Correlación y regresión

Contenidos adaptados del libro “Probability and statistics for computer scientists, Second edition, M. Baron” (Capítulo 11)

Contenido

1. Objetivos
2. Introducción
3. Coeficiente de correlación (de Pearson)
4. Modelo de regresión
5. Regresión lineal simple
6. Regresión lineal múltiple
7. Resumen

1. Objetivos

- Comprender los conceptos de correlación y regresión (RA1)
- Aplicar técnicas de regresión y correlación para establecer relaciones entre variables (RA4)

2. Introducción

- El resultado de la inferencia estadística puede ser
  - Un intervalo de confianza o conjunto de posibles valores para las propiedades estimadas, junto con una probabilidad de error cometido
  - La comprobación de una hipótesis sobre los valores para las propiedades estimadas, junto con una probabilidad de error cometido
  - Un modelo de predicción (regresión) del valor de una propiedad a partir de otra.
- En esta presentación se estudian los modelos de regresión

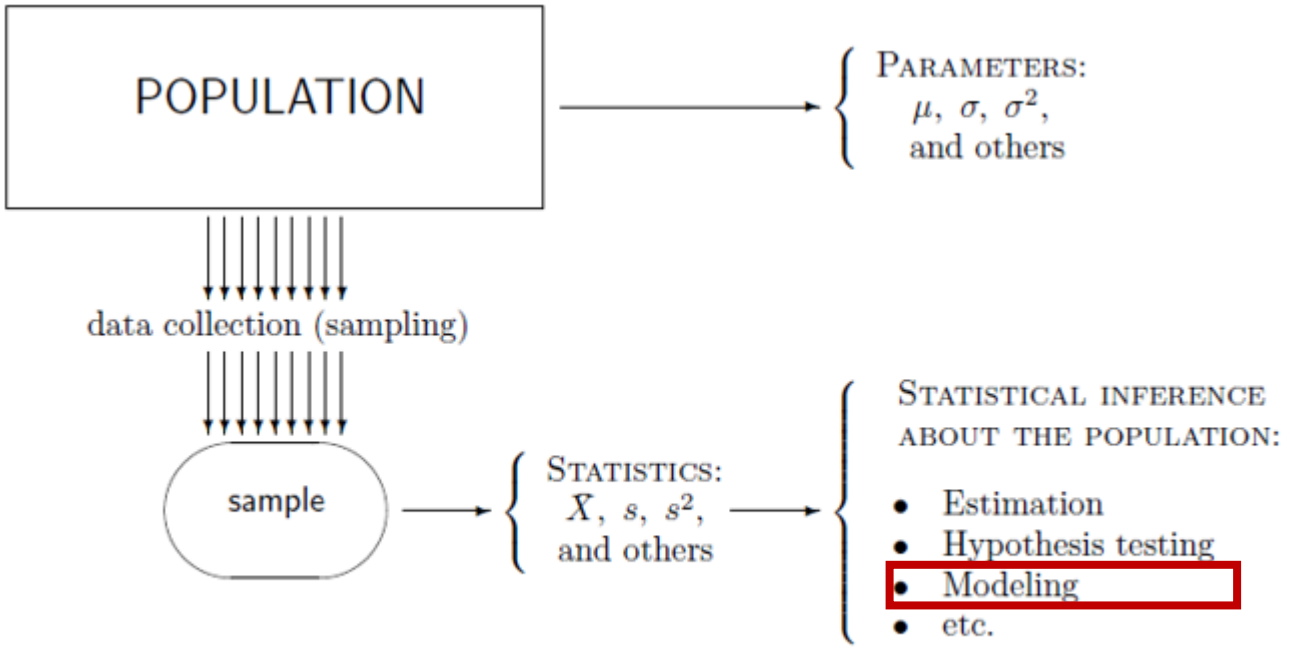


FIGURE 8.1: *Population parameters and sample statistics.*

2. Introducción
Utilidad

- Muchas variables observadas en la vida real están correlacionadas.
- El tipo de su relación a menudo se puede expresar en una forma matemática llamada regresión.
- Establecer y probar tal relación nos permite:
  - comprender las interacciones, causas y efectos entre variables;
  - predecir variables no observadas basadas en las observadas;
  - determinar qué variables afectan significativamente a la variable de interés

2. Introducción
Definiciones

- **Correlación**: “Correspondencia o relación recíproca entre dos o más cosas o series de cosas” (RAE).
  - En Estadística, se refiere al grado y signo de la relación entre dos (o más) variables (estadísticas o aleatorias)
  - La correlación se mide con el **coeficiente de correlación**, que es un valor entre -1 y +1.
- **Regresión**: “Retrocesión o acción de volver hacia atrás” (RAE)
  - En Estadística, el término "regresión" fue acuñado por Francis Galton en el siglo XIX para describir el fenómeno biológico de que las alturas de los descendientes de ancestros altos tienden a regresar hacia abajo, hacia un promedio normal (un fenómeno conocido como regresión hacia la media).
  - Actualmente, la regresión en Estadística se refiere a la forma de correlación entre dos (o más) variables, es decir a la forma de obtener (regresar a) la media de los posibles valores de una de las variables a partir de un valor dado de la otra.
  - La regresión se representa mediante una función matemática (**modelo de regresión**).

3. Coeficiente de correlación (de Pearson)

- El coeficiente de correlación (de Pearson) es una medida de dependencia lineal entre dos variables cuantitativas que representan dos propiedades de los elementos o individuos de una población (peso, altura, edad, etc.)
- Cuando se aplica a toda la población, se denomina coeficiente de correlación poblacional (ρ).
- Cuando se aplica a una muestra de la población, se denomina coeficiente de correlación muestral (r).

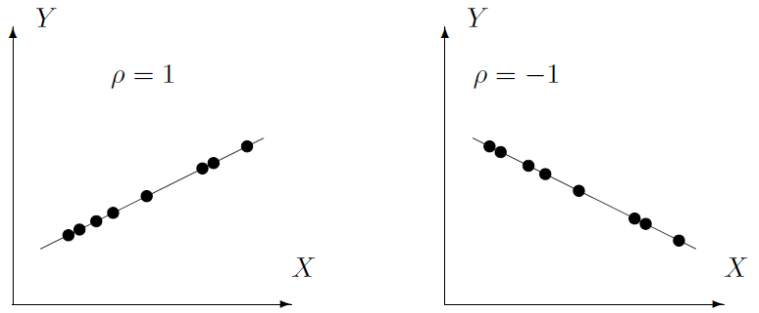


FIGURE 3.5: *Perfect correlation:* $\rho = \pm 1$.

3. Coeficiente de correlación (de Pearson) Cálculo del coeficiente de correlación muestral

- Dada una muestra de pares de valores: $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ de dos variables X e Y
  - $r =$ coeficiente de correlación muestral $= \frac{s_{xy}}{s_x s_y}$
- Donde
  - $s_{xy} =$ covarianza muestral $= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
  - $s_x =$ desviación estándar muestral de X $= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
  - $s_y =$ desviación estándar muestral de Y $= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$
- Valores posibles
  - $-1 \leq r \leq 1$
- Valores de r cercanos a:
  - 1 indican una fuerte correlación lineal positiva
  - – 1 muestran una fuerte correlación lineal negativa
  - 0 muestran una correlación débil o ninguna correlación
- $|r| = 1$ es posible sólo cuando todos los valores de X e Y se encuentran en una línea recta.

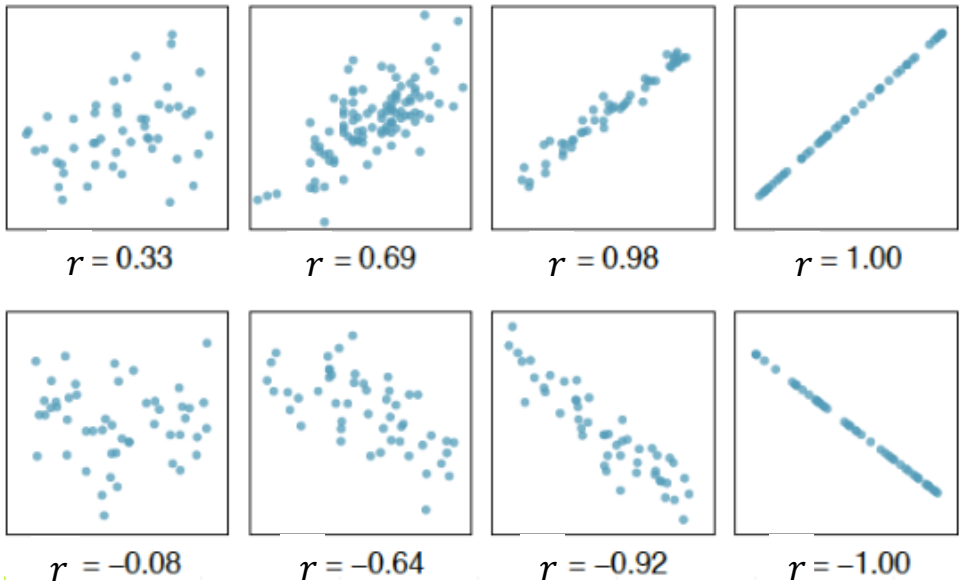


Imagen: RPubs

4. Modelo de regresión

- **Un modelo de regresión** relaciona una variable **respuesta** con una o varias variables **predictoras** (o predictores)
- Si Y es una variable respuesta y $X^{(1)}, \dots, X^{(k)}$ son predictores:
- El **modelo de regresión** de Y sobre $X^{(1)}, \dots, X^{(k)}$ es la función G que representa la esperanza condicional de Y para unos valores dados de los predictores:
  - $G\left(x^{(1)}, \dots, x^{(k)}\right) = E\left\{Y | X^{(1)} = x^{(1)}, \dots, X^{(k)} = x^{(k)}\right\}$
  - Es una función de $x^{(1)}, \dots, x^{(k)}$ cuya forma se puede estimar a partir de los datos de una muestra.
- Si sólo hay un predictor, se denomina **modelo de regresión simple**
  - $G(x) = E\{Y | X = x\}$

4. Modelo de regresión
Ejemplo 11.1 (Población mundial)

- Variable respuesta = Población del mundo
- Predictor = Año
- Modelo de regresión simple de la población mundial sobre el año:
  - $Población \approx G(año)$
- La población aumenta cada año, y su crecimiento es casi lineal, por lo que *G* es una función lineal.
- Las previsiones para el año 2020 pueden estimarse calculando $G(2020)$.

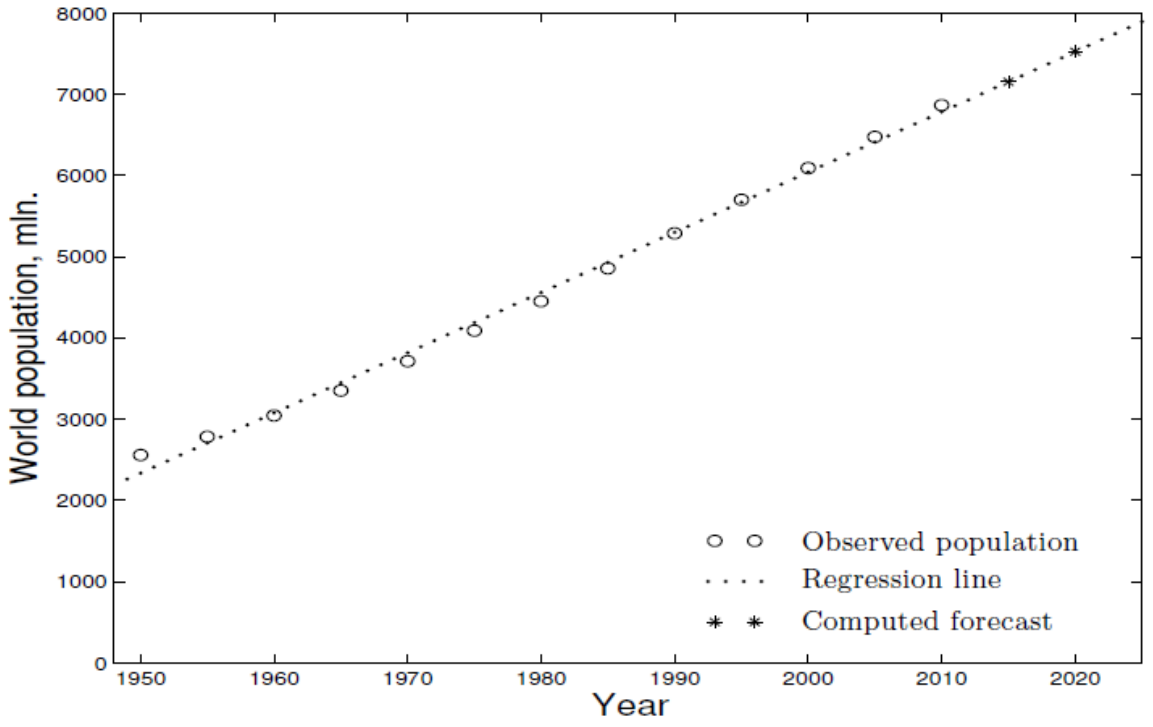


FIGURE 11.1: *World population in 1950–2010 and its regression forecast for 2015 and 2020.*

4. Modelo de regresión
Ejemplo 11.2 (Precios de viviendas)

- Variable respuesta = precio de la vivienda
- Predictor = superficie de la vivienda
- Modelo de regresión simple del precio sobre la superficie:
  - $Precio \approx G(superficie)$
- La tendencia no parece lineal, por lo que *G* no es una función lineal.
- El pronóstico para el precio de una vivienda con una superficie dada se puede estimar, pero la estimación no será tan precisa como en el ejemplo 11.1.

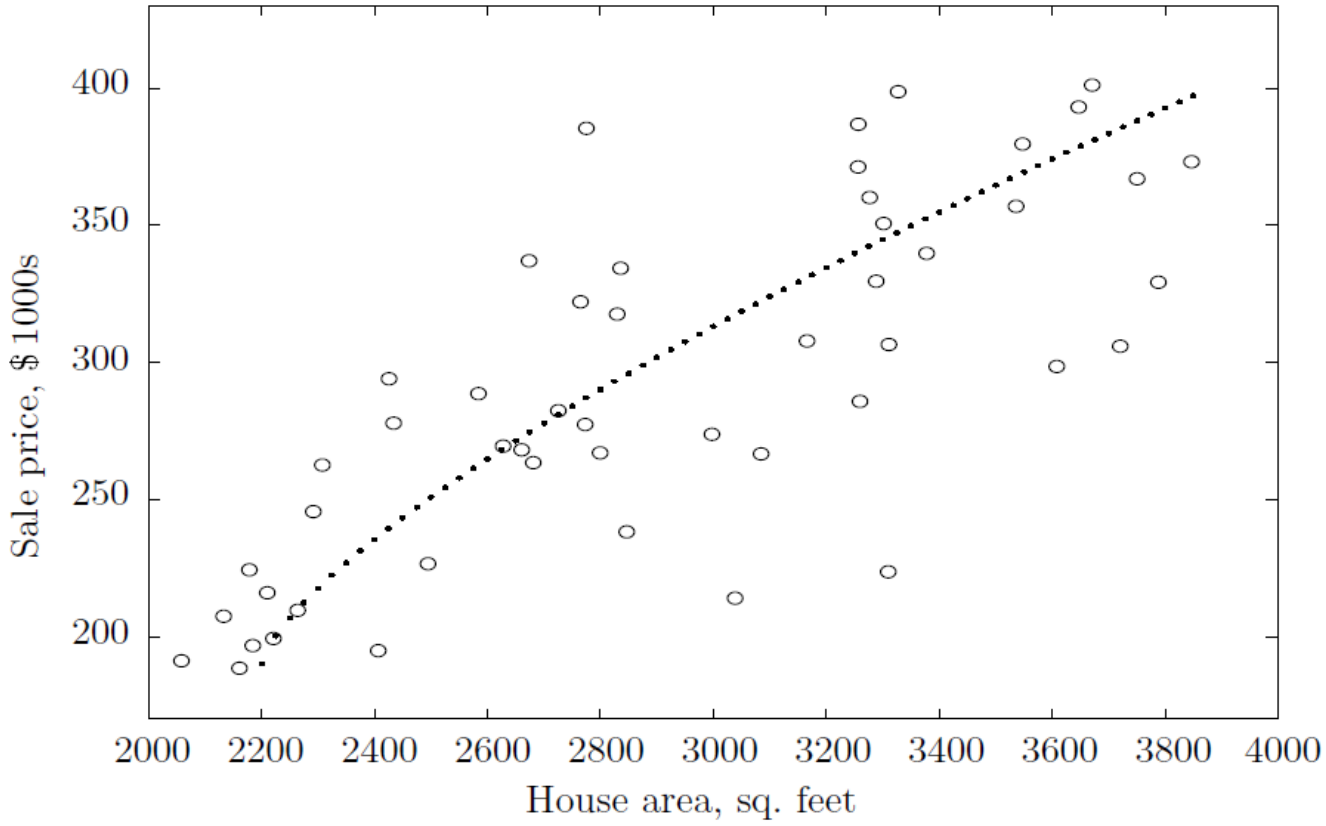


FIGURE 11.2: *House sale prices and their footage.*

4. Modelo de regresión
Regresión múltiple

- Para mejorar la estimación en el ejemplo 11.2, se pueden tener en cuenta otros factores:
  - el número de dormitorios y baños, la superficie del jardín, el salario promedio de los vecinos del barrio, etc.
- Si todas las variables añadidas son relevantes para fijar el precio de una vivienda, el modelo tendrá un ajuste más cercano y proporcionará predicciones más precisas.
- Los modelos de regresión con varios predictores son modelos de regresión múltiple.

5. Regresión lineal simple

- Un modelo de regresión lineal simple asume que la esperanza condicional es una función lineal de x
  - $G(x) = E\{Y \mid X = x\} = \beta_0 + \beta_1 x$
- Se denomina **recta de regresión**
- Si calculásemos el valor (y) de la variable Y a partir de un valor (x) del predictor X usando la recta de regresión, el valor correcto de Y sería el de la esperanza que representa G(x) más un error aleatorio ε que se comete si el resultado no está en la recta
  - $y = G(x) + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$
- Se puede suponer que el error ε es una variable aleatoria Normal con esperanza 0 y varianza desconocida.

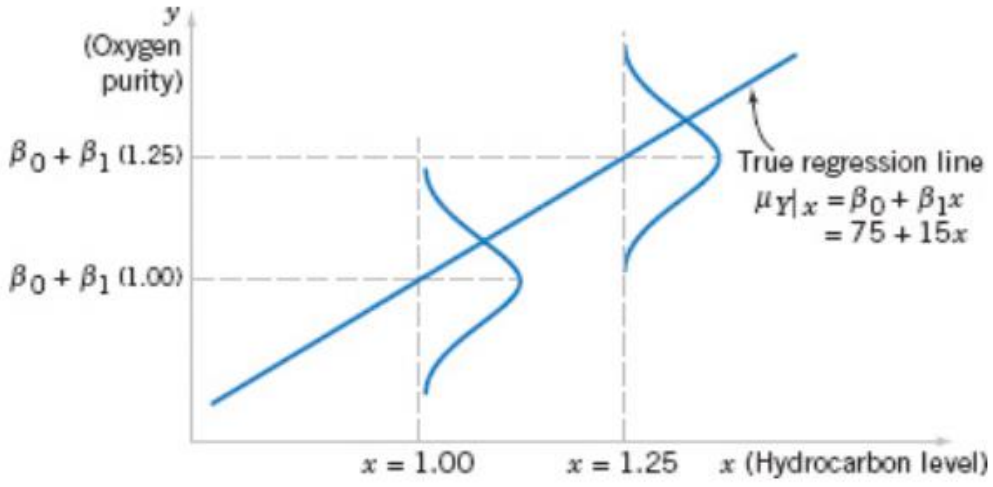
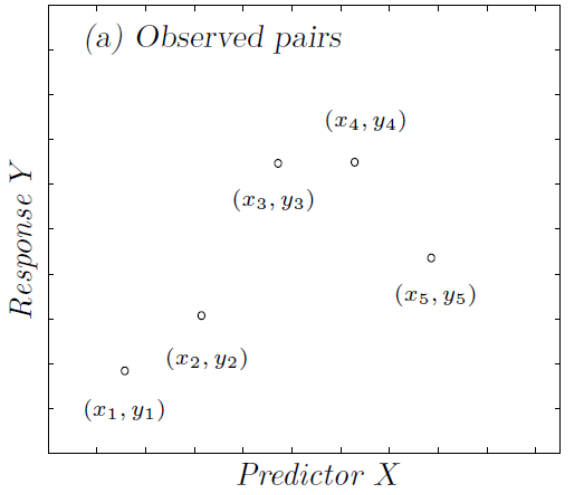
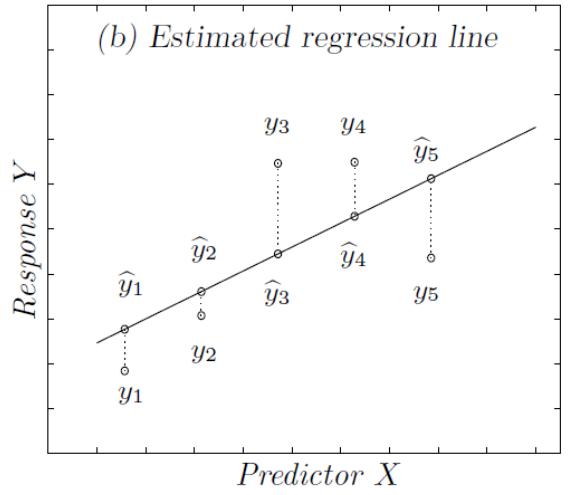


Imagen: Runger y Montgomery

5. Regresión lineal simple
Método de mínimos cuadrados

- No se conocen β_0 y β_1, pero se pueden estimar a partir de una muestra de datos
- Se utiliza el método de mínimos cuadrados para minimizar el error ε o distancia a la recta de regresión, y poder estimar el valor de y para un x dado
  - $\hat{y} = \hat{G}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
- El método de los mínimos cuadrados establece cómo calcular los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ para minimizar el error.





5. Regresión lineal simple
Estimación de la variable respuesta

- Para estimar el valor de la variable Y para un valor dado de X, podemos usar como estimador el modelo de regresión:
  - $\hat{y} = \hat{G}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
- Llamamos b_0 y b_1 a los estimadores de β_0 y β_1, que pueden calcularse de la siguiente forma, según el método de mínimos cuadrados:
  - $\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x}$
  - $\hat{\beta}_1 = b_1 = \frac{s_{xy}}{s_x^2}$
- Donde
  - $s_{xy} = \text{covarianza muestral} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
  - $s_x^2 = \text{varianza muestral} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- Por tanto, se puede estimar el valor de la variable respuesta (y)
  - $\hat{y} = b_0 + b_1 x$

5. Regresión lineal simple
Estimación usando el coeficiente de correlación

- Si se dispone del valor del coeficiente de correlación muestral r, el valor b_1 de la recta de regresión puede calcularse como
  - $b_1 = \frac{s_{xy}}{s_x^2} = r\left(\frac{s_y}{s_x}\right)$
- Donde
  - $s_x =$ desviación estándar muestral de X $= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
  - $s_y =$ desviación estándar muestral de Y $= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$

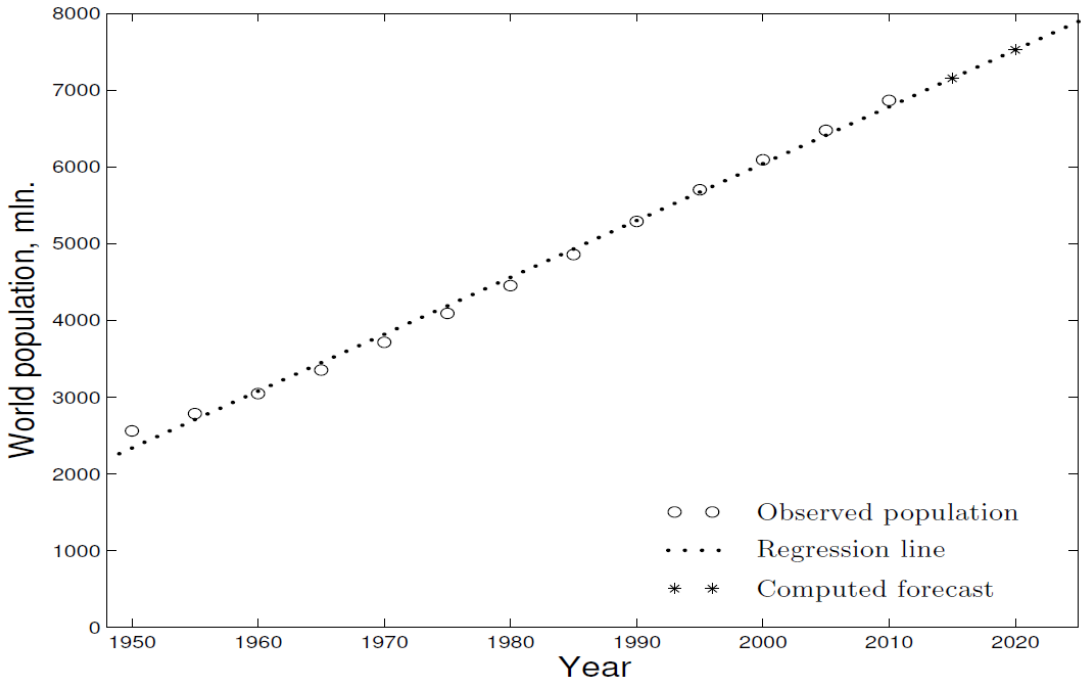
5. Regresión lineal simple
Ejemplo 11.3 (Población mundial)

- Según la Base Internacional de Datos de la Oficina del Censo de los Estados Unidos, la población mundial crece según esta tabla.
- ¿Cómo podemos utilizar estos datos para predecir la población mundial en los años 2015 y 2020?

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287	2015	?
1970	3712	1995	5700	2020	?

5. Regresión lineal simple
Ejemplo 11.3 (Solución)

- $X = Año \rightarrow Muestra: x = (1950, 1955, 1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, 2010)$
- $Y = Población\ mundial \rightarrow Muestra: y = (2558, 2782, 3043, 3350, 3712, 4089, 4451, 4855, 5287, 5700, 6090, 6474, 6864)$
- $\bar{x} = \frac{\sum_{i=1}^{13} x_i}{13} = 1980$
- $\bar{y} = \frac{\sum_{i=1}^{13} y_i}{13} = 4558.1$
- $s_x^2 = \frac{\sum_{i=1}^{13} (x_i - 1980)^2}{13-1} = 379.2$
- $s_{xy} = \frac{\sum_{i=1}^{13} (x_i - 1980)(y_i - 4558.1)}{13-1} = 28104.2$
- $b_1 = \frac{s_{xy}}{s_x^2} = \frac{28104.2}{379.2} = 74.1$
- $b_0 = \bar{y} - b_1 \bar{x} = 4558.1 - (74.1)(1980) = -142201$
- $\hat{y} = \hat{G}(x) = b_0 + b_1 x = -142201 + 74.1x$
- $\hat{G}(2015) = -142201 + 74.1 \cdot 2015 = 7152$ millones de personas
- $\hat{G}(2020) = -142201 + 74.1 \cdot 2020 = 7523$ millones de personas



5. Regresión lineal simple
Coeficiente de determinación (R-cuadrado)

- El coeficiente de determinación (denominado R-cuadrado o R^2) es la proporción de la variación total explicada por el modelo de regresión.
- Siempre está entre 0 y 1, con valores altos que generalmente sugieren un buen ajuste del modelo.
- En la regresión lineal simple, el coeficiente de determinación es igual al coeficiente de correlación muestral elevado al cuadrado
  - $R^2 = r^2$
- Ejemplo 11.3
  - $R^2 = r^2 = \frac{(s_{xy})^2}{s_x^2 s_y^2} = \frac{(28104.2)^2}{(379.2)(2092943)} = 0.995 \text{ o } 99.5\%$
  - s_{xy} y s_x^2 se calcularon anteriormente, y $s_y^2 = \frac{\sum_{i=1}^{13}(y_i - 4558.1)^2}{13-1} = 2092943$
  - El valor del 99.5% del coeficiente de determinación indica que el modelo es un ajuste muy bueno, aunque alguna parte del 0,5 % restante de la variación total todavía puede mejorarse agregando términos no lineales al modelo.

5. Regresión lineal simple
Contrastes de hipótesis

- Para evaluar la bondad del ajuste de un modelo de regresión lineal, se pueden contrastar hipótesis sobre los parámetros del modelo, y se pueden construir intervalos de confianza.
- Para probar hipótesis sobre β_0, β_1 o sobre el coeficiente de correlación (ρ), se debe suponer que el componente de error en el modelo tiene una distribución normal:
  - $H_0: \beta_0 = \text{valor} \quad \text{vs} \quad H_A: \beta_0 \neq \text{valor}$
  - $H_0: \beta_1 = \text{valor} \quad \text{vs} \quad H_A: \beta_1 \neq \text{valor}$
  - $H_0: \rho = \text{valor} \quad \text{vs} \quad H_A: \rho \neq \text{valor}$
  - …
- Nota: Se recomienda leer la sección 11.2 del libro de referencia, y también el capítulo 11 del libro Estadística aplicada y probabilidad para ingenieros, 5ª edición, de G. Runger y D. Montgomery

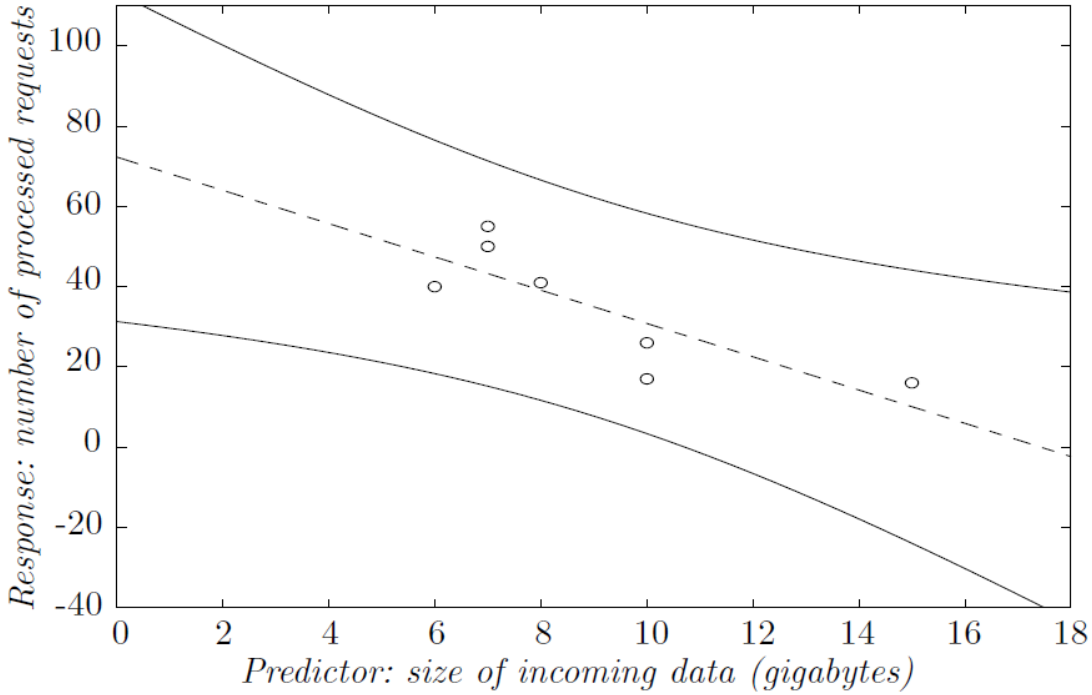


FIGURE 11.5: *Regression prediction of program efficiency.*

5. Regresión lineal simple
Ejemplos de la vida real (Artículos)

- Ergonomía:
  - Limon, M., et al. (2023). Development of new shoe sizing system for women based on regression analysis of foot shapes. International Journal of Industrial Ergonomics, 94, 103408.
- Geografía:
  - Suravi, r., et al. (2023). A study on comparative analysis to predict the future population growth in India, Bangladesh, and Pakistan. International Journal of Mathematics and Computer, Vol. 13, Issue 1, 1–16.
- Redes sociales:
  - Ramadan, A. (2023). The Relationship between Social Media use and Depression Symptoms in Jazan region In the Point of View of the Social Work Profession. Egyptian Journal of Social Work.
- Economía:
  - Rehman, M. (2022). Correlation of Workplace surveillance with Psychological Health, Productivity, and Privacy of employees. International Journal of Scientific & Engineering Research, Vol.13, Issue 11.

5. Regresión lineal simple
Ejercicios propuestos

- Ejercicios 11.1, 11.2(a), 11.3(a), 11.8(a,b), 11.10(a) del libro
  - La respuesta de los ejercicios 11.2 y 11.10 está disponible en el libro

6. Regresión lineal múltiple

- Un modelo de regresión lineal múltiple relaciona una variable respuesta Y con varios predictores $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$.
  - $\hat{y} = \hat{G}(x) = b_0 + b_1 x^{(1)} + b_2 x^{(2)} + \cdots + b_k x^{(k)}$
- Como en el caso de la regresión lineal simple, se puede utilizar el método de mínimos cuadrados para calcular $b_1, b_2, \ldots, b_k$ y el valor en el origen b_0 que minimice el error o distancia a la función de regresión.

- Nota: Se recomienda leer la sección 11.3 del libro de referencia, y también el capítulo 12 del libro Estadística aplicada y probabilidad para ingenieros, 5ª edición, de G. Runger y D. Montgomery

6. Regresión lineal múltiple
Ejemplos de la vida real (Artículos)

- Ingeniería del software:
  - Sharma, A., et al. (2020). Linear regression model for agile software development effort estimation. In 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) (pp. 1-4). IEEE.
- Ingeniería civil:
  - Plebankiewicz, E., et al. (2023). Analysis and Prediction of Universities' Buildings' Renovation Costs Using a Regression Model. Applied Sciences, 13(1), 401.
- Marketing:
  - Dewi, I., et al. (2023). The Factors That Influence TikTok Popularity As A Digital Marketing Technique To Grow Customer Engagement. International Journal of Economics, Business and Innovation Research, 2(02), 103-111.
- Agricultura:
  - Han, X., et al. (2023). Effects of Meteorological Factors on Apple Yield Based on Multilinear Regression Analysis: A Case Study of Yantai Area, China. Atmosphere, 14(1), 183.

7. Resumen

- Los modelos de regresión proporcionan métodos para estimar relaciones matemáticas entre una o varias variables predictoras y una variable de respuesta.
- Los resultados se utilizan para explicar el comportamiento de la respuesta y para predecir su valor para cualquier nuevo conjunto de predictores.
- El método de mínimos cuadrados se utiliza para estimar los parámetros de regresión.
- Para evaluar la bondad del ajuste de un modelo de regresión lineal, se pueden probar hipótesis sobre los parámetros del modelo, y se pueden construir intervalos de confianza.