Extension Questions! **Week 5**

Q1 By defining Y as the sum of outcomes in n iid Bernoulli trials, we can say that $Z = Y/n$ is the sample proportion an as estimate ($\hat{p}$) for the proportion in the overall population. Given this knowledge of Z ($\hat{p}$) derive the following distribution for two proportion ~~under the a rest together that they are from the same population~~

$$\hat{p}_1 - \hat{p}_2 \overset{a}{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n}}\right)$$

Hint: First consider one proportion and think about how you can represent n Bernoulli trials with the normal distribution

Hint: You will need to use the central limit theorem to show this and it is quite satisfying to do so.

→ Why might we prefer to use $\hat{p}$ over the binomial distribution? Surely the binomial would be more accurate.

Q2 We often say that the linear probability model (LPM) is a bad model since the probabilities predicted can fall outside of the interpretable [0,1] range. It also gives us homogeneous treatment effects for varying values of regressors which may not be desired. However... under certain conditions an LPM's prediction will always fall in the [0,1] range. What are these conditions?

Hint: You may need to do some research if you don't know it

Q3 It is not possible to find a closed form solution for $\hat{\beta}$ in the logistic regression model. However, it is possible to derive a condition which $\hat{\beta}$ should be chosen to satisfy. (In practice, coefficients are determined by numerical methods so the condition may not perfectly hold but anyway...)

Take a logistic regression of the form below:

$$p(Y_i = 1 \mid x_i) = \Lambda(\beta x_i) = \frac{e^{\beta x_i}}{1 + e^{\beta x_i}}, \quad (\alpha \text{ is not included})$$

a) By constructing the likelihood function, considering the log-likelihood function, and then maximising this w.r.t. β, show that the following condition holds for $\hat{\beta}$. <u>Note:</u> this is maximum likelihood estimation (MLE)

$$\text{for } \hat{\beta}: \quad \sum_{i=1}^{n} \varepsilon_i x_i = 0, \quad \text{note: } \varepsilon_i = y_i - \Lambda(\beta x_i)$$

b) Now consider the more general case of:

$$p(Y_i = 1 \mid x_i) = \Lambda(\alpha + \beta x_i) = \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}$$

Derive the first order conditions that must be satisfied with $\hat{\alpha}, \hat{\beta}$. Note that you cannot solve for $\hat{\alpha}$ and $\hat{\beta}$, $\rightarrow$ if you've got to this point then I'm sure you know that anyway...