

9

1. Identifique quien es la variable de respuesta  $Y$  y quién es la variable regresora  $X$  y escriba el modelo de regresión.

La variable de respuesta  $Y$  es la resistencia de corte, ya que esta variable busca o explica.  
Mientras  $X$  es la edad ya que se utiliza para la predicción.

#### MODELO DE REGRESION

$$y_i = \beta_0 X_i + \varepsilon_i \quad \times \quad Y = \beta_0 + \beta_1 X$$

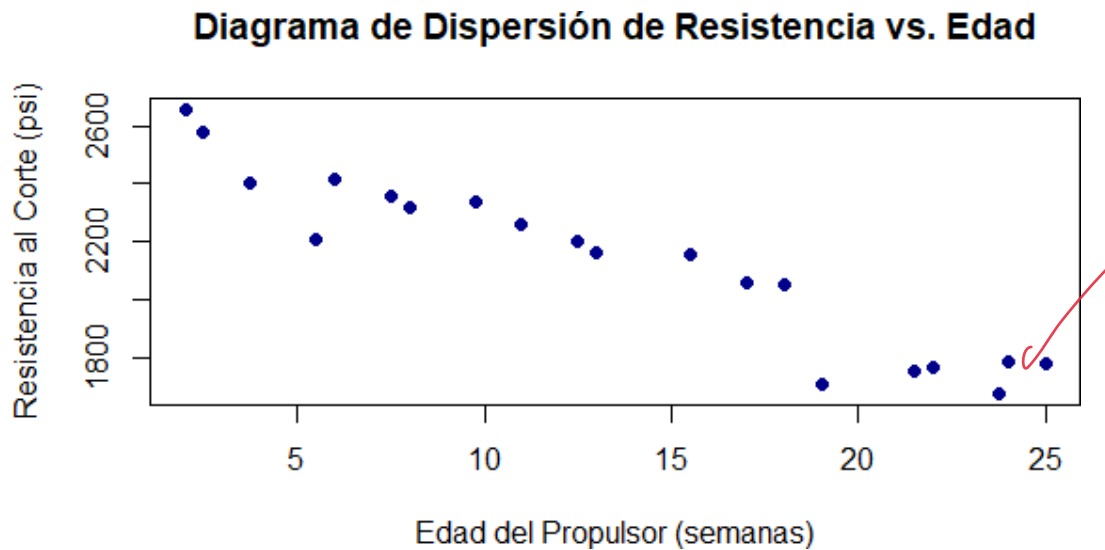
$Y_i$  es la Resistencia al corte para la observación  $i$ .

$\beta_0$  es el intercepto (resistencia promedio cuando la edad es cero). *no necesariamente*

$\beta_1$  es la pendiente (cambio promedio en la resistencia por cada semana de aumento en la edad).

$\varepsilon_i$  es el término de error aleatorio, con  $\varepsilon_i \sim N(0, \sigma^2)$

2. Grafique el diagrama de dispersión de los datos



3. Obtenga los estimadores para  $\theta_0$  y  $\theta_1$  y escriba la ecuación de la recta ajustada.

Para los cálculos de los estimadores mínimos cuadrados  $\theta_0$  y  $\theta_1$  se utilizaron los datos proporcionados.

Estimador de Intercepto  $\theta_0 = 2627.82$

Estimador de pendiente  $\beta_1 = -37.15$

Para la ecuación de la recta de regresión Ajustada:

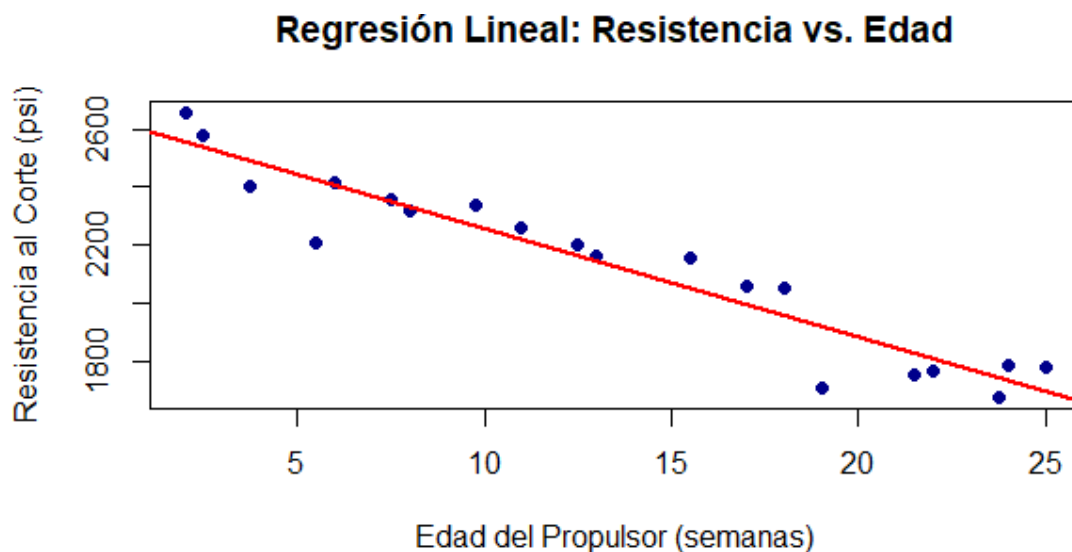
$$\beta_0 = Y - \beta_1 X$$

$$Y = \beta_0 + \beta_1 X$$

$$Y = 2627.82 - 37.15X$$

Esto nos dice que  $\beta_0$  que la resistencia promedio es de 2627.82 psi cuando la edad del propulsor de cero semanas y la pendiente  $\beta_1$  nos dice que, por cada aumento de una semana, la resistencia de corte promedio se estima disminuye en 37.15 psi.

4. Grafique la recta de la regresión junto con los datos. ¿Qué tan bueno cree que es el ajuste?



$$R^2 = 0.9221$$

Gracias a que se realizó estadísticamente el coeficiente de determinación que es de 0.9221 (92.21%) esto indica la variación total en la resistencia al corte. Siendo un valor cercano a 1 por lo que el modelo lineal es una buena aproximación a la relación entre variables. Por otro lado, la gráfica de dispersión nos muestra que los puntos están mínimamente dispersos y se agrupan de manera muy ajustada a la línea de regresión.

5. Efectué la prueba de significancia de la regresión para un nivel  $\alpha=0.05$ . Escriba el valor del  $p$  – valor. ¿Qué conclusiones puede hacer sobre  $\beta_1$ ?

Para este problema tenemos que realizar una hipótesis donde tenemos que  $\beta_1 = 0$  y  $\beta_1 \neq 0$ , donde si es igual a 0 no existe una relación lineal significativa por lo que la edad no afecta la resistencia, pero si es diferente de cero existe una relación significativa por lo que la edad si afecta la resistencia.

COEFICIENTE	ESTIMACION ( $\beta$ )	ERROR ESTANDAR	ESTADISTICO t	p-valor
INTERCEPTO $\beta_0$	2627.82	46.80	56.15	<<0.001
PENDIENTE $\beta_1$	-37.15	2.55	-14.59	<<0.001

Al tener esta información podemos ver que el coeficiente  $\beta_1$  es significativo por lo que es diferente de cero, teniendo una relación lineal significativa.

6. Suponga que se tienen tres lotes del propulsor tipo 1, con 5, 10 y 15 semanas de edad respectivamente. ¿Cuál es la estimación para la resistencia según el modelo de regresión (para cada lote)?

Para la estimación de la resistencia utilizaremos de nuevo la ecuación de la recta de regresión:

$$Y = 2627.82 - 37.15X$$

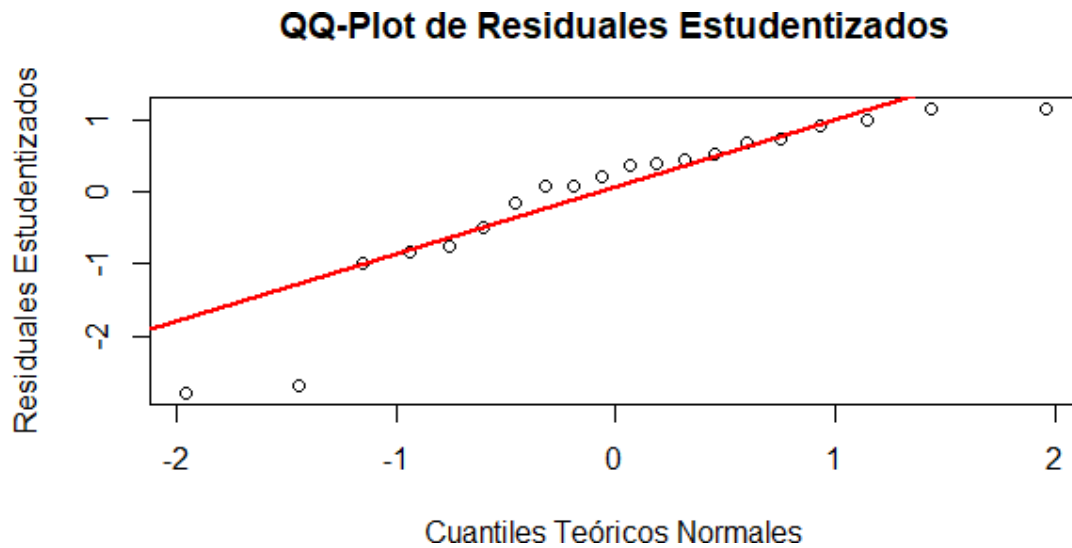
EDAD (X SEMANAS)	CALCULO ( )	RESISTENCIA ESTIMADA (Y psi)
5.0	$2627.82 - 37.15(5.0)$	2442.07 ✓
10.0	$2627.82 - 37.15(10.0)$	2256.32 ✓
15.0	$2627.82 - 37.15(15.0)$	2070.57 ✓

7. Calcule el valor del coeficiente de determinación (Múltiple R Squared). Según este coeficiente. ¿qué tan bueno es el ajuste de la regresión?

Para  $R^2$  tenemos que 0.9221 indica que el 92.21% de la variable observada en la resistencia del corte y es explicada por el modelo de regresión lineal de la edad. Por lo que este porcentaje es alto y está cerca del número 100%.

8. VERIFICACIÓN DE SUPUESTOS DEL MODELO. (Obtenga primero los residuales)

- a) NORMALIDAD. Grafique los residuales contra los cuantiles de una normal (qqnorm, qqline). ¿Se satisface este supuesto?



En la gráfica nos muestra que la mayoría de los círculos se alinean muy cerca de la recta por lo que las desviaciones son limitadas, aceptando que la distribución de los errores sigue una distribución normal con la que podemos validar las inferencias del modelo de regresión.

b) MEDIA CERO, VARIANZA CONSTANTE E INDEPENDENCIA. Grafique los residuales contra los predichos para verificar los tres supuestos (predict, rstudent).

¿Observa alguna anomalía?

Media cero: No se encuentra ninguna anomalía ya que el promedio de los residuales es cero, como se esperaba de los mínimos cuadrados, esto se observa en los puntos que

están distribuidos por encima y por debajo de la línea de referencia.

Varianza constante: Tampoco se encuentra anomalía ya que es constante en todo el rango de predicción.

Independencia: No se observa anomalías, se observa que los residuales parecen independientes entre sí, esto quiere decir que no se muestra un patrón como curvas, ondas o tendencias.

#### 9. PUNTOS ATÍPICOS E INFLUYENTES.

- a) Utilizando la gráfica anterior, ¿se observan puntos que puedan considerarse como atípicos (outliers)?

Se puede observar algunos residuales grandes como se puede observar en el punto cerca de -2.5 en el eje Y y el punto cerca de -2400 en el eje X, o el punto cerca de 1.0 en el eje Y y 1700 en el eje X, **ninguno de ellos supera consistentemente el umbral típico** para ser considerado un *outlier* significativo (generalmente residuales estudentizados fuera del rango de  $\pm 3$ )

- b) Utilizando la distancia de Cook, verifique si hay puntos influyentes.

Para la distancia de Cook se necesitan varios umbrales para identificar los puntos influyentes, un umbral que se puede utilizar es

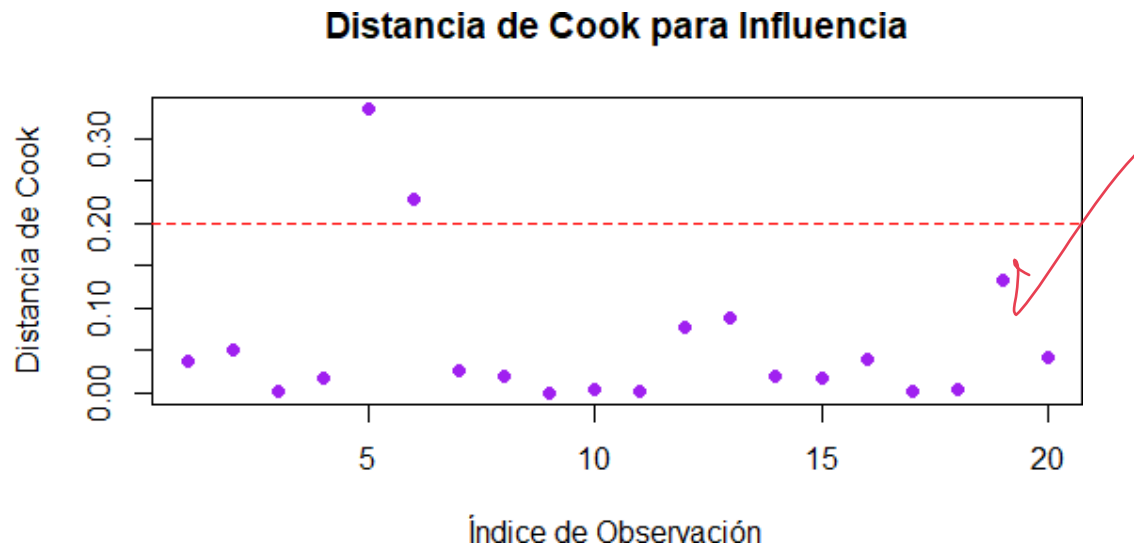
Para muestras pequeñas necesitamos ( $n < 50$ )

$$D_i > 4/(n - k - 1)$$
$$\text{Umbral sensible} \approx \frac{4}{20-1-1} = \frac{4}{18} = 0.222$$

Si se observa la línea de la gráfica está situada en 0.20 que es un valor muy cercano al umbral sensible.

Se puede observar que el punto 6 se puede considerar como influyente, ya que supera la línea de referencia que es 0.20 teniendo un valor aproximado de 0.32.

Otro punto podría ser el 5 que también se acerca al umbral.



10. Escriba una conclusión general para este problema.

En conclusión, podemos observar que la regresión lineal confirma la hipótesis estadística donde nos dice que la antigüedad del impulsor es el factor principal par determinar su resistencia al corte. Mientras el modelo de regresión  $Y = 2627.82 - 37.15X$ , revela una relación inversa  $p < 0.001$ . Por lo que la resistencia disminuye de manera constante en 37.15 psi por cada semana que transcurre. Tenemos que  $R^2 = 0.9221$ , explica más del 92% de la variabilidad en la resistencia. Se observó algo influyente el Punto 6 que requiere atención, aunque no invalida las conclusiones generales.

