多元共線性是指多元迴歸分析中，自變項之間有相關存在的一種現象，是一種程度的問題(degree of matters)，而不是全有或全無(all or none)的狀態。

Linear Regression Analysis (3): Multicollinearity & Variables Selection

杜裕康

國立台灣大學公共衛生學院
流行病學與預防醫學研究所

Multicollinearity（線性重合）：有兩個以上的自變數互為高度相關的現象。

yukangtu@ntu.edu.tw

Outline

- What is multicollinearity and why is it a problem?
- How do we detect this problem?
- What are the consequences of multicollinearity?
- What are the solutions to multicollinearity?
- Algorithms for variable selection
- Use your knowledge not algorithms to select variables

Collinearity and Multicollinearity

- One of the assumptions in regression model is that the explanatory variables are not exactly linearly related.
  - If they are, then not all parameters are estimable
- In mathematical terms, this means that in the regression model with *p* explanatory variables, the following relation holds:

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p = 0$$

- In plain English, this means that no explanatory variable in the model can be expressed as a linear combination of other explanatory variables in the model

Multicollinearity

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p = 0$$

- If this relation holds, there is perfect multicollinearity in the model
  - For example, $x_1 = 5 + x_2 + 2x_3$
- When the explanatory variables are highly intercorrelated but not *perfectly* multicollinear, the individual parameters are not estimable with sufficient precision (because of high standard errors).

Multicollinearity

- Multicollinearity is one of the most misunderstood problems in multiple regression
- Several measures for multicollinearity suggested in the literature (variance-inflation factors, condition numbers)
- All these are not very useful, because they are all based on the correlation structure of the explanatory variables only.
- Multicollinearity is only one of several factors determining high standard errors
- High intercorrelations among the explanatory variables are neither necessary nor sufficient to cause problems

Multicollinearity

- Textbooks usually offer following solutions to multicollinearity problem:
  - Dropping variables
  - Collect more data
  - Principal component regression
  - Partial least squares regression
- However, they may or may not solve the problem

如果共線性是因為而有高度相關

用降維的方法

Perfect Multicollinearity

完全相關(perfect correlated)：即兩個自變數的相關係數為1。

- When one of explanatory variable x_i is a linear combination of other explanatory variables in the model, the rank of design matrix $\mathbf{X}$ is no longer a full rank matrix

X平方的意思

- Consequently, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not full ranked either
- Remember that to estimate regression coefficient $\mathbf{b}$, we need to obtain the inverse matrix of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

求反舉證

5的倒數是1/5,5X1／5=1

- But $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ does not exist, because the determinant of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is zero. Consequently, $\mathbf{b}$ is not estimable in the usual sense

行列數為0,有線性關係時無法計算，代表有多餘的變項

Perfect Multicollinearity

- Perfect multicollinearity is rare, because it usually means at least one variable is redundant.
- For example, body height can be measured in centimeters or feet. Including both will cause multicollinearity.
- However, it may occur by design, i.e. we believe the linear combination of variables has a unique interpretation
- For example, in periodontology, clinical attachment is the sum of probing pocket depth and gingival recession
- In epidemiology, a famous example is the Age-Period-Cohort model, where Period = Age + Cohort

Perfect Multicollinearity

- Note that the problem with regression coefficients in a model with perfect multicollinearity is not that there is no solution but there is no unique solution
- The inverse matrix of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, does not exist, but there are indefinite numbers of generalized inverse matrix of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{g}$
- One of them is the Moore-Penrose generalized inverse matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}$, which is related to principal component regression discussed later

What is Multicollinearity

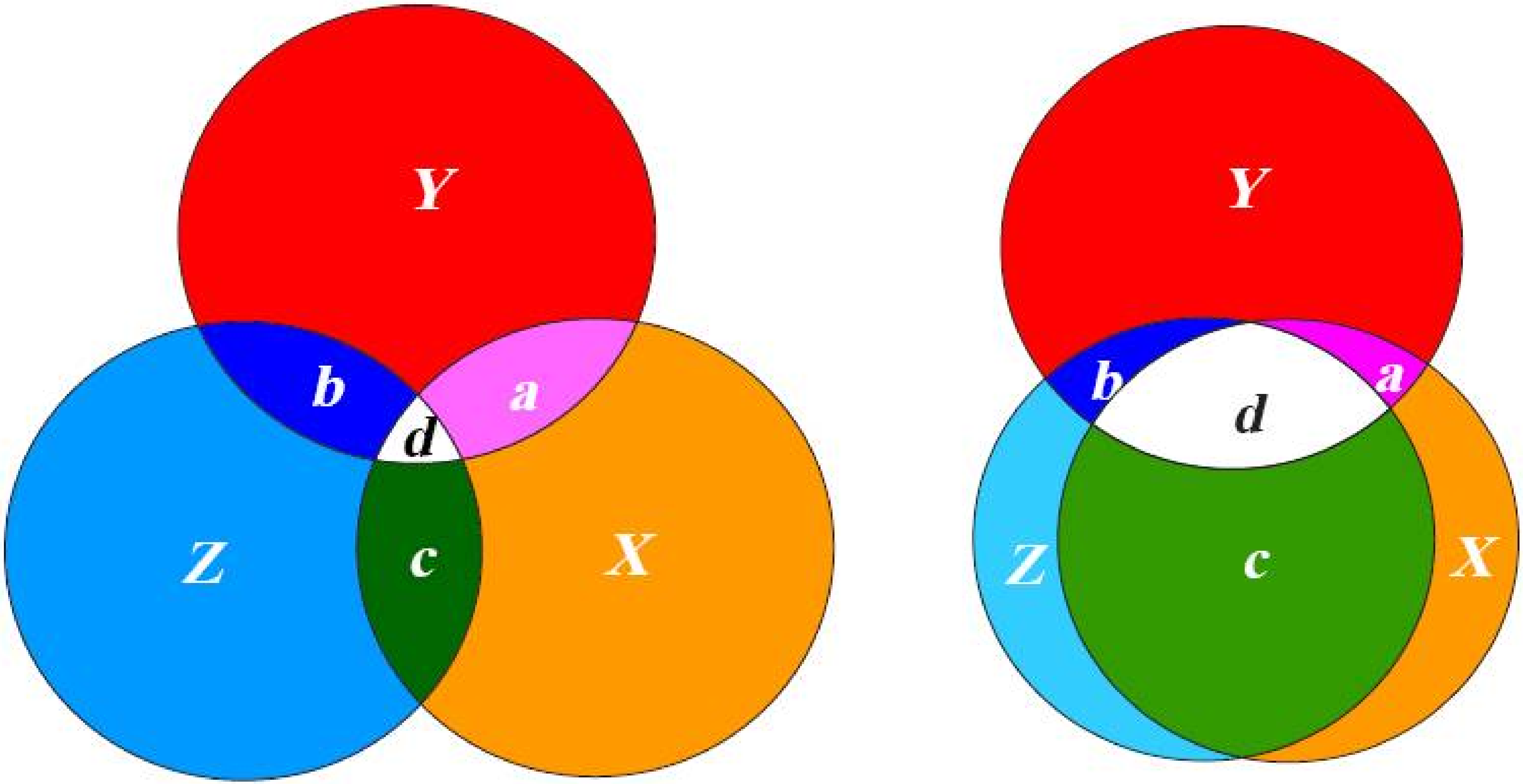
對係數的解釋會有不良影響

- Very often the data we use in multiple regression analysis cannot give decisive (significant) answers to the questions we pose.
- This is because the standard errors are very high or the t-ratios are very low.
- This sort of situation occurs when the explanatory variables display little variation and/or high intercorrelations. This is known as multicollinearity.
- It becomes difficult to disentangle the separate effects of each of the explanatory variables on the explained variable

Venn Diagrams for Multicollinearity

估計獨特性時困難

重疊越多，競爭過度激烈，AB會變小

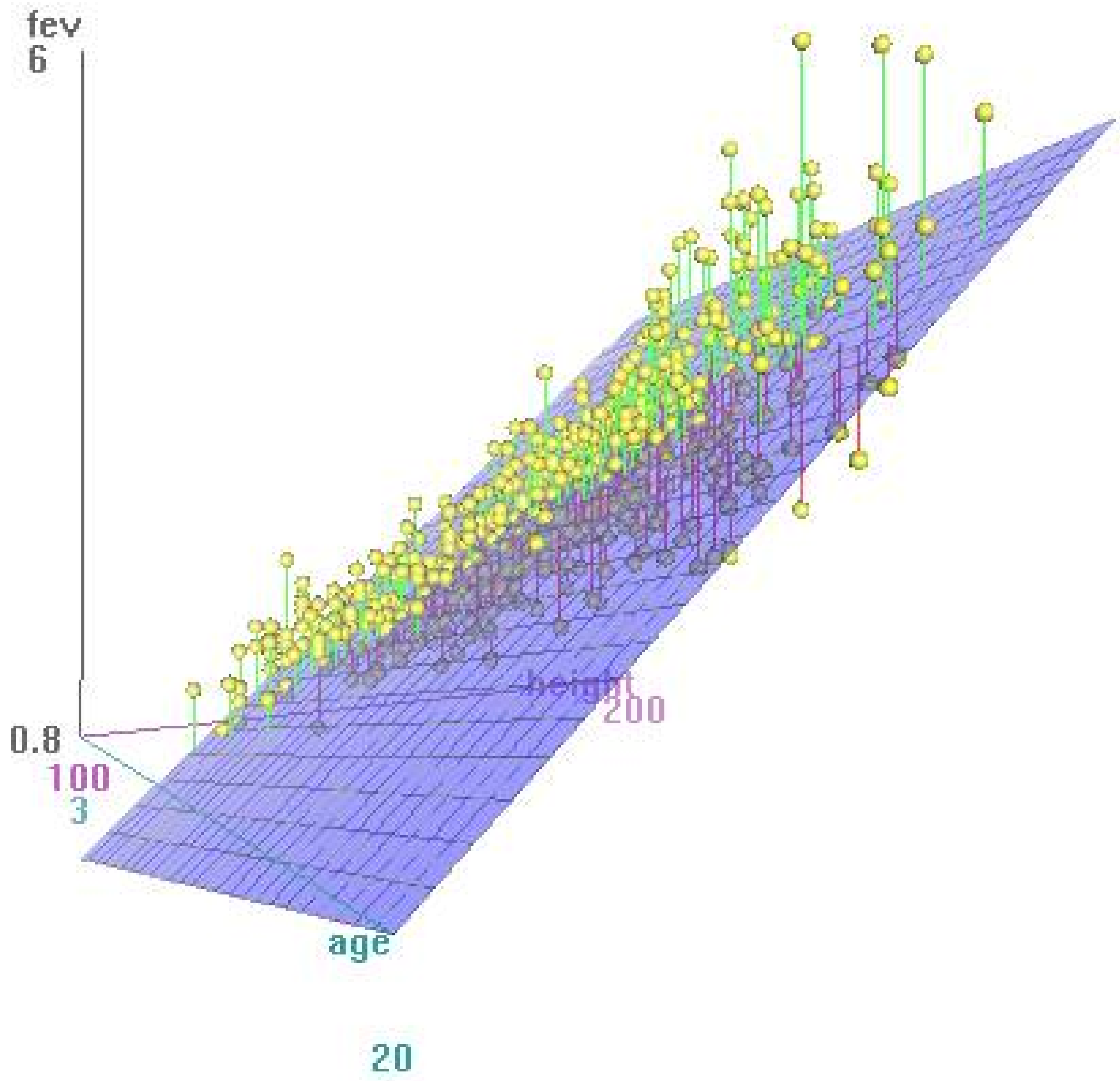


Consequence of Multicollinearity

- As inter-correlated explanatory variables competes for their contributions to explaining the variance of *y*
- Their "independent" contributions after accounting the contributions of others may become much smaller
- In the extreme scenario, the direction of relationship between explanatory variables and y may be reversed, i.e. change in the signs of regression coefficients from simple to multiple regression

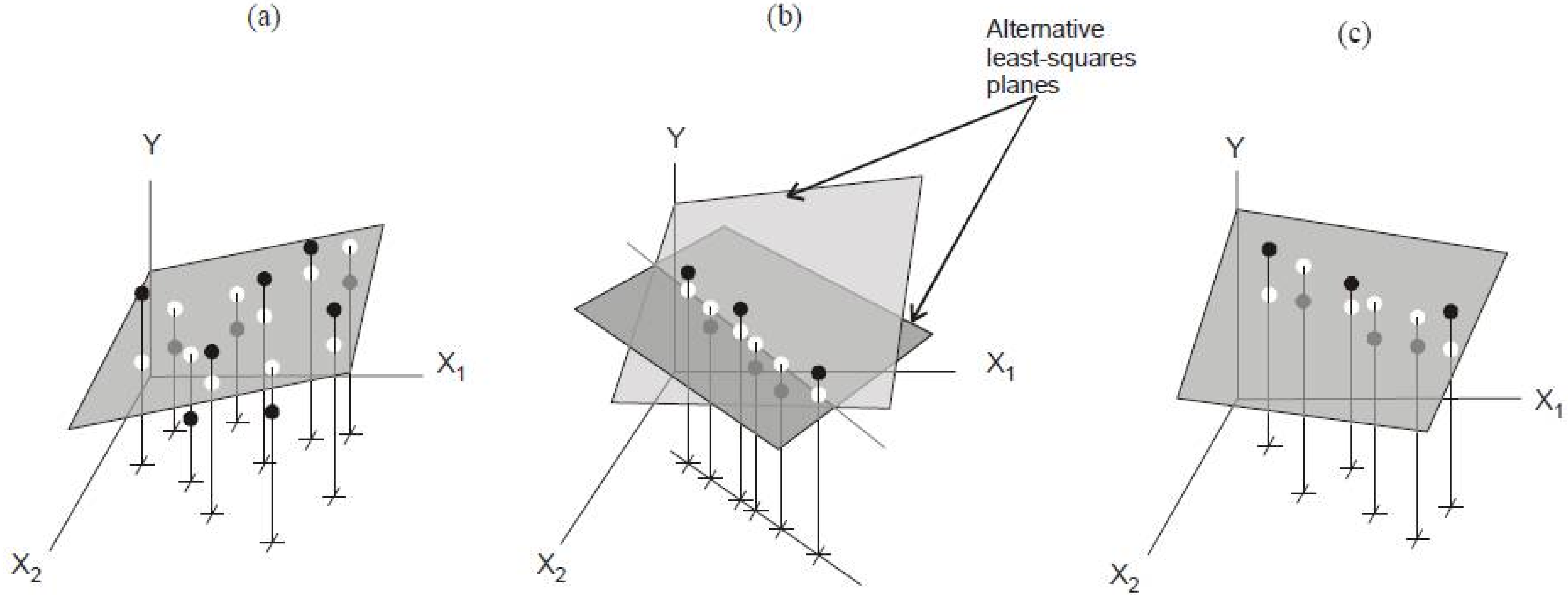
搶人搶太兇，關係會反轉，獨特貢獻會變小

每個點為觀測值，找出最好的平面，點到線的距離最平均



Consequence of Multicollinearity

X1X2太近，相關性太高，有誤差
貢獻性太高對迴歸模型最大的影響在



Taken from Prof John Fox's book "Applied Regression Analysis" 2nd edition, p.310

Consequence of Multicollinearity

- A major concern with collinearity is what happens to the variance of regression coefficient b_i for x_i.
- For $y = b_0 + b_1x_1 + b_2x_2 + e$, the variance of b_1 is given by:

$$Var(b_1) = \sigma^2 \left(\frac{1}{\sum(X_1 - \bar{X}_1)^2}\right)\left(\frac{1}{1 - r_{12}^2}\right)$$

X1X2的相關係數

- Where σ^2 is residual error variance, r_{12}^2 is the correlation between x_1 and x_2, $\sum(X_1 - \bar{X}_1)^2$ is the sum of squares for x_1

Consequence of Multicollinearity

$$Var(b_1) = \sigma^2 \left(\frac{1}{\sum (X_1 - \bar{X}_1)^2} \right) \left(\frac{1}{1 - r_{12}^2} \right)$$

- So $Var(b_1)$ increases, if

和模型有關 σ^2 ↑, i.e. residual error variance increases
  - $\sum(X_1 - \bar{X}_1)^2$ ↓, variance of x_1 decreases
  - r_{12}^2 ↑, the correlation between x_1 and x_2 increases
- Consequently, collinearity increases the variance of estimated regression coefficients

如果X範圍越大，估計出來的關係越精準
變異數就越小
相關係數就越小
兩者不能為1或－1,不然會無解

Consequence of Multicollinearity

- We now extend our discussion to multiple regression
$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + e$
- The variance of b_i is given by:

$$Var(b_i) = \sigma^2 \left(\frac{1}{\sum (X_i - \bar{X})^2} \right) \left(\frac{1}{1 - R_i^2} \right)$$

- Where σ^2 is residual error variance, R_i^2 is the correlation of determination for x_i regressed on other explanatory variables in the model, $\sum (X_i - \bar{X}_i)^2$ is the sum of squares for x_i

RJ平方越大，代表X提供對Y的貢獻，變異更大
訊息重覆程度更大

Consequence of Multicollinearity

解釋相關性

$$Var(b_i) = \sigma^2 \left(\frac{1}{\sum (X_i - \bar{X})^2} \right) \left(\frac{1}{1 - R_i^2} \right)$$

- So $Var(b_i)$ increases, if
  - σ^2 ↑, i.e. residual error variance increases
  - $\sum (X_i - \bar{X}_i)^2$ ↓, variance of x_i decreases
  - R_i^2 ↑, the correlation of determination for x_i **regressed on other explanatory variables** in the model increases
- Multicollinearity increases R_i^2 , and when x_i is a linear combination of some of the explanatory variables, $R_i^2 = 1$, $Var(b_i)$ is singular for all variable involved

Measures of Multicollinearity: Pairwise Correlations

- Calculate the pairwise correlations between each pair of explanatory variables

- If any correlation is greater than, say, 0.7, these two explanatory variables are highly collinear, and including both in the model may cause problems

文字

- However, no high correlation does not necessarily mean there is no problem of multicollinearity, e.g.

- If $x_1 = x_2 + x_3 + x_4 + x_5 + x_6$, and x_2 to x_6 are independent, we shall only find small correlations between x_1 and other variables

Measures of Multicollinearity: VIF

- It is important to be familiar with two measures that are often suggested in the discussion of multicollinearity : the **variance inflation factor** (VIF) and the **condition number** (CN). 看的是資料的結構
- The VIF is defined as $VIF(\hat{b}_1) = \dfrac{1}{1 - R_i^2}$

VIF越大越不好

- where R_i^2 is the correlation of determination for x_i regressed on other explanatory variables in the model.
- Note that VIF is part of the equation for $Var(b_i)$

Measures of Multicollinearity: VIF

- A large VIF suggests multicollinearity may be a problem
- Usually, VIF > 10 is considered a serious problem (but this cut-off value is arbitrary anyway)
- Another related measure called **tolerance** is defined as:

$$tolerance = 1 - R_i^2 = \frac{1}{VIF}$$

VIF越大越不好
tolerance越小越不好

Case Study: Mineral Loss in Patients Receiving Parenteral Nutrition

- Hospitalised patients who are not able to eat normally must be given intravenous nutrients, a process called parenteral nutrition.
- This process leads to significant cariuresis, i.e. loss of calcium via the urine.
- The degree of urine calcium loss varies considerably, but sometimes, this loss is greater than what has been given in the intravenous fluid.

Case Study: Mineral Loss in Patients Receiving Parenteral Nutrition

- Therefore, some of the lost calcium comes from the bones, which play a vital role in maintaining the balance of calcium in the blood.
- Cariuresis can be affected by the composition of parenteral nutrition solutions.
- Urinary Calcium excretion is related to intravenous Calcium intake and other ions such as phosphates in the intravenous fluids (Lipkin et al 1988 AJCN 515-523).

- Urinary calciums (Ca_U), dietary calciums (Ca_D), dietary protein level (P_D), urinary sodium (Na_U) and glomerular filtration rates (GFR; glomeruli are small interwined loops in the kidney responsible for production of urine, and GFR is an indicator of kidney function) are measured.
- The aim of statistical analysis is to investigate the relationship between Ca_U and the other four variables.

Data

Subject	cau	cad	gfr	nau	pd	Subject	cau	cad	gfr	nau	pd
1	220	554	63	54	80	15	43	208	50	3	41
2	182	303	43	99	48	16	42	182	46	35	50
3	166	287	45	14	53	17	37	0	42	54	0
4	162	519	58	55	70	18	29	125	69	1	42
5	137	249	53	37	82	19	24	0	36	20	0
6	136	142	70	37	36	20	22	170	30	7	54
7	128	184	85	110	56	21	15	0	66	1	0
8	113	150	49	78	25	22	3	0	25	0	0
9	100	383	56	79	101	23	1	82	16	2	27
10	90	391	51	1	76	24	36	0	66	83	0
11	75	90	91	134	28	25	31	125	40	6	42
12	71	0	58	29	0	26	21	0	45	16	0
13	60	279	39	1	41	27	12	0	23	3	0
14	60	249	41	17	37						

Ca與其他變數為正向且為顯著的

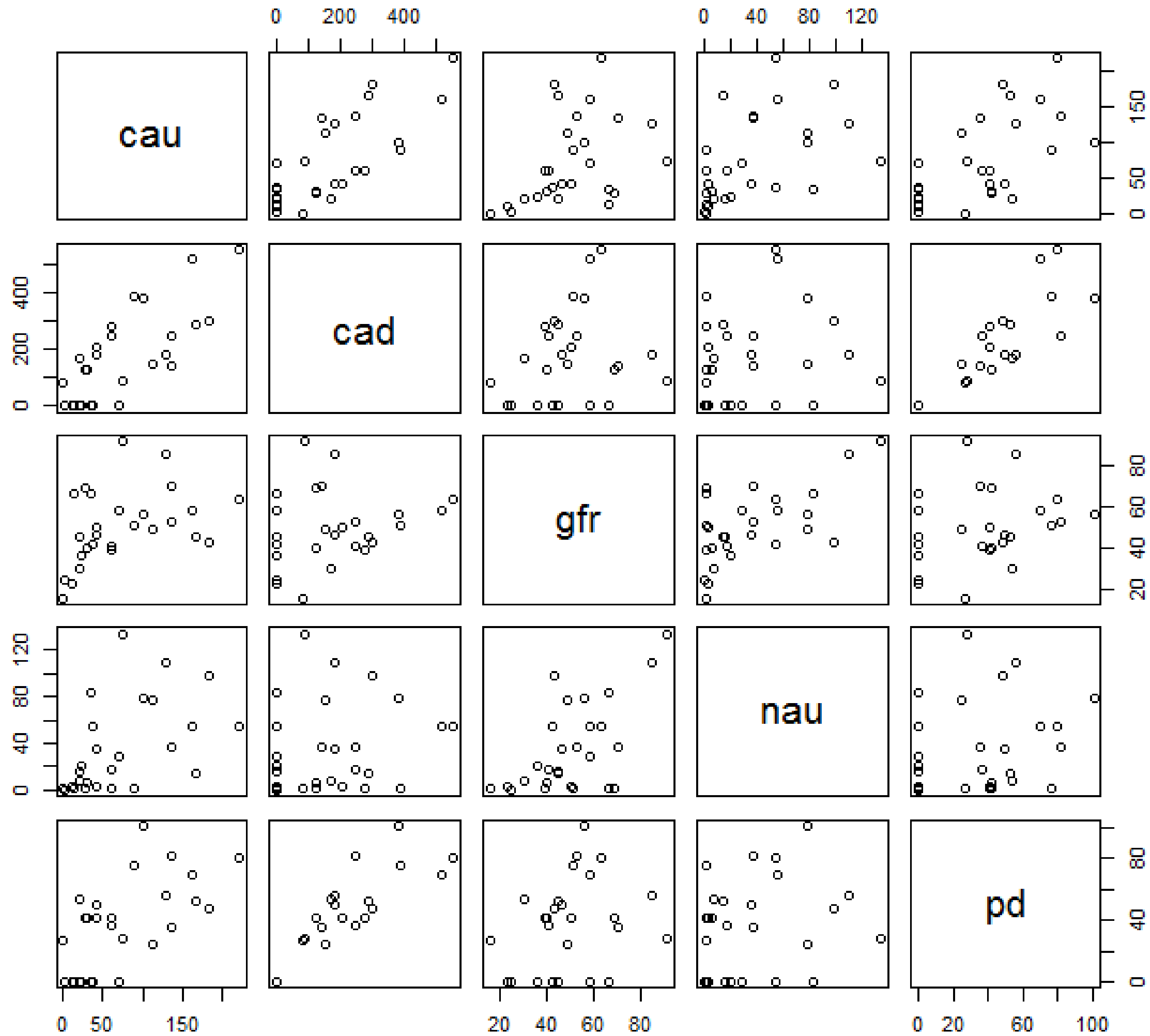
	Mean	SD		Ca_U	Ca_D	GFR	Na_U	P_D
Ca_U (mg/12 h)	74.67	61.12	r	1	0.758	0.410	0.493	0.634
			p		<0.001	0.034	0.009	<0.001
Ca_D (mg/12 h)	173.04	160.69	r	0.758	1	0.161	0.164	0.882
			p	<0.001		0.421	0.414	<0.001
GFR (ml/min)	50.22	17.66	r	0.410	0.161	1	0.616	0.212
			p	0.034	0.421		0.001	0.288
Na_U (meq/12 h)	36.15	38.58	r	0.493	0.164	0.616	1	0.183
			p	0.009	0.414	0.001		0.362
P_D (g/day)	36.63	29.83	r	0.634	0.882	0.212	0.183	1
			p	<0.001	<0.001	0.288	0.362	

1

2

3

```
> pairs(cariuresis[,2:6])
```



當變數多時，不再是XY之間的關係，也有XX之間的影響

數據上有太大的轉變時需考量共線性

共線性：會讓互相競爭會讓顯著變不顯著

- Multiple regression shows that Ca_D, GFR and Na_U all have positive associations with urinary calcium, but there is a negative relation between Ca_U and P_D. Although this negative regression coefficient (-0.51) is not statistically significant, it is contradictory to the positive correlation between Ca_U and P_D in simple regression.

新的係數

飲食的Ca係數變大，其他變小

	Simple Regression		Multiple Regression				
	Coef	*SE*	*Coef*	*SE*	*t*	*P-value*	*Beta*
Ca_D	0.29	0.05	0.35	0.09	3.88	0.00	0.91
GFR	1.42	0.63	0.43	0.49	0.88	0.39	0.12
Na_U	0.78	0.28	0.50	0.22	2.25	0.04	0.31
P_D	1.30	0.32	-0.51	0.48	-1.06	0.30	-0.25

- Can we therefore conclude that the level of dietary proteins is not associated with excretion of calcium in urine after adjusting for the level of dietary calcium, urinary sodium and GFR?
- Moreover, the relationship between excretion of calcium in urine and GFR is no longer statistically significant.

```
> library(car)
> vif(lm1)
     cad      gfr      nau       pd 
4.539028 1.646257 1.621366 4.618214 
```

VIF不超過10，不表示一定是OK的
要整體的看，不能依賴其一的指標

- As nutrition rich in protein is also rich in calcium (r = 0.882), it becomes difficult to differentiate their independent contribution to the excretion of calcium in urine.
- Consequently, whilst P_D has the second highest positive correlation with Ca_U, it turns out to have a non-significant negative regression coefficient in multiple regression.

- Although the change in the direction of relationship between Ca_U and P_D might indicate a serious problem of collinearity, the variance inflation factor (VIF) is only 4.5 for Ca_D, 4.6 for P_D and 1.6 for GFR and Na_U.
- Nevertheless, the change in sign is an indicator for collinearity, and the interpretation of results merits caution.

Solutions to Multicollinearity: Dropping Variables

- This may work sometimes, but in a complex scenario, it is difficult to determine which one to drop

把P值最大的地方拿掉，重跑一次－》GFR

- For example, we remove GFR from the model, but this does not seem to solve our problem

```
> lm2<-lm(cau~cad+nau+pd, data=cariuresis)
> summary(lm2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.51620   11.54849   0.911 0.371945
cad          0.33922    0.08821   3.846 0.000824 ***
nau          0.61457    0.17587   3.494 0.001954 **
pd          -0.45762    0.47676  -0.960 0.347111
```

拿掉GFR之後protein還是負值的

Solutions to Multicollinearity: Dropping Variables

- The sign of the regression coefficient for P_D is still negative, indicating multicollinearity is still an issue.
- This illustrates that in multiple regression, multicollinearity usually involves a group of variables rather than a pair of variables.
- Note that most software packages automatically drop variables that are perfectly collinear until the remaining variable are no longer perfectly collinear.

Solutions to Multicollinearity: Dropping Variables

- It only works when multicollinearity occurs by chance due to sampling error
- New data may therefore reduce the collinearity
- It does not work, if variables are multicollinear due to design or biology

Solutions to Multicollinearity: PCA and PLS

用降緯的方式來驗證

少用這些方式，只有在基因體研究時會用

- Two often suggested alternative methods to OLS regression are principal component regression and partial least squares regression
- Both are rarely used in biostatistics but now popular with bioinformatics
- Both are data dimension reduction techniques, i.e. they can only use partial information in **X** to estimate the relationships between explanatory variables and *y*.
- For a brief introduction, see Chapter 6 in my book “Statistical Thinking in Epidemiology”.

Variable Selection

變數的選擇：把不重要的拿掉，留著重要的（與Y的關係越大的）

- It should be based on your substantive knowledge in the research area not on statistical algorithms
- It should not be based on P-values either
- A statistically non-significant variable does not necessarily mean that it is clinically non-significant
  - A non-significant may not contribute to the total R^2, but it may result in a more precise, unbiased estimate of regression coefficients
  - Also remember : when sample size is sufficiently large, any variable can be statistically significant

不同演算法的結果不一樣，永遠沒有最好的

- Algorithms are only useful when there is a large set of candidate predictor variables.
- Goal is to choose a small subset from the larger set so that the resulting regression model is **simple**, yet have **good predictive ability**.
- i.e. the aim is to construct a regression that can predict the outcome variable, even if the selected variables have no causal relationships with the outcome

Two Basic Methods of Selecting Predictors

- **Stepwise regression**: Enter and remove predictors, in a stepwise manner, until there is no justifiable reason to enter or remove more.
- **Best subsets regression**: Select the subset of predictors that do the best at meeting some well-defined objective criterion.

Stepwise Regression: the Idea

- Start with no predictors in the "**stepwise model.**"
- At each step, enter or remove a predictor based on partial *F*-tests (that is, the *t*-tests).
  - Forward selection specifies an **Alpha-to-Enter** (α_E = 0.1 or any arbitrary value) significance level.
  - Backward selection specifies an **Alpha-to-Remove** (α_R = 0.1 or any arbitrary value) significance level.
- Stop when no more predictors can be justifiably entered or removed from the stepwise model.

Forward Stepwise Regression

把P個因素分別與Y值做迴歸分析，選出最好的

1. Fit each of the one-predictor models, that is, regress y on x_1, regress y on x_2, ... regress y on x_{p-1}.

2. The first predictor put in the stepwise model is the predictor that has the **smallest *t*-test *P*-value** (below $\alpha_E = 0.1$).

3. If no *P*-value < 0.1, stop, i.e. no variable is selected.

把最好的再丟回原本的模型中，看哪個P值最小
找出與X1無關，但與Y有高相關者

Forward Stepwise Regression:

4. Suppose x_1 was the "best" one predictor.
5. Fit each of the two-predictor models with x_1 in the model, that is, regress y on (x_1, x_2), regress y on (x_1, x_3), ..., and y on (x_1, x_{p-1}).
6. The second predictor put in stepwise model is the predictor that has the **smallest *t*-test *P*-value** (below $\alpha_E = 0.1$).
7. If no P-value < 0.1, stop.
8. Repeat the above process until no variable in the data satisfies the criteria

Backward Stepwise Regression:

先把大家都丟進去，再篩選，直到P值＞0.1

1. First enter all the variables into the model and then remove the variable with the largest *P*-value
2. Then re-fit the model with the remaining explanatory variable and remove the variable with the largest *P*-value
3. Repeat the above process and stop when all the variables in the model have *P*-value > 0.10

There is hierarchical stepwise regression which starts with forward stepwise selection but allow the previously included variable to be dropped when its P-value is greater than, say, 0.1.

Example: Mineral Loss in Patients Receiving Parenteral Nutrition

- All the algorithms select the same two variables for the final model: Ca_D and Na_U. The model R^2 is 0.715, which is quite close to 0.735 for the full model

Source	Coefficient	SE	t	P -value
cad	0.696	0.110	6.302	< 0.0001
nau	0.379	0.110	3.433	0.002

- Does this mean GFR and P_D are not clinically important?

Caution about Stepwise Regression!

如果變數不多，不清楚變數間彼此關係時，慎選驗證法

ANOVA變數多的時候，要刪掉型1錯誤

- Do not jump to the conclusion ...
  - that all the important predictor variables for predicting *y* have been identified, or
  - that all the unimportant predictor variables have been eliminated.
- Many *t*-tests for $\beta_k = 0$ are conducted in a stepwise regression procedure.
- The probability is high ...
  - that we included some unimportant predictors
  - that we excluded some important predictors

Drawbacks of stepwise regression

- The final model is not guaranteed to be optimal in any specified sense.
- The procedure yields a single final model, although in practice there are often several equally good models.
- It doesn’t take into account a researcher’s knowledge about the predictors.
- Therefore, *R* does not implement any of these algorithms!

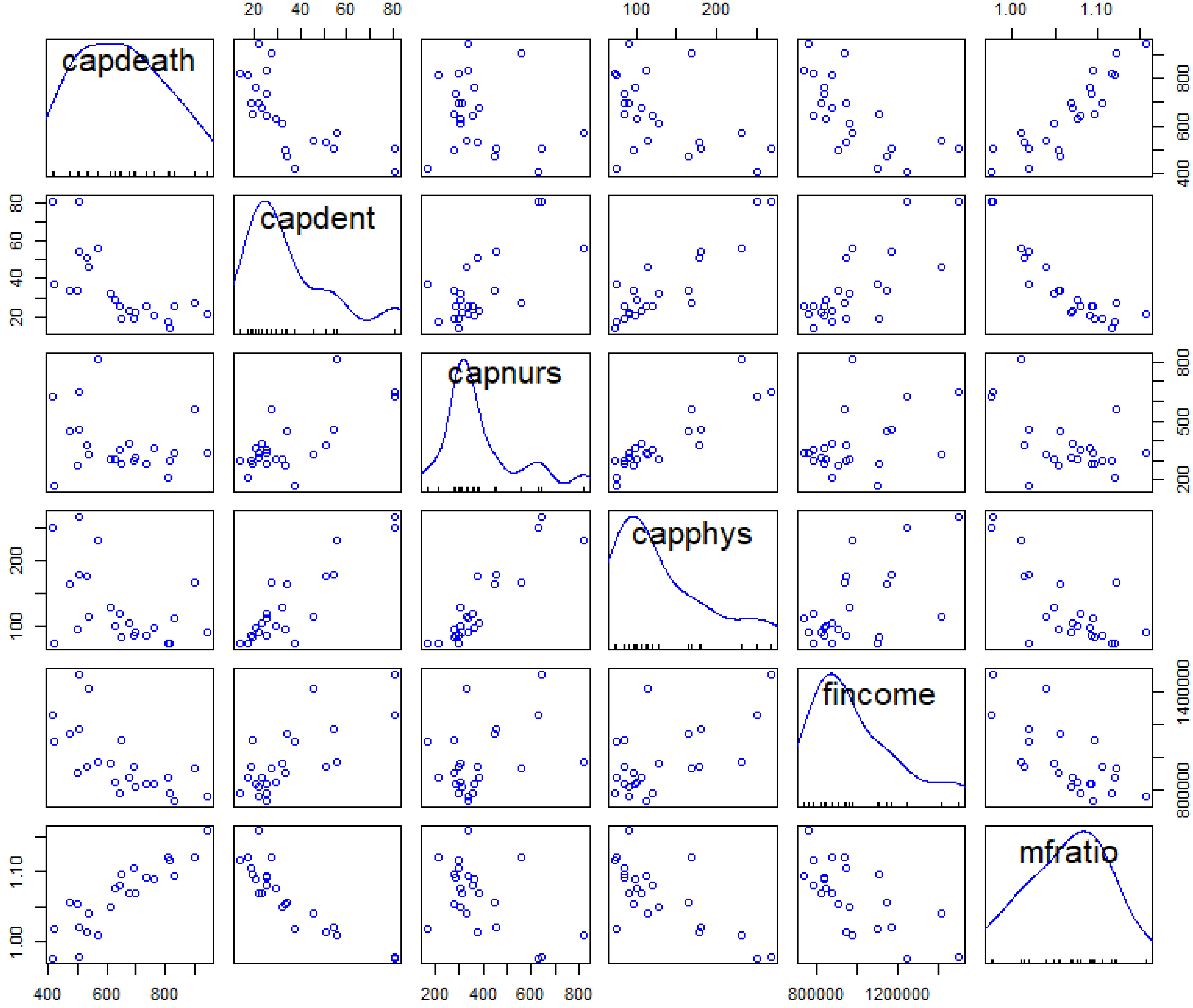
Case Study 2: Mortality Rates in Taiwan

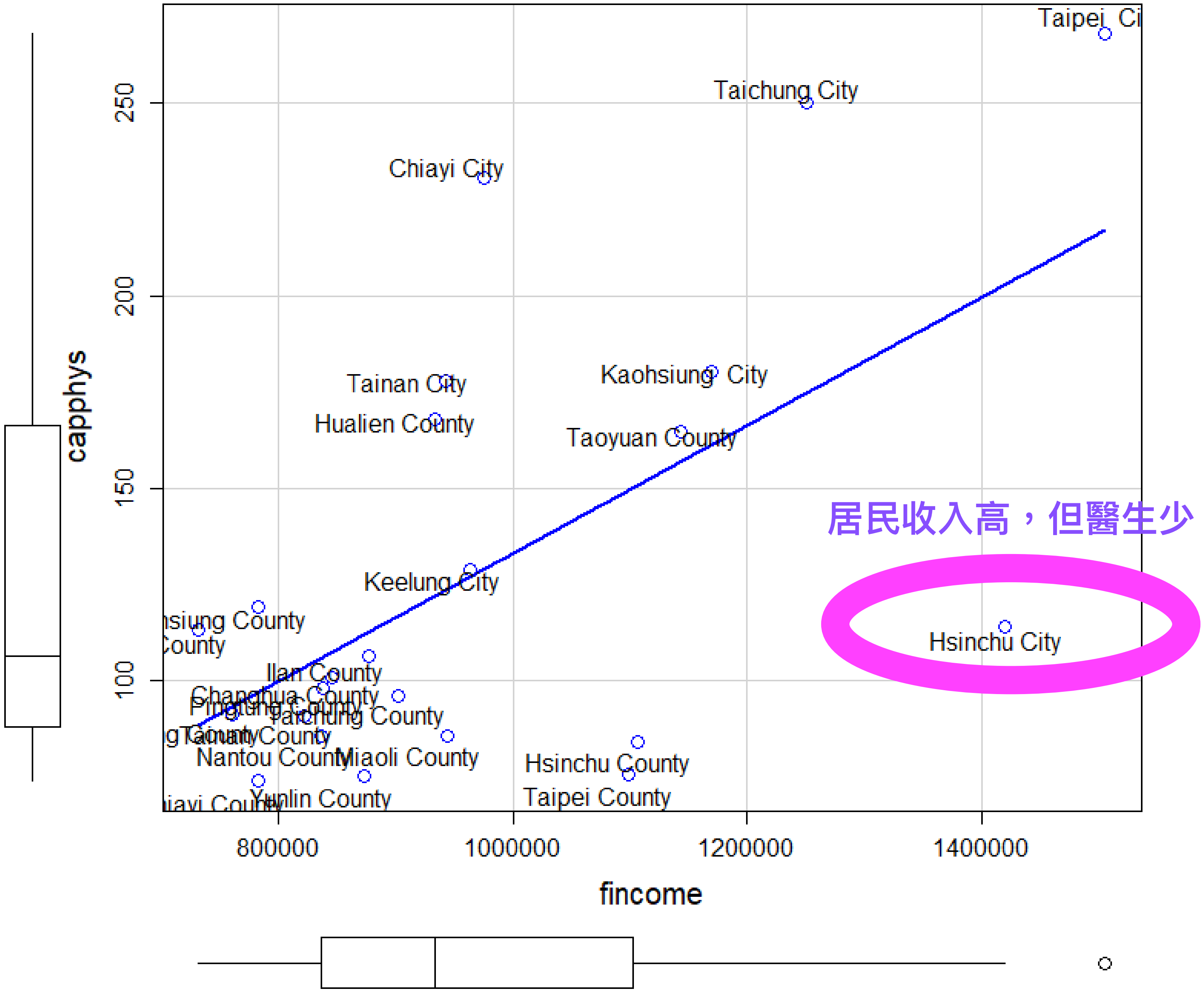
- Data on the number of deaths, size of population, the numbers of medical doctors, dentists, nurses in 23 districts in Taiwan in 2001
- We converted these numbers into the mortality rate, the number of medical doctors, dentists, nurses per 100,000 people
- We also include family income (fincome) and male to female ratio (mfratio) into the analysis
- The aim is to predict the mortality rate

Data

District	capdeath	capphys	capdent	capnurs	fincome	mfratio
Taipei City	505.2	267.9	81.1	646.4	1505506	0.98
Kaohsiung City	503.2	180.2	54.5	457.4	1169386	1.02
Taipei County	419.9	76.0	37.4	167.9	1098408	1.02
Ilan County	678.6	106.6	23.0	385.2	877199	1.07
Taoyuan County	474.5	164.7	33.8	447.3	1143050	1.06
Hsinchu County	652.1	84.2	19.2	281.5	1106603	1.10
Miaoli County	695.1	85.7	18.6	298.7	944066	1.10
Taichung County	499.8	96.2	33.6	276.0	902725	1.05
Changhua County	629.5	100.9	29.0	306.7	845708	1.08
Nantou County	733.8	85.7	25.5	285.2	836262	1.09
Yunlin County	813.0	75.5	17.1	210.5	873803	1.12
Chiayi County	818.2	74.1	13.5	301.0	783819	1.12
Tainan County	697.0	91.1	22.1	310.2	821843	1.07
Kaohsiung County	645.9	119.5	25.2	353.7	783331	1.08
Pingtung County	764.5	98.3	20.7	363.9	838027	1.09
Taitung County	946.3	91.4	21.6	333.9	760989	1.16
Hualien County	905.3	168.1	27.4	561.7	933484	1.12
Penghu County	834.1	113.3	25.3	338.9	732306	1.09
Keelung City	611.8	128.8	32.1	305.1	964170	1.05
Hsinchu City	537.1	114.1	45.8	330.8	1419946	1.04
Taichung City	411.2	250.1	80.5	627.9	1251144	0.98
Chiayi City	569.5	230.6	55.8	817.3	975877	1.01
Tainan City	531.6	177.8	51.2	377.0	943224	1.01

其他都是負向關係，唯性別是正向




Taipei Ci
Taichung City
Chiayi City
Tainan City
Hualien County
Kaohsiung City
Taoyuan County
居民收入高，但醫生少
Keelung City
Hsinchu City
Ilan County
Hsinchu County
Taipei County
Nantou County
Miaoli County
Yunlin County
capphys
fincome
100
150
200
250
800000
1000000
1200000
1400000

Pairwise Correlations

Pearson correlations:

	capdeath	capdent	capnurs	capphys	fincome	mfratio
capdeath	1.0000	-0.6922	-0.2244	-0.4715	-0.6757	0.8883
capdent	-0.6922	1.0000	0.6905	0.8939	0.7535	-0.8909
capnurs	-0.2244	0.6905	1.0000	0.9000	0.3928	-0.5037
capphys	-0.4715	0.8939	0.9000	1.0000	0.5947	-0.7307
fincome	-0.6757	0.7535	0.3928	0.5947	1.0000	-0.6992
mfratio	0.8883	-0.8909	-0.5037	-0.7307	-0.6992	1.0000

Number of observations: 23

Pairwise two-sided p-values:

	capdeath	capdent	capnurs	capphys	fincome	mfratio
capdeath		0.0003	0.3034	0.0231	0.0004	<.0001
capdent	0.0003		0.0003	<.0001	<.0001	<.0001
capnurs	0.3034	0.0003		<.0001	0.0637	0.0143
capphys	0.0231	<.0001	<.0001		0.0028	<.0001
fincome	0.0004	<.0001	0.0637	0.0028		0.0002
mfratio	<.0001	<.0001	0.0143	<.0001	0.0002	

Regression Model

- We now regress capdeath on capdent, capnurs, capphys, fincome and mfratio

$$capdeath = b_0 + b_1 capdent + b_2 capnurs + b_3 capphys + b_4 fincome + b_5 mfratio$$

```
Coefficients:
                  Estimate     Std. Error t value  Pr(>|t|)
(Intercept) -3432.79050192   710.46634071  -4.832  0.000156 ***
capdent         3.24967919     2.75638933   1.179  0.254651
capnurs         0.15941837     0.23779026   0.670  0.511598
capphys         0.08075682     1.00223090   0.081  0.936719
fincome        -0.00016719     0.00009682  -1.727  0.102318
mfratio      3810.99988668   624.43183519   6.103 0.0000117 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.27 on 17 degrees of freedom
Multiple R-squared:  0.8824,  Adjusted R-squared:  0.8478
F-statistic: 25.51 on 5 and 17 DF,  p-value: 0.0000002434
```