Advanced Dataanalysis and Statistical Modelling: Assignment 1

Spring 22, 02424
March 11, 2022

s170303, Andreas Engly
s174434, Mads Esben Hansen

Danmarks
Tekniske Universitet

Introduction

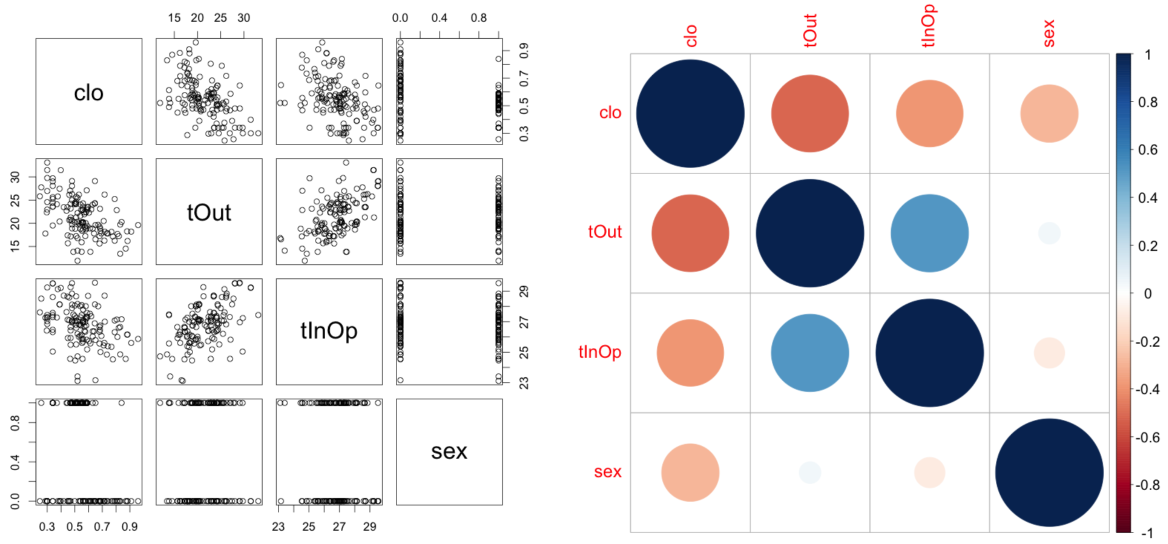
This is the answer for the first of three mandatory assignments for the course 02424. The report concerns itself with data of the level of clothing people are wearing at the office. Table 1 is an overview of the given data which will be used throughout the report.

Table 1 – Overview of available data.

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)

Problem A

The first part of the report concerns itself with modeling of clothing insulation level based on indoor operating temperature and outdoor temperature (we ignore sebjId and day). Figure 1 shows an analysis of the dependence between the variables. From figure 1 (b) it is clear that there is some correlation between temperature and level of clothing. In fact, it seems there is a negative correlation, meaning people tend to wear less clothes on hot days. Further the sex parameter has been encoded s.t. $sex = 1$ means male and $sex = 0$ female. It is seen from figure 1 that there is a small negative correlation. It indicates that men in general are wearing less clothes than women. None of the variables display a clear non-linear relationship.



(a) Raw variable comparison.

(b) Correlation between variables.

Figure 1 – Analysis of variables (sex=1 male, sex=0 female).

In order to find the optimal linear model for predicting amount of clothing, we need a notion of comparison between two models. We will follow the approach outlined by Madsen 2010 [1] in section 3.6. For normal distributed variables the *deviance* is given by definition 3.3.

$$D(\boldsymbol{y};\boldsymbol{\mu}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{1}$$

As we initially assume $\boldsymbol{\Sigma} = \boldsymbol{I}$, the deviance is the residual sum of squares (RSS). Hence, by looking at the relative change as we adjust the parameter space we can infer if the model has changed significantly. As the residuals are assumed to follow a normal distribution, it follows that the squared residuals follow a chi-squared distribution. As chi-square random variables are closed under addition, the ratio between RSS of the two models will follow a F-distribution. Therefore, we can perform a F-test (named after its inventor Ronald Fisher, who is also the inventor of the F-distribution). Specifically, to compare model 1 and 0 we use

$$F(\boldsymbol{y}) = \frac{\mathrm{D}\left(p_1(\boldsymbol{y});p_0(\boldsymbol{y})\right)/\left(m_1-m_0\right)}{\mathrm{D}\left(\boldsymbol{y};p_1(\boldsymbol{y})\right)/\left(n-m_1\right)} \tag{2}$$

where $p_0(\boldsymbol{y})$ and $p_1(\boldsymbol{y})$ denote the projections onto parameter subspace Ω_0 and Ω_1 respectively. Furthermore, $F(\boldsymbol{y}) \sim \mathcal{F}(m_1-m_0, n-m_1)$ where m_i is the number of parameters in model i and n is the number of measurements.

Now we are ready to find the optimal model. We choose to use *type II* method for model selection. This means that we start off with the *full* model with the corresponding *null hypothesis.*

$$\mathcal{H}_0 : \boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{X\beta} \text{ where } \boldsymbol{\beta} \in \boldsymbol{\Omega_0} = \mathbb{R}^8 \tag{3}$$

The model has the following parameterization.

$$\texttt{clo} = a_i + b_{1,i}\cdot\texttt{tOut} + b_{2,i}\cdot\texttt{tInOp} + b_{3,i}\cdot\texttt{tOut}\cdot\texttt{tInOp} + \varepsilon, \qquad i \in \{\text{male},\text{female}\} \tag{4}$$

where the noise $\varepsilon \sim \mathcal{N}(0,\sigma^2 I)$. We now remove the parameters that changes the model the least according to our relative measure of deviances until the model is *significantly* different from the full model. Formally, we continue until we observe a test statistic that is unlikely on a $\alpha = 0.05$ significance level. There is one additional rule: We will not remove any parameters that is currently in the model in a higher order. In other words, we start by the higher order terms. By introducing alternative hypothesis with restricted parameter spaces, we get the following chain.

$$\mathcal{H}_i : \boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{X\beta} \text{ where } \boldsymbol{\beta} \in \boldsymbol{\Omega_i} = \mathbb{R}^{8-i} \text{ for } i \in \{1,2,..,8\} \tag{5}$$

After following the procedure, we end up accepting $\mathcal{H}_4$ (and rejecting $\mathcal{H}_3$) leading us to the following restricted model.

$$\texttt{clo} = a_i + b_1\cdot\texttt{tOut} + b_{2,i}\cdot\texttt{tInOp} + \varepsilon, \qquad i \in \{\text{male},\text{female}\} \tag{6}$$

where $\varepsilon \sim \mathcal{N}(0,\sigma^2 I)$. Notice we went from a model with 8 parameters to one with 5. Table 2 shows the estimated values for the parameters in equation 6. Notice both males and females wear less clothes when it is hot. Further notice that females vary their amount of clothes more than their male counterparts based on the indoor temperature.

Table 2 – OLS parameter estimates.

Variable	**Value**	`sd`
α_{male}	0.849	0.314
α_{female}	2.13	0.317
`tOut`	-0.0122	0.00302
`tInOp`$_{male}$	-0.00289	0.0120
`tInOp`$_{female}$	-0.0475	0.0129

Now the question is, whether the model is a *good* model or not. To give us an indication of this, we will address the residuals of the model. Figure 2 shows the residuals against all 4 variables. The first thing we notice is that the model does not capture all variance in the clothing variable. That is, it does not explain the amount of clothing particularly well.

To investigate if we can improve the model, we must turn our focus to the dependency between residuals and predictors. Looking at `tInOp` and `tOut` there does not seem to be any strong dependency, linear or otherwise. Finally, we have have the dependency on sex. It seems the error in general is larger for females than males. To check if there is in fact a difference we can perform an F-test with the hypothesis $\mathcal{H}_0 : \sigma^2_{\text{male}} = \sigma^2_{\text{female}}$. Doing so yields a p-value $= 3.742e{-}5$ indicating that the variance of females and males are not the same.

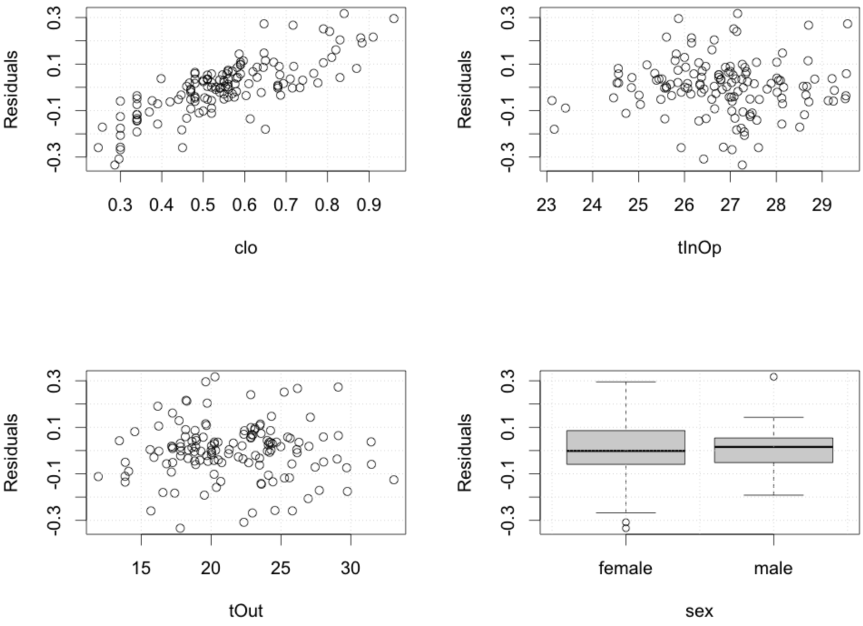


Figure 2 – Residual analysis of the model.

Introducing Weights

It is indicated from the residuals that the assumption on the variance structure is not correct. Specifically, the assumption that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ does not hold.

$$\text{Var}[\epsilon_i | \boldsymbol{x}_i] = f(\boldsymbol{x}_i) \neq \sigma^2, \quad \sigma \in \mathbb{R}, i \in \{\text{male}, \text{female}\} \tag{7}$$

Instead we will use $\varepsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal matrix with separate values for males and females.

The model is estimates as described in appendix A and the weight matrix, $\boldsymbol{\Sigma}$, is estimated using the relaxation algorithm described in appendix B. Finally, the weight matrix, $\boldsymbol{\Sigma}$, converges where $\sigma_{male} = 0.084$ and $\sigma_{female} = 0.1436408$. Table 3 shows the WLS estimates of the model from equation 6.

Table 3 – WLS parameter estimates.

Variable	**Value**	`sd`
α_{male}	0.852	0.221
α_{female}	2.22	0.364
`tOut`	-0.0122	0.00254
$\texttt{tInOp}_{male}$	-0.00289	0.00853
$\texttt{tInOp}_{female}$	-0.0475	0.0142

To illustrate the relationship between clothing and outdoor temperature, we condition on the mean value of the indoor temperature. Figure 3 shows that women can be expected to wear more clothes than men for any level of outdoor temperature. It is worth to notice that the prediction intervals for women are wider than for men. This is because we account for the difference in variance when weighting the observations. Figure 4 shows what happens if you instead were to keep `tOut` constant and vary `tInOp`. Notice the slope of the mean for males and females are different.

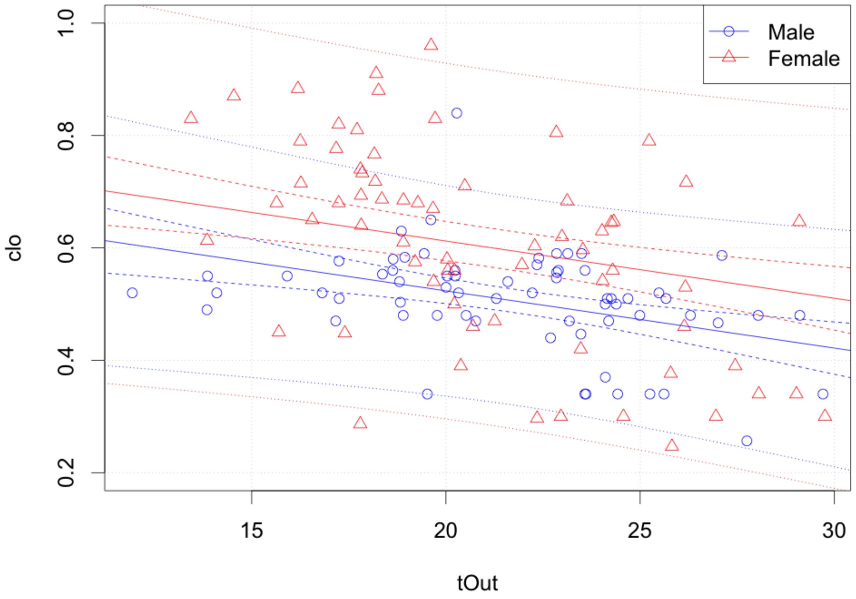


Figure 3 – WLS with `tInOp` constant.

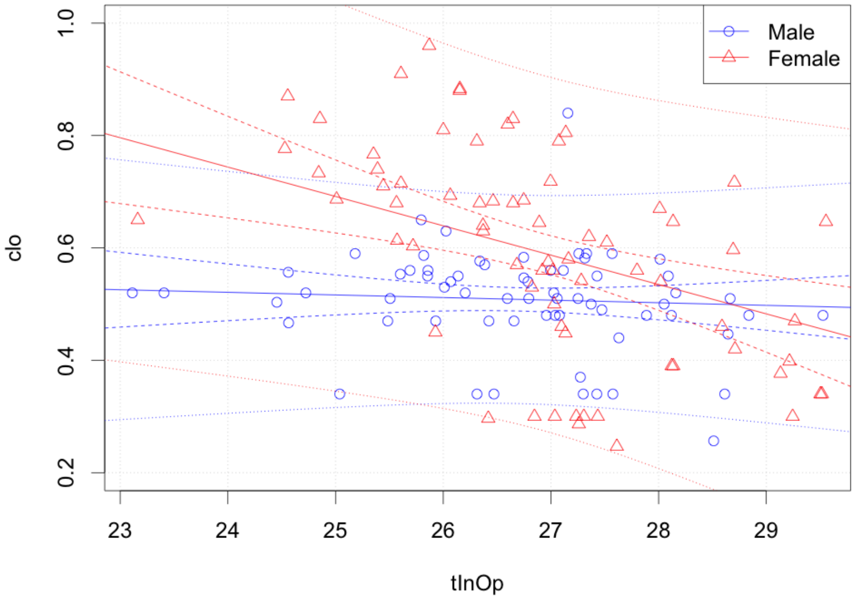


Figure 4 – WLS with `tOut` constant.

To investigate the performance of the model, we can again plot the residuals against the variables. Figure 5 shows the residuals of the WLS model against all the variables. Overall it is very similar to the previous residual plot. The only difference is that now our assumptions regarding residual errors aligns with the difference in residual variance between the sexes.

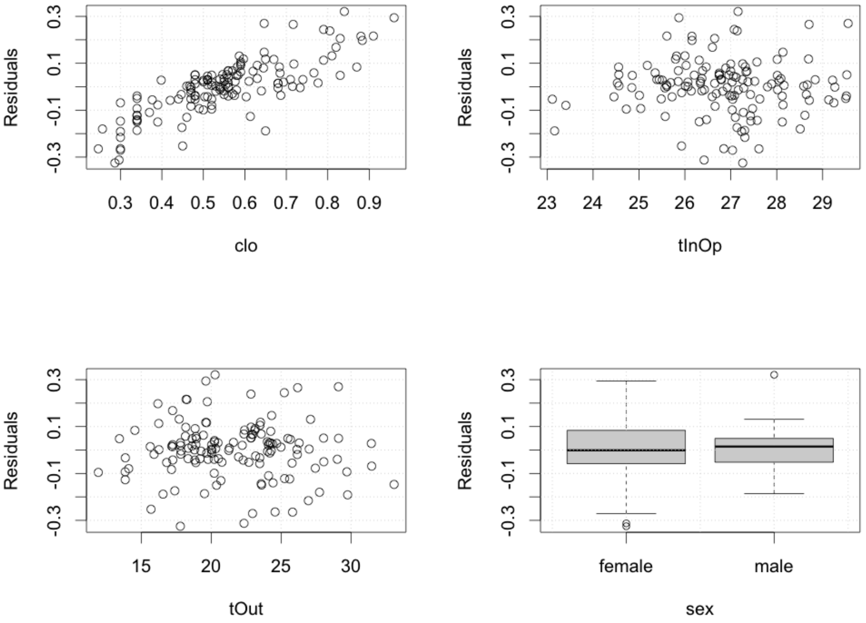


Figure 5 – Residual analysis of the WLS model.

Until now we have completely disregarded `subjId`. We now want to investigate if it is reasonable to do so. Figure 6 show residuals plotted against subjId. By visual inspection, it seems the residuals might belong to different distributions. We will therefore perform a one-way analysis of variance (ANOVA). Since we have assumed different variance for male and female, we will perform the analysis for male and female individually. The analysis tells us, if we can statistically reject the hypothesis that the residuals for the subjIds come from the same distributions (one for male and one for female). Doing the ANOVA yields $p - value = 0.095$ for males and $p - value = 0.29$ for females. Hence, there is no reason to believe the residuals do not come from the same distribution.

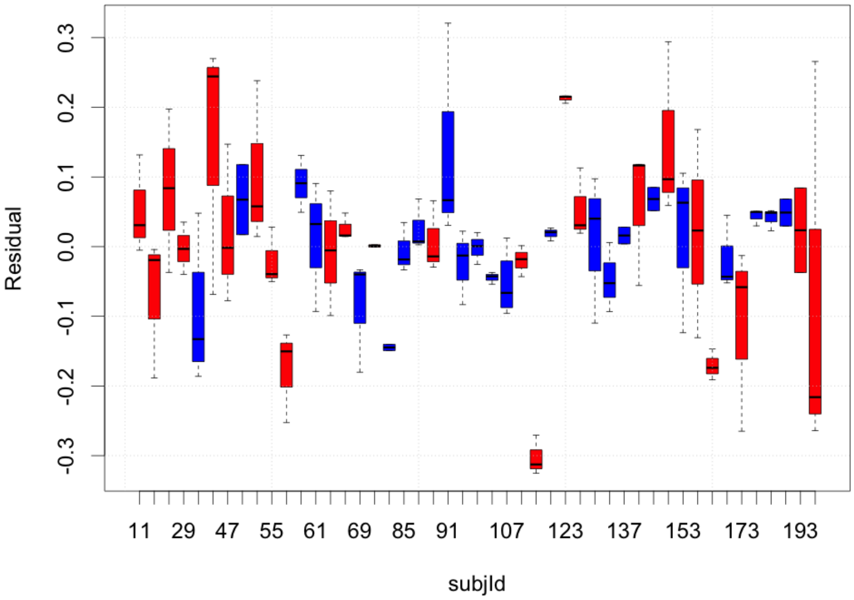


Figure 6 – Residuals against `subjId`.

Problem B

In this section we will develop a model that includes subject Id. Since sex is a given directly from the subject id we will not use sex in this model. Again, we use type II for model selection, we start off with the largest possible model.

$$\texttt{clo} = a_i + b_{1,i} \cdot \texttt{tOut} + b_{2,i} \cdot \texttt{tInOp} + b_{3,i} \cdot \texttt{tOut} \cdot \texttt{tInOp} + \varepsilon, \qquad i \in \texttt{subjId} \tag{8}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Notice this model does in fact have 0 degrees of freedom. This complicates model selection, since we would essentially be trying to divide by 0. Therefore we assume at at least one parameter can be removed. Since we have decided on using type II, we remove the parameter with highest degree and arrive at the following model.

$$\texttt{clo} = a_i + b_{1,i} \cdot \texttt{tOut} + b_{2,i} \cdot \texttt{tInOp} + b_3 \cdot \texttt{tOut} \cdot \texttt{tInOp} + \varepsilon, \qquad i \in \texttt{subjId} \tag{9}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. We are now able to run the algorithm. Afterwards we are left with the model

$$\texttt{clo} = a_i + b \cdot \texttt{tOut} + \varepsilon, \qquad i \in \texttt{subjId} \tag{10}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. So not only does the subject id dependence on the indoor temperature go away, indoor temperature is no longer significant at all! We are left with a fairly simple model: Individual intercept for each subject id, and common linear dependence on outdoor temperature. Figure 7 shows a graphical representation of the fitted model.

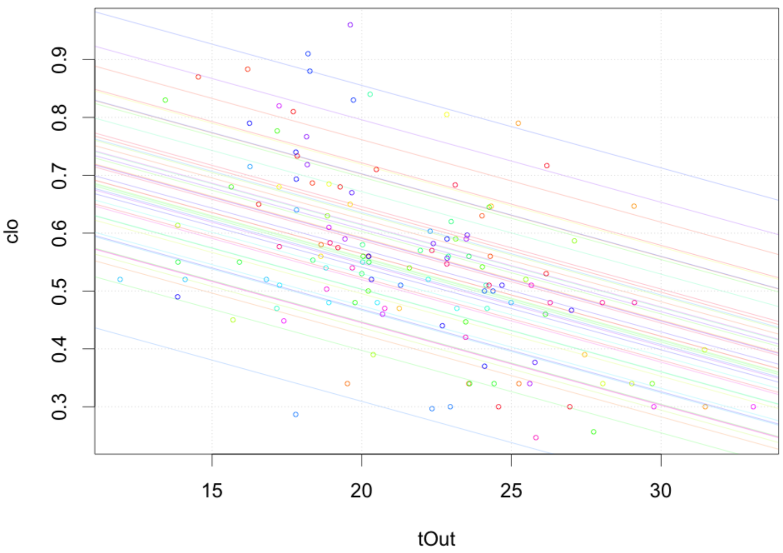


Figure 7 – Model given in equation 10

Upon visual inspection it seems that most of the estimated intercepts lie relatively close with some deviating substantially. This is further backed up by figure 8 which shows the distribution of the intercepts. Ultimately, we have a model where the relationship between outdoor temperature and amount of clothing is equal for all. However, each subject has an individual bias that is roughly normal distributed.

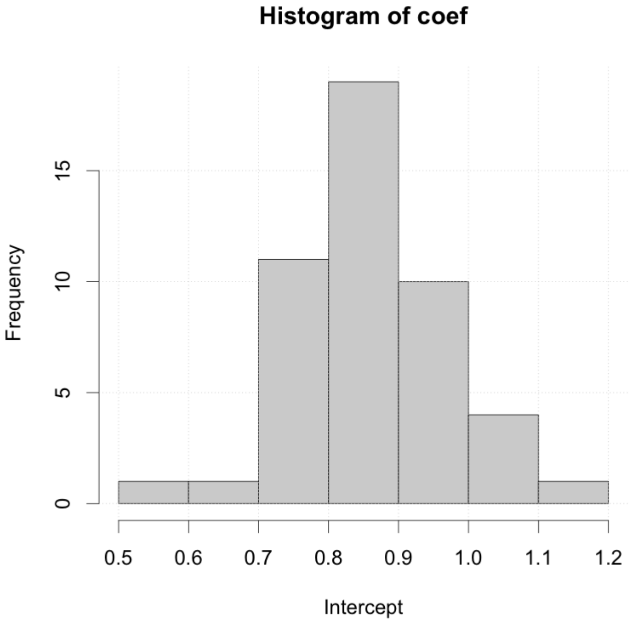


Figure 8 – Distribution of intercepts.

In the case where a specific group of subjects is measured, this model is able to predict the personal behaviour of each subject. However, if a new subject is added to the group the model is not able to estimate the individual bias of said subject, and hence not able to predict their amount of clothing. Further, since the model assumes equal dependence on outdoor temperature for all subject, a subject that dresses differently could be predicted poorly.

Problem C

The data we have used so far is an aggregated version of a larger data-set that includes multiple observations per day. In this part we will use the full data-set, which includes a number of observations (around 6) for each day that a subject visits the lab.

We will now fit a GLM model (like in problem A and problem B) on the full data set. Once again, we use type II for model selection starting off with the full model. Since sex can be derived directly from subject id, we will consider two *lines* of models. One where sex is used as a factor (as in problem A), and one where subjId is used as a factor (as in problem B). When the optimal model for each *line* is found, we will compare these two. Let us start with the model using sex as a factor. We start with the *full* model.

$$\texttt{clo} = a_i + b_{1,i} \cdot \texttt{tOut} + b_{2,i} \cdot \texttt{tInOp} + b_{3,i} \cdot \texttt{tOut} \cdot \texttt{tInOp} + \varepsilon, \qquad i \in \{male, female\} \tag{11}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Assessing the deviance between models when removing the only 3. order term we see that no terms can be removed. Therefore we will simply use the full model.

Now onto the model using subjId. Again, we start with the *full* model.

$$\texttt{clo} = a_i + b_{1,i} \cdot \texttt{tOut} + b_{2,i} \cdot \texttt{tInOp} + b_{3,i} \cdot \texttt{tOut} \cdot \texttt{tInOp} + \varepsilon, \qquad i \in \texttt{subjId} \tag{12}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. When testing for significance between models, we determine that we are allowed to remove 1 term before significantly reducing the accuracy of the model (it fails its F-test). Hence we get the model

$$\texttt{clo} = a_i + b_{1,i} \cdot \texttt{tOut} + b_{2,i} \cdot \texttt{tInOp} + b_3 \cdot \texttt{tOut} \cdot \texttt{tInOp} + \varepsilon, \qquad i \in \texttt{subjId} \tag{13}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Now we have found the optimal model using sex and subject id respectively, but which is better? We once again perform an F-test to see if they are significantly different. We see that the second model, given in equation 13, is significantly better than the model that use sex (even when accounting for the lower degree of freedom). However, as discussed in problem B, a model that uses subject id will not be able to predict clothing for any new person.

Finally, when fitting the model we assume that all residuals are iid. However, with basic knowledge about human behaviour it is fair to say that people often wear the same clothes throughout the day. Therefore there could be a strong correlation between measurements of the same subject id on the same day which is not accounted for in the model. In fact, when the intra-day correlation goes to 1, the optimal model will go towards the model found in section A and B, since this would essentially mean that we only use 1 measurement per day.

Conclusion

This assignment was divided into 3 parts. First we looked at modelling the amount of clothes a person is wearing based on outdoor temperature, indoor temperature and sex. We found that males and females have a different bias, i.e. for same temperature they tend to wear a different amount of clothing (females wearing more). Based on indoor temperature females tend to vary their amount of clothing more than males. There is no indication that males and females clothe differently in relation to outdoor temperature. However, since outdoor and indoor temperature are positively correlated, it can be difficult to separate the effect of these.

Secondly, we modelled amount of clothes based on personal preference instead of sex, i.e. unique id used to group and sex disregarded. Doing so showed that people tend to adjust their amount of clothing similarly based on outdoor temperature. It also showed that people have quite diverse biases, i.e. a person that wears more clothes than average on a given day, usually does so on any other day. The bias seemed to roughly follow a normal distribution.

Thirdly, we tried generating a model using 6 measurements for each day (only used 1 until now). This resulted in a quite extensive model with many parameters and personal dependence (`subjId`). However, there may be an argument that it is necessary to include intra-day correlation in the model to avoid over-fitting.

Ultimately, the conclusion must be: If you want to forecast the amount of clothes someone will wear based on temperature, it is best to know their personal preference; however, knowing their sex will also get you a long way.

References

[1] H. Madsen and P. Thyregod, *Introduction to General and Generalized Linear Models*, 2010.

[2] H. Madsen, *Time Series Analysis*, 2008.

A Derivation of Maximum Likelihood Estimates

We assume that the inherent uncertainty in our model can be modelled by the multivariate normal distribution.

$$\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\Sigma) \tag{14}$$

The *probability density function* for the multivariate normal distribution is given by the following.

$$f_Y(y) = \frac{1}{(\sqrt{2\pi})^n \sigma^n \sqrt{\det(\Sigma)}} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})] \tag{15}$$

Then we introduce the *likelihood function.*

$$L(\boldsymbol{\mu}, \sigma^2; \boldsymbol{y}) = \frac{1}{(\sqrt{2\pi})^n \sigma^n \sqrt{\det(\boldsymbol{\Sigma})}} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})] \tag{16}$$

To ease the calculation of derivatives, we take the logarithm since it can be shown that it preserves the extremas of the join density function. Further, we replace $\boldsymbol{\mu}$ with a linear function $\boldsymbol{\mu} = \boldsymbol{X\beta}$.

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\sigma^2}; \boldsymbol{y}) &= -n \cdot \log(\sqrt{2\pi}) - n \cdot \log(\sigma^2) - \log\sqrt{\det(\boldsymbol{\Sigma})} - \frac{1}{2\sigma^2} \cdot (\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \\ &= -n \cdot \log(\sqrt{2\pi}) - n \cdot \log(\sigma^2) - \log\sqrt{\det(\boldsymbol{\Sigma})} - \frac{1}{2\sigma^2} \cdot (\boldsymbol{y}-\boldsymbol{X\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\end{aligned}$$

Then we can compute the score function by finding the derivative of the log-likelihood function w.r.t. $\boldsymbol{\beta}$.

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\sigma^2}; \boldsymbol{y})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}(-\frac{1}{2\sigma^2} \cdot (\boldsymbol{y}-\boldsymbol{X\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})) && (17) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}(-\frac{1}{2\sigma^2} \cdot (\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1})(\boldsymbol{y}-\boldsymbol{X\beta})) && (18) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}(-\frac{1}{2\sigma^2} \cdot (\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X\beta})) && (19) \\ &= -\frac{1}{\sigma^2} \cdot (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X\beta}) && (20)\end{aligned}$$

Since $\boldsymbol{\beta}$ is unknown, we effectively compute an estimate. Therefore, we replace $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$. Setting the score function equal to 0 yields the normal equation.

$$\begin{aligned}-\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} &= 0 \\ \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\end{aligned}$$

Hence, finding $\boldsymbol{\beta}$ satisfying that the parenthesis is 0 will maximize the likelihood. The statistical moments can then be derived by replacing $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$.

$$\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{\beta}}] &= E[(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})] \\
&= \boldsymbol{\beta} + ((\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1})E[\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta}
\end{aligned}$$

It is seen from above that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator. Then we can compute the variance.

$$\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{\beta}}] &= Var[(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})] \\
&= ((\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1})Var[\boldsymbol{\epsilon}]((\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1})^T \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}
\end{aligned}$$

B Relaxation Algorithm

In some cases, the i.i.d. assumption for the residuals (i.e. $\mathbf{\Sigma} \neq \boldsymbol{I}$) is violated. The relaxation algorithm is used to find an estimate for $\mathbf{\Sigma}$. It is described in *Madsen 2008* [2].

Algorithm 1 Relaxation Algorithm

```
procedure RELAXATION($\boldsymbol{y}, \boldsymbol{X}$)                ▷ Computes covariance structure
    $\mathbf{\Sigma} \leftarrow \boldsymbol{I}$
    while $\mathbf{\Sigma}$ has not converged do                ▷ When structure keeps changing
        $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \mathbf{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \mathbf{\Sigma}^{-1} \boldsymbol{y}$
        $\boldsymbol{\epsilon} \leftarrow \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$
        $\mathbf{\Sigma} \leftarrow D[\boldsymbol{\epsilon}]$
    return $\mathbf{\Sigma}$                ▷ Best estimate for residual structure is $\mathbf{\Sigma}$
```

C R-code

```
#### Problem A ----
setwd( '/Users/mads/Google␣Drev/Skole/Uni/10_semester/02424/Assignment1'
   ↪ )

if (file.exists("clothingFull.csv") & file.exists("clothingSum.csv")) {
  print("Files␣are␣acessable.␣You␣can␣proceed!")
  clothingFull <- read.csv("./clothingFull.csv", stringsAsFactors = FALSE
     ↪ )
  clothingSum <- read.csv("./clothingSum.csv", stringsAsFactors = FALSE)
}

# Make an exploratory analysis of the data
library(corrplot)
clothingSum$sex[clothingSum$sex == 'male'] <- 1;
clothingSum$sex[clothingSum$sex == 'female'] <- 0;

clothingSum$sex <- as.numeric(clothingSum$sex);
clothingSum$X <- NULL;
clothingSum$subjId <- NULL;
clothingSum$obs.no <- NULL;
clothingSum$day <- NULL;

png("figures/cor-raw.png", width = 800, height = 800, pointsize = 24)
pairs(clothingSum)
dev.off()

png("figures/cor-heat.png", width = 800, height = 800, pointsize = 24)
corrplot(cor(clothingSum))
dev.off()

## GLM fit
clothingSum <- read.csv("./clothingSum.csv", stringsAsFactors = FALSE)

fit0 <- lm(clo ~ tOut*tInOp*sex, data = clothingSum)

drop1(fit0,test="F")
fit1 <- update(fit0,.~.-tOut:tInOp:sex)

drop1(fit1,test="F")
fit2 <- update(fit1,.~.-tOut:tInOp)

drop1(fit2,test="F")
fit3 <- update(fit2,.~.-tOut:sex)

drop1(fit3,test="F")
```

```
## Lets just drop one significant one for show
fit4 <- update(fit3,.~.-tInOp:sex)

## And we are done, now lets compare the models with anova
anova(fit0,fit1) ## ok
anova(fit0,fit2) ## ok
anova(fit0,fit3) ## ok
anova(fit0,fit4) ## not ok, full model is significantly better!

summary(fit3) ## final model

anova(fit0)
library("car")
anova(fit3)
Anova(fit3, type = "II")
Anova(fit3, type = "III")

## Compute param sd
Y = clothingSum$clo
X = matrix(NA,nrow=nrow(clothingSum), ncol=5)
X[,1] = clothingSum$sex == "male"
X[,2] = clothingSum$sex != "male"
X[,3] = clothingSum$tOut
X[,4] = clothingSum$tInOp * (clothingSum$sex == "male")
X[,5] = clothingSum$tInOp * (clothingSum$sex != "male")

Sigma = diag(rep(1,nrow(X)))
Beta = solve(t(X) %*% solve(Sigma) %*% X) %*% t(X) %*% solve(Sigma) %*% Y
eps = Y - (X %*% Beta)
df = 131

Beta
sqrt(diag(solve(t(X) %*% solve(Sigma) %*% X))) * sqrt(sum(eps^2/df))

## Make plots
png("figures/res_anal.png", width = 1000, height = 800, pointsize = 24)
par(mfrow=c(2,2))
plot(residuals(fit3) ~ clothingSum$clo, ylab = "Residuals", xlab = "clo")
grid()
plot(residuals(fit3) ~ clothingSum$tInOp, ylab = "Residuals", xlab = "
   ↪ tInOp")
grid()
plot(residuals(fit3) ~ clothingSum$tOut, ylab = "Residuals", xlab = "tOut
   ↪ ")
grid()
plot(residuals(fit3) ~ as.factor(clothingSum$sex), ylab = "Residuals",
   ↪ xlab = "sex")
grid()
```

```
dev.off()

## Weighted analysis
Sigma = rep(1,nrow(clothingSum))
for (i in 1:10) {
  res.male = residuals(fit3)[clothingSum$sex == "male"]
  res.female = residuals(fit3)[clothingSum$sex != "male"]
  Sigma[clothingSum$sex == "male"] = var(res.male)
  Sigma[clothingSum$sex != "male"] = var(res.female)

  fit3 = lm(clo ~ tOut + tInOp + sex + tInOp:sex,data = clothingSum, weights = (Sigma^-1))
}

## Make sigma to matrix
Sigma = diag(Sigma)

## Compute param
Beta = solve(t(X) %*% solve(Sigma) %*% X) %*% t(X) %*% solve(Sigma) %*% Y
eps = Y - (X %*% Beta)

Beta
sqrt(diag(solve(t(X) %*% solve(Sigma) %*% X))) #* sqrt(sum(eps^2/df)) sigma included in estimates of Sigma


## Plots
fit.conf = predict(fit3, newdata = data.frame(clo = seq(0,1,length.out = 2*100),
                                         tInOp = mean(clothingSum$tInOp),
                                         tOut = rep(seq(10,35, length.out = 100),2),
                                         sex = as.factor(c(rep("male",100),rep("female",100)))),
                 interval = 'confidence')

png("figures/model_fin.png", width = 1000, height = 800, pointsize = 24)
par(mfrow = c(1,1))
plot(clothingSum$clo[clothingSum$sex=="male"] ~ clothingSum$tOut[clothingSum$sex=="male"], col = "blue", ylim = c(0.2,1),
     ylab = "clo", xlab = "tOut")
points(clothingSum$clo[clothingSum$sex!="male"] ~ clothingSum$tOut[clothingSum$sex!="male"], col = "red", pch = 2)
lines(fit.conf[1:100,1] ~ seq(10,35, length.out = 100), col = "blue")
lines(fit.conf[1:100,2] ~ seq(10,35, length.out = 100), col = "blue", lty = 2)
lines(fit.conf[1:100,3] ~ seq(10,35, length.out = 100), col = "blue", lty
  ↪  = 2)
lines(fit.conf[1:100,2]-1.96*sd(res.male) ~ seq(10,35, length.out = 100),
  ↪  col = "blue", lty = 3)
lines(fit.conf[1:100,3]+1.96*sd(res.male) ~ seq(10,35, length.out = 100),
  ↪  col = "blue", lty = 3)
lines(fit.conf[101:200,1] ~ seq(10,35, length.out = 100), col = "red")
lines(fit.conf[101:200,2] ~ seq(10,35, length.out = 100), col = "red",
  ↪ lty = 2)
lines(fit.conf[101:200,3] ~ seq(10,35, length.out = 100), col = "red",
  ↪ lty = 2)
lines(fit.conf[101:200,2]-1.96*sd(res.female) ~ seq(10,35, length.out =
  ↪ 100), col = "red", lty = 3)
lines(fit.conf[101:200,3]+1.96*sd(res.female) ~ seq(10,35, length.out =
  ↪ 100), col = "red", lty = 3)
grid()
legend("topright", c("Male","Female"),pch=c(1,2),lty=1, col = c("blue","
  ↪ red"))
dev.off()

fit.conf = predict(fit3, newdata = data.frame(clo = seq(0,1,length.out =
  ↪ 2*100),
                     tOut = mean(clothingSum$tOut),
                     tInOp = rep(seq(10,35, length.
                       ↪ out = 100),2),
                     sex = as.factor(c(rep("male"
                       ↪ ,100),rep("female",100)))
                       ↪ ,
         interval = 'confidence')

png("figures/model_fin2.png", width = 1000, height = 800, pointsize = 24)
par(mfrow = c(1,1))
plot(clothingSum$clo[clothingSum$sex=="male"] ~ clothingSum$tInOp[
  ↪ clothingSum$sex=="male"], col = "blue", ylim = c(0.2,1),
  ylab = "clo", xlab = "tInOp")
points(clothingSum$clo[clothingSum$sex!="male"] ~ clothingSum$tInOp[
  ↪ clothingSum$sex!="male"], col = "red", pch = 2)
lines(fit.conf[1:100,1] ~ seq(10,35, length.out = 100), col = "blue")
lines(fit.conf[1:100,2] ~ seq(10,35, length.out = 100), col = "blue", lty
  ↪  = 2)
lines(fit.conf[1:100,3] ~ seq(10,35, length.out = 100), col = "blue", lty
  ↪  = 2)
lines(fit.conf[1:100,2]-1.96*sd(res.male) ~ seq(10,35, length.out = 100),
  ↪  col = "blue", lty = 3)
lines(fit.conf[1:100,3]+1.96*sd(res.male) ~ seq(10,35, length.out = 100),
  ↪  col = "blue", lty = 3)
lines(fit.conf[101:200,1] ~ seq(10,35, length.out = 100), col = "red")
lines(fit.conf[101:200,2] ~ seq(10,35, length.out = 100), col = "red",
   ↪ lty = 2)
lines(fit.conf[101:200,3] ~ seq(10,35, length.out = 100), col = "red",
   ↪ lty = 2)
lines(fit.conf[101:200,2]-1.96*sd(res.female) ~ seq(10,35, length.out =
   ↪ 100), col = "red", lty = 3)
lines(fit.conf[101:200,3]+1.96*sd(res.female) ~ seq(10,35, length.out =
   ↪ 100), col = "red", lty = 3)
grid()
legend("topright", c("Male","Female"),pch=c(1,2),lty=1, col = c("blue","
   ↪ red"))
dev.off()

png("figures/res_anal2.png", width = 1000, height = 800, pointsize = 24)
par(mfrow=c(2,2))
plot(eps ~ clothingSum$clo, ylab = "Residuals", xlab = "clo")
grid()
plot(eps ~ clothingSum$tInOp, ylab = "Residuals", xlab = "tInOp")
grid()
plot(eps ~ clothingSum$tOut, ylab = "Residuals", xlab = "tOut")
grid()
plot(eps ~ as.factor(clothingSum$sex), ylab = "Residuals", xlab = "sex")
grid()
dev.off()


## subjId
par(mfrow = c(1,1))
png("figures/subjId_a.png", width = 1000, height = 800, pointsize = 24)
aux.col = c("red","blue")
boxplot(eps ~ clothingSum$subjId, ylab = "Residual", xlab = "subjId", col
   ↪  = aux.col[(clothingSum$sex == "male")+1])
grid()
dev.off()

## ANOVA
anova(lm(eps[clothingSum$sex == "male"] ~ clothingSum$subjId[clothingSum$
   ↪ sex == "male"])) # male
anova(lm(eps[clothingSum$sex == "female"] ~ clothingSum$subjId[
   ↪ clothingSum$sex == "female"])) #female



#### Problem B ----
setwd( '/Users/mads/Google␣Drev/Skole/Uni/10_semester/02424/Assignment1'
   ↪ )

if (file.exists("clothingFull.csv") & file.exists("clothingSum.csv")) {
  print("Files␣are␣acessable.␣You␣can␣proceed!")
  clothingFull <- read.csv("./clothingFull.csv", stringsAsFactors = FALSE
     ↪ )
  clothingSum <- read.csv("./clothingSum.csv", stringsAsFactors = FALSE)
}

clothingSum$subjId = as.factor(clothingSum$subjId)

## Test type II
fit0 <- lm(clo ~ tOut*tInOp*subjId, data = clothingSum)

drop1(fit0,test="F")
fit1 <- update(fit0,.~.-tOut:tInOp:subjId)
summary(fit1)
drop1(fit1,test="F")
fit2 <- update(fit1,.~.-tInOp:subjId)

drop1(fit2,test="F")
fit3 <- update(fit2,.~.-tOut:subjId)

drop1(fit3,test="F")
fit4 <- update(fit3,.~.-tOut:tInOp)

drop1(fit4,test="F")
fit5 <- update(fit4,.~.-tInOp)

drop1(fit5,test="F")
anova(fit1,fit5)

## Final model:
summary(fit5)

png("figures/subjid.png", width = 1000, height = 800, pointsize = 24)
plot(clothingSum$clo ~ clothingSum$tOut, cex = 0, xlab = "tOut", ylab = "
   ↪ clo")
id = unique(clothingSum$subjId)
theta = fit5$coefficients[2]
coef = fit5$coefficients[-2]
coef[2:length(coef)] = tail(coef,-1)+coef[1]
cols = rainbow(length(coef))
for (ii in 1:length(coef)) {
  points(clothingSum$clo[clothingSum$subjId == id[ii]] ~ clothingSum$tOut
     ↪ [clothingSum$subjId == id[ii]], col = cols[ii], cex = 0.5)
  abline(a=coef[ii], b = theta, col = cols[ii], lwd = 0.3)
}
grid()
dev.off()
```

```
png("figures/intercept_b.png", width = 800, height = 800, pointsize = 24)
hist(coef,xlab = "Intercept")
grid()
dev.off()


#### Problem C ----
setwd( '/Users/mads/Google␣Drev/Skole/Uni/10_semester/02424/Assignment1'
   ↪ )

if (file.exists("clothingFull.csv") & file.exists("clothingSum.csv")) {
  print("Files␣are␣acessable.␣You␣can␣proceed!")
  clothingFull <- read.csv("./clothingFull.csv", stringsAsFactors = FALSE
     ↪ )
  clothingSum <- read.csv("./clothingSum.csv", stringsAsFactors = FALSE)
}

clothingFull$subjId = as.factor(clothingFull$subjId)

## Version 1 (based on sex)
fit0 <- lm(clo ~ tOut*tInOp*sex, data = clothingFull)
drop1(fit0,test="F")
## It seems we are already done, fit0 is final model!
fit.v1 = fit0
summary(fit.v1)

## Version 2 (based on sex)
fit0 <- lm(clo ~ tOut*tInOp*subjId, data = clothingFull)
drop1(fit0,test="F")
fit1 <- update(fit0,.~.-tOut:tInOp:subjId)

drop1(fit1,test="F")
## It seems we are already done, fit1 is final model!
fit.v2 = fit1
summary(fit.v2)

## Lets compare version 1 and 2
anova(fit.v1,fit.v2)
# Version 2 is much better!

## Final model:
summary(fit5)
```