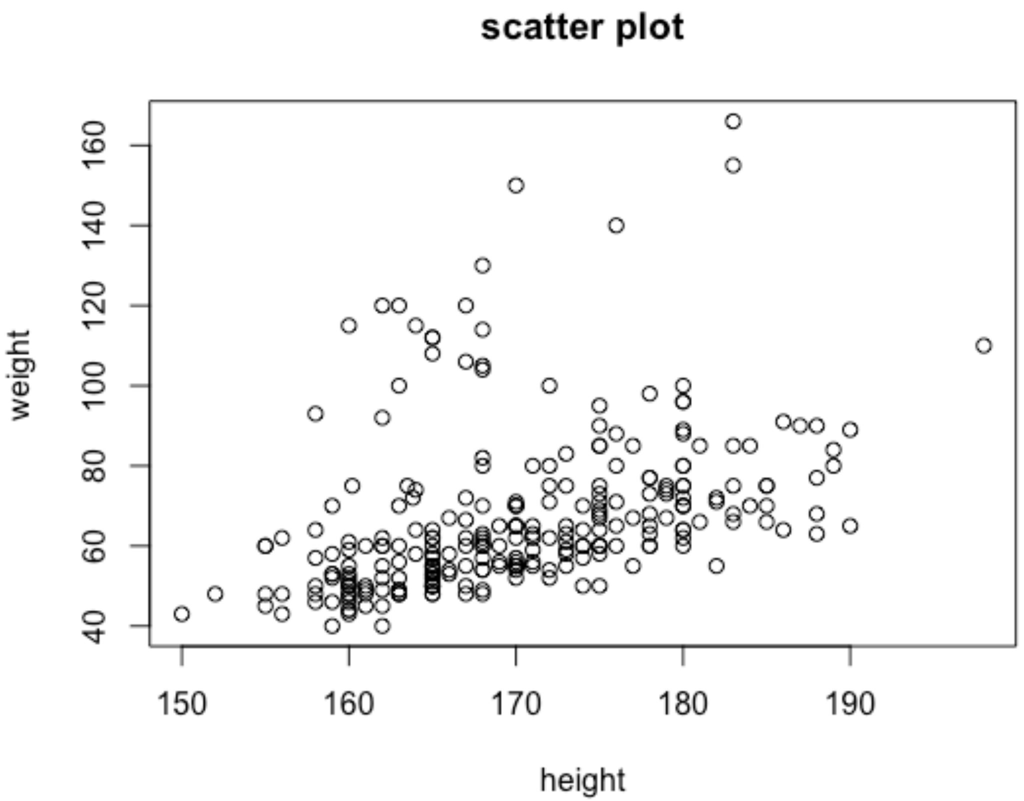
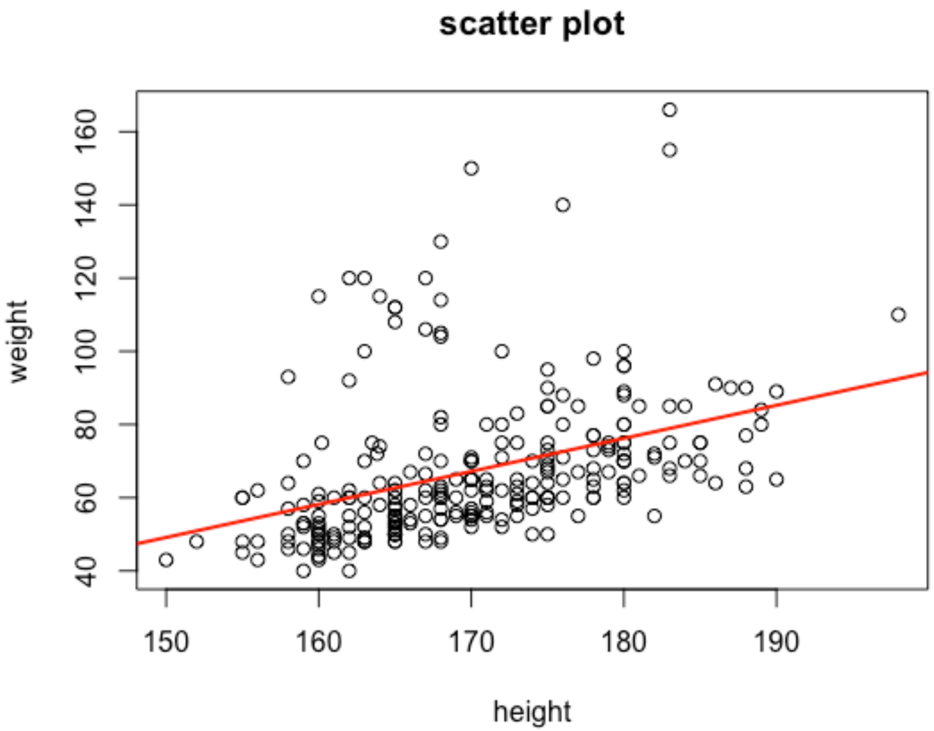
回归分析

简单线性回归

简单线性回归


scatter plot
weight
height
150
160
170
180
190
40
60
80
100
120
140
160


scatter plot
weight
height
150
160
170
180
190
40
60
80
100
120
140
160

研究中国人 身高和体重是否呈现 线性关系：

如果是能否建出 回归模型 体重为Y变量，身高是X变量

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

根据建立的回归模型 给定一个中国人身高 预测出他的体重

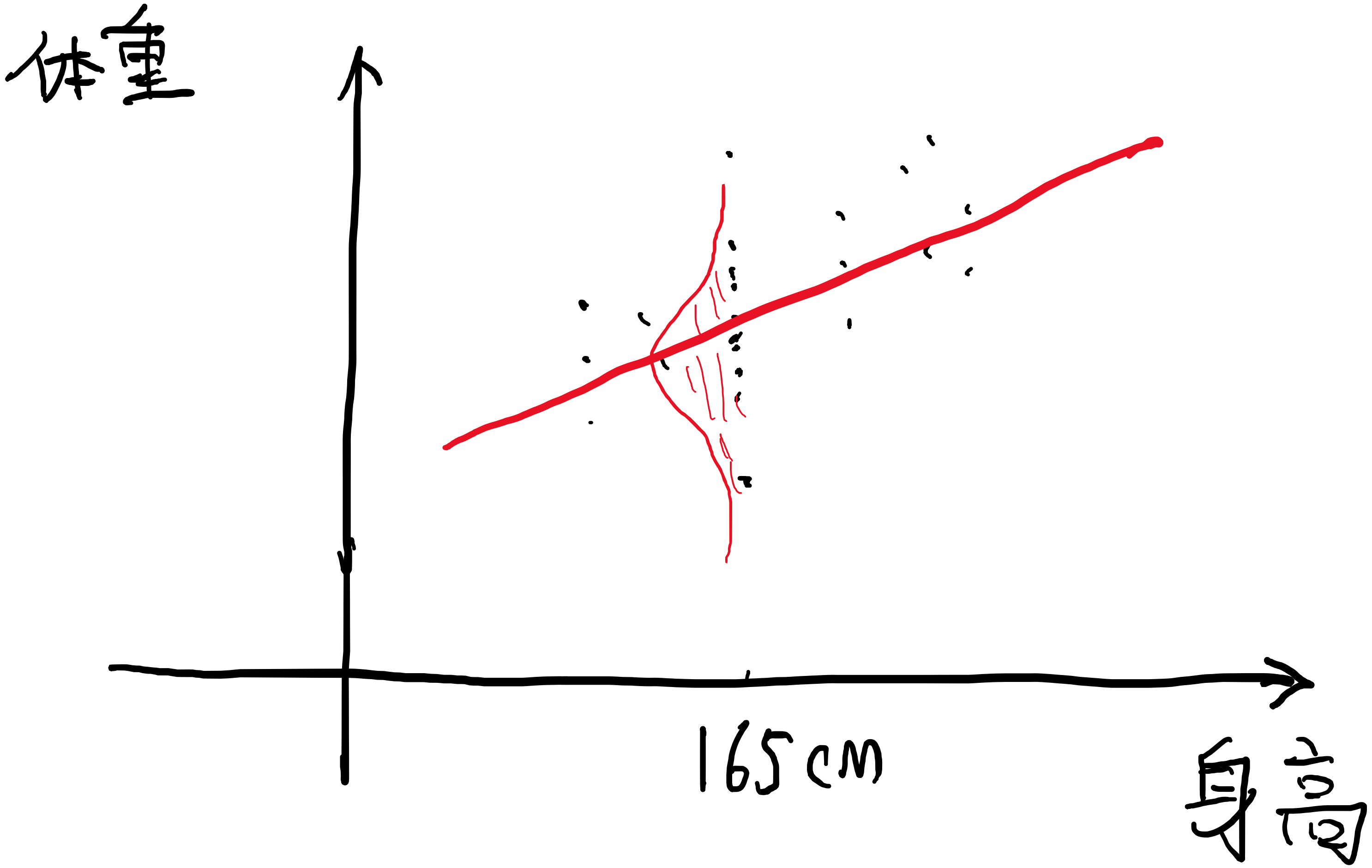
研究中国人 身高和体重是否呈现 线性关系：

如果是能否建出 回归模型 体重为Y变量，身高是X变量

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

$$\varepsilon \sim Normal\,(0, \sigma^2)$$
$$cor(\varepsilon_i, \varepsilon_j) = 0$$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\varepsilon}$$

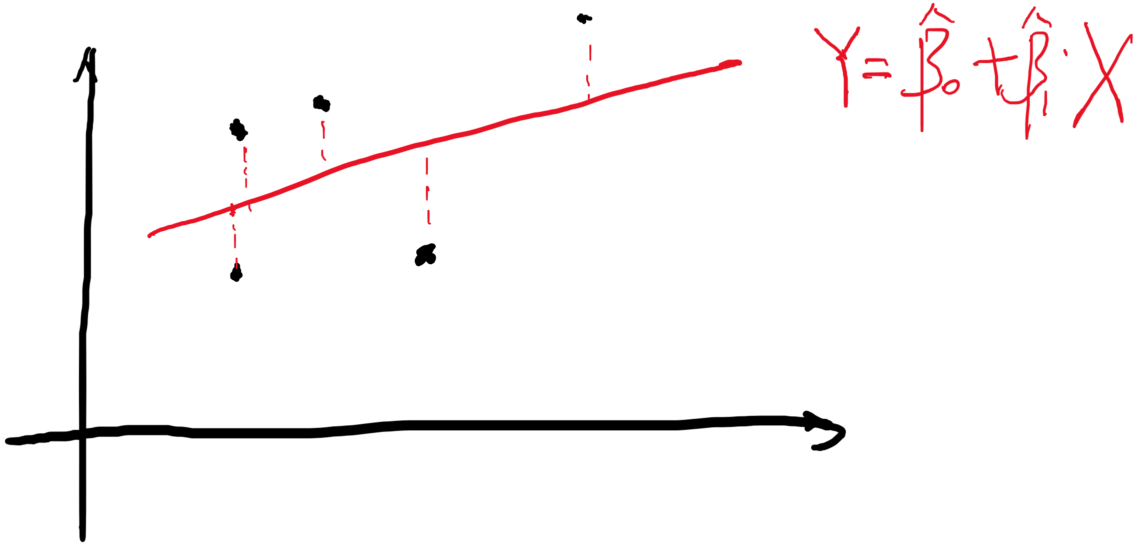

体重
165cm
身高

研究中国人 身高和体重是否呈现 线性关系：

如果是能否建出 回归模型 体重为Y变量，身高是X变量

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

残差：



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

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

残差平方和（RSS）

residual sum of squares (RSS)

SSE= $\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$

or equivalently as

SSE= $\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$

随机误差项：

$\varepsilon_i \sim N(0, \sigma^2)$

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

最小二乘法估计：残差平方和（RSS）最小

residual sum of squares (RSS)

SSE= $\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$

or equivalently as

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname{argmin} (\mathrm{RSS})$$

最小二乘法估计：残差平方和（RSS）最小

residual sum of squares (RSS)

SSE= $$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\mathrm{RSS} = (y_1-\hat{\beta}_0-\hat{\beta}_1 x_1)^2+(y_2-\hat{\beta}_0-\hat{\beta}_1 x_2)^2+\ldots+(y_n-\hat{\beta}_0-\hat{\beta}_1 x_n)^2.$$

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

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

$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]$$

where $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

where $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

如果原假设, $\beta_1 = 0$ 那么 $Y = \beta_0 + \epsilon,$

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

区间估计：

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

p值法估计：

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

区间估计：

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

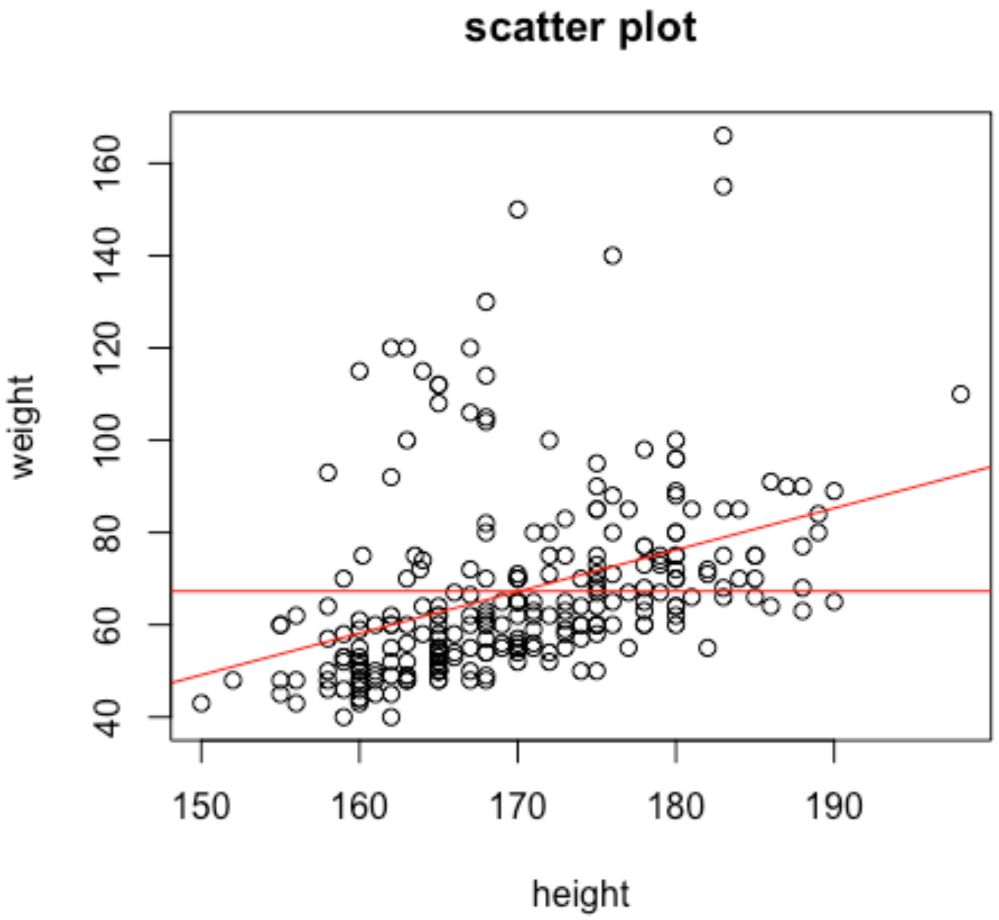
p值法估计：

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

- This will have a t-distribution with $n-2$ degrees of freedom, assuming $\beta_1 = 0$.

F 检验法

$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

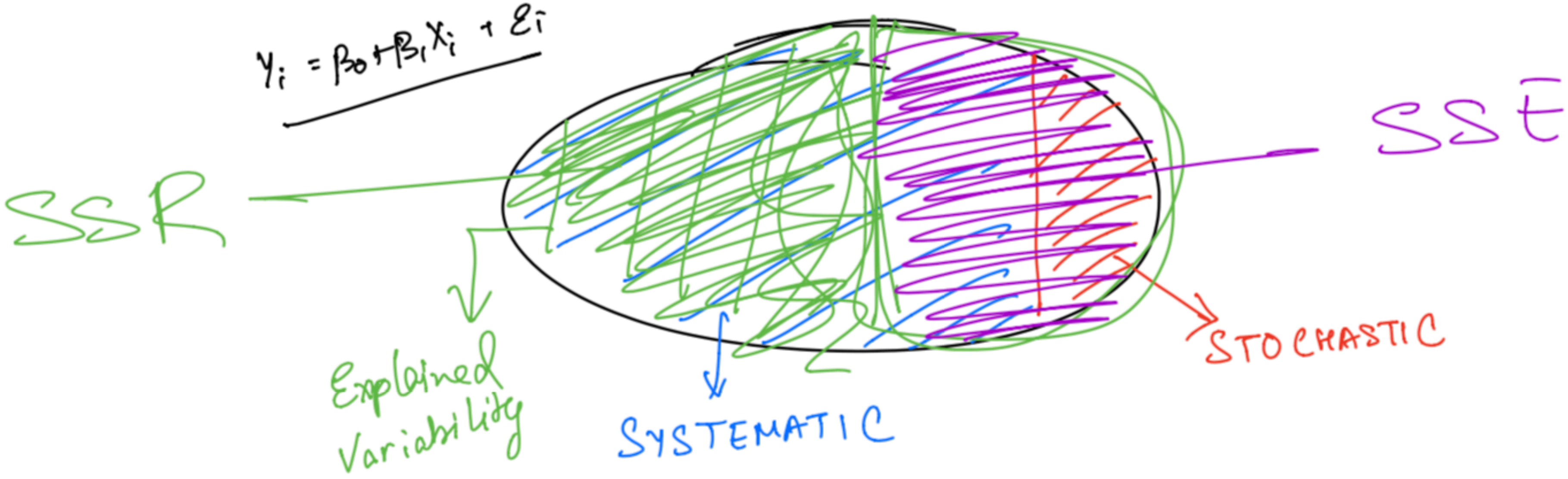


$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

SST= $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

有俩部分组成，分别是 随机部分 RSS(Residual sum square) = SSE(Sum square error) = $\sum_{i=1}^{n} (\widehat{y}_i - y_i)^2$

被解释部分SSR，在简单线性回归 SSR = $\sum_{i=1}^{n} (\widehat{y}_i - \bar{y}_i)^2$



Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS)	F
SSR	1	$MSR=\frac{SSR}{1}$	$F=\frac{MSR}{MSE}$
SSE	n - 2	$MSE=\frac{SSE}{n-2}$	
SST	n - 1		

根据科克**伦**定律：

$$\frac{SSR}{1\sigma^2} = \frac{MSR}{\sigma^2} \rightarrow \chi^2\text{（df=1）}$$

$$\frac{SSE}{(n-2)\sigma^2} = \frac{MSE}{\sigma^2} \rightarrow \chi^2\text{（df=n-2）}$$

$$F^* = \frac{MSR}{MSE} \sim F \text{ distribution} \left(\text{n-df}, \text{d-df}\right)$$

F 检验法：

$$H_0 : \beta_1 = 0$$

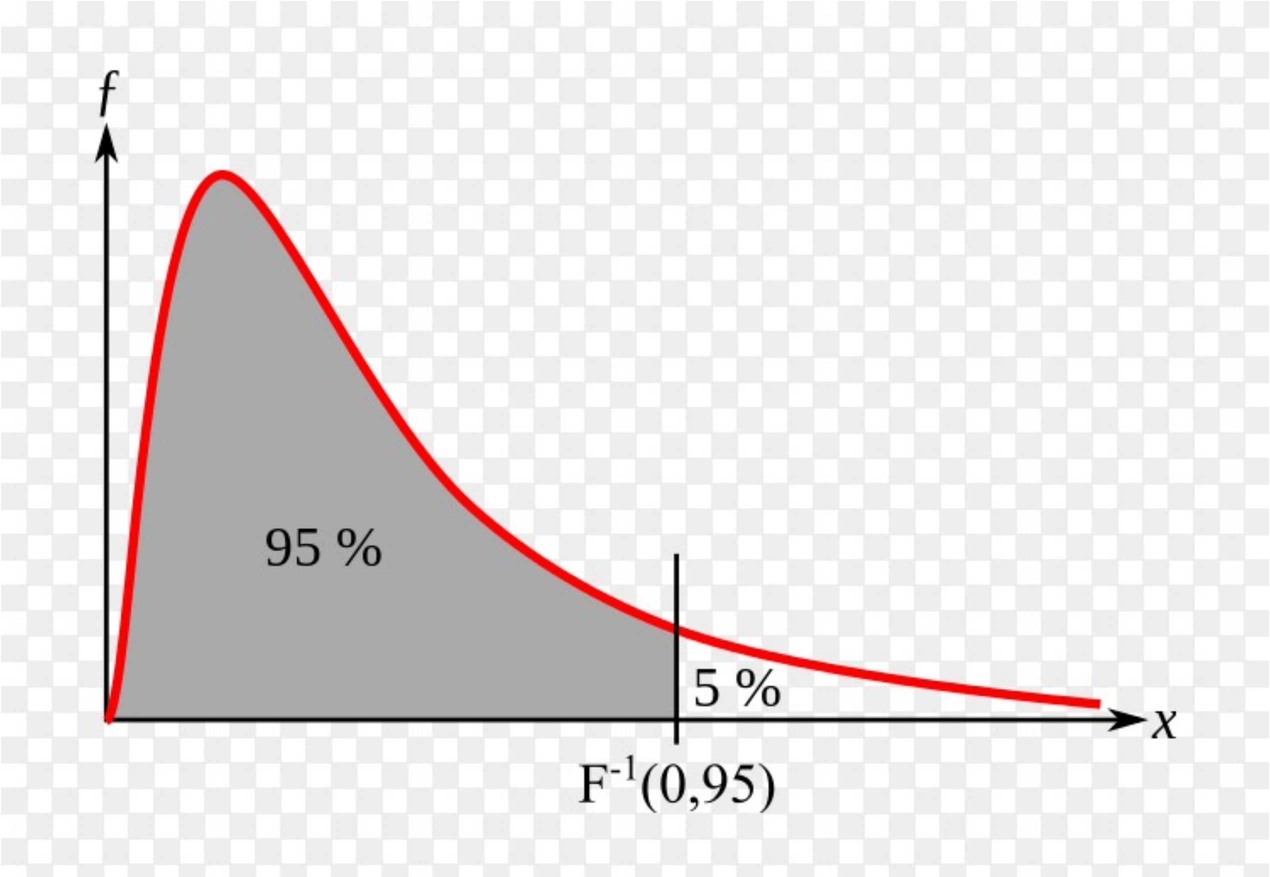
versus

$$H_A : \beta_1 \neq 0,$$

$$\frac{SSR}{1\,\sigma^2} = \frac{MSR}{\sigma^2} \rightarrow \chi^2$$

$$\frac{SSE}{(n-2)\,\sigma^2} = \frac{MSE}{\sigma^2} \rightarrow \chi^2$$

$$F^* = \frac{MSR}{MSE} \sim F \text{ distribution} \begin{pmatrix} n\text{-df}, \\ d\text{-df} \end{pmatrix}$$



研究中国人 身高和体重是否呈现 线性关系：

如果是能否建出 回归模型 体重为Y变量，身高是X变量

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

根据建立的回归模型 给定一个中国人身高 预测出他的体重

研究中国人 身高和体重是否呈现 线性关系：

如果是能否建出 回归模型 体重为Y变量，身高是X变量

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\varepsilon}$$

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

总体中国人身高体重数据 建立的回归模型

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\varepsilon}$$

262个样本建出的回归模型

点估计：

区间估计：

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]$$

where $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$

95%的把握 $\beta_i \in (\widehat{\beta}_i \pm t * SE(\widehat{\beta}_i))$

点估计预测： 中国人身高为165cm的 体重：

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

区间估计预测：对于165cm的人群 给出95%的置信区间、预测区间

$$\hat{y}_0 \pm t_{df=(n-2)}(1-\alpha/2) SE(\hat{y}_0)$$

1. $SE(\hat{y}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$ 置信区间
2. $SE(\hat{y}_0) = \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$ 预测区间