重点：线性回归

2022年3月13日 星期日 21:49

① $r=\frac{Cov(x,y)}{\sqrt{S_x^2 S_y^2}}$ (相关系数)

② $b_1=r\frac{S_y}{S_x}$

③ R^2：决定系数，越高越好
(回归关系可以解释因变量多少变异)

correlation analysis
连续变量
两组均为正态
等方差
- 符合 → Pearson
- 不符 (non-parametric) → Spearman

cor(x,y,method="pearson")
cor(x,y,method="spearman")
默认method='pearson'

① 代码过程 (一元)
先 cor(x,y) 看相关系数
fit ← lm(y~x, data=) → 看 Adjusted R-squared 大小 (higher is better)
residuals(fit) 看残差
predict(fit, data.frame(x=c(1,2,3…)))
※ summary(fit) 看结果

② 代码 (二元)
fit2 ← lm(data, formula = y~x+I(x^2))

③ 代码 (非线性)
Ⓐ fit3 ← nls(data, y~exp(r*x), start = list(r=), alg = "plinear")
Ⓑ 转化为 log10

回归诊断
① residuals are normally distributed
plot(fit,2) Q-Q图
② errors are independent
plot(fit,1) → 残差点与预测值毫不相干
③ 同方差 homoscedasticity (同②)
残差不随预测值↑而↑
④ hist(resid(model))
看残差正态性

画图
Ⓐ geom_smooth(method='lm', formula=y~x, se=F/T)
geom_smooth(method='gam', formula=y~x+I(x^2), se=F/T)

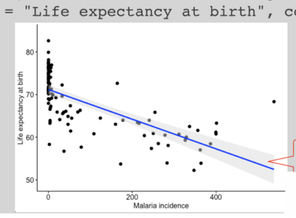
ggpmisc 中 stat_poly_eq，在图上标表达式

```
stat_poly_eq(aes(label = paste(..eq.label.., ..adj.rr.label.., sep ='~~~~')),
             color = "red",
             label.x = 0.25,
             geom="label",
             formula = y~x+I(x^2),
             parse = T)
```

Ⓑ abline(fit) 一元
※ lines(x, fitted(fit2)) 二元
(x轴) (y轴)

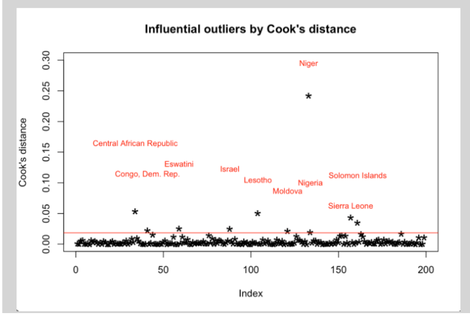
Ⓒ ggpubr 中 ggscatter

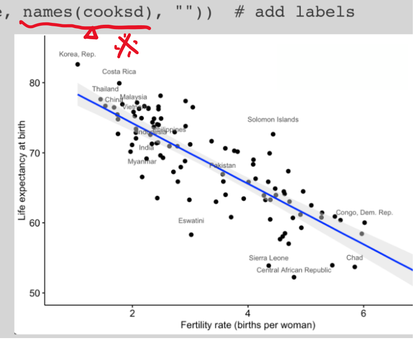
```
>ggpubr::ggscatter(WBd, x = "malaria_incidence", y = "lifeexpectancy", add = "reg.line",
                   add.params = list(color = "blue", fill = "lightgray"), conf.int = TRUE) +
  labs(x = "Malaria incidence", y = "Life expectancy at birth", colour = "")
```



Ⓐ 画 outliers 图

```
>cooksd <- cooks.distance(lm(lifeexpectancy ~ fertility, WBd))
>sample_size <- nrow(WBd)
# Not NAs
>not_nas <- which(!is.na(WBd$lifeexpectancy) & !is.na(WBd$fertility))
# Plot Cook's distance
>plot(cooksd, pch = "*", cex = 2, main = "Influential outliers by Cook's distance",
      ylim=c(0,1), ylab = "Cook's distance")
>abline(h = 4/sample_size, col = "red")  # add cutoff line
>text(x = 1:length(cooksd), y = cooksd + 0.05, col = "red", cex = 0.8,
      labels = ifelse(cooksd > 4 / sample_size, names(cooksd), ""))  # add labels
```





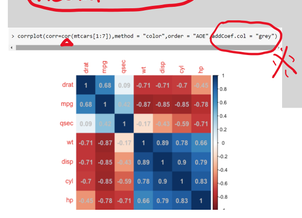
Ⓑ corrplot 包中 corrplot 函数

the graph is added to an existing plot

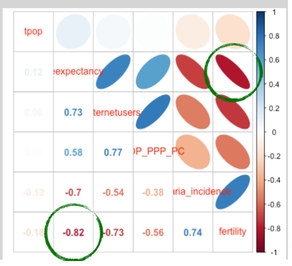
```
>corrplot::corrplot(cor(WBd[withNAs == 0, 3:8]), order = "AOE",
  method = "ellipse", type = "upper", tl.pos = "d")→文本标签在对角线
>corrplot::corrplot(cor(WBd[withNAs == 0, 3:8]), order = "AOE",
  add = TRUE, method = "number", type = "lower", diag = FALSE,
  tl.pos = "n", cl.pos = "n")
```

tl.pos = "n" → null
cl.pos = "n" → 图例位置

method = "color"



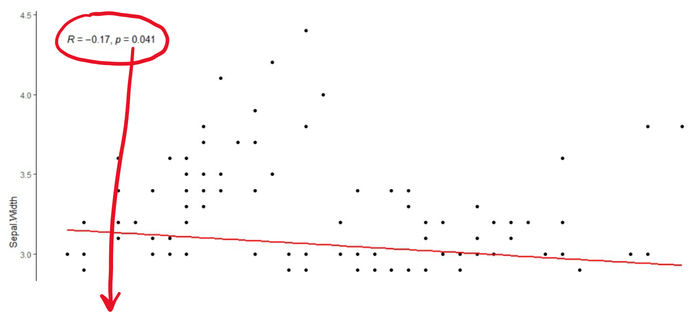
※



Ⓒ 称相关系数 + p值

```
stat_cor(data = iris,
         method = "spearman")
```

"pearson"



讨论两变量是否相关必须讨论显著性水平，不谈P值🙅谈相关系数大小是无意义的，两者之间的相关关系可能只是偶然因素引起的，所以我们要对两个变量之间的相关关系的显著性水平进行判断；

采用假设检验的方法：
原假设H0： R=0 两变量之间不存在线性关联
备择假设H1： R不等于0，两变量之间存在线性关联