Alif Ahsanil Satria Follow
Sep 26, 2021 · 6 min read · Listen

The Virtue of Causal Inference — Part 1 : An Introduction

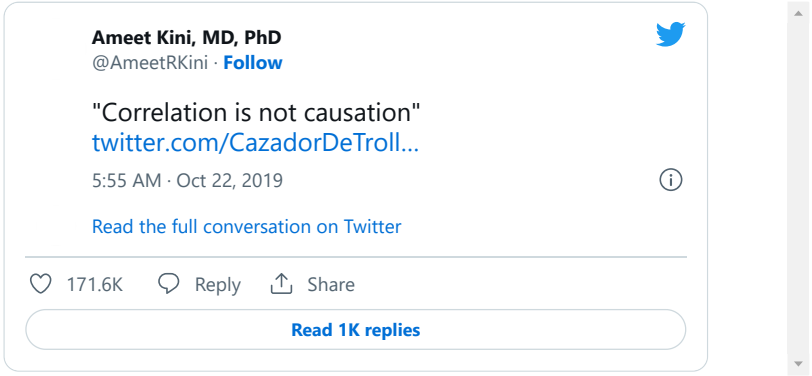
Definition of Causal Effect

As a human being, we are already familiar with cause-and-effect concept in daily basis. As an example, if you eat too much, then you will gain weight fast. Another example is if you are too egocentric, then no one want to become your friend. Intuitively, we reason about causal effect as comparing the outcome when an action A is taken vs the outcome when the action A is withheld. If the two outcomes different, we say that the action A has a causal effect on the outcome. Otherwise, it doesn't. Usually, scientists refer action A as an intervention, an exposure, or a treatment.

Next, I will introduce you into two terms: individual causal effect and average causal effect. Individual causal effect means the comparison of outcome when one doing A vs not doing A. Average causal effect means that the average outcome when sample of our interest doing A vs not doing A. Since it's impossible to track causal effect at individual-level and also the point of causal inference is to infer causal effect at population scale from the sample, we refer 'causal effect' as average causal effect.

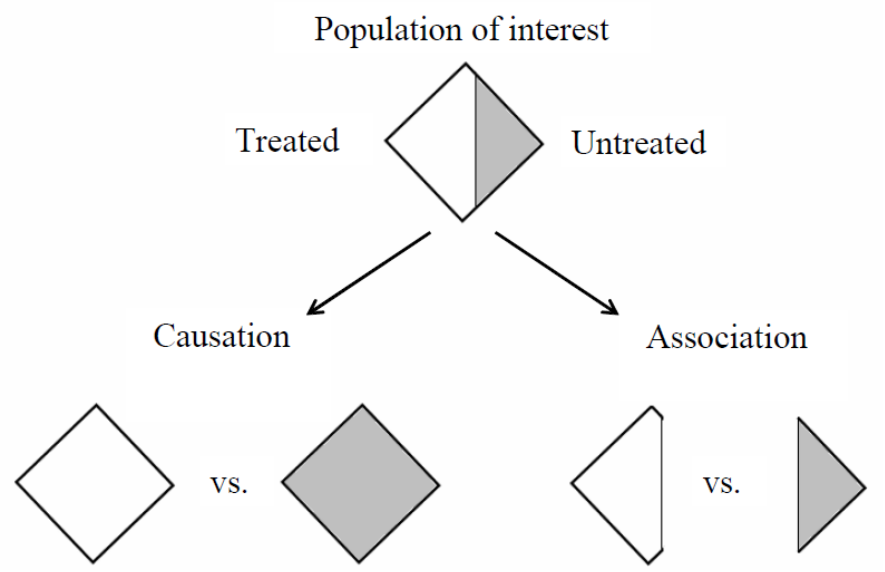
Causation vs Association

Also, as we grow older, we are familiar with term causation and correlation/association, even the differences between them. For example, consider this simple demonstration:



Intuitively, you can see that the train doesn't stop or move because of the man pushing or pulling. Rather, it's merely an association, not causation.

So, lets move forward to more formal definition of causation vs association. Consider this illustration:



causation-association differences

The definition of causation implies a contrast between the whole white diamond (all individuals treated) and the whole grey diamond (all individuals untreated), whereas association implies a contrast between the white (treated) and the grey (untreated) areas of the original diamond. Inferences about causation are concerned with what if questions, such as “what would be the outcome if everybody had been treated?” and “what would be the outcome if everybody had been untreated?”, where as inference about association are concerned with “what is the outcome in the treated?” and “what is the outcome in the untreated?”. The point is that in causation, we use the same group of people (in our population of interest) under two different circumstances (treated vs untreated), whereas in association, we use two different group of people, splitting them to treated and untreated.

In real-world applications, we can never observe the same data with and without treatment. This is the **fundamental problem of causal inference**, so we use association to measure causation. For example, in covid-19 vaccine trials, the experimenter randomly split volunteers into two group: the treatment and placebo/no treatment.

(More Explanation) Why Just Finding Correlation Isn’t Enough?

Consider this simple example

```
data.head()
```

	Location	Loc	Population	MedianAgeMarriage	Marriage	Marriage SE	Divorce	Divorce SE	WaffleHouses
0	Alabama	AL	4.78	25.3	20.2	1.27	12.7	0.79	128
1	Alaska	AK	0.71	25.2	26.0	2.93	12.5	2.05	0
2	Arizona	AZ	6.33	25.8	20.3	0.98	10.8	0.74	18
3	Arkansas	AR	2.92	24.3	26.4	1.70	13.5	1.22	41
4	California	CA	37.25	26.8	19.1	0.39	8.0	0.24	0

Waffle houses data across USA states

Take a look on marriage rate and divorce rate column

```
data[["Location", "marriage_rate", "divorce_rate"]].head(10)
```

	Location	marriage_rate	divorce_rate
0	Alabama	20.2	12.7
1	Alaska	26.0	12.5
2	Arizona	20.3	10.8
3	Arkansas	26.4	13.5
4	California	19.1	8.0
5	Colorado	23.5	11.6
6	Connecticut	17.1	6.7
7	Delaware	23.1	8.9
8	District of Columbia	17.7	6.3
9	Florida	17.0	8.5

Marriage Rate (%) and Divorce Rate (%) data for across USA states

Now, use simple regression to find relationship strength between marriage_rate and divorce_rate

```
import statsmodels.formula.api as smf

model1 = smf.ols("divorce_rate ~ marriage_rate", data=data).fit()
model1.summary()
```

OLS Regression Results

Dep. Variable:	divorce_rate	**R-squared:**	0.140
Model:	OLS	**Adj. R-squared:**	0.122
Method:	Least Squares	**F-statistic:**	7.793
Date:	Sun, 26 Sep 2021	**Prob (F-statistic):**	0.00751
Time:	07:10:43	**Log-Likelihood:**	-96.645
No. Observations:	50	**AIC:**	197.3
Df Residuals:	48	**BIC:**	201.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.0840	1.313	4.632	0.000	3.443	8.725
marriage_rate	0.1792	0.064	2.792	0.008	0.050	0.308

Omnibus:	0.332	**Durbin-Watson:**	1.728
Prob(Omnibus):	0.847	**Jarque-Bera (JB):**	0.417
Skew:	0.179	**Prob(JB):**	0.812
Kurtosis:	2.731	**Cond. No.**	112.

regression between marriage_rate (x) and divorce_rate (y)result

Take a look on marriage_rate coefficient (0.1792) and p-value (0.008). If we take p-value < 0.05 as threshold for statistically significance decision, this means that there is some (positive) relationship between marriage_rate and divorce_rate.

But, take a time to think. Does higher marriage_rate imply higher divorce_rate? Although prior divorce, someone must be in married status, still not entirely sure that high marriage rate must be correlated with high divorce. Also, give a thought that high marriage rate indicating high cultural valuation of marriage and therefore being associated with low divorce rate. So, what to do next?

Take a look again on the data and think how about MedianAgeMarriage (age at marriage). MedianAgeMarriage is also a good predictor of divorce rate because higher age at marriage associated with less divorce. Give a thought that younger people change faster than older people and are therefore more likely to grow incompatible with a partner. Then, MedianAgeMarriage can have an indirect effect by influencing marriage_rate. If people get married earlier, then the marriage_rate may rise, because there are more young people. Thus, MedianAgeMarriage associated positively with marriage_rate.

So, which is it? is there a direct effect of marriage_rate to divorce_rate or rather just MedianAgeMarriage associated with both divorce_rate and marriage_rate, creating a spurious relationship between marriage_rate and divorce_rate?

To tackle this problem, we need multivariable regression model. This can address a useful descriptive question: Is there any additional value in knowing a variable, once already know all of the other predictor variables. So, once you fit a multiple regression with marriage_rate and MedianAgeMarriage as predictor, the model addresses 2 questions:

1. After know marriage_rate , what additional value is there in also knowing MedianAgeMarriage (represented by regression coefficient MedianAgeMarriage)
2. After know MedianAgeMarriage, what additional value is there in also knowing marriage_rate (represented by regression coefficient marriage_rate)

Let's fit the multiple regression model

```
import statsmodels.formula.api as smf

model2 = smf.ols("divorce_rate ~ marriage_rate + MedianAgeMarriage", data=data).fit()
model2.summary()
```

OLS Regression Results

Dep. Variable:	divorce_rate	**R-squared:**	0.363
Model:	OLS	**Adj. R-squared:**	0.336
Method:	Least Squares	**F-statistic:**	13.42
Date:	Sun, 26 Sep 2021	**Prob (F-statistic):**	2.46e-05
Time:	10:07:35	**Log-Likelihood:**	-89.114
No. Observations:	50	**AIC:**	184.2
Df Residuals:	47	**BIC:**	190.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	36.8766	7.661	4.814	0.000	21.465	52.289
marriage_rate	-0.0569	0.081	-0.706	0.484	-0.219	0.105
MedianAgeMarriage	-0.9996	0.246	-4.065	0.000	-1.494	-0.505

Omnibus:	2.545	**Durbin-Watson:**	1.903
Prob(Omnibus):	0.280	**Jarque-Bera (JB):**	1.595
Skew:	-0.301	**Prob(JB):**	0.450
Kurtosis:	3.635	**Cond. No.**	1.20e+03

Multiple Regression Model with marriage_rate and MedianAgeMarriage as predictor

Take a look on:

1. marriage_rate coefficient (-0.0569) and corresponding p-value (0.484).
2. MedianAgeMarriage coefficient (-0.9996) and corresponding p-value (0.000….)

This implies that once we know MedianAgeMarriage, there is little or no additional predictive power in also knowing marriage_rate in that State. Compare this result to model1 result (which marriage_rate coefficient is 0.1792). So, the answer is just MedianAgeMarriage creating a spurious relationship between marriage_rate and divorce_rate. We refer MedianAgeMarriage to as confounding variable because it distort our true relationship estimation between marriage_rate and divorce_rate.

To conclude, based on this demonstration, just finding correlation isn't enough to infer causal relationship between predictor and outcome. So, what to do next? before explain causal effect concept further and how its characteristics is different with association, let me explain you two kind of ways we can get data (thus, may want estimate the causal relationship on that data): randomized experiments and observational studies.

Randomized Experiments vs Observational Studies

Randomized Experiments is when someone split the object of experiment into two groups: treatment and control (nontreatment) randomly and see for differences in outcome interest between these two groups, so the effect of treatment to the outcome will be known. The easiest example for this is A/B testing (often used in startups) and vaccine trials (covid-19 vaccine is one of the examples). This allow greatest reliability and validity of statistical estimates of treatment effect. If conducted correctly, association got from this way is also causation.

The problem is that is all causal relationship problem can be solved with randomized experiment? no. Many scientific studies are not experiments. Much human knowledge is derived from observational studies. Think evolution, tectonic plates, global warming, or astrophysics, even how human learned that hot coffee may cause burns. Also, give a thought how human learned that smoking is dangerous for health. It isn't impossible to assign randomly between group of smoking and nonsmoking people. Thus, we use observational studies.

Observational studies is when someone learn the treatment effect based on what happened in the past, thus we refer to as observation. The example of observational studies is divorce rate case as explained above. Observational studies, likewise randomized experiment, also induces many problems. One of the problem is confounding variable bias, as explained above.

References:

[1] https://www.goodreads.com/book/show/9416017-causal-inference

[2] https://www.goodreads.com/book/show/26619686-statistical-rethinking