Pontificia Universidad Católica de Chile
Facultad de Matemáticas - Departamento de Estadística
Profesora: Ana María Araneda
Ayudantes: Eduardo Vásquez y Vanesa Reinoso
Correos: evasquezt@uc.cl y vcreinoso@mat.uc.cl

EPG3306 - Métodos Estadísticos I
Ayudantía 4
23 de abril del 2022

Contenidos:

- Regresión Lineal Simple.
- Calidad del Modelo.

---

1. Considere el modelo de regresión lineal simple visto en clases.
    - Demuestre que el promedio de los residuos $\epsilon_i = y_i - \hat{y}_i$ es 0.
    - Demuestre que el punto $(\bar{x}, \bar{y})$ pertenece a la recta ajustada.

2. La base de datos `poverty.txt` contiene información sobre los 50 estados de Estados Unidos y el Distrito de Columbia. Las variable que se interesa estudiar es `Brth15to17` la cual indica la tasa de natalidad del año 2002 por cada 1000 mujeres de 15 a 17 años. Una variable explicativa podría ser la tasa de pobreza, que es el porcentaje de la población del estado que vive en hogares con ingresos inferiores al nivel de pobreza definido por el gobierno federal, la cual se encuentra codifica como `PovPct` en la base. Responda las siguientes preguntas con la ayuda del software `R` de ser necesario.

    (a) Cargue la base de datos. Identifique la variable respuesta y la variable explicativa y obtenga un gráfico de dispersión de las variables con la función `plot.`

    (b) A partir de la información que entrega el gráfico, ¿es adecuado medir la asociación entre las variables a través del coeficiente de correlación lineal? Si es así, pídalo a `R` utilizando `cor(x,y)` e interprete su valor.

    (c) Indique aquí el modelo y sus supuestos.

    (d) Obtenga las estimaciones de Mínimos Cuadrados de β_0 y β_1, de acuerdo a las fórmulas derivadas en clases:

    $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

    (e) Ajuste ahora el modelo utilizando la sintaxis: `modelo = lm(y ~ x)`. Compruebe los valores estimados de β_0 y β_1 que obtuvo en el apartado anterior. Puede pedir los coeficientes ajustados utilizando el comando `coef(modelo)`.

(f) Plantee las hipótesis pertinentes para chequear la significancia de la pendiente estimada. Obtenga el estadístico del test:

$$t_0 = \frac{|\hat{\beta}_1|}{\hat{\sigma}/\sqrt{S_{xx}}}$$

Recuerde que $\hat{\sigma}^2 = \frac{SCE}{(n-2)}$. Para obtener los residuos, puede utilizar `resid(modelo)`.

(g) Obtenga el valor-p del test y concluya en términos del problema.

(h) Compare sus resultados con los que entrega el comando `summary(modelo)` e interprete.

(i) Considere el estado de California. ¿Cuál es la tasa de natalidad ajustada o predicha por el modelo? Compárela con el valor observado.

3. En nutrición, es importante estudiar el consumo de grasas en ciertas poblaciones. La base de datos `grasas.csv` contiene información sobre 156 paises y su porcentaje de ingesta de grasas de bebidas alchólicas, productos de origen animales, leche (excluyendo la mantequilla) y productos de origen vegetal. Además, la tasa de obesidad, desnutrición e ingsta de carne por país. Estas observaciones fueron recolectadas entre el 2020 y 2021 aproximadamente.

   (a) Cargue la base de datos e investigue sobre la asociación entre distintas variables de la base de datos con gráficos de dispersión y/o cálculo de correlación.

   (b) Escoja un modelo a estudiar y escribalo señalando la variable y, la covariable x y los supuestos del modelo.

   (c) Ajuste el modelo en `R` añadiendo la recta ajustada. Interprete los parámetros.

   (d) Señale las hipótesis adecuada para analizar la significancia de la pendiente estimada. Obtenga el estadístico del test:

   $$F_0 = \frac{SCR}{MCE}.$$

   calcule el valor-p y compare sus resultados con la tabla anova.

   (e) Calcule el coeficiente de determinación:

   $$R^2 = \frac{SCR}{SCT}$$

   y compare su resultado con el entregado por el comando `summary`. Interprete su valor.