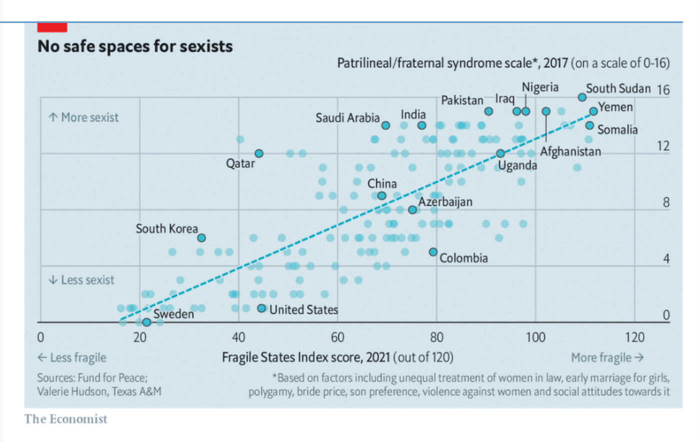
Methods of Applied Statistics I

STA2101H F LEC9101



Week 3

September 29 2021

Today

Start Recording

Wed

1. Upcoming events, HW 3
   Office Hour ~~Monday Oct 4 7pm-8.30pm~~ $\longrightarrow$ Tuesday Oct 5 7pm - 8 pm, Zoom 4-5[30]
2. Linear Regression Part 3: recap, checking model assumptions, collinearity, model-building, $p > n$
3. In the News

Today

1. Upcoming events, HW 3
   Office Hour ~~Monday Oct 4 7pm-8.30pm~~ $\longrightarrow$ Tuesday Oct 5 7pm - 8 pm, Zoom
2. Linear Regression Part 3: recap, checking model assumptions, collinearity, model-building, $p > n$
3. In the News $\longleftarrow$ (Des of Expt.)
4. Third hour – HW 1 Comments

Upcoming

- Non-Central Squared Copulas: Properties and Applications
- Thursday 3.30 Link

About Bouchra Nasri

Dr. Nasri is Assistant Professor in Statistics at the Scholl of Public Health of Univesité de Montréal. Her research interests are dependence modelling, time series, and more recently spatial modelling. The main applications targeted by her research projects are related to climate change, public health and infectious diseases modelling. Dr. Nasri is an associate director of the new infectious diseases network OMNI-RÉUNIS.

- Friday Oct 1 Toronto Data Workshop Zoom link

Toronto Data Workshop this Friday at noon (Toronto time) focuses on the recent Canadian election, with presentations from

- Professor David Andrews on elections forecasting;
- Professor Daniel Rubenson on the Canadian Election Study;
- Johnson Vo on his model of the 2021 election; and
- Eric Zhu, Brian Diep, Ashely (Jing Yuan) Zhang, Kristin (Xi Yu Huang), and Tanvir Hyder on their model of the 2021 election.

Link: https://utoronto.zoom.us/j/84277066292
Meeting ID: 842 7706 6292
Passcode: data_4_lyf

HW 3

Linear regression recap

- Analysis of variance: $\underset{-n\bar{y}^2}{\overset{\sum y_i^2}{y^T y}} = (y - X\hat{\beta})^T (y - X\hat{\beta}) + \hat{\beta}^T X^T X \hat{\beta} - n\bar{y}^2$

LM-1,2
& Ch 3

Source	DF	SS	MS
$\boxed{\text{Regression}}$	$p-1$	SS_{REG}	$RegMS = SS_{REG}/(p-1)$
Residual	$n-p$	RSS	$ResMS = RSS/(n-p)$
Total (corrected)	$n-1$	TSS	

SSE, SS_{Em}, $SS(\hat{\beta})$

$\sum(y_i - \hat{y}_i)^2$

$\sum(y_i - \bar{y})^2$

$$F = \frac{RegMS}{ResMS} \sim F_{p-1,n-p} \quad \text{under } H_0: \beta = 0 \ (\text{excl. } \beta_0)$$

- regression SS can be further partitioned

depends on the order

$\varepsilon_i \sim N(0, \sigma^2)$

Analysis of variance:

data(prostate)

order really matters

```
anova(model1)
Analysis of Variance Table

Response: lpsa
          Df Sum Sq Mean Sq  F value    Pr(>F)
lcavol     1 69.003  69.003 137.4962 < 2.2e-16 ***
lweight    1  5.949   5.949  11.8531 0.0008832 ***
age        1  0.420   0.420   0.8369 0.3627958
lbph       1  1.069   1.069   2.1302 0.1479839
svi        1  5.952   5.952  11.8594 0.0008806 ***
lcp        1  0.129   0.129   0.2576 0.6130533
gleason    1  0.708   0.708   1.4098 0.2382837
pgg45      1  0.526   0.526   1.0476 0.3088604
Residuals 88 44.163   0.502
---
```

only in

SSerror

$\frac{69}{.5} = 137.5$

```
> summary(model1)
Call:
lm(formula = lpsa ~ ., data = prostate)
Coefficients:
             Estimate Std. Error t value Pr(>|t
(Intercept)  0.669337   1.296387   0.516  0.606
lcavol       0.587022   0.087920   6.677 2.11e-
lweight      0.454467   0.170012   2.673  0.008
age         -0.019637   0.011173  -1.758  0.082
lbph         0.107054   0.058449   1.832  0.070
svi          0.766157   0.244309   3.136  0.002
lcp         -0.105474   0.091013  -1.159  0.249
gleason      0.045142   0.157465   0.287  0.775
pgg45        0.004525   0.004421   1.024  0.308

Residual standard error: 0.7084 on 88 degrees o
```

all in

$1.024^2 = 1.04$

Analysis of variance:

```
anova(model1)
Analysis of Variance Table

Response: lpsa
          Df Sum Sq Mean Sq  F value    Pr(>F)
lcavol     1 69.003  69.003 137.4962 < 2.2e-16 ***
lweight    1  5.949   5.949  11.8531 0.0008832 ***
age        1  0.420   0.420   0.8369 0.3627958
lbph       1  1.069   1.069   2.1302 0.1479839
svi        1  5.952   5.952  11.8594 0.0008806 ***
lcp        1  0.129   0.129   0.2576 0.6130533
gleason    1  0.708   0.708   1.4098 0.2382837
pgg45      1  0.526   0.526   1.0476 0.3088604
Residuals 88 44.163   0.502
---
```

```
> drop1(model1)
Single term deletions

Model:
lpsa ~ lcavol + lweight + age + lbph + svi + lc
    pgg45
        Df Sum of Sq    RSS     AIC
<none>               44.163 -58.322
lcavol   1   22.3721 66.535 -20.567
lweight  1    3.5861 47.749 -52.749
age      1    1.5503 45.713 -56.975
lbph     1    1.6835 45.847 -56.693
svi      1    4.9355 49.099 -50.046
lcp      1    0.6740 44.837 -58.853
gleason  1    0.0412 44.204 -60.231
pgg45    1    0.5258 44.689 -59.174
```

"all in" ≡ summary

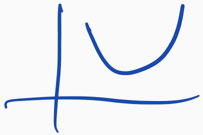
? lcavol explains ~% of variability in y ?

- same principle can be used to test for sets of variables
- or for testing any linear constraint on β

$A\beta = c$

$H_0: A\beta = c$

known matrix

- 

$F_{1,\nu} \equiv t^2_\nu$

- numerator degrees of freedom for F-statistic depend on the rank of A
- sometimes only an F-test can be used to assess the effect of an explanatory variable

when?

$\beta_1 = 0$ A rank 1

$\beta_1 = \beta_2 = 0$ A " 2

$\beta_1 = \beta_2 = 7$ " "

linear in α

IJALM] factor, variables

$\boxed{P(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3}$

$$g_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1+\frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

t_ν df

$-\infty < t < \infty$

$$g_{1,\nu}(f) = \frac{\Gamma\left(\frac{1+\nu}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(\frac{1}{\nu}\right)^{1/2} f^{-\frac{1}{2}} \left(1+\frac{1}{\nu}f\right)^{-\left(\frac{1+\nu}{2}\right)}$$

density of $F_{1,\nu}$

$f > 0$

$$T_\nu^2 \overset{d}{=} F_{1,\nu}$$

$\underset{\sim}{x}_\pi$ scramble of $\underset{\sim}{x}$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$\begin{pmatrix} x_1 & y_1 \\ \vdots & \\ x_n & y_n \end{pmatrix}$$

- §3.3: permutation test – doesn't rely on normal assumption

$H_0 : \beta_1 = 0$

$\hat\beta_0 \quad \hat\beta_1$

$\hat\beta^*_{1,1} \quad \cdots \quad \hat\beta^*_{1,p}$

$\hat\beta_1$

- §3.5: confidence intervals for β_j and regions for $(\beta_j, \beta_{j'})$ `confint; ellipse` see Fig 3.2

$\hat\beta_j \pm t^{\alpha/2}_{n-p} \cdot \hat{se}_j$

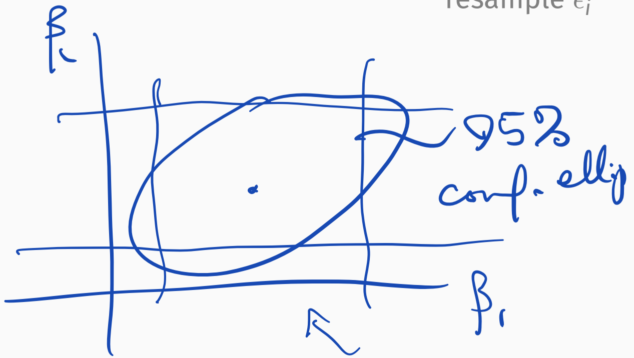
- §3.6: bootstrap inference for β_j resample $\hat\epsilon_i$

$\hat\varepsilon_i \quad i = 1, \ldots, n \quad y_i - \hat y_i$

$b = 1, \ldots, B$

$\hat\varepsilon_i^{*b} \sim$ sample with replacement from $\hat\varepsilon_1, \ldots, \hat\varepsilon_n$

$i = 1 \cdots n$



Model checking

`plot(model1)`

$y_i = x_i^T \beta + \varepsilon_i$ $\varepsilon_i \sim N(0, \sigma^2)$

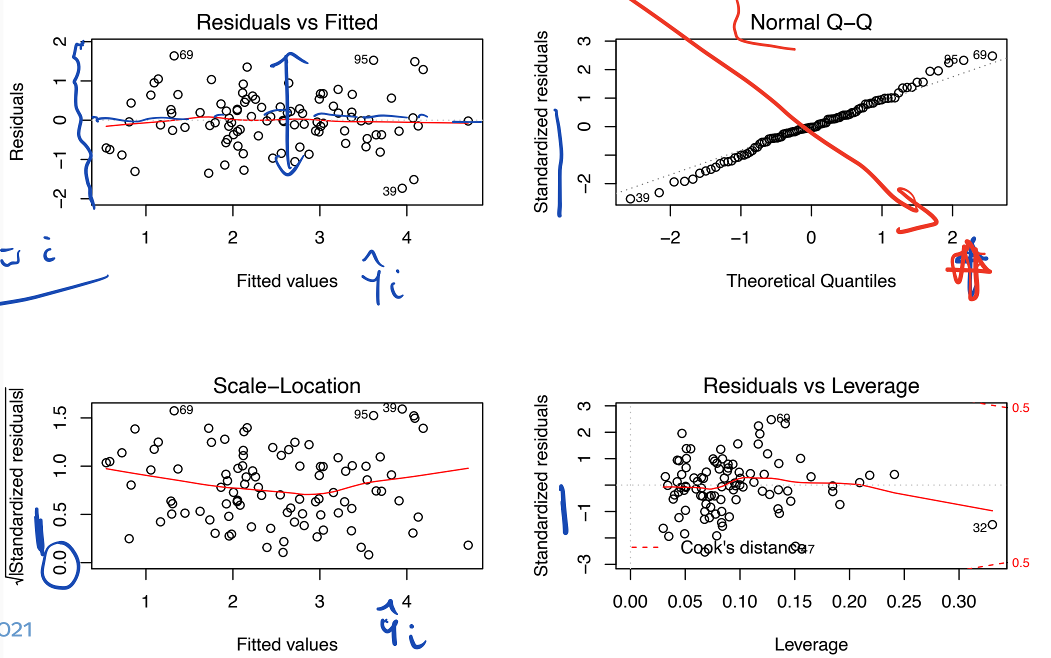
$\hat{\varepsilon}_i = y_i - \hat{y}_i$

$\sum \hat{\varepsilon}_i = 0$

$E(\varepsilon_i) = 0$ σ^2 not changing in i

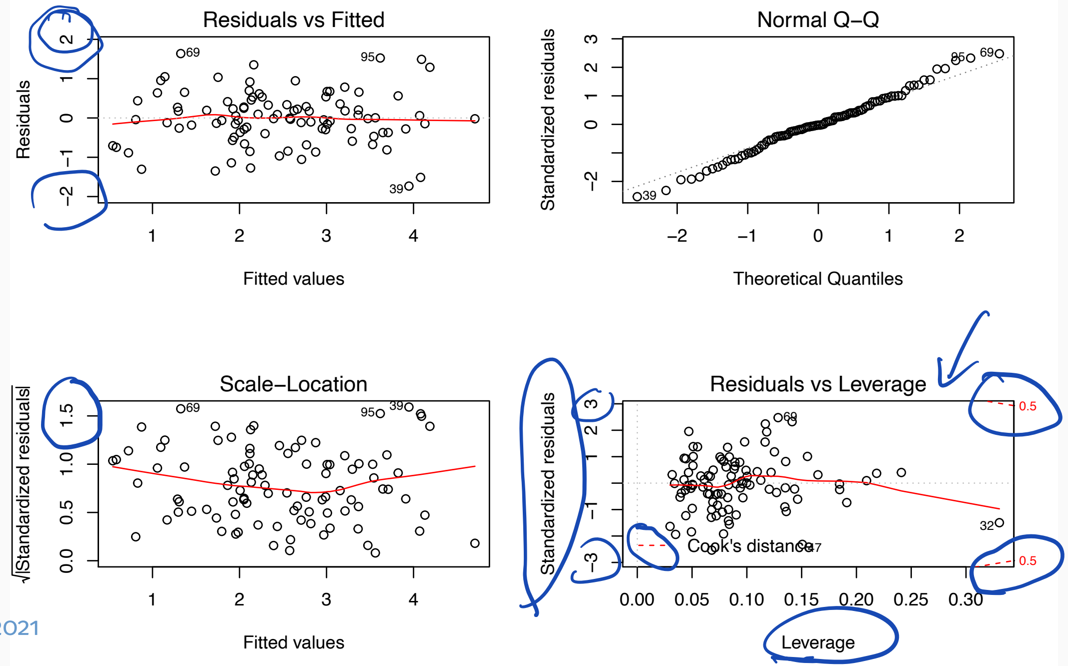
$\sqrt{|\hat{\varepsilon}_i / \_\_|} = r_i$

"more refined version of ↑"



Model assumptions

$$\sqrt{\frac{|y_i-\hat{y}_i|}{\tilde{\sigma}(1-h_{ii})^{1/2}}}$$



… Model checking

- residuals: $\hat{\epsilon}_i =$ $y_i - \hat{y}_i = y_i - x_i^T \hat{\beta}$

... Model checking

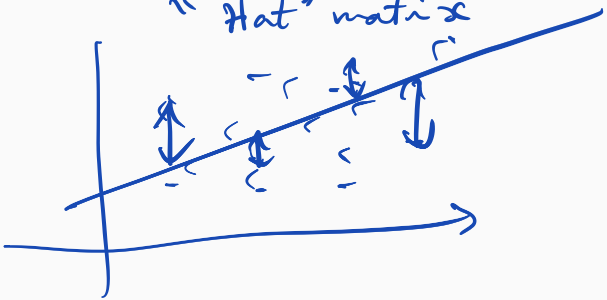
(Ass.)

$\varepsilon_i \sim N(0, \sigma^2)$ ind't

- residuals: $\underline{\hat{\epsilon}_i} =$ $y_i - \hat{y}_i$
- Var($\hat{\epsilon}_i$) $=$ $\sigma^2(1 - h_{ii})$ $\quad \operatorname{cov}(\hat{\underline{\varepsilon}}) = \sigma^2 (I - H)$

H hat matrix

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{\text{"Hat" matrix}} y$$



… Model checking

- residuals: $\hat{\epsilon}_i =$

- $\mathrm{Var}(\hat{\epsilon}) =$
- i.e. don't all have the same variance

… Model checking

- residuals: $\hat{\epsilon}_i =$

- $\text{Var}(\hat{\epsilon}) =$
- i.e. don't all have the same variance

- hat matrix $H =$

... Model checking

- residuals: $\hat{\epsilon}_i =$

- $\text{Var}(\hat{\epsilon}) =$
- i.e. don't all have the same variance

- hat matrix $H =$ $X(X^T X)^{-1} X^T$

- standardized residuals: $r_i =$ $\dfrac{y_i - \hat{y}_i}{\tilde{\sigma}(1 - h_{ii})^{1/2}}$

… Model checking

- residuals: $\hat{\epsilon}_i =$

- $\text{Var}(\hat{\epsilon}) =$
- i.e. don't all have the same variance

- hat matrix $H =$

- standardized residuals: $r_i =$

- Cook's distance $C_i =$

… Model checking

- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$

… Model checking

- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$

- $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H), \quad \text{Var}(y_j - \hat{y}_j) = \sigma^2(1 - h_{jj})$

$0 < h_{jj} < 1, \Sigma h_{jj} = p$

- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$

- $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H), \quad \text{Var}(y_j - \hat{y}_j) = \sigma^2(1 - h_{jj})$ $0 < h_{jj} < 1, \Sigma h_{jj} = p$
- i.e. don't all have the same variance

... Model checking

- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$

- $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H), \quad \text{Var}(y_j - \hat{y}_j) = \sigma^2(1 - h_{jj})$ $\qquad 0 < h_{jj} < 1, \Sigma h_{jj} = p$
- i.e. don't all have the same variance

- hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ $\qquad Hy = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = X\hat{\beta} = \hat{y}$

… Model checking

- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$

- $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H)$, $\text{Var}(y_j - \hat{y}_j) = \sigma^2(1 - h_{jj})$ $0 < h_{jj} < 1, \Sigma h_{jj} = p$
- i.e. don't all have the same variance

- hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ $Hy = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = X\hat{\beta} = \hat{y}$

- standardized residuals: $r_i = \dfrac{\hat{\epsilon}_i}{\tilde{\sigma}(1 - h_{ii})^{1/2}}$ $\boxed{\text{approx var 1}}$

… Model checking

- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$

- $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H), \quad \text{Var}(y_j - \hat{y}_j) = \sigma^2(1 - h_{jj})$ $\mathbb{R}^n$ $0 < h_{jj} < 1, \Sigma h_{jj} = p$
- i.e. don't all have the same variance

- hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ $\qquad Hy = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = X\hat{\beta} = \hat{y}$ $\qquad \underset{n\times n}{H^T H} = H$

$n \times 1 \quad \in \mathbb{R}^p$

- standardized residuals: $r_i = \dfrac{\hat{\epsilon}_i}{\tilde{\sigma}(1 - h_{ii})^{1/2}}$ approx var 1

- Cook's distance $C_i = \dfrac{(\hat{y} - \hat{y}_{-i})^{\mathrm{T}}(\hat{y} - \hat{y}_{-i})}{p\tilde{\sigma}^2} = \dfrac{r_i^2 h_{ii}}{p(1 - h_{ii})}$ measure of influence

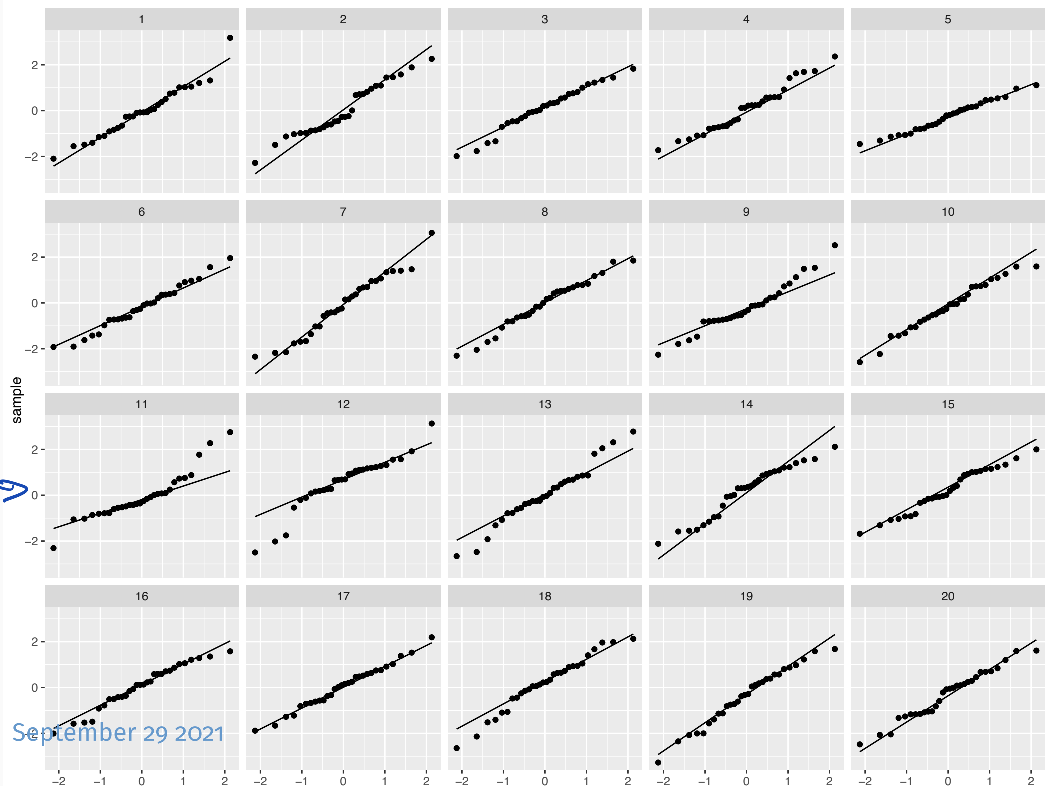
high leverage or high residual

$h_{ii} \approx \frac{p}{n}$

… Model checking

- standard diagnostics check for non-constant variance, influential observations
- and for normality of residuals — using `qqnorm`
- assumption of independence across i may be more important
- but more difficult to assess
- exception: observations collected over time LM-2, §6.1.3, LM-1 §4.1.3

Aside on normal plots



11

```
library(ggplot2); library(nullabor); library(tidyverse)
df5_frame <- data.frame(x = rt(30, df = 5))
lineup_df5_data <- lineup(
  method = null_dist("x", dist = "norm", params = list(mean = 0, sd = 1)),
  true = df5_frame, n=12)

lineup_df5_data %>%
  ggplot(aes(sample = x)) +
  geom_qq_line() +
  geom_qq() +
  facet_wrap(~ .sample)
```

Model structure ? $X\beta$?

$y_i = x_i^T\beta + \varepsilon_i$

$\varepsilon_i \sim \underline{\underline{N}}(0, \underline{\underline{\sigma^2}})$

h_{ii} influence

case

- Model $\boxed{y = X\beta + \epsilon}$, alternatively,
- $\boxed{E(y \mid X) = X\beta;}$ $\boxed{\mathrm{Var}(Y \mid X) = \sigma^2 I}$
- plots of y against each column of x can be helpful
- `for(i in 1:8){plot(prostate[,i],prostate[,9]... }`
- added variable plots can be more helpful
- plot residuals from y on X_{-j} against residuals from x_j on X_{-j}

partial regression plots

slope of this line is $\hat{\beta}_j$

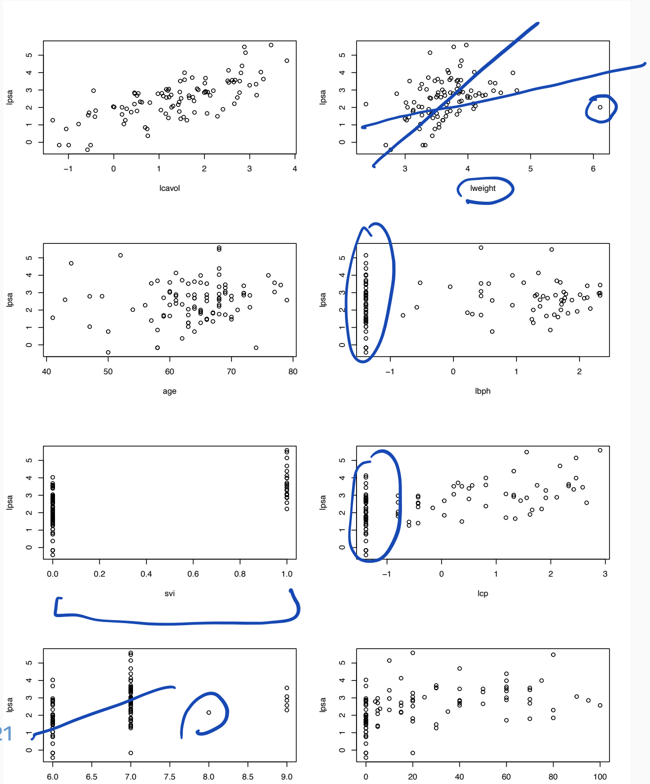
$j = 1, \dots, p$ $\quad y_{n\times 1}$ $\quad X_{n\times p}$ matrix

$\uparrow$ jth c.

$\hat{y} = X_{(-j)}\hat{\beta}_{(-j)}$ $\quad x_j = X_{-j}\hat{\beta}X_{-(j)}$

$y_i - \hat{y}_i$ $\quad x_{ji} - \hat{x}_{ji}$ $\leftarrow$ simple linear regression

Prostate data



Partial residual plots

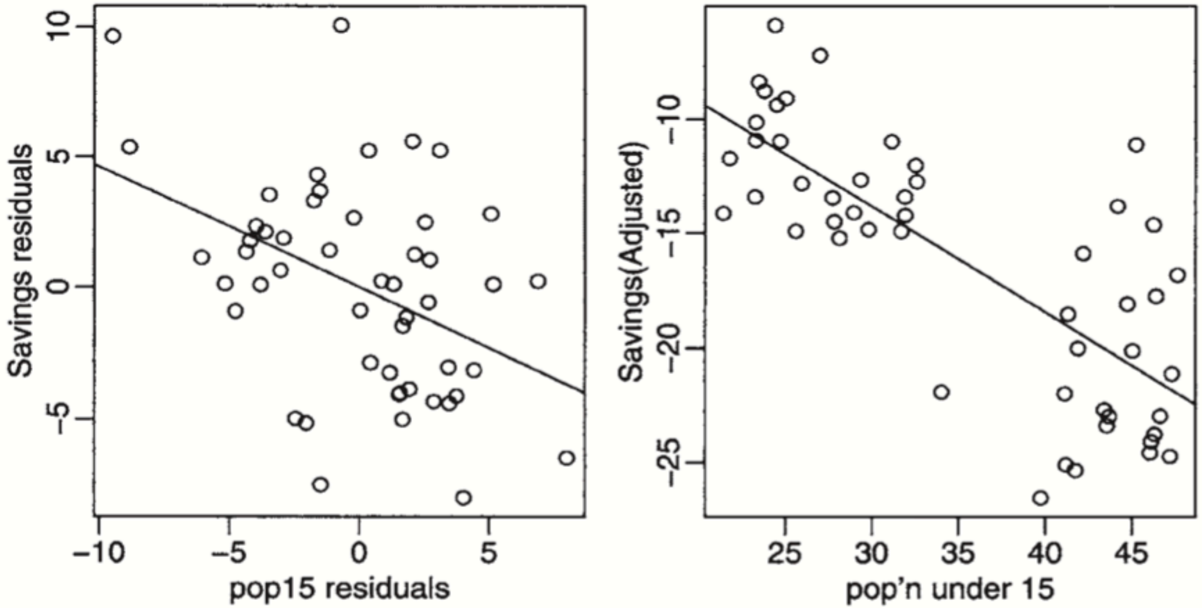


Figure 4.13 *Partial regression (left) and partial residual (right) plots for the savings data.*

Partial residual plots

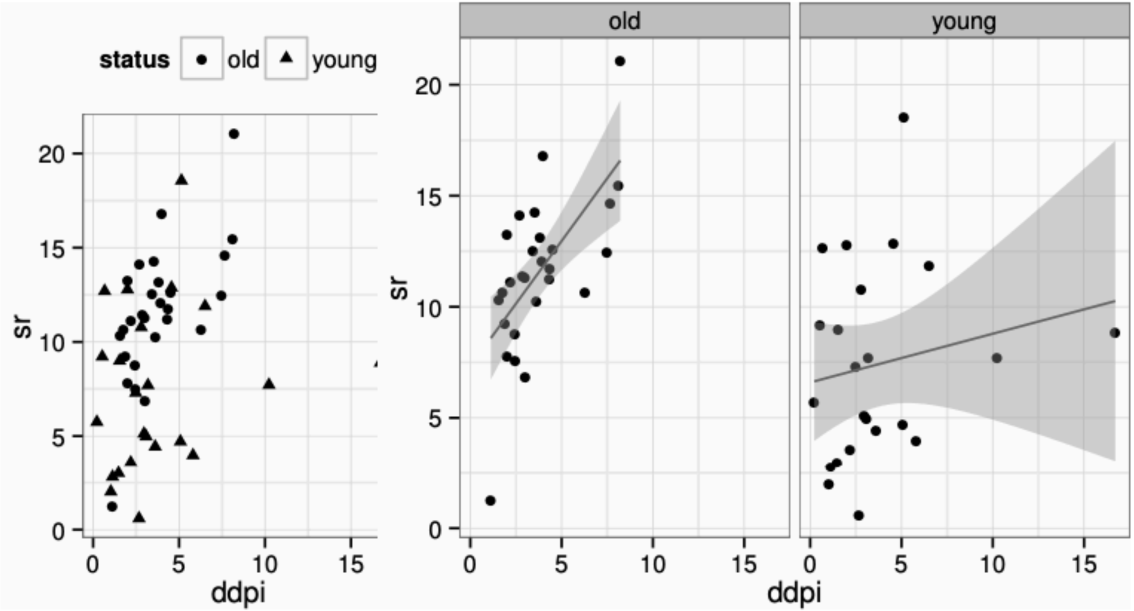


Figure 6.14 Introducing another dimension to diagnostic plots. Shape is used denote the status variable on the left while faceting is used on the right.

So many techniques

Read Chapter 6 of LM-2 or Chapter 4 of LM-1, replicating the results

Read Section 8.6 of SM, working through the algebra PhD, Stats

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \dots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates
- example: health effects of air pollution
- measurable increase in mortality on high-pollution days
- measurable increase in mortality on high-temperature days
- high temperatures and high levels of pollutants tend to co-occur

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \dots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates
- example: health effects of air pollution
- measurable increase in mortality on high-pollution days
- measurable increase in mortality on high-temperature days
- high temperatures and high levels of pollutants tend to co-occur +++
- mathematically, $X^{\mathrm{T}}X$ is nearly singular, or at least ill-conditioned, so calculation of its inverse is subject to numerical errors
- if $p > n$ then $X^{\mathrm{T}}X$ not invertible, no LS solution ridge, Lasso more next week

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?

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 ...

- 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

← ass^n $\qquad (X^TX)^{-}X^Ty$

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

  model

  assuming ...

- error in expected response different from

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

Model Selection

- “analyses should be as simple as possible, but no simpler”

Model Selection

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

Model Selection

- “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, ...

- “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

Model Selection

- “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 $y = \beta_0 + \beta_2 x^2 + \epsilon$

Model Selection

- “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: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$

Model Selection

- "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 $x = 0/1$

- “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 $x = 0/1$
- $y = \beta_0 + \beta_1 \sin(2\pi x) + \beta_2 \cos(2\pi x) + \beta_3 \sin(4\pi x) + \beta_4 \cos(4\pi x) + \epsilon$

Model Selection

- "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* $y_t = \beta_0 + \alpha_2 y_{t-2} + \epsilon$

... 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

```
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
```

- Criterion-based procedures

most widely used

- Criterion-based procedures — most widely used
- AIC, BIC, Mallows C_p, R_a^2 — RSS: residual sum of squares

- Criterion-based procedures — most widely used
- AIC, BIC, Mallows C_p, R_a^2 — RSS: residual sum of squares
-

$$AIC = n\log(RSS/n) + 2p$$

- Criterion-based procedures — most widely used
- AIC, BIC, Mallows C_p, R_a^2 — RSS: residual sum of squares
- $$AIC = n\log(RSS/n) + 2p$$
- $$BIC = n\log(RSS/n) + \log(n)p$$

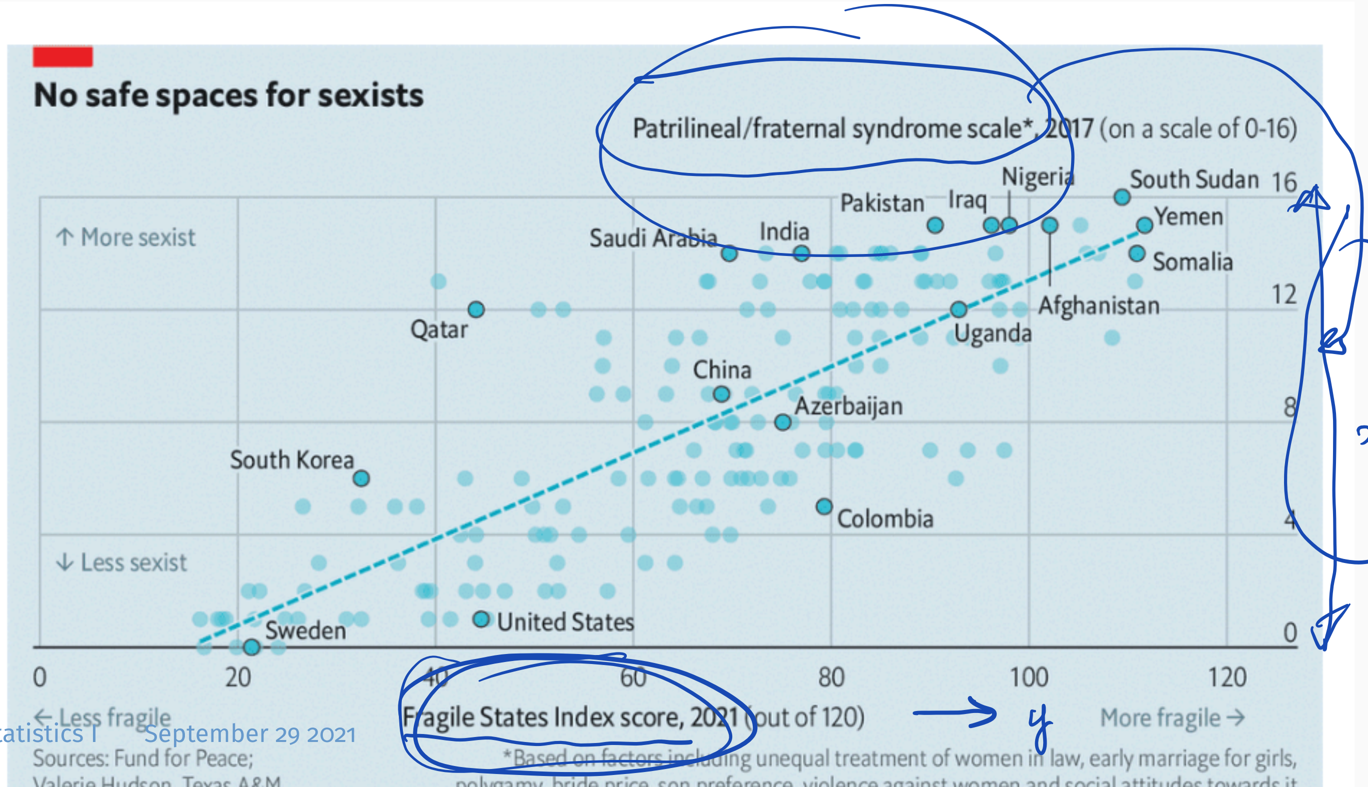
- Criterion-based procedures — most widely used
- AIC, BIC, Mallows C_p, R_a^2 — RSS: residual sum of squares
- $$AIC = n\log(RSS/n) + 2p$$
- $$BIC = n\log(RSS/n) + \log(n)p$$
- $$C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$$

… Model Selection

- Criterion-based procedures — most widely used
- AIC, BIC, Mallows C_p, R_a^2 — RSS: residual sum of squares
- $$AIC = n\log(RSS/n) + 2p$$
- $$BIC = n\log(RSS/n) + \log(n)p$$
- $$C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$$
- $$R_a^2 = 1 - \frac{\tilde{\sigma}^2_{model}}{TSS/(n-1)}$$

- Criterion-based procedures — most widely used
- AIC, BIC, Mallows C_p, R_a^2 — RSS: residual sum of squares
- $$AIC = n\log(RSS/n) + 2p$$
- $$BIC = n\log(RSS/n) + \log(n)p$$
- $$C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$$
- $$R_a^2 = 1 - \frac{\tilde{\sigma}^2_{model}}{TSS/(n-1)}$$
- SM has yet another version AIC_c which may be better than AIC for linear models

- Criterion-based procedures — most widely used
- AIC, BIC, Mallows C_p, R_a^2 — RSS: residual sum of squares
- $$AIC = n\log(RSS/n) + 2p$$
- $$BIC = n\log(RSS/n) + \log(n)p$$
- $$C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$$
- $$R_a^2 = 1 - \frac{\tilde{\sigma}^2_{model}}{TSS/(n-1)}$$
- 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


No safe spaces for sexists
Patrilineal/fraternal syndrome scale*, 2017 (on a scale of 0-16)
↑ More sexist
↓ Less sexist
South Sudan
Nigeria
Iraq
Pakistan
Yemen
India
Saudi Arabia
Somalia
Afghanistan
Qatar
Uganda
China
Azerbaijan
South Korea
Colombia
Sweden
United States
16
12
8
4
0
0
20
40
60
80
100
120
Fragile States Index score, 2021 (out of 120)
← Less fragile
More fragile →
x
y
Sources: Fund for Peace;
Valerie Hudson, Texas A&M
*Based on factors including unequal treatment of women in law, early marriage for girls,
polygamy, bride price, son preference, violence against women and social attitudes towards it

- "In "The First Political Order: How Sex Shapes Governance and National Security Worldwide", Ms Hudson, Ms Bowen and Ms Nielsen rank 176 countries on a scale of 0 to 16 for what they call the "patrilineal/fraternal syndrome". This is a composite of such things as unequal treatment of women in family law and property rights, early marriage for girls, patrilocal marriage, polygamy, bride price, son preference, violence against women and social attitudes towards it"

- "Ms Hudson and her co-authors tested the relationship between their patrilineal syndrome and violent political instability. They ran various regressions on their 176 countries, controlling for other things that might foster conflict, such as ethnic and religious strife, colonial history …"

- “They did not prove that the syndrome caused instability – that would require either longitudinal data that have not yet been collected or natural experiments that are virtually impossible with whole countries”

- “But they found a strong statistical link. The syndrome explained three-quarters of the variation in a country’s score on the Fragile States index compiled by the Fund for Peace, a think-tank in Washington.” ??

- Book website
- Blog

$-.75$ $\{$ 1 1 0 0 1 1 1 0

$.33$

$\bar{x} = \frac{\# 1\text{'s}}{97}$

$1 - \hat{p}$

$-\hat{p}$

Ω

$$\lim_{n\to\infty} \frac{\text{a.var } \hat{\beta}_{LS}}{\text{a.var } \hat{\beta}_{ML}} \qquad \frac{\sigma^2 (X^T X)^{-1}}{\sigma^2 (X^T X)^{-1} \frac{\pi}{2}} \qquad ?$$

$\text{a.var } \hat{\beta}_{ML} =$

$= \cancel{E}$ $\quad$ $\cancel{i_i(\beta) = 2E\left\{\frac{\partial \ell}{\partial \beta}\frac{\partial \ell}{\partial \beta^T}\right\}\frac{1}{n}}$

$\to [i(\beta)]^{-1}$

$\hat{\beta}_{ML} = \text{med}(|\hat{\varepsilon}_i|)$

$?\text{true}?$ $\xrightarrow{?}$ $i(\beta) = \text{cov}_\beta\left(\frac{\partial \ell}{\partial \beta}\right)$

$= E_\beta\left\{\frac{\partial \ell}{\partial \beta}\left(\frac{\partial \ell}{\partial \beta}\right)^T\right\}$ $\Big\}$ lik. theory

$$\arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i^T\beta| \triangleq \hat{\beta}_{ML}$$

$$\min_{\mu} \int |x-\mu| f(x)\,dx = \text{median of } f(\cdot)$$

$$f(y|x)f(x)$$

$$\min_{\beta}: \sum_{i=1}^{n} + \sqrt{(y_i - x_i^T\beta)^2} = \ell(\beta)$$

$$\text{var}(\hat{\beta}) = E\{\text{var}(\hat{\beta}|x)\} + \text{var}\{E(\hat{\beta}|x)\}$$

- “Examining approximately 176 nations, we examined whether national outcomes such as conflict, terrorism, poverty, and so forth, were significantly associated with a subordinative first political order, while controlling for background factors such as level of urbanization, levels of ethnic fractionalization, colonial history, and so forth.
- “Holding these characteristics constant, is that subordinative order strongly related to national outcomes? In all we examined 122 national outcome measures related to conflict, stability, governance, prosperity, health, demographics, education, environmental preservation, and social progress.
- “Across all 122 outcome variables, the subordination of women was both significant and the explanatory factor with the largest or second largest effect size over 70% of the time.

- common objectives

- 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

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

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

… design of studies

- 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