2 Causal interpretation of regression. Least Squares.

2.1 Regression and Causality

A variable x_1 can be said to have a causal effect on the response variable y if the latter changes when all other inputs are held constant. We can write a full model for the response variable y as:

$$y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$$

where x_1 and $\mathbf{x_2}$ are the observed variables, $\boldsymbol{\varepsilon}$ is an $\ell \times 1$ unobserved random factor and h is a functional relationship.

Definition 2.1.1: Causal effect

In the model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ the **causal effect** of x_1 on y is

$$C(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon}) = \nabla_1 h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon}),$$

the change in y due to a change in x_1, holding $\mathbf{x_2}$ and $\boldsymbol{\varepsilon}$ constant.

Note:-

This is just a definition, and does not necessarily describe causality in a fundamental or experimental sense. It might be more appropriate to label this a structural effect (the effect within the structural model).

Example. Suppose firms have Cobb-Douglas production functions:

$$y = AK^{\alpha}L^{\beta}$$

where K, L are observed capital and labour, A is an unobserved production technology and y is output. Here $x_1 = K, x_2 = L, \varepsilon = A$. Then the causal effect of capital on output is

$$C(K, L, A) = y'(K, L, A) = \alpha AK^{\alpha-1}L^{\beta}.$$

Even for firms with identical inputs, this effect differs due to unobserved A.

Sometimes it is useful to write this relationship as a potential outcomes function

$$y(x_1) = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$$

where the notation implies that $y(x_1)$ is holding $\mathbf{x_2}$ and $\boldsymbol{\varepsilon}$ constant. A popular example arises in the analysis of treatment effects with a binary regressor x_1. Let $x_1 = 1$ indicate treatment (e.g., a medical procedure) and $x_1 = 0$ indicate non-treatment. In this case $y(x_1)$ can be written

$$y(0) = h(0, x_2, \boldsymbol{\varepsilon}), \;\; y(1) = h(1, x_2, \boldsymbol{\varepsilon})$$

where $y(0)$ and $y(1)$ are known as the latent outcomes associated with non-treatment and treatment, respectively. The causal effect of treatment for the individual is the change in their health outcome due to treatment; the change in y as we hold both x_2 and $\boldsymbol{\varepsilon}$ constant:

$$C(x_2, \boldsymbol{\varepsilon}) = y(1) - y(0).$$

This is random as both potential outcomes $y(0)$ and $y(1)$ are different across individuals.

Example. Suppose there are two individals Yinfeng and Charles, and both have the possiblity of being a PhD graduate or dropping out. Suppose Yinfeng would earn £8/hour without a PhD and £12/hour as a PhD grad, while Charles would earn £20/hour without and £30/hour with a PhD. The causal effect of a PhD on wages is £4/hour for Yinfeng and £10/hour for Charles.

In a sample, we cannot observe both outcomes from the same individual, we only observe the realised value. As the causal effect varies across individuals and is not observable, it cannot be measured on the individual level. We therefore focus on aggregate causal effects, in particular what is known as the average causal effect.

Definition 2.1.2: Average causal effect

In the model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ the **average causal effect** of x_1 on y conditional on $\mathbf{x_2}$ is

$$\begin{aligned} ACE(x_1, \mathbf{x_2}) &= \mathbb{E}(C(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \,|\, x_1, \boldsymbol{x_2}) \\ &= \int_{\mathbb{R}^\ell} \nabla_1 h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \end{aligned}$$

where $f(\boldsymbol{\varepsilon} \,|\, x_!, \boldsymbol{x_2})$ is the conditional density of $\boldsymbol{\varepsilon}$ given $x_1, \boldsymbol{x_2}$.

Example. In the Cobb-Douglas example, the ACE of capital on output will be:

$$ACE(K, L) = \mathbb{E}(\alpha A K^{\alpha-1} L^\beta | K, L) = \alpha \mathbb{E}(A|K, L) K^{\alpha-1} L^\beta$$

Example. Considering again Yinfeng and Charles, suppose half our population are Yinfeng's and the other half Charles's, then the average causal effect of a PhD is $(10 + 4)/2 =$£7/hour. This is not the individual causal effect, it is the average of the causal effect across all individuals in the population.

We can think of $ACE(x_1, \mathbf{x_2})$ as the average effect in the general population. When we conduct regression analysis we might hope that regression reveals the ACE, i.e.: what is the relationship between $ACE(x_1, \mathbf{x_2})$ and the regression derivative $\nabla_1 m(x_1, \mathbf{x_2})$? The model $h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ implies that the CEF is

$$\begin{aligned} m(x_1, \boldsymbol{x_2}) &= \mathbb{E}(h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \,|\, x_1, \boldsymbol{x_2}) \\ &= \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon}, \end{aligned}$$

the average causal equation, averaged over the conditional distribution of the unobserved component $\boldsymbol{\varepsilon}$.
Applying the marginal effect operator [1], the regression derivative is:

$$\begin{aligned} \nabla_1 m(x_1, \boldsymbol{x_2}) &= \int_{\mathbb{R}^\ell} \nabla_1 h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} + \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \\ &= ACE(x_1, \boldsymbol{x_2}) + \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \end{aligned}$$

In general we see that the regression dervative does not equal the average causal effect. They are only equal in the special case when the second term equals zero, which occurs when the conditional

[1] Alexei uses $\frac{\partial}{\partial x_1}$ throughout, this is equivalent to the marginal effect operator used here with continuous x_1.

density of $\boldsymbol{\varepsilon}$ given $(x_1, \boldsymbol{x_2})$ does not depend on x_1 ($\nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) = 0$). When this condition holds then the regression derivative equals the ACE, which means that regression analysis can be interpreted causally, in the sense that it uncovers average causal effects.

Definition 2.1.3: Condiional Independence Assumption (CIA)

Conditional on $\mathbf{x_2}$, the random variables x_1 and $\boldsymbol{\varepsilon}$ are statistically independent.

The CIA implies $f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) = f(\boldsymbol{\varepsilon} \,|\, \boldsymbol{x_2})$ does not depend on x_1, and thus $\nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) = 0$. Thus the CIA implies that the regression derivative equals the ACE.

Theorem 2.1.1. In the structural model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$, the CIA implies

$$\nabla_1 m(x_1, \boldsymbol{x_2}) = ACE(x_1, \boldsymbol{x_2})$$

the regression derivative equals the average causal effect for x_1 on y conditional on $\mathbf{x_2}$.

Example (Nerlove: Returns to scale in electriciy supply)**.** Nerlove investigated returns to scale in a regulated industry (U.S. electricity) using Cobb-Douglas production. The market had the following features:

1. Privately owned local monopolies supply electricity on demand
2. These local monopolies face competitive factor prices
3. Electricity prices are set by the government

Notably Y is exogenously given (by consumer demand). Nerlove assumes firms pick K, L to minmise the cost of producing $Y = AK^\alpha L^\beta$, i.e. K, L both depend on A, Y, in particular $f(A|K, L)$ depends on K. Thus a regression of Y on K, L will not identify the ACE.

$$\min_{K,L} p_K K + p_L L \;\; s.t. \;\; Y = AK^\alpha L^\beta$$

The Lagrangian and FOCs for this problem are:

$$\mathcal{L} = p_K K + p_L L + \lambda(Y - AK^\alpha L^\beta)$$

$$\frac{\partial \mathcal{L}}{\partial K} = p_K - \lambda \alpha A K^{\alpha-1} L^\beta = 0, \quad \frac{\partial \mathcal{L}}{\partial L} = p_L - \lambda \beta A K^\alpha L^{\beta-1} = 0$$

$$\Rightarrow K = \frac{\alpha p_L}{\beta p_K} L$$

We can subsititute this into the production function to solve for L and K, giving:

$$TC = p_K \left(\frac{\alpha p_L}{\beta p_K} \left(\frac{Y}{A \left(\frac{\alpha p_L}{\beta p_K} \right)^\alpha} \right)^{\frac{1}{\alpha+\beta}} \right) + p_L \left(\left(\frac{Y}{A \left(\frac{\alpha p_L}{\beta p_K} \right)^\alpha} \right)^{\frac{1}{\alpha+\beta}} \right)$$

$$TC = p_L \left(\frac{Y \left(\frac{p_L \alpha}{p_K \beta} \right)^{-\alpha}}{A} \right)^{\frac{1}{r}} \left(\frac{r}{\beta} \right) = r \alpha^{-\alpha/r} \beta^{-\beta/r} A^{-1/r} Y^{1/r} p_K^{\alpha/r} p_L^{\beta/r}$$

Taking logs we obtain the following log-linear relationship for each firm:

$$\log(TC_i) = \mu_i + \frac{1}{r}\log(Y_i) + \frac{\alpha}{r}\log(p_K) + \frac{\beta}{r}\log(p_L)$$

where $\mu_i = \log[r(A_i\alpha^\alpha\beta^\beta)^{-\frac{1}{r}}]$. Coefficients in this equation are elasticities, for example $\frac{\beta}{r}$ is the elasticity of total cost with respect to the wage rate, i.e.: the percentage change in in total cost when the wage rate changes by 1%. The degree of returns to scale (the reciprocal of the output elasticity of total costs), is independent of the level of output.
To estimate this define $\mu \equiv \mathbb{E}[\mu_i]$, $\varepsilon_i \equiv \mu - \mu_i$ so $\mathbb{E}[\varepsilon_i] = 0$, firms with positive ε_i are high-cost firms.

$$\log(TC_i) = \beta_0 + \beta_1 log(Y_i) + \beta_2\log(p_K) + \beta_3\log(p_L),$$

where

$$\beta_0 = \mu, \beta_1 = \frac{1}{r}, \beta_2 = \frac{\alpha}{r}, \beta_3 = \frac{\beta}{r}$$

This equation is overidentified, the 4 coefficients are not free parameters, they are a function of three technology parameters (α, β, μ). Clearly $\beta_2 + \beta_3 = 1$ (as expected, cost function is linearly homogenous in factor prices). To fix this we can subtract p_L from each side and consider relative prices:

$$\log\left(\frac{TC_i}{p_L}\right) = \beta_0 + \beta_1 log(Y_i) + \beta_2\log\left(\frac{p_K}{p_L}\right)$$

To test constant returns to scale ($r = 1$), just t-test $\beta_1 = 1$ in this restricted model.

2.2 Estimating population regression by least squares

If CIA holds, regression captures the causal effect of x's on y. However even if it doesn't, it still provides the best predictor of y given x's. We assume the regression function $\mathbb{E}[y|x] = m(x)$ is parametrised by a finite dimensional vector $\beta = [\beta_1, ..., \beta_k]^T$, so that estimating the population regression $m(x;\beta)$ is equivalent to estimating β. One approach to estimation is using the analogy principle.

Definition 2.2.1: Analogy principle

Consider finding an estimator that satisfies the same properties in the sample that the parameter satisfies in the population; i.e., seek to estimate $\beta(P)$ with $\beta(P_n)$ where P_n is the empirical distribution which puts mass $\frac{1}{n}$ at each sample point. Note this distribution converges uniformly to P.

In the regression context:

$$\beta = \arg\min_b \mathbb{E}(y - m(x;b))^2$$

The sample analgoue of expectation is the average:

$$\hat{\beta} = \arg\min_b \frac{1}{n}\sum_{i=1}^{n}(y - m(x;b))^2$$

When $m(x;b)$ is linear in b, the method is called OLS. We assume that the pbservations of the data (y_i, x_i) are independent and come from the same joint distribution. Let

$$\underbrace{X_i}_{K\times 1} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{iK} \end{bmatrix}, \underbrace{\beta}_{K\times 1} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix},$$

$$\underbrace{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \underbrace{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \underbrace{X}_{n\times K} = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix}.$$

When our model contains a constant, one of the columns of X will contain only ones. Our linear model can thus be represented as:

$$Y = X\beta + \varepsilon$$

When estimating we select the $\hat{\beta}$ such that the sum of squared residuals ($e'e$) is minimised[1].

$$\begin{aligned} e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= (y' - \hat{\beta}'X')(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

Where $y'X\hat{\beta} = (y'X\hat{\beta})' = \hat{\beta}'X'y$ since the transpose of a scalar is itself.

Note:-

Matrix differentiation

$$\frac{\partial \mathbf{a}'\mathbf{b}}{\partial \mathbf{b}} = \frac{\partial \mathbf{b}'\mathbf{a}}{\partial \mathbf{b}} = \mathbf{a} \tag{2.1}$$

when $\mathbf{a}$ and $\mathbf{b}$ are $K \times 1$ vectors.

$$\frac{\partial \mathbf{b}'\mathbf{A}\mathbf{b}}{\partial \mathbf{b}} = 2\mathbf{A}\mathbf{b} = 2\mathbf{A}'\mathbf{b} \tag{2.2}$$

when $\mathbf{A}$ is any symmetric matrix.

$$\frac{\partial 2\mathbf{b}'\mathbf{X}'\mathbf{y}}{\partial \mathbf{b}} = \frac{\partial 2\mathbf{b}'(\mathbf{X}'\mathbf{y})}{\partial \mathbf{b}} = 2\mathbf{X}'\mathbf{y} \tag{2.3}$$

and

$$\frac{\partial \mathbf{b}'\mathbf{X}'\mathbf{X}\beta}{\partial \mathbf{b}} = \frac{\partial 2\mathbf{A}\beta}{\partial \mathbf{b}} = 2\mathbf{A}\beta = 2\mathbf{X}'\mathbf{X}\beta \tag{2.4}$$

when $\mathbf{X}'\mathbf{X}$ is a $K \times K$ matrix.

Solving for the minimum:

$$\begin{aligned} \frac{\partial e'e}{\partial \hat{\beta}} &= -2X'y + 2X'X\hat{\beta} = 0 \\ &\Rightarrow X'X\hat{\beta} = X'y \\ &\Rightarrow (X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y \\ &\Rightarrow I_K\hat{\beta} = (X'X)^{-1}X'y \\ &\Rightarrow \hat{\beta} = (X'X)^{-1}X'y \end{aligned}$$

Here we have assumed that the inverse of $X'X$ exists, i.e. X is full rank[2]. To check this is a minimum, take second derivative which gives us $2X'X$ which is clearly positive semi-definite (when X is full rank). Note that $X'X$ is always square ($k \times k$) and always symmetric.

[1] Note that $e \neq \varepsilon$, residuals e are observed, whilst disturbances ε are unobserved.

[2] The inverse of $X'X$ may not exist, it does not exist in the following two cases: 1) When $n < k$; we have more independent variables than observations 2) One or more of the independent variables are a linear combination of the other variables i.e. perfect multicollinearity.

We can further show that $X'e = 0$, consider the normal form equations $X'X\hat{\beta} = X'y$:

$$\begin{aligned}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= \mathbf{X}'(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}) \\ (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{e} \\ \mathbf{X}'\mathbf{e} &= \mathbf{0}\end{aligned}$$

Proposition 2.2.1 (Properties of OLS). From $X'e$=0 we can derive a number of properties.

1. The observed values of X are uncorrelated with the residuals.
2. The sum of the residuals is zero.
3. The sample mean of the residuals is zero.
4. The regression hyperplane passes through the sample means of observables.
5. The predicted values of y are uncorrelated with the residuals.

Where 2-5 hold when the regression includes a constant term.

Proof. Using $X'e = 0$

1. $\mathbf{X}'\mathbf{e} = 0$ implies that for every column $\mathbf{x}_k$ of $\mathbf{X}$, $\mathbf{x}_k'\mathbf{e} = 0$. In other words, each regressor has zero sample correlation with the residuals. Note that this does not mean that $\mathbf{X}$ is uncorrelated with the disturbances; we'll have to assume this.
2. If there is a constant, then the first column in $\mathbf{X}$ (i.e. $\mathbf{X}_1$) will be a column of ones. This means that for the first element in the $\mathbf{X}'\mathbf{e}$ vector (i.e. $\mathbf{X}_{11}e_1 + \mathbf{X}_{12}e_2 + \ldots + \mathbf{X}_{1n}e_n$), to be zero, it must be the case that $\sum_i e_i = 0$.
3. This follows straightforwardly from the previous property i.e. $\bar{e} = \frac{\sum e_i}{n} = 0$.
4. This follows from the fact that $\bar{e} = 0$. Recall that $e = y - \mathbf{X}\hat{\boldsymbol{\beta}}$. Dividing by the number of observations, we get $\bar{e} = \bar{y} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}} = 0$. This implies that $\bar{y} = \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}$.
5. $\hat{y}'e = (\mathbf{X}\hat{\boldsymbol{\beta}})'e = \hat{\boldsymbol{\beta}}'\mathbf{X}'e = 0$

□