Taller de problemas GRUPO inferencia 2023 MAT3 GIN

Estadística Inferencial

Contenidos

1 Taller Problemas evaluable 22-23: Estadística Inferencial

Valor 14 puntos. Todos los apartados valen 1 punto.

Se trata de resolver los siguientes problemas y cuestiones en un fichero Rmd y su salida en un informe en html, word o pdf.

1.1 Problema 1: Regresión lineal simple. 7 puntos.

Consideremos los siguientes datos

```
x=c(-2,-1,2,0,1,2)
y=c(-7, -5,  5, -3,  3.0,  4)
```

1. Calcular manualmente haciendo una tabla los coeficiente de la regresión lineal de y sobre x.
2. Calcular los valores $\hat{y}_i = b_0 + b_1 \cdot x_1$ para los valores de la muestra y el error cometido.
3. Calcular la estimación de la varianza del error.
4. Resolver manualmente el contraste $\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$, calculando el p-valor.
5. Calcular SST, SSR y SSE.
6. Calcular el coeficiente de regresión lineal r_{xy} y el coeficiente de determinación R^2. Interpretad el resultado en términos de la cantidad de varianza explicada por el modelo
7. Comprobar que los resultados son los mismos que los obtenidos con la función `summary(lm(y~x))`.

1.2 Problema 2: Distribución de los grados de un grafo de contactos. 3 puntos

The marvel chronology project es una web que ha recopilado las apariciones de los personajes Marvel en cada uno de los cómics que se van publicando.

En el artículo Marvel Universe looks almost like a real social network se estudió la red de contactos de los personajes del Universo Marvel de la serie de cómics books. Dos personajes tienen relación si han participado

en al menos un mismo cómic; a semejanza del Oracle of Bacon donde se relacionan los actores de las películas de Hollywood que han participado en al menos una película juntos.

Si construimos el grafo de asociado a esas relaciones el grado de cada carácter (personaje) será el número de ortos caracteres (personajes) con los que ha colaborado. Cuando más importante es el personaje más colaboraciones tiene.

Los grados de cada caracteres están en el fichero `degree_Marvel_characters.cvs`. Según algunos estudios la distribución de los grados de los grafos de contactos sigue una ley potencial frecuencia grado $k = \beta_0 \cdot grado^{\beta} 1$ si eliminamos los 20 más pequeños.

```
data=read_csv("datasets/degree_Marvel_characters.csv")
```

Se pide:

1. Cargad los datos. Calcular las frecuencias de los grados, es decir el número de caracteres que tienen 1, 2 ,3 .... colaboradores para cada grado (número de colaboraciones) observado.

2. Ajustar un modelo lineal, potencial y exponencial a la relación entre $y =$ "frecuencia del grado" y $x = grado$ dibujar las gráficas de ajuste de cada modelo con gráficos semi-log y log-log si es necesario.
3. Para el mejor modelo calcular los coeficientes en las unidades originales y escribir la ecuación del modelos.

1.3 Problema 3: Longitud reviews mallorca AirBnb 2022. 4 puntos

El siguiente código cuenta cuantas palabras hay en un la variable `commnets` del fichero `reviews.csv` de los comentario a cada apartamento de Mallorca extraído de la web Inside AirBnb que recoge datos de los alquileres vacacionales por zonas del mundo de la web de alquiler de apartamentos vacacionales AirBnb. Se puede leer con el siguiente código y contar el número de palabras con la `stringr::str_count`.

```
read_csv("datasets/reviews.csv")->reviews
```

```
## Rows: 342750 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): reviewer_name, comments
## dbl  (3): listing_id, id, reviewer_id
## date (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
names(reviews)
```

```
## [1] "listing_id"    "id"            "date"          "reviewer_id"
## [5] "reviewer_name" "comments"
```

```
library(stringr)
#str_count(str, pattern = "")
str_count(str=reviews$comments[1],pattern ="\\s+")
```

```
## [1] 78
```

Es habitual que la frecuencia de la longitud de los comentarios, es decir cuantos comentarios tienen 5, 6, 7 palabras y sus frecuencias siguen una ley que puede ser: lineal, exponencial o potencial. Como hemos hecho en el tema de regresión lineal calcular se trata de calcular y dibujar los tres modelos y decidir cuál es el más ajustado.

Se pide:

1. Calcular las longitudes de todos los comentarios (utilizar funciones como `mutate`, `arrange`, `filter`....) y las frecuencias de cada longitud y filtrar (con la función `filter`) solo los comentarios con **MÁS de 20 palabras y MENOS de 800** y guardarlos en una tibble con dos columnas N_{words}= número de palabras y $Frec$=frecuencia absoluta de las palabras.

2. Calcular los tres modelos lineal $Freq = \beta_0 + \beta_1 \cdot N_{words}$, potencial $Freq = \beta_0 \cdot (N_{words})^{\beta_1}$ y exponencial $Freq = \beta_0 \cdot \beta_1^{N_{words}}$.

3. Repetir el ajuste anterior pero sustituyendo el la variable N_{words} por el rango u orden de N_{words}.