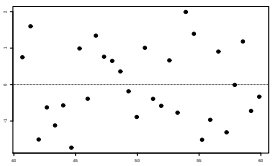
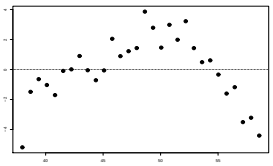
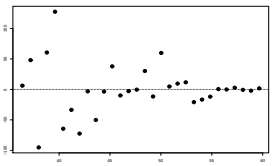
1. (a) A horizontal band of points with no obvious pattern (points just reflect random scatter).



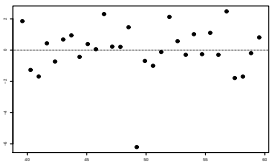
(b) There should be a clear trend (usually curved) present.



(c) There should be a clear funnel pattern with the spread of the points decreasing as $\hat{y}$ increases (spread decreases from left to right).



(d) There should be an isolated point at the bottom of the graph.



2. A high leverage point has an unusual set of values for the explanatory variables. High leverage points have the *potential* to have a big impact on the fitted model but may not. Influential points are points that actually *do* have a big impact on the fitted model.

3. An observation that is not causing any problems should have a covariance ratio of ≈ 1. Deleting a point that has an unusual covariance ratio will affect the standard errors and covariances of the estimated coefficients. As a result the standard errors for predictions will also be affected.

4. (a) i. The fitted line in each panel would have slope = 0 (a horizontal line) but the intercept could be different for different panels.
   ii. The fitted lines would all have the same slopes but could have different intercepts (parallel lines).
   iii. The fitted lines could have different slopes and different intercepts.

(b) We should consider the `pre:diet` interaction first. The last line in the table gives moderate evidence that this term should be kept in the model. If we keep the interaction in the model, we should keep the main effects of `pre` and `diet` as well.

(c) The fitted lines for different diets are:

diet	fitted linear relationship
1	$\mathrm{E}(\texttt{post}) = 137.63 + 0.233\,\texttt{pre}$
2	$\mathrm{E}(\texttt{post}) = 195.74 + 0.045\,\texttt{pre}$
3	$\mathrm{E}(\texttt{post}) = 223.73 - 0.023\,\texttt{pre}$
4	$\mathrm{E}(\texttt{post}) = 276.60 - 0.233\,\texttt{pre}$

To compare diets we need to compare the values of $\mathrm{E}(\texttt{post})$ produced by these fitted lines for values of `pre` that cover the range of values that were observed in the data (approximately 150 to 300). A good way of doing this would be to plot the 4 fitted lines on a common graph (for $\texttt{pre} = 150$ to $\texttt{pre} = 300$).

5. (a) Leaving unnecessary regressors in the model will inflate the size of prediction errors (on average) and also increases the hazards associated with extrapolation. In addition, they can affect the estimated coefficients for the important regressors.

(b) The Cp plot suggests that we should be looking at a model with 2 regressors.

- $p = 2$ is the first time that Cp is close to or below the $p + 1$.
- For $p > 2$, Cp increases by about 1 for each additional variable which is what we would expect if all the important regressors are already in the model.

There are two $p = 2$ models that perform about as well as each other: (i) X_1 and X_5, and (ii) X_2 and X_5. Either of these models would be reasonable. If we had to choose, in the absence of additional information we would select (i) as it fits slightly better. Model (i) also has the smallest value of `sigma2` and the largest value of `adjRsq` among the candidate models. However, the best 3 variable model (X_1, X_3, and X_5) performs almost as well in terms of these criteria and so it could also be considered as a reasonable choice.