应用回归分析

上海财经大学　统计与管理学院

基本信息

- 教室：四教206 (武川路校区)
- 时间:　周四 13:20 - 16:10　　(3 ×12 ＝ 36课时)
- 答疑：周二16:00-17:00或预约
- 办公室：统计与管理学院2205;
- 电话：65901206
- 电邮：huang.tao@mail.shufe.edu.cn
- 课件：见学校BB教学资源平台

考核形式

- 考勤：5%
- 作业：5*6%= 30%
- 期末：65%
-
- **学术诚实**

  涉及学生的学术不诚实问题主要包括作业及考试的作弊；抄袭；伪造或不当使用在校学习成绩；未经老师允许获取、利用考试材料。对于学术不诚实的最低惩罚是考试给予0分。其它的惩罚包括报告学校相关部门并按照有关规定进行处理。

章节内容 Chapters

章节内容 Chapters

第一章 散点图与回归

❖ 章节概括：

- 回归分析研究依赖性
- 数据的拟合、推断与预测；
  模型的构建、选择和推断
- 线性回归模型　（统计学数学）
- 数据图形化　　（散点图）
- 二维散点图、散点图矩阵

例子：身高的遗传性

- 1893–1898, E. S. Pearson，UK
- 1375 组母女的身高

- 身高是否有遗传性？
- 怎样刻画遗传性？

……

一、二维散点图

数据

- 自变量　predictor　X
- 因变量　response　Y
- 单个观测　(x, y)
- 样本　(xi, yi), i = 1, . . . , n,
- 二维散点图描述　X与Y 间的关系

一、二维散点图

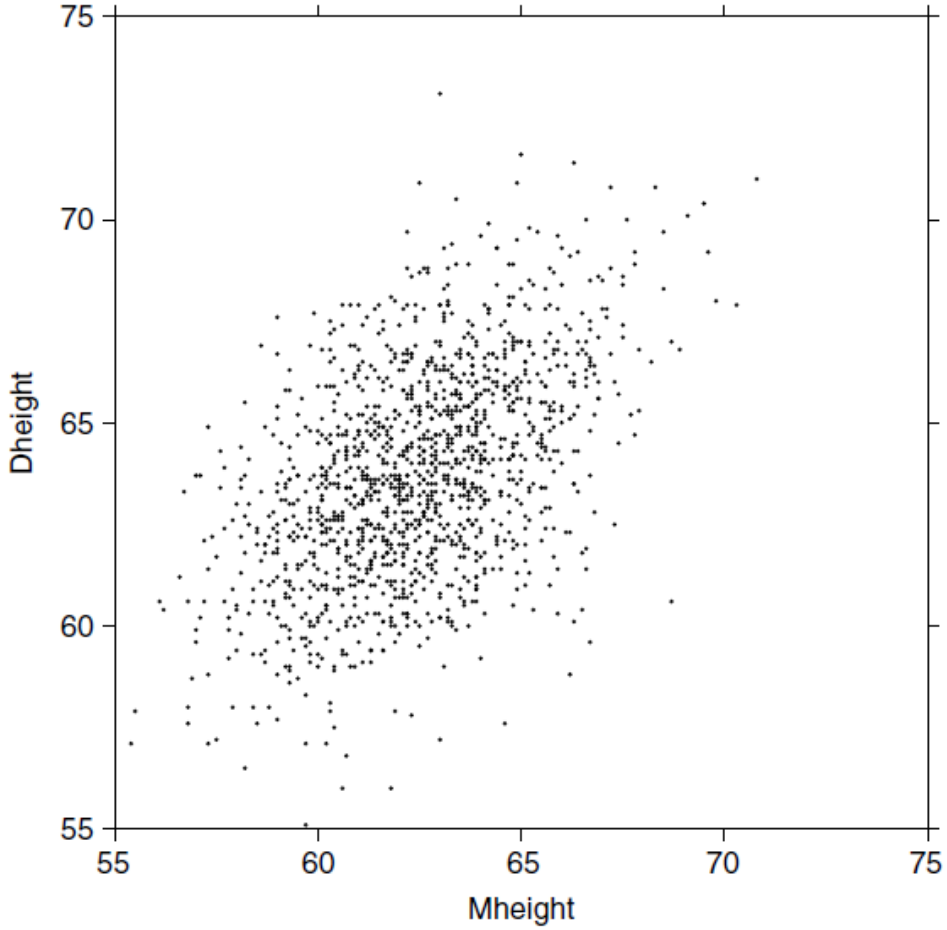
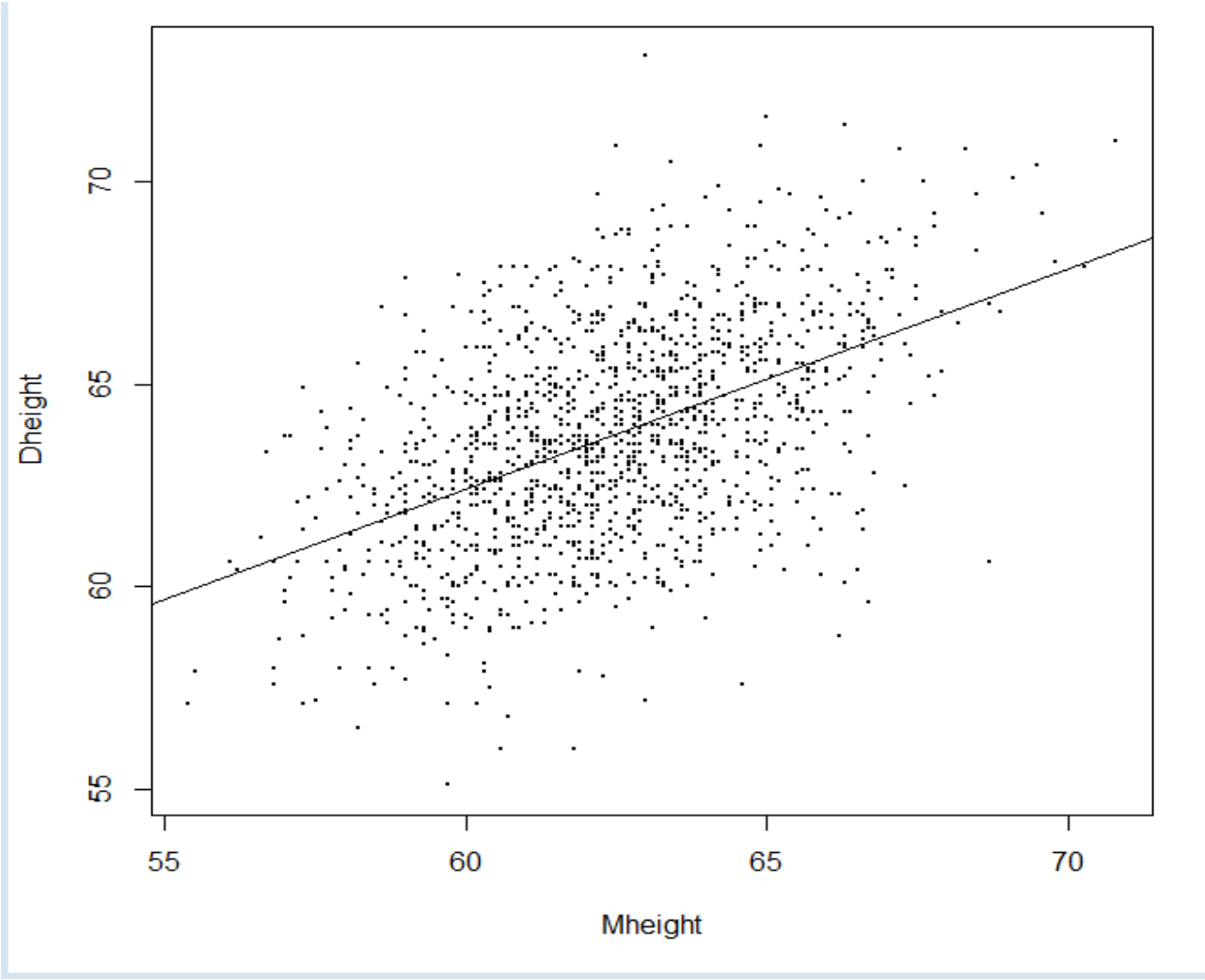


FIG. 1.1 Scatterplot of mothers' and daughters' heights in the Pearson and Lee data. The original data have been jittered to avoid overplotting, but if rounded to the nearest inch would return the original data provided by Pearson and Lee.

R 简介

- 电脑演示

Pearson & Lee Data



Forbes Data I

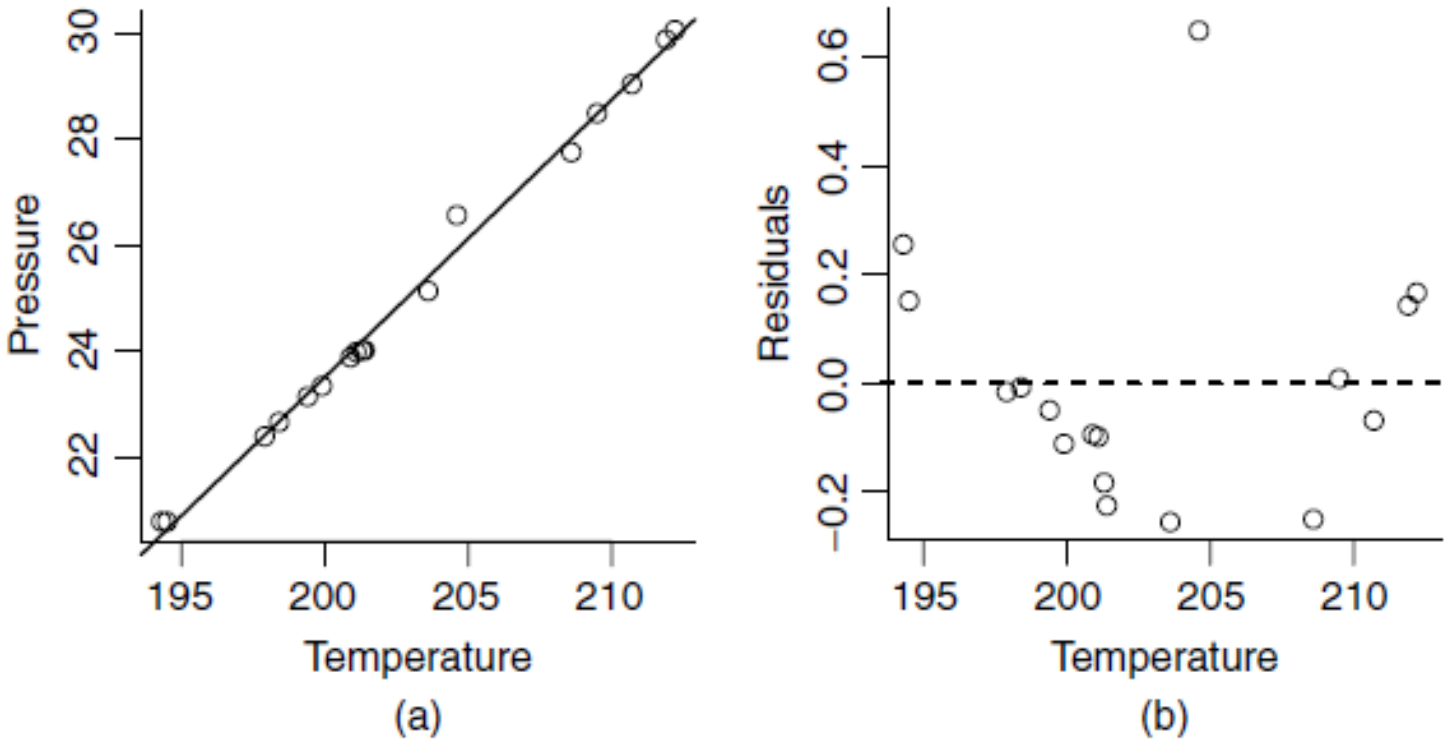


FIG. 1.3 Forbes data. (a) *Pressure* versus *Temp*; (b) Residuals versus *Temp*.

Forbes Data II

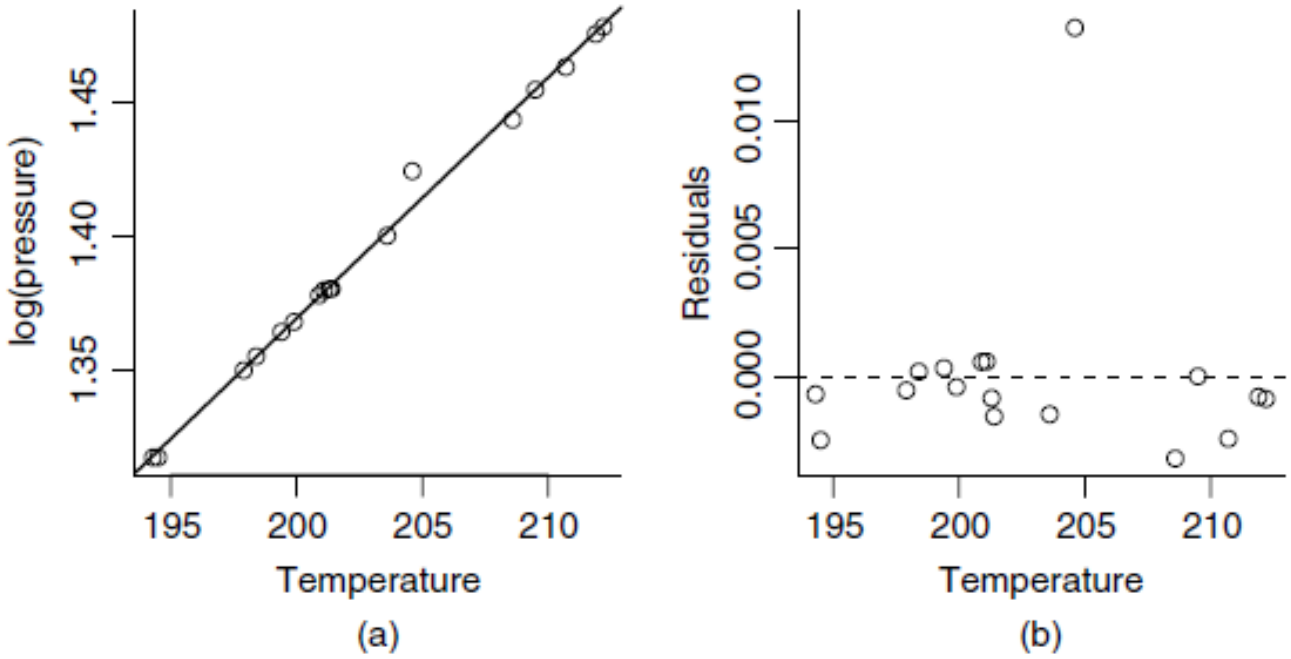


FIG. 1.4 (a) Scatterplot of Forbes' data. The line shown is the OLS line for the regression of log(*Pressure*) on *Temp*. (b) Residuals versus *Temp*.

Smallmouth Bass Data

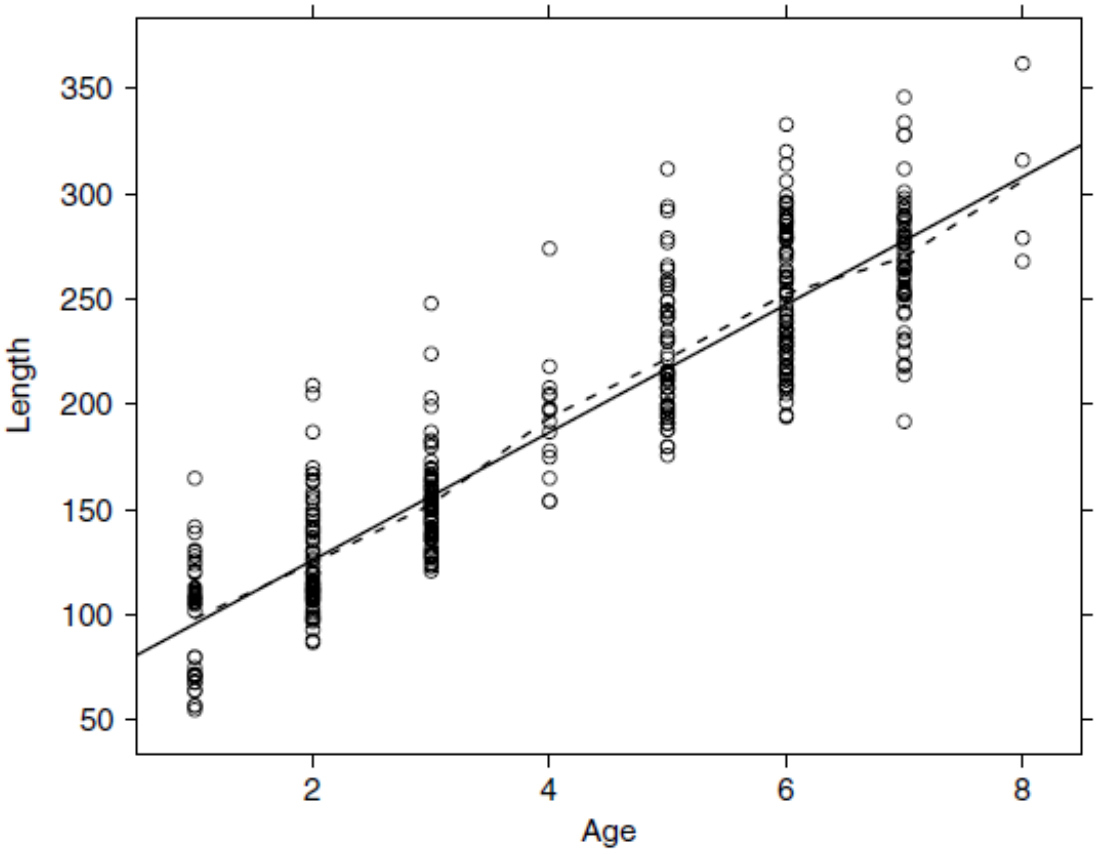


FIG. 1.5 *Length* (mm) versus *Age* for West Bearskin Lake smallmouth bass. The solid line shown was estimated using ordinary least squares or OLS. The dashed line joins the average observed length at each age.

Snowfall Data

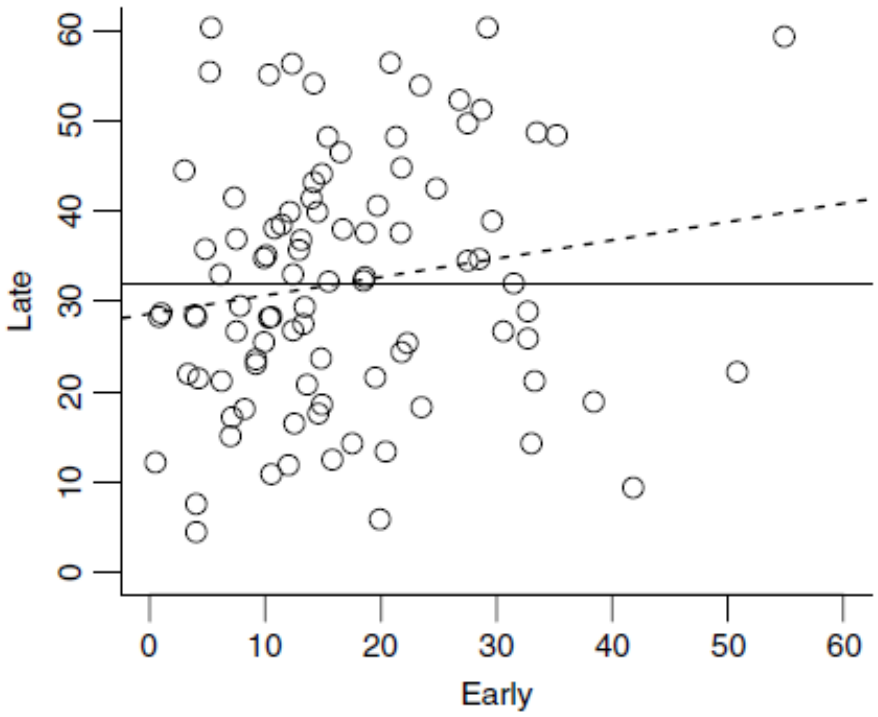


FIG. 1.6 Plot of snowfall for 93 years from 1900 to 1992 in inches. The solid horizontal line is drawn at the average late season snowfall. The dashed line is the best fitting (ordinary least squares) line of arbitrary slope.

Turkey Data

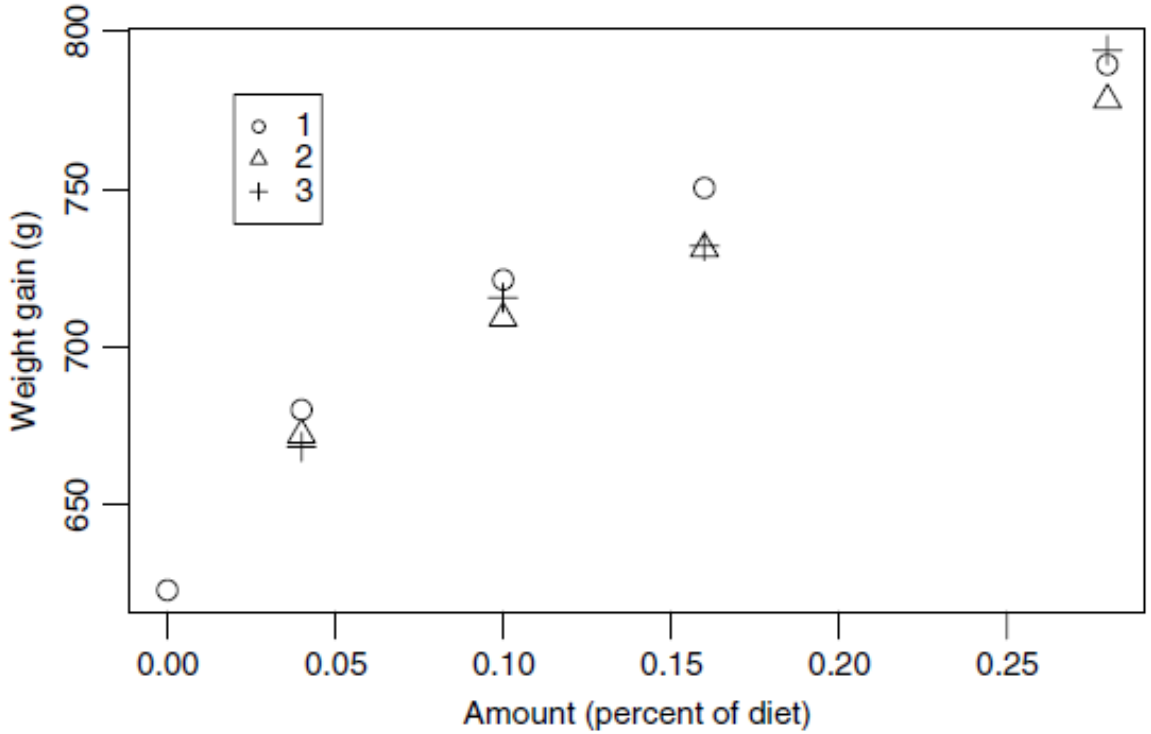


FIG. 1.7 Weight gain versus *Dose* of methionine for turkeys. The three symbols for the points refer to three different sources of methionine.

二、回归模型

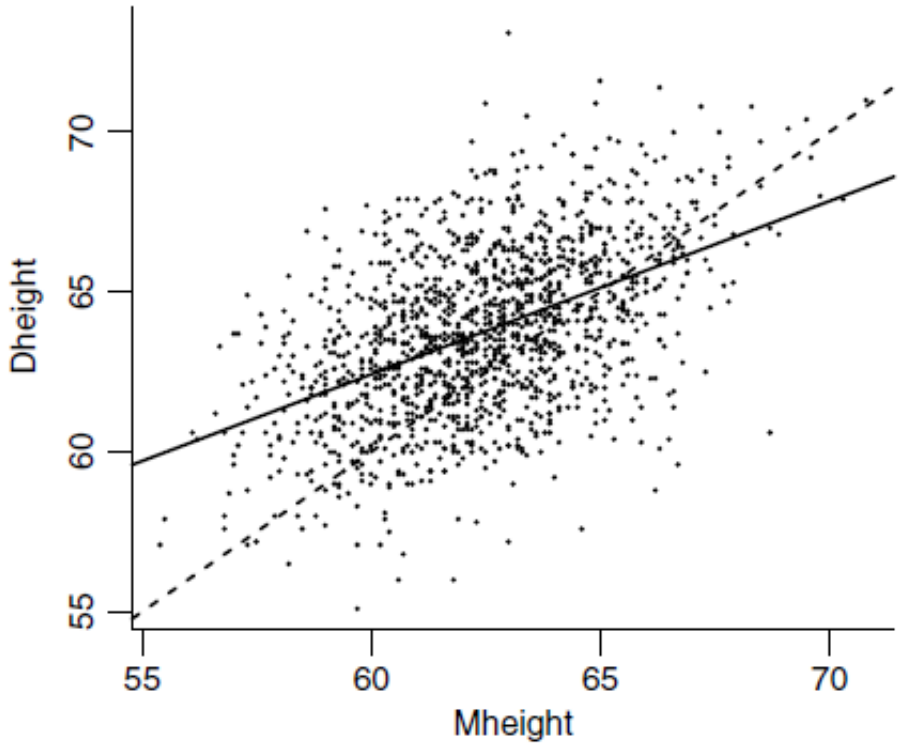


FIG. 1.8 The heights data. The dashed line is for E(*Dheight*|*Mheight*) = *Mheight*, and the solid line is estimated by OLS.

二、回归模型

❖ 回归模型：

- 均值函数
- 方差函数
- 残差

均值函数 Mean Function

- $\mathrm{E}(Y|X=x)$

- 描述Y的平均值如何随着X的变化而变化；
- 参数函数（线形函数）、非参函数
- $\mathrm{E}(Dheight|Mheight=x)=\beta_0+\beta_1 x$
- 参数（截距、斜率）
- 估计、检验、预测

方差函数 Variance Function

- $\mathrm{Var}(Y|X=x)$

- 描述Y的分散度如何随着X的变化而变化；
- 常数、不随着X的变化而变化
- $\mathrm{Var}(Y|X=x)=\sigma^2$
- 方差：未知、参数
- 估计、检验

残差 Residual

$$Residual = Y - \hat{Y}$$

- 观测值与预测值之间的差别
- 随机性

回归分析

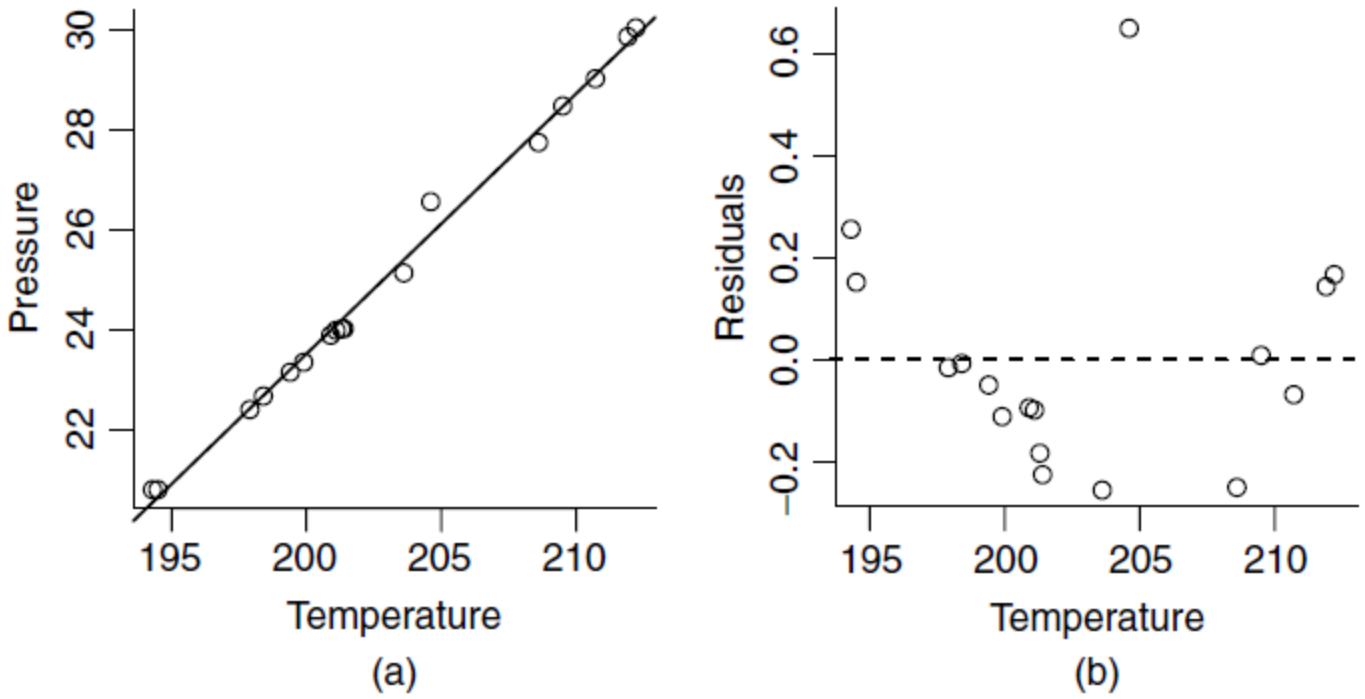


FIG. 1.3 Forbes data. (a) *Pressure* versus *Temp*; (b) Residuals versus *Temp*.

三、散点图初析

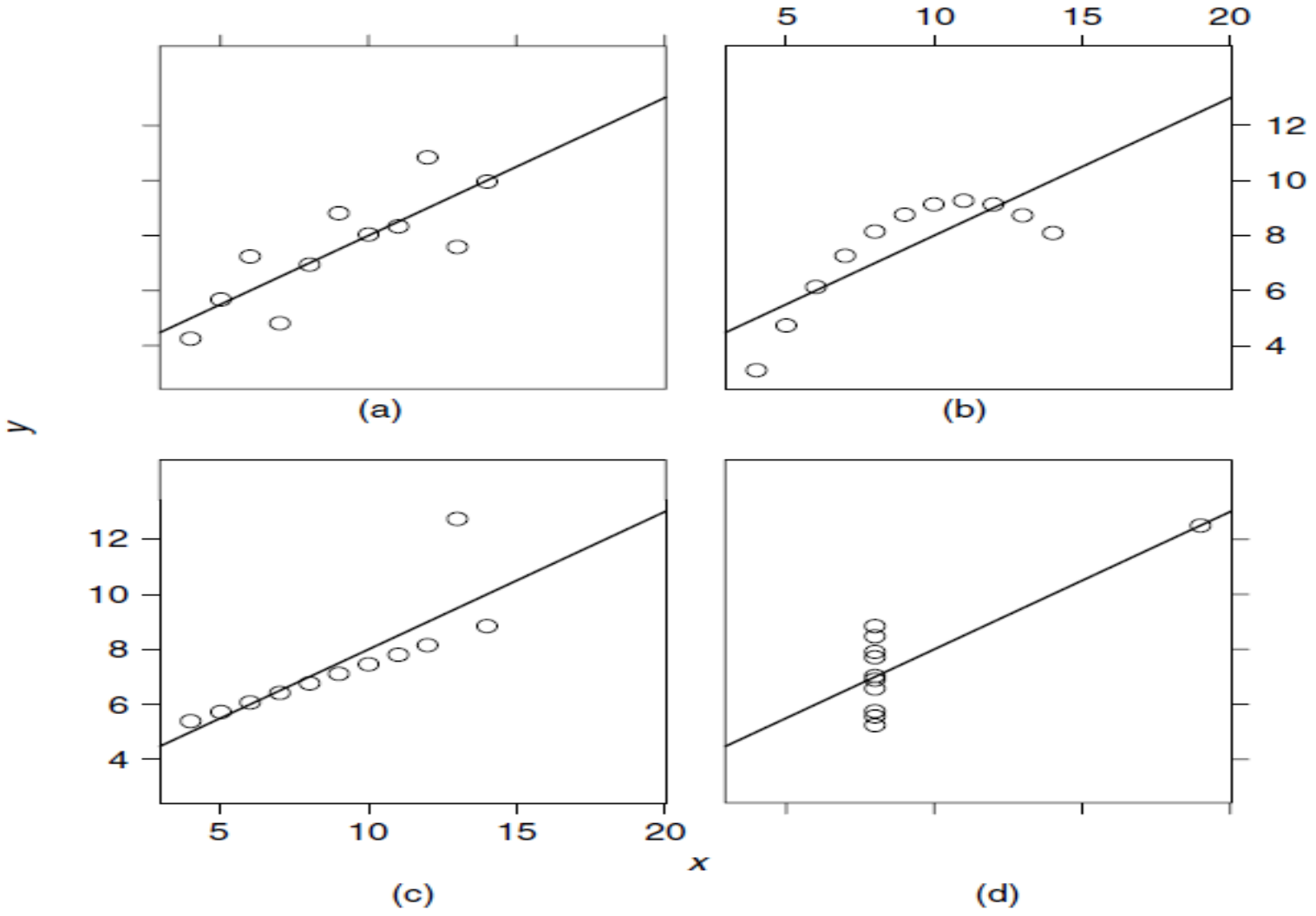


FIG. 1.9 Four hypothetical data sets (from Anscombe, 1973).

三、散点图初析

- **变量转换 (variable transformation)**
- **非线性回归 (nonlinearity)**
- **离异值 (outlier)**
- **……**

变量转换

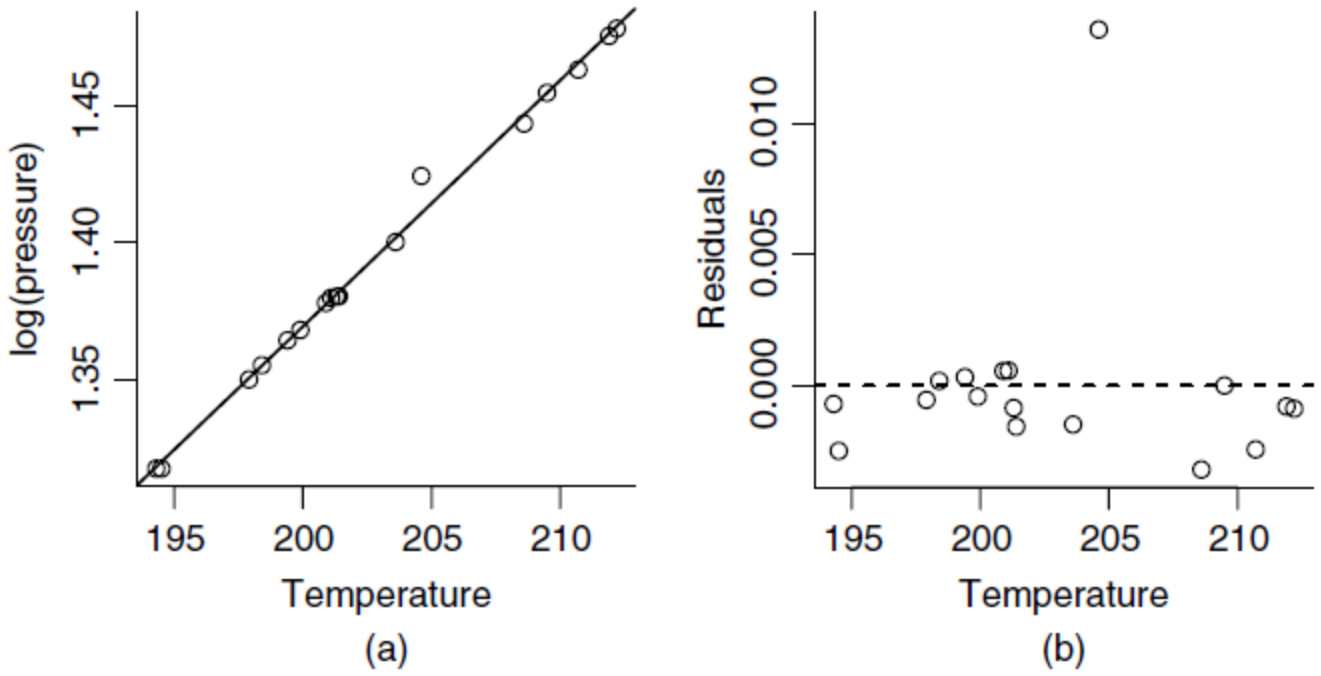


FIG. 1.4 (a) Scatterplot of Forbes' data. The line shown is the OLS line for the regression of log(*Pressure*) on *Temp*. (b) Residuals versus *Temp*.

四、散点图矩阵

- **散点图：　　一对一**
- **散点图矩阵：　一对多（多个自变量）**

- **边际效应与联合效应**

例子：耗油量分析

TABLE 1.2 Variables in the Fuel Consumption Data[a]

Drivers	Number of licensed drivers in the state
FuelC	Gasoline sold for road use, thousands of gallons
Income	Per person 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
State	State name
Fuel	$1000 \times Fuelc/Pop$
Dlic	$1000 \times Drivers/Pop$
log(*Miles*)	Base-two logarithm of *Miles*

Source: "Highway Statistics 2001," http://www.fhwa.dot.gov/ohim/hs01/index.htm.

[a]All data are for 2001, unless otherwise noted. The last three variables do not appear in the data file but are computed from the previous variables, as described in the text.

四、散点图矩阵

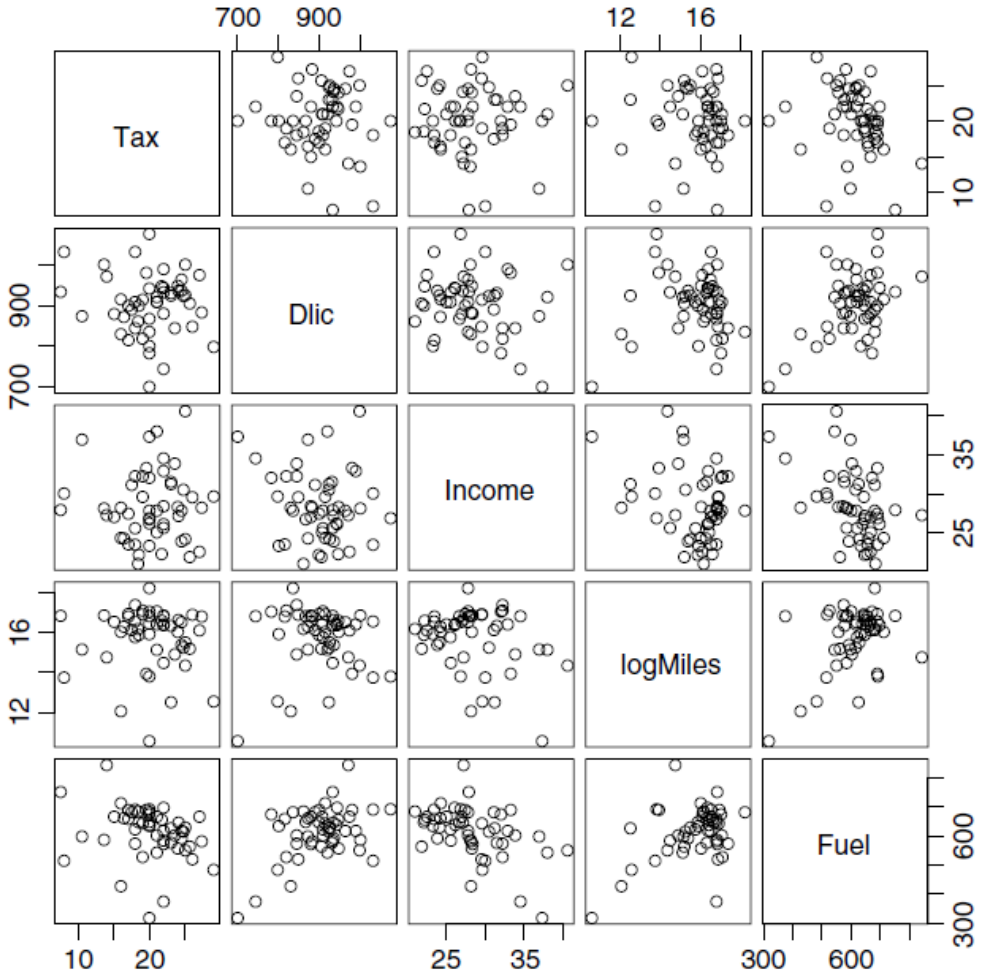


FIG. 1.11 Scatterplot matrix for the fuel data.

五、总结

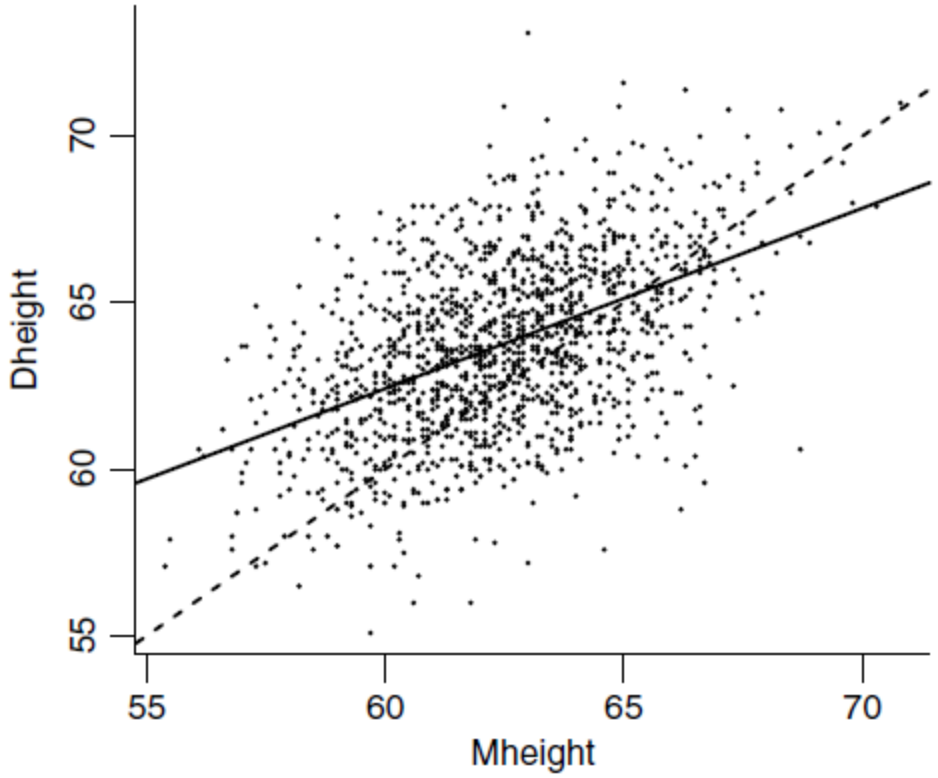


FIG. 1.8 The heights data. The dashed line is for E(*Dheight*|*Mheight*) = *Mheight*, and the solid line is estimated by OLS.

Thank You !