UADY
UNIVERSIDAD AUTÓNOMA DE YUCATÁN
"Luz, Ciencia y Verdad"

FACULTAD DE MATEMÁTICAS

Minería de Datos

Unidad IV: Métodos de Predicción

ADA 11: *Regresión Lineal.*

Licenciatura en Actuaria.

Integrantes:
- Álvarez Herrera Samantha
- Ciau Puga Abigail
- Colonia Espinosa Cindy
- Fernández Caro Frida
- Padilla Jiménez Meybor
- Sobrino Bermejo Samantha

M.C. Ernesto Guerrero Lara

Fecha de entrega: Lunes 18 de mayo de 2020

INSTRUCCIÓN: Se desea determinar un modelo para estimar la esperanza de vida de las personas de una ciudad en función de diversos factores. La base de datos "múltiple.csv" contiene información de dichas variables. Ajusta un modelo de regresión múltiple validando los supuestos y predice la esperanza de vida para una combinación de las variables que permanezcan en el modelo final.

Paso 1: ANÁLISIS DESCRIPTIVO

Análisis individual de las variables.

La base cuenta con nueve variables de interés, con 50 observaciones.

- Esperanza de vida

Resumen estadístico para esp_vida

Mínimo	67.96
1er cuarto	70.12
Mediana	70.67
Media	70.88
3er cuarto	71.89
Máximo	73.60
Rango	5.64
Desviación estándar	1.3424

❖ **Esperanza de vida**

La variable *Esperanza de vida*, en años es de escala de razón.Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de los datos de [67.96,73.60], la media de *Esperanza de Vida* es de 70.88 con una desviación estándar de 1.3424.

- Habitantes

Resumen estadístico para Habitantes

Mínimo	365
1er cuarto	1080
Mediana	2838
Media	4246
3er cuarto	4968
Máximo	21198
Rango	20833
Desviación estándar	4464.49

❖ **Habitantes**

La variable *Habitantes*, en millones de dólares, es de escala de razón.
Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de los datos de [365,21198], la media de *Habitantes* es de 4246 con una desviación estándar de 4464.491

- Ingresos

Resumen estadístico para Ingresos

Mínimo	3098
1er cuarto	3993
Mediana	4519
Media	4436
3er cuarto	4814
Máximo	6315
Rango	3217
Desviación estándar	614.47

❖ **Ingresos**

La variable *Ingresos* es de escala de razón. Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de los datos de [3098,6315], la media de *Ingresos* es de 4436 con una desviación estándar de 614.4699.

- Analfabetismo

Resumen estadístico para Analfabetismo

Mínimo	0.5
1er cuarto	0.625
Mediana	0.950
Media	1.170
3er cuarto	1.575
Máximo	2.800
Rango	2.3
Desviación estándar	0.6095

- Asesinatos

Resumen estadístico para Asesinatos

Mínimo	1.400
1er cuarto	4.350
Mediana	6.850
Media	7.378
3er cuarto	10.675
Máximo	15.100
Rango	13.7
Desviación estándar	3.6915

- Universitarios

Resumen estadístico para Universitarios

Mínimo	37.80
1er cuarto	48.05
Mediana	53.25
Media	53.11
3er cuarto	59.15
Máximo	67.30
Rango	29.5
Desviación estándar	8.077

- Heladas

Resumen estadístico para Heladas

Mínimo	0
1er cuarto	66.25
Mediana	114.50
Media	104.46
3er cuarto	139.75
Máximo	188
Rango	188
Desviación estándar	51.98

- Área

Resumen estadístico para Área

Mínimo	1049
1er cuarto	36985
Mediana	54277
Media	70736
3er cuarto	81163
Máximo	566432
Rango	565383
Desviación estándar	85327.3

❖ **Analfabetismo**

La variable *Analfabetismo* es de escala de razón. Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de datos de [0.5,2.800], la media de *Analfabetismo* es de 1.170 con una desviación estándar de 06095

❖ **Asesinatos**

La variable *Asesinatos* es de escala de razón. Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de [1.400,15.100], la media de los *Asesinatos* es de 7.378 con una desviación estándar de 3.6915.

❖ **Universitarios**

La variable *Universitarios* es de escala de razón. Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de [37.80,67.30], la media de *Universitarios* es de 53.11 con una desviación estándar de 8.0769.

❖ **Heladas**

La variable *Heladas* es de escala de razón. Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de [0.00,188.00], la media de *Heladas* es de 104.46 con una desviación estándar de 51.9809.

❖ **Área**

La variable *Área* es de escala de razón. Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de [1049,566432], la media de *Área* es de 70736 con una desviación estándar de 85237.3

- Densidad de Población

Resumen estadístico para densidad_pobl

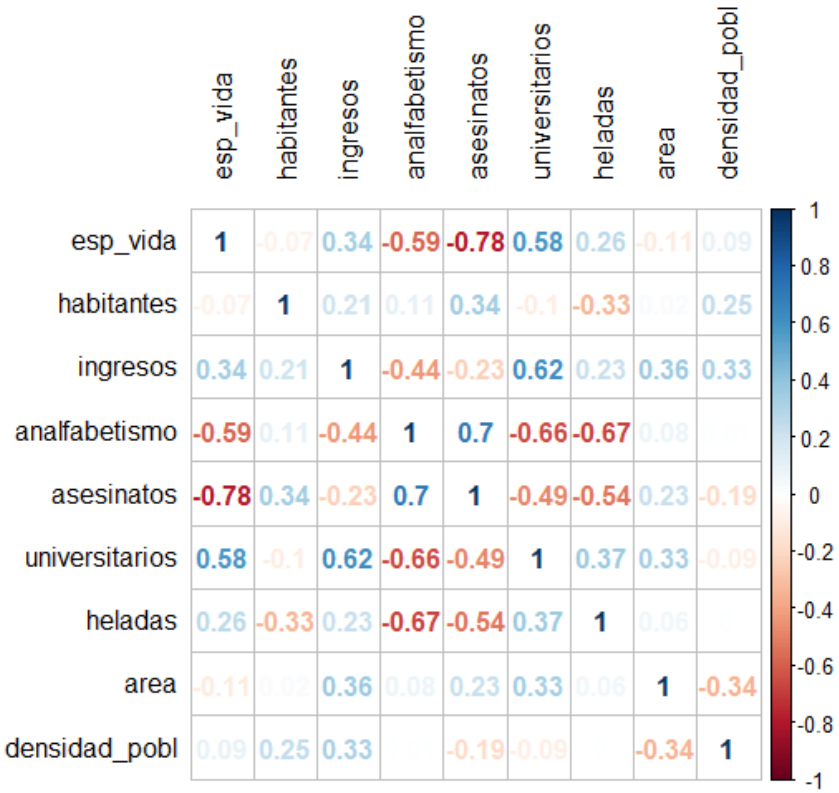
Mínimo	0.6444
1er cuarto	25.3352
Mediana	73.0154
Media	149.2245
3er cuarto	144.2828
Máximo	975.0033
Rango	974.359
Desviación estándar	221.006

❖ **Densidad poblacional**

La variable *Densidad poblacional* es de escala de razón. Después de hacer un análisis descriptivo con un software estadístico se tiene que en un rango de [0.6444,9075.0033], la media de *Densidad poblacional* es de 149.2245 con una desviación estándar de 221.006

Paso 2: SELECCIÓN DE VARIABLES

Si observamos los elementos fuera de la diagonal, r_{ij}, de la matriz de $X'X$. Si los regresores x_i y x_j son casi linealmente dependientes $|r_{ij}|$ será cercano a uno. Con ayuda de R podemos obtener dicha matriz, en la que se observan los coeficientes de correlación de Pearson para cada par de variables.



	esp_vida	habitantes	ingresos	analfabetismo	asesinatos	universitarios	heladas	area	densidad_pobl
esp_vida	1	-0.07	0.34	-0.59	-0.78	0.58	0.26	-0.11	0.09
habitantes	-0.07	1	0.21	0.11	0.34	-0.1	-0.33	0.02	0.25
ingresos	0.34	0.21	1	-0.44	-0.23	0.62	0.23	0.36	0.33
analfabetismo	-0.59	0.11	-0.44	1	0.7	-0.66	-0.67	0.08	-0.07
asesinatos	-0.78	0.34	-0.23	0.7	1	-0.49	-0.54	0.23	-0.19
universitarios	0.58	-0.1	0.62	-0.66	-0.49	1	0.37	0.33	-0.09
heladas	0.26	-0.33	0.23	-0.67	-0.54	0.37	1	0.06	-0.01
area	-0.11	0.02	0.36	0.08	0.23	0.33	0.06	1	-0.34
densidad_pobl	0.09	0.25	0.33	-0.07	-0.19	-0.09	-0.01	-0.34	1

Se ha decidido seleccionar el modelo mediante un procedimiento del tipo por segmentos, conocido como proceso de **eliminación hacia atrás** (en inglés Backward elimination procedure), en el cual se comienza con un modelo que incluya todos los k=8 regresores candidatos iniciales.

Primeramente, analizamos el modelo con **todas las variables regresoras**:

Modelo ajustado 1:

Esperanza de vida= 69.9506 + 0.0000648019*habitantes + 0.000270072*ingresos + 0.302861*analfabetismo - 0.328648*asesinatos + 0.0429077*universitarios - 0.0045802*heladas - 0.00000155776*area - 0.00110478*densidad_pobl

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
Intercepto	69.9506	1.84294	37.956	0.0000
habitantes	0.0000648019	0.0000300113	2.15925	0.0367
ingresos	0.000270072	0.00030869	0.874897	0.3867
analfabetismo	0.302861	0.402356	0.752719	0.4559
asesinatos	-0.328648	0.0494057	-6.65202	0.0000
universitarios	0.0429077	0.0233181	1.8401	0.0730
heladas	-0.0045802	0.00318923	-1.43614	0.1585
area	-0.00000155776	0.00000191438	-0.813716	0.4205
densidad_pobl	-0.00110478	0.000731185	-1.51094	0.1385

Por lo que, utilizando un nivel de significancia del 5% para el análisis, “analfabetismo” es la primera variable considerada para ser eliminada del modelo. De esta manera, ahora nuestro modelo será el siguiente:

Modelo ajustado 2:

Esperanza de vida= 70.981 + 0.0000567492*habitantes + 0.000190107*ingresos - 0.312221*asesinatos + 0.0365192*universitarios - 0.00605943*heladas - 8.63762E-7*area - 0.000861213*densidad_pobl

Analysis of Variance Table

Model 1: esp_vida ~ habitantes + ingresos + analfabetismo + asesinatos + universitarios + heladas + area + densidad_pobl

Model 2: esp_vida ~ habitantes + ingresos + asesinatos + universitarios + heladas + area + densidad_pobl

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	41	22.068			
2	42	22.373	-1	-0.30497	0.5666 0.4559

En la tabla ANOVA para estos dos modelos vemos que como la significación de cambio en F, o la probabilidad de F-para-eliminar a, fue mayor o igual a 0.05, entonces se elimina a del modelo. Continuando con el proceso, consideramos los resultados obtenidos para el modelo 2:

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
Intercepto	70.981	1.22751	57.8254	0.0000
habitantes	0.0000567492	0.0000278946	2.03441	0.0483
ingresos	0.000190107	0.000288334	0.659327	0.5133
asesinatos	-0.312221	0.044095	-7.08064	0.0000
universitarios	0.0365192	0.0216064	1.69021	0.0984
heladas	-0.00605943	0.00249883	-2.42491	0.0197
area	-8.63762E-7	0.00000166905	-0.517517	0.6075
densidad_pobl	-0.000861213	0.000652295	-1.32028	0.1939

Por lo que "área" es la segunda variable para considerar. Así:

Modelo ajustado 3:

Esperanza de vida= 71.309 + 0.0000581066*habitantes + 0.00013236*ingresos - 0.320761*asesinatos + 0.0349908*universitarios - 0.00619119*heladas - 0.0007324*densidad_pobl

Observamos los resultados del anova entre el modelo 2 y el modelo 3:

Analysis of Variance Table

Model 1: esp_vida ~ habitantes + ingresos + asesinatos + universitarios + heladas + area + densidad_pobl

Model 2: esp_vida ~ habitantes + ingresos + asesinatos + universitarios + heladas + densidad_pobl

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	42	22.373			
2	43	22.516	-1	-0.14267	0.2678 0.6075

Como la significación de cambio en F, o la probabilidad de F-para-eliminar esta variable, fue mayor o igual a 0.05, entonces "área" se elimina a del modelo. Considerando el modelo 3:

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
Intercepto	71.309	1.04222	68.4201	0.0000
habitantes	0.0000581066	0.0000275336	2.11039	0.0407
ingresos	0.00013236	0.000263594	0.502134	0.6181
asesinatos	-0.320761	0.0405409	-7.91203	0.0000
universitarios	0.0349908	0.0212206	1.64891	0.1065
heladas	-0.00619119	0.00246457	-2.51208	0.0158
densidad_pobl	-0.0007324	0.000597783	-1.22519	0.2272

Por lo que "ingresos" fue eliminada para ajustar nuestro cuarto modelo:

Modelo ajustado 4:

Esperanza de vida= 71.418 + 0.0000608327*habitantes - 0.316047*asesinatos + 0.0423318*universitarios - 0.00599877*heladas - 0.000586356*densidad_pobl

Observamos los resultados del anova entre el modelo 3 y el modelo 4:

Analysis of Variance Table

Model 1: esp_vida ~ habitantes + ingresos + asesinatos + universitarios + heladas + densidad_pobl

Model 2: esp_vida ~ habitantes + asesinatos + universitarios + heladas + densidad_pobl

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	43	22.516			
2	44	22.648	-1	-0.13203	0.2521 0.6181

Nuevamente, la significación de cambio en F, o la probabilidad de F-para-eliminar la variable, fue mayor o igual a 0.05, entonces "ingresos" se elimina. Considerando el modelo 4:

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
Intercepto	71.418	1.01066	70.6645	0.0000
habitantes	0.0000608327	0.0000267627	2.27304	0.0280
asesinatos	-0.316047	0.0391025	-8.08253	0.0000
universitarios	0.0423318	0.0152499	2.77587	0.0081
heladas	-0.00599877	0.00241381	-2.48518	0.0168
densidad_pobl	-0.000586356	0.0005178	-1.1324	0.2636

La siguiente variable eliminada sería "densidad de población", así:

Modelo ajustado 5:

Esperanza de vida = 71.0271 + 0.00005014*habitantes - 0.300149*asesinatos + 0.0465822*universitarios - 0.00594329*heladas

Obtenemos el anova entre los modelos 4 y 5:

Analysis of Variance Table

Model 1: esp_vida ~ habitantes + asesinatos + universitarios + heladas + densidad_pobl

Model 2: esp_vida ~ habitantes + asesinatos + universitarios + heladas

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	44	22.648			
2	45	23.308	-1	-0.66005	1.2823 0.2636

Vemos que la significación de cambio en F, o la probabilidad de F-para-eliminar la variable fue mayor o igual a 0.05, entonces "densidad de población" se elimina.

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
Intercepto	71.0271	0.952853	74.5415	0.0000
habitantes	0.00005014	0.00002512	1.99602	0.0520
asesinatos	-0.300149	0.0366095	-8.19867	0.0000
universitarios	0.0465822	0.0148271	3.1417	0.0030
heladas	-0.00594329	0.00242087	-2.45502	0.0180

Del modelo 5 tenemos que la variable con mayor valor p es “habitantes”, ajustando este modelo:

Modelo ajustado 6:

Esperanza de vida = 71.0364 - 0.283065*asesinatos + 0.0499487*universitarios - 0.00691173*heladas

```
Residual standard error: 0.7427 on 46 degrees of freedom
Multiple R-squared:  0.7127,         Adjusted R-squared:  0.6939
F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

Obtenemos el anova entre los modelos 5 y 6:

Analysis of Variance Table

Model 1: esp_vida ~ habitantes + asesinatos + universitarios + heladas

Model 2: esp_vida ~ asesinatos + universitarios + heladas

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	45	23.308			
2	46	25.372	-1	-2.0636	3.9841 0.05201 .

Como la significación de cambio en F, o la probabilidad de F-para-eliminar “habitantes” fue mayor o igual a 0.05, entonces esta variable se elimina.

De nuestro modelo actual tenemos que todas las variables que lo forman son significativas:

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
Intercepto	71.0364	0.983262	72.2456	0.0000
asesinatos	-0.283065	0.0367313	-7.70637	0.0000
universitarios	0.0499487	0.0152011	3.28586	0.0020
heladas	-0.00691173	0.00244748	-2.82402	0.0070

Analysis of Variance Table

Response: esp_vida

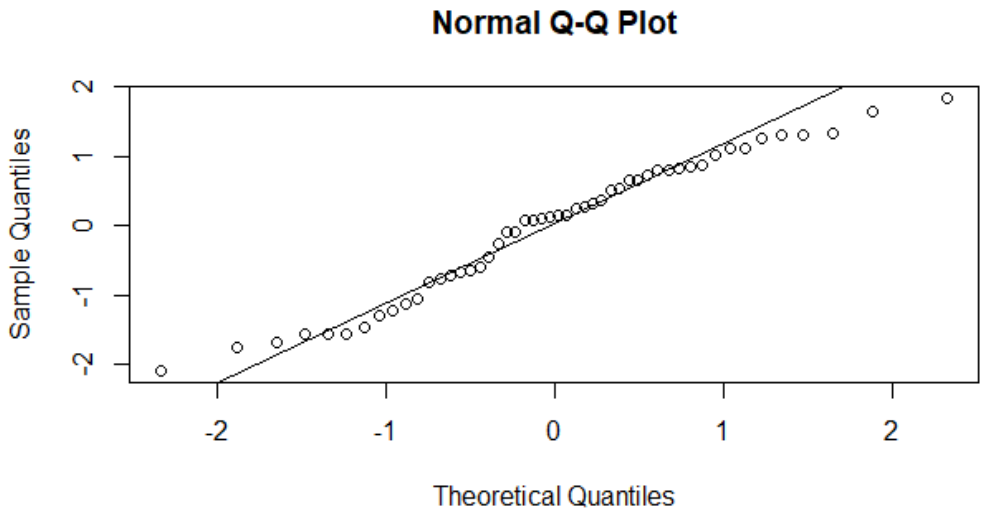
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
asesinatos	1	53.838	53.838	97.6103	5.971e-13 ***
universitarios	1	4.691	4.691	8.5050	0.005459 **
heladas	1	4.399	4.399	7.9751	0.006988 **

En el ANOVA es posible observar que todas las variables regresoras tienen un valor-p menor que el nivel de significancia del 5%, por lo que si existe una relación lineal entre ellas y la esperanza de vida. Adicionalmente, como el valor p del intercepto no fue mayor que 0.05, no se rechaza H_0 y se concluye que este difiere significativamente de cero, por lo tanto, como todas las variables regresoras resultaron significativas consideraremos a este modelo como el modelo final.

EXPLORACIÓN DE SUPUESTOS CORRESPONDIENTES AL ε_i

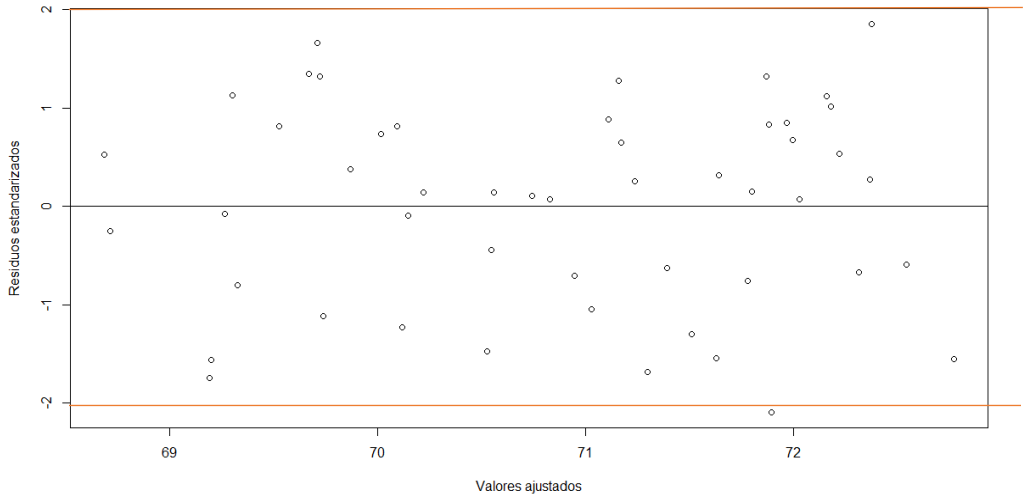
Prueba de normalidad

El siguiente grafico muestra la grafica de probabilidad normal, en el cual es posible observar q ue la mayoría de los residuales se encuentran cerca de la línea que marca la probabilidad norm al, aunque también existen algunos que se encuentran lejos de ella siendo así que el grafico no s sugiere que los residuales se comportan de manera normal.



Se realizo la prueba de Shapiro-Wilks, cuya hipótesis nula nos indica que la distribución se co mporta de manera normal, en la cual se obtuvo un p-valor de 0.1201 por lo tanto, no se rechaz ó la hipótesis nula indicando que los residuales se distribuyen de manera normal (W = 0.9631 2, p-value = 0.1201).

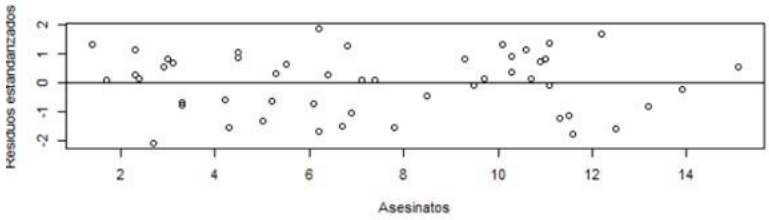
Linealidad, homocedasticidad e independencia de los residuales.

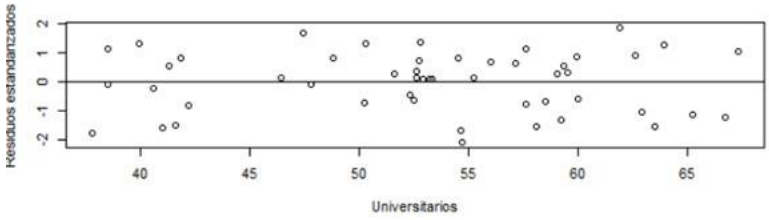


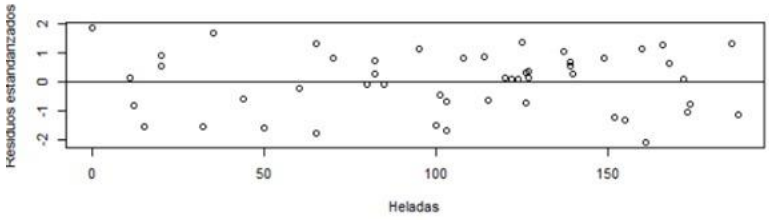
En el gráfico anterior, el de valores predichos vs residuos, a simple vista podemos observar que los residuales se encuentran dispersos de manera equitativa arriba y debajo de la línea cero, lo que nos indica que el supuesto de linealidad se cumple.

Con respecto al supuesto de independencia, es posible observar que no se encuentra algún patrón marcado que nos indique que los residuales no cumplen con este supuesto, también es posible observar que todos los puntos, salvo uno, se encuentran de forma aleatoria dentro de la franja entre – 2 y 2.

Se analizan las gráficas de los residuos en función de los valores correspondientes de cada variable regresora. Vemos que, para las tres gráficas, la mayoría de los puntos se encuentran dentro de la franja horizontal, salvo unos datos atípicos que no presentan diferencias significativas en la toma de decisiones, por lo que se cumple el supuesto de igualdad de varianzas.







PREDICCIÓN

Con los supuestos comprobados y en orden, el modelo final y los intervalos de confianza con n ivel de significancia del 5% para los regresores e intercepto son los siguientes:

Parámetro	*Estimación*	*2.5%*	*Estadístico T*
Intercepto	71.0364	69.0571	73.01558
asesinatos	-0.283065	-0.357001	-0.209128
universitarios	0.0499487	0.019350	0.080546
heladas	-0.00691173	-0.011838	-0.001985

Modelo:

Esperanza de vida = 71.0364 - 0.283065*asesinatos + 0.0499487*universitarios - 0.00691173*heladas

```
Residual standard error: 0.7427 on 46 degrees of freedom
Multiple R-squared:  0.7127,         Adjusted R-squared:  0.6939
F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

Los intervalos de confianza nos indican cuanto esperamos que aumente o disminuya la esperanza de vida cuando una variable regresora aumenta, por ejemplo, para el número de asesinatos, se esperaría que la esperanza de vida disminuya en un intervalo de -0.3570 a -0.2091 años cuando el número de asesinatos aumenta en una unidad.

El coeficiente de determinación R-cuadrada fue igual a 0.7127 por lo tanto, el 71.27% de la variabilidad de la esperanza de vida queda explicada por el modelo que contiene a las variables asesinatos, universitarios y heladas.

Creímos interesante ver la comparación de ciudades hipotéticas cuyos datos fueran cercanos a los mínimos, máximos y medias de las variables significativas.

```
Predicciones:
Ciudad  asesinatos universitarios heladas      fit      lwr      upr
1               2             67       1   73.80990  72.07968  75.54012
2               2             38       1   72.36139  70.58364  74.13913
3               7             53     104   70.98338  69.47329  72.49348
4              15             67     185   68.85829  67.05185  70.66474
5              15             38     185   67.40978  65.67048  69.14908
```

Como primera y segunda predicción tomamos una ciudad que tuviera un número de asesinatos y heladas cercanas al mínimo de estas variables y que su diferencia fuera que tuviera un número de universitarios cercano al máximo y mínimo de esta variable. La ciudad 1 es la que tiene el número de universitarios cercano al máximo y podemos ver la esperanza de vida pronosticada fue de 73.80990 y para la ciudad 2, la cual tiene un número de universitarios cercano al mínimo, fue de 72.36149, la diferencia es de 1.44851 años entre ellas dos lo cual no se podría considerar como grande.

La tercera predicción fue de una ciudad que tuviera en todas las variables números cercanos a la media en todas ellas, es decir, una ciudad promedio en todos los aspectos. La predicción resulto ser de 70.98338 lo cual, comparando con la primera ciudad, se puede observar una diferencia de 2.82652, es decir, casi 3 años de diferencia en la esperanza de vida.

Como cuarta y quinta predicción tomamos una ciudad que tuviera un número de asesinatos y heladas cercanas al máximo de estas variables y que su diferencia fuera que tuviera un número de universitarios cercano al máximo y mínimo de esta variable. La ciudad 4 es la que tiene el número de universitarios cercano al máximo y podemos ver la esperanza de vida pronosticada fue de 68.85829 y para la ciudad 2, la cual tiene un número de universitarios cercano al mínimo, fue de 67.40978, la diferencia es de 1.44851 años entre ellas dos lo cual no se podría considerar como grande.

Comparando la ciudad 1 con la 4 podemos ver que la diferencia en la esperanza de vida es de 4.95161 años, es decir, casi 5 lo años lo cual sí se puede considerar como grande. En base a estas comparaciones reafirmamos que el numero de asesinatos de una ciudad es la variable que más influye en la esperanza de vida de las ciudades.