Bài giảng 8c: Phân tích bằng biểu đồ hộp và tương quan

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Ton Duc Thang University, Vietnam

Dữ liệu PISA (Schools)

```
setwd("~/Dropbox/World Bank 2014/Data for 2015
  workshop")

# pisa = read.csv("~/Dropbox/World Bank 2014/Data for
  2015 workshop/PISA DATA.csv", header=T)

sc = read.csv("~/Dropbox/World Bank 2014/Data for 2015
  workshop/SCHOOL DATA (VN).csv", header=T)

attach(sc)
```

REGION	TYPE	AREA	SCHOOLID	SCHSIZE	SC09Q11	SC03Q01	SC04Q01	SC05Q01	CLSIZE	COMPWEB	PCGIRLS	SCMATEDU	SMRATIO	STRATIO
CENTRAL	PUBLIC	URBAN	1	1804	93	Small Town	One Other	>50	53		0.557	-1.053	120.267	18.89
NORTH	PUBLIC	URBAN	2	1586	84	Town	Two or More	36-40	38	1	0.505	-0.5214	105.733	18.23
SOUTH	PUBLIC	RURAL	3	604	32	Village	No Others	41-45	43	0	0.533	-1.962	151	18.585
SOUTH	PUBLIC	URBAN	4	568	99	Small Town	No Others	36-40	38	1	0.586	-1.2473	35.5	5.737
CENTRAL	PUBLIC	URBAN	5	1078	65	Small Town	One Other	41-45	43	1	0.552	0.224	98	16.211
NORTH	PUBLIC	URBAN	6	1232	37	Small Town	Two or More	41-45	43	1	0.594	0.0288		33.297
NORTH	PRIVATE	URBAN	7	1280	48	Small Town	Two or More	46-50	48	0.096	0.453	-1.6925	182.857	26.667
NORTH	PUBLIC	RURAL	8	1379	57	Village	One Other	31-35	33	1	0.558	-1.4581	125.364	24.193
SOUTH	PUBLIC	URBAN	9	2332	114	City	Two or More	>50	53	1	0.579	-0.1702	137.176	20.103
NORTH	PUBLIC	URBAN	10	1051	100	City	One Other	16-20	28		0.669	-1.2473	61.824	10.304
CENTRAL	PUBLIC	RURAL	11	150	16	Village	Two or More	<16	13		0.467	0.7524	50	9.375
CENTRAL	PUBLIC	URBAN	12	1872	106	City	Two or More	>50	53	1	0.599	0.224	110.118	17.66
CENTRAL	PUBLIC	URBAN	13	890	51	Village	No Others	31-35	33	1	0.491	-0.5214	84.762	16.036
SOUTH	PUBLIC	RURAL	14	634	43	Village	Two or More	36-40	38		0.525	-0.6941	105.667	14.744
SOUTH	PUBLIC	URBAN	15	2890	79	Town	One Other	41-45	43	1	0.376	-0.1702	231.2	35.46
CENTRAL	PUBLIC	URBAN	16	1335	69	Village	Two or More	41-45	43	1	0.55	-1.962	111.25	19.209
NORTH	PUBLIC	RURAL	17	1552	70	Village	One Other	>50	53	0.227	0.541	-1.053	91.294	20.832
NORTH	PUBLIC	URBAN	18	896	81	City	Two or More	>50	53	0.455	0.623	-0.1702	61.793	10.667
NORTH	PUBLIC	REMOTE	19	1243	69	Village	No Others	36-40	38	1	0.558	-0.3482	138.111	17.264
SOUTH	PUBLIC	URBAN	20	1737	48	City	Two or More	>50	53	0.3	0.472	0.018	217.125	32.774
NORTH	PUBLIC	RURAL	21	632	39	Village	No Others	41-45	43	0.5	0.486	0.224	90.286	16
NORTH	PUBLIC	RURAL	22	1400	70	Village	Two or More	41-45	43	1	0.511	0.4606	140	20
CENTRAL	PUBLIC	RURAL	23	316	23	Village	No Others	<16	13	1	0.465	-0.1702	79	12.64

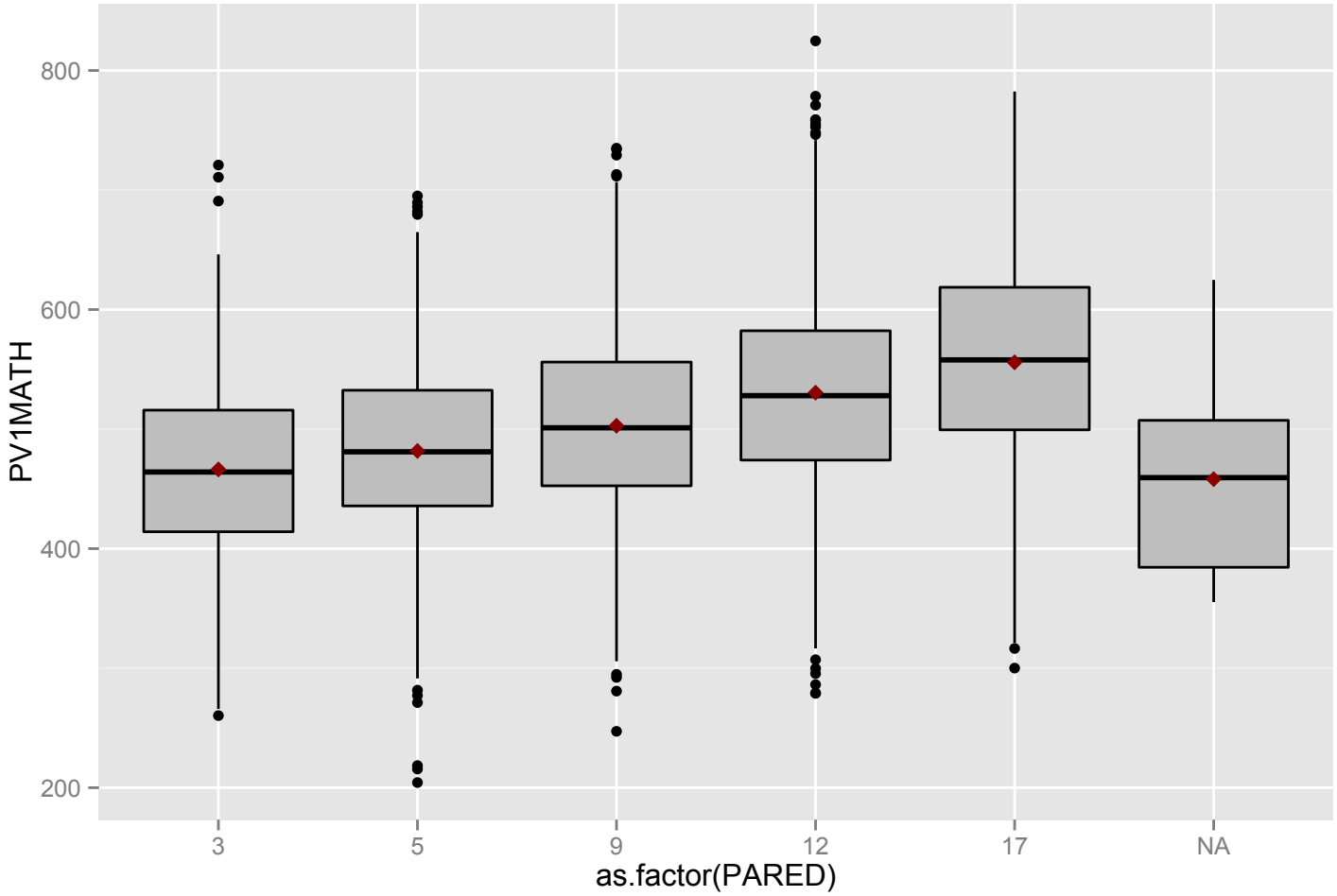
Chuẩn bị data

```
sc$Group[sc$REGION=="NORTH" & sc$AREA=="URBAN"] = "North/Urban"
sc$Group[sc$REGION=="NORTH" & sc$AREA=="RURAL"] = "North/Rural"
sc$Group[sc$REGION=="NORTH" & sc$AREA=="REMOTE"] = "North/Remote"
sc$Group[sc$REGION=="CENTRAL" & sc$AREA=="URBAN"] = "Central/Urban"
sc$Group[sc$REGION=="CENTRAL" & sc$AREA=="RURAL"] = "Central/Rural"
sc$Group[sc$REGION=="CENTRAL" & sc$AREA=="REMOTE"] = "Central/Remote"
sc$Group[sc$REGION=="SOUTH" & sc$AREA=="URBAN"] = "South/Urban"
sc$Group[sc$REGION=="SOUTH" & sc$AREA=="RURAL"] = "South/Rural"
sc$Group[sc$REGION=="SOUTH" & sc$AREA=="REMOTE"] = "South/Remote"

attach(sc)
```

Biểu đồ hộp

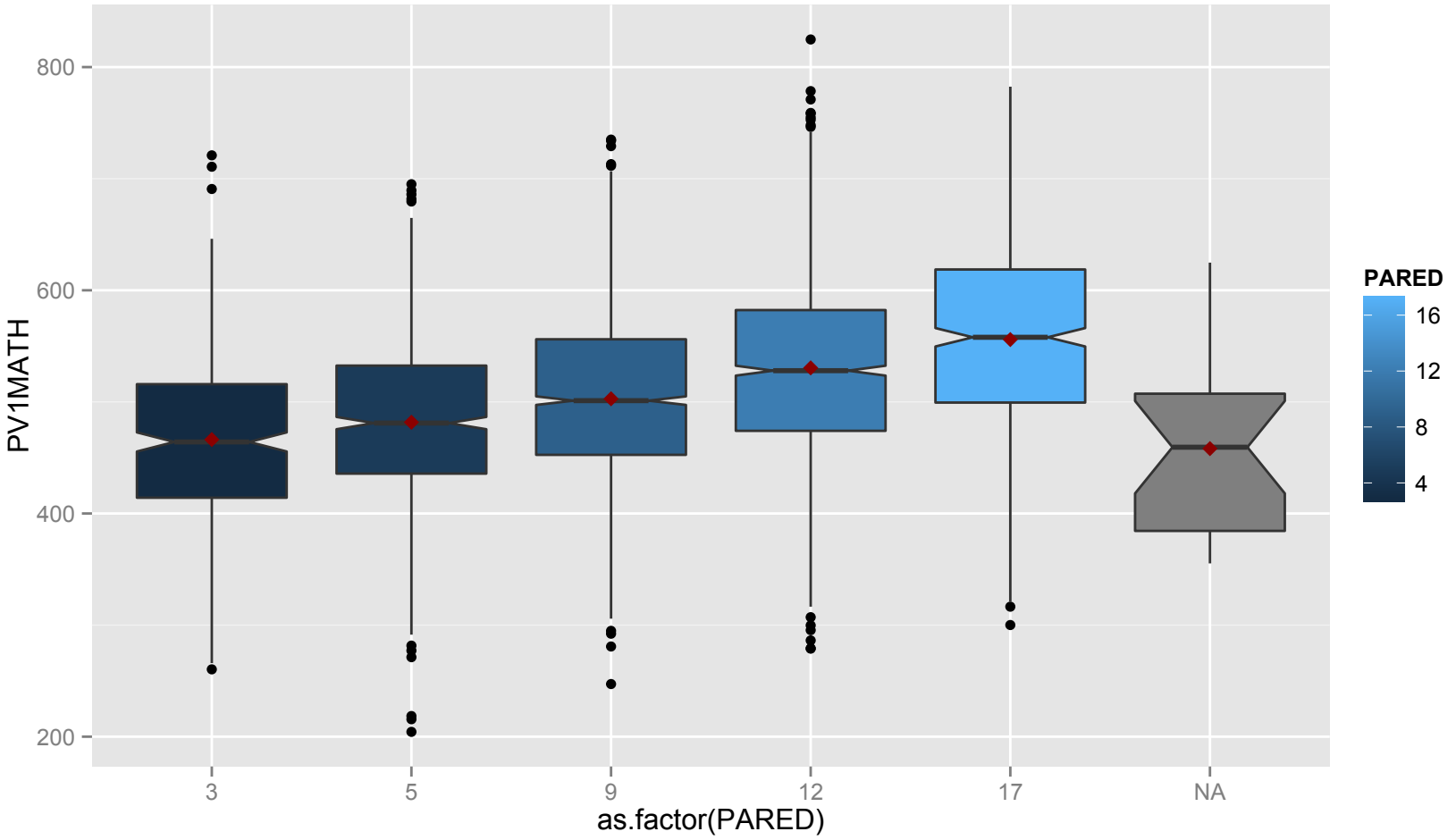
Biểu đồ hộp cơ bản

```
p = ggplot(data=dat,
aes(x=as.factor(PARED), y=PV1MATH))

p = p + geom_boxplot(fill="grey",
color="black")

p = p + stat_summary(fun.y=mean,
colour="darkred", geom="point",
shape=18, size=3)
```


800
600
400
200
PV1MATH
3
5
9
12
17
NA
as.factor(PARED)

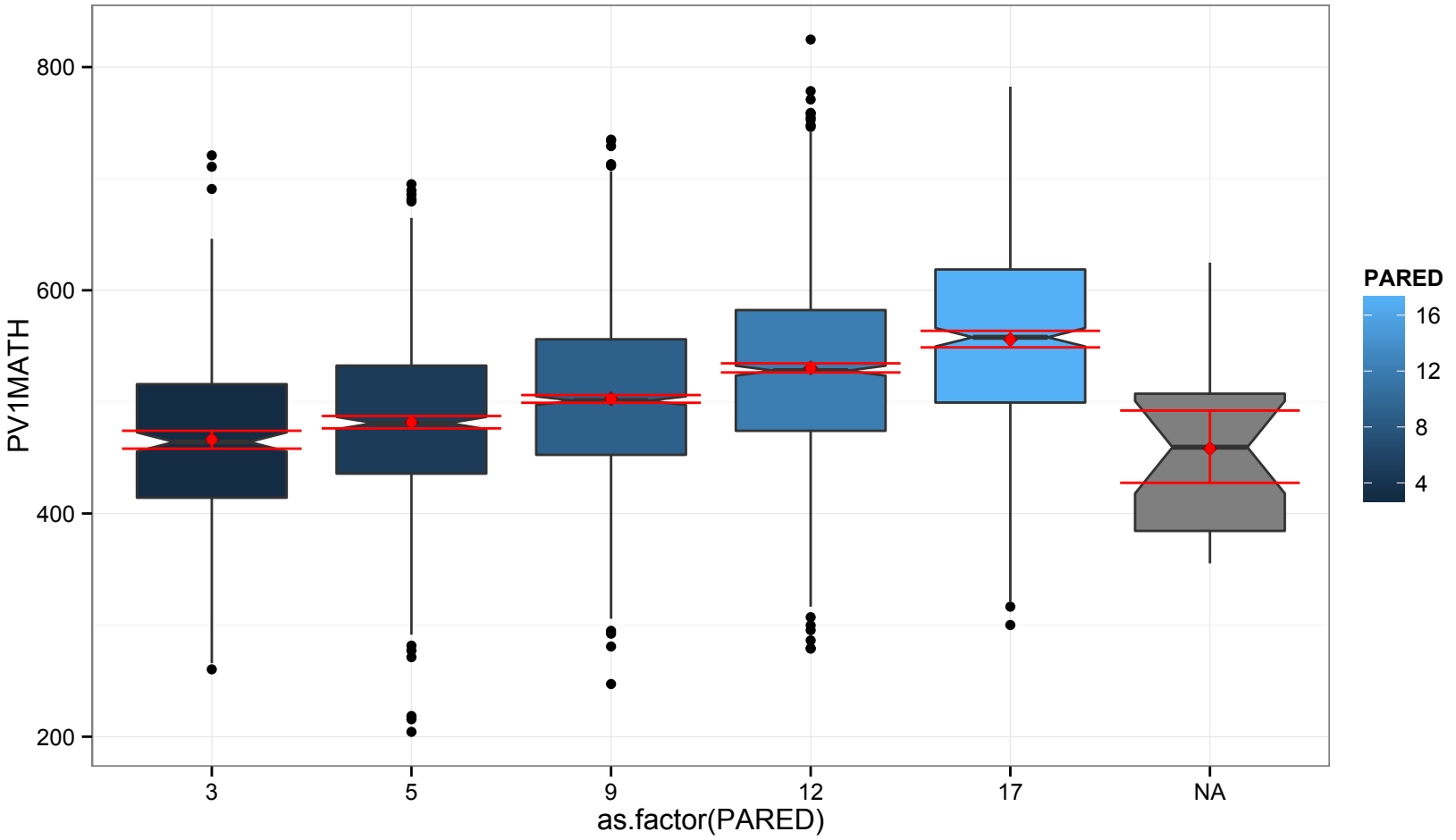
Thêm màu và số trung bình

```
fun_mean = function(x)
{   return(data.frame(y=mean(x),label=mean(x,
na.rm=T)))}

p = ggplot(data=dat, aes(x=as.factor(PARED),
y=PV1MATH, fill=PARED))

p = p + geom_boxplot(notch=T)

p = p +   stat_summary(fun.data=fun_mean,
geom="point", colour="darkred", shape=18,
size=3)
```


800
600
400
200
PV1MATH
3
5
9
12
17
NA
as.factor(PARED)
PARED
16
12
8
4

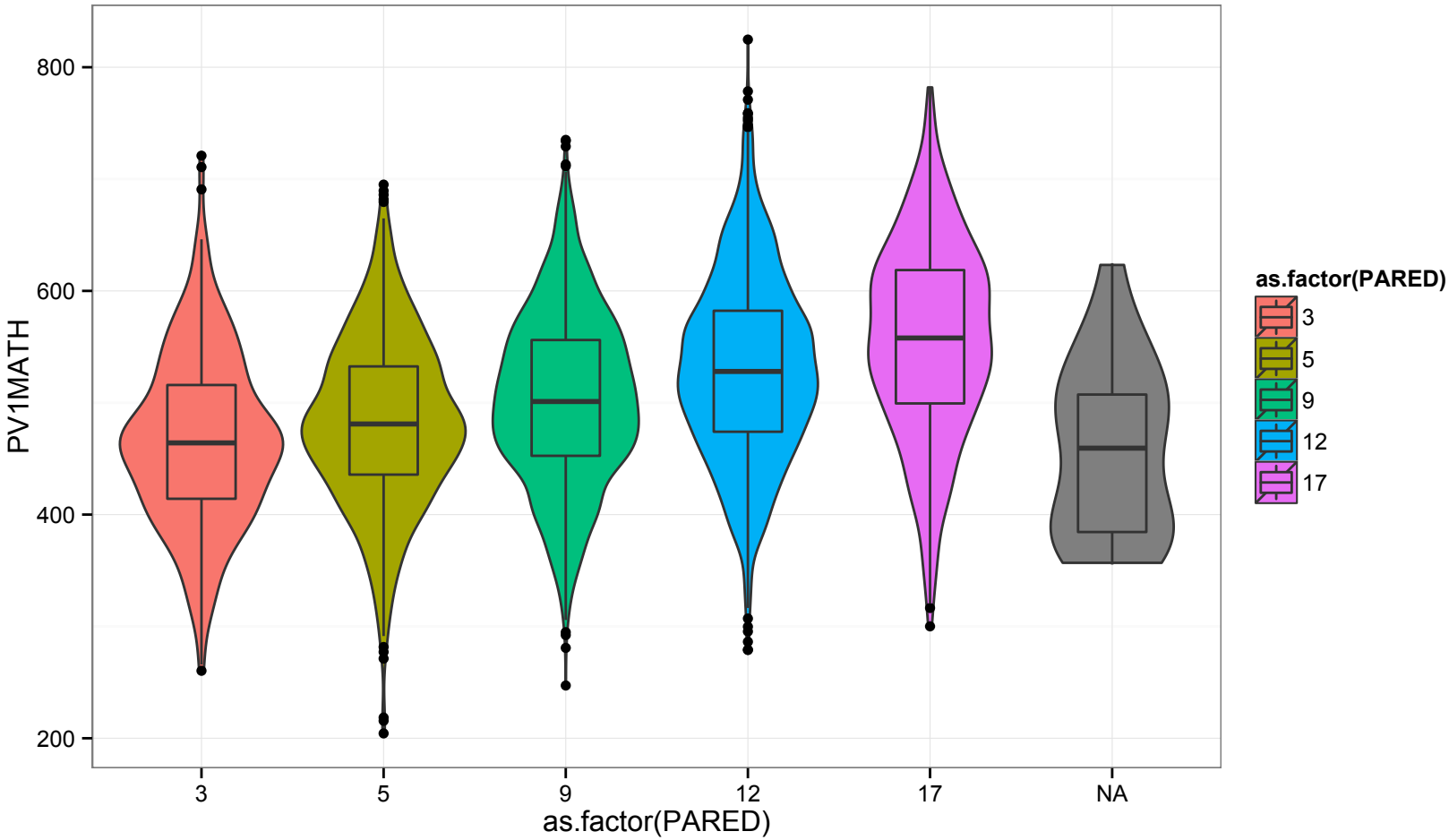
Thêm khoảng tin cậy 95%

```
p = ggplot(data=dat, aes(x=as.factor(PARED),
y=PV1MATH, fill=PARED))

p = p + geom_boxplot(notch=T)

p = p +   stat_summary(fun.data=fun_mean,
geom="point", colour="darkred", shape=18, size=3)

p = p + stat_summary(fun.data=mean_cl_boot,
geom="errorbar", colour="red")

p = p + stat_summary(fun.y=mean, geom="point",
colour="red")

p = p + theme(axis.text.x=element_text(angle=30))

p + theme_bw()
```


800
600
400
200
PV1MATH
3
5
9
12
17
NA
as.factor(PARED)
PARED
16
12
8
4

Biểu đồ violin

```
p = ggplot(data=dat, aes(x=as.factor(PARED),
y=PV1MATH, fill=as.factor(PARED)))

p = p + geom_violin()

p = p + geom_boxplot(width=0.5,
position=position_dodge(width=0))

p = p + theme(legend.position="none")

p + theme_bw()
```


800
600
400
200
PV1MATH
3
5
9
12
17
NA
as.factor(PARED)
as.factor(PARED)
3
5
9
12
17

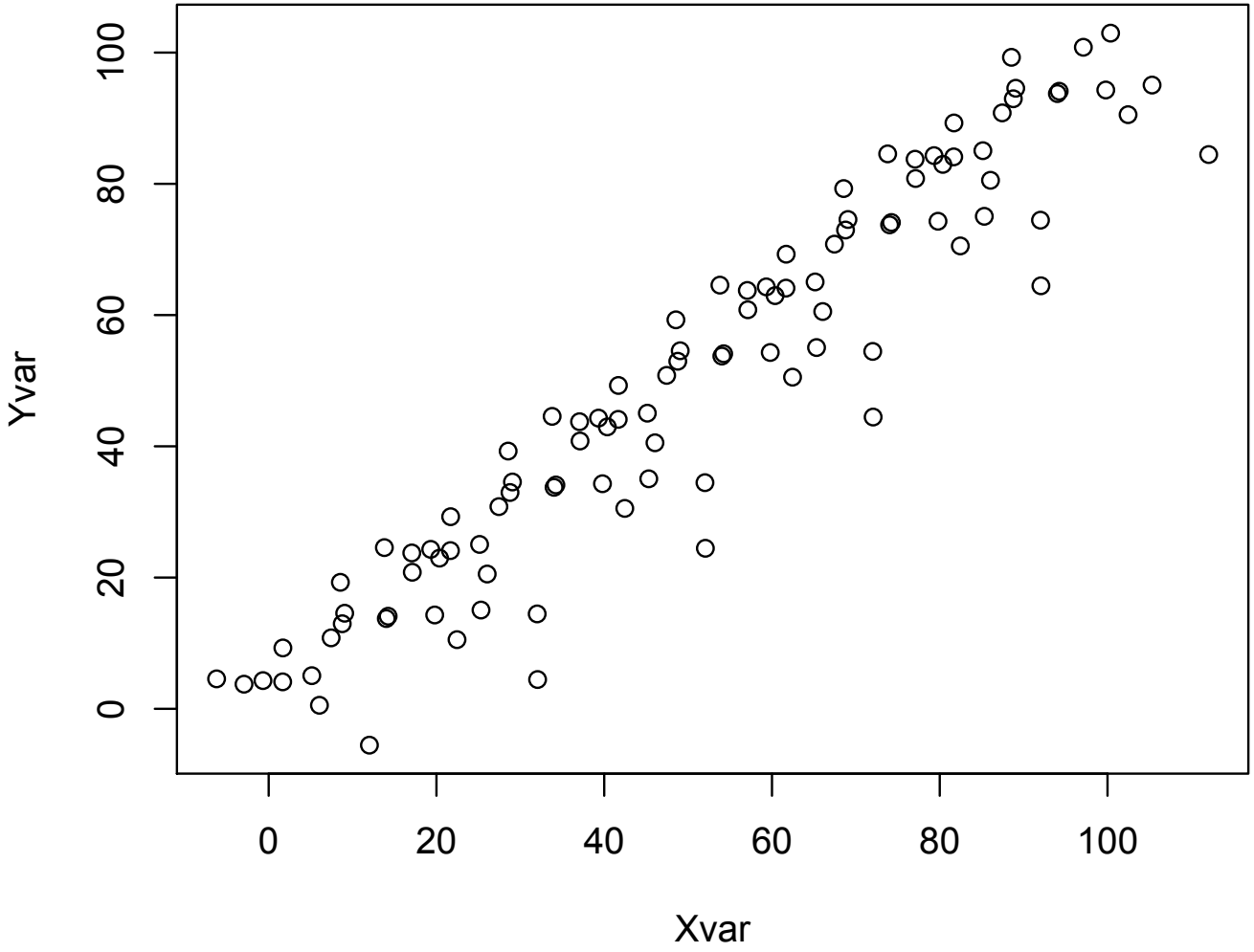
Biểu đồ tương quan

Biểu đồ tương quan

- Mô tả tương quan giữa 2 biến liên tục
- Có thể smooth bằng nhiều hàm số
- Có thể thể hiện cho nhiều nhóm

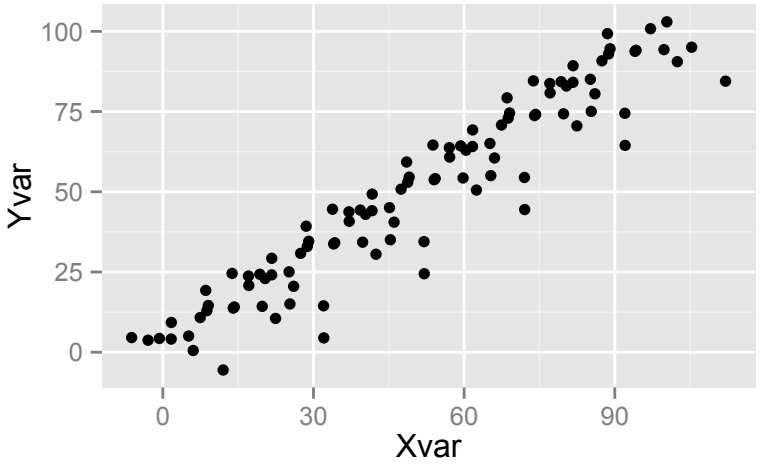
Dữ liệu mô phỏng

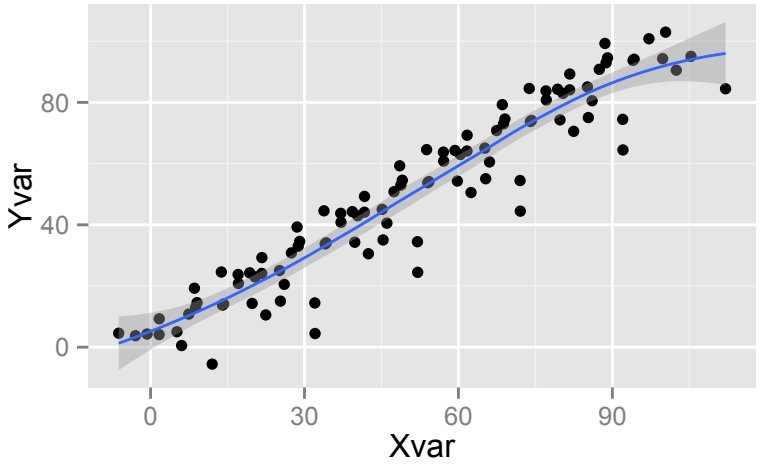
```
Condition = rep(c("A", "B"), each=50)

Yvar = 1:100 + rnorm(10, sd=5)

Xvar = 1:100 + rnorm(20, sd=5)

Data = data.frame(Condition, Yvar, Xvar)

head (Data)

plot(Yvar ~ Xvar)
```

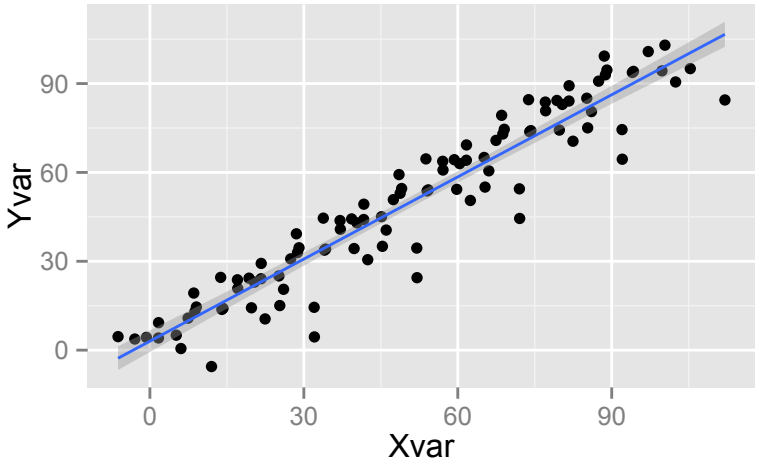

Yvar
0
20
40
60
80
100
Xvar
0
20
40
60
80
100

Biểu đồ tương quan với ggplot2

```
p = ggplot(Data, aes(x=Xvar, y=Yvar))

p1 = p + geom_point(shape=16)

p2 = p + geom_point(shape=16) + geom_smooth()

p3 = p + geom_point(shape=16) +
geom_smooth(method="lm")

library(gridExtra)

grid.arrange(p1, p2, p3, ncol=2)
```

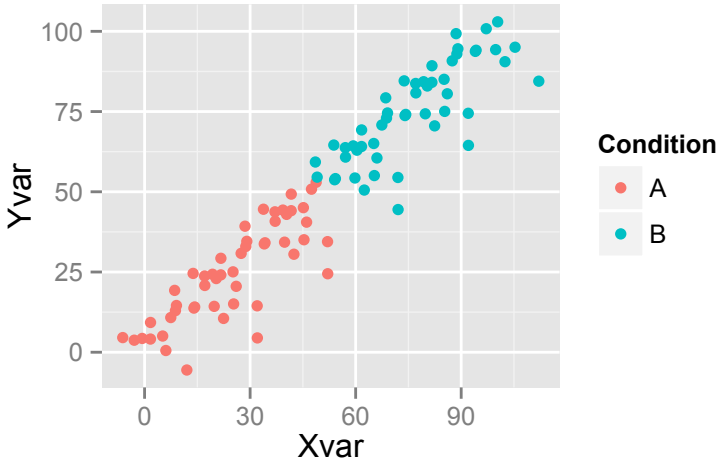

100
75
50
25
0
Yvar
0
30
60
90
Xvar

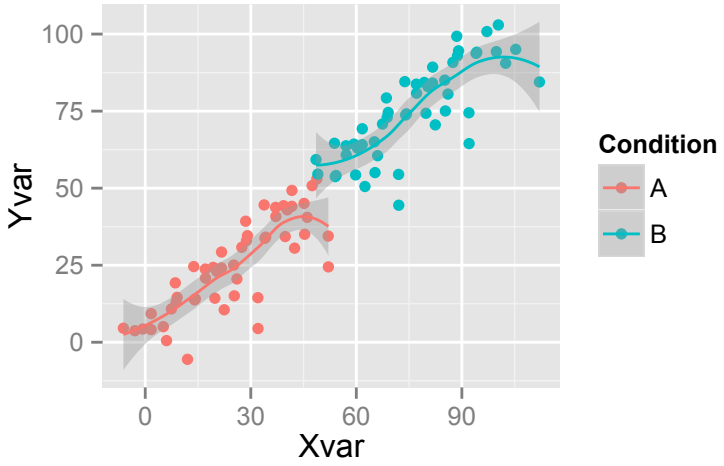

80
40
0
Yvar
0
30
60
90
Xvar

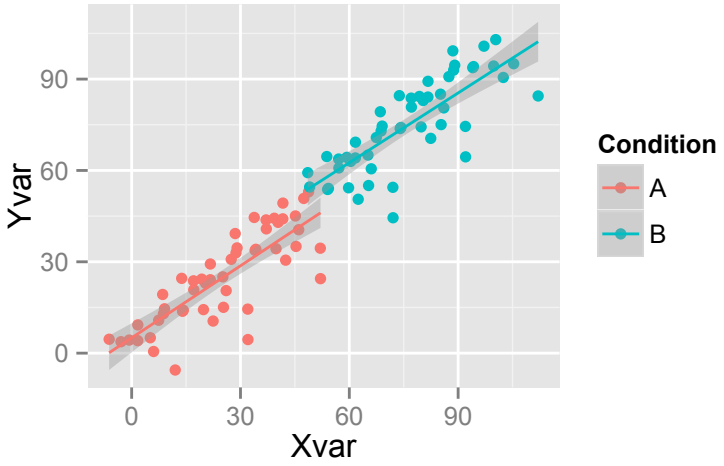

90
60
30
0
Yvar
0
30
60
90
Xvar

Nhưng chúng ta có 2 nhóm (A, B)

```
p = ggplot(Data, aes(x=Xvar, y=Yvar,
col=Condition))

p1 = p + geom_point(shape=16)

p2 = p + geom_point(shape=16) + geom_smooth()

p3 = p + geom_point(shape=16) +
geom_smooth(method="lm")

library(gridExtra)

grid.arrange(p1, p2, p3, ncol=2)
```


100
75
50
25
0
Yvar
0
30
60
90
Xvar
Condition
A
B

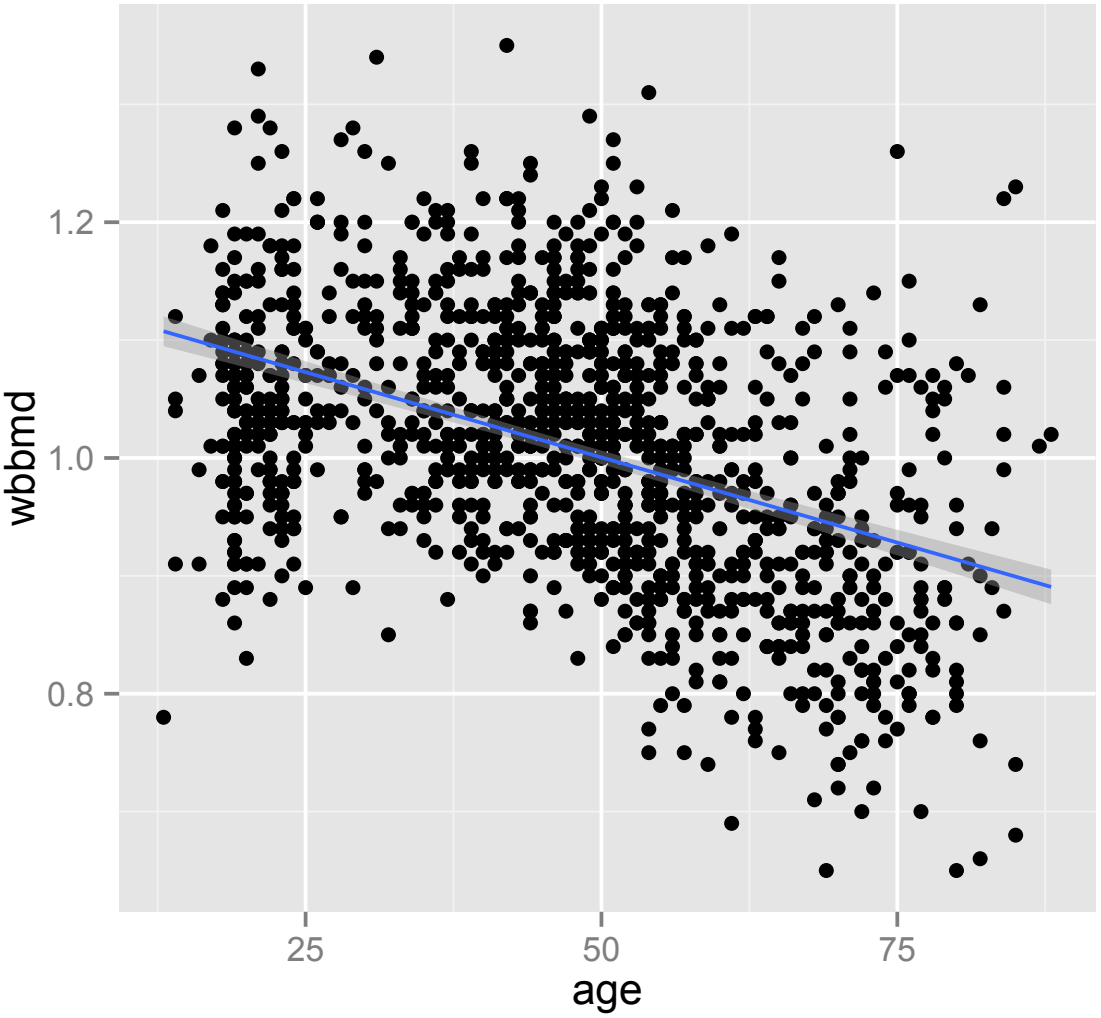

100
75
50
25
0
Yvar
0
30
60
90
Xvar
Condition
A
B


90
60
30
0
Yvar
0
30
60
90
Xvar
Condition
A
B

ggplot2 rất có ích trong việc khám phá

- Dữ liệu nghiên cứu thường phân chia theo nhóm
- Nếu chỉ vẽ một nhóm chung rất khó phát hiện xu hướng và khác biệt giữa các nhóm
- Có thể dùng ggplot2 để vẽ theo nhóm trong CÙNG một biểu đồ

```
p = ggplot(ob, aes(x=age, y=wbbmd))
p + geom_point() + geom_smooth(method="lm")
```



```
p = ggplot(ob, aes(x=age, y=wbbmd))
p + geom_point(aes(col=gender))
```

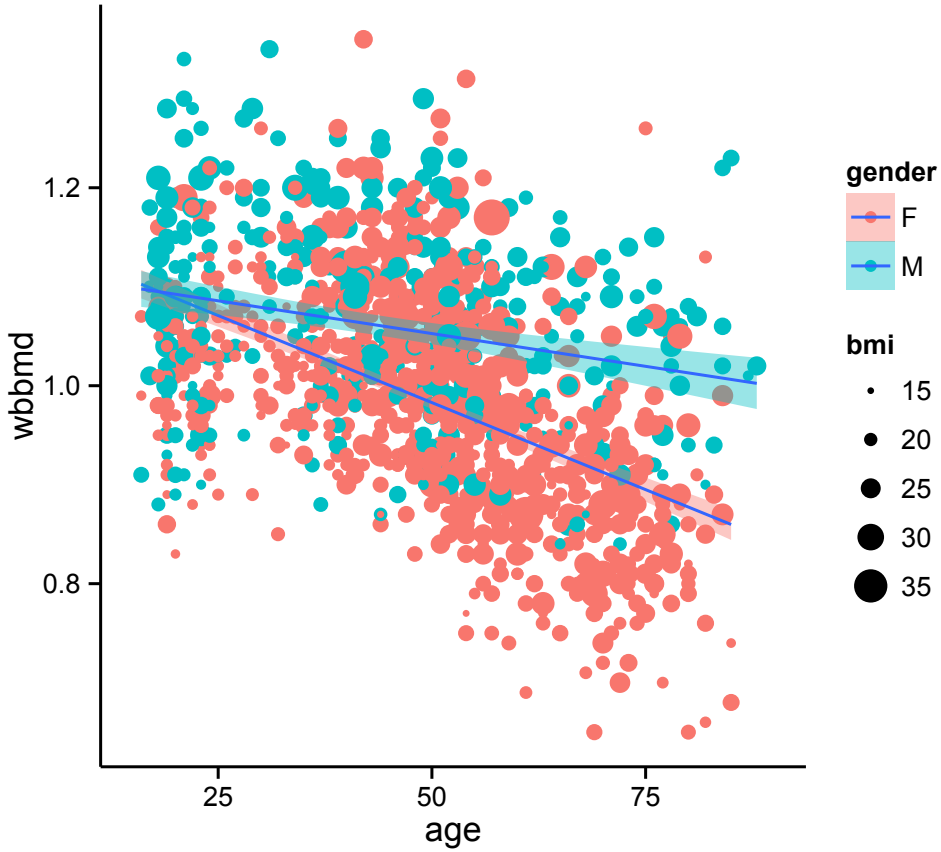
```
p = ggplot(ob, aes(x=age, y=wbbmd))
p + geom_point(aes(col=gender, size=bmi))
```

```
p = ggplot(ob, aes(x=age, y=wbbmd))
p + geom_point(aes(col=gender, size=bmi)) +
geom_smooth()
```

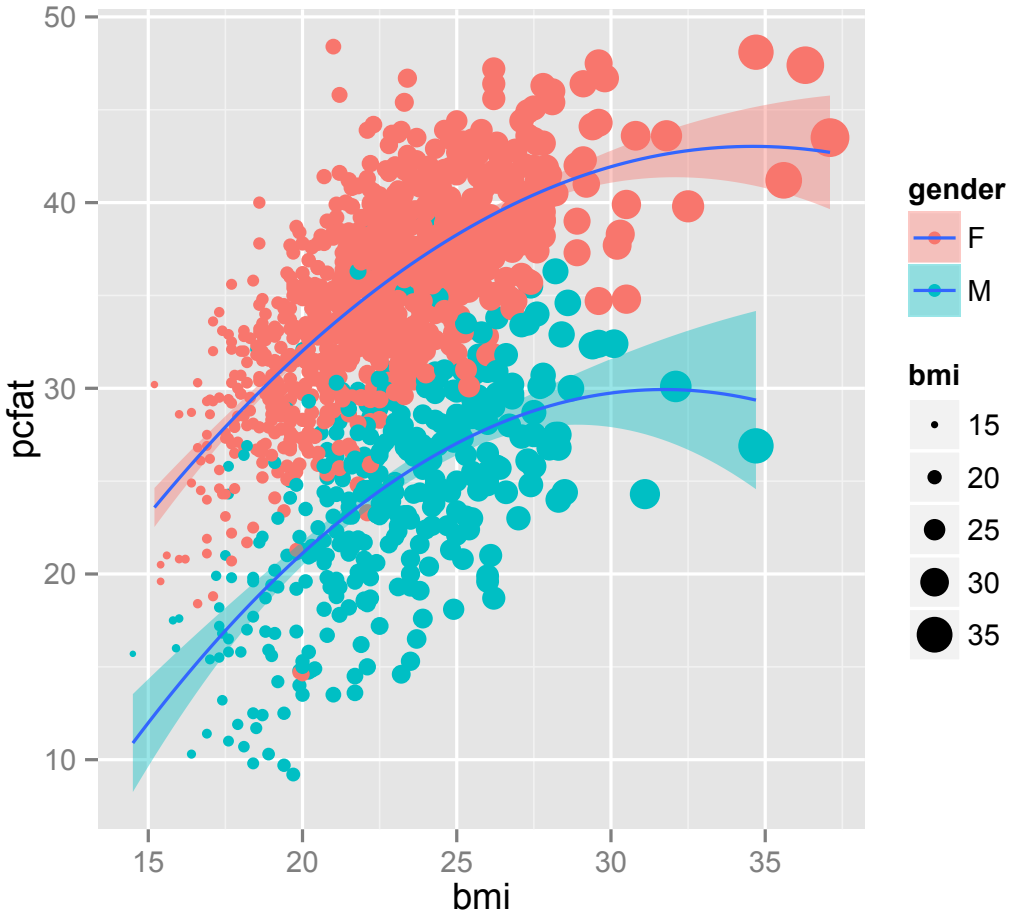
```
p = ggplot(ob, aes(x=age, y=wbbmd,
fill=gender))

p + geom_point(aes(col=gender, size=bmi)) +
geom_smooth()
```

```
p = ggplot(ob, aes(x=age, y=wbbmd,
fill=gender))

p + geom_point(aes(col=gender, size=bmi)) +
geom_smooth(method="lm")
```

```
p = ggplot(ob, aes(x=age, y=wbbmd,
fill=gender))

p + geom_point(aes(col=gender, size=bmi)) +
geom_smooth(method="lm") + theme_bw() +
theme_classic()
```



```
p = ggplot(fm, aes(x=bmi, y=pcfat,
fill=gender))

p + geom_point(aes(col=gender, size=bmi)) +
geom_smooth(method="lm", formula=y~x+I(x^2))
```

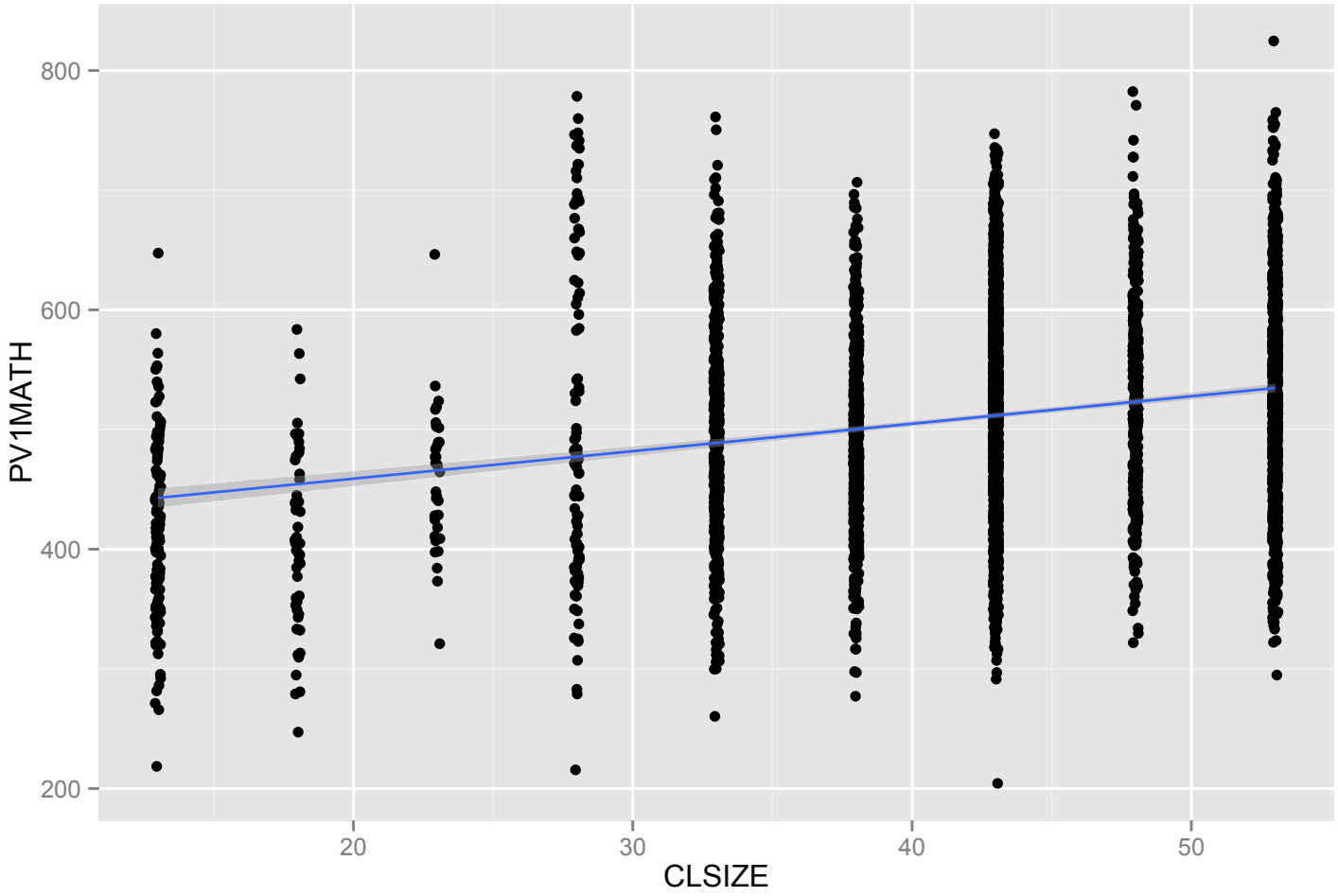


Dữ liệu PISA: trường và học sinh

```
setwd("~/Dropbox/World Bank 2014/Data for
  2015 workshop")

pisa = read.csv("~/Dropbox/World Bank 2014/
  Data for 2015 workshop/PISA DATA temp.csv",
  header=T)

sc = read.csv("~/Dropbox/World Bank 2014/Data
  for 2015 workshop/SCHOOL DATA (VN).csv",
  header=T)

dat = merge(pisa, sc, by="SCHOOLID")

attach(dat)
```

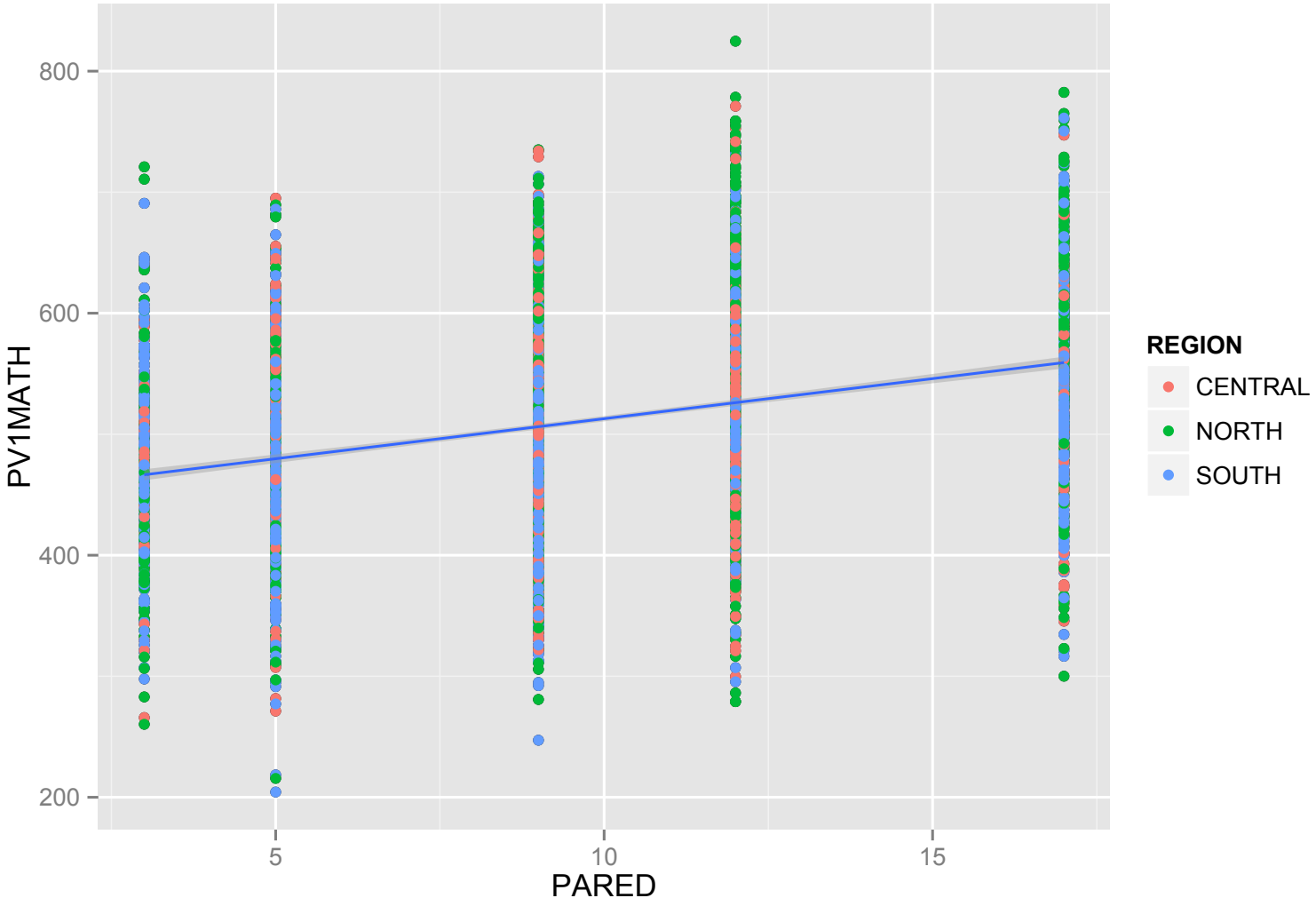
Tương quan giữa CLSIZE và PV1MATH

```
p = ggplot(dat, aes(x=CLSIZE, y=PV1MATH))

p = p +
geom_point(position=position_jitter(w=0.1,
h=0))

p = p + geom_smooth(method="lm")
```


800
600
400
200
PV1MATH
20
30
40
50
CLSIZE

Thêm màu ...

```
p = ggplot(dat, aes(x=PARED, y=PV1MATH))

p = p + geom_point(aes(color=REGION))

p = p + geom_smooth(method="lm", se=T)
```


800
600
400
200
PV1MATH
5
10
15
PARED
REGION
CENTRAL
NORTH
SOUTH

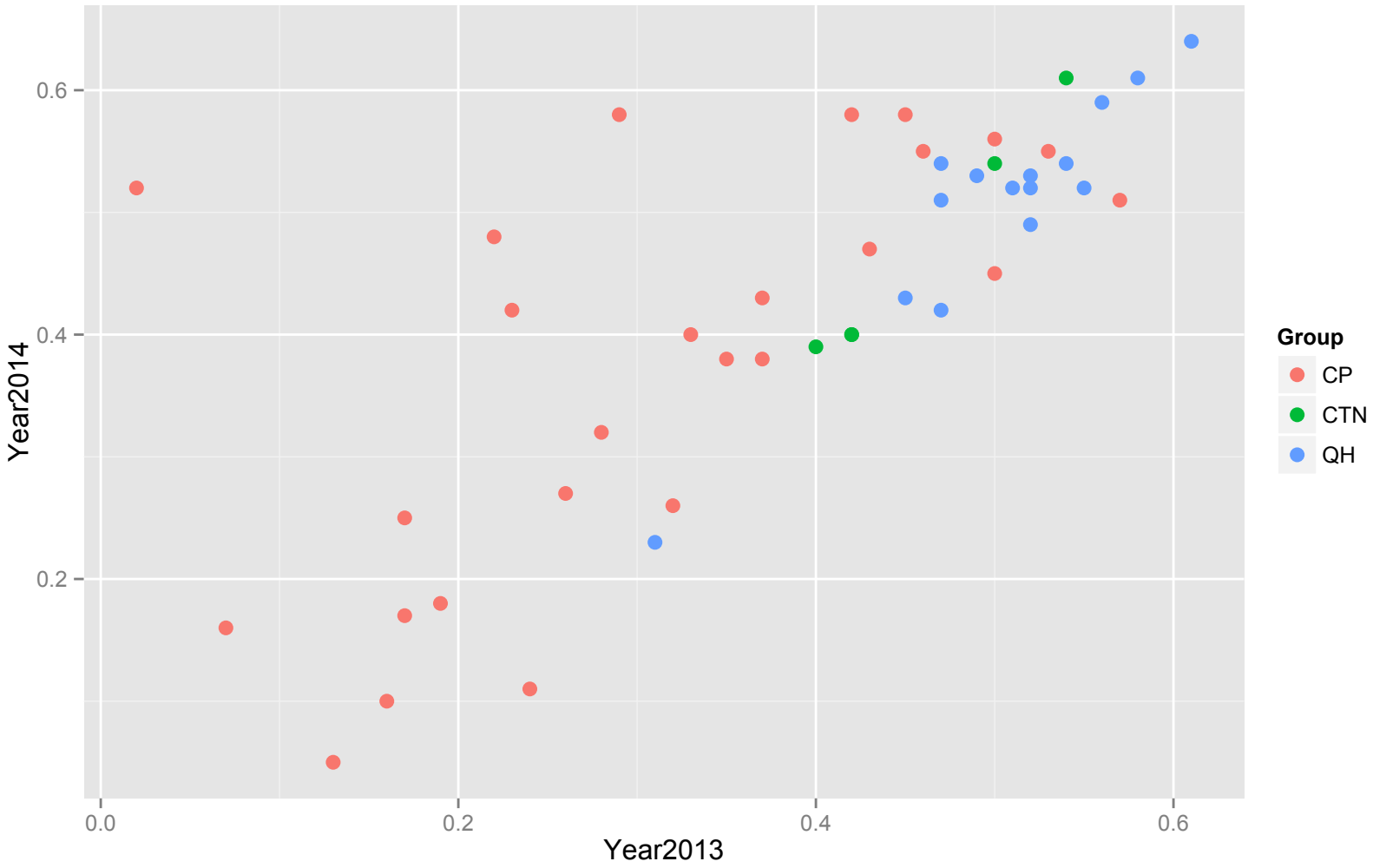
Biểu đồ tương quan với text

- Dữ liệu: lấy phiếu tín nhiệm 2013, 2014
- Mục tiêu: Xem xét sự thay đổi về điểm tín nhiệm

Name	Group	Year2014	Year2013	Change	Cao-2013	TN-2013	Th_p-2013	Cao-2014	TN-2014	Th_p-2014
NTDoan	CTN	0.54	0.5	0.04	263	215	13	302	168	15
DLThang	CP	0.58	0.29	0.29	186	198	99	362	91	28
BQVinh	CP	0.58	0.42	0.16	231	205	46	351	112	20
NXPhuc	CP	0.58	0.45	0.13	248	207	35	356	103	26
NVGiau	CP	0.56	0.5	0.05	273	204	15	317	155	12
PBMinh	CP	0.55	0.46	0.09	238	233	21	320	146	19
PQHien	CP	0.55	0.53	0.02	291	189	11	315	148	20
NVBinh	CP	0.52	0.02	0.5	88	194	209	323	118	41
PQThanh	CP	0.51	0.57	-0.06	323	144	13	313	129	41
NTDung	CP	0.48	0.22	0.26	210	122	160	320	96	68
VDDam	CP	0.47	0.43	0.04	215	245	29	257	196	32
TDQuang	CP	0.45	0.5	-0.05	273	183	24	264	166	50
DTDung	CP	0.44						247	197	41
HTHai	CP	0.43	0.37	0.06	186	261	44	225	226	34
TDDung	CP	0.42	0.23	0.19	131	261	100	236	201	48
VVNinh	CP	0.4	0.33	0.07	167	264	59	202	246	35

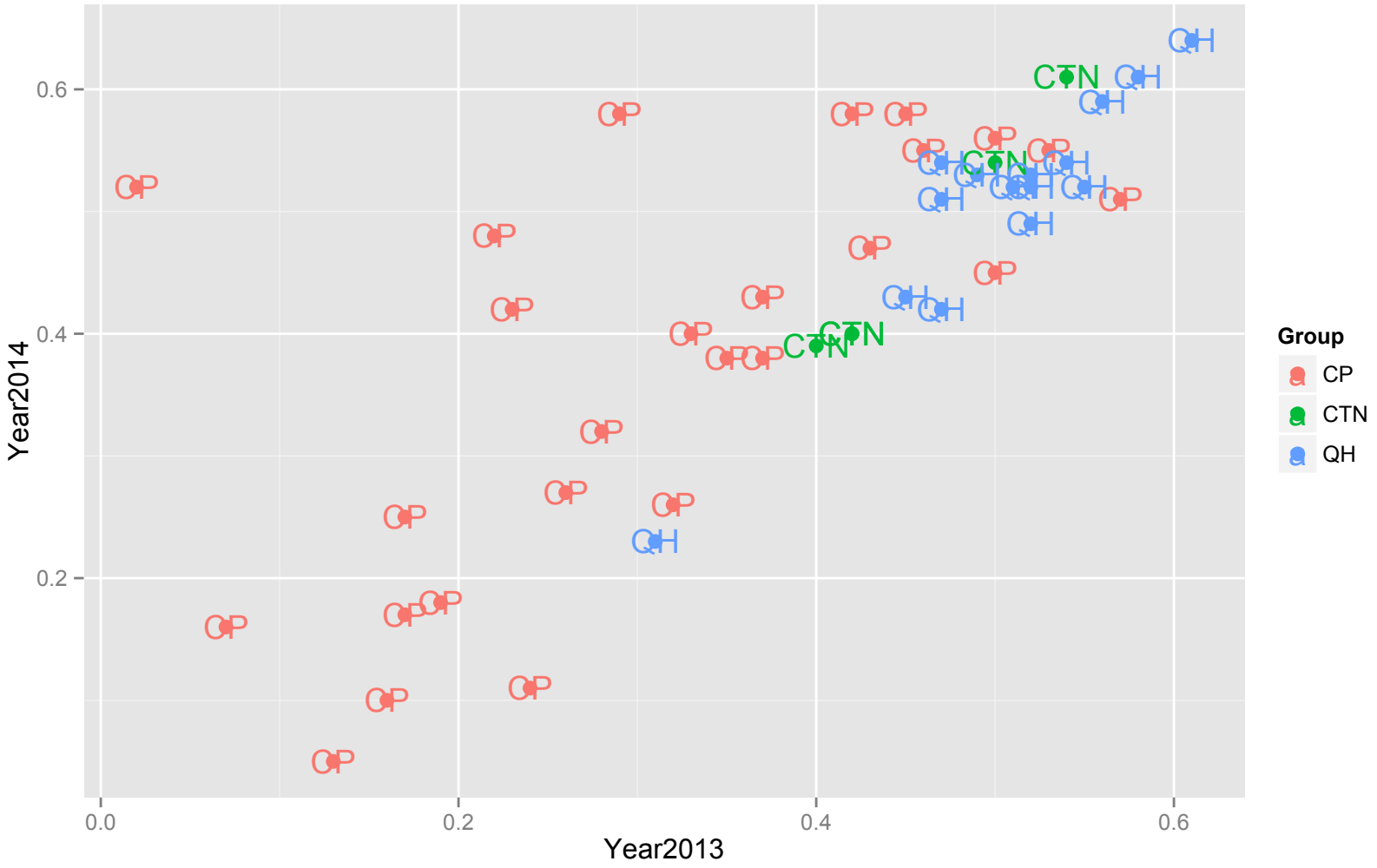
Dùng ggplot2 để xem dữ liệu

```
tn = read.csv("~/Dropbox/World Bank 2014/Data for 2015 workshop/Lay phieu tin nhiem.csv", header=T)

p = ggplot(tn, aes(x=Year2013, y=Year2014, color=Group))

p = p + geom_point(shape=16, size=3)

p
```


0.6
0.4
0.2
Year2014
0.0
0.2
0.4
0.6
Year2013
Group
CP
CTN
QH

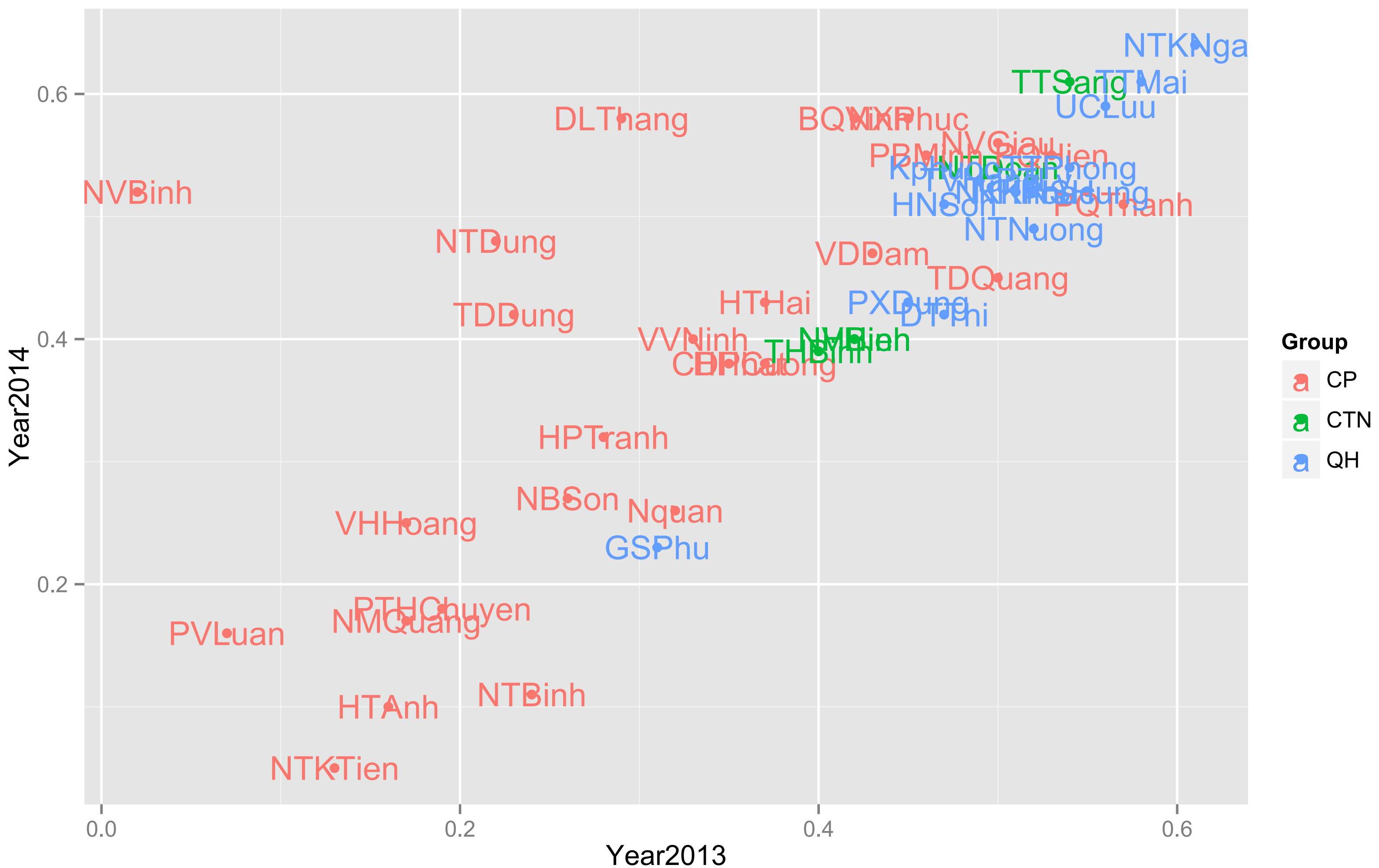
Thêm nhóm

```
p = ggplot(tn, aes(x=Year2013, y=Year2014,
color=Group))

p = p + geom_point(shape=16, size=3)

p1 = p + geom_text(aes(label=Group))
```


Year2014
Year2013
0.0
0.2
0.4
0.6
Group
CP
CTN
QH

Thêm tên

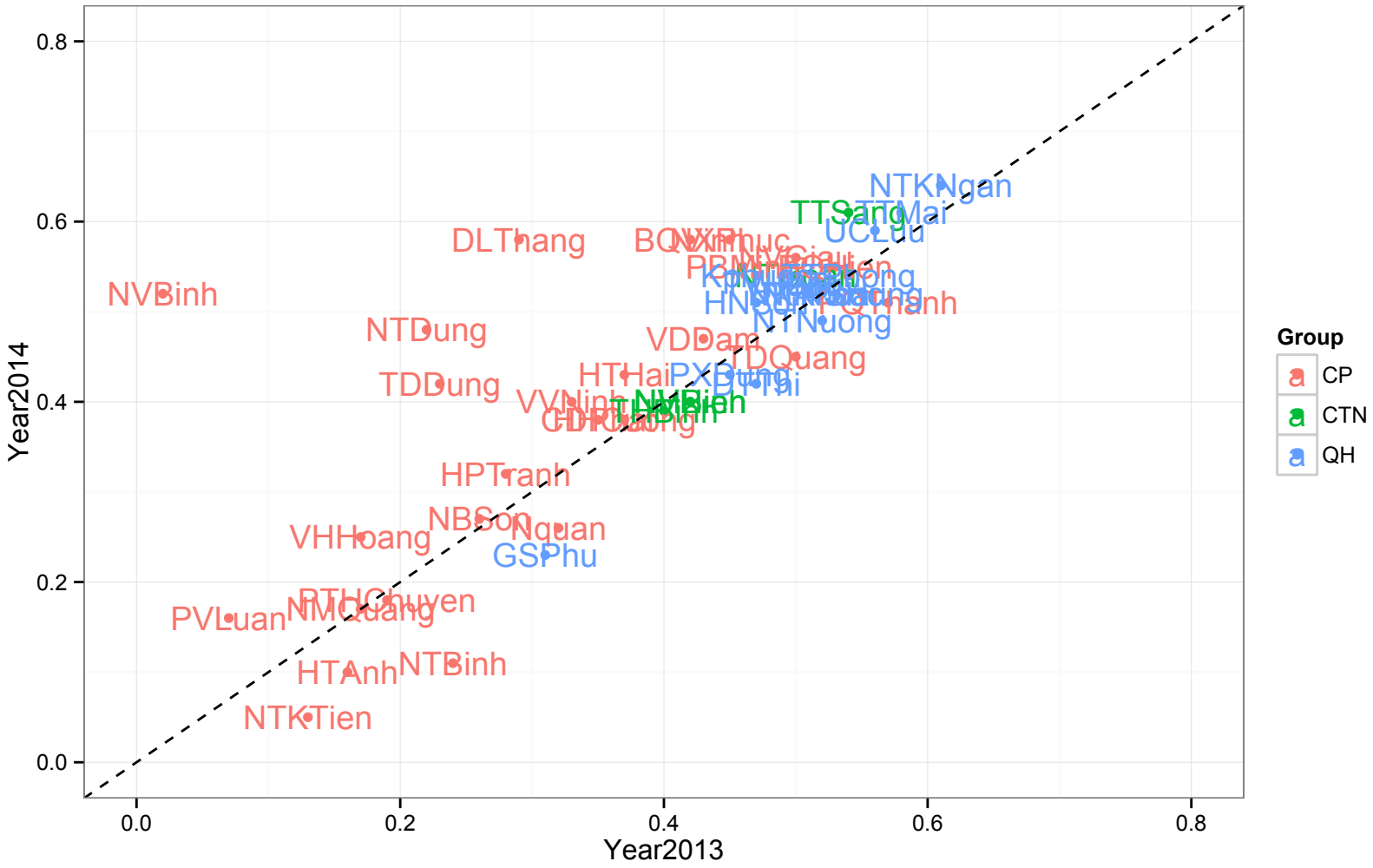
```
p = ggplot(tn, aes(x=Year2013, y=Year2014,
color=Group))

p = p + geom_point(shape=16, size=2)

p2 = p + geom_text(aes(label=Name), size=5)

p2
```



Year2014
Year2013
0.0
0.2
0.4
0.6
Group
CP
CTN
QH
NTKNga
TTSang
TTMai
UCLuu
DLThang
NVBinh
NTDung
TDDung
VDDam
TDQuang
HTHai
PXDung
DTThi
NTNuong
HNSon
PQThanh
VVNinh
HPTranh
NBSon
Nquan
VHHoang
GSPhu
PVLuan
NMQuang
PTHChuyen
HTAnh
NTBinh
NTKTien

Thêm đường tham chiếu (reference line)

```
p = ggplot(tn, aes(x=Year2013, y=Year2014,
color=Group))

p = p + geom_point(shape=16, size=2)

p2 = p + geom_text(aes(label=Name), size=5)

p2 = p2 + xlim(0,0.8) + ylim(0,0.8)

p2 + geom_abline(slope=1, size=0.5, lty=2)

p2 + theme_bw()
```


Year2014
Year2013
0.0
0.2
0.4
0.6
0.8
NTKNgan
TTSang
TTMai
UCLuu
DLThang
NVBinh
NTDung
VDDam
TDQuang
HTHai
TDDung
HPTranh
NBSon
VHHoang
Nquan
GSPhu
PVLuan
HTAnh
NTBinh
NTKTien
NTNuong
Group
CP
CTN
QH