ANCOVA & Variables Selection

Due: Monday 11/01 (11.00PM)
Submission: On Gradescope

Part I: Practice Questions

1. Using the *teengamb* data set with `gamble` as the response and the other variables as predictors. Implement the following variable select methods to determine the best model:

   (a) Backward elimination

   (b) AIC

   (c) Adjusted R^2

   (d) Mallows C_p

2. Using the *teengamb* data set with `gamble` as the response and the other variables as predictors. Investigate the possibility of interactions between `sex` and the other predictors.

Part II: Homework Questions – to be submitted

1. In an experiment to investigate the effect of color of paper (blue=0, green=1) on response rates for questionnaires distributed by the "windshield method" in supermarket parking lots, 26 representative supermarket parking lots were chosen in a metropolitan area. Blue questionnaires were randomly assigned to 9 lots, green questionnaires were randomly assigned to 10 lots. It has been suggested to the investigator that size of parking lot might be a useful variable, so they decided to use it as a continuous predictor. The response is the average response rates.

   (a) State

      (i) the regression line that corresponds to the blue questionnaires.

      (ii) the regression line that corresponds to the green questionnaires.

   (b) Test whether or not the interaction term is statistically significant. State the hypotheses, decision rule and conclusion.

   (c) Does the response rate vary according to the questionnaire color? Justify your answer.

   The data can be found in the `questionnaire.csv` file on Moodle.

2. The Major League Baseball data set contains data for 322 major league players from the 1986 and 1987 seasons. We are interested in predicting the *Salary* of a player. We have available the following predictors:

`AtBat` (Number of times at bat in 1986)	`Hits` (Number of hits in 1986)
`HmRun` (Number of home runs in 1986)	`Runs` (Number of runs in 1986)
`RBI` (Number of runs batted in in 1986)	`Walks` (Number of walks in 1986)
`Years` (Number of years in the major leagues)	`CAtBat` (Number of times at bat during his career)
`CHits` (Number of hits during his career)	`CHmRun` (Number of home runs during his career)
`CRuns` (Number of runs during his career)	`CRBI` (Number of runs batted in during his career)
`CWalks` (Number of walks during his career)	`PutOuts` (Number of put outs in 1986)
`Assists` (Number of assists in 1986)	`Errors` (Number of errors in 1986)

The data set is called `Hitters` and can be found in the `ISLR` library.

(a) What is the *optimal* number of parameters according to

(i) the adjusted R^2 criterion,

(ii) the Mallows C_p criterion,

(iii) the AIC,

(iv) the BIC.

(b) We then use the `step` function for model selection. Explain which variables are removed and in which order.