Examen 1, versión A. Regresión lineal simple

Gonzalo Pérez, Dioney Rosas y Rabindranath Durán

18 de noviembre de 2020

El examen se deberá subir al classroom antes de las 10:00 AM del 25 de noviembre de 2020. Todas las preguntas valen 1 punto, excepto la pregunta 8 que tiene un valor de 1.5.

Favor de argumentar con detalle las respuestas.

NOTA. En caso de que se identifiquen respuestas iguales en otros examenes, se procederá a la anulación de los examenes involucrados.

NOTA. Incluir el(los) nombre(s) completo(s) de la(s) persona(s) que está(n) resolviendo los ejercicios, entregar sólo un archivo, con letra legible y el número de hoja con el formato (1/n), con n el número total de hojas del archivo.

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. Este modelo está dado por:

$$y_i = \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

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

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

i. Encuentre el estimador de β obtenido por el método de mínimos cuadrados, $\widehat{\beta}$.
ii. Encuentre la expresión de la varianza de $\widehat{\beta}$.
iii. Demuestre que $\widehat{\beta}$ es el BLUE de β, es decir, que es el mejor estimador lineal e insesgado de β:
    a. $\widehat{\beta}$ es un estimador lineal e insesgado de β y
    b. su varianza es mínima dentro del conjunto de estimadores lineales insesgados de β.

2. Problema Anova. Equivalencia con la prueba t para comparar dos poblaciones.

Sea $X_1, \ldots, X_n$ una m.a. de la distribución $N(\mu_x, \sigma^2)$ y $Y_1, \ldots, Y_m$ 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:

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

Considere la estadística t asociada a la prueba t antes mencionada.

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 $N(\mu_y, \sigma^2)$. Consdere el modelo de regresión lineal simple:

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

donde $\epsilon_1, \epsilon_2, \ldots, \epsilon_{n+m}$ son variables independientes tal que $\epsilon_j \sim N(0, \sigma^2)\ \ \forall\ \ j = 1, \ldots, n+m$. Asuma que las primeras n observaciones son las que tienen valor $Z = -1$ y la variable W corresponde a $X_1, \ldots, X_n$ y el resto son las que tienen valor $Z = 1$ y la variable W corresponde a $Y_1, \ldots, Y_m$. Demuestre que este modelo de regresión implica los mismos supuestos que los considerados para la prueba t dando la relación entre los parámetros μ_x y μ_y con β_0 y β_1.

ii. En términos de los parámetros del modelo de regresión lineal simple en i), indique cómo se deben escribir las hipótesis

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

Además dé la expresión de la estadística asociada a la prueba que se usaría para contrastar estas hipótesis en el contexto del modelo de regresión lineal simple.

iii. Demuestre que la estadística encontrada en ii) es equivalente a la estadística t asociada a la prueba t.

iv. Repita los incisos i) a iii) pero considerando ahora que Z es tal que $Z = 1$ si la observación es de la población con distribución $N(\mu_x, \sigma^2)$ y $Z = 0$ si la observación es de la población con distribución $N(\mu_y, \sigma^2)$.

3. Datos Iris

Considere los datos *iris* en el paquete *datasets* de R, ver por ejemplo: https://es.wikipedia.org/wiki/Conjunto_de_datos_flor_iris (https://es.wikipedia.org/wiki/Conjunto_de_datos_flor_iris). Los datos corresponden a un estudio que tenía por objeto cuantificar la variación morfológica de la flor Iris de tres especies (Iris setosa, Iris virginica e Iris versicolor). En cada muestra de cada especie se midieron cuatro rasgos: el largo y ancho del sépalo y pétalo, en centímetros. Se supone que los datos se recolectaron en una misma pastura de forma aleatoria, el mismo día y medidos al mismo tiempo por la misma persona y con el mismo aparato.

Para lo que sigue, considere sólo la muestra de Iris versicolor e Iris setosa.

i. Usando un modelo de regresión lineal simple indique si hay evidencia estadística para argumentar en favor de la siguiente afirmación: “En promedio el largo de los sépalos de la especie Iris versicolor es mayor al correspondiente a la especie Iris setosa”. Realice un análisis descriptivo e indique el modelo que se ajusta, verificando los supuestos e interpretando los resultados.
ii. Usando un modelo de regresión lineal simple indique si hay evidencia estadística para argumentar en favor de la siguiente afirmación: “En promedio el ancho de los sépalos de la especie Iris versicolor es mayor al correspondiente a la especie Iris setosa”. Realice un análisis descriptivo e indique el modelo que se ajusta, verificando los supuestos e interpretando los resultados.

4. 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 Ejercicio4A.csv contiene la información: Y es el número total de anticuerpos y Med es una variable con dos niveles dependiendo si se aplicó o no el nuevo medicamento. Se sabe que tener mayores anticuerpos evita que se desarrolle 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 más anticuerpos que los que sólo recibieron placebo.

i. Realice un análisis descriptivo de los datos
ii. Escriba la prueba asociada para argumentar en favor o no de la afirmación de la compañía. Para esto deberán 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 mucho, 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 incluyendo esta nueva información. Comente lo que observe analizando si las conclusiones en iii) se pueden **atribuir** al medicamento.
v. Dependiendo de lo observado en iv) y si considera necesario, repita los incisos ii) y iii) y concluya.

5.

Suponga que x_1 y x_2 son dos variables para las cuales se tienen observaciones en una muestra aleatoria de tamaño n: $x_{11}, \ldots, x_{1n}$ y $x_{21}, \ldots, x_{2n}$, respectivamente. Suponga que se ajusta el modelo de regresión

$$x_{1i} = \beta_0 + \beta_1 x_{2i} + \epsilon_i,$$

donde $E(\epsilon_i) = 0,\ V(\epsilon_i) = \sigma^2$ y $Cov(\epsilon_i, \epsilon_j) = 0\ \forall\ i \neq j;\ \ i, j = 1, \ldots, n$, obteniendose los estimadores por mínimos cuadrados $\widehat{\beta}_0$ y $\widehat{\beta}_1$, respectivamente. Ahora suponga que se ajusta el modelo

$$x_{2i} = \beta_0^* + \beta_1^* x_{1i} + \epsilon_i^*,$$

donde $E(\epsilon_i^*) = 0,\ V(\epsilon_i^*) = \sigma^{*2}$ y $Cov(\epsilon_i^*, \epsilon_j^*) = 0\ \forall\ i \neq j;\ \ i, j = 1, \ldots, n$, obteniendose los estimadores $\widehat{\beta}_0^*$ y $\widehat{\beta}_1^*$. Muestre que si r es el coeficiente de correlación lineal de Pearson entre x_1 y x_2, entonces

$$r^2 = \widehat{\beta}_1^* \widehat{\beta}_1.$$

Recuerde que el coeficiente de correlación lineal de Pearson entre x y y, r_{xy}, se define como

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

6.

Suponiendo que x y y son variables que siguen una distribución normal bivariada con coeficiente de correlación $\rho = \rho_{xy}$, la prueba para contrastar las hipótesis

$$\text{“}H_0 : \rho = 0 \quad vs \quad H_a : \rho \neq 0\ \text{”}$$

es de interés, pues en caso de rechazar H_0 se puede decir que x y y no son independientes. Para realizar esta prueba se usa la siguiente estadística:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

donde r denota la correlación lineal de Pearson r_{xy}. Se puede verificar que esta estadística sigue una distribución t_{n-2} bajo H_0.

a. Demuestre que

$$\hat{\beta}_1 = \left[\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}\right]^{1/2} \times r,$$

donde $\hat{\beta}_1$ es el estimador de β_1 en el modelo de regresión:

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

b. Demuestre que $t^* = t$, donde t es la estadística usada para realizar la prueba

$$\text{“}H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0\ \text{”},$$

es decir,

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

Concluya que las pruebas de hipótesis tienen también la misma regla de decisión sobre H_0, aunque la interpretación y supuestos son diferentes.

7.

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 del peso del huevo mayor en el peso del huevo menor. Comente sobre el ajuste del modelo, es decir, si parece correcto y que se cumplen los supuestos.
ii. Pruebe si la pendiente de la regresión difiere significativamente (estadísticamente) de la unidad. Interprete.
iii. Posteriormente se observa el peso de los huevos de una nueva nidada, observándose un peso de 75 y 115 gramos. Usando un intervalo adecuado, comente sobre la sospecha de que la nidada de huevos no proviene de pinguinos $Macaroni$.

```
x=c(79, 93, 100, 105, 101, 96, 96, 109, 70, 71, 87)
y=c(133, 148, 164, 171, 165, 159, 162, 170, 127, 133, 148 )

Datos7=data.frame(cbind(x,y))
kable(Datos7)
```

x	y
79	133
93	148
100	164
105	171
101	165
96	159
96	162
109	170
70	127
71	133
87	148

8.

En una gran universidad se seleccionó al azar a 8 estudiantes de economía y se les aplicó una encuesta. Dos de las preguntas fueron: (1) ¿Cuál es su puntaje de GPA en el semestre anterior? (variable G) y (2) En promedio ¿Cuántas horas a la semana pasó durante el último semestre en el bar X? (variable H). El bar X es un lugar muy conocido por los estudiantes.

```
Est=c(1 , 2 , 3 , 4 , 5 , 6 , 7, 8 )
G=c(3.6, 2.2, 3.1, 3.5, 2.9, 2.6, 3.9, 2.6)
H=c(3, 15, 8, 9, 12, 12, 4, 16)

Datos8=data.frame(cbind(Est,G,H))
Datos8
```

Est <dbl>	G <dbl>	H <dbl>
1	3.6	3
2	2.2	15
3	3.1	8
4	3.5	9
5	2.9	12
6	2.6	12
7	3.9	4
8	2.6	16

8 rows

i. ¿Dirías que un modelo de regresión lineal simple serviría para describir la relación entre G (puntaje del GPA) y H (horas a la semana en el bar)? Considera a la variable G como la variable dependiente y argumenta.
ii. Ajusta el modelo de regresión lineal simple. Da la ecuación de la recta ajustada.
iii. Encuentra $\hat{\sigma}^2$, las desviaciones estándar estimadas para $\hat{\beta}_0$ y $\hat{\beta}_1$, los intervalos de confianza para β_0 y β_1, y el R^2. ¿Qué puedes decir al respecto?
iv. Realiza la prueba t para contrastar $H_0 : \beta_1 = 0 \quad vs \quad H_a : \beta_1 \neq 0$. Comenta.
v. Obtén la tabla ANOVA y realiza la prueba asociada con una significancia de .05.
vi. Suponga ahora que un nuevo estudiante de economía pasa 15 horas a la semana en el bar durante las dos primeras semanas de clase. Calcule un intervalo para su puntaje de GPA en su primer semestre si continua pasando 15 horas a la semana en el bar.
vii. ¿Cuál sería la variación promedio del puntaje de GPA al aumentar en una hora a la semana la estancia en el bar? Dé un intervalo de confianza al 90%.
viii. Suponga que un grupo de 5 estudiantes acudirá al bar en promedio 10 horas a la semana durante el siguiente semestre, ¿Cuál será el puntaje de GPA promedio de los cinco estudiantes para el próximo semestre?. Calcule un intervalo.
ix. Describa en general cuál es el puntaje de GPA promedio de los estudiantes que asisten 8 horas a la semana al bar.

x. Suponga que un estudiante por cuestiones de una beca, requiere obtener un puntaje mayor a 3.3. Entre sus planes está pasar alrededor de 5 horas en el bar, pues tiene la hipótesis de que en promedio los estudiantes que pasan esas horas en el bar sí logran obtener un puntaje mayor a 3.3. ¿Con los datos existe evidencia en favor de la afirmación del estudiante?

xi. Suponga que al lado del bar X existe una cafeteria. Un analista supone que, en principio, el puntaje promedio de GPA de los estudiantes que no acuden al bar es igual al puntaje promedio de los que no acuden a la cafetería. Sin embargo, cree por su experiencia que el patrón observado en el cambio promedio del puntaje de GPA al aumentar una hora de estancia en la cafetería es sólo de la mitad del observado al aumentar una hora de estancia en el bar. Es decir, al analista le parece plausible usar un modelo de regresión del estilo:

$$y_i = \alpha + \frac{\beta}{2} x_i^* + \epsilon$$

Donde x^* es el promedio en horas a la semana que los estudiantes pasan en la cafetería durante un semestre y y el correspondiente puntaje de GPA. Los demás parámetros asociados a este modelo son los mismos a los asociados al modelo original. Dado que no se cuentan con observaciones para este estudio particular y sólo se tiene el supuesto del analista, se deciden usar los estimadores $\hat{\alpha}$ y $\hat{\beta}$. Calcule un intervalo de confianza al 95% para el porcentaje de GPA promedio de los estudiantes que acuden 8 horas a la semana al cafetería.

9.

Considere los datos en la base *infectionrisk.txt* y las variables: $y =$ riesgo de infección (InfctRsk) y $x =$ promedio de estancia en un hospital (Stay), sólo los datos de las regiones 1 y 2 (Region==1 | Region==2). Después de una investigación minuciosa, los responsables de la base de datos indican que todos los valores parecen reflejar la esperanza del riesgo de infección para los diferentes valores de x, es decir, que no se debe eliminar ninguna observación.

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
   b. Para transformar la variable X, probar con transformaciones Box-Tidwell u otras conocidas como log() o exp().

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: El riesgo de infección de los pacientes cuando tienen una estancia de 10 es en general menor a 3.

10.

Considere los datos en la base *Stereo.csv* y las variables: $y =$ calidad del sonido (SOUND) y $x =$ costo (COST), ambos datos tomados de una muestra aleatoria de equipos de sonido.

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
   b. Para transformar la variable X, probar con transformaciones Box-Tidwell u otras conocidas como log() o exp().

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 cliente que comprará un equipo de sonido a tener una idea de la calidad de sonido que estará observando en su equipo de costo 400.