Lecture 0

§1 Basics of regression

1. 为什么选择 regression

① understanding & interpretation：理解变量是如何相互联系的

② prediction：根据新的 features 进行准确的 prediction

2. 怎样确定一个 regression model

① random component：描述 response values 的 distribution

e.g. Normal / Binomial / Poisson ....

② Systematic component：explanatory variables 和 mean of response 间的 mathematical relationship

e.g. 通常情况下，关于 parameters 线性的 regression models 的 systematic component 形式为：

$$E[y] = \mu = f(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

③ Notations：

- explanatory variables 的数量 p：$x_1, \cdots, x_p$
- regression parameters 的数量 p'：若 regression 含有 constant term β_0，则 $p' = p+1$；否则 $p' = p$

3. Random component 的选取

① 考虑 the scale of measurement，即 continuous 还是 categorical？

② 若为 categorical，考虑有多少个 categories，它们是 nominal 还是 ordinal？

③ 考虑 distribution 的 shape，可以通过 frequency tables，dot plots，histograms，或其他 graphical methods 得出.

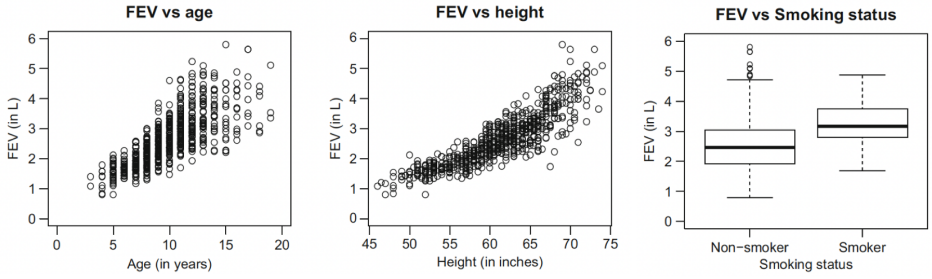
4. Systematic component 的选取

考虑变量之间的 association：

① 对于 categorical variables，使用 cross tabulations

② 对于 continuous variables，使用 scatter plots

③ 对于 continuous response grouped by categorical variables，使用 side-by-side box plots.



5. Model selection 的标准

① Accuracy：能准确描述 systematic 和 random components

② Parsimony：model 需要尽可能 simple

③ covariants 的数量 & covariants 的 function form 的选取

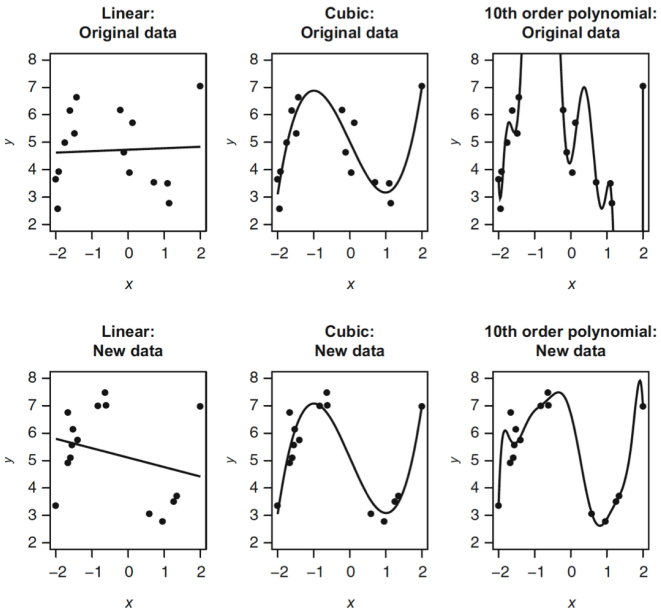


Fig. The top panels show the models fitted to some data; the bottom panels shows the models fitted to data randomly generated from the same model used to generate the data in the top panels.

b. 模型确定后的操作

① Parameter estimation
- 估计 covariants 对 response 的 main / interaction effect

② Inference 和 intepretation
- 计算 covariant effect 的 CI
- 对 covariant effect 进行 hypothesis testing

③ Interpretation
- explanatory variable 多大程度上 / 从哪个方向上能解释 outcome.

④ Model diagosis
- 检测 model 的 adequacy，多大程度上 fits / summarizes the data.