A First Simulation Example on Designing and Assessing a Regression Function

- Generate one small (10 observations) training dataset from binormal distribution (representing a response and a predictor pair $(X,Y)'$) with mean vector $\left(0,0\right)'$, unit variance, and $\rho = 0.8$.
- Fit the data to a linear model, calculate the apparent MSE given by $\frac{1}{N}RSS$.
- Plot the linear model, the data, and the best regression function (which is the conditional expectation of a bivariate normal $(X,Y)'$, given by: $\mathrm{E}\left[Y|X=x\right] = \mu_Y + \frac{\sigma_{YX}}{\sigma_X^2}\left(x-\mu_X\right)$) on the same graph.
- Generate a large data set (1000 observations to represent the population) from the same distribution and calculate the true error rate (the MSE or the risk) of your model on this data set; we denote this by $err_{\mathbf{tr}}$. This is the performance conditional on the training set above. Obtain the performance of best regression function as well; denote it by err^*.
- Do Monte-Carlo (MC) simulation by repeating the above for 500 training sets of the same size (10 observations), with the same large testing set (the 1000 observations) to produce the following:
  - a one plot showing the best regression function (in bold), and the 500 linear models (in gray); do not show the datasets. You will see how the model varies with varying the training dataset.
  - 500 MSE values, one for each trained model. Obtain the mean and the variance of them, i.e., $\mathrm{E}_{MC}\left[err_{\mathbf{tr}}\right] = \frac{1}{M}\sum_{m=1}^{M} err_{\mathbf{tr}_m} \cong \mathrm{E}_{\mathbf{tr}}\left[err_{\mathbf{tr}}\right]$ and $\mathrm{Var}_{MC}\left[err_{\mathbf{tr}}\right] = \frac{1}{M-1}\sum_m \left(err_{\mathbf{tr}_m} - \mathrm{E}_{MC}\left[err_{\mathbf{tr}_m}\right]\right)^2 \cong \mathrm{Var}_{\mathbf{tr}}\left[err_{\mathbf{tr}}\right]$, where $M =$ 500 here.
- Repeat the above 10 more times with different training-set sizes, e.g., 20, 40, 80, 100, 200, 300, 400, 500, 700, 1000, to produce one figure having 3 plots vs. the training set size $n_{\mathbf{tr}}$. The three plots are:
  - err^* (of course is a constant with $n_{\mathbf{tr}}$, and obtained by testing the best regression function once on the 1000-observation testing set)
  - $\mathrm{E}_{\mathbf{tr}}\left[err_{\mathbf{tr}}\right]$.
  - $\mathrm{Var}_{\mathbf{tr}}\left[err_{\mathbf{tr}}\right]$.
- What do you observe? Report all of the results, plots, and your comments.