Model answers to Part B of 2005 Final Exam

Question1 (a)

Anova tables: terms are added one at a time, comparing the current model with the previous model. Terms are added in the order they appear in the model formula. Advantage: quick and does not involve much computation. Disadvantage: Result depends on ordering of terms in the model, p-values not reliable (multiple tests).
APR: Calculates various measures of "model goodness" for each possible regression model, and displays results ordered by R2 for each possible model size. Advantage: lokks at all possible models. Disadvantage: computationally intensive.
Stepwise: Finds a good model by adding and subtracting variables, tying to decrease the AIC each time. Advantage: doesn't require much computation, and is better than the anova approach (considers more models). Disadvantage: not all models considered.

Q1(b) Different criteria are

R^2: only good for comparing models with the same number of variables.

Adjusted R^2: This discounts the R^2 for model size, so can be used to compare models with different numbers of variables. The bigger the adjusted R^2, the better the model.

s^2: residual variance, the smaller the residual variance, the better the model. Gives the same result as adjusted R^2.

Cp: A measure of how well the model predicts. Cp small and approximately equal to p is a good model. Not as good as CV.

CV: Cross-validation. This divides up the data into test and training sets, fits the model using the training set, and estimates the prediction error using the test set. This is done several times and the estimates averaged. Good models are those with small prediction errors. The best method we used.

AIC: Penalises the RSS for big models. Related to Cp.

BIC: Penalises the RSS for big models more severely than AIC. Tends to select simpler models than AIC.

Q1(c) The model picked by the Adjusted R^2 and s^2 criteria is the model deleting CYL. The model picked by Cp is the model deleting CYL and TRANS. The model picked by AIC. BIC and CV is the model deleting COMP, TORQ, CYL and TRANS.
We pick this last model , as AIC, BC and CV are generally better model selection techniques.

Overfitting (putting too many variables) usually results in a predictor whose error is inflated. Underfitting results in a biased predictor.

Question 2 (a). The saturated model is a model that fits a separate success probability to each covariate pattern, with no restriction on the form of the probability. The estimate of the probability is r_i/n_i where r_i is the number of successes and n_i the number of trials for the ith covariate pattern.

The Null model is the model that fits a single identical probability to all the covariate patterns.

The logistic model fits probabilities determined by the logistic curve (i.e. where the log-odds are linear in the covariates.)

The null deviance is $2(\text{Log } L_{SAT} - \text{Log } L_{NULL})$ where $\log L_{SAT}$ is the maximum of the log-likelihood under the saturated model (i.e with the saturated-model probability substituted into the log-likelihood) and $\log L_{NULL}$ is the maximum of the log-likelihood under the null model (i.e with the null-model probability substituted into the log-likelihood) .

The residual deviance is $2(\text{Log } L_{SAT} - \text{Log } L_{MOD})$ where $\log L_{MOD}$ is the maximum of the log-likelihood under the logistic model (i.e with the logistic-model probability substituted into the log-likelihood) .

(b) Inconsistent performance would be when the probabilities of a successful shot changed from game to game. Conversely, consistent performance would be when the probabilities are the same for each game. This is the null model. We can test for consistent performance by seeing if the null model fits, i.e if the null deviance is sufficiently small.

The p-value is just a little bit over 0.05, suggesting that there is not much evidence against the null model. It seems that the press were a bit hard on O'Neal.

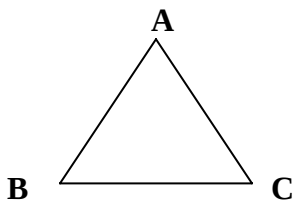
(c) Over-dispersion is when the number of successes is more variable than would be the case under the binomial distribution. (Under-dispersion is the reverse). Over dispersion would occur if the throws were not independent. This could happen if the player gained confidence if successful (or lost confidence if not successful). This it is possible that the chance of success might increase after previous successes, resulting in dependence and overdispersion.

Question 3 (a).

Suppose we have a Poisson regression model for the contingency table. We can split the Poisson means up into main effects and interactions. The corresponding multinomial model is one whose probabilities are proportional to the Poisson means.
Various kinds of independence models for the multinomial probabilities correspond to submodels of the saturated Poisson regression model as follows:

A, B, C independent corresponds to **`y ~ A + B + C`**
A independent of B and C corresponds to **`y ~ A + B*C`**
A, B conditionally independent given C corresponds to **`y ~ A*C + B*C`**
Homogeneous association corresponds to **`y ~ A*B + A*C + B*C`**

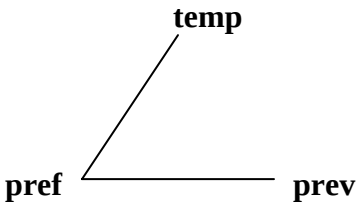
(b) Association graphs are graphs where each factor is represented by a node, and two factors are joined by a line if there is an interaction between them. The models **`y ~ A*B*C`** and **`y ~ A*B + A*C + B*C`** both have the same graph:



(c) (i) The 3-factor interaction has a p-value of 0.095. If we interpret this as insignificant, which seems reasonable, then both the temp:prev and 3-factor interactions are zero. This gives the model

`count ~ temp*pref + pref*prev`

which is the model of conditional independence of previous use and temperature given preference. This model has graph



(ii) This suggests we can collapse the table over temperature. In the collapsed table, previous use and preference seem to be independent.

(iii) Since preference for brand X is a binary variable, we could do a logistic regression using preference for brand X as the response. This would be modelling the conditional distribution of preference given the other variables. The contingency table approach is modelling the joint distribution.