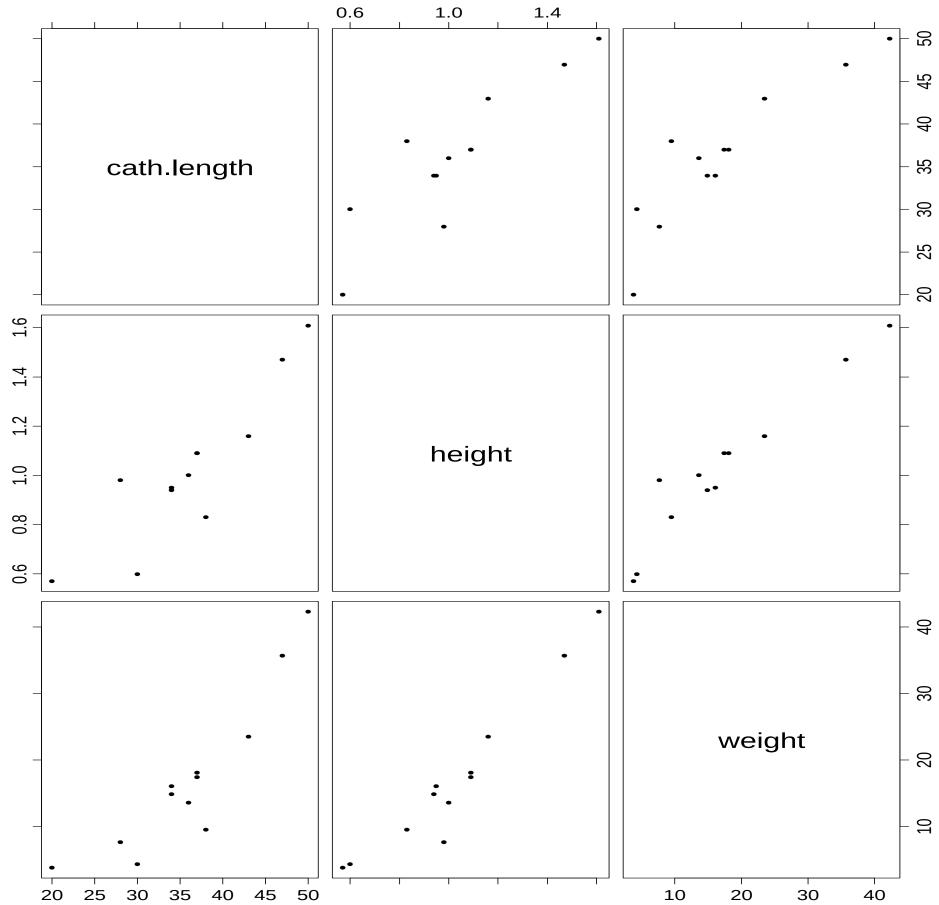
Note: No more than 2 or 3 sentences should be required to answer any part of a question.

1. Heart catheterisation involves inserting a Teflon tube (or catheter) into a major vein at the femoral (thigh) region and moving it into the heart. The catheter is then used to provide information concerning the function of the heart. When this procedure is used on children, the physician must guess the proper length of the catheter.

   In a small study of 12 patients (children), the proper length of the catheter was determined by checking with a fluoroscope (x-ray) that the catheter tip had reached the aortic valve. Each patient's height (in metres) and weight (in kg) were also recorded to see if these could be useful in predicting catheter length (in centimetres).

Obs.	cath. length	height	weight	Obs.	cath. length	height	weight
1	37	1.09	18.1	7	37	1.09	17.4
2	50	1.61	42.3	8	20	0.57	3.8
3	34	0.95	16.1	9	34	0.94	14.9
4	36	1.00	13.6	10	30	0.60	4.3
5	43	1.16	23.5	11	38	0.83	9.5
6	28	0.98	7.7	12	47	1.47	35.7



Linear regression was used to model catheter length as a function of: (i) height and weight, (ii) height only, and (iii) weight only. The S-plus output for each model is given below.

model (i) height and weight

```
              Value Std. Error t value Pr(>|t|)
(Intercept) 20.6652  8.3949     2.4616  0.0361
     height  7.8126 13.6247     0.5734  0.5804
     weight  0.4350  0.3500     1.2429  0.2453

Residual standard error: 3.782 on 9 degrees of freedom
Multiple R-Squared: 0.825
F-statistic: 21.22 on 2 and 9 degrees of freedom, the p-value is 0.0003919
```

model (ii) height only

```
              Value Std. Error t value Pr(>|t|)
(Intercept) 11.4978  4.1167     2.7930  0.0190
     height 24.0868  3.8677     6.2277  0.0001

Residual standard error: 3.883 on 10 degrees of freedom
Multiple R-Squared: 0.795
F-statistic: 38.78 on 1 and 10 degrees of freedom, the p-value is 9.785e-05
```

model (iii) weight only

```
              Value Std. Error t value Pr(>|t|)
(Intercept) 25.3413  1.9255    13.1606  0.0000
     weight  0.6279  0.0934     6.7189  0.0001

Residual standard error: 3.653 on 10 degrees of freedom
Multiple R-Squared: 0.8187
F-statistic: 45.14 on 1 and 10 degrees of freedom, the p-value is 5.24e-05
```

1. (a) Notice that for model (i) neither height nor weight are statistically significant. However, for model (ii) height is significant and for model (iii) weight is significant. Briefly explain how this can occur. [3 marks]

   (b) Explain (1 or 2 sentences each) what the following lines from the output for model (i) indicate: [2 marks each]

      i. `Multiple R-Squared: 0.825`

      ii. `F-statistic: 21.22 on 2 and 9 degrees of freedom,`
      `the p-value is 0.0003919`

   (c) For model (i), the influence measures for the 6th observation are:

```
   .Intercept.  height weight dffits cov.ratio cooks.d  hats inf
 6     1.643   -1.941  2.056 -2.214     1.426   1.334  0.618  *
```

[6 marks in total]

      i. This observation had the largest value in the "hats" column. What does this tell us?

      ii. How is the information provided by "cooks.d" different to that provided by "hats"?

      iii. Explain what the DFBETAS value for weight (2.056) indicates.

      iv. Explain what the DFFITS value (-2.214) indicates.

2. (a) Explain what is meant by multicollinearity and identify **one** problem that can occur when multicollinearity is present in data. [3 marks]

   (b) Identify the diagnostic quantity used to detect multicollinearity and briefly explain how it is used (i.e. what values indicate a multicollinearity problem). [3 marks]

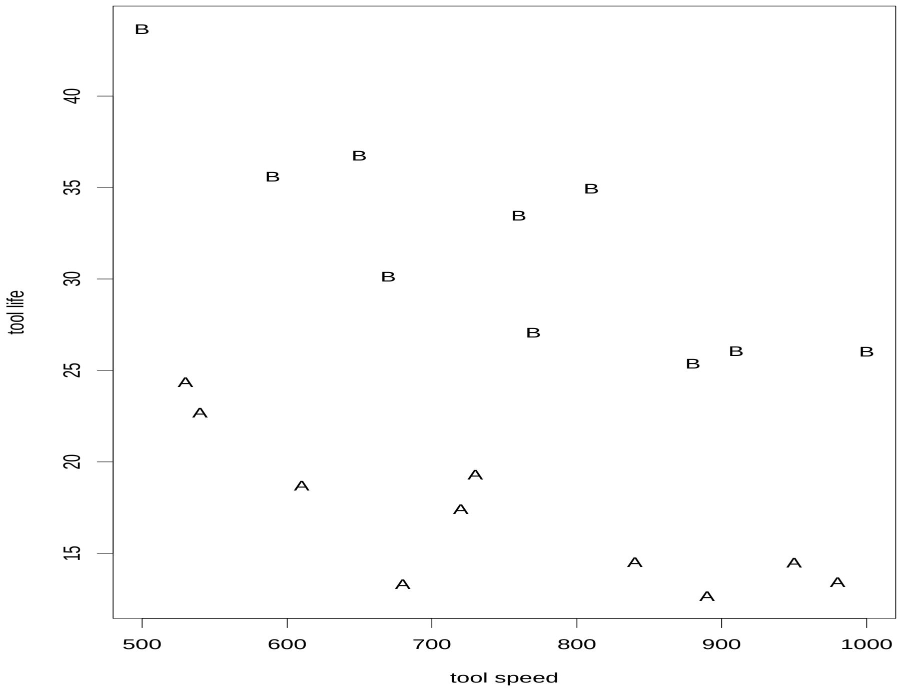
3. (a) Why is Mallow's C_p statistic more useful that R^2 in selecting a model that contains a subset of the possible regressors? [2 marks]

   (b) Briefly explain how you would use Mallow's C_p statistic to identify good subset models. [3 marks]

4. A mechanical engineer wishes to relate the effective life of a cutting tool on a lathe to lathe speed and type of cutting tool used. She collects the following data.

tool life (hours)	lathe speed (RPM)	tool type	tool life (hours)	lathe speed (RPM)	tool type
18.73	610	A	30.16	670	B
14.52	950	A	27.09	770	B
17.43	720	A	25.40	880	B
14.54	840	A	26.05	1000	B
13.44	980	A	33.49	760	B
24.39	530	A	35.62	590	B
13.34	680	A	26.07	910	B
22.71	540	A	36.78	650	B
12.68	890	A	34.95	810	B
19.32	730	A	43.67	500	B

A scatter plot of the data indicates that different regression lines (tool life versus lathe speed) are required for the two tool types.



For each of the following questions **you do not need to give the S-plus commands** – for (a) and (b) just indicate what terms you would include in the model.

4. (a) How would you fit a model that had same the slope but different intercepts (parallel lines) for the two tool types? [2 marks]

   (b) How you would fit a model that had different slopes and different intercepts for the two tool types? [2 marks]

   (c) From the scatter plot, it appears that different intercepts are required but it is not clear whether different slopes are also needed. Briefly explain how you would formally determine whether a model that has different slopes is needed? [2 marks]