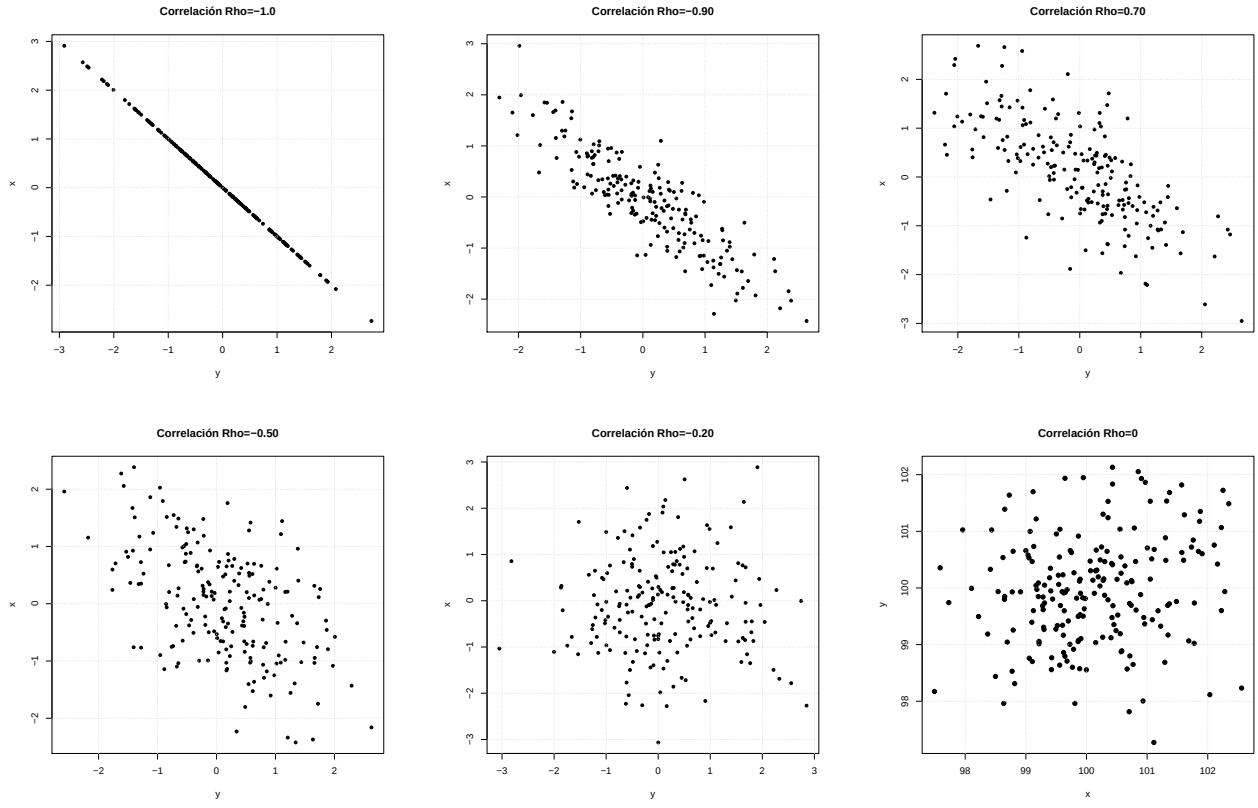
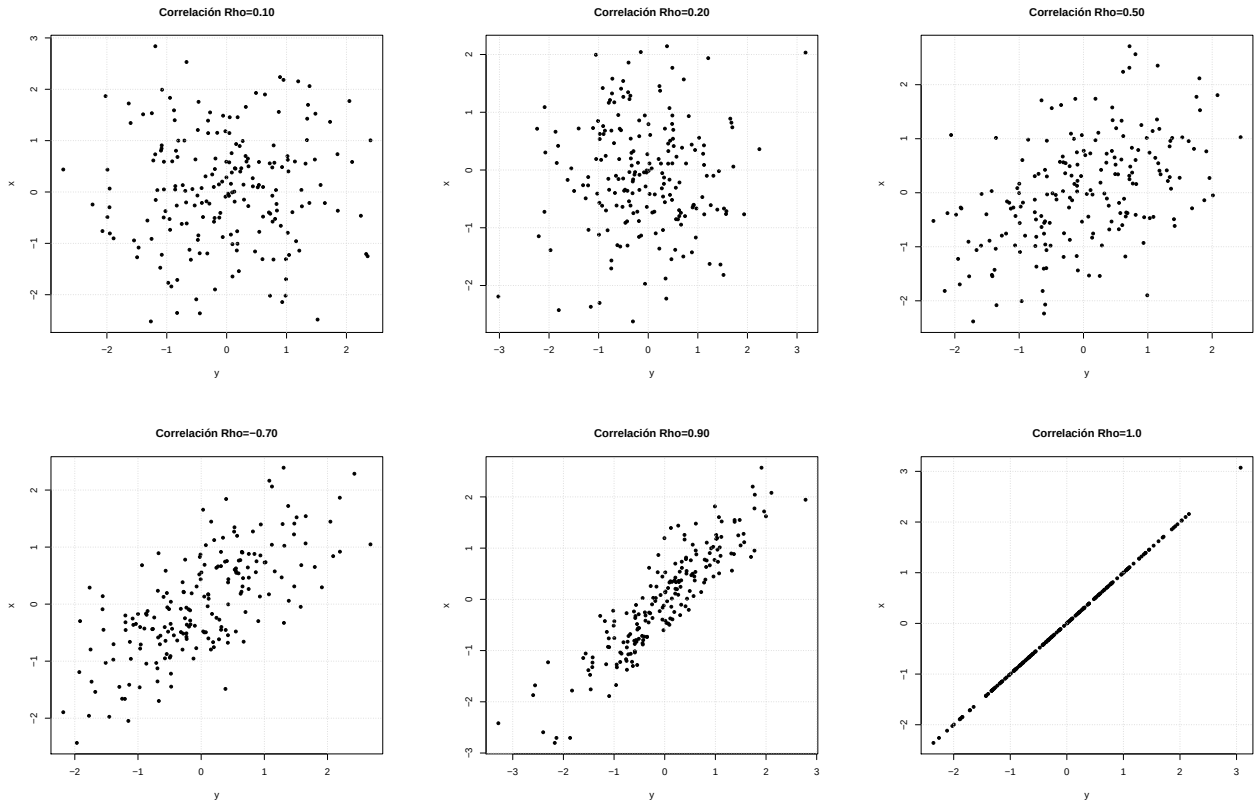
Modelo de Regresión lineal

En esta unidad estudiaremos la relación que puede existir entre dos variables o una variable y un conjunto de variables, mediante la construcción de un modelo lineal que represente dicha relación. Se partirá de un modelo general (modelo de Regresion Lineal Multiple) y se tratará como caso particular el modelo de dos variables también llamado Modelo de Regresión Lineal Simple MRLS.

En unidades anteriores se estudiaron relaciones entre dos variables aleatorias, las cuales se puede medir a través del coeficiente de correlación (ρ), el cual puede presentar valores desde -1 hasta 1 como se representa en las siguientes graficas :





La primera grafica (primera fila y primera columna) y la ultima grafica (cuarta fila y tercera columna) representan relaciones perfectas entre dos variables, que tan solo se presentan cuando la relación es constante $Y = a + bX$. A medida que el coeficiente de correlación aumenta la nube de puntos cambia, permitiendo ver una gama de posibilidades que pueden ser interpretadas dependiendo el signo como relaciones negativas o positivas (que se evidencian en la pendiente β_0) del modelo y por la magnitud desde relaciones muy fuertes a relaciones muy débiles (cerca de cero)

El modelo de regresión poblacional está determinado por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

Donde:

- Y,representa una la variable dependiente, tambien llamada variable respuesta o variable predicha.
- $X_1, X_2, \ldots, X_k$, corresponden a un conjunto de variables independientes o predictoras.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta k$ son un conjunto de constantes o parámetros del modelo (constantes) que deben ser estimadas a partir del conjunto de valores obtenidos en una muestra.

- u corresponde a una variable aleatoria no observable que representa todas las variables que no se incluyen en el modelo y corresponde a un componente aleatorio producto del azar de la naturaleza. Sin la presencia de esta variable en la estructura de la relación, el modelo formulado corresponderá a una relación matemática.

Un caso particular del modelo, ocurre cuando solo tiene una variable independiente, en este caso el modelo toma el nombre de Modelo de Regresión Lineal Simple con la siguiente estructura:

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

En este caso es posible realzar una representación gráfica del modelo mediante una linea recta donde β_0 corresponde al intercepto y β_1 a la pendiente.
A partir de la obtención de una muestra de tamaño n se puede obtener el modelo estimado, cuya ecuación esta dada por:

$$\widehat{E[y|X]} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

Donde : $[b_0, b_1, \ldots, b_k]$ representa el vector de estimadores de los coeficientes del modelo.

Como cada vez que se toma una muestra, por razones del azar los objetos medidos son diferentes, es claro que al realizar un procedimiento de estimación los valores de estos son diferentes, ratificando esto que $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ son variables aleatorias

Modelo de Regresión lineal - enfoque matricial

Para representar el modelo de manera matricial, se puede partir de la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \quad i = 1, 2, \ldots, n$$

Los valores $\beta_0, \beta_1, \ldots, \beta_k$ deben cumplir las siguientes igualdades:

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + u_1 \\
Y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + u_2 \\
Y_3 &= \beta_0 + \beta_1 x_{13} + \beta_2 x_{23} + \cdots + \beta_k x_{k3} + u_3 \\
\vdots &\quad \vdots \quad \vdots \quad \vdots \\
Y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + u_n
\end{aligned}
$$

En forma matricial las podemos escribir

$$y = X\beta + u$$

Donde:

$$y = [y_1, y_2, \ldots, y_n]^T$$

$$\beta = [\beta_0, \beta_1, \ldots, \beta_k]^T$$

$$
X = \begin{bmatrix}
1 & x_{11} & x_{21} & \ldots & x_{k1} \\
1 & x_{12} & x_{22} & \ldots & x_{k2} \\
1 & x_{13} & x_{23} & \ldots & x_{k3} \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
1 & x_{1n} & x_{2n} & \ldots & x_{kn}
\end{bmatrix}
$$

$u = [u_1, u_2, \ldots, u_n]^T$

Estimación de los coeficientes del modelo

El método Minimos Cuadrados Ordinarios (MCO) se basa en la minimización de la suma de los residuales ($\widehat{u}$), los cuales se pueden despejar de la ecuación del modelo $\widehat{y} = X\widehat{\beta} + \widehat{u}$. Como los errores son variables no observables, el método hace uso de sus respectivos estimadores llamados residuales:

$$\widehat{u} = \widehat{y} - X\widehat{\beta}$$

Ahora, la suma de cuadrado de los resiuales se puede obtener al multiplicar el vector $\widehat{u}^T\widehat{u}$

$$SCRes = \sum_{i=1}^{n} \widehat{u}_i^2 = \widehat{u}^T\widehat{u} = \begin{bmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \vdots \\ \widehat{u}_n \end{bmatrix} [\widehat{u}_1, \widehat{u}_2, \ldots, \widehat{u}_n]$$

$$\begin{aligned} SCRes &= [\widehat{y} - Xb]^T[\widehat{y} - Xb] \\ &= y^Ty - y^TXb - b^TX^Ty + b^TX^TXb \\ &= y^Ty - 2bX^Ty + b^TX^TXb \end{aligned}$$

Para encontrar los valores óptimos de los coeficientes, se debe derivar parcialmente con respecto al vector β e igualarlo a cero

$$\frac{\partial SCRes}{\partial \beta} = -2X^T + 2X^TXb = 0$$

De la ecuación anterior se despeja b ($\widehat{\beta}$)

$$\begin{aligned} 2X^T &= 2X^TXb \\ (X^TX)^{-1}X^Ty &= (X^TX)^{-1}X^TXb \\ (X^TX)^{-1}X^Ty &= b \end{aligned}$$

El estimador MCO de los coeficientes b es:

$$\widehat{\beta}_{MCO} = b = (X^TX)^{-1}X^Ty$$

Es claro que para poder realizar este proceso es condición necesaria que la matriz (X^TX) sea invertible. En caso contrario los estimadores MCO no se pueden hallar. La versión del método MCO a partir de sumatorias se encuentra desarrollada en el texto guía (Walpole-2012 pp.395)

Ejemplo 1: Un fabricante de equipos de aire acondicionado tiene problemas en la etapa de montaje, debido principalmente a la falta de una biela, pues debido a su peso no satisface las especificaciones establecidas para el producto por sobrepeso. Para reducir este costo,

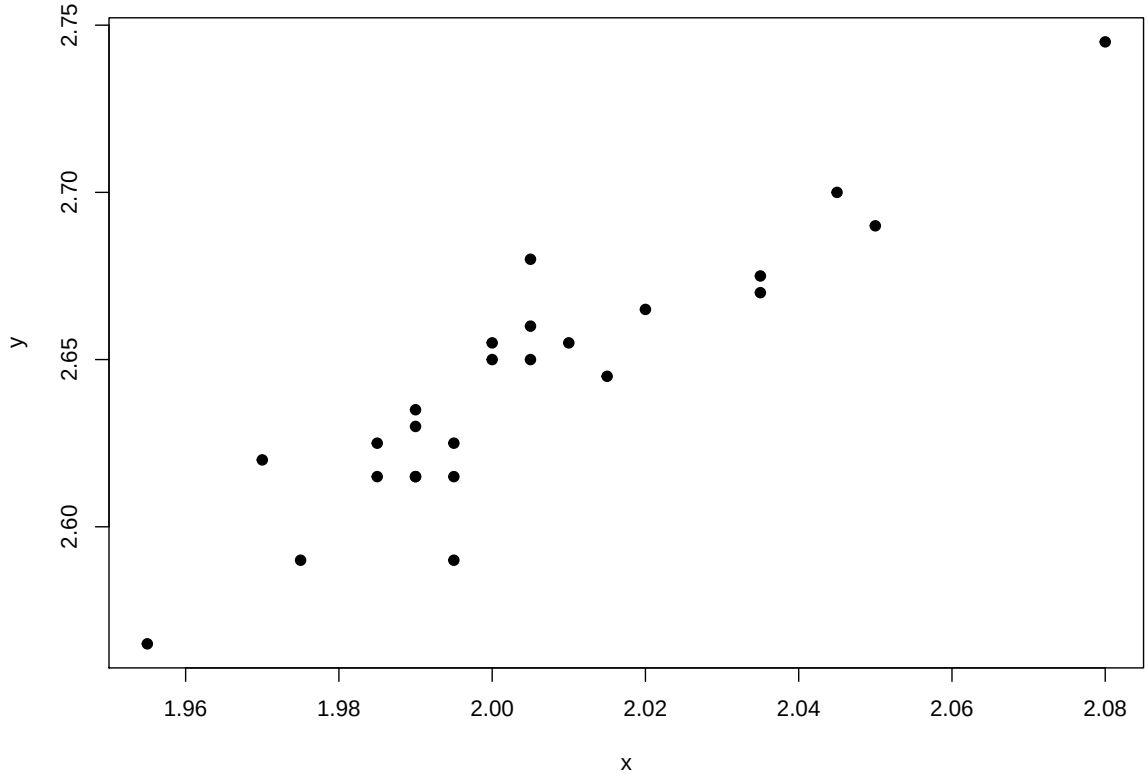
el fabricante estima que estudiando la relación entre el peso de la barra en bruto y su peso final, es probable encontrar una relación entre ellas y de esta manera reducir el problema mediante la detección de elementos que probablemente no cumplan con las condiciones exigidas para el producto.

En el estudio se midieron un total de 25 (x, y) pares de barras, siendo X el peso de la pieza colada en espera de ser procesada (materia prima) y Y el peso de una biela terminada que forma parte del producto final. Se requiere estimar un modelo que permita predecir el valores del peso futuro de la biela terminada como función del peso bruto del bloque de metal.

Los datos recogidos se muestran a continuación:

Antes de realizar la estimación debemos visualizar mediante un diagrama de dispersión la relación lineal entre las variables.

barra Numero	peso bruto	peso final	barra Numero	peso bruto	peso final
1	2.745	2.080	14	2.635	1.990
2	2.700	2.045	15	2.630	1.990
3	2.690	2.050	16	2.625	1.995
4	2.680	2.005	17	2.625	1.985
5	2.675	2.035	18	2.620	1.970
6	2.670	2.035	19	2.615	1.985
7	2.665	2.020	20	2.615	1.990
8	2.660	2.005	21	2.615	1.995
9	2.655	2.010	22	2.615	1.990
10	2.655	2.000	23	2.590	1.975
11	2.650	2.000	24	2.590	1.995
12	2.650	2.005	25	2.565	1.955
13	2.645	2.015			



En la grafica se puede observar que existe una posible relación positiva entre las variables, la cual se puede representar a través de una linea recta con pendiente positiva.
Inicialmente se plantea el modelo formado por una variable respuesta (y) y una variable independiente (x), una variable aleatoria no observable o error (u) y dos coeficientes β_0 y β_1 y un conjunto de 25 observaciones.

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad \text{con} \quad i = 1, 2, 3, \ldots 25$$

Este modelo representa 25 igualdades que originan un sistema matricial $y = X\beta + u$, el cual se resuelve mendiante MCO con el siguiente resultado:

$$
\begin{aligned}
b &= (X^T X)^{-1} X^T y \\
&= \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{bmatrix} \\
&= \begin{bmatrix} 25{,}00 & 50{,}1200 \\ 50{,}12 & 100{,}4986 \end{bmatrix}^{-1} \begin{bmatrix} 66{,}0800 \\ 132{,}5007 \end{bmatrix} \\
&= \begin{bmatrix} 222{,}4160 & -110{,}9218 \\ -110{,}9218 & 55{,}3281 \end{bmatrix} \begin{bmatrix} 66{,}0800 \\ 132{,}5007 \end{bmatrix} \\
\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} &= \begin{bmatrix} 0{,}03753679 \\ 1{,}29971229 \end{bmatrix}
\end{aligned}
$$

Mediante operaciones matriciales podemos obtener los estimadores de los coeficientes así:

```
y=c(2.745, 2.700, 2.690, 2.680, 2.675, 2.670, 2.665, 2.660, 2.655, 2.655,
    2.650, 2.650, 2.645, 2.635, 2.630, 2.625, 2.625, 2.620, 2.615, 2.615,
    2.615, 2.615, 2.590, 2.590, 2.565)

x=c(2.080, 2.045, 2.050, 2.005, 2.035, 2.035, 2.020, 2.005, 2.010, 2.000,
    2.000, 2.005, 2.015, 1.990, 1.990, 1.995, 1.985, 1.970, 1.985, 1.990,
    1.995, 1.990, 1.975, 1.995, 1.955)

unos=rep(1,25)
X=matrix(c(unos,x),nrow=25)

# Estimador MCO - Enfoque matricial
b=solve(t(X)%*%X)%*%(t(X)%*%y)

> b
[,1]
[1,] 0.03753679
[2,] 1.29971229
```

Finalmente el modelo estimado por MCO corresponde a la ecuación:

$$\widehat{y} = 0{,}03753679 + 1{,}29971229 \ x$$

El signo esperado para el coeficiente b_1, está de acuerdo con los resultados encontrados. Como se ha mencionado anteriormente b_1 corresponde a la estimación de la pendiente de la recta de regresión muestral y su signo positivo indica que para un mayor peso de la barra antes de iniciar el proceso, el producto final tendrá un mayor peso.
Para realizar la estimacion de los coeficientes en R se debe correr el siguiente codigo:

```
regresion=lm(y~x)
```

Esta instrucción produce el siguiente resultado

```
> # Estimacin por MCO en R
> summary(regresion)
Call:
lm(formula = y ~ x)
Residuals:
Min        1Q        Median  3Q       Max
-0.040463 -0.011457 0.002044 0.007534 0.036540
Coefficients:
Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 0.03754   0.23810    0.158    0.876
x           1.29971   0.11875   10.945    1.36e-10 ***
---
Signif.codes:0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Residual standard error: 0.01597 on 23 degrees of freedom
Multiple R-squared: 0.8389,  Adjusted R-squared: 0.8319
F-statistic: 119.8 on 1 and 23 DF, p-value: 1.356e-10
```

En el listado para cada estimacion se presenta una prueba-t individual sobre la significancia de cada coeficiente asi:
$Ho : \beta_0 = 0$
$Ha : \beta_0 \neq 0$
$EdeP : T = 0{,}158$
valor-p:$0{,}876$

Indicando que no se rechaza la hipótesis nula y por tanto se asume que Ho es verdad , $\beta_0 = 0$. En este caso se recomienda que aunque la prueba de hipotesis indique que el intersecto es no significativo, se debe dejar en el modelo, pues el no incluirlo puede generar que la o las pendientes estimadas presenten un sesgo.

Para el caso de la pendiente los resultados son los siguientes.
$Ho : \beta_1 = 0$
$Ha : \beta_1 \neq 0$
$EdeP : T = 10{,}945$
valor-p:$0{,}0000$

Lo cual ratifica que el modelo es valido y existe una relación lineal entre la variable de la barra inicial y el peso del producto terminado

Además de los coeficientes estimados y de la validación de sus respectivas pruebas de hipótesis individuales, existe una medida para determinar el grado de explicación de la variable dependiente que es estimada por el modelo ($R^2 = 83{,}9\,\%$). En este caso el modelo explica el 83.9 % del comportamiento de y o de su variación.

Finalmente la salida reporta el valor del estadístico F (119.8) y su valor-p asociado (0.0000), correspondiente al análisis de varianza o ANOVA , útil en análisis de regresión lineal múltiple. Para el caso del modelo de regresión lineal simple el análisis de varianza arroja los mismos resultados que la hipótesis individual para β_1

Validación de supuestos

El método de MCO, no requiere ningún supuesto de tipo estadístico, pues es un método que se basa en la optimización matemática. Sin embargo cuando se requiere realizar inferencia - construcción de intervalos de confianza o la realización de pruebas de hipótesis asociadas con los resultados obtenidos - es necesario plantear y verificar el cumplimiento de supuestos estadístico sobre las variables que intervienen en el modelo. Algunos de ellos se realizan sobre las variables independientes y otros sobre la variable no observable error

(u). Haremos énfasis sobre el cumplimiento de los siguientes supuestos, pues su violación genera estimaciones ineficientes y sesgados.

S1. $E[u] = 0$. El modelo esta completo. Este supuesto garantiza que al hacer estimaciones el componente aleatorio desaparezca. Se valida mediante el calculo de la media de los residuales.

S2. $V[u] = \sigma^2$. La varianza de los errores es constante. El supuesto de HOMOSCEDASTICIDAD o varianza constante hace que los estimadores MCO sean eficientes. Este supuesto se pude validar mediante la realización de otra regresión donde se toma como regresora una proxis de la varianza (u^2) y como variables independientes las mismas del modelo. Esta prueba se conoce como Test de White. En caso de que los coeficientes asociados con las variables independientes sean no significativos, este resultado será indicio de que la varianza es constante.

$$\widehat{u}^2 = \alpha_0 + \alpha_1 x + \epsilon$$

S3. $Cor[u_i, u_{i-1}] = 0$. Los errores no están auto-correlacionados. El supuesto de no autocorrelación de errores se puede verificar mediante el estadístico Durbin-Watson. Para determinar si se rechaza o no el cumplimiento de este supuesto se debe construir una región de rechaz a partir de los datos obtenidos en la table de D-W

S4. $u \sim N(0, \sigma^2)$. Para verificar este supuesto se puede construir una grafica de los residuales $qqnorm$ y una prueba de hipotesis de normalidad

Es necesario antes de utilizar el modelo, verificar el cumplimiento de los anteriores supuestos.

Analisis de los resultados

Listado 1. Supuesto de modelo completo

```
> summary(residuos)
Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
-2.593880 -0.734629 0.131571 -0.002484 0.487185 2.335921

> t.test(residuos)
One Sample t-test
data: residuos
```

```
t = -0.012286, df = 24, p-value = 0.9903
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-0istado\\.4196776 0.4147105
sample estimates:
mean of x
-0.00248355
```

Por un lado el promedio de los residuales es cercano a cero y por otro lado la prueba de hipótesis Ho: $\mu_u = 0$ vs Ha: $\mu_u \neq 0$, no se rechaza, asumimos que el valor esperado de los errores es cero.

Listado 2 Supuesto de no autocorrelación de errores

```
# Es necesario cargar el paquete
library(lmtest)
dwtest(y~x)
> Durbin-Watson test
> data: y ~ x
> DW = 1.518, p-value = 0.07668
> alternative hypothesis: true autocorrelation is greater than 0
```

Los resultados indican que no se cumple el supuesto de no autocorrelacion, los errores estan correlacionados

Listado 3 Supuesto de normalidad de los errores

```
shapiro.test(residuos)
> Shapiro-Wilk normality test
> data: residuos
> W = 0.9658, p-value = 0.5415
```

El valorp correspondiente a la prueba de Shapiro-Wilk indica que los errores se pueden asumir como normales

Listado 4

```
# Supuesto de varianza constante de errores o homoscedasticidad
```

```
summary(lm(residuos^2 ~ x))

Residuals:
Min      1Q  Median      3Q     Max
-1.1515 -0.8039 -0.4418 -0.0125 5.6504

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.855    24.420  0.854   0.402
x            -9.913    12.180 -0.814   0.424

Residual standard error: 1.637 on 23 degrees of freedom
Multiple R-squared: 0.028, Adjusted R-squared: -0.01426
F-statistic: 0.6625 on 1 and 23 DF, p-value: 0.424
```

Para la verificación del supuesto de homoscedasticidad indica que la variable independiente no esta relacionada con la varianza. Se puede suponer la varianza de los errores es constante.

Codigo completo en R

```
y=c(2.745, 2.700, 2.690, 2.680, 2.675, 2.670, 2.665, 2.660, 2.655,
    2.655, 2.650, 2.650, 2.645, 2.635, 2.630, 2.625, 2.625, 2.620,
    2.615, 2.615, 2.615, 2.615, 2.590, 2.590, 2.565)
x=c(2.080, 2.045, 2.050, 2.005, 2.035, 2.035, 2.020, 2.005, 2.010,
    2.000, 2.000, 2.005, 2.015, 1.990, 1.990, 1.995, 1.985, 1.970,
    1.985, 1.990, 1.995, 1.990, 1.975, 1.995, 1.955)
# Descriptivas
summary(Y); summary(X)
# Diagrama de dispersin X vs Y
plot(y,x, main="Diagrama de dispersin",xlab="X:Peso inicial", ylab="Y:
    Peso final", pch=19)
# Estimacin por MCO
regresion=lm(y~x)
summary(regresion)
confint(regresion)
plot(y~x, xlab = "Peso inicial", ylab = "Peso final", pch=19)
abline(regresion)
# Analisis de la varianza79po'
anova(regresion)
# Residuos
residuos <- rstandard(regresion)
valores.ajustados <- fitted(regresion)
plot(valores.ajustados, residuos)
# Diagnostico de normalidad
qqnorm(residuos, main="Normalidad de los Residuales")
qqline(residuos)
plot(y,x, xlab = "peso-F", ylab = "peso-b")
abline(y~x)
# validacin de supuestos del modelo
residuos=rstandard(regresion)
valores.ajustados <- fitted(regresion)
plot(valores.ajustados, residuos)
dwtest(y~x)
bgtest(y~x)
```

```
resettest(y~x , power=2, type="regressor")
shapiro.test(residuos)
```