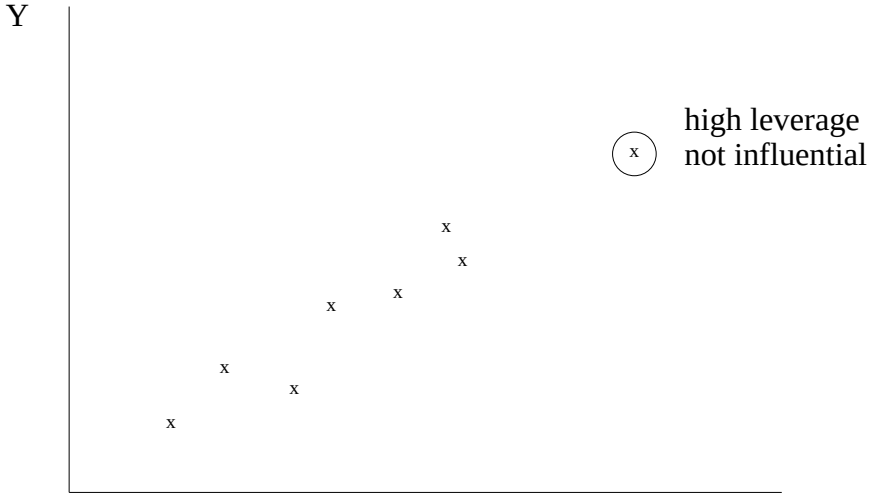
1. (a) A high leverage point is unusual in terms of its X-values. It has the potential to have a big impact on the fitted model. An influential point is an observation that does have a large impact on the fitted model.

i.



ii.

(b) R^2 always increases when an additional regressor is added to any model. As a result, maximising R^2 would always select the model that contains all the possible regressors. Mallow's C_p will penalise a model that has unnecessary regressors. Thus it is more useful for subset selection.

(c) i. Cook's distance is an overall measure of how much the fitted coefficients change if an observation is delete. This observation is having a larger overall effect on the regression surface than any other observation.

ii. If this observation were deleted then the fitted regression surface would shift by 1.32 standard errors at this point.

(d) When fitting a glm the constructed variable plot is used to investigate whether a power transformation of a numeric regressor will improve the model.

The slope of a line drawn through the points will indicate a suitable power transformation for that explanatory variable.

$$\lambda = slope + 1$$

(e) For contingency table data, the dissimilarity index indicates how close a fitted model is to the sample data. It represents the proportion of observations that would need to be moved to different cells for the model to fit perfectly. It is useful if there are a large number of observations and as a result interactions that are of little practical importance are showing up as being statistically significant. If the dissimilarity index for the model that does not include these interactions is small (say < 0.03) it indicates this simpler model does a good job of describing the data.

2. (a)
- mean %BF is higher (on average) for females than males
- mean %BF seems to increase slightly with increasing weight and height
- male athletes tended to be taller and heavier than female athletes
- strong relationship between `Ht` and `Wt`

(b) i. 76% of variability in log(%BF) can be explained by the fitted model

ii. strong evidence that not all coefficients are 0, i.e. the model has predictive value.

(c) The p-value corresponding to `Ht` is quite large. However, the `Sex:Ht` interaction is significant. As a result, we prefer to retain both `Sex` and `Ht` in the model.

(d) We can write out separate models for male and female athletes

$$\text{male: } log(BF) = 1.98 - 0.0139Ht + 0.0215Wt$$

$$\text{female: } log(BF) = 1.67 - 0.0017Ht + 0.0215Wt$$

For fixed levels of `Sex` and `Ht`, log(`BF`) increases by 0.0215 for each increase of 1kg in weight.

For male athletes, log(`BF`) decreases more rapidly as `Ht` increases than for female athletes.

Note that intercept values are misleading since for any reasonable values of `Wt` and `Ht` the predicted log (`BF`) for males will be less than that for females.

(e) The key feature I would like to communicate is the interaction between `Sex` and `Ht`. Either of the fo llowing plots is useful in revealing the nature of this interaction:

i. Plot `BF` or log(`BF`) vs `Ht` using `Sex` as a plotting symbol

ii. A trellis plot of `BF` or log(`BF`) vs `Ht` conditioned on `Sex` (or conditioned on both `Sex` and `Wt`)

3. (a) In general, as `time` increases breakdown strength decreases. This trend is quite small for `temp`=180 but becomes more pronounced as `temp` increases and is very strong for `temp`=275. The trend looks like an ex ponential decay - this is most evident for the plot for `temp`=275.

   The plot also indicates less variability between observations at combinations of `time` and `temp` where the m ean breakdown strength is low (breakdown strength $\downarrow$ as `temp` $\uparrow$).

   (b) I do anticipate problems with the suggested model.

   i. the trellis plot indicates that variability is related to the mean value of the response and that the relationsh ip is non-linear. This suggests it may be necessary to transform the response (log?).

   ii. the relationship between `time` and breakdown strength seems much stronger at high temperatures. This sugges ts that the `time:temp` interaction may be needed in the model.

   (c) The p-value for the `period:activity` interaction is very small giving strong evidence that those 2 factors are not independent.

   Fit a model that does not contain the interaction term. Look at the deviance residuals to determine which counts are not compatible with the assumption of independence.

   (d)
$$\widehat{odds} = \frac{\hat{\pi}}{1-\hat{\pi}} = \frac{38}{34} = 1.12$$

   (e)
$$\widehat{odds\ ratio} = \frac{\widehat{odds}\ morning}{\widehat{odds}\ evening} = \frac{1.12}{10/69} = 7.71$$

   (f)
$$log(7.71) \pm 1.96\sqrt{\frac{1}{10}+\frac{1}{69}+\frac{1}{38}+\frac{1}{34}}$$

$$(1.234, 2.85)$$

   Take exponentials to get interval for odds ratio

$$(3.43, 17.31)$$

4. (a) Test statistic is

$$\begin{aligned} \chi^2_0 &= \text{Null dev - Res dev} \\ &= 28.28 - 20.32 = 7.96 \end{aligned}$$

$$\begin{aligned} \text{p-value} &= Pr(\chi^2 \geq 7.96) \\ &\quad \text{where } \chi^2 \sim \text{Chi-square}(df = 1) \end{aligned}$$

(b) i. $logit(\hat{\pi}) = 15.04 - 0.232 \times Temp$

ii. $\hat{\pi} = \frac{exp(15.04 - 0.232 \times Temp)}{1 + exp(15.04 - 0.232 \times Temp)}$

(c) i. For each increase of 1°F in `Temp`, $logit(\hat{\pi})$ decreases by 0.232

ii. For each increase of 1°F in `Temp`, odds of thermal distress decrease by a factor of $exp(0.232) = 0.793$

(d) At 31°F the fitted model gives

$$\begin{aligned} logit(\hat{\pi}) &= 15.04 - 0.232 \times 31 = 7.848 \\ \hat{\pi} &= \frac{exp(7.848)}{1 + exp(7.848)} = 0.9996 \end{aligned}$$

We are predicting at a value of `Temp` well outside the range of the data.

(e) For $\pi = 0.9$, $logit(\pi) = log\frac{\pi}{1-\pi} = log(9) = 2.197$ Need to find `Temp` so that

$$15.04 - 0.232 \times \texttt{Temp} = 2.20$$

$$\texttt{Temp} = \frac{2.20 - 15.04}{-0.232} = 55.3°F$$

(f) The only thing that really stands out is that point 21 has an unusually large Cook's Distance. We should first check to make sure there has not been an error in recording this observation. Then we should consider what effect del eting this point would have on the fitted point. Obs 21 corresponds to the highest temperature for which at least one primary o-ring suffered thermal distress. Thus if it were deleted the relationship between `Temp` and thermal di stress would become more pronounced.

5. (a) Fit a log-linear model using counts as the response. The model must contain all main effects and use a stepwise procedure to determine which interactions are significant. The association graph is constructed by joining all pairs of factors that occur in any significant interaction.

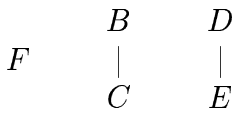
   (b) i. F and B (or any other factor)
   ii. B and A given C
   iii. B and C (A,C), (A,D), (A,E), (D,E)

   (c) Two factors are conditionally independent if they are only related through another factor(s). They only become independent if we fix the level of the third factor(s).

   In the above graph B and A are only related through C. So if we fix the level of C, B and A become independent.

   (d) Yes, this is sensible. B is not directly related to any factor in the group A,D,E,F and so it is permissable to collapse on these.

   (e) If we restrict our attention to smokers we are fixing the level of factor A. As a result the graph becomes



   Thus we have 3 independent sets of factors: F, (B,C) and (D,E).