# 第二章：一元线性回归

马学俊(主讲) 杜悦(助教)

苏州大学

数学科学学院

https://xuejunma.github.io/

# Outline

1. 引言
2. 一元线性回归模型
3. 参数$\beta_0$，$\beta_1$的估计
   - 普通最小二乘估计
   - 最大似然估计
4. 最小二乘估计的性质
5. 回归方程的显著性检验
6. 残差分析
7. 回归系数的区间估计
8. 预测和控制
9. 作业

# 引言

- $F(x) = P(X \le x); -\infty < x < \infty$
- $f_X(x) = f(x) = F'(x)$.
- 
$$
\begin{aligned}
\mathbf{E}(X) &= \sum_x xp(x) \\
\mathbf{E}(X) &= \int_{-\infty}^{\infty} xf(x)dx
\end{aligned}
$$
- $E(X)$ is the "center" of a distribution (or its r,v,) in the sense that

$$
\min_b \mathbf{E}(X-b)^2 = \mathbf{E}[X-\mathbf{E}X]^2.
$$

# Conditional Expectation

- Conditional Expectation of $X$ when $Y$ is given as $y$ is that
  - $\mathbf{E}(X|Y=y) = \sum_x x p_{X|Y}(X|Y)$ for discrete r.v.
  - $\mathbf{E}(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx$ for cont. r.v.
  - Interpretation: Note that $X|Y=y$ is a new r.v., $\mathbf{E}(X|Y=y)$ is the expectation on this r.v.
- Law of Total Expectation [1]

$$\mathbf{E}\Big[\mathbf{E}(X|Y)\Big] = \mathbf{E}(X)$$

- Law of Total Variance

$$Var(X) = Var\Big[\mathbf{E}(X|Y)\Big] + \mathbf{E}\Big[Var(X|Y)\Big]$$

[1] Statistical Inference 2nd Edition by George Casella Roger L. Berger

**Theorem 4.4.3**

If $X$ and $Y$ are any two r.vs, then

$$\mathbf{E}(X) = \mathbf{E}\Big[\mathbf{E}(X|Y)\Big]$$

Proof:

$$\begin{aligned}\mathbf{E}X &= \int\int xf(x,y)dxdy \\ &= \int\Big[\int xf(x|y)dx\Big]f_Y(y)dy \\ &= \int \mathbf{E}(X|y)f_Y(y)dy = \mathbf{E}\Big[\mathbf{E}(X|Y)\Big]\end{aligned}$$

In general, the conditional expectation $\mathbf{E}[X|Y]$ can by defined as a r.v. $g(Y)$ such that

$$\mathbf{E}[(X-g(Y))^2] = \inf_{\text{among all reasonable function } h} \mathbf{E}[(X-h(Y))^2]$$

or $\mathbf{E}[X|Y]$ is the function of $Y$ which is "closest" to $X$ in terms of mean square error.

**Theorem 4.4.7**

For any two random variables $X$ and $Y$,

$$Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$$

provided that the expectation exist.

Proof:

$$\begin{aligned}
Var(X) &= \mathbf{E}\Big\{ [X - \mathbf{E}(X|Y) + \mathbf{E}(X|Y) - \mathbf{E}X]^2 \Big\} \\
&= \mathbf{E}\Big\{ [X - \mathbf{E}(X|Y)]^2 + [\mathbf{E}(X|Y) - \mathbf{E}X]^2 \\
&\quad + 2[X - \mathbf{E}(X|Y)][\mathbf{E}(X|Y) - \mathbf{E}X] \Big\} \\
&= \mathbf{E}\{[X - \mathbf{E}(X|Y)]^2\} + \mathbf{E}\{[\mathbf{E}(X|Y) - \mathbf{E}X]^2\} \\
&= \mathbf{E}[Var(X|Y)] + Var[(\mathbf{E}X|Y)]
\end{aligned}$$

## Theorem 4.4.7

For any two random variables $X$ and $Y$,

$$Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$$

provided that the expectation exist.

Proof:

$$\begin{aligned}
Var(X) &= \mathbf{E}\Big\{ [X - \mathbf{E}(X|Y) + \mathbf{E}(X|Y) - \mathbf{E}X]^2 \Big\} \\
&= \mathbf{E}\Big\{ [X - \mathbf{E}(X|Y)]^2 + [\mathbf{E}(X|Y) - \mathbf{E}X]^2 \\
&\quad + 2[X - \mathbf{E}(X|Y)][\mathbf{E}(X|Y) - \mathbf{E}X] \Big\} \\
&= \mathbf{E}\{[X - \mathbf{E}(X|Y)]^2\} + \mathbf{E}\{[\mathbf{E}(X|Y) - \mathbf{E}X]^2\} \\
&= \mathbf{E}[Var(X|Y)] + Var[(\mathbf{E}X|Y)]
\end{aligned}$$

$$\mathbf{E}\Big\{ 2[X - \mathbf{E}(X|Y)][\mathbf{E}(X|Y) - \mathbf{E}X] \Big\} = \mathbf{E}[\mathbf{E}(Z|Y)]$$

# Real data

- The Current Population Survey (CPS) is a monthly survey of about 57,000 U.S. households conducted by the Bureau of the Census of the Bureau of Labor Statistics [cps09mar] [2].
- `female`: 1 if female, 0 otherwise
- `earnings`: total annual wage and salary earnings
- `hours`: number of hours worked per week
- `week`: number of weeks worked per year

cps09mar

```
> rm(list=ls())
> dat20 <- read.table("cps09mar.txt", head=TRUE, fileEncoding="
> head(dat20)
  age female hisp education earnings hours week union uncov reg
1  52      0    0        12   146000    45   52     0     0
2  38      0    0        18    50000    45   52     0     0
3  38      0    0        14    32000    40   51     0     0
4  41      1    0        13    47000    40   52     0     0
5  42      0    0        13   161525    50   52     1     0
6  66      1    0        13    33000    40   52     0     0
```

[2] https://www.ssc.wisc.edu/~bhansen/econometrics/

- $X = \frac{earnings}{hours*week}$
- $Y = female$
- Homework: **Write R code to check the following equations.**

$$\mathbf{E}\Big[\mathbf{E}(X|Y)\Big] = \mathbf{E}(X)$$

$$Var(X) = Var\Big[\mathbf{E}(X|Y)\Big] + \mathbf{E}\Big[Var(X|Y)\Big]$$

cps09mar

```
y <- dat20$earnings/(dat20$hours*dat20$week)
logy <- log(y)
plot(density(log(y)))
index_men <- which(dat20$female==0)
index_women <- which(dat20$female==1)
logy_mem <- log(y[index_men])
logy_women <- log(y[-index_men])
plot(density(logy_mem), ylim=c(0, 0.8), pch=1)
points(density(logy_women), col="red", lty=2, pch=3)
legend(-8, 0.6, c("man", "woman"), col = c(1,2),  lty = c(1, 2)
         pch = c(1, 3))
```

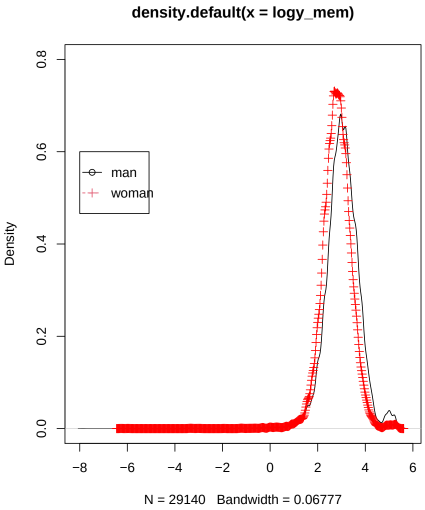


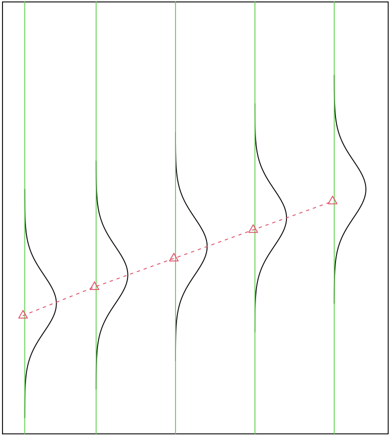

Figure: The density of $Y|female$

**cps09mar**

```
> mean(logy)
[1] 2.946185
> var(logy)
[1] 0.456827
> pp <- c((1-mean(dat20$female)),   mean(dat20$female))
> mean_menwomean <- c( mean(logy_mem), mean(logy_women))
> var_menwomean <- c( var(logy_mem), var(logy_women))
> resu <- data.frame(pp, mean_menwomean, var_menwomean)
> rownames(resu) <- c("men","women")
> resu
             pp mean_menwomean var_menwomean
men   0.5742777       3.045938     0.4956187
women 0.4257223       2.811624     0.3729886
>
```

# 讨论

- 分析female对$Y$的影响
- 如果female 是连续变量



R

```r
y <- seq(-4,4,0.1)
x <- dnorm(y)
plot((x+0.1),y,xlim=c(0,4.5), ylim=c(-4,10),xlab="",
      type="l", ylab="", xaxt="n", yaxt="n")
abline(v=min(x+0.1), col=3)
points((x+1),(y+1),type="l")
abline(v=min(x+1), col=3)
points((x+2),(y+2),type="l")
abline(v=min(x+2), col=3)
points((x+3),(y+3),type="l")
abline(v=min(x+3), col=3)
points((x+4),(y+4),type="l")
abline(v=min(x+4), col=3)
index <- which(x==max(x))
meanx <- x[index] + c(0.1, seq(1:4))
meany <- y[index] + c(0, seq(1:4))
points(meanx-0.42 ,meany-0.42,  pch=2, col=2, type="o", lty=2)
```

# 一元线性回归模型

一元线性回归模型的数学形式为：

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

通常假定：

$$\begin{cases} E(\epsilon) = 0 \\ var(\epsilon) = \sigma^2 \end{cases}$$

对式(1)两端求条件期望，得到回归方程：

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

# 一元线性回归模型经验方程

如果获得n组样本观测值$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，样本模型：

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \qquad i = 1, 2, \ldots, n \tag{2}$$

满足：

$$\begin{cases} E(\epsilon_i) = 0 \\ var(\epsilon_i) = \sigma^2 \end{cases} \quad i = 1, 2, \ldots, n$$

对(2)两端分别求期望和方差，得到：

$$E(y_i) = \beta_0 + \beta_1 x_i, \quad var(y_i) = \sigma^2, \quad i = 1, 2, \ldots, n$$

$E(y_i) = \beta_0 + \beta_1 x_i$从平均意义上表达了变量$y$与$x$的统计规律性。
用$\hat{\beta}_0, \hat{\beta}_1$分别表示$\beta_0, \beta_1$的估计值，获得$y$关于$x$的一元线性经验回归方程

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

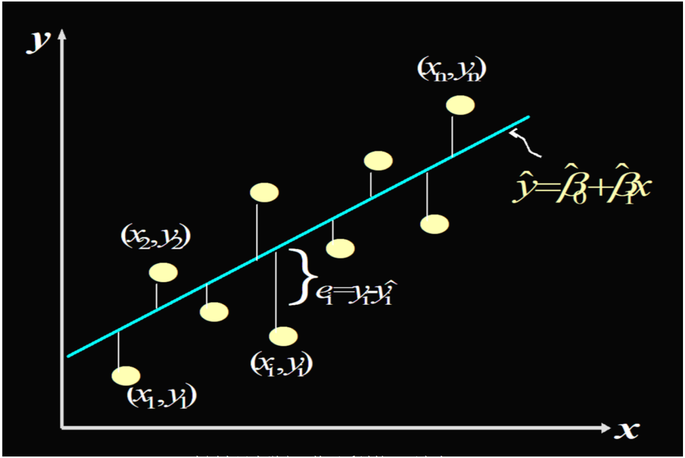
# 普通最小二乘估计

普通最小二乘估计(Ordinary Least Square Estimation,简记为OLSE)就是寻找参数$\beta_0, \beta_1$的估计值使离差平方和达到极小

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \min_{\beta_0, \beta_1} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

- $\hat{\beta}_0, \hat{\beta}_1$称为$\beta_0, \beta_1$的最小二乘估计
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$为$y_i$的回归拟合值，简称回归值或拟合值
- $e_i = y_i - \hat{y}_i$为$y_i$的残差。

# 普通最小二乘估计




从几何关系上看，残差平方和$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$从整体上刻画了$n$个样本观测点$(x_i, y_i), i = 1, \ldots, n$到回归直线$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$距离的长短。

# 普通最小二乘估计

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0}\Big|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1}\Big|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \end{cases}$$

可以得到残差的性质： $\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases}$

整理后得到正规方程组

$$\begin{cases} n\hat{\beta}_0 + (\sum_{i=1}^{n} x_i)\hat{\beta}_1 = \sum_{i=1}^{n} y_i \\ (\sum_{i=1}^{n} x_i)\hat{\beta}_0 + (\sum_{i=1}^{n} x_i^2)\hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

# 普通最小二乘估计

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{cases}$$

其中$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$。
记：

$$L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2$$

$$L_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n(\bar{x}\bar{y})$$

则：

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} \end{cases}$$

# 例2 Homework R code

$$\bar{x} = \frac{49.2}{15} = 3.28, \quad \bar{y} = \frac{396.2}{15} = 26.413$$

$$L_{xx} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 196.16 - 15 \times (3.28)^2 = 34.784$$

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 1470.65 - 1299.536 = 171.114$$

得到：

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 26.413 - 4.919 \times 3.28 = 10.279 \\ \hat{\beta}_1 = L_{xy}/L_{xx} = 171.114/34.784 = 4.919 \end{cases}$$

于是得到回归方程：

$$\hat{y} = 10.275 + 4.919x$$

# 最大似然估计

最大似然估计（maximum likelihood estimation，简记为MLE）是利用总体的分布密度或概率分布的表达式及其样本所提供的信息求未知参数估计量的一种方法。似然函数并不局限于独立同分布的样本。

- 连续型随机变量：似然函数是样本的联合密度函数
- 离散型随机变量：似然函数是样本的联合概率函数

对于一元线性回归模型参数的最大似然估计，如果已经得到样本观测值$(x_i, y_i), i = 1, \ldots, n$，那么在假设$\epsilon_i \sim N(0, \sigma^2)$时，$y_i$服从如下正态分布：

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$y_i$的分布密度为：

$$f_i(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}[y_i - (\beta_0 + \beta_1 x_i)]^2 \right\}$$

$y_1, y_2, \ldots, y_n$的似然函数为：

$$
\begin{aligned}
L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f_i(y_i) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\}
\end{aligned}
$$

取对数似然函数为：

$$
\ln(L) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2
$$

至此与最小二乘原理相同。

$$
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 (\text{有偏估计})
$$

$$
\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 (\text{无偏估计})
$$

# 最小二乘估计的性质

- 线性

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} y_i$$

- 无偏性

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

$$E(\hat{\beta}_1) = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} E(y_i)$$

$$= \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i) = \beta_1$$

其中用到$\sum(x_i - \bar{x}) = 0, \sum(x_i - \bar{x})x_i = \sum(x_i - \bar{x})^2$

## 最小二乘估计的性质

- $\beta_0, \beta_1$的方差

$$var(\hat{\beta}_1) = \sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 var(y_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$var(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\sigma^2$$

$$cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{\bar{x}}{L_{xx}}\sigma^2$$

$$\hat{\beta}_0 \sim N(\beta_0, \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\sigma^2)$$

$$\hat{\beta}_1 \sim N(\beta_0, \frac{\sigma^2}{L_{xx}})$$

# 需要条件

高斯-马尔可夫条件

$$\begin{cases} E(\epsilon_i) = 0, \quad i = 1, 2, \ldots, n \\ cov(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2, i = j \\ 0, i \neq j \end{cases} \quad i, j = 1, 2, \ldots, n \end{cases}$$

# $t$检验

$t$检验用于检验回归系数的显著性，检验的原假设是：

$$H_0: \beta_1 = 0$$

对立假设是：

$$H_1: \beta_1 \neq 0$$

由

$$\hat{\beta}_1 \sim N(\beta_0, \frac{\sigma^2}{L_{xx}})$$

当原假设$H_0: \beta_1 = 0$成立时，有

$$\hat{\beta}_1 \sim N(0, \frac{\sigma^2}{L_{xx}})$$

构造$t$统计量

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}}$$

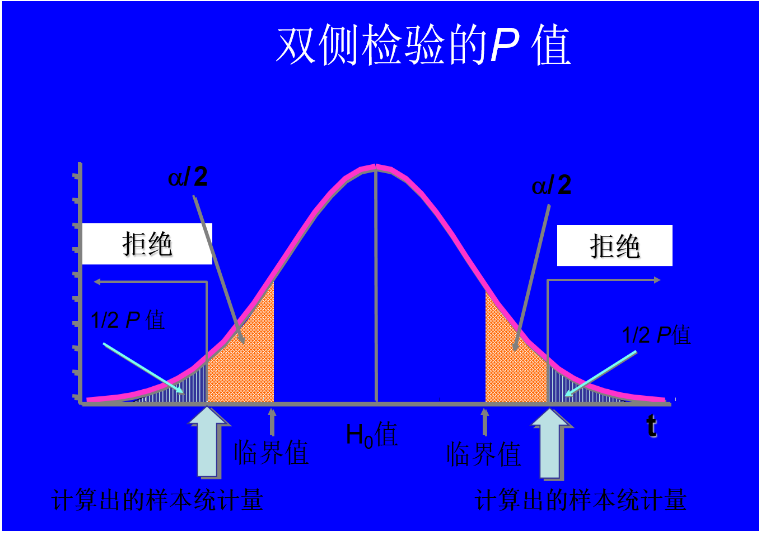
其中

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

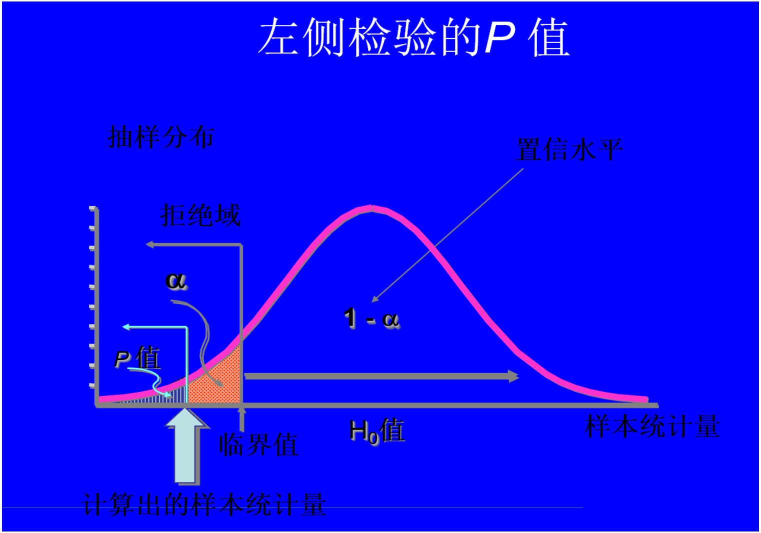
# $P - value$

- $P$ 值即显著性概率值Significence Probability Value
- 是当原假设为真时得到目前的样本以及更极端样本的概率，所谓极端就是与原假设相背离
- 它是用此样本拒绝原假设所犯弃真错误的真实概率，被称为观察到的(或实测的)显著性水平

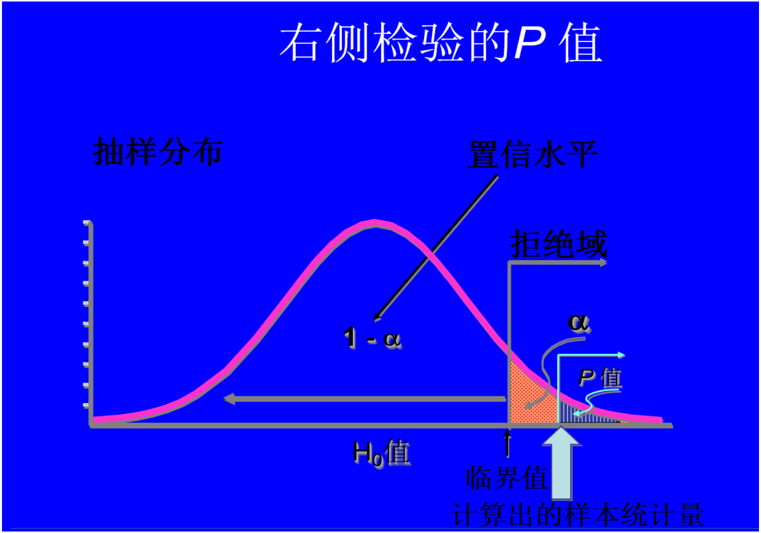
# $P - value$


双侧检验的P 值
α/2
α/2
拒绝
拒绝
1/2 P 值
1/2 P值
临界值
H0值
临界值
t
计算出的样本统计量
计算出的样本统计量

# $P-value$



# $P - value$



# $P-value$

利用$P$值进行检验的决策准则

- 若$p$值$> \alpha$，不能拒绝$H_0$
- 若$p$值$< \alpha$，拒绝$H_0$

双侧检验$p$值$= 2\times$单侧检验$p$值

# $F$检验

$F$检验是根据平方和分解式，直接从回归效果检验回归方程的显著性。
平方和分解式是：

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

$$SST=SSR+SSE$$

总的离差平方和：$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$，
回归平方和：$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，
残差平方和：$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$。

# $F$检验

构造F统计量如下

$$F = \frac{SSR/1}{SSE/(n-2)}$$

Figure: 一元线性回归方差分析表

| 方差来源 | 自由度 | 平方和 | 均方 | F值 | P值 |
|---|---|---|---|---|---|
| 回归 | 1 | $SSR$ | $SSR/1$ | $\frac{SSR/1}{SSE/(n-2)}$ | $P(F>$F值$)$ $=P$值 |
| 残差 | $n$-2 | $SSE$ | $SSE/$（$n$-2） | | |
| 总和 | $n$-1 | $SST$ | | | |

# 相关系数的显著性检验

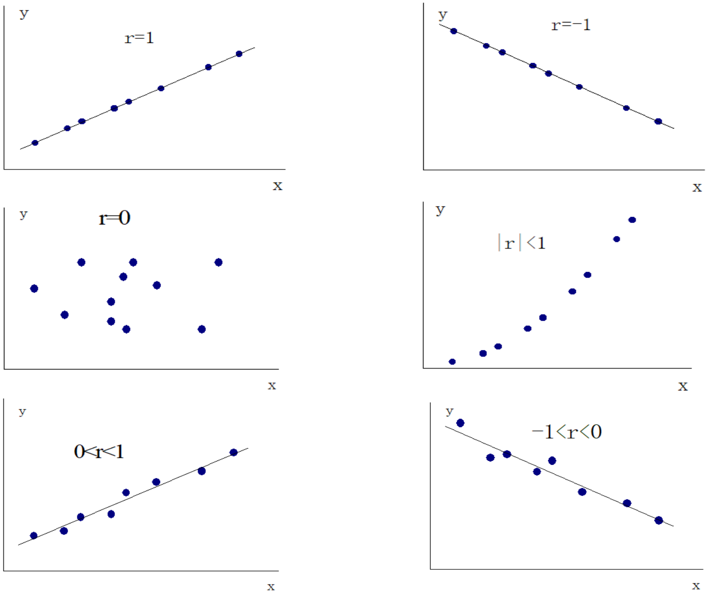
相关系数的显著性检验

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{L_{xx}}{\sqrt{L_{xx}L_{yy}}} = \hat{\beta}_1\sqrt{\frac{L_{xx}}{L_{yy}}}$$

$r$为$x$与$y$的简单相关系数，简称相关系数。

# 相关系数直观意义图



# 相关系数的显著性检验

两变量间相关程度的强弱分为以下几个等级：

- 当$|r| \geq 0.8$时，视为高度相关;
- 当$0.5 \leq |r| \leq 0.8$，视为低度相关;
- 当$0.3 \leq |r| \leq 0.5$时，视为低度相关;
- 当$|r| < 0.3$时，表明两个变量之间的相关程度极弱，在实际应用中可视为不相关。

# 三种检验的关系

对于一元线性回归，这三种检验的结果是完全一致的。

$$H_0: \beta = 0 \qquad t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}}$$

$$H_0: \rho = 0 \qquad t = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}}$$

$$H_0: \text{回归检验} \qquad F = \frac{SSR/1}{SSE/(n-2)}$$

# 决定系数

回归平方和与总离差平方和之比定义为决定系数，也称为判定系数、确定系数。记为$r^2$

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

可以证明

$$r^2 = \frac{SSR}{SST} = \frac{L_{xy}^2}{L_{xx}L_{yy}} = (r)^2$$

决定系数$r^2$是一个反映直线与样本观测拟合优度的相对指标，是因变量的变异中能用自变量解释的比例。其数值在$0 \sim 1$之间，可以用百分数表示。

# 残差概念与残差图

残差：　$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

误差：　$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$

残差项$e_i$是误差项$\epsilon$的估计值。

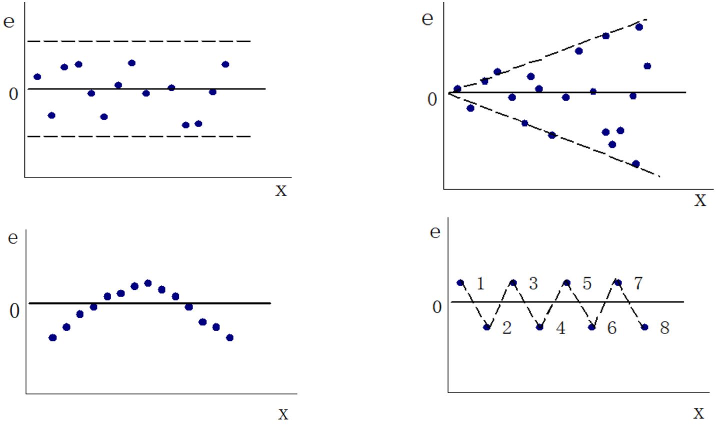



Figure: 残差图

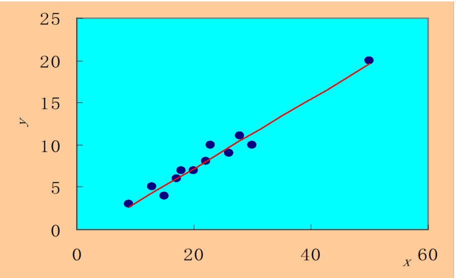
# 残差的性质

- 性质1： $E(e_i)=0$ 证明：

$$E(e_i)=E(y_i)-E(\hat{y}_i)=(\beta_0+\beta_1x_i)-(\hat{\beta}_0+\hat{\beta}_1x_i)=0$$

- 性质2： $var(e_i)=\left[1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{L_{xx}}\right]\sigma^2=(1-h_{ii})\sigma^2$

其中$h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{L_{xx}}$称为杠杆值。




存在高杠杆率观测值的散点图

# 残差的性质

- 性质3：残差满足约束条件

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases}$$

这表明残差$e_1, e_2, \ldots, e_n$的相关的，不是独立的。

# 改进的残差

标准化残差：

$$ZRE_i = \frac{e_i}{\hat{\sigma}}$$

学生化残差：

$$SRE_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

# 回归系数的区间估计

由$\hat{\beta}_1 \sim N(\beta_0, \frac{\sigma^2}{L_{xx}})$可得

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{L_{xx}}}{\hat{\sigma}}$$

服从自由度为$n-2$的$t$分布

$$P(|\frac{(\hat{\beta}_1 - \beta_1)\sqrt{L_{xx}}}{\hat{\sigma}}| < t_{\alpha/2}(n-2)) = 1 - \alpha$$

上式等价于

$$P(\hat{\beta}_1 - t_{\alpha/2}\frac{\hat{\sigma}}{L_{xx}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2}\frac{\hat{\sigma}}{L_{xx}}) = 1 - \alpha$$

即得到$\beta_1$的置信度为$1-\alpha$的置信区间为：

$$(\hat{\beta}_1 - t_{\alpha/2}\frac{\hat{\sigma}}{L_{xx}}, \hat{\beta}_1 + t_{\alpha/2}\frac{\hat{\sigma}}{L_{xx}})$$

# 单值预测

单值预测就是用单个值作为因变量新值的预测值。

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$E(\hat{y}_0) = E(y) = \beta_0 + \beta_1 x_0$$

# 区间预测

区间预测就是对于给定的显著水平$\alpha$，找到一个区间$(T_1, T_2)$，使对应于某特定的$x_0$的实际值$y_0$以$1-\alpha$的概率被区间$(T_1, T_2)$包含，用公示表示：

$$P(T_1 < y_0 < T_2) = 1-\alpha$$

一、因变量新值的区间预测
首先要给出估计值$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$的分布，在正态性假设下$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$服从正态分布，其期望为$E(\hat{y}_0) = \beta_0 + \beta_1 x_0$，计算其方差

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 = \sum_{i=1}^{n} [\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}}] y_i$$

$$var(\hat{y}_0) = \sum_{i=1}^{n} [\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}}] var(y_i) = [\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}] \sigma^2$$

从而得

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

# 因变量新值的区间预测

记$h_{00} = \frac{1}{n} + \frac{(x_0-\bar{x})^2}{L_{xx}}$

$$\hat{y}_0 \sim N(\beta_0 + \beta_1 x_0, h_{00}\sigma^2)$$

$y_0$与$\hat{y}_0$独立

$$var(y_0 - \hat{y}_0) = var(y_0) + var(\hat{y}_0) = \sigma^2 + h_{00}\sigma^2$$

于是

$$y_0 - \hat{y}_0 \sim N(0, (1 + h_{00})\sigma^2)$$

进而可知统计量

$$t = \frac{y_0 - \hat{y}_0}{\sqrt{1 + h_{00}}\hat{\sigma}} \sim t(n-2)$$

# 因变量新值的区间预测

$$P\left(\left|\frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\hat{\sigma}}\right| \le t_{\alpha/2}(n-2)\right)=1-\alpha$$

由此可以求得$y_0$的置信度为$1-\alpha$的置信区间为：

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{1+h_{00}}\hat{\sigma}$$

当样本量n较大，$|x_0-\bar{x}|$较小时，$h_{00}$接近0，$y_0$的置信度为95%的置信区间近似为：

$$\hat{y}_0 \pm 2\hat{\sigma}$$

# 因变量新值的平均值的区间预测

由于$E(y_0) = \beta_0 + \beta_1 x_0$是常数，

$$\hat{y}_0 - E(y_0) \sim N(\beta_0 + \beta_1 x_0, (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}})\sigma^2)$$

可得置信水平为$1 - \alpha$的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{h_{00}}\hat{\sigma}$$

# 控制问题

控制问题相当于预测问题的反问题。给定$y$的预期范围$(T_1, T_2)$，如何控制自变量$x$的值才能以$1-\alpha$的概率保证

$$P(T_1 < y < T_2) = 1-\alpha$$

通常用近似的预测区间来确定$x$

$$\begin{cases} \hat{y}(x) - 2\hat{\sigma} > T_1 \\ \hat{y}(x) + 2\hat{\sigma} < T_2 \end{cases}$$

把$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ 代入求得
当$\hat{\beta}_1 > 0$时

$$\frac{T_1 + 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \frac{T_2 - 2\hat{\sigma} - b\hat{e}ta_0}{\hat{\beta}_1}$$

当$\hat{\beta}_1 < 0$时

$$\frac{T_2 - 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \frac{T_1 + 2\hat{\sigma} - b\hat{e}ta_0}{\hat{\beta}_1}$$

# 作业

2 p.50 2.8, 2.11

3 p.51 编程计算