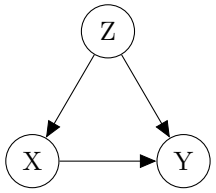
COMP3224: Homework 3

1 Exercises

1. Show how the existence of a confounder introduces bias in regression coefficients. An example model is as follows:

```
def confounder_data(n):
    z = 1.0+np.random.randn(n, 1)
    nx = np.random.randn(n, 1)
    x = 3.0*z + 3.0 + nx
    ny = np.random.randn(n, 1)
    y = 2.5*x - 2.0*z + ny + 2.0
    return z, x, y
```



You may use a library such as `scikit-learn` to call a regressor. Choose a set of features as input and regress the output variable Y on the features. For example, if you want to predict $Y = w_0 + w_x X + w_z Z$, the feature set is (X, Z) and you can call a library function thus:

```
from sklearn import linear_model
linreg = linear_model.LinearRegression()
(zs, xs, ys) = confounder_data(200) # 200 points simulated
features = np.concatenate([xs, zs], 1)
linreg.fit(features, ys)
```

You should choose your own set of parameters to replace the values that I've chosen. Show that all the conditions necessary for backdoor adjustment are satisfied and consequently recover the strength of the causal connection between $X \to Y$.

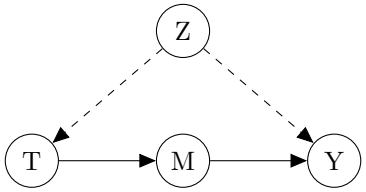
2. In the above function, make 2 changes to the function that generates confounded data.

   (a) replace the Gaussian random variable Z by one drawn from a binomial distribution

(b) introduce an additional, interaction term of the form $w_{xz}ZX$ on the right hand side of Y so that $Y = w_0 + w_x X + w_z Z + w_{xz} ZX$.

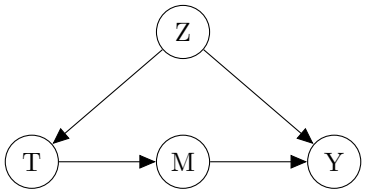
Show how you can recover the strength of the causal connection by using the backdoor adjustment. Interpret the extra term in terms of the regression coefficients. Demonstrate your results graphically by showing the nature of the fit to data from the function.

3. Create a similar function (you may choose binary-valued or continuous random variables) for the SCM with a mediator variable M as shown:



The dotted edges indicate that the variable Z is hidden. Explain that the conditions for the front-door adjustment are satisfied, so the relevant causal estimand, the causal strength of $T \to Y$ can be obtained from observational data. Estimate this quantity.

4. Using the same function as in the previous question but now assume that the variable Z is observed.



Show that the back-door adjustment can now be performed. Compare the causal coefficient estimated by the backdoor adjustment to that from the previous, front-door estimation. By varying the strength of the additive noise variables $\epsilon_Z, \epsilon_T, \epsilon_M$ and ϵ_Y in $T = f_T(Z, \epsilon_T)$ etc., evaluate under what conditions which method provides more reliable estimates.

5. Define a simple function representing $X \to Y$ where the exogenous noise term ϵ_Y is either (1) a uniform distribution or (2) a Gaussian

distribution. Fit a regression function $\widehat{y}(x)$ of Y on X to minimise the sum of the squares of the residuals $r_{\widehat{y}(x)} := (\widehat{y}(x_n) - y_n)$ for the data pairs (x_n, y_n). For the same data, now fit a regression $\widehat{x}(y)$ to minimise the sum of squares of the residuals $r_{\widehat{x}(y)} := (\widehat{x}(y_n) - x_n)$. Draw histograms of $r_{\widehat{y}(x)}$ and $r_{\widehat{x}(y)}$. Comment on any differences you notice between the histograms for the two choices of noise variable ϵ_Y. You should refer to the suggested reading[1] to inform your commentary.

2 Submission details

The report should not be more than 5-6 pages long. Data from your simulations should be presented as evidence for the claims that you make. The claims should be backed up by concepts used in causal inference. You may submit your code/notebooks in a zip file if you wish. Marks will be given for clearly presenting the ideas, from motivating the task to evaluating the validity of the methods, in the light of what has been covered in the lectures and in the readings.

20 Marks.

[1] Chapter 4 *et al. Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press, 2017, link on the notes page.