Total marks = 30

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

1. What problem(s) is each of the following diagnostic plots used to detect? In each case, briefly describe what pattern indicates the problem is present.

   (a) A plot of residuals versus fitted values.

   (b) A plot of leverage (h_{ii}'s) versus squared residuals. [5 marks]

2. Vital lung capacity (vlc) is a measure of the volume of air that a person's lungs can hold. As part of an investigation of the effects of smoking on respiration, vital lung capacity (litres) was measured for eight smokers, eight non-smokers, and eight ex-smokers. Age (years) was also recorded as it known to affect vital lung capacity.

SMOKERS		EX-SMOKERS		NON-SMOKERS	
vlc	age	vlc	age	vlc	age
3.8	52	3.9	50	4.3	42
3.2	48	4.6	44	5.3	20
4.6	30	3.3	62	3.7	52
2.9	65	4.8	30	3.3	55
3.5	50	3.7	52	4.8	45
4.8	42	4.2	34	4.4	34
4.0	26	3.6	30	5.2	26
4.9	38	4.3	28	4.0	36

   (a) One model that could be used for this data is

$$\mathrm{E(vlc)} = \beta_0 + \beta_{\mathrm{ex}}\mathrm{I_{ex}} + \beta_{\mathrm{non}}\mathrm{I_{non}} + \beta_{\mathrm{age}}\mathrm{age}$$

   where $\mathrm{I_{ex}}$ and $\mathrm{I_{non}}$ are dummy variables whose values are determined by the smoking status of the subject as follows:

status	$\mathrm{I_{ex}}$	$\mathrm{I_{non}}$
smoker	0	0
ex-smoker	1	0
non-smoker	0	1

   This model assumes that for each of the three categories (smoker, ex-smoker, non-smoker), E(vlc) = intercept + slope × age. For each category, write down the expressions for the intercept and the slope in terms of the β's. [3 marks]

(b) Splus was used to fit the model from (a).

```
               Value Std. Error  t value Pr(>|t|)
(Intercept)   5.6446   0.4016    14.0569   0.0000
        Iex  -0.0131   0.2381    -0.0552   0.9565
       Inon   0.2112   0.2411     0.8762   0.3913
        age  -0.0383   0.0083    -4.6099   0.0002

Residual standard error: 0.4741 on 20 degrees of freedom
```

i. The estimated coefficient for age is $\hat{\beta}_{\text{age}} = -0.0383$. What does this indicate about the relationship between vital lung capacity and age?

ii. Create a 95% confidence interval for β_{age}. The necessary critical value from the Student's t-distribution is $t_{20}(0.975) = 2.086$. What information do we get from this confidence interval, that we don't get from $\hat{\beta}_{\text{age}}$? [3 marks]

(c) The model from (b) was used to make predictions for a 40 year old smoker: a 95% confidence interval for E(vlc) is (4.0, 4.7) and a 95% prediction interval for vlc is (3.3, 5.4). Explain the difference between these two intervals. Also give a simple reason why the prediction interval is wider than the confidence interval. [4 marks]

(d) Consider the following statement:

> The output in (b) indicates that smoking has no effect on vital lung capacity.

Do you agree with this statement? Explain your response. [4 marks]

3. One indicator of the aerobic fitness of a person is their oxygen consumption measured as millilitres of oxygen consumed per kilogram of body weight per minute. Measurements of oxygen consumption and of six possible explanatory variables were taken for 31 individuals. The goal was to determine whether it was possible to predict oxygen consumption using the other variables.

The variables measured were:

oxy: oxygen consumption (ml per kg per minute)

age: age (years)

wt: weight (kilograms)

run: time to run 1.5 miles (minutes)

rest.p: resting pulse rate

run.p: pulse rate at end of run

max.p: maximum pulse rate during run

S-plus was used to fit the full regression model:

```
> summary(oxy.fit,correlation=F)

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  102.9051   12.4118     8.2909    0.0000
        age   -0.2267    0.0999    -2.2691    0.0325
         wt   -0.0740    0.0547    -1.3533    0.1886
        run   -2.6293    0.3848    -6.8327    0.0000
     rest.p   -0.0214    0.0661    -0.3233    0.7492
      run.p   -0.3701    0.1199    -3.0860    0.0051
      max.p    0.3036    0.1366     2.2235    0.0359

Residual standard error: 2.318 on 24 degrees of freedom
Multiple R-Squared: 0.8486
F-statistic: 22.41 on 6 and 24 degrees of freedom, the p-value is 9.794e-09
```

(a) What information do you get from the following parts of the output:

i. The R^2 value.

ii. The P-value for the overall F-test.

iii. The P-values for the individual t-tests.

Would you recommend using this model to predict oxygen consumption? Briefly, explain your response. [4 marks]

(b) The following variance inflation factors were obtained for the model in (b). Note that these are ordered: age, wt, run, rest.p, run.p, and max.p.

```
> VIF
[1] 1.513279 1.155913 1.591232 1.415495 8.437230 8.743509
```

What do these tell us about the regressors? [3 marks]

(c) The S-plus output from `all.poss.regs` and a plot of the Mallow's C_p statistic are given below. Use these to identify a suitable model(s) for this data. Justify your choice(s). [4 marks]

	rssp	sigma2	adjRsq	Cp	age	wt	run	rest.p	run.p	max.p
1(#1)	218.625	7.539	0.734	13.683	0	0	1	0	0	0
1(#2)	715.896	24.686	0.130	106.217	0	0	0	1	0	0
1(#3)	716.721	24.715	0.129	106.370	0	0	0	0	1	0
1(#4)	772.780	26.648	0.061	116.802	1	0	0	0	0	0
2(#1)	200.903	7.175	0.747	12.385	1	0	1	0	0	0
2(#2)	203.227	7.258	0.744	12.817	0	0	1	0	1	0
2(#3)	217.055	7.752	0.727	15.390	0	0	1	0	0	1
2(#4)	217.334	7.762	0.727	15.442	0	1	1	0	0	0
3(#1)	160.985	5.962	0.790	6.957	1	0	1	0	1	0
3(#2)	161.776	5.992	0.789	7.104	0	0	1	0	1	1
3(#3)	186.030	6.890	0.757	11.617	1	0	1	0	0	1
3(#4)	195.355	7.235	0.745	13.352	1	1	1	0	0	0
4(#1)	138.994	5.346	0.812	4.864	1	0	1	0	1	1
4(#2)	156.438	6.017	0.788	8.111	1	1	1	0	1	0
4(#3)	156.847	6.033	0.788	8.187	0	1	1	0	1	1
4(#4)	160.499	6.173	0.783	8.866	1	0	1	1	1	0
5(#1)	129.536	5.181	0.817	5.105	1	1	1	0	1	1
5(#2)	138.815	5.553	0.804	6.831	1	0	1	1	1	1
5(#3)	155.542	6.222	0.781	9.944	1	1	1	1	1	0
5(#4)	156.644	6.266	0.779	10.149	0	1	1	1	1	1
6(#1)	128.974	5.374	0.811	7.000	1	1	1	1	1	1