回归分析

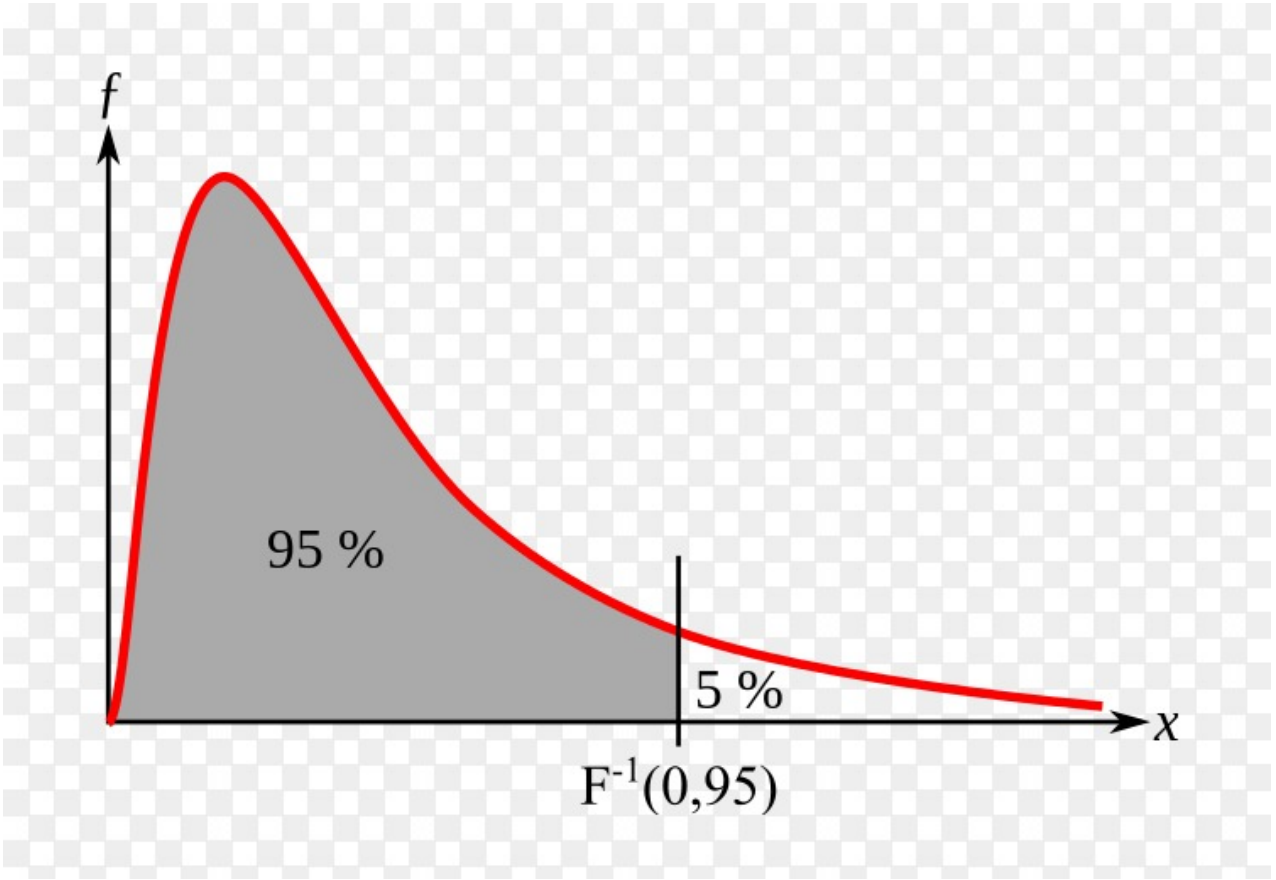
多元线性回归

F 检验法：

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

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

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



H_0：房间数对于log(价格)不显著。 $<=>\beta_1 = 0$

H_a：房间数对于log(价格)显著。 $<=>\beta_1 \neq 0$

H_0：在房间数变量存在情况下，浴室数对于log(价格)不显著。

H_a：在房间数变量存在情况下，浴室数对于log(价格)显著。

H_0：在房间和浴室变量存在情况下，平米对于log(价格)不显著。

H_a：在房间和浴室变量存在情况下，平米对于log(价格)显著。

总体F检验

H_0：房间数，厕所数，平米对于log(价格)都不显著。
H_a：房间数，厕所数，平米对于log(价格)至少有一个变量显著。

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
H_a：β_1，β_2，β_3至少有一个不为0

预测

$$
\begin{aligned}
Var[\hat{Y}|\boldsymbol{x}_0] &= Var[\boldsymbol{x}_0'\hat{\beta}] \\
&= \boldsymbol{x}_0' \; Var[\hat{\beta}] \; \boldsymbol{x}_0 \\
&= \sigma^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0
\end{aligned}
$$

- Suppose we want a $100(1-\alpha)\%$ confidence interval for the predicted mean value of log *price* ($E[Y|\boldsymbol{x}_0]$):

$$
\hat{Y}|\boldsymbol{x}_0 \pm t_{df=n-p}(1-\alpha/2) \; s\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}
$$

处理多重共线性

- pearson 相关系数只能**观测单**个**变量间**的相关性

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu X)(Y-\mu Y))}{\sigma_X \sigma_Y}$$

```
> cor(multi.data)
                price  bedrooms bathrooms sqft.living
price       1.0000000 0.3095577 0.5121808   0.6803845
bedrooms    0.3095577 1.0000000 0.5260057   0.5941190
bathrooms   0.5121808 0.5260057 1.0000000   0.7585961
sqft.living 0.6803845 0.5941190 0.7585961   1.0000000
```

处理多重共线性
方差膨胀系数(variance inflation factor)

- 想要判断多个变量对单个变量间的相关性

不要Y，只在X变量间跑回归（辅助回归）

- 比如想要研究 bedrooms、bathrooms 俩个变量和sqft.living是否相关

1. model =lm(sqft.living ~ bedrooms+bathrooms)
2. $\mathrm{VIF} = \frac{1}{1-model\$R^2}$
3. 如果大于10就认为 bedrooms、bathrooms 和sqft.living相关
4. 就要去掉sqft.living变量避免多重共线性。

处理多重共线性
方差膨胀系数(variance inflation factor)

- 想要判断多个变量对单个变量间的相关性

不要Y，只在X变量间跑回归（辅助回归）

- 比如想要研究 bedrooms、bathrooms 俩个变量和sqft.living是否相关

1. model =lm(sqft.living ~ bedrooms+bathrooms)
2. $\mathrm{VIF} = \frac{1}{1-model\$R^2}$
3. 如果大于10就认为 bedrooms、bathrooms 和sqft.living相关
4. 就要去掉sqft.living变量避免多重共线性。