Análisis de Datos con el Sistema Estadístico R

Lic. Patricia Vásquez Sotero

Análisis de regresión lineal: Diagnosis y pronóstico

Contenidos

- **Modelo de regresión lineal simple**
  - Diagnosis
    - Análisis de los residuos del modelo
      - Análisis gráfico y confirmatorio
      - Observaciones influentes
  - Predicción

DIAGNOSIS DEL MODELO DE REGRESIÓN LINEAL SIMPLE

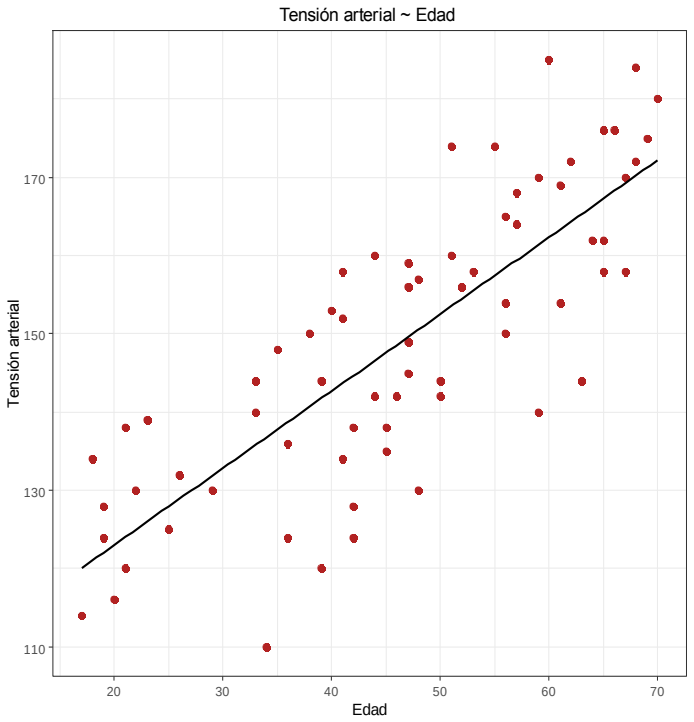
Representación gráfica de un modelo

Si bien la función plot(lm) es una forma muy rápida de obtener los gráficos. A continuación se describe como obtener las mismas representaciones mediante el sistema gráfico ggplot2. Más información en https://www.statmethods.net/advgraphs/ggplot2.html.

Ejemplo. Retomando el ejemplo de pacientes.

Ejemplo en código R

```
# gggplot2: mapea las variables al diseño,
  con gráficas primitivas a usar y se ocupa
  de los detalles.
require(ggplot2)
ggplot(mapping = aes(edad, tas)) +
  geom_point(color = "firebrick", size = 2) +
  labs(title  =  'Tensión arterial ~ Edad ',
  x='Edad', y='Tensión arterial') +
  geom_smooth(method = "lm", se = FALSE,
  color = "black") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



Diagnóstico del modelo

Verificar condiciones para aceptar el modelo

- plot(modelo) -> Análisis de los residuos (distribución, variabilidad…)
- shapiro.test(modelo$residuals) -> Test de hipótesis de Shapiro Wilk para el análisis de normalidad
- bptest(modelo) -> Test de contraste de homocedasticidad Breusch-Pagan
- influence.measures(modelo) -> Detección de observaciones influyentes
- influencePlot(modelo) -> Visualización de observaciones influyentes
- outlierTest(modelo) -> Test de detección de outliers
- rstudent(modelo) -> Cálculo de residuos estudentizados

Donde, modelo es el objeto que contiene resultados de la regresión.

Diagnóstico del modelo

Análisis de los residuos

Si las hipótesis del modelo son ciertas, entonces los residuos son aproximadamente

1. **Normales**
2. **Media cero**
3. **Independientes**
4. **Varianza constante**
5. **No hay residuos atípicos**

Podemos utilizar contrastes y gráficos para ver si hay EVIDENCIA CLARA en contra de alguna de las hipótesis

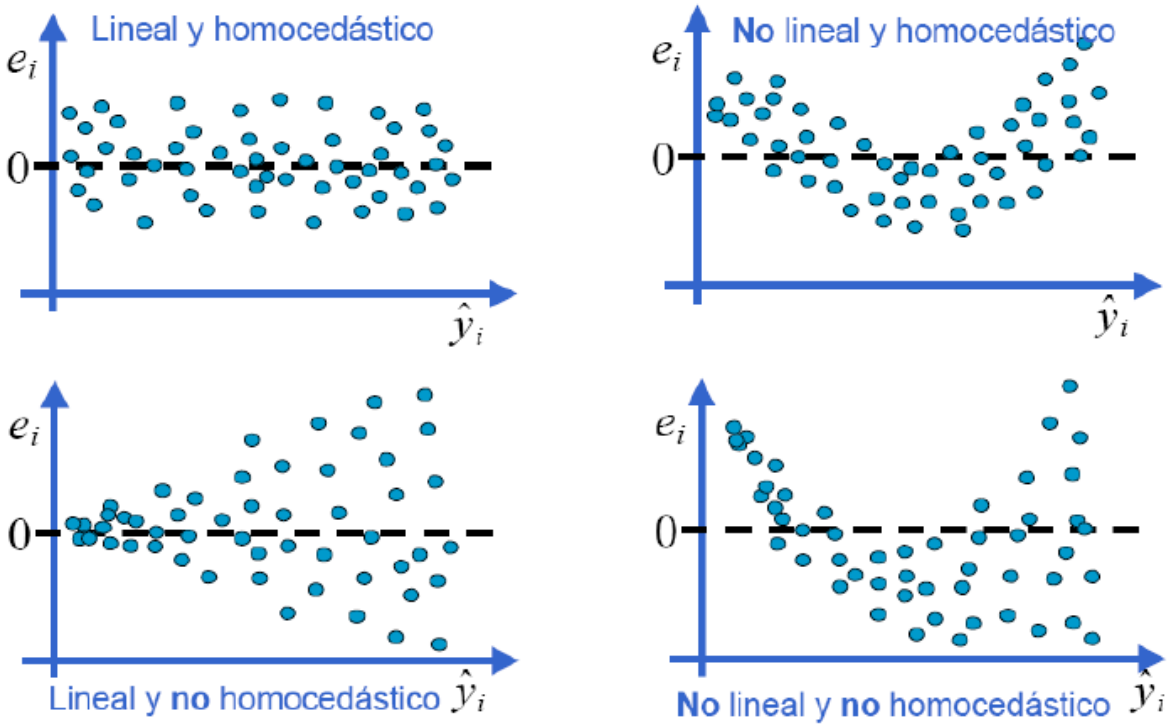
- Normalidad
  - Histograma de los *residuos tipificados*
  - Q-Q plot de los *residuos tipificados*
  - Test de K-S de los *residuos tipificados*
- Linealidad / Homocedasticidad
  - Diagrama de dispersión de los *residuos tipificados* frente a los *valores pronosticados/ajustados*

Tienen que estar entre -2 y 2, en una nube de puntos sin forma

Análisis de los residuos

Relación lineal entre variable dependiente e independiente: Se calculan los residuos para cada observación y se representan (scatterplot). Si las observaciones siguen la línea del modelo, los residuos se deben distribuir aleatoriamente entorno al valor 0.

¿Se cumplen las hipótesis del modelo?



Análisis de los residuos

R incluye una serie de gráficos que permiten hacer un diagnóstico del modelo ajustado

Gráficas

- Valores ajustados vs Residuos
  Deberían ser independientes
- QQ-Plot de los Residuos tipificados
  Deberían ajustar a una línea recta
- Valores ajustados vs Raíz cuadrada residuos
  Permite localizar valores atípicos y ver si para valores ajustados grandes hay desviaciones grandes (p. ej.)
- *Leverage* vs Residuos tipificados

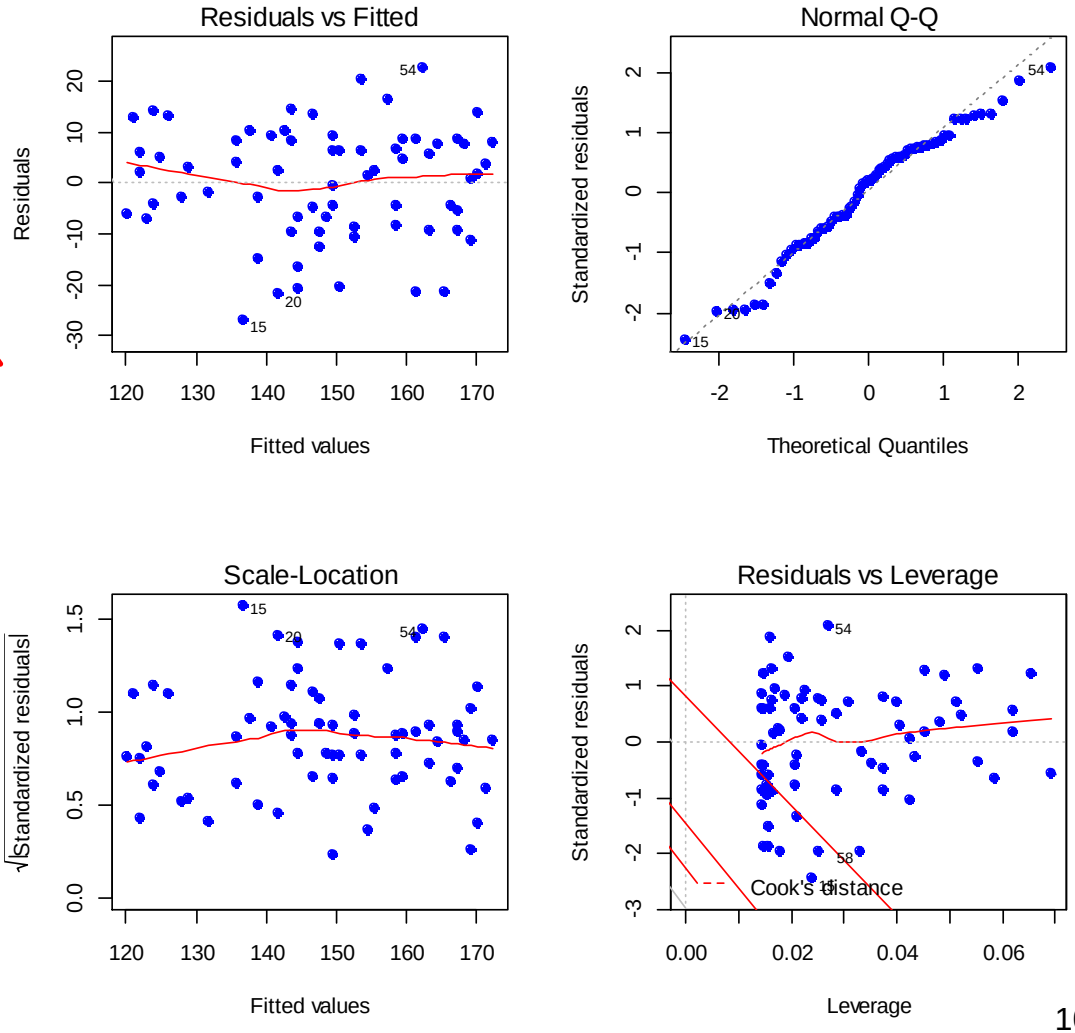
Ejemplo en código R - Ejemplo pacientes

```
par(mfrow=c(2,2))
plot(regresion, col=red)
```

Análisis de los residuos

Resultados

En este gráfico observamos que los residuos van entorno a valor central, excepto por los casos 15, 20 y 54.



Análisis de los residuos

Ejemplo. Pacientes.

Ejemplo en código R

```
# Contraste de hipótesis (normalidad de los residuos)
shapiro.test(regresion$residuals)
# Test de Breush-Pagan (homocedasticidad de los residuos)
library(lmtest)
bptest(regresion)
```

Resultados d análisis confirmatorios

```
> shapiro.test(regresion$residuals)

        Shapiro-Wilk normality test

data:  regresion$residuals
W = 0.97262, p-value = 0.1339

> bptest(regresion)

        studentized Breusch-Pagan test

data:  regresion
BP = 0.0635, df = 1, p-value = 0.801
```

Los resultados evidencia el cumplimiento de la normalidad y homocedasticidad de los residuos con un nivel significancia del 5%.

Análisis de los residuos

Ejemplo. Pacientes.

Ejemplo en código R

```
# La función lm() calcula y almacena los valores predichos por el modelo y
      los residuos
pacientes$prediccion <- regresion$fitted.values
pacientes$residuos <- regresion$residuals
pacientes$resiest <- rstudent(regresion) # Cálculo de residuos estudentizados
head(pacientes)
```

Resultados: Valores pronosticados de *tas* y residuos brutos y estudentizados

	tas	edad	prediccion	residuos	resiest
1	114	17	120.0732	-6.073152	-0.5645529
2	134	18	121.0567	12.943289	1.2111571
3	124	19	122.0403	1.959731	0.1810748
4	128	19	122.0403	5.959731	0.5517978
5	116	20	123.0238	-7.023828	-0.6497053
6	120	21	124.0074	-4.007386	-0.3692364

Análisis de los residuos

Ejemplo. Pacientes.

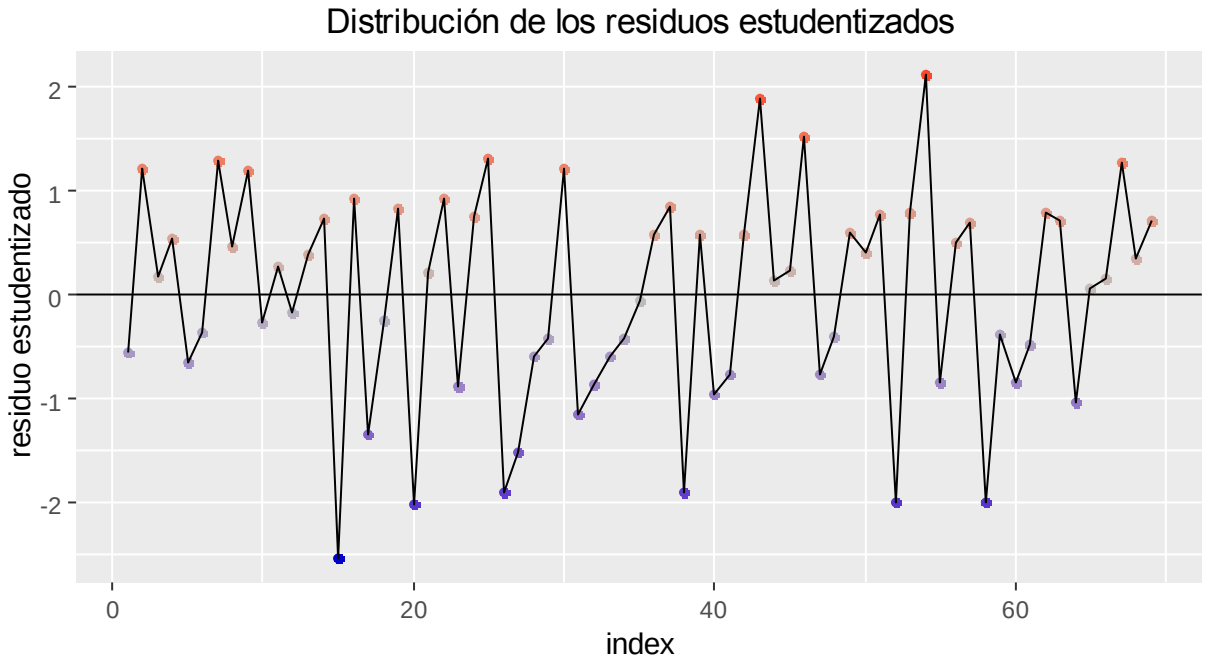
Ejemplo en código R

```
# Análisis gráfico de autocorrelación de los residuos
ggplot(data = pacientes, aes(x = seq_along(regresion$rstudent),
    y = regresion$rstudent)) +
geom_point(aes(color = regresion$rstudent)) +
scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
geom_line(size = 0.1) +
labs(title = "Distribución de los residuos estudentizados", x = "index", y =
    "residuo estudentizado")+
geom_hline(yintercept = 0) +
theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```

Análisis de los residuos

Ejemplo. Pacientes.

Resultados



En este caso, la normalidad de los residuos podemos aceptarla, y tampoco parecen seguir una clara tendencia según el orden de registro de las observaciones, y la condición de homocedasticidad parece cumplirse.

Análisis de los residuos

Observaciones influyentes

Para analizar en qué medida pueda estar influyendo una u otras observaciones, reajustaremos el modelo excluyendo posibles observaciones sospechosas. Dependiendo de la finalidad del modelo, la exclusión de posibles outliers debe analizarse con detalles, ya que estas observaciones podrían ser errores de medida, pero también podrían representar casos interesantes.

Ejemplo en código R

```
# Sobre los residuos estudentizados
which(abs(pacientes$resiest) > 3)
library(car)
summary(influence.measures(model = regresion))
influencePlot(model = regresion)
```

Análisis de los residuos

Observaciones influyentes

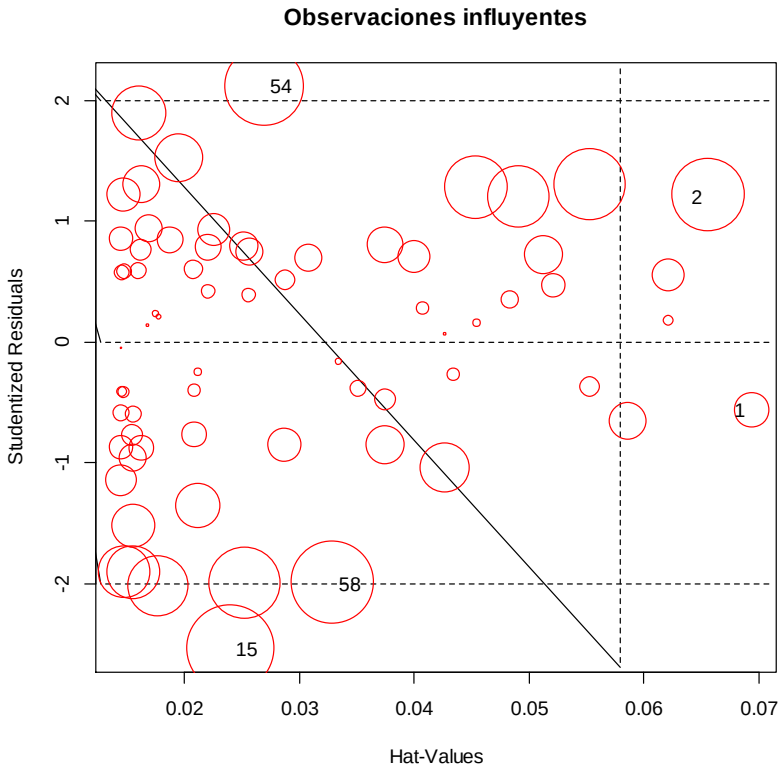
Ejemplo. Pacientes.

Resultados

```
Potentially influential observations of
        lm(formula = tas ~ edad) :

   dfb.1_ dfb.edad dffit cov.r   cook.d hat
1   -0.15    0.14   -0.15  1.10_*  0.01   0.07
3    0.05   -0.04    0.05  1.10_*  0.00   0.06
15  -0.33    0.25   -0.40  0.88_*  0.07   0.02
```

```
> influencePlot(model = regresion, main='O
      StudRes        Hat       CookD
1  -0.5645529 0.06934678 0.01199655
2   1.2111571 0.06564532 0.05117373
15 -2.5423418 0.02400465 0.07349217
54  2.1215300 0.02692760 0.05918355
58 -1.9969254 0.03288871 0.06491082
```



Como el porcentaje de casos influyentes no supera el 20% del total de casos podríamos considerarlos en el análisis; sin embargo, habría que revisarlos.

PREDICCIÓN

Predicción

Predicción

- A veces interesa predecir la respuesta para cierto valor de una variable.
- La predicción es

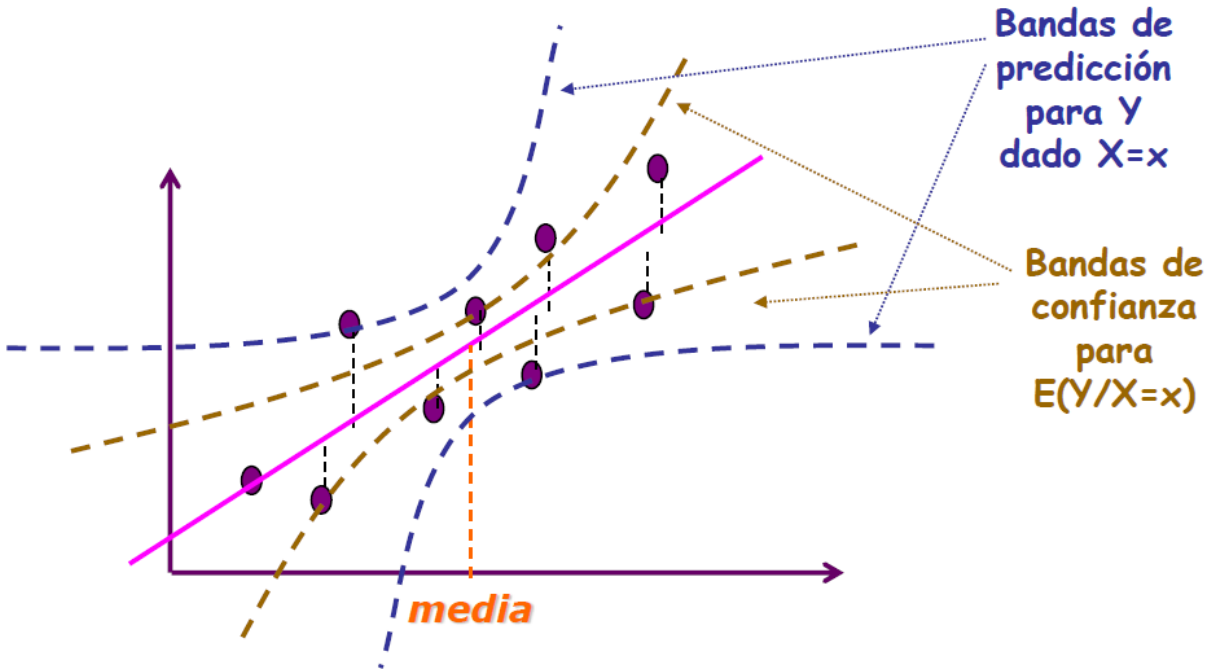
$$\hat{y_{pred}} = \hat{\alpha} + \hat{\beta} x_{pred}$$

- ¿Cuál es la incertidumbre acerca de la predicción? En general, mayor que la incertidumbre de los datos:

$$\hat{y}_{pred} \pm t_{n-2} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n}\right)}$$

Predicción

Gráficamente: Bandas de confianza y de predicción



Los dos bandas tienen la misma forma, siempre más estrechas en la media de las *x* donde hay más información

Predicción

Para poder representar el intervalo de confianza a lo largo de todo el modelo se recurre a la función predict() para predecir valores que abarquen todo el eje X. Se añaden al gráfico los límites superiores e inferiores calculados para cada predicción.

Ejemplo. Pacientes.

Ejemplo en código R

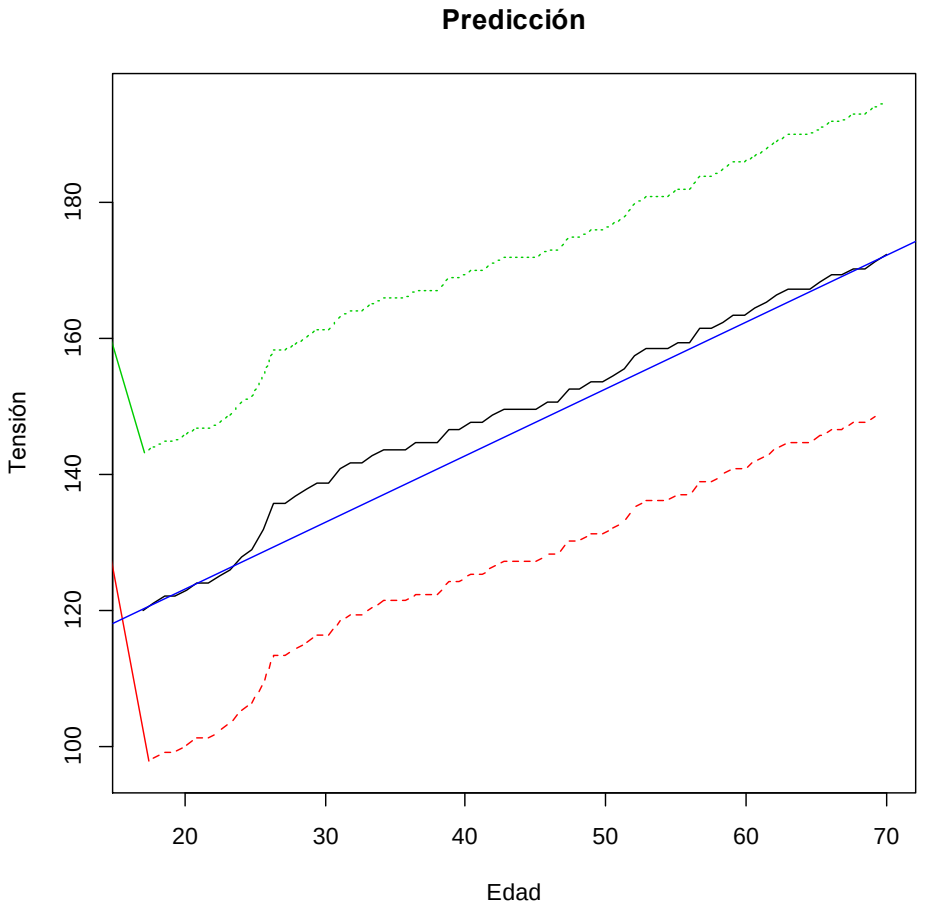
```
x0 <- seq(min(pacientes$edad), max(pacientes$edad), length = 69)
datos <- data.frame(tension = x0)
pred.ip <- predict(regresion, datos, interval = "prediction",
se.fit = TRUE, data = pacientes)
head(pred.ip$fit) # Muestra la primera parte de un objeto

matplot(x0, pred.ip$fit, type = "l", xlab = "Edad", ylab = "Tensión")
abline(regresion, col = 'blue')
```

Predicción

Resultado

	fit	lwr	upr
1	120.0732	97.17408	142.9722
2	121.0567	98.19730	143.9161
3	122.0403	99.21921	144.8613
4	122.0403	99.21921	144.8613
5	123.0238	100.23979	145.8079
6	124.0074	101.25903	146.7557



Comunicación constante con la Escuela del INEI

Correo de la Dirección Técnica de la ENEI
Sr. Eduardo Villa Morocho (Eduardo.villa@inei.gob.pe)

Coordinación Académica
Sra. María Elena Quirós Cubillas (Maria.Quiros@inei.gob.pe)

Correo de la Escuela del INEI
enei@inei.gob.pe

Área de Educación Virtual
Sr. Gonzalo Anchante (gonzalo.anchante@inei.gob.pe)