Práctica 1

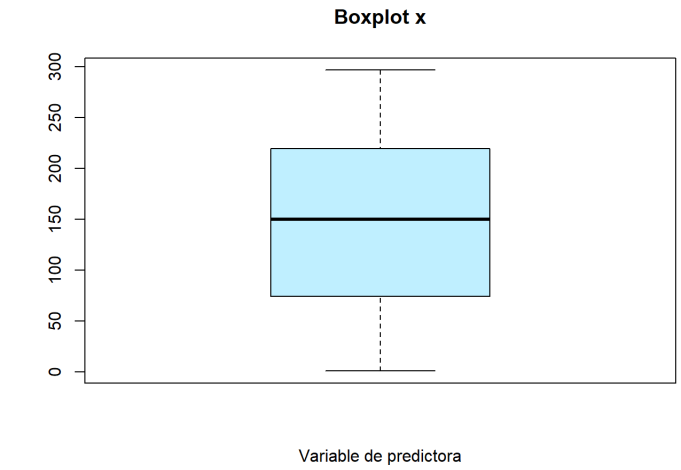
Agustin Riquelme y Heriberto Espino

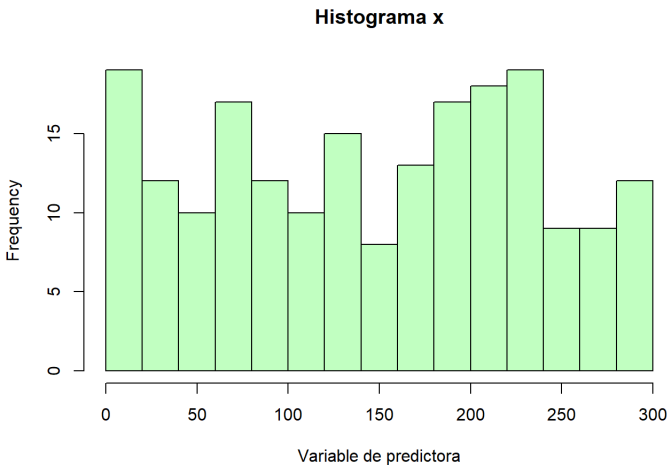
Con el objetivo de sintetizar las líneas de código lo más posible, se ha realizado una función la cuál necesitará como argumentos la variable de respuesta y predictora para el modelo y el nivel de significancia α, esta función será usada a lo largo del presente documento y con el objetivo de visualizar los datos de manera más "amigable", el código de la función se ocultará para el reporte pero permanecerá en el formato markdown para su correcta ejecución.

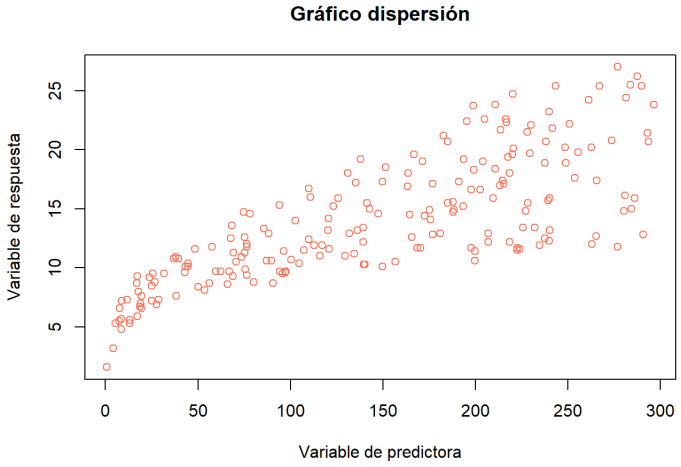
A continuación se mostrarán los resultados obtenidos para cada modelo de regresión líneal simple, considerando a la variable sales como la variable de respuesta, y los gastos de marketing en cada una de las plataformas serían las variables de predicción, pues las ventas pueden tener relación con el medio por el cuál se realiza el marketing y en ese caso, podríamos predecir las ventas a partir de una campaña de marketing ya propuesta para ver si es óptima o no.

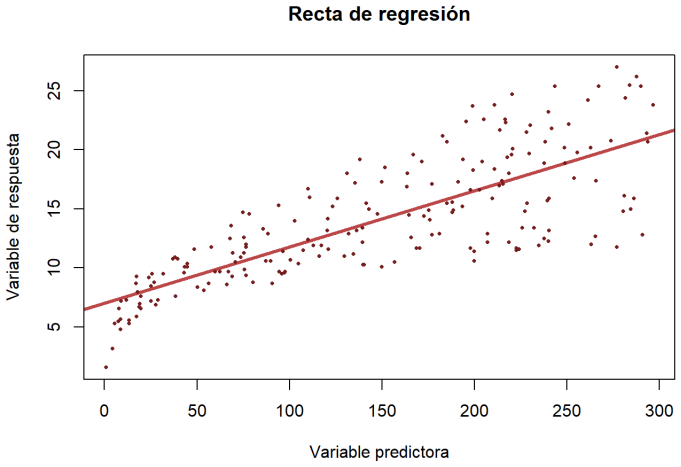
Para el modelo de regresión conformado por las ventas y el marketing de TV se obtuvo lo siguiente:

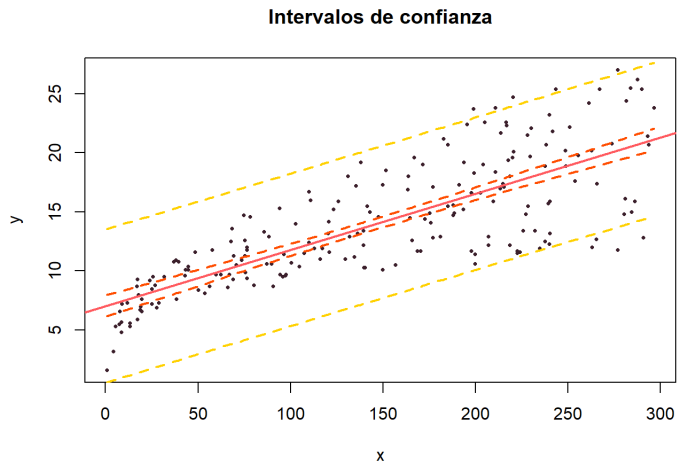
```
linealmodel(tv, sales, 0.05)
```

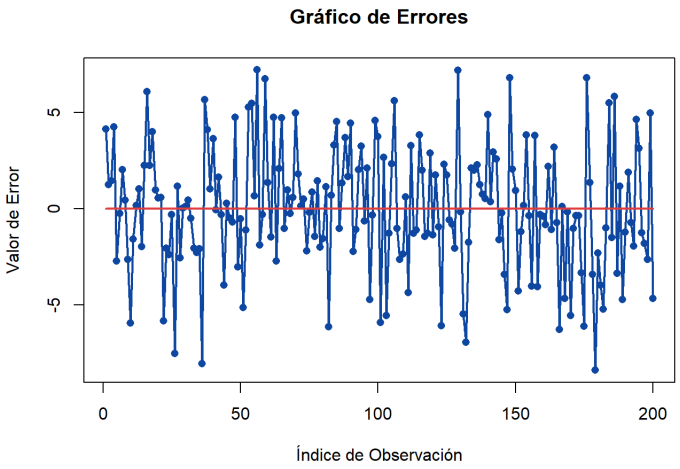


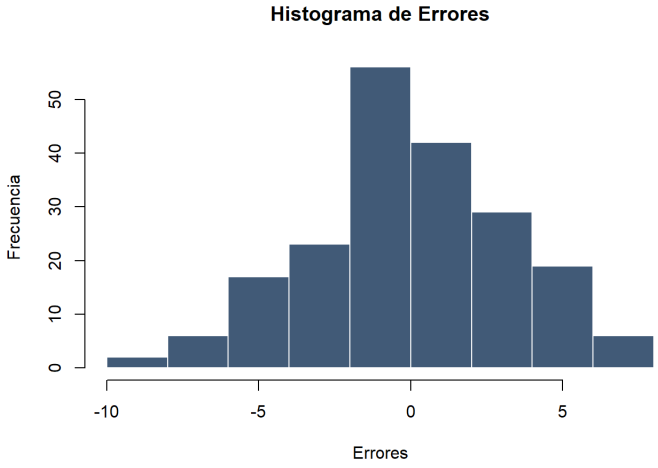


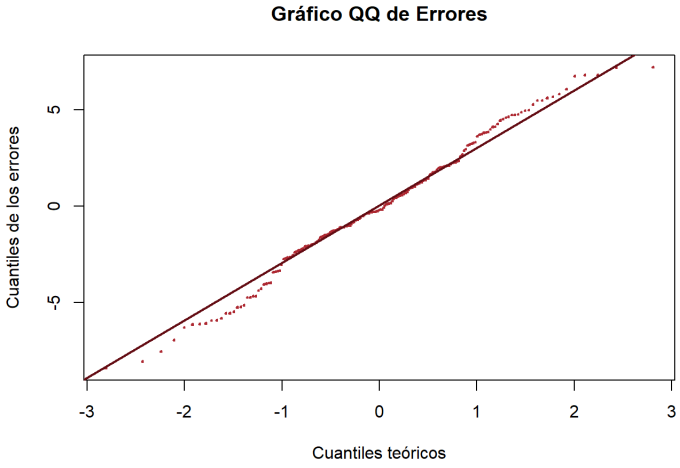












Informacion del modelo

Datos	Valores
β0	7.0325935491277
β1	0.0475366404330198
S	3.25865636865046
S cuad	10.6188413289462
cv	2.21613232136999
df	198
qti	-1.97201747783631
qtd	1.97201747783631
tc	17.6676256008755
p value	1.46738970019465e-42
Int. conf.	0.0422307160326923
	0.0528425648333472
Prom. err.	2.54980603892749
Var. err.	4.03129857236752
P. Hip. β1 Est.	Rechazamos H0
P. Hip. β1 Pvalue	Rechazamos H0

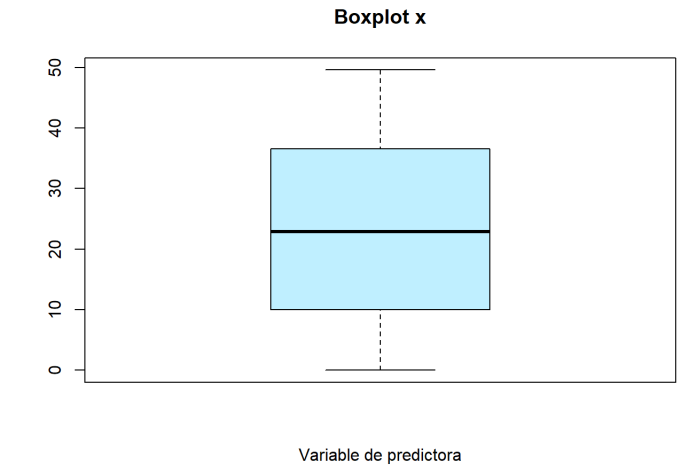
Las tres primeras gráficas son representantes de nuestro vector que almacena todos los presupuestos publicitarios (en miles de dólares) para TV. Principalmente, viendo sólo el boxplot podriamos intuir que la manera en la que se distribuyen los datos es una normal, pues el 50% de los datos se encuentra en el centro y la mediana está justo al centro del cuadrado, además, el valor máximo y el mínimo se encuentran a la misma distancia que los cuartiles, por lo que todo indicaría que se trata de una distribución normal, sin embargo, al momento de realizar el histograma de los mismos podemos comprobar que no se trata de la distribución pensada. Sin embargo, por el tercer gráfico, podemos decir que existe una corelación lineal positiva entre las ventas y el presupuesto hacia TV. Finalmente en el cuarto gráfico podremos corroborar que la recta de regresión muestra lo pensado.

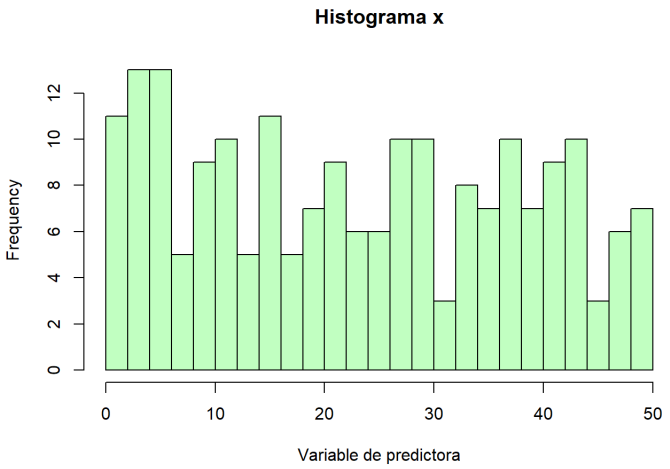
Seguido de estos gráficos tendremos el de los intervalos de confianza para los valores medios y predicción de la variable de respuesta podemos observar que conforme los valores crecen, el modelo ya no se ajusta consistentemente, pues existen datos en la muestra que se salen de los intervalos de confianza, y si nos fijamos en el gráfico de dispersión que muestra la relación entre las variables, si tuvieramos que graficar dos líneas rectas que encerraran todos los puntos, estas no tendrían la misma pendiente y serían en forma de cono, por lo que conforme crecen los valores del modelo, más se separan entre sí, indicando que para valores altos, el modelo no se ajusta lo suficiente.

Por otro lado, los gráficos de los errores nos muestran un error máximo aproximado de 8 por lo que podemos observar en la gráfica de línea, y posteriormente al calcular el promedio de los errores, nos da 2.54 que es un valor relativamente bajo, que podría indicar que el modelo a pesar de no tener un ajuste en los datos más grandes, aún así obtiene resultados precisos para los demás valores. La distribución de los errores está sesgada a la derecha y para el gráfico qq, los residuos intermedios del modelo, son los que más se ajustan a una distribución normal.

Nuestros betas son la manera para crear nuestra línea de regresión, donde con beta 1 al ser bajo representa que conforme nuestra variable predictora aumenta, la de respuesta igual pero en una proporción mucho menor, nuestra varianza residual es de 10, y tenemos un coeficiente de variación de 2, indicando que no existe una gran variabilidad relativa comparada con nuestra media, el intervalo de confianza de beta 1 nos explica entre que intervalo existe con un nivel de significancia alpha, nuestro beta 1, y como podemos observar, se encuentra cercano al cero pero ese número no esta en el intervalo, por lo que podemos decir con mayor confianza que rechazamos la hipótesis nula, indicando que nuestro beta1 es distinto de cero.

```
linealmodel(radio, sales, 0.05)
```


Boxplot x
0
10
20
30
40
50
Variable de predictora


Histograma x
Frequency
0
2
4
6
8
10
12
0
10
20
30
40
50
Variable de predictora

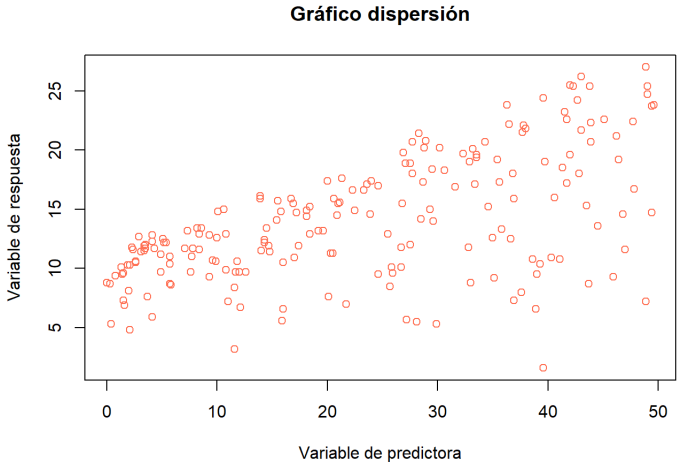
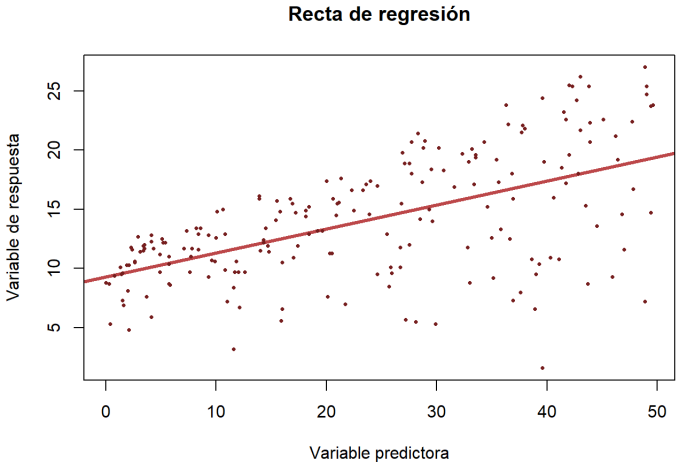
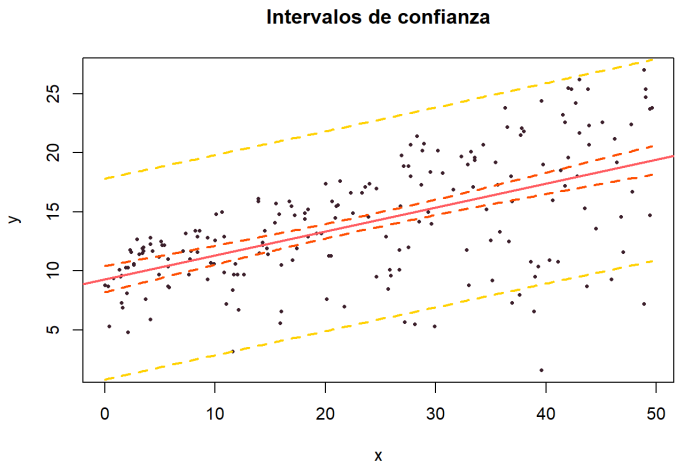
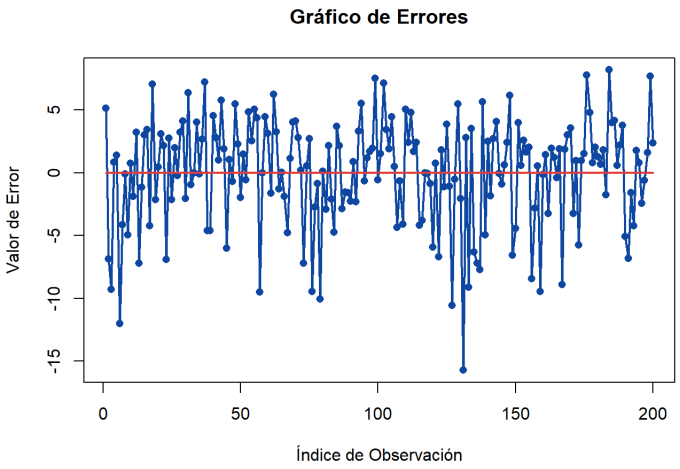
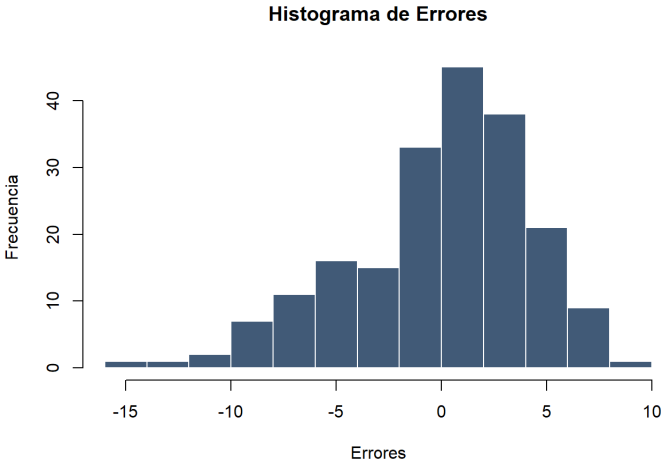


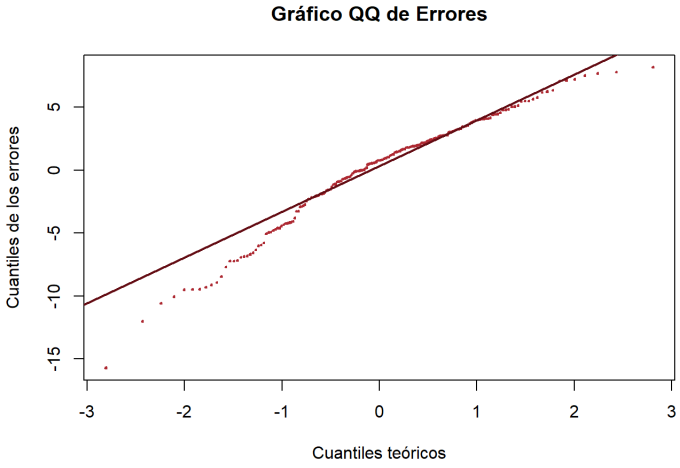
Gráfico dispersión
Variable de respuesta
5
10
15
20
25
0
10
20
30
40
50
Variable de predictora


Recta de regresión
Variable de respuesta
5
10
15
20
25
0
10
20
30
40
50
Variable predictora









Informacion del modelo

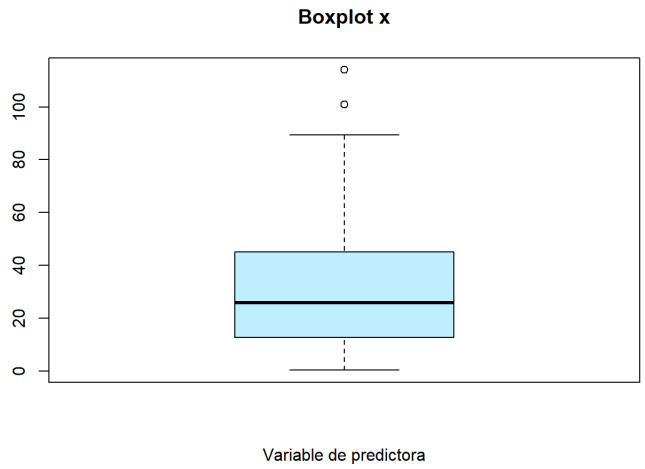
Datos	Valores
β0	9.31163809515829
β1	0.20249578339244
S	4.27494435490106
S cuad	18.2751492375005
cv	18.3757924471332
df	198
qti	-1.97201747783631
qtd	1.97201747783631
tc	9.92076547282495
p value	4.35496600176698e-19
Int. conf.	0.162244330504869
	0.24274723628001
Prom. err.	3.32021879764207
Var. err.	7.10406520694444
P. Hip. β1 Est.	Rechazamos H0
P. Hip. β1 Pvalue	Rechazamos H0

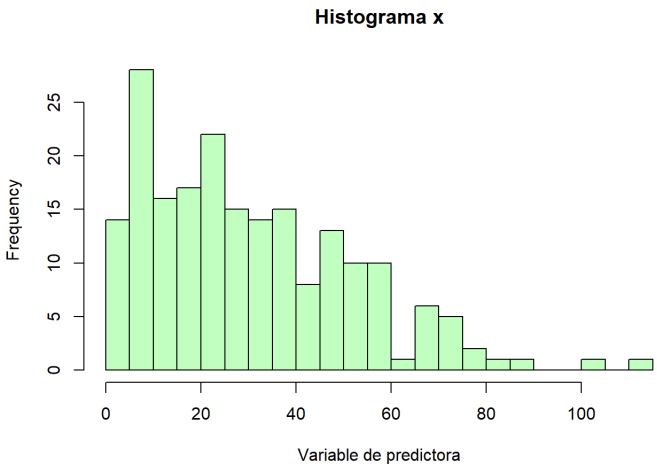
Con nuestros datos del nuevo modelo con radio como variable predictora, podemos ver en el boxplot que nuestros datos se encuentran sesgados a la izquierda, y en el histograma podemos ver de manera similar que los datos se encuentran un poco más condensados del lado izquierdo. Sin embargo no se aproximan a una distribución conocida. Y por lo aprendido en modelo anterior, podemos ver que en este caso se encuentran más dispersos los datos, incluso existiendo uno que otro outlier en la relación. Al graficar la línea de regresión podemos ver que los puntos en esta ocasión se encuentran más lejanos de la misma y se espera un promedio de errores mayor que el modelo anterior.

Para los intervalos de confianza podemos observar que estos encierran mayormente los datos excepto por algunos outliers del modelo, sin embago estos intervalos parecen estar más amplios que el pasado, pues los datos se alejan más de la línea de regresión. Los errores en los gráficos podemos observar que sí están más alejados en la parte inferior, mientras que en la parte superior a la línea, observamos comportamientos de errores similares a los del modelo pasado, en este caso el histograma de dichos errores igualmente esta sesgado a la derecha y para el caso del ajuste a una distribución normal, dichos errroes se aproximan más a una normal conforme los presupuestos crecen.

En este caso,, nuestro beta 1 es un valor más alto, indicando que un aumento en la variable predictora, aumentan en mucha mayor proporción los resultados de nuestra variable de respuesta, sin embargo nos encontramos con una varianza mayor al estar los datos más dispersos entre sí, casi el doble que el modelo pasado, y de la misma manera el coeficiente de variación indica una alta variabilidad relativa con respecto de la media. El intervalo de confianza para beta 1 se encuentra entre 0.16 y 0.25 con un nivel de significacncia alpha, queriendo decir en conjunto con el p valor que se rechaza la hipótesis nula que indica que beta 1 puede ser cero. Como deducimos, los errores son mayores, pues su promedio es 3.32, sin embargo se esperaba que por la forma de la gráfica y la aproximación de la línea, estos serían mayores.

```
linealmodel(newsp, sales, 0.05)
```


Boxplot x
0
20
40
60
80
100
Variable de predictora


Histograma x
Frequency
0
5
10
15
20
25
0
20
40
60
80
100
Variable de predictora

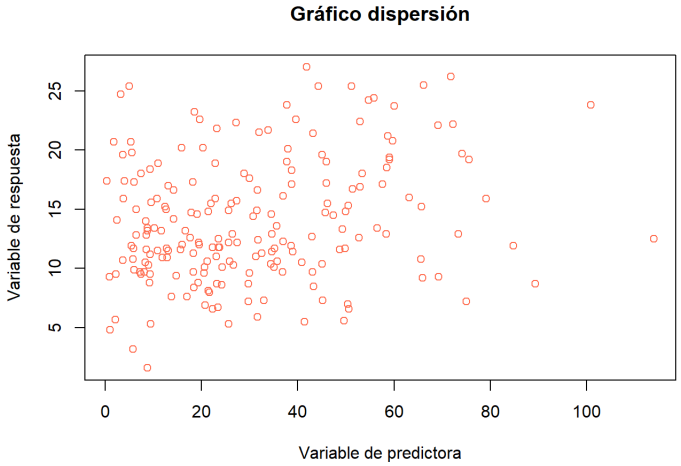
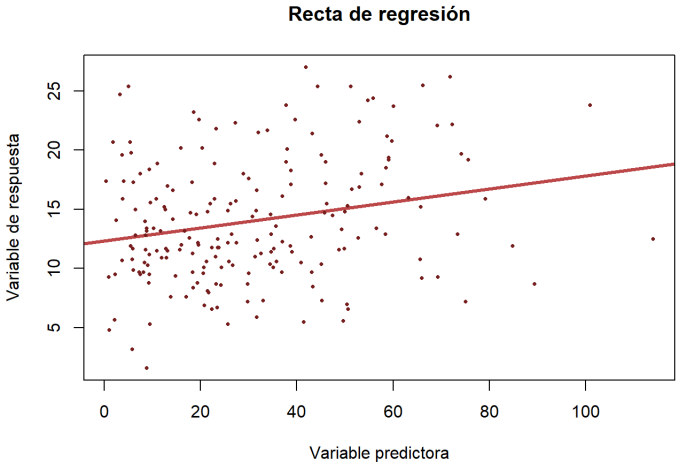
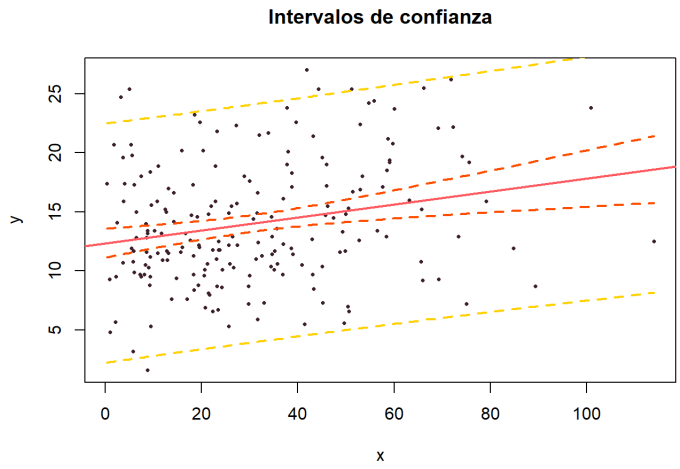
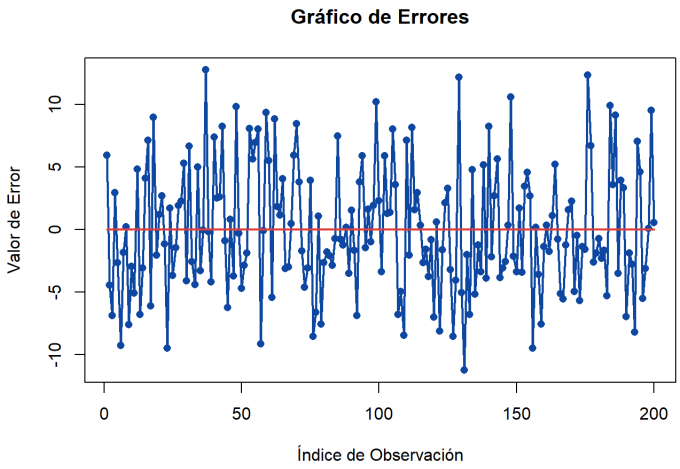
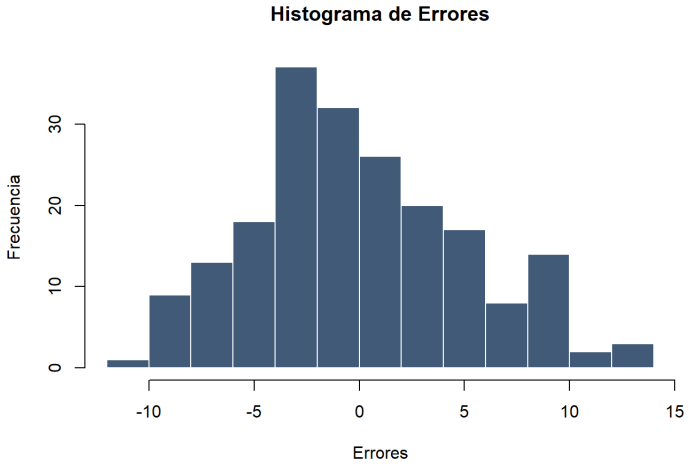


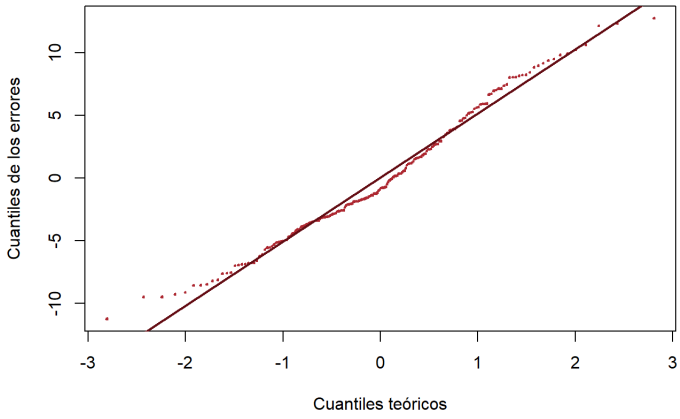
Gráfico dispersión
Variable de respuesta
5
10
15
20
25
0
20
40
60
80
100
Variable de predictora


Recta de regresión
Variable de respuesta
5
10
15
20
25
0
20
40
60
80
100
Variable predictora









Informacion del modelo

Datos	Valores
β0	12.3514070692782
β1	0.0546930984722734
S	5.09248036652019
S cuad	25.9333562833936
cv	16.667147890686
df	198
qti	-1.97201747783631
qtd	1.97201747783631
tc	3.29959074363342
p value	0.00114819586888821
Int. conf.	0.0220054852243413
	0.0873807117202054
Prom. err.	4.14655974383849
Var. err.	8.52267840335456
P. Hip. β1 Est.	Rechazamos H0
P. Hip. β1 Pvalue	Rechazamos H0

Para nuestro último modelo donde la variable predictora corresponde a los gastos publicitarios en periodicos, podemos observar dos grandes outliers por el box plot que se encuentran muy alejados de nuestros datos, a su vez podemos decir que se encuentran los datos muy sesgados a la izquierda y en el histograma podemos observar que sí sucede de esta manera, observando que la mayoría de los presupuestos son bajos para este tipo de publicidad. En este caso sí podemos observar aún más una mayor variabilidad en los datos, pues aún se encuentran más dispersos los puntos, incluso más que los modelos pasados, incluso diciendo que este tendrá los errores más altos por su separación de la línea de regresión. Los outliers vistos corresponden que los gastos son mayores a lo que se observa generalmente

Para el caso de los intervalos de confianza, podemos observar que los intervalos dado igualmente se ajustan pero dejan algunos datos superiores fuera de dichos intervalos, pero como predijimos, los errores son bastante altos, pues varian y llegan hasta más de 10, errores que en promedio son de 4.14 que sí es el error promedio máximo de los tres modelos. El histograma de estos errores muestra un comportamiento sesgado a la izquierda y observamos por el gráfico qq que este es el modelo que menos se ajusta a una distribución normal.

Como todos los datos anteriores, la varianza de este modelo es la más alta, por lo que nuestro datos no parecen depender mucho del modelo, se ajustan muy poco y no parecen tener un impacto sobre la tendencia de los datos. La variación promedio relativa con respecto a la media igual es bastante alta indicando una poca confiabilidad en el modelo. Beta 1 es un valor pequeño, es decir, no cambia en mucha proporción conforme aumentas la variable predictora y además con esto y el p valor rechazamos de nuevo la hipótesis nula y decimos que beta 1 es distinto de cero