1. (a) i. If this observation is deleted, it would have a substantial impact on the fitted coefficients overall.

   ii. If this observation is deleted the residual deviance would decrease by 4.6

   iii. If this observation is deleted the fitted coefficient for that explanatory variable would change by 1.08 standard errors.

   (b) i. The dissimilarity index is useful when it is thought that some of the terms in a model may not be of practical significance even though they are statistically significant. This usually occurs when the number of observations is very large.

   ii. The dissimilarity index measures what percentage of data would need to be redistributed in a contingency table in order for a model to fit exactly. We fit the model without the terms we think are not important and calculate the dissimilarity index. If this number is small (say $< 3\%$) then we use the simplified model.

   (c) Estimated odds ratio is:

$$\text{odds ratio} = \frac{\text{odds(case}| \geq 80\text{g)}}{\text{odds(case}| < 80\text{g)}}$$

$$\widehat{\text{odds ratio}} = \frac{96 \times 666}{104 \times 109} = 5.64$$

95% Confidence Interval for log(odds ratio) is:

$$\begin{aligned} \log(5.64) &\pm 1.96\sqrt{\frac{1}{96} + \frac{1}{666} + \frac{1}{104} + \frac{1}{109}} \\ = 1.73 &\pm 0.34 = (1.39,\ 2.07) \end{aligned}$$

95% Confidence Interval for odds ratio is:

$$(\exp(1.39), \exp(2.07)) = (4.01, 7.92)$$

We can say with 95% confidence that the odds of developing esophageal cancer for the ≥ 80 g alcohol consumption group are between 4.01 and 7.92 times the odds for the < 80 g alcohol consumption group.

   (d) i. Logistic form of model:

$$\hat{\pi} = \frac{\exp(-5.34 + 0.00155 \times \text{load})}{1 + \exp(-5.34 + 0.00155 \times \text{load})}.$$

Logit form of model:

$$\log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = -5.34 + 0.00155 \times \text{load}.$$

   ii. If $\hat{\pi} = 0.50$ then logit $\hat{\pi} = \log 1 = 0$ Thus we need to find the value of load that gives $0 = -5.34 + 0.00155 \times \text{load}$. Therefore load $= 5.34/0.00155 = 3445$ psi.

2. (a) The overall F-test indicates strong evidence against all the coefficients being 0. This indicates that at least some of the variables will be useful for predicting the response. Many of the t-tests are not significant indicating that not all the variables are required in the regression model. The R^2 statistic indicates that 78% of the variability in the response is explained by this model.

   (b) i. Transforming the response affects the:
      A. linearity of the regression surface.
      B. the constant variance assumption.
      C. the normality assumption.

   ii. A. Linearity is checked by a plot of residuals vs fitted values. A non-linear trend indicates non-linearity.
      B. Constant variance is also checked by a plot of residuals vs fitted values. A funnel shaped plot indicates non-constant scatter.
      C. Normality is checked by a Normal Probability plot of residuals. A non-linear trend indicates non-normality.

   (c) Multicollinearity means there is one or more near linear relationships between the explanatory variables. If multicollinearity is present it makes model selection more difficult since there may be a number of subset models that all work about as well as each other. These VIF's indicate that Po1 and Po2 are strongly related to other variables (probably each other) and that GDP ad Ineq are also related to other variables.

   (d) We want to choose models that are close to the minimum value of Cp and/or where the points first approach the p + 1 vs p line. For this dataset it is clear that we need at least 6 variables. We might use 7 variables but there does not seem to be any reason to use > 7. The best 6 variable model includes M, Ed, Po1, V2, Ineq and Prob as regressors. The 3 best 7 variable models all contain these variables so we can be confident that these should be in the regression model. In addition we might add 1 of GDP, V1 and pop.

3. (a) i. For each hospital wt is a constant (doesn't depend on gest) but the constants can be different.

   ii. There is a linear relationship between wt and gest and this relationship is the same for all 3 hospitals.

   iii. There is a linear relationship between wt and gest for each hospital. The lines all have the same slope but can have different intercepts.

   iv. There is a linear relationship between wt and gest for each hospital. The lines can have different slopes as well as different intercepts.

   (b) $\mathrm{E(wt)} = \beta_0 + \beta_B I_B + \beta_C I_C + \beta_g \mathrm{gest} + \beta_{B:g} \mathrm{gest} I_B + \beta_{C:g} \mathrm{gest} I_C$
   β_0 = intercept for hospital A
   β_g = slope for hospital A
   β_B = intercept for hospital B - intercept for hospital A
   β_C= intercept for hospital C - intercept for hospital A
   $\beta_{B:g}$ = slope for hospital B - slope for hospital A
   $\beta_{C:g}$ = slope for hospital C - slope for hospital A

   (c) • line for hosp tests$H_o : \beta_B = \beta_C = 0$ for the model that just contains I_B and I_C
   • line for gest tests$H_o : \beta_g = 0$ given that I_B and I_C are in the model

• line for hosp:gest tests$H_o : \beta_{B:g} = \beta_{C:g} = 0$ provided that I_B , I_C , and gest are in the model.

There is strong evidence that the interaction terms are needed in addition to the main effects. If the interaction is in the model we should have the main effects in as well. Thus model (iv) should be used.

(d) Hospital A : E(wt) = -0.39 + 0.048 × gest
Hospital B : E(wt) = -1.96 + 0.109 × gest
Hospital C : E(wt) = -1.14 + 0.076 × gest

The values of gest for this data set range from 27 weeks to 36 weeks. The fitted model produces these values:

hosp	27 weeks	36 weeks
A	E(wt) = 0.91	E(wt) = 1.34
B	E(wt) = 0.98	E(wt) = 1.96
C	E(wt) = 0.91	E(wt) = 1.60

We can see that for a gestation age of 27 the birth weights are all very close but for a gestation age of 36 weeks E(wt) for hospital B is clearly higher than for hospital C which is clearly higher than for hospital A.

4. (a) i. For each H_o we use a χ^2 test.
For (1):

$$\begin{aligned} \chi^2_o &= \text{null deviance - residual deviance} \\ &= 1208.2 - 64.4 = 1143.8 \\ \text{P-value} &= \Pr(\chi^2_2 \geq 1143.8) = \text{very small} \end{aligned}$$

For (2):

$$\begin{aligned} \chi^2_o &= \text{residual deviance} = 64.4 \\ \text{P-value} &= \Pr(\chi^2_{37} \geq 64.4) = 0.003 \end{aligned}$$

ii. The first test indicates quite clearly that this model is better than the Null model. The second test indicates that we should be able to do better than this model - it does not adequately describe the data.

(b) It tests $H_o : \beta_{dose^2} = 0$ given both dose and month are in the model. The value indicates strong evidence against H_o and thus dose^2 should be kept in the model.

χ^2 test:

$$\begin{aligned} \chi^2_o &= \text{res dev from the previous model} - \text{res dev this model} \\ &= 64.36 - 110.10 = 26.26 \\ \text{P-value} &= \Pr(\chi^2_1 \geq 26.26) \end{aligned}$$

(c) i. $\text{odds ratio} = \dfrac{\text{odds(cancer}|\text{24 months)}}{\text{odds(cancer}|\text{12 months)}} = \exp(12 \times 0.2708) = 25.8$

ii. $\text{odds ratio} = \dfrac{\text{odds(cancer}|\text{dose} = 0.5)}{\text{odds(cancer}|\text{dose} = 1.5)} = A$

First find:
$\Delta\text{logit}\,\hat{\pi} = (3.41 \times 1.5 - 0.984 \times 1.5^2) - (3.41 \times 0.5 - 0.948 \times 0.5^2) = 1.4501$
Then:
$A = \exp(1.4501) = 4.26$

(d) Deviance Residuals:
- detects outliers
- no real problem (11 a bit high)

Leverage Plot
- detects high leverage points
- 40 is quite high

Cooks Distance
- detects points that have a big overall impact on fitted coefficients
- 40 is very high

Deviance Changes
- detects points that make an unusually large contribution to the residual deviance
- 11 is somewhat large

Should investigate observation 40 (and possibly observation 11). Try fitting the model after deleting these observations and see how much it changes.

5. (a) Use Poisson regression with counts as the response and S, G, D, C and A as categorical explanatory variables. Use a stepwise procedure to identify which interactions are needed in the model. The association graph is produced by joining any pair of factors that occur in the same active interaction.

(b) S is directly related to G and C
S is independent of A conditional on C
S is independent of D conditional on C and G

(c) Two factors are conditionally independent if they are independent provided that the levels of one or more other factors are fixed. Thus in part (a) S and A are independent if we only consider each level of C separately but are related to each other if we consider all levels of C together. If two factors are independent (unconditionally) then they are not related under any circumstance.

(d) i. yes
ii. no
iii. yes (but this is not very useful since A is independent of S)

We can collapse on A as it is independent of all other factors. This leaves a 3-way table for S, G and D. We can collapse on any factor not directly related to both other factors (i.e. on S or D but not on G).