Nonparametric Method: Take-Home Exam 2

Mai-Anh Dang | Student ID: 21608631 | TSE M2 EEE 2017-18

November 7, 2017

1 Exercise 1

From the dataset *anglin.gencay*.1996.*csv*. The single-index model below describes the relationship between the price of the house and observed characteristics,

$$p_i = g(x_i'\beta) + \epsilon_i, E(\epsilon|x) = 0$$

(a) Conditions on β, that β identified

Suppose that:

$$\begin{aligned} g^*(\gamma + \delta v) &= g(v), \forall v \\ E[p|X = x] = g(x'\beta) &= g^*(\gamma + x'\beta\delta) \end{aligned}$$

Then, β is not identified. Therefore, the Scale should be normalized (restriction on δ), so in $\beta = (\beta_1, \beta_2, ..., \beta_d)'$, we should set: $\beta_1 = 1$. It means that β_i is interpreted as the effect of x_i on p_i, relative to x_1.

In the context of this dataset, most of regressors are discrete, but at least, there is *lot* variables supposed to be continuous. It should be followed with the conditions that β associated with the discrete covariates must not divide the support of $x'\beta$ into disjoint subsets.

(b) Define the density-weighted average derivative estimator (dWADE) for β

Besides this condition of β, to ensure the identification of β, we need further condition of x and g(.):

- As the above, γ also needs to be normalized. x contains no constant component, and not multicollinearity.
- g(.) is differentiable and has no constant on the support of $x'\beta$
- $g(.)$ should not be periodic function.

For $x = (x_1, x_2, ..., x_{11})$, in which x_1 is continuous. Assume that $g(.)$ is differentiable (as the discussed conditions):

$$\begin{aligned} & E[p|x] = g(x'\beta)(*) \\ \implies & \frac{dE[p|x]}{dx} = \beta g'(x'\beta) \\ \implies & E_x\left[W(x)\frac{dE[p|x]}{dx}\right] = \beta E_x[W(x)g'(x'\beta)] \end{aligned}$$

$E_x[W(x)g'(x'\beta)$ is a scalar, we identify β up-to-scale:

$$\beta \propto E_x\left[W(x)\frac{dE[p|x]}{dx}\right]$$

For the density-weighted ADE, we set: $W(x) = f(x)$, which is the density function of x. Let δ is observationally equivalent to β up-to-scale normalization

$$\begin{aligned}\delta &= E_x\left[W(x)\frac{dE[p|x]}{dx}\right] = E_x\left[f(x)\frac{dE[p|x]}{dx}\right] \\ &= \int \frac{dE[p|x]}{dx} f(x)^2 dx = -2E\left[p\frac{df(x)}{dx}\right]\end{aligned}$$

The dWADE estimate of β is the vector $\hat{\beta}_{dWADE}$:

$$\hat{\beta}_{dWADE} = \frac{-2}{n}\sum_{i=1}^{n} p_i \frac{d\hat{f}_{-i}(x_i)}{dx} \tag{1}$$

where $\hat{f}_{-i}(.)$ is the leave-one-out kernel estimator of the density, $\hat{f}_{-i}(x_i) = \frac{1}{n-1}\frac{1}{h_n^d}\sum_{j=1,j\neq 1}^{n} K\left(\frac{x_i - x_j}{h_n}\right)$ so:

$$\frac{d\hat{f}_{-i}(x_i)}{dx} = \frac{1}{n-1}\frac{1}{h_n^{d+1}}\sum_{j=1,j\neq 1}^{n} K'\left(\frac{x_i - x_j}{h_n}\right)$$

Substitute into (1):

$$\hat{\beta}_{dWADE} = -\frac{2}{n(n-1)}\frac{1}{h_n^{d+1}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} p_i K'\left(\frac{x_i - x_j}{h_n}\right)$$

Interpretations

- From the equation (*), which may have $f(x)$ in the denominator as $E[p|x] = \int p\frac{f(p,x)}{f(x)}dy$, we cancel out the $f(x)$ by using the density $f(x)$ as the weight ($W(x) = f(x)$).
- We do not have any random variable in denominator that makes the computation easier.
- By the "average" of kernel estimators, we obtain $\sqrt{n}(\hat{\beta}_{dWADE} - \beta)$ is asymptotically normal, and has faster rates of convergence ($n^{-1/2}$).
- Similar to the analysis in the Take-home exam 1, the estimator of the derivative density function will have the faster rate of convergence. Also, we can use the higher order of kernel for even faster rate of convergence.

(c) Estimate β using dWADE

Below, the procedure for estimating $\hat{\beta}_{dWADE}$ and the results will be presented. Transform all variables into logarithm, except dummies. Our objective equation is:

$$\hat{\beta}_{dWADE} = -\frac{2}{n(n-1)}\frac{1}{h_n^{d+1}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} p_i \underbrace{K'\left(\frac{x_i - x_j}{h_n}\right)}_{\texttt{mvdker(Xi, Xj, h)}} \tag{2}$$

Procedure

- `ker(v)`: use the gaussian kernel for each univariate $K(v_{ij,k}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{v_{ij,k}^2}{2}\right)$, $v_{ij,k}$ is a scalar, i.e. $v_{ij,k} = \frac{x_{i,k} - x_{j,k}}{h_n}$, when $i, j \in (1, 2, ..., n); k \in (1, 2, ..., 11)$.
- `mvker(V)`: V is a vector, $V_{ij}' = (v_{ij,1}, v_{ij,2}, ..., v_{ij,11})$, the multivariate kernel as the product of kernels $K(V_{ij}) = K(v_{ij,1})K(v_{ij,2})...K(v_{ij,11})$.
- `dker(v)`: the derivative of univariate gaussian kernel $K'(v_{ij,k}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{v_{ij,k}^2}{2}\right)(-v_{ij,k})$.

- `mvdker(Xi, Xj, h)`: From $x_i' = (x_{i,1}, x_{i,2}, ..., x_{i,11})$ and $x_j' = (x_{j,1}, x_{j,2}, ..., x_{j,11})$, we obtain $V_{ij}' = \left(\frac{x_{i,1}-x_{j,1}}{h_n}, ..., \frac{x_{i,11}-x_{j,11}}{h_n}\right)$

$$K'(V_{ij}) = \left(\frac{dK(V_{ij})}{dv_{ij,1}}, \frac{dK(V_{ij})}{dv_{ij,2}}, ..., \frac{dK(V_{ij})}{dv_{ij,11}}\right)'$$

  For element k of the vector:

$$\frac{dK(V_{ij})}{dv_{ij,k}} = \frac{K(v_{ij,1})K(v_{ij,2})...K(v_{ij,11})}{K(v_{ij,k})}K'(v_{ij,k})$$

  or `mvdker[i] = dker(v[i]) * mvker(V) / ker(v[i])`

- For the bandwidth, I use the *'kedd' package* for computing the bandwidth choice in the kernel estimation of the derivatives density. The *lot* is chosen for this computation, as it is the only continuous variate. Several methods in this package have been tried, with the results not much different from each other. I will report the results using the bandwidth from complete cross-validation (CCV) with the derivative order 1, $\mathbf{h_{ccv} = 0.18788}$.

- Finally, go over n observations and accumulate values by the formula (2): `p[i] * mvdker(x[i,], x[j,], hn)`.

Results and Comments

The results of $\hat{\beta}_{dWADE}$ and $\hat{\beta}_{OLS}$ with/without intercept is presented in **Table 1**. The coefficient is standardized to $\hat{\beta}_{lot}$, the dependent variable is $log(price)$. All other variables are transformed to logarithm, except dummies.

Table 1: Compare the results of OLS and dWADE

	$\hat{\beta}_{OLS}$	$\hat{\beta}_{OLS}$	$\hat{\beta}_{dWADE}$
(Intercept)	Yes	No	No
lot size	1	1	1
no. of bedrooms	0.359	0.194	0.255
no. of bathrooms	0.841	0.073	0.007
no. of stories	0.294	0.100	4.5e-06
driveway	0.361	-0.035	2.2e-06
recreational room	0.190	-0.056	-6.8e-06
finished basement	0.337	0.151	5.3e-06
water heating	0.584	0.158	1.3e-06
central air-con	0.539	0.025	5.7e-06
garage places	0.159	-0.051	5.1e-06
preferred neighborhood	0.425	0.008	3.6e-06

- The coefficients should be interpreted in term of scale normalization (i.e. the effect in relative to the effect of lot size).

- While $\hat{\beta}_{OLS}$ presents the marginal effect of regressors on the dependent variable, $\hat{\beta}_{dWADE}$ shows the effect of regressors in the index $x_i'\hat{\beta}$, which is associated with the p_i through the function $g(.)$, assumed to be a non-decreasing function. Thus, it is not very insightful to compare the magnitude of $\hat{\beta}_{dWADE}$ and $\hat{\beta}_{OLS}$, yet the sign of coefficient is still interesting to consider.

- Comparing with the $\hat{\beta}_{OLS}$, the estimate of $dWADE$ is not very contradict. The OLS without the intercept, equivalent to running a linear regression of p on $x' = (x_1, x_2, ..., x_{11})$, $\hat{\beta}_{OLS}$ has 03 negative signs for *driveway*, *recreational_room*, and *garage_places*. Meanwhile, $\hat{\beta}_{dWADE}$ only have the negative effect of *recreational_room*. On the other hand, in $\hat{\beta}_{OLS}$ with the intercept, all coefficients are positive. Most of the coefficients of $\hat{\beta}_{dWADE}$ is also positive. We can say that the estimate of $dWADE$ is more consistent with the estimates of OLS with intercept, which is probably the better OLS estimates.
- Noticing that the magnitude and the sign of $\hat{\beta}_{OLS}$ changes dramatically with and without the intercept, it signals that the $\hat{\beta}_{OLS}$ might be inaccurate when we force the model perfectly linear through the origin.

2 Exercise 2

On the dataset `GDP.xlsx`, with the GDP values (in billions of dollars) from 191 countries. We denote: $log(GDP_{2005}) = X_{05}$ and $log(GDP_{2016}) = X_{06}$.

(a) Construct a 95% confidence interval for the densities of X_{05} and X_{16}

(i) Bootstrap procedure using the test statistic: $T_n(x) = \hat{f}(x) - f(x)$

- Compute $\hat{f}(x)$ for X_{05} and X_{16}, $\hat{f}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)$. The bandwidths to estimate the densities are chosen by *Least-Squared Cross-Validation* , being under-smoothed to reduce the bias.
- Re-sampling $\{X_i^{05}\}_{i=1}^n$ and $\{X_i^{16}\}_{i=1}^n$, with *replacement* to get a bootstrap sample $\{X_i^{*05}\}_{i=1}^n$ and $\{X_i^{*16}\}_{i=1}^n$.
- Create $B = 500$ bootstrap samples. For each bootstrap sample, compute the estimate of densities: $\hat{f}^*(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i^*}{h_n}\right)$, then compute: $T_n^*(x) = \hat{f}^*(x) - \hat{f}(x)$.
- From B bootstrap $T_n^*(x)$, get the critical values at quantile $1-\alpha/2$, i.e. $P\left[|T_n^*| \geq c_{1-\alpha/2}^*\right] = \alpha$
- We have $-c_{1-\alpha/2}^* \leq T_n \leq c_{1-\alpha/2}^*$

$$\implies \hat{f}(x) - c_{1-\alpha/2}^* \leq f(x) \leq \hat{f}(x) + c_{1-\alpha/2}^*$$

(ii) Bootstrap procedure with Asymptotically pivotal test statistic

- Similar to the above procedure, yet for each bootstrap sample, compute the pivotal test statistic: $T_n^{*pivot} = \frac{T_n^*(x)}{s_n^*(x)} = \frac{\hat{f}^*(x)-\hat{f}(x)}{s_n^*(x)}$, in which $s_n^{*2}(x) = \frac{1}{nh_n^2}\sum_{i=1}^{n} K\left(\frac{x-X_i^*}{h_n}\right)^2 - \frac{\hat{f}^*(x)^2}{n}$.
- In the process, there are cases when $\hat{s^*}_n^2(x)$ is very close to 0, thus the $T_n^{*pivot}(x)$ turns to be infinitive. I solve this problem by eliminate these values.
- From B bootstrap $T_n^{*pivot}(x)$, get the critical values at quantile $1-\alpha/2$.
- We have $-c_{1-\alpha/2}^{*pivot} \leq T_n^{pivot} \leq c_{1-\alpha/2}^{*pivot}$:

$$\begin{aligned} \implies \quad & \hat{f}(x) - c_{1-\alpha/2}^{*pivot} \cdot \hat{s}_n(x) \leq f(x) \leq \hat{f}(x) + c_{1-\alpha/2}^{*pivot} \cdot \hat{s}_n(x) \\ where \quad & \hat{s}_n^2(x) = \frac{1}{nh_n^2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)^2 - \frac{\hat{f}(x)^2}{n} \end{aligned}$$

The estimate of $\hat{f}^{GDP}(x)$ and its 95% confidence intervals by test-statistic and asymptotically pivotal test statistic is presented in **Figure 1**.

Comment

- The confidence interval of the test statistic and the asymptotically pivotal test statistics are not much different. Considering the sample size $n = 191$, it is supposed that the CI of the test statistic is quite reliable. In this case, both CIs is acceptably accurate, and relatively equivalent.
- The pivotal test statistic is refined from the test statistic T_n. By studentizing each bootstrap estimate, we have a faster rate of convergence, so the CI of pivotal test statistic behave more accurately. In fact, the pivotal bootstrap CI is slightly more conservative (i.e. the interval is wider), comparing to the simple bootstrap test statistic.

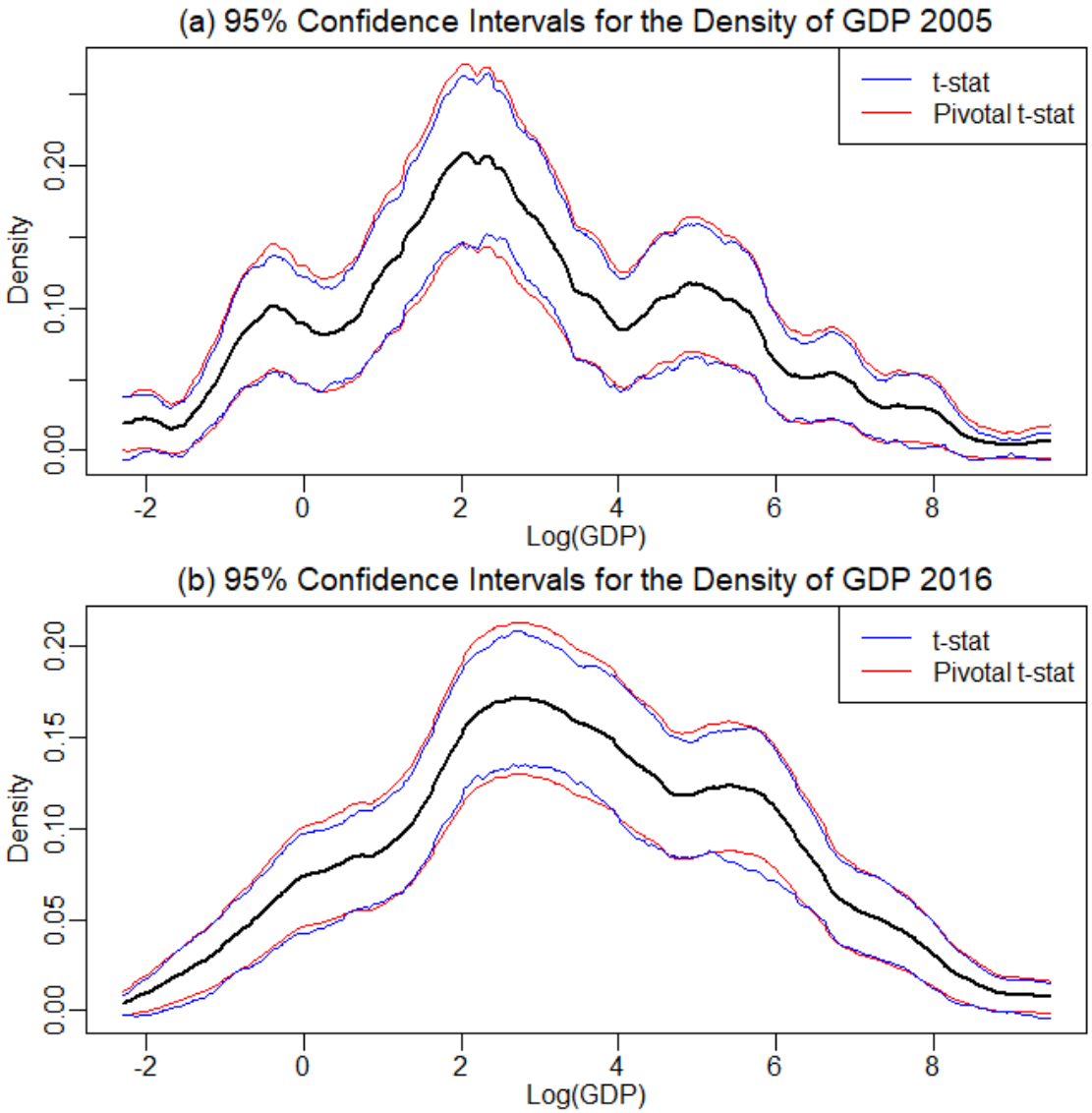

Figure 1: 95% confidence intervals for the densities of X_{05} and X_{16}

(b) Using bootstrap CIs to test hypotheses

The pivotal bootstrap CIs are used to test the null hypotheses visually. For null hypotheses (i)-(iv) in **Figure 2**, the null distributions heavily mismatch the CIs of the testing distributions. We reject all the null hypotheses. For the null hypothesis (v) $H_0 : X_{05} = X_{16}$, the visual evidence is less clear to make a conclusion. The distribution of X_{16} slightly shift to the right, which means that the GDP of countries is 2016 increases, comparing to 2005. However, it is not significantly out of the 2005 confidence interval. Thus, we fail to reject H_0.

3 Exercise 3

From the dataset `engel.dta`, with 235 observations on annual household income and annual household food expenditure. Take $Y = log(income)$, $X = log(expenditure)$. Outliers is eliminated, the histogram of X and Y is in **Figure 3**.

The model to estimate is:

$$Median[Y|X] = g(X)$$

(a) Bhattacharya and Gangopadhyay (1990) Estimator

Procedure

- Based on the histogram of X and Y, we choose the grid of $5.5 \leq x_0 \leq 7.5$ and $6 \leq y_0 \leq 8$.
- Choose an arbitrary bandwidth $h_n = 0.2$, estimate the kernel density estimator for X, $\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \mathbf{1}(|X_i - x_0| \leq h_n/2)$
- With $\alpha = 0.5$ and $b_n(x_0) = nh_n\hat{f}(x_0)$, we identify the effective α for each x_0 in the chosen grid:
$$\alpha_{effective}(x_0) = \frac{int[\alpha \cdot b_n(x_0)]}{b_n(x_0)}$$
- For each pair of (x_0, y_0) belongs to $(5.5 \leq x_0 \leq 7.5) \times (6 \leq y_0 \leq 8)$:
$$\hat{F}(y_0|x_0) = \frac{1}{nh_n\hat{f}(x)} \sum_{i=1}^{n} \mathbf{1}(Y_i \leq y_0)\mathbf{1}(|X_i - x_0| \leq h_n/2)$$
- Get fitted $\widehat{Median}[Y|X] = inf\left[y_0 : \hat{F}(y_0|x_0) \geq \alpha_{effective}(x_0)\right]$

(b) Quantile Regression by `npqreg` in `np` package

The estimate of $\widehat{Median}[Y|X]$ is also obtained by the `npqreg` functions. The estimate from *BG90* and *npqreg* will be plotted in **Figure 4**.

(c) Interpretations and Comments

Compare BG90 and npqreg plots

The estimates by $BG90$ is less smooth. The reason is the $BG90$ using the uniform kernel, $K(v) = \mathbf{1}(|v| \leq 1/2)$, rather than the smooth kernel. It is to avoid the CDF of higher quantiles will cross the CDF of lower quantiles, due to smoothing. Up to the tails, two estimates seems to be apart, as the nearest neighborhood approach might find less supporting points in this region. Nevertheless, the plots of two estimates are very close to each other.

Interpret conditional median regression

The results show the relation of median of Y condition on X. In particular, the relation of Annual income of household at median-level (50th-quantile) with the Annual Household Expenditure in Food. As the illustration in the plot, it is close to be linear. The income is higher for families spending more in food, which is intuitive and reasonable for the middle-income group.

However, the similar pattern is not necessary true for other quantiles. For example, in the 75th-quantile of income, households in this group could be considerably rich, the positive relation between income and food expenditure might become weaker, or even no effect is found. Food is necessity goods, that over a certain threshold, the marginal utility of food will be declined. The increase in income might not be associated with the increase in food expenditure. For that reason, the quantile regression is desirable to investigate different effects in different quantile groups.

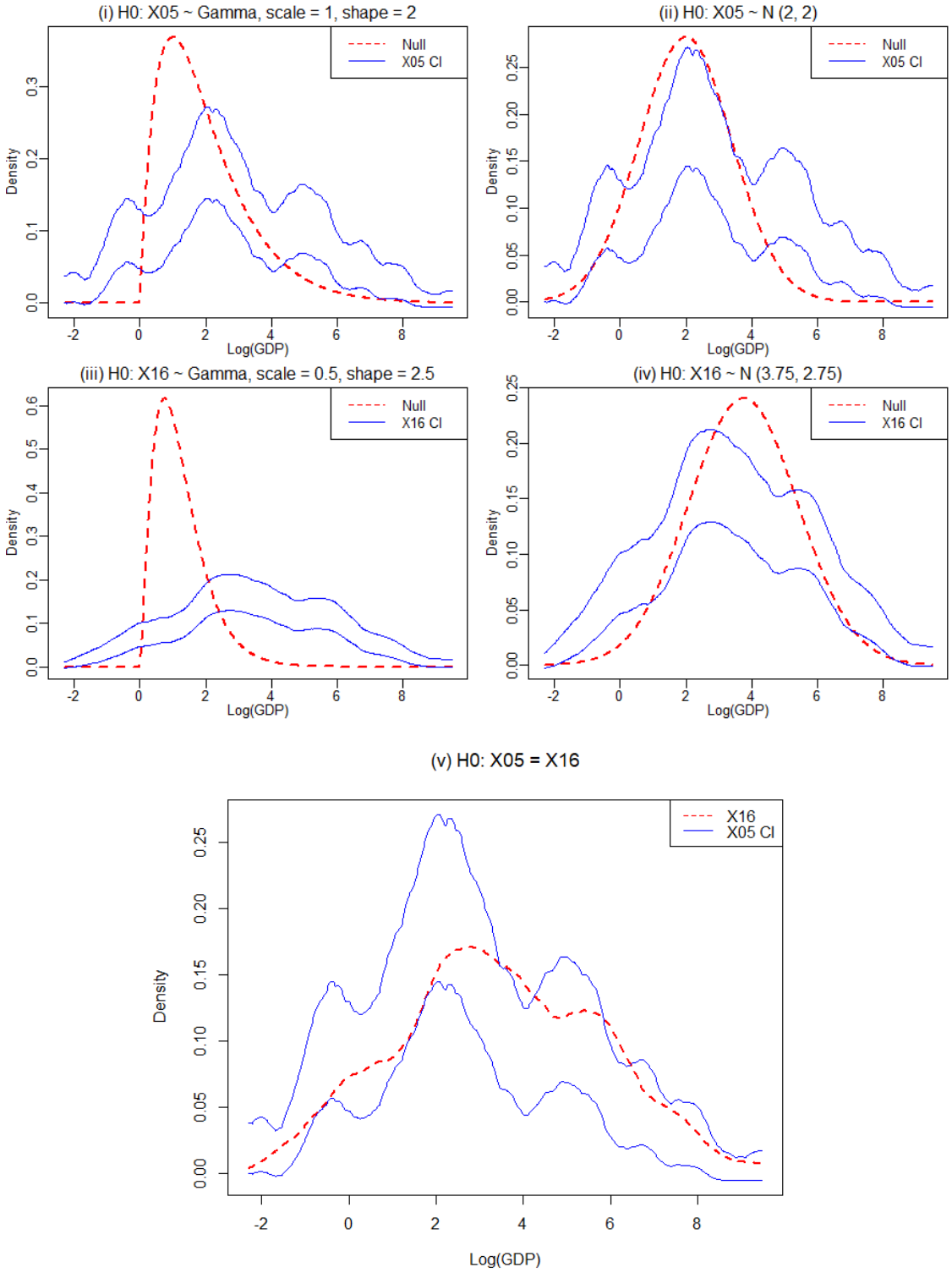

Figure 2: Testing Null Hypotheses by 95% bootstrap pivotal CI of the densities of X_{05} and X_{16}

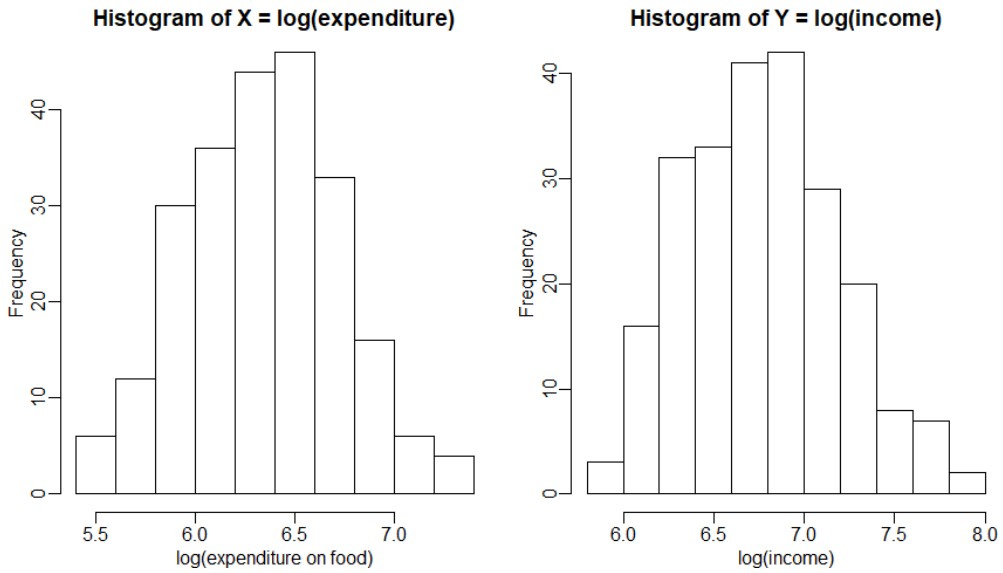

Figure 3: Histrogram of $log(income)$ and $log(expenditure)$

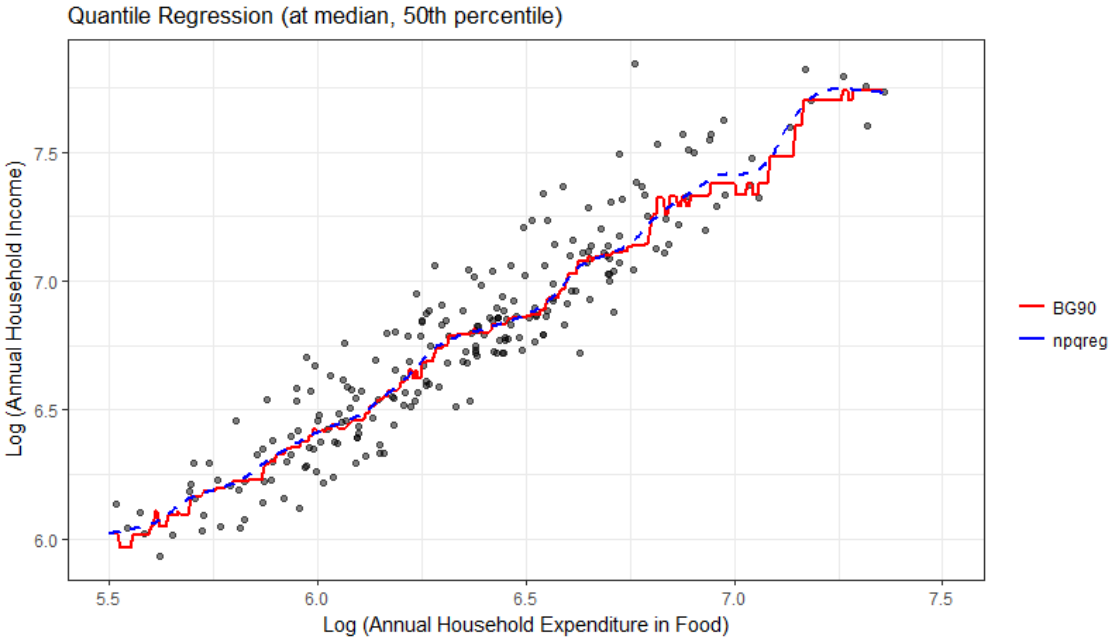

Figure 4: The Estimate of $Median[Y|X] = g(X)$