**20 November 2020**

**ROC (Receiver operating characteristic) curve**

$$
\begin{aligned}
\text{POS}(z) &= \sum_{i \in \mathcal{L}_1} I(\hat{\pi}_i \geq z) \\
\text{FALNEG}(z) &= \sum_{i \in \mathcal{L}_1} I(\hat{\pi}_i < z) \\
\text{FALPOS}(z) &= \sum_{i \in \mathcal{L}_0} I(\hat{\pi}_i \geq z) \\
\text{NEG}(z) &= \sum_{i \in \mathcal{L}_0} I(\hat{\pi}_i < z) \\
\text{Sensitivity}(z) &= \frac{\text{POS}(z)}{\text{number of observations in } \mathcal{L}_1} \\
1 - \text{Specificity}(z) &= \frac{\text{FALPOS}(z)}{\text{number of observations} in \ \mathcal{L}_0}
\end{aligned}
$$

The ROC curve is a plot of sensitivity against (1 - specificity).

$$
\begin{array}{lllllll}
0.9 & < & \text{area} & < & 1 & \Rightarrow & \text{Excellent} \\
0.8 & < & \text{area} & < & 0.9 & \Rightarrow & \text{Good} \\
0.7 & < & \text{area} & < & 0.8 & \Rightarrow & \text{Fair} \\
0.6 & < & \text{area} & < & 0.7 & \Rightarrow & \text{Pass} \\
0.5 & < & \text{area} & < & 0.6 & \Rightarrow & \text{Fail}
\end{array}
$$

where area under the curve is equal to $\frac{\text{nc} + 0.5 * \text{tied}}{\text{t}}$

- nc (number of concordant) $= \sum_z \text{POS}(z) * \text{NEG}(z)$
- nd (the number of disconcordant) $= \sum_z \text{FALNEG}(z) * \text{FALPOS}(z)$
- tied (number of tied pairs) $= \sum_z \sum_{i \in \mathcal{L}_1} I(\hat{\pi}_i = z) * \sum_{i \in \mathcal{L}_0} I(\hat{\pi}_i = z)$
- t (total number of pairs) is equal to nc + nd + tied

**Chapter 4**

**Sequential Variable Selection Procedures**

Forward selection

- The initial model contains only a constant term.

- Use the test statistic

$$F = \frac{ResS.S.|_{H_0}}{ResS.S.|_{H_A}} * (n - p') - n + p'$$

- Add the variable which would have the largest $F$ statistic of any of the variables that are not already in the model.

  $\Rightarrow$ Since $ResS.S.|_{H_0}$ is the same at the same step, choose the variable which would have the smallest $ResS.S.|_{H_A}$ (or the largest $RegS.S.|_{H_A}$ or the largest $R^2|_{H_A}$) to add into the model.

- The process continues until, at some stage, the candidate regressor for entry does not exceed a preselected $F_{IN} = F_{\alpha_{IN},1,n-p'}$.

Backward elimination

- The initial model contains all regressors (independent variables).

- Use the test statistic

$$F = \frac{ResS.S.|_{H_0}}{ResS.S.|_{H_A}} * (n - p') - n + p'$$

- Remove the variable which has the smallest $F$ value for all the variables in that step

  $\Rightarrow$ Since $ResS.S.|_{H_A}$ is the same no matter what the null hypothesis is, choose the variable which would have the smallest $ResS.S.|_{H_0}$ (or the largest $RegS.S.|_{H_0}$ or the largest $R^2|_{H_0}$) to remove from the model.

- The process continues until the candidate regressor for removal exceeds the preselected $F_{OUT} = F_{\alpha_{OUT},1,n-p'}$.

Stepwise regression

- The initial model contains only a constant term.
- For each step
  1. Forward: add the variable which would have the largest $F$ statistic of any of the variables that are not already in the model if the $F$ value exceeds a preselected $F_{IN}$
  2. Backward: removed the variable which has the smallest $F$ value for all the variables in that step if the $F$ value does not exceed a preselected $F_{OUT}$
- The process continues until
  1. the candidate regressor for entry does not exceed a preselected $F_{IN}$
  2. the regressors for entry and removal are the same one.