Regression Analysis

Model Selection

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Prediction Risk Estimation

About This Lesson

Bias-Variance Tradeoff

- **Variable Selection:** Bias vs. Variance
  - Many covariates
    - Low bias, high variance
  - Few covariates
    - High bias, low variance

- **Prediction Risk:** Measure of the Bias-Variance Tradeoff

$$R(S) = \frac{1}{n}\sum_{i=1}^{n} E(\widehat{\boldsymbol{Y}}_i(S) - \boldsymbol{Y}_i^*)^2$$

with $\widehat{\boldsymbol{Y}}_i(S)$ the fitted response for submodel S and $\boldsymbol{Y}_i^*$ the future observation

We cannot obtain the prediction risk because we do not have the future observations.

How to estimate?

Training Risk

- Replace future observations with actual observations

$$R_{\mathrm{tr}}(S) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{Y}}_i(S) - \boldsymbol{Y}_i)^2$$

  with $\hat{\boldsymbol{Y}}_i(S)$ the fitted response for submodel S and $\boldsymbol{Y}_i$ the actual observation

- Uses data twice (data snooping): downward bias in prediction risk estimate

- Always prefers/selects larger/more complex model

➔ Correcting for the bias

$$R_{\mathrm{tr}}(S) + Complexity\ Penalty$$

Variable Selection Criteria

➔ **Correcting for the bias:** $R_{\text{tr}}(S) + Complexity\ Penalty$

➔ **Selection criteria** differ through the complexity penalty as follows**:**

- **Mallow's Cp** with $Complexity\ Penalty = \dfrac{2|S|\hat{\sigma}^2}{n}$

  where |S| is the model size (number of predictors) and $\hat{\sigma}^2$ is the estimated variance based on the full model.

- **Akaike Information Criterion (AIC)** with $Complexity\ Penalty = \dfrac{2|S|\sigma^2}{n}$ where |S| is the model size and σ^2 is the true variance.

> - For AIC, we need to replace σ^2 with an estimate (from the full model or from the S submodel).

Variable Selection Criteria (cont'd)

➔ **Correcting for the bias:** $R_{\mathrm{tr}}(S) + Complexity\ Penalty$

➔ **Selection criteria** differ through the complexity penalty as follows**:**

- **Bayesian Information Criterion (BIC)** with

$$Complexity\ Penalty = \frac{|\mathrm{S}|\sigma^2 \log(n)}{n}$$

where |S| is the model size and σ^2 is the true variance

- For BIC, we need to replace σ^2 with an estimate (from the full model or from the S submodel)
- BIC penalizes complexity more than other approaches
  - Preferred in model selection for prediction

Variable Selection Criteria (cont'd)

➔ **Correcting for the bias:** $R_{\text{tr}}(S) + Complexity\ Penalty$

- **Leave-one-out Cross Validation**

$$R_{\text{CV}}(S) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\boldsymbol{Y}}_{(i)}(\text{S}) - \boldsymbol{Y}_i\right)^2$$

where $\widehat{\boldsymbol{Y}}_{(i)}(S)$ is the i-th predicted value from the S submodel without the i-th observation

- **Leave-one-out Cross Validation Approximation**

$$\hat{R}_{\text{CV}}(\text{S}) \approx R_{\text{tr}}(\text{S}) + \frac{2|\text{S}|\hat{\sigma}^2(\text{S})}{n}$$

where $\hat{\sigma}^2(S)$ is the estimated variance based on the S submodel.

Generalized Linear Models

Training Risk for Generalized Linear Models (including for logistic regression and Poisson regression)

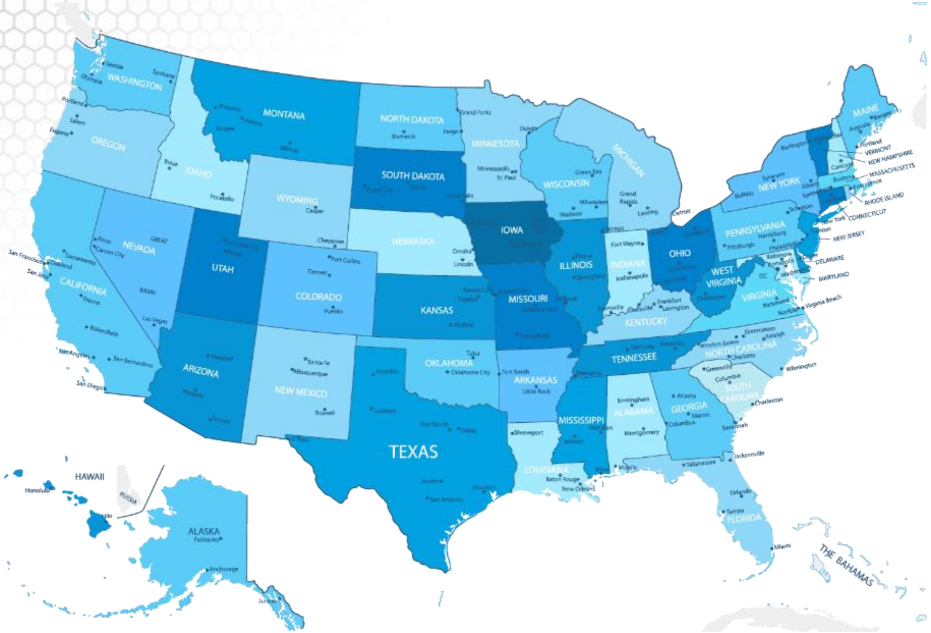
$$R_{\text{tr}}(\mathrm{S}) = \frac{1}{n}\sum_{i=1}^{n}(2\boldsymbol{Y}_i\log[\boldsymbol{Y}_i/\widehat{\boldsymbol{Y}}_i(S)]+2(n_i-\boldsymbol{Y}_i)\log[(n_i-\boldsymbol{Y}_i)/(n_i-\widehat{\boldsymbol{Y}}_i(\mathrm{S}))])$$

where $\widehat{\boldsymbol{Y}}_i(S)$ the fitted response for submodel S and $\boldsymbol{Y}_i$ the actual observation

➔ **Correcting for the bias:** $R_{\text{tr}}(\mathrm{S}) + \mathit{Complexity\ Penalty}$

- AIC & BIC are commonly used for model selection for GLMs

Ranking States by SAT Performance



SAT Mean Score by State – Year 1982
790 (South Carolina) – 1088 (Iowa)

- ***Which variables are associated with state average SAT scores?***
- ***After accounting for selection biases, how do the states rank?***
- ***Which states perform best for the amount of money they spend?***

Model Selection Criteria Using R

```
library(CombMSC)
n = nrow(datasat)

## full model
c(Cp(regression.line, S2=summary(regression.line)$sigma^2),
AIC(regression.line, k=2), AIC(regression.line, k=log(n)))
[1]   7.016756 471.698197 486.994381

## reduced model
c(Cp(regression.red, S2=summary(regression.line)$sigma^2),
AIC(regression.red, k=2), AIC(regression.red, k=log(n)))
[1]  29.67045 490.59880 498.24689
```

- Mallow's Cp: $\hat{\sigma}$ = 24.86 is the estimated standard deviation for the full model
  - Use the estimated variance, $\hat{\sigma}$^2, as the S2 parameter value
- BIC: Similar to AIC, but the AIC complexity is further penalized by log(n)
- The values of the three criteria are different and not comparable
- The full model is better according to all three criteria

Summary