TMA4255 Applied Statistics
Exercise 4

In Table 1 data for 31 trees of a certain kind in a national park in the US are given. Three variables are measured for each tree. These are:

D: The diameter of the tree measured in inches 1.5 m above ground level
H: The height of the tree measured in feet
V: The volume of the tree measured in cubic feet

Obs.	D	H	V	Obs.	D	H	V
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

Table 1: The data set

The problem if one wants to measure the volume of a tree is that it has to be cut down. Therefore it is of interest to develop a model that can be used to estimate the tree volume without having to cut it. The data set is available at the Exercises tab at the course www-page.

MINITAB: Click Open worksheet and then `data4.MTW`.
R: `ds=read.csv("http://www.math.ntnu.no/~erikblys/TMA4255-V14/Data/data4.csv")`

a) Assume that $\mathrm{E}(V_i) = \beta_0 + \beta_1 H + \beta_2 D$, and fit this model.

MINITAB: If C1 contains the volume, C2 the height and C3 the diameter, a regression analysis can be performed by the command MTB:REGRESS C1 2 C2 C3 C4 C5 or by the menus. Residuals will then be in C4 and the fitted values in C5.
`R: fit=lm(Volume~.,data=ds)`. Standardized residuals by `rstandard(fit)` and (exernally) studentized residuals by `rstudent(fit)`.

What does the fitted model look like? Can it be used for all values of D and H? Assume that the residuals are independent and $N(0, \sigma^2)$.

Test if the model explains a significant amount of the variation in the response. Should both variables be included in the model? Choose the level of significance yourself. Plot the standardized (or studentized) residuals against the fitted values and against each of the predictor variables. Comment.

b) Let us assume that we want to add the IQ of the lumberjack that cut down the tree as a covariate in the model in a). This should for obvious reasons not be a good predictor for the volume of the tree. To mimic this situation we simulate (as explained in Exercise 1) new data to resemble the IQ of different lumberjacks by drawing data from the normal distribution with mean 100 and standard deviation 16, and since we have 31 trees we simulate 31 observations.

Then add this new covariate to the model in a) and fit the new model. Look at the model fit (coefficients and p-values) and also at R^2 and R^2_{adj}. What do you observe?

c) Based on the formula for the volume of a cylinder, it was suggested to introduce the variable $X = D^2H$ and regress V against it. Perform this regression. Decide if the model should be fitted with or without the intercept.

Now consider the model without the intercept. Check the residuals.

(Theory:) Derive a theoretical expression for the least squares estimator of the slope and find also an expression for its variance.

Assume that the residuals are independent and $N(0, \sigma^2)$ and find a 95% prediction interval for the volume when $D = 15$ and $H = 80$. Perform the calculations by hand and check your results by comparing them to using the statistical software.

R: (since a bit difficult to get newdata on correct format)

```
x <- ds$Diameter^2*ds$Height
fitx2 <- lm(ds$Volume~x-1) # NB: -1 to remove the intercept
plot(fitx$fitted,rstudent(fitx))
plot(rstudent(fitx))
qqnorm(rstudent(fitx))
newobs <- data.frame("x"=15^2*80)
predict(fitx2,newdata=newobs,interval="prediction")
```

d) Based on the relationship between the volume, diameter and height given in c) another statistician suggested that one should model the relationship between the logarithms of the three variables. Do you expect a constant term in this model?

Perform the regression analysis. What are the differences between this model and the model in c) with respect to the residuals being independent $N(0, \sigma^2)$ in a linear regression model? Look at the plots of the residuals and comment.

Use software to find a 95% prediction interval for lnV when $D = 15$ and $H = 80$. From this, find a 95% prediction interval for V. Compare with c).