STAT 302 - Chapter 4: Additional Topics in Regression - Part 1

Harsha Perera

Key Topics

- Techniques for Choosing Predictors (Model Selection)
- Best Subsets
- Mallow's C_P
- Alternative criteria for Model Selection : AIC and BIC
- Backward Elimination
- Forward Selection and Stepwise Regression

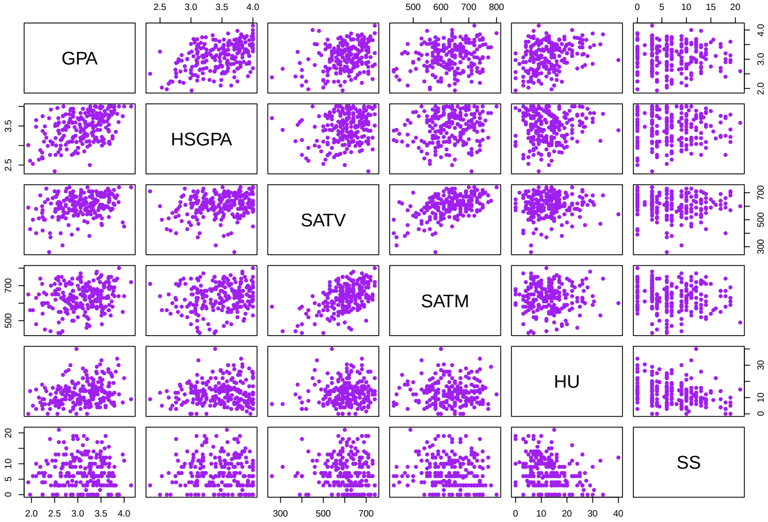
Techniques for Choosing Predictors (Model Selection)

- Model selection can be used to find which factors really affect the mean response
  - i.e. a model that predicts the response
- Importance of model selection
  - In regression we look for a simple model with higher R^2
  - R^2 will increase with the number of predictors
    - Inclusion of higher number predictors results in complex model
    - A higher R^2 does not mean all predictors contribute to response
  - We can create new predictors from existing predictors
  - Our objective should be to find a model which contains a set of predictor(s) that contribute to response with a
    - higher R^2
    - simple set of predictors

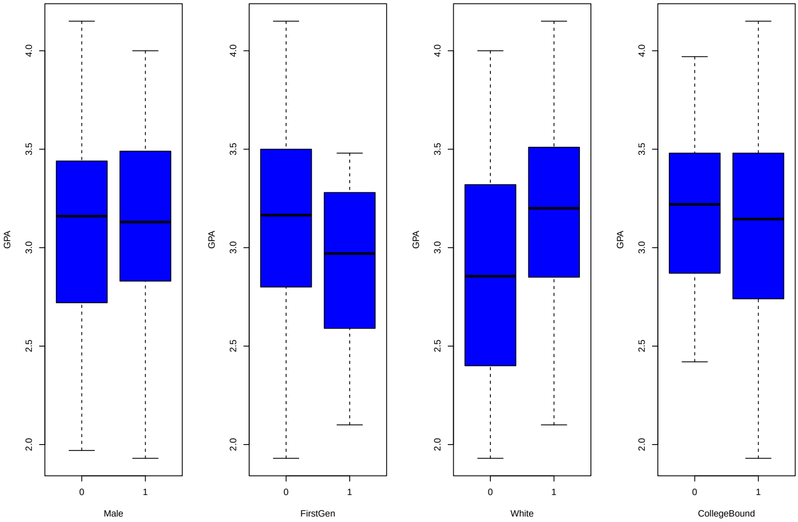
Example: First-year GPA

- The data set **FirstYearGPA** contains measurements on 219 college students
- Response: GPA - after one year of college
- Predictors:
  - HSGPA: High school GPA
  - SATV: Verbal/critical reading SAT score
  - SATM: Math SAT score
  - Male: 1 for male, 0 for female
  - HU: Number of credit hours earned in humanities courses in high school
  - SS: Number of credit hours earned in social science courses in high school
  - FirstGen: 1 if the student is the first to attend school in his/her family
  - White: 1 for white and 0 for others
  - CollegeBound: 1 if attended a high school where $> 50\%$ of students intend to go on to college

First-year GPA: Continuous Predictors



First-year GPA: Categorical Predictors. . .



First-year GPA: Model Selection

- Objective: to find the which of the nine predictors really affect the mean response
- For simplicity we will only consider main effects
  - Interaction terms
  - Polynomial terms will not be considered

Best Subset Regression

- This methods work well when number of predictor are not too large
- Checks all possible models
- But returns the best set of models among all possible models
- You may need to call $library(leaps)$
- In R: we can use the regsubsets() function to obtain best subset

First-year GPA: Best subset regression

- Model with all nine predictors has the highest R^2 - 34.96%
- A model with 6 predictors has the highest R^2_{adj} - 32.85%
  - Predictors: HSGPA, SATV, Male, HU, SS, White
- Another option would be the model with 5 predictors - 32.83%
  - Predictors: HSGPA, SATV, HU, SS, White

First-year GPA: Fitted model with six predictors

```
##
## Call:
## lm(formula = GPA ~ HSGPA + SATV + Male + HU + SS + White, data = FirstYearGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06228 -0.26731  0.05287  0.27230  0.85843
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.5466634  0.2835072   1.928   0.0552 .
## HSGPA       0.4829491  0.0714659   6.758 1.33e-10 ***
## SATV        0.0006945  0.0003449   2.013   0.0453 *
## Male        0.0541049  0.0526937   1.027   0.3057
## HU          0.0167958  0.0038181   4.399 1.72e-05 ***
## SS          0.0075702  0.0054421   1.391   0.1657
## White       0.2045215  0.0685954   2.982   0.0032 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3814 on 212 degrees of freedom
## Multiple R-squared:  0.347,  Adjusted R-squared:  0.3285
## F-statistic: 18.78 on 6 and 212 DF,  p-value: < 2.2e-16
```

- Variable Male and SS is insignificant at 5% level of significance
- 5 predictor model seems to be more sensible (excluding variable Male)

Mallow's C_p

- Both R^2 and R^2_{adj} evaluate a model based on the predictors that were already included in the model
- None of these measures takes into account what information might be available in the other potential predictors that aren't in the model.
- Mallow's C_p overcomes this problem

$$C_p = \frac{SSE_m}{MSE_k} + 2(m+1) - n$$

- k: all possible predictors
- m: number of predictors in the subset
- n: sample size

Mallow's C_p

- We need to look for a model with lower C_p
- MSE_k and n are constants
- When we add a predictor to model SSE_m decreases but $m+1$ increases
- If the predictor contributes to the response, the decrease in SSE_m is substantial in comparison to increase in $m+1$
- In general we consider models: $C_p < m+1$

First-year GPA: Mallow's C_p

- A model with 5 predictors has the lowest C_p - 3.8924
  - Predictors: HSGPA, SATV, HU, SS, White
- Another option would be the model with 4 predictors C_p- 3.9005
  - Predictors: HSGPA, SATV, HU, White
- 4 predictor model is more sensible
  - All predictors significant at 5% level
  - Simpler model is always preferred over a complex model

Alternative criteria for model selection

- AIC (Akaike's information criterion) and BIC (Bayesian information criterion)
- Similar to C_p smaller values are preferred for both AIC and BIC
- Both methods account for number of predictors in the model and how well the response is explained

Model Building Strategies

- We will be considering three model building strategies

1. **Backward Elimination**: Start with the full model and try to drop terms, down to some smaller model.
2. **Forward Selection**: Start with the smallest model (e.g., null model) and try to add terms up to some larger model.
3. **Stepwise Regression**: Try adding and dropping terms, staying between a smallest and largest model.

Backward Elimination

- Start by fitting the full model (model with all possible predictors)
- Identify the terms for which the individual t-test produces the largest p-values
  - If the largest p-value is greater than 0.05, eliminate the term and re-fit the model
  - If the largest p-value is smaller than 0.05, all predictors are significant

Backward Elimination...

- We can replace the p-value with other selection criteria such as C_p, AIC and BIC
- We can eliminate predictors based on C_p, AIC or BIC (minimize) rather than relying on significance of the predictors
- We can eliminate predictors by looking at the largest drop in C_p, AIC or BIC until we reach a point that C_p, AIC or BIC does not get smaller

Forward Selection

- Start with the model with no predictors
  - Find the best single predictor which has the largest correlation with the response
- Add the new predictor to the model
  - Fit the model
  - Find the p-value of individual t-test
  - If p-value < 0.05: keep the predictor in the model
    - Repeat the above steps and try each of the remaining predictors
  - If p-value > 0.05: stop and discard the predictor
    - No predictors will be added to the model

Stepwise Regression

- Combines features of both forward selection and backward elimination
- Begins with forward selection step
- Once a predictor is added to model, perform backward elimination