ECONOMETRÍA I

Tema 2: El Modelo de Regresión Lineal Simple

Patricia Moreno
Juan Manuel Rodriguez Poo
Alexandra Soberon
Departamento de Economía

Modelo de Regresión Lineal Simple

- El Modelo de Regresión Lineal Simple nos permite explicar y en términos de x.
- Sea

$$y = \beta_0 + \beta_1 x + u,$$

donde

  - y: variable dependiente, endógena, explicada o regresando...
  - x: variable independiente, exógena, explicativa, de control, regresor...
  - β_0 y β_1: parámetros poblacionales.
  - u: término de error o perturbación no observable.

Modelo de Regresión Lineal Simple

- β_1: parámetro de pendiente. Mide la relación entre x e y, es decir, cómo cambia y cuando se producen modificaciones en x.
- β_0: término constante. Es el valor de y cuando x y u son cero.
- Si todos los demás factores contenidos en u se mantienen constantes ($\Delta u = 0$), x tiene un efecto lineal sobre y, es decir,

$$\Delta y = \beta_1 \Delta x \qquad \text{si} \quad \Delta u = 0.$$

Modelo de Regresión Lineal Simple: ejemplos

$$y = \beta_0 + \beta_1 x + u.$$

- **Cultivo de soja y fertilizante**: Si y =cosecha y x = cantidad de fertilizante, el término de error (u) recoge factores como:
  - calidad de la tierra.
  - lluvia.
- **Ecuación salarial simple**: Si y =salario y x =años de estudio, el término de error (u) recoge factores no observables como:
  - experiencia laboral.
  - capacidad o habilidad.
  - antigüedad en la empresa.

- **Linealidad en los parámetros**: $y = \beta_0 + \beta_1 x + u$,
  es decir, un cambio unitario en x tiene el mismo efecto sobre y con independencia del valor inicial de x, i.e.

$$\Delta x = 1 \Longrightarrow \Delta y = \beta_1, \qquad \forall x, \Delta u = 0.$$

- **Media condicional cero**: $E(u|x) = 0\ \forall x$. Para cualquier valor de x, la media del término de error no observable es siempre la misma e igual a cero.

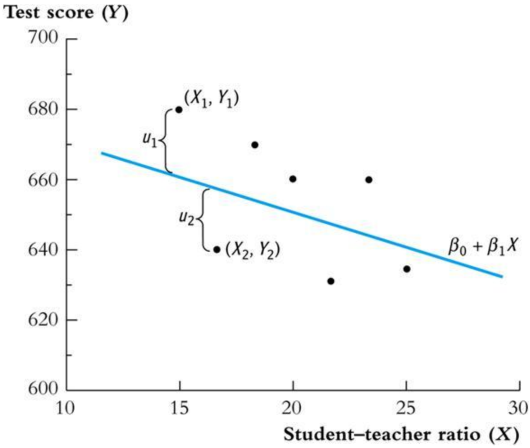
$$E(u|x) = E(u) = 0.$$

Función de regresión poblacional lineal

- El supuesto $E(u|x) = E(u) = 0$ nos lleva a
$$E(y|x) = \beta_0 + \beta_1 x.$$
- Esta expresión nos proporciona el valor de la **función de regresión poblacional**. En este caso es lineal.
- Nos indica cómo varía el valor medio de y ante cambios en x, es decir,
$$\frac{\partial E(y|x)}{\partial x} = \beta_1$$
- Así, $E(y|x) = \beta_0 + \beta_1 x$ es la parte explicada por x y u es la parte no explicada por x.

Función de Regresión Poblacional

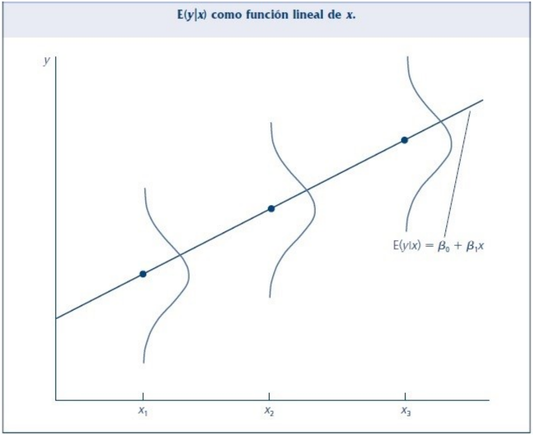
Calificaciones en el examen vs. ratio estudiantes-maestros



Fuente: Stock & Watson (2011)

Función de Regresión Poblacional

$E(y|x)$ es una función lineal de x, donde para cualquier x la distribución de y está centrada alrededor de $E(y|x)$.



Fuente: Wooldridge (2005).

Mínimos Cuadrados Ordinarios (MCO)

- **Idea básica de la regresión**: estimar los parámetros poblacionales (β_0, β_1) a partir de un conjunto de datos.
- Siendo $\{(x_i, y_i) : i = 1, \cdots, n\}$ una muestra aleatoria de tamaño n tomada de una población desconocida, para cada observación i tenemos

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- **Supuesto necesario estimadores OLS**: $E(u|x) = E(u) = 0$ que también implica $Cov(x, u) = E(xu) = 0$.
- ¿Por qué? Recuerda de la probabilidad básica que

$$Cov(x, y) = E(xy) - E(x)E(y).$$

Mínimos Cuadrados Ordinarios (MCO)

- Dado que $u = y - \beta_0 - \beta_1 x$, es posible expresar los parámetros desconocidos en términos de x, y, β_0 y β_1. Así, obtenemos dos **restricciones de momento**:

$$E\left[y - \beta_0 - \beta_1 x\right] = 0$$
$$E\left[x(y - \beta_0 - \beta_1 x)\right] = 0.$$

- **Método de los momentos**: proceso de estimación que implica la imposición de restricciones de momentos poblacionales sobre los momentos muestrales.
- ¿Qué implica todo esto? Sea una muestra aleatoria de observaciones de y, $\{y_i\}_{i=1}^n$, para estimar $E(y)$ utilizamos la media muestral $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$.

MCO: método de los momentos

- **Objetivo**: elegir los valores de los parámetros poblacionales que aseguran que las versiones muestrales de las restricciones de momento son ciertas:

$$n^{-1}\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i\right)=0$$
$$n^{-1}\sum_{i=1}^{n}x_i\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i\right)=0$$

- Dada la definición de la media muestral y aprovechando las propiedades básicas de la suma, podemos reescribir la primera condición como

$$\overline{y}=\widehat{\beta}_0+\widehat{\beta}_1\overline{x},$$

Reemplazando la expresión de $\widehat{\beta}_0$

$$\sum_{i=1}^{n} x_i \left(y_i - \left(\overline{y} - \widehat{\beta}_1 \overline{x} \right) - \widehat{\beta}_1 x_i \right) = 0$$

y agrupando términos,

$$\sum_{i=1}^{n} x_i \left(y_i - \overline{y} \right) = \widehat{\beta}_1 \sum_{i=1}^{n} x_i \left(x_i - \overline{x} \right).$$

Utilizando las relaciones $\sum_{i=1}^{n} (x_i - \overline{x}) x_i = \sum_{i=1}^{n} (x_i - \overline{x})^2$ y $\sum_{i=1}^{n} (y_i - \overline{y}) x_i = \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})$,

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

- Los **estimadores MCO** resultantes son:

$$
\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x}, \\
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.
\end{aligned}
$$

Ejemplo práctico

- **Objetivo**: estudiar la relación entre el rendimiento sobre el capital de una empresa y el sueldo de los directores generales de dicha empresa (CEO).
- **Variables disponibles**:
  - *salary*: sueldo anual de los CEOs.
  - *roe*: rendimiento del capital medio de las empresas.
- **Modelo**: $salary = \beta_0 + \beta_1 roe + u$
- **Recta de regresión**: $\widehat{salary} = 963{,}191 + 18{,}501 roe$.
- ¿Cuál es el cambio predicho en *salary* ante un aumento unitario en *roe*?
- ¿Cuál es el sueldo predicho si $roe = 30$?

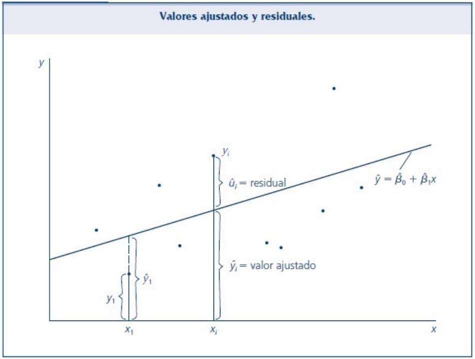
Propiedades estimador MCO de la pendiente

- El estimador de la pendiente $\widehat{\beta}_1$ es simplemente la covarianza muestral entre x y y dividida entre la varianza muestral de x.
- Si x e y están correlacionadas positivamente, $\widehat{\beta}_1$ será positiva.
- Si x e y están correlacionadas negativamente, $\widehat{\beta}_1$ será negativa.
- Para obtener $\widehat{\beta}_1$ la condición necesaria es que x varíe a lo largo de la muestra.

Función de regresión muestral

Una vez obtenidos los estimadores podemos calcular el valor ajustado (predicho):

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$



Fuente: Wooldridge (2005)

- **Criterio MCO**: escoger $\widehat{\beta}_0$ y $\widehat{\beta}_1$ tal que la suma de residuos al cuadrado, $E(u_i^2)$, sea lo más pequeña posible.
- **Residuos**: diferencia existente entre el valor observado y_i y el valor ajustado o predicho $\widehat{y}_i$:

$$\widehat{u}_i = y_i - \widehat{y}_i = y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)$$

- Análogo muestral del modelo del problema a minimizar $E(u_i^2)$

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (\widehat{u}_i)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2.$$

MCO: criterio de minimización de residuos

- Las condiciones de primer orden son

$$\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = 0,$$
$$\sum_{i=1}^{n} X_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = 0.$$

- Estas condiciones de primer orden son el análogo muestral de las condiciones de primer orden poblacionales, es decir,

$$E(u) = 0,$$
$$Cov(x,u) = 0.$$

Propiedades de los estimadores MCO

- La suma (y la media muestral) de los residuos MCO son cero,

$$\sum_{i=1}^{n} \widehat{u}_i = 0.$$

- La covarianza muestral entre los regresores y los residuos MCO es cero,

$$\sum_{i=1}^{n} x_i \widehat{u}_i = 0.$$

- La covarianza muestral entre los valores ajustados y los residuos MCO es cero,

$$\sum_{i=1}^{n} \widehat{y}_i \widehat{u}_i = 0.$$

Propiedades de los estimadores MCO

- El punto $(\overline{x}, \overline{y})$ siempre se encuentra sobre la recta ajustada, es decir
$$\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}.$$
- y_i puede ser descompuesto entre el valor ajustado y su residuo asociado, que están incorrelacionados, es decir,
$$y_i = \widehat{y}_i + \widehat{u}_i.$$

- **Coeficiente de determinación (R^2)**: se interpreta como el porcentaje de la variación muestral de y que es explicado por x y se calcula como la fracción de la suma total de los cuadrados (SST) que es explicada por el modelo:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$

donde

  - Suma Total de los Cuadrados (SST): $SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$.
  - Suma Explicada de los Cuadrados (SSE): $SSE = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$.
  - Suma de los Cuadrados de los Residuos (SSR): $SSR = \sum_{i=1}^{n} \widehat{u}_i^2$.

- $SST = SSE + SSR$.

Bondad de ajuste

- $SST = SSE + SSR$. Prueba:

$$
\begin{aligned}
\sum_{i=1}^{n} (y_i - \overline{y})^2 &= \sum_{i=1}^{n} [(y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})]^2 \\
&= \sum_{i=1}^{n} [\widehat{u}_i + (\widehat{y}_i - \overline{y})]^2 \\
&= \sum_{i=1}^{n} \widehat{u}_i^2 + 2\sum_{i=1}^{n} \widehat{u}_i(\widehat{y}_i - \overline{y}) + \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 \\
&= SSR + 2\sum_{i=1}^{n} \widehat{u}_i(\widehat{y}_i - \overline{y}) + SSE,
\end{aligned}
$$

donde sabemos que $\sum_{i=1}^{n} \widehat{u}_i(\widehat{y}_i - \overline{y}) = 0$.

- ¿Cómo cambios en las unidades de medida de la variable dependiente o de la independiente afecta a las estimaciones MCO?

$$\widehat{salary} = 0{,}963191 + 0{,}018501 roe$$

- Cambio en (Y): si Y es multiplicada por $c = 1000$,

$$\widehat{salary} = 963{,}191 + 18{,}501 roe$$

- Cambio en (X): si definimos $roedec = roe/100$,

$$\widehat{salary} = 963{,}191 + 1850{,}1 roedec$$

Unidades de medida y forma funcional

- Las relaciones lineales no son suficientes para describir las relaciones económicas.
- **Modelo lineal**: implica que el incremento de y ante cambios en x es siempre igual, independientemente del nivel de x:

$$salary = \beta_0 + \beta_1 roe + u$$

- **Elasticidad constante (log-log)**: la relación entre x e y se establece en términos de incrementos relativos.

$$log(salary) = \beta_0 + \beta_1 log(roe) + u,$$

donde β_1 es la elasticidad de *salary* respecto a *roe*. Así,

$$\%\Delta salary = \beta_1 \%\Delta roe.$$

- **Semielasticidad (log-level)**: el cambio se produce en términos porcentuales.

$$log(salary) = \beta_0 + \beta_1 roe + u$$

Si $\Delta u = 0$, entonces $(100 \cdot \beta_1)$ es la semielasticidad de *salary* respecto a *roe*,

$$\%\Delta salary = (100 \cdot \beta_1)\Delta roe.$$

- **Semielasticidad (level-log)**: controlamos incrementos relativos de x,

$$salary = \beta_0 + \beta_1 log(roe) + u,$$

donde $\Delta Y = \frac{\beta_1}{100}\%\Delta x$.

Modelo	Var.Dep	Var.Indep	Interpretación
Nivel-nivel	Y	X	$\Delta Y = \beta_1 \Delta X$
Nivel-log	Y	$log(X)$	$\Delta Y = (\beta_1/100)\, \%\Delta X$
Log-nivel	$log(Y)$	X	$\%\Delta Y = (100\beta_1)\Delta X$
Log-log	$log(Y)$	$log(X)$	$\%\Delta Y = \beta_1\, \%\Delta X$

Insesgadez de los estimadores MCO

- **Supuestos necesarios**:
  - **Linealidad**: el modelo poblacional es lineal en parámetros, $y = \beta_0 + \beta_1 x + u$.
  - **Muestreo aleatorio**: Contamos con una muestra aleatoria de tamaño n, $\{(x_i, y_i) : i = 1, 2, \cdots, n\}$. Así, el modelo muestral es $y_i = \beta_0 + \beta_1 x_i + u_i$.
  - **Media condicional cero**: $E(u|x) = 0$ de modo que $E(u_i|x_i) = 0$.
  - **Variación muestral de la variable explicativa**: no todos los valores muestrales de x son iguales.
- Para analizar el sesgo del estimador MCO es necesario reescribirlo en términos de los parámetros poblacionales.

- Estimador de la pendiente:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{s_x^2}$$

- Numerador:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x})y_i &= \sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i + u_i) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})\beta_0 + \sum_{i=1}^{n}(x_i - \overline{x})\beta_1 x_i + \sum_{i=1}^{n}(x_i - \overline{x})u_i \\
&= \beta_0 \sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1 \sum_{i=1}^{n}(x_i - \overline{x})x_i + \sum_{i=1}^{n}(x_i - \overline{x})u_i \\
&= \beta_1 s_x^2 + \sum_{i=1}^{n}(x_i - \overline{x})u_i,
\end{aligned}$$

dado que $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ y $\sum_{i=1}^{n}(x_i - \overline{x})x_i = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

- De este modo,

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})u_i}{s_x^2}.$$

- Definiendo $d_i = (x_i - \overline{x})$,

$$\widehat{\beta}_1 = \beta_1 + \left(\frac{1}{s_x^2}\right)\sum_{i=1}^{n} d_i u_i.$$

- Finalmente,

$$E(\widehat{\beta}_1) = \beta_1 + \left(\frac{1}{s_x^2}\right)\sum_{i=1}^{n} d_i E(u_i) = \beta_1$$

Insesgadez de los estimadores MCO

- Los estimadores MCO de β_0 y β_1 son insesgados.
- La prueba de insesgadez de los estimadores MCO depende de los 4 supuestos. Si uno sólo falla entonces los estimadores MCO no son necesariamente insesgados.
- La insesgadez es una descripción de los estimadores. En concreto, nos indica si está <u>cerca</u> o <u>lejos</u> del verdadero parámetro.

Insesgadez de los estimadores MCO

- La propiedad de insesgadez es una descripción de los estimadores que nos indica que si disponemos de un nº finito de muestras de tamaño *n* de la misma población y estimamos el mismo modelo con cada una de las muestras:
  - tendremos una distribución de valores estimados de β_j, con una realización numérica distinta para cada muestra.
  - la media de la distribución de dichos valores estimados de β_j coincidirá con el parámetro poblacional β_j.

Varianza de los estimadores MCO

- Además de saber que la distribución muestral de $\widehat{\beta}_1$ está centrada en β_1, también es relevante saber qué tanto puede esperarse que $\widehat{\beta}_1$ se aleje, en promedio, de β_1.
- Así seremos capaces de elegir el mejor estimador de una amplia gama de estimadores insesgados.

Varianza de los estimadores MCO

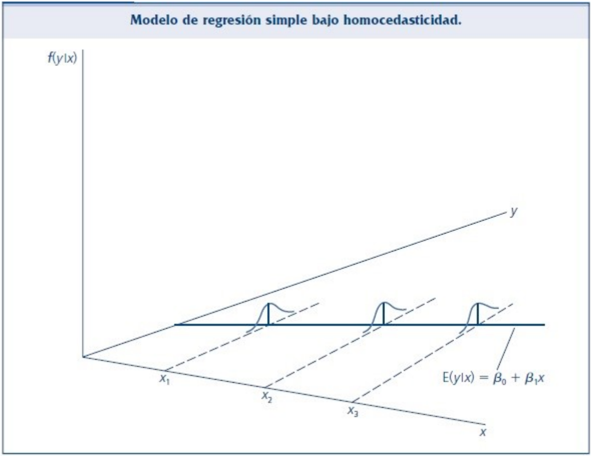
- Supuesto adicional: $Var(u|x) = \sigma^2$ (**homocedasticidad**) y dado que $Var(u|x) = E(u^2|x) - [E(u|x)]^2$ y $E(u|x) = 0$,

$$\sigma^2 = E(u^2|x) = E(u^2) = Var(u).$$

- Así, σ^2 es la varianza incondicional, también conocida como la varianza del error.

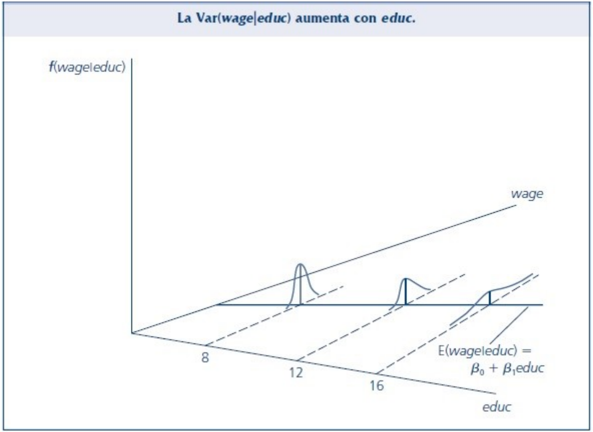
- σ (la raíz cuadrada de la varianza del error) es conocida como la desviación estándar del error.

Varianza del estimador OLS: homocedasticidad



Fuente: Wooldridge (2005)

Varianza del estimador OLS: heterocedasticidad



Fuente: Wooldridge (2005)

Varianza del estimador OLS

$$
\begin{aligned}
Var(\widehat{\beta}_1) &= Var\left(\beta_1 + \left(\frac{1}{s_x^2}\sum_{i=1}^{n} d_i u_i\right)\right) \\
&= \left(\frac{1}{s_x^2}\right)^2 Var\left(\sum_{i=1}^{n} d_i u_i\right) \\
&= \left(\frac{1}{s_x^2}\right)^2 \sum_{i=1}^{n} d_i^2 Var(u_i) \\
&= \left(\frac{1}{s_x^2}\right)^2 \sum_{i=1}^{n} d_i^2 \sigma^2 = \sigma^2 \left(\frac{1}{s_x^2}\right)^2 \sum_{i=1}^{n} d_i^2 \\
&= \sigma^2 \left(\frac{1}{s_x^2}\right)^2 s_x^2 = \frac{\sigma^2}{s_x^2}
\end{aligned}
$$

Varianza del estimador OLS

- Cuanto mayor sea la varianza del error (σ^2) mayor será la varianza del estimador de la pendiente $Var(\widehat{\beta}_1)$.
- A medida que aumenta la variabilidad de las x_i, la varianza del estimador de la pendiente $Var(\widehat{\beta}_1)$ será menor.
- Como resultado, a medida que se incrementa el tamaño de la muestra n también aumentará la varianzación total de las x_i obteniendo una menor $Var(\widehat{\beta}_1)$.
- **Problema**: la varianza del error (σ^2) es desconocida.

Estimación de la varianza del error σ^2

- Dado que los errores, u_i, no pueden ser observados nosotros nos conocemos la varianza del error, σ^2.
- Lo que podemos observar son los residuos de modo que podemos usarlos para estimar la varianza del error:

$$\begin{aligned}\widehat{u}_i &= y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \\ &= u_i - \left(\widehat{\beta}_0 - \beta_0\right) - \left(\widehat{\beta}_1 - \beta_1\right) x_i.\end{aligned}$$

- Reemplazando los errores por los residuos, un estimador factible de σ^2 sería

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i^2.$$

- **Problema**: Este estimador es factible pero sesgado. Como los residuos verifican $n^{-1}\sum_{i=1}^{n}\widehat{u}_i = 0$ y $n^{-1}\sum_{i=1}^{n}\widehat{u}_i x_i = 0$, sólo hay $(n-2)$ residuos independientes (**grados de libertad**).
- Un estimador insesgado factible de σ^2 es

$$\widehat{\sigma}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{SSR}{(n-2)}.$$

- Tanto $\widetilde{\sigma}^2$ como $\widehat{\sigma}^2$ son estimadores consistentes de σ^2. Para tamaños muestrales moderados es irrelevante cuál de los dos utilizar dado que siempre que n no sea muy pequeño nos proporcionarán estimaciones numéricas muy similares.

Estimación del error estándar de la regresión σ

- En el caso poblacional vimos que la función de esperanza condicional $E(y|x)$ es la proyección lineal de y dado x, en el sentido de que minimiza $E(u^2)$.
- Por analogía, en el caso de la estimación MCO a partir de una muestra de los coeficientes del modelo de regresión clásico tenemos $Var(u|x) = \sigma^2$ de modo que $sd(\widehat{\beta}_1) = \frac{\sigma}{s_x}$.
- De este modo, $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$ es el Error Estándar de la regresión.
- Reemplazando σ por $\widehat{\sigma}$ obtenemos el **Error Éstándar** de $\widehat{\beta}_1$,

$$se(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{1/2}}.$$

- **Supuestos**:
  - Linealidad en los parámetros del modelo.
  - Muestreo aleatorio.
  - Variación muestral en la variable explicativa.
  - Media condicional cero.
  - Homocedasticidad.
- **Teorema**: En el contexto del modelo de regresión lineal, bajo los 4 supuestos anteriores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son los estimadores de menor varianza entre los estimadores lineales e insesgados.

Regresión a través del origen

- En ocasiones se impone la restricción de que cuando $x = 0$ el valor esperado de y es cero. Ej: si el ingreso es cero ($x = 0$), la recaudación de impuestos de las rentas del trabajo y debe ser cero.
- **Regresión a través del origen**:

$$\widetilde{y} = \widetilde{\beta}_1 x.$$

- **Problema de suma de cuadrados residual**:

$$\sum_{i=1}^{n} \left(y_i - \widetilde{\beta}_1 x_i \right)^2 .$$

- **Estimador MCO**:

$$\widetilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$