PROBLEMS

Due Friday Feb. 3, 2023 @ 4pm

1. For each of the following pairs of "true" functions $f(\mathbf{x})$ and model functions $\widehat{f}(\mathbf{x};\boldsymbol{\beta})$, determine: (i) whether the model function is linear in the parameters $\boldsymbol{\beta}$; (ii) whether there is underfitting in the model; and (iii) if there is no underfitting, what are the true model parameters $\boldsymbol{\beta}$?

   (a) $f(x) = (1+x)^2$ and $\widehat{f}(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$.

   (b) $f(x) = 1/(2-x) - 2$ and $\widehat{f}(x;a_0,a_1,b_0,b_1) = (a_0 + a_1 x)/(b_0 + b_1 x)$.

   (c) $f(\mathbf{x}) = (x_1 - x_2)^2$ and $\widehat{f}(\mathbf{x};a,b_1,b_2,c_1,c_2) = a + b_1 x_1 + b_2 x_2 + c_1 x_1^2 + c_2 x_2^2$.

   (d) $f(\mathbf{x}) = 3\ln(x_1 x_2/x_3)$ and $\widehat{f}(\mathbf{x};\boldsymbol{\beta}) = \sum_{j=1}^{3} \beta_j \ln(x_j)$

2. A medical researcher is trying to choose between three model orders d. To evaluate the models, she uses 10-fold cross validation, which gave the following results.

Model order	Training $\overline{\text{RSS}}_d$	Test $\overline{\text{RSS}}_d$	Test $\text{stdv}(\text{RSS}_d)$
$d=1$	2.0	2.01	0.25
$d=2$	0.7	0.72	0.06
$d=3$	0.65	0.70	0.07

   Note that the last column is the *unbiased* version of the sample standard deviation, i.e., $\text{stdv}(\text{RSS}_d) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\text{RSS}_{d,k} - \overline{\text{RSS}}_d)^2}$, not the standard error. Which model should be selected based on the "one standard error rule"?

3. Consider a random variable x with $\mathbb{E}\{x\} = \mu$ and $\text{var}\{x\} = v$, and another random variable ϵ with $\mathbb{E}\{\epsilon\} = 0$ and $\text{var}\{\epsilon\} = \sigma^2$, where ϵ and x are independent. Finally, suppose $y = x + \epsilon$. Evaluate the following expectations:

   (a) $\mathbb{E}\{x^2\}$

   (b) $\mathbb{E}\{xy\}$

   (c) $\mathbb{E}\{y^2\}$

   (d) $\mathbb{E}\{y|x\}$

   (e) $\mathbb{E}\{y^2|x\}$

4. Suppose that some training data $\{(x_i, y_i)\}_{i=1}^n$ was generated from a noisy linear model

$$y_i = f(x_i) + \epsilon_i, \quad f(x_i) = \beta x_i.$$

For prediction, we assume the linear model

$$\widehat{y} = \widehat{f}(x; \widehat{\beta}) = \widehat{\beta} x,$$

with $\widehat{\beta}$ computed using a least-squares (LS) fit to the training data. Finally, we wonder how our prediction will fare on test data (x, y) generated via

$$y = f(x) + \epsilon.$$

You can assume that $\{\epsilon_i\}$ and ϵ are mutully independent, zero-mean, σ^2-variance random variables that are independent of $\{x_i\}$ and x.

(a) Derive an expression for the LS $\widehat{\beta}$ in terms of $\{(x_i, y_i)\}_{i=1}^n$ by first writing an expression for RSS and then zeroing its partial derivative with respect to $\widehat{\beta}$.

(b) Write $\widehat{\beta}$ in terms of β and $\{(x_i, \epsilon_i)\}_{i=1}^n$.

(c) What is the bias $\mathbb{E}\{\widehat{y} - y \,|\, x\}$? Is the predictor biased or unbiased? **For this and remaining parts, you can assume that the training $x_i = 1$ for all i**, but please keep the test x generic.

(d) What is the mean-squared error $\mathbb{E}\{(\widehat{y} - y)^2 \,|\, x\}$?

(e) What is the variance of the prediction, $\text{var}\{\widehat{y} \,|\, x\}$?

(f) Compute the irreducible error by subtracting the prediction variance and the squared bias from the MSE. Does it turn out as expected?