```
# Multivariate Statistics: Exercise 4

library(robustbase)

attach(milk)

# 1(a): Least Squares
LS<- lm(X4~X5)
plot(X5,X4)
abline(coefficients(LS),col='red')

# 1(b): Least Trimmed Sum of Squares
LTS <- ltsReg(X4~X5) # Least S
i0LTS <- LTS$lts.wt==0
plot(X5,X4, col=ifelse(i0LTS, "red", "black"))
abline(coefficients(LTS),col='green')
text(X5[i0LTS],X4[i0LTS], row.names(milk)[i0LTS], cex=0.6, pos=4, col="red")

# 1(c): "maximum likelihood type"
MM <- lmrob(X4~X5)
i0MM <- MM$rweights==0
plot(X5,X4, col=ifelse(i0MM, "red", "black"), cex=0.5 + MM$rweights)
abline(coefficients(MM),col='green')
text(X5[i0MM],X4[i0MM], row.names(milk)[i0MM], cex=0.6, pos=4, col="red")

LTS$lts.wt - MM$rweights
row.names(milk)[i0LTS]
row.names(milk)[i0MM]

# 1(d)
plot(LS)
# P1: point 70 outside of +/- 2.5 band
# P2: 70 doesn't fit normal assumption
# P3: 70 far away from rest
# P4: cook's distance of 70 large -> leverage point

# QUESTIONS:
# What is sqrt(standardized residuals).
# Shouldn't it the sqrt from the abs value?

# alternative plot
# par(mfrow=c(2,2),ask=F)
# plot(LM,ask=F)

plot(LTS)
# P1: 1,2,41,70,75 don't fit normal assumption
# P2,3: residuals of 1,2,41,70,75 are outside of +/- 2.5 band
# P4: 1,2,41,75 -> vertical outliers; 14,15 -> good leverage points;
#     70 -> bad leverage point

plot(MM)
# P1: 4 vertical outliers (no numbers); 1 bad LP
# P2: 5 points don't fit normal assumption
# P3: 4 points that seem to have weight 0
# P4: 5 points out of borders (why ~+/-0.5?)
# P5: 4 points far away from rest

# 2(a)
LS.full<- lm(X4~., data=milk)
```

```
LTS.full<- ltsReg(X4~., data=milk)
MM.full<- lmrob(X4~., data=milk)

plot(X4~LS.full$fitted.values)
abline(0,1,col='green')

plot(X4~LTS.full$fitted.values)
abline(0,1,col='green')

plot(X4~MM.full$fitted.values)
abline(0,1,col='green')

# When using all variables as predictors even LS looks not that bad.

# 2(b)
LS.R2 <- 1-sum((X4-LS.full$fitted.values)^2)/sum((X4-mean(X4))^2)

LTS.mean <- sum(LTS$lts.wt*X4)/sum(LTS$lts.wt)
LTS.R2 <- 1-sum(LTS$lts.wt*(X4-LS.full$fitted.values)^2)/sum(LTS$lts.wt*(X4-LTS.mean)^2)

MM.mean <- sum(MM$rweights*X4)/sum(MM$rweights)
MM.R2 <- 1-sum(MM$rweights*(X4-LS.full$fitted.values)^2)/sum(MM$rweights*(X4-LTS.mean)^2)

LS.R2
LTS.R2
MM.R2
# What is the difference to (adjusted) R-squared in summary?

# 2c
summary(LS.full)
# For a model with only one exlanatory variable (predictor) X3 would be the best choice
# Adjusted R-squared is pretty high: 0.85
# Waht is the difference to multiple R-squared?

summary(LTS.full)
# For a model with only one predictor X3 would be again the best choice
# significant predictors: X3,X5,X6,X7
# Pretty good fit: Adjusted R-squared 0.97

summary(MM.full)
# For a model with only one predictor X3 would be again the best choice
# significant predictors: X3,X5,X6,X7
# Pretty good fit: Adjusted R-squared 0.97
# outlier indentification: 1,2,41,70

# 2(d)
plot(LS.full)
# P1: residuals of 1,2,41 are high but within +/- 2.5 band
# P2: Q-Q-plot: 1,2,41 do not fit normal assumption
# P3: Again 1,2,41 are far away from the rest
# P4: Cook#s distance large for 70

plot(LTS.full)
# P1: QQ-plot: 1,2,41,70 do not fit normal assumption
# P2,3: residuals of 1,2,41,70 are outside of +/- 2.5 band
# P4: 1,2,41 -> vertical outliers; a lot of good leverage points;
#     70 -> bad leverage point

plot(MM.full)
# P1: 3 vertical outliers (no numbers); some good LP; 1 bad LP
```

```
# P2: 4 points don't fit normal assumption
# P3: 4 points that seem to have weight 0
# P4: 4 points out of borders (why ~+/-0.5?)
# P5: 4 points far away from rest
```