Tarea 1, versión B. Regresión lineal simple

Gonzalo Pérez, Sheyla Barradas y Luis Genaro Coria

Semestre 2024-2

La tarea se deberá subir al classroom antes de las 11:59 PM del 10 de abril de 2024. Todas las preguntas tienen un valor de 1.5 puntos.

Favor de argumentar con detalle las respuestas.

NOTA. En caso de que se identifiquen respuestas iguales en otras tareas, se procederá a la anulación de las tareas involucradas.

NOTA. Incluir el(los) nombre(s) completo(s) de la(s) persona(s) que está(n) resolviendo los ejercicios. Equipos de máximo cuatro integrantes.

Usar una confianza de 95 % o una significancia de .05 en los casos en donde no se requiera otro nivel de forma explícita. En el caso de realizar alguna transformación de las variables, se tiene que hacer explícita la variable que se usa y la interpretación en las pruebas de hipótesis o intervalos de confianza.

1. Regresión a través del origen.

Ocasionalmente, un modelo en donde el valor del intercepto es conocido a priori y es igual a cero puede ser apropiado. Supongamos que además se considera el posible uso de una regresión ponderada, es decir, el modelo está dado por:

$$y_i = \beta x_i + \varepsilon_i, \qquad i = 1, ..., n,$$

donde $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ son variables independientes tal que $\varepsilon_i \sim N\left(0, \frac{\sigma^2}{w_i}\right) \ \forall \ i = 1, ..., n.$

En general σ^2 es desconocida, pero en lo que sigue suponga que es conocida. Además suponga que

$$w_i = \frac{1}{x_i^2}, \ i = 1, ..., n.$$

I) Encuentre el estimador de β obtenido por el método de máxima verosimilitud, $\widehat{\beta}$.
II) Encuentre la expresión de la varianza de $\widehat{\beta}$.
III) Demuestre que $\widehat{\beta}$ es el UMVUE de β, es decir, que es el mejor estimador insesgado de β.

2.

Considere el modelo de regresión

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

donde $E(\epsilon_i) = 0, \ V(\epsilon_i) = \sigma^2$ y $Cov(\epsilon_i, \epsilon_j) = 0 \ \forall \ i \neq j; \ \ i, j = 1, ..., n.$

Calcular $V(e_i)$, donde $e_i = y_i - \widehat{y}_i$ y $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ con $\widehat{\beta}_0$ y $\widehat{\beta}_1$ los estimadores de los parámetros del modelo.

Hint: Se puede usar que $V(A - B) = V(A) + V(B) - 2Cov(A, B)$ y que $\widehat{y}_i$ se puede escribir como una combinación lineal de las $y_{i's}$.

3. Expresión alternativa para R^2

Considere el coeficiente de correlación muestral o de Pearson para dos variables X y Y:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{(\sum_{i=1}^{n}(x_i - \overline{X})^2 \sum_{i=1}^{n}(y_i - \overline{Y})^2)^{1/2}}.$$

Considere el modelo de regresión

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

a. Demuestre que:

$$R^2 = r_{xy}^2.$$

Hint: Puede usar lo encontrado en la expresión (77) de las notas.

b. Demuestre que $t^* = t$, donde t es la estadística usada para contrastar "$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$" :

$$t = \frac{\widehat{\beta}_1}{\sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{X})^2}}}.$$

Por otra parte, $t^* = \dfrac{r_{xy}\sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$ es la estadística usada para contrastar "$H_0 : \rho = 0 \quad vs \quad H_a : \rho \neq 0$"

cuando (X, Y) sigue una distribución normal bivariada con coeficiente de correlación $\rho = \rho_{xy}$.

Hint: Puede usar la relación en la expresión (68) de las notas.

4. Problema Anova. Equivalencia con la estimación considerando dos poblaciones normales.

Sea $X_1, ..., X_n$ una m.a. de la distribución $N(\mu_x, \sigma^2)$ y $Y_1, ..., Y_m$ una m.a. de la distribución $N(\mu_y, \sigma^2)$, ambas muestras aleatorias son independientes entre sí. La prueba t se usa bajo este contexto para contrastar, por ejemplo:

$$H_0 : \mu_x = \mu_y \qquad vs \qquad H_a : \mu_x \neq \mu_y.$$

Además, se puede verificar que los estimadores máximos verosímiles para μ_x y μ_y son $\overline{X}$ y $\overline{Y}$, respectivamente.

Demuestre que los estimadores para μ_x y μ_y que se obtienen asumiendo un modelo de regresión lineal simple también son iguales a $\overline{X}$ y $\overline{Y}$, respectivamente. Es decir:

I. Considere una variable Z tal que: $Z = 1$ si la observación es de la población con distribución $N(\mu_x, \sigma^2)$ y $Z = -1$ si la observación es de la población con distribución $N(\mu_y, \sigma^2)$. Considere el modelo de regresión lineal simple:

$$w_j = \beta_0 + \beta_1 z_j + \varepsilon_j,$$

donde $\varepsilon_1, \varepsilon_2, ..., \varepsilon_{n+m}$ son variables independientes tal que $\varepsilon_j \sim N(0, \sigma^2) \ \ \forall \ \ j = 1, ..., n+m$. En este modelo los valores de las variables X y Y componen la variable W, asumiendo que las primeras n observaciones son las que tienen valor $Z = 1$ y el resto son las que tienen valor $Z = -1$. Indique cuál es la esperanza de W para cada valor de la variable Z, es decir, $E(W; Z = 1)$ y $E(W; Z = -1)$, haciendo énfasis en indicar la relación que esto implica entre los parámetros μ_x y μ_y con β_0 y β_1.

II. A partir de los estimadores de los parámetros del modelo de regresión lineal simple en I), desarrolle las expresiones de $\widehat{E}(W; Z = 1)$ y $\widehat{E}(W; Z = -1)$ para dejarlas en términos de x_i y y_i.

5. Problema ANOVA. Medicamentos

Suponga que una empresa farmacéutica está ofreciendo al gobierno un nuevo medicamento para tratar a pacientes con la enfermedad Covid-19. El costo del medicamento es considerable y para tomar una buena decisión se han acercado a usted para analizar los datos que ha compartido la empresa farmacéutica. El archivo Ejercicio5B.csv contiene la información siguiente: Y es un índice de carga viral y Med es una variable con dos niveles dependiendo si se aplicó o no el nuevo medicamento. Se sabe que tener una menor carga viral está relacionado con una menor probabilidad de desarrollar una versión grave de la enfermedad y la empresa afirma que eso se logra al aplicar el medicamento, pues los pacientes que recibieron el medicamento tienen menor carga viral que los que sólo recibieron placebo.

I. Realice un análisis descriptivo y/o la visualización de los datos
II. Escriba la prueba asociada para argumentar en favor o no de la afirmación de la compañía. Para esto deberá indicar qué modelo podría usar y cuales son los supuestos de éste.
III. Lleve a cabo la prueba de hipótesis, justificando que los supuestos del modelo que está usando son válidos. Dé la interpretación de los resultados.
IV. Suponga ahora que dado que el costo del medicamento es considerable, le han vuelto a preguntar si los resultados en el inciso III) son contundentes. Para esto, usted ha decidido analizar más el proceso de generación de los datos y ha platicado con los empleados de la farmacéutica, logrando que le compartan una nueva variable $Edad$. Realice un análisis descriptivo y/o visualización de los datos incluyendo esta nueva información. Comente lo que observe analizando si las conclusiones en III) se pueden **atribuir** sólo al medicamento.
V. Dependiendo de lo observado en IV) y si considera necesario, repita los incisos II) y III) con un conjunto de datos donde el efecto se pueda **atribuir** sólo al medicamento y concluya.

Hint: Recuerde que para poder **atribuir** un efecto a algún factor se deben comparar poblaciones homogéneas, es decir, que no exista otro factor oculto que pudiera estar asociado con las diferencias o no diferencias que se observen.

6. Uso del modelo de regresión lineal simple

Los *pingüinos Macaroni* ponen nidadas de dos huevos de tamaño diferente. El peso en gramos de los huevos de 11 nidadas se presenta en la tabla de abajo.

I. Ajuste la recta de regresión para estimar el peso promedio del huevo mayor (y) dado el peso del huevo menor (x). Comente sobre el ajuste del modelo, es decir, si parece correcto y si se cumplen los supuestos.
II. Los investigadores tienen la sospecha de que en promedio se puede decir que la diferencia entre el peso mayor y el peso menor es constante (es decir, no depende del peso del huevo menor observado). Usando el modelo en I) realice una prueba de hipótesis para responder la pregunta de los investigadores, describiendo con detalle las hipótesis que se contrastan.
III. Posteriormente se observa el peso de los huevos de una nueva nidada, observándose un peso de 70 y 145 gramos. Usando un intervalo adecuado, comente sobre la sospecha de que la nidada de huevos sí proviene de pinguinos $Macaroni$.

```r
x=c(79, 93, 100, 105, 101, 96, 96, 109, 70, 71, 87)
y=c(123, 138, 154, 161, 155, 149, 152, 160, 117, 123, 138 )
Datos6=data.frame(cbind(x,y))
kable(t(Datos6)) %>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

x	79	93	100	105	101	96	96	109	70	71	87
y	123	138	154	161	155	149	152	160	117	123	138

7.

Considere los datos en la base *performance.csv* y las variables: y = academic performance of the school (api00) y x = percentage of students receiving free meals (meals). Estos datos corresponden a una muestra aleatoria de 400 escuelas primarias en California, en donde por escuela se realizaron mediciones que tienen que ver con su desempeño en 2000.

I. Ajustar un modelo de regresión lineal simple. Verificar los supuestos a partir de este modelo. Deberá indicar para cada supuesto qué gráfica o prueba sirve para argumentar el cumplimiento o no del supuesto.

II. En caso de que alguno de los supuestos no se satisfaga en I), realizar modificaciones a las variables para encontrar un modelo en donde sí se satisfagan los supuestos:

   a. Para transformar la variable Y, probar con transformaciones Box-Cox u otras conocidas como log() o exp().
   b. Para transformar la variable X, probar con transformaciones Box-Tidwell u otras conocidas como log() o exp().
   c. Recuerde que siempre se puede sumar una constante (e.g. +1) para hacer positivas a las variables.

   Al finalizar, deberá indicar el modelo de regresión lineal simple que se ajustará, haciendo explícito qué variables fueron transformadas y cómo. También deberá indicar para cada supuesto del modelo de regresión qué gráfica o prueba sirve para argumentar su cumplimiento.

III. En una misma gráfica incluir los puntos en escala original, la recta de regresión del modelo en I) y la curva del modelo en II).

IV. Interpretar R^2 y la prueba anova del modelo en II).

V. Con el modelo final ayude a un investigador a argumentar a favor o en contra de la hipótesis: "A mayor porcentaje de comidas gratis en la escuela es menor el desempeño de la escuela".