Ch 3: Regression, Observations and Interventions

Prashant Shekhar, PhD
Assistant Professor of Data Science
Department of Mathematics
Embry-Riddle Aeronautical University, FL, USA
Email: shekharp@erau.edu

Outline

1. Regression
2. Causal perspective on statistical control
3. Regression and structural models

Regression

Linear Regression

- **Linear regression** is a basic data-fitting algorithm that can be used to predict the expected value of a dependent (target) variable, Y, given values of some predictor(s), X. Formally, this is written as:

$$\hat{Y}_{X=x} = \mathbb{E}[Y|X = x]$$

  - In the preceding formula, $\hat{Y}_{X=x}$ is the predicted value of Y given that X takes the value(s) x.
  - $\mathbb{E}[\cdot]$ is the expectation operator (average)
  - X can be multidimensional. In that case X is represented as a matrix, with dimensions $n \times d$, where n is the number of observations and d is the dimension. In multiple dimension case, curve fitting is called **multiple regression**.
- In **multiple regression** with k predictors X_1, X_2, X_3,...,X_k, each predictor, X_j has its respective coefficient, β_j. β_j represents the relative contribution of X_j to the change in the predicted target, $\hat{Y}$, holding everything else constant.

Linear Regression: $X->Y$ or $Y->X$

1. We begin with a very simple model:

$$y = \alpha + \beta x \tag{1}$$

Here:
- y, x: Observed Y,X variable.
- α,β: Intercept and slope

2. However, we can also fit the reverse model:

$$x = \alpha + \beta y \tag{2}$$

3. For both models (1) and (2), we get valid values of α and β, showing the lack of understanding of causal link. i.e., whether $X->Y$ or $Y->X$?

CODE

Refer to the [Python notebook] for implementation of:
- Data generation for regression
- Fitting both regression models in statsmodel (a python statistical library)

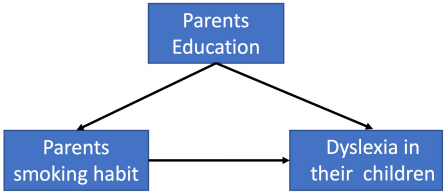
Regression

- Causal attributions become even more complicated in multiple regression, where each additional predictor can influence the relationship between the variables in the model.
- For instance, the learned coefficient for variable X might be 0.63, but when we add variable Z to the model, the coefficient for X changes to -2.34. A natural question in such cases is: if the coefficient has changed, what is the true effect here?
- Let's take a look at this issue from the point of view of statistical control.

Causal perspective on statistical control

Statistical Control (and confounding)

- **Definition**: Statistical control for a variable refers to the process of using it as one of the predictors in a regression model.



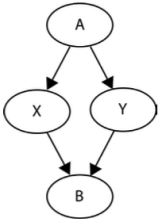
- Consider the SCM above for ' understanding whether parents smoking influences the risk of dyslexia in their children'.
- However, higher education of parents can be a factor that influences their smoking habits (reduces), as well as how much attention they devote to their child's reading and writing.
- Thus, parents education level is potentially a **confounding variable** in analyzing the impact of their smoking habits on risks of dyslexia in their children.

Statistical Control (and confounding)

- *But how do we actually know whether a variable does lead to confounding?* In some cases, we can refer to previous research to find an answer or at least a hint. In other cases, we can rely on our intuition or knowledge about the world.
- For example, Becker and colleagues (Becker et al., 2016) shared a set of 10 recommendations on how to approach statistical control. Some of their recommendations are as follows (the original ordering is given in parentheses):
  - If you are not sure about a variable, don't use it as a control (1).
  - Use conceptually meaningful control variables (3).
  - Conduct comparative tests of relationships between the independent variables and control variables (7).
  - Run results with and without the control variables and contrast the findings (8).
- Recommendations (1) and (3) discourage adding variables to the model. However *Not including a variable in the model might also lead to confounding and spuriousness*. This is because there are various patterns of independence structure possible between any three variables.

Statistical Control (and confounding)

- Consider the SCM (although not shown, there might be inherent exogenous variables)



- From the model structure, we can clearly see that X and Y are causally independent. There's no arrow between them, nor is there a directed path that would connect them indirectly.
- Let's fit four models and analyze which variables, when controlled for, lead to spurious relationships between X and Y:
  1. $Y = f(X)$
  2. $Y = g(X, A)$
  3. $Y = h(X, B)$
  4. $Y = k(X, A, B)$

```
Model 1: [1,x]
Params: [0.011 0.94 ]
p-vals: [0.246 0.   ]
Signif: [False  True]

Model 2: [1,x,a]
Params: [3.000e-03 1.000e-03 1.997e+00]
p-vals: [0.702 0.959 0.   ]
Signif: [False False  True]

Model 3: [1,x,b]
Params: [ 0.    -2.     1.333]
p-vals: [0. 0. 0.]
Signif: [ True  True  True]

Model 4: [1,x,a,b]
Params: [ 0.    -2.    -0.     1.333]
p-vals: [0. 0. 0. 0.]
Signif: [ True  True  True  True]
```

CODE

Refer to the [Python notebook] for implementation of:

- Data generation with the given SCM
- Fitting multiple regression models and comparing significance.

Things to be noted

The only model that recognized the causal independence of X and Y correctly (large p-value for X, suggesting the lack of significance) is the second model $Y = g(X, A)$. This clearly shows us that all other statistical control schemes led to invalid results, including the model that does not control for any additional variables. *Why did controlling for A work while all other schemes did not? There are three elements to the answer*:

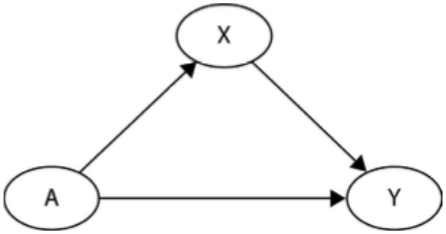
1. First, A is a confounder between X and Y and we need to control for it in order to remove confounding.
2. Second, X, Y, and B form a pattern that we call a **collider** or **immorality**. This pattern enables the flow of information between the parent variables (X and Y in our case) when we control for the child variable (B in our example). This is exactly the opposite of what happened when we controlled for A!
3. Third, not controlling for any variable leads to the same result in terms of the significance of X as controlling for A and B. This is precisely because the effects of controlling for A and controlling for B are exactly the opposite from a structural point of view and they cancel each other out!

Regression and structural models

Linear Regression vs SCMs

- In causal literature, the names **structural equation model (SEM)** and **structural causal model (SCM)** are sometimes used interchangeably.
- Linear regression is a model that allows us to quantify the (relative) strength of a (linear in parameters) relationship between two or more variables.
- There is no notion of causal directionality in linear regression, and in this sense, we don't know which direction (if any) is the causally correct one. This condition is known as **observational equivalence**.
- Linear regression can be used to estimate causal effects, given that we know the underlying causal structure (which allows us to choose which variables we should control for) and that the underlying system is linear in terms of parameters. In the previous example $Y = g(X, A)$ was found to have the right set of controls, which was intuitive from the data SCM.

Another SCM (Linear in parameters but Non-linear in data)



Here we will consider the functional assignments:

- $A \sim N(0, 1)$
- $X := 2A + 0.7 * N(0, 1)$
- $Y := 2A + 3X + 0.75X^2$

Another SCM (Linear in parameters but Non-linear in data)

- Note that our functional assignment contained not only X but also X^2. This is the simplest way to introduce non-linearity into a linear regression model. Also, note that the model is still linear in parameters (we only use addition and multiplication).
- Additionally, we included A in the model. The reason for this is that A is a confounder in our dataset and – as we learned before – we need to control for a confounder in order to get unbiased estimates.
- Since we are controlling for the right confounders, we should be able to recover the correct causal effects, even though we have non-linearity in data. Here causal effect refers to the learnt model weights.

CODE

Refer to the [Python notebook] for implementation of:

- Data generation with the given SCM
- Fitting multiple regression models and comparing significance.

Results

OLS Regression Results

Dep. Variable:	y	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	2.942e+33
Date:	Sun, 26 May 2024	Prob (F-statistic):	0.00
Time:	02:42:13	Log-Likelihood:	1.5551e+05
No. Observations:	5000	AIC:	-3.110e+05
Df Residuals:	4996	BIC:	-3.110e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.497e-15	1.29e-16	27.014	0.000	3.24e-15	3.75e-15
x	3.0000	1.49e-16	2.02e+16	0.000	3.000	3.000
x^2	0.7500	1.62e-17	4.63e+16	0.000	0.750	0.750
a	2.0000	3.15e-16	6.35e+15	0.000	2.000	2.000

Omnibus:	3731.287	Durbin-Watson:	0.684
Prob(Omnibus):	0.000	Jarque-Bera (JB):	91524.281
Skew:	-3.342	Prob(JB):	0.00
Kurtosis:	22.865	Cond. No.	26.0

Since we controlled for the confounder, the correct weight (causal effect of X on Y) have been estimated with high significance levels ($p < 0.05$).

Regression and causal effects

- For a functional assignment of form: $Y = 7X$, a unit increase in X, will increase Y by 7.
- If we have multiple predictors: X_1, X_2 and X_3, each of their respective coefficients will have a analogous interpretation.
- When we add a transformed version of a variable to the model (as we did by adding X^2), the interpretation becomes slightly less intuitive. For example for the functional assignment:

$$Y = \beta_1 X_1 + \beta_2 X_1^2$$

the causal effect of X_1 on Y can be obtained by using derivative:

$$\frac{dY}{dX_1} = X_1 + 2\beta_2 X_1$$

- The causal effect is no longer quantified solely by its coefficient. When the effect of a variable on the outcome depends on this variable's value, we say that the effect is **heterogeneous**.

In conclusion

- The causal interpretation of linear regression only holds when there are no spurious relationships in your data. This is the case in two scenarios:
  1. When you control for a set of all necessary variables (sometimes this set can be empty).
  2. When your data comes from a properly designed randomized experiment
- Any time you run regression analysis on arbitrary real-world observational data, there's a significant risk that there's hidden confounding in your dataset and so causal conclusions from such analysis are likely to be (causally) biased.
- Additionally, remember that in order to obtain a valid regression model, a set of core regression assumptions should be met
  1. Linearity in parameters.
  2. Homoscedasticity of variance.
  3. Independence of observations.
  4. Normality of Y for any fixed value of X.

Linear Regression assumptions

The relationship between the independent variable(s) X and the dependent variable Y is assumed to be linear. Mathematically, the model is expressed as:

$$Y = X\beta + \varepsilon, \quad \text{where } \varepsilon \sim N(0, \sigma^2 I)$$

where ε is the error term. *A linear model is a valid model for your data if following assumptions are satisfied*:

- **Linearity in parameter**: The Parameters $[\beta_0, \beta_1, ..]$ are separable from X and have degree 1.
- **Homoscedasticity of variance**: The variance of the errors should be constant across all values of the independent variable(s):

$$\text{Var}(\varepsilon_i) = \sigma^2, \quad \forall i$$

- **Independence of observations**: Y_i is independent of Y_j
- **Normality of Y for any fixed X**: Observations at a fixed x are normally distributed. This assumption ensures valid hypothesis tests and confidence intervals:

$$Y|X \sim N(X\beta, \sigma^2 I)$$