

*Regression discontinuity design (RDD)

In statistics, econometrics, political science, epidemiology, and related disciplines, a **regression discontinuity design (RDD)** is a quasi-experimental pretest-posttest design that elicits the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments in which randomization is unfeasible. First applied by Donald Thistlethwaite and Donald Campbell to the evaluation of scholarship programs, the RDD has become increasingly popular in recent years.

Example

The intuition behind the RDD is well illustrated using the evaluation of merit-based scholarships. The main problem with estimating the causal effect of such an intervention is the endogeneity of performance to the assignment of treatment (e.g. scholarship award): Since high-performing students are more likely to be awarded the merit scholarship *and* continue performing well at the same time, comparing the outcomes of awardees and non-recipients would lead to an upward bias of the estimates. Even *if* the scholarship did not improve grades at all, awardees would have performed better than non-recipients, simply because scholarships were given to students who were performing well ex ante.

Despite the absence of an experimental design, a RDD can exploit exogenous characteristics of the intervention to elicit causal effects. If all students above a given grade—for example 80%—are given the scholarship, it is possible to elicit the local treatment effect by comparing students around the 80% cut-off: The intuition here is that a student scoring 79% is likely to be very similar to a student scoring 81%—given the pre-defined threshold of 80%, however, one student will receive the scholarship while the other will not. Comparing the outcome of the awardee (treatment group) to the counterfactual outcome of the non-recipient (control group) will hence deliver the local treatment effect. Another RDD example seem is related to the impact of electronic voting on null votes and health outcomes.

The major benefit of using non-parametric methods in a RDD is that they provide estimates based on data closer to the cut-off, which is intuitively appealing. This reduces some bias that can result from using data farther away from the cutoff to estimate the discontinuity at the cutoff. More formally, local linear regressions are preferred because they have better bias properties and have better convergence. However, the use of both types estimation, if feasible, is a useful way to argue that the estimated results do not rely too heavily on the particular approach taken.

**Non-parametric estimation

The most common non-parametric method used in the RDD context is a local linear regression. This is of the form:

$$Y = \alpha + \tau D + \beta_1(X - c) + \beta_2 D(X - c) + \epsilon,$$

where c is the treatment cut-off and D is a binary variable equal to one if $X \geq c$.

Letting h be the bandwidth of data used, we have $c - h \leq X \leq c + h$.

Different slopes and intercepts fit data on either side of the cutoff. Typically either a rectangular kernel (no weighting) or a triangular kernel are used. Research favors the triangular kernel but the rectangular kernel has a more straightforward interpretation.

Stepwise regression

Instead of imposing a particular model of analysis suggested from theory, we can implement a statistical stepwise variable selection procedure to determine which socio-economic variables are more statistically important to explain each social outcome variable. We can apply it to discrete endogenous variable models (for example logistic regression on poverty) or continuous such as an OLS log-linear mincerian regression.

In statistics, **stepwise regression** is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Usually, this takes the form of a sequence of F -tests or t -tests, but other techniques are possible,

The frequent practice of fitting the final selected model followed by reporting estimates and confidence intervals without adjusting them to take the model building process into account should suggest caution. One interesting outcome of the stepwise procedure is to indicate the order of importance of different variables used.

Stepwise on Poverty and Infrastructure with Externality Related Variables

We provide an example of Stepwise regression including In the selection process we included variables that capture externality effects from infrastructure. This is done by including in the regressions the mean of these variables across geographic areas. The idea is to see how much a subdivision of the 27 units of the federation into three or four areas each namely rural, urban non metropolitan or capital of the state. In the case of the states that include one of the 11 major Brazilian metropolitan cities we include a finer division between capital and suburbs for these metro regions. Given the difference in economies or diseconomies of scale between cities sizes. The idea is that beyond individual impacts at the household level, what our neighbors and other community members have in terms of infrastructure use may also affect ours respective social outcomes. For example, if there is a widespread diffusion of landline or cell phones in my region of residence the value of my phone line increases due to network scales, given the fixed cost of intercity connections. Following a different strand the effects of electricity access at the community level may also improve my social outcomes through better work opportunities or school or health services. Transportation use on the other extreme imply a common good congestion problem where the excessive use of infrastructure generates a negative externality on all users. The order of variable selection is indicative of the relevance reached by each explanatory variable.

In the case of the proportion of the poor, the six infrastructure variables are significant in descending order: communication, internet, transportation, water, electricity and sewerage. Two of the externality related variables also presented statistically significant impacts, namely mean transportation time and mean electricity coverage.

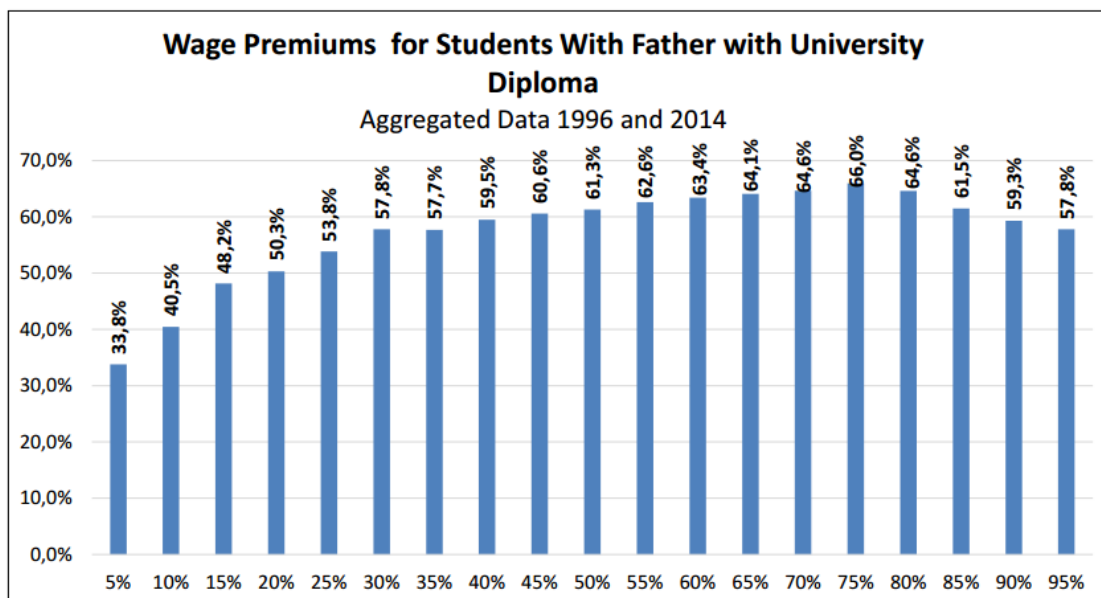
Quantile regression is a type of regression analysis used in statistics and econometrics. Whereas the method of least squares results in estimates that approximate the conditional *mean* of the response variable given certain values of the predictor variables, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. It is useful to depict the impacts of different variables along the income distribution. Fort example, what are the returns to schooling at the basis of the income distribution using a mincerian type quantile regression.

One advantage of quantile regression, relative to the ordinary least squares regression, is that the quantile regression estimates are more robust against outliers in the response measurements.

The mathematical forms arising from quantile regression are distinct from those arising in the method of least squares. The method of least squares leads to a consideration of problems in an inner product space, involving projection onto subspaces, and thus the problem of minimizing the squared errors can be reduced to a problem in numerical linear algebra. Quantile regression does not have this structure, and instead leads to problems in linear programming that can be solved by the simplex method.

Example: Mincerian Regression the dummy of the father with university degree on labor earnings:

QUANTILE REGRESSIONS

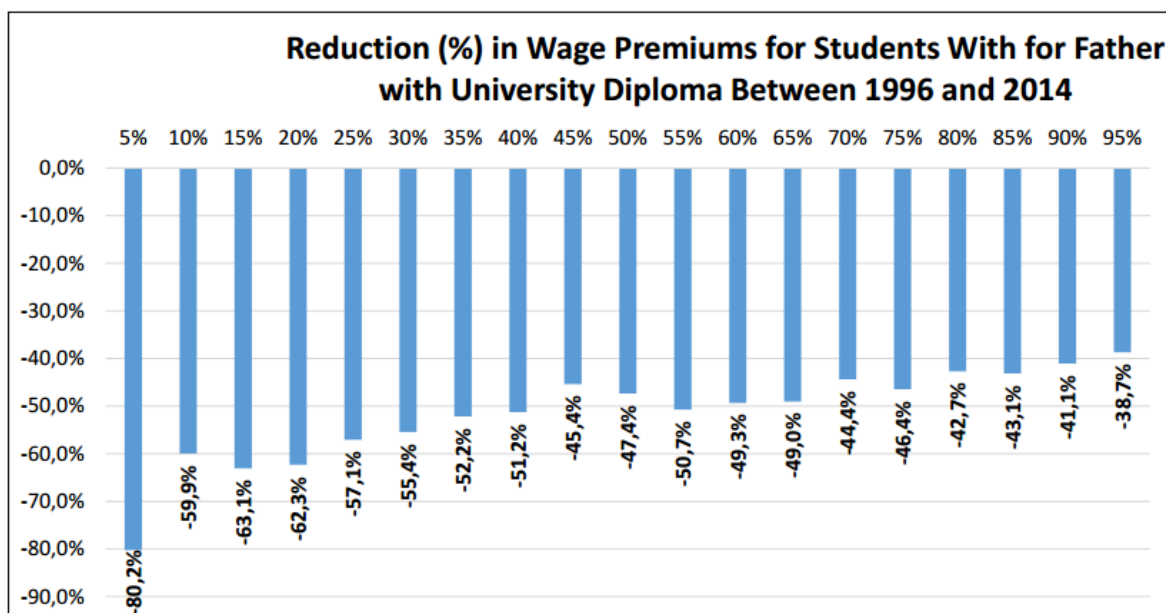


Source: FGV Social/CPS with PNAD/IBGE microdata of 1996 and 2014

➤ Results show a lower bonus for the basis of the income distribution

Additionally interacting time dummy (2014 (basis 1996)) with dummy of the father with university degree on labor earnings:

QUANTILE REGRESSIONS



Source: FGV Social/CPS with PNAD/IBGE microdata of 1996 and 2014

➤ Results show a higher bonus reduction for the basis of the income distribution

***Box Metodológico – Regressão Quantílica

Uma análise de regressão simples tem como objetivo investigar o relacionamento entre duas ou mais variáveis, tendo como ponto de partida entender se certa(s) variável(is) influencia(m) a variável que se quer explicar e de que forma.

O diferencial da técnica de regressão quantílica é que ela possibilita analisar o poder explicativo das variáveis independentes (que influenciam a variável que se quer explicar) sobre a variável dependente (a que se quer explicar) em diferentes quantis da distribuição condicional. Dizemos que um estudante tirou uma nota no τ -ésimo quantil se ele se saiu melhor que a proporção τ de alunos, mas pior que a proporção $(1 - \tau)$. Logo, metade dos alunos foi melhor que o aluno na mediana da distribuição, enquanto a outra metade foi pior. Portanto, os quantis representam subgrupos de mesma proporção do total de dados disponíveis e podem assumir diferentes formatos. Um exemplo de quantil seria o formato em percentil, que divide os dados disponíveis em 100 quantis. Assim, pode-se mensurar, por exemplo, qual a diferença no poder explicativo da inflação sobre a queda recente na renda do trabalho para os 5% mais pobres ou para os 25% mais ricos, dada a distribuição de renda do trabalho de todos os indivíduos da amostra.

Formalmente, qualquer variável aleatória real X pode ser caracterizada por uma função de distribuição acumulada:

$$F(x) = P(X \leq x)$$

Utilizando a função inversa da distribuição acumulada no ponto τ , sendo $0 < \tau < 1$:

$$F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\}$$

Temos o τ -ésimo quantil da variável aleatória X . A mediana, por sua vez, seria definida como $Q(1/2)$. No entanto, podemos definir o quantil de uma outra forma, que é essencial no entendimento dos modelos de regressão quantílica. Seja Y com função de distribuição acumulada F . Estamos interessados no valor m que minimiza $E[Y - m]$. Esse valor é a mediana de Y . Esse resultado pode ser generalizado para todos os quantis. Dada a função de perda:

$$\rho_\tau(u) = u(\tau - I(u < 0)), \quad 0 < \tau < 1$$

Em que I é a função indicadora, buscamos encontrar \hat{x} , um preditor de X , que minimize a perda esperada e represente o τ -ésimo quantil. Então temos,

$$E[\rho_\tau(X - \hat{x})] = (\tau - 1) \int_{-\infty}^{\hat{x}} (x - \hat{x}) dF(x) + \tau \int_{\hat{x}}^{\infty} (x - \hat{x}) dF(x).$$

Diferenciando esta expressão em relação a \hat{x} e igualando a zero, temos:

$$(1 - \tau) \int_{-\infty}^{\hat{x}} dF(x) - \tau \int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau = 0.$$

Com essa definição de quantil podemos seguir para a definição de regressão quantílica. Dada uma variável aleatória Y com n observações, a média amostral é definida pela seguinte minimização:

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^n (y_i - \mu)^2.$$

Vimos que se quisermos prever \hat{x} via a função de perda descrita anteriormente, F pode ser descrita por:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$$

E a minimização da perda esperada é:

$$\int \rho_\tau(x - \hat{x}) dF_n(x) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(x_i - \hat{x})$$

Dessa forma, o τ -ésimo quantil resolve o problema de minimização a seguir:

$$\min_{q \in \mathbb{R}} \sum_{i=1}^n \rho_\tau(y_i - q)$$

Sendo q igual a cada valor de x presente na amostra. Se a intenção é especificar o quantil condicional de Y dado x como uma função linear nos parâmetros de $Q_\tau(Y|x) = \mathbf{x}'\boldsymbol{\beta}(\tau)$, em que $\boldsymbol{\beta}(\tau)$ é um vetor de parâmetros, basta encontrar $\hat{\boldsymbol{\beta}}(\tau)$ que minimize:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i' \boldsymbol{\beta})$$

Se o interesse é estudar diversos quantis da distribuição condicional de Y , supondo relações lineares do tipo:

$$y_i = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip} + u_i$$

Sendo u_i variáveis aleatórias independentes e identicamente distribuídas com o τ -ésimo quantil igual a zero, temos que o τ -ésimo quantil condicional de Y/X é:

$$Q_\tau(Y|x) = \beta_0(\tau) + \beta_1(\tau)x_1 + \cdots + \beta_p(\tau)x_p.$$