Non- and semiparametric tests for conditional independence in two-way contingency tables

Gery Geenens

Institut de Statistique
Université catholique de Louvain
Belgium

Young Researchers Day
February 1, 2008, Louvain-la-Neuve

A real data example: actuarial science

Data

- 14505 car insurance policies in a portfolio
- for each insured, we know the **sex** and the notified **claims** last year (1997)
- contingency table :

		sex		
		men	women	
claims	no	8324	4638	12962
	yes	1034	509	1543
		9358	5147	14505

Of interest

Is there an association between sex and accidents ?

A real data example: actuarial science

Data

- 14505 car insurance policies in a portfolio
- for each insured, we know the **sex** and the notified **claims** last year (1997)
- contingency table :

		sex		
		men	women	
claims	no	8324	4638	12962
	yes	1034	509	1543
		9358	5147	14505

Of interest

Is there an association between sex and accidents ?

A real data example: actuarial science

Chi-square test: $\chi^2 = 4.58$, df=1, p-value=0.03

$\Rightarrow$ rejection of independence hypothesis

A real data example: actuarial science

Chi-square test: $\chi^2 = 4.58$, df=1, p-value=0.03

$\Rightarrow$ rejection of independence hypothesis

Questions

- is this a *direct* association?
- if not, what are the factors which imply this association?
  ex.: **power** of the insured vehicle

A real data example: actuarial science

Chi-square test: $\chi^2 = 4.58$, df=1, p-value=0.03

$\Rightarrow$ rejection of independence hypothesis

Questions

- is this a *direct* association?
- if not, what are the factors which imply this association?
  ex.: **power** of the insured vehicle

A real data example: actuarial science

Chi-square test: $\chi^2 = 4.58$, df=1, p-value=0.03

$\Rightarrow$ rejection of independence hypothesis

Questions

- is this a *direct* association?
- if not, what are the factors which imply this association?
  ex.: **power** of the insured vehicle

Aim

Take into account, and remove if possible, the effect of other variables

A real data example: actuarial science

Chi-square test: $\chi^2 = 4.58$, df=1, p-value=0.03

$\Rightarrow$ rejection of independence hypothesis

Questions

- is this a *direct* association?
- if not, what are the factors which imply this association?
  ex.: **power** of the insured vehicle

Solution

Work **conditionally** to the other variables

A real data example: actuarial science

Chi-square test: $\chi^2 = 4.58$, df=1, p-value=0.03

$\Rightarrow$ rejection of independence hypothesis

Questions

- is this a *direct* association?
- if not, what are the factors which imply this association?
  ex.: **power** of the insured vehicle

Solution

Work **conditionally** to the other variables

$\Rightarrow$ need for **conditional independence** tests

Two-way contingency tables

- R and S two categorical variables, with r and s levels
- sample of n individuals
- n_{ij} =number of individuals for which $R = i$ and $S = j$
- $n_{i.} = \sum_j n_{ij}$ $\quad n_{.j} = \sum_i n_{ij}$ $\quad n_{..} = \sum_{i,j} n_{ij} = n$

Two-way contingency table

		S						
		1	$\cdots$	j	$\cdots$	s	total
	1	n_{11}		$\cdots$		n_{1s}	$n_{1.}$
R	i	$\vdots$		n_{ij}		$\vdots$	$\vdots$
	r	n_{r1}		$\cdots$		n_{rs}	$n_{r.}$
	total	$n_{.1}$		$\cdots$		$n_{.s}$	$n_{..} = n$

Usual chi-square test of independence

Cell probabilities

$$\pi_{ij} = \mathbb{P}(R = i, S = j)$$

Independence hypothesis

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall\,(i,j)$$

Asymptotics

$$U^2 \xrightarrow{H_0} \chi^2_{(r-1)(s-1)}$$

ML estimates

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \quad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

χ^2 divergence

$$U^2 = n \sum_{i,j} \frac{\left(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}\right)^2}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

Gap

implicitly assumed homogeneous population

Usual chi-square test of independence

Cell probabilities

$$\pi_{ij} = \mathbb{P}(R = i, S = j)$$

Independence hypothesis

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall\,(i,j)$$

Asymptotics

$$U^2 \xrightarrow{H_0} \chi^2_{(r-1)(s-1)}$$

ML estimates

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \quad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

χ^2 divergence

$$U^2 = n \sum_{i,j} \frac{\left(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}\right)^2}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

Gap

implicitly assumed homogeneous population

Usual chi-square test of independence

Cell probabilities

$$\pi_{ij} = \mathbb{P}(R = i, S = j)$$

Independence hypothesis

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall\,(i,j)$$

Asymptotics

$$U^2 \xrightarrow{H_0} \chi^2_{(r-1)(s-1)}$$

ML estimates

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \quad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

χ^2 divergence

$$U^2 = n \sum_{i,j} \frac{\left(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}\right)^2}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

Gap

implicitly assumed homogeneous population

Usual chi-square test of independence

Cell probabilities

$$\pi_{ij} = \mathbb{P}(R = i, S = j)$$

Independence hypothesis

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall\,(i,j)$$

Asymptotics

$$U^2 \xrightarrow{H_0} \chi^2_{(r-1)(s-1)}$$

ML estimates

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \quad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

χ^2 divergence

$$U^2 = n \sum_{i,j} \frac{\left(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}\right)^2}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

Gap

implicitly assumed homogeneous population

Usual chi-square test of independence

Cell probabilities

$$\pi_{ij} = \mathbb{P}(R = i, S = j)$$

Independence hypothesis

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall\,(i,j)$$

Asymptotics

$$U^2 \xrightarrow{H_0} \chi^2_{(r-1)(s-1)}$$

ML estimates

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \quad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

χ^2 divergence

$$U^2 = n \sum_{i,j} \frac{\left(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}\right)^2}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

Gap

implicitly assumed homogeneous population

Usual chi-square test of independence

Cell probabilities

$$\pi_{ij} = \mathbb{P}(R = i, S = j)$$

Independence hypothesis

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall\,(i,j)$$

Asymptotics

$$U^2 \xrightarrow{H_0} \chi^2_{(r-1)(s-1)}$$

ML estimates

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \quad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

χ^2 divergence

$$U^2 = n \sum_{i,j} \frac{\left(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}\right)^2}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

Gap

implicitly assumed homogeneous population

Conditional independence

Extra characteristics $X = (X^{(1)}, \ldots, X^{(p)})$

could be associated with R, S, or both

$\Rightarrow$ affect the whole dependence structure of the table

$\Rightarrow$ test **conditional** independence between R and S, **given** X

Conditional cell probabilities

$$\pi_{ij}(x) = \mathbb{P}(R = i, S = j | X = x)$$

Conditional independence hypothesis

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \quad \forall (i,j) \forall x \in S_X$$

$\Rightarrow$ testing for H_0 requires reliable estimates of

$$\pi(x) = (\pi_{11}(x), \pi_{12}(x), \ldots, \pi_{rs}(x))^t$$

Conditional independence

Extra characteristics $X = (X^{(1)}, \ldots, X^{(p)})$

could be associated with R, S, or both

$\Rightarrow$ affect the whole dependence structure of the table

$\Rightarrow$ test **conditional** independence between R and S, **given** X

Conditional cell probabilities

$$\pi_{ij}(x) = \mathbb{P}(R = i, S = j | X = x)$$

Conditional independence hypothesis

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \quad \forall (i,j) \forall x \in S_X$$

$\Rightarrow$ testing for H_0 requires reliable estimates of

$$\pi(x) = (\pi_{11}(x), \pi_{12}(x), \ldots, \pi_{rs}(x))^t$$

Conditional independence

Extra characteristics $X = (X^{(1)}, \ldots, X^{(p)})$

could be associated with R, S, or both

$\Rightarrow$ affect the whole dependence structure of the table

$\Rightarrow$ test **conditional** independence between R and S, **given** X

Conditional cell probabilities

$$\pi_{ij}(x) = \mathbb{P}(R = i, S = j | X = x)$$

Conditional independence hypothesis

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \quad \forall (i,j) \forall x \in S_X$$

$\Rightarrow$ testing for H_0 requires reliable estimates of

$$\pi(x) = (\pi_{11}(x), \pi_{12}(x), \ldots, \pi_{rs}(x))^t$$

Conditional independence

Extra characteristics $X = (X^{(1)}, \ldots, X^{(p)})$

could be associated with R, S, or both

$\Rightarrow$ affect the whole dependence structure of the table

$\Rightarrow$ test **conditional** independence between R and S, **given** X

Conditional cell probabilities

$$\pi_{ij}(x) = \mathbb{P}(R = i, S = j | X = x)$$

Conditional independence hypothesis

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \quad \forall (i,j) \forall x \in S_X$$

$\Rightarrow$ testing for H_0 requires reliable estimates of

$$\pi(x) = (\pi_{11}(x), \pi_{12}(x), \ldots, \pi_{rs}(x))^t$$

Conditional independence

Extra characteristics $X = (X^{(1)}, \ldots, X^{(p)})$

could be associated with R, S, or both

$\Rightarrow$ affect the whole dependence structure of the table

$\Rightarrow$ test **conditional** independence between R and S, **given** X

Conditional cell probabilities

$$\pi_{ij}(x) = \mathbb{P}(R = i, S = j | X = x)$$

Conditional independence hypothesis

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \quad \forall (i,j) \forall x \in S_X$$

$\Rightarrow$ testing for H_0 requires reliable estimates of

$$\pi(x) = (\pi_{11}(x), \pi_{12}(x), \ldots, \pi_{rs}(x))^t$$

Conditional independence

Extra characteristics $X = (X^{(1)}, \ldots, X^{(p)})$

could be associated with R, S, or both

$\Rightarrow$ affect the whole dependence structure of the table

$\Rightarrow$ test **conditional** independence between R and S, **given** X

Conditional cell probabilities

$$\pi_{ij}(x) = \mathbb{P}(R = i, S = j | X = x)$$

Conditional independence hypothesis

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \quad \forall\,(i,j)\,\forall x \in S_X$$

$\Rightarrow$ testing for H_0 requires reliable estimates of

$$\pi(x) = (\pi_{11}(x), \pi_{12}(x), \ldots, \pi_{rs}(x))^t$$

Conditional probabilities as regression functions

Define the random vector $Z = (Z^{(11)}, Z^{(12)}, \ldots, Z^{(rs)})^t$, with

$$Z^{(ij)} = \mathbf{1}(R = i, S = j)$$

Regression functions

$$\pi_{ij}(x) = \mathbb{E}\left(Z^{(ij)} | X = x\right)$$

$\Rightarrow$ use of regression methods to estimate $\pi(x) = \mathbb{E}(Z|X = x)$

Data

$$\{(X_k, Z_k)\}_{k=1}^n \in S_X \times \{z \in \{0,1\}^{rs} : \textstyle\sum_q z^{(q)} = 1\}$$

Conditional probabilities as regression functions

Define the random vector $Z = (Z^{(11)}, Z^{(12)}, \ldots, Z^{(rs)})^t$, with

$$Z^{(ij)} = \mathbf{1}(R = i, S = j)$$

Regression functions

$$\pi_{ij}(x) = \mathbb{E}\left(Z^{(ij)} | X = x\right)$$

$\Rightarrow$ use of regression methods to estimate $\pi(x) = \mathbb{E}(Z|X = x)$

Data

$$\{(X_k, Z_k)\}_{k=1}^n \in S_X \times \{z \in \{0,1\}^{rs} : \textstyle\sum_q z^{(q)} = 1\}$$

Conditional probabilities as regression functions

Define the random vector $Z = (Z^{(11)}, Z^{(12)}, \ldots, Z^{(rs)})^t$, with

$$Z^{(ij)} = \mathbf{1}(R = i, S = j)$$

Regression functions

$$\pi_{ij}(x) = \mathbb{E}\left(Z^{(ij)} | X = x\right)$$

$\Rightarrow$ use of regression methods to estimate $\pi(x) = \mathbb{E}(Z|X = x)$

Data

$$\{(X_k, Z_k)\}_{k=1}^n \in S_X \times \{z \in \{0,1\}^{rs} : \textstyle\sum_q z^{(q)} = 1\}$$

Nadaraya-Watson estimator of $\pi(x)$ ($p = 1$)

Let K be a kernel function and h a bandwidth

NW estimator of $\pi_{ij}(x)$

$$\hat{p}_{ij}(x) = \frac{\sum_{k=1}^{n} K_h(x - X_k) Z_k^{(ij)}}{\sum_{k=1}^{n} K_h(x - X_k)}$$

Nadaraya-Watson estimator of $\pi(x)$ ($p = 1$)

Let K be a kernel function and h a bandwidth

NW estimator of $\pi_{ij}(x)$

$$\hat{p}_{ij}(x) = \frac{\sum_{k=1}^{n} K_h(x - X_k) Z_k^{(ij)}}{\sum_{k=1}^{n} K_h(x - X_k)}$$

NW estimator of $\pi(x)$

$$\hat{p}(x) = \mathcal{Z}\mathcal{K}_h(x)$$

Nadaraya-Watson estimator of $\pi(x)$ ($p = 1$)

NW estimator of $\pi(x)$

$$\hat{p}(x) = \mathcal{Z}\mathcal{K}_h(x)$$

with

$$\mathcal{K}_h(x) = \left(\frac{K_h(x - X_1)}{\sum_{k=1}^{n} K_h(x - X_k)}, \ldots, \frac{K_h(x - X_n)}{\sum_{k=1}^{n} K_h(x - X_k)} \right)^t$$

and

$$\mathcal{Z} = \begin{pmatrix} Z_1^{(11)} & Z_2^{(11)} & \cdots & Z_n^{(11)} \\ Z_1^{(12)} & Z_2^{(12)} & & \\ \vdots & & \ddots & \vdots \\ Z_1^{(rs)} & & \cdots & Z_n^{(rs)} \end{pmatrix}$$

Nadaraya-Watson estimator of $\pi(x)$ ($p = 1$)

NW estimator of $\pi(x)$

$$\hat{p}(x) = \mathcal{Z}\mathcal{K}_h(x)$$

Remark

- weighted average $\Rightarrow$ $\hat{p}_{ij}(x) \in [0, 1] \quad \forall x, \forall (i, j)$
- if common $h \, \forall (i, j)$, then automatically

$$\sum_{ij} \hat{p}_{ij}(x) = 1 \qquad \forall x$$

$\Rightarrow$ natural way of doing (adaptive to the setting)

Nadaraya-Watson estimator of $\pi(x)$ ($p = 1$)

NW estimator of $\pi(x)$

$$\hat{p}(x) = \mathcal{Z}\mathcal{K}_h(x)$$

Asymptotic properties

Usual asymptotic properties of NW and multinomial sampling:

$$\sqrt{nh}\,(\hat{p}(x) - \pi(x) - b(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\nu_0}{f_X(x)}\left(\text{diag}(\pi(x)) - \pi(x)\pi(x)^t\right)\right)$$

with $\nu_0 = \int K^2(x)dx$ and $b(x) = O(h^2)$

Nadaraya-Watson estimator of $\pi(x)$ ($p = 1$)

NW estimator of $\pi(x)$

$$\hat{p}(x) = \mathcal{Z}\mathcal{K}_h(x)$$

Asymptotic properties

Usual asymptotic properties of NW and multinomial sampling:

$$\sqrt{nh}\,(\hat{p}(x) - \pi(x) - b(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\nu_0}{f_X(x)}\left(\mathrm{diag}(\pi(x)) - \pi(x)\pi(x)^t\right)\right)$$

with $\nu_0 = \int K^2(x)dx$ and $b(x) = O(h^2)$

Undersmoothing

$$h = o(n^{-1/5}) \Rightarrow (nh)^{1/2} b(x) = o(1)$$

Nadaraya-Watson estimator of $\pi(x)$ ($p = 1$)

NW estimator of $\pi(x)$

$$\hat{p}(x) = \mathcal{Z}\mathcal{K}_h(x)$$

Asymptotic properties

Usual asymptotic properties of NW and multinomial sampling:

$$\sqrt{nh}\,(\hat{p}(x) - \pi(x)) + o(1) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\nu_0}{f_X(x)}\left(\mathrm{diag}(\pi(x)) - \pi(x)\pi(x)^t\right)\right)$$

with $\nu_0 = \int K^2(x)dx$ and $b(x) = O(h^2)$

Undersmoothing

$$h = o(n^{-1/5}) \Rightarrow (nh)^{1/2} b(x) = o(1)$$

Testing for conditional independence ($p = 1$)

To test

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \qquad \forall (i,j), \forall x \in S_X$$

Testing for conditional independence ($p = 1$)

To test

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \qquad \forall (i,j), \forall x \in S_X$$

Idea

Base the test statistic on the integrated χ^2-divergence between $\hat{p}_{ij}(x)$ and $\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)$

Testing for conditional independence ($p = 1$)

Idea

Base the test statistic on the integrated χ^2-divergence between $\hat{p}_{ij}(x)$ and $\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)$

pointwise χ^2-divergence for x fixed

$$V^2(x) = \frac{nh\hat{f}(x)}{\nu_0} \sum_{i,j} \frac{\left(\hat{p}_{ij}(x) - \hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)\right)^2}{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}$$

Testing for conditional independence ($p = 1$)

Idea

Base the test statistic on the integrated χ^2-divergence between $\hat{p}_{ij}(x)$ and $\hat{p}_{i.}(x)\hat{p}_{.j}(x)$

pointwise χ^2-divergence for x fixed

$$V^2(x) = \frac{nh\hat{f}(x)}{\nu_0} \sum_{i,j} \frac{\left(\hat{p}_{ij}(x) - \hat{p}_{i.}(x)\hat{p}_{.j}(x)\right)^2}{\hat{p}_{i.}(x)\hat{p}_{.j}(x)}$$

Asymptotics

$$V^2(x) \xrightarrow{H_0} \chi^2_{(r-1)(s-1)} \qquad \forall x \in S_X$$

Testing for conditional independence ($p = 1$)

Integrated χ^2-divergence

$$\int V^2(x) f(x) dx = \mathbb{E}(V^2(X)) \simeq \frac{1}{n} \sum_{k=1}^{n} V^2(X_k) \doteq V^2$$

Testing for conditional independence ($p = 1$)

Integrated χ^2-divergence

$$\int V^2(x) f(x) dx = \mathbb{E}(V^2(X)) \simeq \frac{1}{n} \sum_{k=1}^{n} V^2(X_k) \doteq V^2$$

Testing for conditional independence ($p = 1$)

Integrated χ^2-divergence

$$\int V^2(x) f(x) dx = \mathbb{E}(V^2(X)) \simeq \frac{1}{n} \sum_{k=1}^{n} V^2(X_k) \doteq V^2$$

Theorem

Under mild conditions, if $h \sim n^{-\beta}$, $\beta \in]2/9, 1/2[$,

$$\frac{1}{\sqrt{h}} \frac{\nu_0}{\sqrt{2(r-1)(s-1)\phi_0 N_0}} (V^2 - (r-1)(s-1)) \xrightarrow{H_0} \mathcal{N}(0,1)$$

with $\nu_0(u) = (K * K)(u)$, $N_0 = \int \nu_0^2(u) du$ and $\phi_0 = \int f_X^2(x) dx$

Testing for conditional independence ($p = 1$)

Integrated χ^2-divergence

$$\int V^2(x) f(x) dx = \mathbb{E}(V^2(X)) \simeq \frac{1}{n} \sum_{k=1}^{n} V^2(X_k) \doteq V^2$$

Theorem

Under mild conditions, if $h \sim n^{-\beta}$, $\beta \in]2/9, 1/2[$,

$$\frac{1}{\sqrt{h}} \frac{\nu_0}{\sqrt{2(r-1)(s-1)\phi_0 N_0}} (V^2 - (r-1)(s-1)) \xrightarrow{H_0} \mathcal{N}(0,1)$$

with $\nu_0(u) = (K * K)(u)$, $N_0 = \int \nu_0^2(u) du$ and $\phi_0 = \int f_X^2(x) dx$

Testing for conditional independence ($p = 1$)

Integrated χ^2-divergence

$$\int V^2(x) f(x) dx = \mathbb{E}(V^2(X)) \simeq \frac{1}{n} \sum_{k=1}^{n} V^2(X_k) \doteq V^2$$

Theorem

Under mild conditions, if $h \sim n^{-\beta}$, $\beta \in]2/9, 1/2[$,

$$\frac{1}{\sqrt{h}} \frac{\nu_0}{\sqrt{2(r-1)(s-1)\phi_0 N_0}} (V^2 - (r-1)(s-1)) \xrightarrow{H_0} \mathcal{N}(0,1)$$

with $\nu_0(u) = (K * K)(u)$, $N_0 = \int \nu_0^2(u) du$ and $\phi_0 = \int f_X^2(x) dx$

Testing for conditional independence ($p = 1$)

Theorem

Under mild conditions, if $h \sim n^{-\beta}$, $\beta \in]2/9, 1/2[$,

$$\frac{1}{\sqrt{h}} \frac{\nu_0}{\sqrt{2(r-1)(s-1)\phi_0 N_0}} (V^2 - (r-1)(s-1)) \xrightarrow{H_0} \mathcal{N}(0,1)$$

Asymptotic rejection criterion of level α

$$V^2 > (r-1)(s-1) + z_{1-\alpha} \frac{\sqrt{h}}{\nu_0} \sqrt{2(r-1)(s-1)\hat{\phi}_0 N_0}$$

Simulation scenario

Scenario

$r = s = 2, \quad p = 1, \quad X \sim U_{[-2,2]}$

Marginals and dependence

- $\pi_{1.}(x) = \exp(-x^2), \quad \pi_{2.}(x) = 1 - \pi_{1.}(x)$
- $\pi_{.1}(x) = \frac{\exp(-x)}{1+\exp(-x)}, \quad \pi_{.2}(x) = 1 - \pi_{.1}(x)$
- $\pi_{ij}(x) = \pi_{i.}(x)\pi_{.j}(x) + (2\delta_{ij} - 1)\gamma\Delta(x) \qquad \forall(i,j)$
- $\gamma = 0$ (independence), $\gamma = 0.1, 0.3, 0.5, 1$

Simulations

- $n = 50, 100, 500, 1000$
- 500 Monte-Carlo replications in each case

Simulation results

Results

$\alpha = 0.05$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
$\gamma = 0$	0.072	0.066	0.058	0.054
$\gamma = 0.1$	0.098	0.102	0.144	0.222
$\gamma = 0.3$	0.132	0.222	0.760	0.968
$\gamma = 0.5$	0.336	0.534	0.996	1
$\gamma = 1$	0.858	0.998	1	1

$\Rightarrow$ level and power satisfactory

Back to example

Test for the conditional independence between sex and claims, given the power of the insured vehicle

$$V^2 = 0.79, \quad p\text{-value} = 0.21 \Rightarrow \text{no reject of } H_0$$

Conclusion

When driving a vehicle of the **same power**, men and women are exposed to the **same risk**

Back to example

Test for the conditional independence between sex and claims, given the power of the insured vehicle

$$V^2 = 0.79, \quad p\text{-value} = 0.21 \Rightarrow \text{no reject of } H_0$$

Conclusion

When driving a vehicle of the **same power**, men and women are exposed to the **same risk**

The multivariate case ($p > 1$)

Theorem

Under mild conditions, if $h \sim n^{-\beta}$, $\beta \in]\frac{2}{p+8}, \frac{1}{2p}[$,

$$\frac{1}{\sqrt{h^p}} \frac{\nu_0 \sqrt{\Gamma(p/2)}}{\sqrt{4(r-1)(s-1)\pi^{p/2}\phi_0 N_0}} (V^2 - (r-1)(s-1)) \xrightarrow{H_0} \mathcal{N}(0,1)$$

with $\nu_0(u) = (K * K)(u)$, $N_0 = \int_0^2 u^{p-1}\nu_0^2(u)du$ and $\phi_0 = \int f_X^2(x)dx$

Motivation for a semiparametric procedure

Gaps of the nonparametric procedure

1. "curse of dimensionality"
2. discrete covariates not allowed

$\Rightarrow$ develop of a semiparametric procedure

$\Rightarrow$ Single-Index Models idea

Motivation for a semiparametric procedure

Gaps of the nonparametric procedure

1. "curse of dimensionality"
2. discrete covariates not allowed

$\Rightarrow$ develop of a semiparametric procedure

$\Rightarrow$ Single-Index Models idea

Motivation for a semiparametric procedure

Gaps of the nonparametric procedure

1. "curse of dimensionality"
2. discrete covariates not allowed

$\Rightarrow$ develop of a semiparametric procedure

$\Rightarrow$ Single-Index Models idea

Motivation for a semiparametric procedure

Gaps of the nonparametric procedure

1. "curse of dimensionality"
2. discrete covariates not allowed

$\Rightarrow$ develop of a semiparametric procedure

$\Rightarrow$ Single-Index Models idea

Single-Index assumption

SIM assumption

There exist $\theta_0 \in \Theta \subset \{\theta \in \mathbb{R}^p : \theta^{(1)} = 1\}$ and rs functions $g_{ij} : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, r$, $j = 1, \ldots, s$ such that

$$\pi_{ij}(x) = g_{ij}(\theta_0^t x) \qquad \forall\, x \in S_X,\ \forall\, (i,j).$$

Single-Index assumption

SIM assumption

There exist $\theta_0 \in \Theta \subset \{\theta \in \mathbb{R}^p : \theta^{(1)} = 1\}$ and rs functions $g_{ij} : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, r$, $j = 1, \ldots, s$ such that

$$\pi_{ij}(x) = g_{ij}(\theta_0^t x) \qquad \forall\, x \in S_X,\ \forall\, (i,j).$$

Interpretation

The set of covariates X influences the joint distribution of R and S only through the index $\theta_0^t X$

Single-Index assumption

SIM assumption

There exist $\theta_0 \in \Theta \subset \{\theta \in \mathbb{R}^p : \theta^{(1)} = 1\}$ and rs functions $g_{ij} : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, r$, $j = 1, \ldots, s$ such that

$$\pi_{ij}(x) = g_{ij}(\theta_0^t x) \qquad \forall\, x \in S_X,\ \forall\, (i,j).$$

Twolfold estimation

1. index coefficients vector θ_0
2. univariate link functions $\{g_{ij}\}$

Single-Index assumption

SIM assumption

There exist $\theta_0 \in \Theta \subset \{\theta \in \mathbb{R}^p : \theta^{(1)} = 1\}$ and rs functions $g_{ij} : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, r$, $j = 1, \ldots, s$ such that

$$\pi_{ij}(x) = g_{ij}(\theta_0^t x) \qquad \forall\, x \in S_X,\ \forall\, (i,j).$$

Twolfold estimation

1. index coefficients vector θ_0
2. univariate link functions $\{g_{ij}\}$: no course of dimensionality

Estimation of the index coefficients vector θ_0

Idea

Adaptation of the most popular estimation methods of the index in classical SIM

Classical SIM

one unknown link function

SI Conditional probabilities

rs unknown link functions, with $\sum_{ij} g_{ij} \equiv 1$

Studied estimators

- Semiparametric Maximum Likelihood estimator (SML)
- Semiparametric Least Squares estimator (SLS)
- Average Derivatives estimator (ADE)
- Sliced Inverse Regression estimator (SIR)

Estimation of the index coefficients vector θ_0

Idea

Adaptation of the most popular estimation methods of the index in classical SIM

Classical SIM

one unknown link function

SI Conditional probabilities

rs unknown link functions, with $\sum_{ij} g_{ij} \equiv 1$

Studied estimators

- Semiparametric Maximum Likelihood estimator (SML)
- Semiparametric Least Squares estimator (SLS)
- Average Derivatives estimator (ADE)
- Sliced Inverse Regression estimator (SIR)

Estimation of the index coefficients vector θ_0

Idea

Adaptation of the most popular estimation methods of the index in classical SIM

Classical SIM

one unknown link function

SI Conditional probabilities

rs unknown link functions, with $\sum_{ij} g_{ij} \equiv 1$

Studied estimators

- Semiparametric Maximum Likelihood estimator (SML)
- Semiparametric Least Squares estimator (SLS)
- Average Derivatives estimator (ADE)
- Sliced Inverse Regression estimator (SIR)

Estimation of θ_0: theoretical results

Under appropriate conditions,

SML

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SML})$$

SLS

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SLS})$$

ADE

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{ADE})$$

SIR

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SIR})$$

Remarks

1. parametric rates of convergence
2. $\Sigma_{SML} = \Sigma_{SLS} \simeq$ semiparametric efficiency bound
3. in practice, SLS estimator gives the best results

Estimation of θ_0: theoretical results

Under appropriate conditions,

SML

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SML})$$

SLS

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SLS})$$

ADE

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{ADE})$$

SIR

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SIR})$$

Remarks

1. parametric rates of convergence
2. $\Sigma_{SML} = \Sigma_{SLS} \simeq$ semiparametric efficiency bound
3. in practice, SLS estimator gives the best results

Estimation of θ_0: theoretical results

Under appropriate conditions,

SML

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SML})$$

SLS

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SLS})$$

ADE

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{ADE})$$

SIR

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SIR})$$

Remarks

1. parametric rates of convergence
2. $\Sigma_{SML} = \Sigma_{SLS} \simeq$ semiparametric efficiency bound
3. in practice, SLS estimator gives the best results

Estimation of θ_0: theoretical results

Under appropriate conditions,

SML

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SML})$$

SLS

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SLS})$$

ADE

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{ADE})$$

SIR

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{SIR})$$

Remarks

1. parametric rates of convergence
2. $\Sigma_{SML} = \Sigma_{SLS} \simeq$ semiparametric efficiency bound
3. in practice, SLS estimator gives the best results

Estimation of θ_0: Semiparametric Least Squares

Semiparametric Least Squares Estimator

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_k (Z_k - \hat{g}^\theta(\theta^t X_k))^t (Z_k - \hat{g}^\theta(\theta^t X_k))$$

with

$$\hat{g}^\theta(\theta^t X_k) = \frac{\sum_{k' \neq k} Z_{k'} K_1\left(\frac{\theta^t X_{k'} - \theta^t X_k}{h_1}\right)}{\sum_{k' \neq k} K_1\left(\frac{\theta^t X_{k'} - \theta^t X_k}{h_1}\right)}$$

Estimation of the link functions

NW estimators of the link functions

From any root-n estimator $\hat{\theta}$,

$$\hat{g}_{ij}^{\hat{\theta}}(u) = \frac{\sum_k K_h(u - \hat{\theta}^t X_k) Z_k^{(ij)}}{\sum_k K_h(u - \hat{\theta}^t X_k)}$$

Estimation of the link functions

NW estimators of the link functions

From any root-n estimator $\hat{\theta}$,

$$\hat{g}_{ij}^{\hat{\theta}}(u) = \frac{\sum_k K_h(u - \hat{\theta}^t X_k) Z_k^{(ij)}}{\sum_k K_h(u - \hat{\theta}^t X_k)}$$

Remark

No effect of the estimation of θ_0 on the asymptotic distribution of $\hat{g}_{ij}$:

$$(nh)^{1/2}\left(\hat{g}_{ij}^{\hat{\theta}}(u) - g_{ij}(u)\right) = (nh)^{1/2}\left(\hat{g}_{ij}^{\theta_0}(u) - g_{ij}(u)\right) + o_P(1)$$

Estimation of the link functions

NW estimators of the link functions

From any root-n estimator $\hat{\theta}$,

$$\hat{g}_{ij}^{\hat{\theta}}(u) = \frac{\sum_k K_h(u - \hat{\theta}^t X_k) Z_k^{(ij)}}{\sum_k K_h(u - \hat{\theta}^t X_k)}$$

Asymptotic properties

If $h = o(n^{-1/5})$, if SI assumption holds,

$$\sqrt{nh}\left(\hat{g}^{\hat{\theta}}(\hat{\theta}^t x) - \pi(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\nu_0}{f_X(x)}\left(\text{diag}(\pi(x)) - \pi(x)\pi(x)^t\right)\right)$$

Consequence

Consequence

If the SI assumption holds, the conditional independence test can be performed as such from $\hat{g}^{\hat{\theta}}$, as p was equal to 1

Consequence

Consequence

If the SI assumption holds, the conditional independence test can be performed as such from $\hat{g}^{\hat{\theta}}$, as p was equal to 1

Advantages

1. no matter the dimension of X, univariate rate of convergence
2. if there is at least one continuous covariate, $\theta_0^t X$ is continuously distributed
   $\Rightarrow$ some covariates could be discrete (dummy variables)

Consequence

Consequence

If the SI assumption holds, the conditional independence test can be performed as such from $\hat{g}^{\hat{\theta}}$, as p was equal to 1

Advantages

1. no matter the dimension of X, univariate rate of convergence
2. if there is at least one continuous covariate, $\theta_0^t X$ is continuously distributed
   $\Rightarrow$ some covariates could be discrete (dummy variables)

Testing the Single-Index assumption

To test

$H_0 : \exists \theta_0 \in \Theta, g : \mathbb{R} \to [0,1]^{rs} : \pi(x) = g(\theta_0^t x) \quad \forall x \in S_X$

vs.

$H_1 : \exists x \in S_X : \pi(x) \neq g(\theta_0^t x) \quad \forall \theta_0 \in \Theta, \forall g : \mathbb{R} \to [0,1]^{rs}$

Idea

Under H_0, for any positive bounded weight function w,

$$\mathbb{E}\left(w(X)(\pi(X) - g(\theta_0^t X))^t(\pi(X) - g(\theta_0^t X))\right) = 0$$

Example: $w(x) = f_0^2(\theta_0^t x) f(x)$

Testing the Single-Index assumption

To test

$H_0 : \exists \theta_0 \in \Theta, g : \mathbb{R} \to [0,1]^{rs} : \pi(x) = g(\theta_0^t x) \quad \forall x \in S_X$

vs.

$H_1 : \exists x \in S_X : \pi(x) \neq g(\theta_0^t x) \quad \forall \theta_0 \in \Theta, \forall g : \mathbb{R} \to [0,1]^{rs}$

Idea

Under H_0, for any positive bounded weight function w,

$$\mathbb{E}\left(w(X)(\pi(X) - g(\theta_0^t X))^t(\pi(X) - g(\theta_0^t X))\right) = 0$$

Example: $w(x) = f_0^2(\theta_0^t x) f(x)$

Testing the Single-Index assumption

To test

$H_0 : \exists \theta_0 \in \Theta, g : \mathbb{R} \to [0,1]^{rs} : \pi(x) = g(\theta_0^t x) \quad \forall x \in S_X$

vs.

$H_1 : \exists x \in S_X : \pi(x) \neq g(\theta_0^t x) \quad \forall \theta_0 \in \Theta, \forall g : \mathbb{R} \to [0,1]^{rs}$

Idea

Under H_0, for any positive bounded weight function w,

$$\mathbb{E}\left(w(X)(\pi(X) - g(\theta_0^t X))^t(\pi(X) - g(\theta_0^t X))\right) = 0$$

Example: $w(x) = f_0^2(\theta_0^t x) f(x)$

Testing the Single-Index assumption

Test statistic

$$T_n = \frac{1}{C_n^4} \sum (Z_k - Z_{k'})^t (Z_{k''} - Z_{k'''}) K_{h_1}\left(\hat{\theta}^t X_k - \hat{\theta}^t X_{k'}\right) \times K_{h_1}\left(\hat{\theta}^t X_{k''} - \hat{\theta}^t X_{k'''}\right) L_{h_2}(X_k - X_{k''})$$

with K a univariate kernel, L a p-variate kernel, h_1 and h_2 some bandwidths, and $\hat{\theta}$ the SLS estimator of θ_0

Theorem

Under appropriate conditions, under H_0,

$$n h_2^{p/2} T_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \omega^2)$$

with $\omega^2 = 2\nu_0 \mathbb{E}\left((1 - g(\theta_0^t X)^t g(\theta_0^t X)) f_0^4(\theta_0^t X) f(X)\right)$

Testing the Single-Index assumption

Test statistic

$$T_n = \frac{1}{C_n^4} \sum (Z_k - Z_{k'})^t (Z_{k''} - Z_{k'''}) K_{h_1}\left(\hat{\theta}^t X_k - \hat{\theta}^t X_{k'}\right) \times K_{h_1}\left(\hat{\theta}^t X_{k''} - \hat{\theta}^t X_{k'''}\right) L_{h_2}(X_k - X_{k''})$$

with K a univariate kernel, L a p-variate kernel, h_1 and h_2 some bandwidths, and $\hat{\theta}$ the SLS estimator of θ_0

Theorem

Under appropriate conditions, under H_0,

$$n h_2^{p/2} T_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \omega^2)$$

with $\omega^2 = 2\nu_0 \mathbb{E}\left((1 - g(\theta_0^t X)^t g(\theta_0^t X)) f_0^4(\theta_0^t X) f(X)\right)$

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)

Concluding remarks

- independence test between two categorical variables, given a set of extra covariates
- integrated χ^2 divergence between estimated joint conditional distribution and product of estimated conditional marginal distributions
- conditional distributions are nonparametrically estimated
- semiparametric Single-Index assumption to avoid curse of dimensionality

$\Rightarrow$ adaptation of the most popular estimation methods in classical SIM and test for this assumption

- asymptotically normal test statistic
- good behaviour in practice (level, power)