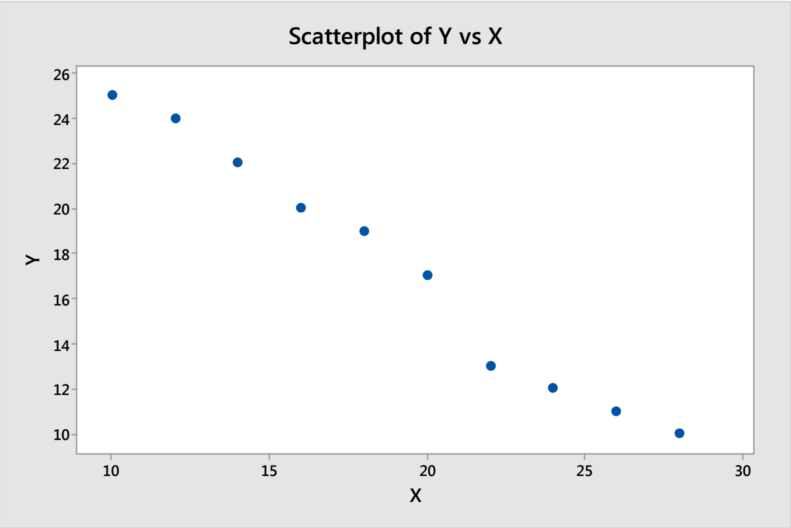
MATU9D2 : Practical Statistics — Spring 2017

Solutions to Practical 8 : Minitab

Question 1

(a) (Plot and Correlation in Practical 7 – extra is to interpret p value)



Subjective Impression

Negative linear relationship

Correlation: X, Y

```
Pearson correlation of X and Y = -0.991
P-Value = 0.000
```

r = -0.991 (as the data is almost all on a straight line in the negative direction close to -1)

Question 1 (continued)

Regression Analysis: Y versus X

```
Analysis of Variance

Source      DF   Adj SS   Adj MS  F-Value  P-Value
Regression   1  270.912  270.912   417.76    0.000
  X          1  270.912  270.912   417.76    0.000
Error        8    5.188    0.648
Total        9  276.100
```

Error /Residual Variance = 0.648

```
Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.805286  98.12%     97.89%      97.15%


Coefficients

Term         Coef  SE Coef  T-Value  P-Value
Constant   34.515    0.880    39.23    0.000
X         -0.9061   0.0443   -20.44    0.000  1
```

$H_0 : \alpha = 0 \quad H_1 : \alpha \neq 0$

Observed Test Statistic = 39.23 p<0.001

So can reject Ho in favour H1 at 1% level : intercept is significantly different to zero

$H_0 : \beta = 0 \quad H_1 : \beta \neq 0$

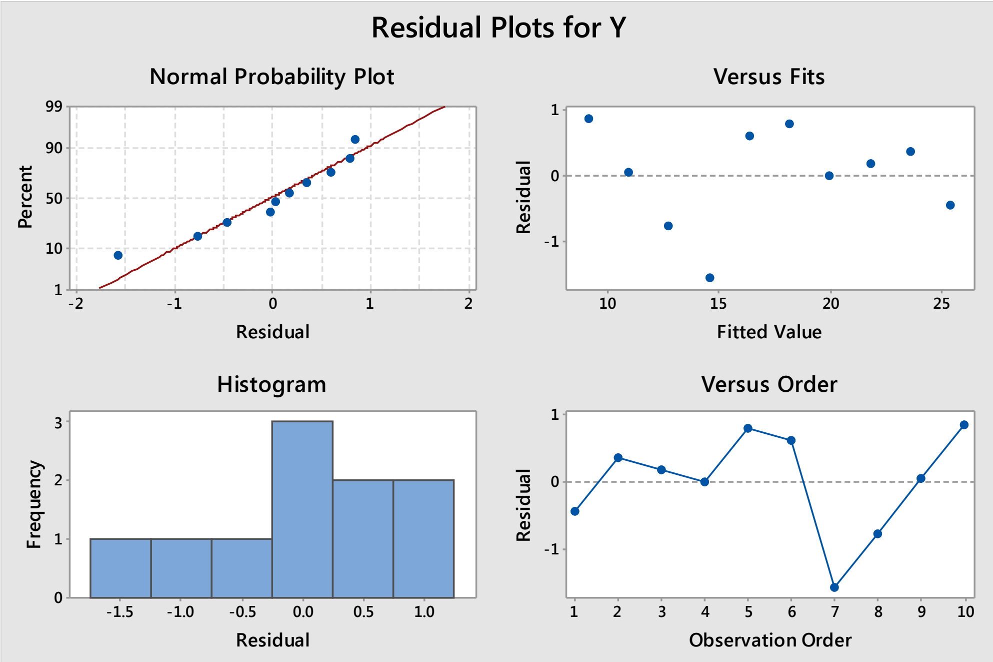
Observed Test Statistic = -20.44 p<0.001

So can reject Ho in favour H1 at 1% level : slope is significantly different to zero i.e. significant relationship

```
Regression Equation

Y = 34.515 - 0.9061 X
```

Fitted Line : y = 34.5 -0.906 x

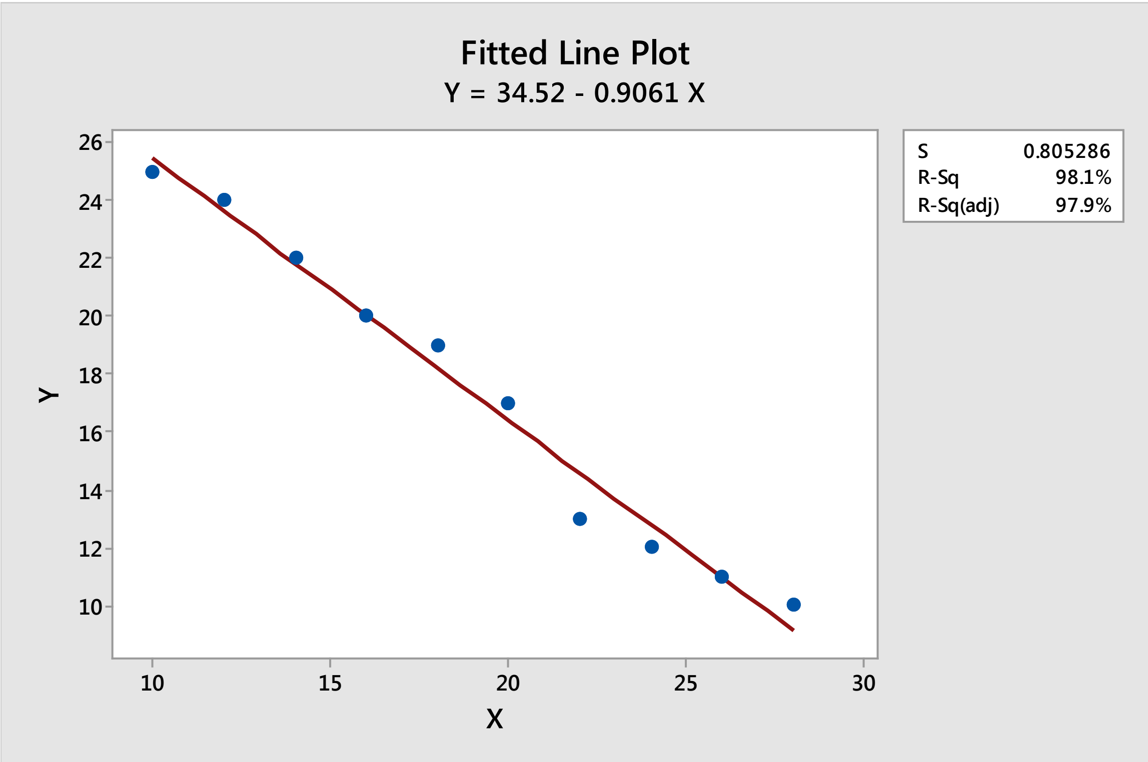
```
Fits and Diagnostics for Unusual Observations

Obs       Y     Fit   Resid  Std Resid
  7  13.000  14.582  -1.582      -2.10  R

R  Large residual
```



Validate Assumptions

Top Left : Can assume normality since graph is approximately linear.

Top Right : Slight problem with assumption of constant variance as points not evenly spread about zero



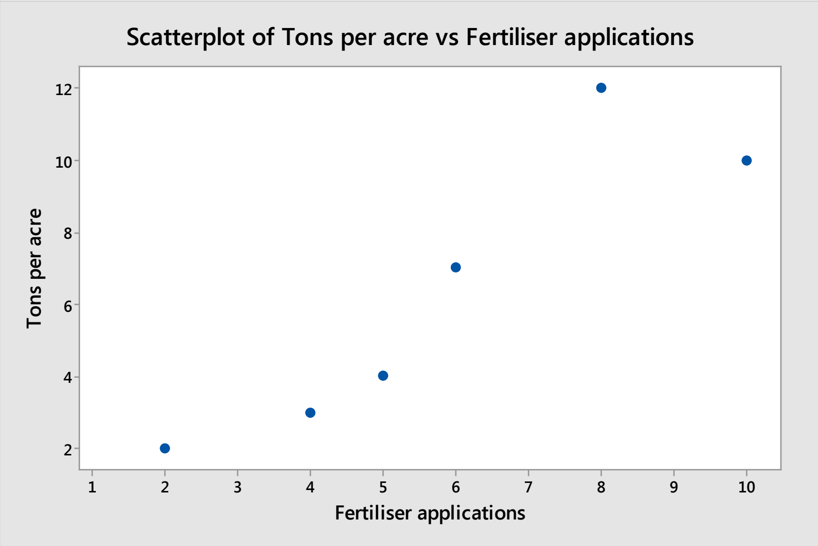
R^2 = 98.1% so 98.1% of the variability in y is explained by the linear relationship with x

Green lines show 95% Prediction Interval for single future observations of x
Red lines show 95% Confidence Intervals for mean y values for given x values

Both are relatively narrow showing a good relationship in practice

Question 2

Note that assumptions invalid so these conclusions based on a linear relationship are not valid!!



Subjective Impression

Positive non-linear relationship

However, not a 'smooth' curve like the plots in the notes – so go ahead with linear relationship but know that is probably not very good

Regression Analysis: Tons per acre versus Fertiliser applications

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	67.07	67.072	18.81	0.012
Fertiliser applications	1	67.07	67.072	18.81	0.012
Error	4	14.26	3.565		
Total	5	81.33			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.88820	82.47%	78.08%	52.17%

R^2 = 49.4% so not a 'good' linear relationship

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.14	1.89	-0.61	0.578	
Fertiliser applications	1.282	0.295	4.34	0.012	1.00

Regression Equation

Tons per acre = -1.14 + 1.282 Fertiliser apps

$H_0 : \beta = 0 \quad H_1 : \beta \neq 0$

Observed Test Statistic = 2.21 p=0.78

So cannot reject Ho in favour H1 at 1% level : slope is not significantly different to zero i.e. not significant relationship

Prediction for Tons per acre

Regression Equation

Tons per acre = -1.14 + 1.282 Fertiliser applications

Variable	Setting
Fertiliser applications	7

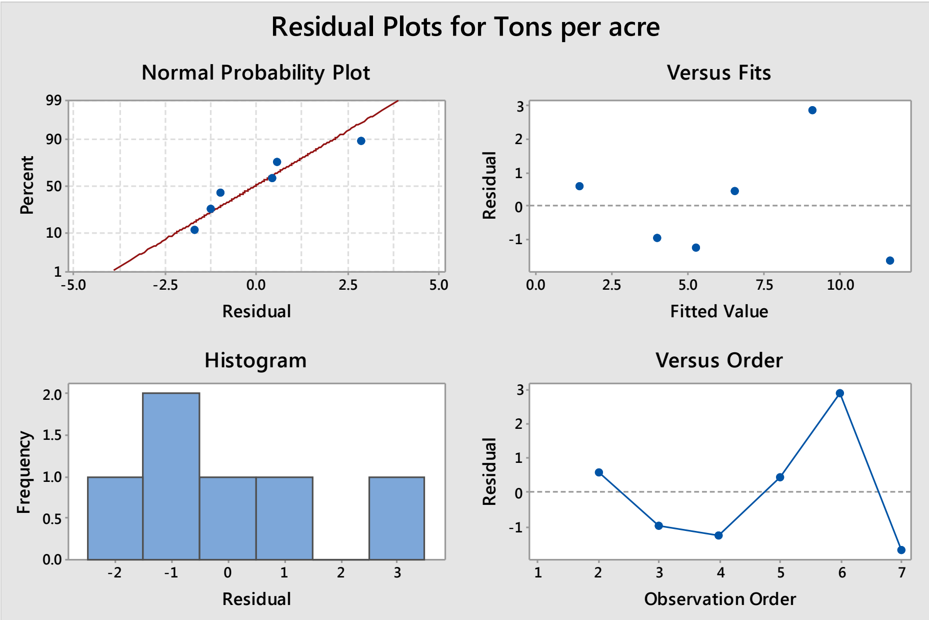
Fitted Line : Yield = 2.24 +0.815 Fertiliser Applications

Fit	SE Fit	95% CI	95% PI
7.82857	0.844430	(5.48406, 10.1731)	(2.08571, 13.5714)

95% Confidence Interval for Mean Yield when 7 applications is 4.64 to 11.24 tons per acre

95% Prediction Interval for Yield for single set of 7 applications is -0.15 to 16.04 tons per acre

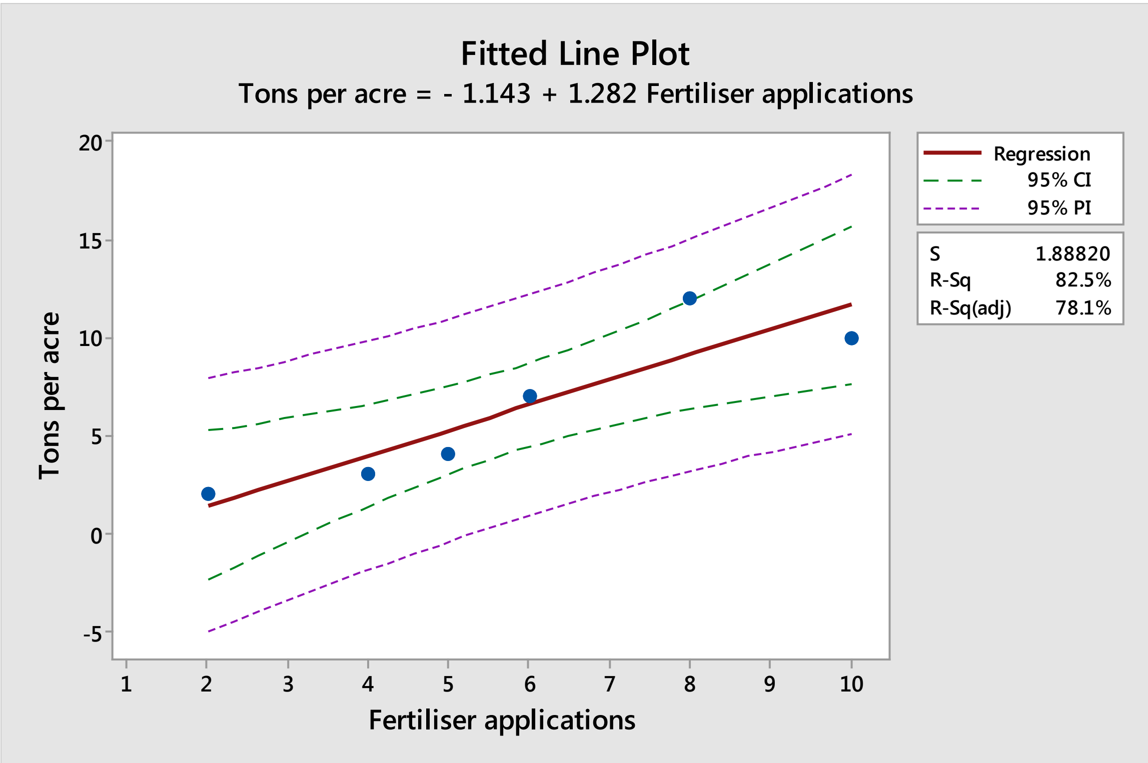
Both very wide : not a good model!!



Validate Assumptions

Top Left : Can assume normality since graph is approximately linear.

Top Right : Problem with assumption of constant variance as points not evenly spread about zero



R^2 = 49.4% so 49.4% of the variability in Yield is explained by the linear relationship with Number of Fertiliser Applications

Purple lines show 95% Prediction Interval for single future observations of x
Green lines show 95% Confidence Intervals for mean y values for given x values

Both are very showing a ‘bad’ and not very useful relationship in practice

Question 3

(a) Regression Analysis: Y versus X2

```
Analysis of Variance

Source          DF   Adj SS   Adj MS  F-Value  P-Value
Regression       1  434.528  434.528   290.37    0.000
  X2             1  434.528  434.528   290.37    0.000
Error            8   11.972    1.496
  Lack-of-Fit    4    3.972    0.993     0.50    0.743
  Pure Error     4    8.000    2.000
Total            9  446.500


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
1.22329  97.32%     96.98%      95.56%


Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  7.523    0.854     8.81    0.000
X2        3.932    0.231    17.04    0.000  1.00


Regression Equation

Y = 7.523 + 3.932 X2
```

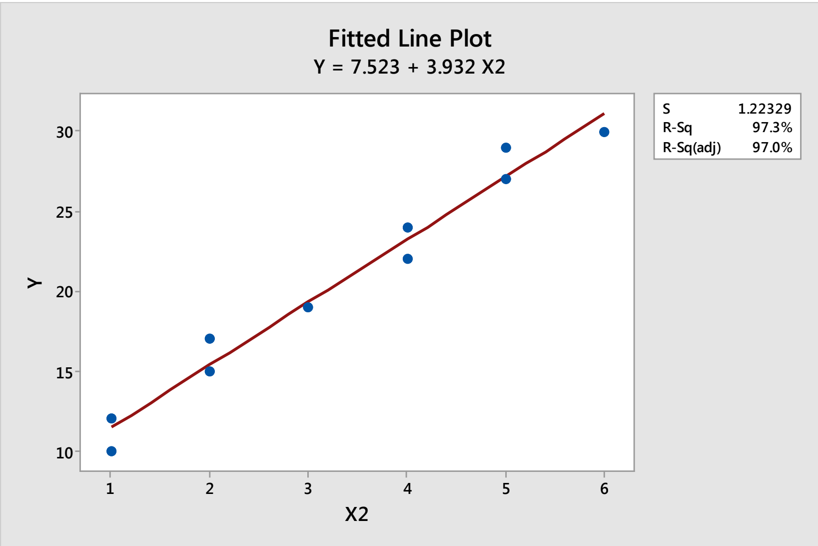
$H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$

Observed Test Statistic = 17.04   p<0.001

So can reject H_0 in favour H_1 at 1% level : slope is significantly different to zero i.e. a significant relationship

R^2 = 97.3% so a 'good' linear relationship

Fitted Line : Y = 7.52 + 3.93 X2



(b) Regression Analysis: Y versus X3

```
Analysis of Variance

Source      DF   Adj SS   Adj MS  F-Value  P-Value
Regression   1  444.512  444.512  1788.89    0.000
  X3         1  444.512  444.512  1788.89    0.000
Error        8    1.988    0.248
Total        9  446.500


Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.498483  99.55%     99.50%      99.16%


Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   33.267    0.341    97.69    0.000
X3        -2.3212   0.0549   -42.30    0.000  1.00


Regression Equation

Y = 33.267 - 2.3212 X3


Fits and Diagnostics for Unusual Observations

Obs       Y     Fit   Resid  Std Resid
 10  30.000  30.945  -0.945      -2.34  R

R  Large residual
```

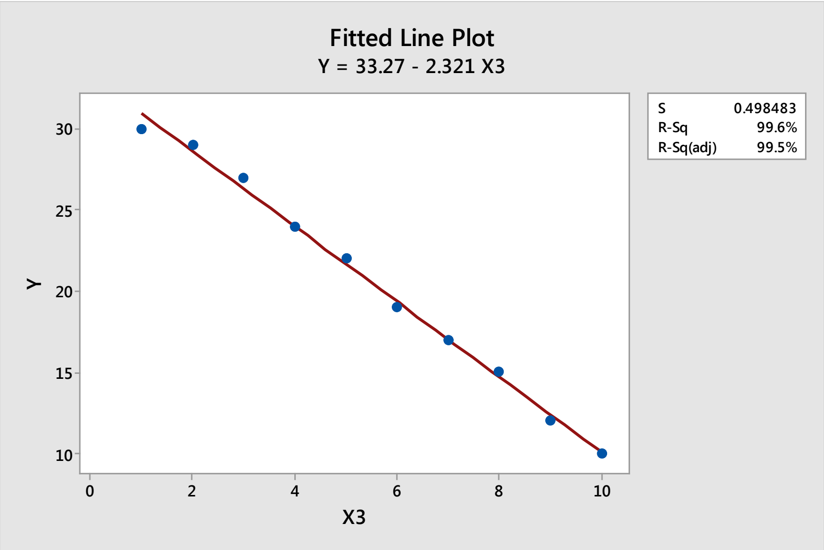
$H_0 : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0$

Observed Test Statistic = -42.30 $p<0.001$

So can reject H_0 in favour H_1 at 1% level : slope is significantly different to zero i.e. a significant relationship

R^2 = 99.6% so a 'good' linear relationship

Fitted Line : Y=33.3 -2.32 X3



Y vs X3 is the 'better' relationship since higher R^2

So the better 1 variable model is Y = 33.3 – 2.32 X3

Note that I should also have checked the assumptions – you must always do this!! Results are only valid if the assumptions are valid.