Variables cuantitativas, relaciones entre estas, y el LS regression

Aniel Nieves-González

Instituto de estadísticas

Variables cuantitativas y relaciones entre estas (Section 2.1, 2.2)

Recuerde que informalmente...

- Una variable aleatoria (cuantitativa) es una asociación entre los miembros del espacio muestral (conjunto de objetos que nos interesa estudiar) y los números reales.

Variables cuantitativas y relaciones entre estas (Section 2.1, 2.2)

Definition (Response variable)

Una variable que mide el resultado de un estudio es un response variable. También es llamada **variable dependiente**.

Definition (Explanatory variable)

Variable que explica o influencia los cambios en el response variable. También es llamada **variable independiente**.

Variables cuantitativas y relaciones entre estas (Section 2.1, 2.2)

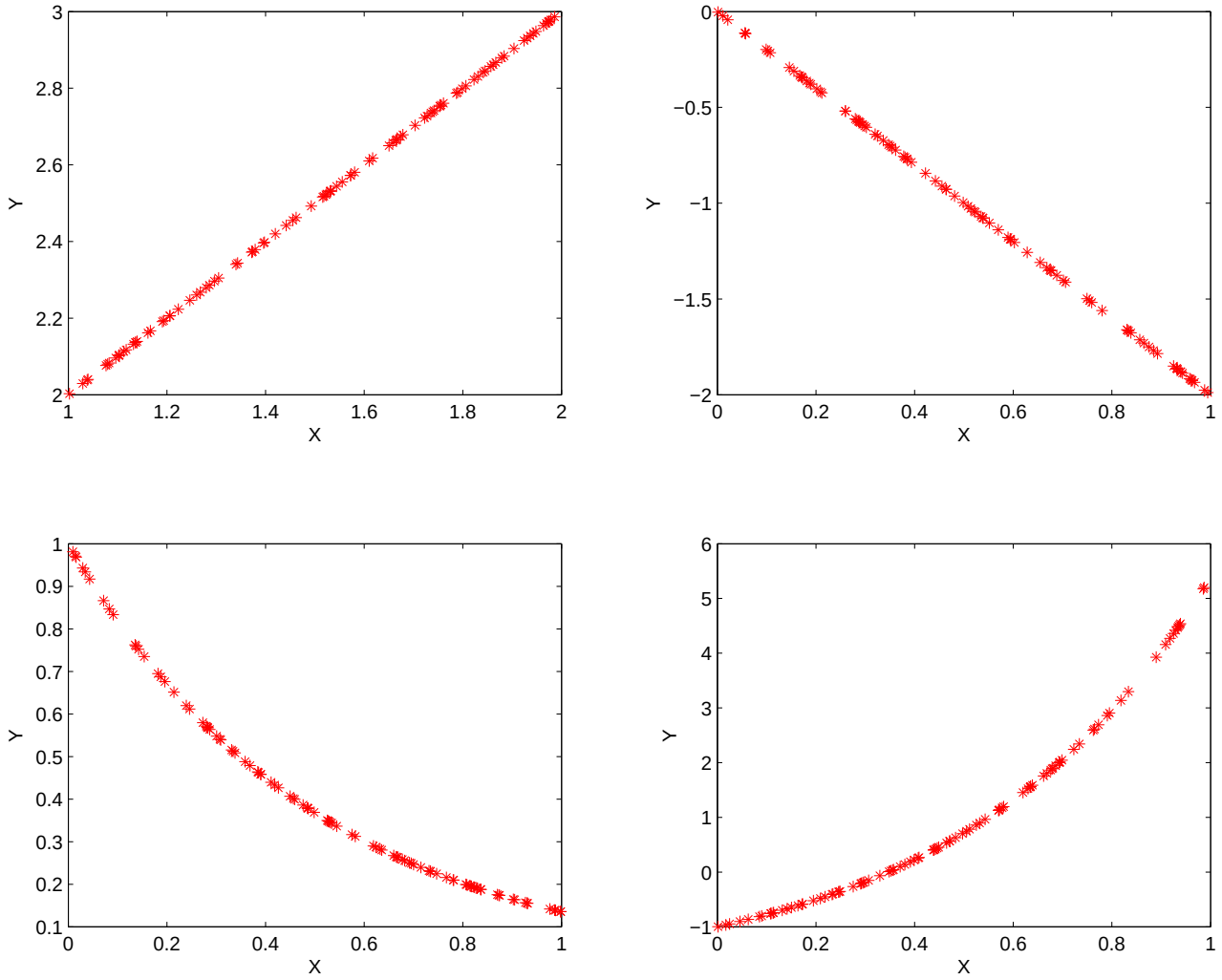
Definition (Scatter Plot)

Gráfica que muestra la relación entre dos variables aleatorias medidas sobre el mismo conjunto de individuos.

- Dos variables (var.) están **asociadas positivamente** si valores sobre el prom. de una acompañan a valores sobre el promedio de otra, y valores por debajo del prom. tienden a ocurrir juntos.

- Dos variables (var.) están **asociadas positivamente** si valores sobre el prom. de una acompañan a valores sobre el promedio de otra, y valores por debajo del prom. tienden a ocurrir juntos.
- Dos var. están **asociadas negativamente** si valores sobre el prom. de una acompañan a valores bajo el promedio de otra, y vice versa.

Section 2.1, 2.2



El concepto de **correlación** mide cuantitativamente la dirección y fuerza de una relación *lineal* entre dos variables.

Definition (Correlación)

Suponga q. tiene n datos de las variables X y Y. Denotamos dichos datos como $\{x_1, \ldots, x_n\}$ y $\{y_1, \ldots, y_n\}$ respectivamente. Sea $\bar{X}$ y $s_x$la media aritmética y la desviación estándar para los datos de X. Análogamente $\bar{Y}$ y s_y para con Y. La **correlación**, denotada como r, entre X y Y es

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{s_x} \right) \left(\frac{y_i - \bar{Y}}{s_y} \right)$$

Observe:

- r no hace distinción entre var. dependente e independiente.

Observe:

- r no hace distinción entre var. dependente e independiente.
- r solo aplica a variables cuantitativas.

Observe:

- r no hace distinción entre var. dependente e independiente.
- r solo aplica a variables cuantitativas.
- r no tiene unidades ("dimensionless").

Observe:

- r no hace distinción entre var. dependente e independiente.
- r solo aplica a variables cuantitativas.
- r no tiene unidades ("dimensionless").
- $r > 0 \Rightarrow$ relación lineal positiva

Observe:

- r no hace distinción entre var. dependente e independiente.
- r solo aplica a variables cuantitativas.
- r no tiene unidades ("dimensionless").
- $r > 0 \Rightarrow$ relación lineal positiva
- $r < 0 \Rightarrow$ relación lineal negativa

Observe:

- r no hace distinción entre var. dependente e independiente.
- r solo aplica a variables cuantitativas.
- r no tiene unidades ("dimensionless").
- $r > 0 \Rightarrow$ relación lineal positiva
- $r < 0 \Rightarrow$ relación lineal negativa
- $|r| \leq 1$ (i.e. $-1 \leq r \leq 1$).

Observe:

- r no hace distinción entre var. dependente e independiente.
- r solo aplica a variables cuantitativas.
- r no tiene unidades ("dimensionless").
- $r > 0 \Rightarrow$ relación lineal positiva
- $r < 0 \Rightarrow$ relación lineal negativa
- $|r| \leq 1$ (i.e. $-1 \leq r \leq 1$).
- $r \approx 0 \Rightarrow$ relación lineal débil.

Observe:

- r no hace distinción entre var. dependente e independiente.
- r solo aplica a variables cuantitativas.
- r no tiene unidades ("dimensionless").
- $r > 0 \Rightarrow$ relación lineal positiva
- $r < 0 \Rightarrow$ relación lineal negativa
- $|r| \leq 1$ (i.e. $-1 \leq r \leq 1$).
- $r \approx 0 \Rightarrow$ relación lineal débil.
- $r \approx -1$ o $r \approx 1$ implica relación lineal fuerte.

Por último: r mide fuerza de relaciones **lineales** solamente. Y al igual que $\bar{X}$ no es resistente a outliers.

Ejemplo 1: Calculando r

Suponga los siguientes datos y calcule r.

rapidez (speed)	20	30	40	50	60
MPG	24	28	30	28	24

Sea $X \equiv$ 'rapidez en mph', $Y \equiv$ 'MPG', y $n = 5$.

Ejemplo 1: Calculando r

Suponga los siguientes datos y calcule r.

rapidez (speed)	20	30	40	50	60
MPG	24	28	30	28	24

Sea $X \equiv$ 'rapidez en mph', $Y \equiv$ 'MPG', y $n = 5$.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{20+30+40+50+60}{5} = \frac{200}{5} = 40$$

Ejemplo 1: Calculando r

Suponga los siguientes datos y calcule r.

rapidez (speed)	20	30	40	50	60
MPG	24	28	30	28	24

Sea $X \equiv$ 'rapidez en mph', $Y \equiv$ 'MPG', y $n = 5$.

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 = \frac{(-20)^2 + (-10)^2 + (0) + (-10)^2 + (10)^2}{4}$$

$$= \frac{400 + 100 + 0 + 100 + 400}{4} = \frac{1000}{4} = 250$$

$$\Rightarrow s_x = 15.81$$

Ejemplo 1: Calculando r

Suponga los siguientes datos y calcule r.

rapidez (speed)	20	30	40	50	60
MPG	24	28	30	28	24

Sea $X \equiv$ 'rapidez en mph', $Y \equiv$ 'MPG', y $n = 5$.
Análogamente:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{134}{5} = 26.8$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{Y})^2 = \frac{7.84 + 1.44 + 10.24 + 1.44 + 7.84}{4}$$
$$= \frac{28.8}{4} = 7.2$$

$$\Rightarrow s_y = 2.683$$

Ejemplo 1: Calculando r

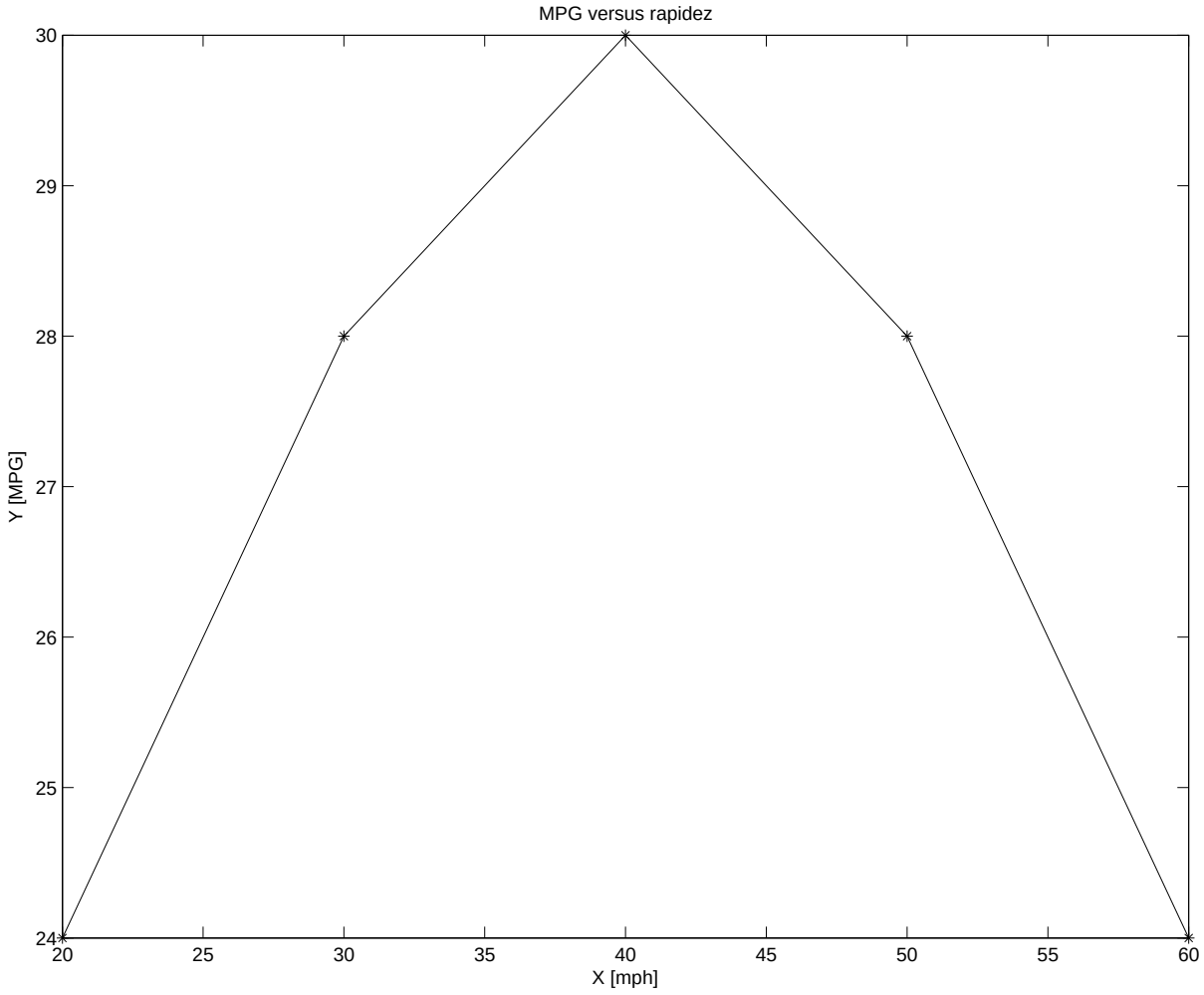
Con $\bar{X} = 40$, $\bar{Y} = 26.8$, $s_x = 15.81$, y $s_y = 2.683$ tenemos

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{s_x}\right)\left(\frac{y_i - \bar{Y}}{s_y}\right)$$
$$= \frac{(-20)(-2.8) + (-10)(1.2) + 0 + (10)(1.2) + (20)(-2.8)}{4(15.81)(2.683)} = 0$$

Por tanto tenemos una relación lineal débil (no-lineal realmente).

Ejemplo 1: Calculando r

Si miramos una gráfica de los datos tenemos...



En R podemos calcular la correlación (r) ejecutando:

- x = c(20,30,40,50,60)
- y = c(24,28,30,28,24)
- r = cor(x,y)

Least-squares regression (regresión por cuadrados mínimos), sección 2.3

En esencia el problema de cuadrados mínimos (o least squares (LS) problem o LS regression) es un problema de minimización, el cual se puede escribir como:

Least-squares regression (regresión por cuadrados mínimos), sección 2.3

En esencia el problema de cuadrados mínimos (o least squares (LS) problem o LS regression) es un problema de minimización, el cual se puede escribir como:

$$\min_{\mathbf{p}} \sum_{i=1}^{n} (F(x_i, \mathbf{p}) - y_i)^2 + \epsilon$$

donde:

- $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ son los datos (note las dos variables X y Y)

Least-squares regression (regresión por cuadrados mínimos), sección 2.3

En esencia el problema de cuadrados mínimos (o least squares (LS) problem o LS regression) es un problema de minimización, el cual se puede escribir como:

$$\min_{\mathbf{p}} \sum_{i=1}^{n} (F(x_i, \mathbf{p}) - y_i)^2 + \epsilon$$

donde:

- $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ son los datos (note las dos variables X y Y)
- $\mathbf{p}$ son los parametros (constantes del modelo matemático).

Least-squares regression (regresión por cuadrados mínimos), sección 2.3

En esencia el problema de cuadrados mínimos (o least squares (LS) problem o LS regression) es un problema de minimización, el cual se puede escribir como:

$$\min_{\mathbf{p}} \sum_{i=1}^{n} (F(x_i, \mathbf{p}) - y_i)^2 + \epsilon$$

donde:

- $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ son los datos (note las dos variables X y Y)
- $\mathbf{p}$ son los parametros (constantes del modelo matemático).
- $F(x_i, \mathbf{p})$ es el modelo matématico. Note que F depende de los datos y de los parámetros.

Least-squares regression (regresión por cuadrados mínimos), sección 2.3

En esencia el problema de cuadrados mínimos (o least squares (LS) problem o LS regression) es un problema de minimización, el cual se puede escribir como:

$$\min_{\mathbf{p}} \sum_{i=1}^{n} (F(x_i, \mathbf{p}) - y_i)^2 + \epsilon$$

donde:

- $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ son los datos (note las dos variables X y Y)
- $\mathbf{p}$ son los parametros (constantes del modelo matemático).
- $F(x_i, \mathbf{p})$ es el modelo matématico. Note que F depende de los datos y de los parámetros.
- En este caso el modelo matemático representará a la relación entre las dos variables X y Y.

Least-squares regression (regresión por cuadrados mínimos), sección 2.3

En esencia el problema de cuadrados mínimos (o least squares (LS) problem o LS regression) es un problema de minimización, el cual se puede escribir como:

$$\min_{\mathbf{p}} \sum_{i=1}^{n} (F(x_i, \mathbf{p}) - y_i)^2 + \epsilon$$

donde:

- $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ son los datos (note las dos variables X y Y)
- $\mathbf{p}$ son los parametros (constantes del modelo matemático).
- $F(x_i, \mathbf{p})$ es el modelo matématico. Note que F depende de los datos y de los parámetros.
- En este caso el modelo matemático representará a la relación entre las dos variables X y Y.
- ϵ representa error.

Least-squares regression (regresión por cuadrados mínimos), sección 2.3

$$\min_{\mathbf{p}} \sum_{i=1}^{n} (F(x_i, \mathbf{p}) - y_i)^2 + \epsilon$$

donde:

- $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ son los datos (note las dos variables X y Y)
- $\mathbf{p}$ son los parametros (constantes del modelo matemático).
- $F(x_i, \mathbf{p})$ es el modelo matématico. Note que F depende de los datos y de los parámetros.
- En este caso el modelo matemático representará a la relación entre las dos variables X y Y.
- ϵ representa error.
- Lo que minimizamos es la distancia vertical (a lo largo de eje de y) entre modelo y datos.

Least-squares regression

Para propósitos del curso, F (el modelo) será lineal, esto es,

$$F(x_i, \alpha, \beta) = \alpha x_i + \beta \quad \text{función lineal}$$

Least-squares regression

Para propósitos del curso, F (el modelo) será lineal, esto es,

$$F(x_i, \alpha, \beta) = \alpha x_i + \beta \quad \text{función lineal}$$

Definition (Regresión lineal)

Una regresión lineal es una linea recta (¡mire arriba!) que describe como una variable dependiente cambia a medida que cambia la variable independiente.

Least-squares regression

Definition (Regresión lineal)

Una regresión lineal es una linea recta (¡mire arriba!) que describe como una variable dependiente cambia a medida que cambia la variable independiente.

Observe que:

- Para la regresión por LS es importante distinguir entre var. independiente y dependiente.

Least-squares regression

Definition (Regresión lineal)

Una regresión lineal es una linea recta (¡mire arriba!) que describe como una variable dependiente cambia a medida que cambia la variable independiente.

Observe que:

- Para la regresión por LS es importante distinguir entre var. independiente y dependiente.
- La pendiente (slope) del LS regression es

$$\alpha = r\frac{s_y}{s_x} = \left(\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{s_x}\right)\left(\frac{y_i - \bar{Y}}{s_y}\right)\right)\frac{s_y}{s_x}$$

  y el intercepto en y (y-intercept) es

$$\beta = \bar{Y} - \alpha\bar{X}$$

  (La demostración va más alla de los objetivos del curso.)

Least-squares regression

Observe también que:

- La linea de regresión siempre pasa por $\left(\bar{X}, \bar{Y}\right)$.

Least-squares regression

Observe también que:

- La linea de regresión siempre pasa por $\left(\bar{X}, \bar{Y}\right)$.
- r^2 es la fracción de variación en Y que se explica por la regresión lineal de Y en X. Esto es:

$$
\begin{aligned}
r^2 &= \frac{\text{variacion en } \hat{y} \text{ a medida que cambia x}}{\text{variacion total en } y} \\
&= \frac{\text{suma de cuadrados debido a regresión}}{\text{suma total de cuadrados}} \\
&= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2}
\end{aligned}
$$

donde $\hat{y}_i$ es el la predicción obtenida de la linea de regresión ($\hat{y}_i = \alpha x_i + \beta$).

Least-squares regression

Observe también que:

- La linea de regresión siempre pasa por $\left(\bar{X}, \bar{Y}\right)$.
- r^2 es la fracción de variación en Y que se explica por la regresión lineal de Y en X. Esto es:

$$\begin{aligned} r^2 &= \frac{\text{variacion en } \hat{y} \text{ a medida que cambia x}}{\text{variacion total en } y} \\ &= \frac{\text{suma de cuadrados debido a regresión}}{\text{suma total de cuadrados}} \\ &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2} \end{aligned}$$

  donde $\hat{y}_i$ es el la predicción obtenida de la linea de regresión ($\hat{y}_i = \alpha x_i + \beta$).
- Note que para el caso de perfecta correlación lineal ($r = 1$ o $r = -1$) $r^2 = 1$. O sea que r^2 mide que tan exitosa fue la correlación.

Least-squares regression: residual

Definition (Residual)

$$\text{residual} = y_i - \hat{y} \quad \text{para cada} \quad i$$

donde y_i es al i-ésimo dato y $\hat{y} = \alpha x + \beta$ (esto es la predicción del linear LS regression).

Note que la media de los residuales es cero.

Least-squares regression: residual

Los residuales pueden ilustrarse en una gráfica de residual vs. variable independiente. En esta nos fijamos:

Least-squares regression: residual

Los residuales pueden ilustrarse en una gráfica de residual vs. variable independiente. En esta nos fijamos:

- Patrones curvos: estos implican relación no lineal.

Least-squares regression: residual

Los residuales pueden ilustrarse en una gráfica de residual vs. variable independiente. En esta nos fijamos:

- Patrones curvos: estos implican relación no lineal.
- Dispersión decresiente o cresiente sobre la media: indican predicciones menos certeras a medida que x crece/decrese.

Least-squares regression: residual

Los residuales pueden ilustrarse en una gráfica de residual vs. variable independiente. En esta nos fijamos:

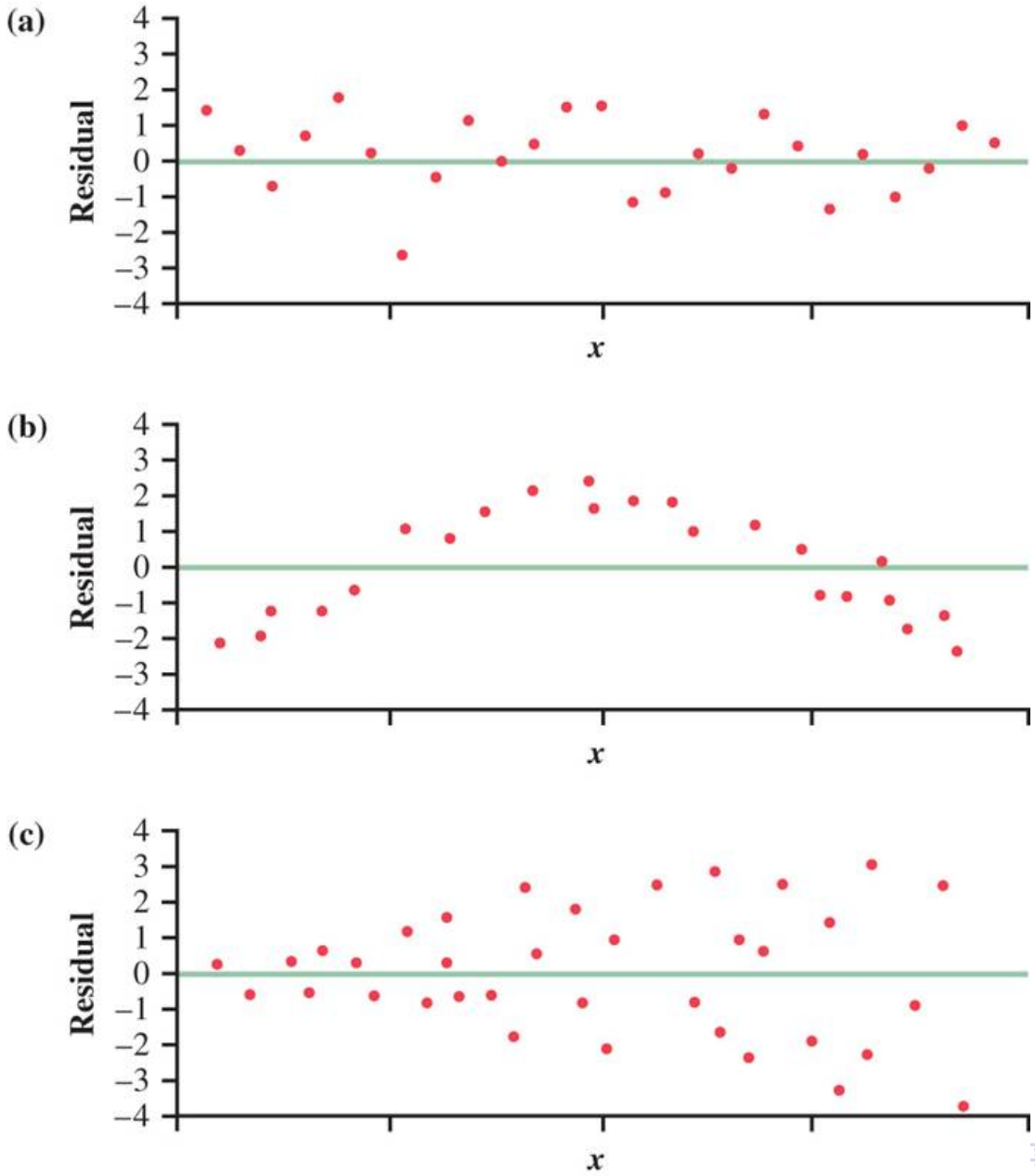
- Patrones curvos: estos implican relación no lineal.
- Dispersión decresiente o cresiente sobre la media: indican predicciones menos certeras a medida que x crece/decrese.
- Puntos individuales con residual grande: outliers

Least-squares regression: residual

Los residuales pueden ilustrarse en una gráfica de residual vs. variable independiente. En esta nos fijamos:

- Patrones curvos: estos implican relación no lineal.
- Dispersión decresiente o cresiente sobre la media: indican predicciones menos certeras a medida que x crece/decrese.
- Puntos individuales con residual grande: outliers
- Puntos extremos a lo largo de eje de x pero sin residual grande: observación influyente (para el LSP).

Fig. 2.18 del texto



Ejemplo:

Considere el "stock market". Algúnos piensan que el comportamiento del stock market a principios del año predice lo que ocurriá el resto del año. Ahora, suponga que X es la variable independiente y Y la dependiente donde:

$X \equiv$ taza de cambio en el stock market index en enero

$Y \equiv$ taza de cambio en el stock market index en el resto del año

Ejemplo:

Los datos en los últimos 38 años se resumen como sigue:
$\bar{X} = 0.0175$, $\bar{Y} = 0.0907$, $s_x = 0.0536$, $s_y = 0.1535$, y $r = 0.596$

Ejemplo:

Los datos en los últimos 38 años se resumen como sigue:
$\bar{X} = 0.0175$, $\bar{Y} = 0.0907$, $s_x = 0.0536$, $s_y = 0.1535$, y $r = 0.596$

1. Calcule la linea de regresión:

$$\alpha = r\frac{s_y}{s_x} = (0.596)\frac{0.1535}{0.0536} = 1.7068$$
$$\beta = \bar{Y} - \alpha\bar{X} = 0.0907 - (1.7068)(0.0175) = 0.0608$$

$$\Rightarrow \hat{y}_i = 1.7068x_i + 0.0608$$

Ejemplo:

Los datos en los últimos 38 años se resumen como sigue:
$\bar{X} = 0.0175$, $\bar{Y} = 0.0907$, $s_x = 0.0536$, $s_y = 0.1535$, y $r = 0.596$

1. Calcule la linea de regresión:

$$\alpha = r\frac{s_y}{s_x} = (0.596)\frac{0.1535}{0.0536} = 1.7068$$
$$\beta = \bar{Y} - \alpha\bar{X} = 0.0907 - (1.7068)(0.0175) = 0.0608$$

$$\Rightarrow \hat{y}_i = 1.7068x_i + 0.0608$$

2. Calcule r^2. ¿Cómo interpretamos?

Ejemplo:

Los datos en los últimos 38 años se resumen como sigue:
$\bar{X} = 0.0175$, $\bar{Y} = 0.0907$, $s_x = 0.0536$, $s_y = 0.1535$, y $r = 0.596$

1. Calcule la linea de regresión:

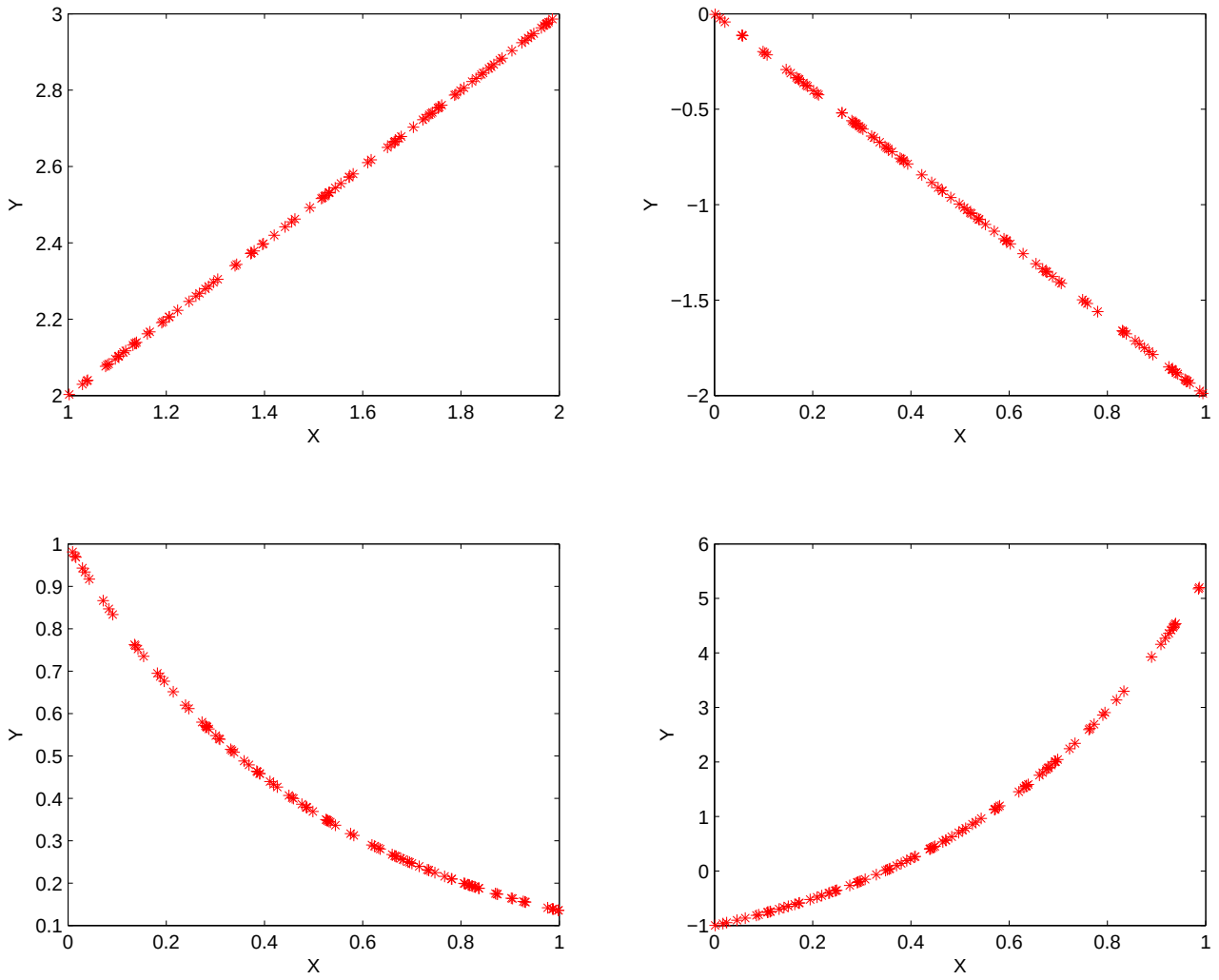
$$\alpha = r\frac{s_y}{s_x} = (0.596)\frac{0.1535}{0.0536} = 1.7068$$
$$\beta = \bar{Y} - \alpha\bar{X} = 0.0907 - (1.7068)(0.0175) = 0.0608$$

$$\Rightarrow \hat{y}_i = 1.7068x_i + 0.0608$$

2. Calcule r^2. ¿Cómo interpretamos? $r^2 = 0.3552$. O sea un 35.5% de la variación en Y se explica (linealmente) por la variación en X.

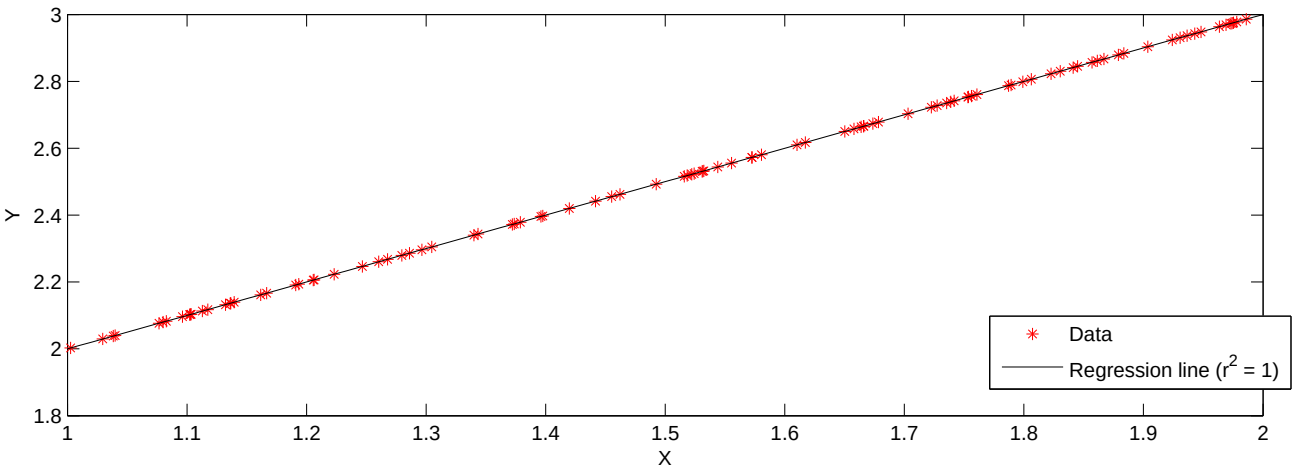
Otro ejemplo

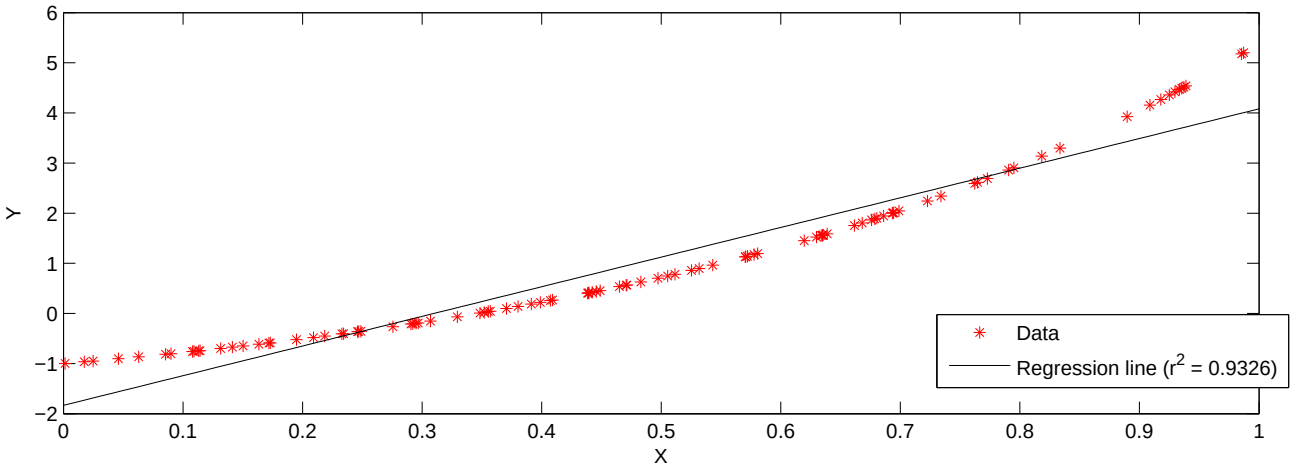
Considere los datos del primer y del cuarto panel del ejemplo 1:



Otro ejemplo

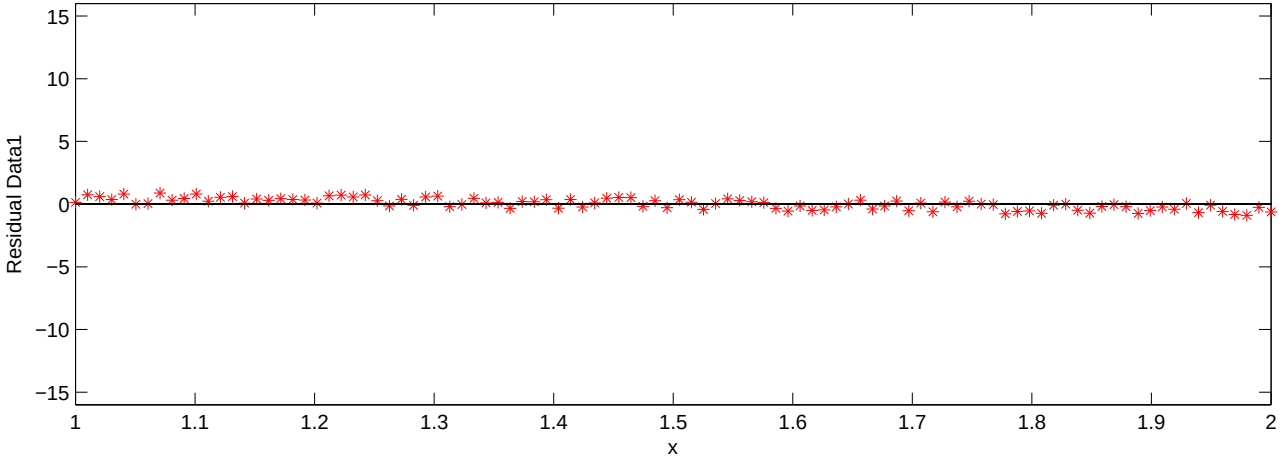
Usando "software" obtenemos lo siguiente

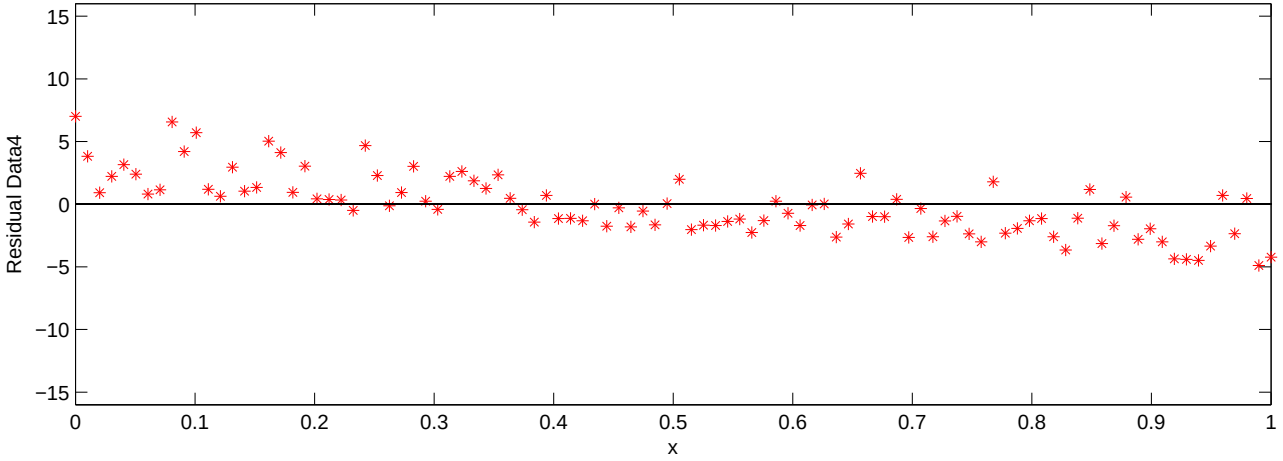




Otro ejemplo

Usando “software” obtenemos lo siguiente





Otro ejemplo

Aunque el ejemplo se hizo usando la palataforma de programación MATLAB, existen otras herramientas:

- lenguage de programación R.

Otro ejemplo

Aunque el ejemplo se hizo usando la palataforma de programación MATLAB, existen otras herramientas:

- lenguage de programación R.
- Excell (spreadsheet)

Otro ejemplo

Aunque el ejemplo se hizo usando la palataforma de programación MATLAB, existen otras herramientas:

- lenguage de programación R.
- Excell (spreadsheet)
- Calc (spreadsheet).

Detalles sobre correlación y regresión lineal

1. Regresión y correlación describen relaciones **lineales**.

Detalles sobre correlación y regresión lineal

1. Regresión y correlación describen relaciones **lineales**.
2. No son medidas robustas, sino sensitivas a outliers y valores extremos.

Detalles sobre correlación y regresión lineal

1. Regresión y correlación describen relaciones **lineales**.
2. No son medidas robustas, sino sensitivas a outliers y valores extremos.
3. La regresión se construye para el rango donde hay datos. ¡Cuidado con la extrapolación! Esto es, hacer predicciones más allá del rango para el cual hay datos para la var. ind.

Detalles sobre correlación y regresión lineal

1. Regresión y correlación describen relaciones **lineales**.
2. No son medidas robustas, sino sensitivas a outliers y valores extremos.
3. La regresión se construye para el rango donde hay datos. ¡Cuidado con la extrapolación! Esto es, hacer predicciones más allá del rango para el cual hay datos para la var. ind.
4. Correlaciones basadas en datos que son promedios tienden a resultar en mayor correlación que si se usan dato de individuos.

Detalles sobre correlación y regresión lineal

1. Regresión y correlación describen relaciones **lineales**.
2. No son medidas robustas, sino sensitivas a outliers y valores extremos.
3. La regresión se construye para el rango donde hay datos. ¡Cuidado con la extrapolación! Esto es, hacer predicciones más allá del rango para el cual hay datos para la var. ind.
4. Correlaciones basadas en datos que son promedios tienden a resultar en mayor correlación que si se usan dato de individuos.
5. Regresión y correlación se enfocan en relaciones entre **dos** variables. Pero en ocaciones la relación se explica por una tercera variable (lurking variable) que no estuvo considerada en el estudio.

Detalles sobre correlación y regresión lineal

1. Regresión y correlación describen relaciones **lineales**.
2. No son medidas robustas, sino sensitivas a outliers y valores extremos.
3. La regresión se construye para el rango donde hay datos. ¡Cuidado con la extrapolación! Esto es, hacer predicciones más allá del rango para el cual hay datos para la var. ind.
4. Correlaciones basadas en datos que son promedios tienden a resultar en mayor correlación que si se usan dato de individuos.
5. Regresión y correlación se enfocan en relaciones entre **dos** variables. Pero en ocaciones la relación se explica por una tercera variable (lurking variable) que no estuvo considerada en el estudio.
6. Asociación (correlación) no implica causalidad (mire ejem. 2.18).

Preguntas...