Text as Data: Homework 1

Question 1: Using Python

a) Install `Python`, `Rstudio`, and `R markdown`

b) Using `Python` write to a `.txt` file:

```
- Hello World
```

With a for loop, write the numbers 1-100 to the same text file

c) Close the .txt file, turn it in with your homework

Question 2: Properties of Random Variables

a) Suppose X is a random variable, with $E[X] = \mu$ and $\text{var}(X) = \sigma^2$. Show that $c = E[X]$ minimizes

$$E[(X - c)^2]$$

Why does this suggest $E[X]$ is a "good" guess for the value of X?

b) Suppose Y and Z are random variables, with joint density $f(y, z)$.

i) How do we obtain the marginal distribution of Y, $f_Y(y)$ (should be an expression involving an integral)

ii) How do we obtain the marginal distribution of Z, $f_Z(z)$ (should be an expression involving an integral)

iii) Show that if Y and Z are independent, $E[YZ] = E[Y]E[Z]$

Question 3: Finding Critical Values for a Function

Suppose we have a function $f : \Re \rightarrow \Re$, $f(x) = \sin(x)$.

a) Using `R` plot $\sin(x)$ for $x \in [-2\pi, 2\pi]$ (here π is the mathematical constant).

b) What is $f^{'}(x)$ (first derivative at x)? Using `R` plot it over $[-2\pi, 2\pi]$

c) What is $f^{''}(x)$ (second derivative at x)? Using `R` plot it over $[-2\pi, 2\pi]$

We say that x^* is a critical value for a function if $f^{'}(x^*) = 0$. We can find x^* algebraically. Or, we can use a computational approach.
We discussed the Newton-Raphson approach in class on Thursday. We're going to write our own implementation of the algorithm in `R` and apply it to find the critical values of $f(x) = \sin(x)$

d) Suppose that we have current guess for the root x_t. Then the updated guess, x_{t+1} is given by

$$x_{t+1} \quad = \quad x_t - \frac{f^{'}(x_t)}{f^{''}(x_t)}$$

where $f^{'}(x_t)$ is the first derivative evaluated at x_t and $f^{''}(x_t)$ is the second derivative evaluated at x_t.
Write a function in `R` that provides the update step for some value x_t if $f(x) = \sin(x)$.

e) The Newton-Raphson algorithm continues to update until the size of the update step drops below a threshold. Using the `while` command in R, write a loop that continues updating until the change,($|x_{t+1} - x_t|$) drops below 10^{-5}.

f) Place the while loop in a function that returns the converged value x_{final}

g) Use your function with initial guesses $-2, -1, 1, 2$. What values do you obtain? Now examine the behavior close to 0. Why is it so unstable?

Table 1: Pseudo Code for Newton Raphson (To assist in developing your function)

- x_0 = initial guess
- Do while change > tolerance:

  $x_{t+1} = x_t - \frac{f^{'}(x_t)}{f^{''}(x_t)}$

  change $= |x_{t+1} - x_t|$
- return x_{t+1}

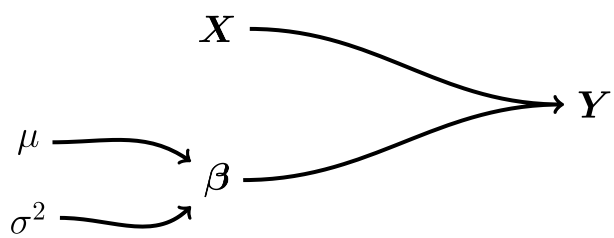
Problem 3: Probit Regression with a Prior

Suppose that we assume the following data generation process

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= \Phi(\boldsymbol{X}_i\boldsymbol{\beta}) \\ \beta_j &\sim \text{Normal}(\mu, \sigma_j^2) \end{aligned}$$

with $\boldsymbol{X}_i = (1, x_i)$ for all i ($i = 1, \ldots, N$), $\boldsymbol{\beta} = (\beta_1, \beta_2)$, and $\Phi(\cdot)$ is the cumulative normal distribution function.

We might equivalently write a directed acyclic graph as,



This is very similar to the model described in class, but now we have added a *prior* on $\boldsymbol{\beta}$. This slightly alters the objective function:

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) &\propto p(\boldsymbol{\beta}|\mu, \sigma^2) \times p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}) \qquad (1) \\ &\propto \prod_{j=1}^{2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\beta_j - \mu)^2}{2\sigma^2}\right) \times \prod_{i=1}^{N} \Phi(\boldsymbol{X}_i\boldsymbol{\beta})^{Y_i}(1 - \Phi(\boldsymbol{X}_i\boldsymbol{\beta})^{1-Y_i} \end{aligned}$$

In this problem, we will examine how the prior on $\boldsymbol{\beta}$, and in particular the values we set for μ and σ^2, alters our inferences about $\boldsymbol{\beta}$.

a) Analytically, write out the $\log(p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}))$.

b) In `R` create a function for the log of Equation 1.

c) Using the synthetic data and the optim guide from class, use `optim` to find $\widehat{\boldsymbol{\beta}}$ with $\mu = 0$ and $\sigma^2 = 1000$

d) Set $\mu = 1$ and then vary σ^2. Using a `for` loop, store estimates of how β_2 changes as you vary σ^2 from 10 to 0.01. Plot β_2 against σ^2 and describe what happens as σ^2 varies.