COURSE

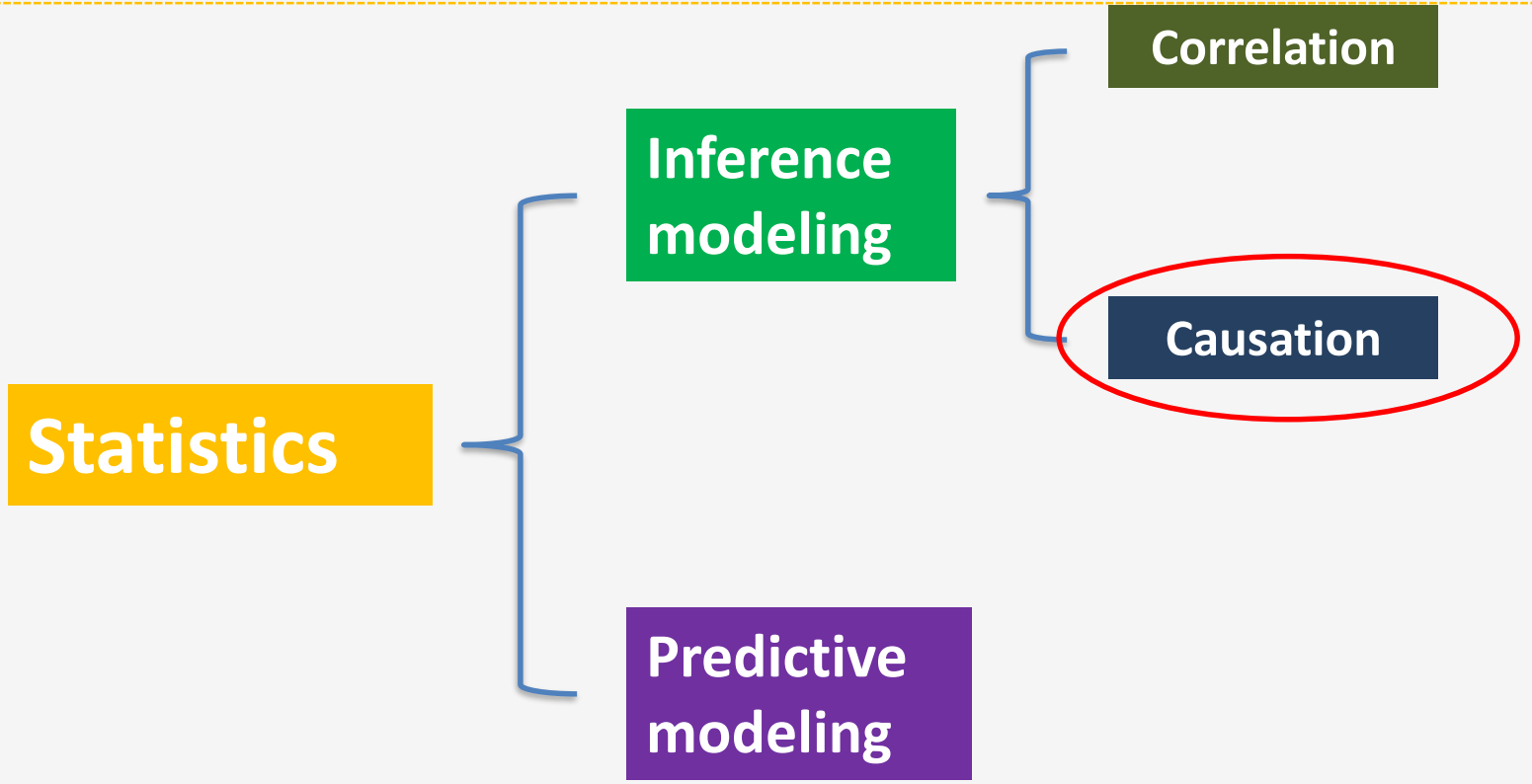
Xây dựng hồi quy đa biến

Lớp phân tích thống kê

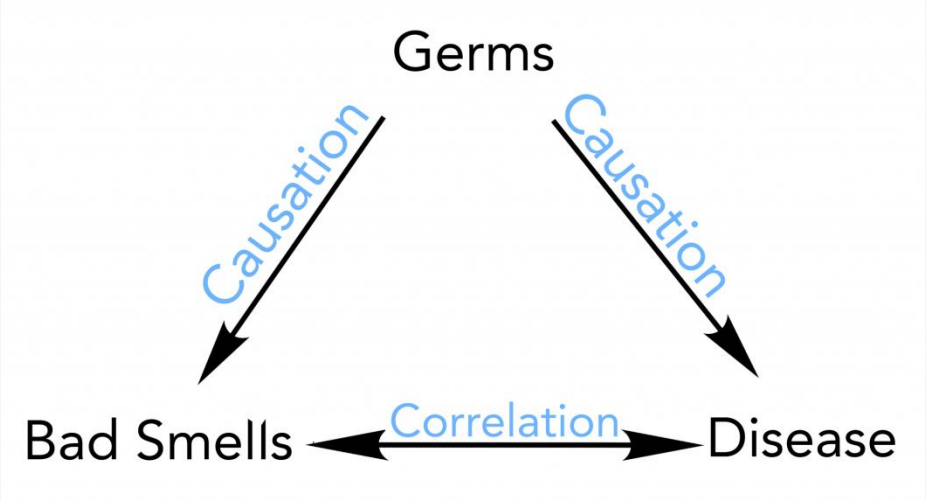
Khương Quỳnh Long
Hà Nội, 06-08/06/2020

Mục tiêu

- Một số chỉ số trong lựa chọn mô hình
- Nguyên tắc chung trong xây dựng hồi quy đa biến
- Một số phương pháp xây dựng hồi quy đa biến
- Ưu điểm, nhược điểm


Statistics
Inference modeling
Correlation
Causation
Predictive modeling

Correlation ≠ Causation



Source of bias

- Sai lệch tiềm tàng trong mô hình
  - ✓ Sai lệch hệ thống (systematic)
  - ✓ Không hiệu chỉnh nhiễu
  - ✓ Interaction/Effect modification
  - ✓ Cộng tuyến (collinearity)
  - ✓ …

“All model are wrong, but some are useful” !

Hồi quy đa biến

- Phương trình:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

- Mô hình y theo $x_1, x_2, \ldots, x_n$
- **Kiểm soát yếu tố gây nhiễu**
  - ✓ Chỉ khi nào yếu tố gây nhiễu được thêm vào mô hình

Mô hình quá phức tạp/dư thừa

glucose	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exercise	-2.841198	1.074594	-2.64	0.008	-4.948304	-.7340926
HT	-1.87393	.974953	-1.92	0.055	-3.785657	.0377968
age	-.1069152	.0833915	-1.28	0.200	-.2704326	.0566021
nwhite	.5645015	1.624027	0.35	0.728	-2.619955	3.748958
smoking	-1.023867	1.542842	-0.66	0.507	-4.049132	2.001399
drinkany	2.075594	1.056013	1.97	0.049	.0049212	4.146266
physact						
somewhat less active	-1.115604	2.202061	-0.51	0.612	-5.433493	3.202284
about as active	2.279466	2.151929	1.06	0.290	-1.940123	6.499055
somewhat more active	.8533755	2.258217	0.38	0.706	-3.574626	5.281377
much more active	-.5739344	2.574254	-0.22	0.824	-5.621634	4.473765
globrat						
Fair	2.793907	3.54875	0.79	0.431	-4.164624	9.752438
Good	1.46192	3.570144	0.41	0.682	-5.53856	8.462399
Very good	1.929168	3.70859	0.52	0.603	-5.342783	9.201118
Excellent	1.204172	4.352672	0.28	0.782	-7.33072	9.739064
medcond	-.4383873	1.056299	-0.42	0.678	-2.50962	1.632846
htnmeds	.4946444	1.315293	0.38	0.707	-2.084435	3.073724
statins	-.6432222	1.048295	-0.61	0.540	-2.698761	1.412317
diabetes	43.30088	1.763215	24.56	0.000	39.84349	46.75826
dmpills	13.77868	2.146727	6.42	0.000	9.569294	17.98807
insulin	24.43578	2.151594	11.36	0.000	20.21685	28.65471
weight	-.0685387	.1021843	-0.67	0.502	-.2689058	.1318285
BMI	.6045992	.2503422	2.42	0.016	.1137182	1.09548
waist	-.0394408	.12716	-0.31	0.756	-.2887811	.2098995
WHR	26.24673	11.55603	2.27	0.023	3.587205	48.90625
tchol	.0745073	.0420354	1.77	0.076	-.0079174	.156932
LDL	-.0571034	.0445027	-1.28	0.200	-.1443661	.0301593
TG	.0222467	.0096515	2.31	0.021	.0033217	.0411717
SBP	-.0034041	.0321762	-0.11	0.916	-.0664965	.0596883
DBP	-.0144425	.062711	-0.23	0.818	-.1374087	.1085238
_cons	59.27812	10.96702	5.41	0.000	37.77354	80.78269

Cách 1: Khái niệm + thống kê

Biến gây nhiễu (khái niệm)

- Yếu tố bên ngoài, tác động đồng thời lên yếu tố phơi nhiễm và kết cuộc
- Không nằm trên đường từ phơi nhiễm → kết cuộc

Biến gây nhiễu

- Theo thống kê: so sánh hệ số phương trình và sai số (standard error) của z → y
  - ✓ Rule of thumb:
    - Nếu z làm thay đổi beta > 10% (hoặc 15%) → z là biến gây nhiễu **Giữ Z trong mô hình**
    - Không thay đổi beta, ↓SE → Z là 1 predictor của y **Tùy tình huống**
    - Tăng SE → Z và X cộng tuyến (collinear) **Loại Z khỏi mô hình**

➔ Khái niệm + thống kê

Biến gây nhiễu

Data FEV

- FEV ~ smoke

FEV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Smoke	.7107189	.1099426	6.46	0.000	.4948346	.9266033
_cons	2.566143	.0346604	74.04	0.000	2.498083	2.634202

- FEV ~ smoke + Age ?

FEV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Smoke	-.2089949	.0807453	-2.59	0.010	-.3675476	-.0504421
Age	.2306046	.0081844	28.18	0.000	.2145336	.2466755
_cons	.367373	.0814357	4.51	0.000	.2074647	.5272814

FEV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Smoke	.7107189	.1099426	6.46	0.000	.4948346	.9266033
_cons	2.566143	.0346604	74.04	0.000	2.498083	2.634202

FEV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Smoke	-.2089949	.0807453	-2.59	0.010	-.3675476	-.0504421
Age	.2306046	.0081844	28.18	0.000	.2145336	.2466755
_cons	.367373	.0814357	4.51	0.000	.2074647	.5272814

-**Thay đổi beta** (>10%) = tuổi là biến gây nhiễu
-**Không đổi beta**, ↓ SE = tuổi là 1 predictor khác
-**Tăng SE** của beta = tuổi và hút thuốc cộng tuyến

Biến gây nhiễu

Data FEV

- FEV ~ smoke + age + gender?

FEV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Smoke	-.1539741	.0779766	-1.97	0.049	-.3070905	-.0008577
Age	.2267942	.0078845	28.76	0.000	.2113121	.2422763
Gender	.3152733	.0427104	7.38	0.000	.2314063	.3991403
_cons	.2377708	.0802279	2.96	0.003	.0802337	.3953079

Cộng tuyến (Collinear)

- Xảy ra khi 2 hay nhiều biến độc lập trong mô hình đa biến liên quan chặt với nhau
- **Đánh giá**
  - ✓ Kiểm tra mối liên quan giữa các biến độc lập
    - Chọn 1 trong các biến có liên hệ chặt với nhau
  - ✓ Xây dựng mô hình hồi quy của các biến độc lập
    - i.e., mỗi mô hình cho 1 biến độc lập (Ví dụ X1 ~ X2)
    - Tính R^2 (kiểm tra nếu R^2 lớn)
    - Tính Variance Inflated Factor – VIF (rule of thumb: >10 hoặc > 5)
- **Giải quyết**
  - ✓ Loại bớt biến liên quan với nhau
  - ✓ Các loại hồi quy khác: e.g., ridge regression, lasso, elastic net ….

Cộng tuyến (Collinear)

- **Giữ biến nào loại biến nào???**
  - ✓ Biến phơi nhiễm chính > các biến độc lập khác
  - ✓ Dựa vào kinh nghiệm
  - ✓ Biến nào được đo lường chính xác hơn
  - ✓ Kiểm tra độ phù hợp mô hình

Đánh giá độ phù hợp mô hình (model fit)

- **Hồi quy Tuyến tính**
  - ✓ Partial F test (nested models)
  - ✓ AIC/BIC (nested models hoặc non-nested models)
- **Hồi quy Logistic**
  - ✓ Likelihood ratio test (nested models)
  - ✓ AIC/BIC (nested models hoặc non-nested models)

Hồi quy tuyến tính đa biến

- **Phương trình:**

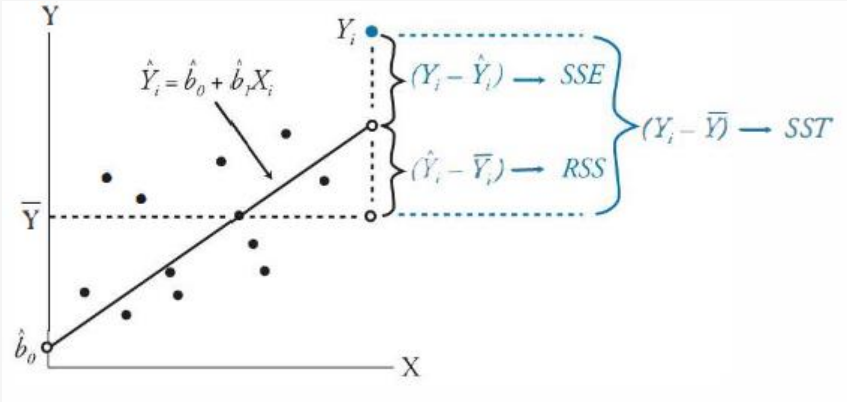
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

- Diễn giải: Khi các x_2, ..., x_n không thay đổi, biến x_1 thay đổi 1 đơn vị thì biến y thay đổi bao nhiêu đơn vị?
- Tiên lượng: Với các thông tin của x_1, x_2, ... x_n thì y là bao nhiêu?
- **Kiểm soát yếu tố gây nhiễu**

Partial F-test

- F stat = ($SSe_{(reduce)} - SSe_{(full)}$) / $MSe_{(full)}$

- Nested models only
  - ✓ Full: FEV ~ Smoke + Age
  - ✓ Reduce: FEV ~ Smoke
- Giả thuyết
  - ✓ H_0: $Sse_{(full)} = Sse_{(reduce)}$
  - ✓ H_a: $Sse_{(full)} < Sse_{(reduce)}$



Partial F-test

- ✓ Full: FEV ~ Smoke + Age
- ✓ Reduce: FEV ~ Smoke

```
ssc install ftest // Cài đặt câu lệnh tính Partial F Test (chỉ cần cài 1 lần duy nhất)
reg FEV Smoke
estimate store M1 // Lưu mô hình M1
reg FEV Smoke Age
estimate store M2 // Lưu mô hình M2
ftest M1 M2       // Partial F-test
```

```
. ftest M1 M2
Assumption: M1 nested in M2

F(  1,     651) =    793.90
      prob > F =    0.0000
```

AIC/BIC

- **AIC (Akaike's Information Criterion)**
  - ✓ AIC = 2 x (Số biến - log-likelihood)
  - ✓ Khi đưa biến mới vào thì log-likelihood sẽ tăng --> cân nhắc giữa số biến và likelihood
  - ✓ Penalize mô hình nhiều biến số & ít ý nghĩa
  - ✓ AIC càng nhỏ → mô hình càng phù hợp
- **BIC (Bayesian Information Criterion)**
  - ✓ BIC = log(n)*Số biến – 2*log-likelihood
  - ✓ Tương tự AIC
  - ✓ Conservative hơn AIC
- **Nested models hoặc non-nested models**
  - ✓ Trong nested model, AIC/BIC cho kết quả tương tự partial F test

AIC/BIC

```
reg FEV Smoke
estat ic
reg FEV Smoke Age
estat ic
estat vif // kiểm tra VIF
```

```
. estat ic

Akaike's information criterion and Bayesian information criterion

-------------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+-----------------------------------------------------------------
           . |        654  -834.1935   -553.167       3   1112.334   1125.783
-------------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] BIC note.
```

Tóm tắt 4 bước

1. Xác định câu hỏi nghiên cứu
2. Explore dữ liệu
3. Xây dựng mô hình ban đầu
4. Kiểm tra giả định

B1: Xác định câu hỏi nghiên cứu

- Mục tiêu của mô hình?
- Biến outcome?
- Biến giải thích?
- Các biến khác có trong mô hình:
  - ✓ Đo lường/code như thế nào?
  - ✓ Biến nào quan trọng?

B2: Explore dữ liệu

- **Univariate distribution**
  - ✓ Phân phối của biến số
  - ✓ Missing data?
  - ✓ Out of range….
- **Bivariate analysis**
  - ✓ Mối liên quan giữa các biến

Missing data

x1	x2	x3	X1 + x2+ x3
1	.	0	.
1	.	1	.
1	1	0	ok
1	0	1	ok
.	0	1	.
.	0	.	.
.	1	0	.

B3: Xây dựng mô hình ban đầu

- **2 cách:**
  - ✓ Forward: từ mô hình đơn giản → thêm biến
  - ✓ Backward: từ mô hình phức tạp → loại bớt biến
- **Cân nhắc thêm/loại biến**
  - ✓ Biến gây nhiễu? (khái niệm + thống kê)
  - ✓ Partial F test
  - ✓ AIC/BIC
- **Nếu có cộng tuyến**
  - ✓ Bỏ bớt các biến có liên hệ chặt với nhau
  - ✓ Which is the best? → AIC/BIC…
- **Kiểm tra interaction/effect modification**
  - ✓ Stratify
  - ✓ Interaction term & kiểm tra bằng Partial F test/AIC/BIC

B4: Kiểm tra giả định

LINE

1. **L**inear: Quan hệ tuyến tính giữa biến độc lập và phụ thuộc
2. **I**ndependence: Các sai số là độc lập
3. **N**ormality: Sai số của ước lượng có phân phối bình thường
4. **E**qual variance: Phương sai đồng nhất (homoscedasticity)

Hồi quy logistic đa biến

- Hồi quy logistic:

$$\log(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

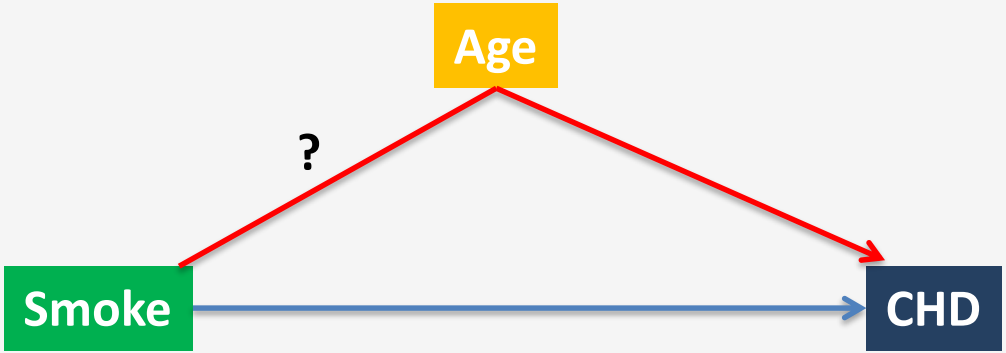
- Xây dựng mô hình và giải thích dựa vào “log-odds”
- **Kiểm soát yếu tố gây nhiễu**

Hồi quy logistic đa biến

- Quy trình tương tự hồi quy tuyến tính
- **Xác định biến gây nhiễu:**
  - ✓ Khái niệm
  - ✓ Sự thay đổi beta & SE (Chú ý: beta = log(odds), không phải OR)
- **So sánh mô hình**
  - ✓ Likelihood ratio test # Partial F test (nested models)
  - ✓ AIC/BIC

Hồi quy logistic đa biến

- Data “WCGS.dta”



- Về lý thuyết tuổi ảnh hưởng smoke?

Hồi quy logistic đa biến

chd ~ smoke

chd	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smoke	.6298628	.1337217	4.71	0.000	.3677731	.8919525
_cons	-2.76362	.1041517	-26.53	0.000	-2.967753	-2.559486

chd ~ smoke + age

chd	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smoke	.6381599	.13472	4.74	0.000	.3741135	.9022062
age	.0751775	.0113917	6.60	0.000	.0528502	.0975048
_cons	-6.321315	.5614499	-11.26	0.000	-7.421737	-5.220894

- Thay đổi beta – log(odds): <10%
- Mô hình nào “tốt” hơn?
  - ✓ Likelihood ratio test/AIC/BIC

Likelihood ratio test

- Tương tự Partial F test
- -2*log-likelihood ~ chi^2 ➔ p value

✓ Full: $\log(odds_{CHD})$ ~ Smoke + Age
✓ Reduce: $\log(odds_{CHD})$ ~ Smoke

```
logit chd smoke
estimate store M1  // Lưu mô hình M1
logit chd smoke age
estimate store M2  // Lưu mô hình M2
lrtest M1 M2       // likelihood ratio test
```

Giả thuyết

✓ H_0: log-likelihood$_{(full)}$ = log-likelihood$_{(reduce)}$
✓ H_a: log-likelihood$_{(full)}$ > log-likelihood$_{(reduce)}$

```
. lrtest M1 M2

Likelihood-ratio test                           LR chi2(1)  =     43.15
(Assumption: M1 nested in M2)                   Prob > chi2 =    0.0000
```

AIC/BIC

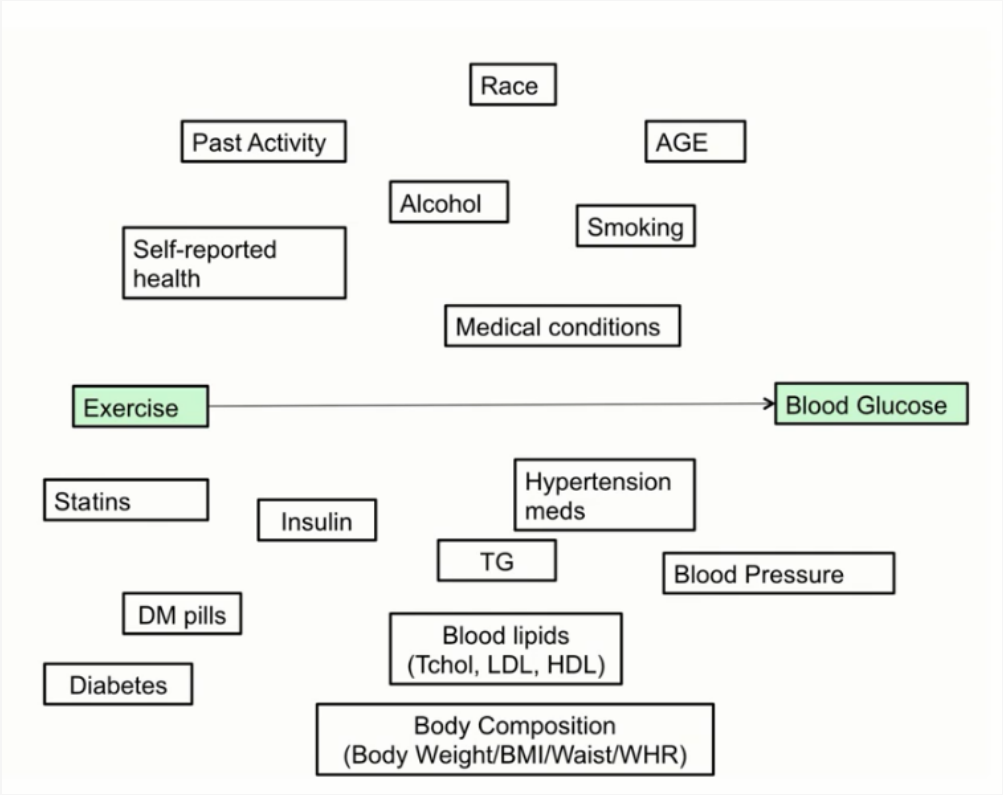
```
logistic chd smoke age
estat ic
```

```
. estat ic

Akaike's information criterion and Bayesian information criterion

-------------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+-----------------------------------------------------------------
           . |      3,154  -890.6219  -857.6172       3   1721.234   1739.404
-------------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] BIC note.
```

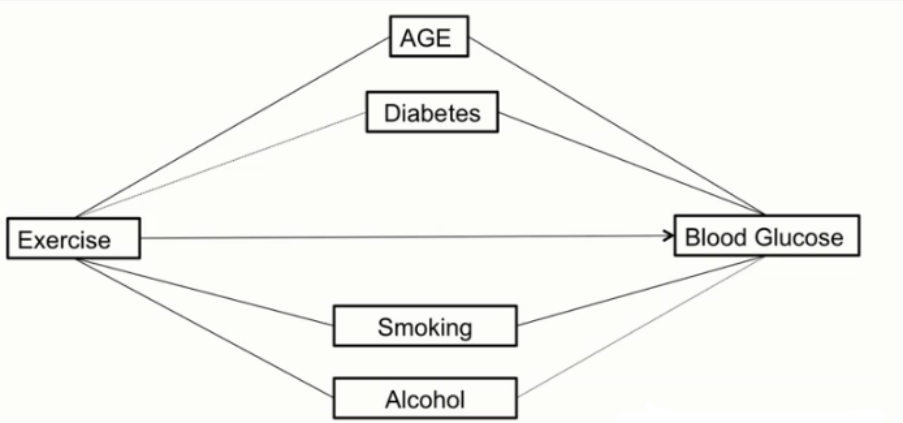
Bài tập nhóm

- HERS data (“HERS.dta”)
- Nghiên cứu thử nghiệm lâm sàng đánh giá liệu pháp điều trị hormone thay thế trong phòng ngừa nhồi máu cơ tim và tử vong. Dữ liệu được thu thập ở baseline trên 2763 phụ nữ mãn kinh có bệnh nền là CHD
- Mục đích (của ví dụ này)
  - ✓ Có mối liên quan giữa luyện tập thể dục đến đường huyết?

HERS data


Race
Past Activity
AGE
Alcohol
Smoking
Self-reported health
Medical conditions
Exercise
Blood Glucose
Statins
Insulin
Hypertension meds
TG
Blood Pressure
DM pills
Blood lipids (Tchol, LDL, HDL)
Diabetes
Body Composition (Body Weight/BMI/Waist/WHR)

Framework



4 bước xây dựng mô hình

- Bước 1:
- Bước 2:
- Bước 3:
- Bước 4:

Một số phương pháp khác

Một số phương pháp lựa chọn biến

- Dựa vào kinh nghiệm/y văn
- Dựa vào p-value[1]
  - ✓ Phân tích đơn biến
  - ✓ Những biến có p < 0.2 (0.25…) ở đơn biến → mô hình đa biến ban đầu
  - ✓ Những biến p <0.05 trong mô hình đa biến được giữ lại
  - ✓ Những biến loại được kiểm tra lại bằng lrtest

1. Stanley Lemeshow David W. Hosmer Jr., Rodney X. Sturdivant (2013) Applied Logistic Regression, 3rd Edition

Một số phương pháp lựa chọn biến

- Stepwise
  - ✓ Dựa vào một số chỉ số (p-value, AIC/BIC…)
  - ✓ 2 loại
    - ✓ Forward
    - ✓ Backward
      - B1: Fit mô hình với toàn bộ biến được chọn
      - B2: Loại biến có giá trị p-value cao nhất
      - B3: Fit lại mô hình với biến đã loại ở bước 2
      - B4…Bn: loại dần các biến đến khi tất cả các biến có p-value dưới ngưỡng

```
xi: stepwise, pr(0.2): reg glucose exercise HT age
nwhite smoking drinkany i.physact i.globrat medcond
htnmeds statins diabetes dmpills insulin weight BMI
waist WHR tchol LDL TG SBP DBP
```

Không được khuyến cáo sử dụng

```
p = 0.9158 >= 0.2000  removing SBP
p = 0.8225 >= 0.2000  removing _Iphysact_5
p = 0.8118 >= 0.2000  removing _Iglobrat_5
p = 0.7579 >= 0.2000  removing _Iglobrat_3
p = 0.7560 >= 0.2000  removing waist
p = 0.7228 >= 0.2000  removing nwhite
p = 0.7300 >= 0.2000  removing DBP
p = 0.6852 >= 0.2000  removing htnmeds
p = 0.6927 >= 0.2000  removing _Iglobrat_4
p = 0.6348 >= 0.2000  removing medcond
p = 0.6140 >= 0.2000  removing _Iphysact_2
p = 0.5661 >= 0.2000  removing statins
p = 0.4987 >= 0.2000  removing smoking
p = 0.3252 >= 0.2000  removing weight
p = 0.2949 >= 0.2000  removing _Iglobrat_2
p = 0.2525 >= 0.2000  removing age
p = 0.2501 >= 0.2000  removing _Iphysact_4
p = 0.2299 >= 0.2000  removing LDL
```

Source	SS	df	MS		
Model	2071501.37	11	188318.306	Number of obs	= 2,740
Residual	1756580.42	2,728	643.907778	F(11, 2728)	= 292.46
				Prob > F	= 0.0000
				R-squared	= 0.5411
				Adj R-squared	= 0.5393
Total	3828081.79	2,739	1397.62022	Root MSE	= 25.375

glucose	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exercise	-2.615241	1.01427	-2.58	0.010	-4.604056	-.6264258
HT	-1.823858	.9707389	-1.88	0.060	-3.727316	.0796
TG	.0275715	.0082211	3.35	0.001	.0114513	.0436916
WHR	22.21973	6.620853	3.36	0.001	9.237334	35.20212
diabetes	43.1877	1.739287	24.83	0.000	39.77725	46.59816
drinkany	2.244258	1.012678	2.22	0.027	.258564	4.229953
dmpills	13.92781	2.132756	6.53	0.000	9.745826	18.10979
_Iphysact_3	2.323166	1.033013	2.25	0.025	.297599	4.348734
BMI	.3722795	.0957505	3.89	0.000	.1845287	.5600303
tchol	.0249485	.012281	2.03	0.042	.0008675	.0490294
insulin	24.74155	2.122927	11.65	0.000	20.57884	28.90425
_cons	55.88684	6.362311	8.78	0.000	43.41141	68.36228

Một số phương pháp lựa chọn biến

- **Bayesian**
  - ✓ Xác suất biến X xuất hiện trong mô hình
  - ✓ Xác suất mô hình ABC xuất hiện
  - ✓ Bayesian model averaging (BMA)
- **Directed Acyclic Graphs (DAGs)**
  - ✓ Dựa hoàn toàn vào mô hình khái niệm
  - ✓ https://gitlab.com/LongKhuong/dags