# Homework 5

1. Say a linear regression model is used to estimate the profit that our company makes on contracts. Our only predictor variable is the department handling the contract, which takes values A, B, C, D, and E. We take department A to be the baseline category, so that the department variable is recoded into the variables $1_{(\text{dept}=B)}$, $1_{(\text{dept}=C)}$, $1_{(\text{dept}=D)}$, $1_{(\text{dept}=E)}$. The intercept estimate is $\hat{\beta}_0 = 1.21$, and the estimated regression coefficients for the four department indicator variables are $\hat{\beta}_1 = -0.0025$, $\hat{\beta}_2 = 1.02$, $\hat{\beta}_3 = 0.317$, and $\hat{\beta}_4 = 0.0074$, respectively.

   (a) **Interpret the value of $\hat{\beta}_4$.**

   (b) **Predict the profit that the company will make for a contract handled by department C.** I'm looking for a point prediction here (single number); show your work.

2. A dataset with 3 variables has been split into test and non-test data, and the non-test data consist of 5 observations:

| | | | | | |
|---|---|---|---|---|---|
| $v_1$ | -2.0 | 0.8 | 0.3 | -0.8 | 0.0 |
| $v_2$ | 1.0 | 1.4 | -0.9 | -1.3 | -0.3 |
| $v_3$ | 1 | 0 | 1 | 0 | 1 |

   **Use leave-one-out cross-validation on the non-test data to decide whether to use 1-nearest-neighbors or 3-nearest neighbors, for the purpose of predicting $v_3$ based on the values of $v_1$ and $v_2$.** Complete this question with only your calculator (i.e., no statistical software), since you need to be able to complete questions like this one in a testing setting. Show all work. **Report what model you selected and why.**

3. We want to fit a spline model to the following data. Specifically, you should use a cubic spline.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| y | -0.7 | -0.1 | 0.9 | 1.7 | 3.1 | 4.2 | 4.8 | 4.9 | 5.0 |
| x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

   (a) **Say that we want to use two knots; where should we locate them?** Use the method we discussed in class.

(b) We know that a spline model can be rewritten in the form of a multiple linear regression model, with a different set of predictors. **Calculate the values of those new predictors, for the first and last observation in the dataset.** Show your work.

4. **What would be the form of a quartic spline (i.e., the degree-4 analogue of a cubic spline? Write out the model. Also explain what smoothness properties this model has at the knots (is it continuous, and if it has continuous derivatives state how many). Why do you think that this model is not used very often, relative to cubic spline models?**

5. Say we have a cubic spline model, and put two knots at the same location (e.g., $\xi_1 = \xi_2$). **Is the resulting cubic spline model identifiable? Give a clear justification why or why not.**

6. Consider a training dataset with only one continuous predictor variable $X$ and a continuous outcome variable $Y$, and the three observations given below. Consider a regression tree with only two leaf nodes, where the interior node corresponds to the split $X < 0.5$.

| | | | |
|---|---|---|---|
| y | -0.7 | -0.1 | 0.9 |
| x | 0.2 | -0.9 | 1.1 |

**What is the per-node increase in RSS obtained by pruning the tree at the root node (in this example, the only interior node)?** Show your work.