Sebastián Valencia Calderón 201111578
Edgar Daniel Gómez 201317343

Punto 1

Diferentes organizaciones, entre las que se incluye The American Journal of Psychiatry, han presentado como una problemática creciente la sustitución de las interacciones sociales por las interacciones a través de medios digitales (redes sociales), lo que de acuerdo con los expertos está correlacionado con problemas psicológicos (estrés, ansiedad, adicción). Con el propósito de realizar un estudio formal en cierta universidad, se propuso como primera etapa determinar si el tiempo, en minutos, que un estudiante dedica a visitar las redes sociales se puede modelar como una variable aleatoria con una distribución de probabilidad conocida, particularmente, si se asemeja a una distribución normal. Para el desarrollo de esta primera etapa, se hizo seguimiento al tiempo que una muestra aleatoria de 150 estudiantes dedica en las redes sociales diariamente. Los resultados de este seguimiento se presentan en el archivo de excel *DatosRedesSociales.xlsx* .

(a) **(2 puntos)** Plantee la prueba de hipótesis nula y alterna de la prueba.

$$H_0 : \text{Tiempo en redes} = \mathcal{N}(121{,}22, 20{,}033^2) \qquad H_1 : \neg H_0$$

(b) **(12 puntos)** Identifique el estadístico de prueba con su distribución y calcúlelo. Utilice 10 clases equiprobables.

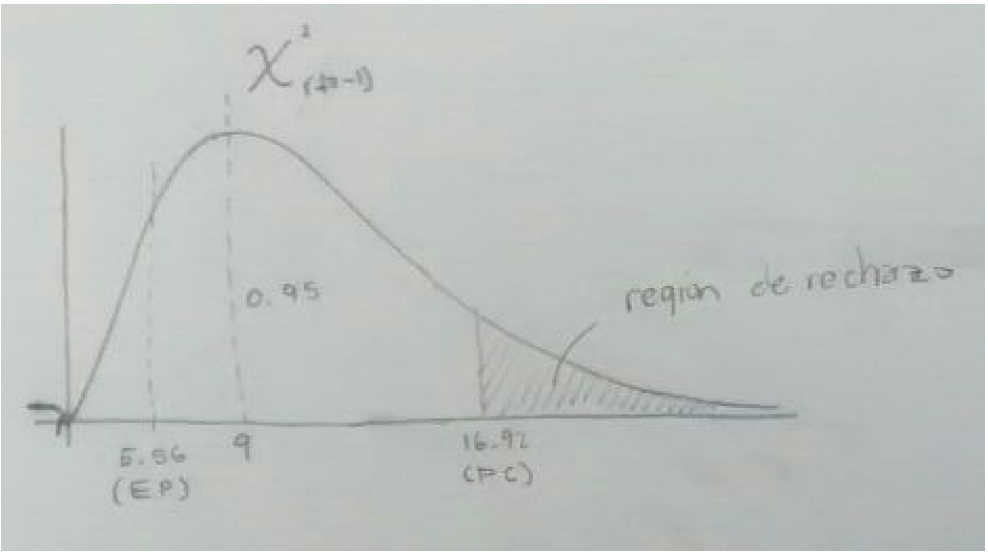
$$\mathbf{EP} \sim \mathcal{X}^2_{(10,1)} = \sum_{i=1}^{10} \frac{\hat{o}_i - \hat{e}_i}{\hat{e}_i}$$

$\hat{o}$ es el valor observado y $\hat{e}$ es e valor esperado.

A continuación se muestran los intervalos y el estadístico de prueba.

Clases equiprobables		estadarizado								
Limite Inferior	Limite Superior	inf	sup	Probabilidad	Observado	Esperado	obs-esp	(obs-esp)^2	(obs-esp)^2/esp	Estadistico de prueba
69	79.70	0.00457	0.01911	0.01453	2	2.18017	-0.18017	0.03246	0.01489	5.5569
79.7	90.40	0.01911	0.06197	0.04286	6	6.42929	-0.42929	0.18429	0.02866	
90.4	101.10	0.06197	0.15761	0.09564	20	14.34579	5.65421	31.97008	2.22853	
101.1	111.8	0.15761	0.31910	0.16149	18	24.22365	-6.22365	38.73384	1.59901	
111.8	122.5	0.31910	0.52547	0.20638	32	30.95642	1.04358	1.08906	0.03518	
122.5	133.2	0.52547	0.72509	0.19961	29	29.94182	-0.94182	0.88703	0.02963	
133.2	143.9	0.72509	0.87121	0.14613	24	21.91893	2.08107	4.33085	0.19759	
143.9	154.6	0.87121	0.95217	0.08096	12	12.14362	-0.14362	0.02063	0.00170	
154.6	165.3	0.95217	0.98611	0.03394	4	5.09116	-1.09116	1.19064	0.23386	
165.3	176	0.98611	0.99688	0.01077	3	1.61494	1.38506	1.91838	1.18789	
					150	150				

(c) **(4 puntos)** Utilizando un nivel de significancia del 5 %, realice una gráfica de la distribución del estadístico de prueba, identifique el punto crítico y la región de rechazo. Adicionalmente, ubique el valor del estadístico de prueba.



(d) **(2 puntos)** Concluya en términos del problema.

Al comparar el estadístico de prueba con la región crítica, se observa que el estadístico de prueba, es menor al punto crítico, por lo que no está en la región de rechazo. Entonces, existe evidencia estadística para no rechazar la hipótesis nula, y de hecho, se puede afirmar que la muestra se distribuye normal.

Sebastián Valencia Calderón 201111578
Edgar Daniel Gómez 201317343

Punto 2

Diana es una comerciante exitosa que se caracteriza por utilizar métodos estadísticos como soporte a la toma de decisiones. En esta ocasión, ella realizó un estudio con el propósito de predecir el valor total de las ventas mensuales (variable aleatoria Y), en millones de pesos, utilizando los gastos en publicidad mensual, en millones de pesos, como la variable dependiente X. Para ello, recolectó información de 12 sucursales de empresas similares. A continuación se presenta el resumen de los datos obtenidos por el comerciante:

$$n = 12 \qquad \bar{X} = 34{,}17 \qquad \bar{Y} = 453{,}75$$

$$\sum_{i=1}^{n} y_i^2 = 2512925 \qquad \sum_{i=1}^{n} x_i^2 = 15650 \qquad \sum_{i=1}^{n} y_i x_i = 191325$$

(a) **(2 puntos)** Escriba el modelo de regresión lineal que le permite predecir el valor total de las ventas semanales en función de la publicidad semanal.

El modelo de regresión lineal simple, para la situación en cuestión, consiste en predecir (o modelar) las ventas mensuales, haciendo uso del gasto mensual de publicidad. Es decir, usando el modelo algebraico de una ecuación de primer orden, cuyo significado geométrico es una linea recta en $\mathbb{R}2$, se desea interpretar la significancia de la variable X para determinar el comportamiento de la variable Y. Lo que se desea, es hallar los valores β_0 y β_1 en la siguiente ecuación:

$$\hat{y} = \beta_0 + \beta_0 x + \epsilon$$

Donde $\hat{y}$ representa la variable aleatoria Y y x la variable aleatoria X, y ϵ el error aleatorio.

(b) **(4 puntos)** Utilizando el método de mínimos cuadrados ordinarios, encuentre de manera teórica la ecuación para estimar los coeficientes del modelo de regresión lineal planteado en el literal a.

Las suposiciones generales para el desarrollo analítico del modelo son:

$$\mathbb{E}(\epsilon_i) = 0 \quad Var(\epsilon_i) = \sigma^2 \quad Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall i, j = 1, 2, ..., n \wedge i \neq j$$

$$\mathbb{E}(\hat{y}) = \mathbb{E}(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x$$

El método de mínimos cuadrados, minimiza la suma de los cuadrados de las desviaciones de las observaciones y el promedio de las mismas. Es decir:

$$\epsilon = y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

$$\sum i = 1^n \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2$$

Para obtener los estimadores β_0, β_1, se diferencia esta última expresión con respecto a cada uno de estos y se iguala la derivada parcial a cero.

$$\frac{\partial \sum e_i^2}{\partial \beta_0} = -2 \times \left[\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right] = 0$$

$$\frac{\partial \sum e_i^2}{\partial \beta_1} = -2 \times \left[\sum_{i=1}^{n} x_i \times \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right] = 0$$

A partir de esto, se obtienen las siguientes expresiones:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \qquad \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = \hat{\beta}_0 + \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Al sustituir $\hat{\beta}_0$ en:

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Se obtiene:

$$\sum_{i=1}^{n} x_i y_i = \left(\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} \right) \times \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Por lo tanto, hallar β_0 y β_1 es ahora trivial.

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right) \times \left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

(c) **(4 puntos)** Estime los parámetros del modelo de regresión lineal simple planteado en el literal a. Realice una interpretación de cada uno de los coeficientes estimados.

Al multiplicar por n, se facilita el cálculo de la expresión:

$$\beta_1 = \frac{12(191325) - 12(34{,}17)(473{,}75)}{12(15650) - (34{,}17 \times 12)^2} = 3{,}1544$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} = 345{,}96$$

Se puede concluir, que los gastos de publicidad influyen de manera positiva sobre las ventas de la compañía, esto a través de los coeficientes estimados en el literal anterior. Es decir, a mayores gastos en publicidad, mayores serán las ventas.

(d) **(2 puntos)** Estime la tabla ANOVA correspondiente al modelo de regresión lineal simple planteado en el litera a.

$$SCE = 6338{,}438 \wedge SCT = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)}{n} \wedge SCT = SCR + SCT$$

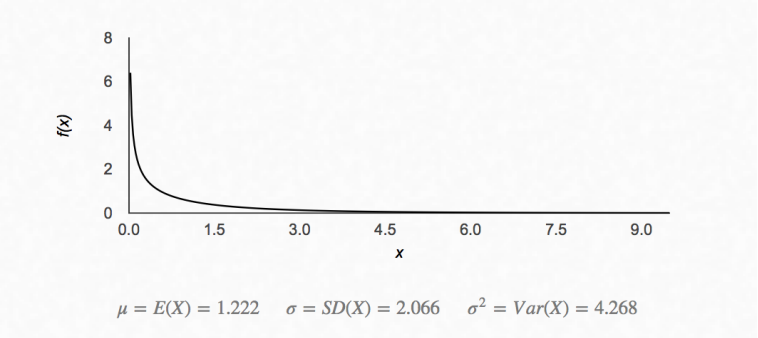
$$SCR = 2512925 - 453{,}75 = 2512471{,}25 \Rightarrow SCT = 2518809{,}688$$

ANOVA				
Model	Sum of squares	gl	Mean Square	F
Regression	2512471.25	1	MCR_{reg}	$\frac{MCR_{reg}}{\left(\frac{\Sigma(y_i-\hat{y}_i)^2}{10}\right)}$
Residual	6338.438	10	$\frac{\Sigma(y_i-\hat{y}_i)^2}{10}$	
Total	2518809.688	11	$\frac{\Sigma(y_i-\bar{y})^2}{11}$	

(e) **(2 puntos)** Plantee las hipótesis nula y alterna de la prueba de significancia global. Defina el estadístico de prueba y concluya en términos del contexto del problema; utilice el criterio del p-valor.

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

$$\textbf{EP} : \frac{SCR}{(SCE/(n-2))} \sim F(1, 11) = \frac{2512471{,}25}{(6338{,}438)/10}$$



Claramente, el estadístico de prueba está en la cola derecha, es decir, se rechaza la hipótesis nula.

(f) **(1 punto)** Estime el valor del coeficiente de determinación, e interprételo.

$$R^2 = \frac{SCR}{SCT} \in [0, 1] \Rightarrow R^2 = \frac{2512471{,}25}{2518809{,}688} = 0{,}9974$$

Dado que R^2 es muy cercano a uno, se concluye que el modelo es globalmente sinificativo para explicar o predecir la variable y haciendo uso de x.

(g) **(16 puntos)** Plantee una prueba de significancia individual para el intercepto y para la pendiente. Para cada una, plantee las hipótesis nula y alterna y defina el estadístico de prueba y concluya en términos del contexto del problema; utilice el criterio del p-valor.

Para β_1

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

$$\textbf{EP} = t_0 = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{s_{xx}}} \wedge S = \sqrt{\frac{s_{yy} - \hat{\beta}_1 s_{xy}}{n-2}}$$

$$s_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 5269{,}4 \wedge s_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 42256{,}28$$

$$S = 50{,}63$$

$$s_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 1638{,}93$$

El estadístico de prueba:

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{s_{xx}}}} = 2{,}52$$

La región de rechazo está dada por:

$$t_0 < t_{(0{,}025,10)} = -2{,}228 \vee t_0 > t_{(0{,}975,10)} = 2{,}228$$

Conclusión:

Debido a que t_0 cae en la región de rechazo, se rechaza la hipótesis nula y se concluye que β_1 es distinto a cero.

Para β_0

$$H_0 : \beta_0 = 0 \qquad H_1 : \beta_0 \neq 0$$

$$\textbf{EP} = t_0 = \frac{\hat{\beta}_0 - \beta_0}{S/\sqrt{\frac{\sum i=1^n x_i^2}{n \times s_{xx}}}} = 7{,}66$$

La región de rechazo está dada por:

$$t_0 < t_{(0,025,10)} = -2,228 \vee t_0 > t_{(0,975,10)} = 2,228$$

Conclusión:

Debido a que t_0 cae en la región de rechazo, se rechaza la hipótesis nula y se concluye que β_0 es distinto a cero.

(h) **(8 puntos)** Calcule un intervalo de confianza ($\alpha = 0,05$) para el intercepto y para el coeficiente asociado a la variable independiente. Interprete cada uno de los intervalos.

Para β_1

$$\hat{\beta}_1 \pm t_{(\alpha/2,10)} \left(\frac{S}{\sqrt{s_{xx}}} \right) \Rightarrow 3,1544 \pm 2,228 \times \left(\frac{50,63}{\sqrt{1638,93}} \right)$$

$$\textbf{IC} : [5,94, 0,37]$$

Para β_0

$$\hat{\beta}_1 \pm t_{(\alpha/2,10)} S \left(\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n s_{xx}}} \right) \Rightarrow 345,96 \pm 2,228 \left(50,63 \sqrt{\frac{15650}{12 \times 1638,93}} \right)$$

$$\textbf{IC} : [446,58, 245,334]$$

(i) **(4 puntos)** Diana decidió invertir 40 millones de pesos en publicidad, y las ventas mensuales fueron de 480 millones de pesos. ¿Cuál es el valor que Diana había pronosticado para las ventas? ¿Cuál es el residuo asociado a este pronóstico?

Dado que $y = 345,96 + 3,1544 \times x$. Entonces el valor de ventas asociado a 40 como gastos de publicidad es: $345,96 + 3,1544 \times 40 = 472,13$. El error residual ϵ es $480 - 472,13 = 7,87$

Punto 3

Descargue de Sicuaplus el archivo de Excel con el nombre *DatosRegresion.xls*. Allí encontrará información relacionada con la contaminación atmosférica en 38 ciudades de Estados Unidos. Las variables que se presentan en el archivo son las siguientes:

- Contenido de SO_2 (Dióxido de Azufre) en el aire. Esta variable se mide en microorganismos por metro cúbico.
- Número de fábricas con más de 20 empleados.
- Número de habitantes (miles).
- Velocidad media del viento al año (millas por hora).
- Precipitación media anual (litros por pulgada).
- Número medio de días con lluvia al año.

Estime el modelo de regresión lineal múltiple para la variable Contenido de SO_2 en el aire usando todas las variables descritas. Para esto de respuesta a los siguientes literales:

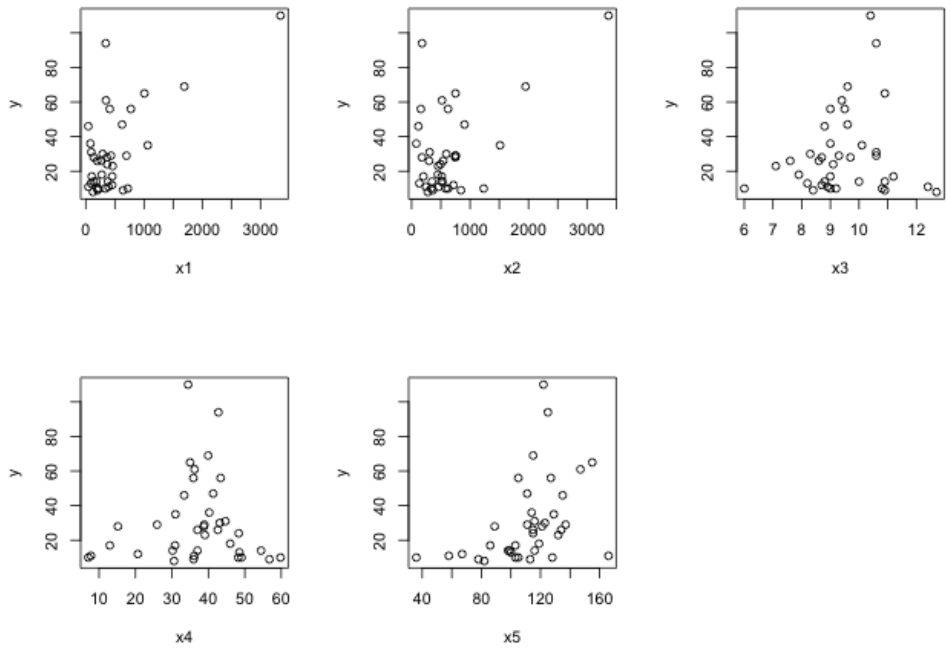
(a) **(2 puntos)** Escriba a continuación el modelo de regresión lineal que le permite predecir el Contenido de SO_2 en el aire con base en las variables independientes.

Y	Contenido deSO_2 en el aire (microorganismos por metro cúbico)
X_1	Número de fábricas con más de 20 empleados
X_2	Numéro de habitantes (miles)
X_3	Velocidad media del viento al año (millas por hora)
X_4	Precipitación media anual (Litros por pulgada)
X_5	Número medio de días con lluvia al año

El modelo es:

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5$$

(b) **(10 puntos)** Realice una gráfica de dispersión entre cada una de las variables independientes y la dependiente. Indique, para cada caso, si de manera visual se puede percibir una relación lineal.



En la gráfica se muestra la relación entre cada una de las variables predictoras y la variable independiente, en $x2$, $x3$, $x4$ y $x5$ se observa una estrecha relación lineal, mientras que en $x1$ no tanto.

(c) **(4 puntos)** Con ayuda de SPSS, presente a continuación la tabla de ANOVA, y la tabla de coeficientes.

A continuación, se muestra la tabla ANOVA, con los valores por cada una de las variables predictoras (los valores de regresión son la suma de estos, los totales, la suma de estos anteriores y los residuales). Las tablas fueron realizadas usando el lenguaje de programación R.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	9403.34	9403.34	38.57	0.0000
x2	1	3630.47	3630.47	14.89	0.0005
x3	1	41.72	41.72	0.17	0.6819
x4	1	109.57	109.57	0.45	0.5074
x5	1	845.94	845.94	3.47	0.0717
Residuals	32	7802.22	243.82		

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.0293	20.6876	0.92	0.3645
$x1$	0.0721	0.0162	4.45	0.0001
$x2$	-0.0469	0.0155	-3.02	0.0049
$x3$	-1.7276	2.0105	-0.86	0.3966
$x4$	-0.1336	0.2581	-0.52	0.6083
$x5$	0.2481	0.1332	1.86	0.0717

La tabla ANOVA es:

ANOVA				
Model	Sum of squares	gl	Mean Square	F
Regression	14031.01	5	14031.01	57,55
Residual	7802.22	32	243,82	
Total	21833.23	37	301,37	

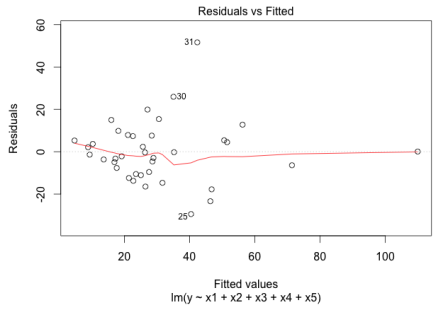
(d) **(4 puntos)** Para aquellos coeficientes que resultan significativos, presente la interpretación correspondiente a cada uno.

Los coeficientes significativos son según la prueba anterior y la tabla ANOVA, x_1 y x_2. Lo cuál significa que el nñumero de fabricas con más de 20 empleados y el número de habitantes, cuyos coeficientes de significancia β_1 y β_2 son respectivamente 0.0721 y -0.04669 contribuyen positiva y negativamente respectivamente sobre la predicción del contenido de SO_2 en el aire.

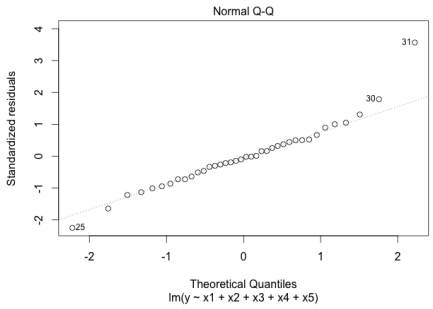
(e) **(2 puntos)** Estime el valor del coeficiente de determinación, e interprételo.

El coeficiente de determinación R^2, es 0.6426. Es decir, el modelo es significativo para predecir, es decir, el modelo explica y usando las variables usadas y posee gran capacodad explicativa y predictiva.

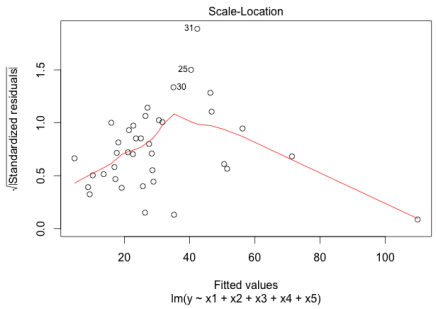
(f) **(8 puntos)** Verifique los supuestos del modelo de regresión lineal múltiple. Para ello, utilizando SPSS, realice una prueba de bondad de ajuste sobre los errores estimados del modelo de regresión lineal múltiple.



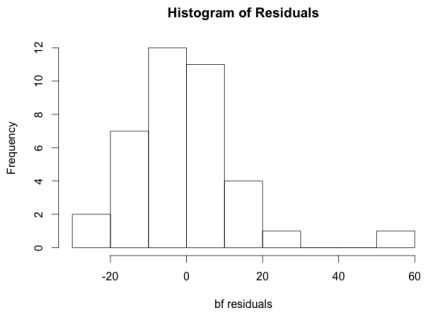
(a)



(b) Ajuste de cuantiles del error



(c)



(d) Histograma de los errores

Figura 1

Por medio de la gráfica anterior, se concluye que el error distribuye normal con media cero y varianza constante.