1. You should sketch rough graphs for each of the following.

   (a) **1:** Hwt~1 identical lines, 0 slope.

   2: Hwt~Sex different intercepts, 0 slope.

   3: Hwt~Bwt identical lines (same slopes, same intercepts)

   4: Hwt~Bwt+Sex parallel lines (same slopes, different intercepts)

   5: Hwt~Sex*Bwt different lines (different slopes, different intercepts)

   (b) The trellis plot looks as if there is different intercepts and slopes for each sex. This indicates that Model 5 describes the data adequately.
   Model 5 is Hwt$\sim$ Sex + Bwt + Sex:Bwt
   An F-test is done to see if the interaction term can be dropped
   H_0:Bwt:Sex is not needed

$$\begin{aligned} F_0 &= \frac{(ReSS_{SUB} - ReSS_{FULL})/d}{ResMS_{FULL}} \\ &= \frac{(299.37 - 291.05)/1}{291.05/140} \\ &= 4.0 \end{aligned}$$

   P-value < 0.05

   This means there is evidence against the null hypothesis that the coefficient for the interaction is 0. Hence the interaction term is needed in the model, therefore both main effects are also kept in the model. This provides evidence that Model 5 adequately describes the data.

   (c) In the tables the variables are added sequentially. Hence in the first table Sum of Squares (SS) for Sex represents SS explained by Sex if Bwt is ignored (not in the model). In the second table SS for Sex represents SS explained by Sex if Bwt is taken into account (included in the model).

   (d) The fitted model for male cats is

$$\begin{aligned} \text{Hwt} &= 2.9813 - 4.1654 + (2.6364 + 1.6763) \times \text{Bwt} \\ &= -1.1841 + 4.3127 \times \text{Bwt} \end{aligned}$$

   MBwt is estimating the difference between the coefficient of Bwt for male cats and the coefficient of Bwt for female cats.

   The heart weight of a male cat weighing 2.75kg is predicted as follows:

$$\begin{aligned} \text{Hwt} &= -1.1841 + 4.3127 \times 2.75 \\ &= 10.68 \end{aligned}$$

2. (a) A generalised linear model should be used as the response is binary (an individual either has CHD or doesn't have CHD).

The differences between linear models and generalised linear models are:
- with GLM's a wide range of different distributions for the response can be assumed (linear models always assume a Normal response).
- GLM's estimate coefficients by maximising the likelihood function (linear model use least squares).
- GLM's use residual deviance (and Chi-square tests based on the residual deviance) to compare possible models (linear models use the residual SS and F-tests based on the residual SS).
- Diagnostics for GLM's are based on Pearson and Deviance residuals (linear models only use 1 type of residual $r_i = y_i - \hat{y}_i$.)

(b) As BP increases, R/N also increases. As Chol increases, R/N increases. It appears that the 2 variables may interact so a model with BP, Chol and BP:Chol would be an appropriate initial model. Since the response is binary we would fit a logistic regression model.

(c) Since the p-value for the interaction term is large, it doesn't need to be retained in the model. The H_0 being tested is that the interaction term is not needed in the model. There is no evidence against this.

(d) The outlier is most likely to be the point that has Chol 200-219 and BP 147-166 since this point does not follow the general pattern seen in the plots.

(e)

$$\begin{aligned} \text{logit}\hat{\pi} &= -3.482 + 0.532 + 1.344 \\ &= -1.605 \end{aligned}$$

$$\begin{aligned} \hat{\pi} &= \frac{\exp(\text{-}1.605)}{1+\exp(\text{-}1.605)} \\ &= 0.167 \end{aligned}$$

BP> 166 and Chol> 260 gives logit $\hat{\pi}$ = -3.4819 + 1.2004 + 1.3441 = -0.9374

$$\begin{aligned} \hat{\pi} &= \frac{\exp(\text{-}0.9374)}{1+\exp(\text{-}0.9374)} \\ &= 0.281 \end{aligned}$$

This is greater than 25 %.

3. (a) A logistic regression model is being fitted. The theoretical form for the model is:

$$\text{logit}(\pi) = \beta_0 + \beta_1 \times \text{Fib} + \beta_2 \times \text{Glob}$$

where π = probability an individual has ESR$\geq$ 20 mm/h.

(b) The terms are added sequentially in the tables. In the first model the deviance for Globulin is determined taking into account that Fibrinogen is already in the model. In the second model the deviance for Globulin is determined ignoring Fibrinogen.

(c) It appears as if the protein Fibrinogen affects the response. In both models the p-values for Fibrinogen are very small, and so there is strong evidence to suggest that it is needed in the model whether Globulin is included in the model or not. The p-values for Globulin are both large and hence indicate that Globulin is not needed in the model.

(d) Points 23 and possibly 15 appear to be influential. We should first check to make sure that an error was not made in entering these points. If not, then assess the effect that deleting these points will have on the fitted model. If the impact is substantial the results with these points in the data and with these points removed should both be reported. Note: we should not just automatically delete these points.

(e) Use model esr.m3
If $\pi > 0.80$ then $\text{logit}(\pi) > \log(0.80/(0.20)) = 1.3862$. We just need to plug this into model esr.m3 and solve for Fib.

$$\begin{aligned} 1.3862 &= -7.4657 + 2.1179 \times \text{Fib} \\ \text{Fib} &= \frac{1.3862 + 7.4657}{2.1179} \\ &= 4.18 \end{aligned}$$

Note that esr.m3 indicates that $\text{logit}(\hat{\pi})$ increases as Fib increases. Hence if Fib> 4.18 then $\hat{\pi} > 0.80$.

4. (a) i. This plot provides a check for linearity. It also allows constant variance to be checked.
ii. These plots allow a check of whether the explanatory variables need transforming. If the plot shows a non-linear trend then a transformation is required.
iii. A plot of residuals vs indices provides a check for outliers. If the points were collected over time it can also be sued to check for serial correlation.
iv. A plot of residuals vs lagged residuals provides a check for a serial correlation. If a linear trend is present this indicates a problem.

(b) A high leverage point is a point which is extreme in terms of its values for the regressors. It attracts the fitted line or plane. This can distort the fitted coefficients and the standard errors for the fitted coefficients. If a high leverage point also has a large error it will cause a substantial change in the least squares line.

5. (a) A case could be made for using either model. Both models appear reasonably linear (the smoothed curves are being affected by 1 or 2 points at the extremes of the plot in each case) although model 1 may be slightly better in this regard. Model 1 suffers from an obvious funnel effect but is better in terms of Normality and would be easier to interpret. Neither model has glaring problems in terms of influential points or outliers. I would probably use model 2 (constant variance is more important than Normality) but you could certainly make a good case for model 1. A better option than either of these may be to use weighted least squares and the model formula from model 1.

(b) If Model 1 is chosen in (a):
Normality - This is justified as the points lie fairly much on the straight line.
Linearity - Since the plot of the residuals vs fitted values shows a fairly linear relationship, the assumption is valid.
Constant variance - This assumption is not justified as there is a funnel effect apparent in the plot of residuals vs predicted values.

If model 2 is chosen in (a):
Normality - This assumption does not appear to be valid as the points do not fall on the straight line. The data appears to be skewed.

Linearity - The plot of residuals vs fitted values shows some indication of a non-linear relationship but this is mainly due to points at the extremes. This assumption is questionable.

Constant variance - The plot of residuals vs predicted values may indicate a small funnel effect but not enough to be of concern.

(c) For either model chosen the analysis indicates that using a cooling tower increases costs. For model 1 using a cooling tower increases the cost by \$72.6537 million. In model 2 the coefficient for CT is 0.1739, so this figure is the increase in log(cost). As result cost would increase by a factor of $\exp(0.1739) = 1.18$.

Both models also indicate that building a plant in the North East increases costs. For model 1 the increase in cost is \$135.6183 million. For model 2 the increase in log(cost) is 0.3114 which means that cost increases by a factor of $\exp(0.3114) = 1.37$.

(d) For model 1:

$$\begin{aligned} C &= -8533.9104 + 125.7379D + 72.6537CT - 8.5374N + .4624S + 135.6183NE \\ &= -8533.9104 + 125.7379(71) + 72.6537(1) - 8.5374(5) + 0.4624(800) + 135.6183(1) \\ &= 929 \end{aligned}$$

For Model 2:

$$\begin{aligned} \log(\mathrm{C}) &= -17.5713 + 0.2734D + 0.1739CT - 0.1373\log(\mathrm{N}) + 0.7446\log(\mathrm{S}) + 0.3114 \\ &= -17.5713 + 0.2734(71) + 0.1739(1) - 0.1373\log(5) + 0.7446\log(800) + 0.3114 \\ &= 7.081786 \\ C &= \exp(7.081786) \\ &= 1190 \end{aligned}$$