STA 314, Review problems 2 (Lecture 3)

Multiple choice questions can have any number of correct answers (including zero).

Q1 *Assume that* `D` *is a data frame that contains 4 columns with names* `Y, X, V, W`*. For each of the following specifications, write down the regression function that corresponds to the* `lm` *call in* ***R****.*

(a) `lm( Y ~ . , data = D)`

(b) `lm( Y ~ . - V - W, data = D)`

(c) `lm( Y ~ X*W + V, data = D)`

(d) `lm( Y ~ I(cos(X)) + I(V^3), data = D)`

Q2 *Assume that* `D` *is a data frame that contains 4 columns with names* `Y, X1, X2, X3`*. For each of the following regression functions, decide if the can be formulated as a linear model, i.e. in the form*

$$f(x) = b_1 + b_2 g_1(x) + ... + b_{L+1} g_L(x).$$

If yes, write down L *and* $g_1, ..., g_L$ *and the* `lm` *call in* `R` *that you would use to fit those functions.*

(a) $f(x) = b_1 + b_2 x_1 x_2$

(b) $f(x) = b_1 + b_2 x_1 x_2 x_3 + b_3 x_3$

(c) $f(x) = b_1 + b_2 x_1 - b_3 e^{x_2}$

(d) $f(x) = b_1 + b_2 x_1 - b_3 e^{b_4 x_2}$

Q3 Using the notation from lectures, show that the Hessian matrix (the matrix of mixed second order partial derivatives) of the function

$$b \mapsto \|\mathbf{X}b - Y\|_2^2$$

is given by $\mathbf{X}^\top \mathbf{X}$ and show that this matrix is strictly positive definite if $\mathbf{X}$ has full rank.

Q4 Consider the following two linear models

$$\begin{aligned} &\text{model 1:} \quad f(x) = b_1 + b_2 x_1 + b_3 \sin(x) \\ &\text{model 2:} \quad f(x) = b_1 + b_2 x_1 + b_3 \sin(x) + b_4 x_1 x_2 \end{aligned}$$

Will one of the two models always have a smaller training error? Justify your answer.

Q5 Consider a simple linear regression with one-dimensional predictor x. Write down an example of a data set with $n = 5$ points where $R^2 = 1$ and an example where $R^2 = 0$.

Q6 Assume that you have a data set (x_i, y_i) with $x_i \in R$, $n = 10$ points where all x_i are different and not all points (x_i, y_i) lie on a line. Which of the following statements are true?

(a) There exists a k such that k-nn regression has smaller training error than simple linear regression.

(b) Simple linear regression will have a larger training error than k-nn regression no matter what k is.

(c) Simple linear regression and k-nn regression will have the same training error no matter how k is chosen.

Q7 Which of the following are true?

(a) Adding more predictors to a linear model will never increase the training error.

(b) Adding more predictors to a linear model will never increase the test error.

(c) Adding more predictors to a linear model will typically decrease the training error.

(d) Adding an interaction to a linear model will never increase the training error.

(e) Adding an interaction to a linear model never have an effect on the training error.