Model Building

Part 4: Prediction Models

STAT 705: Regression and Analysis of Variance

Prediction vs. Estimation

- Linear models can be used for estimation or prediction
- Equations for the models can be the same, but our method of assessing a 'good' model is different
- For estimation models
  - Assess 'goodness' via AIC, SBC, Adjusted R^2, etc.
  - These measure how well the model fits the sample data
- For prediction models
  - Need a measure for how well the model will predict new observations
  - One option: PRESS (PREdicted Sum of Squares)

Deleted Residuals

- For each observation
  - Temporarily delete the observation
  - Fit the model using the remaining $n-1$ observations
  - Use the fitted model to predict the response for this observation
    - Notation: Put the subscript in parenthesis, $\hat{Y}_{(i)}$
  - The *deleted residual* is the difference between the observed and predicted values

    $r_{(i)} = Y_i - \hat{Y}_{(i)}$
- Repeat this for every observation

PRESS Statistic

- PREdiction Sum of Squares
  - Sum of the squared deleted residuals
  - $\text{PRESS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{(i)} \right)^2$
- Software uses a shortcut for the calculations
  - No need to fit n separate regression models
- Small PRESS values are desirable
  - Small prediction errors

SAS Implementation

Use the ‘press’ option on the model statement

```
  proc reg data=senic outest=pressinfo;
  WithNurses:    model InfRisk = Stay CulRatio XRay Nurses
                                 Services MedSch Reg1 Reg2 Reg3 / press;
  WithoutNurses: model InfRisk = Stay CulRatio XRay
                                 Services MedSch Reg1 Reg2 Reg3 / press;
  output press=press;
  run;
proc print data=pressinfo; run;
```

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	_PRESS_	Intercept	Stay	CulRatio	XRay
1	WithNurses	PARMS	InfRisk	0.90912	104.997	-0.14941	0.26708	0.051606	0.011606
2	WithoutNurses	PARMS	InfRisk	0.91120	104.444	-0.32966	0.27469	0.052336	0.011326

Obs	Nurses	Services	MedSch	Reg1	Reg2	Reg3	InfRisk
1	.001304167	0.017633	-0.61838	-1.07183	-0.74377	-0.77185	-1
2	.	0.025459	-0.50274	-1.10697	-0.76674	-0.75937	-1

Interpreting PRESS

- For the model that predicts Infection Risk (in the SENIC dataset)
  - PRESS = 104.997 when 'Nurses' is used as a predictor
  - PRESS = 104.444 when 'Nurses' is not used
- Smaller PRESS ⇨ better predictive ability
- In this case, excluding a predictor *improves* the predictive ability (but not by much)

ADDING MORE PREDICTORS IS SOMETIMES DETRIMENTAL TO THE MODEL

Model Validation

- There is no assurance that a model that is a good fit to the existing data will also be successful for future predictions
- There could be
  - influential factors that were unknown during model building
  - a different correlation structure among the predictors
- The key idea: Test the model in the environment in which it is going to perform
- This is especially important for observational studies
- PRESS is one method for assessing the predictive ability of the model
  - PRESS is sometimes called ‘leave one out’ cross validation, LOOCV

K-fold Cross Validation

- Similar to PRESS, but operates on groups of observations instead of individual observations
- Split the observations into K groups
- For each group
  - Temporarily remove this group from the data
  - Fit the model using the observations in the other groups
  - Predict the response value of the observations that were removed
  - Calculate the residuals
- Do this for all the groups
- Calculate the sum of the squared residuals

Data Splitting

- Another way to assess predictive ability of the model
- Requires a lot of data
    - At least 30 observations; more for complex models
- Split observations into two groups
    - Training set (about 2/3 of all observations)
        - Use these observations to estimate the regression equation
    - Prediction set (about 1/3 of all observations)
        - Apply the estimated regression equation to these observations
        - Predict the value for the response
        - Calculate the residuals

Mean Square for Prediction

- Analogous to MSE, but specifically designed to assess the predictive ability

$$MSPR = \frac{1}{n^*}\sum_{k=1}^{n^*}\left(Y_k - \hat{Y}_k\right)^2$$

- Y_k is the observed response for the k^{th} observation <u>in the prediction set</u>, k = 1, 2, . . . ,n*
- $\hat{Y}_k$ is the predicted response for the k^{th} observation in the prediction set
  - Use the regression equation estimated with the training set.
- MSPR is a measure of the error, so smaller is better

What You Should Know

- How to assess model adequacy
  - Models for Estimation vs. Models for Prediction
- Understand *why* we need different methods of assessing the adequacy of prediction models
- For the PRESS statistic
  - Describe how it is calculated
  - Use SAS to calculate it
  - Interpret the results
- How is the MSPR different from the sum of squared residuals in K-fold cross validation?