Bonus Questions

Question 4:

a: On average, we will use 10% ($\frac{1}{10}$) of available observations to make the prediction. This is because X is uniformly distributed, so in any given range of values such as the range [0.55,0.65], that makes up 10% of the values, we will most likely see 10% of the available observations.

b: On average, we will use 1% ($\frac{1}{100}$) of available observations to make the prediction. The logic follows from part a, because X1 and X2 are uniformly distributed, but now, on average, we will only use 10% of X1 observations and 10% of X2 observations. Because we are looking for observations that are within 10% of X1 and 10% of X2, we can multiply X1 and X2 to find the probability of a point being in both of these ranges. $0.1*0.1 = 0.01 = $1%.

c: This follows the same logic as part b, but in part b we used 0.1^2, and here we will use 0.1^{100} (because p =100, and we want observations within 10% of all 100 of these features). So our likliehood is 0.1^{100} or $\frac{1}{10^{100}}$.

d: As we can see in part A-C, as p increases, the amount of observations used decreases exponentially, so that when p gets very large, there are an extremely small fraction of observations that are close. This will lead to nearest neighbors that are not actually near each other because if we define being near eachother as being within 10% of eachother in each dimension, we will have hardly any neighbors when we look at all of the dimensions together.

e: For p=1, the hypercube would need to cover 10% of the length to capture 10% of the data points, so its length would be 0.1. For p = 2, the hypercube (square) area would need to cover 10% of the area, so $length * width = 0.1 = length^2$, so the side lengths would need to be $0.1^{\frac{1}{2}} = 0.316$. For a p = 100 hypercube, its volume would be found by multiplying all 100 sides, and setting this equal to .1, to find the side lengths when volume covers 10% of observation. In this case, $0.1 = length^{100}$. So $0.1^{1/100} = 0.977$ = length of each side. As we can see, the length needed to cover 10% of the data increases as p increases, showing that the "nearest neighbors" at high p values are not actually very near to each other, when compared to low p values.

Question 12:

a:

p = $\frac{exp(B_0+B_1x)}{1+exp(B_0+B_1x)}$

odds = $\frac{p}{1-p}$

1 - p = $1 - \frac{exp(B_0+B_1x)}{1+exp(B_0+B_1x)}$

$\frac{p}{1-p}$ simplifies to $exp(B_0 + B_1x)$

Answer:

$ln(exp(B_0 + B_1x)) = B_0 + B_1x$

b:

p = $\frac{exp(\alpha_{orange0}+\alpha_{orange1}x)}{exp(\alpha_{orange0}+\alpha_{orange1}x)+exp(\alpha_{apple0}+\alpha_{apple1}x)}$

1 - p = $1 - \frac{exp(\alpha_{orange0}+\alpha_{orange1}x)}{exp(\alpha_{orange0}+\alpha_{orange1}x)+exp(\alpha_{apple0}+\alpha_{apple1}x)}$

$odds = \frac{p}{1-p} = \frac{exp(\alpha_{orange0}+\alpha_{orange1}x)}{exp(\alpha_{apple0})+\alpha_{apple1}x}$

Answer:

ln(odds) = $\frac{\alpha_{orange0}+\alpha_{orange1}x}{\alpha_{apple0}+\alpha_{applee1}x}$

c:

We do not know the exact coefficient estimates but we can set

$\frac{\alpha_{orange0}}{\alpha_{apple0}} = B_0 =$ 2

and

$\frac{\alpha_{orange1}}{\alpha_{apple1}} = B_1 =$ -1

d:

$B_0 = \frac{\alpha_{orange0}}{\alpha_{apple0}} = \frac{1.2}{3} =$ 0.4

$B_1 = \frac{\alpha_{orange1}}{\alpha_{apple1}} = \frac{-2}{.6} =$ -3.333

e:

I expect them to agree 100% of the time. Softmax is used for multiclass data, whereas logistic regression is used for binary data Because this example has only two classes (it is binary), the softmax and logistic approaches should return the same predicted labels (this is why we can even predict one model's coefficients from the other model as seen above).