Prof. Andreas Orth
High Integrity Systems, 2nd semester: Data Mining
Assignment 5   week 5

Exercise 1:
For the shoesize/bodysize example in Assignment 4 construct the following matrices:

```
> shoesize<-c(38,38,39,39,40,40,41,41,42,42)
> bodysize<-c(153,161,167,169,173,176,182,181,188,189)
```

X is a matrix with 2 columns, one of which is filled with 1's, and the other is the vector of x-values (shoesize-values),
X^T, which then has two rows (the first of which is just 1's) consisting of 10 values,
X^TX, whichis a 2 by 2 matrix,
$(X^TX)^{-1}$, which of course is also 2 by 2,
X^Ty where y is the vector of y-values (bodysize-values) – X^Ty will be a 2-vector,
$b = (X^TX)^{-1} X^Ty$, this will also be a 2-vector of coefficients,
$y_{pred} = X(X^TX)^{-1} X^Ty$, this will of course be a 10-vector of bodysize-predictions.

Now calculate the residual standard error using
```
  ss_res<-sum((ypred - y)^2)
  ms_res<-ss_res / 8
  rse<-sqrt(ms_res)
```

Using the two diagonal elements of $(X^TX)^{-1}$, calculate the standard error of each of the two coefficients (don't forget to take sqare roots).

Now calculate a t-value for each coefficient (i.e coefficient / its standard error), and from this a p-value, using the t-distribution, `pt(...,8,lower.tail=FALSE)` for 8 degrees of freedom.

Exercise 2:
Replay the above example using the `lm( )`-function.
Do the p-values coincide? Why or why not?

Do the following diagnostic plots:
(a) x-values on the x-axis and residuals on the y-axis,
(b) predictions on the x axis and residuls on the y-axis,
```
   plot(fitted(fm), resid(fm), main="Residuals vs Fitted",
   xlab="Fitted values", ylab="Residuals")
```
(c) predictions on the x axis and observed values on the y-axis,
(d) histogram of the residuals,
(e) normal scores plot to check for skewness, kurtosis and outliers
```
   qqnorm(resid(fm), main="Residuals Rankit Plot")
```