# TMA4255 Applied Statistics
# Solution to Exercise 6

**a)**

We perform a regular linear regression

```
The regression equation is
Volume = - 58.0 + 4.71 Diameter + 0.339 Height

Predictor        Coef       StDev          T        P
Constant      -57.988       8.638      -6.71    0.000
Diameter       4.7082      0.2643      17.82    0.000
Height         0.3393      0.1302       2.61    0.014

S = 3.882       R-Sq = 94.8%     R-Sq(adj) = 94.4%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      7684.2      3842.1    254.97    0.000
Residual Error    28       421.9        15.1
Total             30      8106.1

Source       DF      Seq SS
Diameter      1      7581.8
Height        1       102.4

Unusual Observations
Obs   Diameter     Volume         Fit   StDev Fit    Residual    St Resid
31       20.6     77.000      68.515       1.850       8.485       2.49R

R denotes an observation with a large standardized residual
```

The fitted model is then

$$\hat{V} = -58 + 4.71D + 0.339H. \tag{1}$$

We see that for small $D$ and $H$, $\hat{V}$ is negative, and this is physically not right.

We assume the error terms to be independent and normally distributed. We test the hypothesis

$$H_0: \quad \beta_D = \beta_H = 0 \tag{2}$$

against

$$H_1: \quad \text{at least one} \neq 0. \tag{3}$$

From the print-out we see that

$$P(F_{2,28} \geq 254.97) = 0,000 \tag{4}$$

Which means that we reject $H_0$ and claim that the model has a significant degree of explanatory power. We look at the residuals plotted against the fitted values and the two explanatory variables, given in Figure 1.

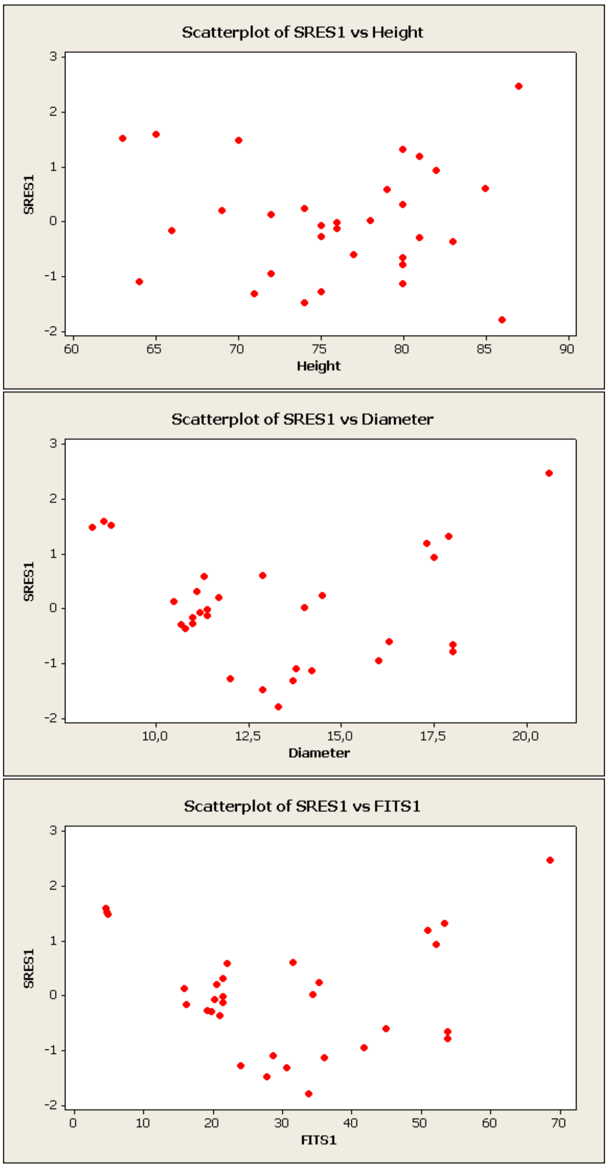


Figure 1: Residual plot, a)

The residuals plotted against the height looks independent, while the two other plots have a hint of U-shape and thereby dependence. This is unfortunate for the model we have chosen and we should consider looking for a better model.

**b)** We introduce simulated data to this model - that has no relationship with the response, to see how this influences our model and fit. This means that the results will differ for each student simulating data (unless the data are simulated with the same seed).

The data for IQ was put into column C4 by Calc-Random data-Normal and choosing 31 data points in C4 with mean 100 and sd 16.

```
The regression equation is
Volume = - 56,8 + 4,70 Diameter + 0,350 Height - 0,0175 IQ


Predictor       Coef  SE Coef      T      P
Constant     -56,832    9,217  -6,17  0,000
Diameter      4,6961   0,2699  17,40  0,000
Height        0,3496   0,1345   2,60  0,015
IQ          -0,01752  0,04296  -0,41  0,687


S = 3,94095   R-Sq = 94,8%   R-Sq(adj) = 94,3%
```

We see that all the regression coefficient estimates have changed, and that IQ is not significant (you may get a different results since you have simulated other IQ-data that in the print-out above).

The $R^2$ is unchanged at 94.8%, but the $R^2_{adj}$ has decreased from 94.4 to 94.3%. You may get a slightly different result, and you may even see an increase in $R^2$. You may try simulating data once more and fit again, to evaluate the difference in the results.

The adjusted coefficient of determination is defined by

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1}. \tag{5}$$

This is an adjusted version of the coefficient of determination, and the coefficient of determination, $R^2$, indicates how much of the variation in the data that are explained by the model. The adjusted oefficient of determination takes into account the number of parameters fitted. It will always be the case that adding a new variable (even if is is only noise) will increase the $R^2$ or keep it unchanged, but not necessarily increase or change the $R^2_{adj}$. Note that Minitab gives this in percent.

**c)** We perform the regression analysis with the new model.

```
The regression equation is
Volume = - 0.298 + 0.00212 D^2*H

Predictor         Coef        StDev          T        P
Constant       -0.2977       0.9636      -0.31    0.760
D^2*H       0.00212437   0.00005949      35.71    0.000

S = 2.493       R-Sq = 97.8%     R-Sq(adj) = 97.7%
```

```
Analysis of Variance

Source            DF        SS        MS        F       P
Regression         1    7925.8    7925.8  1275.27   0.000
Residual Error    29     180.2       6.2
Total             30    8106.1

Unusual Observations
Obs     D^2*H    Volume        Fit   StDev Fit    Residual    St Resid
31      36919    77.000     78.133       1.416      -1.133      -0.55 X

X denotes an observation whose X value gives it large influence.
```

We see from the p-value of the constant that is no reason to include a constant. We do the analysis without the constant term.

```
The regression equation is
Volume = 0.00211 D^2*H

Predictor        Coef       StDev          T        P
Noconstant
D^2*H      0.00210810  0.00002722      77.44    0.000

S = 2.455
Analysis of Variance

Source            DF        SS        MS        F       P
Regression         1     36144     36144  5996.41   0.000
Residual Error    30       181         6
Total             31     36325

Unusual Observations
Obs     D^2*H    Volume        Fit   StDev Fit    Residual    St Resid
31      36919    77.000     77.830       1.005      -0.830      -0.37 X

X denotes an observation whose X value gives it large influence.

Predicted Values

Fit  StDev Fit        95.0% CI             95.0% PI
37.946      0.490   (  36.945,  38.947)  (  32.833,  43.059)
```

The residual plots are given in Figure 2. The residuals plotted against the fitted values shows that the variance increases slightly towards the right. This indicates that our assumptions of equal variance might not hold for this model.

We will derive a theoretical expression for the least squares estimators of the slope when no intercept is present. Let

$$SSE = \sum_{i=1}^{n} (y_i - b_1 x_i)^2. \tag{6}$$

We differentiate with respect to $b_1$ and set it equal to zero. The expression becomes

$$\frac{\partial SSE}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_1 x_i) x_i = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}. \tag{7}$$

The variance is given by

$$\mathrm{Var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 \sigma^2}{(\sum_{i=1}^{n} x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}. \tag{8}$$

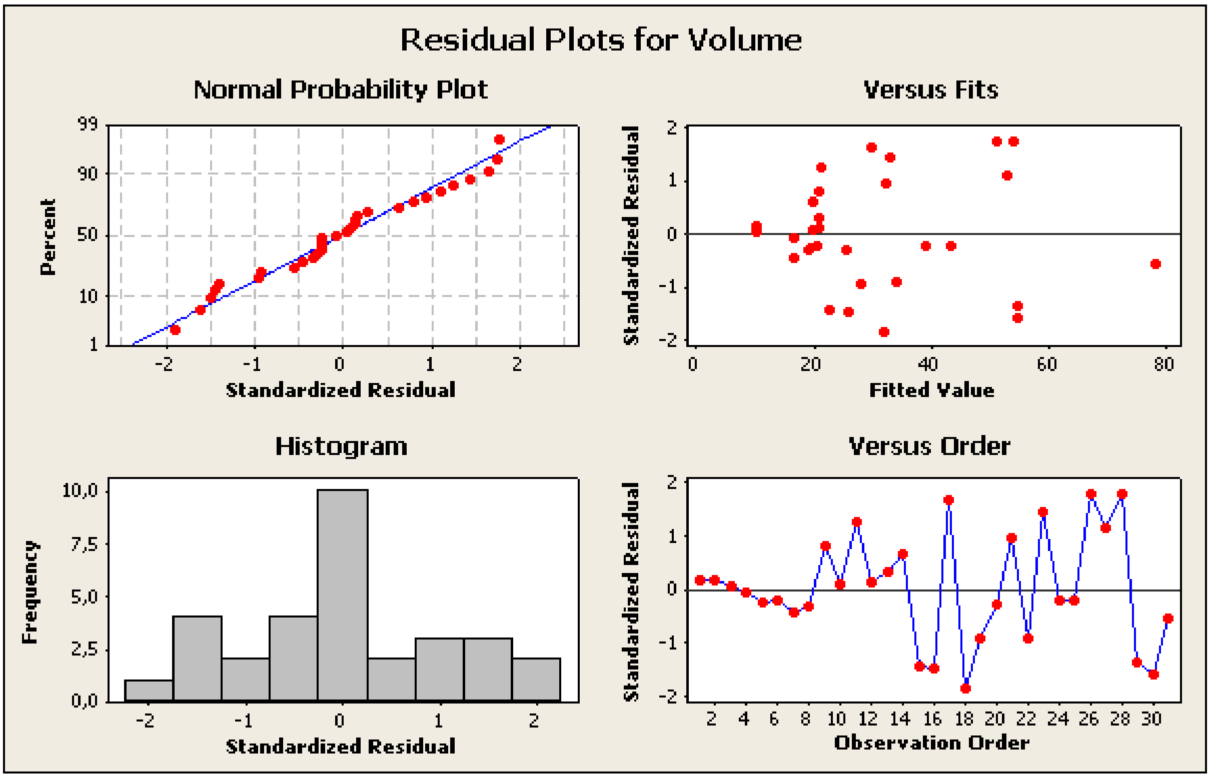


Figure 2: Residual plots v)

We will find the prediction interval when $D = 15$ and $H = 80$. Predicted value is $\hat{y}_0 = 15^2 \cdot 80 \cdot 0.002108 = 37.98$.

$$S_D(y_0 - \hat{y}_0) = \sqrt{\hat{\mathrm{Var}}(Y_0) + \hat{\mathrm{Var}}(\hat{Y}_0)} = \sqrt{S^2 + (D^2 \cdot H)^2 \cdot \mathrm{Var}(\hat{\beta})}, \tag{9}$$

Which gives $S_D(y_0 - \hat{y}_0) = 2.504$. From the table we have $t_{0.025,30} = 2.042$. This gives a 95% prediction interval for $y_0$ of $37,98 \pm 2,504 \cdot 2,042 = [32.87 \quad 43.09]$. This is the same as was given by the software.

**d)** The new model is expressed by $E(V) = konst \cdot D^2 H$. The logarithm of this expression is

$$\ln\, E(V) = \ln\, konst + 2\ln\, D + \ln\, H. \tag{10}$$

From this expression we see that it is natural to include the constant term.

Linear regression in the can be expressed mathematically as

$$V = konst \cdot D^2 H + \epsilon. \tag{11}$$

This implies an additive error model. The linear regression in the logarithmic model becomes

$$\ln\, V = \ln\, konst + 2\ln\, D + \ln\, H + \epsilon \Leftrightarrow V = konst \cdot D^2 H \cdot \epsilon, \tag{12}$$

which implies a multiplicative error model. The residual plots are given in Figure 3. The residual plots look better than for the model with $D^2 H$.
We perform the analysis in Minitab, and get

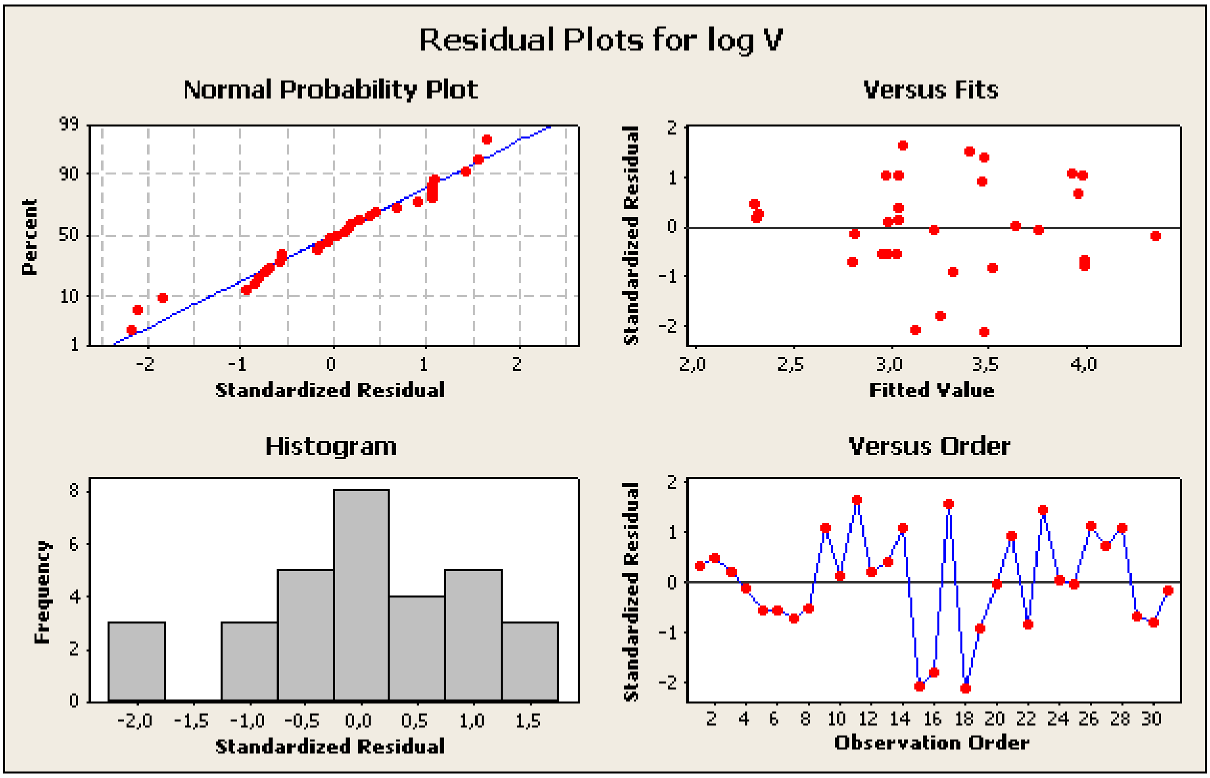


Figure 3: Residual plots d)

```
The regression equation is
LogVolume = - 6.63 + 1.98 LogDiameter + 1.12 LogHeight

Predictor        Coef       StDev          T        P
Constant      -6.6316      0.7998      -8.29    0.000
LogDiame      1.98265     0.07501      26.43    0.000
LogHeigh       1.1171      0.2044       5.46    0.000

S = 0.08139     R-Sq = 97.8%     R-Sq(adj) = 97.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      8.1232      4.0616    613.19    0.000
Residual Error    28      0.1855      0.0066
Total             30      8.3087

Source       DF      Seq SS
LogDiame      1      7.9254
LogHeigh      1      0.1978

Unusual Observations
Obs    LogDiame    LogVolum         Fit    StDev Fit     Residual     St Resid
15         2.48      2.9497      3.1182       0.0154      -0.1686       -2.11R
18         2.59      3.3105      3.4751       0.0288      -0.1645       -2.16R

R denotes an observation with a large standardized residual
```

```
Predicted Values

Fit  StDev Fit         95.0% CI              95.0% PI
3.6326     0.0182   (  3.5953,  3.6700)  (  3.4618,  3.8035)
```

When we transform back we get the prediction interval [31.87   44.85]. We see that the interval is a hint wider than for the model in c).