Regresión Lineal Simple - Semana 01

Johnatan Cardona Jiménez
jcardonj@unal.edu.co
Profesor Asistente - Escuela de Estadística
Universidad Nacional de Colombia, Sede Medellín

Semestre 02-2024

Introducción

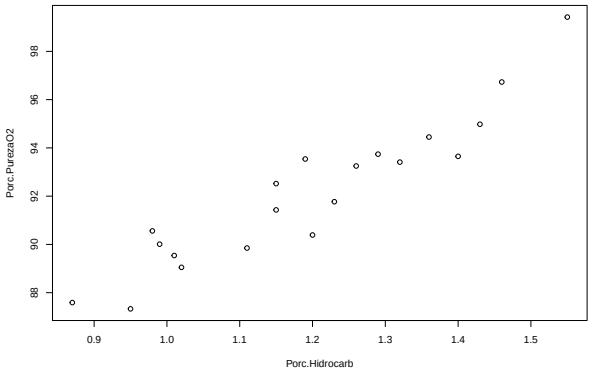
En muchas ocasiones es posible diseñar experimentos estadísticos controlados, en los cuáles es factible el estudio simultáneo de varios factores, aplicando procedimientos de aleatorización apropiados, en lo que se conoce como diseño y análisis de experimentos.

Sin embargo, en muchas ocasiones sólo se cuenta con un conjunto de datos sobre los cuáles es difícil esperar que hayan sido observados en condiciones estrictamente controladas, y de los cuáles también en pocas ocasiones se tienen réplicas para calcular el error experimental.

Cuando se enfrenta la situación anterior un camino apropiado es aplicar los **métodos de regresión**, que permiten establecer asociaciones entre variables de interés, donde la relación usual no es necesariamente de causa - efecto. En principio, consideramos una asociación lineal entre una variable respuesta Y y una variable predictora X.

Ejemplo

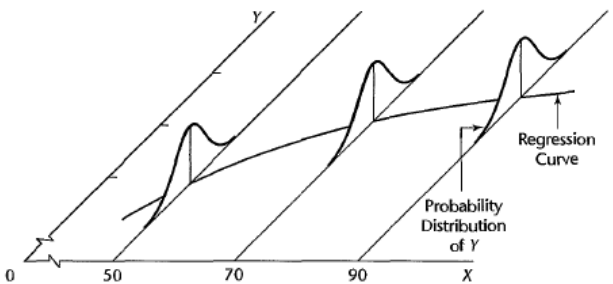
Se tienen datos de un proceso de destilación química donde se desea establecer la relación entre la pureza del oxígeno producido (Y, en %) y el porcentaje de hidrocarburos presentes en el condensador principal de la unidad de destilación (X). Veamos una gráfica de dispersión de los datos:



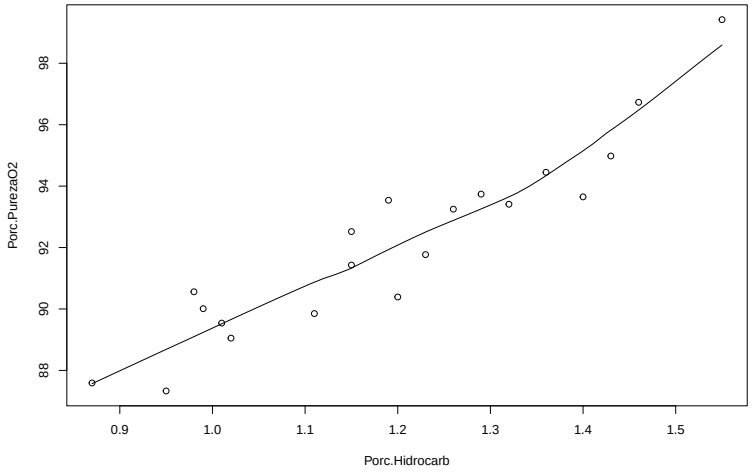
Significados de la regresión lineal

La regresión lineal tiene dos significados:

1. (***Enfoque probabilístico - Método de máxima verosimilitud***) Podemos verla a partir de la distribución conjunta de las variables X e Y, en la cual podemos definir la distribución condicional de $Y|X$, esto es $f(Y|X)$, y determinar $E(Y|X)$. En este caso la regresión pretende ajustar la curva correspondiente a $E(Y|X)$.



2. (***Enfoque no probabilístico - Método de mínimos cuadrados***) Dado un conjunto de pares de datos (X, Y), puede asumirse una forma funcional para la curva de regresión y tratar de ajustarla a los datos minimizando el error de ajuste.

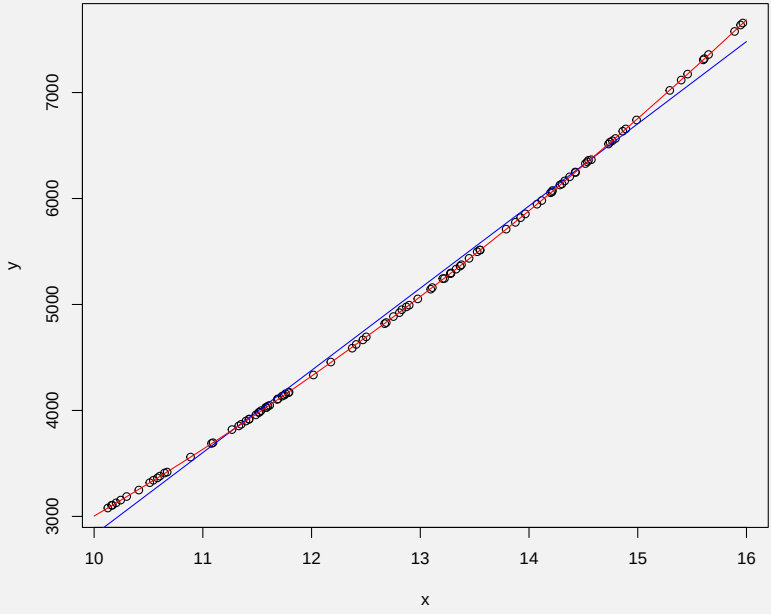


Supuestos bajo un enfoque probabilístico

- La variable respuesta Y es una variable aleatoria cuyos valores se observan mediante la selección de los valores de la variable predictora X en un intervalo de interés (región de diseño o región de observación).
- Por lo anterior, la variable predictora X no es considerada como variable aletatoria, sino como un conjunto de valores fijos que representan los puntos de observación, que se seleccionan con anticipación y se miden sin error.
  Sin embargo, si esto último no se cumple, el método de estimación de mínimos cuadrados ordinarios para los parámetros del modelo de regresión puede seguir siendo válidos si los errores en los valores de la variable predictora son pequeños en comparación con los errores aleatorios del modelo ε_i.

- Los datos observados (x_i, y_i), $i = 1, \ldots, n$, constituyen una muestra representativa de un medio acerca del cual se desea generalizar.
- **El modelo de regresión es lineal en los parámetros. Es decir, ningún parámetro de la regresión aparece como el exponente o es dividido o multiplicado por otro parámetro.**
  Sin embargo, la línea de ajuste puede tener una curvatura (no ser lineal en X y/o en Y). Cuando no es lineal en Y, mediante una transformación conveniente es posible aplicar las técnicas de regresión lineal sobre esta nueva variable.

```r
#Modelo simulado
n = 100; #Numero de observaciones
x = runif(n, 10, 16); #Simular variable predictora
x2 = x^2
#Intercepto, pendiente y desviacion estandar
beta0 = 4; beta1 = 30; sigma = 4
#Modelo verdadero: Cuadratico
y = beta0 + beta1*x^2 + rnorm(n, 0, sigma)
modelo1 = lm(y ~ x)
modelo2 = lm(y ~ x2)
para1 = modelo1$coef
para2 = modelo2$coef
plot(x, y)
curve(para1[1] + x*para1[2], 10, 16, col = "blue",
      add = TRUE)
curve(para2[1] + x^2*para2[2], 10, 16, col = "red",
      add = TRUE)
```


y
7000
6000
5000
4000
3000
10
11
12
13
14
15
16
x

- Los errores aleatorios $\varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \ldots, n$.
- Los errores aleatorios ε_i son estadísticamente independientes.
  Por tanto:
  $$COV(\varepsilon_i, \varepsilon_j) = 0, \forall_{i \neq j}, \quad COV(Y_i, Y_j) = 0, \forall_{i \neq j}.$$
- La varianza de los errores aleatorios es $\sigma^2, \forall_{i=1,2,\ldots,n}$ (supuesto de varianza constante, pero desconocida).
  Dado que los valores X_i de la variable predictora no son considerados aleatorios y que los errores son independientes, la varianza de los Y_i también es $\sigma^2, \forall i$ y por tanto este parámetro es independiente del punto de observación (es decir, del valor de X_i).

En el caso que este último supuesto no pueda aplicarse, entonces el método de regresión empleado será el de mínimos cuadrados ponderados.

En resumen, los supuestos del error en el modelo de regresión lineal simple se pueden expresar como:

$$\varepsilon_i \overset{\text{iid.}}{\sim} N(0\ ,\ \sigma^2), i = 1, 2, \ldots, n$$

donde, iid. es la abreviación de independiente e idénticamente distribuido.

Estos supuestos tienen como consecuencia directa en la respuesta que:

$$Y_i | X_i \overset{\text{ind.}}{\sim} N\Big(\beta_0 + \beta_1 X_i\ ,\ \sigma^2\Big)$$

donde, ind. es la abreviación de independiente distribuido.

Nomenclatura Modelo de Regresión Lineal Simple

- Y: Variable respuesta o variable dependiente.
- X: Variable predictora, variable independiente o variable regresora.
- ε: Error aleatorio
- β_0, β_1, σ^2: Parámetros de la regresión. β_0 es el intercepto, β_1 es la pendiente de la línea recta y σ^2 es la varianza.
- $\widehat{\beta}_0$: Estimador del parámetro β_0.
- $\widehat{\beta}_1$: Estimador del parámetro β_1.
- $\hat{\sigma}^2$: Estimador del parámetro σ^2.
- e: Residual, es una estimación del error aleatorio.
- $\widehat{Y}$: Es la estimación de $E(Y|X)$ ó $\mu_{Y|X}$.

Estimación por mínimos cuadrados ordinarios (MCO)

- Para una selección preliminar de la variable predictora en un modelo de regresión simple es conveniente realizar el diagrama de dispersión (Y vs. X) y mirar si existe una tendencia funcional (lineal, polinomial, etc) en la nube de puntos.
- Si la nube de puntos parece mejor ajustada por una curva hay que buscar una transformación apropiada en X y/o Y;
- Debe tenerse claro que el método de mínimos cuadrados es un método numérico de ajuste de curvas, y no un método estadístico. La estadística opera a partir de los supuestos distribucionales asignados en el modelo de regresión.

Método de MCO

El objetivo del método de MCO es obtener estimaciones de los parámetros de regresión, es decir hallar valores de β_0 y β_1 que minimicen la suma de los cuadrados de los errores $S(\beta_0, \beta_1)$ definida a partir de (2.1) como:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \Big[Y_i - (\beta_0 + \beta_1 X_i) \Big]^2$$

A los valores que minimizan esta expresión se les conoce como estimadores de mínimos cuadrados y se les denota $\widehat{\beta}_0$ y $\widehat{\beta}_1$.

*** Nota:** ***En este proceso de estimación no aparece el parámetro σ^2, pues éste hace parte de los supuestos distribucionales, los cuales no son necesarios en la aplicación de MCO.***

Valor de los estimadores MCO

Dados los pares de observaciones $(x_1, y_1), \ldots, (x_n, y_n)$, hallar β_0 y β_1 que minimicen a $S(\beta_0, \beta_1)$ implica resolver el siguiente sistema de ecuaciones:

$$\left.\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0}\right|_{\widehat{\beta}_0, \widehat{\beta}_1} = 0$$

$$\left.\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1}\right|_{\widehat{\beta}_0, \widehat{\beta}_1} = 0$$

De lo cual surgen las denominadas ecuaciones normales:

$$\sum_{i=1}^{n} y_i = n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Y de éstas se obtiene que las estimaciones por mínimos cuadrados de los parámetros son:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i\, y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} y_i\,(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Sumas de cuadrados y de productos cruzados

- Suma de cuadrados corregidos en x:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

- Suma de cuadrados corregidos en y:

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

- Suma de productos cruzados corregidos:

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

NOTA: $\widehat{\beta}_1$ puede ser expresado en función de S_{xy} y de S_{xx} así:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Estimación por máxima verosimilitud (ML)

- El método de mínimos cuadrados produce estimadores lineales insesgados para los parámetros de la recta y puede ser usado para la estimación de parámetros de un modelo de regresión lineal sin consideraciones distribucionales sobre los errores.
- Sin embargo, para poder aplicar pruebas de hipótesis y construir intervalos de confianza (procedimientos inferenciales), es necesario asumir supuestos distribucionales los cuales deben ser posteriormente validados. Considerando para el modelo de regresión lineal simple los supuestos de normalidad, independencia y varianza constante para los errores, podemos usar el método de **estimación de máxima verosimilitud (MLE)**.

Sean $(x_1, y_1), \ldots, (x_n, y_n)$ los n pares de datos observados, entonces el modelo de regresión lineal simple es:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \ i = 1, 2, \ldots, n.$$

A la variable aleatoria ε_i, se le asignan los siguientes supuestos distribucionales:

$$\varepsilon_i \overset{\text{iid}}{\sim} N\left(0\ ,\ \sigma^2\right),\ i = 1, 2, \ldots, n,$$

Con base en lo anterior y asumiendo que los niveles o valores en que X es observada son fijos, se obtiene que

$$Y_i|X_i \overset{\text{ind.}}{\sim} N\Big(E\left[Y_i|X_i\right]\ ,\ \sigma^2\Big)$$

con

$$E\left[Y_i|X_i\right] = \beta_0 + \beta_1 X_i$$

Sean $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ y $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, entonces la función de verosimilitud $L\left(\beta_0, \beta_1, \sigma^2 \middle| \mathbf{x,y}\right)$ se define a partir de la densidad conjunta de las observaciones, $f\left(y_1, \ldots, y_n \middle| \beta_0, \beta_1, \sigma^2\right)$, que por la condición de independencia es igual al producto de las densidades de probabilidad marginales. Por tanto, podemos escribir,

$$
\begin{aligned}
L\left(\beta_0, \beta_1, \sigma^2 \;\middle|\; \mathbf{x,y}\right) &= f\left(y_1, \ldots, y_n \;\middle|\; \beta_0, \beta_1, \sigma^2\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \\
&= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]
\end{aligned}
$$

El objetivo es hallar los parámetros desconocidos $\beta_0, \beta_1, \sigma^2$, que maximicen L, o equivalentemente, que maximicen $\ell = \ln L$ (el logaritmo natural de L).

$$\ell = \ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Observe que para cualquier valor de σ^2 fijo, ℓ es maximizada como una función de β_0 y β_1 por aquellos valores $\widetilde{\beta}_0$ y $\widetilde{\beta}_1$ que minimizan $S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ y así, los estimadores MLE $\widetilde{\beta}_0$ y $\widetilde{\beta}_1$ son iguales a los respectivos estimadores de mínimos cuadrados, $\widehat{\beta}_0$ y $\widehat{\beta}_1$.

Para hallar el estimador MLE para σ^2 substituimos $\widehat{\beta}_0$ y $\widehat{\beta}_1$ en $\ln L$, y hallamos σ^2 que maximiza

$$-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

de donde obtenemos como estimador MLE de σ^2 a

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Resumiendo, bajo el modelo de regresión lineal normal, es decir, con errores independientes e idénticamente distribuidos $N(0, \sigma^2)$, los estimadores de mínimos cuadrados para β_0 y β_1 coinciden con los estimadores de máxima verosimilitud. Ahora con las estimaciones obtenidas podemos desarrollar procedimientos inferenciales: intervalos de confianza y procedimientos de prueba de hipótesis.

También puede demostrarse que los estimadores MLE insesgados son de mínima varianza cuando son comparados con todos los posibles estimadores lineales insesgados. Además son consistentes, es decir, a medida que aumenta el tamaño de muestra, la diferencia entre éstos y los respectivos parámetros se aproxima a cero.

Estimación de la varianza σ^2

Puede demostrarse que bajo los supuestos del modelo en relación a los errores, la esperanza del estimador de máxima verosimilitud de σ^2 es:

$$E\left[\widetilde{\sigma}^2\right] = \left(\frac{n-2}{n}\right)\sigma^2,$$

por tanto $\widetilde{\sigma}^2$ no es un estimador insesgado de σ^2, aunque si es asintóticamente insesgado, esto es, $\lim_{n\to\infty} E\left[\widetilde{\sigma}^2\right] = \sigma^2$. Sin embargo, a partir de $\widetilde{\sigma}^2$ se puede obtener un estimador insesgado de la varianza, así:

$$\widehat{\sigma}^2 = \left(\frac{n}{n-2}\right)\widetilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2}$$

que cumple $E\left[\widehat{\sigma}^2\right] = \sigma^2$.

Ecuación de regresión ajustada y residuales del modelo

Al tener estimados los parámetros del modelo de regresión lineal simple (por mínimos cuadrados o máxima verosimilitud), entonces se puede realizar una estimación de la respuesta media $E[Y|X] = \mu_{Y|X}$, a través del modelo ajustado, así:

$$\widehat{\mu}_{Y|x_i} = \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \bar{y} + (x_i - \bar{x})\,\widehat{\beta}_1.$$

A esta ecuación se le conoce como la ecuación de regresión ajustada, que en este caso corresponde a una recta ajustada.

A las diferencias entre los valores observados de la respuesta y_i y los valores ajustados por el modelo de regresión $\widehat{y}_i$ (obtenidos de la ecuación de regresión ajustada) se les conoce como los residuales del modelo. Esto es, $e_i = y_i - \widehat{y}_i$ es el i-ésimo residual del modelo, que es una estimación del i-ésimo error aleatorio, ε_i.

Los residuales del modelo tienen gran importancia ya que ellos determinan que tan bueno fue el ajuste del modelo y permitirán más adelante realizar las validaciones de los supuestos realizados sobre los errores aleatorios.

Propiedades de los estimadores de Máxima Verosimilitud

Bajo los supuestos considerados respecto a los errores tenemos que:

1. $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son combinaciones lineales de las variables aleatorias $Y_1, \ldots, Y_n$, pues estos pueden escribirse como:

$$\widehat{\beta}_0 = \sum_{i=1}^{n} m_i \, Y_i,$$

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i \, Y_i,$$

donde:

$$m_i = \frac{1}{n} - \bar{x}\, c_i$$

$$c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

Se puede demostrar a través de cálculos directos que:

$$\sum_{i=1}^{n} c_i = 0, \qquad \sum_{i=1}^{n} c_i\, x_i = 1,$$

$$\sum_{i=1}^{n} m_i = 1, \qquad \sum_{i=1}^{n} m_i\, x_i = 0,$$

$$\sum_{i=1}^{n} c_i^2 = \frac{1}{S_{xx}}, \qquad \sum_{i=1}^{n} m_i^2 = \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}.$$

Además, como $Y_1, \ldots, Y_n$ son variables normales e incorrelacionadas, entonces $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son variables aleatorias normales.

2. El valor esperado de los estimadores, es:

$$E\left[\widehat{\beta}_0\right] = E\left[\sum_{i=1}^{n} m_i\, Y_i\right] = \sum_{i=1}^{n} m_i\, E\left[Y_i\right]$$
$$= \sum_{i=1}^{n} m_i(\beta_0 + \beta_1 x_i)$$
$$= \beta_0 \sum_{i=1}^{n} m_i + \beta_1 \sum_{i=1}^{n} m_i\, x_i = \beta_0$$

$$E\left[\widehat{\beta}_1\right] = E\left[\sum_{i=1}^{n} c_i\, Y_i\right] = \sum_{i=1}^{n} c_i\, E\left[Y_i\right]$$
$$= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i\, x_i = \beta_1$$

3. La varianza de los estimadores, es:

$$V\left[\widehat{\beta}_0\right] = V\left[\sum_{i=1}^{n} m_i\, Y_i\right] = \sum_{i=1}^{n} m_i^2\, V\left[Y_i\right]$$
$$= \sum_{i=1}^{n} m_i^2\, \sigma^2$$
$$= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n S_{xx}}$$

$$V\left[\widehat{\beta}_1\right] = V\left[\sum_{i=1}^{n} c_i\, Y_i\right] = \sum_{i=1}^{n} c_i^2\, V\left[Y_i\right]$$
$$= \sum_{i=1}^{n} c_i^2 \sigma^2$$
$$= \frac{\sigma^2}{S_{xx}}$$

4. La varianza de la respuesta ajustada en un valor dado $X = x_i$, es:

$$
\begin{aligned}
V\left[\widehat{Y}_i\right] &= V\left[\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right] \\
&= V\left[\sum_{j=1}^{n}(m_j + x_i\, c_j)\, Y_j\right] \\
&= \sum_{j=1}^{n}(m_j + x_i\, c_j)^2\, V(Y_j) \\
&= \sigma^2 \sum_{j=1}^{n}\left[\frac{1}{n} + (x_i - \bar{x})c_j\right]^2 \\
&= \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]
\end{aligned}
$$

5. La covarianza (cov) entre los estimadores de los parámetros es:

$$\begin{aligned}
\operatorname{cov}\left[\widehat{\beta}_0\ ,\ \widehat{\beta}_1\right] &= \operatorname{cov}\left[\sum_{i=1}^{n} m_i\, Y_i\ ,\ \sum_{i=1}^{n} c_i\, Y_i\right] \\
&= \sum_{i=1}^{n} m_i\, c_i \operatorname{cov}\left[Y_i\ ,\ Y_i\right] + \sum_{i=1}^{n}\sum_{j\neq i}^{n} m_i\, c_j \operatorname{cov}\left[Y_i\ ,\ Y_j\right] \\
&= \sum_{i=1}^{n} m_i\, c_i\, V\left[Y_i\right] \\
&= \sigma^2 \sum_{i=1}^{n} m_i\, c_i \\
\operatorname{cov}\left[\widehat{\beta}_0\ ,\ \widehat{\beta}_1\right] &= -\frac{\sigma^2 \bar{x}}{S_{xx}}
\end{aligned}$$

6. La covarianza entre la variable respuesta y su correspondiente estimador en un valor dado $X = x_i$ es:

$$\begin{aligned}
\text{cov}\left[Y_i\ ,\ \widehat{Y}_i\right] &= \text{cov}\left[Y_i\ ,\ \widehat{\beta}_0 + \widehat{\beta}_1 x_i\right] \\
&= \text{cov}\left[Y_i\ ,\ \sum_{j=1}^{n}(m_j + x_i\, c_j)\, Y_j\right] \\
&= (m_i + x_i\, c_i)\text{cov}\left[Y_i\ ,\ Y_i\right] + \sum_{j\neq i}^{n}(m_j + x_i\, c_j)\text{cov}\left[Y_i\ ,\ Y_j\right] \\
&= \sigma^2(m_i + x_i\, c_i) \\
&= \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]
\end{aligned}$$

7. La suma de los residuales del modelo de regresión con intercepto es siempre cero:

$$\sum_{i=1}^{n} e_i = 0$$

8. La suma de los valores observados y_i es igual a la suma de los valores ajustados $\widehat{y}_i$:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$$

9. La línea de regresión siempre pasa a través del centroide de los datos $(\bar{x}, \bar{y})$.
10. La suma de los residuales ponderados por el correspondiente valor de la variable predictora es cero:

$$\sum_{i=1}^{n} x_i \, e_i = 0$$

11. La suma de los residuales ponderados por el correspondiente valor ajustado es siempre igual a cero:

$$\sum_{i=1}^{n} \widehat{y}_i \, e_i = 0$$

Inferencias sobre los parámetros del modelo de regresión

Inferencias sobre el intercepto β_0

1. **Intervalos de confianza**
   Se puede demostrar que bajo los supuestos del modelo de regresión, se cumple que:

$$T = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\frac{\widehat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n S_{xx}}}} \sim t_{n-2} \tag{2.2}$$

   con t_{n-2} la variable aleatoria t-Student con $n-2$ grados de libertad.

Por tanto un intervalo de confianza del $(1-\alpha)\%$ para β_0 es:

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2} \times \sqrt{\frac{\widehat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}$$

donde $t_{\alpha/2,n-2}$ es el percentil $(1-\alpha/2)$ de la distribución t-Student con $n-2$ grados de libertad.

2. **Prueba de Hipótesis sobre la significancia del intercepto**
   Para probar si β_0 es significativamente distinto de cero:

$$H_0 : \ \beta_0 = 0$$
$$H_1 : \ \beta_0 \neq 0$$

El estadístico de prueba es la ec. (2.2) y el valor observado de éste (T_0) se halla reemplazando β_0 por 0. Se rechaza H_0 si $|T_0| > t_{\alpha/2,n-2}$.

Inferencias sobre la pendiente β_1

1. Intervalos de confianza

Se puede demostrar que bajo los supuestos del modelo de regresión, se cumple que:

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}} \sim t_{n-2} \tag{2.3}$$

Por tanto un intervalo de confianza del $(1-\alpha)\%$ para β_1 es:

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2} \times \sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}$$

2. Prueba de Hipótesis sobre la significancia de la pendiente

Para probar si β_1 es significativamente distinto de cero:

$$H_0: \ \beta_1 = 0$$
$$H_1: \ \beta_1 \neq 0$$

El estadístico de prueba es la ec. (2.3) y el valor observado de éste (T_0) se halla reemplazando β_1 por 0. Se rechaza H_0 si $|T_0| > t_{\alpha/2,n-2}$.

NOTA: Note que si la pendiente es significativa, entonces el modelo de RLS entre la predictora y la respuesta, también lo es.