1. (a) i. A plot of residuals vs. fitted values. A plot of residuals vs. explanatory variables or a partial residual plot. In each case a non-linear trend indicates a non-linear regression surface.

   ii. A normal plot of residuals. A non-linear trend indicates non-Normality.

   iii. A plot of residuals versus lagged residuals. A positive or negative trend indicates serial correlation.

   (b) Multicollinearity means that there is a near linear relationship between the regressors. It makes the model selection more difficult as there may be several models that all perform approximately as well as each other. The value of a VIF for a regressor that is involved in a near linear relationship will be larger than otherwise. VIF's from 5 - 10 indicate moderate multicollinearity, whereas values > 10 indicate severe multicollinearity.

   (c) For a case-control study a sample of people who had the disease and a sample of controls would be selected and each sample would be classified by smoking habits. For a cohort study a sample of smokers and a sample of non-smokers would be selected and these groups followed over time to see how many developed the disease. For the case-control study we cannot calculate Pr(disease|smoking) or the relative risk between non-smokers and smokers. However we can calculate the odds ratio:

$$\frac{\text{odds(disease|smoker)}}{\text{odds(disease|non-smoker)}}.$$

   So the odds ratio should be used.

   (d)

$$\widehat{\text{odds ratio}} = \frac{53 \times 497}{19 \times 829} = 1.67$$

$$\log(1.67) \pm 1.96\sqrt{\frac{1}{19} + \frac{1}{53} + \frac{1}{497} + \frac{1}{829}}$$

$$0.513 \pm 0.536$$

$$(-0.022, 1.050)$$

$$(\exp(-0.022), \exp(1.050)) = (0.98, 2.86)$$

2. (a) Disagree. This statement ignores the interaction. The model says that if climb is fixed at a value c then for each increase of 1km in distance the expected time increases by $7.89 + 9.10 \times c$ minutes.

   (b) i. Non-constant variance which is indicated by the increasing trend in the first plot and the increasing scatter of the squared residuals in the second plot.

   ii. Based on the output a power transformation of $y^{(p)}$ where p = 1 - 1.28 = -0.28 should be used. Anything from p = 0 (log) to p = -0.5 would be reasonable.

   iii. The linearity of the regression surface and the Normality of the response distribution would also be affected. An alternative remedy would be to use weighted least squares.

   (c) i. A large h_{ii} indicates that the observation is unusual in its x-values and thus has the potential of having a large influence on the fitted model. A large Cook's Distance means the observation has a large influence on the fitted model.

   ii. Points 7, 11 and 35 have a big impact on the fitted model. They are also all unusual in terms of their x-values. I would try deleting these 3 points and assess how the model would change if they were dropped.

   (d) i. The fitted value for the 7^{th} observation would change by 5.126 standard errors if this observation were dropped.

   ii. The coefficient for Distance would change by -2.511 standard errors if this observation were dropped.

3. (a) i. 83.17% of the variability in Time for this data set can be explained by this model.

   ii. Very strong evidence against the coefficients for the explanatory variables all being 0. This means that the model has some predictive value.

   (b) Since many of the P-values for the individual t-tests are large, we should be able to drop some of the explanatory terms. I would start by dropping the 3-way interaction, and refitting the model. Then I would try dropping some 2-way interactions.

   (c) This model indicates that time depends on age and alcohol but not on gender. The inclusion of the alcohol:age interaction indicates that the effect of age depends on the level of alcohol and the effect of alcohol depends on the level of age. For alcohol = yes we can write the regression between time and age as:

$$\text{E(time)} = 13.42 + 0.49 \times \text{age}$$

   For alcohol = no:

$$\text{E(time)} = 16.12 + 0.19 \times \text{age}$$

   These models indicate reaction times increase with age and the rate of increase is faster if alcohol has been consumed.

To the evaluate the effect of alcohol we need to plug in values for age:

	age		
	20	40	60
alc = yes	23.2	33.0	42.8
alc = no	20.0	23.7	27.5

We can see that reaction times are higher if alcohol has been consumed and that the size of the effect that alcohol has on time increases with age.

(d) The model in (c) seems reasonable. The model consists of a linear relationship between time and age which has a different slope and intercept if alcohol has been consumed. Since Sex is not included, the model indicates that these relationships should be the same for both females and males. This is compatible with the coplot: (i) the relationship between time and age is approximately linear in each panel, (ii) the plots for Sex=M are very similar to those for Sex=F and (iii) the slope appears to be different for the plots with alcohol=Y from those with alcohol=N.

There is only a small reduction in R^2 between the full model and the reduced model. For the reduced model since the interaction term is significant we should keep the term for alcohol in the model.

4. (a) i. H_0: Coefficients for sbp, dbp and ht are all 0.
   ii. $\chi^2_0 = 135.52 - 134.85 = 0.67$
   iii. Pr value = $\Pr(\chi^2_3 \geq 0.67)$
   iv. No evidence against H_0.

(b) For each increase of 1 year in age the odds of a coronary incident get multiplied by $\exp(0.053) = 1.05$.

A 95% CI is obtained from

$$\exp(0.053 \pm 1.96 * 0.0208) = \exp(0.012, 0.094)$$

$$(1.01, 1.10)$$

(c)

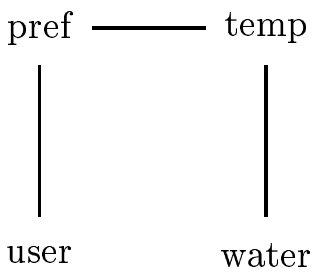
$$\text{logit}\,\hat{\pi} = -9.256 + 0.053 \times 40 + 0.00652 \times 300 + 0.0175 \times 180 = -2.03$$

$$\hat{\pi} = \frac{\exp(-2.03)}{1 + \exp(-2.03)} = 0.116$$

(d) The deviance residual plot is used to detect outliers. Leverage plot detects observations with unusual x-values. Cook's Distance and Deviance Changes plots detect influential points. Cook's Distance measures the change in the fitted coefficients if that point were dropped and Deviance Changes indicates the change in the residual deviance.

There are no real problems indicated by these plots.

5. (a) The association graph is:



i. None present.

ii. pref and temp.

iii. pref is conditionally independent of water.

(b) Two factors are conditionally independent if they are independent given that the level of a third factor is fixed. For this example pref and water are conditionally independent. Thus if we only consider the results for one of the levels of temp (high or low) pref and water are independent. However if we combine the results, pref and water are not independent.

(c) It is sensible to collapse the table for pref and user but not for pref and water. For pref and water we can collapse on user and then consider the relationship between pref and water for each level of temperature.

(d) i. Use logistic regression to model the probability that the consumer prefers M over X. The response is the proportion who choose M for each combination of the explanatory variables (water, user and temp). Fit a logistic regression model using the totals for each covariate pattern as weights.

ii. We would expect that user and temp would be in the logistic regression model. The Poisson model indicates that pref is related to user and temp but not water. Further, there are no 3-way interactions involving pref in the Poisson model so we wouldn't expect any 2-way interactions in the logistic model.