STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 4: Drawing Conclusions

Will curve up the mark later at the end of the term

The first may be harder, the next one may contain more calculation

(possibility: take the higher score of two quizzes)

Midterm: Oct. 24 2 h 5~6 problems

Parameter Interpretation

- meaning of parameter estimates:
- e.g., $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = 15 + 3X_1 + 4X_2 - 2X_3$
- coefficient for X_1 is 3, meaning: an increase of 1 unit in X_1 will be associated with an increase of 3 units in Y, when other are held constant no interaction b/c of potential correlation
- will a change in X_1 affect other X's in this model? It'll change
- association concluded from an observational study Sth. you cannot manipulate $\not\Rightarrow$ causation (possible from a randomized experiment)
- it is possible that the sign of a parameter estimate can change if a new variable is added

Parameter Interpretation - con't

- Berkeley Guidance Study Data, consider $n = 70$ girls
  Y : $soma$ - body type, 1 to 7 (thin to fat)

sometimes, if the objects have an order, we treat it as a continuous distribution.

they're somehow related

$$
\begin{aligned}
WT2 &= \text{weight at age 2} \\
WT9 &= \text{weight at age 9} \\
WT18 &= \text{weight at age 18}
\end{aligned}
$$

the difference reduces the relation

$$
\begin{aligned}
DW9 &= WT9 - WT2 \\
DW18 &= WT18 - WT9
\end{aligned}
$$

- sometimes we use meaningful linear contrasts instead of the original predictors to enhance interpretability

Parameter Interpretation - con't

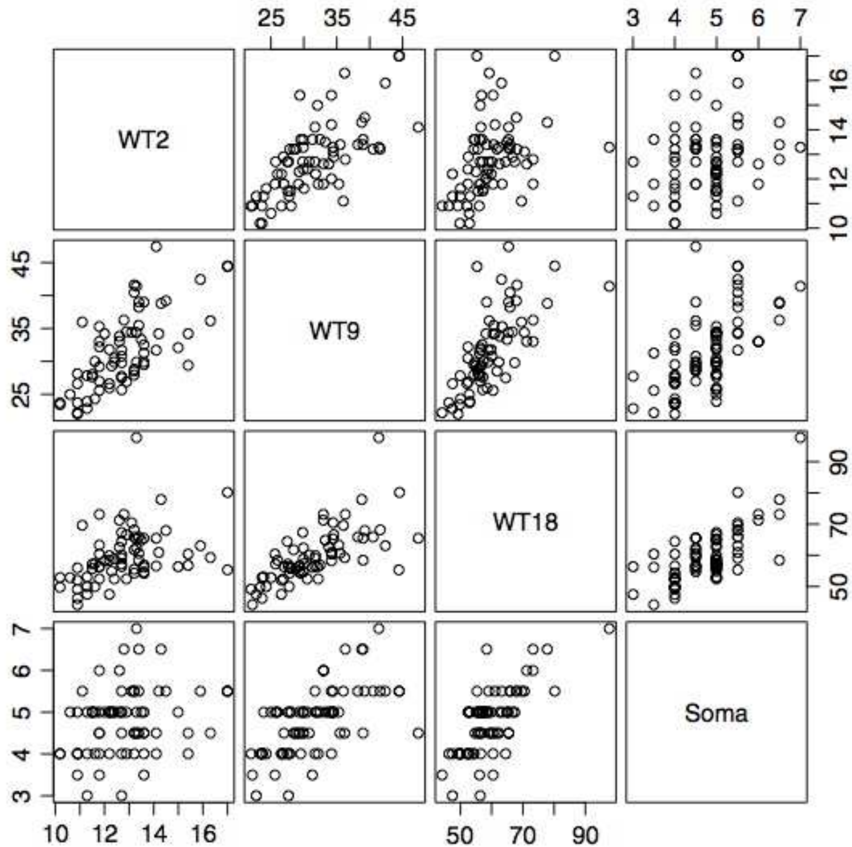


FIG. 4.1 Scatterplot matrix for the girls in the Berkeley Guidance Study.

Parameter Interpretation - con't

Term	Model 1	Model 2	Model 3
(intercept)	1.5921	1.5921	1.5921
$WT2$	-0.2256	-0.0111	-0.2256
$WT9$	0.0562		0.0562
$WT18$	0.0483		0.0483
$DW9$		0.1046	NA
$DW18$		0.0483	NA

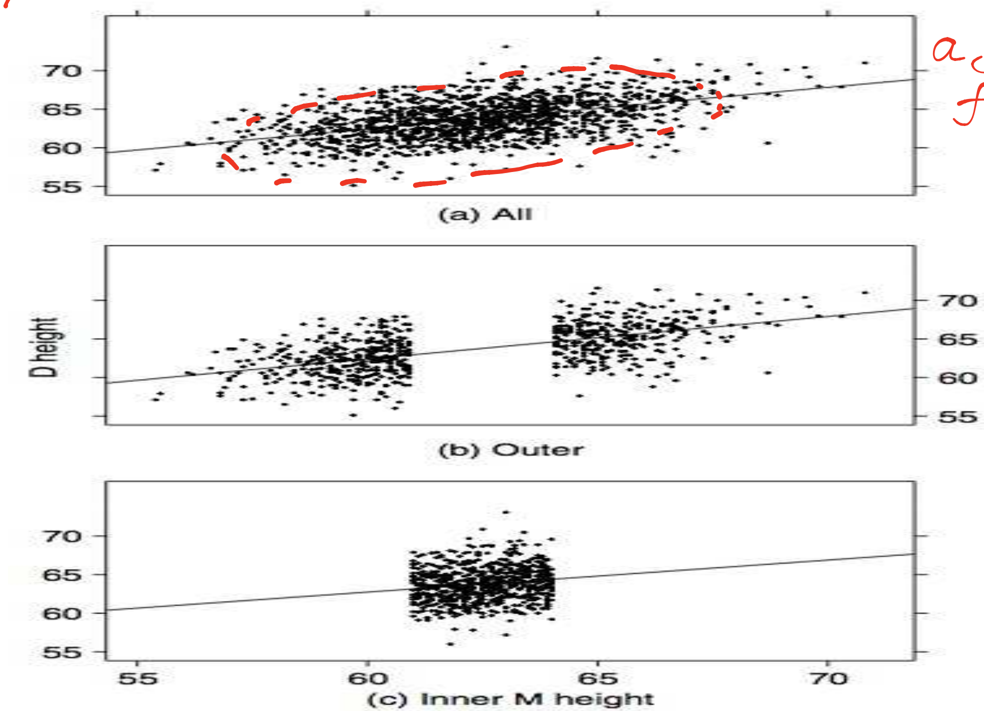
Typo here: Model 1 & 3 should be identical

- same model, different parameterization
  $\Rightarrow R^2$, $\hat{\sigma}^2$ are identical, but estimates and t-values are not
- $WT2$: significant in Model 1 (is -0.2256 surprising?) but not in Model 2 (which makes more sense?) -0.2256 does. For model 2: the coefficient is estimated in another linear relation. (Co-linearity b/w WT2 & DW9 is weak
- why is 0.0483 for $WT18$ and $DT18$ identical? why NA in Model 3?
  If hold fixed then one increment in DW18
  → one increment in WT18 (linearly equivalent)

More on R^2

- Fig 4.2(a): R^2 =0.24 Fig 4.2(b): R^2 =0.37 Fig 4.2(c): R^2 =0.027
- random sampling is important for R^2 to make sense

conclusion



a good model fit

FIG. 4.2 Three views of the heights data.

More on R^2 - con't

- R^2 can be meaningless for some situations
-

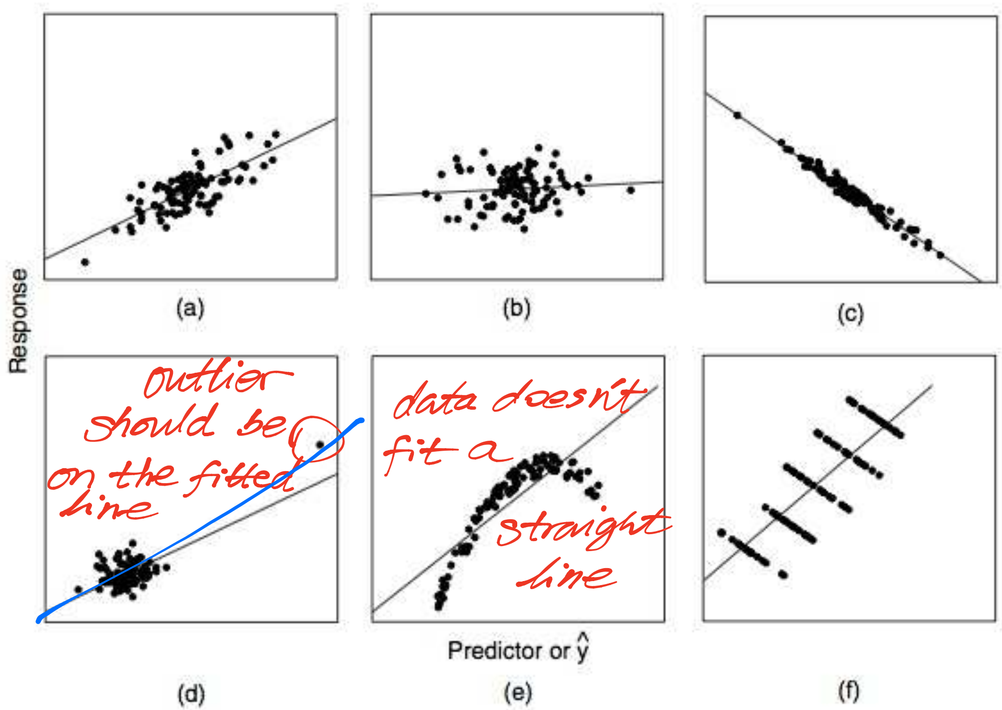


FIG. 4.3 Six summary graphs. R^2 is an appropriate measure for a–c, but inappropriate for d–f.

Sampling from Normal Population

- data: $(x_1, y_1), \cdots, (x_n, y_n)$
- $\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$
- what is the conditional distribution of y_i given x_i?
- $y_i|x_i \sim \mathrm{N}\left(\mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x), \sigma_y^2(1-\rho_{xy}^2)\right)$
- define $\beta_0 = \mu_y - \beta_1\mu_x$, $\beta_1 = \rho_{xy}\frac{\sigma_y}{\sigma_x}$, $\sigma^2 = \sigma_y^2(1-\rho_{xy}^2)$
- $y_i|x_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$
- $\hat{\mu}_x = \bar{x}$, $\hat{\mu}_y = \bar{y}$, $\hat{\sigma}_x^2 = \frac{SXX}{n-1}$, $\hat{\sigma}_y^2 = \frac{SYY}{n-1}$, $\hat{\rho}_{xy} = \frac{SXY}{\sqrt{SXX \cdot SYY}}$
- plug-in to get $\hat{\beta}_0$, $\hat{\beta}_1 \Longrightarrow$ OLS estimates

Conditional mean

Conditional Var

$\boxed{\beta_0 = \mu_y - \rho_{xy}\frac{\sigma_y}{\sigma_x}\mu_x}$

$\underline{E(y_i|x_i) = \beta_0 + \rho_{xy}\frac{\sigma_y}{\sigma_x}x_i}$

$E(y_i|x_i) = \beta_0 + \beta_1 x_i$

$\hat{\mu}_y = \bar{Y}$, $\hat{\mu}_x = \bar{X}$

$\hat{\sigma}_y = S_y$, $\hat{\sigma}_x = S_x$

$\hat{\rho}_{xy} = r_{xy}$

$\hat{\beta}_0 = \bar{y} - r_{xy}\frac{S_y}{S_x}\bar{x}$

$\hat{\beta}_1 = r_{xy}\frac{S_y}{S_x} = \frac{SXY}{SXX}$

How to Handle Missing Data?

- first we need to understand why some data are missing
- "missing at random" (MAR) is the easiest to handle
- MAR: probability of missing does not depend on its value
- two simple strategies: deleting and guessing
- more advanced method: imputation - need statistical modeling

Computationally Intensive Methods

Get some rough idea.

- suppose $X_1, \cdots, X_n \sim \mathrm{N}(\mu, \sigma^2)$
- what is $\mathrm{Var}(\bar{X})$?
- what is $\mathrm{Var}(\tilde{X})$, where $\tilde{X}$ is the median of $X_1, \cdots, X_n$?
- what is $\mathrm{Var}(\bar{X} + \tilde{X}^2)$?
- we can use computers instead of calculus
- suppose $y_1, \cdots, y_n$ from the distribution G
- want to construct a 95% C.I. for the median
- two cases: G is known and G is unknown

Case (i): G is known

- four steps:
  1. obtain a sample $y_1^*, \cdots, y_n^*$ from G
  2. compute the median and store its value
  3. repeat Steps 1 and 2 many times
  4. suppose we repeat 1000 times, so we have 1000 medians. Then a 95% C.I. for the median of G is

     $$(25^{\text{th}} \text{ smallest}, 25^{\text{th}} \text{ largest})$$

- it can be extremely difficult to generate from G. have you heard about Monte Carlo?
- but typically unrealistic to assume G is known

Case (ii): G is unknown

- only one change
- replace Step 1 by:
  obtain a sample $y_1^*, \cdots, y_n^*$ by drawing n data points from $y_1, \cdots, y_n$ with replacement
- yes, some of the entries will be repeated
- this method is called bootstrap
  (sounds familiar? Pirates of Caribbean!)