多元-04-2020270026

2020270026 王姿文

3/22/2021

1.

已知(X,Y)是二随机变量，$(x_i,y_i),i=1,\ldots,n$是其样本，$S=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$，下证$E(S)=Cov(X,Y)$:

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

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

$=\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\bar{y}-\sum_{i=1}^{n}y_i\bar{x}+n\bar{x}\bar{y}$

$=\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}-\frac{\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i}{n}+\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}$

$=\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}$

$(n-1)E(S)=E(\sum_{i=1}^{n}X_iY_i)--\frac{1}{n}E(\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i)$

$=nE(XY)-\frac{1}{n}[nE(XY)+n(n-1)E(X)E(Y)]$

$=(n-1)[E(XY)-E(X)E(Y)]$

$=(n-1)Cov(X,Y)$

$\Rightarrow E(S)=Cov(X,Y)$

2.

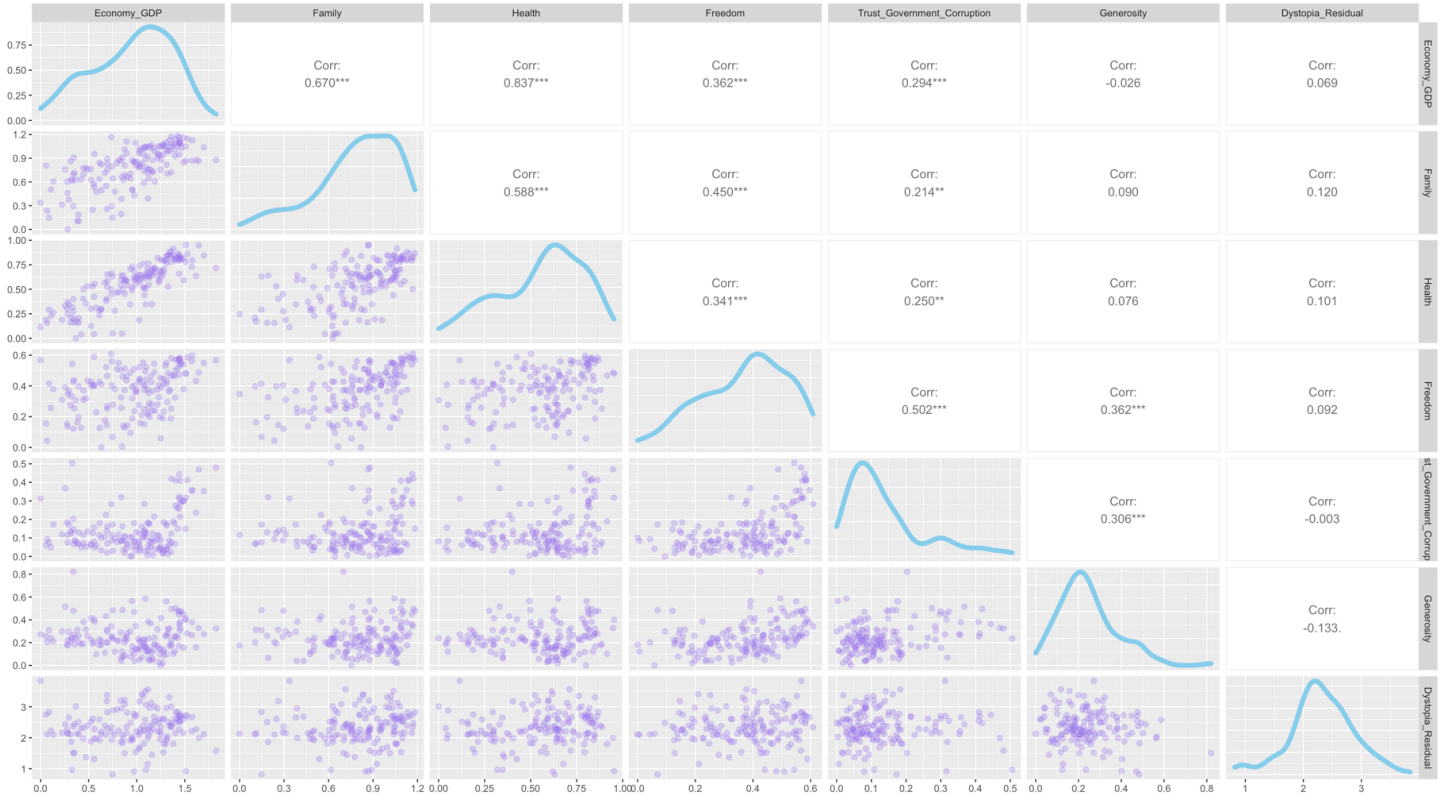
数据来自Kaggle:World Happiness Report (https://www.kaggle.com/unsdsn/world-happiness)，描述不同国家的幸福指数，此处任意挑选2016的数据来绘制简单的探索性资料分析。

```
happy <- read_csv("archive/2016.csv")
happy <- happy %>%
  rename('Happiness_Rank' = 'Happiness Rank',
         'Happiness_Score'='Happiness Score',
         'Lower_Confidence_Interval'='Lower Confidence Interval',
         'Upper_Confidence_Interval'='Upper Confidence Interval',
         'Economy_GDP'='Economy (GDP per Capita)',
         'Health'='Health (Life Expectancy)',
         'Trust_Government_Corruption' = 'Trust (Government Corruption)',
         'Dystopia_Residual'='Dystopia Residual')
dt <- head(happy)
kbl(dt) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size = 7)
```

Country	Region	Happiness_Rank	Happiness_Score	Lower_Confidence_Interval	Upper_Confidence_Interval	Economy_GDP	Family	Health	Freedom	Trust_Government_Corruption	Generosity	Dystopia_Residual
Denmark	Western Europe	1	7.526	7.460	7.592	1.44178	1.16374	0.79504	0.57941	0.44453	0.36171	2.73939
Switzerland	Western Europe	2	7.509	7.428	7.590	1.52733	1.14524	0.86303	0.58557	0.41203	0.28083	2.69463
Iceland	Western Europe	3	7.501	7.333	7.669	1.42666	1.18326	0.86733	0.56624	0.14975	0.47678	2.83137
Norway	Western Europe	4	7.498	7.421	7.575	1.57744	1.12690	0.79579	0.59609	0.35776	0.37895	2.66465
Finland	Western Europe	5	7.413	7.351	7.475	1.40598	1.13464	0.81091	0.57104	0.41004	0.25492	2.82596
Canada	North America	6	7.404	7.335	7.473	1.44015	1.09610	0.82760	0.57370	0.31329	0.44834	2.70485

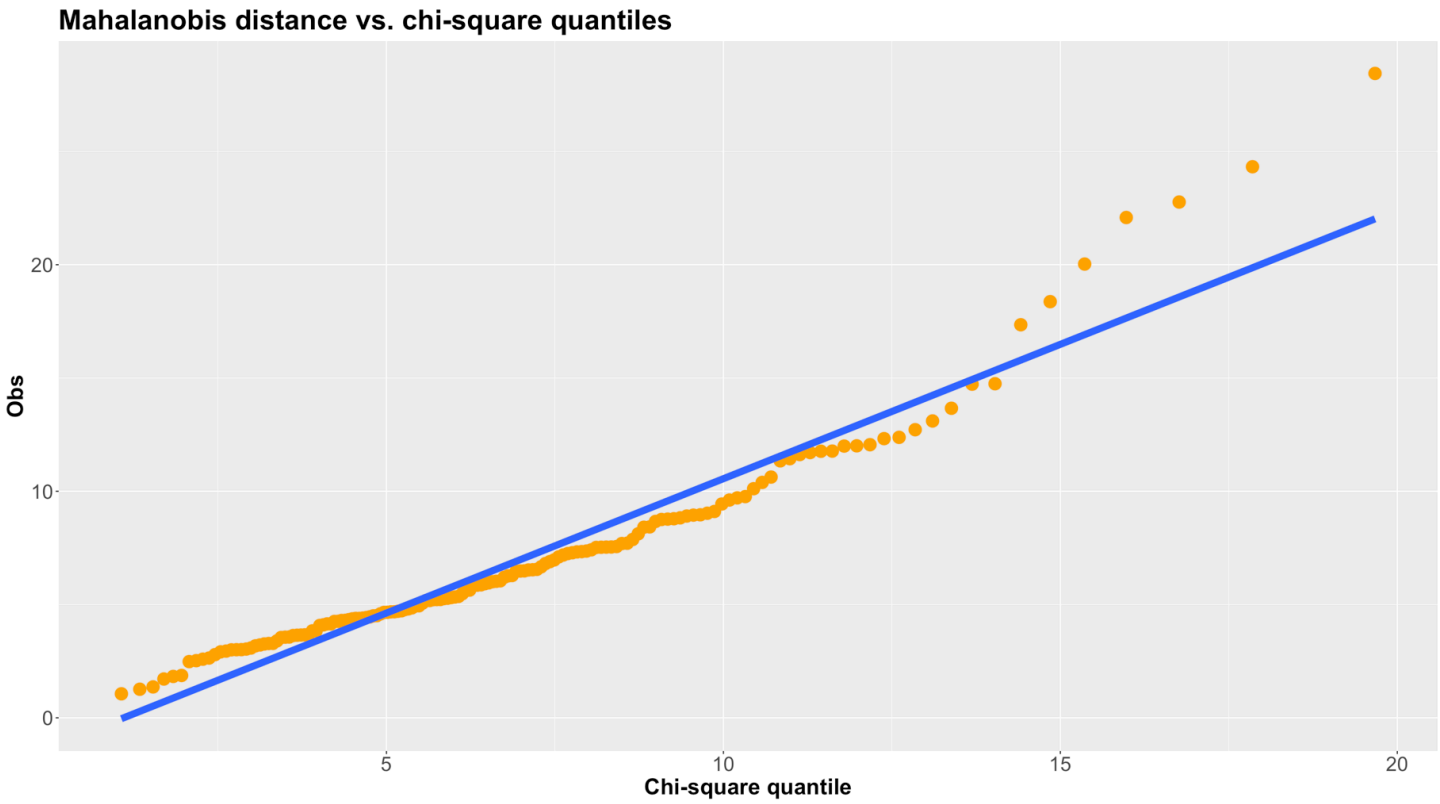
取其中 `Economy_GDP ~ Dystopia_Residual` 的连续型变量，一共七个变量，下图为七个变量间的**scatter plot**和**density plot**。
可以看出分布虽有些偏态，但都是正态的形状。

```
happy_con <- select_if(happy, is.numeric)

d <- as.matrix(happy_con[,5:11])
n <- nrow(d)
p <- ncol(d)

wrap_1<-wrap(ggally_points,size=2,color="mediumpurple2",alpha=0.3)
wrap_2<-wrap(ggally_densityDiag,size=2,color="skyblue")
wrap_3 <- wrap(ggally_cor, size = 40, color = "darkgrey", fontface = "bold")
ggpairs(happy_con[,5:11],
        lower = list(continuous = wrap_1),
        diag = list(continuous = wrap_2),
        higher = list(continuous = wrap_3))
```



接着以曼哈顿距离和卡方来绘制**广义QQ plot**，确实是有偏态但基本符合正态。

```
mah <- mahalanobis(d, colMeans(d), var(d)) ## p=6, 应该近似独立chi^2(6)分布
y <- sort(mah)
x <- qchisq((1:n)/(n+1), p)
dt <- data.frame(x,y)
ggplot(dt, aes(x, y)) +
  geom_point(color = 'orange',size=5) +
  labs(title ='Mahalanobis distance vs. chi-square quantiles',
       x='Chi-square quantile',y='Obs') +
  geom_smooth(method='lm', formula= y~x,se = FALSE,size=3) +
  theme(plot.title = element_text(size=25, face="bold"),
        axis.title = element_text(size=20, face="bold"),
        axis.text = element_text(size=18))
```



做正态转换后的结果符合正态。

```
z <- qnorm(pchisq(mah, 7))
shapiro.test(z)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  z
## W = 0.96819, p-value = 0.001096
```

以下无论用哪个检验也都分显著符合正态分布，检验结果搭配一开始绘制的图，确实符合正态分布。

```
d <- t(as.matrix(happy_con[,5:11]))
mvnormtest::mshapiro.test(d)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.95187, p-value = 0.00003113
```

```
mvShapiroTest::mvShapiro.Test(as.matrix(happy_con[,5:11]))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(happy_con[, 5:11])
## MVW = 0.95442, p-value < 0.00000000000000022
```