Estadística Aplicada. Lab3: Regresión lineal y Correlación

Dr. Edlin Guerra Castro, ENES Mérida, UNAM

26/10/2021

Parte 1

1. Importa los datos del archivo *reg.csv* y explora su estructura.
2. Obtén un gráfico de la relación de y con respecto a x. Responde a las siguientes preguntas:
   a) ¿Existe variabilidad en el eje de las x?
   b) ¿De qué tamaño es la variabilidad en el eje de las y? ¿Cómo podrías medirla?
   c) ¿Crees que el valor de y depende de los valores que toma x?
3. Obtén un gráfico de la relación entre y2 y x, y colócalo junto al anterior. Responde a las siguientes preguntas:
   a) ¿De qué tamaño es la variabilidad en el eje de las y2?
   b) Compara la variabilidad de y con la de y2.
   c) ¿Crees que el valor de y2 depende de los valores que toma x?
4. Ajuste um modelo lineal para la relación y~x, usando la función 'lm', y grábalo con el nombre mod1. Obtén la tabla de ANOVA y responde a las siguientes preguntas.
   a) Identifica el o los valores de la tabla que representa la variación en y que es debida al cambio en x.
   b) Identifica el o los valores de la tabla que representa la variación en y que es debida al error de muestreo.
   c) Identifica el valor de la tabla que representa la variación total de y.
   d) ¿Cómo se obtiene el valor de F, y qué representa?
   e) ¿Cuál es la probabilidad de obtener un valor de F como el obtenido si la Ho fuese verdadera?
   f) ¿Cómo interpretas los resultados de este análisis en términos de la relación entre x y y?
5. Ajuste um modelo lineal para la relación y2~x, usando la función 'lm', y y grábalo con el nombre mod2. Obtén la tabla de ANOVA y responde a las siguientes preguntas:
   a) ¿De qué tamaño es la variabilidad en y2 debida a la regresión en este modelo comparada con el anterior?
   b) ¿De qué tamaño es la variabilidad en y2 debida al error comparada con la variabilidad total en esa misma variable?
   c) ¿Cómo interpretas los resultados en términos de la relación entre las variables x y y2?
6. Considerando que los grados de libertad son iguales en ambos modelos, ¿cuál es el valor de F a partir del cuál la Ho de ambas pruebas debe ser rechazada? (PISTA: usa funciones de la distribución Fdist).
7. Elabora dos gráficos con los datos de los dos modelos anteriores. Representa la linea de regresión estimada por el ajuste del primer modelo, y otra que represente la media general total de y. Para hacer esto usa la función 'abline'. Haz lo mismo para el segundo.
8. Utilizando la función 'summary' obtén el resumen de los análisis efectuados.
   a) ¿Cuáles fueron las estimativas de la pendiente y el intercepto de la recta en el modelo mod1?
   b) ¿Cuáles fueron estas estimativas en el modelo mod2?

c) ¿Cuáles fueron las estimativas de 'sigma' para ambos modelos?
d) ¿Puedes obtener este último dato a partir de las tablas de ANOVA?

9. Con los resultados obtenidos en 'summary', calcula los intervalos de confiança (95%) de los parámetros alfa y beta de la regresión que resultó significativa.

   a) ¿Cuál será el valor que usarás como error estándar?
   b) ¿Cuál el valor de t correcto?
   c) Copia el siguiente código y compara con los resultados que tu obtuviste. confint(mod1)
   d) Verifica si el IC de la pendiente para el mod2 inlcuye el 0.

Parte 2

Para poner en práctica los conceptos sobre correlación introducidos en las clases teóricas, trabajaremos con los datos del ejemplo 'iris' (de R). Edgar Anderson colectó datos del largo y ancho de pétalos y sépalos (cm) de 50 flores individuales de cada una de tres especies (Iris setosa, Iris virginica y Iris versicolor) para cuantificar su variación morfológica. Más tarde Ronald Fisher desarrolló el modelo lineal discriminante para distinguirlas.

1. Explora la estructura de la tabla iris, el tipo de variables que fueron medidas, y el número de réplicas para cada variedad de planta.

2. Crea un nuevo objeto que contenga únicamente las medidas de la variedad versicolor. Obtén un gráfico XY de las dos medidas del sépalo para esta variedad.

3. Obtén el valor de r de Pearson entre dichas medidas, usando la función 'cor'. Invierte los ejes y obtén un nuevo gráfico.

   a) ¿Cambia la fuerza de la asociación con la inversión de los ejes?
   b) ¿Cambia la correlación?

4. Obtén un gráfico donde todas las 4 variables se correlacionen con todas las variables medidas en las 4 especies. (PISTA: experimenta la función 'plot' o la función 'pairs' sobre el objeto iris). Responde:

   a) ¿Cuál es el par de variables que visualmente tienen la asociación más fuerte? ¿Cuál más débil?
   b) Confirma tus respuestas numéricamente, usando la función 'cor'.
   c) Copia el siguiente código para obtener todas las correlaciones a la vez.
   d) Explica el caso donde se considera la asociación con la variable Species.

5. Verifica cuáles de estas asociaciones son significativas, usando la función 'cor.test':

6. Copia el siguiente código y describe cómo cambia la variable YY para cada unidad de cambio en XX.

```
kk <- data.frame(xx=1:30,yy=round((c(1:30)+rnorm(30,3,1))^3,2))
plot(kk$xx,kk$yy)
```

```
a) ¿Es la fuerza de la asociación entre XX y YY similar para cualquier rango de valores
de las variables?
b) ¿Cómo describirías la asociación entre las dos variables para el intervalo de XX entre 0-15?
c) ¿Cómo la describirías para el intervalo de XX entre 15-30?
```

7. Obtén un valor de correlación que describa la relación entre las dos variables, y aplica una prueba estadística al dicho valor, interpretando el resultado.

   a) ¿Consideras que dicho número representa la asociación para todos los valores de XX?
   b) ¿Consideras que este valor sería representativo para explicar la forma como XX y YY se asocian?

8. Experimenta aplicando una transformación de raíz cúbica a la variable YY y repite el gráfico.

   a) Obtén un valor de correlación que describa la relación entre las dos variables, y confirma si este es significativo.

b) Compara los resultados de la prueba con aquellos del numeral 6.

9. Experimenta aplicando la correlación de Spearman y Kendall a los datos sin transformar usando el argumento 'method' de la función 'cor'. Compara los resultados en términos de la fuerza de asociación, la significancia de la prueba y si su representatividad.