1. (a) Residuals versus fitted values:
   - non-linearity is indicated by a curved trend in the plot
   - non-constant variance is indicated by the spread of the residuals changing with fitted values (usually a funnel shaped plot)
   - outliers are indicated by points that are isolated (alone) at the top or bottom of the plot

   (b) Leverage versus squared residuals:
   - high leverage points are indicated by points that are isolated at the top of the plot
   - outliers are indicated by points that are isolated at the right edge of the plot
   - influential points occur in the upper right corner of the plot

2. (a)

status	intercept	slope
smoker	β_0	β_{age}
ex-smoker	$\beta_0 + \beta_{\text{ex}}$	β_{age}
non-smoker	$\beta_0 + \beta_{\text{non}}$	β_{age}

   (b) i. According to this model, for each increase of 1 year in age, vlc decreases by 0.0383 litres on average. This is true for all 3 categories (I_{ex} and I_{non} fixed).

   ii. 95% confidence interval

$$-0.0383 \pm 2.086 \times 0.0083 = (-0.056,\ -0.021)$$

   This interval gives us a range of plausible values for β_{age}.

   (c) The confidence interval is the range of plausible values for the mean vlc of all 40 year old smokers. The prediction interval is the range of plausible values for the vlc of a single 40 year old smoker. The PI is wider since it reflects both variability between individuals and estimation error whereas the CI only reflects estimation error.

   (d) This statement is too strong. The analysis only indicates that the data is compatible with smoking having no effect on vlc. Note that the standard errors for $\hat{\beta}_{\text{age}}$ and $\hat{\beta}_{\text{age}}$ are both approximately 0.24 litres. Thus the data is also consistent with a reasonably large smoking effect. Small data sets can often only identify variables that have a large impact on the response and thus no evidence of an effect should not be interpretted as no effect.

   Note: you received 3.5 marks out of 4 if you argued that β_{non} might become significant if I_{ex} is dropped from the model. This is true but not the primary issue.

3. (a) i. The R^2 value indicates that 85% of the variability in oxy, for this dataset, can be explained by the fitted model.

   ii. The low P-value for the overall F-test indicates the model has some predictive power. I.e. there is strong evidence that the coefficients are not all 0.

iii. The P-values for the individual t-tests indicate that rest.p is not needed in the model provided all other variables are in the model. Also wt is not needed if all other variables are in the model.

I would not recommend using this model to predict oxygen consumption as it contains unnecessary predictors. We should consider dropping run.p (and then possibly wt) from the model.

(b) The VIF's are used to detect the presence of multicollinearity. Since the values for age, wt, run, and rest are all low (close to 1) these are not involved in near linear relationships with other variables. The VIF's for run.p and max.p are moderately high (between 5 and 10) indicating these 2 variables are linearly related.

(c) The plot of C_p indicates at least 4 regressors are needed in the model as this is smallest p where C_p is close to $p+1$. There is one 4-regressor model and two 5-regressor models that are close to the $C_p = p+1$. These models all contain age, run, run.p, and max.p and so we can be confident that these regressors are needed in the model. Note that the plot is consistent with just these 4 variables being active, since all the points that are close to the $C_p = p+1$ line correspond to models that contain all these variables.

However, we may want to entertain the the best 5-regressor model which contains wt as the additional regressor. This model has the highest adjusted R^2 and lowest $\hat{\sigma}^2$ (just slightly better than the 4-regressor model). If expert knowledge suggests wt should be an important regressor we should use this model, otherwise I would choose the 4-regressor model.