Total marks = 25

Notes: Short answers are preferable to long answers (answers in "point form" are acceptable).

1. A plot of residuals versus fitted values (r_i versus $\widehat{y}_i$) can identify several different problems with a fitted regression model. For each of the following situations describe what you would expect the r_i versus $\widehat{y}_i$ plot to look like (include a simple sketch):

   (a) No problems with the fitted model.

   (b) A non-linear relationship between the response and the regressors.

   (c) The variance of the response *decreases* as its mean *increases*.

   (d) An error in entering the data resulted in one of the observations having a much smaller value for the response than it should have.

   [4 marks]

2. The hat matrix diagonals (h_{ii}'s) are used to identify points that have "high leverage" and leave-one-out diagnostics (Cook's Distance, DFBETAS, DFFITS) are used to identify influential points. Explain the difference between an observation that has high leverage and one that is influential.

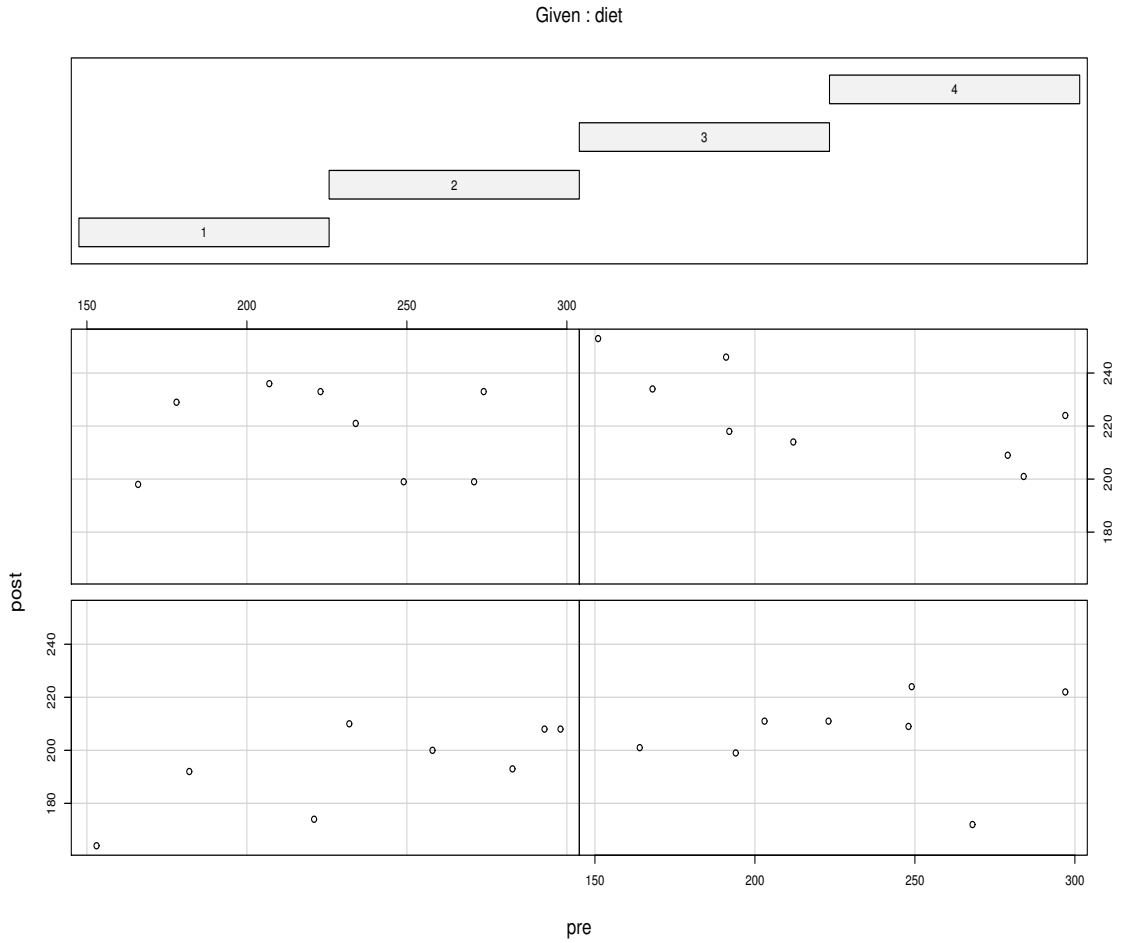
   [3 marks]

3. The covariance ratio is one of the leave-one-out diagnostics covered in STATS 330. What value (approximately) does the covariance ratio take for an observation that is not causing any problems with the fitted model? What does an unusual value of the covariance ratio indicate?

   [3 marks]

4. The data in the table below are cholesterol levels of 32 women subjects who participated in a study of the effect of diet on cholesterol levels. Four diets were investigated and eight women were allocated (randomly) to each diet. The cholesterol level of each participant was measured before the start of the study (`pre`) and was measured again after eight weeks on the allocated diet (`post`).

Diet 1		Diet 2		Diet 3		Diet 4	
post	pre	post	pre	post	pre	post	pre
174	221	211	203	199	249	224	297
208	298	211	223	229	178	209	279
210	232	201	164	198	166	214	212
192	182	199	194	233	223	218	192
200	258	209	248	233	274	253	151
164	153	172	268	221	234	246	191
208	293	224	249	199	271	201	284
193	283	222	297	236	207	234	168



We want to create a regression model that relates the post-diet measurements (`post`) to the pre-diet measurements (`pre`) and to the allocated diet (`diet`).

(a) There are several models that could be used for this data. The coplot should reflect which of these might be suitable. For example, the `lm(post~pre)` command in R produces a model that uses the same fitted line to describe the relationship between `post` and `pre` for each level of `diet`. If this model is suitable, then the coplot should be consistent with having the same fitted line (same intercept and same slope) for each panel.

Describe what each of the following models imply about the fitted lines for each panel of the coplot.

i. `lm(post~diet)`

ii. `lm(post~diet+pre)`

iii. `lm(post~diet+pre+diet:pre)` [4 marks]

(b) The output from `anova` for model iii from (a) is:

```
> chol.fit<-lm( post~pre+diet+pre:diet,data=chol.df)
> anova(chol.fit)
Analysis of Variance Table
Response: post
          Df Sum Sq Mean Sq F value   Pr(>F)
pre        1  130.4   130.4  0.5347 0.471699
diet       3 4465.3  1488.4  6.1034 0.003084 **
pre:diet   3 2334.8   778.3  3.1913 0.041733 *
Residuals 24 5852.9   243.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Explain why this table indicates that `pre`, `diet`, and `pre:diet` should all be kept in the model.

[2 marks]

(c) The output from `dummy.coef` for the model from (b) is:

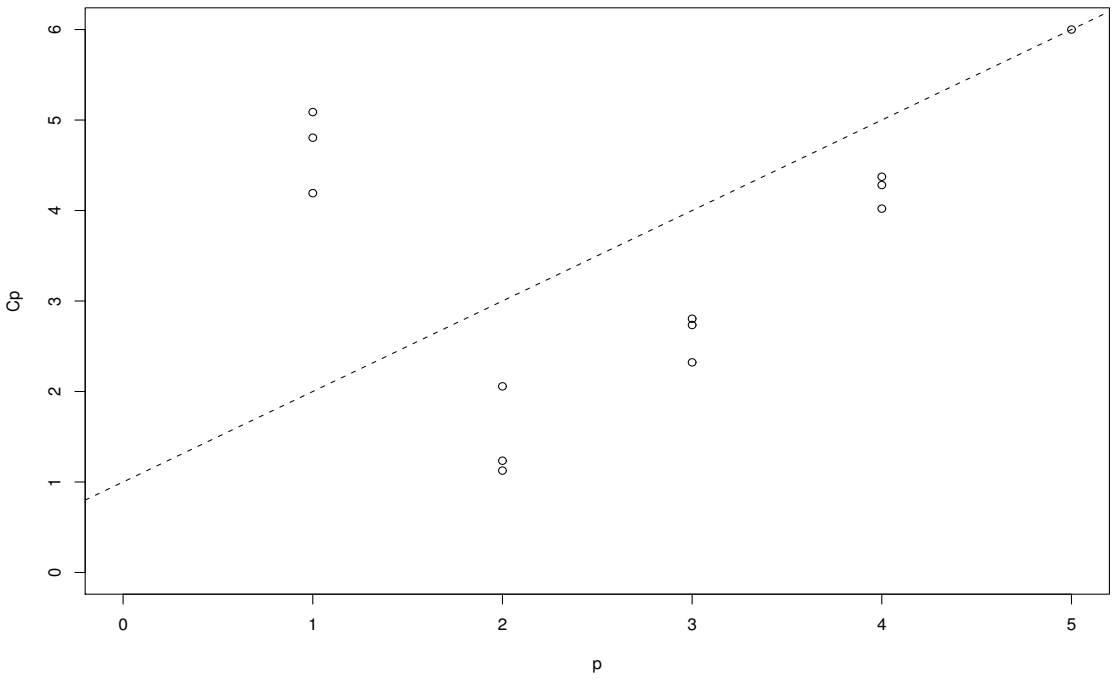
```
> dummy.coef(chol.fit)
Full coefficients are

(Intercept):      137.6323
pre:             0.2333029
diet:                    1          2          3          4
                   0.00000   58.10377   86.10068  138.97097
pre:diet:                1          2          3          4
                 0.0000000 -0.1882805 -0.2565348 -0.4665758
```

Use this output to write down the fitted linear relationship between `post` and `pre` for each of the diets. The goal of this study was to see if any diet results in lower cholesterol levels than the other diets. Explain how you would use the relationships you wrote down to compare diets. Note that you do not need to actually compare the diets – just explain how you would do it. [4 marks]

5. Consider the following output from the `all.poss.regs` and the corresponding plot of Mallow's C_p versus p.

```
      rssp  sigma2 adjRsq    Cp X1 X2 X3 X4 X5
1 2831.244 257.386  0.618 4.192  0  0  1  0  0
1 2962.855 269.350  0.601 4.805  0  0  0  0  1
1 3023.500 274.864  0.592 5.088  1  0  0  0  0
2 1743.943 174.394  0.741 1.126  1  0  0  0  1
2 1767.359 176.736  0.738 1.235  0  1  0  0  1
2 1943.826 194.383  0.712 2.057  0  0  1  0  1
3 1571.452 174.606  0.741 2.322  1  0  1  0  1
3 1659.995 184.444  0.727 2.735  1  1  0  0  1
3 1674.938 186.104  0.724 2.804  0  1  1  0  1
4 1506.775 188.347  0.721 4.021  1  0  1  1  1
4 1562.805 195.351  0.710 4.282  1  1  1  0  1
4 1582.396 197.800  0.707 4.373  0  1  1  1  1
5 1502.321 214.617  0.682 6.000  1  1  1  1  1
```



(a) Briefly explain why it is not a good idea to leave unnecessary regressors in a regression model. [2 marks]

(b) Which model or models would you consider as being reasonable for this situation? Explain how you arrived at your choice(s). [3 marks]