UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

MAESTRÍA EN CIENCIAS DE LA INFORMACIÓN Y LAS TELECOMUNICACIONES

Herramientas matemáticas para el manejo de la información

Taller #3

Salazar Ortiz, Jaiver
jesalazaro@correo.udistrital.edu.co
20221495012

Forero Castro, Diego
ddforefoc@correo.udistrital.edu.co
20221495005

Angel Villarreal, Sergio
slangelv@correo.udistrital.edu.co
20221395001

FECHA: *16 de enero de 2024*
FECHA DE ENTREGA: *16 de enero de 2024*

1. PRIMERA PARTE

PUNTO 1

Un Ingeniero Catastral y Geodesta, quiere establecer una relación lineal, en la que el valor del metro cuadrado de unos inmuebles sea explicada por las variables distancia (en metros) al centro comercial cercano y valor comercial del inmueble, con un 95

N	Valor de metro cuadrado.	Distancia al centro comercial.	Valor comercial inmueble.
1	101423	96	49429499
2	115277	94	51305103
3	122570	102	51099623
4	125809	106	50772703

SOLUCIÓN 1

Realizando el proceso en excel se tiene lo que se observa en la figura 1.1.

	A	B	C	D	E	F	G	H	I
1	Resumen								
2									
3	*Estadísticas de la regresión*								
4	Coeficiente d	0.99906983							
5	Coeficiente d	0.99814053							
6	R^2 ajustado	0.9944216							
7	Error típico	809.148648							
8	Observacione	4							
9									
10	ANÁLISIS DE VARIANZA								
11		*Grados de libert*	*ma de cuadrad*	*dio de los cua*	*F*	*alor crítico de F*			
12	Regresión	2	351447257	175723629	268.394453	0.04312152			
13	Residuos	1	654721.535	654721.535					
14	Total	3	352101979						
15									
16		*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
17	Intercepción	-440235.04	28237.7777	-15.5902863	0.04077852	-799030.025	-81440.0554	-799030.025	-81440.0554
18	Variable X 1	1242.20729	86.324857	14.3899142	0.04416968	145.345978	2339.06859	145.345978	2339.06859
19	Variable X 2	0.0085467	0.00056345	15.1684518	0.04190934	0.00138736	0.01570604	0.00138736	0.01570604
20									
21									
22									
23	Análisis de los residuales								
24									
25	*Observación*	*onóstico para*	*Residuos*						
26	1	101475.966	-52.9655358						
27	2	115021.776	255.224043						
28	3	123203.258	-633.25829						
29	4	125378	430.999783						
30									

Figura 1.1: Datos obtenidos mediante excel

PUNTO 2

El vector de datos de la variables dependientes es:

La variable dependiente es:	Valor metro cuadrado			
Los datos son:	101423	115277	122570	125809

SOLUCIÓN 2

PUNTO 3

La matriz X de la variables independientes e intercepto son:

SOLUCIÓN 3

Intercepto	Distancia centro comercial	Valor comercial inmueble
1	96	49429499
1	94	51305103
1	102	51099623
1	106	50772703

PUNTO 4

Completar las siguientes tablas:

SOLUCIÓN 4

.

R^2	0.9981					
ANOVA	Grados de libertad	Suma de cuadrados	Cuadrados medios	F	Valor crítico de F	Valor P
Regresiónl	2	351447257	175723628.5	268.39	199.5	0.043122
Residuos	1	654721.5	654721.5			
Total	3	352101978.5				

PUNTO 5

En el contexto del problema, la tabla ANOVA indica que _____

SOLUCIÓN 5

.

En el contexto del problema, la tabla ANOVA indica que, a un nivel de significancia del 95 %, las variables distancia centro comercial y valor comercial inmueble son significativas para explicar el valor del metro cuadrado , es decir, brindan la información necesaria para estimar el valor del metro cuadrado del inmueble.

Parametros	Coeficientes	Inferior 95 %	Superior 95 %
Intercepto	-440200	-799030	-81440.06
Distancia centro comercial	1242	145.3460	2339.069
Valor comercial inmueble	0.008547	0.001387359	0.01570604

PUNTO 6

En el contexto del problema, R^2 indica

SOLUCIÓN 6

En el contexto del problema, R cuadrado indica que casi el 100 % de la variabilidad alrededor de la media de la variable respuesta (valor del metro cuadrado) es explicada por el modelo, es decir, las dos variables explicativas (distancia centro comercial y valor comercial inmueble), esto significa, que el modelo se ajusta muy bien a los datos, es decir, es un buen modelo

PUNTO 7

La relación lineal es:_____

SOLUCIÓN 7

es: valor metro cuadrado= -440200+1242* distancia centro comercial + 0.008547* valor comercial inmueble

PUNTO 8

Si un inmueble tiene como valor comercial 50000000 y se encuetra a 100 metros del centro comercial, se espera que el metro cuadrado m^2 valga:_____ y un intervalo de confianza es _____

SOLUCIÓN 8

Si un inmueble tiene como valor comercial 50000000 y se encuetra a 100 metros del centro comercial, se espera que el metro cuadrado valga 111320.7 y un intervalo de confianza es (98864.69;123776.7)

PUNTO 9

Los residuales encontrados son:

SOLUCIÓN 9

RESIDUALES	-52.96554	255.22404	-633.25829	430.99978

- ¿Existe evidencia estadística para decir que los residuales son normales?. Indicarlo a través de la tabla:?

PRUEBA USADA	Shapiro-Wilk
HIPOTESIS	H_0: los valores de los residuales son una muestra aleatoria simple de una distribución normal. H_1: los valores de los residuales no son una muestra aleatoria simple de una distribución normal.
SIGNIFICANCIA	95 %
ESTADISTICO DE PRUEBA EVALUADO	$W = 0{,}93567$
CRITERIO DE DECISION	Si W>W_0, entonces se rechaza H_0, donde W_0 es un valor que se obtiene de la tabla de la distribución del estadístico Shapiro-Wilk, en este caso W_0=0.992 (para un Alpha del 95 %)
CONCLUSION	Como no hay evidencia estadística suficiente para rechaza H0 (W<W0) entonces se concluye que los valores de los residuales provienen de una distribución normal.

PUNTO 10

En el caso que los residuales sean normales, indicar su distribución de probabilidad especificando sus parámetros:_____

SOLUCIÓN 10

Media: 0, Varianza: 218240.5, que se obtienen de la media muestral y la varianza muestral insesgada de los residuos, ya que estos son estimadores insesgados de la media y varianza poblacionales.

2. SEGUNDA PARTE

PUNTO 1

Usando la técnica de Bootstrapping estimar el valor esperado de exp(X) y la probabilidad de que X>3, sabiendo que para X solo se cuenta con la siguiente muestra 3,3,2,2,1,5,4,4,3,2. En cada respuesta, además de la estimación, se debe incluir confianza y error.

SOLUCIÓN 1

.
Para estimar el valor esperado de exp(x) se usa el siguiente código en Rstudio:

```
valores=c(3,3,2,2,1,5,4,4,3,2)
set.seed(200) # Setting the seed for replication purposes
sample.size <- 10 # Sample size
n.samples <- 20000 # Number of bootstrap samples
bootstrap.results <- c()
for (i in 1:n.samples)
{
  obs <- sample(1:sample.size, replace=TRUE)
  bootstrap.results[i] <- mean(exp(valores[obs]))
}
length(bootstrap.results)
summary(bootstrap.results)
sd(bootstrap.results)
hist(bootstrap.results, # Creating an histogram
     col="#d83737", # Changing the color
     xlab="Mean", # Giving a label to the x axis
     main=paste("Means of 2000 bootstrap samples"))
quantile(bootstrap.results, c(0.025, 0.975))
```

Lo cual genera el gráfico ilustrado en la figura 2.1, además de ello podemos obtener el valor esperado, error, e intervalos de confianza mediante los siguientes códigos:

```
> quantile(bootstrap.results, c(0.025, 0.975))
    2.5%    97.5%
13.27022 63.00737
> sd(bootstrap.results)/sqrt(2000)
[1] 0.2981887
> summary(bootstrap.results)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.725  24.281  32.561  34.061  42.914  89.882
```

Por tanto se concluye que el valor esperado es 34.061, con intervalos de confianza 13.27022 63.00737, y error 0.2981887

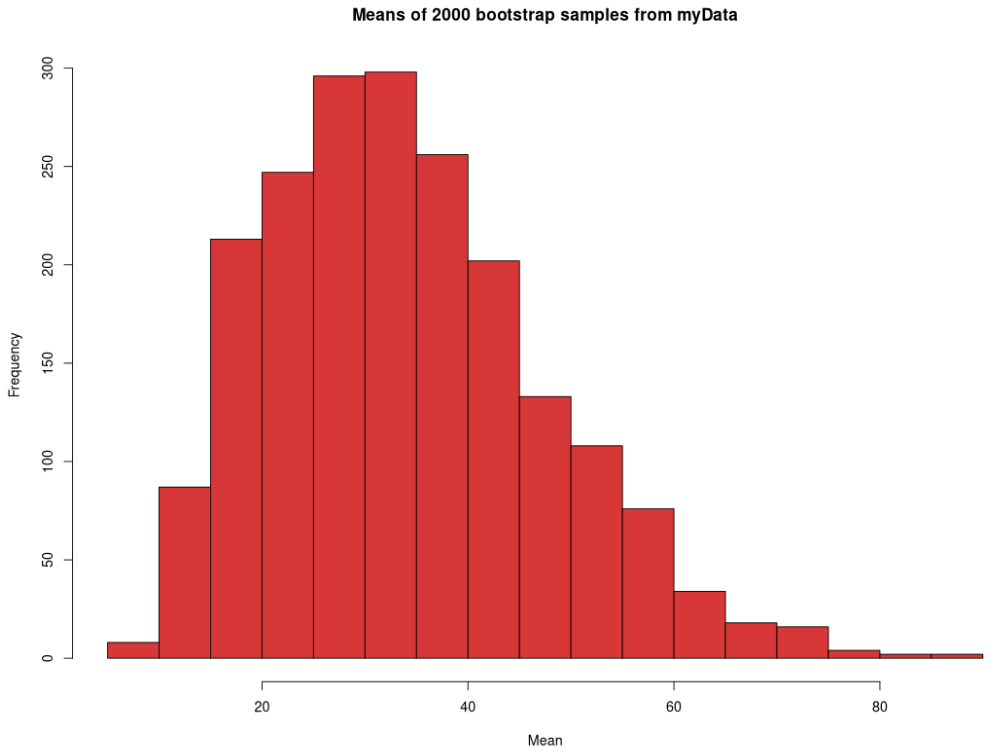

Figura 2.1

Para estimar el valor de la probabilidad de que x>3 se usa el siguiente código en Rstudio:

```
valores=c(3,3,2,2,1,5,4,4,3,2)
set.seed(200) # Setting the seed for replication purposes
sample.size <- 10 # Sample size
n.samples <- 20000 # Number of bootstrap samples
bootstrap.results <- c()
for (i in 1:n.samples)
{
  obs <- sample(1:sample.size, replace=TRUE)
  bootstrap.results[i] <- mean(valores[obs] > 3)
}
length(bootstrap.results)
summary(bootstrap.results)
sd(bootstrap.results)
hist(bootstrap.results, # Creating an histogram
     col="#d83737", # Changing the color
     xlab="Mean", # Giving a label to the x axis
     main=paste("Means of 2000 bootstrap samples"))
```

Lo cual genera el gráfico ilustrado en la figura 2.2, además de ello podemos obtener el valor esperado, error, e intervalos de confianza mediante los siguientes códigos:

```
> quantile(bootstrap.results, c(0.025, 0.975))
 2.5% 97.5%
  0.4    1.0
> sd(bootstrap.results)/sqrt(2000)
[1] 0.00325584
> summary(bootstrap.results)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1000  0.6000  0.7000  0.6997  0.8000  1.0000
```

Por tanto se concluye que el valor de probabilidad para x>3 es 0.6997, con intervalos de confianza 0.4 1.0 , y error 0.00325584.

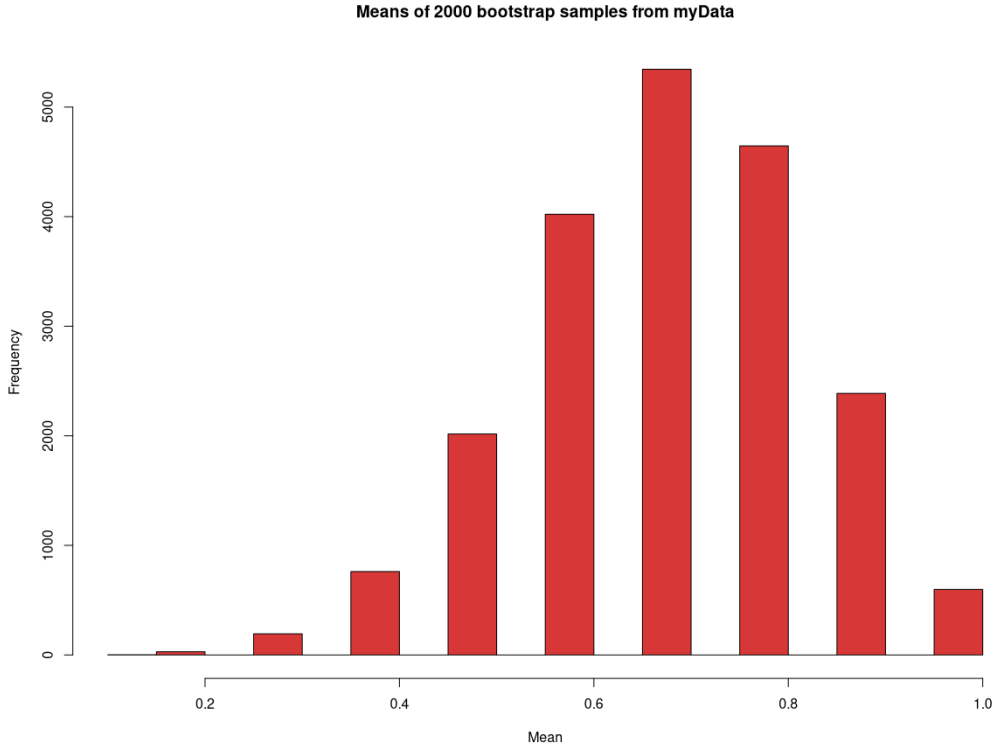

Figura 2.2

PUNTO 2

Usando la técnica MCMC Gibbs Sampler, estimar el valor esperado de la variable X, el valor esperado de Y, el valor esperado de COS(X)+Y, la correlación de X e Y y la probabilidad conjunta de que X>1e Y >0,6. Incluir confianza y error en cada estimación. Asuma que Y binomial con 5 ensayos y proporción X e X es uniforme continua (0,1)

SOLUCIÓN 2

En este caso, se busca estimar el valor esperado de X, el valor esperado de Y, el valor esperado de cos(X)+Y, la correlación de X e Y y la probabilidad conjunta de que X >0.6 e Y >1, usando la técnica MCMC Gibbs Sampler y suponiendo Y binomial con 5 ensayos y proporción X. Además, se supone X con distribución uniforme continua (0.1). Por lo tanto, se considera el siguiente par de distribuciones:

$$\begin{cases} Y|X \sim & Bin(n, X) \\ X \sim & Uni(a, b) \end{cases}$$

donde n=5,a=0 y b = 1 Además, se conoce que la distribución uniforme en el intervalo [0.1] es un caso particular de la distribución beta, con parámetros a = 1 y 4 =1. Por lo tanto, la ecuación (7) se puede reescribir como:

$$\begin{cases} Y|X \sim & Bin(n, X) \\ X \sim & Uni(\alpha, \beta) \end{cases}$$

donde n=5, $\alpha = 1$ y β=1, Por otra parte, se tiene que:

$$X|Y \sim Be(y + \alpha, n - y + \beta) \tag{2.1}$$

Una vez se han formulado las distribuciones conjuntas $X|Y$ e $Y|X$, so realizan los cálenlos implementando MCMC Gibbs sampler en RStudio. Para tal fin, se desarrolló el siguiente código:

```
##Se definen los parametros iniciales
Nsim = 10000;
n = 5;
a = 1;
b = 1;
Y = X = array(0, d n = c(Nsim,1));
X[1] = rbeta(1,a,b)
Y[1] = rbinos(1,n,X[1]);
## Se obtienen las cadenas de Markov usando Gibbs sampler
for (i in 2:1Msim)
{Y[i] = rbinom(1,m,X[i - 1])
X[i] = rbeta(i,a + X[i],n - X[i] +b)}
ts. plot(x)## Se grafican los resultados
ts.plot(Y)
E_X = mean(X) ## Valor esperado de X, error e intervalo
error_X = 2*sqrt(var (1)/Msim) ## Confianza del 95%
inf_X= E_X - error_X
Sup_x = E_X + error_X
## Valor esperado de X, error e intervalo
E_Y = mean(Y)
error_Y = 2*sqrt(var (1)/Msim) ## Confianza del 95%
inf_Y= E_Y - error_Y
Sup_Y = E_Y + error_Y
## Valor esperado de cos(X)+Y, error e intervalo
E_cosX_mas_Y = mean(cos(X) + Y)
error_cosX_mas_Y = 2*sqrt(var((cos(x)) + Y)/Nisim) ## 95%
inf cosX_mas_Y = E_cosX_masY - error_cosX_mas_Y
sup cosX_mas_Y = E_cosX_masY + error_cosX_mas_Y
## Se calcula la correlaci n
correlaci n = cor(X,Y)
## Probabilidad conjunta, error e intervalo
prob = length(which((X > 0.6) & (Y > 1)))/Nsim
error_prob = 2*sqrt(prob*(1 - prob)/Nsim) #
#95%, para proporcion
inf_prob = prob - error_prob
sup_prob = prob + error_prob
```

Los comentarios en el código describen el procedimiento llevado a. cabo en el MCMC Gibbs sampler. La Figuras 6a y 6b presentan las cadenas de Markov obtenidas en el

proceso para las variables A e Y, respectivamente. Finalmente, la Tabla 3 presenta el resumen de los resultados obtenidos para las estimacio- nes usando MCMC Gibbs sampler. En general, se observan errores pequeños en las esti ciones, dado que se generó un número considerable de valores para la simulación (Nsim. 10000, para cada cadena de Markov). Además, se observa muy poca (casi cero) correlación