Clase 3
Regresión Lineal Simple

Análisis Avanzado de Datos

Gabriel Sotomayor

Recordatorio de la clase anterior

Correlación

La correlación mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. La correlación se simboliza con la letra r.

Si tenemos datos de dos variables x e y para n individuos. Los valores para el primer individuo son x_1 e y_1, para el segundo son x_2 e y_2, etc. Las medias y las desviaciones típicas de las dos variables son $\bar{x}$ y s_x para los valores de x, e $\bar{y}$ y s_y para los valores de y. La correlación r entre x e y es:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Características de la correlación (I)

Simetría en las Variables: La correlación no distingue entre variables explicativas y respuesta; es indiferente cuál se llame x o y.

Requisito Cuantitativo: Las dos variables deben ser cuantitativas para que los cálculos de la correlación tengan sentido. No se puede calcular la correlación entre una variable cuantitativa y una categórica.

Independencia de Unidades: Como la correlación utiliza valores estandarizados, no cambia si se modifican las unidades de medida de las variables. La correlación es un valor sin unidades.

Significado del Signo:

- Correlación positiva: Indica una asociación positiva entre las variables.
- Correlación negativa: Indica una asociación negativa.

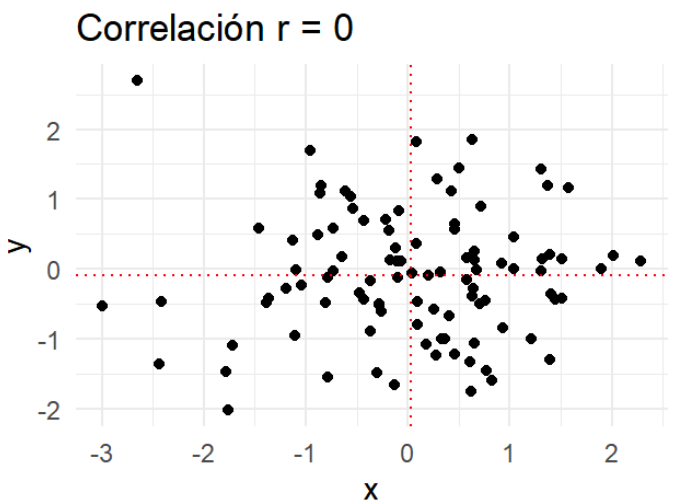
Características de la correlación (II)

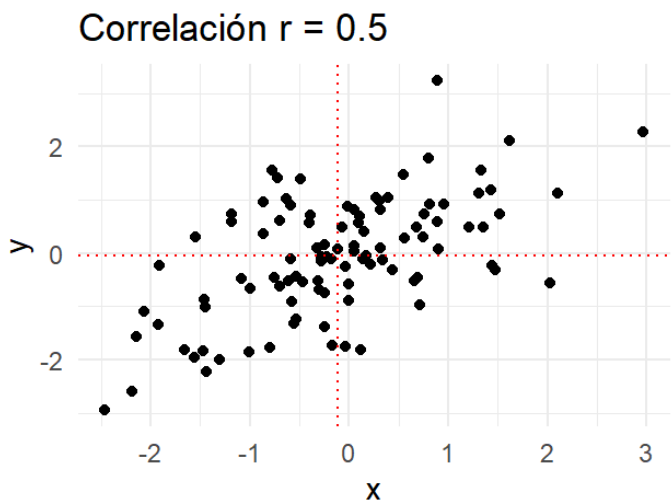
Rango de la Correlación: La correlación siempre toma valores entre −1 y 1.
- Cercanía a 0: Indica una relación lineal débil.
- Cercanía a ±1: Indica una relación lineal fuerte. Un valor de ±1 indica una relación lineal perfecta.

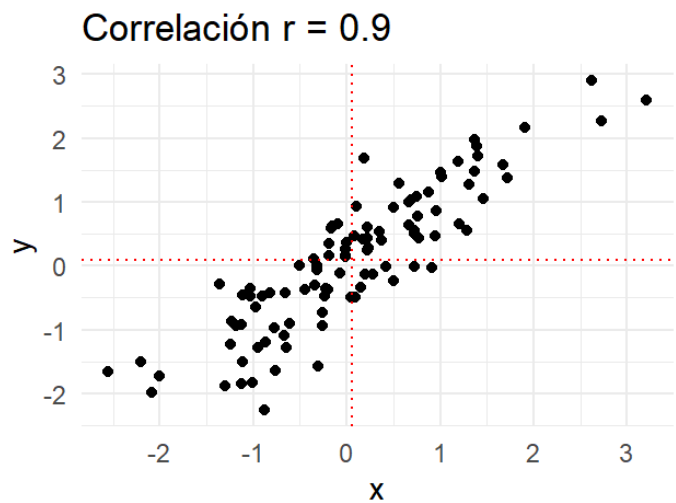
Limitación a Relaciones Lineales: La correlación sólo mide la fuerza de relaciones lineales, no describe adecuadamente las relaciones curvilíneas, aunque estas sean fuertes.

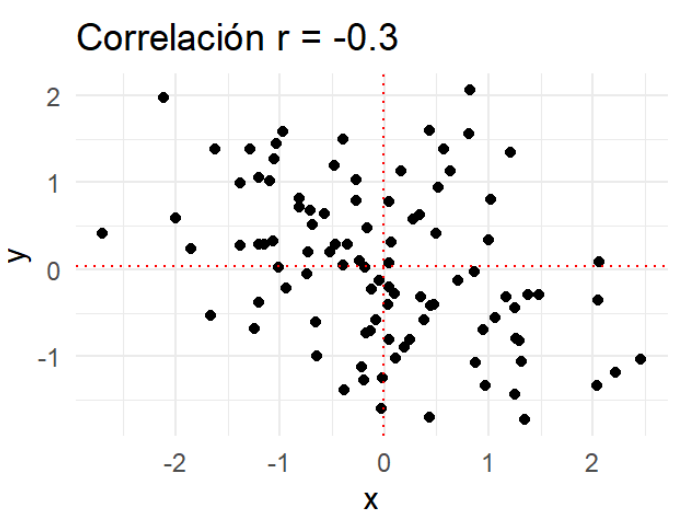
Sensibilidad a Observaciones Atípicas: La correlación puede verse fuertemente afectada por valores atípicos, lo que puede distorsionar la percepción de la relación entre las variables. Es importante utilizar la correlación con precaución cuando se detectan atípicos.

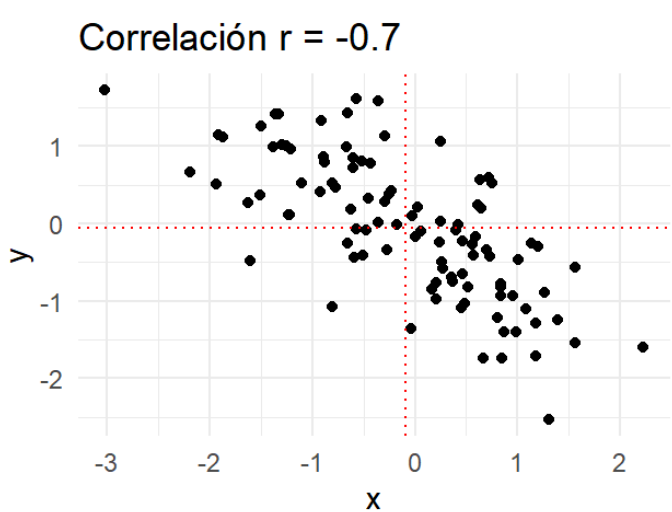
Graficos de dispersión y correlación

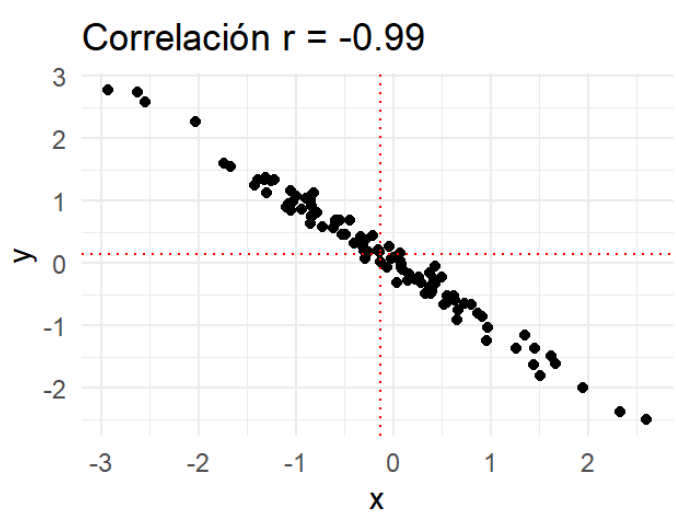












Evaluaciones

Tarea 1: 2 de septiembre (la pauta se subirá hoy)

- Gestión de datos
- Estadística bivariada

Prueba 1: 9 de Septiembre

- Uso de modelos en ciencias sociales
- Estadística bivariada
- Regresión lineal simple

Objetivo de la sesión

Introducir y aplicar la regresión lineal simple para entender la relación entre una variable explicativa y una de respuesta.

Concepto de Regresión Lineal Simple

La regresión lineal simple se utiliza para describir la relación entre dos variables, una independiente (explicativa) y una dependiente (respuesta), mediante una recta de regresión. A diferencia de la correlación si asume una direccionalidad.

Ejemplo: Brecha Salarial de Género y Años de Escolaridad

Consideremos el ejemplo de la brecha salarial de género de una comuna y cómo podría estar influenciada por el promedio de años de escolaridad de la población. La idea es entender cómo varía la brecha salarial en función de los años de escolaridad a través de una recta de regresión.

Medias condicionales

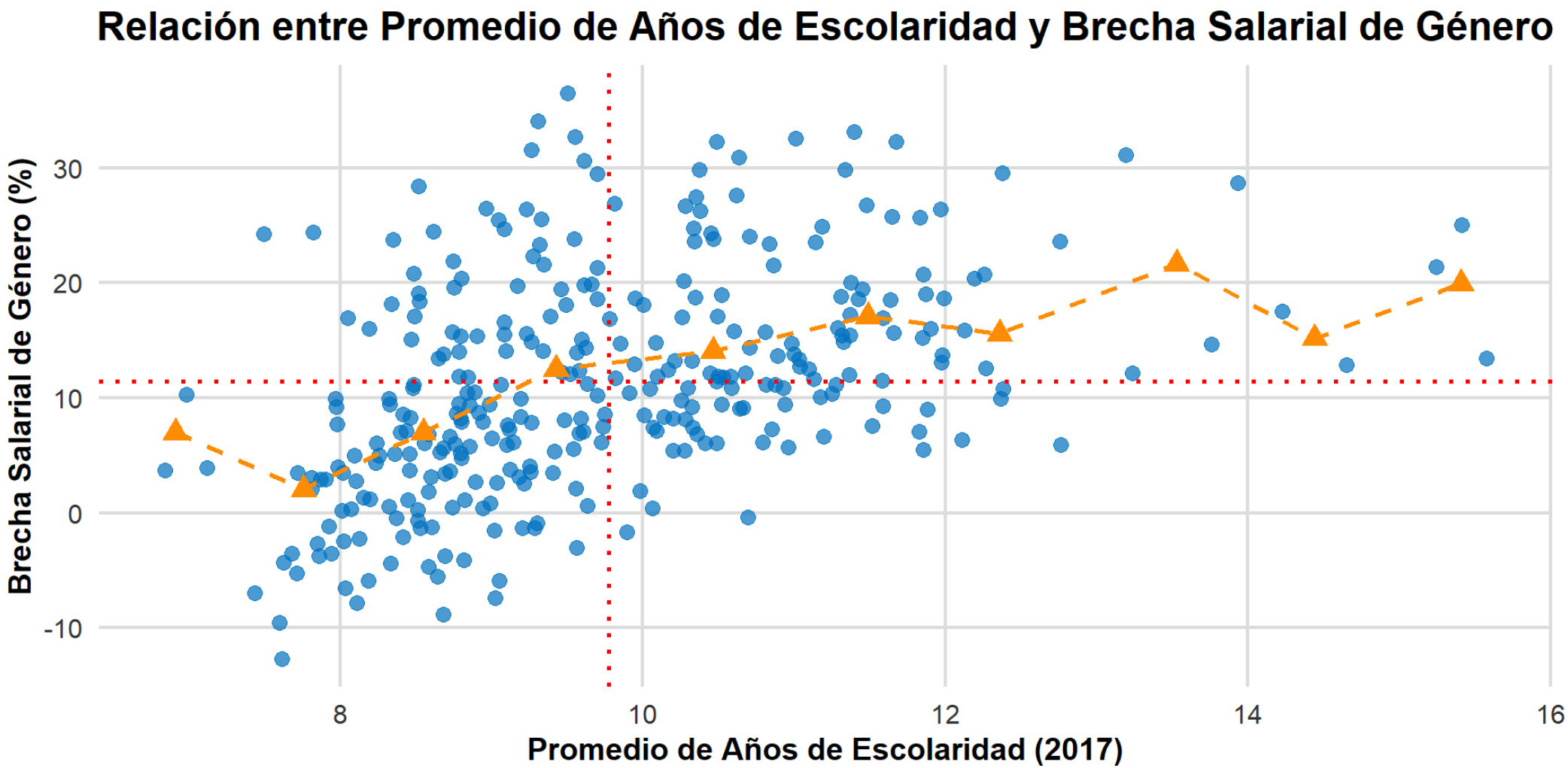
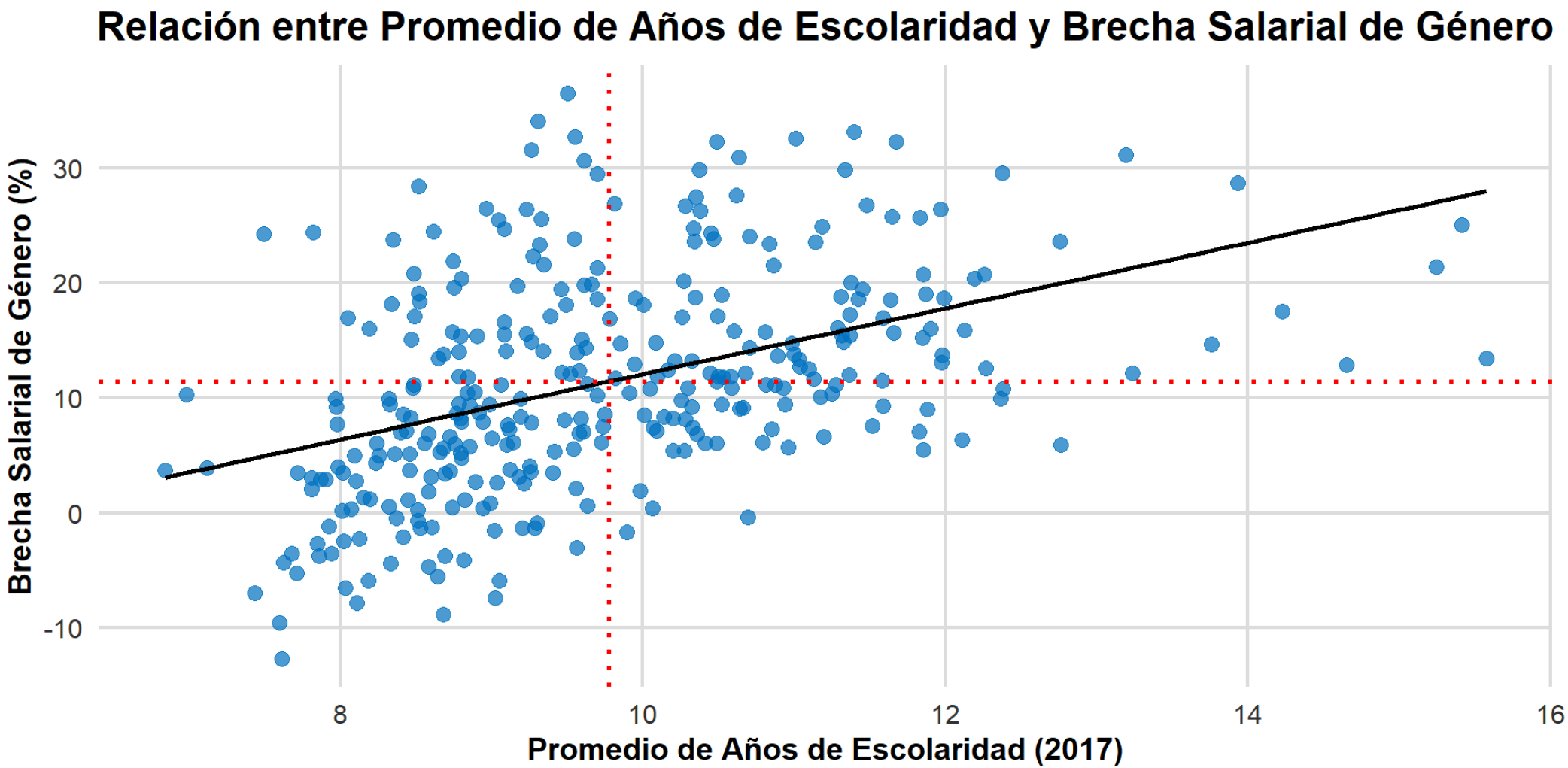


Gráfico de regresión



Recta de Regresión Mínimo-Cuadrática

La recta de regresión mínimo-cuadrática de y con relación a x es la recta que hace que la suma de los cuadrados de las distancias verticales de los puntos observados a la recta sea lo más pequeña posible.

Fórmula General

La recta de regresión se expresa como:

$$\hat{y} = a + bx$$

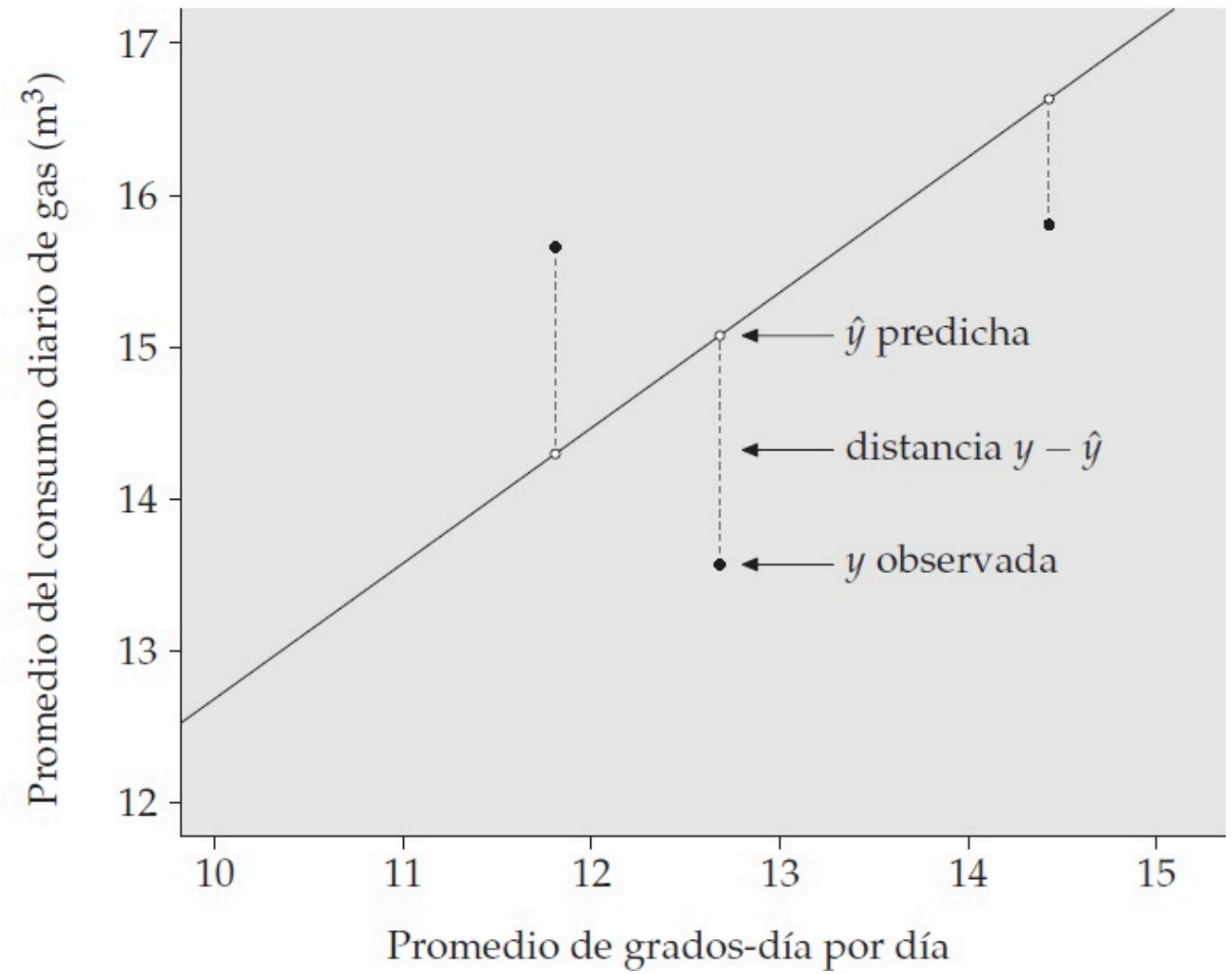
Pendiente b: Indica el cambio promedio en la variable respuesta$y$por cada unidad de cambio en la variable explicativa x.
Ordenada en el origen a: Representa el valor predicho de$y$cuando x= 0. Sólo tiene significado estadístico cuando x toma valores cercanos a 0.

Estimación de mínimos cuadrados

Para cada observación, halla la distancia vertical de cada punto del diagrama de dispersión a la recta. La regresión mínimo-cuadrática hace que la suma de los cuadrados de estas distancias sea lo más pequeña posible.

Ninguna recta podrá pasar exactamente por todos los puntos del diagrama de dispersión (a no ser que haya correlación perfecta). Queremos que las distancias verticales de los puntos a la recta sean lo más pequeñas posible.

Visualización de la distancia entre la recta y los casos



Cálculo de la Pendiente y Ordenada

Para calcular la pendiente by la ordenada en el origen a, se utilizan las siguientes fórmulas:

$$b = r\frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

Donde:

r es la correlación entre x y y.

s_x y s_y son las desviaciones estándar de x y y.

$\bar{x}$ y $\bar{y}$ son las medias de x y y, respectivamente.

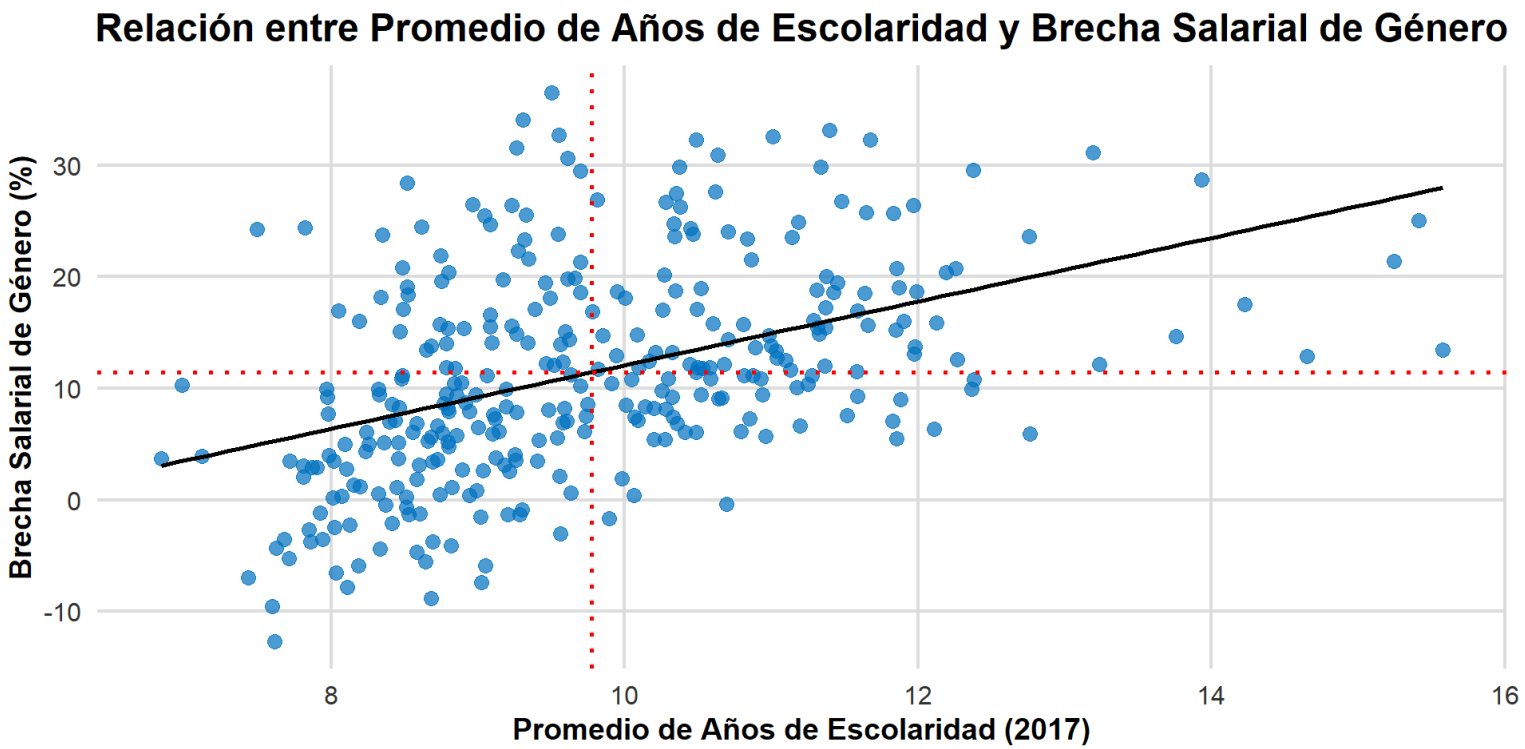
Ejemplo con Datos de Brecha Salarial y Escolaridad

```r
# Cálculo de la pendiente y ordenada
r <- cor(datos$brecha, datos$promedio_anios_escolaridad25_2017)
sx <- sd(datos$promedio_anios_escolaridad25_2017)
sy <- sd(datos$brecha)
mean_x <- mean(datos$promedio_anios_escolaridad25_2017)
mean_y <- mean(datos$brecha)

b <- r * (sy / sx)
a <- mean_y - b * mean_x

paste("Pendiente:", round(b, 2), "| Ordenada:", round(a, 2))
```

```
[1] "Pendiente: 2.85 | Ordenada: -16.45"
```

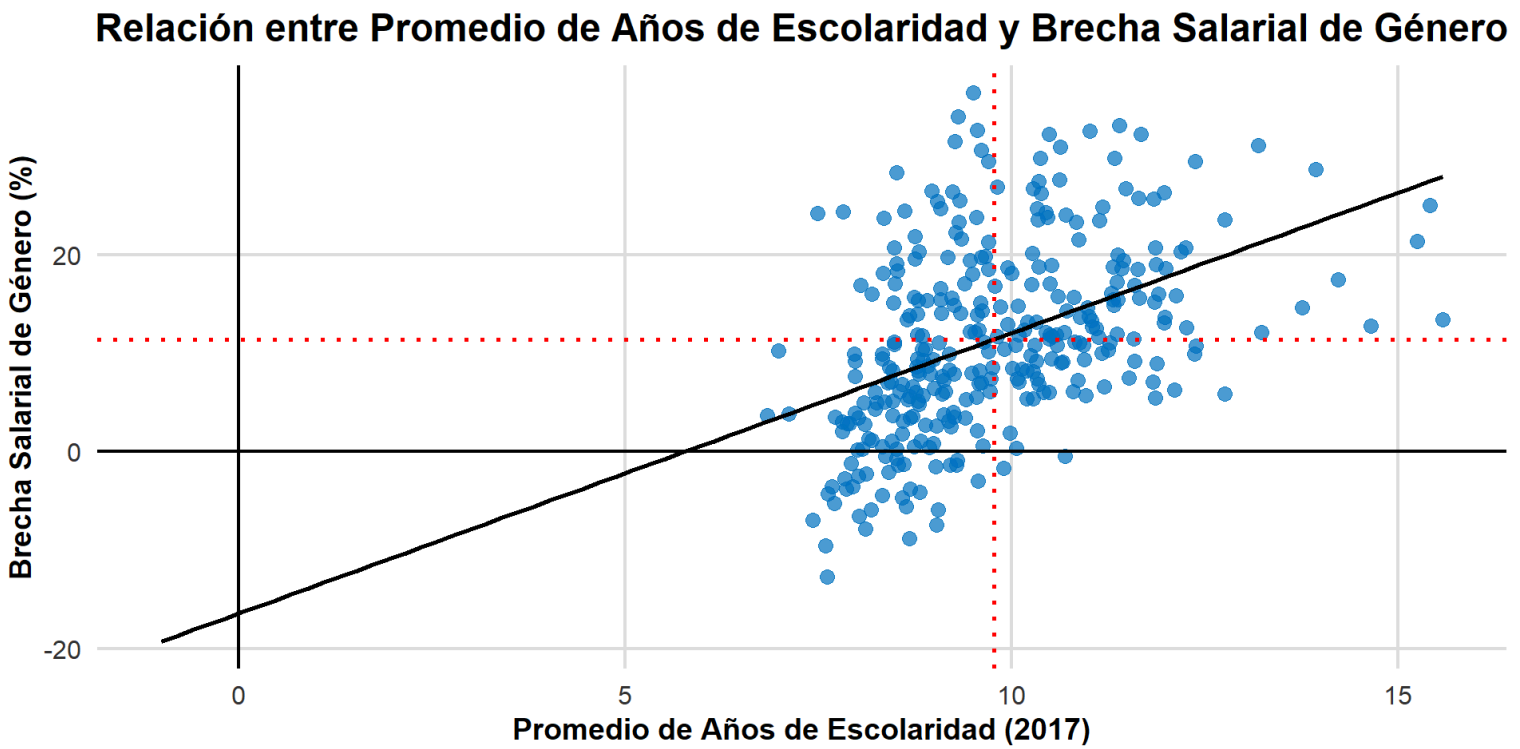
Considerando que en las comunas tenemos una correlación de 0.44 entre la brecha salarial de género y los años de escolaridad, con una desviación estándar de sy = 9.48 para la brecha salarial y de sx = 1.48 para los años de escolaridad, y medias de mean_y = 11.42% y mean_x = 9.77 años respectivamente, la pendiente y la ordenada se calculan como b = 2.85 y a = -16.45.

Gráfico de regresión



$$\hat{brecha} = -16.451 + 2.852 * escolaridad$$

Gráfico de regresión



$$\hat{brecha} = -16.451 + 2.852 * escolaridad$$

Características de la Regresión

- **Distinción entre variable explicativa y variable respuesta**:
  - La regresión mínimo-cuadrática considera sólo las distancias verticales de los puntos a la recta.
  - Cambiar los papeles de las dos variables resulta en una recta de regresión distinta.

Conexión entre correlación y regresión: - La pendiente de la recta de regresión mínimo-cuadrática se calcula como:

$$b = r\frac{s_y}{s_x}$$

- A lo largo de la recta de regresión:
  - Un cambio de una desviación típica en x provoca un cambio de r desviaciones típicas en y.
  - Cuando $r = 1$ o $r = -1$, el cambio en y predicho es igual al cambio en x.
  - Si $-1 \leq r \leq 1$, el cambio en y es menor que el cambio en x.
  - A menor correlación, menor es la predicción de y en respuesta a x.

Características de la Regresión

- **Punto de paso de la recta de regresión**:
  - La recta de regresión mínimo-cuadrática siempre pasa por el punto $(\bar{x}, \bar{y})$.
  - La recta de regresión se describe completamente con $\bar{x}$, s_x, $\bar{y}$, s_y y r.
  - **Correlación r y la fuerza de la relación lineal**:
  - El cuadrado de la correlación, r^2, indica la fracción de la variación de y explicada por la recta de regresión.
  - r^2 se utiliza para medir la calidad de la predicción proporcionada por la regresión.
- **Relación entre r y r^2**:
  - Una correlación perfecta ($r = \pm 1$) implica que $r^2 = 1$, lo que significa que toda la variación de y se explica por la relación lineal con x.
  - Si $r = \pm 0.7$, entonces $r^2 = 0.49$, indicando que aproximadamente la mitad de la variación se explica con la relación lineal.