11.67 Estadística Aplicada - Trabajo Práctico

1 Búsqueda del set de datos

- Buscar un set de datos que contenga entre 3 y 6 variables, de las cuales 1(una) debe ser categórica y el resto deben ser numéricas. La cantidad de observaciones debe estar entre 100 y 500. Si los datos originales tienen más de 500 observaciones, se deberá tomar una muestra aleatoria.
- Identificar si el set de datos consiste en una serie de tiempo o en una muestra de tipo transversal.
- Enviar la propuesta de set de datos al docente asignado antes de comenzar el trabajo.

2 Análisis descriptivo

- Para cada una de las variables numéricas se deberá hacer un análisis descriptivo que incluya todos los aspectos vistos en el ejercicio `T2.1` realizado en clase. No olvidar incluir una definición clara de las variables.
- El grupo deberá elegir una de las variables numéricas y hacer un análisis gráfico de cómo cambia la distribución de los datos en los distintos niveles de la variable categórica. Se sugiere hacer un análisis similar al que se pide en los incisos a. y b. del ejercicio `T2.4`.
- Producir un programa en `R` que permita ejecutar los cálculos y gráficos que el grupo considere necesarios, partiendo de la importación de datos. **Importante:** Usar únicamente las funciones y librerías vistas en clase, a menos que el grupo proponga armar alguna función propia.
- **Importante:** Se deberá trabajar creando un *proyecto* en `R`. Investigar la opción `File > New Project... > New Directory` , que permite armar una carpeta local que contendrá un archivo `.Rproj`. Allí se podrán incluir todos los scripts del trabajo y los archivos de datos. Sobre esta carpeta se trabajará a lo largo del curso.

3 Producción de informe

- Armar la base del informe. Debe tener carátula e índice al principio y una sección de referencias bibliográficas al final. En el cuerpo de este documento se añadirán sucesivamente todas las etapas del trabajo, a medida que avance el curso.
- A esta altura deberá incluirse solamente lo hecho en los puntos 1 y 2. Resumir los resultados en tablas y gráficos. **Importante:** Incluir todos los aspectos de análisis que el grupo considere importantes. Tener en cuenta que no se aceptarán resultados sueltos.
- Se deberá conservar un formato claro y unificado. Se evaluará la calidad de presentación de los contenidos. No olvidar citar claramente la fuente de los datos y otras referencias necesarias.

4 Primera entrega

El grupo debe entregar una carpeta titulada `Proyecto Grupo ...` que tenga lo siguiente: 1) El archivo `.RProj` que debe ejecutar la sesión de `R Studio`, 2) El archivo `Punto2.R` con el programa realizado en el punto 2, 3) El archivo del informe completo, en formato **.pdf**, 4) El o los archivos con los datos originales, en formato `.csv`, `.xlsx` o `.txt`.

5 Elección de cantidades estimables

Para la variable *principal* del data-set, el grupo deberá elegir 2(dos) cantidades estimables, q_1 y q_2, entre las siguientes opciones:

- Una medida de posición dada por un cuantil.
- Una medida de forma de la distribución: Coeficiente de variación, coeficiente de asimetría o coeficiente de curtosis.
- La probabilidad de un evento dado por la variable aleatoria, o la proporción poblacional de individuos que cumplen cierta condición elegida por el grupo.

6 Formulación de estimadores

Para cada cantidad q_i el grupo debe definir dos estimadores $\hat{q}_{i1}$ y $\hat{q}_{i2}$ para compararlos. Las opciones son:

- Un estimador paramétrico y uno no paramétrico.
- Un estimador de máxima verosimilitud y otro de ajuste de momentos basado en la misma distribución.
- Dos estimadores de máxima verosimilitud basados en distribuciones distintas.

Se debe escribir la expresión de cada estimador como variable aleatoria en función de la muestra y justificar claramente la elección de los modelos de distribución propuestos para estimadores paramétricos. Comentar si los estimadores formulados son sensibles o no a la presencia de datos atípicos.

- **Importante:** Aquellos casos en los que la variable principal se vea afectada sensiblemente por las distintas categorías de los datos, evaluar junto con los docentes la posibilidad de implementar un modelo de mezcla de distribuciones para los estimadores paramétricos.

7 Implementación y aplicación de los estimadores

El grupo debe programar en `R` los cuatro estimadores propuestos, mediante funciones que tengan como argumento el vector de datos `x` y devuelvan la estimación correspondiente. Por ejemplo, la función `estimador_11(x)` debe realizar $\hat{q}_{11}$ para cualquier vector de datos. Aplicar las funciones a los datos del trabajo, para obtener las estimaciones correspondientes. Además, estimar el error estándar de cada uno de los estimadores con un método adecuado.

8 Producción de informe y segunda entrega

Resumir el trabajo realizado en una nueva sección del informe. Incluir todas las observaciones cualitativas que el grupo considere apropiadas. Armar una tabla que contenga las estimaciones calculadas para q_1 y q_2, y estimaciones del error estándar de cada uno de los estimadores propuestos e interpretar los resultados. Actualizar el informe en la carpeta del proyecto, e incluir en un nuevo archivo `.R` los programas necesarios para esta entrega.

9 Bondad de ajuste de distribuciones

En el punto 6, el grupo tuvo que elegir al menos 1(un) modelo de distribución para obtener estimadores paramétricos. Se debe estudiar la bondad de ajuste de esos modelos con los siguientes criterios:

- Comparacion de log-verosimilitudes máximas.
- Construcción e interpretación de QQ-Plots.
- Gráficos donde se compare la distribución empírica con las distribución ajustada con cada modelo.
- Gráficos donde se compare, para el caso de variables continuas, el histograma con la densidad ajustada por cada modelo.

Importante: Si el grupo eligió un sólo modelo de distribución para obtener los estimadores, se deberá elegir uno más para realizar la comparación de ajuste. Si el grupo eligió un sólo modelo de distribución para obtener los estimadores, se deberá elegir uno más para realizar la comparación de ajuste. Consultar a los docentes.

10 Intervalos de confianza

Para las dos cantidades estimables elegidas en el punto 5 y usando para cada una el estimador que el grupo crea conveniente en función de todo lo realizado hasta este punto, se deben obtener intervalos de confianza del 90 %. Explicar claramente cuál es la metodología usada. **Importante:** Justificar adecuadamente la elección final del estimador en cada caso.

11 Producción de informe y tercera entrega

Resumir el trabajo realizado en una nueva sección del informe. Incluir todas las observaciones cualitativas que el grupo considere apropiadas. Actualizar el informe en la carpeta del proyecto, e incluir un nuevo archivo `.R` los programas necesarios para esta entrega.

12 Regresión - Análisis exploratorio

Tomando la variable estudiada en las secciones anteriores como explicada y el resto de las variables del dataset como explicativas, se deberán probar distintas combinaciones de variables para ajustar modelos de regresión. En cada caso, se calcularán las siguientes métricas:

- Coeficiente de determinación R^2 ajustado.
- Varianza residual.
- Determinante de la matriz de correlaciones.
- Coeficiente C_p de Mallows.
- Suma de cuadrados de la predicción (PRESS) obtenida por validación cruzada.

Se deben usar estas métricas como criterio para elegir el mejor modelo. Explicar cómo se usa cada métrica para realizar la elección. La cantidad de combinaciones podrá variar de grupo a grupo, pero deberá haber un mínimo de 3(tres). Para las variables elegidas en el punto anterior.

13 Regresión - Diagnóstico

Para un modelo que contenga las variables elegidas en el punto anterior, realizar un diagnóstico de los siguientes aspectos: 1) Supuesto de linealidad de la regresión, 2) Supuesto de normalidad de los errores, 3) Supuesto de homocedasticidad de los errores, 4) Supuesto de independencia de los errores, 5) Outliers y puntos influyentes. Para cada punto el grupo deberá utilizar las herramientas y métricas vistas en clase. En caso de necesitar medidas para corregir posibles desviaciones en los supuestos, investigar sobre la transformación de Box-Cox. Consultar a los docentes.

14 Regresión - Validación del modelo

plantear el modelo lineal de regresión definitivo, indicando los supuestos teóricos. Ajustar los parámetros por cuadrados mínimos. Hacer intervalos de confianza del 90 % para los coeficientes de regresión. ¿Qué observa?

15 Aplicación

Elegir una de las siguientes opciones de aplicación del modelo final:

- Realizar un intervalo de confianza o un test de hipótesis para alguno de los coeficientes de regresión, interpretando su significado.
- Realizar un intervalo de predicción para la variable respuesta, dado un valor fijo para las variables explicativas (propuesto por el grupo).
- Realizar un intervalo de confianza para la media de la variable respuesta, dado un valor fijo para las variables explicativas (propuesto por el grupo).

Justificar la elección.

16 Informe y entrega final

Resumir el trabajo realizado, incluyendo todas las apreciaciones cualitativas necesarias, en una última sección del informe.