简单回归模型

2 简单线性回归模型

简单线性模型研究两个变量的线性关系，是线性回归模型的最简单情形。相比于相关分析它能提供的变量之间关系的具体描述，即回归方程。此外，简单线性回归模型也为理解一般的多变量回归模型提供了帮助（后者以矩阵和向量表达而不易直观理解）。

2.1 总体模型

假设 x, y 都是一维随机变量，y 是响应变量，x 是自变量。简单线性模型假设

$$y = a + bx + \epsilon, \epsilon \sim (0, \sigma^2),\ \epsilon \perp\!\!\!\perp x, \tag{2.1}$$

其中 a, b, σ^2 为未知参数。该模型表明响应变量部分由 x 线性决定，不能解释的部分以 ϵ 代表，通常称为误差，其中 x 的线性函数部分称为均值函数或回归函数:

$$E(y|x) = E(a + bx + \epsilon|x) = a + bx + E(\epsilon|x) = a + bx,$$

它是 x 的线性函数，也是参数的线性函数。同理知道条件方差为常数

$$\mathrm{var}(y|x) = \mathrm{var}(\epsilon|x) = \mathrm{var}(\epsilon) = \sigma^2.$$

下面考察模型中参数的含义。因为 $\epsilon = y - a - bx$ 与 x 独立，从而不相关，所以 ϵ 是 y 关于 x 的去相关化，$cov(x, \epsilon) = 0$, 因此

$$b = \Sigma_{yx}/\Sigma_{xx} = \rho\sqrt{\Sigma_{yy}/\Sigma_{xx}} = \rho\frac{\sigma_y}{\sigma_x},$$

其中 $\sigma_x = \sqrt{\Sigma_{xx}}, \sigma_y = \sqrt{\Sigma_{yy}}$ 为标准差。注意 $b \propto \rho$ 代表了相关性大小, 且 $b = 0 \Leftrightarrow \rho = 0$。又因 $E(\epsilon) = E(y - a - bx) = 0$，知

$$a = \mu_y - b\mu_x = \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x.$$

所以斜率 b 由 x, y 的二阶矩决定，而截距 a 除了二阶矩也与 x, y 的期望有关。

模型 (2.1) 两边同时求方差

$$var(y) = \mathrm{var}(bx) + \mathrm{var}(\epsilon)$$

即

$$\Sigma_{yy} = b^2\Sigma_{xx} + \sigma^2 = \Sigma_{yx}^2/\Sigma_{xx} + \sigma^2$$

所以 y 中所不能被 x 解释的方差即误差方差

$$\sigma^2 = \Sigma_{yy} - \Sigma_{yx}^2/\Sigma_{xx} = \Sigma_{yy\bullet x}.$$

而 x 所能解释的方差 $\mathrm{var}(bx) = \Sigma_{yx}^2/\Sigma_{xx}$ 在总方差 Σ_{yy} 中的占比，即决定系数

$$R^2 = (\Sigma_{yx}^2/\Sigma_{xx})/\Sigma_{yy} = \rho^2.$$

2.1.1 回归效应

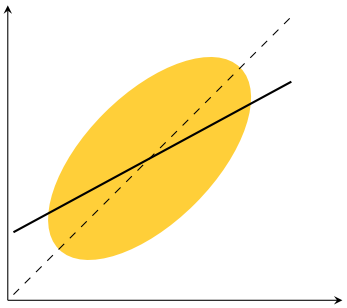
参数 a, b 代入回归函数得

$$E(y|x) = a + bx = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

即

$$E\left(\frac{y-\mu_y}{\sigma_y}\middle|x\right) = \rho \times \frac{x-\mu_x}{\sigma_x}$$

如果 $x = \mu_x + k\sigma_x$，则 $E(y|x) = \mu_y + k\rho\sigma_y$，当 $k > 0$ 时，$E(y|x) \le \mu_y + k\sigma_y$；当 $k < 0$ 时，$E(y|x) \ge \mu_y + k\sigma_y$，无论何种情况，$E(y|x)$ 比 x 更接近其中心。响应变量这种向中心接近的趋势称为回归效应。



2.1.2 矩估计

引入记号

$$s_{aa} = \sum(a_i - \bar{a})^2 = \sum(a_i - \bar{a})a_i,$$
$$s_{ab} = \sum(a_i - \bar{a})(b_i - \bar{b}) = \sum(a_i - \bar{a})b_i.$$

如果基于样本数据得到 x, y 的均值估计 $\bar{\mu}_x = \bar{x}, \bar{\mu}_y = \bar{y}$、方差估计 $\hat{\Sigma}_x x = s_x^2 = s_{xx}/(n-1), \hat{\Sigma}_{yy} = s_y^2 = s_{yy}/(n-1)$ 和协方差估计 $\hat{\Sigma}_{xy} = s_{xy}/(n-1)$，那么我们可以应用矩估计方法得到 a, b 的矩估计

$$\hat{b} = s_{yx}/s_{xx}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x},$$

后面我们将介绍最小二乘法估计参数。矩估计方法和最小二乘法得到的估计基本相同，前者简单，后者具有几何直观。

2.2 最小二乘估计

> 我们在计算方差和进行统计推断的时候，自变量将视作是给定的常数，即在自变量给定的条件下进行统计推断。

2.2.1 回归系数的 LS 估计

假设样本 $(x_i, y_i) \in R^2, i = 1, ..., n$ iid 满足简单回归模型2.1，即

$$y_i = a + bx_i + \epsilon_i,$$

其中 $\epsilon_i, i = 1, ..., n$ iid $\sim (0, \sigma^2)$ 不可观测且 ϵ_i 与 x_i 独立。最小二乘法通过极小化误差平方和求解回归系数 a, b

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

对 a, b 分别求导并令导数为 0，得

$$\begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0 \end{cases}$$

称为正则方程，由此解得 b, a 的最小二乘 (LS) 估计

$$\hat{b} = s_{xy}/s_{xx}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x},$$

注意，LS 估计与与前述矩估计完全相同。

备注 2.1 改写斜率 b 的 LS 估计

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})^2 (y_i - \bar{y})/(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \sum w_i \frac{y_i - \bar{y}}{x_i - \bar{x}},$$

其中 $w_i = (x_i - \bar{x})^2 / \sum_{j=1}^{n} (x_j - \bar{x})^2, \sum w_i = 1$. 因此 $\hat{b}$ 是各个样本点的斜率估计 $\hat{b}_i = (y_i - \bar{y})/(x_i - \bar{x})$ 的加权平均。

命题 2.1 (1) LS 估计的无偏性: $E(\hat{a}) = a, E(\hat{b}) = b.$

(2) LS 估计的方差：给定 $\mathbf{x} = (x_1, ..., x_n)^\top$ 的条件下，

$$\mathrm{var}\left(\begin{array}{c} \hat{a} \\ \hat{b} \end{array} \middle| \mathbf{x} \right) = \left(\begin{array}{cc} \sigma^2/n + \bar{x}^2\sigma^2/s_{xx} & -\bar{x}\sigma^2/s_{xx} \\ -\bar{x}\sigma^2/s_{xx} & \sigma^2/s_{xx} \end{array} \right)$$

证明: （1）因为 ϵ_i 与 x_i 独立，所以 $E(\epsilon_i|x_i) = E(\epsilon_i) = 0$, 进而 $E(y_i|x_i) = E(a + bx_i + \epsilon_i|x_i) = a + bx_i$, 从而

$$E(\hat{b}|\mathbf{x}) = \sum_{i=1}^{n}(x_i - \bar{x})E(y_i|x_i)/s_{xx} = b$$

所以 $E(\bar{b}) = E(E(\hat{b}|\mathbf{x})) = b$。另外 $E(\bar{y}|\mathbf{x}) = a + b\bar{x}$, 所以 $E(\hat{a}|\mathbf{x}) = E(\bar{y} - \hat{b}\bar{x}|\mathbf{x}) = a$，进而 $E(\hat{a}) = a$。

(2) 因为 $\mathrm{var}(y_i|\mathbf{x}) = \mathrm{var}(\epsilon_i|\mathbf{x}) = \mathrm{var}(\epsilon_i) = \sigma^2$, 所以

$$\mathrm{var}(\hat{b}|\mathbf{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2\mathrm{var}(y_i|\mathbf{x})/s_{xx}^2 = \sigma^2/s_{xx}.$$

其它证明类似。 ■

备注 2.2 命题2.1说明，自变量的样本方差 $s_x^2 = s_{xx}/(n-1)$ 越大，斜率的最小二乘估计 $\hat{b}$ 的方差越小，估计越精确。这说明在可以设定或改变自变量的设计比如随机化试验中，我们应该使自变量的取值尽量分散。一般的多重线性模型也有类似结论。

2.2.2 拟合值、残差

给定回归系数 a, b 的 LS 估计，对自变量 x_i, 对应的响应 y_i 的条件期望 $a + bx_i$ 的估计

$$\hat{y}_i = \widehat{a + bx_i} \triangleq \hat{a} + \hat{b}x_i$$

称为 y_i 的拟合值。观测值 y_i 与其拟合值之差

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

称为残差，代表了观察值 y_i 与期望值估计 $\hat{y}_i$ 之间的差异, 因此可以看作是误差 ϵ_i 的预测。由正则方程，我们有

$$\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} e_i x_i = 0,$$

由此知

$$\sum_{i=1}^{n} e_i \hat{y}_i = \sum_{i=1}^{n} (\hat{a} + \hat{b}x_i)e_i = 0.$$

如果记 $\mathbf{e} = (e_1, ..., e_n)^\top, \mathbf{x} = (x_1, ..., x_n)^\top, \mathbf{1} = (1, ..., 1)^\top, \hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_n)^\top$, 那么上述三个方程的几何含义为

$$\mathbf{e} \perp \mathbf{x}, \mathbf{e} \perp \mathbf{1}, \mathbf{e} \perp \hat{\mathbf{y}}.$$

从几何投影的观点来看，上述事实是显然的（参见第 * 章）。

2.2.3 残差平方和与回归平方和

由 $e_i = y_i - \hat{y}_i$, 所有残差之和为 0 蕴含了

$$\bar{y} = \sum_{i=1}^{n} y_i/n = \sum_{i=1}^{n} \hat{y}_i/n = \bar{\hat{y}},$$

即所有拟合值的平均值等于所有响应的平均值。所有拟合值的平方和

$$s_{\hat{y}\hat{y}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

称为回归平方和，度量了拟合值的变化波动大小，而这些拟合值的一条直线上，给回归平方和可理解为自变量变化导致的响应变量的波动程度。响应变量的平方和 $s_{yy} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ 称为总平方和，可以分解如下

$$\begin{aligned} s_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y} + e_i)^2 \\ &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 = s_{\hat{y}\hat{y}} + s_{ee} \end{aligned}$$

容易验证

$$s_{\hat{y}\hat{y}} = \sum (\hat{a} + \hat{b}x_i - (\hat{a} + \hat{b}\bar{x})^2 = \hat{b}^2 s_{xx} = s_{xy}^2/s_{xx}.$$

而所有残差 $e_1, ..., e_n$ 的平方和（称为残差平方和）

$$RSS = s_{ee} = \sum_{i=1}^{n} e_i^2 = s_{yy} - s_{xy}^2/s_{xx},$$

代表了与自变量无关的随机波动程度，即总平方和中自变量所不能解释的部分。定义样本决定系数为总平方和（或总方差）中自变量所能解释的比例：

$$R^2 = \frac{s_{\hat{y}\hat{y}}}{s_{yy}} = \frac{s_{xy}^2}{s_{xx}s_{yy}} = r^2.$$

等于 (x_i, y_i) 样本相关系数的平方。

2.2.4 误差方差估计

因为 $E(\epsilon_i^2) = \sigma^2$, 我们预期 $E(e_i^2) \approx \sigma^2$, 因此定义$\sigma^2$ 的最二乘估计为残差平方和的平均

$$\hat{\sigma}^2 = RSS/(n-2) = \sum_{i=1}^{n} e_i^2/(n-2).$$

为什么不是除以 n 或 $n-1$? 除以 $n-2$ 可以保证无偏性。下面证明 $\hat{\sigma}^2$ 是 σ^2 的无偏估计。

引理 2.2 $RSS = s_{yy} - s_{xy}^2/s_{xx} = s_{\epsilon\epsilon} - s_{x\epsilon}^2/s_{xx}$.

证明: (1) 前面已证 $RSS = s_{yy} - SS_{reg} = s_{yy} - s_{xy}^2/s_{xx}$.

(2) 下面证第二个等式。首先注意到 $\hat{b} = s_{xy}/s_{xx} = \sum(x_i - \bar{x})(a + bx_i + \epsilon_i)/s_{xx} = b + \sum(x_i - \bar{x})\epsilon_i/s_{xx} = b + s_{x\epsilon}/s_{xx}$, 以及 $\hat{y}_i = \bar{y} + \hat{b}(x_i - \bar{x})$ 和 $\bar{y} = \sum(a + bx_i + \epsilon_i)/n = a + b\bar{x} + \bar{\epsilon}$, 所以

$$\begin{aligned} e_i &= y_i - \hat{y}_i = a + bx_i + \epsilon_i - (\bar{y} + \hat{b}(x_i - \bar{x})) \\ &= b(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon} - \hat{b}(x_i - \bar{x}) = \epsilon_i - \bar{\epsilon} - (x_i - \bar{x})s_{x\epsilon}/s_{xx}. \end{aligned}$$

所以

$$\begin{aligned} RSS &= \sum e_i^2 = \sum(\epsilon_i - \bar{\epsilon} - (x_i - \bar{x})s_{x\epsilon}/s_{xx})^2 \\ &= \sum(\epsilon_i - \bar{\epsilon})^2 + \sum(x_i - \bar{x})^2 s_{x\epsilon}^2/s_{xx}^2 - 2\sum(\epsilon_i - \bar{\epsilon})(x_i - \bar{x})s_{x\epsilon}/s_{xx} \\ &= s_{\epsilon\epsilon} - s_{x\epsilon}^2/s_{xx}. \end{aligned}$$

■

命题 2.3 $E(\hat{\sigma}^2) = \sigma^2$.

证明 1: 由引理 2.2, $RSS = s_{yy} - s_{xy}^2/s_{xx} = s_{yy} - \hat{b}^2 s_{xx}$。下面在给定 $\mathbf{x}$ 条件下, 分别求 s_{yy} 和 $\hat{b}^2$ 的条件期望。注意到 $E(y_i^2|\mathbf{x}) = \text{var}(y_i|\mathbf{x}) + (Ey_i|\mathbf{x})^2 = \sigma^2 + (a + bx_i)^2$, $E(\bar{y}^2|\mathbf{x}) = \text{var}(\bar{y}|\mathbf{x}) + (E\bar{y}|\mathbf{x})^2 = \sigma^2/n + (a + b\bar{x})^2$, 所以

$$\begin{aligned} E(s_{yy}|\mathbf{x}) &= E(\sum y_i^2 - n\bar{y}^2|\mathbf{x}) = \sum E(y_i^2|\mathbf{x}) - nE(\bar{y}|\mathbf{x})^2 \\ &= n\sigma^2 + \sum(a + bx_i)^2 - n[\sigma^2/n + (a + b\bar{x})^2] \\ &= (n-1)\sigma^2 + b^2 s_{xx}. \end{aligned}$$

另外, $E(\hat{b}^2|\mathbf{x}) = var(\hat{b}|\mathbf{x}) + (E\hat{b}|\mathbf{x})^2 = \sigma^2/s_{xx} + b^2$, 所以

$$E(RSS|\mathbf{x}) = (n-1)\sigma^2 + b^2 s_{xx} - s_{xx}(\sigma^2/s_{xx} + b^2) = (n-2)\sigma^2,$$

从而 $E(RSS) = (n-2)\sigma^2$, 所以 $E\hat{\sigma}^2 = E(RSS)/(n-2) = \sigma^2$。 ■

证明 2: 利用引理 2.2所给的 RSS 的第二个表达更容易证明。显然 $E(s_{\epsilon\epsilon}) = (n-1)\sigma^2$, 另外, $E(s_{x\epsilon}^2|\mathbf{x})^2 = E[(\sum_{i=1}^n (x_i - \bar{x})\epsilon_i)^2|\mathbf{x}] = \sum_{i=1}^n (x_i - \bar{x})^2 E\epsilon_i^2 = s_{xx}\sigma^2$。 ■

2.2.5 LS 估计的方差估计

命题2.1求出了回归系数估计的方差，其中含有未知参数 σ^2, 代入 (plug-in) 其估计即得到回归系数 LS 的方差的估计，特别地, $\text{var}(\hat{b}|\mathbf{x}) = \sigma^2/s_{xx}$ 的估计为

$$\widehat{\text{var}}(\hat{b}|\mathbf{x}) = \hat{\sigma}^2/s_{xx},$$

其平方根称为标准差（se, standard error)

$$se(\hat{b}|\mathbf{x}) = \sqrt{\widehat{\text{var}}(\hat{b}|\mathbf{x})} = \hat{\sigma}/\sqrt{s_{xx}}.$$

截距项的 LS 估计 $\hat{a}$ 的方差同样地可以估计。

2.3 统计推断

简单线性模型中斜率 (自变量的回归系数)b 代表了相关性, 而截距项通常不是我们感兴趣的，因此我们这里只考虑 b 的统计推断。下面我们将假设误差服从正态分布。

定理 2.4 假设简单回归模型 $y_i = a + bx_i + \epsilon_i$, 其中 $\epsilon_1, ..., \epsilon_n$ iid $\sim N(0, \sigma^2)$ 并与自变量独立。假设 $\hat{a}, \hat{b}, \hat{\sigma}^2$ 分别是 a, b, σ^2 的最小二乘估计。

(1) $\sqrt{s_{xx}}(\hat{b} - b)/\sigma \sim N(0,1)$.
(2) $(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$, 且 $\hat{\sigma}^2$ 与 $(\hat{a}, \hat{b})$ 独立。
(3) $\sqrt{s_{xx}}(\hat{b} - b)/\hat{\sigma} \sim t_{n-2}$.

证明: (1) 因为 $\epsilon_i \sim N(0, \sigma^2)$ 且与 x_i 独立，所以 $\mathbf{x} = (x_1, ..., x_n)^\top$ 给定时，$y_i = a + bx_i + \epsilon_i \sim N(a + bx_i, \sigma^2)$，进而 $\hat{b} = \sum_{i=1}^n (x_i - \bar{x})y_i/s_{xx} \sim N(b, \sigma^2/s_{xx})$, 所以 $\sqrt{s_{xx}}(\hat{b} - b)/\sigma|\mathbf{x} \sim N(0,1)$，该分布与条件无关，所以 $\sqrt{s_{xx}}(\hat{b} - b)/\sigma \sim N(0,1)$。

(2) 由引理2.2,

$$\begin{aligned} RSS &= s_{\epsilon\epsilon} - s_{x\epsilon}^2/s_{xx} \\ &= \sum_{i=1}^n \epsilon_i^2 - n\bar{\epsilon}^2 - (\sum_{i=1}^n (x_i - \bar{x})\epsilon_i)^2/s_{xx} \\ &= \sum_{i=1}^n \epsilon_i^2 - \sum_{i=1}^n (\epsilon_i/\sqrt{n})^2 - \left(\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sqrt{s_{xx}}}\right)\epsilon_i\right)^2 \\ &\triangleq ||\boldsymbol{\epsilon}||^2 - (\mathbf{a}^\top \boldsymbol{\epsilon})^2 - (\mathbf{b}^\top \boldsymbol{\epsilon})^2 \end{aligned}$$

其中 $\mathbf{a} = (1,...,1)^\top/\sqrt{n}$, $\mathbf{b} = (x_1 - \bar{x}...,x_n - \bar{x})^\top/\sqrt{s_{xx}}$, $||\mathbf{a}|| = ||\mathbf{b}|| = 1$, 且 $\mathbf{a} \perp \mathbf{b}$, 由第**??**章引理**??**及 $\mathbf{u} = \boldsymbol{\epsilon}/\sigma \sim N(0, I_n)$,

$$(n-2)\hat{\sigma}^2/\sigma^2 = RSS/\sigma^2 = ||\mathbf{u}||^2 - (\mathbf{a}^\top \mathbf{u})^2 - (\mathbf{b}^\top \mathbf{u})^2 \sim \chi_{n-2}^2,$$

且与 $\mathbf{a}^\top \mathbf{u}, \mathbf{b}^\top \mathbf{u}$ 独立。而 $\hat{b} = \sum(x_i - \bar{x})(a + bx_i + \epsilon_i)/s_{xx} = b + \sum(x_i - \bar{x})\epsilon_i/s_{xx} = b + \mathbf{b}^\top \boldsymbol{\epsilon}/\sqrt{s_{xx}}$ 仅与 $\mathbf{b}^\top \mathbf{u}$ 有关，故与 $\hat{\sigma}^2$ 独立, 同理 $\hat{a}$ 与 $\hat{\sigma}^2$ 独立。

(3). 由（1)、(2）的结果和 t 分布的定义易证。 ■

2.3.1 t 检验

考虑零假设 $H_0 : b = b_0$ (b_0 已知)，Wald 检验考察 LS 估计 $\hat{b}$ 与 b_0 之差的标准化

$$t = \frac{\hat{b} - b_0}{se(\hat{b})} = \sqrt{s_{xx}}(\hat{b} - b_0)/\hat{\sigma},$$

其中 $se(\hat{b}) = \hat{\sigma}/\sqrt{s_{xx}}$。由定理2.4, 在 H_0 成立时，$t \sim t_{n-2}$。因此在 α 水平下，检验规则为

$$若|t| > t_{n-2}(\alpha/2), 则拒绝H_0.$$

绝大多数情况下我们关心的假设是 y 与 x 无关,即 $H_0 : b = 0$, 容易验证此时 t 检验统计量实际上是第**??**章定理**??**基于样本相关系数的独立性检验:

命题 2.5 记 r 为 $(x_i, y_i), i = 1,...,n$ 的样本相关系数，则在 $H_0 : b = 0$ 成立时

$$t = \sqrt{s_{xx}}\hat{b}/\hat{\sigma} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}.$$

2.3.2 置信区间

根据定理2.4结论（3），我们构造 b 的置信度为 $1-\alpha$ 的置信区间如下

$$\left[\hat{b}-\frac{\hat{\sigma}}{\sqrt{s_{xx}}}t_{n-2}(\alpha/2),\hat{b}+\frac{\hat{\sigma}}{\sqrt{s_{xx}}}t_{n-2}(\alpha/2)\right]$$

实际问题中常常需要计算回归函数或均值函数 $m(x)=E(y|x)=a+bx$ 的置信区间。对任何给定的 x_0, $m(x_0)$ 的 LS 估计定义为

$$\hat{m}(x_0)=\hat{a}+\hat{b}x_0,$$

其方差

$$\begin{aligned}&\mathrm{var}(m(x_0))=\mathrm{var}(\bar{y}+\hat{b}(x_0-\bar{x}))=\mathrm{var}(\bar{y})+(x_0-\bar{x})^2\mathrm{var}(\hat{b})\\&=\sigma^2(1/n+(x_0-\bar{x})^2/s_{xx}).\end{aligned}$$

不难证明

$$\frac{\hat{m}(x_0)-m(x_0)}{\sigma\sqrt{1/n+(x_0-\bar{x})^2/s_{xx}}}\sim N(0,1),$$

代入与其独立的 $\hat{\sigma}^2$, 有

$$\frac{\hat{m}(x_0)-m(x_0)}{\hat{\sigma}\sqrt{1/n+(x_0-\bar{x})^2/s_{xx}}}\sim t_{n-2},$$

因此对于给定的任一 x_0, $m(x_0)=a+b(x_0)$ 的置信度为 $1-\alpha$ 的置信区间如下

$$\left[\hat{m}(x_0)\pm t_{n-2}(\alpha/2)\hat{\sigma}\sqrt{c(x_0)}\right]$$

其中 $c(x_0)=1/n+(x_0-\bar{x})^2/s_{xx}$。

当 x_0 变化时，上述置信区间形成回归函数的一个置信带(confidence band), 如图2.1所示，置信带在 $\bar{x}$ 处最窄，x_0 距离 $\bar{x}$ 越远，带宽越大。值得一提的是，置信带在竖轴方向关于 $\hat{m}(x_0)$ 对称，而不是在垂直于回归直线的方向，即置信带关于回归直线在垂直于回归直线的方向上并不是对称的。另外当 n 较大的时候，平方根中忽略 $1/n$ 项，则两条置信边界曲线几乎是直线（但不与回归直线 $\hat{m}(x_0)$ 平行）。

特别需要指出的是，在每个固定的 x_0 处截取置信带，得到的置信区间的置信度为 $1-\alpha$，置信带作为整个回归函数的置信估计其覆盖概率可能远小于 $1-\alpha$。Scheffe 同时置信区间将上述置信带适当修正（加宽）可以得到正确的覆盖概率（参见后续“同时置信区间”）。

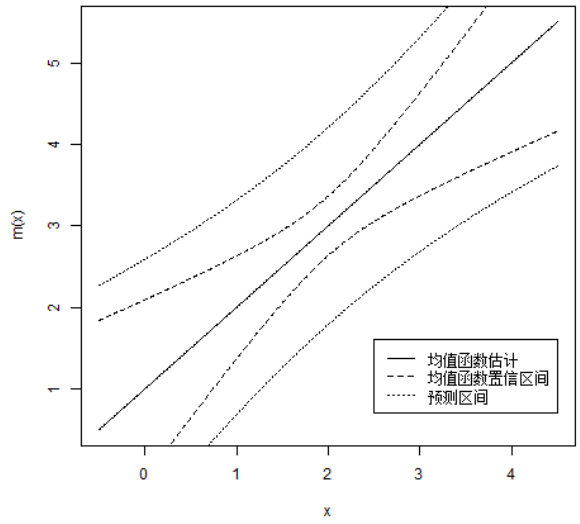


图 2.1: 均值函数的置信带

2.3.3 预测

一个与均值函数估计问题类似但不同的问题是预测。在参数估计问题中估计的对象是未知参数，而预测问题中预测的对象是随机变量。假设 $(x_i y_i), i = 1, ..., n$ iid 满足模型 $y_i = a + bx_i + \epsilon_i$。给定任一自变量 x_0, 我们希望预测其对应的响应 y_0

$$y_0 = a + bx_0 + \epsilon_0,$$

其中 $\epsilon_0, \epsilon_1, ..., \epsilon_n$ iid $\sim (0, \sigma^2)$。因为 $E(y_0|x_0) = m(x_0) = a + bx_0$ 我们仍以 $\hat{y}_0 \stackrel{\Delta}{=} \hat{m}(x_0) = \hat{a} + \hat{b}x_0$ 作为随机变量 y_0 的预测。在前面 $m(x_0)$ 的估计问题中，我们以估计量 $\hat{m}(x_0)$ 与被估计量 $m(x_0)$ 的期望距离平方即方差衡量误差:

$$\text{var}(\hat{m}(x_0))) = E(\hat{m}(x_0) - m(x_0))^2 = \sigma^2 c(x_0),$$

其中 $c(x_0) = 1/n + (x_0 - \bar{x})^2/s_{xx}$。对于预测问题，我们以预测量 $\hat{m}(x_0)$ 与被预测对象 y_0 的预测的期望平方误差作为度量:

$$E(\hat{y}_0 - y_0)^2 = E(\hat{y}_0 - m(x_0))^2 + E(y_0 - m(x_0))^2 = \sigma^2 c(x_0) + \sigma^2$$

比估计误差多了一个被预测对象的方差 $\sigma^2 = var(y_0)$。y_0 的预测区间为

$$[\hat{a} + \hat{b}x_0 \pm t_{n-2}(\alpha)\hat{\sigma}\sqrt{1 + c(x_0)}].$$

图2.1中也演示了预测区间, 可以看到预测区间的长度比均值函数的置信区间要宽很多（当 n 较大的时候, 置信区间长度 $\propto 2\sqrt{c(x_0)} = O(1/\sqrt{n})$, 而预测区间长度 $\propto 2\sqrt{1 + c(x_0)} = O(1)$。

2.3.4 拟合优度：样本决定系数

线性模型拟合数据的好坏以决定系数度量。总体模型 $y = a + bx + \epsilon$ 中我们已经定义决定系数为自变量所能解释的方差在总方差中的比例:

$$R^2 = \frac{\text{var}(bx)}{var(y)} = \frac{\Sigma_{yx}^2/\Sigma_{xx}}{\Sigma_{yy}} = \rho^2,$$

以样本相关系数替代 ρ 即可得到样本决定系数。我们下面直接对拟合方程进行方差分解（或等价地平方和分解）求解样本决定系数。拟合方程

$$y_i = \hat{y}_i + e_i = \hat{a} + \hat{b}x_i + e_i, i = 1, ..., n$$

容易验证 $\sum e_i = 0, \sum y_i = \sum \hat{y}_i$, 拟合方程两边同时减去 $\bar{y} = \bar{\hat{y}}$, 求平方和

$$s_{yy} \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2$$

(两边同时除以 $n - 1$ 即是样本方差的分解公式)，其中回归平方和

$$SS_{reg} = \sum (\hat{y}_i - \bar{y})^2 = \sum (\hat{a} + \hat{b}x_i - \hat{a} - \hat{b}\bar{x})^2 = \hat{b}^2 s_{xx},$$

样本决定系数为回归平方和在总平方和中的占比

$$R^2 = \frac{SS_{reg}}{s_{yy}} = \frac{s_{xy}^2/s_{xx}}{s_{yy}} = r^2.$$

2.4 幂次律

自然界充斥者着大量幂次律 (power law) 的关系, 即

$$y = cx^b, x, y > 0$$

其中 b, c 为参数。有很多科学定律都是幂次律，比如胡可定律, 但更多的是经验分析得到的近似规律, 同样对于理解变量之间的关系有帮助。

幂次律两边同时取对数即得线性方程

$$\log(y) = \log(c) + b \log(x)$$

假设有数据 $(x_i, y_i), i = 1, ..., n$, 它们满足

$$\log(y_i) = \log(c) + b \log(x_i) + \epsilon_i, \epsilon_i \sim (0, \sigma^2),$$

那么基于 $(\log(x_i), \log(y_i)), i = 1, ..., n$, 我们可以求出幂次律中的参数 b, c 的 LS 估计。

Benford 定律

Benford 定律声称自然界很多正实数的首位非 0 数字并不是等可能地分布于 $1,2,...,9$，而是依次下降：首位数字是 1 的概率最大，最不可能的是 9。Benford 定律的首位数字概率分布如下：

$$p(d) = \log_{10}(1 + 1/d), d = 1, ..., 9.$$

这是一个理论分布，如果以幂次率近似，那么最佳逼近为

$$p(d) \approx 0.31/d^{0.86}, d = 1, ..., 9.$$

例 2.1 对于美国 2020 年 3144 个县的人口普查数据，统计各县人口数目的首位数字的样本频率 p 如下 (第二行)：

d	1	2	3	4	5	6	7	8	9
p	0.304	0.185	0.121	0.095	0.075	0.065	0.055	0.051	0.049
$\hat{p}$	0.316	0.172	0.121	0.094	0.077	0.066	0.058	0.051	0.046
$p(d)$	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

拟合线性模型 $\log(p) = a + b\log(d) + \epsilon_d, d = 1, 2, ..., 9$, 得 LS 估计 $\hat{a} = -1.153, \hat{b} = -0.874$, 回归直线估计为 $\log(p) = -1.153 - 0.874\log(d)$, 等价地，得到幂次律

$$p = 0.316/d^{0.874},$$

非常接近于 Benford 定律的幂次逼近 $p(d) \approx 0.31/d^{0.86}$。表格第三行为拟合值，第四行为 Benford 定律的理论值。可以看到经验幂次律很好地拟合了样本数据 p，并与理论分布 $p(d)$ 非常接近。

Zipf 定律

大众语言或文本中，单词的使用频率并不均匀，只有少数单词会频繁出现，而大多数单词只是偶尔出现。比如布朗大学现代美国英语语料库 (Brown coprus) 中排名前三的单词为 the, and, of, 出现频率分别为 7%, 3.5%,2.8%，其中排名第一的单词 the 的使用频率是第二名的 2 倍，第二名是第三名的 1.25 倍，其后的单词频率依次减小，前后差别也越来越小。哈佛大学语言学家 G.K.Zipf 分析了大众语言中最常用的若干单词的使用频率，发现了 Zipf 定律，即单词使用频率与排名存在反比关系（1949 年）

$$p_k \propto 1/k, k = 1, 2, ...$$

其中 p_k 为第 k 个最常用的单词的使用频率。

Zipf 反比定律是关于大众文本的一个规律，但是对于特殊的一类文本，或某个作家的作品进行频率分析可能得到不同的幂次

关系

$$p_k \propto 1/k^{\alpha}, k = 1, 2, ...$$

幂次律在文本分析中有重要的应用，α 代表作者的用词习惯，不同的作者或作品可能有不同的指数 α, 例如，词汇量较小的作者会倾向于大量使用少数单词，从而具有较大的指数 α。

例 2.2 美国联邦文献主要作者麦迪逊的前 10 个高频词频率如下（1/1000）

单词	the	of	to	and	in	a	be	that	it	is
频率	93.7	57.8	35.3	27.6	23.1	20.2	16.5	14.4	13.3	12.8

我们对麦迪逊的高频词汇的使用频率假设模型 $\log(p_k) = a + b\log(k) + \epsilon_k, k = 1, ..., 10$, 其中 k 为排名，p_k 为词频，解得 $\hat{a} = -2.32, \hat{b} = 0.90$, 幂次律为 $p_k = 0.1/k^{0.9}$，拟合情况如图2.2所示。

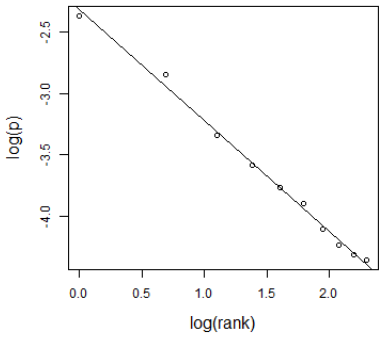


图 2.2: 麦迪逊高频词频率与排名（对数尺度）

另外一方面，对比不同作品的指数可以用于判断是否它们属于同一作者，比如美国著名的联邦论文集（The Federalist Papers）是 18 世纪 80 年代三位美国人哈密尔顿 (A. Hamilton), 麦迪逊 (J. Madison) 和杰伊 (J. Jay) 以 Publilus 为笔名在纽约报纸上发表的一系列政论文章，对美国宪法的制定和联邦政体的建立影响重大，是美国历史上最重要的文献之一。该系列文章共 85 篇，大多数文章的作者已经辨明，但少数文章（12 篇）的作者归属存在争议。学者们对存在争议的 12 篇文章进行了各种词频统计和文风分析，多数结果表明这 12 篇文章的作者是麦迪逊。

Kleiber 定律

生物的重量和能量消耗等与体积有关，而散热速度、代谢速度、承重能力等与面积有关。如果生物是等速生长的，那么其表

面积 S 和体积 V 应该服从所谓平方-立方定律，即

$$S \propto V^{2/3}$$

所以面积的成长速度低于体积的成长速度，从而会造成面积不足的情况（比如不足以散发热量），因此很多生物的生长会出现异速生长现象, 即不同的器官或组织具有不同的生长速度，比如大象的腿会别其它部位生长的更快，以形成更大的横截面面积保证有足够的体重承受能力。瑞士动植物学家 M. Kleiber 于 20 世纪 30 年代发现大多数动物的代谢速率（与表面积成正比）与体重（与体积成正比）服从幂次为 3/4 而不是 2/3 等速生长幂次律

$$\text{代谢速率} \propto \text{体重}^{3/4}$$

其中代谢速率与表面积成正比，单位是卡路里/秒。这表明大多数动物是异速生长的。

例 2.3 R 程序包中的数据集 brains 收集了 62 种哺乳动物体重（BodyWt，单位：千克）和脑重（BrainWt，单位：克）数据，其中有很大的动物，也有非常小的动物。对于这种不在同一个尺度上的数据，通常需要取对数将数据转化到同一尺度上。取对数之后拟合线性模型

$$\log(BrainWt) = a + b\log(BodyWt) + \epsilon$$

得 LS 估计 $\hat{a} = 2.14, \hat{b} = 0.75$, 图2.3表明模型拟合效果良好，从而得到经验幂次律

$$BrwinWt = 8.46 \times BodyWt^{0.75}.$$

与 Kleiber 定律一样，这也是一个 $3/4$ 幂次律。

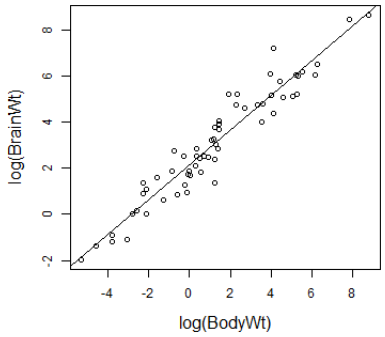


图 2.3: 哺乳动物脑重与体重的关系（对数尺度）

体重指数 BMI

指数 (index) 需要有普适性, 为了衡量一个人的体重是否正常，单独以体重作为指标是不恰当的，因为体重与身高、年龄、性别等因素有关。为了制定一个普遍使用的体重指数，我们需要校正这些干扰因素。国际通用的 BMI(Body Weight Index) 指数是比利时数学家 Quetelet 提出的一个适用于成年人的体重指标

$$BMI = \text{Weight}/\text{Height}^2$$

它在某种意义上消除了身高对体重的影响。但 Quetelet 制定该指标的依据无从查找，该指标不区分性别也不尽合理。下面我们以简单线性回归模型出发，建立一个消除身高因素的体重指数。

例 2.4 假设 (W, H) 为随机抽取的成年人的体重和身高，假设

$$\log(W) = a + b\log(H) + \epsilon, \epsilon \sim N(0, \sigma^2),$$

那么误差 (去相关化)

$$\epsilon = \log(W) - a - b\log(H)$$

代表了消除身高影响后的体重指标，如果再进一步标准化

$$z = \epsilon/\sigma = (\log(W - a - b\log(H))/\sigma \sim N(0, 1)$$

z 就是一个标准化（消除了身高影响）的体重指标，我们可以用标准正态分布的分位数 c 确定某个人是否体重超标。

$$z > c \Leftrightarrow W/H^b > e^{a+c\sigma}.$$

经验数据表明 $b \approx 2$, 从而 BMI 是一个合理的体重指数（数据: http://staff.ustc.edu.cn/ ynyang/2023/data/height-weight.txt）。

不取对数是否可以直接建立如下线性模型?

$$W = a + bH + \epsilon, \epsilon \sim N(0, \sigma^2),$$

这依赖于残差分析是否表明误差可以假设为正态分布。如果该模型是合理的，那么可以定义

$$z = \epsilon/\sigma = (W - a - bH)/\sigma \sim N(0, 1)$$

为体重指数。但显然该指数不够简洁。

2.5 附录