双变量线性回归

教师 黄光辉
hgh@cqu.edu.cn

目录

检验,一把双刃剑.

警惕过度地检验假设.

检验,一把双刃剑.

警惕过度地检验假设.
你对数据愈苛求,数据就会愈多的向你供认;

检验,一把双刃剑.

警惕过度地检验假设.
你对数据愈苛求,数据就会愈多的向你供认;
但是在威逼下得到的供词,在科学询查的法庭上是不容许的.

内容线索

双变量模型的分布假定

$$Y = \beta_0 + \beta_1 X + U$$

双变量模型的分布假定

$$Y = \beta_0 + \beta_1 X + U$$

独立重复观测(Y, X),得到观测数据

$$(Y_1, X_1), (Y_2, X_2), \cdots, (Y_n, X_n).$$

干扰项无法直接观测,假设分布为:
$U_i \sim N(0, \sigma)$, $\{U_i\}$序列独立同分布.

双变量模型的分布假定

$$Y = \beta_0 + \beta_1 X + U$$

独立重复观测(Y, X),得到观测数据

$$(Y_1, X_1), (Y_2, X_2), \cdots, (Y_n, X_n).$$

干扰项无法直接观测,假设分布为:
$U_i \sim N(0, \sigma)$, $\{U_i\}$序列独立同分布.
采用最小二乘法,从上述数据得到模型参数的点估计为:

$$\begin{aligned}
\widehat{\beta_0} &= \overline{Y} - \widehat{\beta_1 X} \\
\widehat{\beta_1} &= \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \\
&= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\
&\quad \text{记号} x_i = X_i - \overline{X}
\end{aligned}$$

参数的分布

$$Y = \beta_0 + \beta_1 X + U$$

$$\widehat{\beta_0} \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{X}}{\sum\left(X_i - \overline{X}\right)^2}\right)\right)$$

$$\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{\sum\left(X_i - \overline{X}\right)^2}\right)$$

$$Z_1 = \frac{\widehat{\beta_0} - \beta_0}{\sigma_{\widehat{\beta_0}}} \sim N(0,1), \quad Z_2 = \frac{\widehat{\beta_1} - \beta_1}{\sigma_{\widehat{\beta_1}}} \sim N(0,1)$$

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u_i^2} \sim \chi^2(n-2).$$

内容线索

β_1的置信区间估计: t统计量

$$\widehat{\beta_1} \sim N(\beta_1, \sigma^2_{\widehat{\beta_1}})$$

β_1的置信区间估计: t统计量

$$\widehat{\beta_1} \sim N(\beta_1, \sigma^2_{\widehat{\beta_1}})$$

$$Z = \frac{\widehat{\beta_1} - \beta_1}{\sigma_{\widehat{\beta_1}}} = \frac{(\widehat{\beta_1} - \beta_1)\sqrt{\sum x_i^2}}{\sigma} \sim N(0,1)$$

β_1的置信区间估计: t统计量

$$\widehat{\beta_1} \sim N(\beta_1, \sigma^2_{\widehat{\beta_1}})$$

$$Z = \frac{\widehat{\beta_1} - \beta_1}{\sigma_{\widehat{\beta_1}}} = \frac{(\widehat{\beta_1} - \beta_1)\sqrt{\sum x_i^2}}{\sigma} \sim N(0,1)$$

$$Z_3 = (n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi(n-2)$$

β_1的置信区间估计: t统计量

$$\widehat{\beta_1} \sim N(\beta_1, \sigma^2_{\widehat{\beta_1}})$$

$$Z = \frac{\widehat{\beta_1} - \beta_1}{\sigma_{\widehat{\beta_1}}} = \frac{(\widehat{\beta_1} - \beta_1)\sqrt{\sum x_i^2}}{\sigma} \sim N(0,1)$$

$$Z_3 = (n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi(n-2)$$

$$t = \frac{Z}{\sqrt{Z_3/(n-2)}}$$

β_1的置信区间估计: t统计量

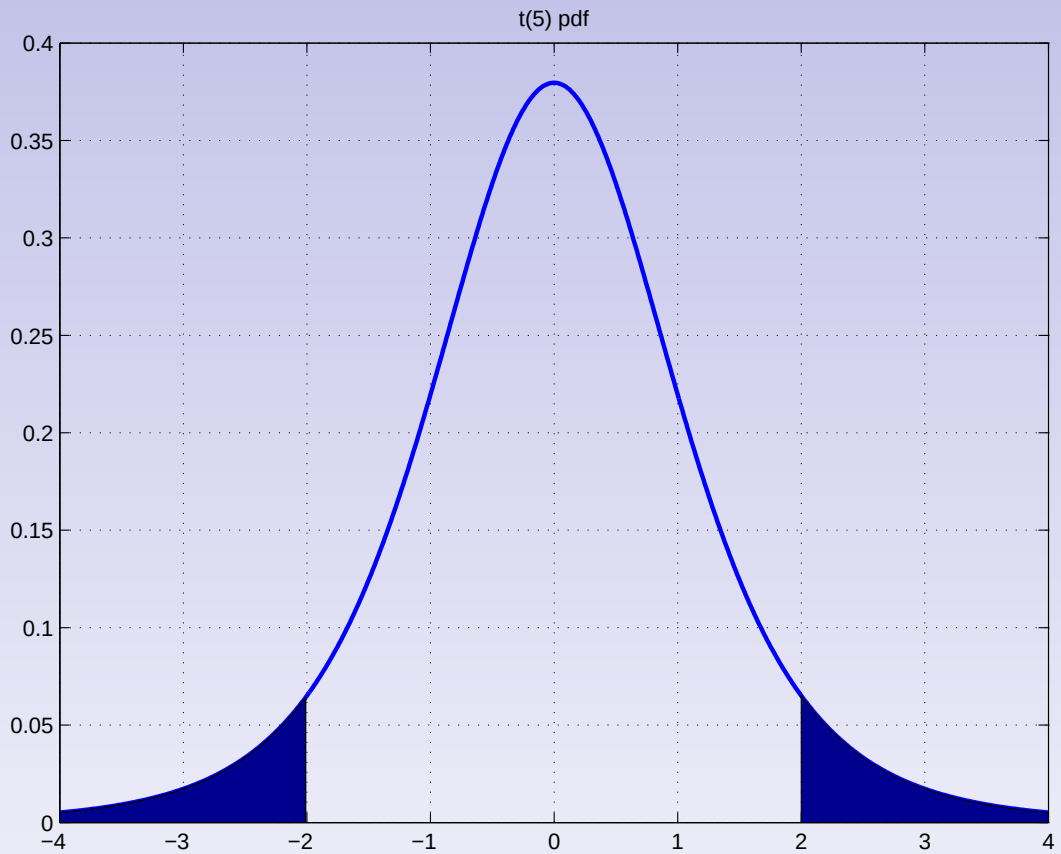
$$\widehat{\beta_1} \sim N(\beta_1, \sigma^2_{\widehat{\beta_1}})$$

$$Z = \frac{\widehat{\beta_1} - \beta_1}{\sigma_{\widehat{\beta_1}}} = \frac{(\widehat{\beta_1} - \beta_1)\sqrt{\sum x_i^2}}{\sigma} \sim N(0,1)$$

$$Z_3 = (n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi(n-2)$$

$$t = \frac{Z}{\sqrt{Z_3/(n-2)}} = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim t(n-2)$$

t分布密度函数,$1-\alpha$置信区间



$1-\alpha$置信区间

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-2) \leq t \leq t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

β_1的置信区间估计

$$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim t(n-2)$$

β_1的置信区间估计

$$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \leq t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

β_1的置信区间估计

$$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \leq t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

$$P\left\{\widehat{\beta_1} - se(\widehat{\beta_1})t_{1-\frac{\alpha}{2}}(n-2) \leq \beta_1 \leq \widehat{\beta_1} + se(\widehat{\beta_1})t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

β_1的置信区间估计

$$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \leq t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

$$P\left\{\widehat{\beta_1} - se(\widehat{\beta_1})t_{1-\frac{\alpha}{2}}(n-2) \leq \beta_1 \leq \widehat{\beta_1} + se(\widehat{\beta_1})t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

β_1的置信区间为:

$$\left(\widehat{\beta_1} - se(\widehat{\beta_1})t_{1-\frac{\alpha}{2}}(n-2), \widehat{\beta_1} + se(\widehat{\beta_1})t_{1-\frac{\alpha}{2}}(n-2)\right)$$

β_0的置信区间估计

$$t = \frac{\widehat{\beta_0} - \beta_0}{se(\widehat{\beta_0})} \sim t(n-2)$$

β_0的置信区间估计

$$t = \frac{\widehat{\beta_0} - \beta_0}{se(\widehat{\beta_0})} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\widehat{\beta_0} - \beta_0}{se(\widehat{\beta_0})} \leq t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

β_0的置信区间估计

$$t = \frac{\widehat{\beta_0} - \beta_0}{se(\widehat{\beta_0})} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\widehat{\beta_0} - \beta_0}{se(\widehat{\beta_0})} \leq t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

$$P\left\{\widehat{\beta_0} - se(\widehat{\beta_0})t_{1-\frac{\alpha}{2}}(n-2) \leq \beta_0 \leq \widehat{\beta_0} + se(\widehat{\beta_0})t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

β_0的置信区间估计

$$t = \frac{\widehat{\beta_0} - \beta_0}{se(\widehat{\beta_0})} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\widehat{\beta_0} - \beta_0}{se(\widehat{\beta_0})} \leq t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

$$P\left\{\widehat{\beta_0} - se(\widehat{\beta_0})t_{1-\frac{\alpha}{2}}(n-2) \leq \beta_0 \leq \widehat{\beta_0} + se(\widehat{\beta_0})t_{1-\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

β_0的置信区间为:

$$\left(\widehat{\beta_0} - se(\widehat{\beta_0})t_{1-\frac{\alpha}{2}}(n-2), \widehat{\beta_0} + se(\widehat{\beta_0})t_{1-\frac{\alpha}{2}}(n-2)\right)$$

干扰项方差σ^2的区间估计

$$\chi^2 = (n-2)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)$$

干扰项方差σ^2的区间估计

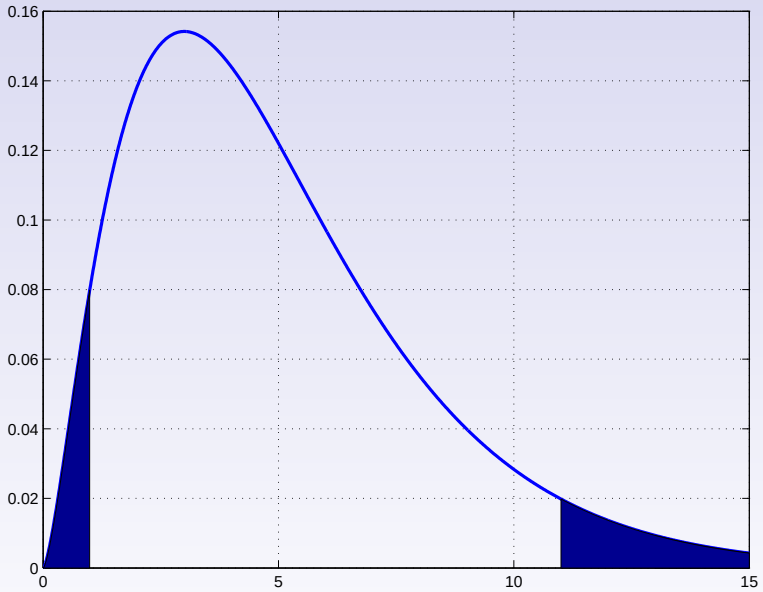
$$\chi^2 = (n-2)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)$$

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n} \widehat{u_i}^2$$

干扰项方差σ^2的区间估计

$$\chi^2 = (n-2)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)$$

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n} \widehat{u_i}^2$$



σ^2的置信区间

$$P\left\{\chi^2_{\frac{\alpha}{2}} \leq \frac{n-2}{\sigma^2}\widehat{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}\right\} = 1-\alpha$$

σ^2的置信区间

$$P\left\{\chi^2_{\frac{\alpha}{2}} \leq \frac{n-2}{\sigma^2}\widehat{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}\right\} = 1-\alpha$$

$$\left(\frac{(n-2)\widehat{\sigma^2}}{\chi^2_{1-\frac{\alpha}{2}}}, \frac{(n-2)\widehat{\sigma^2}}{\chi^2_{\frac{\alpha}{2}}}\right)$$

内容线索

统计量的观测值

β_0和β_1的估计值为$\widehat{\beta_0}$和$\widehat{\beta_1}$。

统计量的观测值

β_0和β_1的估计值为$\widehat{\beta_0}$和$\widehat{\beta_1}$。
统计量的观测值通常出现在什么地方呢？

统计量的观测值

β_0和β_1的估计值为$\widehat{\beta_0}$和$\widehat{\beta_1}$。
统计量的观测值通常出现在什么地方呢?
由大数定律，统计量会出现在自己中心附近区域，
越远离中心，出现的可能性越小。
从而，我们可以断定
$\widehat{\beta_0}$总是出现在真实的β_0附近;
$\widehat{\beta_1}$总是出现在真实的β_1附近。

统计量的观测值

β_0和β_1的估计值为$\widehat{\beta_0}$和$\widehat{\beta_1}$。
统计量的观测值通常出现在什么地方呢?
由大数定律，统计量会出现在自己中心附近区域，
越远离中心，出现的可能性越小。
从而，我们可以断定
$\widehat{\beta_0}$总是出现在真实的β_0附近;
$\widehat{\beta_1}$总是出现在真实的β_1附近。
由此，我们可以回答下列问题
一元线性回归的截距项是否为一个指定的常数β_0
一元线性回归的斜率项是否为一个指定的常数β_1.

参数异常观测的刻画：拒绝域和接受域

$$y = \beta_0 + \beta_1 x + u$$

$$\widehat{\beta_1} \sim N\left(\beta_1, \widehat{\sigma}^2_{\widehat{\beta_1}}\right), \quad t = \frac{\widehat{\beta_1} - \beta_1}{se\left(\widehat{\beta_1}\right)} \sim t(n-2)$$

参数异常观测的刻画：拒绝域和接受域

$$y = \beta_0 + \beta_1 x + u$$

$$\widehat{\beta_1} \sim N\left(\beta_1, \widehat{\sigma}^2_{\widehat{\beta_1}}\right), \quad t = \frac{\widehat{\beta_1} - \beta_1}{se\left(\widehat{\beta_1}\right)} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}} \leq t \leq t_{1-\frac{\alpha}{2}}\right\} = P\left\{-t_{1-\frac{\alpha}{2}} \leq \frac{\widehat{\beta_1} - \beta_1}{se\left(\widehat{\beta_1}\right)} \leq t_{1-\frac{\alpha}{2}}\right\} = 1 - \alpha$$

参数异常观测的刻画：拒绝域和接受域

$$y = \beta_0 + \beta_1 x + u$$

$$\widehat{\beta_1} \sim N\left(\beta_1, \widehat{\sigma}^2_{\widehat{\beta_1}}\right), \quad t = \frac{\widehat{\beta_1} - \beta_1}{se\left(\widehat{\beta_1}\right)} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}} \leq t \leq t_{1-\frac{\alpha}{2}}\right\} = P\left\{-t_{1-\frac{\alpha}{2}} \leq \frac{\widehat{\beta_1} - \beta_1}{se\left(\widehat{\beta_1}\right)} \leq t_{1-\frac{\alpha}{2}}\right\} = 1 - \alpha$$

$$P\left\{\widehat{\beta_1} - t_{1-\frac{\alpha}{2}} se\left(\widehat{\beta_1}\right) \leq \beta_1 \leq \widehat{\beta_1} + t_{1-\frac{\alpha}{2}} se\left(\widehat{\beta_1}\right)\right\} = 1 - \alpha$$

参数异常观测的刻画：拒绝域和接受域

$$y = \beta_0 + \beta_1 x + u$$

$$\widehat{\beta_1} \sim N\left(\beta_1, \widehat{\sigma}^2_{\widehat{\beta_1}}\right), \quad t = \frac{\widehat{\beta_1} - \beta_1}{se\left(\widehat{\beta_1}\right)} \sim t(n-2)$$

$$P\left\{-t_{1-\frac{\alpha}{2}} \leq t \leq t_{1-\frac{\alpha}{2}}\right\} = P\left\{-t_{1-\frac{\alpha}{2}} \leq \frac{\widehat{\beta_1} - \beta_1}{se\left(\widehat{\beta_1}\right)} \leq t_{1-\frac{\alpha}{2}}\right\} = 1 - \alpha$$

$$P\left\{\widehat{\beta_1} - t_{1-\frac{\alpha}{2}} se\left(\widehat{\beta_1}\right) \leq \beta_1 \leq \widehat{\beta_1} + t_{1-\frac{\alpha}{2}} se\left(\widehat{\beta_1}\right)\right\} = 1 - \alpha$$

H_0: 真实的斜率是β_1； H_1: 真实的斜率不是β_1. 接受域为:

$$\left(\widehat{\beta_1} - t_{1-\frac{\alpha}{2}} se\left(\widehat{\beta_1}\right), \widehat{\beta_1} + t_{1-\frac{\alpha}{2}} se\left(\widehat{\beta_1}\right)\right)$$

干扰项方差的假设检验

单侧检验:

$$H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 > \sigma_0^2$$

干扰项方差的假设检验

单侧检验:

$$H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 > \sigma_0^2$$

如果原假设成立，那么

$$\chi^2 = (n-2)\frac{\widehat{\sigma}^2}{\sigma_0^2} \sim \chi^2(n-2).$$

干扰项方差的假设检验

单侧检验:

$$H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 > \sigma_0^2$$

如果原假设成立，那么

$$\chi^2 = (n-2)\frac{\widehat{\sigma}^2}{\sigma_0^2} \sim \chi^2(n-2).$$

支持H_1的证据为

$$\frac{\widehat{\sigma}^2}{\sigma_0^2} > c,$$

干扰项方差的假设检验

单侧检验:

$$H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 > \sigma_0^2$$

如果原假设成立，那么

$$\chi^2 = (n-2)\frac{\widehat{\sigma}^2}{\sigma_0^2} \sim \chi^2(n-2).$$

支持H_1的证据为

$$\frac{\widehat{\sigma}^2}{\sigma_0^2} > c,$$

这个事件出现的概率为$1-\alpha$, 从而我们有拒绝域

$$\left(\widehat{\sigma}^2 : \quad (n-2)\frac{\widehat{\sigma}^2}{\sigma_0^2} > \chi^2_{1-\alpha}\right)$$

干扰项方差的假设检验

单侧检验:

$$H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 > \sigma_0^2$$

如果原假设成立，那么

$$\chi^2 = (n-2)\frac{\widehat{\sigma}^2}{\sigma_0^2} \sim \chi^2(n-2).$$

支持H_1的证据为

$$\frac{\widehat{\sigma}^2}{\sigma_0^2} > c,$$

这个事件出现的概率为$1-\alpha$, 从而我们有拒绝域

$$\left(\widehat{\sigma}^2 : \quad (n-2)\frac{\widehat{\sigma}^2}{\sigma_0^2} > \chi_{1-\alpha}^2\right)$$

也就是

$$\widehat{\sigma}^2 > \frac{\sigma_0^2 \chi_{1-\alpha}^2}{(n-2)}$$

精确的显著性水平：p-value

为了判断一个假设是否显著，选择统计量$T(X_1, X_2, \cdots, X_n)$。观测样本，得到数据$x_1, x_2, \cdots, x_n$,带入统计量，得到观测值$t(x_1, x_2, \cdots, x_n)$.

精确的显著性水平：p-value

为了判断一个假设是否显著，选择统计量$T(X_1, X_2, \cdots, X_n)$。
观测样本，得到数据$x_1, x_2, \cdots, x_n$,带入统计量，得到观测值
$t(x_1, x_2, \cdots, x_n)$.
按照T的分布，计算出概率

$$P\{T > t\} = p$$

p就称为统计量T的p值，它表示在原假设成立时，T统计量超过实际观测值的概率。

精确的显著性水平：p-value

为了判断一个假设是否显著，选择统计量$T(X_1, X_2, \cdots, X_n)$。
观测样本，得到数据$x_1, x_2, \cdots, x_n$,带入统计量，得到观测值
$t(x_1, x_2, \cdots, x_n)$.
按照T的分布，计算出概率

$$P\{T > t\} = p$$

p就称为统计量T的p值，它表示在原假设成立时，T统计量超过实际观测值的概率。
p值大，超过t的事件容易发生，所以原假设成立;
p值小，超过t的事件不容易发生，所以原假设应该被拒绝。

精确的显著性水平：p-value

为了判断一个假设是否显著，选择统计量$T(X_1, X_2, \cdots, X_n)$。
观测样本，得到数据$x_1, x_2, \cdots, x_n$,带入统计量，得到观测值
$t(x_1, x_2, \cdots, x_n)$.
按照T的分布，计算出概率

$$P\{T > t\} = p$$

p就称为统计量T的p值，它表示在原假设成立时，T统计量超过实际观测值的概率。
p值大，超过t的事件容易发生，所以原假设成立;
p值小，超过t的事件不容易发生，所以原假设应该被拒绝。
一般取$0.001, 0.0001$等较小的数作为p值的临界值。

内容线索

回归模型的离差

用相对于中点的离差刻画数据的变异:
因变量的总离差

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

回归直线的离差

$$ESS = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

残差的离差

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

回归模型的离差

用相对于中点的离差刻画数据的变异:
因变量的总离差

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

回归直线的离差

$$ESS = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

残差的离差

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

离差平方和分解公式:

$$TSS = ESS + RSS$$

回归模型拟合优度判定：判定系数

总的离差平方和 = 回归直线的离差平方和 + 残差的离差平方和

回归模型拟合优度判定：判定系数

总的离差平方和 $=$ 回归直线的离差平方和 $+$ 残差的离差平方和

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

回归模型拟合优度判定：判定系数

总的离差平方和 = 回归直线的离差平方和 + 残差的离差平方和

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$r^2 = \frac{ESS}{TSS}$$

称为回归模型的判定系数。
样本相关系数:

$$r = \pm\sqrt{r^2}$$

回归模型显著性判定：方差分析

$$y = \beta_0 + \beta_1 x + u$$

问题：变量x和y之间确实存在线性关系么？

回归模型显著性判定：方差分析

$$y = \beta_0 + \beta_1 x + u$$

问题：变量x和y之间确实存在线性关系么？

$$H_0 : \beta_1 = 0; \quad H_1 : \beta_1 \neq 0.$$

回归模型显著性判定：方差分析

$$y = \beta_0 + \beta_1 x + u$$

问题：变量x和y之间确实存在线性关系么？

$$H_0 : \beta_1 = 0; \quad H_1 : \beta_1 \neq 0.$$

原假设成立时有

$$F = \frac{ESS}{RSS/(n-2)} = \frac{\widehat{\beta_1}^2 \sum \widehat{x_i}^2}{\sum \widehat{u_i}^2/(n-2)} \sim F(1, n-2)$$

一元线性回归方差分析表

变异来源	SS^*	df（自由度）	MSS^*(均方和)
源于回归（ESS）	$\sum \widehat{y}_i^2$	1	$\sum \widehat{y}^2$
源于剩余（RSS）	$\sum \widehat{u}_i^2$	n-2	$\frac{1}{n-2}\sum \widehat{u}_i^2$
TSS	$\sum y_i$	n-1	F

一元线性回归方差分析表

变异来源	SS^*	df（自由度）	MSS^*(均方和)
源于回归（ESS）	$\sum \widehat{y}_i^2$	1	$\sum \widehat{y}^2$
源于剩余（RSS）	$\sum \widehat{u}_i^2$	n-2	$\frac{1}{n-2}\sum \widehat{u}_i^2$
TSS	$\sum y_i$	n-1	F

一元线性回归中，F检验和T检验的作用一样;
多元线性回归中，F检验用于线性结构检验，T检验用于单个系数检验。

回归分析的应用：均值预测

$$Y = \beta_0 + \beta_1 X + U$$

假定$U \sim N(0, \sigma^2)$, 那么$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

$$Y_0 = \beta_0 + \beta_1 X_0 = \mathrm{E}\left[Y | X = X_0\right].$$

回归分析的应用：均值预测

$$Y = \beta_0 + \beta_1 X + U$$

假定$U \sim N(0, \sigma^2)$, 那么$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

$$Y_0 = \beta_0 + \beta_1 X_0 = \mathrm{E}\left[Y | X = X_0\right].$$

$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$是$Y_0 = \beta_0 + \beta_1 X_0$的最优线性无偏估计.

回归分析的应用：均值预测

$$Y = \beta_0 + \beta_1 X + U$$

假定$U \sim N(0, \sigma^2)$, 那么$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

$$Y_0 = \beta_0 + \beta_1 X_0 = \mathrm{E}\left[Y | X = X_0\right].$$

$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$是$Y_0 = \beta_0 + \beta_1 X_0$的最优线性无偏估计.
可以证明,统计量
$\widehat{Y}_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum x_i^2}\right])$

回归分析的应用：均值预测

$$Y = \beta_0 + \beta_1 X + U$$

假定$U \sim N(0, \sigma^2)$, 那么$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

$$Y_0 = \beta_0 + \beta_1 X_0 = \mathrm{E}\left[Y | X = X_0\right].$$

$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$是$Y_0 = \beta_0 + \beta_1 X_0$的最优线性无偏估计.
可以证明,统计量
$\widehat{Y}_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum x_i^2}\right])$
我们有

$$t = \frac{\widehat{Y}_0 - (\beta_0 + \beta_1 x_0)}{se(\widehat{Y}_0)} \sim t(n-2)$$

由此可以得到Y_0的置信区间.

回归结果的评价和分析

用线性模型拟合数据,结果是好,还是不好?

回归结果的评价和分析

用线性模型拟合数据,结果是好,还是不好?
判断的准则:

1. 所估计参数的符号是否与预期的一样;

回归结果的评价和分析

用线性模型拟合数据,结果是好,还是不好?
判断的准则:

1. 所估计参数的符号是否与预期的一样;
2. 统计检验的显著性是否与他人研究的结果一致;

回归结果的评价和分析

用线性模型拟合数据,结果是好,还是不好?
判断的准则:

1. 所估计参数的符号是否与预期的一样;
2. 统计检验的显著性是否与他人研究的结果一致;
3. 从判定系数来看,对变异的解释是否足够明显.

回归结果的评价和分析

用线性模型拟合数据,结果是好,还是不好?
判断的准则:

1. 所估计参数的符号是否与预期的一样;
2. 统计检验的显著性是否与他人研究的结果一致;
3. 从判定系数来看,对变异的解释是否足够明显.

除了上述判断的依据外,我们还需要检验残差序列是否满足正态性假定.

残差序列正态性检验

检验的方法有:

1. 残差直方图分析;
2. 残差正态概率图分析;
3. 雅克-贝拉检验(Jarque-Bera Test)(JB);
4. 安德森-达林检验(Anderson-Darling Normality Test)(A^2).

残差序列正态性检验

检验的方法有:

1. 残差直方图分析;
2. 残差正态概率图分析;
3. 雅克-贝拉检验(Jarque-Bera Test)(JB);
4. 安德森-达林检验(Anderson-Darling Normality Test)(A^2).

JB检验用的较多, 统计量为:

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right]$$

S: 偏态参数;
K:峰态参数.

偏态和峰态的定义

偏态的定义:

$$S = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\sigma^3}$$

偏态和峰态的定义

偏态的定义:

$$S = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\sigma^3}$$

峰态的定义:

$$K = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{\left[\mathrm{E}\left(X-\mu\right)^2\right]^2}$$

偏态和峰态的定义

偏态的定义:

$$S = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\sigma^3}$$

峰态的定义:

$$K = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{\left[\mathrm{E}\left(X-\mu\right)^2\right]^2}$$

1. $K = 3$, 常态, 例如正态分布.
2. $K < 3$, 扁峰, 尾部肥而短.
3. $K > 3$, 尖峰, 尾部细而长.

偏态和峰态的定义

偏态的定义:

$$S = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\sigma^3}$$

峰态的定义:

$$K = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{\left[\mathrm{E}\left(X-\mu\right)^2\right]^2}$$

1. $K = 3$, 常态, 例如正态分布.
2. $K < 3$, 扁峰, 尾部肥而短.
3. $K > 3$, 尖峰, 尾部细而长.

正态分布为常态.

$$H_0 : u_i \sim N(0, \sigma^2)$$

$$JB \xrightarrow{P} \chi^2(2)$$