tInOp)
summary(fitA1)

drop1(fitA1,test="F")
fitA1 <- update(fitA1,.~. -tOut:sex)
summary(fitA1)
####### Done with backwards model selection.

############### Do some residual plots
```

```
par(mfrow=c(3,2))
plot(residuals(fitA1)) #Residuals versus obs number (not subj id)
plot(fitted(fitA1),residuals(fitA1)) #residuals versus fitted values
plot(CS1$clo,residuals(fitA1)) #residuals versus response, this will always be linear!
plot(CS1$tOut,residuals(fitA1)) #residuals versus tOut
plot(CS1$tInOp,residuals(fitA1)) #residuals versus tInOp
plot(CS1$sex,residuals(fitA1)) #residuals versus sex
##### The residuals plotted against sex suggests a small difference in variance.
##### The residuals plotted against obs.nr. suggests the same (need color coding)
##### The boxplot in the initial analysis suggested the same.
# Jan: find variance of residuals for women and men

####### Check for outliers using qqplot
par(mfrow=c(1,1))
qqPlot(fitA1,simulate=FALSE)
range(rstudent(fitA1))
range(rstandard(fitA1))
## which one is the outlier, if any?
which(abs(rstudent(fitA1))==max(abs(rstudent(fitA1))))
par(mfrow=c(2,2))
plot(fitA1)
#### We conclude that there are no outliers here




################# Weighted analysis

ll = 1
vseq = seq(0,10,by=0.01) #Possible weights
for (v in 1:length(vseq)) {
w = c(rep(1,70),rep(vseq[v],66))
fitw <- lm(clo ~ tOut + tInOp + sex + tInOp:sex, weights = w, data = CS1)
#summary(fitw)
ll[v] <- logLik(fitw) #collect the log-likelihood for each weight
}

par(mfrow=c(1,1)) #plot the weights and log-likelihood
plot(vseq,ll,xlab = "variance ratio",ylab="log-likelihood")

# Find the best weight
index = which.max(ll)
optimalweight = vseq[index]
w = c(rep(1,70),rep(optimalweight,66))

# Do a new fit with this weight
fitA2 <- lm(clo ~ tOut + tInOp + sex + tInOp:sex, weights = w, data = CS1)
summary(fitA2)
par(mfrow=c(2,2))
plot(fitA2)
############ p-values get higher, but:

par(mfrow=c(3,2))
```

```
plot(residuals(fitA2)) #Residuals versus obs number (not subj id)
plot(fitted(fitA2),residuals(fitA2)) #residuals versus fitted values
plot(CS1$clo,residuals(fitA2)) #residuals versus response, this will always be linear!
plot(CS1$tOut,residuals(fitA2)) #residuals versus tOut
plot(CS1$tInOp,residuals(fitA2)) #residuals versus tInOp
plot(CS1$sex,residuals(fitA2)) #residuals versus sex


####### Check for outliers
par(mfrow=c(1,1))
qqPlot(fitA2,simulate=FALSE)
range(rstudent(fitA2))
range(rstandard(fitA2))
## which one is the outlier, if any?
which(abs(rstudent(fitA2))==max(abs(rstudent(fitA2))))
par(mfrow=c(2,2))
plot(fitA2)
#### We conclude that 79 is an outlier



# Plot prediction interval
dev.off()
par(mfrow=c(1,2))
tinop <- mean(CS1$tInOp[1:70]) # For given tInOp!
#Sex <- 1 # For given Height!
newdat=data.frame(tOut=CS1$tOut[1:70],tInOp=tinop,sex=CS1$sex[1:70])
pred <- predict(fitA2,se=TRUE, newdata=newdat,interval="prediction")
conf <- predict(fitA2,se=TRUE, newdata=newdat,interval="confidence")

#plot(CS1£tOut[1:69],CS1£clo[1:69])
matplot(CS1$tOut[1:70],pred$fit,type="l",col=c(1,3,3),
        lty=c(1,2,2),main="Female sex",xlab="tOut",ylab="Prediction")


matlines(CS1$tOut[1:70],conf$fit,type="l",col=c(1,3,3),
         lty=c(1,2,2))
points(CS1$tOut[1:70],CS1$clo[1:70])
############## Nr 2
tinop <- mean(CS1$tInOp[71:136]) # For given tInOp!
newdat2=data.frame(tOut=CS1$tOut[71:136],tInOp=tinop,sex=CS1$sex[71:136])
pred2 <- predict(fitA2,se=TRUE, newdata=newdat2,interval="prediction",weights = rep(2.93,136,1))
conf2 <- predict(fitA2,se=TRUE, newdata=newdat2,interval="confidence")

matplot(CS1$tOut[71:136],pred2$fit,type="l",col=c(1,3,3),
        lty=c(1,2,2),main="Male sex",xlab="tOut",ylab="Prediction")


matlines(CS1$tOut[71:136],conf2$fit,type="l",col=c(1,3,3),
         lty=c(1,2,2))
points(CS1$tOut[71:136],CS1$clo[71:136])
```

```
############## See if subject ID can be excluded.
# Load a new subset including subjId
CS3 <- subset(CS, select = -c(day) )
CS3 = CS3[-79,]
CS3$subjId=factor(CS3$subjId)
# Then we look at residuals versus subject ID
par(mfrow=c(1,1))
plot(sort(CS3$subjId),residuals(fitA3)) #residuals versus subjectId
### We can see that the within-subject variance is small, but between-subject is high
# Could be nice to include subject ID

################################### PART B ##########################

########################### Do model selection
CS4 <- subset(CS, select = -c(day,sex) )
CS4$subjId=factor(CS4$subjId)

fitB1 <- lm(clo ~ tOut*tInOp*subjId - tOut:tInOp:subjId, data = CS4)
summary(fitB1)

fitB2 <- update(fitB1,.~. -tOut:subjId)
summary(fitB2)

Anova(fitB2,type = "III")
par(mfrow=c(2,2))
plot(fitB2)
### Do backwards model selection

drop1(fitB2,test="F")
fitB3 <-update(fitB2,.~. -tInOp:subjId)
summary(fitB3)

drop1(fitB3,test="F")
fitB4 <-update(fitB3,.~. -tOut:tInOp)
summary(fitB4)

drop1(fitB4,test="F")
fitB5 <-update(fitB4,.~. -tInOp)
summary(fitB5)
Anova(fitB5,type="III")
par(mfrow=c(2,2))
plot(fitB5)

# Final model has been determined
# Confidence interval on the parameters
# Export to Matlab to get an easy to use plot environment!!! :)
library("datasets")

confint(fitB5)
men <- subset(CS, sex=="male",subjId)
men <- unique(men)
```

```r
women <- subset(CS, sex=="female",subjId)
women <- unique(women)


library(coefplot)

coefplot(fitB5, horizontal = TRUE, innerCI = 0, pointSize =2,numberAngle = -90)

# Prediction interval

# 1. Add predictions
pred.int <- predict(fitB5, interval = "prediction")
mydata <- cbind(CS4, pred.int)

# 2. Regression line + confidence intervals as function of tOut
p <- ggplot(mydata, aes(tOut, clo)) +
  geom_point() +  stat_smooth(method = lm)

# 3. Add prediction intervals
p + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y = upr), color = "red", linetype = "dashed")

# 4. Regression line + confidence intervals as function of subjId
ps <- ggplot(mydata, aes(subjId, clo)) +
  geom_point() +  stat_smooth(method = lm)
# 5. Add prediction intervals
ps + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y = upr), color = "red", linetype = "dashed")


######################## Part C ######################3
#CSfull <- read.csv(file = "clothingfull2.csv") #Sorted according to sex
CSfull <- clothingFull2
View(CSfull)
CS3 <- subset(CSfull, select = -c(subjId, day,X) )
ggpairs(CS3)
corrplot(cor(CS1))

ll = 1
vseq = seq(0,10,by=0.01) #Possible weights
for (v in 1:length(vseq)) {
  W = c(rep(1,419),rep(vseq[v],384))
  fitw <- lm(clo ~ tOut + tInOp + sex + tInOp:sex, weights = W, data = CS3)
  #summary(fitw)
  ll[v] <- logLik(fitw) #collect the log-likelihood for each weight
}
par(mfrow=c(1,1)) #plot the weights and log-likelihood
plot(vseq,ll,xlab = "variance ratio",ylab="log-likelihood")
index = which.max(ll)
optimalweight = vseq[index]
W=c(rep(1,419),rep(optimalweight,384))

fitC1 <- lm(clo ~ tOut*tInOp *sex,weights=W, data = CSfull)
```

```r
summary(fitC1)
drop1(fitC1,test="F")
fitC1 <- update(fitC1,.~. -tOut:tInOp:sex)
summary(fitC1)

drop1(fitC1,test="F")
fitC1 <- update(fitC1,.~. -tOut:tInOp)
summary(fitC1)

drop1(fitC1,test="F")
fitC1 <- update(fitC1,.~. -tInOp:sex)
summary(fitC1)

drop1(fitC1,test="F")
fitC1 <- update(fitC1,.~. -tInOp)
summary(fitC1)
#Model fra opgave 1: fitA1 <- lm(clo ~ tOut + tInOp + sex + tInOp:sex,weights=W, data = CSfull)
summary(fitC1)
par(mfrow=c(2,2))
plot(fitC1)
CS3 = CSfull[-560,]
fitC1 <- lm(clo ~ tOut+ sex + tOut:sex,weights=W, data = CS3)


c0 <- coef(M)
cc <- confint(M, level = 0.95)
b <- drop(barplot(c0,ylim=range(c(cc))))  ## b stores vector of x positions
arrows(b,c0,b,cc[,1],angle=90,length=0.05) ## lower bars
arrows(b,c0,b,cc[,2],angle=90,length=0.05) ## upper bars

Anova(fitA1,fitC1, type = "III")

par(mfrow=c(1,1))
qqPlot(fitC1,simulate=FALSE)
#install.packages("zoom")
#(zoom)
#zm(fitC1)
#Test af mean og std. afivgelse af temperatur for begge dataset
mean(CSfull$tOut)
sd(CSfull$tOut)
mean(CSfull$tInOp)
sd(CSfull$tInOp)
mean(CS$tOut)
sd(CS$tOut)
mean(CS$tInOp)
sd(CS$tInOp)
```