Estadística Aplicada III

Pruebas y Transformaciones para normalidad

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana 4

Transformaciones

Motivación

- La calidad de las inferencias de los métodos que veremos en este curso dependen en mucho de la normalidad de los datos
- Tenemos que verificar, *a forciori*, si las observaciones $\mathbf{X}_i$ violan de manera importante el supuesto de que provienen de una distribución normal.
- ¿Qué sabemos?
  - Todas las combinaciones lineales de normales son normales
  - Contornos de la densidad normal son elipsoides.
  - Las distancias de Mahalanobis tienen distribución χ^2.
- A partir de estas propiedades, podemos construir algunas pruebas.

La mayor parte de las pruebas de normalidad se concentran a una o dos dimensiones, buscando responder las preguntas

- ¿Las distribuciones marginales de $\mathbf{X}$ parecen ser normales?
- ¿scatterplots de pares de observaciones dan una apariencia elíptica como se espera de una población normal?
- ¿Hay valores extremos que deban ser revisados o eliminados de algún modo del conjunto de los datos? ¿Hay puntos influenciales?

En conjuntos de datos de la vida real, es raro encontrar conjuntos de datos que no puedan transformarse a normalidad de algún modo.

Tipos de herramientas

Hay tres tipos de técnicas a considerar para verificar y alcanzar normalidad:

- Herramientas gráficas: qq-plots o gráficas de probabilidad, gráfica de log-densidad,
- Herramientas numéricas: prueba de Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov, Cràmer-von-Mises, prueba tipo Monte Carlo
- Transformaciones: Box-Cox

En la práctica estadística, usualmente esta parte no se publica, pero todos suponen que se aplicaron varias técnicas, más de una en cada categoría.

Herramientas gráficas: Gráficas de probabilidad

- Son gráficas de los percentiles o cuantiles muestrales contra los percentiles de la distribución esperada de los datos: $(X_{(i)}, q_{(i)})$, donde $q_{(i)}$ satisface $P(Z \leq q_{(j)}) = \frac{j-0.5}{n}$
- Si los datos siguen la distribución esperada, se espera observar a los puntos muy cerca de una linea recta.
- Los $qq-$plots son informativos sólo cuando la muestra es moderadamente grande, digamos $n \geq 20$. En pequeñas muestras, éste tipo de gráfica puede llevar a conclusiones erróneas.
- Un indicador cuantitativo asociado a esta gráfica es la correlación entre $X_{(i)}$ y $q_{(i)}$. Esta es la estadística de la prueba muy similar a la de Shapiro-Wilk.
- Es conveniente hacer $qq-$plots de combinaciones lineales de varias características. Una sugerencia es considerar combinaciones con el eigenvector correspondiente al valor propio más grande.

Herramientas gráficas: Distancias de Mahalanobis

- Si el vector **x** tiene distribución normal, entonces

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_{(p)}$$

- Si sustituímos $\boldsymbol{\mu}$ por $\bar{\mathbf{x}}$ y $\boldsymbol{\Sigma}$ por $\mathbf{S}$, esperamos que la relación se siga cumpliendo bajo la normalidad de los datos.
- Si se calcula para cada observación la distancia de Mahalanobis:

$$d_j^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

entonces esperamos que los valores d_j^2 formen una muestra de una $\chi^2_{(p)}$

- Se puede hacer un $qq-$plot de los valores de d_j^2 contra los percentiles de una distribución ji-cuadrada, o bien graficar d_j contra percentiles de $\sqrt{\chi^2_{(p)}}$, que son más fáciles de interpretar.
- La función `delta` que vimos la clase anterior es útil para definir una función en R que haga esta gráfica.
- Adicionalmente, valores d_j^2 grandes indican que tales observaciones pueden ser valores extremos (outliers) conjuntos.

Herramientas gráficas: Gráfica de log-densidad

Sea $X_1, \ldots, X_n$ una muestra aleatoria, y sea

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

un estimado del kernel de la densidad de X. Aquí, $K_h(x) = \frac{1}{h} K(\frac{x}{h})$ y K es el kernel de una densidad conocida y simétrica (normal, t, Cauchy, triangular, etc.); h es un parámetro de suavizamiento (ancho de banda).
Entonces

- Escoge K la distribución normal estándar y $h = 1.06\hat{\sigma} n^{-1/5}$
- Se construye una gráfica con los pares de puntos $(X_i, \log[\hat{f}(X_i)])$
- Los puntos anteriores se comparan con (el log de) una normal con parámetros $\hat{\mu}$ (20 % trimmed mean) y $\hat{\sigma}$ ($MAD * 1.4826 + h^2$). Se usan estimadores robustos para evitar sesgos por valores extremos o datos atípicos en la media y varianza que inflen el estimador empírico de la densidad.

La gráfica está implementada en el script `normplot` debido a Martin L. Hazelton.

Herramientas numéricas: Prueba de Shapiro-Wilk

- Esta prueba se basa en el coeficiente de correlación entre los percentiles muestrales y una función del valor esperado de las estadísticas de orden de una distribución normal (en lugar de los percentiles per se) como vimos en el $qq-$plot.
- En R:

```
shapiro.test(rnorm(100))


Shapiro-Wilk normality test

data:  rnorm(100)
W = 0.98651, p-value = 0.4056
```

Prueba tipo Monte Carlo

- Esta prueba se basa en la gráfica log-densidad.
- Sea

$$e_i = \log(\hat{f}(X_i)) - \log(\hat{\phi}(X_i))$$

y definan la discrepancia total entre valores ajustados y empíricos, $S = \sum e_i^2$. S es la estadística de prueba para la hipótesis: $H_0 : X_1, \dots X_n \sim N$. Se propone una prueba tipo Monte Carlo en el script `normplot`:

  - Genera N muestras independientes de una distribución normal, de tamaño n cada muestra.
  - Calcula $S^{(j)}$ para cada muestra, para $j = 1, \dots, N$
  - Calcula el p-value para aceptar o rechazar la hipótesis, como $p_{MC} = \#(S^{(j)} \geq s_{obs})/N$.

Transformaciones: Box-Cox multivariado. I

- Para buscar una transformación conjunta a la normalidad, se resuelve el siguiente problema de optimización:

$$Max_{\boldsymbol{\lambda}} l(\boldsymbol{\lambda}) = -\frac{n}{2}\log(|\mathbf{S}(\boldsymbol{\lambda})|) + \sum_{k=1}^{p}\left((\lambda_k - 1)\sum_{j=1}^{n}\log(x_{jk})\right)$$

  donde $\mathbf{S}(\boldsymbol{\lambda})$ es la matriz de covarianzas muestral del vector $\mathbf{x}_j^{(\boldsymbol{\lambda})}$.
- Al implementar computacionalmente el método puede ser que no se obtenga una solución razonable. No siempre se tiene éxito.
- Otra posibilidad es resolver marginalmente los valores λ_k, aunque marginales normales no garantiza normalidad conjunta. (busquen un ejemplo de que esto es cierto).

Transformaciones: Box-Cox multivariado. II

- En R hay una implementación del método de Box-Cox multivariado en la biblioteca `car`.

```
library(car)
```

```
Loading required package: carData
```

```
data(iris3)
X <- iris3[,,1]
(m <- powerTransform(X))
```

```
Estimated transformation parameters
  Sepal L.   Sepal W.   Petal L.   Petal W.
0.41655582 1.27294127 0.72864401 0.02443928
```

```
summary(m)
```

```
bcPower Transformations to Multinormality
         Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Sepal L.    0.4166           1      -2.6421       3.4752
Sepal W.    1.2729           1      -0.1900       2.7358
Petal L.    0.7286           1      -0.8958       2.3531
Petal W.    0.0244           0      -0.5009       0.5498

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                                LRT df    pval
LR test, lambda = (0 0 0 0) 3.922044  4 0.41666

Likelihood ratio test that no transformations are needed
                                LRT df     pval
LR test, lambda = (1 1 1 1) 12.9728  4 0.011409
```