STAT1006 Week 11 Cheat Sheet

Lisa Luff

11/4/2020

Contents

Multilinear Model Explanatory Variables

- For multilinear models with lots of explanatory variables, not all will be significant to the model, and some variables might be correlated with each other (multi-collinearity), leading to biased predictions. As such in both cases, these variables should be removed, as most models of observational data are made for the purpose of making a predictive model
- Aiming to find parsimonious models (as few variables as possible while maintaining goodness-of-fit)
- Best model is subjective based on goals
- How to deal with many explanatory variables:
    1. Identify the main objectives of the analysis
    2. Justify the potential inclusion of each variable in the model
    3. Exploratory and graphical analysis using scatterplots and correlations
    - Remove one of each pair of highly collinear variables
    - Consider possible transformations of explanatory variables and/or response variable(Y)
    4. Find a suitable subset of explanatory variables

Variable Selection

- There are several methods to selecting variables to be included, all with their own pros and cons, and there is not one "best" method
    - For any given scenario, you need to decide what is the most appropriate trade-off between model complexity (minimising complexity by minimising variables included in modelling), and goodness-of-fit
- All compare models that include various combinations of variables with each other
    - This means for any given data set, there are 2^m possible models to choose from, where m is the number of explanatory variables
- Subset selection methods:
    - Brute-force
    - Stepwise
        * Forward
        * Backward
        * Both directions
    - Regularisation (not covered)
        * Shrinkage (no variable selection)
        * Shrinkage and selection
- Always need to conduct usual diagnostics on final model

Brute-Force/All Subset Selection

- Compare all possible models with all other possible models
- There will be 2^m models, where m is the number of variables
  - If m becomes too large, we evaluate all possible 2^q subsets, where $q \ll m$
- Subset models will include:
  - Every version of only a single variable being included
  - Then every version of two variables being included
  - So on, and so forth, until you have a model with all p variables included
- These subsets will be evaluated based on a set of criteria, including:
  - Adjusted R^2 - $R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$
  - Akaike Information Criteria - AIC $= nlog\left(\frac{SSE}{n}\right) + 2p$
  - Bayseian Information Criteria - BIC $= nlog\left(\frac{SSE}{n}\right) + plog(n)$
  - Mallow's Comparison - $C_p = \frac{SSE}{\hat{\sigma}^2} + 2p - n$
    - If values for these get out of control, plot their log
- Combining these criteria is considered a good amount of trade-off between 'goodness-of-fit' (small SSE), and the number of variables included in the model
- Models wont always agree on best model, so choose a small selection of models to compare in more detail

Criteria 1: Adjusted Correlation Coefficient

- R^2 - Will always increase as a new variable is added
  - $R^2 = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST}$
- R^2_{adj} - Takes into account the number of variables in the model
  - $R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST} = 1 - \frac{\hat{\sigma}^2}{s^2}$
    - Where $\hat{\sigma}^2 = MSE = \frac{SSE}{n-p}$
      - MSE also takes int oaccount the number of variables in the model
- Can include irrelvant variables, so not to be used alone
- **Want the highest possible R^2_{adj}**

Criteria 2: Information Criteria

- Akaike Information Criteria
  - AIC $= nlog\left(\frac{SSE}{n}\right) + 2p$
- Bayesian Information Criteria
  - BIC $= nlog\left(\frac{SSE}{n}\right) + plog(n)$
- Both are essentially the same thing
- **Want lowest possible value for both**

Criteria 3: Mallow's Comparison

- $C_p = \frac{SSE}{MSE_{full}} + 2p - n$
  - MSE_{full} includes intercept and ALL explanatory
- **Want $E(C_p) \approx p$**

Stepwise Regression

- These variable subset methods carry out a sequential search of the possible regressions models, which leads to evaluating significantly fewer models
- They do not guarantee that the optimal subset will be found based on any criteria, but will give a result in practice
- The models are evaluated using:
  - F statistic
  - AIC
  - BIC
- Stepwise methods include:
  - Forward
  - Backward
  - Both directions
- Be warned:
  - Forward and backward can lead to different models
  - Does not look to minimise SSE, MSE or optimise for R^2_{adj}
  - Does not account for multi-collinearity
  - Generally chooses too many explanatory variables

Forward Selection

- Using F statistic:
  1. Start with the constant mean model $Y = \beta_0 + \varepsilon$
  - No explanatory
  2. Consider all possible models with only one explanatory variable
  - Calculate the F statistic for each compared to the constant mean model
  3. Add the variable from the model that has the highest F statistic
  - **Only if F stat $>$ 4**
  4. Then consider all possible two explanatory variable models
  - Calculate the F statistic comparing with the model chosen in step 3
  5. Repeat step 3
  6. Continue until there are no more models with a F statistic > 4
  7. **Analyse the selected model**
  - Find parameter estimates
  - Run diagnostic tests of the residuals
  - Then use to make the required inferences
- Using AIC:
  - Do the same, but add the variable from the model that yields the smallest AIC
    - **Until the AIC is higher than that of the current model**

Backward Selection

- Using F statistic:
  1. Start with full model containing all K variables
  2. Consider all model with K - 1 variables
  - Calculate the F statistic for each compared with the full model
  3. Remove the variable from the model that has the smallest F statistic
  - **Only if F stat $<$ 2**
  4. Repeat until there are no more variables with a F statistic < 2
  5. **Analyse the selected model**
  - Find parameter estimates
    * Run diagnostic tests of the residuals
    * Then use to make the required inferences
- Using AIC:
  - Do the same, but remove the variable from the model that yields the highest AIC
    * **Until the AIC is lower than that of the current model**

Stepwise Selection/Both Directions

- Same name as for the group, accounts for "both directions"
- Adds and drops at each step of selection

Multi-Collinearity

- When explanatory variables are highly correlated with each other
- Can cause issues when fitting a regression model
  - Increased variance
  - Less accurate predictions due to being biased
- We can show the issue by setting correlation as r_{12}
  - For $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{2i} + \varepsilon_i$
    * $E(\hat{\beta}_2) = \beta_2 + r_{12}\beta_1$
    * $Var(\hat{\beta}_2) = \sigma^2 \left(\frac{1}{1-r_{12}^2}\right)\left(\frac{1}{\sum(X_{2i}-\overline{X}_2)^2}\right)$
  - If the correlation is large, the denomenator becomes small, and variance will be large
- Find correlation between variables, and remove one variable from any pair with high correlation (**$>$ 0.5**)

In R:

- Brute-force -
  - View p-value for all variables -
    print(load(".RData"))
    lm <- lm(y ~ ., data = *data*)
    summary(*lm*)$coefficients
  - Evaluate n best models with up to x variables
    require(leaps)
    subsets <- regsubsets(y ~ ., nbest = n, nvmax = x, data = *data*)
    subsetsummary <- summary(*subsets*)
    $subsetsummaryoutmat * Betterversionofthegraphlibrary(kable)kable(*subsetsummary*$outmat)
    - Will show a list of models, far left number is number of variables, second number is if it's the 1st, 2nd, etc best option for that number of variables
    - Compare the graphs of models
      par(mfrow = c(1, 3))
      # Plot R^2
      plot(1:10, *subsetssummary*$adjr2, **log = "*y*"**) (If using log of values)
      # Plot BIC
      plot(1:10, *subsetssummary*$bic, **log = "*y*"**) (If using log of values)
      # Plot C_p
      plot(1:10, *subsetssummary*$cp, **log = "*y*"**) (If using log of values)
  - Look at p-value for each model with x number of variables
    subsetmatrix <- *subsetssummary*$outmat *lmp* <- lm(formula(paste("y", paste(names(which(*subsetmatrix*[,*x*]=="*")), collapse="+"))),
    data = *data*)
    summary(*lmp*)

  - Extract AIC
    extractAIC(*lm*, k = *df weight*)

- Forward selection -
  - Minimal model
    lm -> lm(y ~ 1, data = *df*)
  - See the list of AIC values
    lm <- lm(y ~ *current model variables*, data = *data*)
    lmforward <- step(*lm*, scope = ~ $x_1 + x_2 + ... + x_n$ (not including variables in current model), direction = "forward")
    - is the current model
  - OR
    step(*lm*, scope = formula(*df*), direction = "forward")
  - Can use trace = 0 to hide the steps
  - OR
    regsubsets(y ~ ., nbest = n, data = *data*, method = "forward")

- Backward selection -
  - Same as forward, but with direction/method = "backward", and don't need scope

- Stepwise selection/both direction selection
  - Same as forward and backward, except direction/method = "both"
    - Don't actually need direction, as "both" is default

- Multi-colinearity
  - Coloured graph showing levels of correlation
    require(corrplot)
    corrplot(cor(*data*[,*columnstoremove*]))
    - To see a single row
      corrplot(cor(*data*[,*columnscinluded*])[*columnsincluded*, *columntoview*, drop = FALSE], cl.pos = 'n', method = 'number')
  - A different, uglier version
    require(lattice)
    splom(~*data*[,*columnstoremove*], groups = *category*, data = *data*, pscales = 0, varname.cex = 0.5)
  - VIF (variance inflation score) **- want to be less than 5**, else possible multicollinearity
    library(car)
    vif(*lm*)