Universidad Simón Bolívar
C03321
Trimestre: Ene-Mar 2018
Estadística para Ingenieros.
Cristopher Rodríguez 14-10944
Germán Robayo 14-10924
Yuni Quintero 14-10880

CAPACITACIÓN

Se realizó un estudio estadístico sobre la base de datos de 300 aspirantes que forman parte de un programa de inserción laboral para personas que hayan desertado del sistema de educación preparatoria formal, con el objetivo de describir e interpretar los datos recopilados. Cada observación contiene información sobre el puntaje en la prueba de capacitación aplicada al final del programa, el ingreso en el hogar del aspirante como porcentaje por encima del nivel de pobreza, nivel educativo efectivo del responsable del hogar del aspirante, número de faltas de disciplina del aspirante durante sus dos últimos años en la escuela, número de semanas transcurridas entre la deserción de la escuela y el inicio del programa de capacitación, número de empleos ocupados desde la deserción de la preparatoria, porcentaje de asistencia durante los últimos dos años en la escuela, ingreso promedio mensual del aspirante en sus anteriores empleos y la zona donde reside el aspirante. Con estas variables se propuso realizar un análisis descriptivo de los datos y determinar el comportamiento de cada variable en función de las demás.. Se estipuló el modelo lineal que mejor explica el puntaje en la prueba de capacitación en función de las demás y su poder predictivo. Finalmente se determinó si existen diferencias entre los puntajes medios de los aspirantes según la clasificación del ingreso de sus anteriores empleos y la zona.

Entre los resultados más relevantes destacan un puntaje promedio de 557,3±262,8293ptos. Las variables que mejor ser relacionan con JTS son ING, EDU, PAS y DIS. Además, no hay suficiente evidencia estadística para afirmar que JTS sigue una distribución normal. El modelo que mejor explica la variable JTS viene representado por: y = -287.7378 + 110.7998(ING) + 8.3967(EDU) – 11.0904(DIS) + 7.9451(PAS). Sin embargo, este no posee un poder predictivo razonable debido a una gran diferencia promedio entre los valores observados y la predicción del modelo. Finalmente se puede afirmar que, a partir de los datos proporcionados, los puntajes medios de los aspirantes según el tipo de empleo y zona no difieren entre sí.

PLANTEAMIENTO DEL PROBLEMA

Según Wackerly, Mendenhall y Scheaffer (2008), la estadística es la teoría de la información cuyo objetivo es la inferencia de las propiedades de la muestra que se considera. Esta ciencia es utilizada en muchas áreas de trabajo actualmente debido a que permite obtener información un poco más realista que para fines prácticos resulta funcionar bastante bien en la mayoría de los casos. En este curso estaremos trabajando con dos enfoques de la estadística: El enfoque descriptivo y el enfoque inferencial.

Para este caso, se realizó un proyecto con el objetivo principal de poner en práctica la teoría vista durante el curso y terminar de afianzar los conocimientos adquiridos. El caso de estudio son los datos de 300 aspirantes que forman parte de un programa de inserción laboral para personas que hayan desertado del sistema de educación preparatoria formal. Cada observación contiene información sobre el puntaje en la prueba de capacitación aplicada al final del programa (JTS), el ingreso en el hogar del aspirante como porcentaje por encima del nivel de pobreza (ING), nivel educativo efectivo del responsable del hogar del aspirante (EDU), número de faltas de disciplina del aspirante durante sus dos últimos años en la escuela (DIS), número de semanas transcurridas entre la deserción de la escuela y el inicio del programa de capacitación (SEM), número de empleos ocupados desde la deserción de la preparatoria (EMP), porcentaje de asistencia durante los últimos dos años en la escuela (PAS), ingreso promedio mensual del aspirante en sus anteriores empleos (IPR) y la zona donde reside el aspirante (zona).

Para poder alcanzar el objetivo principal, se deben cumplir ciertos objetivos específicos representados como sub-problemas a resolver durante el desarrollo de este proyecto.

Primero que todo, se trabajará en el enfoque descriptivo de la estadística haciendo un análisis de este tipo sobre la muestra. Para esto, se deben conocer las variables y en caso de que estas sean cuantitativas se deben calcular parámetros como el mínimo, máximo, media, mediana, etc. Además, a partir de sus histogramas, se interpretará la información que se refleja en los resultados obtenidos. Posteriormente, se realizará un gráfico de dispersión de las variables y su matriz de correlación para inferir el comportamiento de cada variable en función de las demás. Además se realizará una prueba de bondad de ajuste para determinar el comportamiento de la variable JTS. Aunado a esto, se determinará el modelo lineal que mejor explique la variable JTS en función del resto de variables cuantitativas realizando distintos modelos y descartando aquellas variables que no sean significativas para el mismo. Luego, se estudiará el poder predictivo de este a partir de un histograma, diagrama de cajas y resumen estadísticos de los residuos de predicción. Finalmente se determinará las medias de una nueva variable cualitativa que clasifique los

aspirantes según la variable IPR y, a partir de un análisis de varianzas, se estudiará si existen diferencias entre esta nueva clasificación y la zona.

Para realizar estas actividades y cumplir los objetivos prácticos se emplearon las fórmulas y procedimientos de la estadística descriptiva para calcular parámetros a partir de los datos obtenidos en una muestra y posteriormente se analizaron los resultados obtenidos para dar una interpretación de los mismos. Adicionalmente, se utilizaron el lenguaje de programación R y el software RStudio como herramientas de apoyo para generar las tablas y los gráficos correspondientes a cada etapa del desarrollo de este proyecto. Los mismos serán expuestos en este informe para generar una mejor comprensión del análisis y las interpretaciones que serán expuestas.

DESARROLLO

1. Análisis descriptivo de las variables

Tabla 1. Análisis descriptivo completo para cada variable

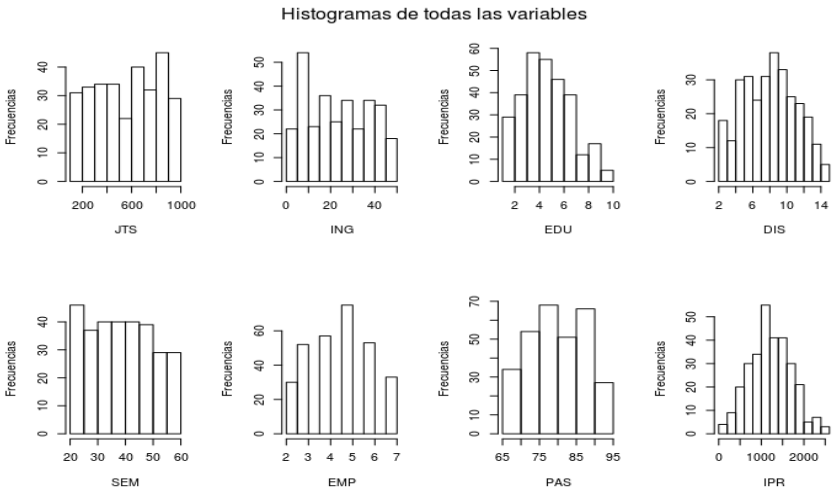
	JTS	ING	EDU	DIS	SEM	EMP	PAS	IPR	Zonas	Frec.
Mínimo	101	1	1	2	20	2	65	97	A	86
Primer Cuartil	336	10	4	6	29	3	74,41	844,1	B	80
Mediana	581,5	24	5	9	39	5	79,67	1188,58	C	67
Promedio	557,3	24,28	5,08	8,48	39,14	4,56	79,82	1213,16	D	67
Tercer Cuartil	795,5	38	6	11	49	6	86,17	1534,34		
Máximo	988	50	10	15	60	7	94,98	2527,92		
Desv. Est.	262,8293	14,02255	2,021688	3,151089	11,309	1,487877	7,56542	496,6928		

En la **Tabla 1** se presenta el análisis descriptivo global de los datos de 300 aspirantes que forman parte de un programa de inserción laboral. En la misma se pueden observar las 9 variables (8 cuantitativas y 1 cualitativa) que fueron consideradas para el estudio:

Puntaje en la prueba de capacitación JTS

Para esta variable (JTS), se puede observar que sus valores están en el rango de 101 a 988, con una media de 557.3 puntos, que está por debajo de la mediana. El primer cuartil tiene un valor de 336, lo cual indica que aproximadamente 25% de los datos son menores a 336. La mediana indica que el 50% de los datos son menores a 581.5, y, el tercer cuartil, indica que el 75% de los datos son menores a 795.5. Los datos están muy bien distribuidos, no hay preferencia por un valor específico. Hay una desviación estándar medianamente alta (262.8293), permitiendo inferir que gran cantidad de los aspirantes obtuvieron un puntaje por encima o por debajo de la media.

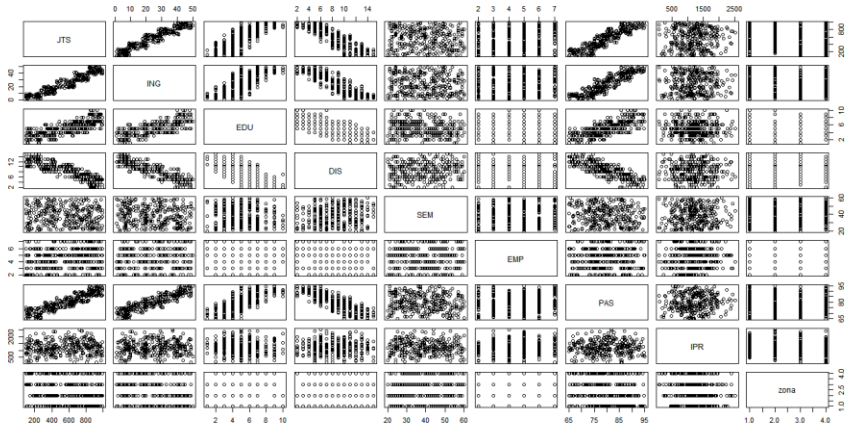
Grafica 1. Histogramas de Frecuencias de las variables



En la **Gráfica 1** se encuentran los histogramas de frecuencias para todas las variables cuantitativas. En el histograma del JTS, las frecuencias son muy similares, lo cual concuerda con nuestra aseveración de que esta uniformemente distribuidos. En la variable ING, gran parte de los aspirantes están entre 0% y 10% del nivel de la pobreza. En el caso de la variable EDU, la mayoría de los responsables del hogar del aspirante no superan los 7 años de educación. Para la variable DIS se afirma que casi todos los aspirantes obtuvieron más de 5 faltas de disciplina. Para SEM existe un hay cercanía entre cada intervalo. La variable EMP se comporta muy similar a la variable DIS al igual que PAS e IPR.

2. Comportamiento entre variables

Gráfica 2. Dispersión de las variables



En la **Gráfica 2** se presenta la dispersión de las variables cuantitativas. Las variables que presentan relación lineal con JTS son ING, EDU y PAS con pendiente positiva y DIS con pendiente negativa. El resto de variables poseen mucho ruido por lo que no presentan una relación lineal con JTS.

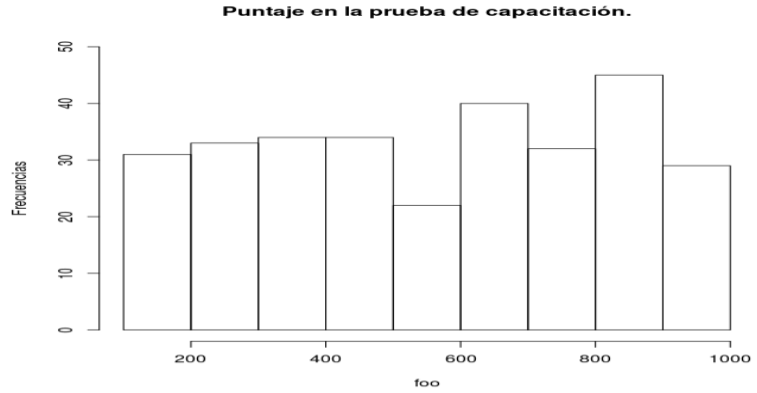
Tabla 2. Correlación de la variable JTS con las demás variables

	ING	EDU	DIS	SEM	EMP	PAS	IPR
JTS	0,944	0,754	-0,9	-0,1	0,119	0,9	0,044

En la **Tabla 2** se presenta la correlación que existe entre la variable JTS y el resto de variables cuantitativas. Las variables con mayor correlación con JTS son ING, DIS y PAS siendo mayor o igual a 0.9, seguido por EDU con 0.754. El resto de variables poseen una muy baja correlación siendo la más alta de ellas EMP con 0.119.

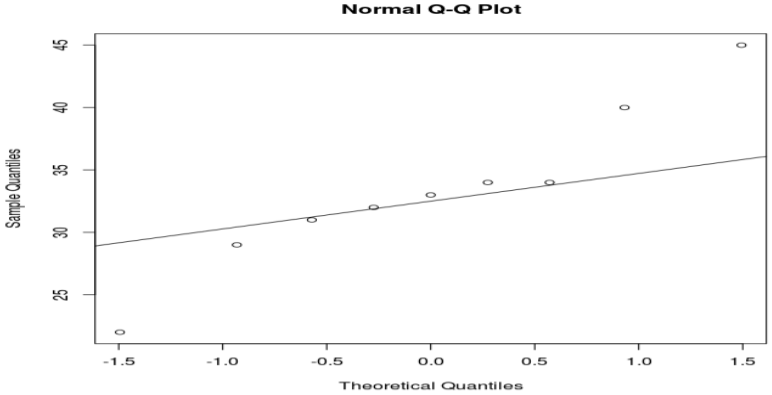
3. Comportamiento de JTS

Grafica 3. Histograma de Frecuencias del puntaje en la prueba de capacitación (JTS)



En la **gráfica 3** se tiene la tabla de frecuencias de los puntajes en la prueba de capacitación (JTS). Se observa que los datos no poseen simetría respecto a un eje ni similar al comportamiento de la curva normal.

Grafica 3. Comportamiento normal de la variable JTS



En la **gráfica 3** se observa el comportamiento normal esperado (representada por la recta). Sin embargo, se observa que los datos no siguen el patrón de la recta.

Realizando una prueba de bondad y ajuste con un nivel de significancia de 95%, se obtuvo un p-valor $1.030287e^{-13}$ el cual permite interpretar que no hay evidencia suficiente para afirmar que los datos de JTS siguen distribución Normal.

4. Modelo Lineal

Siguiendo un proceso de descarte para conseguir el modelo lineal que mejor explica la variable JTS, el modelo que cumple lo anterior es:

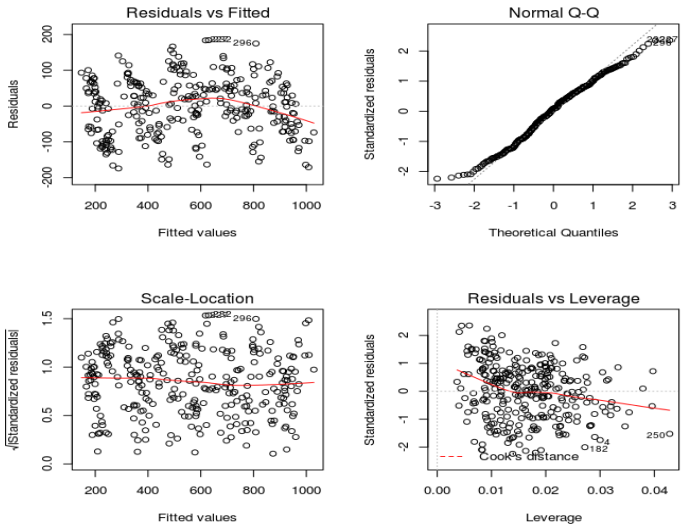
$$y = -287.7378 + 110.7998(ING) + 8.3967(EDU) - 11.0904(DIS) + 7.9451(PAS)$$

Tabla 3. Resumen modelo lineal

	Estimado	Error Standard	Valor de t	Pr(>|t|)	
Intercepto	-287,7378	107,8054	-2,669	0,008028	**
ING	10,7998	0,8792	12,284	< 2e-16	***
EDU	8,3967	3,4668	2,422	0,016036	*
DIS	-11,0904	2,823	-3,929	0,000106	***
PAS	7,9451	1,4076	5,644	3,89E-08	***

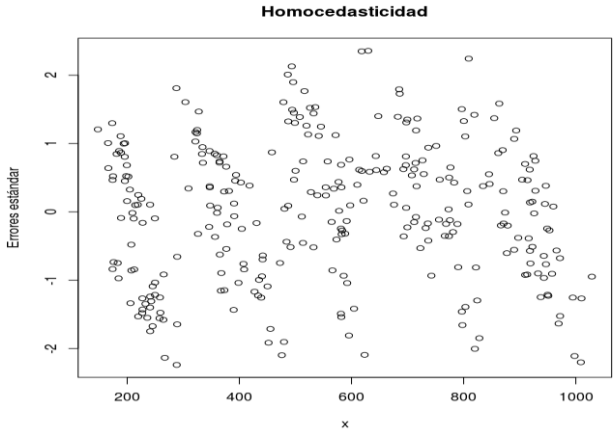
En la **Tabla 3** se observa el resumen del modelo utilizado. Este modelo es el que mejor ajusta la variables JTS debido a que mantiene un R^2 ajustado alto con un valor de 91.11%. Además, durante el proceso de descarte este modelo solo incluye las variables que en el software R se consideran como significativas para el ajuste.

Gráfica 4. Estudio de Residuos del Modelo



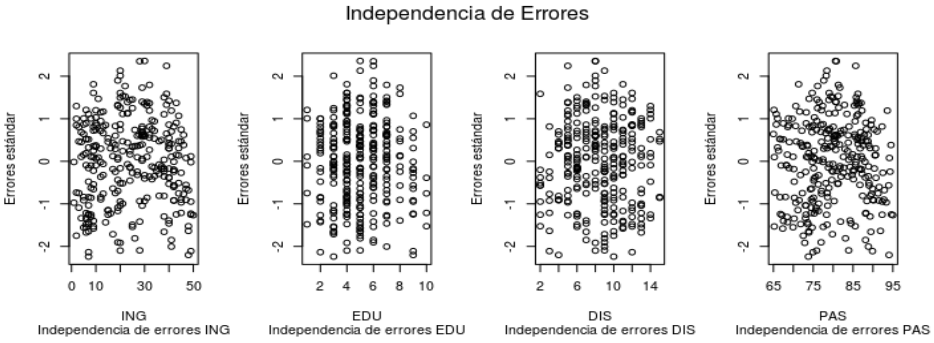
En la **Gráfica 4** se presenta el estudio de residuos del modelo. En el primer y tercer recuadro se puede observar que la comparación entre los residuos, los resultantes del modelo y la estandarización de residuos no sigue ningún patrón. Además, en el segundo recuadro se verifica la normalidad de los errores y podemos afirmar que los residuos del modelo poseen una distribución normal debido a la cercanía de los datos a la recta.

Grafica 5. Homocedasticidad



En la **Gráfica 5** se presenta la Homocedasticidad de los datos. Se observa que no existe patrón alguno, lo cual implica que las varianzas son constantes.

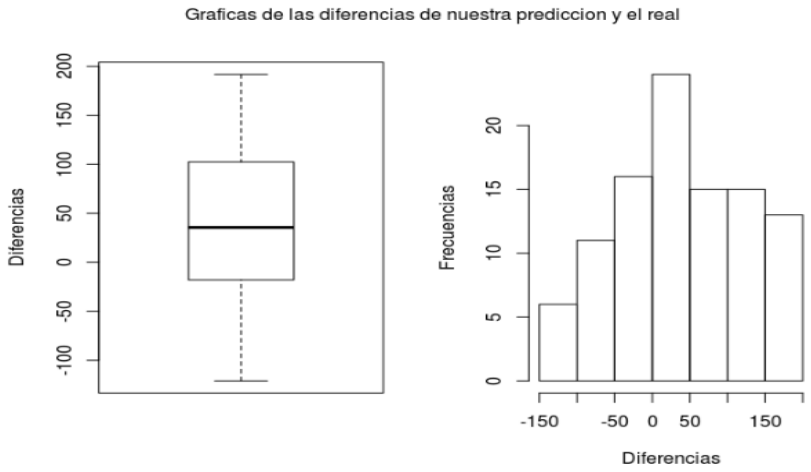
Gráfica 6. Independencia de Errores



En la **Gráfica 6** se comprueba que no existe patrón alguno en ningún recuadro, concluyendo en efecto la independencia de errores de cada variable presente en el modelo

5. Predicción de la variable JTS

Gráfica 7 y 8. Boxplot e Histograma



Las **Gráficas 7 y 8** permiten afirmar que el modelo obtenido en el inciso 4 tiene muy poco poder predictivo para la variable JTS debido a que la diferencia promedio entre los valores observados y la predicción del modelo es aproximadamente de 50. En el histograma se puede observar que la mayoría de las diferencias entre los valores observados y predichos es mayor a 10.

6. Clasificación de los aspirantes según el tipo de empleo.

Tabla 4. Clasificación de los aspirantes según el tipo de empleo y zona

Empleo	Frecuencia
A	73
B	79
C	75
D	73

En la **Tabla 4** se presentan una nueva variable cualitativa que clasifica a los aspirantes según el ingreso promedio mensual de sus anteriores empleos (IPR). Estas se dividen en cuatro categorías:

A: mayor de 1550

B: entre 1200 y 1550

C: entre 840 y 1200

D: menos de 840

Se observa que los aspirantes están distribuidos equitativamente entre las cuatro categorías el mínimo de aspirantes por categoría es de 73 y el máximo de 79.

7. Puntaje medio de los aspirantes según el tipo de empleo y zona.

Tabla 5. Puntaje medio de los aspirantes según el tipo de empleo y zona

	Empleos			
Zona	Empleo A	Empleo B	Empleo C	Empleo D
A	683.5	604.0	490.0	665.0
B	714.0	583.5	623.0	324.0
C	386.0	676.5	413.0	478.0
D	633.5	764.0	598.0	422.0

En la **Tabla 5**, se presenta el puntaje medio de los aspirantes según la clasificación obtenida en la **Tabla 4** y la variable cualitativa *zona*. Se observa que el puntaje más bajo se encuentra en los aspirantes de la zona B y empleo D, mientras que el puntaje mayor se encuentra en los aspirantes de la zona D y empleo B.

8. Estudio de los puntajes medio de los aspirantes según el tipo de empleo y zona.

Tabla 6. Tabla ANDEVA para bloques aleatorizados

Fuente	G.L.	SS	MS	F	Pr(>F)
Bloques	3	38054	12684	0.8210	0.5144
Tratamientos	3	79032	26344	1.7051	0.2350
Error	9	139048	15450		
Total	15	256134			

Los datos de la **Tabla 6** representan un estudio de análisis de varianzas de dos vías para las medias de los aspirantes que se encuentran en la **Tabla 5**. Con un nivel de confianza del 95%, los resultados obtenidos permiten afirmar que, en el caso de los tratamientos, las medias por zona son iguales y, en el caso de bloques, las medias por tipo de empleo también son iguales. Por lo que no importa la zona ni el tipo de empleo por nivel de salario para los aspirantes, los puntajes medios son similares en todos los casos.

CONLUSIONES Y RECOMENDACIONES

A partir del estudio de los datos de 300 aspirantes que forman parte de un programa de inserción laboral se logró un análisis descriptivo e inferencial de la muestra permitiendo cumplir con todos los objetivos y conocer parámetros de cada una de las variables así como gráficas que permiten una mejor interpretación.

Finalmente se puede concluir que los resultados obtenidos muestran un puntaje promedio de 557,3±262,8293ptos. Las variables que mejor ser relacionan con JTS son ING, EDU, PAS y DIS. Además, no hay suficiente evidencia estadística para afirmar que JTS sigue una distribución normal. El modelo que mejor explica la variable JTS viene representado por:

$$y = -287.7378 + 110.7998(ING) + 8.3967(EDU) - 11.0904(DIS) + 7.9451(PAS)$$

Sin embargo, este no posee un poder predictivo razonable debido a una gran diferencia promedio entre los valores observados y la predicción del modelo. Finalmente se puede afirmar que, a partir de los datos proporcionados, los puntajes medios de los aspirantes según el tipo de empleo y zona no difieren entre sí.

BIBLIOGRAFIA

- Dennis D. Wackerly, William Mendenhall, Richard L. Scheaffer - Mathematical statistics with applications - Thomson Brooks_Cole (2008)

- RStudio – Open source and enterprise-ready professional software for R. https://www.rstudio.com/

- R: The R Project for Statistical Computing. https://www.r-project.org/

ANEXOS

Código en R

```
datos <- read.table( 'Capacitacion.txt', header = T )
attach(datos)

#PARTE 1

#Analisis descriptivo

summary(datos)

#Desviaciones estandard

sd(datos$JTS)
sd(datos$ING)
sd(datos$EDU)
sd(datos$DIS)
sd(datos$SEM)
sd(datos$EMP)
sd(datos$PAS)
sd(datos$IPR)

# Histogramas
par(mfrow = c(2,4))
hist(datos$JTS, ylab='Frecuencias', xlab = 'JTS', main = '')
hist(datos$ING, ylab='Frecuencias', xlab = 'ING', main = '')
hist(datos$EDU, ylab='Frecuencias', xlab = 'EDU', main = '')
hist(datos$DIS, ylab='Frecuencias', xlab = 'DIS', main = '')
hist(datos$SEM, ylab='Frecuencias', xlab = 'SEM', main = '')
hist(datos$EMP, ylab='Frecuencias', xlab = 'EMP', main = '')
hist(datos$PAS, ylab='Frecuencias', xlab = 'PAS', main = '')
hist(datos$IPR, ylab='Frecuencias', xlab = 'IPR', main = '')
mtext("Histogramas de todas las variables", side = 3, line =
-2, outer = TRUE)

# Boxplots
par(mfrow = c(2,4))
boxplot(datos$JTS, xlab = 'JTS')
boxplot(datos$ING, xlab = 'ING')
boxplot(datos$EDU, xlab = 'EDU')
boxplot(datos$DIS, xlab = 'DIS')
boxplot(datos$SEM, xlab = 'SEM')
boxplot(datos$EMP, xlab = 'EMP')
boxplot(datos$PAS, xlab = 'PAS')
boxplot(datos$IPR, xlab = 'IPR')
mtext("Boxplots de todas las variables", side = 3, line = -2,
outer = TRUE)
```

```
boxplot(datos)
par(mfrow = c(1,1))

# Grafico de barras para la zona
barplot(table(datos$zona), ylim = c(0, 100), xlab = 'Zonas',
main = 'Frecuencias de la variable zona')
#PARTE 2 graph de dispersion y matriz de correlacion

pairs(datos)
cor(datos[,1:8]) #sin la columna zona ya que es cualitativa

# a primera vista se observa que las variables que mejor se
ajustan con JTS
# son ING, EDU, DIS y PAS

#PARTE 3 Prueba de bondad de ajuste
foo = datos$JTS
hist(foo, ylim=c(0,50), main = 'Puntaje en la prueba de
capacitación.', ylab = 'Frecuencias')
hist(foo, plot = F)

fi = c(31,33,34,34,22,40,32,45,29)
qqnorm(fi)
qqline(fi)

# los datos estan muy alejados de la linea, no sigue una dist
normal

# se realiza la prueba chi cuadrado

(k = length(fi))
n = sum(fi)
mi = c(150,250,350,450,550,650,750,850,950)
(xbarra = sum(fi * mi) / n)
x_barra = rep(xbarra, k)
S_cuadrado = sum(fi * (mi - x_barra) ^ 2) / (n - 1)
(S = sqrt(S_cuadrado))
(pi = pnorm(2 : 10 * 100, xbarra, S) - pnorm(1 : 9 * 100,
xbarra, S))
chi2_obs = sum((fi - n * pi) ^ 2 / ( n * pi))
chi2_obs
r = 2
chi2_alpha = qchisq(1 - 0.05, k- 1 - r)
chi2_alpha
(p_valor = 1 - pchisq(chi2_obs, k - 1 - r))
```

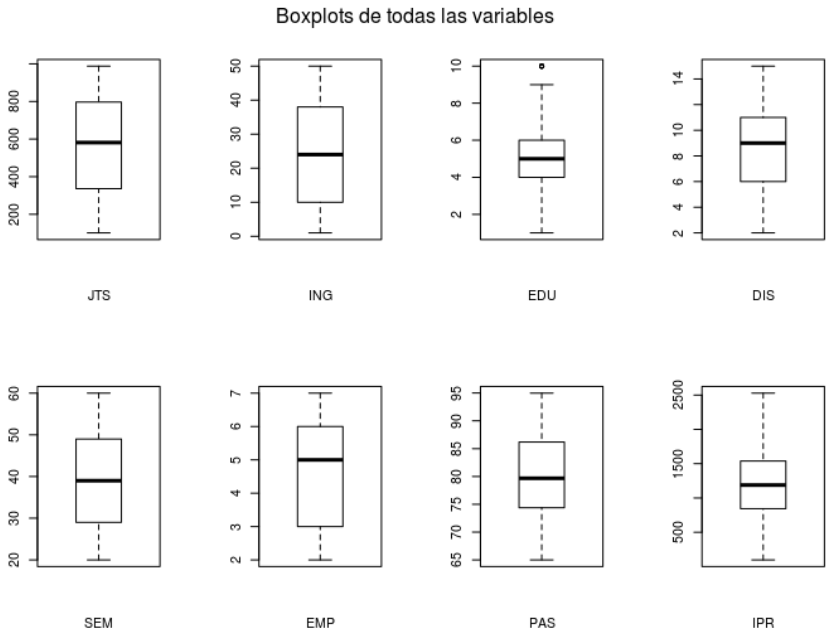
```
#como pvalor=1.030287e-13 < 0.05, se rechaza que tenga dist
normal

#PARTE 4


#como zona es una variable o factor cualitativo, no se
incluye en el estudio

modelo0=lm(JTS ~ ING + EDU + DIS + SEM + EMP + PAS + IPR)
summary(modelo0)
#R2 91.17%

# se procede a quitar la variable SEM ya que no es
significativa
modelo1=lm(JTS ~ ING + EDU + DIS + EMP + PAS + IPR)
summary(modelo1)

#R2 91.18%

# se procede a quitar la variable IPR
modelo2=lm(JTS ~ ING + EDU + DIS + PAS + EMP)
summary(modelo2)

#R2 91.17%

# se procede a quitar la variable EMP
modelo3=lm(JTS ~ ING + EDU + DIS + PAS)
summary(modelo3)
par(mfrow=c(2, 2))
plot(modelo3)
par(mfrow=c(1, 1))
#R2 91.11%, todas las variables son significativas.

#chequeo de normalidad de errores
qqnorm(rstandard(modelo3))
qqline(rstandard(modelo3))
boxplot(datos$ING + datos$EDU + datos$DIS + datos$PAS, ylab =
'Valores del modelo',
        main = 'Boxplot del modelo escogido')

#homocedasticidad

plot(fitted.values(modelo3),   rstandard(modelo3),   main   =
'Homocedasticidad',
     ylab = 'Errores estándar', xlab = 'x')
```

```
# independecia de errores
par(mfrow = c(2,2))
#mtext("My 'Title' in a strange place", side = 3, line = -2,
outer = TRUE)
plot(datos$ING, rstandard(modelo3), sub = 'Independencia de
errores ING',
     xlab = 'ING', ylab = 'Errores estándar')
plot(datos$EDU, rstandard(modelo3), sub = 'Independencia de
errores EDU',
     xlab = 'EDU', ylab = 'Errores estándar')
plot(datos$DIS, rstandard(modelo3), sub = 'Independencia de
errores DIS',
     xlab = 'DIS', ylab = 'Errores estándar')
plot(datos$PAS, rstandard(modelo3), sub = 'Independencia de
errores PAS',
     xlab = 'PAS', ylab = 'Errores estándar')
mtext("Independencia de Errores", side = 3, line = -2, outer
= TRUE)
par(mfrow = c(1,1))

# se observa los valores no siguen patron alguno ni en
homocedasticidad ni en independencia, son discretos

# PARTE 5

pred <- read.table( "Capacitacion_prediccion.txt", header = T
)
attach(pred)

new <- data.frame(pred$JTS, pred$ING, pred$EDU, pred$DIS,
pred$PAS)
JTS1 <- predict(modelo3,new, interval = "confidence")
JTS2 <- predict(modelo3,new, interval = "prediction")

matplot(new$pred.JTS,cbind(JTS1,                JTS2[,-1]),
lty=c(1,2,2,3,3), type="l",
        ylab="predicted JTS", main = 'Bandas de confianza de
nuestro modelo')

Y_ING = pred$ING
Y_EDU = pred$EDU
Y_DIS = pred$DIS
Y_PAS = pred$PAS

#Valores de la prediccion del modelo
y = -287.738 +  10.800*Y_ING + 8.397*Y_EDU - 11.090*Y_DIS +
7.945*Y_PAS
```

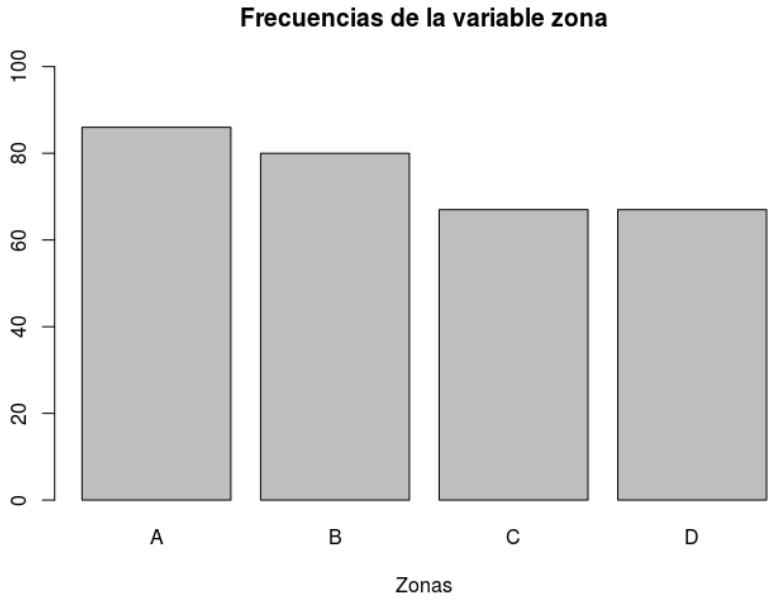
```
#Valores observados
Y_Y = pred$JTS

boxplot(Y_Y-y, main = 'Boxplot de las diferencias real -
predicción')
hist(Y_Y-y, main = 'Diferencias entre real - predicción',
     ylab = 'Diferencias', xlab = 'Frecuencias')


#PARTE 6
summary(datos$IPR)
empleoA = "EmpleoA"
empleoB = "EmpleoB"
empleoC = "EmpleoC"
empleoD = "EmpleoD"
datos["parte6"]      =      cut(datos$IPR,      breaks      =
c(97,840,1200,1550,2530),   labels   =   c(empleoD,   empleoC,
empleoB, empleoA))
summary(datos$parte6)
barplot(table(datos$parte6), ylim = c(0, 100), main = 'Tipo
de Empleo segun salario')
#PARTE 7

medAA = median(datos$JTS[(datos$zona == "A") & datos$parte6
== empleoA])
medAB = median(datos$JTS[(datos$zona == "A") & datos$parte6
== empleoB])
medAC = median(datos$JTS[(datos$zona == "A") & datos$parte6
== empleoC])
medAD = median(datos$JTS[(datos$zona == "A") & datos$parte6
== empleoD])
medBA = median(datos$JTS[(datos$zona == "B") & datos$parte6
== empleoA])
medBB = median(datos$JTS[(datos$zona == "B") & datos$parte6
== empleoB])
medBC = median(datos$JTS[(datos$zona == "B") & datos$parte6
== empleoC])
medBD = median(datos$JTS[(datos$zona == "B") & datos$parte6
== empleoD])
medCA = median(datos$JTS[(datos$zona == "C") & datos$parte6
== empleoA])
medCB = median(datos$JTS[(datos$zona == "C") & datos$parte6
== empleoB])
medCC = median(datos$JTS[(datos$zona == "C") & datos$parte6
== empleoC])
```

```
medCD = median(datos$JTS[(datos$zona == "C") & datos$parte6
== empleoD])
medDA = median(datos$JTS[(datos$zona == "D") & datos$parte6
== empleoA])
medDB = median(datos$JTS[(datos$zona == "D") & datos$parte6
== empleoB])
medDC = median(datos$JTS[(datos$zona == "D") & datos$parte6
== empleoC])
medDD = median(datos$JTS[(datos$zona == "D") & datos$parte6
== empleoD])

# PARTE 8
# Realice un análisis de varianza para estudiar si existen
diferencias entre los puntajes medio de los
# aspirantes según la clasificación del ingreso y zona.
#empleos
medias = c(medAA, medAB, medAC, medAD,
           medBA, medBB, medBC, medBD,
           medCA, medCB, medCC, medCD,
           medDA, medDB, medDC, medDD)
empleos <- gl(4, 4)
zona <- factor(rep(1:4, 4))
anova_formula = medias ~ zona + empleos
xtabs(anova_formula)
mod.lm = lm(anova_formula)
anova(mod.lm)

# Asumiendo un nivel de significancia del 5%, no se puede
rechazar las hipotesis de que son iguales
# las medias de las poblaciones, tanto por zona como por
clasificacion del ingreso.

# PARTE 9
# En el caso de conseguir diferencias significativas, realice
pruebas de medias para determinar cuales
# son las medias que difieren.

# No se consiguieron diferencias significativas, por lo cual
no hace falta realizar
# pruebas de medias.
```

GRÁFICAS



Gráfica 1. Boxplots de todas las variables cuantitativas del estudio.



Gráfica 2. Diagrama de Barras de la variable “zona”.

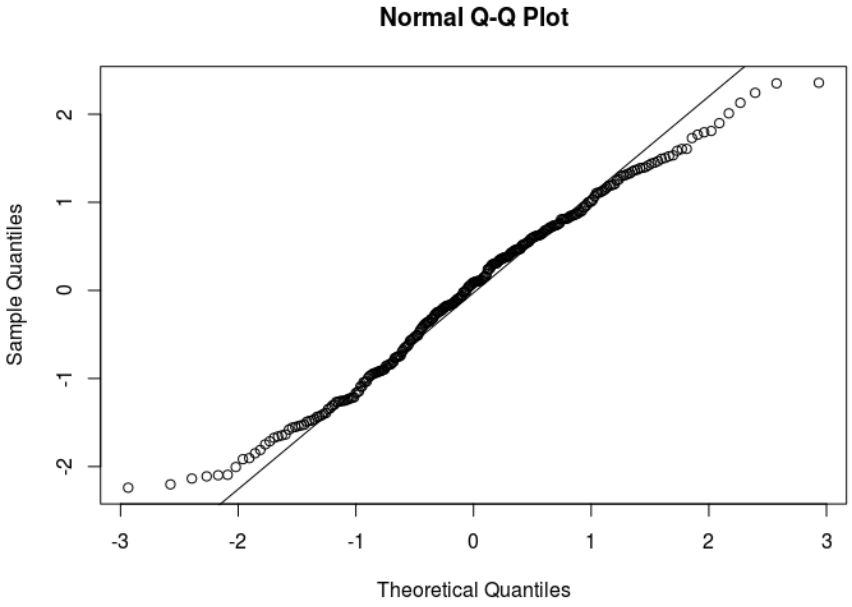


Gráfico 3. Normalidad de los errores estandarizados del modelo escogido.

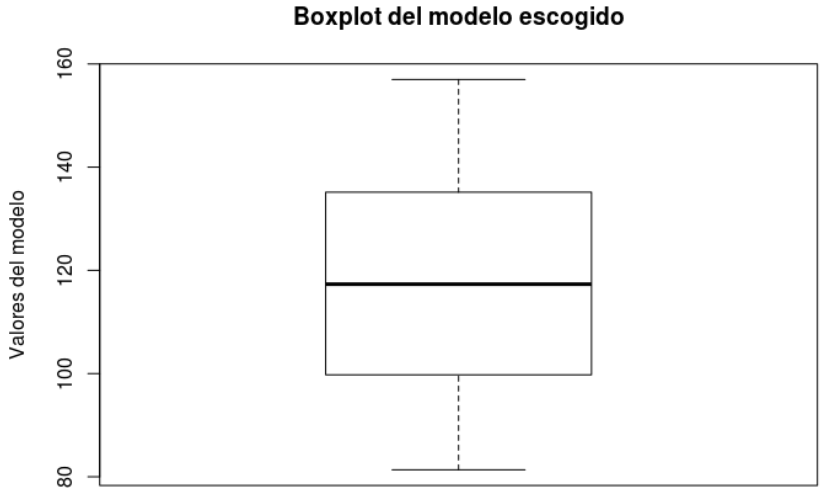


Gráfico 4. Boxplot de nuestro modelo lineal para JTS.

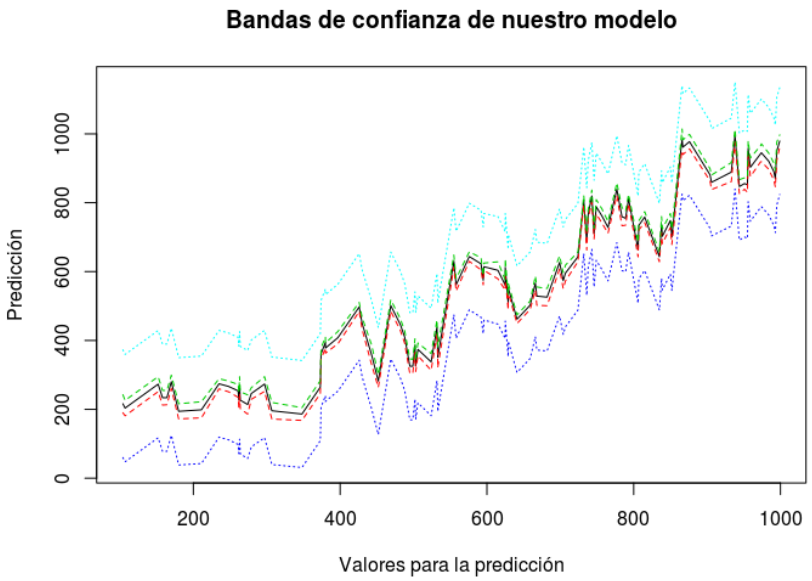


Gráfico 5. Bandas de confianza y predicción del modelo lineal para JTS.

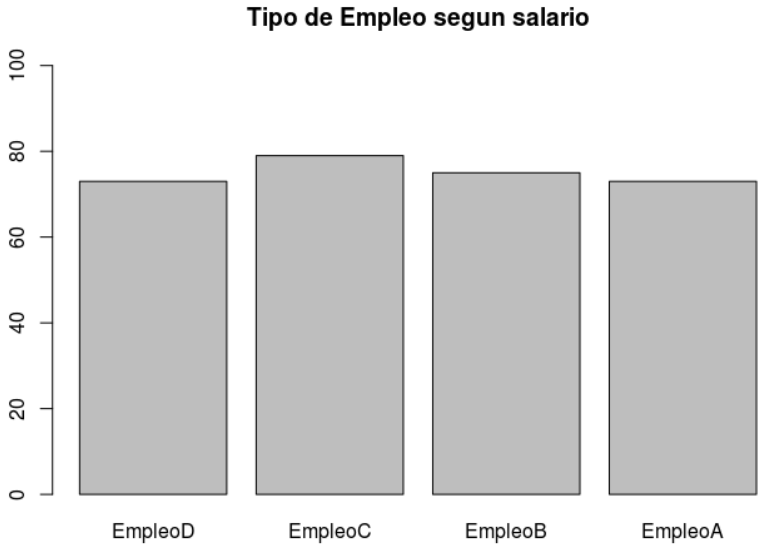


Gráfico 6. Diagrama de Barras para la clasificación según el IPR.