Regression Homework (HW 4)

1. Consider a multiple linear regression model for predicting the level of monetary contribution of members of a nonprofit organization for military veterans. The predictors include the gender of the individual, and an indicator of whether or not the person has pets. **Can the standard multiple linear regression model capture the situation where a woman is more likely to contribute if she has pets than if she doesn't, while a man is less likely to contribute if he has pets than if he doesn't? Clearly explain why or why not.**

2. One method that we have for choosing a subset of the predictors in a linear regression model is to cut the data into training and validation data sets, fit each model on the training data and choose the model that has the smallest estimation error (best accuracy) on the validation data. As a measure of estimation error we have used the root mean squared error (RMSE). Using the RMSE implicitly assumes that underestimation of Y by, e.g., k units has the same cost as overestimation of Y by k units. By "cost" here we are referring to the price of inaccurate prediction in terms of financial cost or cost in human suffering, for instance. An example is if we overestimate the salary that a baseball player needs to be offered in order to have a reasonable chance of convincing him to move between teams. The difference between our estimate and what we could have offered is the financial cost of our inaccurate estimation. Consider the general situation where overestimation of the outcome Y is more costly than underestimation, or vice versa.

Suggest an error measure that is more reasonable than RMSE in this situation. Give an example of a situation where overestimation is more costly than underestimation, or vice versa, and explain why.

3. Consider a linear regression model with 4 predictors, two of which are categorical variables with 3 possible values. We convert each of the categorical variables into 2 dummy binary variables, so that one can consider the model to have $p = 6$ continuous and binary predictors. Then we want to do model selection (choose a subset of the predictors). But for each of the categorical variables, it is reasonable to consider including or excluding all of the dummy variables as a group instead of individually, i.e. restricting our model selection to only consider models that either include both binary variables associated with the categorical predictor, or exclude both binary variables. **If we make this restriction, how many possible linear regression models are under consideration?**

4.

(a) **How many parameters are there in a (standard) multiple linear regression model with** 4 **continuous predictors? Include all unknown quantities that**

characterize the model and have to be estimated.

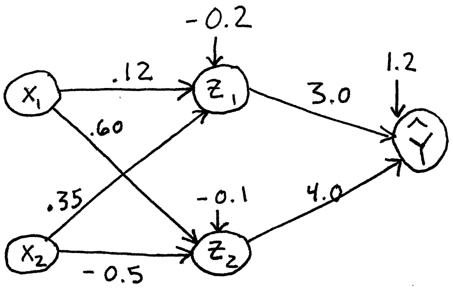
(b) **How about if we include not just the linear terms $\beta_j X_{ij}$, but also all pairwise interaction terms, i.e. terms of the form $\beta_{j,k} X_{ij} X_{ik}$ where $j, k \in \{1, \ldots, p\}$ and $j \neq k$?** Interaction terms like these are a way of capturing interactions between predictors in how they affect the outcome; as we discussed in class, the standard multiple linear regression model cannot capture such interactions. Do not count $\beta_{j,k} X_{ij} X_{ik}$ and $\beta_{k,j} X_{ik} X_{ij}$ as different terms, since they are redundant.

(c) **How about if we include all pairwise *and* three-way *and* four-way etc. interactions, i.e. all K-way interactions for $K \in \{2, \ldots, p\}$?** A three-way interaction, for example, refers to terms of the form $\beta_{j,k,\ell} X_{ij} X_{ik} X_{i\ell}$ for $j, k, \ell \in \{1, \ldots, p\}$ and $j \neq k \neq \ell \neq j$. Again, for any set of k variables only include a single k-way interaction term. **Comment on the practical implications of including all these terms.**

5. (James, Witten, Hastie, and Tibshirani 2013). **Draw an example (of your own invention) of a partition of a two-dimensional predictor space that could result from a regression tree as we have described them in class.** Your example should include at least 6 regions. Draw a decision tree corresponding to this partition. Label all aspects of your figures, including the regions $R_1, R_2, \ldots$, the cut points $t_1, t_2, \ldots$, and so forth.

6. A bank has done an assessment to evaluate the long-term cost c of a defaulted loan, and the long-term benefit b to the bank of a non-defaulted loan customer (b and c are both in units of dollars). Assume that the cost/benefit of turning down a loan applicant is essentially zero. To decide whether to grant a loan, the bank uses a statistical method to estimate the probability that a customer will default on a loan, given some available financial and demographic variables. The bank will grant the loan if the estimated chance of default is below some cutoff. **What cutoff would you recommend that the bank use? Explain why.**

7. Below is a trained neural network for predicting a continuous outcome Y given two continuous predictors X_1, X_2. The activation functions are $\sigma(v) = e^v/(1+e^v)$ for the hidden layer and $\tilde{\sigma}(v) = v$ for the output layer. Predict the value of Y if $X_1 = 5$ and $X_2 = 10$. Show your work.


-0.2
1.2
X1
.12
Z1
3.0
.60
Ŷ
.35
-0.1
4.0
X2
Z2
-0.5