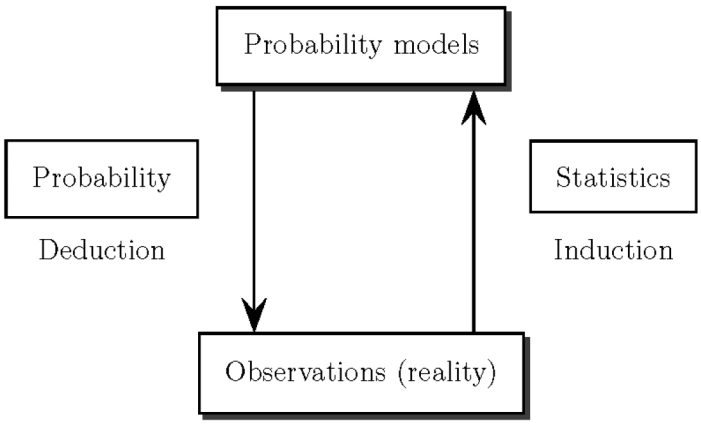
Topic 1 Scatterplots and Regression

-



- Statistical models: 2 components

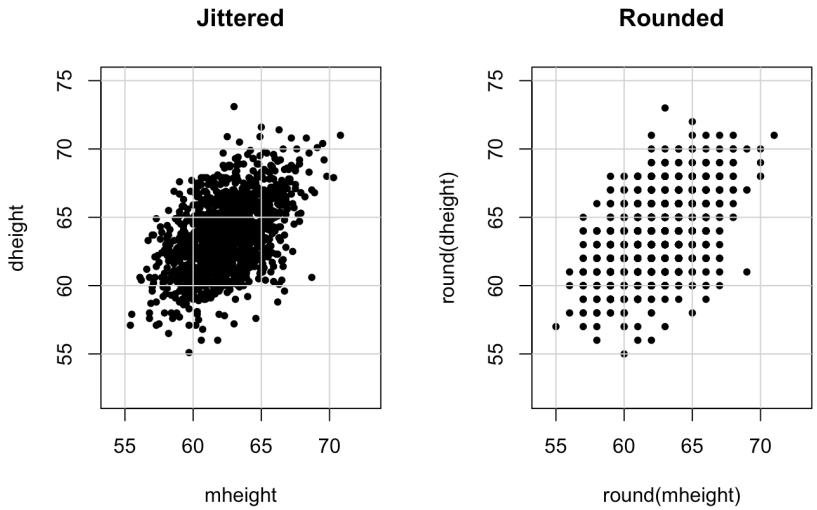
  1. Deterministic -> predictable portion (mechanisms and/or relationships)
  2. Stochastic -> unpredictable portion

  e.g.

  - Linear models
  - Generalized linear models (GLMs): discrete outcome, e.g. logistic ("J" curve) & Poisson
  - Nonparametric regression: generic deterministic component
  - Quantile regression: generic stochastic component
  - etc.
- Principles: *Agnosticism*, *Parisimony*, "*Worrying selectively*"
- e.g. Inheritance of height
  Predictors/independent variables: *X*, here is the height of mothers (`mheight`)
  Response/dependent variable: *Y*, here is the the height of daughters (`dheight`)

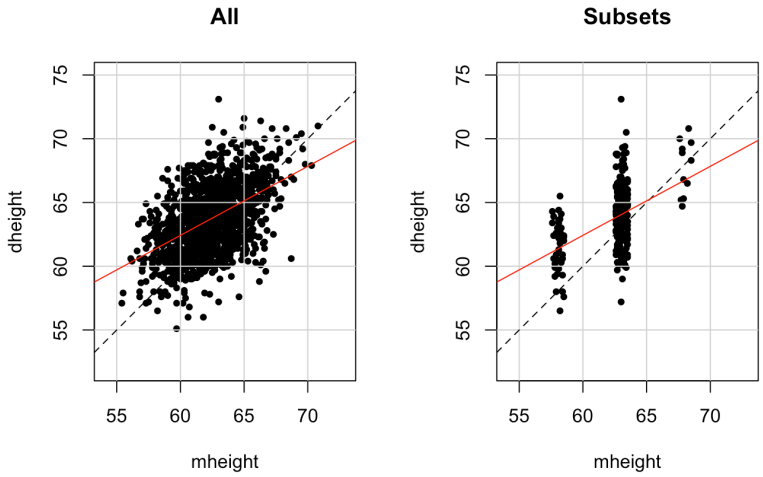
  - Jittered 抖动: "Jittering is the act of adding random noise to data in order to prevent overplotting in statistical graphs. Overplotting can occur when a continuous measurement is rounded to some convenient unit. "
    => 抖动散点图（jitter plot）可以避免常规散点图中点过于重叠的情况，比如我们测试数据中一共1000个数据，分布比较集中，如果数据量再大，就会出现很多点重合的现象。这时候，我们就可以采用抖动散点图，它对散点添加随机的“抖动”效果，将散点适当地沿*x*轴刻度两侧拓展，在一定程度上表示出了局部分布。

- Rounded 四舍五入: original data, so each point in the plot refers to one or more mother-daughter pairs => *overlapping*
- Jittering strategy: add a small uniform random number to each value, (-0.5,+0.5) here
    - in the jittered plot, the scatter of points appears to be more or less elliptically shaped, and more points in the ellipse center rather than the edges => **simple linear regression (SLR)**
    - Jittering increases variability but allows better visualization



- Dashes line: a 45°-line, indicating all mothers and daughters pairs had *exactly* the same height
    - $\beta_0 = 0$ & $\beta_1 = 1$ => daughters have the same height as their mothers on the average for mothers of any height
- *separated points*:
    - *leverage* points 杠杆点: extreme values on the left and right of the horizontal axis are points that are likely to be important in fitting regression models 远离样本空间中心的点（即：自变量x的值是极端值的观测值）
    - *outliers* 异常点: separated points on the vertical axis, here unusually tall or short daughters give their mother's height -> potential outliers
        - PS:
          异常点 (Outlier): 残差很大的点，即：因变量*y*的值是极端值的观测值
          强影响点 (Influential Point)：对模型有较大影响的点，如果删除该点能改变拟合回归方程
        - 异常点不一定是强影响点，强影响点也不一定是异常点
          高杠杆点不一定是强影响点，强影响点也不一定是高杠杆点
        - deal with outliers: the goal is prediction or average?

- One important function of the scatterplot is to decide if we might reasonably assume that the response on the vertical axis is independent of the predictor on the horizontal axis. 散点图的一项重要功能是确定我们是否可以合理地假设垂直轴上的响应独立于水平轴上的预测变量 => only show the points with `mheight` rounding to either 58, 64, or 68 inches => get strips or *slices* => vertical variability be more or less the same for each of the `mheight` fixed values
    - the "power" to pull the lines away from the average line (45 degrees) => individual subgroups



- Red solid line: simple linear regression
    - in R: `lm` = linear model
      the solid line is estimated by the `OLS` method
        - OLS: Ordinary Least Squares 最小二乘法
    - conditional average of the response increases with the predictor, in a seemingly linear way
    - The slope is less than 1 => This is the origin of the name *regression*

- **Mean function 平均函数**:
  the linear regression model stipulates the conditional expectation of the (predictors) give the (predictors) is a linear function of the (predictors)
  Here as:

$$\mathrm{E}(dheight|mheight = x) = \beta_0 + \beta_1 x$$

$$\mathrm{E}(Y|X = x) = \text{a function that depends on the value of } x \qquad (1.1)$$

- This is: a *modeling assumption* =X=> a statement of reality
  ==> simplification and efficiency, with only 2 parameters
  β_1: slope; β_0: intercept

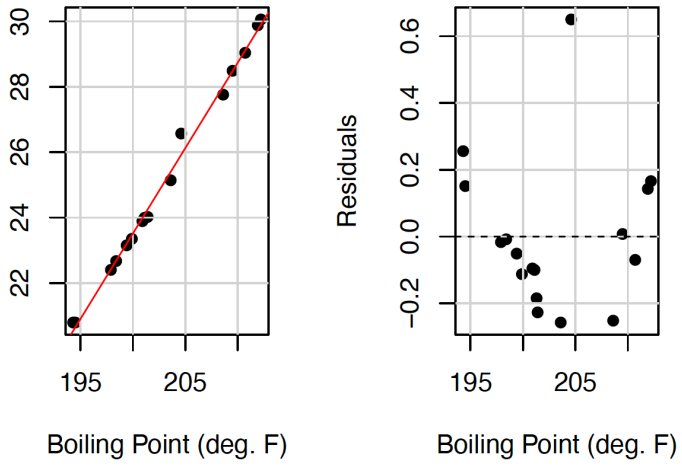
- **Variance function 方差函数**:
  为简单起见，拟合线性回归模型的常见假设是每个x值的方差函数都相同。在这种情况下：

$$\mathrm{Var}(dheight|mleight = x) = \sigma^2$$

$$\mathrm{Var}(Y|X = x) = \sigma^2 \qquad (1.4)$$

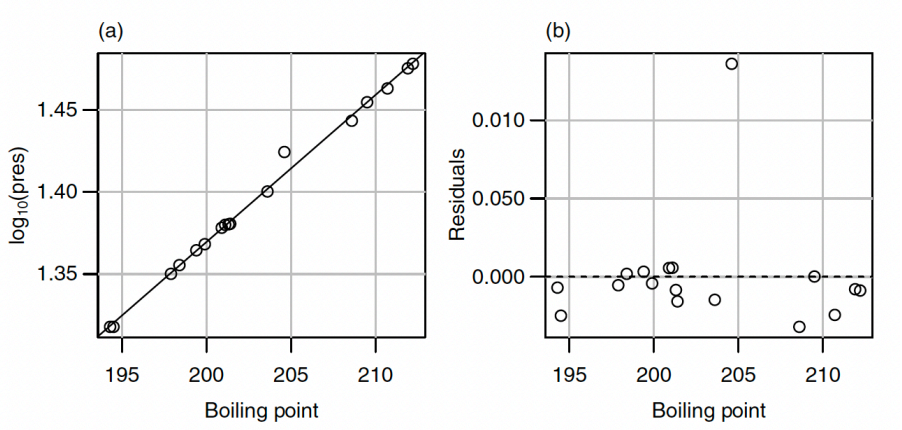
  => not dependent on x
  => This assumption is called *homoscedasticity 异方差*: 指的是一系列的随机变量间的方差不相同

- e.g. Forbes data (James D. Forbes, Scotland, relationship between atmospheric pressure and the water boiling point, $n = 17$)
  Predictors: `pres` ; Responses: `bp`
  - all the points fall almost exactly on a smooth curve => the variability in pressure for a given boiling point is extremely small
    - the mean of pressure given bp could be modeled by a straight line (given a small systematic deviation from the straight line)
    - `residual = pres` - point on the line
      => gain resolution in the plot, since the vertical axis on the left is about 10 inches of mercury while the range on the right is about 0.8 inches of mercury
      => 10/0.8 = 12.5
      => this is clearly a *curvature 曲率* in the residual plot
      => close inspection can reveal nonlinearities
      => close inspection can reveal outliers



  - Log transformation: to avoid curvature
    => log(`pres`) is linearly related to `bp`
    - e.g., a case of log transformation

$$\mathrm{E}(Y|\texttt{Dose} = x) = \beta_0 + \beta_1[1 - \exp(-\beta_2 x)] \qquad (1.3)$$



- e.g. Ft. Collins Snow `ftcollinssnow`
  Predictors: `Early` ; Responses: `Late` <= early & late: early/late season snowfall
  - solid horizontal line: the average late-season snowfall
  - dashes line: `OLS` line

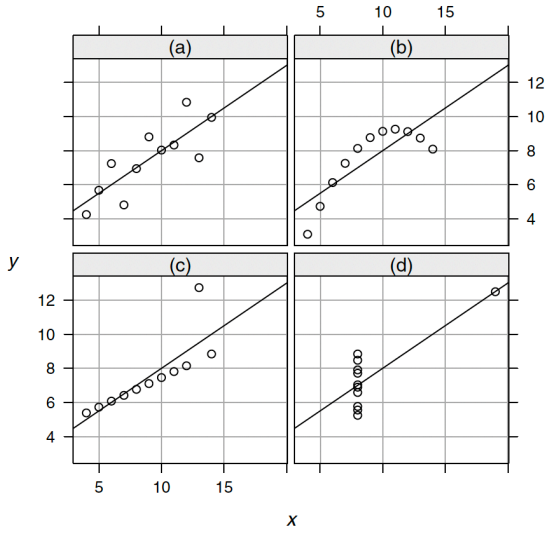
  Results:
  - variance in the data is substantial 数据差异很大
  - relationship between variables =>unclear<= where statistical test may help
    - may be completely *uncorrelated*
    - Statistical test: 最终将通过测试$\beta_1 = 0$ 的假设与 $\beta_1 \neq 0$ 的备择假设来测试

      `Early` 和 `Late` 的独立性
  - linear regression be influenced by a few large observations
  - no mathie theory of significance yet
- SUMMARY GRAPH
  - Being able to fit a linear regression does not imply that the relationship between the variables is linear.
    能够拟合线性回归**并不**意味着变量之间的关系是线性的
  - The relationship may be more complex. The linear regression fit only captures the linear portion of the relationship.
    关系可能更复杂 & 线性回归拟合仅捕获关系的线性部分

  e.g. Anscombe's quarter: artificial dataset, $n = 11$, where SLR mean function is fit
  => each data set leads to an identical summary analysis with the same β_0 & β_1

  <= but the visualization is quite different
  - (a) - Standard: SLR is appropriate

- (b) - Quadratic: SLR is incorrect and that is a smooth curve (maybe quadratic polynomial 二次多项式?)
- (c) - High influence: SLR might be correct for most of the data, but one of the cases is too far from the fitted regression line => *outlier problem*
    - possibly the case that does not match the others should be deleted from the data set => different fitted line
    - Without a context for the data, we cannot judge one line as "correct" and the other as "incorrect"
- (d) - High leverage: not enough info to make a judgment concerning the mean function
    - must distrust an analysis that so heavily dependent upon a single case



- Tools for looking at scatterplots:
    1. *Size*: choose appropriate scales, units and limits
        - goal: to extract all the available information
        - by changing scales, by resizing, or by removing linear trends
    2. *Transformations*: either variable may be transformed
        - e.g. log or X by X^{λ}
        - 由于对数变换如此频繁地使用，将把$\lambda = 0$解释为对应于对数变换
    3. *Nonlinear mean function*: estimate the mean conditional on subsets of the predictor or apply a smoother 平滑器 such as **LOESS**
        - LOESS = locally estimated scatterplot smoothing 局部估计散点图平滑
        - 粗略地说，LOESS smooth通过将直线拟合到最接近x的点的一部分来估计x点处的$E(Y|X = x)$
            - 在该图中使用了0.20的分数 fraction (因为样本量太大)，但更常见的是

将分数设置为2/3左右
    - 通过连接许多x值的$E(Y|X = x)$估计值来获得smoother

  对于接近平均值的 `mheight` ，LOESS平滑和直线几乎完全一致，但对于数据少得多的较大 `mheight` 值，它们的一致性较差

- 平滑器在图的边缘往往不太可靠

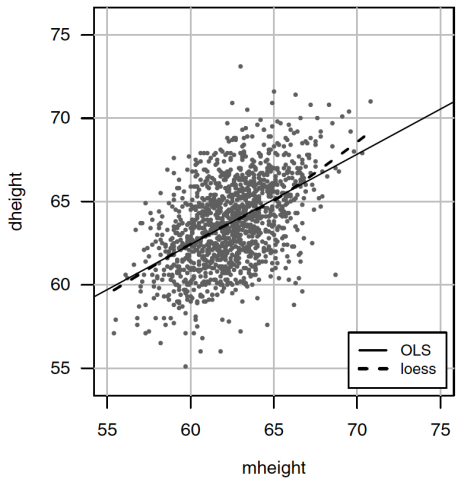
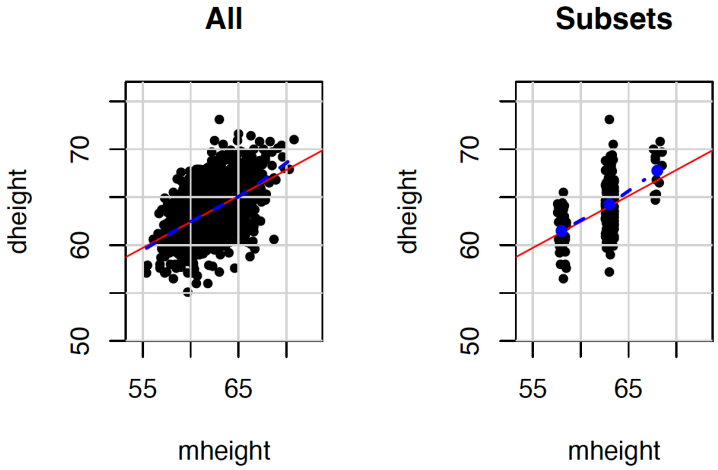


Figure 1.10 Heights data with the OLS line and a loess smooth with span = 0.20.

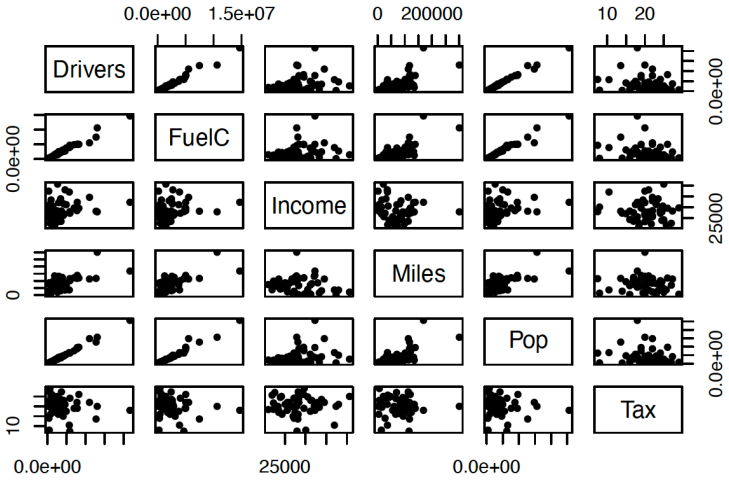


- Linear or nonlinear?
  Model choice dependent on the GOAL: explanatory 解释性 or predictive 预测性
    - 非正式评估：explanatory model 解释模型
        - nonlinear fit & linear fit => similar
        - relationship between the variables may in fact be nonlinear, but we can not be completely sure
        - linear model: captures the relationship well in a SIMPLE model with only 2 parameters
        - nonlinear model: nonparametric, there is no easy way to summarize it in a few parameters or to represent it other than plotting it
        - *Occam's razor*: 在解释数据效果相同的两个模型之间，我们倾向于更**简单**的一个

    - 正式评估：
        - *Explanatory model*: main tool - *testing*
          人们可能会假设这种关系是线性的，并寻找证据证明它不是线性的。 适当的*p*值可以评估观察到实际观察到的非线性程度的概率。
        - *Predictive model*: 人们可以检查哪个模型的预测误差较低
- Scatterplots matrices (pairs plot)
  e.g. fuel consumption data

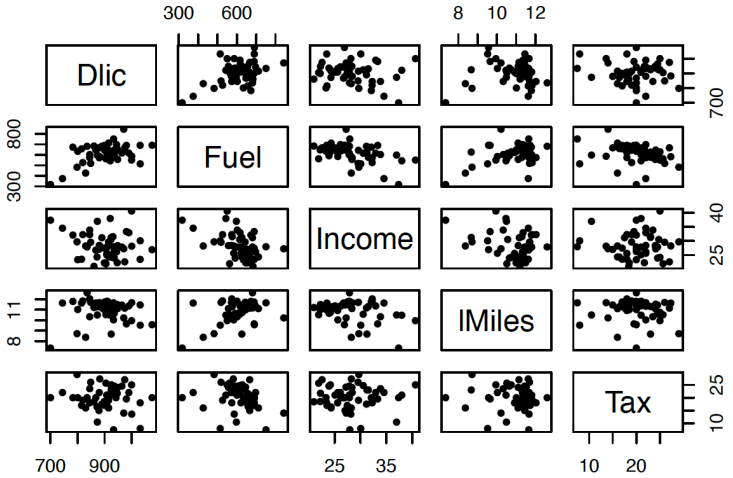
Variable	Description
`Drivers`	Number of licensed drivers in the state
`FuelC`	Gasoline sold for road use, thousands of gallons
`Income`	personal income for the year 2000, in thousands of dollars
`Miles`	Miles of Federal-aid highway miles in the state
`Pop`	2001 population age 16 and over
`Tax`	Gasoline state tax rate, cents per gallon

  => Pairs plot



    - Rages are big: better to use larger units
    - `Drivers` & total fuel consumption are almost proportional to the state's population: 可能更好地考虑每1000名居民的驾驶员数量和燃料消耗
    - `Miles` has a skewed distribution: better to apply a log distribution

  => transformed plots



- SUMMARY:
  - **Always plot the data first**
  - Ranges, graph limits, transformations => matters
  - get an intuitive understanding of the situation in the context of the problem
  - Postulate a plausible model for the data 假设数据的合理模型
  - Evaluate your model mathematically