```r
################################################################
### The aim of this practical is to get familiar with simple ###
### regression analysis using R                            ###
################################################################
# Clear workspace
rm(list=ls())

# Read data in
d <- read.csv("F:\\Opetus\\AAM2017\\Harjoitukset\\AirTemperatureData.csv", sep=";")

### Basic data checks
names(d) # prints out variable names
str(d) # structure of the data frame
summary(d) # summaries the variable's variation

# Plot air temperatures against elevation
plot(d$elev, d$temp, xlab="Elevation", ylab="Air temperature")

# Fit a simple linear regression to examine is there a relationship
# between air temperatures and elevation
model1 <- lm(temp~elev, data=d)

# Object "model1" stores a lot of information
names(model1)

# Extract estimated coefficients and R-squared
summary(model1)

# The model summary can also be stored as an object
summ <- summary(model1)
names(summ)
summ$r.squared

# Conduct anova for significance testing
anova(model1, test="F")

# Add fitted regression line to the existing plot
abline(model1, col="red", lwd=2)

# Add R-squared value to the plot
mtext(side=3, text = paste("R-squared =",
                           round(summ$adj.r.squared, 2)))

# Extract and plot model residuals
res1 <- resid(model1)
hist(res1, main="Model1 residuals", xlab="Error term")

# Calculate and plot mean error
mean(res1)
abline(v=mean(res1), col="blue", lwd=2)

# Plot regression diagnostics
par(mfrow=c(2,2))
plot(model1)

##################################################################
### Let's examine whether air temperature-elevation is curvilinear ###
##################################################################
# Define a model including elevation's second order term
model2 <- lm(temp~elev+I(elev^2), data=d)

# Print model results
summ2 <- summary(model2)
print(summ2)
anova(model2, test="F")

# Is model2 significantly better than model1?
# In another words, are the sum of squared errors
# significantly different between the two models?
anova(model1, model2)

### Plot regression line
# First return to normal plot settings
par(mfrow=c(1,1))
```

```
plot(d$elev, d$temp, xlab="Elevation", ylab="Air temperature")

### Plotting a model with higher order terms is a bit more complicated,
### that with a model consisting only first order terms
# First create a new data frame with elevation values from 0 to 1100
new <- data.frame(elev=seq(0, 1100, 1))

# Use the model to predict the values to the new data
p <- predict(model2, new)

# Add line to existing plot
lines(p, col="blue", lwd=2)

# Add R-squared value to the plot
mtext(3, text = paste("R-squared =",
                      round(summ2$adj.r.squared, 2)))

#################
##### EXTRA #####
#################
### Linear regression model can be used to predict values, when background
### values (here elevation) are known
# Based on our temperature model consisting of 1) first and 2) second
# order polynomial terms, what is air temperature at 3000 m a.s.l.?
predict(model1, data.frame(elev=3000))
predict(model2, data.frame(elev=3000))

plot(d$elev, d$temp, xlab="Elevation", ylab="Air temperature",
     ylim=c(-30, 62), xlim=c(0, 3000),
     main = "Extrapolating outside data range :/")

abline(model1, col="red", lwd=2)

new <- data.frame(elev=seq(0, 3000, 1))
p <- predict(model2, new)
lines(p, col="blue", lwd=2)
```