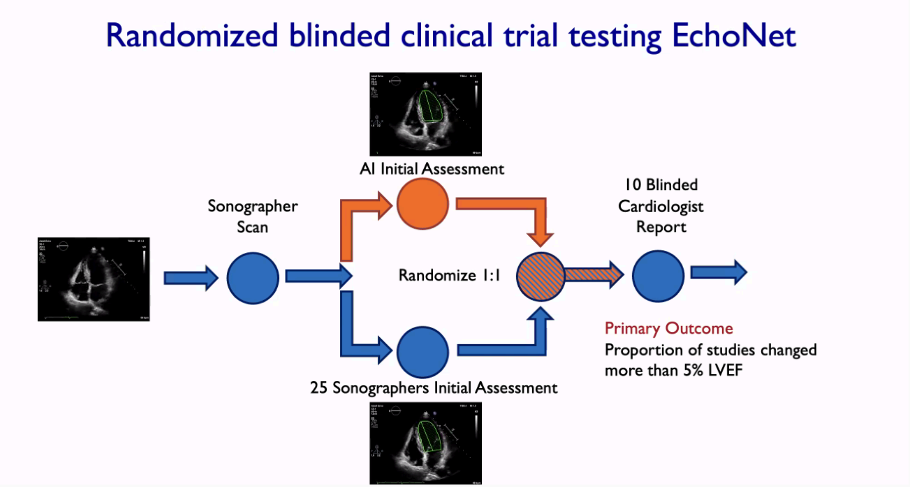
Methods of Applied Statistics I

STA2101H F LEC9101

Week 4

October 5 2022



Today

Start Recording

1. Upcoming events
2. Matrix derivatives
3. Linear Regression Part 4: collinearity, model-building, $p > n$

4. In the News

your choice

HW 2 ~ 100%

find a dataset → Kaggle.com or . . .

data is clean

$$\left[\begin{array}{l} n \simeq (100, 3000) \\ p \simeq (10, 300) \end{array} \right]$$ fairly clean

Today

1. Upcoming events
2. Matrix derivatives
3. Linear Regression Part 4: collinearity, model-building, $p > n$

4. In the News

Office Hour Tuesday 7-8 Zoom

not Monday

Upcoming

- October 7 12.00-13.00 : STAGE International Seminar link
- Teri Manolio, National Human Genome Research Institute
- Genomic Diversity and Genomic Healthcare

Aside: matrix derivatives

Note on Matrix Derivatives

STA 2101F: Methods of Applied Statistics I 2022

The matrix version of the linear model is

$$y = X\beta + \epsilon,$$

where y and $X\beta$, and ϵ are $n \times 1$ vectors; X is an $n \times p$ matrix and β is a $p \times 1$ vector. To find the least squares estimator we minimize

$$SS(\beta) = (y - X\beta)^T(y - X\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2,$$

Aside: partitioning SS

$$(y-\bar{y}\mathbf{1})^{\mathrm{T}}(y-\bar{y}\mathbf{1}) = (y-X\hat{\beta})^{\mathrm{T}}(y-X\hat{\beta}) + \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} - n\bar{y}^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \underbrace{\sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}}\hat{\beta})^2}_{RSS} + \cancel{\hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta}_2}$$

$$\underbrace{+\ \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} - n\bar{y}^2}_{\text{Reg SS}}$$

SS due to model

(ignoring β_0 ($\beta_0 1$)

↑ constant term

$$\begin{bmatrix} 1 & \\ \vdots & X_{(2)} \\ 1 & \end{bmatrix} = X \qquad \begin{pmatrix} \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Linear regression recap

- factor variables: a factor with k levels needs $k-1$ parameters
- linear model assumptions: $\mathrm{E}(y) = X\beta$, $\mathrm{cov}(Y) = \sigma^2 I$, $\quad y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I)$

usual

- true residuals have constant variance
- true residuals are normally distributed
- true residuals are independent

qnorm

- Shapiro-Wilk test? ← QQ plot
- reminder: Q-Q plot of a vector of observations z

$i = 1, \ldots, n$

- y-axis: ordered observations $z_{(1)}, \ldots z_{(n)}$
- x-axis: theoretical quantiles from distribution F: $F^{-1}\{i/(n+1)\} \simeq \mathrm{E}(X_{(i)})$
- SW test is a summary of weighted LS regression from this plot

$X_1, \ldots, X_n$ iid F

... linear regression recap

leverage + outlier

- $H = X(X^T X)^{-1} X^T$ hat matrix $\hat{y} = X\hat{\beta} = Hy$
- h_{ii} measures the leverage of observation i on the fitted model
- $\text{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$
- C_i Cook's distance measures the influence of observation i on the fit

$$\text{SM} \rightarrow C_i = \frac{(\hat{y} - \hat{y}_{-i})^{\mathrm{T}}(\hat{y} - \hat{y}_{-i})}{p\tilde{\sigma}^2} = \frac{r_i^2 h_{ii}}{p(1 - h_{ii})}$$

called D_i in LM-2 §6.2.3, LM-1 §4.2.3

- Durbin-Watson test checks for auto-correlation in residuals
- may be useful if the residual plots seem to oscillate, or if the data are collected over time
- collinearity: if there is a linear relationship among columns of X, then individual coefficients are not well-determined
- check condition number of $X^T X$

LM-2 §7.3; LM-1 §5.3

LM-2 Figure 6.7

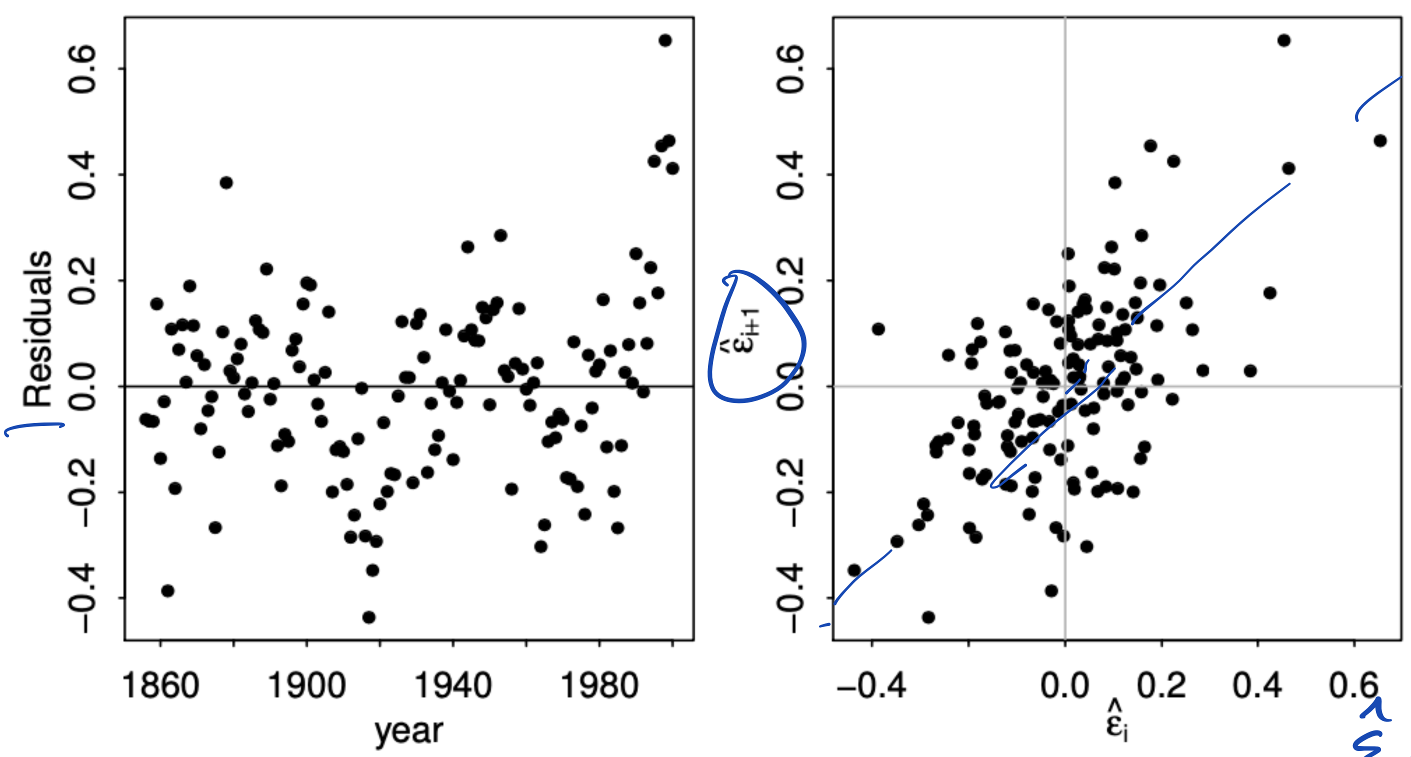


Figure 6.7 Diagnostic plots for correlated errors in the air quality data.

$\hat{\varepsilon}_{i+1}$

$\hat{\varepsilon}_i$

Three tasks related to linear regression

- Estimation of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

  alternatively comparing sub-models using F-tests

- Prediction of y_+, say, given a new vector of explanatory variables x_+ HW 3

  LM-2 Ch.4, LM-1 §3.5, SM §8.3.2

- Model Selection: which explanatory variables do we need for prediction or inference?

(checking the ass., generalizing model, etc.)

- Model Building

Three tasks related to linear regression

- Estimation of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

  alternatively comparing sub-models using F-tests

- Prediction of y_+, say, given a new vector of explanatory variables x_+

  LM-2 Ch.4, LM-1 §3.5, SM §8.3.2

- Model Selection: which explanatory variables do we need for prediction or inference?

These same questions arise in other models such as logistic regression, analysis of survival data, and so on, but the generic linear model is often a good starting point

Three tasks related to linear regression

- **Estimation** of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

alternatively comparing sub-models using F-tests

- **Prediction** of y_+, say, given a new vector of explanatory variables x_+

LM-2 Ch.4, LM-1 §3.5, SM §8.3.2

- **Model Selection:** which explanatory variables do we need for prediction or inference?

These same questions arise in other models such as logistic regression, analysis of survival data, and so on, but the generic linear model is often a good starting point

- **Prediction:** $y_+ = x_+^{\mathrm{T}}\beta + \epsilon; \qquad \hat{y}_+ = x_+^{\mathrm{T}}\hat{\beta}; \qquad \mathrm{var}(\hat{y}_+) = \sigma^2 x_+ (X^{\mathrm{T}}X)^{-1} x_+$

assuming ...

Three tasks related to linear regression

- Estimation of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

  alternatively comparing sub-models using F-tests

- Prediction of y_+, say, given a new vector of explanatory variables x_+

  LM-2 Ch.4, LM-1 §3.5, SM §8.3.2

- Model Selection: which explanatory variables do we need for prediction or inference?

$z \in \mathbb{R}^p$

$\mathrm{var}(Az) = A^T \mathrm{var}\, z\, A$

These same questions arise in other models such as logistic regression, analysis of survival data, and so on, but the generic linear model is often a good starting point

- Prediction: $y_+ = x_+^{\mathrm{T}}\beta + \epsilon$; $\quad \hat{y}_+ = x_+^{\mathrm{T}}\hat{\beta}$; $\quad \mathrm{var}(\hat{y}_+) = \sigma^2 x_+^{T}(X^{\mathrm{T}}X)^{-1}x_+$

  ↑ fitted value
  exp'd resp

  assuming...
  model stays same
  ind't

- error in expected response different from

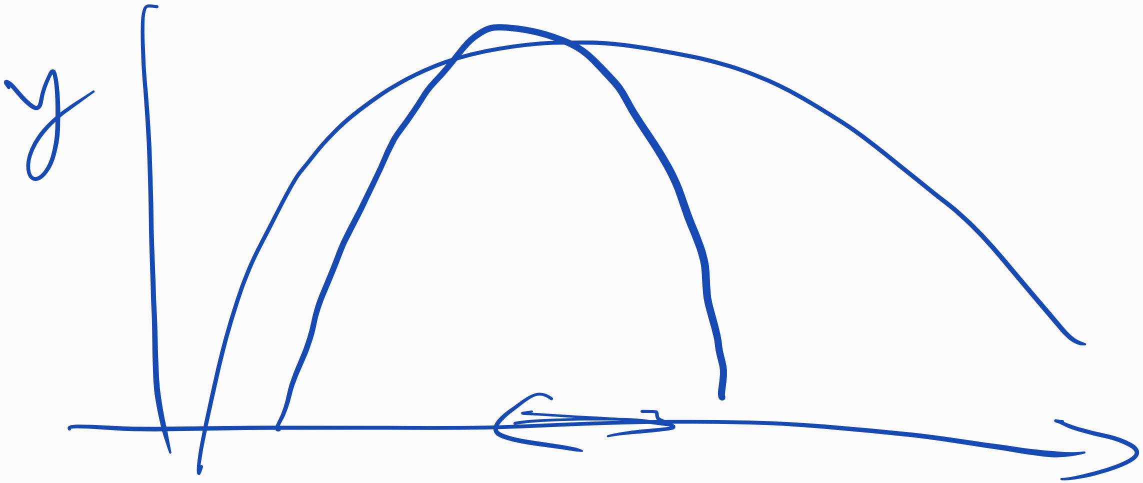
  prediction error $\mathbb{E}(y_+ - \hat{y}_+)^2 = \sigma^2 + \mathrm{var}(\hat{y}_+)$

pred. $\pm \sqrt{\sigma(1 + \ldots)}\, t_\nu$

Model Selection

CD Princ.

- “analyses should be as simple as possible, but no simpler”
- What variables should we keep in the model ?



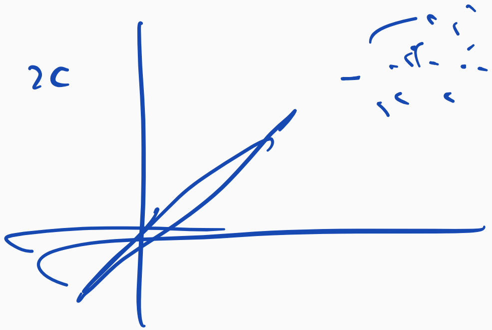
$\beta_0 + \beta_2 x^2$

$\beta_0' + \beta_1' x + \beta_2' x^2$

$x \to (x+a)^2$

$2ax$

$\beta_0 + \beta_1 x$



- "analyses should be as simple as possible, but no simpler"
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the 'highest' level of the hierarchy is removed first
- e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ should *not* be simplified to $y = \beta_0 + \beta_2 x^2 + \epsilon$
- e.g. if interaction terms are included, then main effects and other 2nd-order terms also need to be included: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$
- *not* $\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon}$

unless $x = 0/1$

- "analyses should be as simple as possible, but no simpler"
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the 'highest' level of the hierarchy is removed first
- e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ should *not* be simplified to $y = \beta_0 + \beta_2 x^2 + \epsilon$
- e.g. if interaction terms are included, then main effects and other 2nd-order terms also need to be included: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$
- *not* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$ unless $x = 0/1$
- $y = \beta_0 + \beta_1 \sin(2\pi x) + \beta_2 \cos(2\pi x) + \beta_3 \sin(4\pi x) + \beta_4 \cos(4\pi x) + \epsilon$
- $y_t = \beta_0 + \alpha y_{t-1} + \epsilon$ $y_t = \beta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} \epsilon$ *not* $y_t = \beta_0 + \alpha_2 y_{t-2} + \epsilon$

AR(1) AR(2)

- testing procedures: forward selection, backward selection, stepwise selection
- it is quite common to fit all explanatory variables, and then drop if $p > 0.05$

all possible
subsets
(leaps)
(regsubsets)

… Model Selection

- testing procedures: forward selection, backward selection, stepwise selection
- it is quite common to fit all explanatory variables, and then drop if $p > 0.05$
- if estimates and estimated standard errors don't change very much, may be okay
- if estimates and estimated standard errors change a lot, cause for concern
- if estimates change sign, points to possibly extreme confounding

- testing procedures: forward selection, backward selection, stepwise selection
- it is quite common to fit all explanatory variables, and then drop if $p > 0.05$
- if estimates and estimated standard errors don't change very much, may be okay
- if estimates and estimated standard errors change a lot, cause for concern
- if estimates change sign, points to possibly extreme confounding
- importance of retained explanatory variables probably overstated
- procedures not directly linked to final objectives of prediction or explanation
- tends to pick models that are smaller than desirable for prediction LM-2 10.2, LM-1 8.2
- "should be discouraged" LM-2 10.2

- Criterion-based procedures most widely used
- $AIC = n\log(RSS/n) + 2p$ balance between fit and parsimony

RSS: residual sum of squares

model fit

+ constant

Akaike's info. criterion

compare AIC from 2 models

smaller is better

step and in stepAIC

stepwise regression full ↓ variables out until AIC stops ↓

~~smallest AIC~~

- Criterion-based procedures

most widely used

- $AIC = n\log(RSS/n) + 2p$ balance between fit and parsimony

RSS: residual sum of squares

- $BIC = n\log(RSS/n) + \log(n)p$ choose models with smallest AIC or BIC

chooses smaller models on average

SM $AIC_{corrected}$ ←

- Criterion-based procedures

most widely used

- $AIC = n\log(RSS/n) + 2p$ balance between fit and parsimony

RSS: residual sum of squares

- $BIC = n\log(RSS/n) + \log(n)p$ choose models with smallest AIC or BIC

- $C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$ estimates average MSE of prediction

Mallows C_p

RSS_{full}

- Criterion-based procedures — most widely used
- $AIC = n\log(RSS/n) + 2p$ balance between fit and parsimony

  RSS: residual sum of squares
- $BIC = n\log(RSS/n) + \log(n)p$ choose models with smallest AIC or BIC
- $C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$ estimates average MSE of prediction

  $C_p \approx p$ for variables check text
- $R_a^2 = 1 - \dfrac{RSS/(n-p)}{TSS/(n-1)} = 1 - \dfrac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}$ adjusted R^2
- SM has yet another version AIC_c which may be better than AIC for linear models
- C_p and R_a^2 are only useful for linear models; AIC and BIC more general

  $n\log\left(\frac{RSS}{n}\right) \equiv -2\log L(\hat{\beta})$ $\hat{\beta}_{mle}$

model1 ← lm(lpsa ~ . , prostate)

↑ original data frame

```
step(model1)
...
Step:  AIC=-61.37
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
<none>                 45.526 -61.374
- age      1    0.9592 46.485 -61.352
- lbph     1    1.8568 47.382 -59.497
- lweight  1    3.2251 48.751 -56.735
- svi      1    5.9517 51.477 -51.456
- lcavol   1   28.7665 74.292 -15.871

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi
    0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095
```

larger

"$p < .05$"

- hierarchical principle, testing procedures, criterion-bsed procedures, all provide guidance on how to choose x's
- in a linear regression model

  and extensions
- rote application of any of these methods gives little insight into the structure of the model

Model Building

CD §6.5

- hierarchical principle, testing procedures, criterion-bsed procedures, all provide guidance on how to choose x's
- in a linear regression model

and extensions

- rote application of any of these methods gives little insight into the structure of the model

- Empirical models: "In many fields of study the models used as a basis for interpretation do not have a special subject-matter base, but, rather represent broad patterns of haphazard variation quite widely seen in at least approximate form

$(X\beta + \varepsilon)$

- this is typically combined with a specificaton of the systematic part of the variation, which is often, although not always, the primary focus of interest"
- $\mathrm{E}(y \mid X) = X\beta$

systematic

"Supose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j, say, on the response y. Then the following points are relevant:

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model

"Supose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j, say, on the response y. Then the following points are relevant:

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model
- relatively mechanical methods of choosing which explanatory variables to use may be helpful in preliminary exploration, especially if p is quite large, but are insecure as a basis for final interpretation

"Supose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j, say, on the response y. Then the following points are relevant:

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model
- relatively mechanical methods of choosing which explanatory variables to use may be helpful in preliminary exploration, especially if p is quite large, but are insecure as a basis for final interpretation
- explanatory variables not of direct interest but known to have a substantial effect should be included

… Model Building

CD §7.3

“Supose that, at some point in the analysis, interest is focused on the role of a particular explanatory variable or variables, x_j, say, on the response y. Then the following points are relevant:

- the value, standard error, and interpretation of $\hat{\beta}_j$ depends on the other variables in the model
- relatively mechanical methods of choosing which explanatory variables to use may be helpful in preliminary exploration, especially if p is quite large, but are insecure as a basis for final interpretation
- explanatory variables not of direct interest but known to have a substantial effect should be included
- it may be essential to recognize that several different models are potentially equally effective

Example

SM Eg 8.29

- nuclear plant data — Cox & Snell 1981
- `library(SMPracticals); data(nuclear); View(nuclear); ?nuclear`

Table 8.13 Data on light water reactors (LWR) constructed in the USA (Cox and Snell, 1981, p. 81). The covariates are `date` (date construction permit issued), `T1` (time between application for and issue of permit), `T2` (time between issue of operating license and construction permit), `capacity` (power plant capacity in MWe), `PR` (=1 if LWR already present on site), `NE` (=1 if constructed in north-east region of USA), `CT` (=1 if cooling tower used), `BW` (=1 if nuclear steam supply system manufactured by Babcock–Wilcox), `N` (cumulative number of power plants constructed by each architect-engineer), `PT` (=1 if partial turnkey plant).

	cost	date	T_1	T_2	capacity	PR	NE	CT	BW	N	PT
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1065	0	0	1	0	1	0
3	443.22	67.33	10	85	1065	1	0	1	0	1	0
4	652.32	68.00	11	67	1065	0	1	1	0	12	0
5	642.23	68.00	11	78	1065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
9	457.12	68.42	15	55	822	1	0	0	0	5	0
10	690.19	68.33	12	71	792	0	1	1	1	2	0
11	350.63	68.58	12	64	560	0	0	0	0	3	0
12	402.59	68.75	13	47	790	0	1	0	0	6	0
13	412.18	68.42	15	62	530	0	0	1	0	2	0
14	495.58	68.92	17	52	1050	0	0	0	0	7	0
15	394.36	68.92	13	65	850	0	0	0	1	16	0
16	423.32	68.42	11	67	778	0	0	0	0	3	0
17	712.27	69.50	18	60	845	0	1	0	0	17	0
18	289.66	68.42	15	76	530	1	0	1	0	2	0
19	881.24	69.17	15	67	1090	0	0	0	0	1	0
20	490.88	68.92	16	59	1050	1	0	0	0	8	0
21	567.79	68.75	11	70	913	0	0	1	1	15	0
22	665.99	70.92	22	57	828	1	1	0	0	20	0
23	621.45	69.67	16	59	786	0	0	1	0	18	0
24	608.80	70.08	19	58	821	1	0	0	0	3	0
25	473.64	70.42	19	44	538	0	0	1	0	19	0
26	697.14	71.08	20	57	1130	0	0	1	0	21	0
27	207.51	67.25	13	63	745	0	0	0	0	8	1
28	288.48	67.17	9	48	821	0	0	1	0	7	1
29	284.88	67.83	12	63	886	0	0	0	1	11	1
30	280.36	67.83	12	71	886	1	0	0	1	11	1
31	217.38	67.25	13	72	745	1	0	0	0	8	1
32	270.71	67.83	7	80	886	1	0	0	1	11	1

p-values

(stepwise)

	Full model		Backward		Forward	
	Est (SE)	t	Est (SE)	t	Est (SE)	t
Constant	−14.24 (4.229)	−3.37	−13.26 (3.140)	−4.22	−7.627 (2.875)	−2.66
date	0.209 (0.065)	3.21	0.212 (0.043)	4.91	0.136 (0.040)	3.38
log(T1)	0.092 (0.244)	0.38				
log(T2)	0.290 (0.273)	1.05				
log(cap)	0.694 (0.136)	5.10	0.723 (0.119)	6.09	0.671 (0.141)	4.75
PR	−0.092 (0.077)	−1.20				
NE	0.258 (0.077)	3.35	0.249 (0.074)	3.36		
CT	0.120 (0.066)	1.82	0.140 (0.060)	2.32		
BW	0.033 (0.101)	0.33				
log(N)	−0.080 (0.046)	−1.74	−0.088 (0.042)	−2.11		
PT	−0.224 (0.123)	−1.83	−0.226 (0.114)	−1.99	−0.490 (0.103)	−4.77
Residual SE (df)	0.164 (21)		0.159 (25)		0.195 (28)	

T2

Table 8.14 Parameter estimates and standard errors for linear models fitted to nuclear plants data; forward and backward indicate models fitted by forward selection and backward elimination.

– could also use `stepAIC` or `leaps::regsubsets`

… Example

- transformation of variables: `cost`, `T1`, `T2`, `cap`, `cum.n` all converted to log
- "partly to lead to unit-free parameters whose values can be interpreted in terms of power-law relations between the original variables"
- "Costs are typically relative. Moreover large costs are likely to vary more than small ones. For consistency we also take logs of the other quantitative covariates" SM

- transformation of variables: `cost`, `T1`, `T2`, `cap`, `cum.n` all converted to log
- "partly to lead to unit-free parameters whose values can be interpreted in terms of power-law relations between the original variables"
- "Costs are typically relative. Moreover large costs are likely to vary more than small ones. For consistency we also take logs of the other quantitative covariates" SM

- backward elimination leaves six variables with residual mean square $0.0253 = 0.159^2$; none of the eliminated variables is significant if re-introduced"

- transformation of variables: `cost`, `T1`, `T2`, `cap`, `cum.n` all converted to log
- "partly to lead to unit-free parameters whose values can be interpreted in terms of power-law relations between the original variables"
- "Costs are typically relative. Moreover large costs are likely to vary more than small ones. For consistency we also take logs of the other quantitative covariates" SM

- backward elimination leaves six variables with residual mean square $0.0253 = 0.159^2$; none of the eliminate variables is significant if re-introduced"

- variable PT is unbalanced
- check on the model includes interaction with PT

one variable at a time

bec. PT has only 6 1's

```
nuclear.lm2 <- lm(log(cost) ~ date + log(cap) + ne + ct + log(cum.n) + pt,
data = nuclear)

(Intercept)        date    log(cap)          ne
  -13.26031     0.21241     0.72341     0.24902
         ct  log(cum.n)          pt
    0.14039    -0.08758    -0.22610

 > update(nuclear.lm, .~. + pt*log(cap))$coef
 (Intercept)        date    log(cap)          ne
  -13.08645     0.21044     0.71761     0.24841
         ct  log(cum.n)          pt log(cap):pt
    0.13998    -0.08683    -2.18759     0.29159
```

Interaction

χ^2 term betw

PT & log(cap)

- $y = X\beta + \epsilon$, suppose p very large
- if $p > n$ then $RSS = 0$ with n explanatory variables
- no reduction in complexity; nothing learned about the relationship between y and X

- $y = X\beta + \epsilon$, suppose p very large
- if $p > n$ then $RSS = 0$ with n explanatory variables
- no reduction in complexity; nothing learned about the relationship between y and X

signal

- we expect few variables to be "active", i.e. useful in explaining variation in y $s << n$
- how do we find them? penalized regression

sparse model

regularized

L.S.

non-zero β's

$s << n < p$

30000

1 . - -

1

:

50

- $y = X\beta + \epsilon$, suppose p very large
- if $p > n$ then $RSS = 0$ with n explanatory variables
- no reduction in complexity; nothing learned about the relationship between y and X

- we expect few variables to be "active", i.e. useful in explaining variation in y $\quad s << n$
- how do we find them? penalized regression $\quad$ regularized

tuning parameter

- $$\arg\min_{\beta}\left\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_0\right\}$$

$$||\beta_0|| = \#\{j : \beta_j \neq 0\}$$

- $$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_0\}$$
  $$||\beta_0|| = \#\{j : \beta_j \neq 0\}$$
- non-convex optimization problem; computationally challenging

- $$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_0\}$$
$$||\beta_0|| = \#\{j : \beta_j \neq 0\}$$
- non-convex optimization problem; computationally challenging
- convex relaxation of this is
$$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_1\}$$
$$|||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$$

ridge

$\lambda \|\beta\|_2$ $\lambda \sum \beta_j^2$

- $$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_0\}$$

$$||\beta_0|| = \#\{j : \beta_j \neq 0\}$$

- non-convex optimization problem; computationally challenging
- convex relaxation of this is

$$\arg\min_{\beta}\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||_1\}$$

$$|||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$$

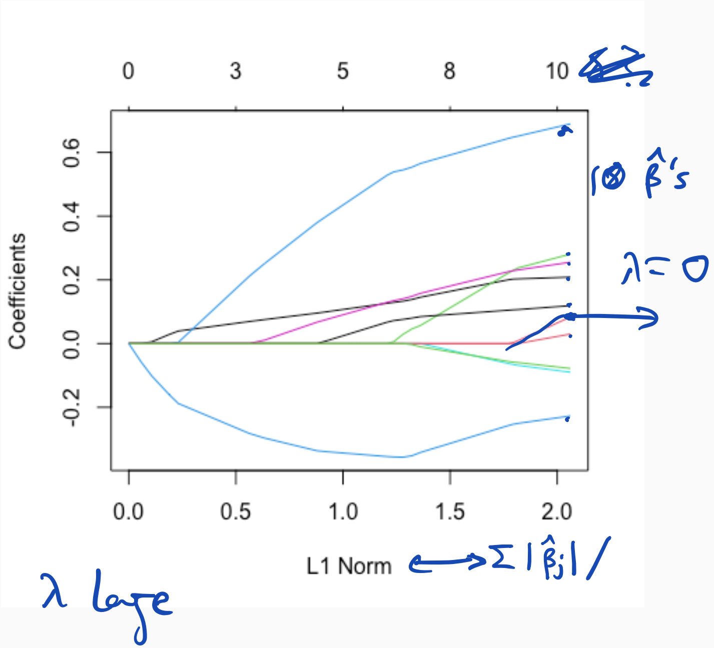
- resulting estimate $\hat{\beta}_\lambda$ called the Lasso estimate
- has the property that many $\hat{\beta}_{\lambda,j}$ are 0
- another route to variable selection

```
> require(glmnet)
> x <- model.matrix(nuclear.lm)
> y <- log(nuclear$cost)
> nuclear.lasso <- glmnet(x,y)
> plot(nuclear.lasso)
> cv.glmnet(x,y)
      Lambda Index Measure      SE Nonzero
min  0.0295     24  0.0427 0.0105       6
1se  0.0621     16  0.0530 0.0142       5
> coef(nuclear.lasso, s = 0.05)
```

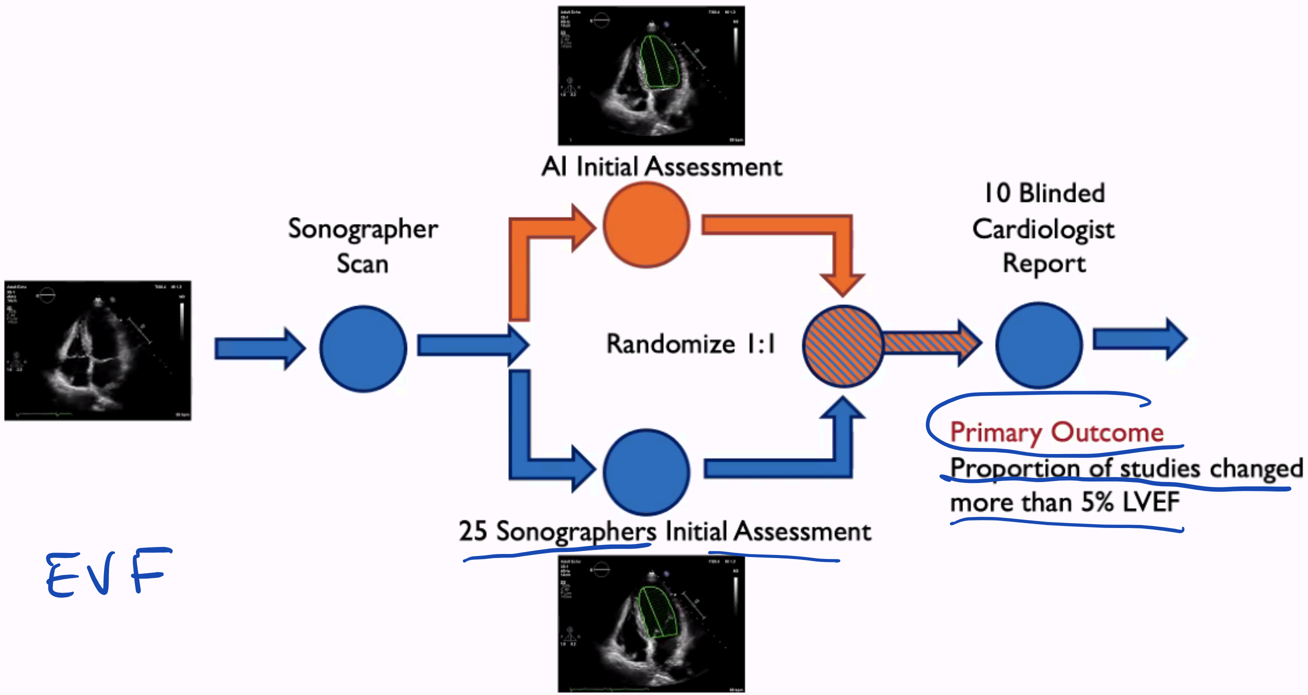
matrix

rep.

whole lecture



Randomized blinded clinical trial testing EchoNet



EVF

Sono AI

A B A

	A	B
person 1	—	
2		—
⋮	—	
	—	
n		—

1   A   B $Y_{B1} - Y_{A1}$

B   A

⋮

This week

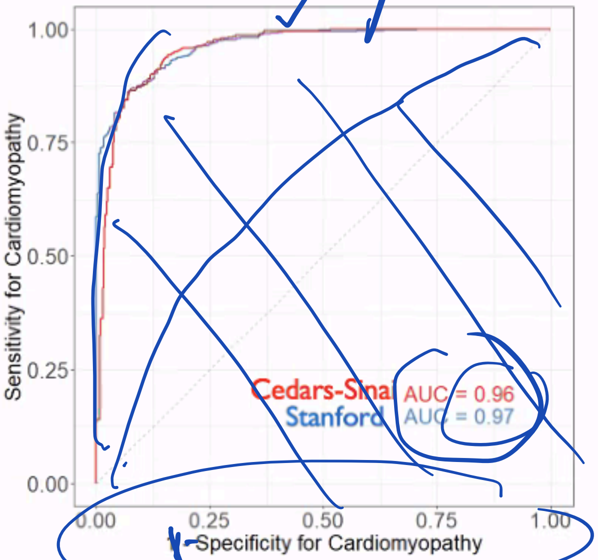
ROC curves

Retrospective evaluation: achieves expert performance

AUC plot

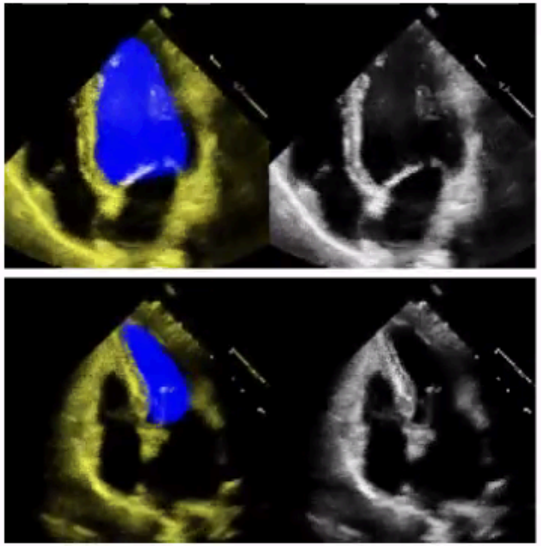
Sensitivity

Predicting heart failure



Evaluation at two hospitals

Examples



EVF

> 50%

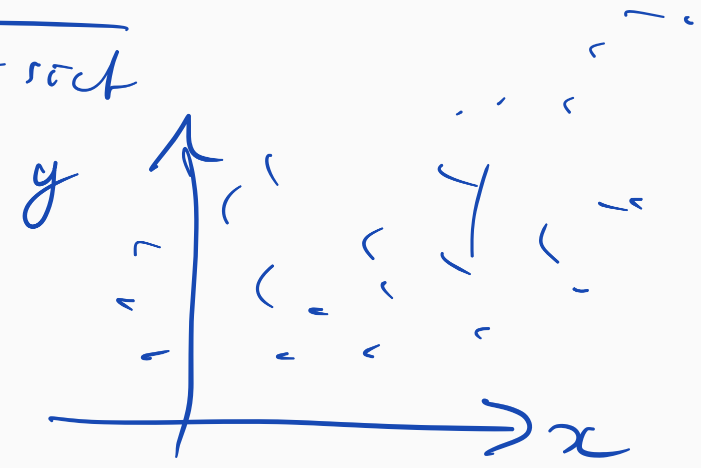
1- specificity

		AI +	AI −
ground truth	ill +	TP	FN
	not ill −	FP	TN

sens. $\dfrac{TP}{TP+FN}$ have dis

spec $\dfrac{TN}{TN+FP}$ ← not ill

$$\frac{\cancel{TN}+FP-\cancel{TN}}{\text{not sick}}$$



$$y = \beta_0 + \underline{\beta_1 x} + \underline{\beta_2 x^2} + \cancel{\beta_3 x^3} + \varepsilon$$

$H_0: \beta_3 = 0$

simpler quad. model

- common objectives $\cancel{\quad}$ $\beta_0 + \beta_2 x^2 + \beta_3 x^3$

Design of Studies

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions

Design of Studies

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions
- to ensure that the scale of effort is appropriate

… design of studies

- we concentrate largely on the careful analysis of individual studies

… design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed

… design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed
- but even there the quality of individual studies remains important

… design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed
- but even there the quality of individual studies remains important
- examples include overviews (such as the Cochrane reviews)

… design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed
- but even there the quality of individual studies remains important
- examples include overviews (such as the Cochrane reviews)
- in some areas new investigations can be set up and completed relatively quickly; design of individual studies may then be less important

- formulation of a plan of analysis

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context

... design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context
- even if pre-specified methods must be used, it is crucial not to limit analysis

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context
- even if pre-specified methods must be used, it is crucial not to limit analysis
- planned analysis may be technically inappropriate

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context
- even if pre-specified methods must be used, it is crucial not to limit analysis
- planned analysis may be technically inappropriate
- more controversially, data may suggest new research questions or replacement of objectives

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern
- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context
- even if pre-specified methods must be used, it is crucial not to limit analysis
- planned analysis may be technically inappropriate
- more controversially, data may suggest new research questions or replacement of objectives
- latter will require confirmatory studies

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...
- split plot experiments have two classes of units of study and analysis

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt

- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...
- split plot experiments have two classes of units of study and analysis
- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...
- split plot experiments have two classes of units of study and analysis
- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible
- the unit of analysis may not be the unit of interpretation – ecological bias

  systematic difference between impact of x at different levels of aggregation

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment

  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...
- split plot experiments have two classes of units of study and analysis
- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible
- the unit of analysis may not be the unit of interpretation – ecological bias

  systematic difference between impact of x at different levels of aggregation
- on the whole, limited detail is needed in examining the variation within the unit of study

Types of observational studies

- secondary analysis of data collected for another purpose

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features
- study of a relationship between features, where individuals may be examined

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features
- study of a relationship between features, where individuals may be examined
    - at a single time point

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features
- study of a relationship between features, where individuals may be examined
    - at a single time point
    - at several time points for different individuals

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features
- study of a relationship between features, where individuals may be examined
    - at a single time point
    - at several time points for different individuals
    - at different time points for the same individual

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features
- study of a relationship between features, where individuals may be examined
    - at a single time point
    - at several time points for different individuals
    - at different time points for the same individual
- experiment: investigator has complete control over treatment assignment

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features
- study of a relationship between features, where individuals may be examined
  - at a single time point
  - at several time points for different individuals
  - at different time points for the same individual
- experiment: investigator has complete control over treatment assignment
- census

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features
- study of a relationship between features, where individuals may be examined
    - at a single time point
    - at several time points for different individuals
    - at different time points for the same individual
- experiment: investigator has complete control over treatment assignment
- census
- meta-analysis: statistical assessment of a collection of studies on the same topic

Avoidance of systematic error

- “distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run”

Avoidance of systematic error

- "distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run"
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units

Avoidance of systematic error

- “distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run”
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis

Avoidance of systematic error

- "distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run"
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis
- can arise by the entry of personal judgement into some aspect of the data collection process

Avoidance of systematic error

- “distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run”
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis
- can arise by the entry of personal judgement into some aspect of the data collection process
- this can often be avoided by randomization and blinding