TMA4255 Applied statistics
Solution exercise 5

Problem a)

Stat → Basic Statistics → Display Descriptive Statistics :

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Vekt	18	3,000	3,000	3,000	2,114	0,498
Alder	18	5,056	5,000	5,000	2,754	0,649
Tilførse	18	4,000	4,000	4,000	2,142	0,505
Temp.	18	2,000	2,000	2,000	0,840	0,198
Trommel	18	3,000	3,000	3,000	0,840	0,198
Y	18	4,556	4,000	4,375	2,121	0,500

Variable	Minimum	Maximum	Q1	Q3
Vekt	0,000	6,000	1,000	5,000
Alder	1,000	10,000	2,750	7,250
Tilførse	1,000	7,000	2,000	6,000
Temp.	1,000	3,000	1,000	3,000
Trommel	2,000	4,000	2,000	4,000
Y	2,000	10,000	3,000	6,000

Stat → Basic Statistics → Correlations :

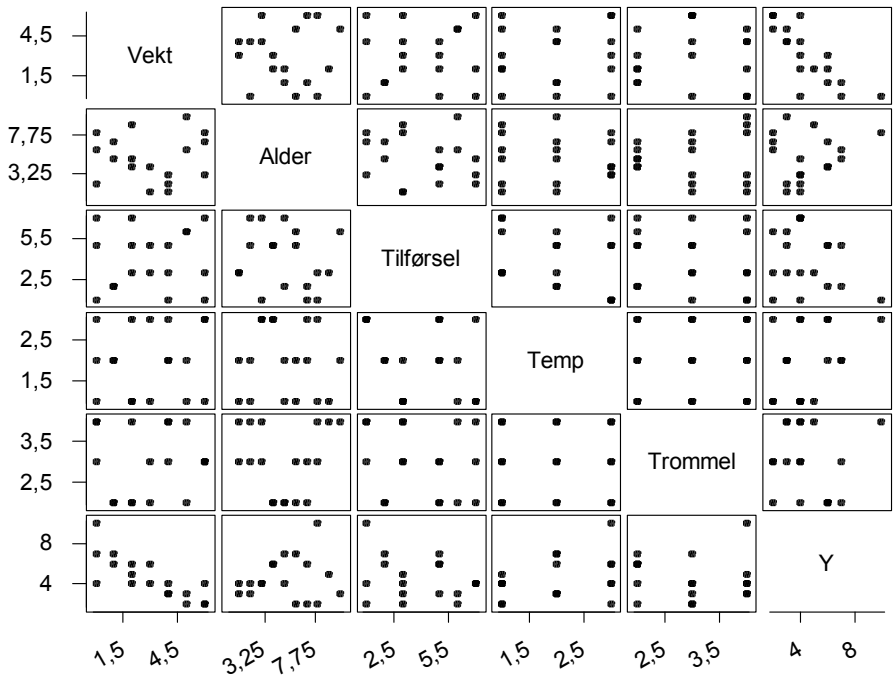
Correlations (Pearson)

	Vekt	Alder	Tilførse	Temp	Trommel
Alder	0,020				
	0,937				
Tilførse	0,052	-0,189			
	0,838	0,451			
Temp	0,099	-0,051	-0,294		
	0,695	0,841	0,236		
Trommel	0,033	0,051	-0,196	0,000	
	0,896	0,841	0,435	1,000	
Y	-0,787	0,145	-0,259	0,363	-0,066
	0,000	0,565	0,299	0,139	0,795

Cell Contents: Correlation
P-Value

From the correlation matrix we see that only Corr(Vekt, Y) is significantly different from zero. (The p-value gives the probability of observing the given values or something more extreme, when the null hypethesis is assumed true)

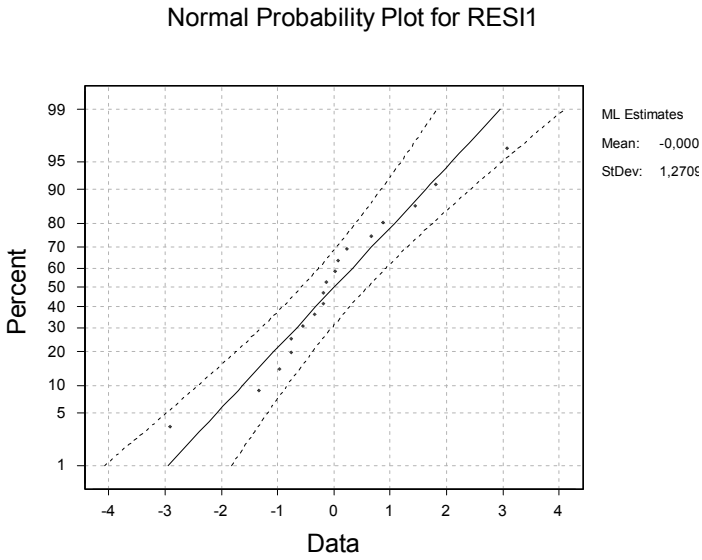
With Graph → Matrix we can plot the different variables against each other. Here we can also see that there is very little dependence between the variables, except Y and Vekt. We also see how Trommel and Temperatur are controlled to give zero correlation.

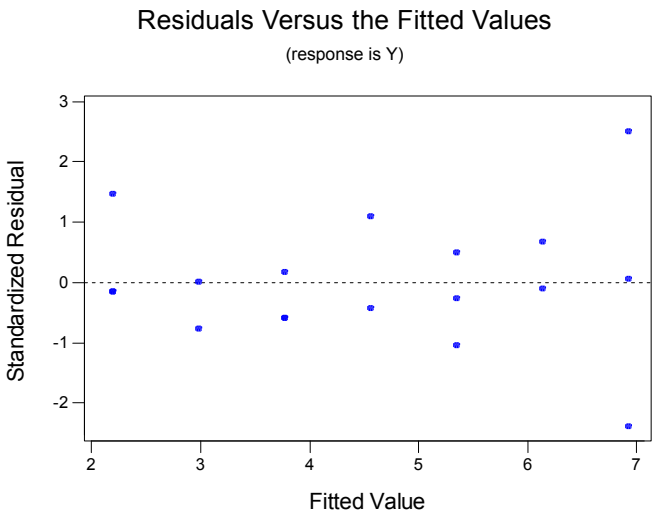


Problem b)

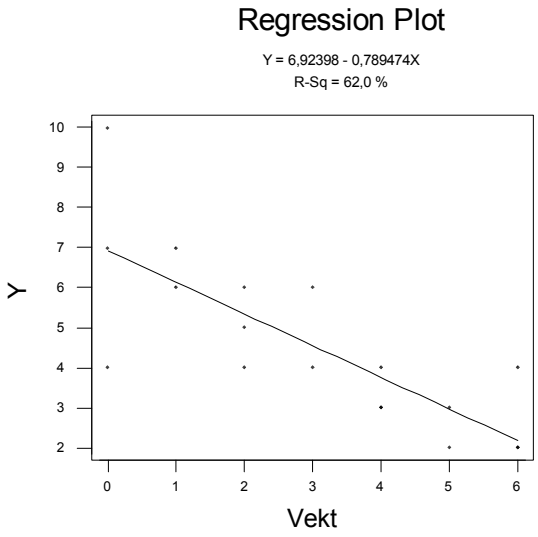
From part a) it seems natural to choose Vekt as an explanatory variable, since Y seems to be independent of the other variables. Stat → Regression gves:

```
The regression equation is
Y = 6,92 - 0,789 Vekt

Predictor         Coef       StDev           T         P
Constant       6,9240      0,5623       12,31     0,000
Vekt          -0,7895      0,1546       -5,11     0,000

S = 1,348       R-Sq = 62,0%     R-Sq(adj) = 59,6%

Analysis of Variance

Source             DF          SS          MS          F         P
Regression          1      47,368      47,368      26,07     0,000
Residual Error     16      29,076       1,817
Total              17      76,444

Unusual Observations
Obs       Vekt           Y         Fit   StDev Fit    Residual    St Resid
  8       0,00      10,000       6,924       0,562       3,076        2,51R
 12       0,00       4,000       6,924       0,562      -2,924       -2,39R
```





The Normal plot suggests that the residuals are normally distributed and from the Residuals vs. Fits they seem to be independent. Therefore the assumptions seem reasonable.



The resulting coefficients of the regression are significantly different from zero, and we can conclude that Vekt has an impact on the response Y. The unusual residuals that MINITAB tells us about, are due to the very high variance of the residuals.

Oppgave c)

Confidence interval for the expectation $\mu_{\hat{Y}|x=x_0}$

We use that $\hat{Y} \sim N(\mu_{\hat{Y}}, \sigma^2_{\hat{Y}})$ where

$$\mu_{\hat{Y}} = E(\hat{Y} \mid x = x_0) = E(\hat{\alpha} + \hat{\beta}x_0) = \alpha + \beta x_0 = \mu_Y$$

$$\sigma^2_{\hat{Y}} = Var(\hat{\alpha} + \hat{\beta}x_0) = Var(\overline{Y} + \hat{\beta}(x_0 - \overline{x})) = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right].$$

σ^2 is unknown, so we estimate it by $S^2 = \frac{1}{n-2}\sum(Y_i - \overline{Y})^2$.

Therefore we can use the T-statistic

$$T = \frac{\hat{Y}_0 - \mu_{\hat{y}}}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim t_{n-2}$$

and the confidence inerval for $\mu_{\hat{Y}|x=x_0}$ becomes $\left[\hat{y}_0 \pm t_{\alpha/2,n-2} \cdot S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right]$

The prediction interval for a new observation $Y^*_{|x_0}$:

We use that $\hat{Y}_0 - Y_0^* \sim N(\mu_{\hat{Y}_0 - Y_0^*}, \sigma^2_{\hat{Y}_0 - Y_0^*})$, wher

$$\mu_{\hat{Y}_0 - Y_0^*} = E\left[\hat{Y}_0 - Y_0^*\right] = \mu_{\hat{Y}_0} - \mu_{Y_0^*} = 0$$

$$\sigma^2_{\hat{Y}_0 - Y_0^*} = \ldots = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)$$

The test statistic is

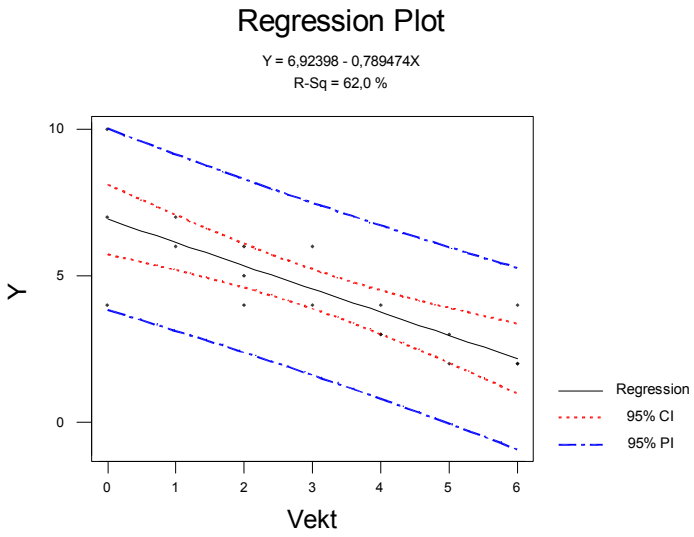
$$T = \frac{\hat{Y}_0 - Y_0^*}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim t_{n-2}$$

and the predition interval for $Y^*_{|x_0}$ becomes

$$\left[\hat{y}_0 \pm t_{\alpha/2,n-2} \cdot S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right]$$

Both intervals are minimal when $x_0 = \overline{x}$

Predition equation plotted with confidence interval in MINITAB



Because the confidence interval only says something about the expectation of the observations this interval is quite narrow. The prediction interval, however is quite wide, and this is du to the big estimated variance.

For Vekt = 1,2,3 we get the following prediction intervals

```
Predicted Values
 Vekt     Fit  StDev Fit          95,0% CI              95,0% PI
   1   6,135      0,443   (   5,195;   7,074)  (   3,126;   9,143)
   3   4,556      0,318   (   3,882;   5,229)  (   1,619;   7,492)
   5   2,977      0,443   (   2,037;   3,917)  (  -0,032;   5,985)
```

Oppgave d)

When we perform the regression we choose to save X'X inverse under storage. That way we can print the covariance matrix by clicking Data->Display Data

Regression Analysis

The regression equation is
Y = 4,90 - 0,831 Vekt + 0,135 Alder - 0,066 Tilførsel + 1,10 Temp
- 0,153 Trommel

Predictor	Coef	StDev	T	P
Constant	4,897	1,438	3,40	0,005
Vekt	-0,8308	0,1108	-7,50	0,000
Alder	0,13454	0,08633	1,56	0,145
Tilførse	-0,0662	0,1189	-0,56	0,588
Temp	1,0971	0,2934	3,74	0,003
Trommel	-0,1530	0,2822	-0,54	0,598

S = 0,9554 R-Sq = 85,7% R-Sq(adj) = 79,7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	65,491	13,098	14,35	0,000
Residual Error	12	10,953	0,913		
Total	17	76,444			

Source	DF	Seq SS
Vekt	1	47,368
Alder	1	1,992
Tilførse	1	2,792
Temp	1	13,071
Trommel	1	0,268

Unusual Observations

Obs	Vekt	Y	Fit	StDev Fit	Residual	St Resid
8	0,00	10,000	8,587	0,646	1,413	2,01R
17	6,00	2,000	3,620	0,562	-1,620	-2,10R

R denotes an observation with a large standardized residual

Data Display

Matrix XPXI1

2,26697	-0,01773	-0,05553	-0,11667	-0,25899	-0,29761
-0,01773	0,01345	-0,00047	-0,00146	-0,00454	-0,00177
-0,05553	-0,00047	0,00817	0,00238	0,00326	-0,00013
-0,11667	-0,00146	0,00238	0,01548	0,01237	0,00746
-0,25899	-0,00454	0,00326	0,01237	0,09429	0,00602
-0,29761	-0,00177	-0,00013	0,00746	0,00602	0,08724

Comment:
We see that the variance of the estimators of the coefficients is not particularly high, and this suggests that coefficients with a high p-value are actually very close to zero.

Problem e)

Stat → Regression → Best Subsets

The best alternative variable selections are listed with to alternatives for each number of variables,with different measures of how good they are.

R-Sq:

Adj. R-Sq:

C-p:

Best Subsets Regression

```
Response is Y
```

Vars	R-Sq	Adj. R-Sq	C-p	s	Vekt	Alder	Tilføre	Trom Temp
1	62,0	59,6	17,9	1,3481	X			
1	13,2	7,8	58,7	2,0366			X	
2	81,6	79,2	3,4	0,96729	X		X	
2	66,7	62,3	15,9	1,3020	X	X		
3	85,1	81,9	2,5	0,90282	X	X		X
3	82,4	78,7	4,7	0,97944	X		X	X
4	85,3	80,8	4,3	0,92907	X	X	X	X
4	85,3	80,8	4,3	0,92969	X	X		X X
5	85,7	79,7	6,0	0,95538	X	X	X	X X

We see that the model with Vekt and Temp get a high score in all the tests, and at the same time it contains a small number of variables.The improvement from having only one variable is also big, while the improvement of adding a third is small.

Oppgave f)

We look at the best subsets as in part e):

Best Subsets Regression

```
Response is Y
```

Vars	R-Sq	Adj. R-Sq	C-p	s	Vekt	Alder	Tilføre	Trom	Vmann	VmTid	VmTemp	VmTrom
1	62,0	59,6	48,2	1,3481	X							
1	55,6	52,8	58,7	1,4571								X
2	87,5	85,8	8,5	0,79849			X			X		
2	81,6	79,2	18,0	0,96729	X		X					

3	89,3	87,0	7,6	0,76602		X		X		X		
3	88,2	85,6	9,3	0,80352		X		X			X
4	93,3	91,2	3,0	0,62997	X		X	X		X	
4	92,5	90,1	4,3	0,66591		X		X		X	X
5	94,2	91,8	3,5	0,60832	X		X	X		X	X
5	94,1	91,6	3,7	0,61451	X	X	X	X		X	
6	94,7	91,9	4,6	0,60468	X	X	X	X		X	X
6	94,5	91,6	4,9	0,61579	X		X	X	X	X	X
7	94,9	91,4	6,3	0,62267	X	X	X	X		X	X	X
7	94,8	91,2	6,4	0,62781	X	X	X	X	X	X	X	
8	95,1	90,7	8,1	0,64830	X	X	X	X	X	X	X	X
8	95,0	90,6	8,2	0,65113	X	X	X	X	X	X	X	X
9	95,1	89,6	10,0	0,68362	X	X	X	X	X	X	X	X

We see that multiplying Vekt and Temp, catches more of the variation. Therefore Temp and Vekt*Temp seems to be the two variables that give the best model. Here we also see that little is gained by adding more variables, even though the 4- variable model suggested in the backward regression gives a good result. The model suggested by the forward regression is not quite as good.

Regression with the variables Vekt*Temp and Temp gives:

Regression Analysis

The regression equation is
Y = 2,20 + 2,39 Temp - 0,392 VmTe

Predictor	Coef	StDev	T	P
Constant	2,1996	0,5010	4,39	0,001
Temp	2,3864	0,2782	8,58	0,000
VmTe	-0,39193	0,04153	-9,44	0,000

S = 0,7985 R-Sq = 87,5% R-Sq(adj) = 85,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	66,881	33,440	52,45	0,000
Residual Error	15	9,564	0,638		
Total	17	76,444			

The residual plots show that the assumptions are correct