1 一元线性回归

1.1 基本概念与假设

回归 (Regression)：F. 高尔顿在研究父子身高遗传问题中提出的.

一元线性回归

涉及 一个 自变量的回归，因变量与自变量之间是 线性关系

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x + \varepsilon \quad (\text{总体}) \\
y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \quad (\text{样本})
\end{aligned}
$$

- 即 y 与 x 的关系由两部分描述，第一部分是由于 x 的变化引起的 y 的线性变化 $(\beta_0 + \beta_1 x)$，另一部分是由其他一切随机因素引起的 (ε).
    - 其他随机因素会有：由于条件的制约没有因入模型的自变量对因变量造成的影响；观测误差；模型设定的误差；其他误差
- 对于每个确定的 x_i 值，y_i 都会有一个分布，而对于许多观测到的 x_i 值，会对应不同的 y_i 的分布，而我们不关心最大的 y 的分布，我们主要关心在 x_i 给定的条件下，某个具体的 x_i 的分布，因而在研究中，我们将 x_i 视为给定的 常数，而不是随机变量，而将其对应的 y_i 视为 随机变量.
- 称描述 y 的 平均值 如何依赖于 x 的方程，称为 回归方程：如一元线性回归方程：

  $$E(y \mid x) = \beta_0 + \beta_1 x$$

  其中 β_1 称为 回归系数，它表示当 x 每变动一个单位时，y 的平均变动值
- 对于给定的样本 (x_1, y_1)，(x_2, y_2)，$\cdots$，(x_n, y_n) 可以对其中的参数作出估计，得到 估计 的回归直线：

  $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{1}$$

  其中 $\hat{y}_i$ 称为 回归拟合值.

基本假定

1. Gauss-Markov 条件：零均值、同方差、不相关

$$
\begin{cases}
E(\varepsilon_i) = 0, \quad i = 1, \cdots, n, \\
\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j, \\ 0, & i \neq j. \end{cases}
\end{cases}
$$

2. 正态 假定：$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \cdots, n, \\ \varepsilon_1, \cdots, \varepsilon_n \text{独立}. \end{cases}$
3. 通常默认满足：$\operatorname{Cov}(\varepsilon_i, x_i) = 0.$

1.2 参数估计与性质

1.2.1 参数估计

记 $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$，$\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$.

1. β_0, β_1 的估计 OLS

核心：要使得残差平方和最小：$\min \quad Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

- 配方法：

  由 (1) 式：$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$，则

$$
\begin{aligned}
Q&=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=\sum_{i=1}^{n}(y_i-\overline{y}+\overline{y}-\hat{\beta}_0-\hat{\beta}_1\overline{x}+\hat{\beta}_1\overline{x}-\hat{\beta}_1x_i)^2\\
&=\sum_{i=1}^{n}(y_i-\overline{y})^2+\sum_{i=1}^{n}(\overline{y}-\hat{\beta}_0-\hat{\beta}_1\overline{x})^2+\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\overline{x})^2-2\hat{\beta}_1\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})
\end{aligned}
$$

  引入记号：$L_{xx}\overset{\Delta}{=}\sum_{i=1}^{n}(x_i-\overline{x})^2$，$L_{xy}\overset{\Delta}{=}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$，$L_{yy}\overset{\Delta}{=}\sum_{i=1}^{n}(y_i-\overline{y})^2$，则

$$
\begin{aligned}
Q&=\hat{\beta}_1^2L_{xx}-2\hat{\beta}_1L_{xy}+L_{yy}+n(\overline{y}-\hat{\beta}_0-\hat{\beta}_1\overline{x})^2\\
&=L_{xx}\left(\hat{\beta}_1^2-2\frac{L_{xy}}{L_{xx}}\hat{\beta}_1+\frac{L_{xy}^2}{L_{xx}^2}\right)-\frac{L_{xy}^2}{L_{xx}}+L_{yy}+n(\overline{y}-\hat{\beta}_0-\hat{\beta}_1\overline{x})^2\\
&=L_{xx}\left(\hat{\beta}_1-\frac{L_{xy}}{L_{xx}}\right)^2-\frac{L_{xy}^2}{L_{xx}}+L_{yy}+n(\overline{y}-\hat{\beta}_0-\hat{\beta}_1\overline{x})^2
\end{aligned}
$$

  当样本给定时，L_{xx}，L_{xy}，L_{yy} 都给定，要想使得 Q 最小，即使得 $\begin{cases}\hat{\beta}_1=\dfrac{L_{xy}}{L_{xx}},\\ \hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}.\end{cases}$

  - 还能得到 $Q_{\min}=L_{yy}-\dfrac{L_{xy}^2}{L_{xx}}$，在 ☞ § 2.3 中的常用二级结论中，会发现这里的三项还有别的表达方式，其实就是 $\mathrm{SSE}=\mathrm{SST}-\mathrm{SSR}$.

- 求导法：

  $Q=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2 \implies$

$$
\begin{cases}\dfrac{\partial Q}{\partial\hat{\beta}_0}=-2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0,\\ \dfrac{\partial Q}{\partial\hat{\beta}_1}=-2\sum_{i=1}^{n}x_i(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0.\end{cases}\tag{2}
$$

  令两个偏导数为 0，即得到了正规方程组，拆开整理，很容易求解出相同的结果.

  - 更重要的是，如果令 $e_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_i$，则由 (2) 式，还有 $\begin{cases}\sum_{i=1}^{n}e_i=0,\\ \sum_{i=1}^{n}x_ie_i=0.\end{cases}$

2. σ^2 的估计：MLE

由于 MLE 估计需要总体分布，由 $\varepsilon_i\sim N(0,\sigma^2)$，得到 $y_i\sim N(\beta_0+\beta_1x_i,\ \sigma^2)$，$i=1,2,\cdots,n$，由于 x_i 被视作常数，那么研究 y_i 的似然函数：

$$
L(y_1,\cdots,y_n;\beta_0,\beta_1,\sigma^2)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2\right\}
$$

对其求关于 β_0,β_1 的最值时，就是对指数部分求导，并令其为 0，方程与求导法一样，因此结果也与 OLS 的结果一样，但是 MLE 还能够给出 σ^2 的估计，对 σ^2 求导并令其为 0：$\hat{\sigma}^2=\dfrac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$，实际上常常使用修正过的估计量：

$$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

1.2.2 估计量的性质与分布

- 在进行下面的讨论前，一定要明白哪些是常数，哪些是随机变量：
  - 常数：x_i，$\overline{x}$，β_0，β_1，L_{xx}，σ^2
  - 随机变量：y_i，$\overline{y}$，ε_i，$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{y}_i$，e_i，$\hat{\sigma}^2$

1. 线性性

由于 $\sum_{i=1}^{n}(x_i-\overline{x})=0$，故 $\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\sum_{i=1}^{n}\frac{x_i-\overline{x}}{L_{xx}}\cdot y_i$.

结合 $\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}$，可以得到：

- $\hat{\beta}_0$，$\hat{\beta}_1$ 都是 y_i 的线性组合，因此都服从正态分布.

2. 无偏性

$\hat{\beta}_0$，$\hat{\beta}_1$ 的无偏性

$$E(\hat{\beta}_1)=\sum_{i=1}^{n}\frac{(x_i-\overline{x})}{L_{xx}}\cdot E(y_i)=\sum_{i=1}^{n}\frac{(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot(\beta_0+\beta_1x_i)$$

- 拆成两项，分别计算，得到 $E(\hat{\beta}_1)=\beta_1$.
- 进而由 $\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_1x_i)=\frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1x_i+\varepsilon_i-\hat{\beta}_1x_i)$ 得到：

$$E(\hat{\beta}_0)=\frac{1}{n}\sum_{i=1}^{n}(\beta_0+x_iE(\beta_1-\hat{\beta}_1))=\beta_0.$$

$\hat{\sigma}^2$ 的无偏性

- 由定义 $e_i=y_i-\hat{y}_i=\beta_0+\beta_1x_i+\varepsilon_i-\hat{\beta}_0-\hat{\beta}_1x_i$，即 $E(e_i)=0$.
- 由 $\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{1}{n-2}\sum_{i=1}^{n}e_i^2$，得：

$$E(\hat{\sigma}^2)=\frac{1}{n-2}\sum_{i=1}^{n}E(e_i^2)=\frac{1}{n-2}\sum_{i=1}^{n}\mathrm{Var}(e_i)$$

由下面的 (4) 式，可得 $\mathrm{Var}(e_i)=\left[1-\frac{1}{n}-\frac{(x_i-\overline{x})^2}{L_{xx}}\right]\sigma^2$，故

$$\begin{aligned}E(\hat{\sigma}^2)&=\frac{1}{n-2}\sum_{i=1}^{n}\left[1-\frac{1}{n}-\frac{(x_i-\overline{x})^2}{L_{xx}}\right]\sigma^2\\&=\frac{\sigma^2}{n-2}\left(n-1-\sum_{i=1}^{n}\frac{(x_i-\overline{x})^2}{L_{xx}}\right)\\&=\sigma^2\end{aligned}$$

3. $\hat{\beta}_0$，$\hat{\beta}_1$，e_i 的方差

由于 $\hat{\beta}_1$ 是 y_i 的线性组合，因此，记 $\hat{\beta}_1=\sum_{i=1}^{n}k_iy_i$，其中 $k_i=\frac{x_i-\overline{x}}{L_{xx}}$

- 由 k_i 的定义，易得 $\sum_{i=1}^n k_i = 0$，$\sum_{i=1}^n k_i x_i = 1$，$\sum_{i=1}^n k_i^2 = \frac{1}{L_{xx}}$.
- 由于我们已知方差的变量是 y_i，ε_i，因此，为了研究二者的方差，要将 $\hat\beta_0$，$\hat\beta_1$ 写成 y_i 的 线性组合 的形式.

$\hat\beta_1$ 的方差

由线性性以及 $y_i \overset{iid.}{\sim} N(\beta_0+\beta_1 x_i,\ \sigma^2)$：

$$\mathrm{Var}(\hat\beta_1) = \sum_{i=1}^n k_i^2 \cdot \mathrm{Var}(y_i) = \frac{\sigma^2}{L_{xx}}$$

$\therefore\ \hat\beta_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{L_{xx}}\right)$

$\hat\beta_0$ 的方差

$$\begin{aligned}\hat\beta_0 = \overline{y} - \hat\beta_1\overline{x} &= \frac{1}{n}\sum_{i=1}^n y_i - \overline{x}\sum_{i=1}^n k_i y_i \\ &= \sum_{i=1}^n\left(\frac{1}{n}-\overline{x}k_i\right)y_i = \sum_{i=1}^n\left(\frac{1}{n}-\overline{x}k_i\right)(\beta_0+\beta_1 x_i+\varepsilon_i) \\ &= \beta_0 + \sum_{i=1}^n\left(\frac{1}{n}-\frac{\overline{x}(x_i-\overline{x})}{L_{xx}}\right)\varepsilon_i\end{aligned}$$

由于诸 ε_i 不相关，且 β_0 为常数，因此

$$\begin{aligned}\mathrm{Var}(\hat\beta_0) &= \sigma^2\sum_{i=1}^n\left(\frac{1}{n}-\frac{\overline{x}(x_i-\overline{x})}{L_{xx}}\right)^2 \\ &= \left(\frac{1}{n}+\frac{\overline{x}^2}{L_{xx}}\right)\sigma^2 = \frac{\sigma^2}{nL_{xx}}\sum_{i=1}^n x_i^2\end{aligned}$$

$\therefore\ \hat\beta_0 \sim N\left(\beta_0,\ \left(\frac{1}{n}+\frac{\overline{x}^2}{L_{xx}}\right)\sigma^2\right)$

$\hat\beta_0, \hat\beta_1$ 的协方差

同样牢牢抓住 $\hat\beta_0, \hat\beta_1$ 都是 y_i 的线性组合这一点：

$$\begin{aligned}\mathrm{Cov}(\hat\beta_0,\hat\beta_1) &= \mathrm{Cov}(\overline{y}-\hat\beta_1\overline{x},\ \hat\beta_1) \\ &= \mathrm{Cov}(\overline{y},\hat\beta_1) - \overline{x}\mathrm{Var}(\hat\beta_1)\end{aligned}$$

其中：

$$\begin{aligned}\mathrm{Cov}(\overline{y},\hat\beta_1) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^n y_i,\ \frac{1}{n}\sum_{i=1}^n k_i y_i\right) = \frac{1}{n^2}\mathrm{Cov}\left(\sum_{i=1}^n y_i,\ \sum_{i=1}^n k_i y_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^n k_i \cdot \mathrm{Var}(y_i) = 0.\end{aligned}$$

所以

$$\mathrm{Cov}(\hat\beta_0,\hat\beta_1) = -\frac{\overline{x}}{L_{xx}}\sigma^2$$

- 进而对于 任意给定 的 x_i：$\hat y_i = \hat\beta_i + \hat\beta_1 x_i$，也是 y_i 的线性组合，因此 $\hat y_i$ 也服从正态分布，
  - $E(\hat y_i) = \beta_0 + \beta_1 x_i = E(y_i)$，即 $\hat y_i$ 是 $E(y_i)$ 的无偏估计.
  - $\mathrm{Var}(\hat y_i) = \mathrm{Var}(\hat\beta_0) + x_i^2\mathrm{Var}(\hat\beta_1) + 2x_i\mathrm{Cov}(\hat\beta_0,\hat\beta_1) = \left[\frac{1}{n}+\frac{(x_i-\overline{x})^2}{L_{xx}}\right]\sigma^2$.

  即

$$\hat{y}_i \sim N\left(\beta_0+\beta_1 x_i,\ \left[\frac{1}{n}+\frac{(x_i-\overline{x})^2}{L_{xx}}\right]\sigma^2\right) \tag{3}$$

- 有时候记 $h_{ii}=\dfrac{1}{n}+\dfrac{(x_i-\overline{x})^2}{L_{xx}}$ 称为 杠杆值，且 $0<h_{ii}<1$，则 $\hat{y}_i \sim N(\beta_0+\beta_1 x_i,\ h_{ii}\sigma^2)$.

e_i 的方差

求 e_i 的方差，核心也是将其化为 y_i 的若干个线性函数之间的关系，使用 y_i 的方差进行处理.

$\operatorname{Var}(e_i)=\operatorname{Var}(y_i-\hat{y}_i)=\operatorname{Var}(y_i)+\operatorname{Var}(\hat{y}_i)-2\operatorname{Cov}(y_i,\hat{y}_i)$

- 单独考虑 $\operatorname{Cov}(y_i,\hat{y}_i)$：

$$\begin{aligned}\operatorname{Cov}(y_i,\ \hat{y}_i) &= \operatorname{Cov}(y_i,\ \hat{\beta}_0+\hat{\beta}_1 x_i)=\operatorname{Cov}(y_i,\ \overline{y}-\hat{\beta}_1\overline{x}+\hat{\beta}_1 x_i)\\ &= \operatorname{Cov}(y_i,\ \overline{y})+\operatorname{Cov}(y_i,\ \hat{\beta}_1(x_i-\overline{x}))\\ &= \frac{1}{n}\sigma^2+\operatorname{Cov}\left(y_i,\ \sum_{i=1}^n\frac{(x_i-\overline{x})^2}{L_{xx}}y_i\right)\\ &= \frac{1}{n}\sigma^2+\operatorname{Cov}\left(y_i,\ \frac{(x_i-\overline{x})^2}{L_{xx}}y_i\right)\\ &= \frac{1}{n}\sigma^2+\frac{(x_i-\overline{x})^2}{L_{xx}}\sigma^2.\end{aligned}$$

由 (3)：$\operatorname{Var}(\hat{y}_i)=\left[\dfrac{1}{n}+\dfrac{(x_i-\overline{x})^2}{L_{xx}^2}\right]\sigma^2$，因此得到

$$\operatorname{Var}(e_i)=\left[1-\frac{1}{n}-\frac{(x_i-\overline{x})^2}{L_{xx}}\right]\sigma^2 \tag{4}$$

使用杠杆值 h_{ii}，还可以表示为 $\operatorname{Var}(e_i)=(1-h_{ii})\sigma^2$

- 在寻找异常值的时候，人们往往认为 $\pm 2\hat{\sigma}$ 或 $\pm 3\hat{\sigma}$ 的残差为异常值，由于 e_i 的方差为 $(1-h_{ii})\sigma^2$，不相等，因此，人们对于残差进行改进，分别提出了 标准化残差 、学生化残差：
  - 标准化残差：$\text{ZRE}_i=\dfrac{e_i}{\hat{\sigma}}$，学生化残差：$\text{SRE}_i=\dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$.
  - 标准化残差使得残差具有可比性，同时学生化残差使得残差的方差相等
  - 在寻找异常值的时候，一般使用学生化残差.

4. 最佳线性无偏估计（BLUE）
   - 若某统计量是 样本观测值 的线性函数，且是无偏估计，则称该统计量是 线性无偏估计 .

G-M 定理：$\hat{\beta}_0,\ \hat{\beta}_1$ 是 β_0,β_1 的最小方差线性无偏估计（最佳线性无偏估计）

证明：$\hat{\beta}_1=\sum_{i=1}^n\dfrac{(x_i-\overline{x})}{L_{xx}}\cdot y_i$ 是 β_1 的 BLUE:

设 β_1 的线性无偏估计为 $\tilde{\beta}_1$，即证明 $\hat{\beta}_1$ 是这些线性无偏估计中方差最小的，可以看作是条件极值问题。

由于 $\tilde{\beta}_1$ 要是线性无偏的，因此，不妨假设 $\tilde{\beta}_1=\sum_{i=1}^n c_i y_i$，则 $E(\tilde{\beta}_1)=\sum_{i=1}^n c_i E(\beta_0+\beta_1 x_i+\varepsilon_i)$，因此得到

约束条件：$\sum_{i=1}^n c_i=0,\ \sum_{i=1}^n c_i x_i=1$. 又 $\operatorname{Var}(\tilde{\beta}_1)=\sigma^2\sum_{i=1}^n c_i^2$，因此即求解条件极值问题：

$$\begin{aligned}\min\quad & \operatorname{Var}(\tilde{\beta}_1)=\sum_{i=1}^n c_i^2\\ \text{s.t.}\quad & \sum_{i=1}^n c_i=0\\ & \sum_{i=1}^n c_i x_i=1\end{aligned}$$

构造 Lagrange 函数：$L(c_1,\cdots,c_n,\lambda,\mu)=\sum\limits_{i=1}^n c_i^2-\lambda\sum\limits_{i=1}^n c_i-\mu\left(\sum\limits_{i=1}^n c_ix_i-1\right)$

令偏导为 0，得到 $n+2$ 个方程：

$$2c_j+\lambda+\mu x_j=0,\ \ j=1,\cdots,n$$
$$\sum_{i=1}^n c_i=0,\quad \sum_{i=1}^n c_ix_i=1$$

对最上面的 n 个方程求和得到：$\lambda+\mu\overline{x}=0$；对上面 n 个方程同乘 x_j 后求和：$2+\lambda n\overline{x}+\mu\sum\limits_{i=1}^n x_i^2=0$，再将 $\lambda=-\mu\overline{x}$ 代入，得到 $\mu=\dfrac{-2}{L_{xx}}$，进而 $\lambda=\dfrac{2\overline{x}}{L_{xx}}$，再代入前 n 个方程，得到 $c_i=\dfrac{x_i-\overline{x}}{L_{xx}}$，即证明了定理.

证明： $\hat\beta_0=\sum\limits_{i=1}^n\left(\dfrac{1}{n}+\dfrac{\overline{x}(x_i-\overline{x})}{L_{xx}}\right)\cdot y_i$ 是 β_0 的 BLUE：

一样的构造条件极值问题，一样的求解方程组方法，不过限制条件变为 $\sum\limits_{i=1}^n c_i=1,\ \sum\limits_{i=1}^n c_ix_i=0.$

1.3 平方和分解

记总偏差平方和为 $\mathrm{SST}=\sum\limits_{i=1}^n(y_i-\overline{y})^2$，$\mathrm{SSE}=\sum\limits_{i=1}^n(y_i-\hat y_i)^2$，$\mathrm{SSR}=\sum\limits_{i=1}^n(\hat y_i-\overline{y})^2$.

证明：$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE}$：

$$\begin{aligned}\mathrm{SST}&=\sum_{i=1}^n(y_i-\overline{y})^2=\sum_{i=1}^n(y_i-\hat y_i+\hat y_i-\overline{y})^2\\&=\mathrm{SSE}+\mathrm{SSR}+2\sum_{i=1}^n(y_i-\hat y_i)(\hat y_i-\overline{y})\end{aligned}$$

单独考虑交叉项，由于 $\hat y_i$ 是与 $\overline{y}$ 是有关系的，具体来说：

$$\hat\beta_1=\frac{\hat y_i-\overline{y}}{x_i-\overline{x}}\tag{5}$$

几何意义就是过数据中心的回归直线斜率. 表明 $\hat y_i-\overline{y}$ 与 $x_i-\overline{x}$ 存在倍数关系，也可以联立 $\begin{cases}\hat y_i=\hat\beta_0+\hat\beta_1x_i,\\ \hat\beta_0=\overline{y}-\hat\beta_1\overline{x},\end{cases}$ 来证明. 故有 $\hat y_i=\overline{y}+\hat\beta_1(x_i-\overline{x})$，这一点在求 ☞ $\mathrm{Cov}(y_i,\hat y_i)$ 时已经用到了，因此交叉项化为：

$$2\sum_{i=1}^n(y_i-\hat y_i)\cdot(\hat y_i-\overline{y})=2\hat\beta_1\sum_{i=1}^n e_i\cdot(x_i-\overline{x})$$

由正规方程组：$\sum\limits_{i=1}^n e_i=0$，$\sum\limits_{i=1}^n e_ix_i=0$，得交叉项为 0.

【注：常用结论】

1. 其实这些平方和在之前的叙述中也出现过，如 SST 就是 L_{yy}，SSE 就是对回归参数进行 OLS 估计时用到的残差平方和.
2. 利用平方和分解的思想，将 $y_i-\overline{y}$ 写作 $y_i-\hat y_i+\hat y_i-\overline{y}$，再利用公式 (5)，还可以得到一些二级结论：

- 对 L_{xy} 使用：

$$
\begin{aligned}
L_{xy} &= \sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sum_{i=1}^{n}(x_i-\overline{x})(y_i-\hat{y}_i+\hat{y}_i-\overline{y}) \\
&= \sum_{i=1}^{n}(x_i-\overline{x})e_i + \sum_{i=1}^{n}(x_i-\overline{x})(\hat{y}_i-\overline{y}) \\
&= \hat{\beta}_1\sum_{i=1}^{n}(x_i-\overline{x})^2 = \hat{\beta}_1 L_{xx}
\end{aligned} \tag{6}
$$

  这个结论也可以直接从最小二乘估计得到：$\hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}}$.

- 对 L_{yy} 使用：
  首先，由平方和分解：$L_{yy} = \mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$，其中含有 $\hat{y}_i - \overline{y}$ 的是 SSR，故

$$
\mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 = \hat{\beta}_1^2 L_{xx} \tag{7}
$$

  进一步，还有 $\mathrm{SSR} = \hat{\beta}_1 L_{xy} = \dfrac{L_{xy}^2}{L_{xx}}$. 这些表达式会在假设检验中用到.

3. 平方和还有各自的分布：
   - $\dfrac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2)$
   - 在 H_0 成立时，才有 $\dfrac{\mathrm{SSR}}{\sigma^2} \sim \chi^2(1)$
   - SSR 与 SSE 和 $\overline{y}$ 独立（或 $\hat{\beta}_1$，SSE，$\overline{y}$ 独立）

   证明需要构造合适的正交矩阵.
4. $\mathrm{SSE} = \mathrm{RSS}$, (error, residue), $\quad \mathrm{SSR} = \mathrm{ESS}$, (regression, explained)

1.4 显著性检验

1.4.1 t 检验

要检验回归系数是否显著，检验假设为：

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

此检验为双侧检验，容易想到 $\hat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{L_{xx}}\right)$，故原假设下有：$\dfrac{\hat{\beta}_1\sqrt{L_{xx}}}{\sigma} \sim N(0,1)$，因此想到基于此构造检验统计量，由于其中的 σ 未知，因此想到使用 $\hat{\sigma}$ 来代替，由于 $\dfrac{\mathrm{SSE}}{\sigma^2} = \dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$，故

$$
t = \frac{\hat{\beta}_1\sqrt{L_{xx}}}{\hat{\sigma}} \sim t(n-2) \tag{8}
$$

拒绝域的形式为 $W = \left\{|t| > t_{1-\frac{\alpha}{2}}(n-2)\right\}$.

1.4.2 F 检验

由 $\dfrac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2)$ 与 $\dfrac{\mathrm{SSR}}{\sigma^2} \sim \chi^2(1)$ 且二者独立，也可以考虑构造 F 统计量.

检验假设仍为 ☞ 1.4.1 中 的假设，当回归模型成立时，被解释的方差 (SSR) 会很大，不被解释的方差 (SSE) 很小，因此

$$
F = \frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)} \sim F(1, n-2) \tag{9}
$$

拒绝域的形式为：$W=\{F>F_{1-\alpha}(1,n-2)\}$

- 特别地，F 检验也可以写在方差分析表中进行：

来源	平方和	自由度	均方	F	P
回归	SSR	1	$\mathrm{SSR}/1$	$\dfrac{\mathrm{SSR}/1}{\mathrm{SSR}(n-2)}$	p
残差	SSE	$n-2$	$\mathrm{SSE}/(n-2)$		
总和	SST	$n-1$			

1.4.3 相关系数检验

决定系数与样本相关系数定义

决定系数 (R^2) R^2 表示的是：能够被解释变量解释的那部分 y_i 的离差平方和占总的离差平方和的比率，定义为

$$R^2=\frac{\mathrm{SSR}}{\mathrm{SST}}=1-\frac{\mathrm{SSE}}{\mathrm{SST}} \tag{10}$$

- 由平方和分解，知 $R^2\in[0,1]$.

样本相关系数 (r)：总体相关系数的表达式为：$\rho=\dfrac{\mathrm{Cov}(x_i,y_i)}{\sqrt{\mathrm{Var}(x_i)}\sqrt{\mathrm{Var}(y_i)}}$，样本相关系数表达式为

$$r=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}=\frac{L_{xy}}{\sqrt{L_{xx}\cdot L_{yy}}} \tag{11}$$

- 由柯西不等式，知 $|r|\leqslant 1$.

关系

1. $R^2=r^2$

证明：由于 R^2 与各个平方和有关系，而平方和与 L_{xx}，L_{xy}，L_{yy} 有关系，因此，根据 (6), (7) 得到思路：

$r^2=\dfrac{L_{xy}^2}{L_{xx}\cdot L_{yy}}$，上下同乘 $\hat{\beta}_1^2$，得：$r^2=\dfrac{\hat{\beta}_1^2L_{xy}^2}{L_{xx}\cdot L_{yy}\cdot\hat{\beta}_1^2}=\dfrac{\mathrm{SSR}^2}{\mathrm{SSR}\cdot L_{yy}}=\dfrac{\mathrm{SSR}}{\mathrm{SST}}$.

2. $r^2=\dfrac{F}{F+(n-2)}$

证明：$F=\dfrac{\mathrm{SSR}}{\mathrm{SSE}/(n-2)}$，$r^2=\dfrac{\mathrm{SSR}}{\mathrm{SSE}+\mathrm{SSR}}=\dfrac{\mathrm{SSR}/\mathrm{SSE}}{1+\mathrm{SSR}/\mathrm{SSE}}$，将 $\dfrac{\mathrm{SSR}}{\mathrm{SSE}}=\dfrac{F}{n-2}$ 代入，即得.

应该注意，其中 $F\sim F(1,n-2)$，第一自由度是 1，因此此处的 F 由 $T\sim t(n-2)\implies T^2\sim F(1,n-2)$ 可以写作 T 统计量的平方.

相关系数检验

检验回归方程是否显著，还可以通过检验相关系数是否显著异于 0 来进行，即

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

因此，检验统计量应该基于样本相关系数 r 进行构造，可以证明（在☞ § 1.4.4 会说明）检验统计量

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\sim t(n-2) \tag{12}$$

当 ρ 显著异于 0 时，$|t|$ 会非常大，因此，拒绝域的形式为 $W=\{|t|\geqslant t_{1-\frac{\alpha}{2}}(n-2)\}$.

1.4.4 三种检验的关系

在一元线性回归中，可以证明以上三种检验完全等价：

- 在回归系数 t 检验中，构造的检验统计量为 (8) : $t=\dfrac{\hat\beta_1\sqrt{L_{xx}}}{\hat\sigma}$；
- 在 F 检验中，构造的统计量为 (9) : $F=\dfrac{\text{SSR}}{\text{SSE}/(n-2)}$；
- 在相关系数 t 检验中，构造的统计量为 (12) : $t=\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

证明：$F=\dfrac{\text{SSR}}{\text{SSE}/(n-2)}=t^2$

首先，由 (7) 得：$\text{SSR}=\hat\beta_1^2L_{xx}$，且由 σ^2 的无偏估计：$\hat\sigma^2=\dfrac{\text{SSE}}{n-2}$，易得：

$$F=\frac{\hat\beta_1^2L_{xx}}{\hat\sigma^2}=\left(\frac{\hat\beta_1\sqrt{L_{xx}}}{\hat\sigma}\right)^2=t^2.$$

证明：$t=\dfrac{\hat\beta_1\sqrt{L_{xx}}}{\hat\sigma}=\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

由 $r=\dfrac{L_{xy}}{\sqrt{L_{xx}\cdot L_{yy}}}$，$r^2=R^2=\dfrac{\text{SSR}}{\text{SST}}$，得到：

$$\begin{aligned}\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}&=\frac{L_{xy}}{\sqrt{\dfrac{L_{xx}\cdot L_{yy}}{n-2}}\cdot\sqrt{\dfrac{\text{SSE}}{\text{SST}}}}=\frac{L_{xy}}{\sqrt{L_{xx}\cdot\dfrac{\text{SSE}}{n-2}}}\\&=\frac{L_{xy}}{\hat\sigma\sqrt{L_{xx}}}=\frac{L_{xy}\sqrt{L_{xx}}}{L_{xx}\hat\sigma}=\frac{\hat\beta_1\sqrt{L_{xx}}}{\hat\sigma}\end{aligned}$$

综上所述，三种检验是等价的.

1.5 区间估计

实际中，我们主要关心回归系数 $\hat\beta_1$ 的精度，主要寻找 β_1 的置信区间：

由 $\hat\beta_1$ 的分布：$\hat\beta_1\sim N\Big(\beta_1,\dfrac{\sigma^2}{L_{xx}}\Big)$，则 $\dfrac{\hat\beta_1-\beta_1}{\sigma/\sqrt{L_{xx}}}\sim N(0,1)$，其中 σ 未知，考虑使用 $\hat\sigma$ 替代，因为 $\dfrac{\text{SSE}}{\sigma^2}=\dfrac{(n-2)\hat\sigma^2}{\sigma^2}\sim\chi^2(n-2)$，故取枢轴量：

$$t=\frac{\hat\beta_1-\beta_1}{\hat\sigma/\sqrt{L_{xx}}}\sim t(n-2)$$

因此，得到 β_1 的置信区间为：$\hat\beta_1\pm\dfrac{\hat\sigma}{\sqrt{L_{xx}}}\cdot t_{1-\frac{\alpha}{2}}(n-2)$.

但是由于 $\hat\beta_0\sim N\left(\beta_0,\Big(\dfrac{1}{n}+\dfrac{\bar x^2}{L_{xx}}\Big)\sigma^2\right)$，也可以很轻易构造出 $\hat\beta_0$ 的置信区间.