STAT1006 Week 12 Cheat Sheet

Lisa Luff

11/6/2020

Contents

Interaction in Regression

- An alternative to variable selection
- When there is multi-collinearity between explanatory variables, rather than remove one of them, treat them as interacting (take collinearity into account, and keep both)
- This is used when:
  - There is a joint effect on the response
    - So their combined correlation is more significant than their individual correlation
    - Or explanatory variables show as being insignificant in the linear model, but there is a high R^2 and R^2_{adj} value
  - And/or both are equally import to take into consideration for the purpose of the experiment

Interaction Model

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$, where
  - β_0 is the intercept
  - β_1 and β_2 and the slopes of explanatory variables 1 and 2
  - β_3 **The Interaction Term** - takes into account the interaction between explanatory variables 1 and 2
- β_3:
  - Positive - means the response is boosted (or enhanced) by large values for both X_1 and X_2
  - Negative - meand the response is reduced (or interfered with) by large values for both X_1 and X_2
    - **Be careful** if the explanatory variables have negative values
- In practice, β_3 will be an additional variable included in model assessments, created by multiplying explanatory variable 1 with explanatory variable 2
  - Models including interaction terms should be compared to a full model of all terms and interaction terms

Predictive Ability of Linear Models

- To ensure models are a good measure of predicted values, you can use two methods:
  - Internal measure - PRESS (Predicted Residual Sum of Squares) - $\hat{y}_{-i}$
  - External measure - RMSEP (Root Mean Squared Error of Prediction) - Test set

PRESS

- Predicted Residual Sum of Squares

1. Leave out the i^{th} observation for y_i (Denoted as $\hat{y}_{-i}$), then fit a model
2. Then use the model to predict the value of the y_i that was left out
3. Calculate the residual $y_i - \hat{y}_{-i}$ and square it
4. Repeat for all values of y_i

- $PRESS = \sum_{i=1}^{n} (Y_i - \hat{y}_{-i})^2 = \sum_{i=1}^{n} \left(\frac{\hat{e}_i}{1-h_{ii}} \right)^2$, where
  - $\hat{e}_i$ is error
  - h_{ii} is the hat martix element to represent leverage
  - **The lower the better**

RMSEP

- Root Mean Squared Error of Prediction

1. Split data randomly into a training set and a test set

- About 80% - 20% split

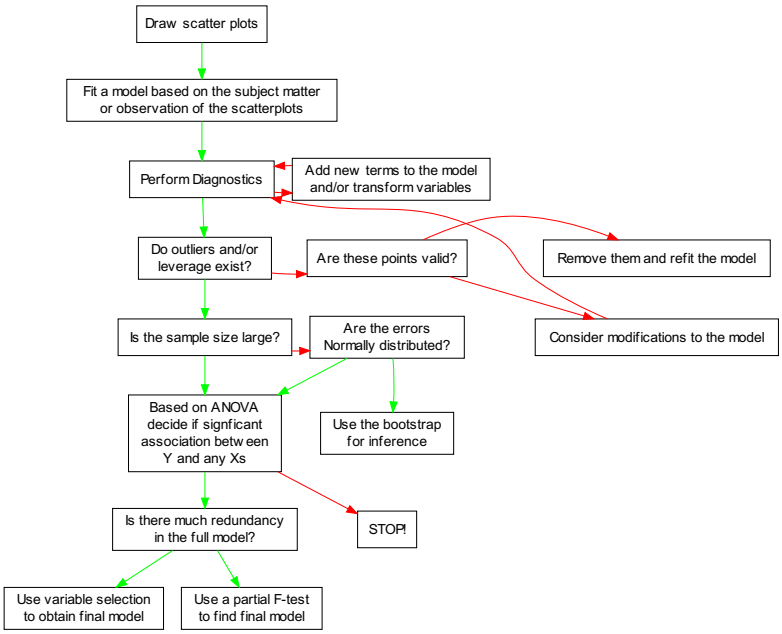
2. Create model from training set
3. Calculate root mean squared error of prediction on test set

- $RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i^t - \hat{y}_i^t)^2}{n_t}}$, where
  - y_i^t and $\hat{y}_i^t$ are the the test values for y_i and $\hat{y}_i$
  - n_t is the number of test values
  - **The lower the better**
    - Keep in mind different set splits will result in different RMSEP values

4. Calculate a plot of the predicted values against the actual values

- The closer to the line, the better the prediction

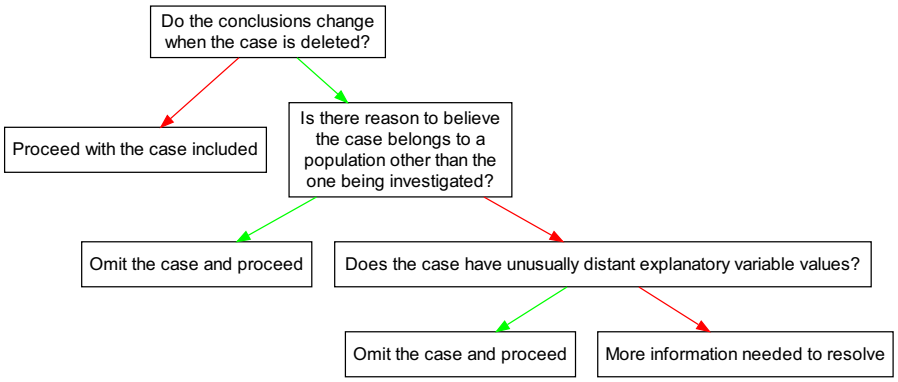
Strategies for Modelling



Bias-Variance Trade-Off

- Bias is the amount of error introduced by approximation using a model that is simpler than real life
- $Var(\epsilon)$ is the irreducible error, which is an upper bound on the accuracy of our prediction
- The general version of our model is $y = f(X) + \varepsilon$
  - The best estimate of the response is $\hat{Y} = \hat{f}(X)$
  - For any new data (x_0, y_0), our prediction is $\hat{Y}_0 = \hat{f}(x_0)$
- This means that for a large number of observations, the average squared prediction error is $Ave(y_0 - \hat{f}(x_0))^2$ (MSE)
  - We want a model that minimises this
- MSE as above can be broken down into $E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$
  - As such, they the directional inverse of each other
    - As the complexity (number of explanatory variables) increase, so will variance and bias will decrease
    - As the complexity decreases, so will variance and bias will increase
- So when we decide a model to go with, we are choosing the amount of trade-off we find acceptable for each
  - There can be many good sets, so this decision needs to be based on what best suits the objective

Strategies for Dealing with Outliers



In R

- PRESS
  - Can use leaps package, but doesn't output fitted models
    press(*modellm*)
  - Otherwise there is a package DAAG with a press function
    press(*modellm*)
- Creating a training and test set
  samplevalues -> sample(nrow(*df*), floor(0.2, * nrow(*df*)))
  test -> *df*[*samplevalues*,]
  training -> *df*[-*samplevalues*,]
- Compare model to prediction
  prediction -> predict(*modellm*, newdata = *test*)
  actual -> *df*$*y*
  plot(*actual*, *prediction*)
  abline(0, 1)

- RMSEP
  RMSEP <- sqrt(sum((*actual* - *prediction*)^2)/length(*actual*))