Modelos no paramétricos y de Regresión

Sofía Villers Gómez Dulce María Reyes Varela

Índice general

Prefacio

Al llegar al curso de Modelos No Paramétricos y de Regresión ya hemos cursado Inferencia estadística en la que estudiamos una serie de métodos de estimación puntual (método de momentos, estimadores de máxima verosimilitud), además aprendimos a evaluar dichos estimadores para encontrar los mejores. Sin embargo, en éste enfoque estadístico se tiene la desventaja de que siempre se trabaja con muestras aleatorias basadas en el supuesto de que siguen cierta distribución conocida, más adelante los conocerán en los ejercicios prácticos.

Algunos de los problemas que tienen las pruebas de hipótesis es que suponen que las observaciones disponibles para el estadístico provienen de distribuciones cuya forma exacta es conocida, aún cuando los valores de algunos parámetros sean desconocidos. En otras palabras, se supone que las observaciones provienen de una cierta familia paramétrica de distribuciones y que se debe hacer inferencia estadística acerca de los valores de los parámetros de dicha familia, comunmente la media, la varianza y en otros casos la proporción.

Objetivos

- Proporcionar a los alumnos, herramientas suficientes para el curso Modelos no paramétricos y de Regresión.
- Reforzar las bases teóricas con contenido electrónico completado con herramientas de R-Studio.
- Dar continuidad al material para el curso *Modelos no paramétricos y de Regresión.*

Este libro fue escrito con bookdown usando RStudio.

Esta versión fue escrita con:

Licencia

- **No additional restrictions**—You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Introducción

Parte I

Un primer vistazo

Capítulo 1

Escalas de Medición

Este capítulo es una traducción de Conover (1998), a continuación se definirán los tipos de escalas que se le asigna a una variable. Hay cuatro escalas básicas las cuales son: **nominal, ordinal, de intervalo y de razón.** A las variables que se consideran con escala ordinal y nominal también se les conoce como **variables categóricas**, mientras que a las variables que se consideran con escala de intervalo y razón se conocen como **variables cuantitativas**.

1.1. Variables categóricas

1.1.1. Escala nominal:

La escala nominal utiliza números simplemente como un medio para separar las propiedades o elementos en diferentes clases o categorías. El número asignado a la observación sirve sólo como un **"nombre"** para la categoría a la que pertenece la observación, de ahí el título **"nominal"**. Por ejemplo, utilizamos la escala nominal de medición cuando definimos una variable aleatoria que equivalía a 1 si una moneda caía **"sol"**, y 0 si la moneda caía **"águila"**. Podríamos, igual bien, haber usado los números **7.3** y **3.9** para representar el sol y el águila, respectivamente. Nuestra elección de 1 y 0 fue principalmente por conveniencia. Otro ejemplo, cuando 12 sujetos se numeran arbitrariamente del 1 al 12, se utiliza una escala de medición nominal y la asignación de los números es una forma de variable aleatoria. Al clasificar objetos según el color, las categorías pueden etiquetarse como 1, 2, 3 o azul, amarillo, rojo o A, B, C. Los números son simplemente nombres de categoría. Los números pueden ser reemplazados por otros números no utilizados, siempre que las categorías permanezcan intactas.

1.1.2. Escala ordinal:

La escala ordinal se refiere a mediciones en las que sólo son relevantes las comparaciones **"mayor"**,**"menor"** o **"igual"** entre ellas. El valor numérico se usa sólo como un medio para organizar los elementos que se miden en orden, de menor a mayor. Ésta capacidad de ordenar los elementos, en función del tamaño relativo de sus medidas, le da el nombre de si algunos de los elementos son iguales entre sí, decimos que existen vínculos. Cuando se le pide a una persona que asigne el número 1 a la más preferida de las tres marcas, el número 3 a la menos preferida y el número 2 a la marca restante, está usando una escala ordinal y está usando los números 1, 2, 3; ésta podría haber usado tres números, digamos 16, 20, 75, siempre y cuando los números se asignen a las marcas de tal manera que el orden relativo del número represente la preferencia relativa de la marca.

1.2. Variables cuantitativas

1.2.1. Escala de intervalo:

La escala de intervalo considera como información pertinente **no sólo** el orden relativo de las mediciones como en la escala ordinal sino también el **tamaño del intervalo** entre mediciones, es decir, el tamaño de la diferencia (en un sentido de resta) entre dos mediciones. La escala de intervalo implica el concepto de una unidad de distancia, y la distancia entre dos mediciones cualquiera puede expresarse como un cierto número de unidades. Un buen ejemplo es la escala por la cual generalmente representamos la temperatura. El aumento de una unidad (grado) de temperatura se define por un cambio particular en el volumen de mercurio en un termómetro; en consecuencia, la diferencia entre dos temperaturas cualquiera puede medirse en **unidades o grados**. El valor numérico real de la temperatura es simplemente una comparación con un punto arbitrario llamado "cero grados". La escala de intervalo requiere un punto cero y una unidad de distancia (no es posible tener este último sin el primero), pero no es importante cómo se definan los "ceros" y la unidad de distancia.

La temperatura ha sido medida de manera bastante adecuada durante algún tiempo por las escalas Fahrenheit y Celsius, que tienen diferentes temperaturas cero y diferentes definiciones de 1 grado o unidad. El principio de las mediciones de intervalo no se viola por un cambio en la escala o la ubicación o ambos.

1.2.2. Escala de razón:

La escala de razón se usa cuando **no sólo el orden y el tamaño del intervalo son importantes**, sino que también la **razón entre dos medidas** es significativa. Si es razonable hablar de que una cantidad es "dos veces" otra cantidad, la escala de proporción es apropiada para la medición, como cuando se miden los rendimientos de cultivos, distancias, pesos, alturas, ingresos, etc. En realidad, la única distinción entre la escala de razón y la escala de intervalo es que la escala de razón tiene una **medida natural** "cero", mientras que la medida de cero se define **arbitrariamente** en la escala de intervalo. Como en la escala de intervalo, la unidad de distancia de la escala de razón se define arbitrariamente.

No existe un acuerdo universal entre los científicos que prefieren usar escalas adicionales, y algunas mediciones no se incluyen claramente en una de las cuatro escalas recién definidas. Por lo tanto, ésta clasificación de escalas de medida puede ser demasiado simplista, pero es suficiente para los propósitos de este curso.

La mayoría de los métodos estadísticos paramétricos habituales requieren una escala de medición de intervalo (o más fuerte). La mayoría de los métodos no paramétricos suponen que la escala nominal o la escala ordinal son apropiadas.

Parte II

Pruebas Binomiales

Introducción

La distribución de probabilidad binomial se introdujo para describir las probabilidades asociadas con el número de caras cuando se lanza una moneda n veces. En su forma más general, cada uno de n ensayos independientes da como resultado "$éxito$", con probabilidad p, o "$fracaso$", con probabilidad $q = 1 - p$. La distribución binomial describe la probabilidad de obtener exactamente "k" éxitos.

Decimos que una variable aleatoria X se distribuye Binomial con parámetros n,p.

$$X \sim Bin(n, p)$$
$$x = 0, 1, 2, ...n$$

Función de Distribución

$$P_X(x) = \binom{n}{p} p^x (1-p)^{n-x}$$

Media

$$\mathbf{E}[X] = np$$

Varianza

$$Var(X) = np(1-p)$$

Capítulo 2

Prueba de Proporciones

Supongamos que un científico desea saber si la tasa de mortalidad va en aumento, si la tasa de pobreza está cambiando o si la tasa de algún grupo cívico está a favor de una política en particular. En muchos casos existe una proporción hipotética p de una población en estudio y una poporción específica p^* y se pretende llevar a cabo una comparación para saber si la proporción hipotética es igual, menor o mayor que la proporción específica p^*. Una prueba de proporciones puede ser útil para ayudar a responder este tipo de preguntas.

2.1. Datos

Los datos consisten en una muestra

$$X_1, X_2, X_3, \cdots, X_n \text{ una muestra aleatoria de tamaño n de la población.}$$

cuyo resultado se puede clasificar en dos y sólo dos categorías "categoría 1" o "categoría 2". El número de observaciones en la categoría 1 es O_1 y el número de observaciones en la categoría 2 es $O_2 = n - O_1$

2.2. Supuestos

1) Los n ensayos son mutuamente independientes.
2) La probabilidad p de que el resultado de cada ensayo caiga en la categoría 1 es la misma en cada uno de los ensayos.

2.3. Estadístico de Prueba

Como nos interesa la probabilidad del resultado "clase 1", dejaremos que el estadístico de prueba T sea el número de veces que el resultado es **"clase 1"**. Es decir, si $n < 20$ utilizar el estadístico

$$T = O_1,$$
$$T \sim Bin(n, p^*),$$

Donde p^* es la probabilidad especificada en la hipótesis nula de nuestra prueba a realizar y n es el tamaño de la muestra.

Por otro lado, si $n \geq 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el siguiente cuantil:

$$t = np + Z_q\sqrt{np(1-p)}$$

Donde Z_q es el cuantil de una distribución normal estándar que se puede obtener en la tabla correspondiente[1].

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis.

2.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_0 : p = p^*,$$
$$vs$$
$$\mathbf{H}_a : p \neq p^*,$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T \leq t_1 \quad o \quad T > t_2$$

Elegimos α_1 y α_2, los tamaños de la cola inferior y superior, respectivamente. El tamaño de la prueba es $\alpha = \alpha_1 + \alpha_2$.

Debemos encontrar t_1 y t_2 tales que:

$$\mathbf{P}[Y \leq t_1] = \alpha_1 \quad \mathbf{P}[Y > t_2] = \alpha_2$$

Donde $Y \sim Bin(n, p^*)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * min\{\mathbf{P}[Y \leq T], \mathbf{P}[Y \geq T]\}$$

Sugerimos que si $n > 20$,el $p - value$ puede obtenerse usando:

$$\mathbf{P}[Y \leq t_{obs}] \approx \mathcal{N}\left(\frac{t_{obs} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

y

$$\mathbf{P}[Y \geq t_{obs}] \approx 1 - \mathcal{N}\left(\frac{t_{obs} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Con $\mathcal{N}(\cdot)$ la función de distribución de la normal estándar.

[1] Véase que de acuerdo al tipo de cuantil (de cola inferior o superior), el signo del cuantil Z_q cambiará.

Caso B (Prueba de cola inferior)

$$\mathbf{H}_0 : p \geq p^*,$$

$$vs$$

$$\mathbf{H}_a : p < p^*,$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T \leq t$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = \alpha$$

Donde $Y \sim Bin(n, p^*)$.
y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T]$$

Sugerimos que si $n > 20$, el $p - value$ puede obtenerse usando:

$$\mathbf{P}[Y \leq t_{obs}] \approx \mathcal{N}\left(\frac{t_{obs} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : p \leq p^*,$$

$$vs$$

$$\mathbf{H}_a : p > p^*,$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T > t$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = 1 - \alpha$$

Donde $Y \sim Bin(n, p^*)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \geq T]$$

Sugerimos que si $n > 20$, el $p - value$ puede obtenerse usando:

$$\mathbf{P}[Y \geq t_{obs}] \approx \mathcal{N}\left(\frac{t_{obs} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

2.5. Intervalos de Confianza

En este enfoque lo que se busca es estimar cotas inferiores y superiores (L, U) tales que el intervalo formado por las mismas contenga al parámetro de interés (en este caso una proporción) con una confianza $1 - \alpha$ especificada por el usuario.

Datos

Los datos a los que se les puede aplicar éste método son del mismo que en las pruebas de hipótesis, es decir, una colección $\{X_i\}_{i=1}^n$ de ensayos Bernoulli independientes en dónde se debe poder suponer que la probabilidad de éxito p se mantiene constante en todos los ensayos.

Supuestos

1) Los n ensayos son independientes.

2) La probabilidad de éxito permanece constante durante todos los ensayos.

La idea detrás del método es encontrar todos los valores de p^* tales que rechacemos la hipótesis nula.

Hipótesis

$$\mathbf{H}_0 : p = p^*,$$

$$vs$$

$$\mathbf{H}_a : p \neq p^*,$$

Procedimiento

1) Fijar el nivel de confianza $1 - \alpha$

2) Hallar p_1 y p_2 tales que:

$$\mathbf{P}[Y \leq O_1 |\ p = p_1] = \alpha_1 \quad y \quad \mathbf{P}[Y \geq O_1 |\ p = p_2] = \alpha_2$$

donde $\alpha = \alpha_1 + \alpha_2$ y $Y \sim Bin(n, p)$.

3) Hacer $L = p_1$ y $U = p_2$

Recordemos que mucha de la teoría abarcada en este material esta basado en el libro Conover (1998) y en dicha referencia se pueden encontrar tablas con los valores de los intervalos de confianza para muestras menores a 30, cuando se tienen más grandes, es decir si $n > 30$, se pueden utilizar las siguientes expresiones para el cálculo de los intervalos:

$$L = \frac{O_1}{n} - Z_{1-\alpha/2}\sqrt{\frac{O_1(n-O_1)}{n^3}} \qquad U = \frac{O_1}{n} + Z_{1-\alpha/2}\sqrt{\frac{O_1(n-O_1)}{n^3}}$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo.

2.6. Ejemplo

Se tienen 20 graduados del Tecnológico de Texas que presentaron el examen general de leyes y 18 de ellos lo pasaron. Si esta muestra es aleatoria y representativa de todos los estudiantes graduados del Tecnológico de Texas, ¿esto prueba que la probabilidad de que un graduado de esa escuela pase el examen general de leyes es más alto que el promedio del estado, que es del 70 %?.

Paso 1 Escribimos la prueba a utilizar,

La prueba a utilizar **Prueba de proporciones caso C cola superior,**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado,

$$\mathbf{H}_0 : \text{La probabilidad de que un graduado pase el examen es menor/igual al } 70\,\%.$$
$$vs$$
$$\mathbf{H}_a : \text{La probabilidad de que un graduado pase el examen es mayor al } 70\,\%.$$

De manera alternativa:

$$\mathbf{H}_0 : p \leq p^* \text{ es decir, } p \leq .70 \quad vs \quad \mathbf{H}_a : p > p^* \text{ es decir, } \quad p > .70$$

Paso 3 Estadístico de prueba,

Utilizaremos el estadístico

$$T = O_1 \text{ número de observaciones de la clase 1.}$$
$$T = 18 \;\; \textit{número de graduados que pasaron el examen}$$
$$T \sim Bin(20, 0.70)$$

Paso 4 Procedimiento completo,

Supuestos:

1. Muestra aleatoria de tamaño 20.
2. Tomaremos como "éxito" al acreditar el examen.
   - $T = O_1 = 18$ número de éxitos.
3. Tomaremos α= 5 %=0.05 el nivel de significancia.
   - $n = 20$ tamaño de la muestra.
   - p^*=70 %=0.70.

Paso 5 Regla de decisión

$$Rechazo \;\; H_0 \;\; si \;\; T > t_2 \;\; y \;\; Rechazo \;\; H_0 \;\; si \;\; p - value < \alpha,$$

ya que $n \geq 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se podemos utilizar:

$$t_2 = np^* + Z_q\sqrt{np^*(1-p^*)}.$$

$$t_2 = 20 * (0.7) + 1.65 * \sqrt{20 * (0.7) * (0.3)} \quad = 17.38.$$

$\therefore$ como $T = 18 > 17 = t_2$ entonces rechazo H_0.

y por otro lado calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \geq T] \ = \ \mathbf{P}[Y \geq 18] = 1 - \mathbf{P}[Y < 18] = 0.035.$$

$\therefore$ como $p - value = 0.035 < \alpha = 0.05$ entonces rechazo H_0.

Paso 6 Conclusión

Existe información suficiente para decir que los alumnos que acreditaron el examen están por arriba del promedio del Estado que es del 70 %, en otras palabras la probabilidad de que un graduado del Tecnológico de Texas pase el examen general de leyes es mayor al 70 %.

2.7. Ejemplo en R-Studio

Ahora haremos la réplica en R.

La estadística de prueba será $T = O_1$

```
# Datos

T=18                                    #Número de éxitos
alpha=0.05                              #Nivel de significancia
n=20                                    #Tamaño de la muestra
p=0.70                                  #Proporción

t=qbinom(.95,n,p)                        #Valor crítico
t
```

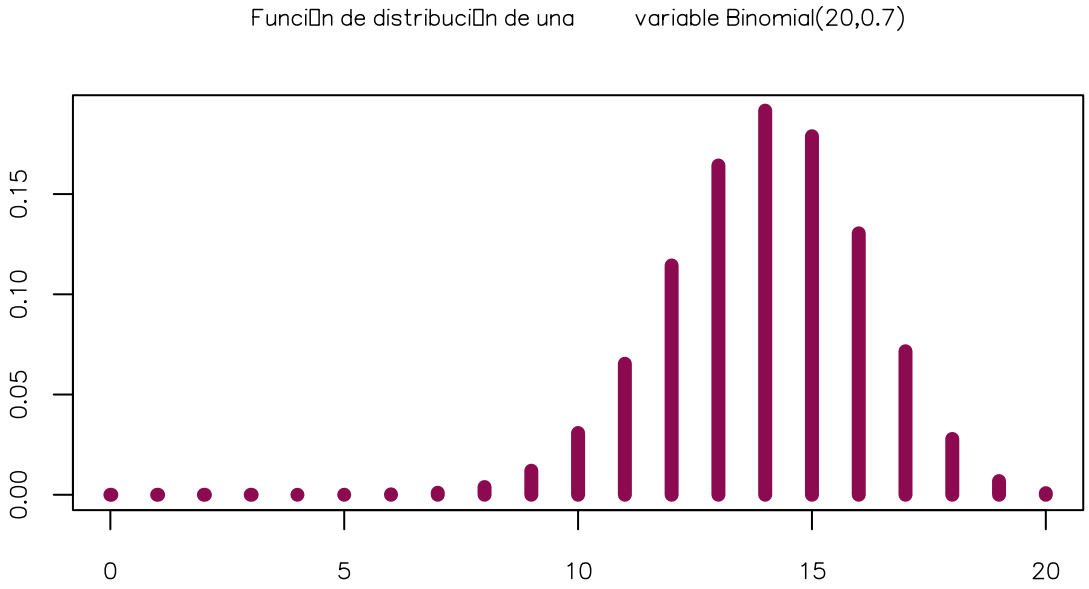
```
[1] 17
```

```
pvalue= 1-pbinom(17,20,0.7)                #P-value
pvalue
```

```
[1] 0.03548313
```

Según el planteamiento de las hipótesis, este es un Caso C (de cola superior), por lo que siguiendo la regla de decisión tenemos que como $T = 18 > 17 = t$, entonces se rechaza H_0 y por lo tanto se concluye que hay información suficiente para decir que la probabilidad de que un graduado del tecnológico de Texas pase el examen es mayor al 70 %.

Graficamos la función binomial



Utilizando la función de R, debemos tener cuidado ya que nuestra muestra es pequeña:

```
binom.test(T,n, p = 0.7, alternative = c("greater"))
```

```
    Exact binomial test

data:  T and n
number of successes = 18, number of trials = 20, p-value = 0.03548
alternative hypothesis: true probability of success is greater than 0.7
95 percent confidence interval:
 0.7173815 1.0000000
sample estimates:
probability of success
                   0.9
```

2.8. Ejercicios

1. Un fabricante de teléfonos móviles afirma que sólo el 5 % de todas las unidades que vende sufre una falla durante el primer mes de operación normal. Una organización de consumidores ha pedido a 45 consumidores que han adquirido estos teléfonos móviles, que reporten cualquier mal funcionamiento durante el primer mes. Al final de éste sólo siete consumidores reportaron mal funcionamiento. Si la organización de consumidores cree que la proporción de teléfonos que sufrirán alguna falla es mayor al valor afirmado por el fabricante. ¿Con una α del 10 % podría la organización sustentar su creencia?
2. Un candidato a través de encuestas propias afirma que el 65 % o más de los votantes están a su favor. Sin embargo, a través de una encuesta a 20 personas, 10 están a su favor. Probar que a un nivel de significancia del 5 %, la hipótesis de que el candidato a sobrestimado los votos a su favor.
3. En una muestra de 150 partidos de básquetbol universitario, el equipo de casa ganó 98 partidos. Realice una prueba para determinar si los datos sustentan la hipótesis de que en el básquetbol universitario el equipo de casa tiene ventaja. ¿A qué conclusión llega con $\alpha = 0.05$?

Capítulo 3

Prueba de Cuantiles

Lo que nos interesa en esta prueba binomial es hacer inferencia sobre los cuantiles de una variable aleatoria. Por ejemplo, muchos examinamos una muestra aleatoria de valores de alguna variable x para ver si la mediana de x es igual a 17 (por ejemplo).

La escala de medición suele ser al menos ordinal para esta prueba de hipótesis, aunque la prueba binomial solo require la escala nominal que es más débil para su medición. Esto se debe a que los cuantiles tienen poco significado con las mediciones de escala nominal.

Recordatorio

El cuantil de valor p de una variable aleatoria X es un número x tal que:

$$\mathbf{P}[X < x] \leq p$$

- En el caso de una variable aleatoria continua se da la igualdad.

3.1. Datos

Los datos consisten en una muestra:

$$X_1, X_2, X_3, \cdots, X_n \quad \text{una muestra aleatoria de tamaño } n \text{ de la población.}$$

3.2. Supuestos

1) Los datos a los que se les aplica esta prueba son una muestra aleatoria independiente e idénticamente distribuida.
2) La escala de medida de las $X_i s$ es al menos ordinal.

3.3. Estadístico de Prueba

Los estadísticos de prueba son:

$$T_1,$$

El número de observaciones en la muestra que son **menores o iguales** al valor x^* sobre el que se va a hacer la hipótesis. /break

Y también se utilizará como estadístico de prueba a:

$$T_2,$$

El número de observaciones en la muestra que son **estrictamente menores** a x^*.

Cuando $n \leq 20$ la distribución nula de éstos estadísticos es $Bin(n, p = p^*)$ con n el tamaño de la muestra y p^* dada en la hipótesis nula (recuerde que queremos hacer pruebas relacionadas con el cuantil x_p de la $v.a.$ en cuestión).

Por otro lado, si $n > 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el cuantil:

$$t = np + Z_q\sqrt{np(1-p)}$$

Donde Z_q es el cuantil de una distribución normal estándar que se puede obtener en la tabla correspondiente.

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

3.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_o : x_p = x^* \quad \text{es equivalente decir :} \quad \mathbf{P}[X \leq x^*] = p^*,$$

$$vs$$

$$\mathbf{H}_a : x_p \neq x^* \quad \text{es equivalente decir :} \quad \mathbf{P}[X \leq x^*] \neq p^*,$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 \leq t_1 \qquad o \qquad T_2 > t_2.$$

Elegimos $\alpha_1, \alpha_2 \geq 0$, tales que $\alpha_1 + \alpha_2 = \alpha$ el tamaño de la prueba y debemos encontrar t_1 y t_2 tales que:

$$\mathbf{P}[Y \leq t_1] = \alpha_1 \qquad y \qquad \mathbf{P}[Y \leq t_2] = 1 - \alpha_2$$

Donde $Y \sim Bin(n, p^*)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * min\{\mathbf{P}[Y \leq T_1], \mathbf{P}[Y \geq T_2].\}$$

Sugerimos que si $n > 20$, el $p - value$ puede obtenerse usando:

$$\mathbf{P}[Y \leq T_1] \approx \mathcal{N}\left(\frac{T_1 - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

y

$$\mathbf{P}[Y \geq T_2] \approx 1 - \mathcal{N}\left(\frac{T_2 - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Con $\mathcal{N}(\cdot)$ la función de distribución de la normal estándar.

Caso B (Prueba de cola inferior)

$$\mathbf{H}_0 : x_p \leq x^* \text{ es equivalente a decir: } \mathbf{P}[X < x^*] \geq p^*,$$

$$vs$$

$$\mathbf{H}_a : x_p > x^* \text{ es equivalente a decir : } \mathbf{P}[X < x^*] < p^*.$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 \leq t_1.$$

Elegimos α, el tamaño de la prueba y debemos encontrar t_1 tal que:

$$\mathbf{P}[Y \leq t_1] = \alpha$$

Donde $Y \sim Bin(n, p^*)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T_1]$$

Sugerimos que si $n > 20$, el $p - value$ podría obtenerse de forma mas sencilla usando la aproximación normal:

$$\mathbf{P}[Y \leq T_1] \approx \mathcal{N}\left(\frac{T_1 - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : x_p \geq x^* \text{ es equivalente a decir : } \mathbf{P}[X < x^*] \leq p^*,$$

$$vs$$

$$\mathbf{H}_a : x_p < x^* \text{ es equivalente a decir : } \mathbf{P}[X < x^*] > p^*.$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_2 > t_2$$

Elegimos α, el tamaño de la prueba y debemos encontrar t_2 tal que:

$$\mathbf{P}[Y \leq t_2] = 1 - \alpha$$

Donde $Y \sim Bin(n, p^*)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \geq T_2]$$

Sugerimos que si $n > 20$, el $p - value$ podría obtenerse de forma mas sencilla usando la aproximación normal:

$$\mathbf{P}[Y \geq T_2] \approx 1 - \mathcal{N}\left(\frac{T_2 - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo.

3.5. Ejemplo

El intervalo de tiempo entre las erupciones del geiser Old Faithful se registra 112 veces, de las cuales 8 son menores a 60 minutos y una es exactamente 60 minutos. Se quiere desea verificar que la mediana del intervalo es mayor a 60 minutos.

Paso 1 Escribimos la prueba a utilizar.

La prueba a utilizar **Prueba de cuantiles caso B cola inferior**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado,

$$\mathbf{H}_0 : \mathbf{P}[X \leq 60] \geq 0.50,$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[X \leq 60] < 0.50.$$

Donde X es el intervalo de tiempo entre las erupciones, suponiendo que ambos intervalos son independientes e identicamente distribuidas.

Paso 3 Estadístico de prueba.

La estadística de prueba será $T_1 = 9$ y $T_2 = 8$. Tomaremos $\alpha = .1$,

$$T_1 = 9 \text{ número de intervalos que son menores o iguales a 60 minutos.}$$

$$T_2 = 8 \text{ número de intervalos que son estrictamente menores a 60 minutos.}$$

$$T_1 \sim Bin(112, 0.50)$$

Paso 4 Procedimiento completo

Supuestos:

1. Muestra aleatoria de tamaño 112.

- Tomaremos $\alpha = 5$ el nivel de significancia.
- $n = 112$ tamaño de la muestra.

- $p^* = 0.50$.

Paso 5 Regla de decisión,

Rechazamos H_0 a un nivel de significancia α, si $T_1 \leq t_1$ y Rechazo H_0 si $p - value < \alpha$.

$$T_1 = 9 \leq 47.3 = t_1,$$

como $n > 20$ resultará más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el estadístico:

$$t_1 = np + Z_q\sqrt{np(1-p)} = (112)(0.5) + (-1.645)\sqrt{112(0.5)(1-0.5)},$$

$$t_1 = 47.3.$$

$\therefore$ Rechazamos H_0.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T_1] = \mathbf{P}[Y \leq 9].$$

Como $n > 20$, el $p - value$ puede obtenerse de una manera más sencilla usando la aproximación normal:

$$\mathbf{P}[Y \leq 9] \approx \mathcal{P}\left[Z \leq \frac{T_1 - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right] =$$

$$\left(Z \leq \frac{9 - (112)(0.5) + 0.5}{\sqrt{(112)(0.5)(1-0.5)}}\right) = \mathbf{P}[Z \leq -8.7876] << 0.0001.$$

$\therefore$ Rechazo H_0 ya que $p - value = 0.0001 << 0.05 = \alpha$. es decir, el $p - value$ es "muy pequeño".

Paso 6 Conclusión,

Existe información suficiente para decir que la mediana del intervalo de erupciones es mayor a 60 minutos.

3.6. Ejemplo en R-Studio

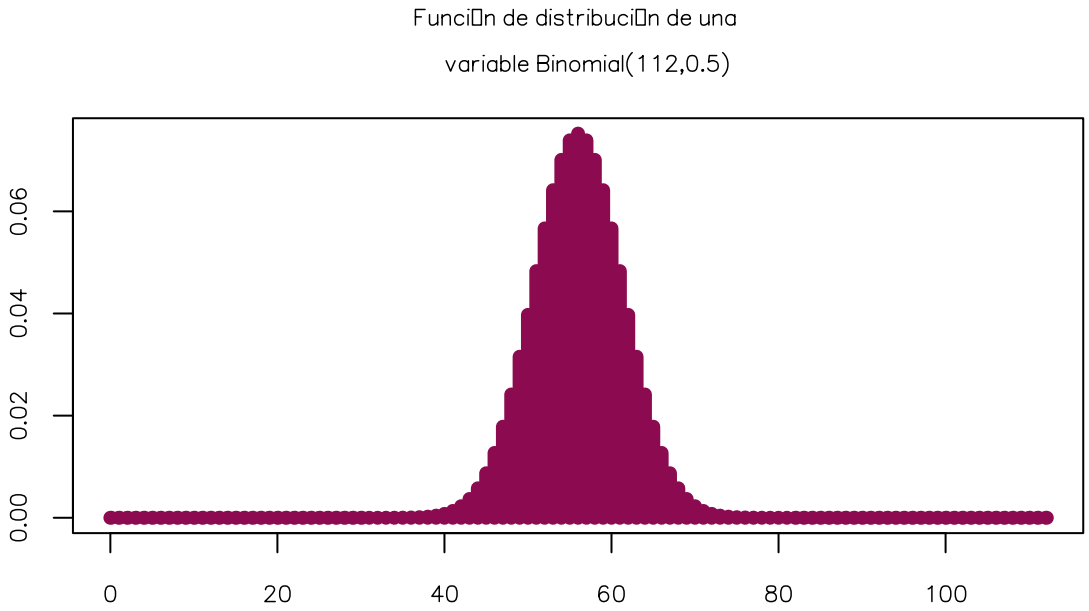
Ahora haremos la réplica en R.

La estadística de prueba será $T_1 = 9$ y $T_2 = 8$. Tomaremos $\alpha = .1$

```
#Datos

T_1=9          #Observaciones menores/iguales a 60 minutos
T_2=8          #Observaciones que son menores estrictamente a 60 minutos
alpha=0.05     #Nivel de significancia
n=112          #Tamaño de la muestra
p=0.5          #Cuantil en este caso la mediana
```

Según el planteamiento de las hipótesis, este es un Caso B (de cola inferior), por lo que siguiendo la regla de decisión, se rechaza H_0 si $T_1 \leq t_1$ donde t_1 será el cuantil que acumule 10 % en la distribución binomial

Podemos graficar la función de distribución:



A continuación calculamos t_1 y el $p-value$:

```
t_1=qbinom(.10,n,p)          #Cuantil a comparar con el estadístico de prueba
t_1
```

```
[1] 49
```

```
pvalue=pbinom(T_1,n,p)       #P-value correspondiente
pvalue
```

```
[1] 1.157256e-21
```

Tenemos que como $T_1 = 9 \leq 49 = t_1$, entonces se rechaza H_0 y por lo tanto se concluye que hay información suficiente para decir que la mediana de los intervalos de tiempo entre erupciones es mayor a 60 minutos.

Al mismo resultado llegamos si en lugar de buscar t_1 calculamos el $p-value$ de la estadística T_1 la cual da un valor muy cercano a cero por lo tanto cae en la región de rechazo y se conlcuye que la mediana de los intervalos de tiempo entre erupciones es mayor a 60 minutos.

Finalmente podemos utilizar la función en R.

```
binom.test(T_1,n,p=0.5,alternative = "less")
```

```
    Exact binomial test

data:  T_1 and n
number of successes = 9, number of trials = 112, p-value < 2.2e-16
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.000000 0.136034
sample estimates:
probability of success
            0.08035714
```

3.7. Ejercicios

1. Una muestra aleatoria de niños de tercer año de secundaria mostró las siguientes observaciones de peso (kg)

64	60	44	54	61
46	66	56	42	62
39	54	68	64	65
75	38	53	58	46

Probar:

a) La mediana de los pesos es 46kg.

b) El cuartíl superior es al menos 60kg.

c) El tercer decil no es mayor a 45kg.

2. La siguiente es una muestra de 15 departamentos nuevos de 2 recamaras con estacionamiento en la colonias Roma, Condesa y Escandón. Los datos están en millones de pesos.

6.4	5	4.2
4.6	4.4	5.6
3.5	3.8	4.5
7.5	5.6	4.2
8.1	5.8	6.3

Probar:

a) Cuando menos 50 % de las observaciones están por debajo de los 4.3 millones.

b) No mas del 20 % de las observaciones tienen un costo mayor a 7 millones.

Capítulo 4

Prueba de Signos

La prueba de signos merece una consideración especial debido a su versatilidad, su gran utilidad y simplicidad. Ésta es una prueba de proporciones cuando el valor específico $p^* = 1/2$ y para los casos de dos colas, cola inferior y cola superior la máxima probabilidad para rechazar la hipótesis nula H_0 se da cuando $p = 1/2$. Frecuentemente la prueba de signos también es apropiada para analizar datos de un vector aleatorio (X, Y) y ver si alguna de sus entradas tiene valores más grandes que la otra. De esta manera, si una variable tiende a tener valores mayores que la otra, se puede utilizar la prueba de signos para determinar si las medias de estas variables son diferentes.

4.1. Datos

Los datos consisten en observaciones bivariadas aleatorias:

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \cdots, (X_{n'}, Y_{n'}).$$

Dentro de cada par de datos en la muestra se clasificará de la siguiente manera:

- Por un signo " + " cuando $X_i < Y_i$
- Por un signo "-" cuando $X_i > Y_i$
- Se omitirán las parejas cuando $Xi = Yi$.

El tamaño de la muestra después de quitar los empates será n.

4.2. Supuestos

1) Las variables aleatorias bivariadas (X_i, Y_i) son mutuamente independientes.

2) La escala de medida es al menos ordinal dentro de cada par.

4.3. Estadístico de Prueba

El estadístico de prueba es:

$$T = Total\ de\ signos\ " + ",$$

La distribución nula de T es una distribución binomial con n el número de parejas de la muestra sin empates y $p = 1/2$.

$$T \sim Bin(n, 1/2).$$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

4.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_0 : \mathbf{P}[obtener\ +] = \mathbf{P}[obtener\ -],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[obtener\ +] \neq \mathbf{P}[obtener\ -].$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T \leq t \quad o \quad T > n - t.$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que:

$$\mathbf{P}[Y \leq t] = \alpha/2.$$

Donde $Y \sim Bin(n, 1/2)$.

Por otro lado, si $n > 20$ resultará más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el siguiente cuantil:

$$t = \frac{1}{2}\left(n + z_{\alpha/2}\ \sqrt{n}\right).$$

Donde $z_{\alpha/2}$ es el cuantil de una distribución normal estándar que se puede obtener en la tabla correspondiente.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * min\{\mathbf{P}[Y \leq T], \mathbf{P}[Y \geq T]\}.$$

Sugerimos que si $n > 20$, el $p - value$ puede obtenerse de forma más sencilla usando la aproximación normal:

$$\mathbf{P}[Y \leq T] \approx \mathcal{N}\left(\frac{2T - n + 1}{\sqrt{n}}\right),$$
$$\mathbf{P}[Y \geq T] \approx 1 - \mathcal{N}\left(\frac{2T - n - 1}{\sqrt{n}}\right)$$

Caso B (Prueba de cola inferior)

$$\mathbf{H}_0 : \mathbf{P}[obtener\ +] \geq \mathbf{P}[obtener\ -],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[obtener\ +] < \mathbf{P}[obtener\ -].$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T \leq t.$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que:

$$\mathbf{P}[Y \leq t] = \alpha.$$

Donde $Y \sim Bin(n, 1/2)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T]$$

Sugerimos que si $n > 20$,el $p - value$ puedría obtenerse de forma más sencilla usando la aproximación normal:

$$\mathbf{P}[Y \leq T] \approx \mathcal{N}\left(\frac{2T - n + 1}{\sqrt{n}}\right).$$

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : \mathbf{P}[obtener\ +] \leq \mathbf{P}[obtener\ -],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[obtener\ +] > \mathbf{P}[obtener\ -].$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T > n - t.$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que:

$$\mathbf{P}[Y \leq t] = \alpha.$$

Donde $Y \sim Bin(n, 1/2)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \geq T].$$

Sugerimos que si $n > 20$, el $p - value$ puedría obtenerse de forma más sencilla usando la aproximación normal:

$$\mathbf{P}[Y \geq T] \approx 1 - \mathcal{N}\left(\frac{2T - n - 1}{\sqrt{n}}\right).$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo.

4.5. Ejemplo

Un grupo de 6 amigos se puso a dieta, en un intento para perder peso obtuvieron los siguientes resultados:

Nombre	**Peso Antes**	**Peso Después**
Ed	174	165
Jim	191	186
Max	188	183
Ray	182	178
$Abdul$	201	203
$Phil$	188	181

El grupo de amigos desea ver si la dieta que están realizando es efectiva.

Recordando que cada par de datos en la muestra se clasificará de la siguiente manera:

- " + " cuando $X_i < Y_i$
- " − " cuando $X_i > Y_i$

y se omitirán las parejas cuando $X_i = Y_i$. Y el tamaño de la muestra después de quitar los empates será n, haremos lo siguiente:

Nombre	**Peso Antes** $\mathbf{X}_i$	**Peso Después** $\mathbf{Y}_i$	**Signo**
Ed	174	165	"
Jim	191	186	"
Max	188	183	"
Ray	182	178	"
$Abdul$	201	203	-"
$Phil$	188	181	"

Paso 1 Escribimos la prueba a utilizar.

La prueba a utilizar **Prueba de signos Caso B cola inferior.**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado,

$$\mathbf{H}_0 : \mathbf{P}[obtener\ +] \geq \mathbf{P}[obtener\ -],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[obtener\ +] < \mathbf{P}[obtener\ -].$$

es decir,

$\mathbf{H}_0$: En promedio los pesos antes de la dieta son mayores a los pesos después de la dieta.

vs

$\mathbf{H}_a$: En promedio los pesos después de las dieta son mayores a los pesos antes de la dieta.

Paso 3 Estadístico de prueba.

Utilizaremos el estadístico

$$T = 1 \ \ \textit{número de signos } \text{"}+\text{"}$$

$$T \sim Bin(6, 1/2)$$

Paso 4 Procedimiento completo.

Supuestos:

1. Muestra aleatoria de tamaño 6.

- Tomaremos como "éxito" los que si perdieron peso, en este caso los signos "+"
- Tomaremos α= 5 % =0.05 el nivel de significancia.
- $n = 6$ tamaño de la muestra.
- p=1/2.

Paso 5 Regla de decisión.

Rechazamos H_0 a un nivel de significancia α si, $T \leq t$

$$T = 1 \leq 1 = t.$$

$\therefore$ Rechazamos H_0.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T] = \mathbf{P}[Y \leq 1] = 0.1093.$$

$\therefore p - value = 0.1093 > 0.05 = \alpha$ entonces no rechazamos H_0.

Nos preguntaremos, ¿Cómo es que pasa esto?, bueno esto se debe a que nuestra muestra es muy pequeña (son 6 datos), claro no siempre vamos a tener muestras así. Pero, debemos tener en cuenta que al tener muestras reducidas le estamos reduciendo potencia a la prueba de hipótesis utilizada, ya que la capacidad de detectar diferencias significativas entre los datos disminuye.

Paso 6 Conclusión.

Existe evidencia suficiente para decir que los pesos después de la dieta son menores a los pesos antes de la dieta.

4.6. Ejemplo en R-Studio

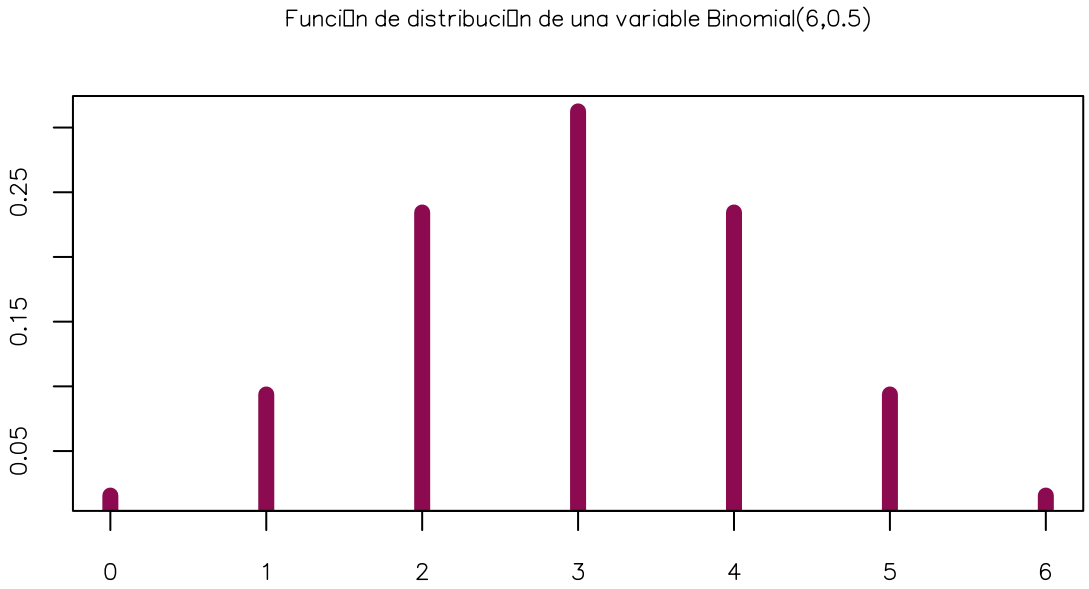
Ahora haremos la réplica en R.

La estadística de prueba será $T = 1$. Tomaremos $\alpha = 5\,\%$

```
#Datos

T=1               #Número de éxitos (+)
alpha=0.05        #Nivel de
n=6               #Tamaño de la muestra
p=0.5             #Para la prueba de signos "p" siempre será 1/2
```

Según el planteamiento de las hipótesis, este es un Caso B (de cola inferior), por lo que siguiendo la regla de decisión, se rechazará H_0 si $T \leq t$ donde t será en cuantil que acumule $5\,\%$ en la distribución binomial

Podemos graficar la función de distribución:



Calculamos t y el $p - value$:

```
t=qbinom(.05,n,p)          #Valor a comparar con nuestro estadístico
t
```

```
[1] 1
```

```
pvalue=pbinom(T,n,p)        #P-value
pvalue
```

```
[1] 0.109375
```

Tenemos que como $T = 1 \leq 1 = t$, entonces se rechaza H_0 y por lo tanto se concluye que hay información suficiente para decir que en promedio los pesos después de la dieta son menores a los pesos antes de la dieta. Por lo tanto la dieta parece funcionar.

Finalmente utilizaremos la función de R:

```
binom.test(T,n,p=0.5,alternative = "less")
```

```
    Exact binomial test

data:  T and n
```

```
number of successes = 1, number of trials = 6, p-value = 0.1094
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.5818034
sample estimates:
probability of success
             0.1666667
```

4.7. Ejercicios

1. En una encuesta anual para determinar si los salarios en el sector federal son proporcionales con los pagos en el sector privado, los trabajadores publicos y privados fueron emparejados tan cerca como fue posible (con respecto al tipo de trabajo, formación académica, años de experiencia, etc.) los salarios se ordenaron en parejas.

Pareja i	**Privado**	**Gobierno**
1	12,500	11,750
2	22,300	20,900
3	14,500	14,800
4	32,300	29,900
5	20,800	20,500
6	19,200	18,400
7	15,800	14,500
8	17,500	17,900
9	23,300	21,400
10	42,100	43,200
11	16,800	15,200
12	14,500	14,200

Probar la hipótesis nula de que los salarios son iguales contra la hipótesis alternativa de que el salario de los trabajadores federales es generalmente menor a la contraparte en el sector privado. Uitiliza $\alpha = 0.1$.

2. Una oficina tiene dos computadoras : A y B. En un estudio del uso del ordenador, la compañía ha recabado registro de las tasas de uso por 5 semanas. La meta es decidir cual computadora se pone bajo un contrato de servicio porque tiene una tasa alta de uso. Con los datos de las tasas de uso de la siguiente tabla se puede hacer una recomendación preliminar respecto a que computadora contratar?

Semanas	**Computadora A**	**Computadora B**
1	15.7	32.4
2	10.8	41.2
3	45	35.1
4	12.3	25
5	8.2	8.2

3. Se tienen dos trituradores de alimentos y se tiene la sospecha de que el aparato B es mas eficiente que el aparato A. Para probar dicha sospecha, se probaron en ambos trituradores diferentes alimentos y se registró el tiempo en minutos que le tomaba a cada aparato convertir el alimento en puré. Los resultados fueron los siguientes.

Alimento	A	B
1	0.5	0.6
2	1	0.9
3	1.2	1.2
4	0.8	0.9
5	0.4	0.5
6	1.5	1.8
7	0.3	0.4
8	1.2	1.4
9	0.7	0.9

Con los datos observados se puede corroborar la sospecha en cuanto a la eficiencia de los trituradores? Use $\alpha = 5\,\%$

Capítulo 5

Prueba Mc Nemar

El objetivo de esta prueba es ver el efecto que tuvo cierto "tratamiento" sobre un sujeto cuya "condición" se observa antes y después del mismo. Por ejemplo se puede usar para analizar el efecto que un debate tiene en la decisión de una asamblea. Para esto se puede hacer una encuesta para registrar la opinión de los miembros de la misma en categorías **a favor** o **en contra** de una propuesta y después de realizado el debate se les vuelve a preguntar su opinión sobre el tema. Nos interesa estudiar a los individuos que cambiaron de opinión, es decir, los que antes estaban **a favor** y ahora están **en contra** y los que estaban **en contra** y ahora están **a favor**.

La manera de estudiar esto es identificando a los individuos encuestados con parejas ordenadas de la forma (0,0),(0,1),(1,0) o (1,1) donde la primera entrada representa la **postura** del individuo antes del debate y la segunda representa su postura después, (por ejemplo, "0" puede representar **en contra** y "1" **a favor**). Luego se ingresan estos datos en una tabla de contingencia de la siguiente forma:

	ahora en contra	**ahora a favor**
antes en contra	$N\acute{u}mero\ de\ (0,0)$	$N\acute{u}mero\ de\ (0,1)$
antes a favor	$N\acute{u}mero\ de\ (1,0)$	$N\acute{u}mero\ de\ (1,1)$

5.1. Datos

Los datos consisten en observaciones bivariadas aleatorias:

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \cdots, (X_{n'}, Y_{n'}).$$

Dentro de cada par de datos en la muestra se tiene X e Y que sólo tomarán los valor 0 y 1.

En general los datos se resumen en una tabla de contingencia de la forma:

	después 0	**después 1**
antes 0	$\mathbf{a} = N\acute{u}mero\ de\ (0,0)$	$\mathbf{b} = N\acute{u}mero\ de\ (0,1)$
antes 1	$\mathbf{c} = N\acute{u}mero\ de\ (1,0)$	$\mathbf{d} = N\acute{u}mero\ de\ (1,1)$

5.2. Supuestos

1) Cada pareja (X_i, Y_i) son mutuamente independientes.

2) La escala de medida es nominal con 2 categorías para toda X_i y Y_i.

3) La probabilidad de que se observe (0,1) es $\geq$ que la de (1,0) o la de (1,0) es $\geq$ (0,1) para todos los elementos de la muestra.

5.3. Estadístico de Prueba

El estadístico de prueba es:

$$T_1 = \frac{(b-c)^2}{(b+c)} \approx \chi^2_{(1)}.$$

Es decir, la distribución de T_1 es aproximadamente $\chi^2_{(1)}$.

Por otro lado, si $b + c \leq 20$, es preferible utilizar el siguiente estadístico:

$$T_2 = b \sim Bin(b + c, p).$$

Es decir, la distribución de T_2 es exactamente $Bin(b + c, p)$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

5.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_0 : \mathbf{P}[X_i = 0, Y_i = 1] = \mathbf{P}[X_i = 1, Y_i = 0],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[X_i = 0, Y_i = 1] \neq \mathbf{P}[X_i = 1, Y_i = 0].$$

Es decir, se quiere ver si la probabilidad de que el individuo pase de 0 a 1 es la misma que la probabilidad de que pase de 1 a 0. En otras palabras, se quiere comprobar si el efecto del tratamiento fue neutro o si hay alguna tendencia tras el mismo.

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_2 \leq t \quad o \quad T_2 > n - t.$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que:

$$\mathbf{P}[Y \leq t] = \alpha/2.$$

Donde $Y \sim Bin(b + c, 1/2)$.

Por otro lado, si $b + c > 20$,

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 > t_{1-\alpha}.$$

Elegimos α, el tamaño de la prueba y debemos encontrar $t_{1-\alpha}$ tal que, $t_{1-\alpha}$ se encuentra en la tabla correspondiente ya que $T_1 \sim \chi^2_{(1)}$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * min\{\mathbf{P}[Y \leq T_2], \mathbf{P}[Y \geq T_2]\}$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo.

5.5. Ejemplo

May y Johnson (1997) publicaron el resultado de un estudio en el cual los investigadores querían determinar el efecto de la hipnosis en reducir el dolor asociado con la venopunción en pacientes juveniles de cáncer. Los datos se observan en la siguiente tabla:

	Después de la hipnosis		
Antes de lahipnosis	**Sin Dolor**	**Con Dolor**	*Total*
Sin Dolor	**a** = 18	**b** = 4	22
Con Dolor	**c** = 12	**d** = 5	17
Total	30	9	39

Paso 1 Escribimos la prueba a utilizar.

La prueba a utilizar **Prueba de Mc Nemar Caso A de dos colas.**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado:

$$\mathbf{H}_0 : \mathbf{P}[X_i = 0, Y_i = 1] = \mathbf{P}[X_i = 1, Y_i = 0],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[X_i = 0, Y_i = 1] \neq \mathbf{P}[X_i = 1, Y_i = 0],$$

es decir,

$$\mathbf{H}_0 : \text{ Se mantiene el sentir dolor después de la hipnosis.}$$

$$vs$$

$$\mathbf{H}_a : \text{ Hay un cambio considerable al no sentir dolor después de la hipnosis.}$$

Paso 3 Estadístico de prueba.

Recordando que $b + c = 4 + 12$ entonces $b + c = 16 \leq 20$; entonces utilizaremos el estadístico:

$$T_2 = b = 4.$$
$$T_2 \sim Bin(16, 1/2).$$

Paso 4 Procedimiento completo.

Supuestos

1. Muestra aleatoria de tamaño n=16.

- Tomaremos α= 5 % =0.05 el nivel de significancia.
- $n = 16$ tamaño de la muestra.
- p=1/2.

Paso 5 Regla de decisión.

Rechazamos H_0 a un nivel de significancia α si $T_2 \leq t$ o $T_2 > n - t$.

Debemos encontrar t tal que:

$$\mathbf{P}[Y \leq t] = \alpha/2 = 0.025.$$

$$T_2 = 4 \leq 5 = t \quad o \quad 5 > 16 - 4.$$
$$T_2 = 4 \nleq 5 = t \quad y \quad 5 \not> 12.$$

$\therefore$ Rechazamos H_0.

Paso 6 Conclusión.

Existe evidencia suficiente para decir que hay un cambio considerable a no sentir dolor después de recibir hipnosis.

5.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

Veamos que $b + c = 4 + 12 = 16 \leq 20$ por lo tanto el estadístico de prueba será:

$T_2 = b = 4$. $T_2 \sim Bin(16, 1/2)$.

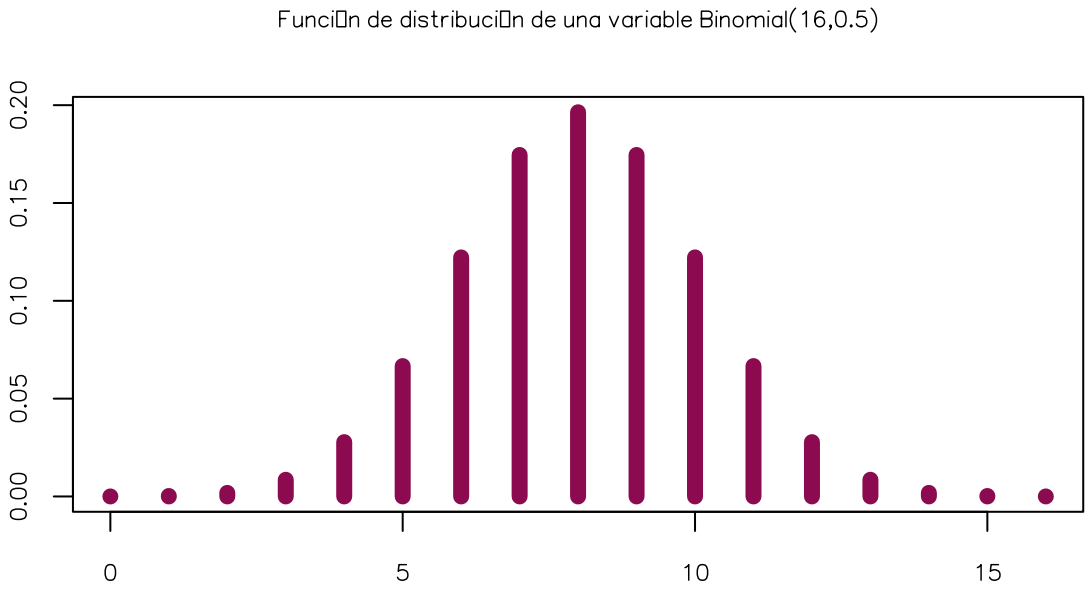
Tomaremos $\alpha = 5\,\%$

```
              Despues hipnosis
Antes hipnosis Sin dolor dolor
     Sin dolor        18     4
     dolor            12     5
```

```
#Datos

T_2=4                    #Es la posición b
n=16                     #Tamaño de la muestra
alpha=0.05               #Nivel de significancia
p=0.5                    #Para la prueba Mc nemar p es 1/2
```

Según el planteamiento de las hipótesis, este es un Caso A (dos colas), por lo que siguiendo la regla de decisión tenemos que rechazaremos H_0 si $T_2 \leq t$ o $T_2 > n - t$ donde t será en cuantil que acumule $5\,\%$ en la distribución binomial.

Podemos graficar la función de distribución:



Calculamos t y el $p-value$:

```
t=qbinom(.05,n,p)        #Valor con el que vamos a comparar el estadístico
t
```

```
[1] 5
```

```
pvalue=2*min(c(pbinom(T_2,n,p), pbinom(T_2,n,p,lower.tail = F)))    #P-value
pvalue
```

```
[1] 0.07681274
```

Tenemos que como $T_2 = 4 \leq 5 = t$, entonces se rechaza H_0 como ya rechazamos en la cola inferior no es necesario probar la cola superior, pero si NO hubieramos rechazado con la primer ecuación debemos probar si $T_2 > n - t$.

Por lo tanto se concluye que hay información suficiente para decir que hay un cambio considerable al no sentir dolor después de la hipnosis.

Finalmente utilizaremos la función en R:

```
binom.test(4,n,p,"t")
```

```
    Exact binomial test

data:  4 and n
number of successes = 4, number of trials = 16, p-value = 0.07681
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.07266204 0.52377082
sample estimates:
probability of success
                  0.25
```

5.7. Ejercicios

1. Se toma una muestra aleatoria de 135 ciudadanos de E.U. y se les preguntó su opinión con respecto a cierta política. El estudio registró a 43 ciudadanos que estaban en contra de esa política. Después de varias semanas, durante las cuales los ciudadanos recibieron cartas informativas, se les volvió a preguntar su opinión; 37 estuvieron en contra, y 30 de las 37 originalmente no estaban en contra de la política. ¿Es significativo el cambio en el número de personas en contra de la política?

2. Un investigador intenta determinar si un fármaco tiene un efecto sobre una enfermedad particular. Se cuenta con la información de los individuos en el estudio con el diagnóstico (enfermedad: presente o ausente) antes del tratamiento, y el diagnóstico después del tratamiento

	Después de tratamiento		
Antes de tratamiento	**Presente**	**Ausente**	Total
Presente	101	121	222
Ausente	59	33	92
Total	160	154	314

Capítulo 6

Prueba de Cox Stuart

La siguiente prueba tiene como objetivo identificar tendencias en series de observaciones. Si $\{X_i\}_{i=1}^n$ es una serie de observaciones consecutivas, una manera de intentar descubrir si hay o no una tendencia es fijarse en las diferencias del tipo $X_{i+c} - X_i$ con $c = \frac{n}{2}$ si n es par y $c = \frac{(n+1)}{2}$ si es impar. Si $\{X_i\}_{i=1}^n$ tuviera una tendencia creciente, esperaríamos que las diferencias $X_{i+c} - X_i$ fuesen, en promedio, positivas, en cambio, si la tendencia fuese decreciente, se esperaría que las $X_{i+c} - X_i$ fueran negativas, en promedio. Note que la elección de c produce las mayores distancias entre índices temporales para diferencias homogéneas de la serie de tiempo.

6.1. Datos

Los datos son observaciones consecutivas de una serie de tiempo $\{X_i\}_{i=1}^n$.

En esta prueba se agrupan en parejas de la forma (X_i, X_{i+c}) a las que se les asigna el signo " + " si $X_{i+c} > X_i$, el signo " − " si $X_{i+c} < X_i$ o "0" si hay empates.

6.2. Supuestos

1) Las variables aleatorias $X_1, X_2, \ldots, X_n$ son mutuamente independientes.
2) La escala de medida de las $X_i s$ es al menos ordinal.
3) Las $X_i s$ son todas idénticamente distribuidas o existe una tendencia creciente o decreciente.

6.3. Estadístico de Prueba

El estadístico de prueba es:

$$T = Total\ de\ signos\ " + ".$$

La distribución nula de T es una distribución binomial con n el número de parejas de la muestra sin empates y $p = 1/2$.

$$T \sim Bin(n, 1/2).$$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

6.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_0 : \mathbf{P}[obtener\ +] = \mathbf{P}[obtener\ -],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[obtener\ +] \neq \mathbf{P}[obtener\ -].$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T \leq t \quad o \quad T > n - t.$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que:

$$\mathbf{P}[Y \leq t] = \alpha/2.$$

Donde $Y \sim Bin(n, 1/2)$.

Por otro lado, si $n > 20$ resultará más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el estadístico:

$$t = \frac{1}{2}\left(n + z_{\alpha/2}\ \sqrt{n}\right).$$

Donde $z_{\alpha/2}$ es el cuantil de una distribución normal estándar que se puede obtener en la tabla correspondiente.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * min\{\mathbf{P}[Y \leq T], \mathbf{P}[Y \geq T]\}.$$

Sugerimos que si $n > 20$, el $p - value$ podrá obtenerse de forma más sencilla usando la aproximación normal:

$$\mathbf{P}[Y \leq T] \approx \mathcal{N}\left(\frac{2T - n + 1}{\sqrt{n}}\right).$$
$$\mathbf{P}[Y \geq T] \approx 1 - \mathcal{N}\left(\frac{2T - n - 1}{\sqrt{n}}\right).$$

Caso B (Prueba de cola inferior)

$$\mathbf{H}_0 : \mathbf{P}[obtener\ +] \geq \mathbf{P}[obtener\ -],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[obtener\ +] < \mathbf{P}[obtener\ -].$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T \leq t.$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que:

$$\mathbf{P}[Y \leq t] = \alpha.$$

Donde $Y \sim Bin(n, 1/2)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T].$$

Sugerimos que si $n > 20$, el $p - value$ podrá obtenerse de forma más sencilla usando la aproximación normal:

$$\mathbf{P}[Y \leq T] \approx \mathcal{N}(\frac{2T - n + 1}{\sqrt{n}}).$$

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : \mathbf{P}[obtener\ +] \leq \mathbf{P}[obtener\ -],$$

$$vs$$

$$\mathbf{H}_a : \mathbf{P}[obtener\ +] > \mathbf{P}[obtener\ -].$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T > n - t.$$

Elegimos α, el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = \alpha$$

Donde $Y \sim Bin(n, 1/2)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \geq T]$$

Sugerimos que si $n > 20$, el $p - value$ podría obtenerse de forma más sencilla usando la aproximación normal:

$$\mathbf{P}[Y \geq T] \approx 1 - \mathcal{N}\left(\frac{2T - n - 1}{\sqrt{n}}\right).$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo.

6.5. Ejemplo

Una hidroeléctrica está muy interesada en seguir las tasas promedio de descarga de agua de las corrientes que lo alimentan. Se tienen los registros mensuales de estas tasas (en pies cúbicos por segundo) durante un período de 24 meses. La empresa sospecha que la tasa está disminuyendo. ¿Podemos corroborar la sospecha con un nivel de significancia del 5 %?

Los datos están en la siguiente tabla, los cuales están emparejados por mes ya que estas tasa de descarga siguen un ciclo anual, se sabe que la tasa de descarga sigue un ciclo anual, por lo que no se puede emparejar las descargas de corriente durante dos meses diferentes; sin embargo, al emparejar los mismos meses consecutivos, se puede investigar la existencia de una tendencia. Los datos son los siguientes:

Mes	**Primer Año**	**Segundo Año**	**Mes**	**Primer Año**	**Segundo Año**
Enero	14.6	14.2	*Julio*	92.8	88.1
Febrero	12.2	10.5	*Agosto*	74.4	80
Marzo	104	123	*Septiembre*	75.4	75.6
Abril	220	190	*Octubre*	51.7	48.8
Mayo	110	138	*Noviembre*	29.3	27.1
Junio	86	98.1	*Diciembre*	16	15.7

Los datos ya están emparejados ahora sólo debemos asignar los signos considerando X_i el primer año y Y_i el segundo año.

Recordando que cada par de datos en la muestra se clasificará por un signo " + " cuando $X_i < Y_i$, por un signo "-" cuando $X_i > Y_i$ y se omitirán las parejas cuando $X_i = Y_i$. Y el tamaño de la muestra después de quitar los empates será n, haremos lo siguiente:

Mes	**Primer Año**	**Segundo Año**	**Signo**
Enero	14.6	14.2	−
Febrero	12.2	10.5	−
Marzo	104	123	+
Abril	220	190	−
Mayo	110	138	+
Junio	86	98.1	+
Julio	92.8	88.1	−
Agosto	74.4	80	+
Septiembre	75.4	75.6	+
Octubre	51.7	48.8	−
Noviembre	29.3	27.1	−
Diciembre	16	15.7	−

Paso 1 Escribimos la prueba a utilizar.

La prueba a utilizar **Prueba Cox-Stuart Caso B cola inferior.**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado:

$$\mathbf{H}_0: \text{ La tasa promedio de descarga de agua no está disminuyendo.}$$

$$vs$$

$$\mathbf{H}_a: \text{ La tasa promedio de descarga de agua está disminuyendo.}$$

Paso 3 Estadístico de prueba.

Utilizaremos el estadístico

$$T = 5 \;\; \textit{número de signos } " + ".$$

$$T \sim Bin(12, 1/2).$$

Paso 4 Procedimiento completo

Supuestos:

1. Muestra aleatoria de tamaño 12.

- Tomaremos como "éxito" los que si están disminuyendo, en este caso los signos "+".
- Tomaremos α= 5 % =0.05 el nivel de significancia.
- $n = 12$ tamaño de la muestra.
- p=1/2.

Paso 5 Regla de decisión.

Rechazamos H_0 a un nivel de significancia α si $T \leq t$.

$$T = 5 \nleq 3 = t.$$

$\therefore$ No rechazamos H_0.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T] = \mathbf{P}[Y \leq 5] = 0.3872.$$

$\therefore p - value = 0.3872 > 0.05 = \alpha$ entonces no rechazamos H_0.

Paso 6 Conclusión.

Existe evidencia suficiente para decir que la tasa promedio de descarga de agua no está disminuyendo.

6.6. Ejemplo en R-Studio

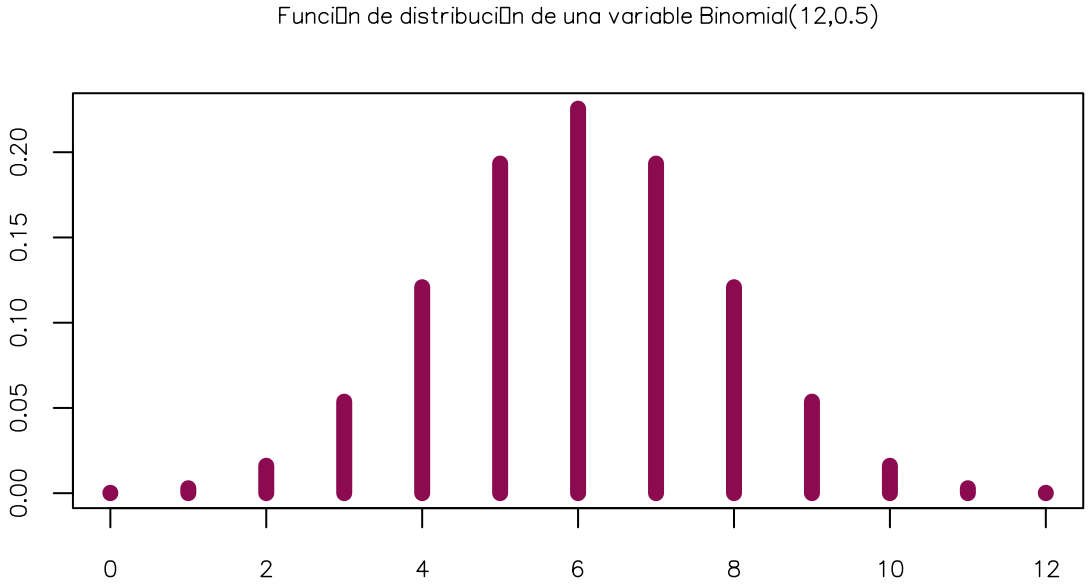
Ahora haremos la réplica en R.

La estadística de prueba será $T = 5$. Tomaremos $\alpha = 5\,\%$

```
#Datos

T_1=5                #Número de éxitos (+)
alpha=0.05          #Nivel de significancia
n=12                 #Tamaño de la muestra
p=0.5               #Para la prueba de cox "p" siempre será 1/2
```

Según el planteamiento de las hipótesis, este es un Caso B (de cola inferior), por lo que siguiendo la regla de decisión se rechaza H_0 si $T \leq t$ donde t será en cuantil que acumule 5 % en la distribución binomial.

Podemos graficar la función de distribución:



Calculamos t y el $p-value$:

```
t=qbinom(.05,n,p)        #Cuantil a comparar con el estadístico
t
```

```
[1] 3
```

```
pvalue=pbinom(T_1,n,p)   #p-value
pvalue
```

```
[1] 0.387207
```

Tenemos que como $T = 5 \not\leq 3 = t$, entonces NO se rechaza H_0 y por lo tanto se concluye que NO hay información suficiente para decir que la tasa promedio de descarga mensual este disminuyendo.

Finalmente utilizaremos la función en R:

```
binom.test(5,n,p=0.5,alternative = "less")
```

```
    Exact binomial test

data:  5 and n
number of successes = 5, number of trials = 12, p-value = 0.3872
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.6847622
sample estimates:
probability of success
             0.4166667
```

```
x=c(14.6,14.2,12.2,10.5,104,123,220,190,110,138,86,98.1,92.8,88.1,74.4,80,75.4,75.6,51.7,48.8,29.3,2


library(randtests)
cox.stuart.test(x)
```

```
    Cox Stuart test
```

```
data:  x
statistic = 4, n = 12, p-value = 0.3877
alternative hypothesis: non randomness
```

6.7. Ejercicios

1. Un programa ecológico sobre la contaminación de un río tomó 5 muestras de agua de diferentes lugares de un río antes y después de dos años, obteniéndose los siguientes resultados. Los números representan la media de la contaminación, donde medidas grandes indican alta contaminación.

Numero de Muestras	**Medidas iniciales**	**Medidas después de 2 años**
1	88.4	87.1
2	81.3	79.4
3	68.4	69.1
4	100.5	91.1
5	93.2	95.3

Se está interesado en saber si el programa de rehabilitación ecológica ha tenido efecto en la reducción de la contaminación. Use $\alpha = 1\,\%$.

2. Se presentan a continuación los tipos de cambio MXN-USD de los últimos 30 días.

24.48 , 24.07 , 23.66 , 23.70 , 23.93 , 24.34 , 24.22 , 23.87 , 23.96 , 24.05 , 23.78 , 23.74 , 23.21 , 22.88 , 22.74 , 22.73 , 22.58 , 22.21 , 22.33 , 22.21 , 22.17 , 22.21 , 22.04 , 21.79 , 21.76 , 21.91 , 21.58 , 21.60 , 21.50 , 21.91

Se desea corroborar la aseveración del presidente respecto a que el peso mexicano está recuperando fuerza. Use $\alpha = 10\,\%$.

Parte III

Prueba de Rango

Introducción

Ya nos empapamos de la pruebas binomiales ahora en esta sección se presentan tres de las pruebas de rangos más utilizadas en modelos no paramétricos, la prueba de Mann y Witney, la prueba de Kruskal-Wallis y la prueba de Igualdad de Varianzas.

Estas pruebas son otros métodos no paramétrico que se usan para determinar si hay diferencia entre poblaciones. El procedimiento se basa en ordenar las observaciones pertenecientes a muestras aleatorias de cada población, asignarles rangos según sus valores y posteriormente construir la prueba sobre dichos rangos.

Capítulo 7

Prueba U-Mann y Witney

Esta prueba está diseñada para determinar si dos muestras han sido extraídas de la misma población. La prueba también es conocida como Mann-Whitney-Wilcoxon o suma de rangos de Wilcoxon.

A diferencia de las pruebas binomiales bivariadas vistas en la sección anterior, ésta no se basa en una muestra por pares. En lugar usa dos muestras independientes, una de cada población a probar.

7.1. Datos

Se considera que los datos provienen de dos muestras aleatorias:

$x_1, x_2, x_3, \cdots, x_n$ una muestra aleatoria de tamaño n de la población.

$y_1, y_2, y_3, \cdots, y_m$ una muestra aleatoria de tamaño m de la población.

De forma conveniente, tomamos $N = n + m$.

Sea:

$F(t)$ la función de distribución de probabilidad de X.

$G(t)$ la función de distribución de probabilidad de Y.

7.2. Supuestos

1) Independencia dentro de cada muestra.
2) Independencia entre muestras.
3) La escala de medida es al menos ordinal.

Pasos a seguir para obtener R(X_{i})

1. Formar una nueva muestra combinada de los $n + m$ datos.
2. Ordenar de menor a mayor.
3. Asignamos el rango correspondiente.
4. Si hay empates entonces asignamos el promedio de los rangos.

7.3. Estadístico de Prueba

Utilizaremos el sieguiente estadístico de prueba:

- **Si no hay empates**

$$T = \sum_{i=1}^{n} R(X_i).$$

- **Si hay empates**

$$T_1 = \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

$$T_1 \sim N(0,1).$$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

7.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_0 : F(x) = G(x) \ \ \forall\, x, \quad \text{de manera equivalente} \quad \mathbf{H}_0 : \mathbf{E}[X] = \mathbf{E}[Y].$$

$$vs$$

$$\mathbf{H}_a : F(x) \neq G(x) \ \ \text{para alguna } x, \quad \text{de manera equivalente} \quad \mathbf{H}_a : \mathbf{E}[X] \neq \mathbf{E}[Y].$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T < t_{\frac{\alpha}{2}} \quad o \quad T > t_{1-\frac{\alpha}{2}}.$$

Donde $t_{\frac{\alpha}{2}}$ el cuantil inferior se busca en las tablas correspondientes a nuestra prueba.

Y $t_{1-\frac{\alpha}{2}}$ el cuantil superior se calcula de la siguiente manera:

$$\omega_p = n(n+m+1) - \omega_{1-p}.$$

donde ω_{1-p} es el cuantil inferior obtenido de la tabla correspondiente a nuestra prueba.

Los cuantiles aproximados en el caso de no tener empates, y n o m mayores a 20, se encuentra la aproximación normal

$$\omega_p = \frac{n(N+1)}{2} + z_p\sqrt{\frac{nm(N+1)}{12}}$$

donde el cuantil z_p se obtiene de la tabla de la distribución normal y calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * \mathbf{P}\left(Z \leq \frac{T + \frac{1}{2} - n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}}\right).$$

Donde Z es la variable aleatoria normal estándar.

- Si ocupamos T_1, rechazamos H_0 si:

$$T_1 < z_{\frac{\alpha}{2}} \quad o \quad T_1 > z_{1-\frac{\alpha}{2}}.$$

Donde $z_{\frac{\alpha}{2}}$ y $z_{1-\frac{\alpha}{2}}$ son cuantiles que se buscan en la tabla de una normal estándar

y calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * min\{\mathbf{P}[Z \leq T_1], \mathbf{P}[Z \geq T_1]\}.$$

Donde Z es la variable aleatoria normal estándar.

Caso B (Prueba de cola inferior)

$$\mathbf{H}_0 : F(x) = G(x) \;\; vs \;\; \mathbf{H}_a : F(x) > G(x), \;\; \text{de manera equivalente} \;\; \mathbf{H}_a : \mathbf{E}[X] < \mathbf{E}[Y]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T < t_\alpha.$$

Donde t_α el cuantil inferior se busca en las tablas correspondientes a nuestra prueba.

Los cuantiles aproximados en el caso de no tener empates, y n o m mayores a 20, se encuentra la aproximación normal

$$\omega_p = \frac{n(N+1)}{2} + z_p\sqrt{\frac{nm(N+1)}{12}},$$

donde el cuantil z_p se obtiene de la tabla de la distribución normal.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}\left(Z \leq \frac{T + \frac{1}{2} - n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}}\right).$$

- Si ocupamos T_1, rechazamos H_0 si:

$$T_1 < z_\alpha.$$

Donde z_α es el cuantil que se buscan en la tabla de una normal estándar

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Z \leq T_1].$$

Donde Z es la variable aleatoria normal estándar.

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : F(x) = G(x) \;\; vs \;\; \mathbf{H}_a : F(x) < G(x), \;\; \text{de manera equivalente} \;\; \mathbf{H}_a : \mathbf{E}[X] > \mathbf{E}[Y].$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T > t_{1-\alpha}.$$

Y $t_{1-\alpha}$ el cuantil superior se calcula de la siguiente manera:

$$\omega_p = n(n+m+1) - \omega_{1-p}$$

donde ω_{1-p} es el cuantil inferior obtenido de la tabla correspondiente a nuestra prueba.

Los cuantiles aproximados en el caso de no tener empates, y n o m mayores a 20, se encuentra la aproximación normal

$$\omega_p = \frac{n(N+1)}{2} + z_p\sqrt{\frac{nm(N+1)}{12}}.$$

donde el cuantil z_p se obtiene de la tabla de la distribución normal.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}\left(Z \geq \frac{T + \frac{1}{2} - n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}}\right).$$

- Si ocupamos T_1, rechazamos H_0 si:

$$T_1 > z_{1-\alpha}$$

Donde $z_{1-\alpha}$ es el cuantil que se buscan en la tabla de una normal estándar

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Z \geq T_1] \;\; o \;\; 1 - \mathbf{P}[Z \leq T_1].$$

Donde Z es la variable aleatoria normal estándar.

Tal vez es demasiada información que procesar así que veremos un ejemplo paso por paso.

7.5. Ejemplo

Una agencia publicitaria está investigando a qué tipo de avisos les prestan más atención los adolescentes. Se observan a 11 adolescentes, a 6 de ellos se les muestran anuncios de comida y a los 5 restantes se les muestran anuncios de bebidas; todos los anuncios tienen una duración similar y a continuación se muestra el registro del tiempo de atención (en segundos) de los 11 adolescentes. Utilizaremos α= 5 %.

Comida	25	41	42	45	47	50
Bebidas	23	28	30	35	38	

Podemos ver que m=6 (en este caso comidas) y n= 5 (en este caso bebidas).

Paso 1 Escribimos la prueba a utilizar.

La prueba a utilizar **Prueba U-Mann-Whitney - Caso A - dos colas.**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado,

$\mathbf{H}_0$: La distribución del tiempo de atención que prestan los adolescentes a los anuncios sobre comida es igual a la distribución de los anuncios de bebidas.

$$vs$$

$\mathbf{H}_a$: La distribución del tiempo de atención que prestan los adolescentes a los anuncios sobre comida es distinta a las distribución de los anuncios de bebidas.

Ordenamos de menor a mayor

Tipo de aviso	B	C	B	B	B	B	C	C	C	C	C
Dato ordenado	23	25	28	30	35	38	41	42	45	47	50
Rango	1	2	3	4	5	6	7	8	9	10	11

Paso 3 Estadístico de prueba.

Como no tenemos empates utilizaremos el estadistico T:

$$T = \sum_{i=1}^{n} R(X_i).$$

Paso 4 Procedimiento completo,

$$T = \sum_{i=1}^{5} R(X_i) = 1 + 3 + 4 + 5 + 6 = 19.$$

Paso 5 Regla de decisión.

Rechazamos H_0 a un nivel de significancia α=.05 si:

$$T < t_{\frac{\alpha}{2}} \quad o \quad T > t_{1-\frac{\alpha}{2}}.$$

El valor del cuantil inferior localizado en las tablas es 19, y por otro lado vamos a calcular el cuantil superior:

$$\omega_p = n(n+m+1) - \omega_{1-p}.$$
$$\omega_p = 5(5+6+1) - 19.$$
$$\omega_p = 41.$$

$$19 \not< 19 \quad o \quad 19 \not> 41.$$

$\therefore$ No rechazo H_0

Calculamos p-value

$$p-value = 2 * \mathbf{P}\left(Z \leq \frac{T+\frac{1}{2}-n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}}\right).$$

$$p-value = 2 * \mathbf{P}\left(Z \leq \frac{19+\frac{1}{2}-5\frac{12}{2}}{\sqrt{\frac{5*6(12)}{12}}}\right) = 0.028.$$

$$p-value = 2 * 0.028.$$

$$p-value = 0.056.$$

Ya que $p-value > \alpha$,

$\therefore$ No rechazo H_0.

Paso 6 Conclusión.

$\therefore$ No hay diferencia significativa en el tiempo de atención de los adolescentes en los anuncios de comida y bebidas.

7.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

El primer paso es ordenar los rangos:

```
R_i=c(1,2,3,4,5,6,7,8.5,8.5,10,11)
```

El segundo paso es calcular T que corresponde a la suma de los rangos asignados a los anuncios de "Bebida".

```
T=1+3+4+5+6  #rangos asignados a los anuncios de bebida
T
```

```
[1] 19
```

```
n=5
m=6
N=n+m
```

Según el planteamiento de las hipótesis, este es un Caso A (de 2 colas), por lo que siguiendo la regla de decisión, se rechazará H_0 si $T \leq t_1$ o $T > t_2$, donde t_1 es el cuantil que acumula 2.5 % de probabilidad en una distribución normal estándar, y t_2 es el cuantil que acumula 97.5 % en la misma distribución.

```
t1=19
t1
```

```
[1] 19
```

```
t2=41
t2
```

```
[1] 41
```

```
pvalue=2*0.028
pvalue
```

```
[1] 0.056
```

Observamos que $T \leq t_1$ por lo tanto rechazaremos H_0 y concluimos que la distribución del tiempo de atención que prestan los adolescentes a los anuncios sobre comida es distinta a la distribución del tiempo de atención prestada a los anuncios de bebidas.

En R la prueba **"wilcox.test"** también realiza esta prueba con un cálculo ligeramente distinto de la estadística de prueba. En el caso de la prueba de Wilcoxon lo que contaremos es el número de ocasiones que la $x_i > y_j$ para todas las combinaciones de $i = 1, ...n$ y $j = 1, ..., m$.

Si las distribuciones son iguales entonces se esperaría que el número de veces que $x_i > y_j$ sea cercano a la mitad de las combinaciones.

```
B=c(23, 28, 30, 35, 38)      #Datos de bebidas
C=c(25, 41, 42, 45, 47, 50) #DAtos de comidas
m1<-wilcox.test(B,C, na.rm=TRUE, paired=FALSE, exact=FALSE)
print(m1)
```

```
    Wilcoxon rank sum test with continuity correction

data:  B and C
W = 4, p-value = 0.05523
alternative hypothesis: true location shift is not equal to 0
```

Con esta prueba se rechaza la hipótesis nula de que ambas distribuciones son iguales. Y se concluye de la misma forma que la prueba de rangos anterior. La distribución del tiempo de atención que prestan los adolescentes a los anuncios sobre comida es distinta a la distribución del tiempo de atención prestada a los anuncios de bebidas.

##Ejercicios

1. Siete estudiantes aprendieron álgebra utilizando el método actual y seis estudiantes aprendieron álgebra según un nuevo método.

Método	**Puntajes**						
Actual	68	72	79	69	84	80	78
Nuevo	64	60	68	73	72	70	

¿Con una α del 5 % podría decir que la efectividad (medida con los punatjes) de ambos métodos es similar?

2. En un laboratorio con entorno controlado, 10 hombres y 10 mujeres fueron evaluados para determinar la temperatura ambiente que encontraron más cómoda. los resultados fueron los siguientes:

Hombres	**Mujeres**
74	75
72	77
77	78
76	79
76	77
73	73
75	78
73	79
74	78
75	80

Suponiendo que estas temperaturas son una muestra aleatoria de la población, ¿la temperatura promedio es la misma para hombres y mujeres?.

Capítulo 8

Intervalo de confianza para la diferencia entre dos medias

8.1. Datos

Los datos consisten en dos muestras aleatorias $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_m$ de tamaño n y m, respectivamente. Sea X y Y las variables aleatorias con la misma distribución que las X_i y las Y_j, respectivamente.

8.2. Supuestos

1) Ambas muestras son muestras aleatorias de sus respectivas poblaciones.
2) Además de la independencia dentro de cada muestra, existe mutua independencia entre las dos muestras.
3) Las dos funciones de distribución de la población son idénticas, excepto por una posible diferencia en los parámetros de ubicación. Es decir, hay una constante d (por ejemplo) tal que X tiene la misma función de distribución que $Y + d$.

8.3. Método

Determine el cuantil $\alpha/2$ ($\omega_{\alpha/2}$) para n y m de las tablas correspondientes a **Prueba Mann-Whitney**, o si n y m son mayores a 20 se ocupa la aproximación del cuantil, donde $1-\alpha$ es el coeficiente de confianza. Note que esto se puede utilizar incluso si hay muchos empates.

Luego calcule k, dado por:

$$k = \omega_{\alpha/2} - n(n+1)/2$$

Para todos los pares posibles (X_i, Y_j), encuentre las k diferencias más grandes $X_i - Y_j$ y encontrar las k diferencias más pequeñas.

Para encontrar las diferencias más grandes y más pequeñas, es conveniente ordenar primero cada muestra, de menor a mayor, y luego formar una matriz de diferencias $X_i - Y_j$ usando las X_s como filas y las Y_s como columnas.

La k-ésima diferencia más grande es el límite superior U y la $k - ésima$ diferencia más pequeña es el límite inferior L.

Entonces el intervalo de confianza es obtenido por:

$$\mathbf{P}[L \leq \mathbf{E}(X) - \mathbf{E}(Y) \leq U] \geq 1 - \alpha$$

Ahora haremos un ejemplo para ilustrar la teoría:

8.4. Ejemplo

Se desea mezclar una masa de pastel hasta que se alcance una consistencia específica. Se mezclan cinco lotes de la mezcla usando la batidora A, y los otros cinco lotes se mezclan usando la batidora B. Los tiempos requeridos para mezclar se dan de la siguinte manera (en minutos):

Batidora A	**Batidora A**
7.3	7.4
6.9	6.8
7.2	6.9
7.8	6.7
7.2	7.1

Se busca un intervalo de confianza del 95 % para la diferencia de medias en los tiempos de mezcla, más específicamente para $\mathbf{E}(X) - \mathbf{E}(Y)$, donde X se refiere a la *batidora A* y Y se refiere a la *batidora B*.

Paso 1

Encontrar cuantil $\alpha/2$ ($\omega_{\alpha/2}$), para nuestro ejemplo n=5, m=5, $\alpha = 0.05$, y buscando en la **Tabla Mann-Whitney** tenemos $\omega_{\alpha/2} = \omega_{0.025} = 18$.

Paso 2

Calculamos k, dado por:

$$k = \omega_{\alpha/2} - n(n+1)/2.$$

$$k = 18 - (5)(6)/2 = 3.$$

Paso 3

Ordenamos las muestras de menor a mayor, las X_s las usaremos como filas y las Y_s las usaremos como columnas para formar la matriz de diferencias $X_i - Y_j$:

X_i Y_j	6.7	6.8	6.9	7.1	7.4
6.9	0.2	0.1	0.0	−0.2	−0.5
7.2	0.5	0.4	0.3	0.1	−0.2
7.2	0.5	0.4	0.3	0.1	−0.2
7.3	0.6	0.5	0.4	0.2	−0.1
7.8	1.1	1.0	0.9	0.7	0.4

Paso 4

Entonces las más grandes y las más pequeñas diferencias son encontradas:

Diferencias Pequeñas	Diferencias Grandes
$6.9 - 7.4 = -0.5$	$7.8 - 6.7 = 1.1$
$6.9 - 7.1 = -0.2$	$7.8 - 6.8 = 1.0$
$7.2 - 7.4 = -0.2 = \mathbf{L}$	$7.8 - 6.9 = 0.9 = \mathbf{U}$

Paso 5

El intervalo de confianza del 95 % (L, U) para la diferencia de medias, es decir, $\mathbf{E}(X) - \mathbf{E}(Y)$ es $(-0.2, 0.9)$.

Capítulo 9

Prueba de Kruskal-Wallis

Se obtienen k $m.a.$ independientes de k distintas poblaciones ($k \geq 2$) y queremos probar la hipótesis nula de que todas las poblaciones tienen la misma distribución contra la alternativa de que algunas de las poblaciones tienden a tener distribución distinta. Un caso particular, si $k = 2$ se tiene la prueba de **Mann-Whitney**.

9.1. Datos

Se tienen k $m.a.$ que pueden tener incluso distintos tamaños, la $i-ésima$ $m.a.$ de tamaño n_i es:

$$X_{i_1},\ X_{i_2},\ \cdots,\ X_{i_{ni}}\ \ (m.a.\ \text{de la}\ \ i-ésima\ \ \text{población})$$

Así:

$Muestra\ 1$	$Muestra\ 2$	$Muestra\ 3$	...	$Muestra\ k$
x_{11}	x_{21}	x_{31}	...	x_{k1}
x_{12}	x_{22}	x_{32}	...	x_{k2}
...	...	...	$\ddots$	...
x_{1n_1}	x_{2n_2}	x_{3n_3}	...	x_{kn_k}

sea N el número total de observaciones:

$$N = \sum_{i=1}^{k} n_i$$

9.2. Supuestos

1) Todas las muestras son muestras aleatorias de sus respectivas poblaciones.
2) Además de la independencia dentro de las muestras, suponemos que hay independencia entre las muestras (es decir, las poblaciones son independientes).
3) Las k poblaciones son idénticas o algunas de las poblaciones tienden a tener valores más grandes que las otras poblaciones.
4) La escala de medida es al menos ordinal.

9.3. Hipótesis

$\mathbf{H}_0$: Las funciones de distribución de las k poblaciones son idénticas.

vs

$\mathbf{H}_a$: Al menos una de las poblaciones tiende a tener valores mayores que

al menos una de las poblaciones.

o equivalentemente :

$\mathbf{H}_a$: Las k poblaciones no tienen medias o medianas idénticas.

Asignación de Rangos

- Asignamos el rango 1 a la observación más pequeña de las N observaciones, el rango 2 a la $2da$ más pequeña, y así sucesivamente hasta llegar a la observación mayor que recibirá el rango N.
- Sea $R(X_{ij})$ el rango asignado a la observación X_{ij}
- Sea R_i la suma de los rangos asignados a la $i-ésima$ muestra (la $i-ésima$ columna):

$$R_i = \sum_{j=1}^{n_i} R(X_{ij}), \quad i = 1, 2, \cdots, k$$

NOTA

Se calcula R_i para cada muestra, si hay observaciones repetidas, se le asigna el promedio de los rangos de las observaciones repetidas.

$Rangos$ $Muestra\ 1$	$Rangos$ $Muestra\ 2$	$Rangos$ $Muestra\ 3$	$\cdots$	$Rangos$ $Muestra\ k$
$R(X_{11})$	$R(X_{21})$	$R(X_{31})$	$\cdots$	$R(X_{k_1})$
$R(X_{12})$	$R(X_{22})$	$R(X_{31})$	$\cdots$	$R(X_{k_2})$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$R(X_{1n_1})$	$R(X_{2n_2})$	$R(X_{3n_3})$	$\cdots$	$R(X_{kn_k})$
R_1	R_2	R_3	$\cdots$	R_k

9.4. Estadístico de prueba

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right)$$

donde N y R_i son como los definimos anteriormente y donde:

$$S^2 = \frac{1}{N-1}\left(\sum_{todos\ los\ rangos} R(X_{ij})^2 - N\frac{(N+1)^2}{4}\right)$$

* **Si no hay empates** S^2 se simplifica a $N(N+1)/12$, y la estadística de prueba se reduce a:

$$T = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

9.5. Regla de decisión

Rechazamos H_0 al nivel de significancia α si $T > \omega_{1-\alpha}$, donde $\omega_{1-\alpha}$ se obtiene de la tabla de cuantiles de la distribución de T (tablas Kruskal-Wallis). En caso de no encontrarse en la tabla, los tamaños de muestra correspondientes se utiliza la aproximación Ji-cuadrada con $k-1$ grados de libertad para el estadístico de prueba.

Es decir, rechazamos H_0 a un nivel α si $T > \chi^2_{(k-1)}(1-\alpha)$

y calculamos el $p-value$ de la siguiente manera el $p-value$ es aproximadamente la probabilidad de una variable aleatoria chi-cuadrado con $k-1$ grados de libertad que excede el valor observado de T.

Múltiples comparaciones

Si y sólo si la hipótesis nula es rechazada, podríamos utilizar el siguiente procedimiento para determinar cuáles pares de poblaciones tienden a ser diferentes.

Podemos decir que las poblaciones i y j parecen ser diferentes si se satisface la siguiente desigualdad:

$$\left|\frac{Ri}{ni} - \frac{Rj}{nj}\right| > t_{1-\alpha/2} \left(S^2 \frac{N-1-T}{N-k}\right)^{1/2} \left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}.$$

Donde R_i y R_j son las sumas de los rangos de las dos muestras, $t_{1-(\frac{\alpha}{2})}$ es el cuantil $1-\frac{\alpha}{2}$ de la distribución t obtenida de las tablas de la distribución con $N-k$ grados de libertad. Este procedimiento es repetido para todas las parejas de poblaciones.

Ahora veamos cómo se hace un ejemplo:

9.6. Ejemplo

Se tienen 4 métodos de cultivo de maíz, para ver si hay diferencia entre uno y otro se utilizan terrenos similares y se aplican los distintos métodos y se ve la cantidad de cosecha por m^2 de tierra de cada terreno obtenido.

Paso 1 Prueba a utilizar **Prueba Kruskal-Wallis**

Método 1	**Método 2**	**Método 3**	**Método 4**
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Paso 2 Planteamiento de hipótesis

Queremos probar:

Hipótesis:

$\mathbf{H}_0$: Los 4 Métodos son idénticos.

vs

$\mathbf{H}_a$: Al menos alguno de los métodos tiende a producir cosechas diferentes (mayores o menores) que los otros métodos.

Paso 3 Estadístico de Prueba:

- N = Número total de observaciones

$$T = \frac{1}{S^2} * \left[\sum_{i=1}^{k} \frac{R_i^2}{n_i} - \frac{N * (N+1)^2}{4} \right].$$

- Donde

$$S^2 = \frac{1}{N-1} * \left[\sum_{Todoslosrangos} R(X_{i,j})^2 - N * \frac{(N+1)^2}{4} \right].$$

- Si NO hay empates:

$$S^2 = \frac{N * (N+1)}{12}.$$

y T se reduce a

$$T = \frac{12}{N * (N+1)} * \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3 * (N+1).$$

ahora regresando al problema:

Paso 4 Planteamiento completo:

Método 1	Rango	Método 2	Rango	Método 3	Rango	Método 4	Rango
83	11	91	23	101	34	78	2
91	23	90	19.5	100	33	82	9
94	28.5	81	6.5	91	23	81	6.5
89	17	83	11	93	27	77	1
89	17	84	13.5	96	31.5	79	3
96	31.5	83	11	95	30	81	6.5
91	23	88	15	94	28.5	80	4
92	26	91	23			81	6.5
90	19.5	89	17				
		84	13.5				
	$R_1 = 196.5$		$R_2 = 153$		$R_3 = 207$		$R_4 = 38.5$

Supuestos:

- Muestra aleatoria de tamaño 34.
- k = 4 Número de Métodos.
- $\alpha = 0.05$ Nivel de significancia.
- $N = 34$ Tamaño de la muestra.
- n_1= 9, n_2= 10, n_3= 7, n_4= 8, (número de registros por categoría).

Como hubo empates utilizaremos el estadístico:

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right).$$

donde N y R_i son como los definimos anteriormente y donde:

$$S^2 = \frac{1}{N-1}\left(\sum_{todoslosrangos} R(X_{ij})^2 - N\frac{(N+1)^2}{4}\right).$$

Regresando a nuestro ejemplo:

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right).$$

- donde N y R_i son como los definimos anteriormente y donde:

$$S^2 = \frac{1}{34-1}\left(\left[(11)^2 + (23)^2 + (28.5)^2 + \ldots + (4)^2 + (6.5)^2+\right] - 34\frac{(34+1)^2}{4}\right).$$

$$S^2 = \frac{1}{33}\left[13621.75 - 34\frac{(35)^2}{4}\right].$$

$$S^2 = 97.25.$$

Y nuestro estadístico es:

$$T = \frac{1}{97.25}\left(\left[\frac{196.5^2}{9} + \frac{153^2}{10} + \frac{207}{7} + \frac{38.5^2}{8}\right] - \frac{34(34+1)^2}{4}\right).$$

$$T = \frac{1}{97.25}\left(12937.71 - 10412.5\right).$$

$$T = \frac{1}{97.25}\left(2525.216964\right).$$

$$T = 25.46$$

Paso 5 Regla de decisión.

- Si $k = 3$ y $n_i \leq 5 \cdots \forall i$ y NO HAY EMPATES, usar tablas Kruskal-Wallis y rechazo H_0 a un nivel de significancia α si $T > \omega^{1-\alpha}$ donde el cuantil se busca en tablas Kruskal-Wallis, en caso de no encontrar el valor se ocupa:
- En general, rechazamos H_0 a un nivel de significancia α si $T > \chi^2_{k-1}(1-\alpha)$.

Por otro lado calculamos:

$$\chi^2_{(3)}(1-0.05) = 7.815.$$

Por lo que $T = 25.46 > 7.815 = \chi^2_{(3)}(0.95)$, entonces rechazo H_0.

Ahora calculamos el $p - value$:

$P[X > T] = 1 - P[X \leq 25.46] = 0.0000123 < \alpha$.

Por lo tanto rechazo H_0 a un nivel de significancia de $\alpha = 0.05$.

Paso 6 Conclusión.

Entonces existe evidencia suficiente para decir que al menos alguno de los métodos tiende a producir cosechas diferentes (mayores o menores) que los otros métodos.

9.7. Ejemplo en R-Studio

Calculamos el estadístico de prueba T

```
##Prueba Kruskal-Wallis

Metod1=c(83,91,94,89,89,96,91,92,90)
Metod2=c(91,90,81,83,84,83,88,91,89,84)
Metod3=c(101,100,91,93,96,95,94)
Metod4=c(78,82,81,77,79,81,80,81)

calif=list(g1=Metod1,g2=Metod2,g3=Metod3,g4=Metod4)

n1=length(Metod1)
n2=length(Metod2)
n3=length(Metod3)
n4=length(Metod4)
N=n1+n2+n3+n4
rangos=rank(c(Metod1,Metod2,Metod3,Metod4))
rangos2=rangos^2
R1=sum(rangos[1:9])
R2=sum(rangos[10:19])
R3=sum(rangos[20:26])
R4=sum(rangos[27:34])

S2=((N*(N+1))/12)
T=(1/S2)*((R1^2/n1)+(R2^2/n2)+(R3^2/n3)+(R4^2/n4))-3*(N+1)
T
```

```
[1] 25.46437
```

Comparamos con el cuantil

```
qchisq(0.950,3)
```

```
[1] 7.814728
```

Utilizando la función de R

```
kruskal.test(calif)
```

```
    Kruskal-Wallis rank sum test

data:  calif
Kruskal-Wallis chi-squared = 25.629, df = 3, p-value = 1.141e-05
```

El valor de la estadísitca de prueba es 25.62 y su correspondiente $p - value$ es mucho menor a 0.01 ($\alpha = 1\,\%$) por lo tanto se rechaza la hipótesis nula y se concluye que al menos uno de los 4 métodos tiene distribución diferente.

Para determinar qué método es el distinto se tendrán que hacer las comparaciones dos a dos.

```
Rendimiento=c(83,91,94,89,89,96,91,92,90,91,90,81,83,84,83,88,91,89,84,101,100,91,
              93,96,95,94,78,82,81,77,79,81,80,81)
Metodo=c(1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4,4,4)
pairwise.wilcox.test(Rendimiento,Metodo)
```

```
    Pairwise comparisons using Wilcoxon rank sum test

data:  Rendimiento and Metodo
```

```
  1      2      3
2 0.0395 -      -
3 0.0385 0.0047 -
4 0.0036 0.0047 0.0047

P value adjustment method: holm
```

En este caso observamos que todos los $p-values$ de las comparaciones dos a dos son menores a 0.05 por lo tanto con un α del 5 % podemos concluir que todos los métodos tienen distribuciones de los rendimientos distintas.

9.8. Ejercicios

1. Una muestra aleatoria de 5 diferentes marcas de focos son probados para medir la duración del foco, y los resultados fueron los siguientes:

A	B	C	D	E
73	84	82	80	85
64	80	79	85	82
67	81	71	82	80
62	77	75	86	
70		80		

¿Los datos observados indican una diferencia significativa entre las marcas? De haber diferencia, ¿qué marcas parecen ser las diferentes? Use $\alpha = 10\,\%$

2. Tres maestros desean comparar las claificaciones que pusieron en el semestre pasado para ver si alguno tiende a dar calificaciones diferentes que los otros.

Calificación	A	B	C
10	4	10	6
9	14	6	7
8	17	9	8
7	6	7	6
6	2	6	1

Con los datos observados se puede corroborar la sospecha de que alguno de los maestros da calificaciones diferentes a los otros? Use $\alpha = 5\,\%$

3. El Hospital "Los Ángeles" realizó un estudio en 4 Dietas a un grupo de 20 individuos; cada Dieta se aplicó aleatoriamente a cada individio. Se registró la pérdida de peso en kg. Realiza la prueba correspondiente para poder concluir si, ¿Hay diferencia significativa entre la efectividad de las Dietas?

Dieta 1	**Dieta 2**	**Dieta 3**	**Dieta 4**
6.1	5	7.6	6.2
4.3	5.6	6.8	8
4.5	7.3	3.9	7.4
2.4	5.7	7.9	4.6
9.1	2.1	5.9	7

Capítulo 10

Prueba de Igualdad de Varianzas

Usualmente para comparar varias poblaciones nos basamos en las medias o varianzas de las poblaciones, en algunas situaciones las varianzas podrían ser nuestro campo de interés. Por ejemplo, se ha afirmado que el efecto de sembrar nubes con yoduro de plata podría aumentar la varianza de la lluvia resultante.

Prueba de Igualdad de Varianzas para 2 poblaciones

10.1. Datos

Consiste en dos muestras aleatorias

Sea $x_1, x_2, x_3, \cdots, x_n$ una muestra aleatoria de tamaño n de la *población* 1.

Sea $y_1, y_2, y_3, \cdots, y_m$ una muestra aleatoria de tamaño m de la *poblacion* 2.

10.2. Supuestos

1) Ambas muestras son aleatorias de su respectiva población.
2) Además de la independencia dentro de cada muestra, existe una independencia mutua entre las dos muestras.
3) La escala de medida es al menos intervalo.

Asignación de Rangos

Modificaremos cada X_i y Y_j con el valor absoluto de la desviación de la media utilizando:

$$U_i = |X_i - \mu_1| \ , \quad i = 1, \cdots, n.$$

$$y$$

$$V_j = |Y_j - \mu_2| \ , \quad j = 1, \cdots, m.$$

Donde μ_1 y μ_2 son las medias de las poblaciones 1 y 2 respectivamente.

Si μ_1 y μ_2 son desconocidas, se usará $\overline{X}$ para μ_1 y $\overline{Y}$ para μ_2.

Asignamos los rangos 1 al $n+m$ a la muestra combinada de U y V de la manera habitual. Si varios valores de U y/o V son exactamente iguales entre sí (empatados), asigne a cada uno el promedio de los rangos que se les habrían asignado si no hubiera habido empates.

Sean $R(U_i)$ y $R(V_j)$ los rangos así asignados.

10.3. Estadístico de Prueba

Se utilizará el siguiente estadístico de prueba:

- **Si no hay empates**

Si no hay empates de la población U con la población V ocupamos:

$$T = \sum_{i=1}^{n} [R(U_i)]^2.$$

La suma de los cuadrados de los rangos asignados a la *población* 1.

- **Si hay empates**

$$T_1 = \frac{T - n\overline{R^2}}{\left[\frac{nm}{N(N-1)} \sum_{i=1}^{N} R_i^4 - \frac{nm}{N-1}(\overline{R^2})^2\right]^{\frac{1}{2}}}.$$

Donde $N = n + m$, $\overline{R^2}$ representa el promedio de los cuadrados de los rangos de ambas muestras combinadas:

$$\overline{R^2} = \frac{1}{N}\left(\sum_{i=1}^{n}[R(U_i)]^2 + \sum_{j=1}^{m}[R(V_j)]^2\right).$$

y $\sum R_i^4$ representa la suma de los rangos elevados a la cuarta potencia:

$$\sum_{i=1}^{N} R_i^4 = \sum_{i=1}^{n}[R(U_i)]^4 + \sum_{j=1}^{m}[R(V_j)]^4.$$

Los cuantiles exactos de la distribución nula de T se obtienen de las tablas correspondientes de la prueba para los casos de no tener empates y $n \leq 10$, $m \leq 10$.

Para muestras de tamaño mayor a 10 utilizamos la siguiente aproximación, basada en la estandarización normal de cuantiles z_p, obtenidas de la tabla de la distribución normal, puede ser usado la aproximación obtenida de cuantiles ω_p para T,

$$\omega_p = \frac{n(N+1)(2N+1)}{6} + z_p\sqrt{\frac{mn(N+1)(2N+1)(8N+11)}{180}}.$$

Donde $N = n + m$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

10.4. Hipótesis

Caso A Prueba de dos colas

$\mathbf{H}_0$: X e Y son identicamente distribuidas, excepto por medias diferentes.

$$vs$$

$$\mathbf{H}_a: \ Var(x) \neq Var(Y).$$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$$T > t_{1-\frac{\alpha}{2}} \quad o \quad T < t_{\frac{\alpha}{2}}.$$

Rechazamos H_0 al nivel de significancia α si T (o T_1 en caso de tener empates) es mayor que el cuantil $1-\frac{\alpha}{2}$ o menor al cuantil $\frac{\alpha}{2}$ encontrados en las tablas correspodientes a la prueba (tablas de rangos al cuadrado). Y en el caso de T_1 en la tabla de distribución normal estándar.

Y calculamos el $p-value$:

$$p-value = 2 * (el\ menor\ p-value\ de\ una\ cola).$$

donde

el $p-value$ de la cola inferior es aproximadamente:

$$p-value = \mathbf{P}\left[Z \leq \frac{T - n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}}\right].$$

y el $p-value$ de la cola superior es aproximadamente:

$$p-value = \mathbf{P}\left[Z \geq \frac{T - n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}}\right].$$

Caso B Prueba de cola inferior

$$\mathbf{H}_0: \ X \text{ e } Y \text{ son identicamente distribuidas, excepto por medias diferentes.}$$

$$vs$$

$$\mathbf{H}_a: \ Var(x) < Var(Y).$$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$T < t_\alpha$ (o T_1 en caso de tener empates) donde, t_α es el cuantil que se busca en tablas de la prueba. Y en el caso de T_1 en la tabla de distribución normal estándar.

y el $p-value$ se calcula:

$$p-value = \mathbf{P}\left[Z \leq \frac{T - n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}}\right].$$

Caso C Prueba de cola superior

$$\mathbf{H}_0 : \ X \text{ e } Y \text{ son identicamente distribuidas, excepto por medias diferentes.}$$

$$vs$$

$$\mathbf{H}_a : \ Var(x) > Var(Y).$$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$T > t_{1-\alpha}$(o T_1 en caso de tener empates) donde, $t_{1-\alpha}$ es el cuantil que se busca en tablas de la prueba. Y en el caso de T_1 en la tabla de distribución normal estándar.

y el $p - value$ se calcula:

$$p - value = \mathbf{P}\left[Z \geq \frac{T - n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}}\right].$$

Ahora vamos a resolver un ejemplo.

10.5. Ejemplo

Una cuenca hidrográfica particular se ha construido extensivamente en los últimos años, con desarrollos de vivienda, represas, etc. una muestra aleatoria de las tasas de flujo de la corriente (pies cúbicos por minuto) para una corriente en esa cuenca hidrográfica se compara con una muestra de las tasas de tiempos anteriores para ver si la variabilidad ha cambiado.

Tasas Actuales	**Tasas Pasadas**
32	39
36	21
41	58
27	46
35	30
48	22
31	17
28	19

¿Es significativa la diferencia en las varianzas? Utiliza α=0.05

Paso 1 Prueba a utilizar **Prueba de Igualdad de Varianzas**

Paso 2 Formulamos las hipótesis

$$\mathbf{H}_0 : \text{Las tasas actuales y las tasas pasadas son idénticamente distribuidas.}$$

$$vs$$

$$\mathbf{H}_a : \ Var(\text{tasas pasadas}) \ \neq \ Var(\text{tasas actuales})$$

Paso 4 Procedimiento completo

Primero como μ_1 y μ_2 son desconocidas se usará $\overline{X}$ para μ_1 y $\overline{Y}$ para μ_2.

Medidas Tasas Actuales **(X)**	Originales Tasas Pasadas **(Y)**	Desviación Actuales **(U)**	Absoluta Pasadas **(V)**	Rangos Actuales	Rangos Pasadas	Rangos Actuales	Al Cuadrado Pasadas
32	39	2.75	7.5	4	8	16	64
36	21	1.25	10.5	2	11	4	121
41	58	6.25	26.5	6	16	36	256
27	46	7.75	14.5	9	14.5(Empate)	81	210.25
35	30	0.25	1.5	1	3	1	9
48	22	13.25	9.5	13	10	169	100
31	17	3.75	14.5	5	14.5(Empate)	25	210.25
28	19	6.75	12.5	7	12	49	144
$\overline{X} = 34.75$	$\overline{Y} = 31.5$					$T = 381$	

T= Sumas del cuadrado de los rangos (Actuales) = 381

$\overline{R^2} = \frac{1}{16}(16 + 4 + 36 + \cdots + 210.25 + 144)$=93.46

$\sum_{i=1}^{N} R_i^4 = (16)^2 + (4)^2 + \cdots + (210.25)^2 + (144)^2$= 243217.125

Paso 3 Estadístico de Prueba

Debido a que encontramos empates utilizaremos el estadístico T_1:

$$T_1 = \frac{T - n\overline{R^2}}{\left[\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^4 - \frac{nm}{N-1}(\overline{R^2})^2\right]^{\frac{1}{2}}}$$

Entonces tenemos:

$$T_1 = \frac{381 - 8(93.46)}{\left[\frac{(8)(8)}{16(15)}(243217.125) - \frac{(8)(8)}{15}(93.46)^2\right]^{\frac{1}{2}}}$$

$$T_1 = \frac{381 - 747.68}{\left[\frac{64}{240}(243217.125) - \frac{64}{15}(93.46)^2\right]^{\frac{1}{2}}}$$

$$T_1 = \frac{-366.68}{[64857.9 - 37268.35]^{\frac{1}{2}}}$$

$$T_1 = \frac{-366.68}{\sqrt{27589.55}}$$

$$T_1 = \frac{-366.68}{166.10}$$

$$T_1 = -2.208282$$

Paso 5 Regla de Decisión

Rechazamos H_0 al nivel de significancia α si:

$$T_1 > t_{1-\frac{\alpha}{2}} \qquad o \qquad T_1 < t_{\frac{\alpha}{2}}$$

Rechazamos H_0 al nivel de significancia α si T_1 (ya que tuvimos empates) es mayor que el cuantil $1 - \frac{\alpha}{2}$ o menor al cuantil $\frac{\alpha}{2}$ encontrados en las tablas correspodientes a la distribución normal estándar.

Regresando a nuestro ejemplo:

$$T_1 > t_{0.975} \quad o \quad T_1 < t_{.025}$$

$$-2.2082 \not> 1.96 \quad o \quad -2.2082 < -1.96$$

$\therefore$ Entonces Rechazo H_0.

Paso 6 Conclusión

Entonces podemos concluir que existe evidencia estadística suficiente para decir que existe diferencia significativa en las varianzas.

10.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
# Datos
TasasActuales = c(32,36,41,27,35,48,31,28)
TasasPasadas = c(39,21,58,46,30,22,17,19)

n=length(TasasActuales)  #Tamaño de TasasActuales
m=length(TasasPasadas)   #Tamaño de TasasPasadas
N=n+m                    #Tamaño Total

# Matriz de TasasActuales y TasasPasadas
datos=matrix(c(TasasActuales,TasasPasadas), byrow=FALSE, ncol=2)
colnames(datos)=c("TasasActuales: X", "TasasPasadas: Y")
datos
```

```
     TasasActuales: X TasasPasadas: Y
[1,]               32              39
[2,]               36              21
[3,]               41              58
[4,]               27              46
[5,]               35              30
[6,]               48              22
[7,]               31              17
[8,]               28              19
```

```
alpha=0.05  #Nivel de significancia

Xbarra=mean(TasasActuales)
cat("Media muestral de X: ", Xbarra, "\n")
```

```
Media muestral de X:  34.75
```

```
Ybarra=mean(TasasPasadas)
cat("Media muestral de Y: ", Ybarra, "\n")
```

```
Media muestral de Y:  31.5
```

```
# Valores absolutos de la desviacion de la media de X
desvX= c()
for(i in 1:n) {
  desvX[i]=abs(TasasActuales[i] - Xbarra)
}

# Valores absolutos de la desviacion de la media de Y
desvY = c()
for(i in 1:n) {
```

```
  desvY[i] = abs(TasasPasadas[i] - Ybarra)
}

# Creamos la matriz U, V
desviacion = matrix(c(desvX,desvY), byrow= FALSE, ncol=2)
colnames(desviacion) = c("U","V")
desviacion
```

```
         U    V
[1,]  2.75  7.5
[2,]  1.25 10.5
[3,]  6.25 26.5
[4,]  7.75 14.5
[5,]  0.25  1.5
[6,] 13.25  9.5
[7,]  3.75 14.5
[8,]  6.75 12.5
```

```
# Asignamos los rangos
rang = rank(desviacion)
#rang

# Los acomodamos en una nueva matriz U, V
rangos = matrix(rang, byrow=FALSE, ncol=2)
colnames(rangos) = c("Rangos U", "Rangos V")
rangos
```

```
     Rangos U Rangos V
[1,]        4      8.0
[2,]        2     11.0
[3,]        6     16.0
[4,]        9     14.5
[5,]        1      3.0
[6,]       13     10.0
[7,]        5     14.5
[8,]        7     12.0
```

```
# Rangos de U al cuadrado
rangosU2 = c()
for(i in 1:n) {
  rangosU2[i] = (rangos[i,1])^2
}
#rangosU2

# Rangos de V al cuadrado
rangosV2 = c()
for(i in 1:n) {
  rangosV2[i] = (rangos[i,2])^2
}
#rangosV2

# Suma de los rangos al cuadrado de U
sumarangosU2 = sum(rangosU2)

# Suma de los rangos al cuadrado de V
sumarangosV2 = sum(rangosV2)

# Nueva matriz con los rangos^2
rangos2 = matrix(c(rangosU2,rangosV2), byrow=FALSE, ncol=2)
```

```
colnames(rangos2) = c("Rangos al cuadrado U", "Rangos al cuadrado V")
rangos2
```

```
     Rangos al cuadrado U Rangos al cuadrado V
[1,]                   16                64.00
[2,]                    4               121.00
[3,]                   36               256.00
[4,]                   81               210.25
[5,]                    1                 9.00
[6,]                  169               100.00
[7,]                   25               210.25
[8,]                   49               144.00
```

```
#Promedio de los rangos^2
promrangos2 = (sumarangosU2+ sumarangosV2)/N

#Rangos de U^4
rangosU4 = c()
for(i in 1:n) {
  rangosU4[i] = (rangosU2[i])^2
}
#rangosU4

#Rangos de V^4
rangosV4 = c()
for(i in 1:n) {
  rangosV4[i] = (rangosV2[i])^2
}
#rangosV4

#Nueva matriz con los rangos^4
rangos4 = matrix(c(rangosU4,rangosV4), byrow=FALSE, ncol=2)
colnames(rangos4) = c("Rangos a la cuarta U", "Rangos a la cuarta V")
rangos4
```

```
     Rangos a la cuarta U Rangos a la cuarta V
[1,]                  256              4096.00
[2,]                   16             14641.00
[3,]                 1296             65536.00
[4,]                 6561             44205.06
[5,]                    1                81.00
[6,]                28561             10000.00
[7,]                  625             44205.06
[8,]                 2401             20736.00
```

```
#Suma de los rangos a la cuarta^4
sumarangos4 = sum(rangosU4) + sum(rangosV4)

#Calculamos el estadístico de prueba T1 porque tenemos empates
T1=(sumarangosU2 - (n*promrangos2))/(sqrt(((      ((n^2)/(N-1))/N)*sumarangos4 )-(((n^2)/(N-1))*(pro

cat("T1 = ", T1, "\n")
```

```
T1 =  -2.208273
```

```
#Regla de decision
#cat("Rechazamos H_0 a un nivel de significancia alpha=", alpha, "si T1>t1 o T1<t2, en donde t1 y t2

# Calculamos los cuantiles t1 y t2
t1 = qnorm(1-(alpha/2),mean=0,sd=1)
```

```
t2 = qnorm(alpha/2,mean=0,sd=1)

#cat("t1 = ", t1, "\n")
#cat("t2 = ", t2, "\n")

if((T1>t1) || (T1<t2)) {
  print("Rechazamos H_0")
} else {
  print("No rechazamos H_0")
}
```

```
[1] "Rechazamos H_0"
```

Ya que utilizamos la estadística T_1, entonces el cuantil que hay que buscar es en una distribución normal. Con $\alpha = 0.05$, hace que $\frac{\alpha}{2} = 0.025$ y $1 - \frac{\alpha}{2} = 0.975$

```
t1=qnorm(.025,0,1)
t1
```

```
[1] -1.959964
```

```
t2=qnorm(.975,0,1)
t2
```

```
[1] 1.959964
```

Como $T_1 = -2.21 \leq -1.96 = t_1$ entonces se cumple la regla de decisión y rechazamos H_0 a un nivel $\alpha = 5\,\%$. Concluimos que existe evidencia suficiente para decir que las varianzas de las tasas de flujo de la corriente son diferentes.

```
pvalue=2*min(pnorm(T1,0,1),(1-pnorm(T1,0,1)))
pvalue
```

```
[1] 0.02722523
```

10.7. Ejercicios

1. Un banco de sangre mantuvo un registro de la frecuencia cardíaca de varios donadores de sangre.

Hombres	Mujeres
58	66
76	74
82	69
74	76
79	72
65	73
74	75
86	67
	68

¿Es la variación entre los hombres significativamente mayor que la variación entre las mujeres? nivel de significancia 5 %

2. Se desea probar que las variaciones de las temperaturas altas en Des Moines son mayores que las variaciones de las temperaturas altas en Spokane, para ello se tomó una muestra de las temperaturas altas diarias durante el verano. Use nivel de significancia 10 %

Des Moines	Spokane
83	78
91	82
94	81
89	77
89	79
96	81
91	80
92	81
82	79
93	80
90	
93	

Capítulo 11

Prueba para más de dos Muestras

Si hay tres o más muestras, esta prueba se modifica fácilmente para probar la igualdad de varianzas. De cada observación se resta su media poblacional (o su media muestral ($\overline{x}$) cuando μ_i es desconocida) y convertimos el signo de la diferencia resultante a " + ", como se acaba de describir para ambas muestras. Los rangos se asignan de menor a mayor y asignamos el promedio de los rangos correspondientes en caso de empate.

Calculamos la suma del cuadrado de los rangos asignados de cada muestra, denotamos por:

$$S_1, S_2, \cdots, S_k \text{ la suma de los rangos de cada muestra.}$$

11.1. Datos

Se tienen k m.a. que pueden tener distintos tamaños, la $j-$ésima m.a, de tamaño n_j es:

$$x_{j1},\ x_{j2},\ x_{j3},\ \cdots,\ x_{jn_j}.$$

Así.

$Muestra\ 1$	$Muestra\ 2$	$Muestra\ 3$	...	$Muestra\ k$
x_{11}	x_{21}	x_{31}	...	x_{k1}
x_{12}	x_{22}	x_{32}	...	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_{1n_1}	x_{2n_2}	x_{3n_3}	...	x_{kn_k}

Sea N el número total de observaciones:

$$N = \sum_{j=1}^{k} n_j.$$

11.2. Hipótesis.

$\mathbf{H}_0$: Las k poblaciones son identicas, excepto por diferencia en las medias.

$$vs$$

$\mathbf{H}_a$: Algunas de las varianzas de las poblaciones no son idénticas entre sí.

11.3. Estadístico de Prueba

El estadístico de prueba que vamos a utilizar es:

- **Cuando tenemos empates**

$$T_2 = \frac{1}{D^2}\left[\sum_{j=1}^{k}\frac{S_j^2}{n_j} - N(\overline{S})^2\right],$$

donde:

- n_j= Número de observaciones en la muestra j.
- $N = n_1 + n_2 + \cdots + n_k$.
- S_j= la suma de los cuadrados de los rangos en la muestra j.
- $\overline{S} = \frac{1}{N}\sum_{j=1}^{k} S_j$ Es el promedio de los cuadrados de todos los rangos.
- $D^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} R_j^4 - N(\overline{S})^2\right]$ y $\sum R_i^4$ representa la suma de los rangos después de elevarlos a la cuarta potencia.
- **Cuando no tenemos empates**

D^2 y $\overline{S}$ se simplifican:

$$D^2 = N(N+1)(2N+1)(8N+11)/180.$$

$$\overline{S} = (N+1)(2N+1)/6.$$

11.4. Regla de decisión

Rechazamos H_0 si:

$$T_2 > t_{1-\alpha}.$$

donde $t_{1-\alpha}$ es el cuantil de la distribución χ^2 con $k-1$ grados de libertad donde el cuantil se localiza en las tablas de dicha distribución.

y calculamos el $p - value$

$$p - value = \mathbf{P}\left[\chi^2 > T_2\right].$$

es la probabilidad de una variable aleatoria χ^2 con $k-1$ grados de libertad sea mayor a el valor observado de T_2.

11.5. Comparación múltiple

Si la hipótesis nula es rechazada, podríamos utilizar el siguiente procediemiento para determinar cuáles pares de poblaciones tienden a ser diferentes.

Podemos decir que las poblaciones i y j parecen ser diferentes si se satisface la siguiente desigualdad:

$$\left|\frac{Si}{ni} - \frac{Sj}{nj}\right| > t_{1-\alpha/2}\left(D^2\frac{N-1-T_2}{N-k}\right)^{1/2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)^{1/2}.$$

Donde $t_{1-\alpha/2}$ es el cuantil $1-\alpha/2$ de la distribución t obtenido en tablas de dicha distribución con $N-k$ grados de libertad.

Ahora resolveremos un ejercicio.

11.6. Ejemplo

Retomando en ejemplo de los 4 métodos de cultivo de maíz, probaremos ahora si las varianzas de la cantidad de cosecha por m^2 de tierra de cada terreno son diferentes.

Método 1	**Método 2**	**Método 3**	**Método 4**
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Paso 1 Prueba a utilizar **Prueba de Igualdad de Varianzas para más de dos muestras.**

Paso 2 Formulamos las hipótesis:

$$\mathbf{H}_0: \text{ Los 4 métodos son idénticos, excepto por diferencias en las medias.}$$

$$vs$$

$$\mathbf{H}_a: \text{ Al menos alguno de los métodos tiene varianzas distintas a al menos alguno de los otros métodos.}$$

Paso 3 Estadístico de Prueba.

Cuando tenemos empates

$$T_2 = \frac{1}{D^2}\left[\sum_{j=1}^{k}\frac{S_j^2}{n_j} - N(\overline{S})^2\right],$$

donde:

- n_j= Número de observaciones en la muestra j.
- $N = n_1 + n_2 + \cdots + n_k$.
- S_j= la suma de los cuadrados de los rangos en la muestra j.
- $\overline{S} = \frac{1}{N}\sum_{j=1}^{k} S_j$ Es el promedio de los cuadrados de todos los rangos.
- $D^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} R_j^4 - N(\overline{S})^2\right]$ y $\sum R_i^4$ representa la suma de los rangos después de elevarlos a la cuarta potencia.

$$T_2 = 6.006.$$

Paso 5 Regla de Decisión.

Rechazamos H_0 al nivel de significancia α si:

$$T_2 > t_{1-\alpha}.$$

donde $t_{1-\alpha}$ es el cuantil de la distribución χ^2 con $k-1$ grados de libertad donde el cuantil se localiza en las tablas de dicha distribución.

Ya que utilizamos la estadística T_2, entonces el cuantil que hay que buscar es en una distribución Ji-cuadrada con $k-1=3$ grados de libertad. Consideremos $\alpha = 0.01$.

$$t_{1-\alpha} = 11.334.$$

Como $T_2 = 6.006 \not> 11.344 = t_{1-\alpha}$ entonces no se cumple la regla de decisión.

$\therefore$Entonces no rechazo H_0.

Paso 6 Conclusión.

Conlcuimos que no hay evidencia suficiente para decir que las distribuciones de los rendimientos de los 4 métodos de cultivo no sean idénticos, excepto por diferencias en las medias.

11.7. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
M1=c(83,91,94,89,89,96,91,92,90)
M2=c(91,90,81,83,84,83,88,91,89,84)
M3=c(101,100,91,93,96,95,94)
M4=c(78,82,81,77,79,81,80,81)
n1=length(M1)
n2=length(M2)
n3=length(M3)
n4=length(M4)
N=n1+n2+n3+n4

U1=sort(abs(M1-mean(M1)))
U2=sort(abs(M2-mean(M2)))
U3=sort(abs(M3-mean(M3)))
U4=sort(abs(M4-mean(M4)))
#U=rbind(U1,U2,U3,U4)
#Asignación de rango
R1=c(3.5,3.5,5,11,12.5,12.5,25,33,34)
R2=c(14,18.5,18.5,20,23.5,23.5,26,28.5,28.5,32)
R3=c(2,6,15,21,27,30,31)
R4=c(1,7,9,9,9,16,17,22)
#rangos al cuadrado
R1s=R1^2
R2s=R2^2
R3s=R3^2
R4s=R4^2
#suma del cuadrado de los rangos
S1=sum(R1s)
S2=sum(R2s)
S3=sum(R3s)
S4=sum(R4s)

S=sum(S1,S2,S3,S4)/N
R4=sum(R1s^2)+sum(R2s^2)+sum(R3s^2)+sum(R4s^2)
D2=(R4-N*S^2)/(N-1)
```

```
T2=(S1^2/n1+S2^2/n2+S3^2/n3+S4^2/n4-N*S^2)/D2
T2
```

```
[1] 6.006228
```

Ya que utilizamos la estadística T_2, entonces el cuantil que hay que buscar es en una distribución Ji-cuadrada con $k-1=3$ grados de libertad. Consideremos $\alpha = 0.01$.

```
t=qchisq(.99,3)
t
```

```
[1] 11.34487
```

Como $T_2 = 6.006 \not> 11.344 = t_{1-\alpha}$ entonces no se cumple la regla de decisión y no rechazamos H_0. Conlcuimos que no hay evidencia suficiente para decir que las distribuciones de los rendimientos de los 4 métodos de cultivo no sean idénticos, excepto por diferencias en las medias.

##Ejercicios

1. Una muestra aleatoria de 5 diferentes marcas de focos son probados para medir la duración del foco, y los resultados fueron los siguientes:

A	B	C	D	E
73	84	82	80	85
64	80	79	85	82
67	81	71	82	80
62	77	75	86	
70		80		

Los datos observados indican una diferencia significativa entre las varianzas de las duraciones por marca? Use $\alpha = 10\,\%$

Parte IV

Tablas de Contingencia

Introducción

Estas son tablas en las que se muestran las frecuencias de observaciones medidas sobre variables que tienen diferentes clases. El objetivo es observar si dos variables son independientes entre si. Para la realización de esta prueba las frecuencias son anotadas en tablas en las cuales cada observación es categorizada en solo una de las clases.

El uso de las tablas de contingencia fue mencionado previamente en la prueba de signos conocida como **McNemar**.

Una **tabla de contingencia** es un arreglo de números naturales en forma de matriz, donde cada número representa conteos o frecuencias.

Por ejemplo:

	Fuma	**No Fuma**	**Total**
Hombre	11	18	29
Mujer	22	13	35
Total	33	31	64

Queremos saber si influye el género en fumar.

Esta prueba nos ayuda para ver si hay alguna asociación entre 2 variables.

Algunos ejemplos:

- La asociación entre el estado nutricional de un estudiante con su desempeño académico.
- Si la preferencia por un refresco es independiente del sexo del consumidor.
- La asociación entre la región geográfica y la inversión financiera.

NOTA:

Si no hay asociación entre las variables decimos que son independientes.

En el caso de ser dependientes el valor de una variable nos ayudará a determinar el valor de la otra.

De forma general, una tabla de contingencia se construirá con r renglones y c columnas, y se llamará tabla de contingencia de $r \times c$. Éstas tablas de contingencia pueden usarse para presentar una tabulación de los datos contenidos en varias muestras, donde los datos representan al menos una escala de medición nominal.

Capítulo 12

Tablas de Contingencia de 2x2

En general una tabla de contingencia de $r \times c$ es un arreglo de números naturales que tiene r renglones y c columnas y por lo tanto tiene rc celdas o lugares para los números.

En este caso particular, r=2 y c=2, llamadas **Tablas de Contingencia de 2x2**.

Una aplicación de las tablas de contingencia de 2×2 surge cuando N objetos o personas, posiblemente seleccionadas aleatoriamente de alguna población, son clasificadas en una o dos categorías antes de aplicar un tratamiento o se produzca un evento.

Después de aplicar el tratamiento los mismos N objetos son nuevamente examinados y clasificados en las dos categorías. La pregunta a responder es ¿El tratamiento altera significativamente la proporción de objetos en cada una de las dos categorías?

El uso de las tablas de contingencia fue introducido anteriormente, y se vio que el procedimiento estadístico apropiado era una variación de la prueba de signos conocida como **McNemar**.

12.1. Datos

Una muestra aleatoria de n_1 observaciones se extrae de una población (o antes de aplicar un tratamiento) y cada observación se clasifica en la clase 1 o 2, los números totales en las dos clases son O_{11} y O_{12} respectivamente, donde $O_{11} + O_{12} = n_1$.

Una segunda muestra aleatoria de n_2 observaciones se extrae de una segunda población (o la primera población después de aplicar algún tratamiento) y el número de observaciones en la clase 1 o 2 es O_{21} y O_{22} respectivamente, donde $O_{21} + O_{22} = n_2$.

Los datos se organizan en la siguiente tabla de contingencia:

	Clase 1	**Clase 2**	**Total**
Población 1	O_{11}	O_{12}	n_1
Población 2	O_{21}	O_{22}	n_2
Total	C_1	C_2	$N = n_1 + n_2$

El número total de observaciones es denotado por N.

12.2. Supuestos

1) Cada muestra es una muestra aleatoria.

2) Las dos muestras son mutuamente independientes.

3) Cada observación puede clasificarse en *Clase* 1 o *Clase* 2.

12.3. Estadístico de Prueba

Si alguna columna total es cero, el estadístico de prueba es definido así:

$$T_1 = 0.$$

En otro caso:

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}} \sim N(0,1).$$

La distribución exacta de T_1 es difícil debido a todas las diferentes combinaciones de valores posibles para $O_{11}, O_{22}, O_{12}, O_{21}$.

Por lo tanto, se utiliza la aproximación de muestra grande, que es la distribución normal estándar cuyos cuantiles se dan en la tabla correspondiente.

12.4. Hipótesis

Caso A Prueba de dos colas

$$\mathbf{H}_0 : p_1 = p_2,$$

$$vs$$

$$\mathbf{H}_a : p_1 \neq p_2.$$

Donde p_1 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 1 y p_2 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 2.

Regla de Decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 < Z_{\frac{\alpha}{2}} \quad o \quad T_1 > Z_{1-\frac{\alpha}{2}}.$$

Donde el cuantil $Z_{\frac{\alpha}{2}}$ y el cuantil $Z_{1-\frac{\alpha}{2}}$ se buscan en las tablas correspondientes de la distribución Normal.

Ahora calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * min\{\mathbf{P}[Z < T_1], \mathbf{P}[Z > T_1]\},$$

donde Z es una normal estándar.

NOTA

Para la hipótesis anterior, también es usual usar T_1^2 en lugar de T_1 como estadístico de prueba. Entonces la región de rechazo es la cola superior de la distribución χ^2 con 1 grado de libertad, obtenidos en la tabla de la Distribución χ^2.

Caso B Prueba de cola inferior

$$\mathbf{H}_0 : p_1 \geq p_2,$$

$$vs$$

$$\mathbf{H}_a : p_1 < p_2.$$

Donde p_1 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 1 y p_2 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 2.

Regla de Decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 < Z_\alpha.$$

Donde el cuantil Z_α se buscan en las tablas correspondientes de la distribución Normal.

Ahora calcularemos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Z < T_1],$$

donde Z es una normal estándar.

Caso C Prueba de cola superior

$$\mathbf{H}_0 : p_1 \leq p_2,$$

$$vs$$

$$\mathbf{H}_a : p_1 > p_2.$$

Donde p_1 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 1 y p_2 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 2.

Regla de Decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 > Z_{1-\alpha}.$$

Donde el cuantil $Z_{1-\alpha}$ se buscan en las tablas correspondientes de la distribución Normal.

Ahora calcularemos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Z > T_1],$$

donde Z es una normal estándar.

Ahora hagamos un ejercicio.

12.5. Ejemplo

En la academia naval se instaló un nuevo sistema de iluminación en las habitaciones guardamarinas. Se informó que el nuevo sistema de iluminación daba como resultado una vista deficiente debido a la tensión continua en los ojos de los guardias marinos. Se consideró un estudio (ficticio). Para probar la siguiente hipótesis nula. Utilizaremos α=0.05

$\mathbf{H}_0$: La probabilidad de que un guardia marino tenga 20-20 (buena visión) es mayor o igual bajo el nuevo sistema de luces que con el viejo sistema de luces.

vs

$\mathbf{H}_a$: La probabilidad de buena visión es menor con el nuevo sistema de luces que con el viejo sistema de luces.

Paso 1 Prueba a utilizar.

Es una prueba **Tablas de contingencia de 2x2.**

Paso 2 Planteamos las hipótesis.

Es decir si planteamos nuestra hipótesis quedarían así:

Caso C Prueba de cola superior

$$\mathbf{H}_0 : p_1 \leq p_2,$$

vs

$$\mathbf{H}_a : p_1 > p_2.$$

Donde p_1 es la probabilidad de elegir al azar un guardia marino con buena visión bajo el viejo sistema de luces y p_2 es la probabilidad de elegir al azar un guardia marino con buena visión con el nuevo sistema de luces.

Paso 3 La tabla de contingencia, en este caso es:

	Buena Visión	**Mala Visión**	**Total**
Luces viejas	$O_{11} = 714$	$O_{12} = 111$	$n_1 = 825$
Luces nuevas	$O_{21} = 662$	$O_{22} = 154$	$n_2 = 816$
Total	$C_1 = 1376$	$C_2 = 265$	$N = n_1 + n_2 = 1641$

Paso 4 Estadístico de Prueba,

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}.$$

$$T_1 = \frac{\sqrt{1641}[(714)(154) - (111)(662)]}{\sqrt{(825)(816)(1376)(265)}}.$$

$$T_1 = 2.9821.$$

Paso 5 Regla de decisión.

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 > Z_{1-\alpha}.$$

$$T_1 = 2.9821 > Z_{0.95} = 1.6449.$$

Donde el cuantil $Z_{1-\alpha}$ se buscan en las tablas correspondientes de la distribución Normal.

$\therefore$ Rechazo H_0.

Ahora calculamos el $p - value$ la probabilidad de que Z sea mayor que el valor observado de T_1, de la tabla de la Distribución Normal.

$$p - value = \mathbf{P}[Z > T_1] = \mathbf{P}[Z > 2.9821] = 1 - P[Z \leq 2.9821] = 0.00143.$$

$$\therefore p - value = 0.00143 < 0.05 = \alpha.$$

$\therefore$ Rechazo H_0. Entonces podemos concluir que la probabilidad de buena visión es menor con el nuevo sistema de luces que con el viejo sistema de luces.

12.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
Guardias=c("LucesViejas","LucesNuevas")
BuenaVision=c(714,662)
MalaVision=c(111,154)
datos=data.frame(Guardias,BuenaVision,MalaVision)
datos
```

```
     Guardias BuenaVision MalaVision
1 LucesViejas         714        111
2 LucesNuevas         662        154
```

```
Guardias=c(Guardias,"Total")
BuenaVision=c(BuenaVision, sum(BuenaVision))
MalaVision=c(MalaVision, sum(MalaVision))
Totales=c(BuenaVision[1]+MalaVision[1],BuenaVision[2]+MalaVision[2],
          BuenaVision[1]+MalaVision[1]+BuenaVision[2]+MalaVision[2])
datos1=data.frame(Guardias,BuenaVision,MalaVision,Totales)
datos1
```

```
     Guardias BuenaVision MalaVision Totales
1 LucesViejas         714        111     825
2 LucesNuevas         662        154     816
3       Total        1376        265    1641
```

```
#Datos

O11=714    #Guardias Buena visión/Luces viejas
O12=111    #Guardias Mala vision/Luces viejas
O21=662    #Guardias Buena vision/Luces nuevas
O22=154    #Guardias Mala vision/Luces nuevas

n1=825      #Total luces viejas
n2=816      #Total luces nuevas
C1=1376     #Total buena vision
C2=265      #Total Mala vision
N=n1+n2     #Tamaño total
#N
```

Calcularemos el Estadístico de prueba:

```
T_1=(sqrt(N)*(O11*O22-O12*O21))/(sqrt(n1*n2*C1*C2))
T_1
```

```
[1] 2.982177
```

Calculamos el cuantil y el $p-value$:

```
cuantil=qnorm(0.95)  #Cuantil a comparar
cuantil
```

```
[1] 1.644854
```

```
pvalue=1-pnorm(T_1)  #Calculamos el p-value
pvalue
```

```
[1] 0.00143103
```

Como el estadístico $T_1 = 2.98$ es mayor a el cuantil $Z_{1-\alpha} = 1.64$ por lo tanto rechazamos H_0 y el $p-value$ es menor a nuestra $\alpha = 5\,\%$; Entonces podemos concluir que la probabilidad de buena visión es menor con el nuevo sistema de luces que con el viejo sistema de luces.

Ahora podemos utilizar la prueba en R:

```
tabla<-matrix(c(714,111,825,662,154,816),ncol=3,byrow=TRUE)
tabla
```

```
     [,1] [,2] [,3]
[1,]  714  111  825
[2,]  662  154  816
```

```
chisq.test(tabla)
```

```

    Pearson's Chi-squared test

data:  tabla
X-squared = 8.8934, df = 2, p-value = 0.01172
```

Capítulo 13

Prueba de Independencia

13.1. Datos

Una muestra aleatoria de tamaño N. Las observaciones en la muestra aleatoria son clasificados de acuerdo a dos criterios, usando el primer criterio cada observación es asociada con uno de los r renglones, y usando el segundo criterio cada observación es asociada con una de las C columnas. Sea O_{ij} el número de observaciones asociadas con el renglón i y la columna j simultáneamente; el número total de observaciones en el renglón i es designado por R_i, (en lugar de n_i como la prueba anterior, para enfatizar que los totales de las filas ahora son aleatorios en lugar de fijos), y en la columna j por C_j. La suma de los números en todas las celdas es N.

Los datos se organizan en la siguiente tabla de contingencia:

	Columna 1	**Columna 2**	...	**Columna c**	**Totales**
Renglón 1	O_{11}	O_{12}	...	O_{1c}	R_1
Renglón 2	O_{21}	O_{22}	...	O_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Renglón r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Totales	C_1	C_2	...	C_c	N

13.2. Supuestos

- La muestra de N observaciones es una muestra aleatoria. (Cada observación tiene la misma probabilidad como cualquier otra observación de ser clasificada en el renglón i y la columna j independientemente de las otras observaciones).
- Cada observación puede ser clasificada en una de las r diferentes categorías de acuerdo al primer criterio y en una de las C diferentes categorías de acuerdo al segundo criterio.

13.3. Estadístico de Prueba

El estadístico de prueba T es obtenido de la siguiente manera:

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \quad Donde \;\; E_{ij} = \frac{R_j C_j}{N}.$$

Una expresión equivalente para T, mas adecuado para el uso de su calculadora es:

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N.$$

Distribución de T

La distribución nula de T es obtenida aproximadamente por la Distribución χ^2 con $(r-1) \times (c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Es decir,

$$T \sim \chi^2_{(r-1)(c-1)}.$$

13.4. Hipótesis

$\mathbf{H}_0$: El evento "*una observación está en la fila i*" es independiente del evento "*esa misma observación está en la columna j*" para todos $i,\ j$.

es decir,

$$\mathbf{H}_0 : \mathbf{P}[\ \textit{renglón}\ \ i\ ,\ \ \textit{columna}\ \ j\ \ ] = \mathbf{P}[\ \textit{renglón}\ \ i\ \ ] * \mathbf{P}[\ \textit{columna}\ \ j\ \ ] \quad \forall\ i,\ j.$$

vs

$$\mathbf{H}_a : \mathbf{P}[\ \textit{renglón}\ \ i\ ,\ \textit{columna}\ \ j\ \ ] \neq \mathbf{P}[\ \textit{renglón}\ i\ \ ] * P[\ \textit{columna}\ j\ \ ] \quad ;\ \textit{para alguna}\ \ i\ ,\ j.$$

Regla de decisión

Rechazo H_0 si $T > \chi^2(1-\alpha)$ con $(r-1) \times (c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Ahora un ejercicio.

13.5. Ejercicio

Se especula que la preferencia del cereal ToastyOs está asociada con el nivel educativo de las personas. Si esto resulta cierto, Tabisco Food, la distribuidora del producto, siente que debería aprovechar este mercado dando un mayor empuje de su campaña de marketing a este segmento de la población. Sin embargo, antes de comprometerse a esta tarea se decidió realizar un análisis objetivo que verifiquen las especulaciones. Para esto se tomó una muestra aleatoria de 500 individuos que han probado el producto con los siguientes resultados:

Preferencia	Sin Universidad	Nivel de Estudios Truncados	Graduados
Gusta	75	90	135
Neutral/No les gusta	25	60	115

Paso 1 Prueba a utilizar **Tablas de Contingencia rxc prueba de Independencia.**

Es una prueba de independencia ya que sólo tenemos una población.

Paso 2 Planteamiento de hipótesis.

$\mathbf{H}_0$: El nivel educativo es independiente de la preferencia hacia el producto.

vs

$\mathbf{H}_a$: El nivel educativo no es independiente de la preferencia hacia el producto.

Paso 3 Estadístico de Prueba.

Para poder calcular el estadístico de prueba no ayudaremos de cierto calculos previos.

Preferencia	Sin Universidad	Nivel de Estudios Truncados	Graduados	Totales
Gusta	75	90	135	300
Neutral/No les gusta	25	60	115	200
Totales	100	150	250	500

Ahora:

$P[Guste\ el\ producto] = \frac{300}{500} = .6 = 60\,\%$ *de la población le gusta el cereal.*

$P[No\ Guste\ el\ producto] = \frac{200}{500} = .4 = 40\,\%$ *de la población no le gusta el cereal.*

$P[Sin\ universidad] = \frac{100}{500} = .2 = 20\,\%$ *de la población no tiene universidad.*

$P[Estudios\ truncados] = \frac{150}{500} = .3 = 30\,\%$ *de la población tiene estudios truncados.*

$P[Graduados] = \frac{250}{500} = .5 = 50\,\%$ *de la población están graduados.*

Bajo H_0 el nivel educativo y la preferencia por el cereal son independientes:

Ahora:

$P[Guste\ el\ producto\ y\ sin\ universidad] = P[Guste\ el\ producto] * P[sin\ universidad] = .6 * .2 = .12.$

$P[Guste\ el\ producto\ y\ Estudios\ truncados] = P[Guste\ el\ producto] * P[Estudios\ truncados] = .6 * .3 = .18.$

$P[Guste\ el\ producto\ y\ Graduados] = P[Guste\ el\ producto] * P[Graduados] = .6 * .5 = .30.$

Análogo para no gustar el producto.

Ahora vamos a calcular los valores esperados:

$$E_{11} = \frac{300 * 100}{500} = 60 \quad E_{12} = \frac{300 * 150}{500} = 90 \quad E_{13} = \frac{300 * 250}{500} = 150$$
$$E_{21} = \frac{200 * 100}{500} = 40 \quad E_{22} = \frac{200 * 150}{500} = 60 \quad E_{23} = \frac{200 * 250}{500} = 100$$

Regresando al Estadístico de Prueba:

$$T = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(75-60)^2}{60} + \frac{(90-90)^2}{90} + \frac{(135-150)^2}{150} +$$
$$\frac{(25-40)^2}{40} + \frac{(60-60)^2}{60} + \frac{(115-100)^2}{100} = 13.125.$$

Paso 4 Procedimiento completo.

Utilizaremos un $\alpha = 0.05$.

Entonces tenemos $\chi^2_{(.95,(2-1)*(3-1))} = \chi^2_{(.95,2)} = 5.99$.

Paso 5 Regla de decisión.

Rechazo H_0 si $T > \chi^2(1-\alpha)$ con $(r-1) \times (c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Como $\chi^2 = 5.99 < T = 13.125$

$\therefore$ Rechazo H_0.

Paso 6 Conclusión.

Existe evidencia suficiente para suponer que hay una relación entre la preferencia del cereal y el nivel educativo.

13.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
Preferencia=c("Gusta","Neutral/No les gusta")
Sin_Universidad=c(75,25)
Nivel_Truncado=c(90,60)
Graduados=c(135,115)
datos=data.frame(Preferencia,Sin_Universidad,Nivel_Truncado,Graduados)
datos
```

```
           Preferencia Sin_Universidad Nivel_Truncado Graduados
1               Gusta              75             90       135
2 Neutral/No les gusta              25             60       115
```

```
#ahora, vamos a calcular los totales y adjuntarlos a la tabla
Preferencia=c(Preferencia,"Totales")
Sin_Universidad=c(Sin_Universidad, sum(Sin_Universidad))
Nivel_Truncado=c(Nivel_Truncado, sum(Nivel_Truncado))
Graduados=c(Graduados,sum(Graduados))
Totales=c(Sin_Universidad[1]+Nivel_Truncado[1]+Graduados[1],Sin_Universidad[2]+Nivel_Truncado[2]+Gra
          Sin_Universidad[1]+Nivel_Truncado[1]+Graduados[1]+Sin_Universidad[2]+Nivel_Truncado[2]+Gr
datos=data.frame(Preferencia,Sin_Universidad,Nivel_Truncado,Graduados,Totales)
datos
```

```
           Preferencia Sin_Universidad Nivel_Truncado Graduados Totales
1               Gusta              75             90       135     300
2 Neutral/No les gusta              25             60       115     200
3              Totales             100            150       250     500
```

```
#ESTADÍSTICO DE PRUEBA
#Primero calcularemos los valores esperados
Esperados=matrix(nrow = 2,ncol = 3)
for (i in 1:2) {
  for (j in 1:3) {
    Esperados[i,j]=(datos$Totales[i]*datos[3,j+1])/datos$Totales[3]
  }
}
#Ahora sí, vamos a calcular el estadístico completo
Totales=matrix(nrow = 2,ncol = 3)
for (i in 1:2) {
  for (j in 1:3) {
    Totales[i,j]=(datos[i,j+1]^2)/(Esperados[i,j])
  }
}
```

```
T1=sum(Totales)-datos$Totales[3]
print(c("Estadístico T1 = ", T1))
```

```
[1] "Estadístico T1 = " "13.125"
```

```
alpha=.05
#Ahora vamos a calcular nuestro cuantil
cuantil=qchisq(1-alpha,2)
print(c("Cuantil Ji-cuadrado con 2 grados de libertad  = ", cuantil))
```

```
[1] "Cuantil Ji-cuadrado con 2 grados de libertad  = "
[2] "5.99146454710798"
```

```
if(T1>cuantil){
  print("Se rechaza $H_0$")
  print(c("Estadístico T1 = ", T1))
  print(c("Cuantil Ji-cuadrado con 2 grados de libertad  = ", cuantil))
}
```

```
[1] "Se rechaza $H_0$"
[1] "Estadístico T1 = " "13.125"
[1] "Cuantil Ji-cuadrado con 2 grados de libertad  = "
[2] "5.99146454710798"
```

Observamos que la estadística de prueba tiene un valor de 13.125 y su correspondiente p-value es menor a 0.05, por lo tanto con $\alpha = 5\,\%$ rechazaremos H_0 y concluimos existe evidencia suficiente para suponer que hay una relación entre la preferencia del cereal y el nivel educativo.

Ahora podemos utilizar la prueba en R:

```
tabla<-matrix(c(75,90,135,25,60,115),ncol=3,byrow=TRUE)
dimnames(tabla)<- list(Preferencia=c("Gustar","Neutral o no les gusta"), Nivel=c("Sin universidad","
                       "Graduados"))


tabla
```

```
                       Nivel
Preferencia             Sin universidad Estudios truncados Graduados
  Gustar                             75                 90       135
  Neutral o no les gusta             25                 60       115
```

```
chisq.test(tabla)
```

```

    Pearson's Chi-squared test

data:  tabla
X-squared = 13.125, df = 2, p-value = 0.001412
```

13.7. Ejercicios

1. Queremos probar si la selección de cierto deporte es independienye del género. Para ellos se les pregunto a 100 hombres y 100 mujeres que deporte entre arquería, boxeo y ciclismo preferian practicar y en la siguiente tabla se resume las respuestas que dieron:

Género	Arquería	Boxeo	Ciclismo	Total
Mujer	35	15	50	**100**
Varon	10	30	60	**100**
Total	**45**	**45**	**110**	**200**

2. En un estudio, llevado a cabo por el INE, se tomó una muestra aleatoria de ciudadanos registrados en el Padrón Electoral. Se obtuvo información sobre el partido por el que votaron para Presidente y si tenían o no estudios universitarios.

X_2	X_1						
SI	*PRI*	*PRI*	*PRI*	*PRI*	*PRI*	*PRI*	*PRI*
NO	*PRI*	*PRI*	*PRI*	*PRI*	*PRI*	*PRI*	*PRI*
SI	*PRI*	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*
NO	*PRI*	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*
SI	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*
NO	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*	*PAN*	

Considere X_1= Partido por el cual votó, X_2=tiene o no estudios universitarios.

Se desea probar si existe asociación entre el partido por el cual votó y si tienen o no estudios universitarios. Plantee, en el contexto del problema, las hipótesis, estadística de prueba y obtenga su conclusión con un nivel de significancia del 5 %

3. En una encuesta telefónica se preguntó a los participantes hasta que grado estaban de acuerdo con la proposición: "se debe prohibir fumar en lugares públicos". Con base en los datos recabados se desea saber si existen diferencias significativas en el grado en el que están de acuerdo hombres y mujeres con respecto a prohibir fumar en lugares públicos.

Sexo	Muy de acuerdo	De acuerdo	Neutral	En desacuerdo	En total desacuerdo
Mujer	41	16	28	27	31
Varon	22	40	14	39	41

Capítulo 14

Tablas de Contingecia de $r \times c$

Como una generalización inmediata de la tabla de contingencia 2×2 mencionadas anteriormente, tenemos la tabla de contingencia con r renglones y c columnas, llamada **tablas de contingencia de** $r \times c$. Estas tablas de contingencia pueden usarse, como en la sección anterior, para presentar una tabulación de los datos contenidos en varias muestras, donde los datos representan al menos una escala de medición nominal, y para probar la hipótesis de que las probabilidades no difieren de muestra en muestra.

Prueba de χ^2 para Tablas de Contingencia (Proporciones)

14.1. Datos

Hay r poblaciones en total, y se extrae una muestra aleatoria de cada población. Supongamos que n_i representa el número de observaciones en la muestra $i-$*ésima* (de la población $i-$*ésima*) para $1 \leq i \leq r$. Cada observación en cada muestra es clasificada en una de las C diferentes categorías.

Sea O_{ij} el número de observaciones de la $i-$*ésima* muestra de la categoría j:

$$n_i = O_{i1} + O_{i2} + \cdots + O_{ic} \qquad \forall \;\; i.$$

Los datos se organizan en la siguiente tabla de contingencia de $r \times c$:

	Clase 1	**Clase 2**	$\cdots$	**Clase c**	**Total**
Población 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
Población 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Población r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
Total	C_1	C_2	$\cdots$	C_c	N

El número total de observaciones de todas las muestras es denotado por N:

$$N = n_1 + n_2 + \cdots + n_r.$$

El número de observaciones en la $j-$*ésima* columna denotada como C_j. Esto es, C_j es el número total de observaciones en la $j-$*ésima* categoría, o clase, de todas las muestras combinadas.

$$C_j = O_{1j} + O_{2j} + \cdots + O_{rj}, \qquad para \;\; j = 1, 2, \ldots, c.$$

14.2. Supuestos

- Cada muestra es una muestra aleatoria.
- Los resultados de las diversas muestras son mutuamente independientes (particularmente entre las muestras, porque la independencia dentro de las muestras es parte del primer supuesto).
- Cada observación puede clasificarse en exactamente una de las categorías o clases C.

14.3. Estadístico de Prueba

El estadístico de prueba T es obtenido de la siguiente manera:

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad Donde \;\; E_{ij} = \frac{n_i C_j}{N}.$$

Mientas el término O_{ij} representa el número de observaciones en la celda (i, j), el término E_{ij} representa el número de observaciones esperadas en la celda (i, j). Si H_0 es realmente verdadera, es decir, si H_0 es cierta el número de observaciones en la celda (i, j) podrían estar cerca a la i-ésima muestra de tamaño n_i multiplicado por la proporción $\frac{C_j}{N}$ de todas las observaciones en la categoría j.

NOTA:

En el caso de 2×2 el estadístico T es equivalente a T^2 visto anteriormente, porque solo se considera la hipótesis alternativa de dos colas.

Una expresión equivalente para T, mas adecuado para el uso de su calculadora es:

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N.$$

Distribución de T

La distribución nula de T es obtenida aproximadamente por la Distribución χ^2 con $(r-1) \times (c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Es decir,

$$T \sim \chi^2_{(r-1)(c-1)}$$

14.4. Hipótesis

Sea la probabilidad de que un valor seleccionado aleatoriamente de la i-ésima población se clasifique en la j-ésima clase, denotado por p_{ij} para $i = 1, 2, \ldots, r$ y $j = 1, 2, \ldots, c$.

$\mathbf{H}_0$: Todas las probabilidades en la misma columna son iguales entre sí.

$$vs$$

$\mathbf{H}_a$: Al menos dos de las probabilidades en la misma columna no son iguales entre sí.

en otros términos:

$$\mathbf{H}_0 : \; p_{1j} = p_{2j} = \cdots = p_{rj} \qquad \forall \; j.$$

vs

$$\mathbf{H}_a: \ p_{ij} \neq p_{kj} \ \text{ para algún par } \ i \ \text{ y } \ k.$$

Regla de decisión

Rechazo H_0 si $T > \chi^2(1-\alpha)$ con $(r-1) \times (c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Ahora aplicaremos los conocimientos en el próximo ejemplo:

14.5. Ejercicio

Una muestra de estudiantes seleccionados aleatoriamente de escuelas secundarias privadas recibió pruebas de rendimiento estandarizadas con los siguientes resultados:

	Puntaje de Prueba				
	0-275	**276-350**	**351-425**	**426-500**	**Totales**
Escuela Privada	6	14	17	9	46
Escuela Pública	30	32	17	3	82
Totales	36	46	34	12	128

Paso 1 Prueba a utilizar: **Tablas de Contingencia de $r \times c$.**

Paso 2 Plantear hipótesis:

$\mathbf{H}_0$: La distribución del puntaje de los estudiantes en la prueba es la misma para la escuela privada como para la escuela pública.

vs

$\mathbf{H}_a$: La distribución del puntaje de los estudiantes en la prueba es distinta para la escuela privada como para la escuela pública.

Para poder calcular nuestro estadístico de prueba podemos auxiliarnos de una pequeña tablita para

- E_{ij} :

	Columnas			
	1	**2**	**3**	**4**
Renglón 1	12.9	16.5	12.2	4.3
Renglón 2	23.1	29.5	21.8	7.7

Paso 3 Estadístico de Prueba.

Para la celda en el *renglón* 1, *columna* 1 tenemos:

$$\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(O_{11}-E_{11})^2}{E_{11}} = \frac{(6-12.9)^2}{12.9} = \frac{47.61}{12.9} = 3.69$$

Si hacemos los calculos celda por celda el resultado es:

$$T = 3.69 + 0.38 + 1.89 + 5.14 + 2.06 + 0.21 + 1.06 + 2.87 = 17.3$$

Paso 4 Procedimiento completo.

Ahora buscaremos el cuantil: $\chi^2(1-\alpha)_{(r-1)(c-1)}$ donde $(r-1)\times(c-1) = (2-1)\times(4-1) = 3$. entonces buscamos $\chi^2_{(3)} = 7.815$.

Paso 5 Regla de desición :

Rechazo H_0 si $T > \chi^2(1-\alpha)$ y $17.3 > 7.815$ entonces Rechazo H_0.

Paso 6 Conclusión.

Entonces la distribución del puntaje de los estudiantes en la prueba es distinta para la escuela privada como para la escuela pública.

14.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
data <- matrix(c(6,14,17,9,46,30,32,17,3,82,36,46,34,12,128),nrow=3,byrow=T,
               dimnames=list("Escuelas"=c("EscuelaPrivada","EscuelaPublica","Totales"),
                             "Puntajes"=c("cat1","cat2","cat3","cat4","Totales")))
data
```

```
               Puntajes
Escuelas         cat1 cat2 cat3 cat4 Totales
  EscuelaPrivada    6   14   17    9      46
  EscuelaPublica   30   32   17    3      82
  Totales          36   46   34   12     128
```

```
#Datos

O11=6
O12=14
O13=17
O14=9
O21=30
O22=32
O23=17
O24=3
n1=46
n2=82
C1=36
C2=46
C3=34
C4=12
N=n1+n2

#Ahora calcularemos los Eij

E11=n1*C1/N
E12=n1*C2/N
E13=n1*C3/N
E14=n1*C4/N
E21=n2*C1/N
E22=n2*C2/N
E23=n2*C3/N
E24=n2*C4/N
```

```
#Ahora calcularemos los (Oij-Eij)^2/Eij

j1=(O11-E11)^2/E11
j2=(O12-E12)^2/E12
j3=(O13-E13)^2/E13
j4=(O14-E14)^2/E14
j5=(O21-E21)^2/E21
j6=(O22-E22)^2/E22
j7=(O23-E23)^2/E23
j8=(O24-E24)^2/E24

#Ahora calculamos el estadístico de prueba

T=sum(j1,j2,j3,j4,j5,j6,j7,j8)
T
```

```
[1] 17.28581
```

```
#El cuantil es de una distribución ji.cuadrada con (r-1)(c-1) grados de libertad en este caso serian

qchisq(0.95,3)
```

```
[1] 7.814728
```

Observamos que la estadística de prueba tiene un valor de 17.28 y su correspondiente $p - value$ es menor a 0.05, por lo tanto con $\alpha = 5\,\%$ rechazaremos H_0 y concluimos existe evidencia suficiente para suponer la distribución de los puntajes en la prueba es distinta entre las escuelas privadas y las escuelas públicas.

Ahora podemos utilizar la prueba en R:

```
tabla<-matrix(c(6,14,17,9,46,30,32,17,3,82),ncol=5,byrow=TRUE)
tabla
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    6   14   17    9   46
[2,]   30   32   17    3   82
```

```
chisq.test(tabla)
```

```

    Pearson's Chi-squared test

data:  tabla
X-squared = 17.286, df = 4, p-value = 0.001701
```

14.7. Ejercicios

1. ¿Qué tan bueno es el servicio que dan las líneas aéreas a sus clientes? En un estudio las evaluaciones dadas por los clientes fueron las siguientes: 3 excelente, 28 bueno, 45 aceptable y 24 malo (BusinessWeek, 11 de septiembre de 2000). En otro estudio sobre las empresas de servicio telefónico, en una muestra de 400 adultos las evaluaciones fueron las siguientes: 24 excelente, 124 bueno, 172 aceptable y 80 malo. ¿La distribución de las evaluaciones a las empresas telefónicas difiere de la distribución de las evaluaciones a las líneas aéreas? Emplee $\alpha = 5\,\%$ ¿Cuál es su conclusión?

2. Estamos interesados en estudiar la fiabilidad de cierto componente informático con relación al distribuidor que nos lo suministra. Para realizar esto, tomamos una muestra de 100 componentes de cada uno de los tres distribuidores que nos sirven el producto comprobando el número de defectuosos en cada lote. La siguiente tabla muestra el número de defectuosos para cada uno de los distribuidores.

	Componentes Defectuosos	Componentes Correctos
Distribuidor 1	16	94
Distribuidor 2	24	76
Distribuidor 3	9	81

Plantea la hipótesis correspondiente y concluye de acuerdo al contexto.

3. Durante las primeras 13 semanas, se registraron las proporciones siguientes de televidentes los sábados de 8 a 9 de la noche: ABC 29 %, CBS 28 %, NBC 25 % e independientes 18 %. Dos semanas después en una muestra de 300 hogares se obtuvieron las audiencias siguientes en sábado por la noche: ABC 95 hogares, CBS 70 hogares, NBC 89 hogares e independientes 46 hogares. Use $\alpha = 5\,\%$ para determinar si han variado las proporciones en la audiencia de televidentes.

Capítulo 15

Prueba de la Mediana

La prueba mediana está diseñada para examinar si varias muestras provienen de poblaciones que tienen la misma mediana. En realidad, la prueba de la mediana no es nueva, es simplemente una aplicación especial de la prueba de ji cuadrado con totales marginales fijos. Sin embargo, es una aplicación muy útil y consideramos que vale la pena un trato especial.

Para probar si varias c poblaciones tienen la misma mediana, se extrae una muestra aleatoria de cada población (la escala de medición es al menos ordinal, o el término "mediana" no tendría sentido). Se construye una tabla de contingencia $2 \times c$ y las dos entidades en la $i-ésima$ columna son los números de observaciones en la $i-ésima$ muestra que están por encima y por debajo de la gran mediana (la mediana de todas las observaciones combinadas). La prueba de χ^2 habitual se aplica luego a la tabla de contingencia.

15.1. Datos

Para cada una de las poblaciones c se obtiene una muestra aleatoria de tamaño n_i, $i = 1, 2, \ldots, c$. Se determina la mediana de la muestra combinada; es decir, se determina el número que excedió aproximadamente la mitad de las observaciones en toda la muestra $N = n_1 + n_2 + \ldots + n_c$. Esto se llama mediana. Sea O_{1i}, el número de observaciones en la i-ésima muestra que excede la mediana y, sea O_{2i} el número en la i-ésima muestra que sea menor o igual a la mediana, organice los conteos de frecuencia en una tabla de contingencia $2 \times c$ de la siguiente manera:

	Muestra 1	**Muestra 2**	$\cdots$	**Muestra c**	**Totales**
>Mediana	O_{11}	O_{12}	$\cdots$	O_{1c}	a
$\leq$ Mediana	O_{21}	O_{22}	$\cdots$	O_{2c}	b
Totales	n_1	n_2	$\cdots$	n_c	N

Sea a el número total de observaciones mas grandes a la mediana en todas las muestras, sea b el número total de observaciones menores o iguales a la mediana. Entonces $a + b = N$ es el número total de observaciones.

15.2. Supuestos

- Cada muestra es una muestra aleatoria.
- Las muestras son independientes entre si.
- La escala de medida es al menos ordinal.
- Si todas las poblaciones tienen la misma mediana, todas las poblaciones tienen la misma probabilidad p de que una observación exceda la mediana.

15.3. Estadístico de Prueba

El estadístico de prueba es obtenido por el reordenamiento del estadístico utilizado en la prueba anterior, notando que $O_{2i} = n_i - O_{1i}$ en el caso especial con 2 renglones.

$$T = \frac{N^2}{ab} * \sum_{i=1}^{c} \frac{(O_{1i} - \frac{n_i a}{N})^2}{n_i}.$$

Para ahorrarnos algunos cálculos:

$$T = \frac{N^2}{ab} * \sum_{i=1}^{c} \frac{O_{1i}^2}{n_i} - \frac{Na}{b}.$$

NOTA:

Si a es exactamente igual a b nuestro estadístico de prueba se simplifica en:

$$T = \sum_{i=1}^{c} \frac{(O_{1i} - O_{2i})^2}{n_i}.$$

Distribución de T

La distribución nula de T es obtenida aproximadamente por la Distribución χ^2 con $(c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Es decir,

$$T \sim \chi^2_{(c-1)}.$$

15.4. Hipótesis

$\mathbf{H}_0$: Todas las c poblaciones tienen la misma mediana.

vs

$\mathbf{H}_a$: Al menos 2 de las poblaciones tienen diferente mediana.

Regla de decisión

Rechazo H_0 si $T > \chi^2(1-\alpha)$ con $(c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

15.5. Comparación Múltiple

Si la hipótesis nula es rechazada, se pueden hacer comparaciones múltiples por parejas entre poblaciones utilizando la prueba de mediana repetidamente en tablas de contingencia de 2×2. En cada comparación se encuentra la mediana de las dos muestras, y el número por encima o por debajo de esa mediana se utiliza en la tabla de contingencia 2×2. Calculamos el estadístico de prueba T para la prueba y si T es más grande que el cuantil $1-\alpha$ de la distribución χ^2 con 1 grado de libertad encontrado en tablas de dicha distribución, entonces decimos que las medianas de esas 2 poblaciones son iguales.

Aplicaremos nuestros conocimientos en el siguiente ejercicio.

15.5.1. Ejercicio

Se asignaron al azar cuatro métodos diferentes de cultivo de maíz a un gran número de parcelas diferentes y se calculó el rendimiento por acre para cada parcela. Los datos son los siguientes:

Método 1	**Método 2**	**Método 3**	**Método 4**
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Para determinar si existe una diferencia en los rendimientos como resultado del método utilizado, se empleó la prueba mediana porque se consideró que una diferencia en las medianas de la población podría interpretarse como una diferencia en el valor del método utilizado.

Paso 1 Prueba a utilizar **Tablas de contingencia, Prueba de la mediana.**

Paso 2 Planteamineto de Hipótesis.

$\mathbf{H}_0$: Todos los métodos tienen el mismo rendimiento medio (mediana) por acre.

vs

$\mathbf{H}_a$: Al menos dos de los métodos difieren con respecto al rendimiento medio(mediana) por acre.

En el conteo rápido revela que hay 34 observaciones en total, por lo que el promedio de las observaciones más pequeñas decimoséptima y decimoctava es la mediana, y se ve que es 89. Luego, para cada método (muestra), el número de valores que exceden 89 y el número que es menor o igual a 89 se registra en la siguiente forma.

	Método 1	**Método 2**	**Método 3**	**Método 4**	**Totales**
>89	6	3	7	0	16
$\leq$ 89	3	7	0	8	18
Totales	9	10	7	8	34

Paso 3 Estadístico de Prueba.

$$T = \frac{N^2}{ab} * \sum_{i=1}^{c} \frac{(O_{1i} - \frac{n_i a}{N})^2}{n_i}.$$

$$T = \frac{(34)^2}{(16)(18)} * \left(\frac{\left[6 - \frac{(9)(16)}{34}\right]^2}{9} + \frac{\left[3 - \frac{(10)(16)}{34}\right]^2}{10} + \frac{\left[7 - \frac{(7)(16)}{34}\right]^2}{7} + \frac{\left[0 - \frac{(8)(16)}{34}\right]^2}{8} \right).$$

$$T = 4.01(0.34 + 0.29 + 1.97 + 1.78) = 17.6$$

Paso 4 Regla de decisión.

Rechazo H_0 si $T > \chi^2(1-\alpha)$ con $(c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Tomaremos un α=0.05, entonces $\chi^2(.95)$ con 3 grados de libertad es 7.815, obtenido en las tablas correspondientes a la Distribución χ^2.

Como $T = 17.6 > 7.815 = \chi^2(.95)$ rechazamos H_0.

Paso 5 Conclusión

Como rechazo H_0, entonces almenos 2 de los métodos difieren con respecto al rendimiento medio (mediana) por acre.

Como se rechazó H_0, se le deja al lector hacer la comparación múltiple.

15.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Datos
Rendimiento=c(83,91,94,89,89,96,91,92,90,91,90,81,83,84,83,88,91,89,84,101,100,91,
              93,96,95,94,78,82,81,77,79,81,80,81)
Metodo=c(1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4,4,4)

#Calcular la mediana
mediana=median(Rendimiento)
mediana
```

```
[1] 89
```

Después de calcular la mediana, contamos cuantas observaciones hay por método por arriba y por abajo de la mediana y con dichas frecuencias construimos la tabla de contingencia.

```
#Construir la tabla de contingencia
table(Metodo[which(Rendimiento>mediana)])
```

```
1 2 3
6 3 7
```

```
table(Metodo[which(Rendimiento<=mediana)])
```

```
1 2 4
3 7 8
```

```
Observados=matrix(c(6,3,7,0,3,7,0,8), nrow = 2, ncol = 4, byrow = T)
rownames(Observados)=c('Mayor_med','MenorIgual_med')
colnames(Observados)=c('Metodo 1','Metodo 2','Metodo 3','Metodo 4')
Observados
```

```
               Metodo 1 Metodo 2 Metodo 3 Metodo 4
Mayor_med             6        3        7        0
MenorIgual_med        3        7        0        8
```

Ya que se tiene la tabla se le puede aplicar la prueba. En este caso observaremos que la prueba nos advierte sobre el uso de la estadísitca Ji-Cuadrada, lo anterior se debe a que las frecuencias en la tabla de contingencia son pequeñas e incluso cero. La prueba alternativa en estos casos es la prueba exacta de Fisher.

```
#Prueba
T1 <- chisq.test(Observados)
T1
```

```
    Pearson's Chi-squared test

data:  Observados
X-squared = 17.543, df = 3, p-value = 0.0005464
```

```
T2 <- fisher.test(Observados)
T2
```

```
    Fisher's Exact Test for Count Data

data:  Observados
p-value = 0.0001631
alternative hypothesis: two.sided
```

Con ambas pruebas se llega a la conclusión de rechazar H_0 con $\alpha = 0.01$ y por lo tanto al menos dos de los métodos difieren con respecto al rendimiento medio (mediana) por m^2.

Podemos ver los valores esperados con:

```
#Valores de las frecuencias esperadas
T1$expected
```

```
               Metodo 1 Metodo 2 Metodo 3 Metodo 4
Mayor_med      4.235294 4.705882 3.294118 3.764706
MenorIgual_med 4.764706 5.294118 3.705882 4.235294
```

Podemos realizar las comparaciones múltiples. A continuación se presenta la comparación de los métodos 2 y 3.

```
#Datos
Rendimiento=c(91,90,81,83,84,83,88,91,89,84,101,100,91,
              93,96,95,94)
Metodo=c(2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3)

#Calcular la mediana
mediana=median(Rendimiento)
mediana
```

```
[1] 91
```

```
table(Metodo[which(Rendimiento>mediana)])
```

```
3
6
```

```
table(Metodo[which(Rendimiento<=mediana)])
```

```
 2  3
10  1
```

```
Observados=matrix(c(0,6,10,1), nrow = 2, ncol = 2, byrow = T)
rownames(Observados)=c('Mayor_med','MenorIgual_med')
colnames(Observados)=c('Metodo 2','Metodo 3')
Observados
```

```
               Metodo 2 Metodo 3
Mayor_med             0        6
MenorIgual_med       10        1
```

```
#Prueba
T2 <- fisher.test(Observados)
```

```
T2
```

```
    Fisher's Exact Test for Count Data

data:  Observados
p-value = 0.0005656
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.0000000 0.2669263
sample estimates:
odds ratio
         0
```

Los métodos 2 y 3 difieren con respecto al rendimiento medio (mediana) por m^2.

15.7. Ejercicios

1. En el campo de entrenamiento 100 reclutas son asignados aleatoriamente en cuatro regimientos con 4 diferentes sargentos. Al final del entrenamiento sólo quedarón 84 reclutas y sus tiempos en el ejercicio de obstáculos fue medido para todos ellos. Los resultados fueron, para el sargento Adams 11 de sus 20 reclutas tuvieron tiempos por arriba de la mediana, para el sargento Baker 8 de sus 22 reclutas tuvieron tiempos por arriba de la mediana, el sargento Callahan 8 de sus 20 reclutas tuvieron tiempos por encima de la mediana y del sargento Davis 15 de 22 tuvieron tiempos arriba de la mediana. Se puede decir con un nivel de significancia del 5 % que existe una diferencia significativa en los tiempos en los ejercicios de obstáculos entre cada uno de los regimientos?

2. Se subastaron varios contratos de explotación de petróleo al mejor postor. Para cada contrato se recibieron una o más ofertas selladas. Pruebe la hipótesis de que los contratos que eventualmente se convirtieron en productores de petróleo tuvieron la misma mediana de ofertas que los contratos que no produjeron petróleo. A continuación se muestra una muestra aleatoria de cada tipo de contrato.

	Número de ofertas en cada contrato de arrendamiento
Productores	$6, 3, 1, 14, 8, 9, 12, 1, 3, 2, 1, 7$
No productores	$6, 2, 1, 1, 3, 1, 2, 4, 8, 1, 2$

Parte V

Bondad de Ajuste

Introducción

Las pruebas de bondad de ajuste se utilizan para decidir si un conjunto de datos (una muestra aleatoria) se ajusta a una función de distribución dada.

Estas pruebas son importantes porque existen métodos estadísticos que se basan en algún supuesto de la distribución de los datos y si tal supuesto no se cumple, el método no es válido. Por ejemplo, el modelo de regresión lineal supone que los errores tienen distribución normal y de no validarse este supuesto entonces la inferencia hecha sobre los parámetros del modelo de regresión carece de sustento estadístico.

La hipótesis nula en este tipo de pruebas es que los datos tienen la distribución requerida. Algunas veces la hipótesis nula específica totalmente a la distribución (es decir también especifica el valor de los parámetros), otras veces sólo especifica de que distribución se trata (sin importar los parámetros).

Las pruebas de bondad de ajuste desarrolladas dependen de la variable aleatoria que se está modelando. Para distribuciones discretas, se comparan las frecuencias esperadas con las observadas (Pruebas Ji-Cuadrada). Para distribuciones continuas, se compara la función de distribución empírica con la de distribución requerida (Pruebas Kolmogorov-Smirnov, Lilliefors, Anderson-Darling)

Capítulo 16

Prueba de la Ji-cuadrada

16.1. Datos

$Sea X_1, \cdots, X_n$ *m.a.* de tamaño "n" que proviene de una distribución $F(x)$ desconocida.

Cada una de las variables se pueden acomodar en alguna clase "k" (o categoría)

16.2. Hipótesis

$$\mathbf{H}_0: \text{ Los datos siguen una distribución } F_0(x).$$

$$vs$$

$$\mathbf{H}_a: \text{ Los datos no siguen una distribución } F_0(x).$$

- Donde $F_0(x)$ es la distribución que se propone.

Es decir:

$$\mathbf{H}_0: \mathbf{P}[\ X \text{ pertenezca a la categoría } j\ ] = \mathbf{P}_j, \quad \text{para toda } j = 1, \ldots, k.$$

$$vs$$

$$\mathbf{H}_a: \mathbf{P}[\ X \text{ pertenezca a la categoría } j\ ] \neq \mathbf{P}_j, \quad \text{para alguna } j = 1, \ldots, k.$$

Procedimiento.

Vamos a buscar las probabilidades de ocurrencia en cada categoría.

Calcular los valores esperados e_j; Donde $e_j = n \times P_j$.

16.3. Estadístico de Prueba

El estadístico de prueba que ocuparemos es:

$$Q = \sum_{j=1}^{k} \frac{(f_j - e_j)^2}{e_j}, \qquad Donde \;\; Q \sim \chi^2_{(k-1)}.$$

Observaciones

- Q es estable cuando el número de observaciones en cada categoría debe ser mayor a 5.
- En caso de que alguna categoría tenga menos de 5 observaciones colapsamos las categorías.
- Si desconocemos los parámetros los estimamos con la muestra (EMV, $Momentos$), con esto se van perdiendo grados de libertad.
- $Q \sim \chi^2_{(k-1-r)}$ Donde k=Número de clases o intervalos, r=Número de parámetros estimados.

Regla de decisión.

Rechazamos H_0 si $Q > q_{teórica}$.

- Cuando $Q = \sum_{j=1}^{k} \frac{(f_j - e_j)^2}{e_j} > \chi^2_{(k-1)}(1-\alpha)$.

16.4. Ejemplo

Un gobierno local tiene registros del número de niños y el número de hogares en el área. Se sabe que el número promedio de niños por hogar es 1.40. Se sugiere que el número de niños por hogar se pueda modelar por una distribución Poisson con parámetro 1.40. Para probar esta hipótesis se toma una muestra de 1000 hogares; los resultados se muestran en la siguiente tabla:

Número de niños	0	1	2	3	4	5+
Número de hogares	273	361	263	78	21	4

NOTA: Como tenemos una categoría con obeservaciones menores a 5 colapsamos la categoría.

Número de niños	0	1	2	3	4+
Número de hogares	273	361	263	78	25

Paso 1 Prueba a utilizar **Prueba Ji-cuadrada.**

Paso 2 Planteamiento de hipótesis:

$\mathbf{H}_0$: El número de niños por hogar sigue una distribución Poisson (1.40).

vs

$\mathbf{H}_a$: El número de niños por hogar no sige una distribución Poisson (1.40).

Necesitamos buscar la Distribución de una variable aleatoria Poisson, con parámetro λ= 1.40.

Nota En R-Studio tenemos dpois(c(0,1,2,3,4),1.40) para la primera categoría.

Nos ayudaremos de la siguiente tablita:

Número de niños	Número de hogares	$P_i = dpois(0:4, 1.4)$	$e_i = n \times P_i$
0	273	0.2465	247
1	361	0.3452	345
2	263	0.2416	242
3	78	0.1127	113
4	25	0.0394	39
	$n = 1000$	$.9854 \approx 1$	$9854 \approx 1000$

Paso 3 Estadístico de prueba.

$$Q = \sum_{j=0}^{k} \frac{(f_j - e_j)^2}{e_j} = \sum_{j=0}^{4} \frac{(f_j - e_j)^2}{e_j} =$$

$$= \frac{(273-247)^2}{247} + \frac{(361-345)^2}{345} + \frac{(263-242)^2}{242} + \frac{(78-113)^2}{113} + \frac{(25-39)^2}{39}.$$

$$T = 21.4605.$$

Paso 4 Regla de decisión.

Rechazamos H_0 si $Q > q_{teórica}$.

- Cuando

$$Q = \sum_{j=1}^{k} \frac{(f_j - e_j)^2}{e_j} > \chi^2_{(k-1)}(1-\alpha).$$

Ocuparemos $\alpha = 0.05$.

$\chi^2_{(k-1)}(1-\alpha) = \chi^2_{(5-1)}(1-0.05) = \chi^2_{(4)}(.95)$.

$$Q = 21.4605 > 9.48 = \chi^2_{(4)}(.95).$$

$\therefore$ Rechazamos H_0.

Paso 5 Conclusión.

$\therefore$ Existe evidencia estadística suficiente para decir que el número de niños por hogar no sigue una Distribución Poisson (1.40).

16.5. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Número de hogares
observados =c(273,361,263,78,21,4)
#Matriz de frecuencias observadas
tabla=matrix(observados,nrow=1)
#Número de niños
num_ninos=c("0","1","2","3","4","5+")
#Asigna nombres a la tabla de cada categoría
dimnames(tabla)=list(NULL,num_ninos)
tabla
```

```
      0   1   2  3  4 5+
[1,] 273 361 263 78 21  4
```

```
#Una de las categorías es menor a 5, asi que colapsamos la categoría 5+

#Número de hogares
observados =c(273,361,263,78,25)
```

```
#Matriz de frecuencias observadas
tabla=matrix(observados,nrow=1)
#Número de niños
num_ninos=c("0","1","2","3","4+")
#Asigna nombres a la tabla de cada categoría
dimnames(tabla)=list(NULL,num_ninos)
tabla
```

```
     0   1   2  3 4+
[1,] 273 361 263 78 25
```

```
#Total
N=sum(observados)

#Calculamos las probabilidades
p=dpois(0:4,1.40)
p
```

```
[1] 0.24659696 0.34523575 0.24166502 0.11277701 0.03947195
```

```
#Calculamos los valores esperados
esperados=N*p
round(esperados,0) #Hacemos esto por que estamos hablando de niños
```

```
[1] 247 345 242 113  39
```

```
residuos=round((observados-esperados)^2/esperados,4)
ji_cal=sum(residuos)
ji_cal
```

```
[1] 21.4605
```

```
#En un principio teníamos 6 categorías pero colapsamos una, entonces
#tenemos 5 categorías menos una que corresponden a nuestros grados de libertad.

ji_teo=qchisq(0.95,df=4)
ji_teo
```

```
[1] 9.487729
```

```
#Como Ji-Cal>Ji_teo rechazamos H0.

p_value=pchisq(ji_cal,df=4,lower.tail=FALSE)
p_value
```

```
[1] 0.0002565777
```

Observamos que la estadística de prueba tiene un valor de 21.46 y su correspondiente p-value es menor a 0.05, por lo tanto con $\alpha = 5\,\%$ rechazaremos H_0 y concluimos existe evidencia suficiente para suponer que el número de niños por hogar no sigue una distribución Poisson(1.40).

Ahora podemos utilizar la prueba de R:

```
chisq.test(c(273,361,263,78,25), p = c(0.2466,0.3452,0.2417,
                                       0.1128,0.03947), rescale.p=TRUE)
```

```
    Chi-squared test for given probabilities

data:  c(273, 361, 263, 78, 25)
X-squared = 20.96, df = 4, p-value = 0.0003226
```

16.6. Ejercicios

1. Se lanza un dado 600 veces se obtuvieron los siguientes resultados.

Número del dado	1	2	3	4	5	6
Observaciones	87	96	108	89	122	98

¿El dado está balanceado (es decir, los datos tienen distribución uniforme con proba 1/6)? Use $\alpha = 0.10$

2. Cierto banco otorga crédito a las personas con una tasa preferencial, de tal manera que los acreditados pueden pagar en cualquier momento desde que piden el préstamo hasta 8 semanas posteriores para que les sea respetada la tasa preferencial. Se seleccionaron aleatoriamente a 1,000 personas y observó su comportamiento de pago, generando de esta manera la siguiente tabla de frecuencia:

Semana	Créditos Pagados
Menos de 1 semana	64
$1 \leq x < 2$	191
$2 \leq x < 3$	283
$3 \leq x < 4$	241
$4 \leq x < 5$	140
$5 \leq x < 6$	51
$6 \leq x < 7$	25
$7 \leq x < 8$	4
8 semanas o más	1

Probar que el pago de estos créditos, sigue una distribución binomial con parámetros $n = 10$ y $p = 0.25$.

Capítulo 17

Prueba Kolmogorov

Comenzaremos con una prueba de bondad de ajuste que fue presentada por Kolmogorov (1933). Esta prueba es quizás la más útil, en parte porque nos proporciona una alternativa, diseñada para datos ordinales, a la prueba χ^2 para bondad de ajuste, que fue diseñada para datos de tipo nominal, y en parte porque la estadística de prueba de Kolmogorov nos permite formar una "banda de confianza" para la función de distribución desconocida.

Una prueba de bondad de ajuste generalmente involucra una muestra aleatoria de alguna distribución desconocida para probar la hipótesis nula de que la función de distribución desconocida es de hecho una función conocida y especificada. Esto es, la hipótesis nula especifica alguna función de distribución $F^*(x)$, tal vez gráficamente como en la figura 1, o tal vez como una función matemática que puede ser graficada. Luego se toma una muestra aleatoria, $X_1, X_2, \ldots, X_n$ de alguna población y se compara con $F^*(x)$ de alguna manera para ver si es razonable decir que, $F^*(x)$ es la verdadera función de distribución de la muestra aleatoria.

Una forma lógica de comparar la muestra aleatoria con $F^*(x)$ es mediante la función de distribución empírica $S(x)$, definida como la fracción de $X_i s$ que son menores o iguales a x, para cada x, $-\infty < x < +\infty$. La función de distribución empírica $S(x)$ es útil como estimador de $F(x)$, la función de distribución desconocida de la $X_i s$. Entonces podemos comparar la función de distribución empírica $S(x)$ con la función de distribución hipotética $F^*(x)$ para ver si hay un buen ajuste. Si no hay un buen ajuste, entonces podemos rechazar la hipótesis nula y concluir que la función de distribución verdadera pero desconocida, $F(x)$, en realidad no está dada por la función $F^*(x)$ en la hipótesis nula.

Figura 1

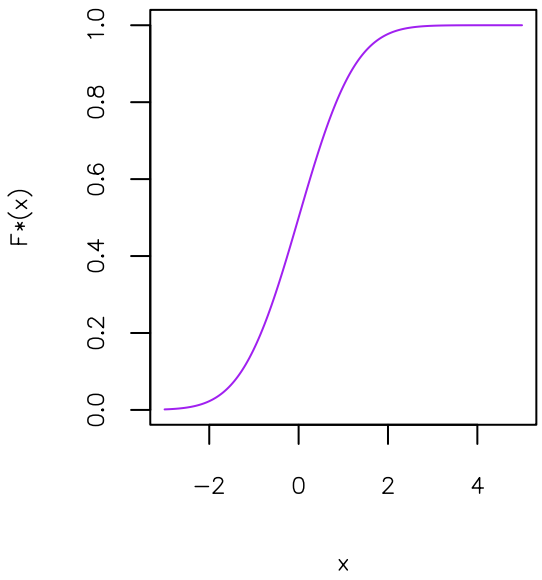
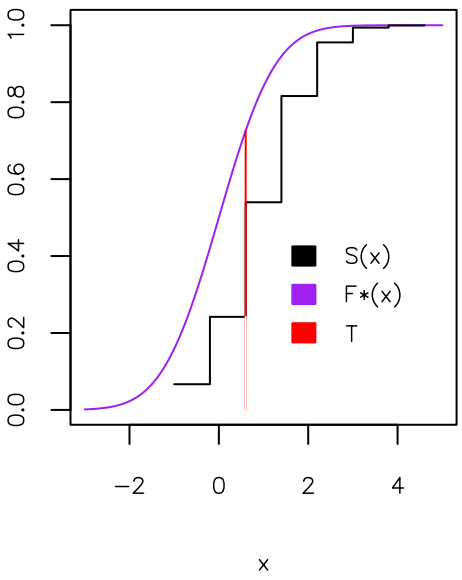


Figura 2



Pero, ¿qué tipo de estadística de prueba podemos usar como medida de la diferencia entre $S(x)$ y $F^*(x)$? Una de las medidas más simples imaginables es la mayor distancia entre los dos gráficos $S(x)$ y $F^*(x)$, medidos en dirección vertical. Ésta es la estadística sugerida por Kolmogorov (1993). Es decir, si la **figura 1** proporciona $F^*(x)$ y se extrae una muestra aleatoria de tamaño 5 de la población, la función de distribución empírica $S(x)$ se puede dibujar en el mismo gráfico junto con $F^*(x)$, como se muestra en la figura 2. Si $F^*(x)$ y $S(x)$ son como se indica, la distancia vertical máxima entre los dos gráficos se produce justo antes del tercer paso de $S(x)$. Esta distancia es de aproximadamente 0.5 en la **figura 2**; por lo tanto, el estadístico Kolmogorov T es igual a 0.5 en este caso. Los valores grandes de T según lo determinado por la correspondiente tabla (Tabla de Cuantiles Kolmogorov) conducen a rechazar $F^*(x)$ como una aproximación razonable a la función de distribución verdadera desconocida $F(x)$.

La prueba de Kolmogorov puede preferirse a la prueba de Ji-Cuadrada para bondad del ajuste si el tamaño de la muestra es pequeño; la prueba de Kolmogorov es exacta incluso para muestras pequeñas, mientras que la prueba de Ji-cuadrada supone que el número de observaciones es lo suficientemente grande como para que la distribución de χ^2 proporcione una buena aproximación como la distribución del estadístico de prueba. Existe controversia sobre ¿qué prueba es la más "poderosa"?, pero la sensación general parece ser que la prueba de Kolmogorov es probablemente más poderosa que la prueba de Ji-cuadrada en la mayoría de las situaciones que involucran datos ordinales.

17.1. Datos

Los datos consisten en una muestra aleatoria $X_1, X_2, \ldots, X_n$ de tamaño n asociada con alguna función de distribución desconocida, denotada por $F(x)$.

17.2. Supuestos

1) La muestra es una muestra aleatoria.

17.3. Estadístico de Prueba

Sea $S(x)$ la función de distribución empírica basada en la muestra aleatoria $X_1, X_2, \ldots, X_n$. El estadístico de prueba es definido diferente para los 3 casos para las hipótesis correspondientes. Sea $F^*(x)$ una función de distribución hipotética completamente especificada.

Caso A (Prueba de 2 colas)

Sea el estadístico de prueba T la mayor distancia vertical entre $S(x)$ y $F^*(x)$(denotado por sup o el supremo).

$$T = \underset{x}{sup}|F^*(x) - S(x)|.$$

Esto se lee "T es igual al supremo para todas las x, del valor absoluto de la diferencia de $F^*(x) - S(x)$."

Caso B (Prueba de 1 cola)

Denotamos el estadístico de prueba T^+ la mayor distancia vertical alcanzada por $F^*(x)$ sobre $S(x)$.

$$T^+ = \underset{x}{sup}[F^*(x) - S(x)].$$

Que es similar a T, a excepción que solo vamos a considerar la mayor diferencia alcanzada por $F^*(x)$ sobre la función $S(x)$.

Caso C (Prueba de 1 cola)

Denotamos el estadístico de prueba T^- la mayor distancia vertical alcanzada por $S(x)$ sobre $F^*(x)$.

$$T^- = \underset{x}{sup}[S(x) - F^*(x)].$$

Que es similar a T, a excepción que solo vamos a considerar la mayor diferencia alcanzada por $S(x)$ sobre la función $F^*(x)$.

17.4. Hipótesis

Caso A (Prueba de 2 colas)

$$\mathbf{H}_0 : \ F(x) = F^*(x) \quad \forall \ x \quad \text{de} \quad -\infty \quad \text{a} \ +\infty$$

$$vs$$

$$\mathbf{H}_a : \ F(x) \neq F^*(x) \quad \text{para al menos un valor de } x.$$

Regla de decisión

Rechazo H_0 a un nivel de significancia α si $T > W$. Donde W es el cuantil $(1 - \alpha)$, obtenido en la tabla correspondiente a nuestra prueba, para la prueba de 2 colas.

Caso B (Prueba de 1 cola)

$$\mathbf{H}_0 : \ F(x) \geq F^*(x) \quad \forall \ x \ \text{ de } \ -\infty \quad \text{a} \quad +\infty$$

$$vs$$

$$Ha : \ F(x) < F^*(x) \quad \text{para al menos un valor de } x.$$

Regla de decisión

Rechazo H_0 a un nivel de significancia α si $T^+ > W$. Donde W es el cuantil $(1 - \alpha)$, obtenido en la tabla correspondiente a nuestra prueba, para la prueba de 1 cola.

Caso C (Prueba de 1 cola)

$$\mathbf{H}_0 : \ F(x) \leq F^*(x) \quad \forall \ x \ \text{ de } \ -\infty \ \text{ a } \ +\infty$$

$$vs$$

$$\mathbf{H}_a : \ F(x) > F^*(x) \quad \text{para al menos un valor de } x.$$

Regla de decisión

Rechazo H_0 a un nivel de significancia α si $T^- > W$. Donde W es el cuantil $(1-\alpha)$, obtenido en la tabla correspondiente a nuestra prueba, para la prueba de 1 cola.

Vamos a aplicar este conocimiento en un ejemplo.

17.5. Ejemplo

Una muestra aleatoria de tamaño 10, es obtenida:

$X_1 = 0.621,\ \ X_2 = 0.503,\ \ X_3 = 0.203,\ X_4 = 0.477,\ X_5 = 0.710,\ X_6 = 0.581,$
$X_7 = 0.329,\ X_8 = 0.480,\ X_9 = 0.554,\ X_{10} = 0.382$

La hipótesis nula es que la función de distribución es una función de distribución uniforme. La expresión matemática de función de distribución hipotética es:

$$\mathbf{F^*(x)=}\begin{cases} 0 & si \quad x < 0 \\ x & si \quad 0 \leq x < 1 \\ 1 & si \quad 1 \leq x \end{cases}$$

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Kolmogorov.**

Paso 2 Planteamiento de Hipótesis:

$$\mathbf{H}_0 :\ F(x) = F^*(x) \quad \forall\ x\ \text{ de }\ -\infty\ \text{ a }\ +\infty$$

$$vs$$

$$\mathbf{H}_a :\ F(x) \neq F^*(x) \quad \text{para al menos un valor de } x.$$

- Donde **F(x)** es la función de distribución desconocida común a las $X_i s$ y $F^*(x)$ se da por la expresión matemática.

$$\mathbf{F^*(x)=}\begin{cases} 0 & si \quad x < 0 \\ x & si \quad 0 \leq x < 1 \\ 1 & si \quad 1 \leq x \end{cases}$$

Paso 3 Estadístico de Prueba.

Calculamos el Estadístico de Prueba:

$$T = \sup_x |F^*(x) - S(x)|.$$

$$T = 0.290.$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba: La siguiente tabla representa los cálculos para encontrar nuestro Estadístico de Prueba T:

Paso 5 Regla de Decisión.

El cuantil W que acumula $1-\alpha$ de probabilidad, usando α=0.05 es $W = 0.409$, encontrado en las tablas correspondientes.

Tenemos que $T < W$, entonces no rechazamos la hipotesis nula.

Paso 6 Conclusión.

Entonces podemos concluir que los datos siguen una distribución uniforme.

i	$X(i)=x$	$F^*(x)$	S_n	$F^*(x)-S_n(x)$	$\|F^*(x)-S_n(x)\|$
1	0.203	0.203	0.1	0.103	0.103
2	0.329	0.329	0.2	0.129	0.129
3	0.382	0.382	0.3	0.082	0.082
4	0.477	0.477	0.4	0.077	0.077
5	0.480	0.480	0.5	-0.020	0.020
6	0.503	0.503	0.6	-0.097	0.097
7	0.554	0.554	0.7	-0.146	0.146
8	0.581	0.581	0.8	-0.219	0.219
9	0.621	0.621	0.9	-0.279	0.279
10	0.710	0.710	1.0	-0.290	0.290

17.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
i = c(1:10) #Representa el numero de nuestra muestra
x = c(0.203,0.329,0.382,0.477,0.480,0.503,0.554,0.581,0.621, 0.710) #Los datos de la muestra
X_i = sort(x) #ordena nuestros datos
F_=c(0.203,0.329,0.382,0.477,0.480,0.503,0.554,0.581,0.621, 0.710)
Sn=c(1/10,2/10,3/10,4/10,5/10,6/10,7/10,8/10,9/10,1)
Tabla = cbind(i,X_i=X_i,F_,Sn,"|F_-Sn|"=abs(F_-Sn))
Tabla
```

```
      i   X_i    F_  Sn |F_-Sn|
 [1,]  1 0.203 0.203 0.1   0.103
 [2,]  2 0.329 0.329 0.2   0.129
 [3,]  3 0.382 0.382 0.3   0.082
 [4,]  4 0.477 0.477 0.4   0.077
 [5,]  5 0.480 0.480 0.5   0.020
 [6,]  6 0.503 0.503 0.6   0.097
 [7,]  7 0.554 0.554 0.7   0.146
 [8,]  8 0.581 0.581 0.8   0.219
 [9,]  9 0.621 0.621 0.9   0.279
[10,] 10 0.710 0.710 1.0   0.290
```

```
EstdPrueba = max(Tabla [,5])
EstdPrueba
```

```
[1] 0.29
```

Observamos que la estadística de prueba tiene un valor de 0.29. El cuantil W que acumula $1-\alpha$ de probabilidad, usando $\alpha = 0.05$ es W $= 0.409$, encontrado en las tablas correspondientes. Por lo tanto tenemos que $T_1 < W$, entonces no rechazamos la hipótesis nula y su correspondiente p-value es mucho mayor a 0.05, por lo tanto con $\alpha = 5\,\%$ no rechazaremos H_0 y concluimos no existe evidencia suficiente para suponer que la distribución de la muestra no es uniforme en el intervalo (0,1).

Ahora haremos la prueba usando la función **"ks.test"** de R.

```
#Prueba
ks.test(X_i,Sn)
```

```
    Two-sample Kolmogorov-Smirnov test

data:  X_i and Sn
D = 0.3, p-value = 0.7869
alternative hypothesis: two-sided
```

17.7. Otro ejemplo en R

Se mencionó que uno de los usos de estas pruebas es para validar el supuesto de normalidad en los modelos de regresión lineal. Veremos ahora un ejemplo en donde los datos a los que se aplica la prueba de bondad de ajuste son los residuales de una regresión lineal simple.

La base de datos **"Loblolly"** en R, contiene información sobre tres características de arboles de pino originarios del sudeste de Estados Unidos. Al ajustar un modelo de regresión lineal simple entre $X =$ "*edad*" y $Y =$ "*altura*", deseamos probar con un nivel de significancia del 1 % que los residuales estandarizados se distribuyen normal estándar.

```
m1=lm(Loblolly$height~Loblolly$age)
x=rstandard(m1)

ks.test(x,"pnorm",0,1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  x
D = 0.10804, p-value = 0.2613
alternative hypothesis: two-sided
```

Observamos que la estadística de prueba tiene un valor de 0.10804 y su correspondiente $p - value$ es 26.13 %, por lo tanto con $\alpha = 5\,\%$ no rechazaremos H_0 y concluimos no existe evidencia suficiente para suponer que la distribución de la muestra no es normal(0,1).

17.8. Ejercicios

1. Dada la siguiente muestra

0.6379, 1.5299, 0.35005, 2.0505, 2.1906, 0.3459, 2.3214, 0.3128 0.6548, 2.4373, 1.803, 2.3674, 1.2716, 0.2566, 0.2513.

Se desea hacer el siguiente contraste:

$$\mathbf{H}_0: \text{ Los datos } \sim LogN(0,1) \quad vs \quad \mathbf{H}_a: \text{Los datos } \nsim LogN(0,1)$$

Realice la prueba de Kolmogorov al 5 % de significancia.

2. Se desea probar la hipótesis de que los tiempos entre las llegadas de los pacientes a un hospital con una emergencia se distribuyen exponencial con media $\bar{x}$. Para ello se registró el tiempo transcurrido entre las llegadas sucesivas de pacientes en una mañana. El tiempo en minutos es el siguiente:

14.3,	38.0,	3.8,
10.8,	6.1,	10.1,
3.6,	6.2,	12.8,
22.1,	4.2,	4.6,
1.5,	3.3,	1.2,
20.0,	7.1,	8.1.

Pruebe la hipótesis con un nivel se significancia del 5 %.

Capítulo 18

Prueba Kolmogorov-Smirnov

En ejercicios prácticos es muy difícil conocer la distribución de una muestra aleatoria, generalmente sólo se tiene la información; ésta hay que procesarla para averiguar si sigue una determinada distribución probabilística, en un primer intento se ajustó mediante la prueba de la Ji-cuadrada, sin embargo, al ser una de las pruebas más sencillas su **"potencia"** al estimar una determinada distribución es baja, es por ello, que se idearon otros métodos y uno de ello es la Prueba de Kolmogorov-Smirnov.

La prueba de Kolmogorov presenta la ventaja de que los datos no deben ser categorizadas para poder realizar estimaciones en su distribución. Al igual que en la prueba de la Ji-Cuadrada, Kolmogorov-Smirnov trabaja con una distribución $F^*(x)$ totalmente especificada, es decir, se debe de tener sospecha de que la muestra aleatoria siga una determinada distribución. De esta manera el objeto de estudio es una muestra $X_1, \ldots, X_n$ de variables aleatorias idénticamente distribuidas, las cuales siguen una distribución desconocida $F(X)$ y se tiene la sospecha de que la muestra sigue una distribución conocida $F^*(x)$.

Para probar la suposición de la distribución $F^*(x)$ se realiza la siguiente contraste:

18.1. Hipótesis

18.1.1. Caso A (Prueba de 2 colas)

Solo será este caso

$$\mathbf{H}_0 : \ F(x) = F^*(x) \quad \forall \ x \ \text{ de } \ -\infty \ \text{ a } \ +\infty$$

$$vs$$

$$\mathbf{H}_a : \ F(x) \neq F^*(x) \quad \text{para al menos un valor de } x.$$

Regla de decisión

Rechazo H_0 a un nivel de significancia α si $D_n > W$. Donde W es el cuantil $(1-\alpha)$, obtenido en la tabla correspondiente a nuestra prueba O para la prueba de 2 colas en las tablas de Kolmogorov.

Donde $F^*(x)$ es una distribución completamente conocida, es decir además de conocer a la familia que pertenece también se conocen sus parámetros.

Lo que se busca es poder medir las distancia entre $F(x)$, la distribución desconocida, con los datos que siguen la función de distribución propuesta y completamente conocida $F^*(x)$. Sin embargo, $F(x)$ al ser desconocida se recurre a la construcción de una distribución empírica la cual se define como:

$$S_n(x) = \frac{\sum_{i=1}^{n} número\ de\ valores\ muestrales\ \leq x}{n}.$$

Es decir, la función empírica mide el número de elementos menores o iguales a la observación X, puede observarse que en el caso continuo, al no haber **"empates"** la función empírica puede ser vista como:

$$S_n(x) = \frac{i}{n} \quad i = 1, \ldots, n.$$

Al tener una distribución desconocida $F(x)$, la función empírica $S_n(x)$ puede ser usada como un estimador insesgado de $F(x)$ pues:

$$\mathbb{E}(S_n(x)) = F(x).$$

La función empírica es de gran importancia ya que gracias al teorema de **Glivenko-Cantelli** se sabe que cuando el tamaño de la muestra tiende a infinito cualquier distribución empírica se aproxima a la distribución real de los datos, la cual, es una distribución completamente especificada. El teorema de **Glivenko-Cantelli**, menciona que al calcular las diferencias de la distribución real y la empírica éstas son cero en cada observación dada, el teorema que se enuncia como:

Sea $X_1, \ldots, X_n$ una muestra aleatoria de distribución $F(x)$ desconocida y sea $S_n(x)$ la función empírica entonces:

$$\sup_x \mid S_n(x) - F(x) \mid \longrightarrow 0.$$

Es decir, conforme mayor sea el tamaño de la muestra, $S_n(x)$ reproduce la verdadera distribución. De esta manera se establece el estadístico de Prueba Dn, el cual no depende de ningún parámetro desconocido, ya que engloba a la distribución empírica y a la distribución propuesta:

$$D_n = \sup_x \mid S_n(x) - F^*(x) \mid = max\ \{\ max\{\ S_n(X_{i-1}) - F^*(X_i)\ \}, max\{\ S_n(X_i) - F^*(X_i)\ \}\ \} \quad \forall\ i.$$

y

$$D_n = \sup_x \mid S_n(x) - F^*(x) \mid = max\ \{\ D^+, D^-\ \}.$$

Donde:

$$D^+ = max\ \{\ S_n(X_i) - F^*(X_i)\ \}.$$
$$D^- = max\ \{\ S_n(X_{i-1}) - F^*(X_i)\ \}.$$

Finalmente, se observa que si H_0 es cierta si $D_n \longrightarrow 0$ ya que las diferencias entre la diferencias entre la función empírica y la propuestas son mínimas, lo que cumple con el **Teorema de Glivenko-Cantelli** ; por lo que hay evidencia para rechazar H_0 cuando $D_n > W$ Donde W es el cuantil que acumula el $1 - \alpha$ de probabilidad de la distribución asociada a D_n la cual puede obtenerse de la tablas correspondientes la cual muestra los cuantiles de la distribución Kolmogorov-Smirnov o la Tabla Kolmogorov para 2 colas.

18.2. Ejemplo

Dada la siguiente muestra

0.6379 1.5299 0.35005 2.0505 2.1906 0.3459 2.3214 0.3128

0.65482.4373 1.803 2.3674 1.2716 0.2566 0.2513

Se desea hacer el siguiente contraste:

$$\mathbf{H}_0: \text{ Los datos } \sim \ LogN(0,1).$$

$$vs$$

$$\mathbf{H}_a: \text{ Los datos } \nsim \ LogN(0,1).$$

Realice la prueba de Kolmogorov-Smirnov al 5 % de significancia.

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Kolmogorov-Smirnov.**

Paso 2 Planteamiento de Hipótesis:

$$\mathbf{H}_0: \text{ Los datos } \sim \ LogN(0,1).$$

$$vs$$

$$\mathbf{H}_a: \text{ Los datos } \nsim \ LogN(0,1).$$

Paso 3 Estadístico de Prueba:

$$D_n = \sup_x \mid S_n(x) - F^*(x) \mid = max \ \{ \ D^+, D^- \ \}.$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba:

1) Se procede a ordenar nuestras observaciones de menor a mayor.
2) Se calcula la función empírica, como no tenemos ningún valor repetido:

$$S_n = \frac{i}{n} = \frac{1}{15}, \frac{2}{15}, \ldots, 1.$$

3) Se calcula la función empírica menos un valor, es decir,

$$S_n = \frac{i-1}{n} = \frac{0}{15}, \frac{1}{15}, \ldots, \frac{14}{15}.$$

4) Se calcula la distribución conocida, es decir, $F^*(x) \quad LogN(0,1)$.
5) Se calcula D^+ que es el resultado de la resta de la distribución conocida menos la distribución empírica, es decir:

$$D^+ = max \ \{ \ S_n(X_i) - F^*(X_i) \ \}.$$

6) Se calcula D^- que es el resultado de la resta de la distribución empírica menos uno menos la distribución conocida, es decir:

$$D^- = max \ \{ \ S_n(X_{i-1}) - F^*(X_i) \ \}.$$

7) Finalmente realizada la tabla, se calcula el máximo de las columnas D^+ y D^- de ésta manera, se tiene la siguiente tabla:
8) Entonces,

i	X	X_i	$S_n(X_i)$	$S_n(X_{i-1})$	$F^*(X_i)$	$D^+=S_n(X_i)-F^*(X_i)$	$D^-=S_n$
1	0.63790	0.25130	0.0666667	0.0000000	0.0836	-0.0169	
2	1.52990	0.25660	0.1333333	0.0666667	0.0869	0.0464	
3	0.35005	0.31280	0.2000000	0.1333333	0.1226	0.0774	
4	2.05050	0.34590	0.2666667	0.2000000	0.1442	0.1224	
5	2.19060	0.35005	0.3333333	0.2666667	0.1472	0.1861	
6	0.34590	0.63790	0.4000000	0.3333333	0.3265	0.0735	
7	2.32140	0.65480	0.4666667	0.4000000	0.3360	0.1306	
8	0.31280	1.27160	0.5333333	0.4666667	0.5949	-0.0615	
9	0.65480	1.52990	0.6000000	0.5333333	0.6647	-0.0647	
10	2.43730	1.80300	0.6666667	0.6000000	0.7222	-0.0555	
11	1.80300	2.05050	0.7333333	0.6666667	0.7636	-0.0302	
12	2.36740	2.19060	0.8000000	0.7333333	0.7835	0.0165	
13	1.27160	2.32140	0.8666667	0.8000000	0.8002	0.0664	
14	0.25660	2.36740	0.9333333	0.8666667	0.8056	0.1277	
15	0.25130	2.43730	1.0000000	0.9333333	0.8135	0.1865	

$$D^+ = max\ \{\ S_n(X_i) - F^*(X_i)\ \} = 0.1865 \quad y \quad D^- = max\ \{\ S_n(X_{i-1}) - F^*(X_i)\ \} = 0.1198.$$

Por lo tanto:

$$D_n = \sup_x \mid S_n(x) - F^*(x) \mid = max\ \{\ D^+, D^-\ \} = max\ \{\ 0.1865, 0.1198\ \} = 0.1865$$

Paso 5 Regla de Decisión.

Este último resultado se compara con la tabla de valores críticos de la Tabla Kolmogorov-Smirnov, para un nivel de significancia $\alpha = 0.05$, $W_{0.05}$=0.338, de esta manera se tiene que $0.338 = W_{0.05} > D_n = 0.1865$, como el estadístico $W_{0.05}$ es mayor a comparación de D_n=0.1865. No Rechazamos H_0.

Paso 6 Conclusión.

Se acepta la prueba de lognormalidad con media 1 y varianza 0, con un nivel de significancia $\alpha = 0.05$. Es decir, Los datos se distribuyen $LogN(0, 1)$.

```
i= 1:15
X= c(0.6379,1.5299,0.35005,2.0505,2.1906,0.3459,2.3214,0.3128,0.6548,2.4373,
               1.803,2.3674,1.2716,0.2566,0.2513)
X_i=c(0.25130,0.25660,0.31280, 0.34590,0.35005,0.63790, 0.65480, 1.27160,
               1.52990,1.80300, 2.05050, 2.19060, 2.32140, 2.36740,2.43730)

Sn=c(1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15,1)
Sn_1=c(0/15,1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15)
F_=c(0.0836,0.0869,0.1226,0.1442,0.1472,0.3265,0.3360,0.5949,
                    0.6647,0.7222,0.7636,0.7835,0.8002,0.8056,0.8135)
D_mas=c(-0.0169,0.0464,0.0774,0.1224,0.1861,0.0735,0.1306,-0.0615,-0.0647,-0.0555,
                -0.0302,0.0165,0.0664,0.1277,0.1865)
D_menos=c(-0.0836,-0.0202,0.0107,0.0558,0.1194,0.0068,0.0640,-0.1282,
          -0.1313,-0.1222,-0.0969,-0.0501,-0.0002,0.0610,0.1198)

Tabla=cbind(i,X_i,Sn,Sn_1,F_,D_mas,D_menos)
Tabla
```

```
      i     X_i         Sn        Sn_1     F_   D_mas D_menos
 [1,] 1 0.25130 0.06666667 0.00000000 0.0836 -0.0169 -0.0836
 [2,] 2 0.25660 0.13333333 0.06666667 0.0869  0.0464 -0.0202
 [3,] 3 0.31280 0.20000000 0.13333333 0.1226  0.0774  0.0107
```

```
 [4,]  4 0.34590 0.26666667 0.20000000 0.1442  0.1224  0.0558
 [5,]  5 0.35005 0.33333333 0.26666667 0.1472  0.1861  0.1194
 [6,]  6 0.63790 0.40000000 0.33333333 0.3265  0.0735  0.0068
 [7,]  7 0.65480 0.46666667 0.40000000 0.3360  0.1306  0.0640
 [8,]  8 1.27160 0.53333333 0.46666667 0.5949 -0.0615 -0.1282
 [9,]  9 1.52990 0.60000000 0.53333333 0.6647 -0.0647 -0.1313
[10,] 10 1.80300 0.66666667 0.60000000 0.7222 -0.0555 -0.1222
[11,] 11 2.05050 0.73333333 0.66666667 0.7636 -0.0302 -0.0969
[12,] 12 2.19060 0.80000000 0.73333333 0.7835  0.0165 -0.0501
[13,] 13 2.32140 0.86666667 0.80000000 0.8002  0.0664 -0.0002
[14,] 14 2.36740 0.93333333 0.86666667 0.8056  0.1277  0.0610
[15,] 15 2.43730 1.00000000 0.93333333 0.8135  0.1865  0.1198
```

```r
EstdPrueba = max(D_mas,D_menos)
EstdPrueba
```

```
[1] 0.1865
```

Por lo tanto:

$$D_n = \sup_x \mid S_n(x) - F^*(x) \mid = max \{ D^+, D^- \} = max \{ 0.1865, 0.1198 \} = 0.1865$$

Lo comparamos con la tabla de valores críticos de la **Tabla Kolmogorov-Smirnov**, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}$=0.338, de esta manera se tiene que $0.338 = W_{0.05} > D_n = 0.1865$, como el estadístico $W_{0.05}$ es mayor a comparación de D_n=0.1865. No Rechazamos H_0.

Podemos concluir aceptando la prueba de lognormalidad con media 1 y varianza 0, con un nivel de significancia $\alpha = 0.05$. Es decir, Los datos se distribuyen $LogN(0, 1)$.

```r
#Ahora podemos utilizar la prueba en R
ks.test(F_,Sn,alternative = "two.sided")
```

```
    Two-sample Kolmogorov-Smirnov test

data:  F_ and Sn
D = 0.2, p-value = 0.9383
alternative hypothesis: two-sided
```

Capítulo 19

Prueba Lilliefors para Normalidad

La prueba de bondad de ajuste de Kolmogorov presentada anteriormente es una buena prueba para ver si una muestra aleatoria tiene alguna función de distribución especificada. La prueba de Kolmogorov está diseñada para usarse solo cuando la función de distribución hipotética está completamente especificada, es decir, cuando no hay parámetros desconocidos que deben estimarse a partir de la muestra. La prueba de bondad de ajuste Ji-cuadrada es lo suficientemente flexible como para permitir que algunos parámetros se estimen a partir de los datos. Simplemente se resta un grado de libertad para cada parámetro estimado descrita anteriormente.

Sin embargo, la prueba de Ji-cuadrada requiere que los datos se agrupen, y dicha agrupación de datos suele ser arbitraria. Además, la distribución del estadístico de prueba se conoce solo aproximadamente, y a veces el poder de la prueba de Ji-cuadrada no es muy bueno. Por estas razones, se buscan otras pruebas de bondad de ajuste, especialmente para distribuciones probadas con frecuencia.

La prueba de Kolmogorov se ha modificado para permitir su uso en varias situaciones en las que los parámetros se estiman a partir de los datos. En realidad, el estadístico de prueba permanece sin cambios, pero se utilizan diferentes tablas de valores críticos. Estas tablas ya no son las mismas para todas las distribuciones; cambian de una distribución hipotética a otra. La prueba sigue siendo una prueba no paramétrica porque la validez de la prueba (el nivel de α) no depende de supuestos no probados con respecto a la distribución de la población; en cambio, la forma de distribución de la población es la hipótesis que se está probando.

La primera modificación a la prueba de Kolmogorov es para probar la hipótesis compuesta de la normalidad. Es decir, la hipótesis nula establece que la población es una de la familia de distribuciones normales sin especificar la media o la varianza de la distribución normal. Ésta prueba fue presentada por primera vez por Hubert Lilliefors (1967). Una característica interesante de esta prueba es que este es uno de los primeros casos en que el compilador se utilizó para generar números aleatorios con el fin de obtener estimaciones precisas de los verdaderos cuantiles de la distribución exacta del estadístico de la prueba y además, aproximar a los parámetros a través del uso de los estimadores puntuales.

19.1. Datos

Los datos consisten en una muestra aleatoria $X_1, X_2, \ldots, X_n$ de tamaño n asociada con alguna función de distribución desconocida, denotada por $F(x)$.

Calculando la media muestral

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

para usarla como una estimación puntual de μ

y calculando como una estimación de σ

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Después calcularemos los valores de muestra "normalizados" Z_i definidos por:

$$Z_i = \frac{X_i - \overline{X}}{s} \quad i = 1, 2, \ldots, n.$$

El estadístico de prueba se calcula a partir de $Z_i s$ en lugar de a partir de la muestra aleatoria original.

19.2. Supuestos

1) La muestra es una muestra aleatoria.

19.3. Hipótesis:

$\mathbf{H}_0$: La muestra aleatoria proviene de una población con distribución normal, con media y desviación estándar desconocidas.

vs

$\mathbf{H}_a$: La función de distribución de las $X_i s$ no es normal.

19.4. Estadístico de Prueba.

Normalmente, el estadístico de prueba es el mismo para la prueba de Kolmogorov de dos colas, definida como la distancia vertical máxima entre la función de distribución empírica de $X_i s$ y la función de distribución normal con media $\mu = \overline{X}$ y desviación estándar $\sigma = s$. Sin embargo, el siguiente método para calcular el estadístico de prueba es un poco más fácil, ya que es equivalente al método indicado. Es decir, el cálculo del estadístico T_1 será en función de la $Z_i s$.

$$T_1 = \sup_x |F^*(x) - S(x)|.$$

Regla de Decisión.

Rechazo H_0 a un nivel de significancia α si $T_1 > W_{1-\alpha}$ donde $W_{1-\alpha}$ es el cuantil obtenido en las tablas correspondientes a nuestra prueba.

19.5. Ejemplo

Los siguientes datos, corresponden a una muestra aleatoria en la que mide la pérdida y ganancia de peso en KG de un grupo después de vacaciones.

$$0.6822,\ 3.994,\ -0.9705,\ -0.5575,\ -2.1532,\ 0.0829,\ 2.9224,\ 0.2425$$
$$-0.4962,\ -0.1621,\ 0.449,\ -0.8827,\ -0.8368,\ -1.5805,\ 0.386.$$

Se desea probar si los datos provienen de una distribución normal con μ y σ desconocidas. Realizar la prueba a un nivel de significancia del 95 %.

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Lilliefors para Normalidad.**

Paso 2 Planteamiento de Hipótesis:

$$\mathbf{H}_0: \text{ La muestra } \sim N(\mu, \sigma^2).$$

$$vs$$

$$\mathbf{H}_a: \text{ La muestra } \nsim N(\mu, \sigma^2).$$

Paso 3 Estadístico de Prueba.

$$T_1 = \sup_z \mid F^*(z) - S(z) \mid.$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba:

1) Se procede a ordenar nuestras observaciones de menor a mayor.
2) Se obtienen los estimadores puntuales de distribución normal con los datos de la muestra,

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 0.07463333$ y $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} = 1.590808$.

3) Después calcularemos los valores de muestra "normalizados" Z_i definidos por:

$$Z_i = \frac{X_i - \overline{X}}{s} \qquad i = 1, 2, \ldots, 15.$$

4) Se calcula la función empírica, como no tenemos ningún valor repetido:

$$S_n = \frac{i}{n} = \frac{1}{15}, \frac{2}{15}, \ldots, 1.$$

5) Se calcula la función empírica menos un valor, es decir,

$$S_n = \frac{i-1}{n} = \frac{0}{15}, \frac{1}{15}, \ldots, \frac{14}{15}.$$

6) Se calcula D^+ que es el resultado de la resta de la distribución conocida menos la distribución empírica, es decir:

$$D^+ = max \{ S_n(Z_i) - F^*(Z_i) \}.$$

7) Se calcula D^- que es el resultado de la resta de la distribución empírica menos uno menos la distribución conocida, es decir:

$$D^- = max \{ S_n(Z_{i-1}) - F^*(Z_i) \}.$$

8) Finalmente realizada la tabla, se calcula el máximo de las columnas D^+ y D^- de ésta manera, se tiene la siguiente tabla:
9) Entonces:

$$D^+ = max \{ S_n(Z_i) - F^*(Z_i) \} = 0.2179 \quad y \quad D^- = max \{ S_n(Z_{i-1}) - F^*(Z_i) \} = 0.1513.$$

Por lo tanto:

$$T_1 = \sup_x \mid S_n(z) - F^*(z) \mid = max \{ D^+, D^- \} = max \{ 0.2179, 0.1513 \} = 0.2179.$$

i	X	X_i	$Z_i=\frac{x_i-\bar{x}}{s}$	$S_n(Z_i)$	$S_n(Z_{i-1})$	$F^*(Z_i)$	$D^+=S_n
1	0.6822	-2.1532	-1.4004	0.0666667	0.0000000	0.0806	-(
2	3.9940	-1.5805	-1.0404	0.1333333	0.0666667	0.1490	-(
3	-0.9705	-0.9705	-0.6569	0.2000000	0.1333333	0.2555	-(
4	-0.5575	-0.8827	-0.6017	0.2666667	0.2000000	0.2736	-(
5	-2.1532	-0.8368	-0.5729	0.3333333	0.2666667	0.2833	(
6	0.0829	-0.5575	-0.3973	0.4000000	0.3333333	0.3455	(
7	2.9224	-0.4962	-0.3588	0.4666667	0.4000000	0.3598	(
8	0.2425	-0.1621	-0.1488	0.5333333	0.4666667	0.4408	(
9	-0.4962	0.0829	0.0051	0.6000000	0.5333333	0.5020	(
10	-0.1621	0.2425	0.1055	0.6666667	0.6000000	0.5420	(
11	0.4490	0.3860	0.1957	0.7333333	0.6666667	0.5775	(
12	-0.8827	0.4490	0.2353	0.8000000	0.7333333	0.5930	(
13	-0.8368	0.6822	0.3819	0.8666667	0.8000000	0.6487	(
14	-1.5805	2.9224	1.7901	0.9333333	0.8666667	0.9632	-(
15	0.3860	3.9940	2.4637	1.0000000	0.9333333	0.9931	(

Paso 5 Regla de Decisión.

Este último resultado se compara con la tabla de valores críticos de la Tabla Lilliefors para Normalidad, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}$=0.2190, de esta manera se tiene que 0.2190 = $W_{0.05} > T_1$ = 0.2179, como el cuantil $W_{0.05}$ es mayor a comparación de T_1=0.2179.

No Rechazamos H_0.

Paso 6 Conclusión.

Podemos concluir que a un nivel de significancia α =0.05, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución normal.

19.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Datos

i= 1:15
X=c(0.6822, 3.994, -0.9705, -0.5575, -2.1532, 0.0829, 2.9224,
    0.2425,-0.4962, -0.1621, 0.449, -0.8827, -0.8368, -1.5805, 0.386)
X_i=c(-2.1532,-1.5805,-0.9705,-0.8827,-0.8368,-0.5575,-0.4962, -0.1621,  0.0829,  0.2425,
      0.3860, 0.4490,  0.6822,2.9224,3.9940)
Z_i=c(-1.4004, -1.0404,-0.6569,-0.6017,-0.5729,-0.3973,-0.3588,-0.1488,0.0051,
0.1055,0.1957,0.2353,0.3819,1.7901,  2.4637) #Calculado xi-xbarra/s
Sn_Zi=c(1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15,1)
Sn_Zi_1=c(0/15,1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15)
F_Zi=c(0.0806, 0.1490,0.2555,0.2736,0.2833,0.3455,0.3598,0.4408,0.5020,0.5420,0.5775,0.5930,
       0.6487,0.9632, 0.9931)
D_mas=c(-0.0139,-0.0156,-0.0555,-0.0069,0.0500,0.0545,0.1068,0.0925,0.0980,0.1246, 0.1558,
        0.2070,0.2179,-0.0298,0.0069)    #Sn(Zi)-F*(Zi)
D_menos=c(-0.0806,-0.0823,-0.1221,-0.0736,-0.0166,-0.0121,0.0402,0.0258,0.0313,
          0.0580,0.0891,0.1403,0.1513,-0.0965,-0.0597)   #Sn(Zi-1)-F*(Zi)

Tabla=cbind(i,X_i,Z_i,Sn_Zi,Sn_Zi_1,F_Zi,D_mas,D_menos)
Tabla
```

```
      i     X_i     Z_i       Sn_Zi    Sn_Zi_1   F_Zi   D_mas D_menos
 [1,]  1 -2.1532 -1.4004 0.06666667 0.00000000 0.0806 -0.0139 -0.0806
```

```
 [2,]  2 -1.5805 -1.0404 0.13333333 0.06666667 0.1490 -0.0156 -0.0823
 [3,]  3 -0.9705 -0.6569 0.20000000 0.13333333 0.2555 -0.0555 -0.1221
 [4,]  4 -0.8827 -0.6017 0.26666667 0.20000000 0.2736 -0.0069 -0.0736
 [5,]  5 -0.8368 -0.5729 0.33333333 0.26666667 0.2833  0.0500 -0.0166
 [6,]  6 -0.5575 -0.3973 0.40000000 0.33333333 0.3455  0.0545 -0.0121
 [7,]  7 -0.4962 -0.3588 0.46666667 0.40000000 0.3598  0.1068  0.0402
 [8,]  8 -0.1621 -0.1488 0.53333333 0.46666667 0.4408  0.0925  0.0258
 [9,]  9  0.0829  0.0051 0.60000000 0.53333333 0.5020  0.0980  0.0313
[10,] 10  0.2425  0.1055 0.66666667 0.60000000 0.5420  0.1246  0.0580
[11,] 11  0.3860  0.1957 0.73333333 0.66666667 0.5775  0.1558  0.0891
[12,] 12  0.4490  0.2353 0.80000000 0.73333333 0.5930  0.2070  0.1403
[13,] 13  0.6822  0.3819 0.86666667 0.80000000 0.6487  0.2179  0.1513
[14,] 14  2.9224  1.7901 0.93333333 0.86666667 0.9632 -0.0298 -0.0965
[15,] 15  3.9940  2.4637 1.00000000 0.93333333 0.9931  0.0069 -0.0597
```

```
EstdPrueba=max(D_mas,D_menos)
EstdPrueba
```

```
[1] 0.2179
```

Por lo tanto:

$$T_1 = \sup_x \mid S_n(z) - F^*(z) \mid = max \{ D^+, D^- \} = max \{ 0.2179, 0.1513 \} = 0.2179$$

Lo vamos a comparar con la tabla de valores críticos de la Tabla Lilliefors para Normalidad, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}$=0.2190, de esta manera se tiene que $0.2190 = W_{0.05} > T_1 = 0.2179$, como el cuantil $W_{0.05}$ es mayor a comparación de T_1=0.2179. Entonces no Rechazamos H_0 y podemos concluir que a un nivel de significancia α =0.05, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución normal.

Ahora podemos utilizar la prueba en R

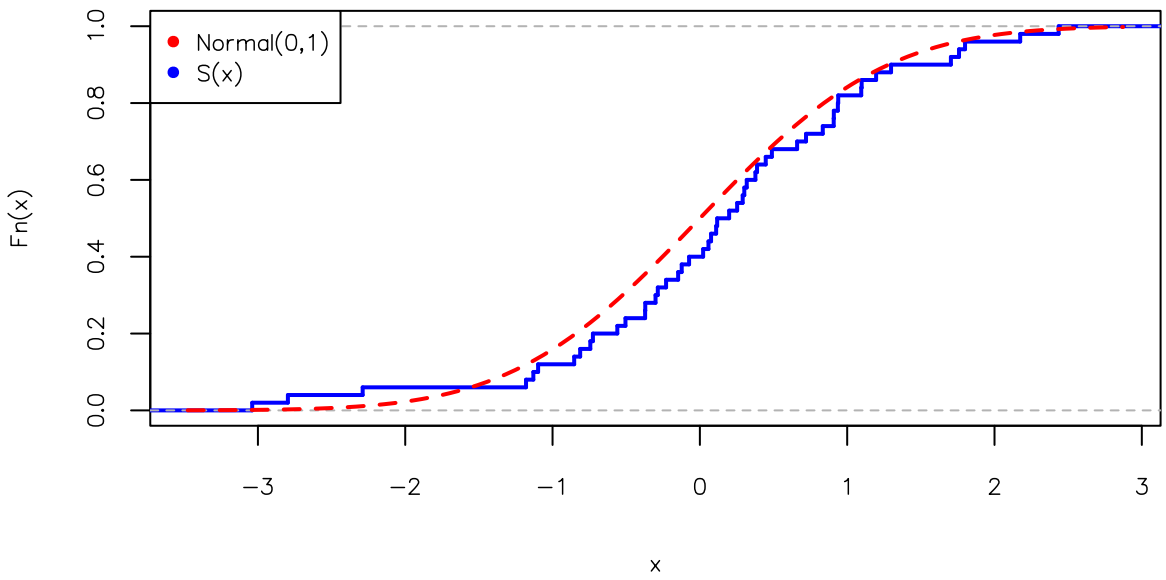
```
library(nortest) #prueba lilliefors
lillie.test(X)
```

```

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  X
D = 0.21793, p-value = 0.05356
```

##Otro ejemplo en R En R fije la semilla 2020, y genere 250 observaciones distribuidas como una $N(0,1)$ y con ella realice:

1. Grafique la función de distribución empírica de las observaciones generadas.
2. Agregar sobre esa misma gráfica, la curva de la distribución verdadera $N(0,1)$.
3. Realizar la prueba Lilliefors de bondad de ajuste para probar que la muestra proviene de una distribución $N(0,1)$.

```
#Semilla y simulación
set.seed(2020)
x=rnorm(50,0,1)
#Gráfico de la función de distribución empírica
plot.ecdf(x,col="blue",verticals = TRUE, do.points = FALSE, lwd=2, main="")
#Gráfico de la distribución verdadera $N(0,1)$
curve(pnorm,add=TRUE, col="red", lty="dashed", lwd=2)
legend("topleft", c("Normal(0,1)","S(x)"), cex=0.8, col=c("red","blue"), pch=c(16,16))
```




```
#Prueba
library(nortest)
T1<-lillie.test(x)
T1
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  x
D = 0.08739, p-value = 0.4436
```

Observamos que la estadística de prueba tiene un valor de 0.08739 y su correspondiente p-value es 44.36 %, por lo tanto no rechazaremos H_0 y concluimos no existe evidencia suficiente para suponer que la distribución de la muestra no es $N(0,1)$. Esperabamos que la prueba diera ese resultado ya que por construcción la muestra proviene de una distribucion $N(0,1)$.

19.7. Ejercicios

1. Se desea probar la hipótesis de que las tasas de interés de un determinado producto financiero tiene el comportamiento de una variable aleatoria como función de distribución normal.

9.1, 5, 7.3, 7.4, 5.5, 8.6, 7, 4.3, 4.7, 8, 4, 8.5, 6.4,6.1, 5.8, 9.5, 5.2, 6.7, 8.3, 9.2.

Capítulo 20

Pueba de Lilliefors Exponencial

Otro problema importante de bondad de ajuste es la prueba para la distribución exponencial con media no especificada. Este problema es importante porque el supuesto de una distribución exponencial con media desconocida tiene muchas aplicaciones, particularmente donde las variables aleatorias bajo estudio representan el tiempo de espera o el tiempo en que ocurre cierto evento. Lilliefors en 1969 desarrolló una prueba análoga a la prueba de Kolmogorov-Smirnov y dió una tabla de valores críticos basados en simulaciones Monte Carlo.

20.1. Datos

Los datos consisten en una muestra aleatoria $X_1, X_2, \ldots, X_n$ de tamaño n que sigue una distribución exponencial con media $\hat{\lambda} = \frac{1}{\overline{X}}$,el cual corresponde al estimador puntual de la media.

Después calcularemos los valores de muestra Z_i definidos por:

$$Z_i = \frac{X_i}{\overline{X}} \quad i = 1, 2, \ldots, n.$$

El estadístico de prueba se calcula a partir de $Z_i s$ en lugar de a partir de la muestra aleatoria. original.

20.2. Supuestos

1) La muestra es una muestra aleatoria.

20.3. Hipótesis

$\mathbf{H}_0$: La muestra aleatoria proviene de una población con distribución exponencial:

$$\mathbf{F(x)} = \begin{cases} 1 - e^{-\frac{x}{\lambda}} & si \quad x > 0 \\ x & si \quad x < 0 \end{cases}$$

Donde λ es un parámetro desconocido.

vs

$\mathbf{H}_a$: La función de distribución de las $X_i s$ no es exponencial.

20.4. Estadistico de Prueba.

El estadístico de prueba está dada por la máxima distancia vertical:

$$T_2 = \underset{x}{sup} \mid F^*(x) - S(x) \mid.$$

Regla de Decisión

Rechazo H_0 a un nivel de significancia α si $T_2 > W_{1-\alpha}$ donde $W_{1-\alpha}$ es el cuantil obtenido en las tablas correspondientes a nuestra prueba.

20.5. Ejemplo

Dada la siguiente muestra:

$$0.4976,\ 1.2514,\ 0.6619,\ 0.561,\ 1.0026,\ 0.3529,\ 0.8595,\ 1.6254,$$
$$1.1514,\ 1.5181,\ 0.8642,\ 0.5206,\ 0.4229,\ 0.9825,\ 1.0183.$$

Se desea probar si la muestra sigue una distribución exponencial con parámetro λ desconocido.

Realizar la prueba con un nivel de significancia del 95 %.

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Lilliefors Exponencial.**

Paso 2 Planteamiento de Hipótesis:

$\mathbf{H}_0$: La muestra aleatoria proviene de una población con distribución exponencial

con λ parámetro desconocido.

vs

$\mathbf{H}_a$: La función de distribución de las $X_i s$ no es exponencial.

Paso 3 Estadístico de Prueba:

$$T_2 = \underset{z}{sup} \mid F^*(z) - S(z) \mid.$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba:

1) Se procede a ordenar nuestras observaciones de menor a mayor.
2) Se obtiene el estimador puntual de la distribución exponencial con los datos de la muestra, por lo que el parámetro λ es calculado como:

$$\hat{\lambda} = \frac{1}{\overline{X}} = \frac{1}{0.88602}; \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{15}\sum_{i=1}^{15} X_i = 0.88602.$$

3) Después calcularemos los valores de muestra Z_i definidos por:

$$Z_i = \frac{X_i}{\overline{X}} \quad i = 1, 2, \ldots, 15$$

i	X	X_i	$Z_i=\frac{X_i}{\bar{x}}$	$S_n(Z_i)$	$S_n(Z_{i-1})$	$F^*(Z_i)$	$D^+=S_n(Z
1	0.4976	0.3529	0.3982	0.0666667	0.0000000	0.3285	-0.2
2	1.2514	0.4229	0.4773	0.1333333	0.0666667	0.3795	-0.2
3	0.6619	0.4976	0.5616	0.2000000	0.1333333	0.4297	-0.2
4	0.5610	0.5206	0.5875	0.2666667	0.2000000	0.4443	-0.1
5	1.0026	0.5610	0.6331	0.3333333	0.2666667	0.4690	-0.1
6	0.3529	0.6619	0.7470	0.4000000	0.3333333	0.5262	-0.1
7	0.8595	0.8595	0.9700	0.4666667	0.4000000	0.6209	-0.1
8	1.6254	0.8642	0.9753	0.5333333	0.4666667	0.6229	-0.0
9	1.1514	0.9825	1.1088	0.6000000	0.5333333	0.6700	-0.0
10	1.5181	1.0026	1.1315	0.6666667	0.6000000	0.6774	-0.0
11	0.8642	1.0183	1.1492	0.7333333	0.6666667	0.6831	0.0
12	0.5206	1.1514	1.2995	0.8000000	0.7333333	0.7273	0.0
13	0.4229	1.2514	1.4123	0.8666667	0.8000000	0.7564	0.1
14	0.9825	1.5181	1.7133	0.9333333	0.8666667	0.8197	0.1
15	1.0183	1.6254	1.8344	1.0000000	0.9333333	0.8403	0.1

4) Se calcula la función empírica, como no tenemos ningún valor repetido:

$$S_n = \frac{i}{n} = \frac{1}{15}, \frac{2}{15}, \ldots, 1.$$

5) Se calcula la función empírica menos un valor, es decir,

$$S_n = \frac{i-1}{n} = \frac{0}{15}, \frac{1}{15}, \ldots, \frac{14}{15}.$$

6) Se calcula D^+ que es el resultado de la resta de la distribución conocida menos la distribución empírica, es decir:

$$D^+ = max\ \{\ S_n(Z_i) - F^*(Z_i)\ \}.$$

7) Se calcula D^- que es el resultado de la resta de la distribución empírica menos uno menos la distribución conocida, es decir:

$$D^- = max\ \{\ S_n(Z_{i-1}) - F^*(Z_i)\ \}.$$

8) Finalmente realizada la tabla, se calcula el máximo de las columnas D^+ y D^- de ésta manera, se tiene la siguiente tabla:

9) Entonces

$$D^+ = max\ \{\ S_n(Z_i) - F^*(Z_i)\ \} = 0.1596 \quad y \quad D^- = max\ \{\ S_n(Z_{i-1}) - F^*(Z_i)\ \} = 0.0930$$

Por lo tanto:

$$T_2 = \sup_z \mid S_n(z) - F^*(z) \mid = max\ \{\ D^+, D^-\ \} = max\ \{\ 0.1596, 0.0930\ \} = 0.1596.$$

Paso 5 Regla de Decisión.

Este último resultado se compara con la tabla de valores críticos de la Tabla Lilliefors Exponencial, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}$=0.2776, de esta manera se tiene que $0.2776 = W_{0.05} > T_2 = 0.1596$, como el cuantil $W_{0.05}$=0.2776 es mayor a comparación de T_2=0.2776.

No Rechazamos H_0.

Paso 6 Conclusión.

Podemos concluir que a un nivel de significancia α=0.05, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución exponencial.

20.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Datos
i=1:15
X=c(0.4976, 1.2514, 0.6619, 0.561, 1.0026, 0.3529, 0.8595, 1.6254,
                            1.1514, 1.5181, 0.8642, 0.5206, 0.4229, 0.9825, 1.0183)
X_i=c(0.3529,0.4229,0.4976,0.5206,0.5610, 0.6619,0.8595,0.8642,0.9825,1.0026,1.0183,
      1.1514,1.2514,1.5181,1.6254)
Z_i=c(0.3982, 0.4773, 0.5616,0.5875,0.6331,0.7470,0.9700,0.9753,1.1088,1.1315,1.1492,
      1.2995,1.4123,1.7133,1.8344)
Sn_Zi=c(1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15,1)
Sn_Zi_1=c(0/15,1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15)
F_Zi=c(0.3285,0.3795,0.4297,0.4443,0.4690,0.5262,0.6209,0.6229,0.6700,0.6774,0.6831,
       0.7273,0.7564,0.8197, 0.8403)
D_mas=c(-0.2618,-0.2462,-0.2297,-0.1776,-0.1357,-0.1262,-0.1542,-0.0896,-0.0700,
        -0.0108,0.0501,0.0726,0.1102,0.1135,0.1596)
D_menos=c(-0.3285,-0.3128,-0.2963,-0.2443, -0.2024,-0.1929,-0.2209,-0.1562,
          -0.1367,-0.0774,-0.0164,0.0059,0.0435,0.0469,0.0930)

Tabla=cbind(i,X_i,Z_i,Sn_Zi,Sn_Zi_1,F_Zi,D_mas,D_menos)
Tabla
```

```
       i    X_i    Z_i      Sn_Zi    Sn_Zi_1   F_Zi   D_mas D_menos
 [1,]  1 0.3529 0.3982 0.06666667 0.00000000 0.3285 -0.2618 -0.3285
 [2,]  2 0.4229 0.4773 0.13333333 0.06666667 0.3795 -0.2462 -0.3128
 [3,]  3 0.4976 0.5616 0.20000000 0.13333333 0.4297 -0.2297 -0.2963
 [4,]  4 0.5206 0.5875 0.26666667 0.20000000 0.4443 -0.1776 -0.2443
 [5,]  5 0.5610 0.6331 0.33333333 0.26666667 0.4690 -0.1357 -0.2024
 [6,]  6 0.6619 0.7470 0.40000000 0.33333333 0.5262 -0.1262 -0.1929
 [7,]  7 0.8595 0.9700 0.46666667 0.40000000 0.6209 -0.1542 -0.2209
 [8,]  8 0.8642 0.9753 0.53333333 0.46666667 0.6229 -0.0896 -0.1562
 [9,]  9 0.9825 1.1088 0.60000000 0.53333333 0.6700 -0.0700 -0.1367
[10,] 10 1.0026 1.1315 0.66666667 0.60000000 0.6774 -0.0108 -0.0774
[11,] 11 1.0183 1.1492 0.73333333 0.66666667 0.6831  0.0501 -0.0164
[12,] 12 1.1514 1.2995 0.80000000 0.73333333 0.7273  0.0726  0.0059
[13,] 13 1.2514 1.4123 0.86666667 0.80000000 0.7564  0.1102  0.0435
[14,] 14 1.5181 1.7133 0.93333333 0.86666667 0.8197  0.1135  0.0469
[15,] 15 1.6254 1.8344 1.00000000 0.93333333 0.8403  0.1596  0.0930
```

```
EstPrueba=max(D_mas,D_menos)
EstPrueba
```

```
[1] 0.1596
```

```
EstdPrueba
```

```
[1] 0.2179
```

El cuantil W que acumula $1-\alpha$ de probabilidad, usando $\alpha = 0.05$ es $W = 0.2776$, encontrado en las tablas correspondientes. Por lo tanto tenemos que $T1 = 0.1596 < W = 0.2776$, entonces rechazamos la hipotesis nula. Y concluimos que hay evidencia suficiente para decir que la muestra no proviene de una distribución exponencial.

```
library(nortest) #prueba lilliefors
lillie.test(F_Zi) #Ocupamos estos datos ya que debemos recordar que
```

```

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  F_Zi
```

```
D = 0.15238, p-value = 0.458
```

```
                    #es una distribución exponencial y F_Zi son datos estandarizados
```

20.7. Ejercicios

1. El gerente de una tienda quiere probar la hipótesis de que los clientes llegan aleatoriamente a su tienda, para ello registro el tiempo transcurrido entre las llegadas sucesivas de clientes en una mañana. El tiempo en minutos es el siguiente:

3.6, 6.2, 12.9, 14.2, 38.0, 3.8,
10.8, 6.1, 10.1, 22.1, 4.2, 4.6,
1.4, 3.3, 8.2.

Pruebe la hipótesis nula de que el tiempo entre las llegadas de los clientes se distribuyen con función de distribución exponencial.

Capítulo 21

Prueba Anderson-Darling

La prueba de Anderson Darling, al igual que la prueba de Lilliefors sirve para probar la hipótesis de que una muestra aleatoria sigue una cierta distribución especificada.

21.1. Datos

Los datos consisten en una muestra aleatoria $X_1, X_2, \ldots, X_n$ de tamaño n asociada con alguna función de distribución desconocida, denotada por $F(x)$.

21.2. Supuestos

- La muestra es una muestra aleatoria.

21.3. Hipótesis

Caso A (Prueba de 2 colas) Solo será este caso

$$\mathbf{H}_0 : \ F(x) = F^*(x) \quad \forall \ x \text{ de } \ -\infty \ \text{ a } \ +\infty$$

$$vs$$

$$\mathbf{H}_a : \ F(x) \neq F^*(x) \quad \text{para al menos un valor de } x.$$

Donde $F^*(x)$ es la distribución teórica que se quiere probar con un nivel de significancia α.

Para probar dicha hipótesis Anderson propone examinar las diferencias al cuadrados entre la distribución empírica de los datos $S_n(X)$ y la distribución teórica propuesta y completamente especificada $F^*(X)$ y luego integrar respecto a la distribución propuesta. A este tipo de pruebas se les conoce como funciones de distribución empíricas cuadráticas **(QEDF)** por sus siglas en inglés. De esta manera la estadística de la prueba **Anderson-Darling** se obtiene de integrar la siguiente función **QEDF**:

$$A_n^2 = n \int_{-\infty}^{\infty} (Sn(X) - F^*(X))^2 \frac{1}{F^*(X)(1 - F^*(X))}.$$

Una característica importante es que se usa la expresión $\frac{1}{F^*(X)(1-F^*(X))}$ debido a que se busca que las colas de distribución tengan un peso cuantificablemente mayor, con la finalidad de detectar diferencias en las colas de la distribución.

21.4. Estadístico de Prueba.

Resolviendo la integral se obtiene la estadístico de la forma:

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\ [\ ln(F^*(X_{(i)})) + ln(1-F^*(X_{(n-i+1)}))\ ].$$

Dado que el estadístico no depende de $F(X)$ y sólo depende de n entonces la distribución asintótica de **Anderson-Darling** es la que se muestra a continuación, asimismo se mostrarán algunos ajustes a la estadística con la finalidad de que la prueba sea más potente para determinados casos:

Caso	**Ajuste en el estadístico**	**0.90**	**0.95**	**0.975**	**0.99**
Todos los parámetros conocidos	A_n^2 *para* $n \geq 5$	1.933	2.492	3.070	3.857
Normal con $N(\overline{X}, S^2)$	$(1+\frac{4}{n}+\frac{25}{n^2})A_n^2$	0.632	0.751	0.870	1.029
Exponencial con $exp(\overline{X})$	$(1+\frac{0.6}{n})A_n^2$	1.070	1.326	1.587	1.943
Weibull con $Weibull(\hat{\alpha}, \hat{\beta})$	$(1+\frac{0.2}{\sqrt{n}})A_n^2$	0.637	0.757	0.877	1.038
Log-logística con $Log-log(\hat{\alpha}, \hat{\beta})$	$(1+\frac{0.25}{\sqrt{n}})A_n^2$	0.563	0.660	0.769	0.906

21.5. Regla de Decisión.

Rechazo H_0 a un nivel de significancia α si $A_n^2 > A_{1-\alpha}$ donde A_n^2 ya tiene el ajuste para cada caso mencionado anteriormente y $A_{1-\alpha}$ es el cuantil obtenido en las tablas correspondientes a nuestra prueba.

Para ejemplificar la prueba de $Anderson-Darling$ veamos el siguiente ejemplo:

21.5.1. Ejemplo

Se desea probar si la siguiente muestra:

$$-4.1302,\ \ 9.315,\ \ 3.9757,\ \ 8.49,\ \ 5.6204,\ \ -6.9098,\ \ -0.1426,\ \ -2.3838,$$
$$-2.0039,\ \ 1.7349,\ \ 5.7442,\ \ 2.7931,\ \ 6.2938,\ \ 11.7337,\ \ -0.1318.$$

Sigue una distribución Normal, para ello se realizará la prueba de Anderson Darling con un nivel de significancia del 5 %.

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Anderson-Darling.**

Paso 2 Planteamiento de Hipótesis:

$$\mathbf{H}_0:\ \text{La muestra}\ \sim\ N(\mu, \sigma^2).$$

$$vs$$

$$\mathbf{H}_a:\ \text{La muestra}\ \nsim\ N(\mu, \sigma^2).$$

Paso 3 Estadístico de Prueba:

i	X	X_i	$F^*(X_i)$	$L_1=\ln(F^*(X_i))$	$L_2=\ln(1-F^*(X_{n-i+1}))$	Qi
1	-4.1302	-6.9098	0.0357	-3.3313	-3.1245	-0.4303
2	9.3150	-4.1302	0.1004	-2.2984	-2.2498	-0.9096
3	3.9757	-2.3838	0.1709	-1.7665	-1.9911	-1.2525
4	8.4900	-2.0039	0.1896	-1.6623	-1.3967	-1.4275
5	5.6204	-0.1426	0.2985	-1.2089	-1.2686	-1.4865
6	-6.9098	-0.1318	0.2992	-1.2066	-1.2408	-1.7948
7	-0.1426	1.7349	0.4304	-0.8430	-0.9095	-1.5189
8	-2.3838	2.7931	0.5094	-0.6743	-0.7123	-1.3866
9	-2.0039	3.9757	0.5973	-0.5153	-0.5628	-1.2218
10	1.7349	5.6204	0.7108	-0.3412	-0.3555	-0.8826
11	5.7442	5.7442	0.7187	-0.3301	-0.3545	-0.9586
12	2.7931	6.2938	0.7525	-0.2842	-0.2103	-0.7583
13	6.2938	8.4900	0.8634	-0.1468	-0.1874	-0.5570
14	11.7337	9.3150	0.8945	-0.1113	-0.1058	-0.3909
15	-0.1318	11.7337	0.9560	-0.0449	-0.0363	-0.1572

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\,[\ ln(F^*(X_{(i)})) + ln(1-F^*(X_{(n-i+1)}))\ ].$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba:

1) Se procede a ordenar nuestras observaciones de menor a mayor.

2) Se obtienen los estimadores puntuales de distribución normal con los datos de la muestra, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 2.66658$ y $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} = 5.313212$.

3) Se calcula la distribución propuesta, en este caso una distribución normal con media $\overline{X}$ y varianza s^2, es decir,$F^*(X)$, para ello se usa la aproximación a una normal estándar

4) Se calcula el primer sumando $ln(F^*(X_{(i)}))$ el cual se denotará como $L1$, después se calculará el segundo sumando $ln(1-F^*(X_{(n-i+1)}))$ el cual se denotará como $L2$

5) Se realiza el sumando de manera puntal, es decir, calcular:

$$Q_i = \left(\frac{2i-1}{n}\right)[\ ln(F^*(X_{(i)})) + ln(1-F^*(X_{(n-i+1)}))\ ] \quad \text{para } i = 1,\ldots,15.$$

6) Y se realiza la siguiente tabla:

7) Finalmente se suma todos los $Q_i s$ y se construye el estadístico:

$$A_n^2 = -n - \sum_{i=1}^{n} Qi = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\,[\ ln(F^*(X_{(i)})) + ln(1-F^*(X_{(n-i+1)}))\ ].$$

$$A_n^2 = -n - \sum_{i=1}^{n} Qi = -15 - (-15.1340) = 15.1340 - 15 = 0.1340.$$

$$A_n^2 = 0.1340.$$

8) Aplicando el ajuste de la tabla anterior:

Caso	**Ajuste en el estadístico**	**0.90**	**0.95**	**0.975**	**0.99**
Normal con $N(\overline{X}, S^2)$	$(1+\frac{4}{n}+\frac{25}{n^2})A_n^2$	0.632	0.751	0.870	1.029

$$(1+\frac{4}{n}+\frac{25}{n^2})A_n^2=(1+\frac{4}{15}+\frac{25}{15^2})A_n^2=0.1846.$$

Paso 5 Regla de Decisión.

Este último resultado se compara con la tabla de valores críticos de la Tabla Anderson-Darling para el caso de Normalidad, para un nivel de significancia $\alpha=0.05$.

Por lo anterior $A_{0.95}$=0.751, de esta manera se tiene que $0.751=A_{0.95}>A_n^2=0.1846$, como el valor crítico $A_{0.95}$ es mayor a comparación de A_n^2=0.1846.

$\therefore$ No Rechazamos H_0.

Paso 6 Conclusión.

Podemos concluir que a un nivel de significancia $\alpha=0.05$, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución normal.

21.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Datos
i=1:15
n=15
X=c(-4.1302, 9.315, 3.9757, 8.49, 5.6204, -6.9098, -0.1426, -2.3838,
                            -2.0039, 1.7349, 5.7442,2.7931,6.2938,11.7337,-0.1318)
X_i=c(-6.9098,-4.1302,-2.3838,-2.0039, -0.1426,-0.1318,1.7349,2.7931,3.9757,5.6204,
      5.7442,6.2938,8.4900,9.3150,11.7337)
F_Xi=c( 0.0357, 0.1004,0.1709,0.1896,0.2985,0.2992,0.4304,0.5094,0.5973,0.7108,
        0.7187,0.7525,0.8634,0.8945,0.9560)
L1=c( -3.3313,-2.2984,-1.7665,-1.6623,-1.2089,-1.2066,-0.8430,-0.6743,-0.5153,-0.3412,-0.3301,
      -0.2842,-0.1468,-0.1113,-0.0449)    #ln(F*(Xi))
L2=c(-3.1245,-2.2498,-1.9911,-1.3967,-1.2686,-1.2408,-0.9095,
-0.7123,-0.5628,-0.3555,-0.3545,-0.2103,-0.1874,-0.1058,-0.0363)  #ln(1-F*(Xn-i+1))
Qi=c(-0.4303,-0.9096,-1.2525,-1.4275,-1.4865,-1.7948,-1.5189,-1.3866,-1.2218,-0.8826,
-0.9586,-0.7583,-0.5570,-0.3909,-0.1572)

#Ahora calcularemos el estadistico de prueba
A_2=-n-(sum(Qi))
A_2
```

```
[1] 0.1331
```

```
#Ahora aplicaremos el ajuste correspondiente

A_2ajust=(1+4/n+25/n^2)*A_2
A_2ajust
```

```
[1] 0.1833822
```

Por la tabla auxiliar tenemos $A_{0.95}=0.751$, de esta manera se tiene que $0.751=A_{0.95}>A_n^2=0.1846$, como el valor crítico $A_{0.95}$ es mayor a comparación de $A_n^2=0.1846$, entonces no rechazamos H_0. Podemos concluir que a un nivel de significancia α=0.05, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución normal.

Ahora podemos utilizar la prueba en R

```
ad.test(X)   #Recordemos que la función no hace el ajuste
```

```
    Anderson-Darling normality test

data:  X
A = 0.13402, p-value = 0.9722
```

Capítulo 22

Otras estadísticas

La prueba de Kuiper esta muy relacionada con la prueba Kolmogorov–Smirnov (o prueba K-S). Como la prueba $K-S$, esta prueba utiliza las estadísticas D^+ y D^- que representan las diferencias más positivas y más negativas entre las distribuciones que se están comparando. La estadística de prueba de Kuiper es:

$$V = D^+ + D^-$$

Con este pequeño cambio, la prueba de Kuiper es tan sensible en las colas como lo es en la mediana de la distribución.

Las pruebas **Anderson–Darling** y **Cramér–von Mises** pertenecen a un grupo llamado **estadísticas EDF cuadráticas**, en donde el término EDF se refiere a que se basan en la función de distribución empírica.

Este grupo de estadísticas está definido de la siguiente manera:

$$n \int_{-\infty}^{\infty} (S(x) - F_0(x))^2 w(x) dF_0(x).$$

En donde la diferencia entre la distribución empírica y la hipótetica está calculada con el término cuadrático y el término $w(x)$ es un peso que se da esas diferencias.

Cuando $w(x) = 1$ entonces se tiene la estadística de Cramér–von Mises; cuando $w(x) = [F_0(x)(1-F_0(x))]^{-1}$ entonces se tiene la estadística de Anderson–Darling, la cual por construcción asigna mayores pesos a observaciones en las colas de la distribución.

En R, la libreria "goftest" contiene las pruebas Anderson–Darling y Cramér–von Mises entre otras.

Retomando el ejemplo de las alturas de los pinos, probaremos ahora que las alturas tienen distribución normal y exponencial.

```
X=Loblolly$height
meanx=mean(X)
meanx
```

```
[1] 32.3644
```

```
sdx=sd(X)
sdx
```

```
[1] 20.6736
```

```
library(goftest)
ad.test(X,null = "pnorm",mean=meanx,sd=sdx)
```

```
    Anderson-Darling test of goodness-of-fit
    Null hypothesis: Normal distribution
    with parameters mean = 32.3644047619048, sd = 20.6736047504145
```

```
    Parameters assumed to be fixed

data:  X
An = 2.7319, p-value = 0.03765
```

```
cvm.test(X,null = "pnorm",mean=meanx,sd=sdx)
```

```
    Cramer-von Mises test of goodness-of-fit
    Null hypothesis: Normal distribution
    with parameters mean = 32.3644047619048, sd = 20.6736047504145
    Parameters assumed to be fixed

data:  X
omega2 = 0.38648, p-value = 0.07825
```

```
ad.test(X,null = "pexp",rate=1/meanx)
```

```
    Anderson-Darling test of goodness-of-fit
    Null hypothesis: exponential distribution
    with parameter rate = 0.0308981428009166
    Parameters assumed to be fixed

data:  X
An = 4.5223, p-value = 0.004895
```

```
cvm.test(X,null = "pexp",rate=1/meanx)
```

```
    Cramer-von Mises test of goodness-of-fit
    Null hypothesis: exponential distribution
    with parameter rate = 0.0308981428009166
    Parameters assumed to be fixed

data:  X
omega2 = 0.79946, p-value = 0.007174
```

22.1. Mas ejercicios

1. Cinco niños de cuarto grado de primaria fueron seleccionados al azar de todos los niños de ese grado en la escuela “15 de septiembre”, para participar en una carrera. Sus tiempos en segundos fueron: 6.3, 4.2, 4.7, 6.0, y 5.7. Probar con la hipótesis de que los tiempos provienen de una distribución uniforme en el intervalo de 4 a 8 segundos.
2. La siguiente muestra aleatoria hace referencia a los rendimientos positivos de cierta acción a lo largo del tiempo.

0.2513, 0.2566, 0.3459, 0.6379, 2.0505, 1.803, 2.1906,1.5299, 0.35005, 0.3128, 1.2726, 2.3674, 2.3214, 2.4373, 0.6548.

¿Usted piensa que la anterior muestra sigue una distribución normal?

Realizar la prueba correspondiente para verificar que su suposición es cierta con un nivel de confianza del 90 %.

3. Se obtuvieron sesenta y dos observaciones de un experimento, y se plantea la pregunta de si dichas observaciones provienen de una distribución normal con media 12 y desviación estándar 3. Ninguna observación se encontró por debajo del cuartil inferior de la distribución, 35 estuvieron por arriba de cuartil superior, 22 tomaron valores menores a la mediana, y 5 estuvieron entre la mediana y el cuartil superior. ¿Es posible concluir que las observaciones provienen de la distribución mencionada?

Parte VI

Regresión Lineal Simple

Introducción

El análisis de regresión es una de las **técnicas estadísticas** de uso más frecuente para analizar datos. Su atractivo y utilidad general son el resultado de usar una ecuación para expresar la relación entre una variable de interés y un conjunto de variables predictorias relacionadas. Para usar bien la regresión se requiere tanto apreciar la teoría que hay detrás como los problemas prácticos que suelen presentarse cuando se emplea ésta técnica con datos de la vida real.

Capítulo 23

Modelo con intercepto

El objetivo principal del **Modelo de Regresión Lineal Simple** es el poder asociar o ajustar a una dispersión de datos una función (en primera instancia se abordará el ajuste mediante una recta) cuyos parámetros dependan directamente de las observaciones con la finalidad de poder resumir, simplificar u obtener propiedades importantes sobre comportamiento de la muestra.

Dicho modelo es el más sencillo de los modelos lineales e involucra una variable de interés y llamada **dependiente** o **respuesta** y su relación con la variable **predictoria** o **independiente** x, estableciendo que la media de la variable dependiente y cambia a razón constante cuando el valor de la variable independiente x crece o decrece.

El modelo de regresión lineal simple en general es:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

donde:

- x es la variable regresora,
- y es la variable de respuesta,
- β_0 ordenada al origen,
- β_1 pendiente del modelo,
- ϵ es un error aleatorio.

Conviene considerar a la variable regresora x como una varible determinista, o bien, una variable controlada por el investigador la cual puede ser medida, mientras que la variable respuesta y es una variable aleatoria. Ahora bien, los datos no caen exactamente sobre una recta por lo que se considera ϵ como un error estadístico, esto es, que es una variable aleatoria que explica por qué el modelo no ajusta exactamente los datos.

Una vez vista la ecuación general del modelo de regresión lineal simple se hablará de algunos supuestos que se deben de cumplir al ajustar una serie de datos, éstas consideraciones hace que en ocaciones carezca de sentido realizar una regresión lineal, sin embargo, no hay que perder de vista que éste es un modelado por lo que algunas características físicas del problema pueden haber sido simplificadas u omitidas.

Definición 2.1 (Supuestos del Modelo de Regresión Simple).

En el modelo de regresión simple se supone que ϵ satisface:

- $\mathbf{E}[\epsilon_i] = 0$
- $\mathbf{Var}(\epsilon_i) = \sigma^2$
- $\mathbf{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall\, i = 1, \ldots, n, \quad j = 1, \ldots, n, \quad i \neq j.$
- $\epsilon \sim N(0, \sigma^2)$

Una vez considerados estos supuestos se pueden obtener resultados aún más importantes.

Por ejemplo, con los supuestos dados es posible calcular la esperanza y la varianza de la variable y_i, dado un valor x_i.

Teorema 2.1 Sea y una variable de interés, denominada variable de respuesta, la cual es relacionada con una variable regresora x, entonces:

a) $\mathbf{E}[y_i] = \beta_0 + \beta_1 x_i$

b) $\mathbf{Var}(y_i) = \sigma^2$

Demostración:

a)

$$\mathbf{E}[y_i] = \mathbf{E}[\beta_0 + \beta_1 x_i + \epsilon_i]$$

La parte aleatoria de y_i es ϵ_i; β_0,β_1 son constantes y x_i ya es un valor dado; por lo que:

$$\mathbf{E}[y_i] = \beta_0 + \beta_1 x_i + \mathbf{E}[\epsilon_i]$$

$$= \beta_0 + \beta_1 x_i + 0$$
$$\therefore \mathbf{E}[y_i] = \beta_0 + \beta_1 x_i.$$

b)

$$Var(y_i) = Var(\beta_0 + \beta_1 x_i + \epsilon i)$$

La parte aleatoria de y_i es ϵ_i; β_0,β_1 son constantes y x_i ya es un valor dado; por lo que $Var(c+\epsilon_i) = Var(\epsilon_i)$ con c constante, de esta manera:

$$Var(y_i) = 0 + 0 + Var(\epsilon_i)$$
$$\therefore Var(y_i) = \sigma^2$$

23.1. Estimación por mínimos cuadrados de los parámetros del modelo

El modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x + \epsilon$$

cuenta con dos parámetros desconocidos, β_0 y β_1, los cuales deben ser estimados a partir de los datos de la muestra. Con la hipótesis de varianza constante sobre los errores, aparece otro parámetro σ^2 desconocido, aunque no está incluido en el modelo también debe ser estimado. Un procedimiento para estimar los parámetros de un modelo lineal simple es el **método de mínimos cuadrados**, que se puede ilustar sencillamente aplicándolo para ajustar una línea recta a $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, se estiman β_0 y β_1 tales que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea mínima.

Definición 2.2 (Residuos). Sea y_i los valores observados, $\hat{y}_i$ los valores estimados mediante la regresión lineal simple. La forma de calcular la desviación de y_i con respecto a su media estimada $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ para un $\hat{\beta}_0$ y $\hat{\beta}_1$ dados es:

$$e_i = y_i - \hat{y}_i$$

donde e_i son los residuos.

Por lo anterior,

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Como se mencionó, lo que se busca es que la diferencia entre todos los valores observados y los valores estimados sea 0, es decir, que la suma de distancias entre y_i y $\hat{y}_i$ sea cero, lo anterior significa que:

$$\sum_{i=1}^{n} e_i = 0$$

Para estimar β_0 y β_1 el **método de mínimos cuadrados** propone minimizar la suma de los cuadrados de los residuos, ya que de ésta manera se minimizan las distancias verticales entre las observaciones reales (y) y las estimadas $(\hat{y})$, ya que entre más cercanas a cero se encuentren las distancias, mejor se ajusta el modelo a los datos. Antes de continuar, es necesario abordar unos cuantos resultados.

Se define a:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \sum_{i=1}^{n} (x_i^2 - 2x_i\overline{x} + \overline{x}_i^2) \\ &= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\overline{x} + \sum_{i=1}^{n} \overline{x}_i^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\overline{x}^2 n + \overline{x}_i^2 n \end{aligned}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \overline{x}^2 n.$$

También definimos:

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} (xi - \overline{x})(y_i - \overline{y}) \\ &= \sum_{i=1}^{n} (x_i - \overline{x}) y_i \\ &= \sum_{i=1}^{n} x_i (y_i - \overline{y}) \\ &= \sum_{i=1}^{n} X_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \ /n \end{aligned}$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$$

Una vez definida la notación, tenemos el siguiente teorema:

Teorema 2.2 (Mínimos Cuadrados). Sea $\hat{\beta}_0$ y $\hat{\beta}_1$ los parámetros que minimizan la suma de cuadrados de la diferencia entre los valores observados y los estimados $\left(\sum_{i=1}^{n} e_i^2\right)$ entonces:

a) $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

b) $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$

Demostración:

a)

Tenemos:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Minimizando la suma de cuadrados, se deriva respecto a $\hat{\beta}_0$

$$\frac{\delta \sum_{i=1}^{n} e_i^2}{\delta \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\frac{\delta \sum_{i=1}^{n} e_i^2}{\delta \hat{\beta}_0} = -2\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{\beta}_0 - \sum_{i=1}^{n}\hat{\beta_1 x_i}\right).$$

Igualando a 0

$$-2\left(\sum_{i=1}^{n} y_i - \hat{\beta}_0 n - \sum_{i=1}^{n}\hat{\beta}_1 x_i\right) = 0$$

$$\sum_{i=1}^{n} y_i - \hat{\beta}_0 n - \sum_{i=1}^{n}\hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\overline{y}n = n\hat{\beta}_0 + \hat{\beta}_1 \overline{x}n$$

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$$

$$\therefore \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.\blacksquare$$

Por lo tanto se obtiene el primer estimador, $\hat{\beta}_0$; para el estimador de β_1 se deriva respecto a $\hat{\beta}_1$:

$$\frac{\delta \sum_{i=1}^{n} e_i^2}{\delta \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i$$

$$= -2\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\hat{\beta}_0 x_i - \sum_{i=1}^{n}\hat{\beta}_1 x_i^2\right).$$

Igualando la derivada a 0 para hallar el punto crítico

$$-2\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\hat{\beta}_0 x_i - \sum_{i=1}^{n}\hat{\beta}_1 x_i^2\right) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\hat{\beta}_0 x_i - \sum_{i=1}^{n}\hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n}\hat{\beta}_0 x_i + \sum_{i=1}^{n}\hat{\beta}_1 x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \overline{x} n + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Sustituyendo $\hat{\beta}_0$ por $\overline{y} - \hat{\beta}_1 \overline{x}$ se tiene:

$$\sum_{i=1}^{n} x_i y_i = \overline{yx} n - \hat{\beta}_1 \overline{x}^2 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i - \overline{yx} n + \hat{\beta}_1 \overline{x}^2 n - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Por la notación tenemos que: $S_{xy} = \sum x_i y_i - n\overline{xy}$ y $S_{xx} = \sum x_i^2 - n\overline{x}^2$

$$\frac{\delta \sum_{i=1}^{n} e_i^2}{\delta \hat{\beta}_1} = -2 \left(S_{xy} - \beta_1 \left(\sum_{i=1}^{n} \hat{x_i^2} - \overline{x}^2 n \right) \right)$$

$$= -2(S_{xy} - \hat{\beta}_1 S_{xx}).$$

Igualando a 0

$$-2(S_{xy} - \hat{\beta}_1 S_{xx}) = 0$$

$$S_{xy} - \hat{\beta}_1 S_{xx} = 0$$

$$\Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}. \blacksquare$$

De esta manera se demuestra el teorema 2.2. Un punto a destacar es que las siguientes ecuaciones:

$$\begin{cases} \hat{\beta}_0 \ n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Son conocidas como las **ecuaciones normales**, que en conjunto forman un sistema de ecuaciones; al resolverlas simultáneamente para $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtiene los estimadores de β_0 y β_1 que se plantean en el teorema anterior.

Un problema del método de mínimos cuadrados es que no proporciona un estimador para σ^2, sin embargo, se obtendrá bajo el supuesto de normalidad que es el siguiente:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

De esta manera se ha cumplido el objetivo que se planteó al inicio, encontrar los estimadores de β_0 y β_1 tal que los residuales fueran igual a 0.

Teorema 2.3 (Diferencia de Residuales) Sea $\hat{\beta}_0$ y $\hat{\beta}_1$ los estimadores de mínimos cuadrados de β_0 y β_1, respectivamente, entonces la suma de las distancias entre y_i y $\hat{y}_i$ es cero, es decir:

$$\sum_{i=1}^{n} e_i = 0$$

Demostración

Tenemos que el residual e_i se calcula como: $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ entonces si sustituimos se tiene:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i$$

$$= n\overline{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \overline{x}$$

$$= n \left[\overline{y} - \hat{\beta}_0 - \hat{\beta}_1 \overline{x} \right]$$

Recordando que la estimación de $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ tenemos:

$$= n \left[\overline{y} - \overline{y} + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 \overline{x} \right]$$

$\therefore \sum_{i=1}^{n} e_i = 0.\blacksquare$

Esto implica que si la suma de residuales es 0, entonces la $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$ y vamos a demostrarlo:

Corolario 1 Sea $\hat{\beta}_0$ y $\hat{\beta}_1$ los estimadores de mínimos cuadrados de β_0 y β_1 respectivamente, si se cumple el teorema anterior, entonces la suma de los valores observados y la suma de los valores ajustados por la regresión son iguales, es decir:

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$$

Demostración

Por definición:

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right)$$

Por la estimación de $\hat{\beta}_1$

$$= \sum_{i=1}^{n} \left(\overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i \right)$$

$$= \sum_{i=1}^{n} \overline{y} - n\hat{\beta}_1 \overline{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$= n\overline{y} - n\hat{\beta}_1 \overline{x} + n\hat{\beta}_1 \overline{x}$$

$$= n\overline{y}$$

Por construcción de $\overline{y}$

$$\therefore \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i.\blacksquare$$

Una consecuencia de lo anterior,

Corolario 2 Sea $\hat{y}_i$ el valor estimado de la forma $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ y el residual $e_i = y_i - \hat{y}_i$, entonces se cumple que:

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

Demostración

Por definición de las $\hat{y}$

$$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i$$

$$= \sum_{i=1}^{n} \left(\hat{\beta}_0 e_i + \hat{\beta}_1 x_i e_i \right)$$

$$= \sum_{i=1}^{n} \hat{\beta}_0 e_i + \sum_{i=1}^{n} \hat{\beta}_1 x_i e_i$$

$$= \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i$$

Y como demostramos anteriormente, la suma de los residuos es cero:

$$= \hat{\beta}_1 \sum_{i=1}^{n} x_i (y_i - \hat{y}_i)$$

$$= \hat{\beta}_1 \sum_{i=1}^{n} \left(x_i y_i - x_i \hat{\beta}_0 - x_i^2 \hat{\beta}_1 \right)$$

$$= \hat{\beta}_1 \sum_{i=1}^{n} \left(x_i y_i - x_i^2 \hat{\beta}_1 - x_i y_i + x_i^2 \hat{\beta}_1 \right)$$

Entonces:

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0. \blacksquare$$

23.2. Propiedades de los estimadores

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen propiedades estadísticas muy importantes. ya que son estimadores insesgados, además son de mínima varianza. La propiedad de insesgamiento se revisará en el siguiente teorema:

Teorema 2.4 Sea $\hat{\beta}_0$, $\hat{\beta}_1$ los estimadores de mínimos cuadrados de β_0 y β_1, respectivamente, entonces los estimadores $\hat{\beta}_0$,$\hat{\beta}_1$ son insesgados. Es decir:

a) $\mathbf{E}\left[\hat{\beta}_1\right] = \beta_1$

b) $\mathbf{E}\left[\hat{\beta}_0\right] = \beta_0$

Demostración

a) Para demostrar el insesgamiento de $\hat{\beta}_1$ usaremos la propiedad de combinación lineal de β_1 :

"Haremos un pequeño paréntesis para demostrarlo"

Tenemos:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Sustituyendo S_{xx} y S_{xy}

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{S_{xx}}$$
$$= \sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{S_{xx}}\right)y_i.$$

Haciendo $a = \frac{(x_i - \overline{x})}{S_{xx}}$ nada depende de y, por lo que a es constante, por consiguiente la combinación lineal de las y_i para $\hat{\beta}_1$ es:

$$\therefore \hat{\beta}_1 = \sum_{i=1}^{n} a y_i.$$

Para la combinación lineal de las y_i para $\hat{\beta}_0$ se tiene:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Sustituyendo la combinación lineal de $\hat{\beta}_1$ y notación de S_{xx}

$$\hat{\beta}_0 = \sum_{i=1}^{n}\frac{y_i}{n} - a y_i \overline{x}$$
$$= \sum_{i=1}^{n}(\frac{1}{n} - a\overline{x})y_i$$
$$= \sum_{i=1}^{n}\left(\frac{1}{n} - \left(\frac{(x_i - \overline{x})}{S_{xx}}\right)\overline{x}\right)y_i.$$

Haciendo $b = \frac{1}{n} - \left(\frac{(x_i - \overline{x})}{S_{xx}}\right)\overline{x}$ nada depende de y, por lo que b es constante, entonces la combinación lineal de y_i para $\hat{\beta}_0$ es:

$$\therefore \hat{\beta}_0 = \sum_{i=1}^{n} b y_i.$$

"Regresando del paréntesis tenemos:"

$$\mathbf{E}\left[\hat{\beta}_1\right] = \mathbf{E}\left[\sum_{i=1}^{n} a y_i\right]$$
$$= \sum_{i=1}^{n} a\mathbf{E}[y_i].$$

Tenemos que $\mathbf{E}[y_i] = \beta_0 + \beta_1 x_i$, sustituyendo:

$$\mathbf{E}\left[\hat{\beta}_1\right] = \sum_{i=1}^{n} a(\beta_0 + \beta_1 x_i)$$
$$= \sum_{i=1}^{n} a\beta_0 + \sum_{i=1}^{n} a\beta_1 x_i$$
$$= \beta_0 \sum_{i=1}^{n} a + \beta_1 \sum_{i=1}^{n} a x_i.$$

Sustituyendo a de la linealidad de β_1 se tiene:

$$\mathbf{E}\left[\hat{\beta}_1\right] = \beta_0 \sum_{i=1}^{n}\left(\frac{(x_i-\overline{x})}{S_{xx}}\right) + \beta_1 \sum_{i=1}^{n}\left(\frac{(x_i-\overline{x})}{S_{xx}}\right) x_i$$

$$= \frac{\beta_0}{S_{xx}}(n\overline{x} - \overline{x}n) + \frac{\beta_1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})x_i.$$

Simplificando y recordando que $\sum_{i=1}^{n}(x_i-\overline{x})x_i = S_{xx}$

$$\mathbf{E}\left[\hat{\beta}_1\right] = 0 + \frac{\beta_1}{S_{xx}}S_{xx}$$

$$\therefore \mathbf{E}\left[\hat{\beta}_1\right] = \beta_1$$

Por lo tanto el estimador $\hat{\beta}_1$ es insesgado. ■

b) Ahora para demostrar el insesgamiento de $\hat{\beta}_0$, se sustituye el estimador $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$, de esta forma tenemos:

$$\mathbf{E}\left[\hat{\beta}_0\right] = \mathbf{E}\left[\overline{y} - \hat{\beta}_1\overline{x}\right]$$

$$= \mathbf{E}\left[\overline{y}\right] - \mathbf{E}\left[\hat{\beta}_1\overline{x}\right]$$

$$= \mathbf{E}\left[\sum_{i=1}^{n}\frac{y_i}{n}\right] - \overline{x}\mathbf{E}\left[\hat{\beta}_1\right]$$

Como el estimador de β_1 es insesgado:

$$= \sum_{i=1}^{n}\frac{1}{n}\mathbf{E}\left(y\right) - \overline{x}\beta_1.$$

y sabemos que $\mathbf{E}(y) = \beta_0 + \beta_1 x$, sustituyendo tenemos:

$$\mathbf{E}\left[\hat{\beta}_0\right] = \sum_{i=1}^{n}\frac{1}{n}(\beta_0 + \beta_1 x_i) - \overline{x}\beta_1$$

$$= \beta_0 + \beta_1\overline{x} - \overline{x}\beta_1$$

$$\therefore \mathbf{E}[\hat{\beta}_0] = \beta_0.$$

Por lo tanto el estimador $\hat{\beta}_0$ es insesgado. ■

Ahora nos preguntaremos por la varianza de los estimadores, analizando qué tan distante está el estimador del parámetro buscado. Es decir, la varianza es el margen de error que obtiene en la estimación de los parámetros.

Teorema 2.5 Sea $\hat{\beta}_0$ $\hat{\beta}_1$ los estimadores puntuales de β_0 y β_1 entonces la varianza de estimación

a) $Var\left(\hat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2$.

b) $Var\left(\hat{\beta}_1\right) = \frac{1}{S_{xx}}\sigma^2$.

Demostración:

a)

$$Var\left(\hat{\beta}_0\right) = Var\left(\overline{y} - \hat{\beta}_1\overline{x}\right)$$

$$= Var(\overline{y}) + Var(\hat{\beta}_1\overline{x}) - 2Cov(\overline{y}, \hat{\beta}_1\overline{x})$$

$$= Var(\overline{y}) + Var(\hat{\beta}_1\overline{x})$$

$$= Var\left(\sum_{i=1}^{n}\frac{y_i}{n}\right) + \overline{x}^2 Var(\hat{\beta}_1)$$

$$= \sum_{i=1}^{n}\frac{1}{n^2}Var(y_i) + \overline{x}^2\frac{\sigma^2}{S_{xx}}$$

$$= \sum_{i=1}^{n}\frac{1}{n^2}\sigma^2 + \overline{x}^2\frac{\sigma^2}{S_{xx}}$$

$$\therefore Var\left(\hat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2.\blacksquare$$

b)

$$Var\left(\hat{\beta}_1\right) = Var\left(\frac{S_{xy}}{S_{xx}^2}\right)$$

$$= \frac{1}{S_{xx}}Var\left(S_{xy}\right)$$

$$= \frac{1}{S_{xx}^2}Var\left(\sum_{i=1}^{n}(x_i - \overline{x})y_i\right)$$

$$= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 Var(y_i)$$

$$= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 Var(e_i)$$

$$= \frac{1}{S_{xx}^2}S_{xx}\sigma^2$$

$$\therefore Var\left(\hat{\beta}_1\right) = \frac{1}{S_{xx}}\sigma^2.\blacksquare$$

Una característica importante de los estimadores obtenidos es que existe una covarianza conjunta entre $\hat{\beta}_0$ y $\hat{\beta}_1$, que veremos a continuación:

Teorema 2.6 Sea β_0,β_1 los estimadores puntuales de β_0 y β_1 entonces la covarianza conjunta de los estimadores es:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sigma^2\left(-\frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$$

Demostración:

Como mencionamos anteriormente, $\hat{\beta}_0, \hat{\beta}_1$ pueden ser expresadas como combinaciones lineales de y por lo que:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \mathbf{E}\left[\left(\hat{\beta}_0 - \mathbf{E}(\hat{\beta}_0)\right)\left(\hat{\beta}_1 - \mathbf{E}(\hat{\beta_1})\right)\right]$$

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \mathbf{E}\left[\left(\sum_{i=1}^{n}a_iY_i - \mathbf{E}(\sum_{i=1}^{n}a_iY_i)\right)\left(\sum_{i=1}^{n}b_iY_i - \mathbf{E}(\sum_{i=1}^{n}b_iY_i)\right)\right].$$

Por linealidad de la esperanza:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n} a_i b_i \mathbf{E}[(Y_i - \mathbf{E}(Y_i))(Y_i - \mathbf{E}(Y_i))].$$

Por propiedades de la esperanza:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n} a_i b_i Var[Y_i].$$

Por el **teorema 2.1** se sabe que la varianza del modelo es:

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n} a_i b_i \sigma^2$$

Sustituyendo $a = \frac{(x_i - \overline{x})}{S_{xx}}$ y $b = \frac{1}{n} - \left(\frac{(x_i - \overline{x})}{S_{xx}}\right)\overline{x}$ se tiene:

$$\begin{aligned}
Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})}{S_{xx}}\left[\frac{1}{n} - \left(\frac{(x_i - \overline{x})}{S_{xx}}\right)\overline{x}\right]\right)\sigma^2 \\
&= \sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})}{nS_{xx}} - \left(\frac{(x_i - \overline{x})}{S_{xx}}\right)^2 \overline{x}\right)\sigma^2 \\
&= \sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})}{nS_{xx}} - \left(\frac{\overline{x}(x_i - \overline{x})}{S_{xx}^2}\right)^2 \overline{x}\right)\sigma^2 \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{nS_{xx}} - \frac{\overline{x}(x_i - \overline{x})^2}{S_{xx}^2}\right)\sigma^2 \\
&= \left(\frac{\overline{x}n - n\overline{x}}{nS_{xx}} - \frac{\overline{x}S_{xx}}{S_{xx}^2}\right)\sigma^2 \\
&= \left(-\frac{\overline{x}}{S_{xx}}\right)\sigma^2
\end{aligned}$$

$$\therefore Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sigma^2\left(-\frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right). \blacksquare$$

Una consecuencia inmediata del **teorema 2.3** y **teorema 2.4**, es que si la suma de residuales es 0, implica que la esperanza del valor observado sea igual a la esperanza del valor estimado, es decir:

Corolario 3 Sea $\hat{\beta}_0, \hat{\beta}_1$ los estimadores de mínimos cuadrados de β_0 y β_1 respectivamente, si se cumple el teorema 2.3, entonces la esperanza de los valores observados y la esperanza de los valores ajustados por la regresión son iguales, es decir:

$$\mathbf{E}\left[\hat{y}_i\right] = \mathbf{E}\left[y_i\right]$$

De esta manera lo que se debe demostrar es que se cumple la siguiente igualdad:

Pd. $\mathbf{E}\left[Y_i - \hat{y}_i\right] = 0$

Demostración:

$\mathbf{E}\left[Y_i - \hat{y}_i\right] = \mathbf{E}\left[Y_i\right] - \mathbf{E}\left[\hat{y}_i\right]$

$= \beta_0 + \beta_1 x_i - \mathbf{E}\left[\hat{y}_i\right]$ Por el teorema 2.1

$= \beta_0 + \beta_1 x_i - \mathbf{E}\left[\hat{\beta}_0 + \hat{\beta}_1 x_i\right]$

$= \beta_0 + \beta_1 x_i - \mathbf{E}\left[\hat{\beta}_0\right] - \mathbf{E}\left[\hat{\beta}_1 x_i\right]$

$= \beta_0 + \beta_1 x_i - \beta_0 - \beta_1 x_i$ Por el teorema 2.5 son insesgados

$\therefore \quad \mathbf{E}\left[Y_i - \hat{y}_i\right] = 0.$ ■

Capítulo 24

Modelo sin intercepto

El modelo anterior es bueno, ya que es un caso general de como se comporta la regresión lineal, sin embargo, si los datos no incluyen el 0 entonces no tendría caso calcular β_0 ya que no se presenta una intersección con el eje y. La manera en la que se construye el modelo de regresión lineal sin intercepto, es similar la construcción con intercepto. Por lo que de igual manera, la mejor forma de estimar la pendiente de la recta sería usando el método de mínimos cuadrados.

En primera instancia, la recta al no tener intersección con el eje y con $y \neq 0$, $\beta_0 = 0$, lo que provoca que la ecuación de la recta de regresión lineal esté conformada por:

$$y = \beta x + \epsilon$$

De igual manera tenemos los mismos supuestos que en la definición 2.1, son:

Supuestos

- $\mathbf{E}[\epsilon_i] = 0$
- $\mathbf{Var}(\epsilon_i) = \sigma^2$
- $\mathbf{Cov}(\epsilon_i, \epsilon_j) = 0 \ \forall i = 1, \ldots, n \ j = 1, \ldots, n \ i \neq j.$
- $\epsilon \sim N(0, \sigma^2)$

Al usar estos supuestos se pueden obtener estadísticos importantes como la media y la varianza de y.

Teorema 2.7 Sea una variable de interés y llamada **dependiente**, relacionada con una variable explicativa x, sin intercepto entonces:

a) $\mathbf{E}[y] = \beta x$.

b) $\mathbf{Var}(y) = \sigma^2$.

Demostración:

a)

$\mathbf{E}[y] = \mathbf{E}[\beta x + \epsilon]$

La estimación es sobre y, como se mencionó, β es constante y x es un valor dado. Por lo que:

$= \beta x + \mathbf{E}[\epsilon]$

$= \beta x + 0$

$\therefore \mathbf{E}[y] = \beta x.$ ■

b)

$Var(y) = Var(\beta x + \epsilon)$

La estimación es sobre y, por lo que β es constante, x ya es un valor dado, así:

$= 0 + Var(\epsilon)$

$\therefore Var(y) = \sigma^2$. ■

24.1. Estimación por mínimos cuadrados de los parámetros del modelo

Para estimar la pendiente, es decir, β. Se debe de construir al estimador de tal manera que la diferencia entre todos los valores observados y los valores estimados sea 0, es decir, que la línea de regresión pase en la parte media de estos valores de dispersión. A este concepto se le conoce como **Residuos** sin intercepto.

Definición 2.3 (Residuos). Sea y_i los valores observados, $\hat{y_i}$ los valores estimados mediante la regresión lineal simple sin intercepto. La forma de calcular la desviación de y_i con respecto a su media estimada $\hat{y_i} = \hat{\beta} x_i$ para un $\hat{\beta}$ dado es:

$$e_i = y_i - \hat{y_i}.$$

donde e_i son los Residuos.

De esta manera los residuos se encuentran de la forma:

$$e_i = y_i - \hat{\beta} x_i.$$

Lo que se busca es que la suma de la diferencia de los valores observados menos los valores estimados sea 0, es decir:

$$\sum_{i=1}^{n} e_i = 0$$

De aquí surge la idea nuevamente de usar el método de Mínimos Cuadrados para estimación de β.

Teorema 2.8 (Mínimos Cuadrados). Si se minimiza la suma de cuadrados de la diferencia de los valores observados y los estimados $\left(\sum_{i=1}^{n} e_i^2\right)$ entonces se tiene como estimador de β a:

- $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

Demostración:

Se busca minimizar $\sum_{i=1}^{0} e_i^2$ por ello:

$$S(\beta) = \sum_{i=1}^{n} e_i^2.$$

Sustituyendo:

$$S(\beta) = \sum_{i=1}^{n} \left(y_i - \hat{\beta} x_i\right)^2$$

Derivando respecto a $\hat{\beta}$

$$\frac{\partial S(\beta)}{\partial \hat{\beta}} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta} x_i\right) x_i$$

Igualando la derivada a 0 para hallar el punto crítico.

$$-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}x_i\right)x_i=0$$

$$\sum_{i=1}^{n}x_iy_i-\hat{\beta}x_i^2=0$$

$$\therefore\ \hat{\beta}=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}$$

Volviendo a derivar para obtener si es máximo o mínimo.

$$\frac{\partial^2 S(\beta)}{\partial\hat{\beta}^2}=\left(-2\sum_{i=1}^{n}\left(y_ix_i-\hat{\beta}x_i^2\right)\right)'$$

$$\therefore\frac{\partial^2 S(\beta)}{\partial\hat{\beta}^2}=2\sum_{i=1}^{n}\left(y_ix_i-\hat{\beta}x_i\right)^{-1}x_i^2\ >0.$$

Por lo tanto es un mínimo, entonces, el estimador de β que minimiza la suma de cuadrados de los residuales es:

$$\hat{\beta}=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}.\ \blacksquare$$

24.2. Propiedades de los estimadores

Después de haber obtenido el estimador del modelo de regresión lineal simple sin intercepto, se demuestran las propiedades que cumple $\hat{\beta}$.

Teorema 2.9 Sea el estimador $\hat{\beta}$ de β insesgado y cumple con:

a) $\mathbf{E}\left[\hat{\beta}\right]=\beta$.

b) $\mathbf{Var}\left[\hat{\beta}\right]=\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2}$

Demostración:

a) Por hipótesis

$$\mathbf{E}\left[\hat{\beta}\right]=\mathbf{E}\left[\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}\right]$$

$$=\frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2}\mathbf{E}[y_i]$$

Por el **teorema 2.7**

$$=\frac{\sum_{i=1}^{n}x_i\beta x_i}{\sum_{i=1}^{n}x_i^2}$$

$$=\frac{\sum_{i=1}^{n}x_i^2\beta}{\sum_{i=1}^{n}x_i^2}$$

$$\therefore\mathbf{E}\left[\hat{\beta}\right]=\beta.\ \blacksquare$$

Por lo tanto el estimador $\hat{\beta}$ es insesgado.

b) Para la varianza se puede definir a $\hat{\beta}$ como:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Es decir,

$$\hat{\beta} = \left(\frac{x_1}{\sum_{i=1}^{n} x_i^2}\right) y_1 + \left(\frac{x_2}{\sum_{i=1}^{n} x_i^2}\right) y_2 + \ldots + \left(\frac{x_n}{\sum_{i=1}^{n} x_i^2}\right) y_n$$

Con ésta notación se tiene que:

$$Var\left(\hat{\beta}\right) = Var\left(\left(\frac{x_1}{\sum_{i=1}^{n} x_i^2}\right) y_1 + \left(\frac{x_2}{\sum_{i=1}^{n} x_i^2}\right) y_2 + \ldots + \left(\frac{x_n}{\sum_{i=1}^{n} x_i^2}\right) y_n\right)$$

$$= \left(\frac{x_1}{\sum_{i=1}^{n} x_i^2}\right)^2 Var(y_1) + \left(\frac{x_2}{\sum_{i=1}^{n} x_i^2}\right)^2 Var(y_2) + \ldots + \left(\frac{x_n}{\sum_{i=1}^{n} x_i^2}\right)^2 Var(y_n)$$

$$= \left(\frac{x_1}{\sum_{i=1}^{n} x_i^2}\right)^2 \sigma^2) + \left(\frac{x_2}{\sum_{i=1}^{n} x_i^2}\right)^2 \sigma^2 + \ldots + \left(\frac{x_n}{\sum_{i=1}^{n} x_i^2}\right)^2 \sigma^2$$

$$= \left(\frac{\sum_{i=1}^{n} x_i^2}{(\sum_{i=1}^{n} x_i^2)^2}\right) \sigma^2$$

$\therefore Var\left(\hat{\beta}\right) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$. ■

Con el método de Mínimos Cuadrados no es posible obtener un estimador para la varianza del modelo, sin embargo, se obtendrá bajo el supuesto de normalidad que es el siguiente:

$$\tilde{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \hat{\beta} x_i\right)^2$$

La cual es equivalente a ser representada por la siguiente ecuación.

$$\tilde{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y})^2.$$

Se elige este estimador ya que cumple con ser insesgado para σ, ésta afirmación se prueba a continuación:

$$\mathbf{E}[\tilde{\sigma}^2] = \mathbf{E}\left[\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y})^2\right]$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \mathbf{E}\left[(y_i - \hat{y})^2\right].$$

Al ser el segundo momento nos conviene utilizar la siguiente igualdad $Var(y_i - \hat{y}) = \mathbf{E}[(y_i - \hat{y})^2] - \mathbf{E}^2[y_i - \hat{y}]$, debido al corolario 3, $\mathbf{E}^2[y_i - \hat{y}] = 0$, por lo que por facilidad se debe calcular la varianza del estimador $Var(y_i - \hat{y}) = \mathbf{E}[(y_i - \hat{y})^2]$:

$$\mathbf{E}[\tilde{\sigma}^2] = \frac{1}{n-1} \sum_{i=1}^{n} \mathbf{E}\left[(y_i - \hat{y})^2\right]$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} Var(y_i - \hat{y})$$

$$= \frac{\sum_{i=1}^{n}}{n-1} Var(y_i - x_i\hat{\beta})$$

$$= \frac{\sum_{i=1}^{n}}{n-1} Var\left(y_i - \frac{x_i \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right)$$

$$= \frac{\sum_{i=1}^{n}}{n-1} Var\left[\left(1 - \frac{x_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}\right) y_i\right]$$

$$= \frac{\sum_{i=1}^{n}}{n-1} \left(1 - \frac{x_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}\right)^2 Var[y_i]$$

$$= \frac{\sum_{i=1}^{n}}{n-1} \left(1 - \frac{x_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \sigma^2$$

$$= \frac{\sum_{i=1}^{n}}{n-1} \left(1 - \frac{x_i \sum_{i=1}^{n} x_i}{S_{xx}}\right)^2 \sigma^2$$

$$= \frac{\sum_{i=1}^{n}}{n-1} \frac{[S_{xx}^2 - 2S_{xx} n\overline{x} x_i + (n\overline{x} x_i)^2]}{S_{xx}^2} \sigma^2$$

$$= \frac{\sum_{i=1}^{n}}{n-1} \frac{[S_{xx}^2 - 2S_{xx} n\overline{x} x_i + n^2\overline{x}^2 x_i^2]}{S_{xx}^2} \sigma^2$$

$$= \frac{\sum_{i=1}^{n}}{n-1} \left[1 - \frac{2n\overline{x} x_i}{S_{xx}} + \frac{n^2\overline{x}^2 x_i^2}{S_{xx}^2}\right] \sigma^2$$

$$= \frac{1}{n-1} \left[1 - \frac{2n\overline{x} \sum_{i=1}^{n} x_i}{S_{xx}} + \frac{n^2\overline{x}^2 \sum_{i=1}^{n} x_i^2}{S_{xx}^2}\right] \sigma^2$$

$$= \frac{1}{n-1} \left[1 - \frac{2n^2\overline{x}^2}{S_{xx}} + \frac{n^2\overline{x}^2 S_{xx}}{S_{xx}^2}\right] \sigma^2$$

$$= \frac{1}{n-1} \left[1 - \frac{2(\sum_{i=1}^{n} x_i)^2}{S_{xx}} + \frac{(\sum_{i=1}^{n} x_i)^2}{S_{xx}^2}\right] \sigma^2$$

$$= \frac{1}{n-1} \left[n + \frac{-2\sum_{i=1}^{n}(x_i)^2 + \sum_{i=1}^{n}(x_i)^2}{S_{xx}}\right] \sigma^2$$

$$= \frac{1}{n-1} \left[n - \frac{\sum_{i=1}^{n}(x_i)^2}{S_{xx}}\right] \sigma^2$$

$$= \frac{1}{n-1} \left[n - \frac{\sum_{i=1}^{n}(x_i)^2}{\sum_{i=1}^{n}(x_i)^2}\right] \sigma^2$$

$$= \frac{1}{n-1} [n-1] \sigma^2$$

$$\therefore \mathbf{E}[\tilde{\sigma}^2] = \sigma^2.$$

Por lo tanto, $\tilde{\sigma}$ es insesgado.■

$Día$	$Facturas$	$Tiempo$
1	149	2.1
2	60	1.8
3	188	2.3
4	23	0.8
5	201	2.7
6	58	1.0
7	77	1.7
8	222	3.1
9	181	2.8
10	30	1.0
11	110	1.5
12	83	1.2
13	60	0.8
14	25	1.0
15	173	2.0
16	169	2.5
17	190	2.9
18	233	3.4
19	289	4.1
20	45	1.2
21	193	2.5
22	70	1.8
23	241	3.8
24	103	1.5
25	163	2.8
26	120	2.5
27	201	3.3
28	135	2.0
29	80	1.7
30	29	1.5

24.2.1. Ejemplo en R-Studio

El gerente del departamento de ventas de la compañía **CALLCENT** desea predecir, de alguna manera, el tiempo promedio que tardarían en procesar un número dado de facturas. Esto con el objetivo de llevar a cabo una buena logística de diversas operaciones dentro de la empresa.

Se ha recolectado, durante un periodo de 30 días, la información sobre el número de facturas procesadas y el tiempo que tomó (en horas):

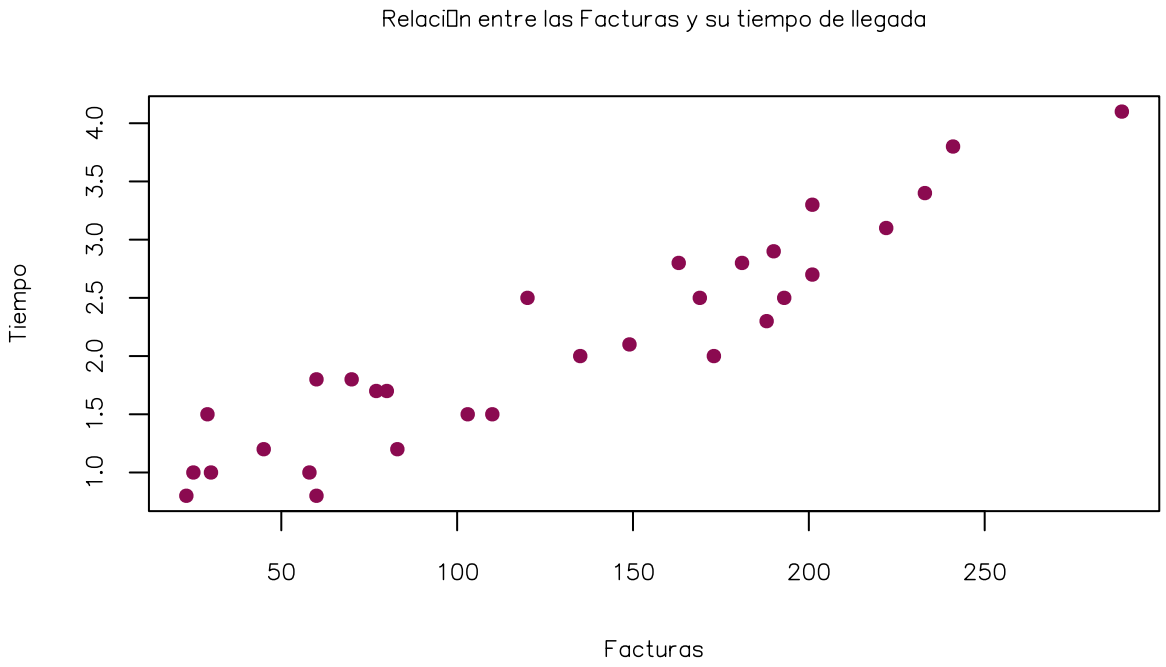
Ahora haremos la réplica en R.

```
Dia=read.table("Problema8T1.csv",sep=",",header=TRUE)
names(Dia)
```

```
[1] "Dia"      "Facturas" "Tiempo"
```

```
attach(Dia)
```

```
plot(Dia$Facturas,Dia$Tiempo,type = "p",
     col="deeppink4",pch=16, xlab="Facturas", ylab="Tiempo",
     main= "Relación entre las Facturas y su tiempo de llegada")
```



Como podemos observar el gráfico nos grita que exite una posible relación lineal entre el número de facturas y el tiempo empleado para éstas. Para confirmar nuestras sospechas vamos a calcular el coefieciente de correlación de Pearson:

```
cor(Tiempo,Facturas)
```

```
[1] 0.9336877
```

Es decir, r=0.9336 indica una fuerte relación lineal positiva entre el número de facturas procesadas y el tiempo. Entonces tiene "sentido" emplear un modelo de regresión lineal simple.

Ahora estimaremos los parámetros β_0 y β_1 con el método de mínimos cuadrados visto.

```
y=Dia$Tiempo
x=Dia$Facturas
y_barra=mean(y)
x_barra=mean(x)
Sxx=sum((x-x_barra)^2)
Sxy=sum((x-x_barra)*(y-y_barra))
beta1=Sxy/Sxx
beta0=y_barra-beta1*x_barra
```

Entonces $\hat{\beta_1}$ será:

```
[1] 0.01129164
```

y $\hat{\beta_0}$ será:

```
[1] 0.6417099
```

Ahora estimaremos el parámetro β con el método de mínimos cuadrados para un modelo sin intercepto.

```
beta_gorro=sum(x*y)/sum(x^2)
```

Entonces $\hat{\beta}$ será:

```
[1] 0.01503001
```

Residuales

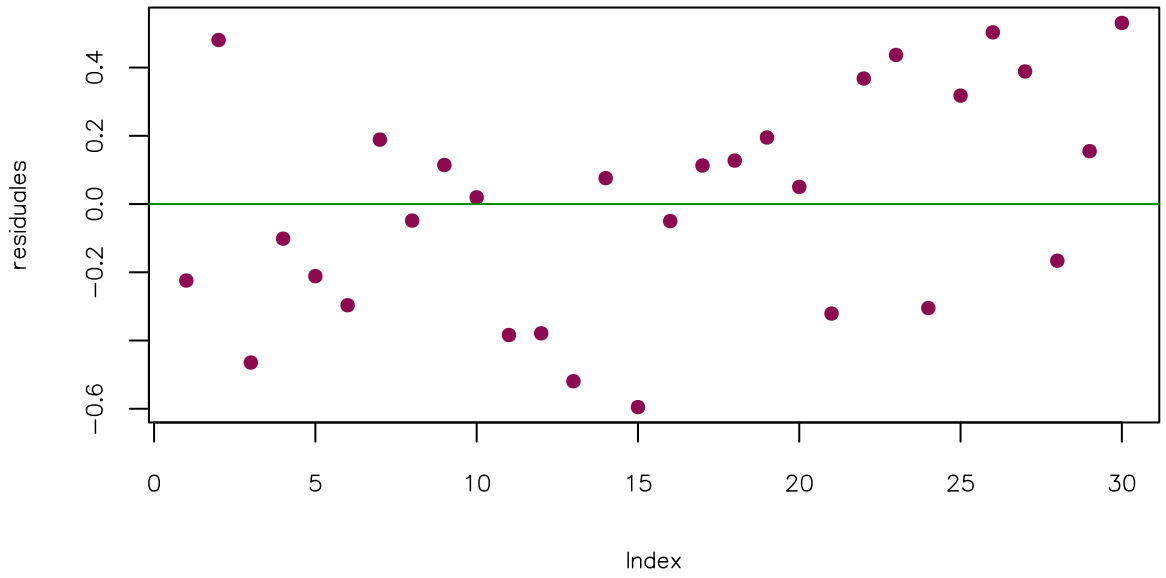
Ahora calcularemos los residuales, es decir la diferencia entre los valores observados y los valores estimados ($e_i = y_i - \hat{y}_i$)

Primero calculamos el vector de los valores estimados $\hat{y}$:

```
y_gorro=beta0+beta1*x
```

Luego los residuales y los graficamos

```
e=y-y_gorro
plot(e,type = "p",pch=16, ylab="residuales",col="deeppink4")
abline(a=0,b=0, col="green4")
```



Para que el modelo propuesto ajuste bien a los datos originales esperaríamos que los residuales estuvieran lo mas cercano al cero (linea amarilla). Mas adelante veremos como usar estos gráficos para verificar algunos de los supuestos del modelo.

Capítulo 25

Intervalos de confianza

Anteriormente hemos obtenido, de manera puntual, las estimaciones de los parámetros desconocidos del modelo de regresión lineal simple. Sin embargo, en ocasiones se puede tener una gran variabilidad en el ajuste de los parámetros, por lo que realizar inferencia puntual no siempre puede ser recomendable, es por ello que desarrollaremos intervalos de confianza para proporcionar estimaciones por intervalo en el cual, el parámetro de interés tenga una alta probabilidad de pertenecer a este conjunto.

25.1. Intervalo para β_0

Dado que el estimador de β_0 es una combinación lineal, de igual forma se tiene una combinación lineal de variables aleatorias normales independientes, por lo que β_0 tiene una distribución normal asociada con media y varianza demostrada en el **teorema 2.4**.

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2\right).$$

Estandarizando:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2}} \sim N(0,1).$$

Como $\frac{(n-2)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{(n-2)}$ se tiene:

$$\frac{\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2}}}{\sqrt{\frac{\frac{(n-2)}{\sigma^2}\hat{\sigma}^2}{n-2}}} \sim t_{(n-2)}$$

Simplificando se obtiene la cantidad pivotal para $\hat{\beta}_0$:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

De esta manera, construyendo el intervalo de confianza con la cantidad pivotal:

$$\mathbf{P}\left[-t^{\alpha/2}_{(n-2)} < \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}} < t^{\alpha/2}_{(n-2)}\right] = 1 - \alpha$$

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}<\hat{\beta}_0-\beta_0<t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}\right]=1-\alpha$$

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}<\beta_0-\hat{\beta}_0<t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}\right]=1-\alpha$$

Sumando $\hat{\beta}_0$ en todas las desigualdades:

$$\mathbf{P}\left[\hat{\beta}_0-t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}<\beta_0<\hat{\beta}_0+t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}\right]=1-\alpha$$

Por lo tanto, el intervalo de confianza $1-\alpha$ para β_0 es:

$$\beta_0\in\left(\hat{\beta}_0-t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}\ \ ,\ \ \hat{\beta}_0+t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}\right).$$

25.2. Intervalo para β_1

Dado que el estimador de β_1 es una combinación lineal, de igual forma se tiene una combinación lineal de variables aleatorias normales independientes, por lo que β_1 tiene una distribución normal asociada con media y varianza demostrada en el **teorema 2.4**.

$$\hat{\beta}_1\sim N\left(\beta_1,\frac{\sigma^2}{S_{xx}}\right)$$

Estandarizando:

$$\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}}\sim N(0,1).$$

Como $\frac{(n-2)}{\sigma^2}\hat{\sigma}^2\sim\chi^2_{(n-2)}$ se tiene:

$$\frac{\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{\sqrt{\frac{\frac{(n-2)}{\sigma^2}\hat{\sigma}^2}{n-2}}}\sim t_{(n-2)}$$

Por lo tanto, simplificando se obtiene una cantidad pivotal para $\hat{\beta}_1$:

$$\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}\sim t_{(n-2)}$$

Construyendo un intervalo de confianza $1-\alpha$ para β_1 se tiene que:

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}<\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}<t_{(n-2)}^{\alpha/2}\right]=1-\alpha$$

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}<\hat{\beta}_1-\beta_1<t_{(n-2)}^{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right]=1-\alpha$$

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} < \beta_1 - \hat{\beta}_1 < t_{(n-2)}^{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right] = 1-\alpha$$

Sumando $\hat{\beta}_1$ en todas las desigualdades:

$$\mathbf{P}\left[\hat{\beta}_1 - t_{(n-2)}^{\alpha/2}\sqrt{\frac{1}{S_{xx}}\hat{\sigma}^2} < \beta_1 < \hat{\beta}_1 + t_{(n-2)}^{\alpha/2}\sqrt{\frac{1}{S_{xx}}\hat{\sigma}^2}\right] = 1-\alpha$$

Por lo tanto, el intervalo de confianza $1-\alpha$ para β_1 es:

$$\beta_1 \in \left(\hat{\beta}_1 - t_{(n-2)}^{\alpha/2}\sqrt{\frac{1}{S_{xx}}\hat{\sigma}^2} \quad , \quad \hat{\beta}_1 + t_{(n-2)}^{\alpha/2}\sqrt{\frac{1}{S_{xx}}\hat{\sigma}^2}\right).$$

25.3. Intervalo para σ^2

Para construir el intervalo de confianza para σ^2 se observa que se posee una cantidad pivotal asociada de la forma:

$$\frac{(n-2)\hat{\sigma}^2_{MC}}{\sigma^2} \sim \chi^2_{(n-2)}.$$

La distribución χ^2 no es una distribución simétrica por lo que se plantean los cuantiles $W_{\alpha/2}$ y $W_{1-\alpha/2}$ que corresponden a la valuación de la $\chi^2_{(n-2)}$ en el cuantil $\alpha/2$ y $1-\alpha/2$, respectivamente.

$$\mathbf{P}\left[W_{\alpha/2} < \frac{(n-2)\hat{\sigma}^2_{MC}}{\sigma^2} < W_{1-\alpha/2}\right] = 1-\alpha$$

Obteniendo el recíproco en ambas partes de las desigualdades se observa que:

$$\mathbf{P}\left[\frac{1}{W_{\alpha/2}} > \frac{\sigma^2}{(n-2)\hat{\sigma}^2_{MC}} > \frac{1}{W_{1-\alpha/2}}\right] = 1-\alpha$$

Reordenando el intervalo de confianza se tiene que:

$$\mathbf{P}\left[\frac{(n-2)\hat{\sigma}^2_{MC}}{W_{\alpha/2}} > \sigma^2 > \frac{(n-2)\hat{\sigma}^2_{MC}}{W_{1-\alpha/2}}\right] = 1-\alpha$$

Por la estimación insesgada propuesta para σ^2 se sabe que $\hat{\sigma}^2_{MC} = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2$ así:

$$\mathbf{P}\left[\frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{W_{1-\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{W_{\alpha/2}}\right] = 1-\alpha$$

Por convención se usa que $W_{1-\alpha/2} = \chi^{2(1-\alpha/2)}_{(n-2)}$ y $W_{\alpha/2} = \chi^{2(\alpha/2)}_{(n-2)}$. De esta manera el intervalo de confianza para σ^2 es:

$$\mathbf{P}\left[\frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{\chi^{2(1-\alpha/2)}_{(n-2)}} < \sigma^2 < \frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{\chi^{2(\alpha/2)}_{(n-2)}}\right] = 1-\alpha,$$

reescribiendo el intervalo de confianza en su forma explícita:

$$\sigma^2 \in \left(\frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{\chi^{2(1-\alpha/2)}_{(n-2)}} \quad , \quad \frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{\chi^{2(\alpha/2)}_{(n-2)}}\right).$$

25.4. Intervalo para el valor esperado y

Después de haber realizado un modelo de regresión lineal; como vimos en el teorema 2.1, el valor esperado de y_i es $\mathbf{E}[y_i] = \beta_0 + \beta_1 x_i$, para toda $i = 1, \ldots, n$, en el caso de que se conozca un nuevo valor x' de la variable regresora x entonces se podrá calcular el valor esperado de y al sustituir los estimadores de β_0 y β_1, respectivamente. Sin embargo, al realizar estas sustituciones se tiene asociada diversas variabilidades, como la desviación estándar de los estimadores. Es por ello que se realizan intervalos de confianza para el valor esperado y con la finalidad de aportar mejores ajustes con un nivel de significancia α.

El valor esperado de y dado que se conoce un nuevo valor x' de x, hace referencia a la esperanza condicional de la forma $\mathbf{E}[y|x = x'] = \beta_0 + \beta_1 x'$, la cual es denotada como $\mu_x = \mathbf{E}[y|x = x']$, sin embargo, al desconocer el valor de β_0, β_1 se realiza la estimación del valor de y usando los estimadores de mínimos cuadrados, es decir, $\mathbf{E}[\widehat{y|x = x'}] = \hat{\beta}_0 + \hat{\beta}_1 x'$, usualmente escrita como $\hat{\mu}_x = \mathbf{E}[\widehat{y|x = x'}]$.

Por ejemplo, suponga que después de estimar un modelo de regresión lineal simple se obtuvo como parámetros $\hat{\beta}_0 = 3$ y $\hat{\beta}_1 = 5$, un año después se observa que el valor de la variable regresora es $x' = 10$, de esta manera el valor esperado dado x' es:

$$\hat{\mu}_x = \hat{\beta}_0 + \hat{\beta}_1 x'.$$

Sustituyendo los valores del ejemplo:

$$\hat{\mu}_x = 3 + 5(10)$$

$$\therefore \hat{\mu}_x = 53.$$

Es decir, el valor esperado de y dado $x' = 10$ es 53 unidades. Debido a que cada y_i es una combinación lineal se sabe que los valores esperados se distribuyen con normalidad es decir:

$$\hat{\mu}_x \sim N\ (\mathbf{E}[\mu_x], Var[\mu_x]).$$

Teorema 2.11 Sea $\hat{\mu}_x = \hat{\beta}_0 + \hat{\beta}_1 x'$ el valor esperado de y dado x' $\left(\mathbf{E}[\widehat{y|x = x'}]\right)$, entonces cumple con las propiedades de esperanza y varianza:

a) $\mathbf{E}[\hat{\mu}_x] = \mu_x$ donde $\mu_x = \beta_0 + \beta_1 x'$.

b) $Var[\hat{\mu}_x] = \left(\frac{1}{n} + \frac{(x' - \overline{x})^2}{S_{xx}}\right)\sigma^2$.

Demostración

a) Para la esperanza se sabe que $\hat{\mu}_x = \mathbf{E}[\widehat{y|x = x'}]$ así sustituyendo se sabe:

$$\mathbf{E}[\hat{\mu}_x] = \mathbf{E}\left[\hat{\beta}_0 + \hat{\beta}_1 x'\right].$$

Por propiedades de la esperanza se tiene:

$$\mathbf{E}[\hat{\mu}_x] = \mathbf{E}\left[\hat{\beta}_0\right] + x'\mathbf{E}\left[\hat{\beta}_1\right].$$

Por el **teorema 2.4** sabemos que los estimadores son insesgados:

$$\mathbf{E}[\hat{\mu}_x] = \beta_0 + \beta_1 x'$$
$$\therefore \mathbf{E}[\hat{\mu}_x] = \mu_x.\blacksquare$$

Por lo que es insesgado para μ_x, es decir, $\mathbf{E}\left[\mathbf{E}[\widehat{y|x = x'}]\right] = \mathbf{E}[y|x = x']$.

b) Para la varianza se tiene:

$$Var[\hat{\mu}_x] = Var\left[\hat{\beta}_0 + \hat{\beta}_1 x'\right]$$

$$Var[\hat{\mu}_x] = Var[\hat{\beta}_0] + Var[\hat{\beta}_1 x'] + 2Cov(\hat{\beta}_0, \hat{\beta}_1 x').$$

Por el **teorema 2.5** se sabe que las varianzas de los estimadores son:

$$Var[\hat{\mu}_x] = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2 + x'^2 Var[\hat{\beta}_1] + 2x' Cov(\hat{\beta}_0, \hat{\beta}_1)$$

$$Var[\hat{\mu}_x] = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2 + \frac{x'^2}{S_{xx}}\sigma^2 - 2x'\frac{\overline{x}\sigma^2}{S_{xx}}$$

$$Var[\hat{\mu}_x] = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} + \frac{x'^2}{S_{xx}} - 2x'\frac{\overline{x}}{S_{xx}}\right)$$

$$Var[\hat{\mu}_x] = \sigma^2\left(\frac{1}{n} + \frac{(x'-\overline{x})^2}{S_{xx}}\right).\blacksquare$$

De esta manera se busca construir un intervalo de confianza para el valor esperado de y dado $x'(\mu_x)$, y sabemos que el valor esperado se comporta de la forma:

$$\hat{\mu}_x \sim N\left(\mu_x, \sigma^2\left(\frac{1}{n} + \frac{(x'-\overline{x})^2}{S_{xx}}\right)\right).$$

Estandarizando $\hat{\mu}_x$ para obtener una normal estándar:

$$\frac{\hat{\mu}_x - \mu_x}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x'-\overline{x})^2}{S_{xx}}\right)}} \sim N(0,1)$$

Como $\frac{(n-2)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{(n-2)}$ se tiene que el cociente entre una normal y una Ji-Cuadrada se distribuye como t de Student:

$$\frac{\frac{\hat{\mu}_x - \mu_x}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x'-\overline{x})^2}{S_{xx}}\right)}}}{\sqrt{\frac{\frac{(n-2)}{\sigma^2}\hat{\sigma}^2}{n-2}}} \sim t_{(n-2)}$$

Simplificando términos se tiene:

$$\frac{\hat{\mu}_x - \mu_x}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x'-\overline{x})^2}{S_{xx}}\right)}} \sim t_{(n-2)}$$

Denotando a $\hat{\sigma}^2_x = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x'-\overline{x})^2}{S_{xx}}\right)$ se tiene la cantidad pivotal para el valor esperado de y dado x

$$\frac{\hat{\mu}_x - \mu_x}{\sqrt{\hat{\sigma}^2_x}} \sim t_{(n-2)}.$$

Una vez hallado el estadístico a usar se construye el intervalo de confianza $(1-\alpha)$x100 para μ_x.

$$\mathbf{P}\left[-t^{\alpha/2}_{(n-2)} < \frac{\hat{\mu}_x - \mu_x}{\sqrt{\hat{\sigma}^2_x}} < t^{\alpha/2}_{(n-2)}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t^{\alpha/2}_{(n-2)}\sqrt{\hat{\sigma}^2_x} < \hat{\mu}_x - \mu_x < t^{\alpha/2}_{(n-2)}\sqrt{\hat{\sigma}^2_x}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma_x^2}} < \mu_x - \hat{\mu}_x < t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma_x^2}}\right] = 1-\alpha$$

$$\mathbf{P}\left[\hat{\mu}_x - t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma_x^2}} < \mu_x < \hat{\mu}_x + t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma_x^2}}\right] = 1-\alpha$$

$$\therefore \mathbf{P}\left[\hat{\mu}_x - t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{(x'-\overline{x})^2}{S_{xx}}\right)\hat{\sigma}^2} < \mu_x < \hat{\mu}_x + t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{(x'-\overline{x})^2}{S_{xx}}\right)\hat{\sigma}^2}\ \right] = 1-\alpha,$$

En su forma más compacta:

$$\mu_x \in \left(\hat{\mu}_x - t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma_x^2}}\ \ ,\ \ \hat{\mu}_x + t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma_x^2}}\right).$$

entonces el intervalo de confianza para el valor esperado de y dado x' es:

$$\beta_0+\beta_1 x' \in \left(\hat{\beta}_0+\hat{\beta}_1 x' - t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{(x'-\overline{x})^2}{S_{xx}}\right)\hat{\sigma}^2}\ \ ,\ \ \hat{\beta}_0+\hat{\beta}_1 x' + t_{(n-2)}^{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{(x'-\overline{x})^2}{S_{xx}}\right)\hat{\sigma}^2}\right).$$

25.5. Intervalo de predicción

La diferencia significativa entre intervalos de predicción e intervalos de confianza para el valor esperado y, es que en el intervalo del valor esperado lo que se busca encontrar es el valor que en promedio se debería obtener y dado que se tiene una observación x, es decir, la observación que cae sobre la recta de regresión de la forma $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$, mientras que en un intervalo de predicción se "predice" valores futuros de y dado que se conoce o se estima un valor x, denotado como x^*, es decir, $y = \beta_0 + \beta_1 x + \epsilon$, por ende la predicción de valores de y es $\hat{y_x} = \hat{\beta}_0 + \hat{\beta}_1 x^* + \epsilon$.

La varianza del valor de predicción que se debe considerar será la varianza del valor esperado $\hat{\mu}_x$ pero además se añade la varianza del modelo de regresión lineal simple, es decir:

$$Var(\hat{y_x}) = Var(\hat{\mu}_x) + Var(y).$$

El cual por el **teorema 2.1** y **teorema 2.11** tenemos:

$$Var(\hat{y_x}) = \left(1 + \frac{1}{n} + \frac{(x^*-\overline{x})^2}{S_{xx}}\right)\sigma^2.$$

Teorema 2.12 La predicción de un valor de y dado que se conoce un valor x^* de la variable regresora x está dado por $\hat{y_x} = \hat{\beta}_0 + \hat{\beta}_1 x^* + \epsilon$, donde ϵ satisface que $\epsilon \sim N(0, \sigma^2)$ e independientemente a β_0, β_1, así la predicción cumple con las siguientes propiedades:

a) $\mathbf{E}[\hat{y_x}] = \beta_0 + \beta_1 x^*$.

b) $Var[\hat{y_x}] = \left(1 + \frac{1}{n} + \frac{(x^*-\overline{x})^2}{S_{xx}}\right)\sigma^2$.

Demostración

a) La esperanza de $\hat{y_x}$ está dada por:

$$\mathbf{E}[\hat{y_x}] = \mathbf{E}\left[\hat{\beta}_0 + \hat{\beta}_1 x^* + \epsilon\right]$$

Por linealidad de la esperanza

$$= \mathbf{E}\left[\hat{\beta_0}\right] + \mathbf{E}\left[\hat{\beta}_1 x^*\right] + \mathbf{E}[\epsilon]$$

Por el **teorema 2.4** e hipótesis

$$= \beta_0 + \beta_1 x^* + 0$$

$$\therefore \mathbf{E}\left[\hat{y_x}\right] = \beta_0 + \beta_1 x^*. \blacksquare$$

b) La varianza de $\hat{y_x}$ está dada por:

$$Var\left[\hat{y_x}\right] = Var\left[\hat{\beta_0} + \hat{\beta_1} x^* + \epsilon\right]$$

$$= Var\left[\hat{\beta_0} + \hat{\beta_1} x^*\right] + Var\left[\epsilon\right] + 2Cov(\hat{\beta_0} + \hat{\beta_1} x^*, \epsilon)$$

Por independencia

$$= Var\left[\hat{\beta_0} + \hat{\beta_1} x^*\right] + Var\left[\epsilon\right] + 2(0)$$

Por el **teorema 2.11** e hipótesis

$$= \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)\sigma^2 + \sigma^2$$

$$\therefore Var\left[\hat{y_x}\right] = \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)\sigma^2. \blacksquare$$

Usando los resultados, la estimación de valores futuros de y se comporta:

$$\hat{y_x} \sim N\left(y_x, \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)\right),$$

donde $y_x = \beta_0 + \beta_1 x^*$. Estandarizando $\hat{y_x}$:

$$\frac{\hat{y_x} - y_x}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)}} \sim N(0,1)$$

Como $\frac{(n-2)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{(n-2)}$:

$$\frac{\frac{\hat{y_x} - y_x}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \sim t_{(n-2)}.$$

Simplificando términos se tiene:

$$\frac{\hat{y_x} - y_x}{\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)}} \sim t_{(n-2)}$$

Denotando a $\sigma_x^2 = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)$ se tiene la cantidad pivotal para el valor de y dado x

$$\frac{\hat{y_x} - y_x}{\sqrt{\sigma_x^2}} \sim t_{(n-2)}.$$

Por lo que construyendo el intervalo de predicción:

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2} < \frac{\hat{y_x} - y_x}{\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)}} < t_{(n-2)}^{\alpha/2}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)} < \hat{y_x} - y_x < t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)} < y_x - \hat{y_x} < t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)} < \beta_0+\beta_1x^* - \hat{\beta}_0 - \hat{\beta}_1x^* - \epsilon < t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)}\right] = 1-\alpha$$

$$\mathbf{P}\left[\hat{\beta}_0+\hat{\beta}_1x^* - t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)} < \beta_0+\beta_1x^* - \epsilon < \hat{\beta}_0+\hat{\beta}_1x^* + t_{(n-2)}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)}\right] = 1$$

Dado que ϵ es una variable aleatoria simétrica y con media 0 se tiene:

$$\beta_0+\beta_1x^*+\epsilon \in \left[\hat{\beta}_0+\hat{\beta}_1x^* - t_{(n-2)}^{\alpha/2}\sqrt{\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)\hat{\sigma}^2}, \hat{\beta}_0+\hat{\beta}_1x^* + t_{(n-2)}^{\alpha/2}\sqrt{\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)\hat{\sigma}^2}\right].$$

25.5.1. Ejemplo

Retomando los datos de la sección anterior, recordemos que el gerente del departamento de ventas de la compañía **CALLCENT** desea predecir, de alguna manera, el tiempo promedio que tardarían en procesar un número dado de facturas. Esto con el objetivo de llevar a cabo una buena logística de diversas operaciones dentro de la empresa.

Se ha recolectado, durante un periodo de 30 días, la información sobre el número de facturas procesadas (en nuestrio caso definimos como nuestra variable x) y el tiempo que tardan las mismas (que hemos definido como nuestra variable y).

Como se mencionó en la teoría para considerar la variabilidad en el ajuste provocada por el uso de una muestra aleatoria, además de estimación puntual, hay que realizar estimación por intervalos.

Intervalos de confianza para β_0 y β_1

Entonces haremos el cálculo de los intervalos al 95 % de confianza para ambos estimadores ($\hat{\beta}_0$, $\hat{\beta}_1$).

Como vimos en la sección anterior, debemos estimar σ^2, recordemos que el estimador es:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

```
[1] 30
```

```
s2_gorro=1/(n-2)*sum((y-y_gorro)^2)
s2_gorro
```

```
[1] 0.1087505
```

ya con $\hat{\sigma}^2$ podemos construir los intervalos de confianza.

- Primero para β_0, sustituyendo los valores en la siguiente ecuación:

$$\beta_0 \in \left(\hat{\beta}_0 - t_{(n-2)}^{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right) \hat{\sigma}^2} \quad , \quad \hat{\beta}_0 + t_{(n-2)}^{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right) \hat{\sigma}^2} \right)$$

```
b0Liminf=beta0-qt(.975,n-2)*sqrt((1/n+x_barra^2/Sxx)*s2_gorro)

b0Limsup=beta0+qt(.975,n-2)*sqrt((1/n+x_barra^2/Sxx)*s2_gorro)
```

Entonces el intervalo del 95 % de confianza para β_0 es:

```
      b0Liminf  b0Limsup
[1,] 0.3912496 0.8921701
```

- Ahora para β_1, sustituyendo los valores en la siguiente ecuación:

$$\beta_1 \in \left(\hat{\beta}_1 - t_{(n-2)}^{\alpha/2} \sqrt{\frac{1}{S_{xx}} \hat{\sigma}^2} \quad , \quad \hat{\beta}_1 + t_{(n-2)}^{\alpha/2} \sqrt{\frac{1}{S_{xx}} \hat{\sigma}^2} \right).$$

```
b1Liminf=beta1-qt(.975,n-2)*sqrt((1/Sxx)*s2_gorro)
b1Limsup=beta1+qt(.975,n-2)*sqrt((1/Sxx)*s2_gorro)
```

Entonces el intervalo del 95 % de confianza para β_1 es:

```
        b1Liminf   b1Limsup
[1,] 0.009615224 0.01296806
```

- Y por último calculamos el intervalo de confianza para σ^2, sustituyendo los valores en la siguiente ecuación:

$$\sigma^2 \in \left(\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\chi_{(n-2)}^{2(1-\alpha/2)}} \quad , \quad \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\chi_{(n-2)}^{2(\alpha/2)}} \right).$$

```
sigLiminf=sum((y-y_gorro)^2)/qchisq(0.975,n-2)
sigLimsup=sum((y-y_gorro)^2)/qchisq(0.025,n-2)
```

Entonces el intervalo del 95 % de confianza para β_1 es:

```
      sigLiminf sigLimsup
[1,] 0.06848759 0.1989182
```

Valor Esperado

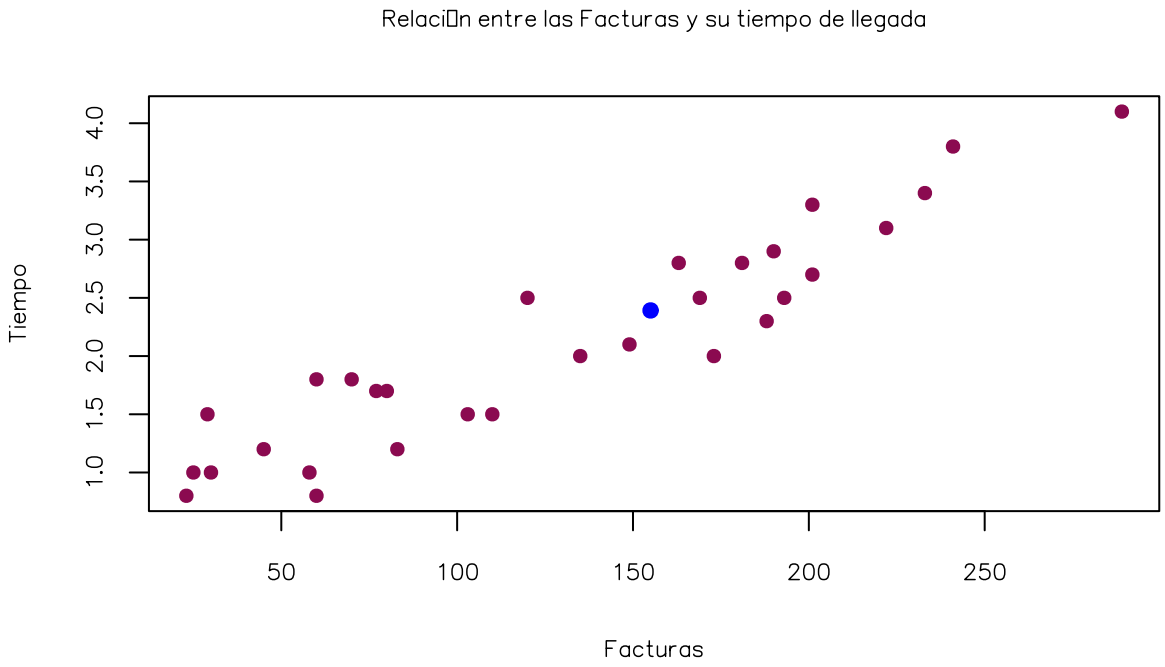
Ahora calcularemos el tiempo en horas promedio esperado para la fabulosa cantidad de 155 facturas.

```
x_fac=155
y_esperado=beta0+beta1*x_fac
```

Entonces el tiempo en horas promedio esperado será:

```
[1] 2.391915
```

Gráficamente se ve así:



Donde el punto azul es nuestro valor esperado.

- Y ahora construiremos su correspondiente intervalo del 99 % de confianza, sustituyendo en la siguiente ecuación:

$$\beta_0 + \beta_1 x' \in \left(\hat{\beta}_0 + \hat{\beta}_1 x' - t^{\alpha/2}_{(n-2)} \sqrt{\left(\frac{1}{n} + \frac{(x' - \overline{x})^2}{S_{xx}} \right) \hat{\sigma}^2} \quad , \quad \hat{\beta}_0 + \hat{\beta}_1 x' + t^{\alpha/2}_{(n-2)} \sqrt{\left(\frac{1}{n} + \frac{(x' - \overline{x})^2}{S_{xx}} \right) \hat{\sigma}^2} \right).$$

```
y_espLiminf=beta0+beta1*x_fac-qt(.995,n-2)*sqrt((1/n+(x_fac-x_barra)^2/Sxx)*s2_gorro)

y_espLimisup=beta0+beta1*x_fac+qt(.995,n-2)*sqrt((1/n+(x_fac-x_barra)^2/Sxx)*s2_gorro)
```

Entonces el intervalo del 99 % de confianza para el valor esperado de y es:

```
intery_esperado=cbind(y_espLiminf,y_espLimisup)
intery_esperado
```

```
     y_espLiminf y_espLimisup
[1,]    2.216224     2.567605
```

- El procedimiento es análogo al del valor esperado se aplica para el cálculo del intervalo de confianza para la predicción

Capítulo 26

Pruebas de hipótesis

El objetivo de la pruebas de hipótesis es poder escoger, a través de métodos estadísticos, la hipótesis capaz de retratar de mejor manera la realidad que se observa en la muestra.

Definición 2.5 Una prueba de hipótesis es una regla de decisión que, cuando se ha obtenido los valores de la muestra observada, lleva a una decisión de aceptar o rechazar la hipótesis nula bajo una cierta consideración.

La hipótesis nula, la sentencia que se pone a prueba es denotada como H_0, la hipótesis alternativa, el complemento de la hipótesis nula, es denotada como H_a.

En el modelo de regresión lineal simple, se desea realizar dos pruebas de hipótesis de gran importancia.

- $\mathbf{H}_0 : \hat{\beta}_1 = 0$ *vs.* $\mathbf{H}_a : \hat{\beta}_1 \neq 0$
- $\mathbf{H}_0 : \hat{\beta}_0 = 0$ *vs.* $\mathbf{H}_a : \hat{\beta}_0 \neq 0$

Estas pruebas son de gran importancia, ya que la primera de ellas es probar si β_1 es igual a cero, es decir, que el modelo no tenga pendiente y sea constante a través del tiempo, con un gran nivel de significancia dado. Si la prueba se rechaza entonces establece que el valor esperado de y depende de x.

La segunda prueba de hipótesis analiza la posibilidad de que $\beta_0 = 0$, si la prueba es aceptada implica que el modelo de regresión simple se comporta sin intercepto. Mientras que si se rechaza la hipótesis nula entonces $\beta_0 \neq 0$, por lo cual el mejor ajuste para la muestra es un modelo de regresión simple con intercepto.

26.1. Pruebas para β_0

Como se observó, la cantidad pivotal para β_0 está determinado por $t^{\alpha/2}_{(n-2)}$ con el estadístico:

$$t^* = \frac{\hat{\beta}_0 - b_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}},$$

donde:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Entonces las pruebas de hipótesis para β_0, en sus tres casos, que el valor de β_0 sea igual, inferior o superior al punto crítico deseado (b_0) tienen las siguientes reglas de decisión, con un nivel de significancia α.

Hipótesis	*Región de rechazo* :
$H_0 : \beta_0 = b_0 \ vs. \ H_a : \beta_0 \neq b_0$	$\lvert t^* \rvert > t^{\alpha/2}_{(n-2)}$
$H_0 : \beta_0 \leq b_0 \ vs. \ H_a : \beta_0 > b_0$	$t^* > t^{\alpha}_{(n-2)}$
$H_0 : \beta_0 \geq b_0 \ vs. \ H_a : \beta_0 < b_0$	$t^* < t^{1-\alpha}_{(n-2)}$

Para ejemplificar

Suponga que se quiere ajustar un modelo con intercepto o sin intercepto. En éste caso, conviene realizar una prueba de hipótesis de dos colas sobre β_0 valuando el punto crítico $b_0 = 0$, es decir, se desea probar que $\beta_0 = 0$, ya que si, con un nivel de significancia α, se rechaza H_0 se tiene evidencia de que el modelo a ajustar debería ser el modelo con intercepto. Detallando la prueba, en este caso en particular, tenemos:

$$\mathbf{H}_0 : \hat{\beta}_0 = 0 \quad vs. \quad \mathbf{H}_a : \hat{\beta}_0 \neq 0$$

La regla de decisión es rechazar H_0 cuando el estadístico $t^* = \frac{\hat{\beta}_0 - b_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}}$ esté contenida en la región de rechazo. En este caso en particular, se tiene lo siguiente:

$$t^* = \frac{\hat{\beta}_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2}}.$$

Por lo que se rechaza H_0 cuando $t^* < -t^{\alpha/2}_{(n-2)}$ o si $t^* > t^{\alpha/2}_{(n-2)}$. Si se rechaza H_0 entonces $\beta_0 \neq 0$ por lo que convendría realizar un modelo con intercepto; si no se rechaza la hipótesis nula con un nivel de significancia α, $\beta_0 = 0$, entonces el modelo sin intercepto sería el más óptimo para el conjunto de datos que se examina.

26.2. Prueba para β_1

Similar al caso anterior, β_1 está determinado por $t^{\alpha/2}_{(n-2)}$ con el estadístico:

$$t^* = \frac{\hat{\beta}_1 - b_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}},$$

donde:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Por lo tanto tenemos la prueba de hipótesis para β_1 en sus tres casos, que el valor de β_1 sea igual, inferior o superior al punto crítico deseado (b_1) y sus respectivas regiones de rechazo con un nivel de significancia α.

Hipótesis	*Región de rechazo* :
$H_0 : \beta_1 = b_1 \ vs. \ H_a : \beta_1 \neq b_1$	$\lvert t^* \rvert > t^{\alpha/2}_{(n-2)}$
$H_0 : \beta_1 \leq b_1 \ vs. \ H_a : \beta_1 > b_1$	$t^* > t^{\alpha}_{(n-2)}$
$H_0 : \beta_1 \geq b_1 \ vs. \ H_a : \beta_1 < b_1$	$t^* < t^{1-\alpha}_{(n-2)}$

Para ejemplificar, suponga que quiere ajustar un modelo de regresión lineal simple, sin embargo, duda si realmente la variable respuesta y depende de la variable regresora x, es decir, sospecha que $\beta_1 = 0$. Para ello se raliza la siguiente prueba de hipótesis valuando β_1 en el punto crítico de interés $b_1 = 0$:

$$\mathbf{H}_0 : \hat{\beta}_1 = 0 \quad vs. \quad \mathbf{H}_a : \hat{\beta}_1 \neq 0$$

La regla de decisión es rechazar H_0 cuando el estadístico

$$t^* = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

Esté contenida en la región de rechazo.Por lo que se rechaza H_0 cuando $t^* < -t^{\alpha/2}_{(n-2)}$ o si $t^* > t^{\alpha/2}_{(n-2)}$. Si se rechaza H_0 entonces $\beta_1 \neq 0$ por lo que la variable respuesta y depende de la variable regresora x; si no se rechaza H_0 con un nivel de significancia α, $\beta_1 = 0$, entonces la variable respuesta y no depende de la variable regresora x, debido a ello un modelo de regresión no sería viable con esas variables.

26.3. Prueba para σ^2

Para σ^2 se tiene las siguientes reglas de decisión con un nivel de significancia α :

Hipótesis	*Región de rechazo* :
$H_0 : \sigma^2 = s \ vs. \ H_a : \sigma^2 \neq s$	$t^* > \chi^{1-\alpha/2}_{(n-2)}$ ó $t^* < \chi^{\alpha/2}_{(n-2)}$
$H_0 : \sigma^2 \leq s \ vs. \ H_a : \sigma^2 > s$	$t^* > \chi^{\alpha}_{(n-2)}$
$H_0 : \sigma^2 \geq s \ vs. \ H_a : \sigma^2 < s$	$t^* < \chi^{1-\alpha}_{(n-2)}$

donde el estadístico t^* :

$$t^* = \frac{(n-2)\hat{\sigma}^2}{s}$$

Con:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

26.4. Análisis de la varianza (ANOVA)

El análisis de la varianza, también llamado ANOVA por sus siglas en ingles (Analysis of Variance), mide la asociación lineal entre la variable respuesta y y la variable regresora x.

La asociación lineal se logra cuando $\beta_1 \neq 0$, ya que en caso contrario, la estimación de y sería la media muestral $\overline{y}$, y la variable regresora x no aportaría información. Entonces, al suponer $\beta_1 = 0$, el ajuste de y estaría dado de la siguiente manera:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Con $\hat{\beta}_1 = 0$:

$$\hat{y} = \hat{\beta}_0.$$

Por el **teorema 2.2** se sustituye la estimación de $\hat{\beta}_0$:

$$\hat{y} = \overline{y} - \overline{x}\hat{\beta}_1.$$

como $\hat{\beta}_1 = 0$:

$$\hat{y} = \overline{y}.$$

Es decir, no hay asociación lineal entre la variable respuesta y con la variable regresora x. Es por ello que se desea demostrar que $\beta_1 \neq 0$, mediante la siguiente prueba de hipótesis:

$$\mathbf{H}_0 : \hat{\beta}_1 = 0 \quad vs. \quad \mathbf{H}_a : \hat{\beta}_1 \neq 0$$

Se puede usar el estadístico de prueba de hipótesis para β_1, con $b1 = 0$. Y se tiene lo siguiente:

$$t^* = \frac{\hat{\beta}_1 - b_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{(n-2)}$$

De esta manera se rechaza la hipótesis nula y se supone que $\beta_1 \neq 0$ cuando $|t^*| > t^{\alpha/2}_{(n-2)}$.

Sin embargo, en el análisis de la varianza, lo que se busca construir es otro estadístico que pueda ser usado cuando se tenga más de una variable explicativa. Es por ello que el matemático y biológo Ronald Fisher (1920), analizó la siguiente igualdad:

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

Lo que busca es fraccionar la distancia del valor observado respecto a la media, por la suma de la distancia del valor estimado a la media más la distancia del valor observado al valor estimado.

Obteniendo el cuadrado de la ecuación anterior:

$$(y_i - \overline{y})^2 = (\hat{y}_i - \overline{y})^2 + (y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \overline{y})(y_i - \hat{y}_i)$$

Sumando sobre todos los valores:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i)$$

Resolviendo la ecuación:

$$2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\overline{y}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Por la **definición 2.2**, se tiene $(y_i - \hat{y}_i) = e_i$:

$$2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\overline{y}\sum_{i=1}^{n} e_i.$$

Por el **teorema 2.3**, se sabe que $\sum_{i=1}^{n} e_i = 0$

$$2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 2\sum_{i=1}^{n}\hat{y}_i e_i$$

Por el **corolario 2**, $\sum_{i=1}^{n}\hat{y}_i e_i = 0$, por lo tanto tenemos:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

De esta manera se construye la **ANOVA**, generalmente se hace un cambio de notación:

- SC_T es la suma de cuadrados, mide la variabilidad de las observaciones del total corregido por la media y es denotado por $SC_T = \sum_{i=1}^{n}(y_i - \overline{y})^2$.
- SC_{reg} es la suma de cuadrados de la regresión, mide la variabilidad de las observaciones y_i y la línea de regresión ajustada, es denotado por $SC_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$.

- SC_{error} es la suma de cuadrados del error, es decir mide la variación residual que queda sin explicar por la línea de regresión, es denotado por $SC_{error} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$.

De esta forma se tiene la siguiente igualdad:

$$SC_T = SC_{reg} + SC_{error}$$

Además se observa que:

- SC_T tiene $n-1$ grados de libertad, ya que la suma $\sum_{i=1}^{n}(y_i - \overline{y})^2$ es el núcleo de $(n-1)S^2 = \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1}$ en el cual al estimar μ con $\overline{y}$ se pierde un grado de libertad.
- SC_{reg} tiene 1 grado de libertad, ya que la regresión sólo tiene una variable indepediente (x).
- SC_{error} tiene $n-2$ grados de libertad, ya que la regresión $SC_{error} = SC_T - SC_{reg}$, en el cual se ha visto que SC_T y SC_{reg} tienen $n-1$ y 1 grado de libertad, respectivamente, así los grados de libertad de $SC_{error} = n-1-1$, entonces SC_{error} tienen n-2 grados de libertad. Sabemos que el modelo de Regresión Lineal con errores normales tiene las siguientes propiedades:
- $\frac{SC_{reg}}{\sigma^2} \sim \chi^2_{(1)}$.
- $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{SC_{error}}{\sigma^2} \sim \chi^2_{(n-2)}$.
- Además SC_{reg} es independiente a SC_{error}.

Debido a un resultado de probabilidad se sabe que si $x \sim \chi^2_{(n)}$ y $y \sim \chi^2_{(m)}$ y si x es independiente a y entonces:

$$\frac{x/n}{y/m} \sim F_{(n,m)}$$

De esta forma se puede aplicar la prueba F de Fisher en el análisis de varianza para probar las hipótesis. La prueba F consiste en dividir SC_{reg} entre sus grados de libertad y éste dividirlo entre SC_{error} que a su vez está dividida entre sus grados de libertad, de esta manera:

$$F = \frac{\frac{SC_{reg}}{1}}{\frac{SC_{error}}{n-2}}$$

Aplicando el resultado:

$$F = \frac{\frac{SC_{reg}}{1}}{\frac{SC_{error}}{n-2}} \sim F_{(1,n-2)}$$

Ahora se ocupará el **Cuadrado Medio** denotado como **CM**, la cual corresponde a la Suma de Cuadrados entre los grados de libertad. Así, se define al cuadrado medio de la regresión CM_{reg} y al cuadrado medio del error CM_{error} como:

$$CM_{reg} = \frac{SC_{reg}}{1} \quad y \quad CM_{error} = \frac{SC_{error}}{n-2}.$$

De igual manera, haciendo el ajuste, la prueba F queda definida como:

$$F = \frac{CM_{reg}}{CM_{error}}.$$

Observamos que $SC_{reg} = \hat{\beta}_1 S_{xy}$ entonces:

$$F = \frac{\hat{\beta}_1 S_{xy}}{CM_{error}}.$$

De esta manera, la región de rechazo para la prueba de hipótesis $H_0 : \beta_1 = 0$ es:

$$\frac{\hat{\beta}_1 S_{xy}}{CM_{error}} > (t_{(n-2)}^{\alpha/2})^2$$

La cual es equivalente a la prueba t.

$$\frac{|\hat{\beta}_1|}{\sqrt{Var(\hat{\beta}_1)}} > t_{(n-2)}^{\alpha/2}.$$

Ya que si $x \sim t_{(n-1)}$ entonces $x^2 \sim F_{(1,n-2)}$.

El siguiente cuadro resume la información anterior, mejor conocida como **Tabla ANOVA:**

	Grados de libertad	*Suma de Cuadrados*	*Cuadrado Medio*	*Prueba F*
Regresión	1	$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \hat{\beta}_1 S_{xy}$	SC_{reg}	$\frac{CM_{reg}}{CM_{error}}$
Error	$n-2$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SC_T - \hat{\beta}_1 S_{xy}$	$\frac{SC_{error}}{n-2}$	–
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \overline{y})^2$	–	–

26.5. Coeficiente de determinación

El coeficiente de determinación, con frecuencia se le asocia a la proporción de la variación explicada por el regresor x, ya que $0 < SC_{reg} < SC_T$ entonces los valores del coeficiente de determinación están entre $0 < R^2 < 1$.

Se define el coeficiente de determinación del modelo de regresión como:

$$R^2 = \frac{SC_{reg}}{SC_T} = 1 - \frac{Sc_{error}}{SC_T}.$$

26.6. Propiedades de R^2

El coeficiente de determinación R^2 satisface las siguientes propiedades:

- $0 \leq R^2 \leq 1$ ya que $0 \leq SC_{error} \leq SC_T$.
- Si $R^2 = 1$ entonces tenemos que $SC_{reg} = SC_T$ y $\frac{SC_{error}}{SC_T} = 0$.
- Si $R^2 = 0$ entonces $SC_{error} = SC_T$.

En particular, se buscan valores de R^2 cercanos a 1, ya que esto indica que la mayor parte de la variabilidad de y es determinada o explicada por el modelo de regresión. La magnitud de R^2 también depende del rango de variabilidad de la variable regresora. En general R^2 aumenta a medida que la propagación de los valores de x aumenta, y disminuye a medida que la propagación de los valores de x disminuyan, siempre que el modelo asumido es correcto.

Valor de R^2	**Tipo de correlación**
$R^2 = 0$	*No hay correlación*
$0 < R^2 < 0.25$	*Correlación muy débil*
$0.25 \leq R^2 < 0.5$	*Correlación débil*
$0.5 \leq R^2 < 0.75$	*Correlación moderada*
$0.75 \leq R^2 < 0.9$	*Correlación fuerte*
$0.9 \leq R^2 < 1$	*Correlación muy fuerte*
$R^2 = 1$	*Correlación perfecta*

26.7. Relación R^2 y la correlación de Pearson

El coeficiente de correlación de Pearson entre $x_1, \ldots, x_n$ y $y_1, \ldots, y_n$ se define como:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

"r" toma valores de -1 a 1, dónde si r toma el valor de 1 decimos que hay una relación lineal positiva, si r toma el valor de -1 decimos que hay una relación lineal negativa y finalmente si toma el valor de 0 decimos que no hay una relación entre variables.

Para el modelo de regresión lineal, en particular, se cumple que:

$$r^2 = R^2.$$

26.7.1. Ejemplo

En las secciones anteriores tomamos los datos de **CALLCENT** y comenzamos a resolver el problema que el gerente nos planteó de poder predecir, de alguna manera, el tiempo promedio que tardarían en procesar un número dado de facturas.

Se ha recolectado, durante un periodo de 30 días, la información sobre el número de facturas procesadas (en nuestrio caso definimos como nuestra variable x) y el tiempo que tardan las mismas (que hemos definido como nuestra variable y).

Verificamos gráficamente que hubiera una relación lineal entre las variables, estimamos los parámetros del modelo de regresión lineal simple con intercepto y sin intercepto. Luego construimos intervalos de confianza para los parámetros estimados, para el valor esperado y la predicción.

Como se mencionó en nuestro capítulo debemos aplicar pruebas estadísticas que nos aseguren que las estimaciones de los parámetros serán distintos de cero.

Prueba de hipótesis para β_0 y β_1

Haremos el cálculo del estadístico de prueba para probar la hipótesis: $\mathbf{H}_0 = \hat{\beta}_0 = 0 \quad vs \quad \mathbf{H}_a \neq 0$ con una confianza del 90 % y sustituimos en la siguiente ecuación:

$$t^* = \frac{\hat{\beta}_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2}}.$$

```
t_est=beta0/sqrt((1/n+x_barra^2/Sxx)*s2_gorro)

t_est
```

```
[1] 5.24827
```

Ahora buscamos el cuantil de la distribución $t - Student$ para comprar nuestro estadístico.

```
qt=qt(.95,n-2)
qt
```

```
[1] 1.701131
```

Como nuestro estadístico (5.2482) es mayor que (1.7011). Entonces rechazamos $\mathbf{H}_0$. Por lo tanto con una confianza del 90 % podemos decir que $\beta_0 \neq 0$.

Ahora haremos la prueba para la hipótesis $\mathbf{H}_0 : \hat{\beta}_1 = 0 \quad vs \quad \mathbf{H}_a : \hat{\beta}_1 \neq 0$ con una confianza del 90 %. Sustituimos en la siguiente ecuación:

$$t^* = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

```
t_est=beta1/sqrt(s2_gorro/Sxx)

t_est
```

```
[1] 13.79718
```

Ahora buscamos el cuantil de la distribución $t - Student$ para comparar nuestro estadístico:

```
qt=qt(.95,n-2)

qt
```

```
[1] 1.701131
```

Como nuestro estadístico (13.7971) es mayor que (1.7011). Entonces rechazamos $\mathbf{H}_0$. Por lo tanto con una confianza del 90 % podemos decir que $\beta_1 \neq 0$.

Con éstas dos pruebas corroboramos que los datos ajustan a un modelo de regresión lineal que considera a la variable facturas y el intercepto, esto debido a que ambos parámetros resultaron ser distintos de cero.

- Si la prueba de hipótesis para β_0 hubiera resultado en no rechazar $\mathbf{H}_0$ entonces, se buscaría ajustar un modelo sin intercepto.
- Si la prueba de hipótesis para β_1 hubiera resultado en no rechazar $\mathbf{H}_0$ entonces, se buscaría ajustar un modelo con otra variable (diferente a las facturas) ya que los datos estarían diciendo que la variable facturas no es significativa para explicar el tiempo promedio de su llegada.

Coeficiente de determinación

Por último vamos a calcular el coeficiente de determinación de nuestro modelo. Como se menciona en nuestro capítulo, este coeficiente mide la proporción de la variación explicada por el modelo en relación a la variación total existente en los datos, y por eso es un valor entre 0 y 1.

Y sustituiremos en la siguiente expresión:

$$R^2 = \frac{\hat{\beta}_1 S_{xy}}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

```
R_2=(beta1*Sxy)/sum((y-y_barra)^2)
R_2
```

```
[1] 0.8717727
```

Lo que quiere decir que nuestro modelo está explicando 87.17 % de la variabilidad observada en la variable facturas. El ideal es tener un modelo que explique el 100 % pero como vimos, debido a que estamos ajustando una recta y esta no pasa por todos los puntos algún error en el ajuste estamos cometiendo.

Capítulo 27

Validación de supuestos

El modelo de regresión lineal es una buena herramienta de estimación, sin embargo, en dicho proceso se hace uso de diversos supuestos que deben cumplirse para que los resultados obtenidos sean acordes a la teoría desarrollada, que hasta el momento no se le había prestado gran atención. Estos supuestos los vimos en la **definición 2.1** para el caso de regresión lineal simple. A manera de resumen y de forma general los supuestos que deben cumplirse son: la esperanza de los errores ϵ tienen media cero, es decir, $\mathbf{E}[\epsilon] = 0$, varianza constante sobre los errores, es decir, $(Var(\epsilon) = \sigma^2)$, los errores no se encuentran correlacionados entre si, es decir, $(Cov(\epsilon_i, \epsilon_j) \forall i \neq j)$, por último, los errores tienen distribución normal con media cero y varianza σ^2, es decir, $\epsilon \sim \mathbf{N}(0, \sigma^2)$.

Es por ello que analizaremos y verificaremos el cumplimiento de los supuestos, la mayoría de estas validaciones se basan bajo el principio del análisis de residuales.

27.1. Análisis de residuales

Anteriormente definimos los residuales como $e_i = y_i - \hat{y_i}$, el cual su nombre deriva de la obtención del residual o diferencia que existe entre la línea de regresión ajustada y los valores observados de la variable respuesta y_i; esta cantidad residual es un buen ajuste debría ser cercana a cero, pues cuando esto sucede se tiene que $y_i \approx \hat{y_i}$ debido a ello se opta por trabajar con los residuales para verificar el cumplimiento de los supuestos.

Muchas pruebas usan como base modificaciones o tipos específicos de residuos, es por esta razón que se indagará sobre los diversos tipos más conocidos.

residuales ordinarios

Los residuales ordinarios o simplemente residuales, miden la diferencia entre la línea de regresión y la variable respuesta y_i, es por ello que se definen como:

$$e_i = y_i - \hat{y_i} \qquad i = 1, 2, \ldots, n.$$

La desventaja que presentan estos residuales, es que depende de cada y_i por lo que observaciones atípicas puede generar grandes residuales, ocasionando que pueda existir gran variabilidad al considerarse los errores en forma conjunta.

residuales estandarizados

Una forma de disminuir esta variabilidad consiste en dividir a los residuales e_i entre la varianza global, es decir:

$$d_i = \frac{e_i}{\sqrt{\sigma^2}}$$

Una propiedad importante es que si e_i sigue una distribución normal, entonces al dividirlo entre la desviación estándar, se tiene que los residuales estandarizados siguen una distribución normal estándar.

residuales estudentizados

Los residuales estudentizados se basan en la idea de involucrar la varianza de cada observación en el cálculo de los residuales, pues teóricamente los residuales no tiene varianza constante.

Como se demostrará en el corolario A (ver *Apéndice*), se sabe que $e = (I - H)\underline{Y}$; en donde $H = X(X'X)^{-1}X'$ la definimos como **matriz sombrero**. De esta manera calculando la varianza de los errores se tiene:

Corolario 6 Sea $\underline{e}$ los residuales del modelo entonces la varianza de $\underline{e}$ está dada por:

$$Var(\underline{e}) = \sigma^2(I - H)$$

Demostración:

$$Var(\underline{e}) = Var((I - H)\underline{Y})$$

$$Var(\underline{e}) = (I - H)'Var(\underline{Y})(I - H)$$

Sabemos que $(I - H)$ es simétrica (ver *apéndice*)

$$Var(\underline{e}) = (I - H)(I - H)Var(\underline{Y})$$

Sabemos quue $(I - H)$ es idempotente (ver *apéndice*)

$$Var(\underline{e}) = (I - H)Var(\underline{Y})$$

Por el **teorema B** (demostrado en *apéndice*)

Teorema B Sea una variable de interés $\underline{Y}$, llamada **dependiente**, relacionada con dos o más variables explicativas$x_1, x_2, \ldots, x_k$, entonces:

a) $\mathbf{E}[\underline{Y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$.

b) $Var(\underline{Y}) = \sigma^2$.

$$\therefore Var(\underline{e}) = \sigma^2(I - H).\blacksquare$$

La varianza de cada residual no es constante para todas las observaciones es por ello que el resultado depende de la siguiente forma:

$$Var(e_i) = \sigma^2(1 - h_{ii})$$

donde h_{ii} correspondiente al $i - ésimo$ elemento de la diagonal de la matriz sombrero H.

Cada residual se divide entre su varianza obteniendo de esta forma lo que se conoce como **residuales estudentizados**, los cuales se definen como:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

Debido a la construcción de los **residuales estudentizados** se logra estandarizar de mejor manera a los residuales del modelo. Si se cumple que cada residual e_i sigue una distribución normal entonces los residuales estudentizados siguen una distribución aproximada a una $t - student$ con $n - k - 1$ grados de libertad.

27.2. Supuesto de normalidad

Como se mencionó, en el modelo de regresión lineal se tiene el supuesto de que los errores tienen distribución normal con media cero y varianza σ^2. Debido a la construcción del modelo, este supuesto puede presentar desviaciones en la distribución, en la cual, serias desviaciones de ésta puede ocasionar que las estimaciones dadas sean erróneas, principalmente se pueden observar en la construcción de intervalos de confianza o pruebas de hipótesis, ya que las cantidades pivotales se basan en la premisa de normalidad en los errores, por lo que desviaciones significativas provocan una mala aproximación o cálculo; además que pruebas de hipótesis subsecuentes basadas en distribuciones como la t de Student o la F de Fisher presentan malas estadísticas para la toma de decisiones.

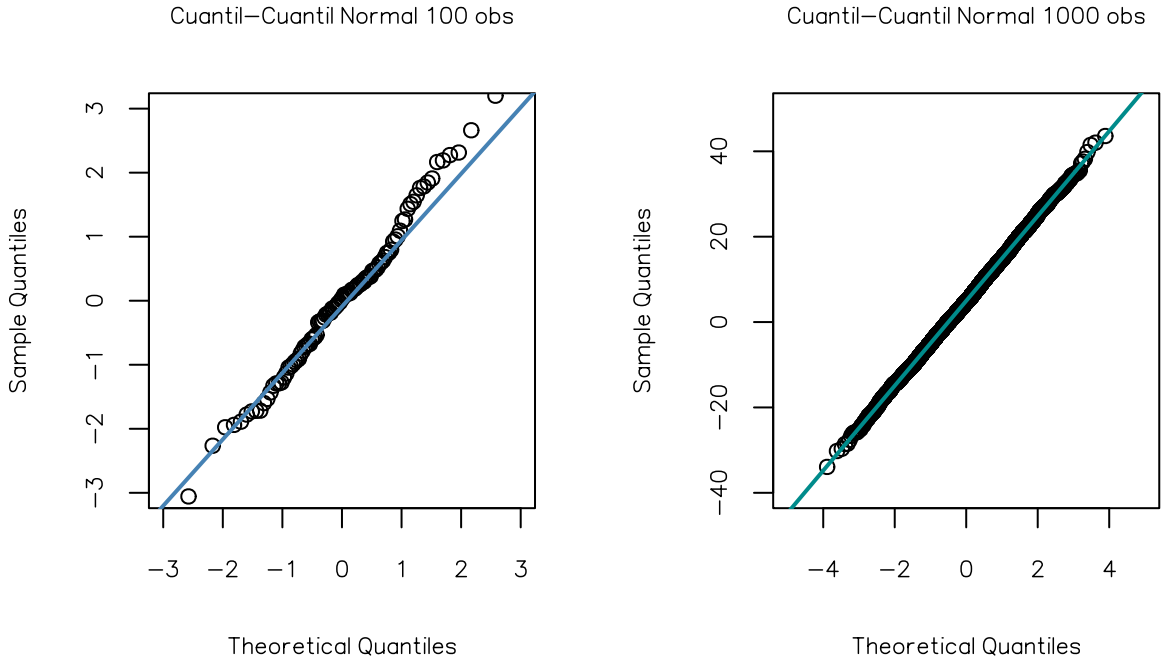
27.2.1. Validación del supuesto de normalidad

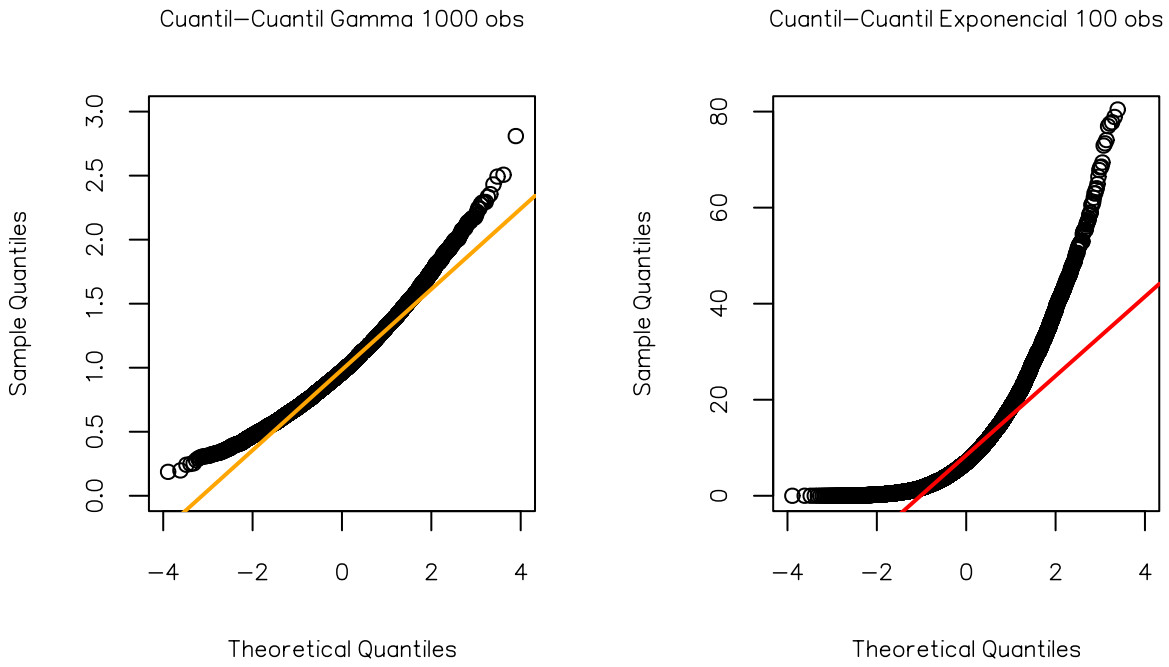
Para validar el supuesto de normalidad existen varios métodos, el primero de ellos es de forma visual, a través de las denominadas gráficas cuantil-cuantil, o también denominadas QQ-plot, ésta gráfica es muy usada para comprobar si una muestra sigue una determinada distribución. El procedimiento se basa en ordenar los residuales (e_i) en orden ascendente los cuales se mostrarán en el eje horizontal X, mientras que el eje vertical Y se muestra al valor esperado de la estadística de orden de una distribución normal. El valor esperado se denota como:

$$\mathbf{E}[e_i] = \phi^{-1}\left[\frac{i-\frac{1}{2}}{n}\right].$$

Donde la función ϕ^{-1} es la distribución inversa de una normal. Si los resultados se comportaran de manera normal la gráfica de cuantiles de los errores pareciera que siguen una marcada línea de 45°, por lo que valores fuera de la línea recta indicarían una distribución no normal.

Diversas distribuciones en gráficas QQ-plot





Lo que podemos observar, es que las dos muestras superiores siguen una distribución normal, sin embargo, la primera tiene pocas observaciones por lo que algunos puntos se encuentran cercanos a la recta azul, línea que representa la distribución normal ideal, pero pocas veces la muestra toca esta línea y las colas presentan mucha variabilidad, la gráfica de la derecha al tener tamaño de muestra mayor, se aprecia que muchos datos caen sobre la recta, presentando ligeras irregularidades en la cola, tanto superior como inferior, por lo que se puede asumir, que las muestras siguen una distribución normal.

Por último, las gráficas inferiores, representan a una muestra con distribución gamma y exponencial, respectivamente, al no seguir una distribución normal, los datos salen completamente de la línea recta marcada, por lo que es evidente que no se comportan con normalidad.

Otro procedimiento para validar el supuesto de normalidad, es mediante pruebas de bondad de ajuste, sin embargo hay que tener cuidado, ya que los errores no son independientes entre si, debido a que están correlacionados, mientras que las pruebas de bondad de ajuste asumen precisamente que las observaciones son independientes entre si. Windfried Stute demostró que pruebas como la Anderson-Darling convergen a la distribución teórica aunque la independencia de los errores no se cumpla, debido a que se basan en el proceso empírico. Sin embargo, las pruebas deben de usarse como una medida de aproximación y no como regla de decisión.

27.3. Supuesto de linealidad

En la construcción del modelo de regresión lineal se asume que la relación entre X_j y Y es lineal, para cada $j \in 1, \ldots, k$, con k el número de variables regresoras con el que fue ajustado el modelo.

Sin embargo, la anterior afirmación no siempre se cumple, es por ello que se valida este supuesto de manera gráfica. Debido a lo complejo que es una gráfica en más de tres dimensiones, se verá que en el modelo de regresión múltiple con k regresores se ajusta un hiperplano de dimensión k.

Cuando $k \geq 4$ se recomienda realizar gráficas individuales para comprobar la linealidad de la variable explicativa X_j y la variable del interés Y.

Aunque este método proporciona una buena aproximación para saber si dos variables son lineales o no, este tipo de análisis puede proporcionar conclusiones erróneas cuando los coeficientes tienen magnitudes distintas ya que se analiza la relación marginal de la variable respuesta con cada variable explicativa. Es por ello que se opta trabajar mediante un análisis de residuales, en este análisis se grafican los errores estandarizados contra los valores observados de cada variable explicativa, en el cual el cumplimiento de la hipótesis daría como resultado ruido blanco con media 0 y varianza σ^2.

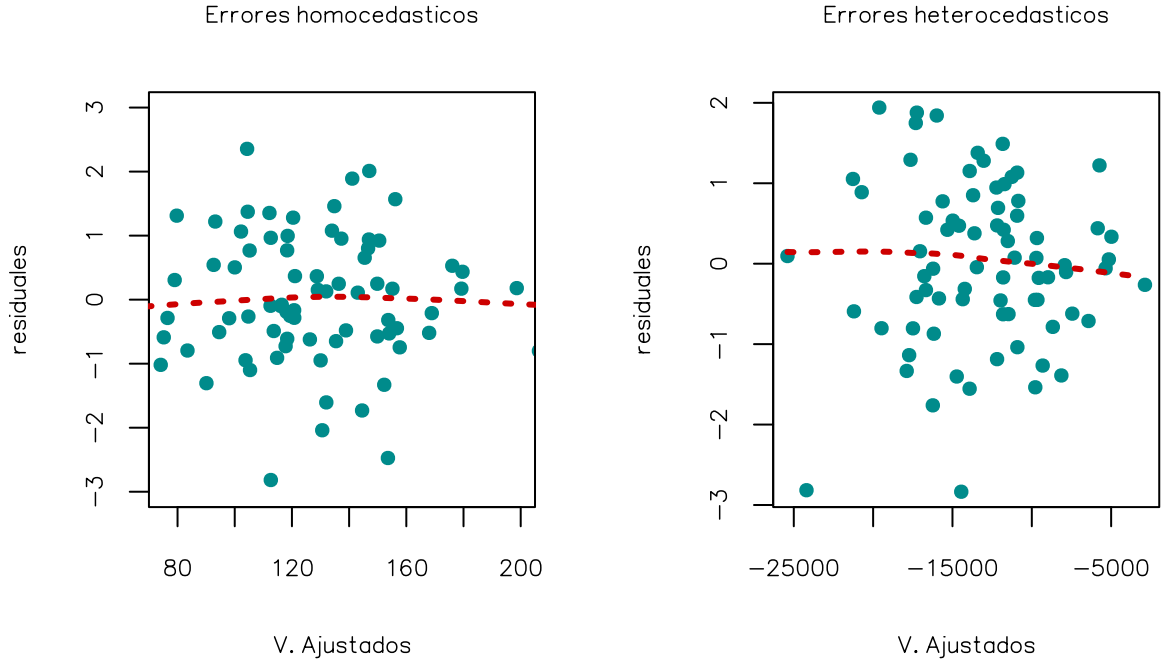
Cuando se detecte problemas de linealidad entre variables explicativas y la variable de interés, el ajuste del modelo es malo, debido a que la varianza presenta problemas en la estimación y por consecuencia estadísticas, se usa σ^2 en su desarrollo lo cual hereda errores en sus cálculos.

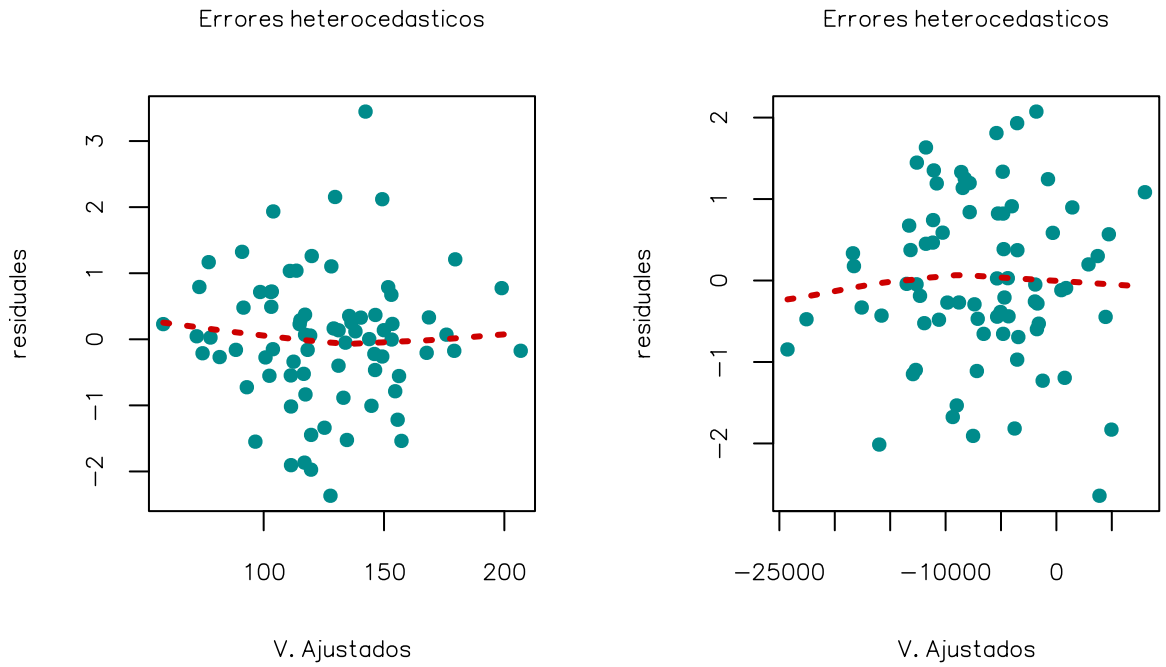
27.4. Supuesto de homocedasticidad

Se dice que una muestra es homocedástica cuando la varianza es constante a lo largo de todas las observaciones, es decir, no varia conforme se presentan nuevas observaciones. Mientras una muestra heterocedástica se presenta cuando hay variaciones de la varianza conforme se presentan nuevas observaciones.

Las desviaciones en el supuesto de homocedasticidad pueden observarse mediante gráficas, la más óptima para el análisis es realizar una gráfica de dispersión en el que se muestre la relación entre los valores ajustados $\underline{\hat{Y}}$ contra los residuales estandarizados d_i. Si la varianza es constante entonces la gráfica fluctuará entre el eje horizontal de manera simétrica, asemejando a una distribución uniforme, y sin seguir algún tipo de patrón, ya que típicamente se considera que la mayor parte de los errores deben estar contenidos en franjas horizontales delimitados por el eje vertical entre $y = -2$ y $y = 2$.

Muestra homocedástica y una Muestra heterocedástica





Como estamos observando, la primera imagen corresponde a una muestra homocedástica pues los errores se distribuyen a lo largo del eje horizontal, además que éstos fluctúan entre $y = -2$ y $y = 2$ distribuidos de una manera simétrica. Mientras que la segunda gráfica muestra que los errores fluctúan entre -2 y 2, sin embargo los resultados siguen un patrón de tender hacia a la media conforme se presentan nuevas observaciones, simulando un megáfono, por lo que se dice que la muestra sigue una tendencia heterocedástica.

Por último las gráficas ubicadas en la parte inferior, muestran como conforme se presentan nuevas observaciones los errores se alejan de la media, por lo que son muestras heterocedásticas.

Existen métodos más formales para probar homocedasticidad mediante pruebas de hipótesis, una de las que se desarrollarán a continuación.

27.4.1. Prueba de Breusch-Pagan

La prueba de Breusch-Pagan fue desarrollada en 1979 por los estadísticos Trevor Breusch y Adrian Pagan, se utiliza para determinar si una muestra presenta problemas de homocedasticidad o heterocedasticidad en un modelo de regresión lineal. El método consiste en analizar si la varianza estimada de los residuales de una regresión depende directamente de los valores obtenidos de las variables independientes, uno de los supuestos de esta prueba es que los errores deben comportarse con normalidad.

La prueba de Breusch-Pagan contrasta como hipótesis nula el cumplimiento de homocedasticidad, por lo que se tiene la siguiente prueba de hipótesis.

$\mathbf{H}_0$: Muestra homocedástica i.e. $\sigma_j^2 = \sigma^2 \quad vs \quad \mathbf{H}_a$: Muestra Heterocedástica i.e. $\sigma_j^2 \neq \sigma^2 \quad \forall\ j = 1, \ldots, n.$

El procedimiento se basa en calcular los residuales estandarizados al cuadrado $\left(\tilde{e}_j^2 = \frac{e_j^2}{\sigma^2}\right)$; Con ello se realiza una regresión lineal tomando como variable respuesta a cada $\tilde{e}_j$ al cuadrado y con variables explicativas dentro del conjunto de variables exógenas Z:

$$\tilde{e}_j^2 = \gamma_0 + \gamma_1 Z_1 + \ldots + \gamma_n Z_n$$

Después se procede a calcular la suma de cuadrados de la regresión del modelo con los errores estandarizados divididos entre 2, $\frac{SC_{reg}}{2}$, donde $SC_{reg} = \underline{Y}'(H - \frac{1}{n}J)\underline{Y}'$; Breusch-Pagan descubrieron que este estadístico sigue asintóticamente una distribución ji-cuadrada con k grados de libertad, siendo k el número de

variables del modelo. Por lo que la región de rechazo para H_0 sucede cuando la estadística $\frac{SC_{reg}}{2}$ es mayor al cuantil de una ji-cuadrada con k grados de libertad con un nivel de significancia α, es decir:

$$\frac{SC_{reg}}{2} > \chi_k^{2(\alpha)}.$$

En otro caso no se tiene evidencia sificiente para rechazar la hipótesis nula.

En R la prueba de Breusch-Pagan puede ser fácilmente implementada, suponga que se tiene una muestra en el que se ha implementado el procedimiento de regresión lineal en R ($lm(Y \sim X)$), por lo que aplicándo el siguiente código se tiene:

```
    studentized Breusch-Pagan test

data:  modelo1
BP = 0.026689, df = 1, p-value = 0.8702
```

Se observa que en el anterior caso particular, la prueba supone como válida la hipótesis nula; la homocedasticidad de muestra debido a que el $p - value$ es alto, (de 0.8702), lo que conlleva a que no se rechace la hipótesis nula con un nivel de significancia $\alpha = 0.05$, por lo que se acepta que la muestra se comporta con homocedasticidad.

27.4.2. Prueba de White

La prueba de White es similar a la prueba de Breusch-Pagan, sin embargo, se considera que ésta prueba es más general pues no requiere que los errores sigan una distribución normal.

La prueba de White fue propuesta por Hilbert White en 1980, como alternativa a la prueba de Breusch-Pagan, el procedimiento es similar, se analiza si la varianza estimada de los residuos de una regresión depende directamente de los valores obtenidos.

El test contrasta como hipótesis nula el cumplimiento de homocedasticidad, por lo que se tiene la siguiente prueba de hipótesis.

$\mathbf{H}_0$: Muestra homocedástica i.e. $\sigma_j^2 = \sigma^2 \quad vs \quad \mathbf{H}_a$: Muestra Heterocedástica i.e. $\sigma_j^2 \neq \sigma^2 \quad \forall\, j = 1, \ldots, n.$

El procedimiento se basa en calcular los residuales estandarizados al cuadrado $\left(\tilde{e}_j^2 = \frac{e_j^2}{\sigma^2}\right)$; Con ello se realiza una regresión lineal tomando como variable respuesta a cada $\tilde{e}_j$ al cuadrado y el producto cruzado de variables explicativas dentro del conjunto de variables exógenas Z:

$$\tilde{e}_j^2 = \gamma_0 + \gamma_1 Z_{1i} + \ldots + \gamma_k Z_{ki} + \gamma_{k+1} Z_{1k}^2 + \ldots + \gamma_{k+k} Z_{1k} Z_{ki} + \gamma k + k + 1 Z_{2k} Z_{1k} + \ldots + \gamma t k Z_{kk}^2 + \epsilon$$

Del anterior ajuste de regresión se procede a calcular el coeficiente de determinación $R^2 = \frac{SC_{reg}}{SC_T}$ y sea n el tamaño de la muestra, entonces la estadística nR^2 sigue asintóticamente una distribución ji-cuadrada con k grados de libertad, siendo k el número de variables del modelo original.Por lo que la región de rechazo para H_0 sucede cuando la estadística nR^2 es mayor al cuantil de una ji-cuadrada con k grados de libertad con un nivel de significancia α, es decir:

$$nR^2 > \chi_k^{2(\alpha}).$$

En otro caso no se tiene evidencia suficiente para rechazar la hipótesis nula.

```
White's Test for Heteroskedasticity:
====================================
```

```
No Cross Terms

H0: Homoskedasticity
H1: Heteroskedasticity

Test Statistic:
11.0525

Degrees of Freedom:
12

P-value:
0.5244
```

Se observa que en el anterior caso particular la prueba de White supone como válida la hipótesis nula la homocedasticidad de la muestra, debido a que el $p - value$ es alto, (de 0.5244), lo que conlleva a que no se rechace la hipótesis nula con un nivel de significancia $\alpha = 0.05$, por lo que se acepta que la muestra se comporta con homocedasticidad.

27.4.3. Ejemplo

En las secciones anteriores tomamos los datos de **CALLCENT** y comenzamos a resolver el problema que el gerente nos planteó de poder predecir, de alguna manera, el tiempo promedio que tardarían en procesar un número dado de facturas.

Se ha recolectado, durante un periodo de 30 días, la información sobre el número de facturas procesadas (en nuestrio caso definimos como nuestra variable x) y el tiempo que tardan las mismas (que hemos definido como nuestra variable y).

Verificamos gráficamente que hubiera una relación lineal entre las variables, estimamos los parámetros del modelo de regresión lineal simple con intercepto y sin intercepto. Luego construimos intervalos de confianza para los parámetros estimados, para el valor esperado y la predicción. Realizamos pruebas de hipótesis sobre los estimadores de los parámetros. Calculamos el coeficiente de determinación y realizamos un análisis de varianza sobre el modelo seleccionado.

Como se mencionó en nuestro capítulo debemos verificar que los supuestoshechos para ajustar este modelo de regresión lineal simple se cumplen.

Esta vez vamos a ocupar la función en R para el modelo de regresión que es **lm()**.

```
M1=lm(Tiempo~Facturas)
M1
```

```
Call:
lm(formula = Tiempo ~ Facturas)

Coefficients:
(Intercept)     Facturas
    0.64171      0.01129
```

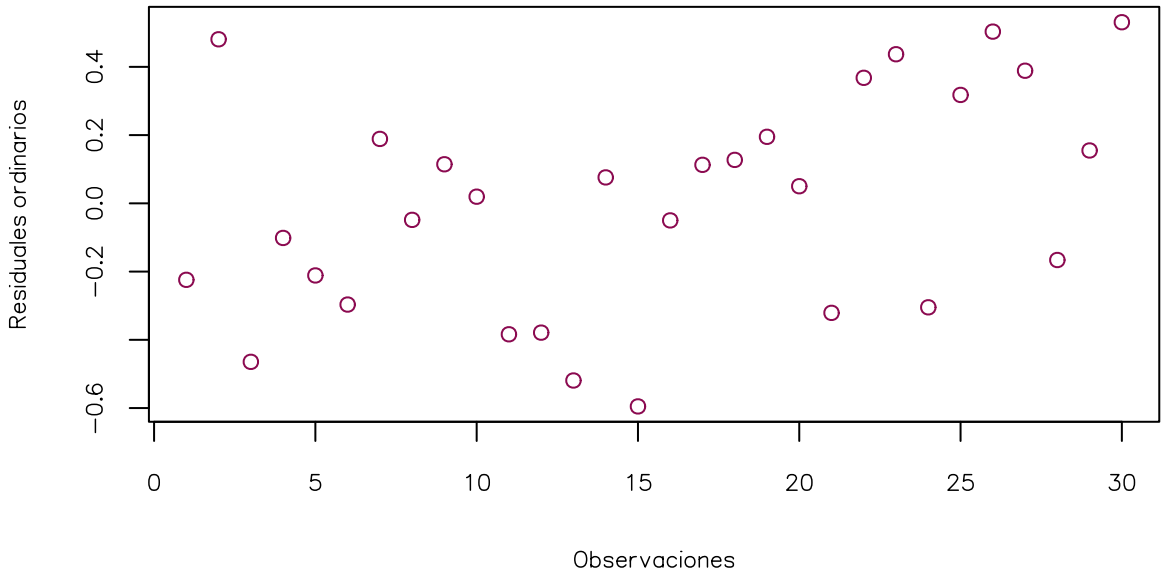
Residuales

Entonces primero calcularemos los diferentes residuales vistos. Se presentaran solamente los primeros 6 residuos de las 30 observaciones y un diagrama de dispersión.

- **Los residuales ordinarios**

```
head(M1$residuals)
```

```
         1          2          3          4          5          6
-0.2241648  0.4807915 -0.4645390 -0.1014177 -0.2113303 -0.2966252
```

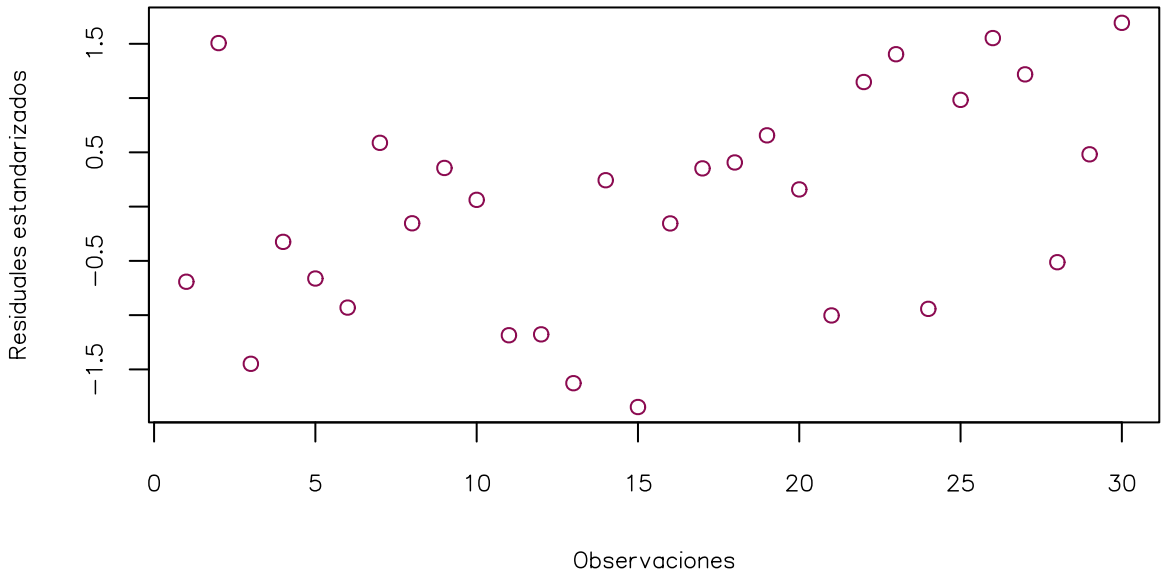


- **Los residuales estandarizados**

```
head(rstandard(M1))
```

```
         1          2          3          4          5          6
-0.6921686  1.5065958 -1.4483299 -0.3248760 -0.6625061 -0.9303669
```

```
plot(rstandard(M1),col="deeppink4",ylab="Residuales estandarizados", xlab="Observaciones")
```

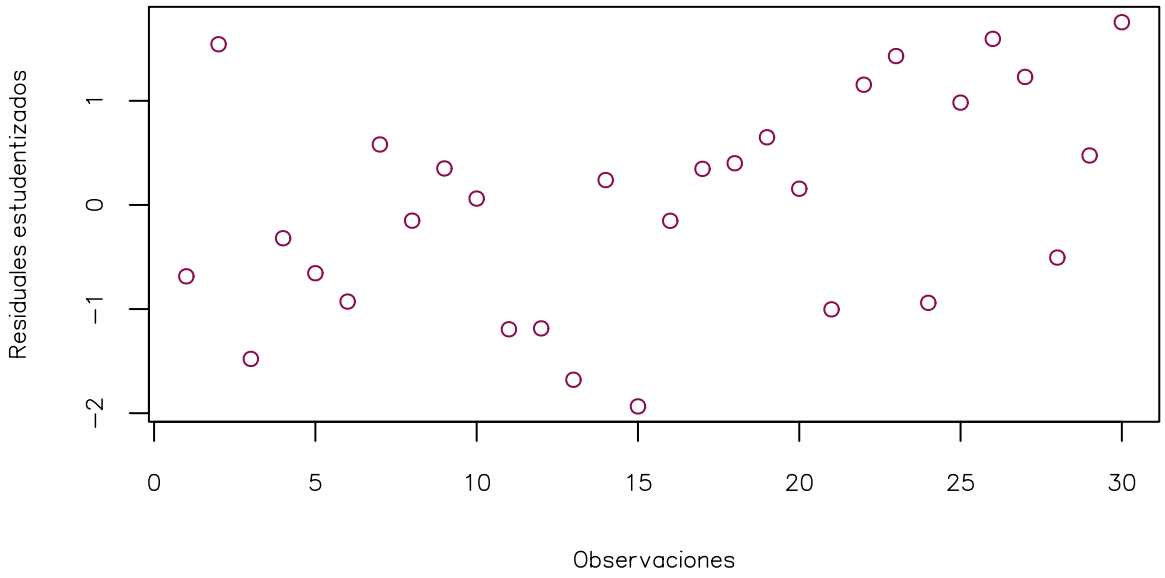


- **Los residuales estudentizados**

```
head(rstudent(M1))
```

```
         1          2          3          4          5          6
-0.6855868  1.5433247 -1.4786993 -0.3196249 -0.6557278 -0.9280596
```

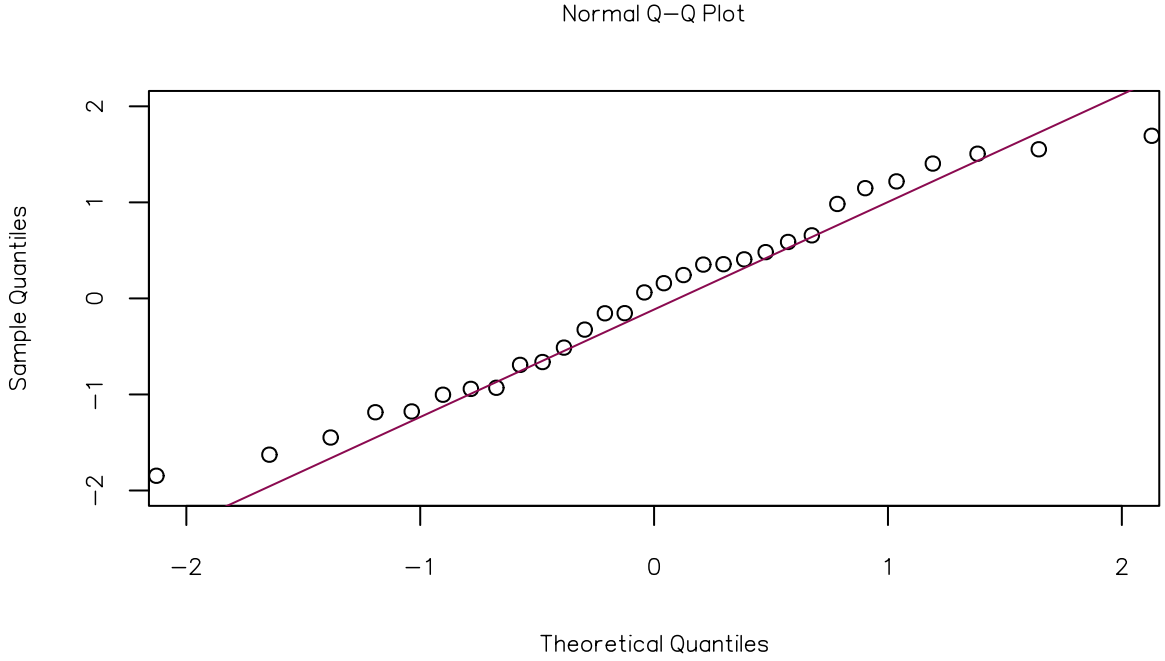
```
plot(rstudent(M1),col="deeppink4",ylab="Residuales estudentizados", xlab="Observaciones")
```



Validación de supuesto de normalidad

Para validar gráficamente la normalidad de los errores debemos graficar los errores contra los cuantiles de la distribución normal. Para esto aplicaremos la función en R **qqnorm()** y con **qqline()** obtenemos la recta diagonal que nos servirá para ver que tan lejos o cerca de la distribución normal están cayendo los residuales del modelo.

```
qqnorm(rstandard(M1),ylim = c(-2,2),xlim = c(-2,2))
qqline(rstandard(M1),distribution = qnorm,col="deeppink4")
```



Podemos observar que la parte central de la distribución si se ajusta a una distribución normal, sin embargo, en los extremos los residuales ya no se comportan como una distribución normal.

Podemos aplicar la prueba de bondad de ajuste **Lilliefors para normalidad** vista en Bondad de Ajuste:

```
nortest::ad.test(rstandard(M1))
```

```
    Anderson-Darling normality test

data:  rstandard(M1)
```

```
A = 0.2675, p-value = 0.6615
```

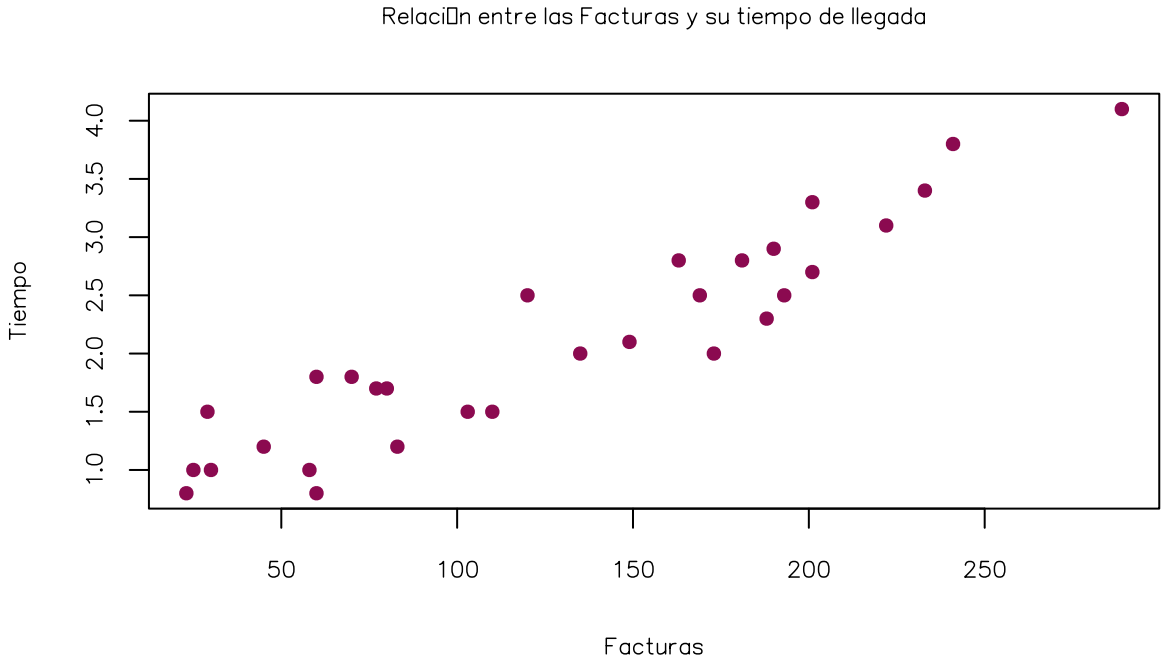
```
lillie.test(rstandard(M1))
```

```

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  rstandard(M1)
D = 0.088454, p-value = 0.7946
```

Como el valor del $p - value$ es mayor al nivel de significancia $\alpha = 0.05$ entonces no rechazamos H_0, es decir nuestros residuales tienen distribución normal.

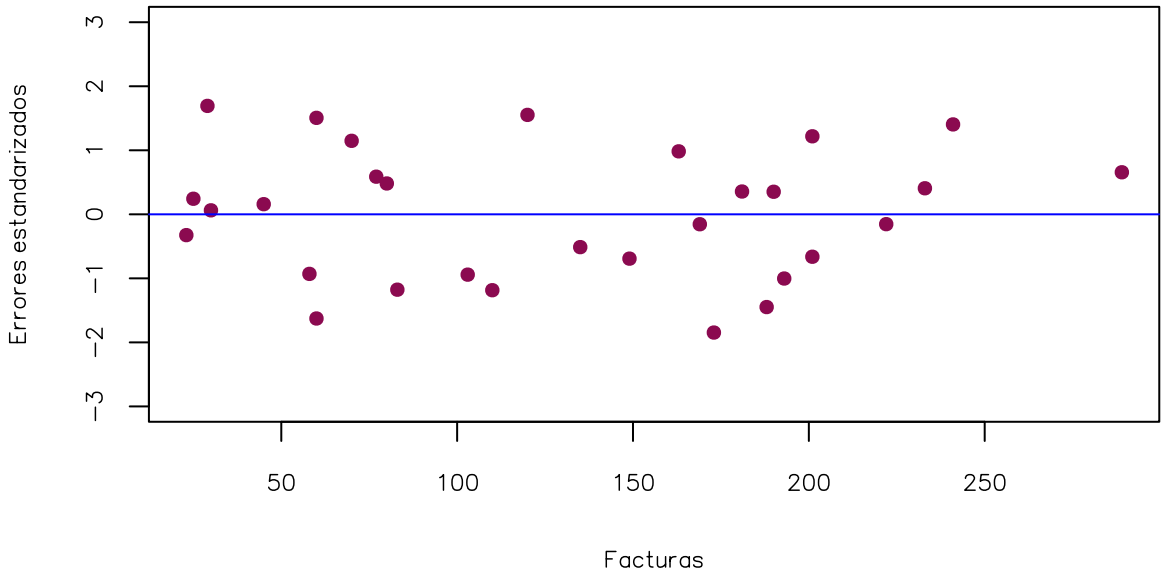
Supuesto de linealidad

El supuesto de linealidad lo verificamos gráficamente haciendo el diagrama de dispersión entre las variables como lo hicimos anteriormente.



Como lo mencionamos en su tiempo al observar nuestros datos nos grita que existe una relación lineal entre las variable facturas y tiempo empleado en ellas.

Como mencionamos en el capítulo, también se pueden graficar los errores estandarizados contra los valores observados de la variable explicativa.

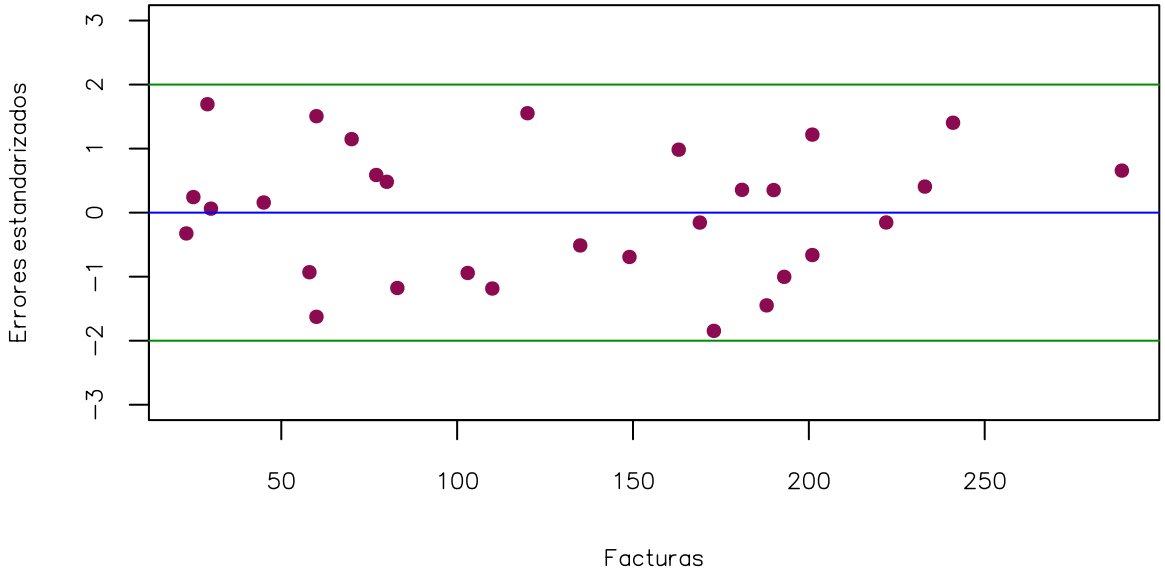


En la gráfica anterior se observa un patrón aleatorio de los residuales estandarizados, esto indica que el modelo lineal es adecuado.

Un punto igual de importante es que no hay presencia de datos atípicos, ya que ningún residual está fuera de las bandas superior e inferior. Datos influyentes tampoco están presentes, pues no hay residuales que estén en alguna dirección lejana a los demás.

Supuesto de Homocedasticidad

Se dice que una muestra es homocedástica cuando la varianza es constante a lo largo de todas las observaciones, es decir, no varia conforme se presentan nuevas observaciones.



- Si la varianza es constante entonces la gráfica fluctuaráentre el eje horizontal de manera simétrica, y sin seguir algún patrón, y se espera que la mayor parte de los errores estén contenidos en franjas horizontales delimitados por el eje entre -2 y 2. En éste ejemplo la dispersión regular de los residuales dentro de las Bandas superior e inferior y que no haya residuales que se alejen tanto de la Banda 0, indican varianza constante.

Adicionalmente aplicaremos las pruebas vistas en el capítulo para tener certeza estadística de la validez del supuesto de homocedasticidad.

Prueba de Breusch-Pagan

```
bptest(M1)
```

```
    studentized Breusch-Pagan test

data:  M1
BP = 0.13226, df = 1, p-value = 0.7161
```

El valor del $p - value$ es mayor, por lo que la hipótesis de homocedasticidad no se rechaza.

Prueba White

```
[1] "Dia"      "Facturas" "Tiempo"
```

```
dataset=data.frame(x, y)
model1= VAR(dataset, p = 1)
whites.htest(model1)
```

```
White's Test for Heteroskedasticity:
====================================

 No Cross Terms

 H0: Homoskedasticity
 H1: Heteroskedasticity

 Test Statistic:
 11.8749

 Degrees of Freedom:
 12

 P-value:
 0.4558
```

El $p - value$ es mayor, por lo que la hipótesis de homocedasticidad no se rechaza.

Supuesto de No Correlación

El estadístico de Durbin-Watson es una estadística de prueba que se utiliza para detectar la presencia de autocorrelación (una relación entre los valores separados el uno del otro por un intervalo de tiempo dado) en los residuales de un análisis de la regresión.

Las hipótesis que se plantean en la prueba de Durbin-Watson es:

$\mathbf{H}_0$: La autocorrelación de los residuales es igual a 0 vs $\mathbf{H}_a$: La autocorrelación de los residuales es $\neq 0$

En R se puede hacer la prueba de Durbin Watson con el comando **dwtest()**.

```
dwtest(M1)
```

```
    Durbin-Watson test

data:  M1
DW = 1.7604, p-value = 0.2558
alternative hypothesis: true autocorrelation is greater than 0
```

De acuerdo con el $p - value$ los residuales no están correlacionados.

Conclusiones

Los supuestos hechos sobre los residuales se cumplen, por lo tanto el modelo propuesto es **totalmente adecuado** para **predecir** el tiempo promedio que tomará procesar un número de facturas dado:

$$\text{Tiempo promedio estimado} = 0.6417 + 0.01129 * \text{Número de Facturas Procesadas}$$

* **Puntos importantes**

1. El gerente del departamento de ventas de **CALLCENT** podrá predecir el tiempo promedio en el que se procesará un número de facturas dado utilizando el modelo ajustado. Se sugiere realizar estimaciones dentro del rango de la dispersión de los datos, de lo contrario la variabilidad aumenta y podría tenerse estimaciones no tan precisas.
2. La ausencia de datos atípicos e influyentes indica que no hay factores que estén afectando el proceso de facturas y su tiempo empleado.
3. La cantidad de horas en la que tardarían en procesar una factura oscila en el intervalo (0.0096 hrs, 0.01296 hrs) a una confianza del 95 %. Puntualmente se estima que las horas requeridas para procesar una factura es 0.01129 hrs.

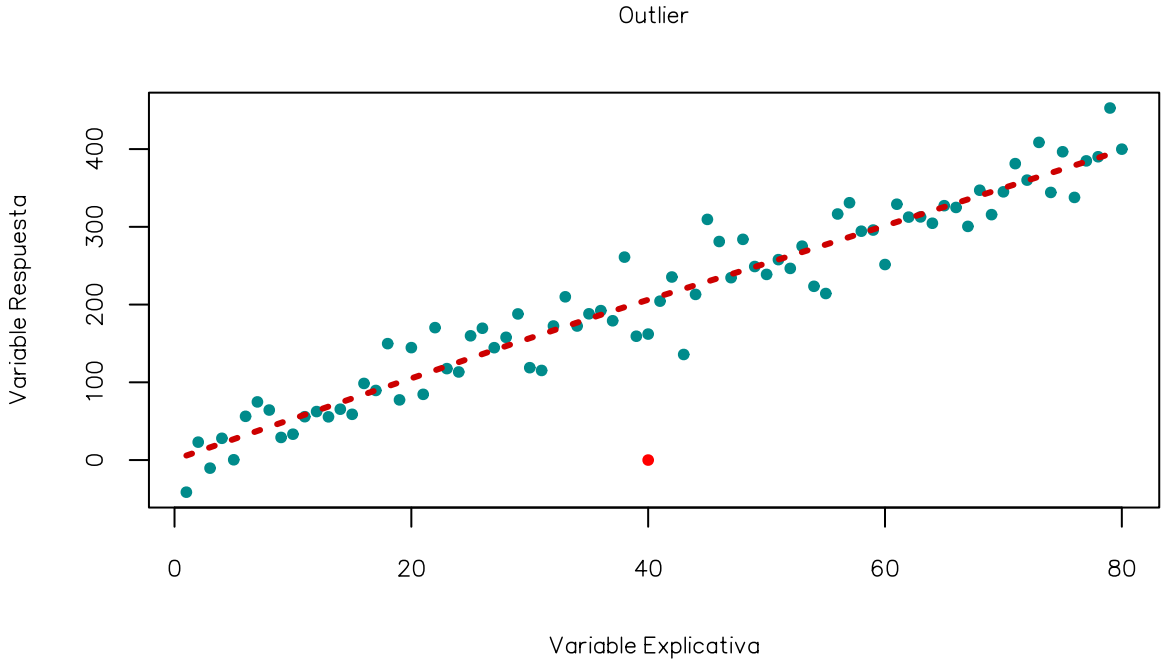
4.- En este caso, el valor de $\hat{\beta}_0$ (intercepto) no tiene una interpretación de acuerdo al contexto del problema.

27.5. Valores outlier e influyentes

Una vez que se ha verificado el cumplimiento de los supuestos en el modelo de regresión, se procede a examinar puntualmente cada observación en búsqueda de valores atípicos o de gran influencia en el modelo.

27.5.1. Valores outlier

Los valores atípicos, también conocidos por la terminología inglesa *outlier*, son observaciones de la muestra aleatoria que no se comportan como el resto de los elementos que conforman el conjunto de datos, gráficamente, la observación con valor atípico no sigue la tendencia que de manera general sigue la muestra aleatoria, lo veremos en la siguiente figura, en el cual el punto rojo sobresale de toda la muestra marcada por puntos azules, por lo que la observación puede ser catalogada como un outlier o valor atípico.

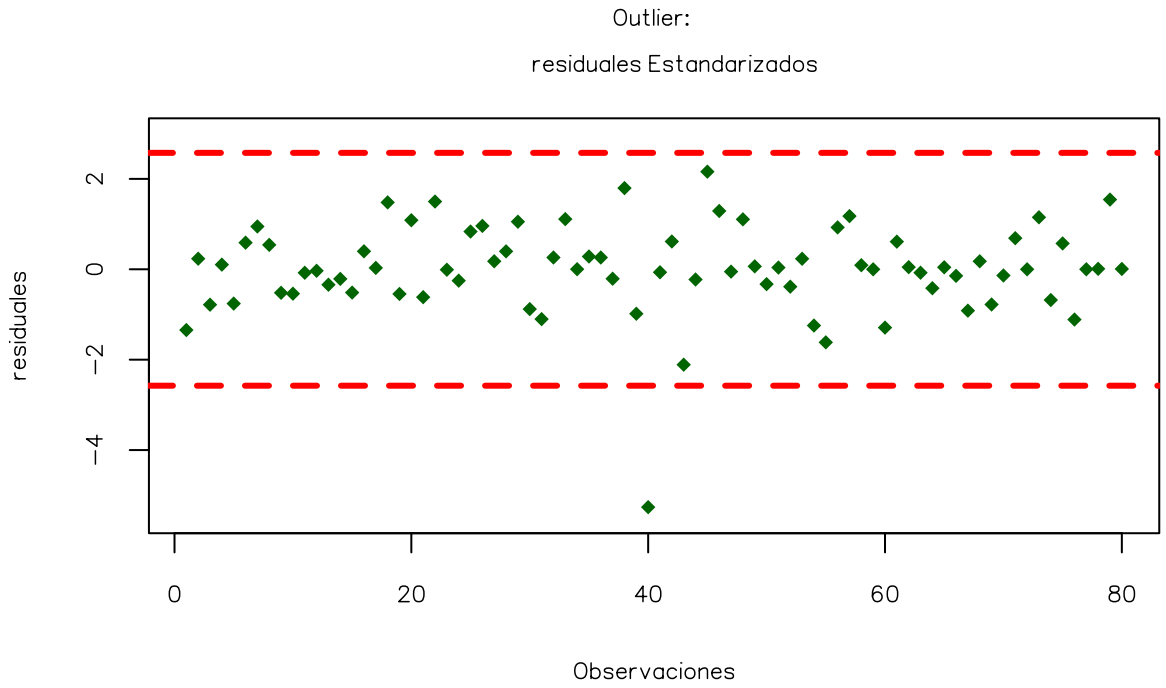


Otra manera de detectar a los posibles valores atípicos es por medio de un análisis de residuales. Dicho análisis consiste en obtener los residuales, ya sean los estandarizados o estudentizados y observar si éstos son mayores o menores a comparación de un punto crítico con nivel de significancia α. Si se escoge trabajar

con los residuales estandarizados sigue una distribución normal con media 0 y varianza σ^2. Por lo que los residuales superiores o inferiores del punto crítico $\pm Z_{1-\alpha/2}$ son considerados como un posible *outlier* con un nivel de significancia α. Por otra parte, si se decide trabajar con los residuales estudentizados entonces el punto crítico está determinado por los residuales que se encuentren por arriba o por abajo de la banda determinada por el cuantil $\pm\, t_{1-\alpha/2,n-k-1}$. Es decir, se tiene evidencia de un valor atípico con nivel de significancia α cuando suceda alguna de las siguientes dos desigualdades:

$$\mid d_i \mid \geq Z_{1-\alpha/2}$$
$$\mid r_i \mid \geq t_{1-\alpha/2,n-k-1}$$

Continuando con el ejemplo anterior, los posibles valores atípicos pueden ser visualizados en la siguiente figura, con un nivel de confianza del 99 %, y usando los residuales estandarizados para el análisis.



Como se mencionó, es más que evidente que el punto atípico corresponde a la observación 40 la cual contrasta y sale de las bandas marcadas por el cuantil de la normal.

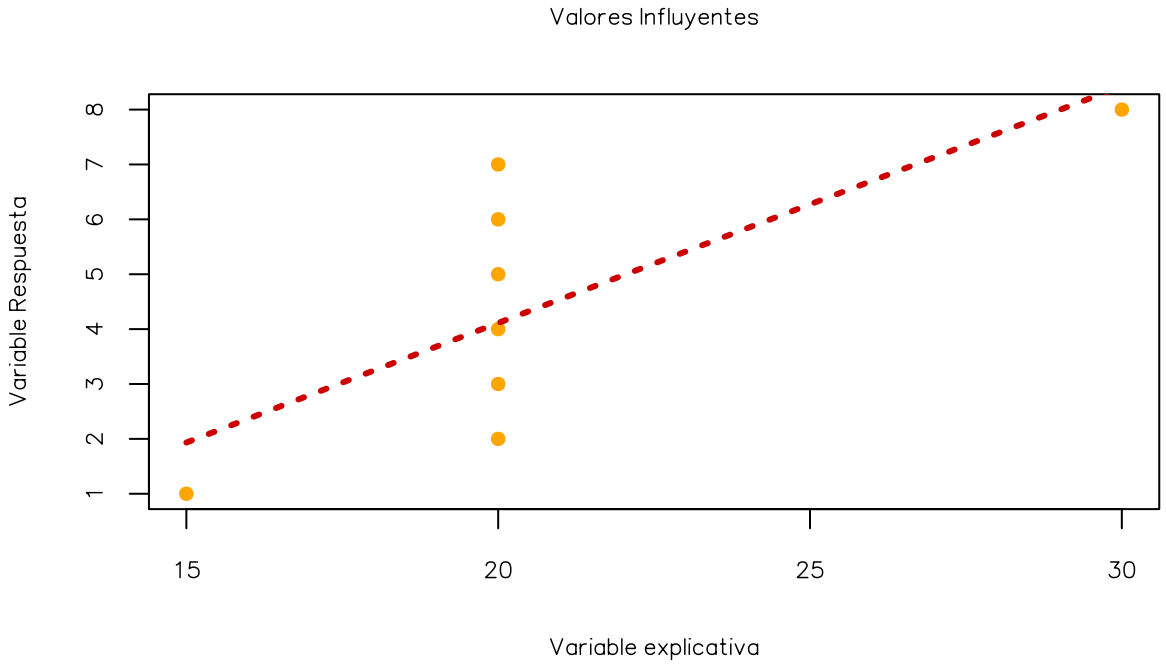
Generalmente se procede a eliminar observaciones atípicas, sin embargo, se recomienda realizar un análisis de influencia de las observaciones que presentan problemas de *outlier*, ya que aunque se trata de puntos atípicos, puede resultar beneficioso para el modelo ya que puede que sean significativas para el modelo, por lo que eliminar estas observaciones puede ocasionar conflictos o desviaciones en la estimación.

27.5.2. Valores influentes

Los valores influentes, son observaciones que tienen una gran influencia en el ajuste del modelo, es decir, remover estas observaciones ocasionaría un cambio drástico en el modelo de regresión, ya que dichas observaciones tienen gran influencia en el cálculo de los estimadores de los parámetros o en las predicciones.

Es por esta propiedad por lo que se busca analizar estos puntos para medir su impacto en el modelo, para identificar si el *outlier* encontrado puede ser eliminado o no de la muestra.

Por ejemplo, imagine que tiene una muestra aleatoria conformada por 8 observaciones, en la cual 6 elementos son iguales y dos con diferente valor, tal como se muestra a continuación:



El modelo tiene asociado una línea de regresión, sin embargo, si se quitan los extremos o valores atípicos, el modelo cambiaría rotundamente.

Un método para identificar la influencia del modelo es a través de los puntos palanca o leverage. El método consiste en examinar la medida entre el punto y el punto medio de los datos, a este punto también se le conoce como *centroide*. Para ello se observa cuales son las observaciones influyentes examinando la matriz H, o matriz sombrero, en el cual se pondrá especial atención a los elementos de la diagonal de la matriz H, se denotará como h_{ii} al i-ésimo elemento de la diagonal H, este último elemento se le denomina como el término de punto palanca o leverage.

Dado que el promedio de los valores leverage es $\frac{\sum_{i=1}^{n} h_{ii}}{n}$, entonces cuando un punto sea mayor que el doble de la media de los puntos palanca, es decir, cuando se cumpla que:

$$h_{ii} \geq 2\ \frac{\sum_{i=1}^{n} h_{ii}}{n}$$

Se puede concluir que dicha observación tiene un punto palanca muy grande. Por lo que se puede concluir que hay evidencia de que se trate de un punto influyente, sin embargo, para afirmar la anterior premisa es necesario el uso de otros métodos estadísticos, uno de ellos es la llamada distancia **Cook**.

La estadística de **Cook** propone calcular la distancia cuadrática entre el modelo ajustado y el modelo ajustado sin la $i-ésima$ observación. La cual puede expresarse como:

$$C_i = \frac{\left(\underline{\hat{Y}} - \underline{\hat{Y}}_{(i)}\right)\left(\underline{\hat{Y}} - \underline{\hat{Y}}_{(i)}\right)}{CM_{error}\sum_{i=1}^{n} h_{ii}} \qquad \forall\ i \in [1, n].$$

Donde $\underline{\hat{Y}}$ hace referencia al modelo ajustado de la forma $X\underline{\hat{\beta}}$ mientras que $\underline{\hat{Y}}_{(i)}$ hace referencia al modelo ajustado sin la i-ésima observación de forma $X\underline{\hat{\beta}}_{(i)}$.

De esta manera, se calcula la distancia de *Cook* para cada observación, con la finalidad de evaluar el cambio del modelo sin la i-ésima observación. Se considera que una observación es influyente si el cambio del modelo con y sin observación varía mucho entre si, aunque no hay una medida establecida, (Hair, Tatham,Anderson y Black, 1998) sugiere que si la distancia de *Cook* de la i-ésima observación es mayor o igual a 1 se tiene evidencia de que la observación analizada tiene gran influencia en el modelo, mientras que para otros autores como (Bollen y Jackman, 1985) mencionan que las distancias mayores que $\frac{4}{n}$ presenta indicios de influencia en el modelo.

Capítulo 28

Modelo de regresión lineal múltiple

Antes de empezar, el siguiente tema "Regresión Lineal Múltiple", queremos hacer referencia que fue obtenido de tesis: "Apoyo a la docencia" Omar (2019).

28.1. Introducción

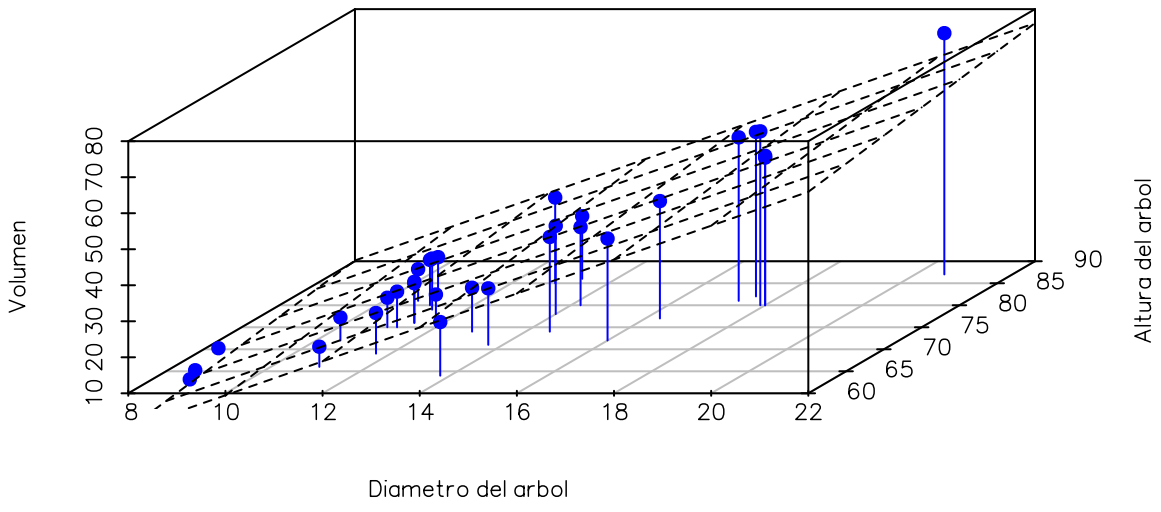
El modelo de regresión lineal simple ajusta una variable explicativa a una variable respuesta; Por su parte, el **Modelo de regresión lineal múltiple** busca hallar el mejor ajuste con dos o más variables regresoras. Es decir, la variable respuesta $\underline{Y}$ depende de k regresores de la forma:

$$\underline{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

En primera instancia no parece ser un gran cambio, sin embargo, es de gran importancia ya que de esta forma se puede estimar de una mejor manera un evento aleatorio, pues en general, un suceso no depende de sólo una acción o variable, sino que es resultado de una serie de diversos eventos o variables.

Es importante mencionar que en un modelo de regresión múltiple se deja de ajustar una línea recta a los datos, en cambio se ajusta un hiperplano.

```
data("trees")
library("scatterplot3d")
s3d <- scatterplot3d(trees, type = "h", color = "blue",
    angle=55, pch = 16, xlab = "Diametro del arbol",ylab = "Altura del arbol", zlab="Volumen")
my.lm <- lm(trees$Volume ~ trees$Girth + trees$Height)
s3d$plane3d(my.lm)
```



El "scatterplot" de arriba, se realizó con la base precargada en R, "$trees$", los datos que componen la muestra se encuentran en un vector de dimensión 3, la cual busca relacionar la variable $\underline{Y}$, con dos variables explicativas X_1 y X_2, en este caso, la variable $\underline{Y}$ hace referencia al volumen de un árbol y la variable X_1 hace referencia al diámetro del tronco del árbol y X_2 denota la altura del árbol, se observa que existe una tendencia, la cual es representada mediante el hiperplano de regresión marcado, en la cual a menor diámetro y menor altura, el volumen del árbol tiende a disminuir.

Debido a que se trabaja con cierto error ϵ en el ajuste de la regresión, es conveniente suponer que se cumplen lo siguientes supuestos:

Definición 3.1 (Supuestos del modelo de regresión múltiple)

El error ϵ_i en el modelo de regresión lineal múltiple cumple:

- $\mathbf{E}[\epsilon_i] = 0$.
- $Var(\epsilon_i) = \sigma^2$.
- $Cov(\epsilon_i, \epsilon_j) = 0, \quad i \neq j \ \ \forall \ \ i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, n.$

Al cumplirse estos supuestos es posible calcular la esperanza y varianza de la variable respuesta $\underline{Y}$ dado un conjunto de valores $x_1, x_2, \ldots, x_k$.

Teorema 3.1 Sea una variable de interés $\underline{Y}$, llamada **dependiente**, relacionada con dos o más variables explicativas o también llamadas regresoras $x_1, x_2, \ldots, x_k$, entonces:

a) $\mathbf{E}[\underline{Y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$.

b) $Var(\underline{Y}) = \sigma^2$.

Demostración:

a) Para la esperanza de $\underline{Y}$ se tiene:

$$\mathbf{E}[\underline{Y}] = \mathbf{E}[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon].$$

La estimación es sobre $\underline{Y}$, como $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son constantes; $x_1, x_2, \ldots, x_k$ son los valores dados, por lo que:

$$\mathbf{E}[\underline{Y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \mathbf{E}[\epsilon].$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + 0$$
$$\therefore \mathbf{E}[\underline{Y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k. \blacksquare$$

b) Para la varianza de $\underline{Y}$ se tiene:

$$Var(\underline{Y}) = Var\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon\right).$$

La estimación es sobre $\underline{Y}$, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son constantes; $x_1, x_2, \ldots, x_k$ son valores dados, por lo que cumple que:

$$Var(\underline{Y}) = 0 + 0 + 0 + \ldots + 0 + Var(\epsilon)$$
$$\therefore Var(\underline{Y}) = \sigma^2. \blacksquare$$

28.2. Modelo de regresión lineal múltiple

El objetivo del modelo de regresión lineal múltiple consiste en modelar $\underline{Y}$ a través de k variables regresoras en n observaciones independientes. Es decir, se tiene el siguiente modelo:

$$\begin{array}{ccccccccccccc} Y_1 & = & \beta_0 & + & \beta_1 x_{11} & + & \beta_2 x_{12} & + & \cdots & + & \beta_k x_{1k} & + & \epsilon_1 \\ Y_2 & = & \beta_0 & + & \beta_1 x_{21} & + & \beta_2 x_{22} & + & \cdots & + & \beta_k x_{2k} & + & \epsilon_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ Y_n & = & \beta_0 & + & \beta_1 x_{n1} & + & \beta_2 x_{n2} & + & \cdots & + & \beta_k x_{nk} & + & \epsilon_n \end{array}$$

El anterior conjunto de igualdades puede ser denotado matricialmente mediante la siguiente igualdad:

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon}$$

donde:

$$\underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

$$\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Sustituyendo en la ecuación anterior, se observa que el modelo de regresión multiple puede ser visto como:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

La dimensión de las matrices señaladas anteriormente se mencionan en el siguiente recuadro:

Matriz	*Dimensión*
$\underline{Y}$	$n \times 1$
X	$n \times (k+1)$
$\underline{\epsilon}$	$n \times 1$
$\underline{\beta}$	$(k+1) \times 1$

Finalmente para definir correctamente el modelo es necesario realizar las siguientes suposiciones acerca de las matrices del modelo de regresión lineal múltiple.

Definición 3.2 Sea X la denominada *matriz diseño* entonces satisface que:

- $X_{n\times(k+1)}$ es el rango completo en la columna, es decir, X es de rango $k+1$

Éste supuesto es importante ya que satisface que $k+1 \leq n$, es decir, **el máximo número de variables con el que se ajusta el modelo no puede ser superior al número de observaciones.**

De igual forma, observe que el supuesto de la varianza de los errores en la definición 3.1, puede reescribirse en forma matricial:

$$Var(\epsilon) = \begin{pmatrix} Var(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & Cov(\epsilon_1, \epsilon_3) & \ldots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_2, \epsilon_1) & Var(\epsilon_2) & Cov(\epsilon_2, \epsilon_3) & \ldots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(\epsilon_n, \epsilon_1) & Cov(\epsilon_n, \epsilon_2) & Cov(\epsilon_n, \epsilon_3) & \ldots & Var(\epsilon_n) \end{pmatrix}$$

Por **definición 3.1**, $Cov(\epsilon_i, \epsilon_j) = 0 \quad i \neq j$

$$Var(\epsilon) = \begin{pmatrix} Var(\epsilon_1) & 0 & 0 & \dots & 0 \\ 0 & Var(\epsilon_2) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & Var(\epsilon_n) \end{pmatrix}$$

Por **definición 3.1**, $Var(\epsilon_i) = \sigma^2$

$$Var(\epsilon) = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

$\therefore Var(\epsilon) = \sigma^2 I_{n\times n}$.■

28.3. Estimación por mínimos cuadrados de los parámetros del modelo

Es necesario dar una estimación de la intersección con el eje $\underline{Y}$, las variables que conforman el hiperplano, es decir, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ respectivamente. La manera en la que se construyen a los estimadores es tal que la diferencia entre todos los valores observados y los valores estimados sea 0, es decir, a éstas diferencias se le conoce como **residuales**, muchos autores también hacen referencia a ellos como **residuos**.

Definición 3.3 (Residuales). Sea y_i los valores observados, y sea $\hat{y}_i$ los valores ajustados de la forma $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki}$ para $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ dados, entonces:

$$e_i = y_i - \hat{y}_i \qquad i = 1, \dots, n.$$

Se les conoce como **residuales.**

Para estimar los valores desconocidos $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ se usa el **método de mínimos cuadrados**, el cual es similar al caso de regresión lineal simple, dicho método propone minimizar la suma de cuadrados de los residuales.

Antes de continuar es necesario ver algunos resultados importantes de equivalencia y notación.

De la **definición 3.3** se sabe que los valores esperados de y_i pueden ser definidos como:

$$\begin{array}{ccccccccccc} \hat{y_1} & = & \hat{\beta}_0 & + & \hat{\beta}_1 x_{11} & + & \hat{\beta}_2 x_{12} & + & \dots & + & \hat{\beta}_k x_{1k} \\ \hat{y_2} & = & \hat{\beta}_0 & + & \hat{\beta}_1 x_{21} & + & \hat{\beta}_2 x_{22} & + & \dots & + & \hat{\beta}_k x_{2k} \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ \hat{y_n} & = & \hat{\beta}_0 & + & \hat{\beta}_1 x_{n1} & + & \hat{\beta}_2 x_{n2} & + & \dots & + & \hat{\beta}_k x_{nk} \end{array}$$

La anterior ecuación puede ser descompuesta en forma matricial de la siguiente manera:

$$\begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_n} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

Por lo tanto, podemos renombrar a las matrices de acuerdo a los elementos que las conforman.

Se tiene la siguiente igualdad para los valores estimados $\underline{\hat{Y}}$.

$$\underline{\hat{Y}} = X\underline{\hat{\beta}}.$$

Ahora por la definición 3.3 y lo anterior tenemos que los residuales se encuentran de la forma:

$$\underline{e} = \underline{Y} - X\underline{\hat{\beta}}.$$

Teorema 3.2 (Mínimos Cuadrados).(MC) Si se minimiza la suma de cuadrados de la diferencia entre los valores observados y los estimados, la cual se expresa matricialmente de la siguiente forma:

$$\underline{e'e}.$$

Entonces se tiene como estimador de $\underline{\beta}$ a:

$$\underline{\hat{\beta}} = (X'X)^{-1} X'\underline{Y}.$$

Demostración:

Se sabe que los residuales están definidos como $\underline{e} = \left(\underline{Y} - X\underline{\hat{\beta}}\right)$ de esta manera, por hipótesis se tiene:

$$\begin{aligned}
\underline{e'e} &= \left(\underline{Y} - X\underline{\hat{\beta}}\right)' \left(\underline{Y} - X\underline{\hat{\beta}}\right) \\
&= \left(\underline{Y}' - \underline{\hat{\beta}}' X'\right) \left(\underline{Y} - X\underline{\hat{\beta}}\right) \\
&= \underline{Y}'Y - \underline{Y}'X\underline{\hat{\beta}} - \underline{\hat{\beta}}' X'\underline{Y} + \underline{\hat{\beta}}' X'X\underline{\hat{\beta}} \\
\underline{e'e} &= \underline{Y}'Y - 2\underline{Y}'X\underline{\hat{\beta}} + \underline{\hat{\beta}}' X'X\underline{\hat{\beta}}.
\end{aligned}$$

Lo anterior se da ya que $\underline{\hat{\beta}}' X'\underline{Y}$ es una matriz de 1×1, es decir, un escalar, y que su transpuesta $(\underline{\hat{\beta}}' X'\underline{Y})' = \underline{Y}'X\underline{\hat{\beta}}$ es el mismo escalar.

Derivando respecto a $\underline{\hat{\beta}}$ para hallar los posibles mínimos se divide la suma matricial de la siguiente forma:

$$\Delta\underline{\hat{\beta}} = \Delta_1\underline{\hat{\beta}} + \Delta_2\underline{\hat{\beta}} + \Delta_3\underline{\hat{\beta}}$$

Procederemos a derivar:

$$\Delta\underline{\hat{\beta}} = \begin{cases} \Delta_1\underline{\hat{\beta}} & \left(\underline{Y}'\underline{Y}\right) & = 0 \\ \Delta_2\underline{\hat{\beta}} & \left(-2\underline{Y}'X\underline{\hat{\beta}}\right) & = \left(-2\underline{Y}'X\right)' = -2X'\underline{Y} \\ \Delta_3\underline{\hat{\beta}} & \left(\underline{\hat{\beta}}' X'X\underline{\hat{\beta}}\right) & = 2X'X\underline{\hat{\beta}} \end{cases}$$

De esta forma se tiene que la derivada respecto a $\underline{\beta}$ es:

$$\Delta\underline{\beta} = -2X'\underline{Y} + 2X'X\underline{\hat{\beta}}.$$

Igualamos la derivada a 0, para hallar un punto crítico:

$$\begin{aligned}
\Delta\underline{\hat{\beta}} &= 0 \\
-2X'\underline{Y} + 2X'X\underline{\hat{\beta}} &= 0 \\
2X'X\underline{\hat{\beta}} &= 2X'\underline{Y}
\end{aligned}$$

Esto se simplifica a:

$$X'X\underline{\hat{\beta}} = X'\underline{Y}$$

Ahora multiplicamos ambos lados por la inversa de $X'X$, es decir, $(X'X)^{-1}$

$$\therefore \underline{\hat{\beta}} = (X'X)^{-1} X'\underline{Y}.$$

Nótese que la inversa $(X'X)^{-1}$ existe porque X es de rango completo en las columnas, como se mencionó en la definición. Por lo que el producto matricial $X'X$ es de rango completo $(k+1)$, es decir $|X| \neq 0$, por lo que se garantiza la existencia de la inversa.

Realizando la segunda derivada, veremos si la función es cóncava o convexa para saber si es mínimo o máximo.

$$\Delta\Delta\underline{\hat{\beta}} = \begin{cases} \Delta\Delta\underline{\hat{\beta}} & -2X'\underline{Y} = 0 \\ \Delta\Delta\underline{\hat{\beta}} & 2X'X\underline{\hat{\beta}} = (2X'X)' = 2X'X \end{cases}$$

Así $\Delta\Delta\underline{\hat{\beta}} = 2X'X$

De esta manera $\Delta\Delta\underline{\hat{\beta}} > 0$ ya que $X'X$ es definida positiva , por consiguiente $(X'X)^{-1}X'\underline{Y}$ es considerado un mínimo. Por lo tanto, $\underline{\hat{\beta}} = (X'X)^{-1} X'\underline{Y}$. es el estimador de mínimos cuadrados del modelo de regresión múltiple.■

Una vez encontrada una estimación a los parámetros desconocidos de $\underline{\beta}$, será conveniente desarrollar algunas variantes en la forma en la que se denota a los residuales, para ello se define a la matriz H como $H = X(X'X)^{-1}X'$. Cabe destacar que la matriz H es conocida como **"matriz sombrero"**, que junto con la matriz $(I - H)$ cumplen con ser matrices idempotentes, es decir, que al elevar las matrices a una potencia dada los valores contenidos en la matriz no se modifican; de igual forma ambas matrices cumplen con ser simétricas, denominadas así ya que al transponer las matrices los valores contenidos en ellas conservan su lugar.

Debemos considerar el siguiente resultado, el cual será importante al desarrollar el siguiente teorema 3.3 ya que demuestra que $(X'X)^{-1}$ es una matriz simétrica.

$$\begin{aligned} [(X'X)^{-1}]' &= [(X'X)']^{-1} \\ &= (X'(X')')^{-1} \\ \therefore [(X'X)^{-1}]' &= (X'X)^{-1}.\blacksquare \end{aligned}$$

Es decir, la inversa de $X'X$ es simétrica, resultado importante en el siguiente teorema:

Teorema 3.3 Sea $H = X(X'X)^{-1}X'$ e $(I - H)$ entonces:

a) Las matrices H e $I - H$ son idempotentes.

b) Las matrices H e $I - H$ son simétricas.

Demostración:

a) Para demostrar la idempotencia de H basta probar que $H^2 = H$, es decir, al elevar la matriz H ésta no se alterará:

$$\begin{aligned} H^2 &= (X(X'X)^{-1}X')(X(X'X)^{-1}X') \\ &= X(X'X)^{-1}X'X(X'X)^{-1}X'. \end{aligned}$$

Transponiendo con la finalidad de simplificar el producto matricial y por el resultado mostrado anteriormente $[(X'X)^{-1}]' = (X'X)^{-1}$ se tiene:

$$\begin{aligned} &= [(X'X)^{-1}X'X(X'X)^{-1}X']'X' \\ &= [(X'X)^{-1}X']'X' \\ &= X(X'X)^{-1}X' \end{aligned}$$

$$\therefore H^2 = H.$$

Por lo tanto H es idempotente. ■

Para probar la idempotencia de $I - H$, ésta será elevada al cuadrado.

$$\begin{aligned}(I-H)^2 &= (I-H)(I-H)\\ &= I - IH - IH + H^2\\ &= I - 2H + H^2.\end{aligned}$$

Por idempotencia de H, $H = H^2$. Por lo tanto:

$$(I-H) = I - 2H + H$$

$$\therefore (I-H)^2 = I - H.$$

b) Para demostrar la simetría de H, se transpondrá la matriz H. Además debemos recordar que $[(X'X)^{-1}]' = (X'X)^{-1}$ así:

$$\begin{aligned}H' &= (X(X'X)^{-1}X')'\\ &= X(X'X)^{-1}X'\\ \therefore H' &= H.\end{aligned}$$

Por lo tanto la matriz H es simétrica.

Para la simetría de $I - H$ se transpone la matriz:

$$(I-H)' = I' - H'.$$

Por simetría de H y de I

$$\therefore (I-H)^2 = I - H$$

Por lo tanto $I - H$ es simétrica. ■

Corolario 4 Sea $\underline{e}$ la matriz de residuales, entonces éstos pueden ser expresados por la siguiente ecuación:

$$\underline{e} = (I-H)\underline{Y}$$

donde I es la matriz identidad, y $H = X(X'X)^{-1}X'$.

Demostración:

Se sabe que los valores estimados son calculados de la siguiente manera:

$$\begin{aligned}\underline{\hat{Y}} &= X\underline{\hat{\beta}}\\ \underline{\hat{Y}} &= X(X'X)^{-1}X'\underline{Y}\\ \underline{\hat{Y}} &= H\underline{Y}.\end{aligned}$$

donde $H = X(X'X)^{-1}X'$. De esta manera calculando la matriz de residuales se tiene:

$$\begin{aligned}\underline{e} &= \underline{Y} - \underline{\hat{Y}}\\ \underline{e} &= \underline{Y} - X\underline{\hat{\beta}}\\ \underline{e} &= \underline{Y} - H\underline{Y}\end{aligned}$$

$$\underline{e} = (I - H)\underline{Y}.\blacksquare$$

Como se mencionó en regresión lineal simple, SC_{error} mide la variación residual que queda sin explicar por la línea de regresión, en el modelo de regresión múltiple es denotada como $SC_{error} = \underline{e}'\underline{e}$, la cual es equivalente a la suma de residuales al cuadrado.

Corolario 5 La suma de cuadrados del error, puede denotarse matricialmente como:

$$SC_{error} = \underline{Y}'(I - H)\underline{Y}.$$

donde:

- "$\underline{Y}$" son los valores observados de la variable respuesta.
- I es la matriz identidad.
- $H = X(X'X)^{-1}X'$.

Demostración:

Se sabe por hipótesis que:

$SC_{error} = \underline{e}'\underline{e}$

Por el **corolario 4**, se puede expresar a los residuales como $\underline{e} = (I - H)\underline{Y}$, sustituyendo:

$$SC_{error} = ((I - H)\underline{Y})'((I - H)\underline{Y})$$

$$= (\underline{Y}'(I - H)')((I - H)\underline{Y})$$

$$= (\underline{Y}'(I - H))((I - H)\underline{Y})$$

$$= \underline{Y}'(I - H)(I - H)\underline{Y}$$

$$= [(I - H)'(I - H)'\underline{Y}]\underline{Y}$$

$$= [(I - H)^2\underline{Y}]'\underline{Y}$$

$$= \underline{Y}'(I - H)'\underline{Y}$$

$$\therefore SC_{error} = \underline{Y}'(I - H)'\underline{Y}.$$

Con los resultados, se procede a examinar las propiedades de los estimadores obtenidos por el método de mínimos cuadrados. Éstas propiedades son agrupadas y enunciadas en el **Teorema de Gauss-Markov**

Teorema 3.4 (Teorema de Gauss-Markov).

En el modelo de **regresión lineal múltiple** $\underline{Y} = X\underline{\beta} + \underline{\epsilon}$, bajo la hipótesis:

- $\mathbf{E}[\underline{\epsilon}] = 0$ y $Var(\underline{\epsilon}) = \sigma^2 I_n$.
- $\mathbf{E}[\underline{Y}] = X\underline{\beta}$ y $Var(\underline{Y}) = \sigma^2 I_n$.
- X de rango completo en las columnas.

El estimador de mínimos cuadrados de $\underline{\beta}$, es el **MELI** (**BLUE** por su abreviación en inglés), el mejor estimador lineal insesgado. Es decir, $\underline{\hat{\beta}}$ es insesgado y además, si $\underline{\tilde{\beta}}$ es otro estimador insesgado, entonces $Var(\underline{\tilde{\beta}}) \geq Var(\underline{\hat{\beta}})$, es decir, $\underline{\hat{\beta}}$ es de mínima varianza.

Demostración:

Para demostrar que el estimador $\underline{\hat{\beta}}$ es insesgado, es necesario probar que el estimador cumple que $\mathbf{E}[\underline{\hat{\beta}}] = \underline{\beta}$, de ésta forma:

$$\mathbf{E}[\underline{\hat{\beta}}] = \mathbf{E}[(X'X)^{-1}X'\underline{Y}]$$

Ya que X son constantes

$$= (X'X)^{-1}X'\mathbf{E}[\underline{Y}]$$

Ya que $\mathbf{E}[\underline{Y}] = X\underline{\beta}$

$$= (X'X)^{-1}X'X\beta$$
$$= I\beta$$

$$\therefore \mathbf{E}[\underline{\hat{\beta}}] = \underline{\beta}.$$

Por lo tanto $\underline{\hat{\beta}}$ es un estimador insesgado para $\underline{\beta}$.■

Para conocer la varianza del estimador $\underline{\hat{\beta}}$ se sabe que:

$$Var(\underline{\hat{\beta}}) = Var\left((X'X)^{-1}X'\underline{Y}\right)$$
$$= (X'X)^{-1}X'Var(\underline{Y})[(X'X)^{-1}X']'$$

Ya que la $Var(\underline{Y}) = \sigma^2 I_n$

$$= (X'X)^{-1}X'[(X'X)^{-1}X']'\sigma^2$$
$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$
$$= \sigma^2 I(X'X)^{-1}$$
$$\therefore Var(\underline{\hat{\beta}}) = \sigma^2(X'X)^{-1}$$

Para comprobar que el estimador $\underline{\hat{\beta}}$ es el estimador insesgado de mínima varianza, se propone a un estimador $\underline{\tilde{\beta}}$ el cual cumple con ser lineal e insesgado. Para ello sea $\underline{\tilde{\beta}}$ un estimador linealmente insesgado para $\underline{\beta}$. Es decir, existe una matriz de $A_{(k+1)\times n}$ tal que $\underline{\tilde{\beta}} = A\underline{Y}$. De esta forma:

$$\mathbf{E}\left[\underline{\tilde{\beta}}\right] = \mathbf{E}\left[A\underline{Y}\right]$$
$$= A\mathbf{E}[\underline{Y}]$$

Ya que $\mathbf{E}[\underline{Y}] = X\underline{\beta}$

$$= AX\underline{\beta}$$
$$\mathbf{E}\left[\underline{\tilde{\beta}}\right] = AX\underline{\beta}$$

Para que sea un estimador insesgado, entonces AX tiene que cumplir: $AX = I$, así:

$$\mathbf{E}\left[\underline{\tilde{\beta}}\right] = I\underline{\beta}$$

$$\therefore \mathbf{E}\left[\tilde{\underline{\beta}}\right] = \underline{\beta}.$$

Para conocer la varianza de $\tilde{\underline{\beta}}$ se tiene:

$$Var\left(\tilde{\underline{\beta}}\right) = Var(A\underline{Y})$$
$$Var\left(\tilde{\underline{\beta}}\right) = AVar(\underline{Y})A'$$
$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2 AA'.$$

Sea C una matriz de dimensión $(k+1)\times n$ tal que $C = A - (X'X)^{-1}X'$. Observe que $CX = 0$ ya que $CX = AX - (X'X)^{-1}X'X = I - I = 0$. De esta forma se tiene la siguiente igualdad:

$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2((X'X)^{-1}X' + C)((X'X)^{-1}X' + C)'$$
$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2((X'X)^{-1}X' + C)\left[((X'X)^{-1}X')' + C'\right]$$

$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2((X'X)^{-1}X' + C)[X(X'X)^{-1} + C']$$
$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2[(X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'C' + CX(X'X)^{-1} + CC']$$
$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2\left[(X'X)^{-1}X'X(X'X)^{-1} + [CX(X'X)^{-1}]' + CX(X'X)^{-1} + CC'\right]$$

Debido a que $CX = 0$

$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2\left[I(X'X)^{-1} + 0 + 0 + CC'\right]$$
$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2[(X'X)^{-1} + CC']$$
$$Var\left(\tilde{\underline{\beta}}\right) = \sigma^2(X'X)^{-1} + \sigma^2 CC'.$$

$$\therefore Var\left(\tilde{\underline{\beta}}\right) = Var\left(\hat{\underline{\beta}}\right) + \sigma^2 CC'$$

Además se observa que CC', es una matriz semidefinida positiva ya que los valores propios de CC' son reales y no negativos, además debido al supuesto de que X es de rango completo para las columnas se cumple que $rango(CC') = rango(X) = k+1$, esto es importante, ya que si no se cumple se tendría una solución no trivial, por lo que 0 podría ser una solución para un egeinvalor por lo que no sería semidefinido positivo. Como $CC' \geq 0$, se observa que:

$$Var\left(\tilde{\underline{\beta}}\right) \geq Var\left(\hat{\underline{\beta}}\right).$$

Por lo que la varianza del estimador propuesto es mayor al obtenido por mínimos cuadrados. Por lo tanto, el estimador de MC de $\underline{\beta}$ es el mejor estimador linealmente insesgado y de mínima varianza.■

Las anteriores propiedades de los estimadores son importantes ya que garantizan que los valores estimados $\hat{\underline{Y}}$, asignan valores que efectivamente recaen en el hiperplano propuesto en el modelo de regresión lineal múltiple.

Teorema 3.5 Sea $\hat{\underline{Y}}$, los valores estimados de Y, de forma que $\hat{\underline{Y}} = X\hat{\underline{\beta}}$, entonces se cumple:

a) $\mathbf{E}[\hat{\underline{Y}}] = X\underline{\beta}$.

b) $Var(\hat{\underline{Y}}) = \sigma^2 H$.

Demostración:

a) Para demostrar la esperanza de los valores estimados, se observa que:

$$\mathbf{E}[\hat{\underline{Y}}] = \mathbf{E}[X\hat{\underline{\beta}}]$$

$$\mathbf{E}[\underline{\hat{Y}}] = X\mathbf{E}[\underline{\hat{\beta}}]$$

$$\therefore \mathbf{E}[\underline{\hat{Y}}] = X\underline{\beta}.\blacksquare$$

b) Para la varianza se tiene:

$$Var(\underline{\hat{Y}}) = Var(X\underline{\hat{\beta}})$$
$$Var(\underline{\hat{Y}}) = XVar(\underline{\hat{\beta}})X'$$
$$Var(\underline{\hat{Y}}) = X\sigma^2(X'X)^{-1}X'$$

$$Var(\underline{\hat{Y}}) = \sigma^2X(X'X)^{-1}X'$$

$$\therefore Var(\underline{\hat{Y}}) = \sigma^2H.\blacksquare$$

Teorema 3.6 Sea $\underline{e}$ los residuales del modelo, de forma $\underline{e} = \underline{Y} - \underline{\hat{Y}}$, entonces cumplen con:

a) $\mathbf{E}[\underline{e}] = 0$.

b) $Var(\underline{e}) = \sigma^2(I - H)$.

Demostración:

$**a)$ Para demostar la esperanza de los residuales, se observa que:

$$\mathbf{E}[\underline{e}] = \mathbf{E}[\underline{Y} - \underline{\hat{Y}}]$$

Por el **corolario 4**

$$\mathbf{E}[\underline{e}] = \mathbf{E}[(I - H)\underline{Y}]$$
$$\mathbf{E}[\underline{e}] = (I - H)\mathbf{E}[\underline{Y}]$$

Por el **teorema 3.4**

$$\mathbf{E}[\underline{e}] = (I - H)X\beta$$

$$\mathbf{E}[\underline{e}] = X\underline{\beta} - HX\underline{\beta}$$
$$\mathbf{E}[\underline{e}] = X\underline{\beta} - X(X'X)^{-1}X'X\underline{\beta}$$
$$\mathbf{E}[\underline{e}] = X\underline{\beta} - \left[X'X(X'X)^{-1}X'\right]'\underline{\beta}$$

$$\mathbf{E}[\underline{e}] = X\underline{\beta} - [IX']'\underline{\beta}$$

$$\mathbf{E}[\underline{e}] = X\underline{\beta} - X\underline{\beta}$$
$$\therefore \mathbf{E}[\underline{e}] = 0.\blacksquare$$

$**b)$ Para la varianza se tiene:

$$Var(\underline{e}) = Var\left(\underline{Y} - \underline{\hat{Y}}\right)$$

Por el **corolario 4**

$$Var(\underline{e}) = Var\left((I-H)\underline{\hat{Y}}\right)$$

$$Var(\underline{e}) = (I-H)Var(\underline{\hat{Y}})(I-H)'$$

Por el **teorema 3.3**

$$Var(\underline{e}) = (I-H)\sigma^2(I-H)$$

Por idempotencia de $I-H$

$$Var(\underline{e}) = \sigma^2(I-H)(I-H)$$

$$\therefore Var(\underline{e}) = \sigma^2(I-H).\blacksquare$$

28.4. Estimación por máxima verosimilitud

Se han usado varios supuestos para poder calcular a los estimadores por medio del método de mínimos cuadrados, sin embargo, para hacer uso de la estimación por máxima verosimilitud se supondrá que los errores se distribuyen como una normal multivariada $\underline{\epsilon} \sim \mathbf{N}_n(O_n, \sigma^2 I_n)$ por lo que el modelo $\underline{Y}$ tiene distribución normal, es decir, $\underline{Y} \sim \mathbf{N}_n(X\underline{\beta}, \sigma^2)$.

Tenemos:

Definición 3.4 Tomando el supuesto de normalidad conjunta para los errores se cumple que:

- ϵ_i es independiente $\forall i$ tal que $i \neq j$.
- $\underline{Y} \sim \mathbf{N}_n(X\underline{\beta}, \sigma^2)$.
- Cada y_i es independiente pero no es idénticamente distribuida.

De esta forma se podrá usar el método de máxima verosimilitud para estimar a los parámetros desconocidos a través de la función de verosimilitud. Al realizar la estimación por éste método se obtendrán resultados parecidos a los obtenidos por mínimos cuadrados.

Teorema 3.7 (Función de verosimilitud).

Sea $\underline{\hat{\beta}}$ y $\hat{\sigma}^2$, los estimadores de $\underline{\beta}$ y σ^2 respectivamente; suponiendo normalidad en los errores $\underline{\epsilon} \sim \mathbf{N}_n(0_n, \sigma^2 I_n)$ y $\underline{Y} \sim \mathbf{N}_n(X\underline{\beta}, \sigma^2)$ entonces la estimación de los parámetros $\underline{\beta}$ y σ^2 por el método de máxima verosimilitud están dados por:

a) $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{Y}$.

b) $\hat{\sigma}^2 = \frac{1}{n}\left(\underline{Y} - X\underline{\hat{\beta}}\right)'\left(\underline{Y} - X\underline{\hat{\beta}}\right)$.

Demostración:

Por hipótesis $\underline{Y} \sim \mathbf{N}_n(X\underline{\beta}, \sigma^2 I_n)$, escribiendo la función de verosimilitud se tiene que:

$$L(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2} exp\left[-\frac{1}{2\sigma^2}(\underline{Y}-\mu)'(\underline{Y}-\mu)\right]$$

$$L(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = (2\pi\sigma^2)^{\sum_{i=1}^{n} -1/2} exp\left[-\frac{1}{2\sigma^2}(\underline{Y}-X\underline{\beta})'(\underline{Y}-X\underline{\beta})\right]$$

$$L(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}(\underline{Y}-X\underline{\beta})'(\underline{Y}-X\underline{\beta})\right].$$

Aplicando logaritmo natural a la función de verosimilitud:

$$lnL(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta})$$

$$lnL(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = -\frac{n}{2}[ln(2\pi) + ln(\sigma^2)] - \frac{1}{2\sigma^2}(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta})$$

$$lnL(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta})$$

a) Derivando respecto a $\underline{\beta}$ para obtener su estimador.

$$\frac{\partial}{\partial \underline{\beta}} lnL(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = -\frac{1}{2\sigma^2} 2(X'X\underline{\beta} - X'\underline{Y}).$$

$$\frac{\partial}{\partial \underline{\beta}} lnL(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = -\frac{1}{\sigma^2}(X'X\underline{\beta} - X'\underline{Y}).$$

Igualando la derivada a 0, para encontrar el punto silla

$$\frac{\partial}{\partial \underline{\beta}} lnL(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = 0$$

$$-\frac{1}{\sigma^2}(X'X\underline{\beta} - X'\underline{Y}) = 0$$

$$(X'X\underline{\beta} - X'\underline{Y}) = 0$$

$$X'X\underline{\beta} = X'\underline{Y}$$

$$\therefore \underline{\hat{\beta}} = (X'X)^{-1}X'\underline{Y}.\blacksquare$$

Ya que X es de rango completo por columnas entonces existe $(X'X)^{-1}$.

b) Derivando respecto a σ^2 para obtener su estimador:

$$\frac{\partial}{\partial \sigma^2} lnL(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta}).$$

Igualando la derivada parcial a 0 para hallar el punto crítico de un posible máximo.

$$\frac{\partial}{\partial \sigma^2} lnL(\underline{\beta}, \sigma^2 \mid \underline{Y}, X) = 0$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta}) = 0$$

$$\frac{1}{2\sigma^4}(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta}) = \frac{n}{2\sigma^2}$$

$$(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta}) = n\sigma^2$$

$$\therefore \hat{\sigma}^2 = \frac{1}{n}\left(\underline{Y} - X\underline{\hat{\beta}}\right)'\left(\underline{Y} - X\underline{\hat{\beta}}\right).\blacksquare$$

Por lo tanto los estimadores de máxima verosimilitud son:

$$\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{Y}$$

$$y$$

$$\hat{\sigma}^2 = \tfrac{1}{n}\left(\underline{Y} - X\underline{\hat{\beta}}\right)'\left(\underline{Y} - X\underline{\hat{\beta}}\right)\blacksquare$$

Se observa que el estimador de $\underline{\beta}$ obtenido por el método de mínimos cuadrados es similar al estimador por máxima verosimilitud, sin embargo, éste último aporta mayor información al proporcionar el estimador para la varianza del modelo $\hat{\sigma}^2$.

De igual forma el estimador $\hat{\sigma}^2_{MV}$ guarda cierta relación con la suma de cuadrados residuales ya que se tiene:

$$\hat{\sigma}^2_{MV} = \frac{1}{n}(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta})$$

Por el **corolario 4**

$$\hat{\sigma}^2_{MV} = \frac{1}{n} SC_{error}$$

Por el **corolario 5**

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\underline{e}'\ \underline{e}$$

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\underline{Y}'(I - H)\underline{Y}.$$

Cabe destacar, que el estimador de $\underline{\beta}$ por máxima verosimilitud hereda todas las propiedades que cumple el estimador de mínimos cuadrados del teorema 3.4, es decir, $\underline{\hat{\beta}}$ es insesgado y de mínima varianza, sin embargo, el método de máxima verosimilitud proporciona una estimación para σ^2 el cual cumple con tener sesgo ($\mathbf{E}[\sigma^2] \neq 0$).

$$\mathbf{E}[\hat{\sigma}^2] = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\underline{Y_i} - \underline{\hat{Y}})'(\underline{Y_i} - \underline{\hat{Y}})\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[(\underline{Y_i} + X\underline{\beta} - X\underline{\beta} - X\underline{\hat{\beta}})'(\underline{Y_i} + X\underline{\beta} - X\underline{\beta} - X\underline{\hat{\beta}})\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{E}[(\underline{Y_i} - X\underline{\beta})(\underline{Y_i} - X\underline{\beta})'] - \mathbf{E}\left[(X\underline{\hat{\beta}} - X\underline{\beta})(X\underline{\hat{\beta}} - X\underline{\beta})'\right]\right)$$

$$= \sigma^2 - \frac{1}{n^2}\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\underline{Y_i} - \underline{\hat{Y}})(\underline{Y_i} - \underline{\hat{Y}})' - \sum_{i=1}^{n}(\underline{Y_i} - \underline{\hat{Y}})(\underline{Y_i} - \underline{\hat{Y}})'\right]$$

$$= \sigma^2 - \frac{n^2 - nk - n}{n}\sigma^2$$

$$\therefore \mathbf{E}[\hat{\sigma}^2] = \frac{n - k - 1}{n}\sigma^2.$$

Por lo tanto el estimador $\hat{\sigma}^2$ no es insesgado.■

El método de mínimos cuadrados no proporciona información acerca de la estimación de varianza del modelo σ^2, es por ello que se propone al estimador:

$$\sigma^2_{MC} = \frac{SC_{error}}{n - k - 1}.$$

Capítulo 29

Intervalos de confianza

En las secciones anteriores se obtuvieron, de manera puntual, las estimaciones de los parámetros desconocidos del modelo de regresión lineal múltiple. Sin embargo, a continuación se mostrará que se pueden realizar intervalos con un nivel de confianza α en donde los parámetros desconocidos tengan una alta probabilidad de pertenecer a este conjunto.

29.1. Intervalo para β_j

Para la construcción del intervalo de confianza para el parámetro desconocido β_j, con $j = 0, 1, \ldots, n$ se mantiene la hipótesis de que los errores se distribuyen como variables aleatorias normales con media cero y varianza $I_n\sigma^2$, como consecuencia $\underline{Y}$ se distribuye de forma normal, con media $X\underline{\beta}$ y varianza $\sigma^2 I_n$, es decir:

$$\underline{Y} \sim \mathbf{N}(X\underline{\beta}, \sigma^2 I_n)$$

De acuerdo con (Montgomery, Peck y Vining, 2012), el estimador $\underline{\hat{\beta}}$ por mínimos cuadrados es una combinación lineal de las observaciones, el cual también se distribuye con normalidad, con el vector medio $\underline{\beta}$ y varianza $\sigma^2(X'X)^{-1}$, lo cual vimos en el teorema 3.4, al igual que pudimos demostrar que el estimador es insesgado $\mathbf{E}[\underline{\hat{\beta}}] = \underline{\beta}$ y varianza $Var(\underline{\beta}) = \sigma^2(X'X)^{-1}$ por lo que de manera conjunta $\underline{\hat{\beta}}$ se distribuye con normalidad de la forma:

$$\underline{\hat{\beta}} \sim \mathbf{N}(\underline{\beta}, \sigma^2(X'X)^{-1})$$

Esto implica que la distribución marginal de cualquier coeficiente de la regresión $\hat{\beta}_j$ asimismo se distribuye normal, con media β_j y varianza $\sigma^2 C_{(j+1)(j+1)}$ donde $C_{(j+1)(j+1)}$ es el j-ésimo elemento de la diagonal de la matriz $(X'X)^{-1}$, es decir, $\hat{\beta}_j$ tiene una distribución normal asociada de la forma:

$$\hat{\beta}_j \sim \mathbf{N}(\beta_j, \sigma^2 C_{(j+1)(j+1)}$$

Normalizando se tiene que:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 C_{(j+1)(j+1)}}} \sim \mathbf{N}(0, 1)$$

Como $\frac{(n-k-1)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-k-1}$, tenemos:

$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 C_{(j+1)(j+1)}}}}{\sqrt{\frac{\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2}}{n-k-1}}} \sim t_{n-k-1}$$

De esta forma se obtiene la cantidad pivotal para β_j sl simplificar la ecuación anterior

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}}} \sim t_{n-k-1}$$

donde $\hat{\sigma}^2 = \frac{SC_{error}}{n-k-1}$. Por lo tanto el intervalo de confianza $1-\alpha$ para β_j es:

$$\mathbf{P}\left[-t_{n-k-1}^{\alpha/2} < \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}}} < t_{n-k-1}^{\alpha/2}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}} < \hat{\beta}_j - \beta_j < t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}}\ \right] = 1-\alpha$$

$$\mathbf{P}\left[-t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}} < \beta_j - \hat{\beta}_j < t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}}\ \right] = 1-\alpha$$

$$\mathbf{P}\left[\hat{\beta}_j - t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}} < \beta_j < \hat{\beta}_j + t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}}\ \right] = 1-\alpha$$

Por lo tanto el intervalo de confianza $1-\alpha$ para β_j es:

$$\beta_j \in \left(\hat{\beta}_j - t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}}\ \ ,\ \ \hat{\beta}_j + t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{(j+1)(j+1)}}\ \right)$$

A manera de ejemplo, para j=1 se tiene el siguiente intervalo de confianza para β_1

$$\beta_1 \in \left(\hat{\beta}_j - t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{22}}\ \ ,\ \ \hat{\beta}_j + t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{22}}\ \right).$$

Y así sucesivamente para $j \in \mathbf{N}$ tal que $j \in [0,k]$ donde k es el número total de variables explicativas con la que se ajustó el modelo de regresión múltiple.

29.2. Intervalo para σ^2

Para hacer inferencias sobre σ^2 se hace notar la siguiente cantidad pivotal:

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} = \frac{SC_{error}}{\sigma^2} \sim \chi^2_{n-k-1}$$

De esta manera, el intervalo de confianza $1-\alpha$ queda definido de la siguiente manera:

$$\mathbf{P}\left[\chi_{n-k-1}^{2(\alpha/2)} < \frac{SC_{error}}{\sigma^2} < \chi_{n-k-1}^{2(1-\alpha/2)}\right] = 1-\alpha$$

$$\mathbf{P}\left[\frac{1}{\chi_{n-k-1}^{2(\alpha/2)}} > \frac{\sigma^2}{SC_{error}} > \frac{1}{\chi_{n-k-1}^{2(1-\alpha/2)}}\right] = 1-\alpha$$

$$\mathbf{P}\left[\frac{SC_{error}}{\chi_{n-k-1}^{2(1-\alpha/2)}} < \sigma^2 < \frac{SC_{error}}{\chi_{n-k-1}^{2(\alpha/2)}}\right] = 1-\alpha$$

Por lo tanto, el intervalo de confianza $1-\alpha$ para σ^2 es:

$$\sigma^2 \in \left(\frac{SC_{error}}{\chi_{n-k-1}^{2(1-\alpha/2)}}\ \ ,\ \ \frac{SC_{error}}{\chi_{n-k-1}^{2(\alpha/2)}}\right).$$

29.3. Intervalos de la respuesta media

La estimación del valor de Y a través de una X dada, se le conoce como **respuesta media** en un determinado punto, es decir, si se conocen los valores de las variables explicativas $x_1, x_2, \ldots, x_k$ en determinado punto, entonces se puede calcular la estimación del valor Y en ese punto, para ello se define al vector de valores conocidos x^* como:

$$x^* = (1 \quad x_1 \quad x_2 \quad \ldots \quad x_k)\,.$$

De esta forma el valor ajustado de Y dado $X = x^*$ es:

$$\mathbf{E}[\ Y \mid X = x^*] = x^* \underline{\beta}$$

Al suponer que $\underline{\beta}$ es un valor desconocido, se propone como estimador de $\underline{\beta}$ al obtenido por máxima verosimilitud, así el valor estimado de Y sería:

$$\mathbf{E}\left[\widehat{Y \mid X = x^*}\right] = x^* \underline{\hat{\beta}}$$

El valor ajustado de Y dado $X = x^*, \mathbf{E}\left[\widehat{Y \mid X = x^*}\right]$, se denota por $\hat{Y^*}$ para facilitar su tratamiento; además se observa que la esperanza y varianza están determinados por las siguientes igualdades, las cuales se demostrarán a continuación:

$$\mathbf{E}[\hat{Y^*}] = x^* \underline{\beta} \qquad Var(\hat{Y^*}) = \sigma^2 x^{*'} (X'X)^{-1} x^*$$

Teorema 3.8 Sea $\hat{Y^*}$, el valor de la respuesta media de Y en un determinado punto, de forma que $\hat{Y^*} = x^* \underline{\hat{\beta}}$, entonces se cumple que:

a) $\mathbf{E}[\hat{Y^*}] = x^* \underline{\beta}$.

b) $\mathbf{Var}(\hat{Y^*}) = \sigma^2 x^{*'} (X'X)^{-1} x^*$.

Demostración:

a) Para demostrar la esperanza del valor esperado, se observa que:

Por hipótesis

$$\mathbf{E}[\hat{Y^*}] = \mathbf{E}[x^* \underline{\hat{\beta}}]$$

Por linealidad de la esperanza

$$\mathbf{E}[\hat{Y^*}] = x^* \mathbf{E}[\underline{\hat{\beta}}]$$

Ya que $\underline{\hat{\beta}}$ es insesgado

$$\therefore \mathbf{E}[\hat{Y^*}] = x^* \underline{\beta}. \blacksquare$$

b) Para la varianza se tiene:

Por hipótesis

$$Var(\hat{Y^*}) = Var(x^* \underline{\hat{\beta}})$$

$$Var(\hat{Y^*}) = x^* Var(\underline{\hat{\beta}}) x^{*'}$$

Por el **Teorema 3.4**

$$Var(\hat{Y^*}) = x^* \sigma^2 (X'X)^{-1} x^{*'}$$
$$\therefore Var(\hat{Y^*}) = \sigma^2 x^{*'} (X'X)^{-1} x^*. \blacksquare$$

Bajo el supuesto de normalidad se tiene que la respuesta media de Y en un determinado punto, sigue una distribución normal con parámetros:

$$\hat{Y^*} \sim \mathbf{N}_n \left(x^* \underline{\beta}, \sigma^2 x^{*'} (X'X)^{-1} x^* \right).$$

Al estandarizar para obtener la cantidad pivotal se obtiene:

$$\frac{\hat{Y^*} - x^* \underline{\beta}}{\sqrt{\sigma^2 x^{*'} (X'X)^{-1} x^*}} \sim \mathbf{N}(0,1)$$

Como $\frac{(n-k-1)}{\sigma^2} \hat{\sigma}^2 \sim \chi^2_{n-k-1}$, al dividir una distribución normal con una distribución χ^2, se obtiene una distribución $t - Student$

$$\frac{\frac{\hat{Y^*} - x^* \underline{\beta}}{\sqrt{\sigma^2 x^{*'} (X'X)^{-1} x^*}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \sim t_{n-k-1}.$$

Simplificando la ecuación anterior se obtiene la cantidad pivotal deseada:

$$\frac{\hat{Y^*} - x^* \underline{\beta}}{\sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*}} \sim t_{n-k-1.}$$

donde $\hat{\sigma}^2 = \frac{SC_{error}}{n-k-1}$. Por lo tanto, sustituyendo a $\hat{Y^*}$ por $x^* \underline{\hat{\beta}}$, el intervalo de confianza $1-\alpha$ para $\mathbf{E}[Y \mid X = x^*]$ es:

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1} < \frac{\hat{Y^*} - x^* \underline{\beta}}{\sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*}} < t^{\alpha/2}_{n-k-1} \right] = 1 - \alpha$$

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1} < \frac{x^* \underline{\hat{\beta}} - x^* \underline{\beta}}{\sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*}} < t^{\alpha/2}_{n-k-1} \right] = 1 - \alpha$$

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1} \sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*} < x^* \underline{\hat{\beta}} - x^* \underline{\beta} < t^{\alpha/2}_{n-k-1} \sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*} \right] = 1 - \alpha$$

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1} \sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*} < x^* \underline{\beta} - x^* \underline{\hat{\beta}} < t^{\alpha/2}_{n-k-1} \sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*} \right] = 1 - \alpha$$

$$\mathbf{P}\left[x^* \underline{\hat{\beta}} - t^{\alpha/2}_{n-k-1} \sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*} < x^* \underline{\beta} < x^* \underline{\hat{\beta}} + t^{\alpha/2}_{n-k-1} \sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*} \right] = 1 - \alpha$$

Por lo tanto, el intervalo de confianza $1-\alpha$ para $\mathbf{E}[Y \mid X = x^*] = x^* \underline{\beta}$ es:

$$x^* \underline{\beta} \in \left(x^* \underline{\hat{\beta}} - t^{\alpha/2}_{n-k-1} \sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*} \quad , \quad x^* \underline{\hat{\beta}} + t^{\alpha/2}_{n-k-1} \sqrt{\hat{\sigma}^2 x^{*'} (X'X)^{-1} x^*} \right).$$

29.4. Intervalos de predicción

La diferencia entre los intervalos de respuesta media y el intervalo de confianza para el valor esperado Y, es que en el intervalo del valor esperado se desea hallar el valor que en promedio se debería obtener al conocer el vector $x^*(x^*\underline{\beta})$, es decir, se busca el valor de la regresión en el punto x^*, mientras que en un intervalo de predicción se "predice" un valor Y dado que se conocen los valores de las variables explicativas, reflejadas en el vector x_o.

De esta manera se desea estimar la observación $Y = x_0\underline{\beta} + \epsilon$, por medio de $\hat{Y}_x = x_0\underline{\hat{\beta}} + \epsilon$, donde $\epsilon \sim \mathbf{N}(0, \sigma^2)$.

Para ello se define al vector de valores conocidos x_0 como:

$x_0 = (1 \;\; x_1 \;\; x_2 \;\; \ldots \;\; x_k)$.

Donde x_i es el valor de la variable explicativa $i \in 1, \ldots, k$ en un determinado momento o punto. Así la varianza de un valor de predicción considera la varianza de la **respuesta media** en un punto dado en el cual además se añade la varianza del modelo es decir:

$$Var(\hat{Y}_x) = Var(\hat{Y^*}) + Var(Y).$$

Por lo que sustituyendo los valores se tiene que:

$$Var(\hat{Y}_x) = \sigma^2 x_0'(X'X)^{-1}x_0 + \sigma^2$$

$$Var(\hat{Y}_x) = \sigma^2 \left(1 + x_0'(X'X)^{-1}x_0\right).$$

El valor esperado de $\hat{Y}_x$ es igual a la esperanza de $\hat{Y^*}$, de esta manera:

$$\mathbf{E}[\hat{Y}_x] = x_0\underline{\beta}.$$

Teorema 3.9 Sea x_0 el vector que contiene los valores observados de las variables explicativas x_i, $i \in 1, \ldots, k$ en un determinado punto y sea $\hat{Y}_x$ la predicción de un valor Y evaluado en el vector x_0 de la forma $\hat{Y}_x = x_0\underline{\hat{\beta}} + \epsilon$, en donde $\epsilon \sim \mathbf{N}(0, \sigma^2)$, entonces se cumple que:

a) $\mathbf{E}[\hat{Y}_x] = x_0\underline{\beta}$

b) $\mathbf{Var}(\hat{Y}_x) = \sigma^2 \left(1 + x_0'(X'X)^{-1}x_0\right)$

Demostración:

a) Para demostrar la esperanza del valor de predicción, se observa que:

Por hipótesis

$$\mathbf{E}[\hat{Y}_x] = \mathbf{E}[x_0\underline{\hat{\beta}} + \epsilon]$$

Por linealidad de la esperanza

$$\mathbf{E}[\hat{Y}_x] = x_0\mathbf{E}[\underline{\hat{\beta}}] + \mathbf{E}[\epsilon]$$

Tenemos que $\hat{\beta}$ es insesgado y por hipótesis

$$\mathbf{E}[\hat{Y}_x] = x_0\underline{\beta} + 0$$

$$\therefore \mathbf{E}[\hat{Y}_x] = x_0\underline{\beta}.\blacksquare$$

b) Para la varianza se tiene:

Por hipótesis

$$Var(\hat{Y}_x) = Var(x_0\underline{\hat{\beta}} + \epsilon)$$

$$Var(\hat{Y}_x) = x_0 Var(\underline{\hat{\beta}})x_0' + Var(\epsilon) + 2x_0 Cov(\underline{\hat{\beta}}, \epsilon)$$

Por el **teorema 3.4** e independencia de ϵ

$$Var(\hat{Y}_x) = x_0\sigma^2(X'X)^{-1}x_0' + \sigma^2 + 0$$

$$\therefore Var(\hat{Y}_x) = \sigma^2\left(1 + x_0'(X'X)^{-1}x_0\right). \blacksquare$$

Bajo el supuesto de normalidad se tiene que la predicción de Y en un determinado punto, sigue una distribución normal con parámetros:

$$\hat{Y}_x \sim \mathbf{N}_n\left(x_0\underline{\beta}, \sigma^2\left(1 + x_0'(X'X)^{-1}x_0\right)\right).$$

De esta manera al suponer normalidad, estandarizando y dividiendo entre $\frac{(n-k-1)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-k-1}$ se obtiene la cantidad pivotal deseada:

$$\frac{\hat{Y}_x - x_0\underline{\beta}}{\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}} \sim t_{n-k-1}.$$

Por lo tanto, el intervalo de confianza $1-\alpha$ para la predicción de Y es:

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1} < \frac{\hat{Y}_x - x_0\underline{\beta}}{\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}} < t^{\alpha/2}_{n-k-1}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1} < \frac{x_0\underline{\hat{\beta}} + \epsilon - x_0\underline{\beta}}{\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}} < t^{\alpha/2}_{n-k-1}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} < x_0\underline{\hat{\beta}} + \epsilon - x_0\underline{\beta} < t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} < x_0\underline{\beta} - x_0\underline{\hat{\beta}} - \epsilon < t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}\right] = 1-\alpha$$

$$\mathbf{P}\left[-t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} < x_0\underline{\beta} - \epsilon - x_0\underline{\hat{\beta}} < t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}\right] = 1-\alpha$$

$$\mathbf{P}\left[x_0\underline{\hat{\beta}} - t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} < x_0\underline{\beta} - \epsilon < x_0\underline{\hat{\beta}} + t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}\right] = 1-\alpha$$

Dado que ϵ es una variable aleatoria simétrica y de media 0 entonces se puede sustituir $-\epsilon$ por ϵ, así:

$$\mathbf{P}\left[x_0\underline{\hat{\beta}} - t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} < x_0\underline{\beta} + \epsilon < x_0\underline{\hat{\beta}} + t^{\alpha/2}_{n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}\right] = 1-\alpha$$

Por lo tanto el intervalo de confianza para la predicción de Y dado el vector x_0, con un nivel de significancia α es:

$$x_0\underline{\beta} + \epsilon \in \left(x_0\underline{\hat{\beta}} - t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} \;\;,\;\; x_0\underline{\hat{\beta}} + t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} \right)$$

$$x_0\underline{\beta} + \epsilon \in \left(\hat{Y} - t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} \;\;,\;\; \hat{Y} + t_{n-k-1}^{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} \right).$$

Capítulo 30

Pruebas de hipótesis

Una vez estimados los parámetros de forma puntual a través del método de mínimos cuadrados o de máxima verosimilitud, nos preguntaremos si realmente la variable regresora $X_j, \quad j \in [1,k]$ donde k es el número total de variables regresoras del modelo, que aportan información o son **significativas** para el modelo, para respondernos ésto contruiremos las pruebas de hipótesis, primero para β_j y después para σ^2 como se detallará a continuación:

30.1. Región de rechazo para β_j

Dado que el estimador de $\underline{\beta}$ es una combinación lineal, de igual forma se tiene una combinación lineal de variables aleatorias normales independientes, por lo que se tiene para $\beta_j \ \forall j \ \in [0,k]$ una distribución asociada normal con parámetros:

$$\beta_j \sim \mathbf{N}\left(\beta_j, \sigma^2 (X'X)^{-1}_{(j+1)(j+1)}\right).$$

Donde $(X'X)_{(j+1)(j+1)}$ es la entrada $j+1$ entrada en la diagonal de la matriz $(X'X)$, además el subíndice j hace referencia al parámetro β_j al que se desea examinar mediante la prueba de hipótesis, normalizando la expresión anterior se tiene:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 (X'X)^{-1}_{(j+1)(j+1)}}} \sim \mathbf{N}(0,1)$$

Como $\frac{(n-k-1)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-k-1}$ tenemos:

$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 (X'X)^{-1}_{(j+1)(j+1)}}}}{\sqrt{\frac{\frac{(n-k-1)}{\sigma^2}\hat{\sigma}^2}{n-k-1}}} \sim t_{n-k-1}$$

Por lo tanto, el estadístico deseado para probar que los tres posibles casos de β_j, que sea igual, inferior o superior a un valor dado b, es la siguiente:

$$t = \frac{\hat{\beta}_j - b}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{(j+1)(j+1)}}}.$$

Usando el estadístico anterior se definen las regiones de rechazo para cada caso de β_j como:

Hipótesis	*Región de rechazo* :
$H_0 : \beta_j = b \; vs. \; H_a : \beta_j \neq b$	$\lvert t \rvert > t^{(\alpha/2)}_{(n-k-1)}$
$H_0 : \beta_j \leq b \; vs. \; H_a : \beta_j > b$	$t > t^{(\alpha)}_{(n-k-1)}$
$H_0 : \beta_j \geq b \; vs. \; H_a : \beta_j < b$	$t < t^{(1-\alpha)}_{(n-k-1)}$

Para ejemplificar lo anterior:

Supongamos que queremos ajustar un modelo de regresión lineal múltiple, sin embargo, dudamos si realmente la variable respuesta $\underline{Y}$ depende de la variable regresora X_5, es decir, sospechamos que $\beta_5 = 0$. Para ello realizaremos la siguiente prueba de hipótesis evaluando β_5 en el punto crítico de interés $b = 0$.

$$\mathbf{H}_0 : \beta_5 = 0 \;\; vs. \;\; \mathbf{H}_a : \beta_5 \neq 0.$$

La regla de decisión, es rechazar H_0 cuando el estadístico $t = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{(j+1)(j+1)}}}$ esté contenida en la región de rechazo. Por lo que se rechaza H_0 cuando $t < -t^{\alpha/2}_{n-k-1}$ o si $t > t^{\alpha/2}_{n-k-1}$.

Si se rechaza H_0 entonces $\beta_5 \neq 0$ por lo que la variable regresora X_5 tiene influencia en la variable respuesta $\underline{Y}$; Si no se rechaza la hipótesis nula con un nivel de significancia α, $\beta_5 = 0$, es decir, la variable regresora X_5 no influye estadísticamente en la variable respuesta $\underline{Y}$, entonces en nuestro modelo de regresión lineal múltiple con la combinación de variables regresoras dada, no es conveniente usar la variable X_5 pues no es significativa para nuestro modelo.

De manera similar, se puede realizar la anterior prueba para cualquier β_j para $j \in \mathbf{N}$ tal que $j \in [0, k]$ donde k es el número total de variables regresoras con las que se ajusta el modelo de regresión múltiple.

30.2. Prueba para σ^2

Para ésta parte necesitaremos las siguientes relaciones de probabilidad:

$$\frac{n-k-1}{\sigma^2} \, \hat{\sigma}^2 \;\; \sim \chi^2_{n-k-1}$$

$$\frac{n-k-1}{\sigma^2} \left(\frac{SC_{error}}{n-k-1} \right) \;\; \sim \chi^2_{n-k-1}$$

$$\frac{SC_{error}}{\sigma^2} \;\; \sim \chi^2_{n-k-1}.$$

Dado que se quiere realizar inferencias sobre σ^2 definimos a la siguiente estadística S como:

$$S = \frac{SC_{error}}{\sigma_0^2}$$

Donde σ_0^2 es el valor crítico que se desea poner a prueba, para verificar a través de las hipótesis, que la varianza del modelo σ^2 es igual, superior o inferior al punto crítico σ_0^2 propuesto.

A continuación, tenemos las siguientes regiones de rechazo:

Hipótesis	*Región de rechazo* :
$H_0 : \sigma^2 = \sigma_o^2 \; vs. \; H_a : \sigma^2 \neq \sigma_0^2$	$S > \chi^{2(1-\alpha/2)}_{(n-k-1)} \;\; o \;\; S < \chi^{2(\alpha/2)}_{(n-k-1)}$
$H_0 : \sigma^2 \leq \sigma_o^2 \; vs. \; H_a : \sigma^2 > \sigma_o^2$	$S > \chi^{2(\alpha)}_{(n-k-1)}$
$H_0 : \sigma^2 \geq \sigma_o^2 \; vs. \; H_a : \sigma^2 < \sigma_o^2$	$S < \chi^{2(1-\alpha)}_{(n-k-1)}$

30.3. Análisis de la varianza (ANOVA)

La idea del método de análisis de la varianza sirve para probar el significado de la regresión.

El análisis de la ANOVA está conformada por varios elementos como:

- SC_T es la suma de cuadrados, la cual sirve para medir la variabilidad de las observaciones del total corregido por la media.
- SC_{reg} es la suma de cuadrados de la regresión, mide la variabilidad en que las observaciones y_i y la línea de regresión.
- SC_{error} es la suma de cuadrados del residual o del error, es decir, mide la variación residual que queda sin explicar por la línea de regresión.

Con éstas variables se cumple con la siguiente igualdad:

$$SC_T = SC_{reg} + SC_{error}.$$

Por el **corolario 5** tenemos que:

$$SC_{error} = \underline{Y}'(I - H)\underline{Y}.$$

Adaptando la suma de residuales a la forma matricial tenemos lo siguiente:

$$SC_T = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$SC_T = \sum_{i=1}^{n}(y_i^2 - \overline{y}^2)$$

$$SC_T = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$$

$$\therefore SC_T = \underline{Y}'\underline{Y} - n\overline{y}^2.$$

Por otro lado tenemos:

$$SC_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

$$SC_{reg} = \sum_{i=1}^{n}(\hat{y_i}^2 - \overline{y}^2)$$

$$\therefore \ SC_{reg} = \hat{\beta}' X' \underline{Y} - n\overline{y}^2. \blacksquare$$

Además se sabe que el modelo de regresión lineal múltiple con errores normales tiene las siguientes propiedades:

- $\frac{SC_{reg}}{\sigma^2} \sim \chi^2_{k+1}$
- $\frac{SC_{error}}{\sigma^2} \sim \chi^2_{n-k-1}$
- SC_{reg} es independiente a SC_{error}

Y podemos obtener los siguientes resultados:

- La suma de cuadrados del total (SC_T) tiene $n-1$ grados de libertad.
- La suma de cuadrados de la regresión (SC_{reg}) tiene k grados de libertad.

- La suma de cuadrados de los errores (SC_{error}) tiene $n-k-1$ grados de libertad.

Se observa que si $X \sim \chi^2_n$, $\underline{Y} \sim \chi^2_m$ y si X es independiente a $\underline{Y}$ entonces:

$$\frac{X/n}{Y/m} \sim F_{n,m}.$$

De esta forma, aplicando la prueba F de $Fisher$ en el análisis de varianza se puede usar como región de rechazo para demostrar la prueba de hipótesis, la cual consiste en dividir SC_{reg} entre sus grados de libertad que se divide entre la razón de la SC_{error} y sus grados de libertad:

$$F = \frac{\frac{SC_{reg}}{k}}{\frac{SC_{error}}{n-k-1}} \sim F_{k,n-k-1}.$$

A manera de notación se usa el **Cuadrado Medio** denotado como **CM** que corresponde a la Suma de Cuadrados entre sus grados de libertad. Así $CM = \frac{SC_{reg}}{k}$ y $CM_{error} = \frac{SC_{error}}{n-k-1}$.

De esta manera podemos reescribir lo anterior como:

$$F = \frac{CM_{reg}}{CM_{error}} \sim F_{k,n-k-1}.$$

Entonces se define a la región de rechazo con un nivel de significancia α para:

$$\mathbf{H}_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$
$$F > F^{(\alpha)}_{k,n-k-1}.$$

La forma en que se resume la información es mostrada en la siguiente tabla mejor conocida como:

Tabla ANOVA

	Grados de libertad	*Suma de Cuadrados*	*Cuadrado Medio*	*Prueba F*
Regresión	k	$\hat{\beta}' X' \underline{Y} - n\bar{y}^2$	$\frac{SC_{reg}}{k}$	$\frac{CM_{reg}}{CM_{error}}$
Error	$n-k-1$	$\underline{Y}'(I-H)\underline{Y}$	$\frac{SC_{error}}{n-k-1}$	–
Total	$n-1$	$\underline{Y}'\underline{Y} - n\bar{y}^2$	–	–

30.4. Coeficiente de determinación

El coeficiente de determinación con frecuencia se le asocia a la proporción de la variación explicada por la variable regresora X, ya que $0 \leq SC_{reg} \leq SC_T$ entonces los valores del coeficiente de determinación están entre $0 \leq R^2 \leq 1$. Cabe mencionar que este coeficiente proporciona una idea muy cercana a la variabilidad que cae sobre la regresión, sin embargo, no hayq ue perder de vista que en primera instancia ésta medida sirve como una aproximación.

Se define el coeficiente de determinación del modelo de regresión como:

$$R^2 = \frac{SC_{reg}}{SC_T} = 1 - \frac{SC_{error}}{SC_T}$$

Los valores cercanos a 1 implican que la mayor parte de la cariabilidad de $\underline{Y}$ está explicada por el modelo de regresión.

30.5. R^2 ajustado

Como se mencionó el coeficiente de determinación es una aproximación, ya que el valor de la medida se aproxima a 1, conforme vamos incluyendo más variables. Es decir, entre más variables en el modelo el coeficiente de determinación supone que va mejorando el ajuste, lo cual no es cierto, entre más variables no siempre se proporcionará un mejor ajuste. Es por ello que se realiza una R^2 ajustada de la siguiente forma:

$$R^2_{adj} = 1 - \frac{CM_{error}}{CM_T}.$$

Capítulo 31

Validación de supuestos

Como vimos anteriormente, debemos validar los supuestos del modelo de regresión lineal múltiple:

- Supuesto de normalidad
- Supuesto de linealidad
- Supuesto de homocedasticidad
- Valores outlier e influyentes
- Supuesto de multicolinealidad

Los primeros enlistados ya se vieron en la sección anterior, lo que vamos a introducir ahora es el supuesto de multicolinealidad.

31.1. Supuesto de multicolinealidad

Se dice que el ajuste del modelo lineal sobre una muestra tiene problemas de multicolinealidad cuando hay una correlación alta entre dos o más variables explicativas, consecuencia de que una variable regresora es linealmente dependiente a alguna o algunas de las demás variables de la matriz diseño X.

Un modelo de regresión lineal con ésta característica provoca un error en la estimación de los parámetros $\underline{\beta}$, ya que al tener variables linealmente dependientes el determinante $\mid X'X \mid \to 0$, lo cual provoca que la inversa $(X'X)^{-1} \to \infty$. Esto provoca el cálculo erróneo de los parámetros, ya que el producto matricial de $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{Y}$ no puede ser aproximado de la mejor manera.

31.2. Detección de multicolinealidad

Debido al problema de estimación de los parámetros $\underline{\beta}$ que ocasiona un modelo con multicolinealidad, se realizan validaciones con la finalidad de tener evidencia suficiente para asumir si el modelo posee o no el problema de dependencia lineal de las variables explicativas.

El método más adecuado para efectuar ésta validación, es encontrar en la matriz de diseño X variables regresoras linealmente dependendientes a través de operaciones matriciales elementales, sin embargo, este procedimiento es tardado conforme el tamaño de la muestra es mayor, además de que computacionalmente el procedimiento no es óptimo, es por ello que se desarrollan diferentes métodos para conocer si el modelo presenta multicolinealidad.

A través de correlaciones

El primer método para detectar la dependencia lineal es analizar el coeficiente de correlación de las variables, ya que si el modelo posee variables altamente correlacionadas entre sí, es probable que el ajuste lineal presente problemas de multicolinealidad debido a la relación estrecha de éstas variables, una

desventaja de éste método es que analiza a través de la matriz de correlaciones de una variable comparada con otra, sin embargo, puede existir una baja correlación entre variables analizándolas una a una pero al analizarlas de dos en dos, o de tres en tres, etcétera, puede encontrarse una alta correlación.

Para analizar de mejor forma la información, se obtiene el determinante de la matriz de correlaciones, si este es muy cercano a cero podría proporcionar indicios de problemas de multicolinealidad, debido a que si hay dos variables que son linealmente dependientes el determinante de la matriz de correlaciones tenderá a cero. Sin embargo, depende del prejuicio del investigador a partir de qué cifra debe ser considerada una cantidad mínima.

A través del índice κ

En 1960, Jacob Cohen, presentó el método del coeficiente kappa. El coeficiente kappa o simplemente κ se basa en el análisis de los eigenvalores $\lambda_1, \lambda_2, \ldots, \lambda_k$ de la matriz $(X'X)$, este procedimiento es llamado *análisis del egeinsistema.*

Debido a que $(X'X)$ es una matriz simétrica, el producto de la matriz diseño es igual a sus valores propios o eigenvalores. De esta manera el método propone observar la proporción del eigenvalor más grande respecto al más pequeño, si la proporción es pequeña, no hay problemas de multicolinealidad pues todos los valores propios son similares, sin embargo, proporciones altas indican gran variabilidad en los valores propios por lo que se tiene indicios de multicolinealidad. De esta manera el coeficiente κ es calculado como:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}.$$

31.3. Ejemplo

Se tiene información sobre las cajetillas vendidas y queremos saber que factores influyen en dichas ventas.

```
cigarros=read.table("datos cigarros.csv",sep=",",header=TRUE)
attach(cigarros)
```

Lo primero que hay que hacer es identificar las variables que hay en la base y realizar diagramas de dispersión para tener una idea de que tipo de relación existe entre las variables.

```
  ESTADO  SCIG  AGE   ED PERFEM PRICE
1      1  89.8 27.0 41.3   51.7  42.7
2      2 121.3 22.9 66.7   45.7  41.8
3      3 115.2 26.3 58.1   50.8  38.5
4      4 100.3 29.1 39.9   51.5  38.8
5      5 123.0 28.1 62.6   50.8  39.7
6      6 124.8 26.2 63.9   50.7  31.1
```

$\mathbf{y} =$ **SCIG** : Cajetillas de cigarro vendidas

$\mathbf{x}_1 =$ **AGE** : Edad promedio

$\mathbf{x}_2 =$ **ED** : Porcentaje de personas con más de 25 años

$\mathbf{x}_3 =$ **PERFEM** : Porcentaje de mujeres

$\mathbf{x}_4 =$ **PRICE** : Precio

```
     ESTADO           SCIG            AGE             ED       
 Min.   : 1.00   Min.   : 65.5   Min.   :22.90   Min.   :37.80  
 1st Qu.:14.75   1st Qu.:105.9   1st Qu.:26.48   1st Qu.:47.80  
 Median :28.50   Median :117.5   Median :27.30   Median :53.90  
 Mean   :28.50   Mean   :120.9   Mean   :27.50   Mean   :53.23  
 3rd Qu.:42.25   3rd Qu.:124.4   3rd Qu.:28.73   3rd Qu.:59.23  
 Max.   :56.00   Max.   :265.7   Max.   :32.30   Max.   :67.30  
     PERFEM          PRICE      
 Min.   :45.70   Min.   :29.00  
```

```
 1st Qu.:50.67   1st Qu.:34.62
 Median :51.00   Median :38.90
 Mean   :50.93   Mean   :38.01
 3rd Qu.:51.50   3rd Qu.:41.33
 Max.   :53.50   Max.   :45.50
```

Ajuste de un modelo de RLM

Ajustaremos un modelo de RLM considerando lo siguiente:

- Emplearemos el método **backward**.
- El criterio de selección del *mejor modelo* estará basado en el **AIC**

El método **backward** (eliminación hacia atrás) consiste en comenzar con un modelo que incluya las k variables regresoras y luego ir eliminando varible por variable de acuerdo a cierto criterio.

El **AIC** (Criterio de información de Akaike) es una medida de la calidad de un modelo estadístico, considera la *bondad de ajuste* del modelo y su complejidad.

$$AIC = -2log(L) + 2 * n_{parámetros}$$

El mejor modelo es aquel que tenga un **AIC** pequeño.

Primer Paso: Relación entre Variables

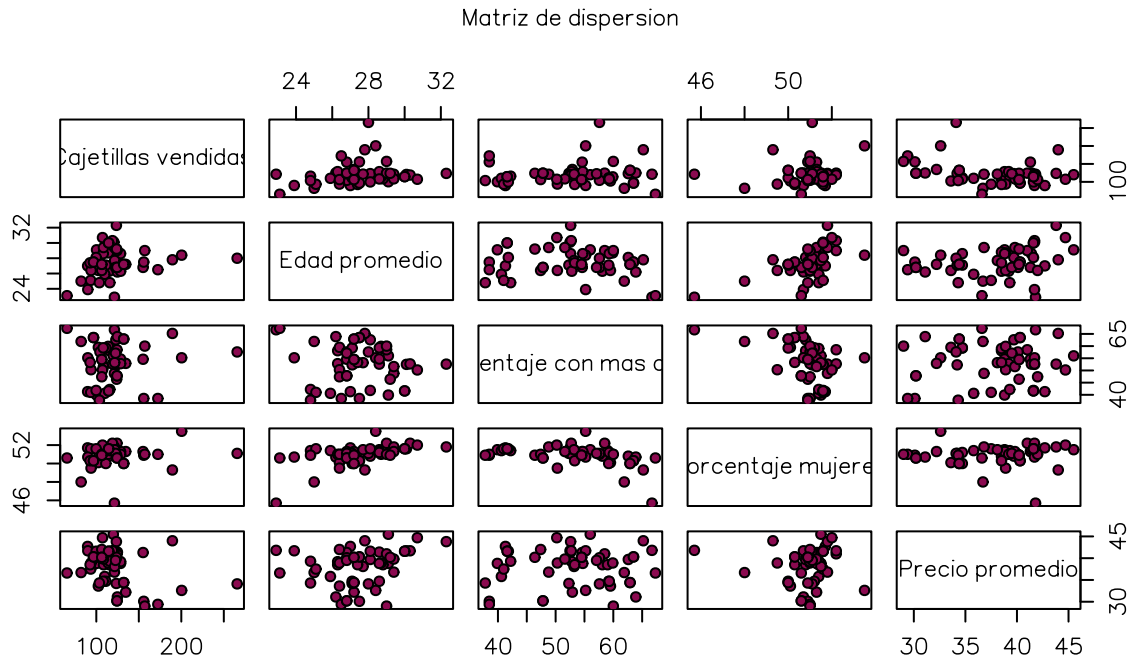
Primero haremos las gráficas de dispersión correspondientes para ver el comportamiento de los datos.

Diagramas de dispersion por parejas: Primero juntamos todas las variables explicativas y la dependiente en un solo objeto para hacer los diagramas de dispersion para cada pareja

```
cigarrosvars=data.frame(cigarros$SCIG,cigarros$AGE,cigarros$ED,cigarros$PERFEM,cigarros$PRICE)
cor(cigarrosvars)
```

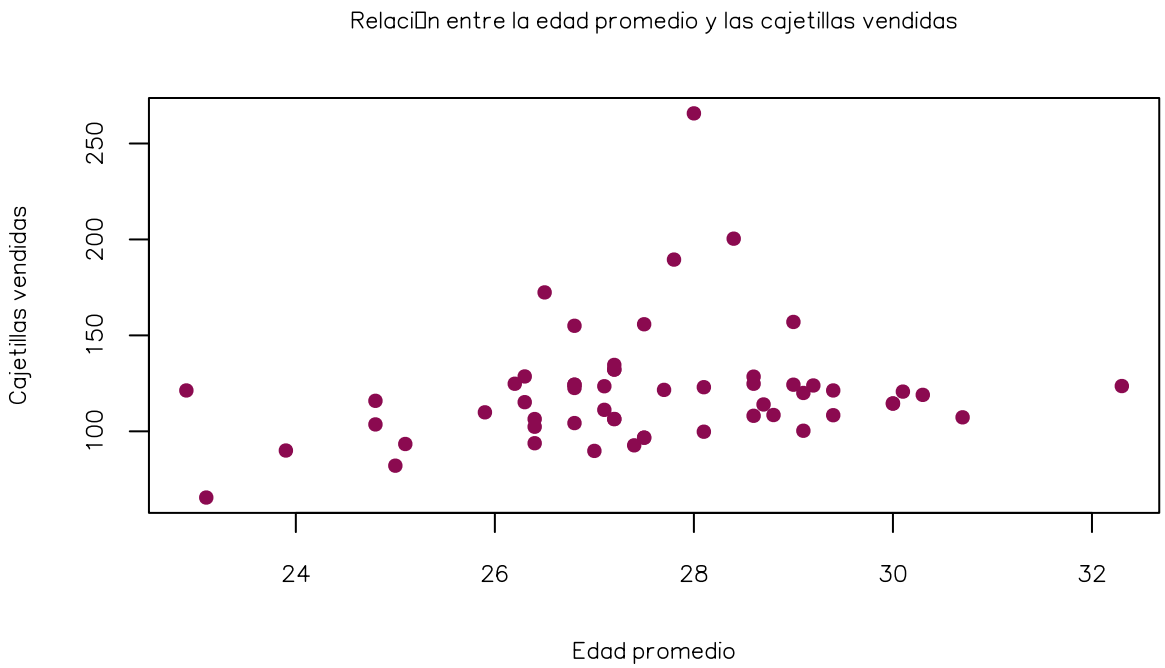
```
               cigarros.SCIG cigarros.AGE cigarros.ED cigarros.PERFEM
cigarros.SCIG      1.00000000    0.2150081  0.05494773      0.14296083
cigarros.AGE       0.21500813    1.0000000 -0.12564933      0.55508730
cigarros.ED        0.05494773   -0.1256493  1.00000000     -0.43495874
cigarros.PERFEM    0.14296083    0.5550873 -0.43495874      1.00000000
cigarros.PRICE    -0.30696253    0.2660560  0.04667213      0.04906057
               cigarros.PRICE
cigarros.SCIG      -0.30696253
cigarros.AGE        0.26605603
cigarros.ED         0.04667213
cigarros.PERFEM     0.04906057
cigarros.PRICE      1.00000000
```

```
pairs(cigarrosvars,main="Matriz de dispersion",labels=c("Cajetillas vendidas", "Edad promedio", "Por
```

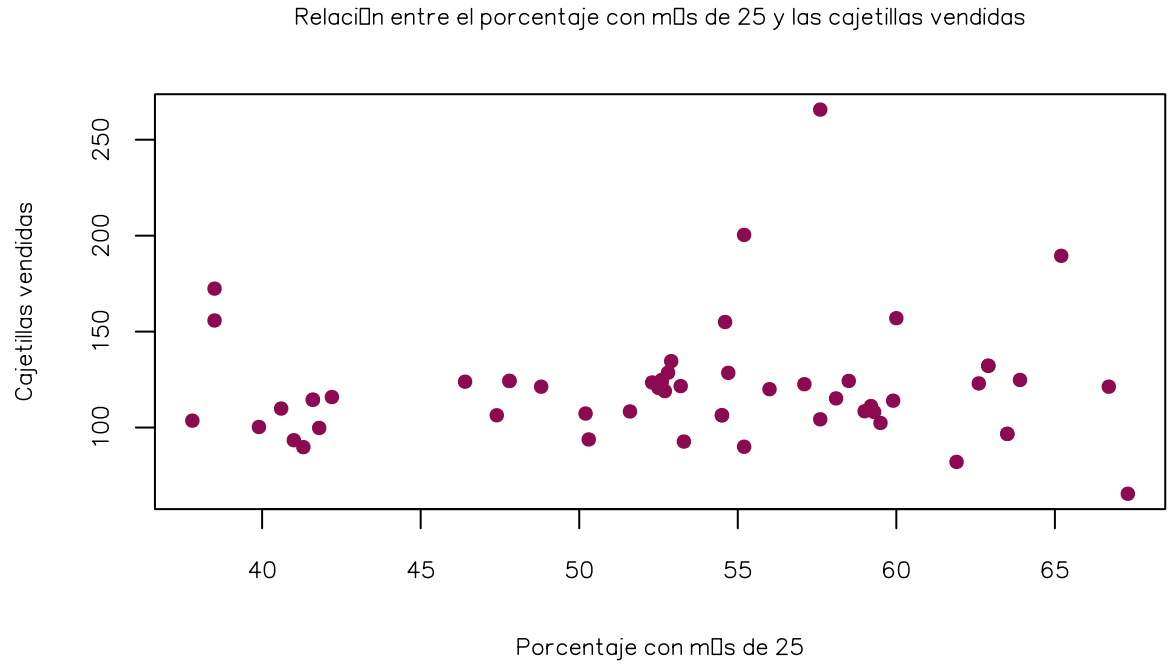


A simple vista "no estamos apreciando toda la información", haremos gráficos dos a dos:

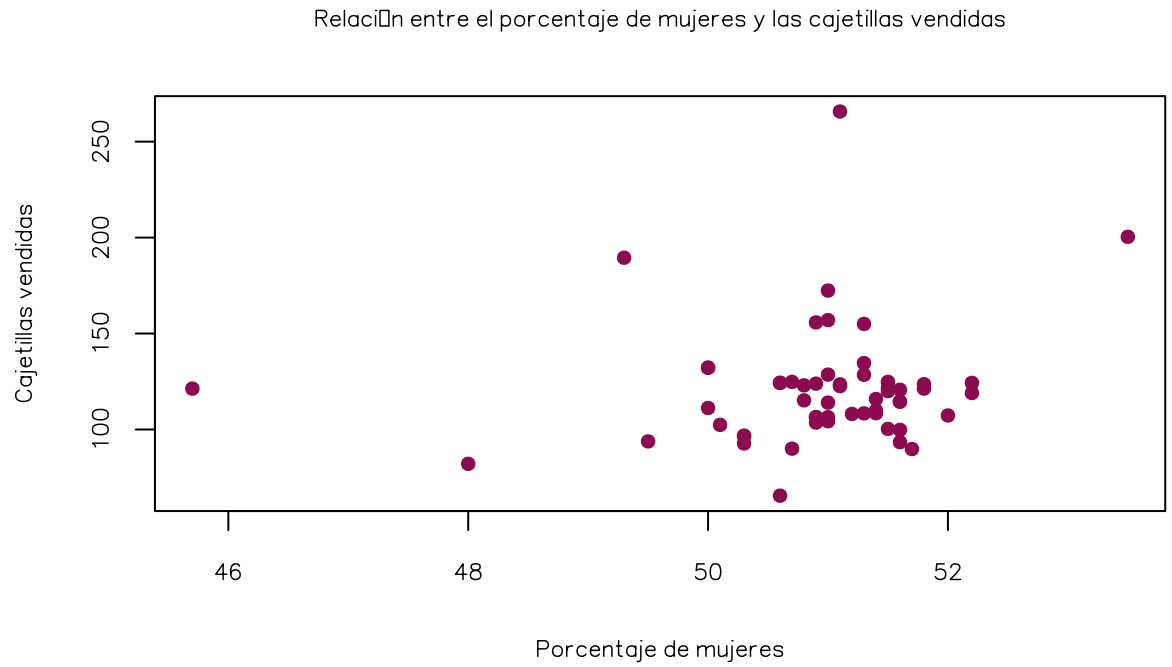
```
plot(cigarros$AGE,cigarros$SCIG,type = "p",col="deeppink4",pch=16,
xlab="Edad promedio", ylab="Cajetillas vendidas",
main= "Relación entre la edad promedio y las cajetillas vendidas")
```



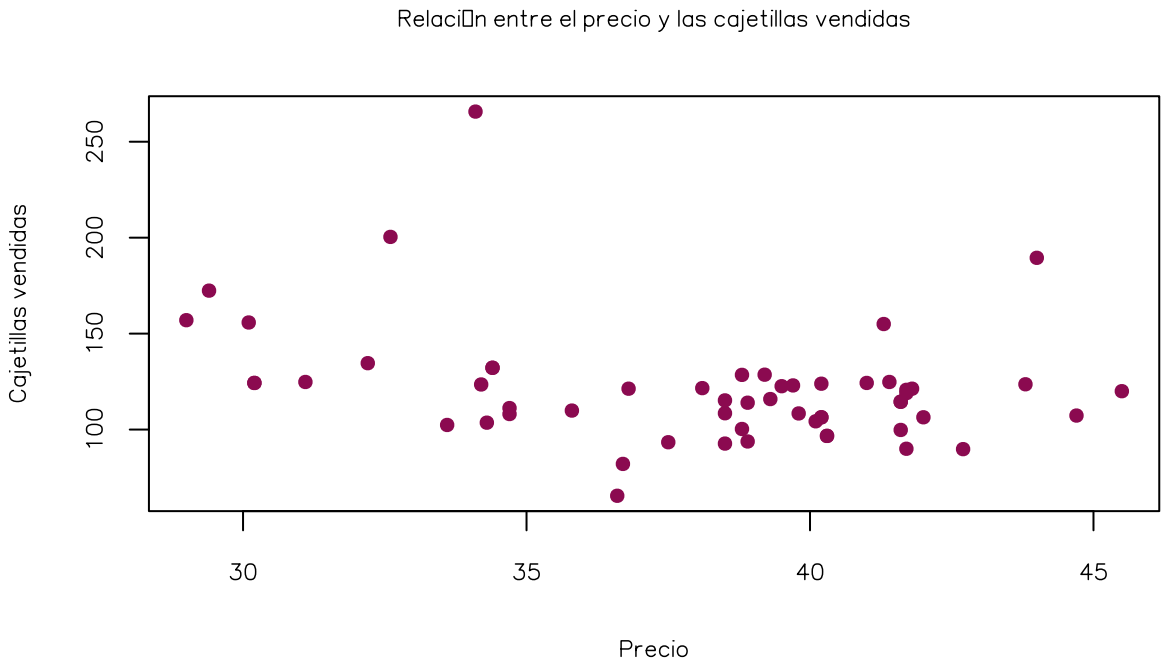

```
plot(cigarros$ED,cigarros$SCIG,type = "p",col="deeppink4",pch=16,
xlab="Porcentaje con más de 25", ylab="Cajetillas vendidas",
main= "Relación entre el porcentaje con más de 25 y las cajetillas vendidas")
```




```
plot(cigarros$PERFEM,cigarros$SCIG,type = "p",col="deeppink4",pch=16,
xlab="Porcentaje de mujeres", ylab="Cajetillas vendidas",
main= "Relación entre el porcentaje de mujeres y las cajetillas vendidas")
```




```
plot(cigarros$PRICE,cigarros$SCIG,type = "p",col="deeppink4",pch=16,
xlab="Precio", ylab="Cajetillas vendidas",
main= "Relación entre el precio y las cajetillas vendidas")
```



Observemos que tanto la variable de promedio de más de 25 y las cajetillas vendidas muestran una relación, al igual que la variable precio.

Veamos ahora la correlación entre las variables:

```
cor(cigarros)
```

```
             ESTADO        SCIG         AGE          ED       PERFEM
ESTADO  1.000000000 -0.11088992  0.05535895 -0.05138058 -0.002021391
SCIG   -0.110889919  1.00000000  0.21500813  0.05494773  0.142960829
AGE     0.055358955  0.21500813  1.00000000 -0.12564933  0.555087298
ED     -0.051380577  0.05494773 -0.12564933  1.00000000 -0.434958738
PERFEM -0.002021391  0.14296083  0.55508730 -0.43495874  1.000000000
PRICE  -0.055232952 -0.30696253  0.26605603  0.04667213  0.049060568
             PRICE
ESTADO -0.05523295
SCIG   -0.30696253
AGE     0.26605603
ED      0.04667213
PERFEM  0.04906057
PRICE   1.00000000
```

Las correlaciones entre las variables que no son las cajetillas vendidas no son “altas” por lo que en principio no parece haber problema de considerar todas las variables dentro del modelo.

Segundo Paso: Generación de Modelos y Elección del Mejor Modelo

Emplearemos comandos de R:

Empezaremos con el modelo inicial, le llamamos así ya que es el modelo más grande que podemos suponer, ya que el modelo explica las cajetillas vendidas en función de las otras 4 variables:

```
modeloini=lm(SCIG~AGE+ED+PERFEM+PRICE, data=cigarros)
summary(modeloini)
```

```

Call:
lm(formula = SCIG ~ AGE + ED + PERFEM + PRICE, data = cigarros)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-43.242 -13.160  -5.053   3.166 128.099

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.0742   231.4589  -0.035  0.97231
AGE           5.2521     2.6733   1.965  0.05492 .
ED            0.5115     0.5389   0.949  0.34699
PERFEM        1.3716     4.8326   0.284  0.77770
PRICE        -2.9599     0.9713  -3.047  0.00365 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.62 on 51 degrees of freedom
Multiple R-squared:  0.2034,    Adjusted R-squared:  0.1409
F-statistic: 3.255 on 4 and 51 DF,  p-value: 0.01878
```

De este primer modelo dado que las pruebas de hipótesis para cada β_i nos dicen que la única variable significativa es β_4 que es para la variable Precio. Y hasta el momento tiene sentido esto.

Ahora si, emplearemos el método backward con el criterio AIC.

```
modelo_backward_AIC=stepAIC(modeloini,direction = "backward")
```

```
Start:  AIC=380.41
SCIG ~ AGE + ED + PERFEM + PRICE

         Df Sum of Sq   RSS    AIC
- PERFEM  1      66.0 41828 378.49
- ED      1     737.8 42500 379.39
<none>                41762 380.41
- AGE     1    3160.6 44923 382.49
- PRICE   1    7604.1 49366 387.77

Step:  AIC=378.49
SCIG ~ AGE + ED + PRICE

        Df Sum of Sq   RSS    AIC
- ED     1     689.2 42517 377.41
<none>                41828 378.49
- AGE    1    5402.7 47231 383.30
- PRICE  1    7813.1 49641 386.08

Step:  AIC=377.41
SCIG ~ AGE + PRICE

        Df Sum of Sq   RSS    AIC
<none>                42517 377.41
- AGE    1    4965.6 47483 381.60
- PRICE  1    7481.7 49999 384.49
```

Ahora observemos el mejor modelo, de acuerdo a lo arrojado en **stepAIC()**.

```
summary(modelo_backward_AIC)
```

```
Call:
lm(formula = SCIG ~ AGE + PRICE, data = cigarros)

Residuals:
    Min      1Q  Median      3Q     Max
-35.828 -15.601  -6.451   3.807 130.691
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.4541    61.0443   1.367  0.17736
AGE           5.3878     2.1656   2.488  0.01603 *
PRICE        -2.9121     0.9536  -3.054  0.00353 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.32 on 53 degrees of freedom
Multiple R-squared:  0.1889,    Adjusted R-squared:  0.1583
F-statistic: 6.174 on 2 and 53 DF,  p-value: 0.003889
```

A través de la tablas ANOVA se tiene:

```
anova(modelo_backward_AIC)
```

```
Analysis of Variance Table

Response: SCIG
          Df Sum Sq Mean Sq F value   Pr(>F)
AGE        1   2423  2423.4  3.0209 0.088005 .
PRICE      1   7482  7481.7  9.3264 0.003529 **
Residuals 53  42517   802.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De los resultados obtenidos podemos refinar el *mejor modelo* obtenido:

```
modelo_refinado=lm(SCIG~0+AGE+PRICE,cigarros)
summary(modelo_refinado)
```

```
Call:
lm(formula = SCIG ~ 0 + AGE + PRICE, data = cigarros)

Residuals:
    Min      1Q  Median      3Q     Max
-30.841 -16.734  -5.385   4.999 131.509

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
AGE      7.809      1.257   6.214 7.73e-08 ***
PRICE   -2.477      0.906  -2.734  0.00845 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.55 on 54 degrees of freedom
Multiple R-squared:  0.9495,    Adjusted R-squared:  0.9476
F-statistic: 507.6 on 2 and 54 DF,  p-value: < 2.2e-16
```

Entonces ya obtuvimos el mejor modelo, que está compuesto por:

$$\text{Cajetillas vendidas} = 7.809 * \text{Edad promedio} - 2.477 * \text{Precio}$$

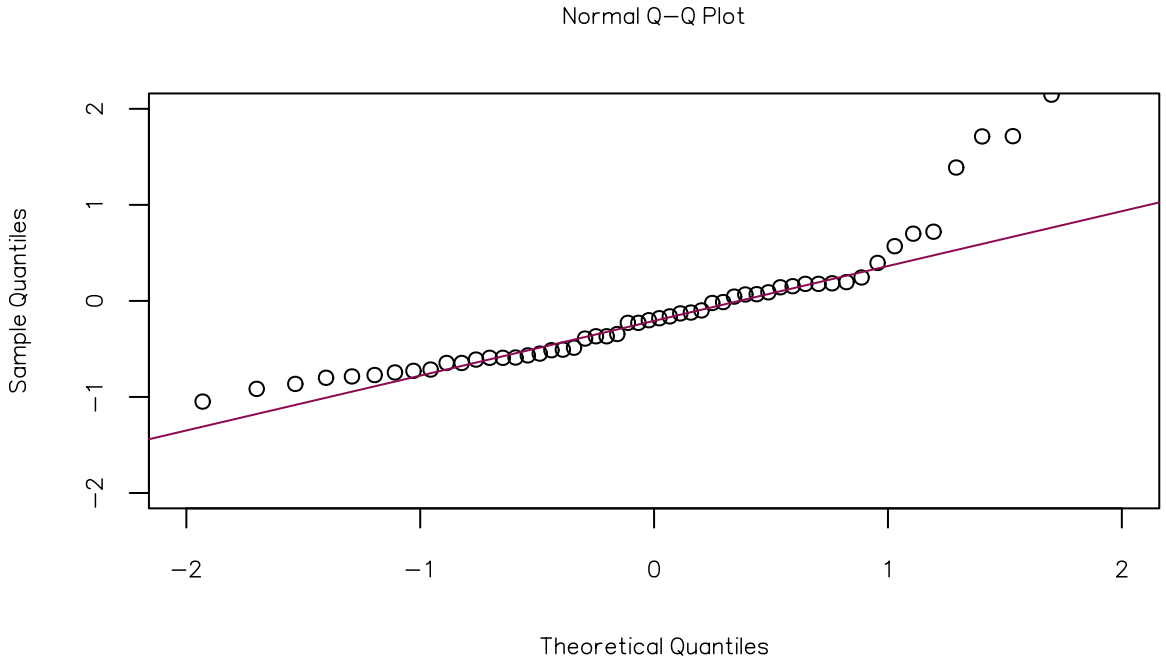
Bueno todavía debemos validar los supuestos:

Tercer Paso Validadción de Supuestos

Siguiendo el orden de nuestros capítulos, para validar gráficamente la normalidad de los errores debemos graficar los errores contra los cuantiles de la distribución normal.Para ello aplicamos la función **qqnorm**

y con **qqline** obtenemos la recta diagonal que nos servirá para ver que tan lejos o cerca de la distribución normal están cayendo los residuales.

```
qqnorm(rstandard(modelo_refinado),ylim = c(-2,2),xlim = c(-2,2))
qqline(rstandard(modelo_refinado),distribution = qnorm,col="deeppink4")
```



Podemos observar que la parte central de la distribución si se ajusta a una distribución normal, sin embargo, en los extremos los residuales ya no se comportan como una distribución normal.

Podemos aplicar la prueba de bondad de ajuste **Lilliefors para normalidad** vista en Bondad de Ajuste:

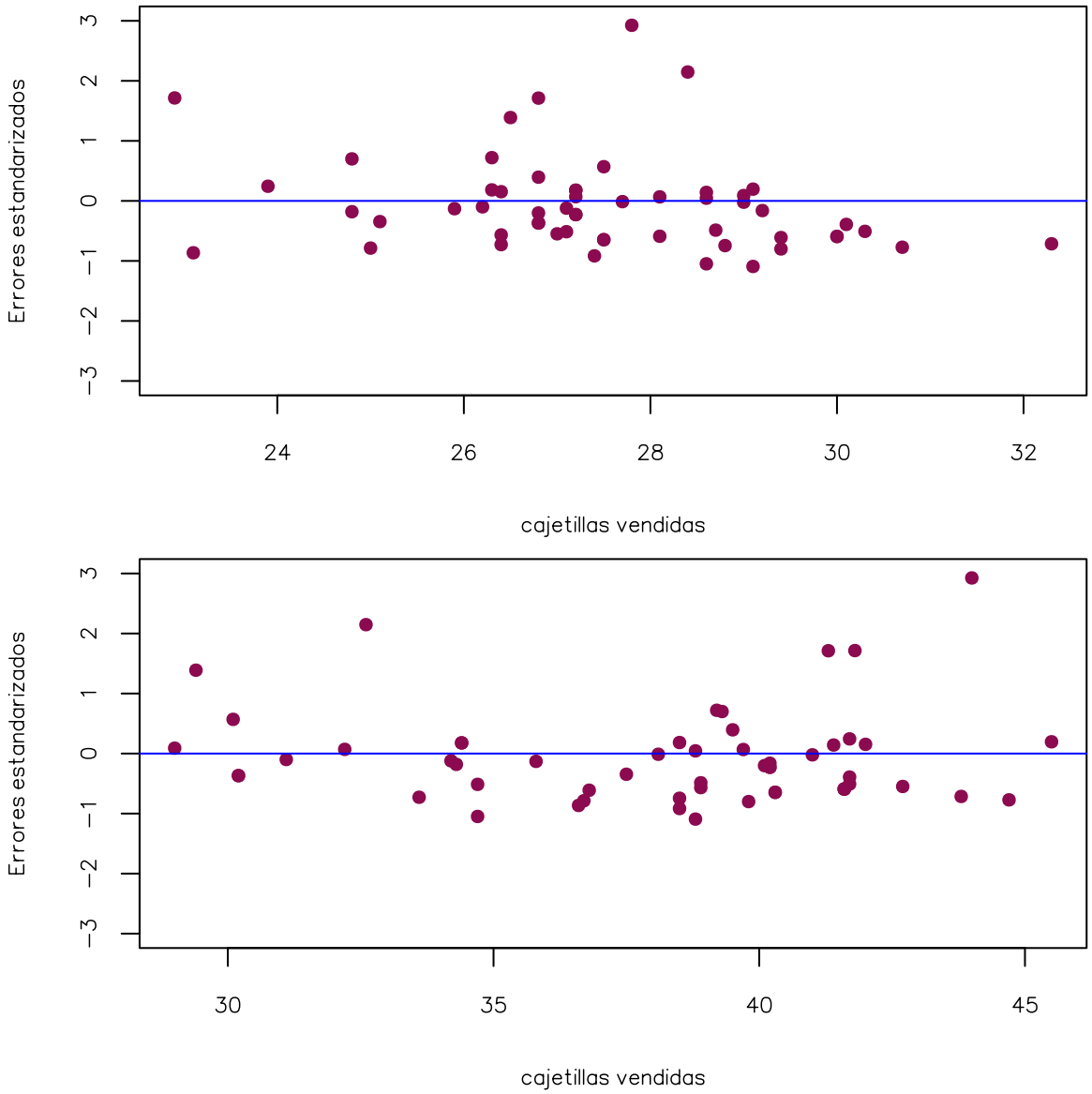
```
lillie.test(rstandard(modelo_refinado))
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  rstandard(modelo_refinado)
D = 0.23318, p-value = 4.065e-08
```

Como el valor del $p - value$ es menor al nivel de significancia $\alpha = 0.05$ entonces rechazamos H_0, es decir nuestros residuales no tienen distribución normal.

Supuesto de Linealidad

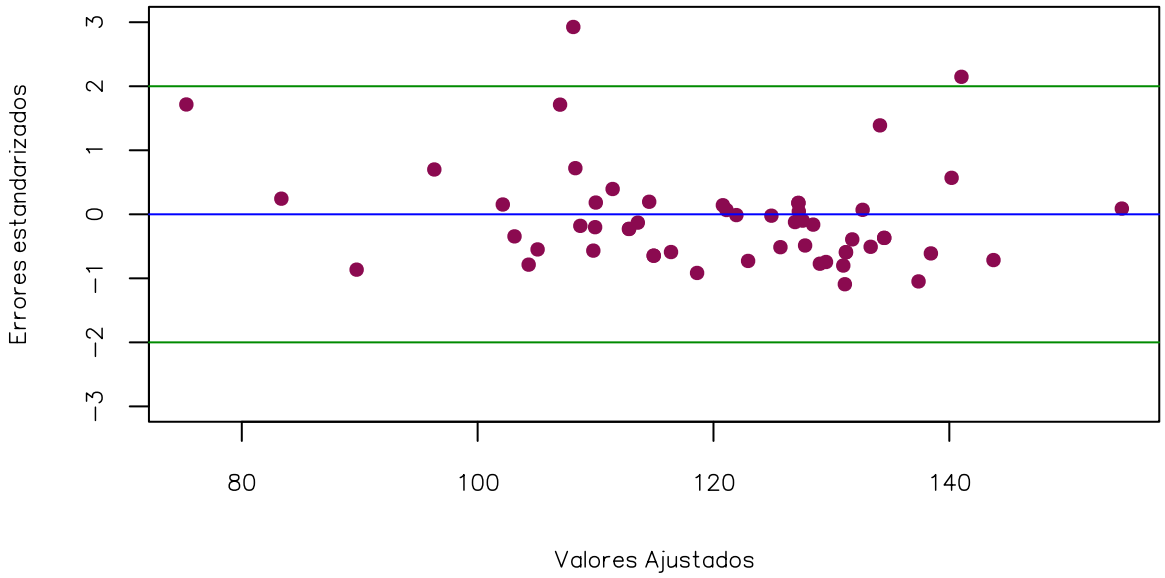
Como se menciona en el capítulo, graficaremos los errores estandarizados contra los valores observados de la variable explicativa.



Para ambas variables, salvo por la posible presencia de algunos valores atípicos que salen de la franja horizontal $(-2, 2)$, el resto de las observaciones parecen distribuirse como ruido blanco.

Supuesto de Homocedasticidad

Se dice que una muestra es homocedástica cuando la varianza es constante a lo largo de todas las observaciones, es decir, no varia conforme se presentan nuevas observaciones.



- Si la varianza es constante entonces la gráfica fluctuaráentre el eje horizontal de manera simétrica, y sin seguir algún patrón, y se espera que la mayor parte de los errores estén contenidos en franjas horizontales delimitados por el eje entre -2 y 2. En éste ejemplo la dispersión regular de los residuales dentro de las Bandas superior e inferior y que no haya residuales que se alejen tanto de la Banda 0, indican varianza constante.

Adicionalmente aplicaremos las pruebas vistas en el capítulo para tener certeza estadística de la validez del supuesto de homocedasticidad.

Prueba White

```
dataset=data.frame(cigarros$AGE,cigarros$PRICE,cigarros$SCIG)
mode1=VAR(dataset,p=1)
whites.htest(mode1)
```

```
White's Test for Heteroskedasticity:
====================================

 No Cross Terms

 H0: Homoskedasticity
 H1: Heteroskedasticity

 Test Statistic:
 39.1790

 Degrees of Freedom:
 36

 P-value:
 0.3291
```

Por el $p - value$, la hipótesis de homocedasticidad no se rechaza.

Supuesto de Multicolinealidad

```
X1=scale(cigarros[,-5])

A=t(X1)%*%X1

kappa=max(eigen(A)$values)/min(eigen(A)$values)
kappa
```

```
[1] 3.151592
```

El coeficiente *kappa* es de 3.15 por lo tanto no tenemos problemas de multicolinealidad.

Capítulo 32

Apéndice

Será conveniente desarrollar algunas variantes en la forma en la que se denota a los residuales, para ello se define a la matriz H como $H = X(X'X)^{-1}X'$,es conocida como **"matriz sombrero"**, que junto con la matriz $(I - H)$ cumplen con ser matrices idempotentes, es decir, que al elevar las matrices a una potencia dada los valores contenidos en la matriz no se modifican; de igual forma ambas matrices cumplen con ser simétricas, denominadas así ya que al transponer las matrices los valores contenidos en ellas conservan su lugar.

Debemos considerar el siguiente resultado, el cual será importante al desarrollar el siguiente teorema A ya que demuestra que $(X'X)^{-1}$ es una matriz simétrica.

$$\begin{aligned}[(X'X)^{-1}]' &= [(X'X)']^{-1} \\ &= (X'(X')')^{-1} \\ \therefore [(X'X)^{-1}]' &= (X'X)^{-1}.\blacksquare\end{aligned}$$

Es decir, la inversa de $X'X$ es simétrica, resultado importante en el siguiente teorema:

Teorema A Sea $H = X(X'X)^{-1}X'$ e $(I - H)$ entonces:

a) Las matrices H e $I - H$ son idempotentes.

b) Las matrices H e $I - H$ son simétricas.

Demostración:

a) Para demostrar la idempotencia de H basta probar que $H^2 = H$, es decir, al elevar la matriz H ésta no se alterará:

$$\begin{aligned}H^2 &= (X(X'X)^{-1}X')(X(X'X)^{-1}X') \\ &= X(X'X)^{-1}X'X(X'X)^{-1}X'.\end{aligned}$$

Tranponiendo con la finalidad de simplificar el producto matricial y por el resultado mostrado anteriormente $[(X'X)^{-1}]' = (X'X)^{-1}$ se tiene:

$$\begin{aligned}&= [(X'X)^{-1}X'X(X'X)^{-1}X']'X' \\ &= [(X'X)^{-1}X']'X' \\ &= X(X'X)^{-1}X'\end{aligned}$$

$$\therefore H^2 = H.$$

Por lo tanto H es idempotente. $\blacksquare$

Para probar la idempotencia de $I - H$, ésta será elevada al cuadrado.

$$\begin{aligned}(I-H)^2 &= (I-H)(I-H)\\ &= I - IH - IH + H^2\\ &= I - 2H + H^2.\end{aligned}$$

Por idempotencia de H, $H = H^2$. Por lo tanto:

$$(I-H) = I - 2H + H$$

$$\therefore (I-H)^2 = I - H.$$

b) Para demostrar la simetría de H, se transpondrá la matriz H. Además debemos recordar que $[(X'X)^{-1}]' = (X'X)^{-1}$ así:

$$\begin{aligned}H' &= (X(X'X)^{-1}X')'\\ &= X(X'X)^{-1}X'\\ &\therefore H' = H.\end{aligned}$$

Por lo tanto la matriz H es simétrica.

Para la simetría de $I - H$ se transpone la matriz:

$$(I-H)' = I' - H'.$$

Por simetría de H y de I

$$\therefore (I-H)^2 = I - H$$

Por lo tanto $I - H$ es simétrica. ■

Corolario A Sea $\underline{e}$ la matriz de residuales entonces estos pueden ser expresados por la siguiente ecuación:

$$\underline{e} = (I-H)\underline{Y}$$

donde I es la matriz identidad, y $H = X(X'X)^{-1}X'$.

Demostración:

Se sabe que los valores estimados son calculados de la siguiente manera:

$$\begin{aligned}\underline{\hat{Y}} &= X\underline{\hat{\beta}}\\ \underline{\hat{Y}} &= X(X'X)^{-1}X'\underline{Y}\\ \underline{\hat{Y}} &= H\underline{Y}.\end{aligned}$$

donde $H = X(X'X)^{-1}X'$. De esta manera calculando la matriz de residuales se tiene:

$$\begin{aligned}\underline{e} &= \underline{Y} - \underline{\hat{Y}}\\ \underline{e} &= \underline{Y} - X\underline{\hat{\beta}}\\ \underline{e} &= \underline{Y} - H\underline{Y}\\ \underline{e} &= (I-H)\underline{Y}.\blacksquare\end{aligned}$$

Teorema B Sea una variable de interés $\underline{Y}$, llamada **dependiente**, relacionada con dos o más variables explicativas$x_1, x_2, \ldots, x_k$, entonces:

a) $\mathbf{E}[\underline{Y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$.

b) $Var(\underline{Y}) = \sigma^2$.

Demostración:

a) Para la esperanza de $\underline{Y}$ se tiene:

$$\mathbf{E}[\underline{Y}] = \mathbf{E}[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon].$$

La estimación es sobre $\underline{Y}$, como $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son constantes; $x_1, x_2, \ldots, x_k$ son los valores dados, por lo que:

$$\mathbf{E}[\underline{Y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \mathbf{E}[\epsilon].$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + 0$$
$$\therefore \mathbf{E}[\underline{Y}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k. \blacksquare$$

b) Para la varianza de $\underline{Y}$ se tiene:

$$Var(\underline{Y}) = Var\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon\right).$$

La estimación es sobre $\underline{Y}$, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son constantes;$x_1, x_2, \ldots, x_k$ son valores dados, por lo que cumple que:

$$Var(\underline{Y}) = 0 + 0 + 0 + \ldots + 0 + Var(\epsilon)$$
$$\therefore Var(\underline{Y}) = \sigma^2. \blacksquare$$

Bibliografía

Conover, W. J. (1998). *Practical nonparametric statistics*, volume 350. John Wiley & Sons.

Omar, R. T. (2019). *Estadística no Paramétrica y Análisis de Regresión. Una introducción para estudiantes de la licenciatura de actuaría.* UNAM.