As usual, you can do this assignment in a group of maximum three people. Objectives of this assignment:

- To learn about penalized regression splines, and various methods for choosing model complexity.
- To learn how to conduct a simulation experiment.

1. Write a *R* program for fitting penalized regression splines with $p = 3$ (i.e., cubic). You can fix the number of knots as 30, and place them equi-spaced within the domain of the data. Our model assumption is $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{e}$, where $E(\boldsymbol{e}) = 0$ and $Cov(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$.

   (a) Implement the cross-validation and generalized cross-validation methods for choosing the smoothing parameter λ.

   (b) Let $\hat{\boldsymbol{f}}_\lambda = \boldsymbol{H}_\lambda \boldsymbol{y}$ be the fitted values, where $\boldsymbol{H}_\lambda$ is the hat/smoother matrix and $\boldsymbol{y}$ is the noisy data. Define the *effective degrees of freedom* for $\hat{\boldsymbol{f}}_\lambda$ as $\mathrm{tr}(\boldsymbol{H}_\lambda)$. Another method for choosing λ is *Corrected AIC*, for which λ is chosen as the minimizer of

   $$\mathrm{AIC}_C(\lambda) = \log \|\boldsymbol{y} - \hat{\boldsymbol{f}}_\lambda\|^2 + \frac{2\{\mathrm{tr}(\boldsymbol{H}_\lambda) + 1\}}{n - \mathrm{tr}(\boldsymbol{H}_\lambda) - 2}.$$

   Implement this λ-selection method.

   (c) Another method for choosing λ: ideally one would like to choose λ so that the L_2 risk is minimized: $\mathrm{risk}(\lambda) = E\|\boldsymbol{f} - \hat{\boldsymbol{f}}_\lambda\|^2$. Of course this cannot be done in practice, as $\boldsymbol{f}$, the true function, is unknown. One way to overcome this issue is to first construct an unbiased estimator for the risk and then minimize this risk estimator instead. Suppose for the moment that σ^2 is known. Prove:

   $$E\|\boldsymbol{y} - \hat{\boldsymbol{f}}_\lambda\|^2 = \|(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{f}\|^2 + \sigma^2\{\mathrm{tr}(\boldsymbol{H}_\lambda \boldsymbol{H}_\lambda^T) - 2\mathrm{tr}(\boldsymbol{H}_\lambda) + n\},$$

   where n is the number of data points. Use this expression, or otherwise, construct an unbiased estimator for $\mathrm{risk}(\lambda)$ (when σ^2 is known). Implement this method – you will first need to obtain an estimate $\hat{\sigma}^2$ for σ^2, and then treat $\hat{\sigma}^2$ as the true σ^2.

   (d) Conduct a simulation study to compare the above four methods for choosing λ. Use the experimental setting from the paper downloadable from Canvas.

—— End of Assignment 5 ——