Q8. Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. K = 1) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

When K=1 in a KNN, the training error rate is 0 because the method at that K level the method is overly flexible and thus won’t give a good prediction. Also, the fact that the average error rate of both the test and the training data set is 18% when the training error rate is 0 implies that the testing error rate when KNN=1 is 36% [(t+0)/2=18% t=36%], which is higher than that in the logistic regression. Therefore, the logistic regression method is more reliable in classifying new observations in this case.

**Q2.** Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide nn and pp.

1. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

This is a **regression**, because we’re looking for quantitative response in the existing trend. It is also an **inference** because we’re trying to understand the correlation between the factors and no prediction is needed. **N=500** since we have a pool of 500 firms and **p=3** (profit, number of employees& industry

1. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

This is a **classification**, because we’re looking for a qualitative response. It is also a **prediction** because we’re trying to predict if the product will be a success or a failure. **N=20** (data on 20 similar products) and **p=13** (price charged for the product, marketing budget, competition price, and ten other variables)

1. We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

This is a **regression,** because we’re looking for quantitative response in the existing trend. It is also a **prediction** because we’re trying to predict the %change. **N=52** (52weeks in 2012) and **p=3** (the % change in the US market, the % change in the British market, and the % change in the German market.)