概率论与数理统计（13）

清华大学

2020 年春季学期

小回顾

- 设 $x_1, \ldots, x_{10}$ 是来自 0-1 总体 $b(1, p)$ 的样本，考虑以下检验问题：

$$H_0 : p = 0.2 \quad \text{vs} \quad H_1 : p = 0.4,$$

取拒绝域为 $W = \{\bar{x} \geqslant 0.5\}$，求该检验犯两类错误的概率。

小回顾

- 设 $x_1, \ldots, x_{10}$ 是来自 0-1 总体 $b(1, p)$ 的样本，考虑以下检验问题：

$$H_0 : p = 0.2 \quad \text{vs} \quad H_1 : p = 0.4,$$

取拒绝域为 $W = \{\bar{x} \geqslant 0.5\}$，求该检验犯两类错误的概率。

- 发生第一类错误的概率：

$$\alpha = P(\bar{x} \geqslant 0.5 | H_0) = \sum_{k=5}^{10} \binom{10}{k} \frac{1}{5^k} (\frac{4}{5})^{10-k} = 0.0328.$$

小回顾

- 设 $x_1, \ldots, x_{10}$ 是来自 0-1 总体 $b(1, p)$ 的样本，考虑以下检验问题：

$$H_0 : p = 0.2 \quad \text{vs} \quad H_1 : p = 0.4,$$

取拒绝域为 $W = \{\bar{x} \geqslant 0.5\}$，求该检验犯两类错误的概率。

- 发生第一类错误的概率：

$$\alpha = P(\bar{x} \geqslant 0.5 | H_0) = \sum_{k=5}^{10} \binom{10}{k} \frac{1}{5^k} (\frac{4}{5})^{10-k} = 0.0328.$$

- 发生第二类错误的概率为：

$$\beta = P(\bar{x} < 0.5 | H_1) = \sum_{k=0}^{4} \binom{10}{k} (2/5)^k (3/5)^{10-k} = 0.6331.$$

一元线性回归

- 变量和变量之间的关系：确定性关系：如函数关系 $y = f(x)$.

一元线性回归

- 变量和变量之间的关系：确定性关系：如函数关系 $y = f(x)$.相关性关系：身高和体重的关系；身高和智商的关系；体重和受欢迎程度的关系；臀部大小和智商的关系.....

一元线性回归

- 变量和变量之间的关系：确定性关系：如函数关系 $y = f(x)$.相关性关系：身高和体重的关系；身高和智商的关系；体重和受欢迎程度的关系；臀部大小和智商的关系.....
- 相关性关系不能用完全确定的函数形式来决定。但在平均意义下有一定的定量关系。确定这种定量关系的具体形式就是回归分析的主要任务。

一元线性回归

- 变量和变量之间的关系：确定性关系：如函数关系 $y = f(x)$.相关性关系：身高和体重的关系；身高和智商的关系；体重和受欢迎程度的关系；臀部大小和智商的关系.....
- 相关性关系不能用完全确定的函数形式来决定。但在平均意义下有一定的定量关系。确定这种定量关系的具体形式就是回归分析的主要任务。
- 线性回归分析即是寻找线性关系：$y = ax + b$.

一元线性回归模型

- 假设 y 与 x 之间有相关性关系，x 为自变量，y 为因变量，我们要寻找的关系应该为

$$f(x) = E(y|x) = \int yp(y|x)\,dx.$$

一元线性回归模型

- 假设 y 与 x 之间有相关性关系，x 为自变量，y 为因变量，我们要寻找的关系应该为

$$f(x) = E(y|x) = \int yp(y|x)\,dx.$$

- 太过一般了。这里考虑的模型为

$$y = f(x) + \epsilon,$$

其中 ϵ 是随机误差，一般假设为 $\epsilon \sim N(0, \sigma^2)$，显然有 $f(x) = E(y|x)$.

一元线性回归模型

- 假设 y 与 x 之间有相关性关系，x 为自变量，y 为因变量，我们要寻找的关系应该为

$$f(x) = E(y|x) = \int yp(y|x)\,dx.$$

- 太过一般了。这里考虑的模型为

$$y = f(x) + \epsilon,$$

其中 ϵ 是随机误差，一般假设为 $\epsilon \sim N(0, \sigma^2)$，显然有 $f(x) = E(y|x)$.

- 进一步假设 $y = \beta_0 + \beta_1 x + \epsilon$,

一元线性回归模型

- 假设 y 与 x 之间有相关性关系，x 为自变量，y 为因变量，我们要寻找的关系应该为

$$f(x) = E(y|x) = \int yp(y|x)\,dx.$$

- 太过一般了。这里考虑的模型为

$$y = f(x) + \epsilon,$$

其中 ϵ 是随机误差，一般假设为 $\epsilon \sim N(0, \sigma^2)$，显然有 $f(x) = E(y|x)$.

- 进一步假设 $y = \beta_0 + \beta_1 x + \epsilon$，假设 x 不是随机量，y 是随机量。即有

$$y \sim N(\beta_0 + \beta_1 x, \sigma^2).$$

一元线性回归模型

- $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$, β_0, β_1 为未知参数。假设收集了数据 $(x_1, y_1), \ldots, (x_n, y_n)$, 同时假设 $y_1, \ldots, y_n$ 相互独立。则具体的一元线性回归模型为

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, 2, \ldots, n, \\ \epsilon_i \text{为独立同分布随机变量}, \epsilon_i \sim N(0, \sigma^2). \end{cases}$$

由数据 (x_i, y_i), $i = 1, \ldots, n$ 可以获得 β_0, β_1 的估计量 $\hat{\beta}_0$, $\hat{\beta}_1$, 则称

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

为 y 关于 x 的经验回归函数，也称回归方程。给定 $x = x_0$, $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 为回归值，或者预测值、拟合值等。

回归系数的最小二乘估计

- 最小二乘方法估计参数 β_0, β_1,

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

寻找 $\hat{\beta}_0$, $\hat{\beta}_1$ 使得 $Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$。

回归系数的最小二乘估计

- 最小二乘方法估计参数 β_0, β_1,

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

寻找 $\hat{\beta}_0, \hat{\beta}_1$ 使得 $Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$。

$$\begin{cases} \partial_{\beta_0} Q = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \partial_{\beta_1} Q = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0. \end{cases}$$

回归系数的最小二乘估计

- 最小二乘方法估计参数 β_0, β_1,

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

寻找 $\hat{\beta}_0$, $\hat{\beta}_1$ 使得 $Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$。

$$\begin{cases} \partial_{\beta_0} Q = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \partial_{\beta_1} Q = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0. \end{cases}$$

$$\begin{cases} n\hat{\beta}_0 + n\bar{x}\hat{\beta}_1 = n\bar{y}, \\ n\bar{x}\hat{\beta}_0 + \sum x_i^2 \hat{\beta}_1 = \sum x_i y_i. \end{cases} \Rightarrow \begin{cases} \hat{\beta}_1 = \frac{l_{xy}}{l_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \end{cases}$$

$$l_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}, \quad l_{xx} = \sum (x_i - \bar{x})^2.$$

参数的最大似然估计

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$，则 $y_i - \beta_0 - \beta_1 x_i$ 可以看作是来自正态总体 $N(0, \sigma^2)$ 的样本。

参数的最大似然估计

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, 则 $y_i - \beta_0 - \beta_1 x_i$ 可以看作是来自正态总体 $N(0, \sigma^2)$ 的样本。似然函数为

$$L = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}.$$

指数似然函数为 $l = \frac{-n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$.

参数的最大似然估计

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, 则 $y_i - \beta_0 - \beta_1 x_i$ 可以看作是来自正态总体 $N(0, \sigma^2)$ 的样本。似然函数为

$$L = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}.$$

指数似然函数为 $l = \frac{-n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$.

$$\begin{cases} \partial_{\beta_0} l = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \partial_{\beta_1} l = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0. \end{cases}$$

参数的最大似然估计

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$，则 $y_i - \beta_0 - \beta_1 x_i$ 可以看作是来自正态总体 $N(0, \sigma^2)$ 的样本。似然函数为

$$L = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}.$$

指数似然函数为 $l = \frac{-n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$.

$$\begin{cases} \partial_{\beta_0} l = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \partial_{\beta_1} l = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0. \end{cases}$$

得到的方程与最小二乘估计中的一样，所以此时最小二乘估计与最大似然估计相同。

参数的最大似然估计

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$，则 $y_i - \beta_0 - \beta_1 x_i$ 可以看作是来自正态总体 $N(0, \sigma^2)$ 的样本。似然函数为

$$L = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}.$$

指数似然函数为 $l = \frac{-n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$.

$$\begin{cases} \partial_{\beta_0} l = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \partial_{\beta_1} l = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0. \end{cases}$$

得到的方程与最小二乘估计中的一样，所以此时最小二乘估计与最大似然估计相同。最小二乘估计不要求偏差服从正态分布。

最小二乘估计的性质

- 定理：1) $\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}})\sigma^2)$, $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{l_{xx}})$;
  2) $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}}\sigma^2$;
  3) 对于给定的 x_0,
  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N(\beta_0 + \beta_1 x_0, (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}})\sigma^2)$.

最小二乘估计的性质

- 定理：1) $\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}})\sigma^2)$, $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{l_{xx}})$;
  2) $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}}\sigma^2$;
  3) 对于给定的 x_0,
  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N(\beta_0 + \beta_1 x_0, (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}})\sigma^2)$.
- $\hat{\beta}_0$, $\hat{\beta}_1$ 分别是 β_0, β_1 的无偏估计。
- $\hat{y}_0$ 是 $\beta_0 + \beta_1 x_0$ 的无偏估计。
- 除 $\bar{x} = 0$ 外，$\hat{\beta}_0$, $\hat{\beta}_1$ 是相关的。
- 若要提高 $\hat{\beta}_0$，$\hat{\beta}_1$ 的估计精度，要增大 n, 和 l_{xx}（即要求 $x_1, \ldots, x_n$ 尽量分散）。

回归方程的显著性检验

- 任意给出 n 对数据 (x_i, y_i) 都可以求出回归方程 $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

回归方程的显著性检验

- 任意给出 n 对数据 (x_i, y_i) 都可以求出回归方程 $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.但这个方程不一定有意义。

回归方程的显著性检验

- 任意给出 n 对数据 (x_i, y_i) 都可以求出回归方程 $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.但这个方程不一定有意义。
- 对于真正的回归方程 $E(y) = \beta_0 + \beta_1 x$, 若 $\beta_1 = 0$, 则无论 $E(y)$ 不随 x 的变化而做线性变化，那么一元线性回归方程意义不大，或称回归方程不显著；若 $\beta_1 \neq 0$, 则称回归方程显著。

回归方程的显著性检验

- 任意给出 n 对数据 (x_i, y_i) 都可以求出回归方程 $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.但这个方程不一定有意义。
- 对于真正的回归方程 $E(y) = \beta_0 + \beta_1 x$, 若 $\beta_1 = 0$, 则无论 $E(y)$ 不随 x 的变化而做线性变化，那么一元线性回归方程意义不大，或称回归方程不显著；若 $\beta_1 \neq 0$, 则称回归方程显著。
- 对于回归方程显著与否，要进性假设检验，其统计假设为:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0.$$

即拒绝零假设 H_0 表示回归方程为显著的。

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0, \quad \text{vs} \quad H_1: \beta \neq 0.$

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0, \quad \text{vs} \quad H_1: \beta \neq 0.$
- 记 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 为 y 在 x_i 处的回归值，$y_i - \hat{y}_i$ 为 x_i 处的残差。

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0, \quad \text{vs} \quad H_1: \beta \neq 0.$
- 记 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 为 y 在 x_i 处的回归值，$y_i - \hat{y}_i$ 为 x_i 处的残差。
- 总偏差平方和：$S_T = \sum (y_i - \bar{y})^2 = l_{yy}.$
- 回归平方和：$S_R = \sum (\hat{y}_i - \bar{y})^2 = \frac{l_{xy}^2}{l_{xx}};$
- 残差平方和：$S_e = \sum (y_i - \hat{y}_i)^2;$

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0, \quad \text{vs} \quad H_1: \beta \neq 0.$
- 记 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 为 y 在 x_i 处的回归值，$y_i - \hat{y}_i$ 为 x_i 处的残差。
- 总偏差平方和：$S_T = \sum (y_i - \bar{y})^2 = l_{yy}.$
- 回归平方和：$S_R = \sum (\hat{y}_i - \bar{y})^2 = \frac{l_{xy}^2}{l_{xx}};$
- 残差平方和：$S_e = \sum (y_i - \hat{y}_i)^2;$
- 一元线性回归的平法和分解：$S_T = S_R + S_e,$

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0, \quad \text{vs} \quad H_1: \beta \neq 0.$
- 记 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 为 y 在 x_i 处的回归值，$y_i - \hat{y}_i$ 为 x_i 处的残差。
- 总偏差平方和：$S_T = \sum (y_i - \bar{y})^2 = l_{yy}.$
- 回归平方和：$S_R = \sum (\hat{y}_i - \bar{y})^2 = \frac{l_{xy}^2}{l_{xx}};$
- 残差平方和：$S_e = \sum (y_i - \hat{y}_i)^2;$
- 一元线性回归的平法和分解：$S_T = S_R + S_e, S_e = l_{yy} - \frac{l_{xy}^2}{l_{xx}}.$

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0, \quad \text{vs} \quad H_1: \beta \neq 0.$
- 记 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 为 y 在 x_i 处的回归值，$y_i - \hat{y}_i$ 为 x_i 处的残差。
- 总偏差平方和：$S_T = \sum (y_i - \bar{y})^2 = l_{yy}.$
- 回归平方和：$S_R = \sum (\hat{y}_i - \bar{y})^2 = \frac{l_{xy}^2}{l_{xx}};$
- 残差平方和：$S_e = \sum (y_i - \hat{y}_i)^2;$
- 一元线性回归的平法和分解：$S_T = S_R + S_e, S_e = l_{yy} - \frac{l_{xy}^2}{l_{xx}}.$
- $E(S_R) = \sigma^2 + \beta_1^2 l_{xx}$, $E(S_e) = (n-2)\sigma^2$,

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0, \quad \text{vs} \quad H_1: \beta \neq 0.$
- 记 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 为 y 在 x_i 处的回归值，$y_i - \hat{y}_i$ 为 x_i 处的残差。
- 总偏差平方和：$S_T = \sum (y_i - \bar{y})^2 = l_{yy}.$
- 回归平方和：$S_R = \sum (\hat{y}_i - \bar{y})^2 = \frac{l_{xy}^2}{l_{xx}};$
- 残差平方和：$S_e = \sum (y_i - \hat{y}_i)^2;$
- 一元线性回归的平法和分解：$S_T = S_R + S_e, S_e = l_{yy} - \frac{l_{xy}^2}{l_{xx}}.$
- $E(S_R) = \sigma^2 + \beta_1^2 l_{xx}$, $E(S_e) = (n-2)\sigma^2$，$\frac{S_e}{n-2}$ 为 σ^2 的无偏估计。
- $S_e/\sigma^2 \sim \chi^2(n-2).$
- 若 H_0 成立，则有 $S_R/\sigma^2 \sim \chi^2(1).$
- S_R 与 S_e, $\bar{y}$ 独立。

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0, \quad \text{vs} \quad H_1: \beta \neq 0.$
- 记 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 为 y 在 x_i 处的回归值，$y_i - \hat{y}_i$ 为 x_i 处的残差。
- 总偏差平方和：$S_T = \sum (y_i - \bar{y})^2 = l_{yy}.$
- 回归平方和：$S_R = \sum (\hat{y}_i - \bar{y})^2 = \frac{l_{xy}^2}{l_{xx}};$
- 残差平方和：$S_e = \sum (y_i - \hat{y}_i)^2;$
- 一元线性回归的平法和分解：$S_T = S_R + S_e, S_e = l_{yy} - \frac{l_{xy}^2}{l_{xx}}.$
- $E(S_R) = \sigma^2 + \beta_1^2 l_{xx}$, $E(S_e) = (n-2)\sigma^2$，$\frac{S_e}{n-2}$ 为 σ^2 的无偏估计。
- $S_e/\sigma^2 \sim \chi^2(n-2).$
- 若 H_0 成立，则有 $S_R/\sigma^2 \sim \chi^2(1).$
- S_R 与 S_e, $\bar{y}$ 独立。
- 若 H_0 成立，则 $F = \frac{S_R}{S_e/(n-2)} \sim F(1, n-2)$。

线性回归方程的显著性检验

- F 检验：$H_0: \beta_1 = 0 \quad \text{vs} \quad H_1: \beta_1 \neq 0.$
- 统计量：

$$F = \frac{S_R}{S_e/(n-2)};$$

- 限定显著水平 $\alpha \in (0,1)$，并确定拒绝区域：

$$W = \{F \geqslant F_{1-\alpha}(1, n-2)\}.$$

即若样本的数据计算的统计量的值 $F_0 \geqslant F_{1-\alpha}$，则线性回归方程显著。

线性回归方程的显著性检验

- F 检验：回归方程的方差分析表：

来源	平方和	自由度	均方	F 比	p 值
回归	$S_R = \frac{l_{xy}^2}{l_{xx}}$	1	$MS_R = S_R$	$F_0 = \frac{MS_R}{MS_e}$	$P(F \geqslant F_0)$
残差	S_e	$n-2$	$MS_e = \frac{S_e}{n-2}$		
总计	$S_T = l_{yy}$	$n-1$			

线性回归方程的显著性检验

- t 检验: $H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0.$
- 选择统计量:

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}} \sim t(n-2)\ (H_0),$$

这里 $\hat{\sigma} = \sqrt{S_e/(n-2)}$, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/l_{xx})$.

- 给定显著水平 $\alpha \in (0,1)$, 给出拒绝域,

$$W = \{|t| \geqslant t_{1-\alpha/2}\}.$$

也就是若由样本数据计算出来的统计量 $|t_0| \geqslant t_{1-\alpha/2}$, 则认为线性回归方程显著。

线性回归方程的显著性检验

- 相关系数检验：$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0.$
- 选择统计量，为样本相关系数：

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}}.$$

$$r^2 = \frac{l_{xy}^2}{l_{xx} l_{yy}} = \frac{S_R}{S_T} = \frac{S_R}{S_R + S_e} = \frac{F}{F + (n-2)}.$$

其中 $F = \frac{S_R}{S_e/(n-2)} \sim F(1, n-2) \quad (H_0).$，且 $|r|$ 关于 F 严格单调增。

- 给定显著水平 $\alpha \in (0, 1)$，给出拒绝域，

$$W = \{|r| \geqslant r_{1-\alpha} = \sqrt{\frac{F_{1-\alpha}(1, n-2)}{F_{1-\alpha}(1, n-2) + n - 2}}\}.$$

也就是若由样本数据计算出来的统计量 $|r_0| \geqslant r_{1-\alpha/2}$，则认为线性回归方程显著。

估计与预测

- 当回归方程经验证是显著后，可用来做估计和预测。

估计与预测

- 当回归方程经验证是显著后，可用来做估计和预测。
  - 当 $x = x_0$ 时，寻找均值 $E(y_0) = \beta_0 + \beta_1 x_0$ 的点估计和区间估计。此为估计问题。

估计与预测

- 当回归方程经验证是显著后，可用来做估计和预测。
  - 当 $x = x_0$ 时，寻找均值 $E(y_0) = \beta_0 + \beta_1 x_0$ 的点估计和区间估计。此为估计问题。
  - 当 x_0 时，y_0 的观测值在什么范围内？若 $P(|y_0 - \hat{y}_0| \leqslant \delta) = 1 - \alpha$, 则称 $[\hat{y}_0 - \delta, \hat{y}_0 + \delta]$ 为 y 的概率为 $1 - \alpha$ 的预测区间。这是预测问题。

估计与预测

- 当回归方程经验证是显著后，可用来做估计和预测。
  - 当 $x = x_0$ 时，寻找均值 $E(y_0) = \beta_0 + \beta_1 x_0$ 的点估计和区间估计。此为估计问题。
  - 当 x_0 时，y_0 的观测值在什么范围内？若 $P(|y_0 - \hat{y}_0| \leqslant \delta) = 1 - \alpha$, 则称 $[\hat{y}_0 - \delta, \hat{y}_0 + \delta]$ 为 y 的概率为 $1 - \alpha$ 的预测区间。这是预测问题。
- $E(y_0)$ 的估计。
  - $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 是 $E(y_0)$ 的无偏估计。
  - $E(y_0)$ 的 $1 - \alpha$ 置信区间：

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N(\beta_0 + \beta_1 x_0, (\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{l_{xx}})\sigma^2).$$

$$\frac{(\hat{y}_0 - Ey_0)/\sqrt{(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}})\sigma^2}}{\sqrt{\frac{S_e}{\sigma^2}/(n-2)}} = \frac{\hat{y}_0 - Ey_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim t(n-2).$$

所以 Ey_0 的置信水平为 $1 - \alpha$ 的置信区间为 $[\hat{y}_0 - \delta, \hat{y}_0 + \delta]$,

$$\delta = t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}, \quad \hat{\sigma} = \sqrt{\frac{S_e}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}.$$

- y_0 的预测区间：$y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$，$y_0$ 与 $\hat{y}_0$ 相互独立，所以

$$y_0 - \hat{y}_0 \sim N(0, [1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}]\sigma^2).$$

$$\frac{y_0 - \hat{y}_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim t(n-2).$$

所以 y_0 的概率为 $1-\alpha$ 的预测区间为

$$[\hat{y}_0 - \delta', \hat{y}_0 + \delta'], \quad \delta' = \hat{\sigma} t_{1-\alpha/2}(n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}.$$

- y_0 的预测区间：$y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$，$y_0$ 与 $\hat{y}_0$ 相互独立，所以

$$y_0 - \hat{y}_0 \sim N(0, [1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}]\sigma^2).$$

$$\frac{y_0 - \hat{y}_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim t(n-2).$$

所以 y_0 的概率为 $1-\alpha$ 的预测区间为

$$[\hat{y}_0 - \delta', \hat{y}_0 + \delta'], \quad \delta' = \hat{\sigma} t_{1-\alpha/2}(n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}.$$

- 预测区间的长度不会为零；当 l_{xx} 大时，预测区间较短，即应该让 $x_1, \ldots, x_n$ 分散。$x_0 = \bar{x}$ 是区间最短。
- 当 n 比较大，且 x_0 离 $\bar{x}$ 比较近时，预测区间近似为 $[\hat{y}_0 - \hat{\sigma} u_{1-\alpha/2}, \hat{y}_0 + \hat{\sigma} u_{1-\alpha/2}]$.

预测区间

- $X_1, \ldots, X_n, X_{n+1}$ 为来自正态总体的样本。考虑 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. 收集到前面 n 个数据。考虑下一个数据的可能范围。

预测区间

- $X_1, \ldots, X_n, X_{n+1}$ 为来自正态总体的样本。考虑 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. 收集到前面 n 个数据。考虑下一个数据的可能范围。
- $X_{n+1} - \bar{X}_n \sim N(0, \sigma^2 + \frac{1}{n}\sigma^2)$, $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1)$.

预测区间

- $X_1, \ldots, X_n, X_{n+1}$ 为来自正态总体的样本。考虑 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. 收集到前面 n 个数据。考虑下一个数据的可能范围。
- $X_{n+1} - \bar{X}_n \sim N(0, \sigma^2 + \frac{1}{n}\sigma^2)$, $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1)$.

$$\frac{X_{n+1} - \bar{X}_n}{s_n\sqrt{1+\frac{1}{n}}} = \frac{\frac{X_{n+1}-\bar{X}_n}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{\frac{(n-1)s_n^2}{(n-1)\sigma^2}}} \sim t(n-1).$$

预测区间

- $X_1, \ldots, X_n, X_{n+1}$ 为来自正态总体的样本。考虑 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. 收集到前面 n 个数据。考虑下一个数据的可能范围。
- $X_{n+1} - \bar{X}_n \sim N(0, \sigma^2 + \frac{1}{n}\sigma^2)$, $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1)$.

$$\frac{X_{n+1} - \bar{X}_n}{s_n\sqrt{1+\frac{1}{n}}} = \frac{\frac{X_{n+1}-\bar{X}_n}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{\frac{(n-1)s_n^2}{(n-1)\sigma^2}}} \sim t(n-1).$$

- $P(\bar{X}_n - t_{1-\frac{\alpha}{2}} s_n\sqrt{1+\frac{1}{n}} \leqslant X_{n+1} \leqslant \bar{X}_n + t_{1-\frac{\alpha}{2}} s_n\sqrt{1+\frac{1}{n}}) = 1-\alpha.$

预测区间

- $X_1, \ldots, X_n, X_{n+1}$ 为来自正态总体的样本。考虑 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. 收集到前面 n 个数据。考虑下一个数据的可能范围。
- $X_{n+1} - \bar{X}_n \sim N(0, \sigma^2 + \frac{1}{n}\sigma^2)$, $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1)$.

$$\frac{X_{n+1} - \bar{X}_n}{s_n\sqrt{1+\frac{1}{n}}} = \frac{\frac{X_{n+1}-\bar{X}_n}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{\frac{(n-1)s_n^2}{(n-1)\sigma^2}}} \sim t(n-1).$$

- $P(\bar{X}_n - t_{1-\frac{\alpha}{2}} s_n\sqrt{1+\frac{1}{n}} \leqslant X_{n+1} \leqslant \bar{X}_n + t_{1-\frac{\alpha}{2}} s_n\sqrt{1+\frac{1}{n}}) = 1-\alpha.$
- 何用？

例子

- 现收集了 16 组合金钢的碳含量 x 与强度 y 的数据，整理得

$$\bar{x} = 0.125,\ \bar{y} = 45.7886,\ l_{xx} = 0.3024,\ l_{xy} = 25.5218,\ l_{yy} = 2432.45$$

例子

- 现收集了 16 组合金钢的碳含量 x 与强度 y 的数据，整理得

$$\bar{x} = 0.125,\ \bar{y} = 45.7886,\ l_{xx} = 0.3024,\ l_{xy} = 25.5218,\ l_{yy} = 2432.45$$

- 建立 y 关于 x 的一元线性回归方程：
$\hat{\beta}_1 = \frac{l_{xy}}{l_{xx}} = \frac{25.5218}{0.3024} = 84.3975$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 35.2389$. 所以回归方程为

$$\hat{y} = 35.2389 + 84.375x.$$

例子

- 现收集了 16 组合金钢的碳含量 x 与强度 y 的数据，整理得

$$\bar{x} = 0.125,\ \bar{y} = 45.7886,\ l_{xx} = 0.3024,\ l_{xy} = 25.5218,\ l_{yy} = 2432.45$$

- 建立 y 关于 x 的一元线性回归方程：
$\hat{\beta}_1 = \frac{l_{xy}}{l_{xx}} = \frac{25.5218}{0.3024} = 84.3975$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 35.2389$. 所以回归方程为

$$\hat{y} = 35.2389 + 84.375x.$$

- $\frac{1}{l_{xx}} = \frac{1}{0.3024} = 3.3069$, $\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} = 0.1142$. 所以

$$\hat{\beta}_0 \sim N(\beta_0, 0.1142\sigma^2), \quad \hat{\beta}_1 \sim N(\beta_1, 3.3069\sigma^2).$$

例子

- 一元线性回归方程的显著性检验:

$$S_T = l_{yy} = 2432.4566,\ S_R = \frac{l_{xy}^2}{l_{xx}} = 2153.9758,\ S_e = S_T - S_R = 278.$$

$$F = \frac{S_R}{S_e/(n-2)} = \frac{2153.9758}{278.4808/14} = 108.2848.$$

若取显著水平为 $\alpha = 0.05$, $F_{0.95}(1, 14) = 4.60$. 所以回归方程显著。

例子

- 一元线性回归方程的显著性检验:

$$S_T = l_{yy} = 2432.4566,\ S_R = \frac{l_{xy}^2}{l_{xx}} = 2153.9758,\ S_e = S_T - S_R = 278.$$

$$F = \frac{S_R}{S_e/(n-2)} = \frac{2153.9758}{278.4808/14} = 108.2848.$$

若取显著水平为 $\alpha = 0.05$, $F_{0.95}(1, 14) = 4.60$. 所以回归方程显著。

- $x = 0.15$ 时 $\hat{y}_0 = 35.2389 + 84.3975 \times 0.15 = 47.8985.$
  - $E(y)$ 的 0.95 置信区间为: $\hat{y}_0 \pm \delta$,
    $\delta = t_{0.975}(n-2)\hat{\sigma}\sqrt{1/n + (x_0 - \bar{x})^2/l_{xx}}$.
  - y 的概率为 0.95 的预测区间为 $\hat{y}_0 \pm \delta'$.
    $\delta' = t_{0.975}(n-2)\hat{\sigma}\sqrt{1 + 1/n + (x_0 - \bar{x})^2/l_{xx}}$.
  - $\hat{\sigma} = \sqrt{S_e/(n-2)}$.

例子

- 数据修正：n 组数据 (x_i, y_i)，已经计算出 $\bar{x}$, $\bar{y}$, l_{xx}, l_{xy}, l_{yy}. 现发现其中一组数据 (x_k, y_k) 记录有错，正确的应该为 (x'_k, y'_k)，怎么办？

例子

- 数据修正：n 组数据 (x_i, y_i)，已经计算出 $\bar{x}$, $\bar{y}$, l_{xx}, l_{xy}, l_{yy}. 现发现其中一组数据 (x_k, y_k) 记录有错，正确的应该为 (x'_k, y'_k)，怎么办？
- 修正各个计算数据：

$$\bar{x}' = \bar{x} + \frac{1}{n}(x'_k - x_k),$$

例子

- 数据修正：n 组数据 (x_i, y_i)，已经计算出 $\bar{x}$, $\bar{y}$, l_{xx}, l_{xy}, l_{yy}. 现发现其中一组数据 (x_k, y_k) 记录有错，正确的应该为 (x'_k, y'_k)，怎么办？
- 修正各个计算数据：

$$\bar{x}' = \bar{x} + \frac{1}{n}(x'_k - x_k),$$

$$\bar{y}' = \bar{y} + \frac{1}{n}(y'_k - y_k),$$

例子

- 数据修正：n 组数据 (x_i, y_i)，已经计算出 $\bar{x}$, $\bar{y}$, l_{xx}, l_{xy}, l_{yy}. 现发现其中一组数据 (x_k, y_k) 记录有错，正确的应该为 (x'_k, y'_k)，怎么办？
- 修正各个计算数据：

$$\bar{x}' = \bar{x} + \frac{1}{n}(x'_k - x_k),$$

$$\bar{y}' = \bar{y} + \frac{1}{n}(y'_k - y_k),$$

$$l_{xx} = \sum_i x_i^2 - n(\bar{x})^2 \Rightarrow l'_{xx} = l_{xx} + (x'_k)^2 - (x_k)^2 - n(\bar{x}')^2 + n(\bar{x})^2,$$

例子

- 数据修正：n 组数据 (x_i, y_i)，已经计算出 $\bar{x}$, $\bar{y}$, l_{xx}, l_{xy}, l_{yy}. 现发现其中一组数据 (x_k, y_k) 记录有错，正确的应该为 (x'_k, y'_k)，怎么办？
- 修正各个计算数据：

$$\bar{x}' = \bar{x} + \frac{1}{n}(x'_k - x_k),$$

$$\bar{y}' = \bar{y} + \frac{1}{n}(y'_k - y_k),$$

$$l_{xx} = \sum_i x_i^2 - n(\bar{x})^2 \Rightarrow l'_{xx} = l_{xx} + (x'_k)^2 - (x_k)^2 - n(\bar{x}')^2 + n(\bar{x})^2,$$

$$l'_{yy} = l_{yy} + (y'_k)^2 - (y_k)^2 - n(\bar{y}')^2 + n(\bar{y})^2,$$

例子

- 数据修正：n 组数据 (x_i, y_i)，已经计算出 $\bar{x}$, $\bar{y}$, l_{xx}, l_{xy}, l_{yy}. 现发现其中一组数据 (x_k, y_k) 记录有错，正确的应该为 (x'_k, y'_k)，怎么办？
- 修正各个计算数据：

$$\bar{x}' = \bar{x} + \frac{1}{n}(x'_k - x_k),$$

$$\bar{y}' = \bar{y} + \frac{1}{n}(y'_k - y_k),$$

$$l_{xx} = \sum_i x_i^2 - n(\bar{x})^2 \Rightarrow l'_{xx} = l_{xx} + (x'_k)^2 - (x_k)^2 - n(\bar{x}')^2 + n(\bar{x})^2,$$

$$l'_{yy} = l_{yy} + (y'_k)^2 - (y_k)^2 - n(\bar{y}')^2 + n(\bar{y})^2,$$

$$l_{xy} = \sum_i x_i y_i - n(\bar{x})(\bar{y}) \Rightarrow l'_{xy} = l_{xy} + x'_k y'_k - x_k y_k + n(\bar{x})(\bar{y}) - n(\bar{x}')(\bar{y}').$$

一元非线性回归

- 非线性函数关系：如
  - 双曲：$\frac{1}{y} = a + \frac{b}{x}$;
  - 幂函数：$y = ax^b$;
  - 指数函数：$y = ae^{bx}$ 或者 $y = ae^{b/x}$;
  - 对数函数：$y = a + b\ln x$;
  - S-形曲线：$y = \frac{1}{a+be^{-x}}$.

一元非线性回归

- 非线性函数关系：如
  - 双曲：$\frac{1}{y} = a + \frac{b}{x}$;
  - 幂函数：$y = ax^b$;
  - 指数函数：$y = ae^{bx}$ 或者 $y = ae^{b/x}$;
  - 对数函数：$y = a + b\ln x$;
  - S-形曲线：$y = \frac{1}{a+be^{-x}}$.
- 通过变量代换，将非线性关系变成线性关系，再进行一元线性回归分析。
  - 双曲：$u = \frac{1}{y}$, $v = \frac{1}{x}$, $u = a + bv$;
  - 幂函数：$u = \ln y$, $v = \ln x$, $u = \ln a + bv$;
  - 指数函数：$u = \ln y$, $u = \ln a + bx$ 或者 $u = \ln y$, $v = \frac{1}{x}$, $u = \ln a + bv$;
  - 对数函数：$y = a + b\ln x$; $v = \ln x$, $y = a + bv$.
  - S-形曲线：$u = \frac{1}{y}$, $v = e^{-x}$, $u = a + bv$.