Práctica sobre otras pruebas de hipótesis y regresión

Giovanni Sanabria Brenes

1. A continuación se presentan los datos obtenidos en una investigación realizada entre estudiantes universitarios, quienes evaluaron el desempeño de alguno de sus profesores. Cada estudiante seleccionó al profesor por evaluar. Se trata de un total de 780 estudiantes:

	Alumnos	Alumnas
Profesores	269	275
Profesoras	59	177

   Con base en estos datos, ¿considera que el sexo del estudiante y el sexo del profesor evaluado son dependientes? $R/$ Valor $P < 0.01$, Si

2. Seguidamente se presentan los datos de 7 bebés sobre el número de días que tienen de nacidos (X) y su peso en kilogramos (Y):

X	2	7	14	21	30	60	90
Y	3.3	3.7	4	5	5.5	8	10

   Suponga que se cumplen las hipótesis de regresión lineal.

   (a) Encuentre una ecuación de regresión lineal para el peso de un bebé en función del número de días de nacido. $R/$ $\hat{y} = 3.162621926 + 0.077507351x$

   (b) ¿Aproximadamente cuánto es el peso promedio de un bebé recién nacido?$R/$ 3.162621926 kg

   (c) ¿Aproximadamente qué porcentaje de variación en el peso de un bebé se debe a otros factores distintos al número de días de nacido? $R/$ 0.4329112%

   (d) Encuentre un intervalo de confianza de 90% para el promedio de pesos de los bebés de 150 días de nacidos. $R/$ $]14.22845786, 15.34899115[$

3. En una pequeña empresa se ha estimado que el aumento promedio en las ventas mensuales es de 5 dólares por cada dólar invertido en publicidad. Para mejorar este valor se implementaron nuevas técnicas de publicidad desde hace unos meses y se han registrado los siguientes datos para algunos de estos meses:

Gastos en publicidad en dólares (X)	50	55	60	64	70
Ventas en dolares (Y)	2040	2070	2093	2120	2171

   Suponga que se cumplen las hipótesis de regresión lineal.

(a) Encuentre la ecuación de regresión lineal para las ventas en función de los gastos en publicidad, a partir de la implementación de las nuevas técnicas de publicidad . $R/\quad \hat{y} = 1716.656146 + 6.390365449x$

(b) A un nivel de significancia del 5%, ¿existe evidencia de que el aumento promedio en las ventas mensuales por cada dólar invertido en publicidad es mayor ahora con las nuevas técnicas en publicidad? $R/\quad$ Si . $t_{obs} = 3.058297412$

4. Se tienen las siguientes observaciones:

$X_1:$	0	1	-1	2
$X_2:$	1	0	2	3
$Y:$	15	10	20	40

Estime la ecuación de regresión de Y como función lineal de $X_1, X_2. R/\quad y = \frac{25}{4} + \frac{15}{4}x_1 + \frac{35}{4}x_2$

5. Se desea estudiar la dependencia entre el grado académico y la aceptación a una reforma tributaria en la ciudad C, para ello se tomaron los datos de 365 individuos, los cuales se resumen en la siguiente tabla:

	Primaria	Secundaria	Universidad
Aceptan la reforma	50	65	70
No aceptan la reforma	75	55	50

Puede concluirse con estos datos que la aceptación de la reforma es independiente del grado académico. $R/\quad$ Se rechaza H_0, hay evidencia en contra de la independencia.

6. En una fábrica se tiene una política de que los empleados reciban un pequeño incentivo por cada pieza elaborada (cada pieza es elaborada de forma conjunta entre los empleados). Sin embargo, se cree que desde la vigencia de la política han aumentado el número de piezas defectuosas fabricadas. Para analizas este problema se tomaran los datos del número de piezas defectuosas que se fabricaron por día (X) y el número de piezas elaboradas diarias (Y) , durante 6 días. Los datos son los siguientes:

X	7	15	11	5	20	10
Y	200	250	220	195	270	225

Suponiendo que se cumplen las hipótesis de regresión.

(a) Encuentre la ecuación de regresión lineal para el número número total de piezas elaboradas por día como función del número de piezas defectuosas que se fabrican por día. $R/\quad \hat{y} = 167.34375 + 5.234375x$

(b) Aproximadamente, ¿cuánto es el aumento promedio en el número de piezas fabricadas diarias por cada pieza defectuosa elaborada? $R/\quad$ 5.234375 piezas por cada pieza defectuosa

(c) Aproximadamente, ¿cuál es el número promedio de piezas elaboradas en un día si ninguna es defectuosa? $R/\quad$ 167.34375 piezas en promedio

(d) Cada empleado de la fábrica recibe un salario diario fijo de 30 mil colones más 100 colones por cada pieza que se fabrique en el día. Es decir, el salario diario de un trabajador es $S = 30000 + 100Y$. Determine un intervalo de confianza del 95% para el salario diario promedio de un trabajador si ese día se fabricaron tres piezas defectuosas. $R/$ $]47247.6 \quad , \quad 49361.8[$

(e) Considera apropiada la política de la fábrica. Explique.

(f) Aproximadamente, ¿qué porcentaje de variación de en el número de piezas elaboradas en un día se debe a otros factores distintos al número de piezas defectuosas diarias? $R/$ Aproximadamente el 2.19%.

(g) Encuentre un intervalo de confianza de 90% para el aumento promedio en el número de piezas fabricadas diarias por cada pieza defectuosa elaborada. $R/$ $]4.398653526, 6.070096474[$

7. Considere los datos de la siguiente tabla:

$X:$	1	2	4	5
$Y:$	7	12	22	32

A partir de estos datos, estime el coeficiente β de la ecuación de regresión $y = \beta x$ utilizando el método de mínimos cuadrados. $R/$ $6.065\,22$

8. Para los datos en la tabla, encuentre una buena ecuación que dé y como función de x. $R/$ $\widehat{y} = x/(0.0988085x - 0.510081)$

x	-10	2	5	8	11	18
y	6	-7	-48	38	21	14

9. Considere los datos de la siguiente tabla:

$X:$	2	4	5	7	10
$Y:$	110	40	20	10	2

(a) Determine una ecuación para el modelo de regresión exponencial $y = \alpha\beta^x$.
$R/$ $\widehat{y} = 280.9609532\,(0.610654856)^x$

(b) Determine un IC del 80% para α. $R/$ $]227.450358, 347.0605979[$

(c) Pruebe, a un nivel de significacia del 10%, si el porcentaje de disminución de Y, por cada unidad de X, es del 40%. $R/$ No hay evidencia en contra de que $\beta = 0.6$.

10. Se desea estimar los coeficientes β_0, β_1 y β_2 en la ecuación $z = \dfrac{x}{\beta_0 x + \beta_1 x^2 + \beta_2 y}$, para lo cual se obtiene la siguiente muestra de cuatro observaciones.

$X:$	1	-2	-4	-2
$Y:$	2	1	0	4
$Z:$	1	2	$-\frac{1}{2}$	$\frac{1}{2}$

(a) Transforme el problema en uno de regresión lineal m ultiple, y escriba la tabla de datos transformados. $R/$ $1/z = \beta_0 + \beta_1 x + \beta_2\,(y/x)\,.$

(b) Escriba el sistema de ecuaciones para el problema de regresión lineal múltiple.

$$R/ \quad \begin{pmatrix} 4 & -7 & -\frac{1}{2} \\ -7 & 25 & 7 \\ -\frac{1}{2} & 7 & \frac{33}{4} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ 4 \\ -\frac{9}{4} \end{pmatrix}$$

(c) ¿Cuáles son las estimaciones de β_0, β_1 y β_2 en el problema original? $R/$ $b_0 = 2$, $b_1 = 1$, $b_2 = -1$

11. En una encuesta se encontró la siguiente información sobre estudiantes de una universidad de distintas áreas:

	Salud	Ciencias básicas	Ciencias sociales
Hace ejercicio	21	12	7
No hacen ejercicio	99	68	48

¿hay evidencia de que la proporción de personas que realizan ejercicio varía entre las áreas de estudio indicadas? $R/$ *No*

12. La tabla siguiente presenta las notas de aprobación en matemática discreta y programación de 6 estudiantes elegidos al azar de Ingeniería en Computación.

Matemática discreta:	35	60	93	65	87	71
Programación:	50	57	95	73	91	80

(a) Encuentre la ecuación de regresión lineal para la nota de matemática discreta como función de los nota de programación. $R/$ $\widehat{y} = 0.833217351x + 17.25794479$

(b) ¿Aproximadamente, que porcentaje de variación en las notas de discreta se debe a otros factores aparte de la nota obtenida en programación? $R/$ 0.082677751

(c) Determine un intervalo de predicción del 95% para la nota en matemática discreta de un estudiante que obtuvo un 80 en programación. $R/$ $]66.04931621, 101.7813495[$

13. Se sabe que las variables x, y están relacionadas por una ecuación $y = \dfrac{x}{(\alpha - 1)\,x + (2 - \beta)}$ donde α y β son constantes. Encuentre estimaciones de α y β a partir de la siguiente tabla. $R/$ $a = 1.52296046$, $b = 3.028528529$

x	1	4	6	20
y	-2	4	3	2

14. En una encuesta sobre la soda comedor EL COMELON, se les preguntó a 200 clientes su opinión sobre la variedad de los alimentos y su nivel de ingreso. Los resultados se resumen en la siguiente tabla de contingencia.

		Nivel de ingreso		
		Bajo	Medio	Alto
Variedad	Poco	3	10	27
	Regular	15	20	50
	Mucha	21	40	14

¿Existe evidencia de que la opinión que tiene un cliente sobre la variedad de los alimentos dependiente de su nivel de ingreso? $R/$ Si, valor $P < 0.05$.

15. Un curso es impartido normalmente por tres profesores. Un estudiante considera que la nota en el curso depende del profesor que lo imparta. Ante esto la Oficina de Registro seleccionó al azar el promedio de 11 estudiantes que aprobaron el curso, los cuales se muestran a continuación

Profesor A	Profesor B	Profesor C
77	96	74
88	84	84
81	75	75
82	80	
78		

Con base a estos datos, ¿Puede afirmarse que la nota promedio en el curso no varía según el profesor que lo imparte? $R/$ Si, $f_{obs} \approx 0.758\,938$

16. Una fabrica recolectó la siguiente información de 8 de sus trabajadores sobre el número de minutos que llegaron tarde al trabajo (X) y el número de piezas defectuosas que fabricaron ese día (Y):

X	2	5	10	15	20	25	30	40
Y	2	4	9	12	15	20	24	30

Suponiendo que se cumplen las hipótesis de regresión.

(a) Encuentre la ecuación de regresión lineal para el número de piezas defectuosas que fabrica un empleado por día como función del número de minutos que llega tarde al trabajo. $R/$ $\hat{y} = a + bx = 0.732887615 + 0.749230606x$

(b) ¿Aproximadamente, ¿cuál es el número promedio de piezas defectuosas que realiza un empleado que llega puntual a su trabajo? $R/$ 0.732887615 piezas en promedio

(c) ¿Aproximadamente, ¿cuánto es el aumento promedio en el número de piezas defectuosas que realiza un empleado por cada minuto que llega tarde al trabajo? $R/$ 0.749230606 piezas por minuto

(d) Encuentre un intervalo de predicción de 95% para el número de piezas defectuosas que realiza un empleado en un día que llega 35 minutos tarde al trabajo$R/$ $]25.00546394, 28.90645371[$

(e) Si un trabajo realiza más de 20 piezas defectuosas en un día será sancionado. ¿Considera probable que si un trabajador llega 35 minutos tarde al trabajo será sancionado? Justifique su respuesta. $R/$ Si

(f) Aproximadamente, ¿qué porcentaje de variación de en el número de piezas defectuosas que elabora un empleado en un día se debe a otros factores distintos al número de minutos que llega tarde a trabajar? $R/$ Aproximadamente el 0,42%.

(g) Un trabajador tiene un salario base diario de 20 mil colones del cuál se le descuenta 200 colones por cada pieza defectuosa que realiza. Es decir, el salario diario de un trabajador es $S = 20000 - 200Y$. Determine un intervalo de confianza del 90% para el salario diario promedio de un trabajador que llega 13 minutos tarde a su trabajo.
$R/$ $]18008.15747 \quad , \quad 17802.68833[$

17. Una investigación indica que cierta enfermedad tuvo su aparición en los primeros días del año 1998. Seguidamente se presenta algunos datos sobre el porcentaje de personas infectadas de la enfermedad en un país C, X años después del 1° de enero del 1998:

Años (X)	0.2	2	5	10	12
Porcentaje (Y)	3%	12.7%	13.4%	13.9%	14%

(a) Justifique por qué el modelo Hiperbólico es un buen modelo de regresión para explicar la relación entre las variables X,Y.

(b) Encuentre la ecuación de regresión del modelo Hiperbólico para Y como función de X. $R/ \quad \widehat{y} = \dfrac{x}{0.062475762x + 0.053998736}$

(c) Encuentre un intervalo de confianza de 90% para α (uno de los parámetros del modelo hiperbólico**,** según la notación vista en el curso) $R/$ $]0.053623507, \quad 0.071328017[$

(d) Encuentre un intervalo de predicción del 95% para el porcentaje de personas del país C que tendrán la enfermedad el 1° de enero del 2011. $R/$ $]10.83451446, 24.41323233[$

18. Un profesor de Estudios Sociales considera que la nota obtenida en su examen final (Y) depende linealmente del número de horas que durme un alumno en la noche antes del examen (x_1) y del número de horas de estudio en la semana previa al examen (x_2). Seguidamente se presentan los datos obtenidos de 4 estudiantes

x_1	4	1	6	8
x_2	4	8	12	15
Y	45	60	80	96

(a) Estime los coeficientes β_0, β_1 y β_2 en la ecuación de regresión lineal múltiple $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. $R/ \quad y = 24.533 + 0.770\,285\, x_1 + 4.313\,65 x_2$

(b) De acuerdo a la ecuación estimada, ¿Qué tiene más peso en la nota: las horas de estudio ó las horas que duerme la noche antes del examen? $R/$ Las horas de estudio.

19. Considere las siguientes observaciones:

x	0	1	2	3
y	5	3	2	1

Estime los coeficientes β_0, β_1 y β_2 en la ecuación $y = \dfrac{1}{\beta_0 + \beta_1 x + \beta_2 x^2}$. $R/ \quad b_0 = 0.214\,995$, $\quad b_1 = -0.018\,355$, $\quad b_2 = 0.091\,675$

20. En el año 1998 inicio el correo basura ("spam") que circula por las redes de un país C. Seguidamente se presenta algunos datos sobre el porcentaje Y de correos Spam que han circulado X años después del 1° de enero del 1998:

Años (X)	0.5	2	5	8	10
Porcentaje (Y)	1%	15%	29%	29.7%	29.9%

(a) Seleccione y justifique un modelo de regresión del porcentaje de correos Spam en función del número de años después de 1998.
$R/$ Hiperbólico

(b) Encuentre la ecuación de regresión del modelo seleccionado en la parte (**a**). $R/$ $y = 1029.977976\,(0.628744589)^{x}$

(c) Encuentre un intervalo de confianza de 90% para α y β, parámetros del modelo escogido en **a.**

	Extremos del IC para α		Extremos del IC para β	
$R/$	izquierdo:	846,4924561	izquierdo:	0,580541892
	derecho:	1253,235777	derecho:	0,68094958

(d) Encuentre un intervalo de confianza del 95% para el porcentaje esperado de correos Spam que han circulado hasta el 1° de enero del 2010

21. Se extraen tres cartas de una baraja ordinaria, con reemplazo, y se registra el número Y de espadas. Después de repetir el experimento 64 veces se registran los siguientes resultados

y	0	1	2	3
Frecuencia	21	31	12	0

¿existe evidencia en contra, con un nivel de significancia de 0.01, de que los datos se pueden ajustar a una distribución binomial $P(y) = C(n, y)\left(\frac{1}{4}\right)^{y}\left(\frac{3}{4}\right)^{n-y}$ para $y = 0, 1, 2, 3$?.$R/$ No, $X_{obs} \approx 2.325\,9$

22. Considere la siguiente tabla de datos, donde x es el número total de años de educación (primaria, secundaria y técnica o universitaria) y y es el ingreso anual en dólares de varias personas.

x	4	4	6	6	8	8	10	12	12	14
y	8280	10516	9212	11744	12405	14664	15336	16908	18347	19512

Encuentre un intervalo de predicción de 95% para el salario anual de una persona con once años de educación $R/$ $]13578.1921, 19312.6079[$

23. Se desea realizar un estudio que relacione el tiempo n en meses de entrenamiento de un estudiante que participará en la Olimpiada Internacional de Matemáticas (OIM) y el tiempo $t$en minutos que tardará resolviendo un problema de la OIM. Según un entrenador, el tiempo mínimo que tarda un estudiante muy entrenado en resolver este tipo de problemas es de 30 minutos. Se obtuvieron los siguientes datos de ambas variables:

Tiempo de entrenamiento (n)	2	3	5	10	20
Tiempo de resolución de un problema (t)	280	195	130	80	50

(a) Justifique por qué el modelo Recíproco es un buen modelo de regresión para t con funcón de n.
$R/$ Recíproco

(b) Encuentre la ecuación de regresión del modelo Recíproco para t como función de n. $R/$ $\widehat{y} = 27.29825291 + \dfrac{505.78203}{x}$

(c) ¿Cuántos meses de entrenamiento se necesitan para esperar resolver un problema de este tipo en dos horas o menos? $R/$ Aproxidamente, al menos 5.46 meses.

(d) Encuentre un intervalo de confianza de 90% para β (uno de los parámetros del modelo Recíproco, según la notación vista en el curso) $R/$]491.6619096, 519.9021503[

(e) Encuentre un intervalo de predicción 90% para el tiempo que tarda un estudiante resolviendo un problema de la IMO si tuvo cuatro meses de entrenamiento.
$R/$]148.0878732, 159.3996476[

(f) Una prueba de la IMO consta de 3 problemas y tiempo de resolución es de 4 horas 30 minutos. Si un estudiante tuvo cuatro meses de entrenamiento, ¿considera que pueda resolver un problema de la prueba?. $R/$ Si

24. Considere los datos de la siguiente tabla:

$X:$	2	3	4	5
$Y:$	3	12	28	45

Se sabe que las variables x y y están relacionadas por la ecuación $y + 5 = 2x^{\beta}$, donde β es una constante desconocida. Use los datos en el cuadro anterior para estimar β, transformando la ecuación en una lineal y usando el criterio de mínimos cuadrados. $R/$ 1.99675.

25. Se desea estimar los coeficientes β_0, β_1 y β_2 en la ecuación $z = \dfrac{\beta_0 x + \beta_1 + \beta_2 y}{x}$, para lo cual se obtiene la siguiente muestra de cuatro observaciones.

$X:$	1	0.5	0.25	0.0625
$Y:$	1	2	3	4
$Z:$	5	15	43	228

(a) Transforme el problema en uno de regresi on lineal m ultiple, y escriba la tabla de datos transformados. $R/$ $z = \beta_0 + \beta_1 (1/x) + \beta_2 (y/x)$

(b) Escriba el sistema de ecuaciones para el problema de regresión lineal múltiple.

$$R/ \quad \begin{pmatrix} 4 & 23 & 81 \\ 23 & 277 & 1081 \\ 81 & 1081 & 4257 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 291 \\ 3855 \\ 15\,173 \end{pmatrix}$$

(c) ¿Cuáles son las estimaciones de 0, 1 y 2 en el problema original? $R/$ $b_0 \approx 3.30165$, $b_1 \approx -2.39463$, $b_2 \approx 4.1095$

26. Actualmente, en el Ministerio de Transportes del país C, se analiza la posibilidad de la apertura en el examen teórico para obtener la licencia de conducir. Existen tres compañías en el extranjero interesadas, con experiencia en la elaboración y aplicación de este tipo de pruebas, las cuales

tienen una escala de 0 a 100. Se compararon los resultados de ciudadanos en las pruebas elaboradas por estas compañías:

Compañía		
C1	C2	C3
60	95	74
40	80	84
79	30	53
		47

A un nivel de significación del 5%, ¿hay evidencia en contra de que los promedios en las pruebas de las tres compañías son iguales? *R/ No*