# Estadística I
# Tema 2: Análisis de datos bivariantes

Grado en Administración de Empresas 08/09

1. a) Distribuciones de frecuencias marginales relativas:

| # de h: \ nota: | Suspenso | Aprobado | Notable | Sobresaliente | D. marg. de # de h |
|---|---|---|---|---|---|
| 2 | 0.20 | 0.15 | 0.08 | 0.03 | 0.46 |
| 3 | 0.12 | 0.07 | 0.02 | 0.02 | 0.23 |
| 4 | 0.04 | 0.10 | 0.02 | 0.00 | 0.16 |
| 5 | 0.00 | 0.05 | 0.05 | 0.05 | 0.15 |
| D. marg. de nota | 0.36 | 0.37 | 0.17 | 0.10 | 1 |

b) Distribuciones de "notaçondicionadas a los distintos valores de "número de horas de estudio":

| Nota \| # horas= 2: | Suspenso | Aprobado | Notable | Sobresaliente | Total |
|---|---|---|---|---|---|
| $fr_{x_i\|y=2}$ | 0.435 | 0.326 | 0.174 | 0.065 | 1 |

| Nota \| # horas= 3: | Suspenso | Aprobado | Notable | Sobresaliente | Total |
|---|---|---|---|---|---|
| $fr_{x_i\|y=3}$ | 0.522 | 0.304 | 0.87 | 0.87 | 1 |

| Nota \| # horas= 4: | Suspenso | Aprobado | Notable | Sobresaliente | Total |
|---|---|---|---|---|---|
| $fr_{x_i\|y=4}$ | 0.250 | 0.625 | 0.125 | 0.000 | 1 |

| Nota \| # horas= 5 | Suspenso | Aprobado | Notable | Sobresaliente | Total |
|---|---|---|---|---|---|
| $fr_{x_i\|y=5}$ | 0.000 | 0.333 | 0.333 | 0.333 | 1 |

Distribuciones de "número de horas de estudioçondicionadas a los distintos valores de "nota":

| # horas\| nota = Suspenso | $fr_{y_j\|x=Sus.}$ |
|---|---|
| 2 | 0.556 |
| 3 | 0.333 |
| 4 | 0.111 |
| 5 | 0.000 |
| Total | 1 |

| # horas\| nota = Aprobado | $fr_{y_j\|x=Apr.}$ |
|---|---|
| 2 | 0.405 |
| 3 | 0.189 |
| 4 | 0.270 |
| 5 | 0.135 |
| Total | 1 |

| # horas\| nota = Notable | $fr_{y_j\|x=Not.}$ |
|---|---|
| 2 | 0.471 |
| 3 | 0.118 |
| 4 | 0.118 |
| 5 | 0.294 |
| Total | 1 |

| # horas\| nota = Sobresaliente | $fr_{y_j\|x=Sob.}$ |
|---|---|
| 2 | 0.3 |
| 3 | 0.2 |
| 4 | 0.0 |
| 5 | 0.5 |
| Total | 1 |

2. a) Distribuciones de frecuencias marginales relativas:

| # de hijos \ renta: | 0-1000 | 1000-2000 | 2000-3000 | $> 3000$ | D. marg. de # de hijos |
|---|---|---|---|---|---|
| 0 | 0.15 | 0.05 | 0.03 | 0.02 | 0.25 |
| 1 | 0.10 | 0.20 | 0.10 | 0.05 | 0.45 |
| 2 | 0.05 | 0.10 | 0.05 | 0.03 | 0.23 |
| $\geq 3$ | 0.02 | 0.03 | 0.02 | 0.00 | 0.07 |
| D. marg. de renta | 0.32 | 0.37 | 0.18 | 0.12 | 1 |

b) Distribución condicionada de $Y|X = 2$:

| renta \| # hijos= 2: | 0-1000 | 1000-2000 | 2000-3000 | $> 3000$ | Total |
|---|---|---|---|---|---|
| $fr_{y_i\|x=2}$ | 0.218 | 0.435 | 0.217 | 0.130 | 1 |

c) Distribución condicionada de $X|1000 < Y < 2000$:

| # hijos\| renta= $1000 < Y < 2000$ | $fr_{x_i\|1000<y<2000}$ |
|---|---|
| 0 | 0.135 |
| 1 | 0.541 |
| 2 | 0.270 |
| $\geq 3$ | 0.054 |
| Total | 1 |

3. a) Distribución conjunta de frecuencias absolutas:

| | | Núm. compras por semana | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| Núm. tarjetas | 1 | 24 | 39 | 27 | 18 | 9 |
| | 2 | 9 | 24 | 24 | 27 | 21 |
| | 3 | 3 | 9 | 18 | 24 | 24 |

b) Distribución marginal de Y:

| Núm. compras por semana | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| $n_j$ | 36 | 72 | 69 | 69 | 54 | 300 |

Media del número de compras por semana:

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{5} y_j \cdot n_j = (0 \cdot 36 + 1 \cdot 72 + 2 \cdot 69 + 3 \cdot 69 + 4 \cdot 54)/300 = 2.11.$$

Varianza del número de compras por semana:

$$s_y^2 = \frac{1}{300}\sum_{j=1^5} y_j^2 \cdot n_j - \overline{y}^2 = (0^2 \cdot 36 + 1^2 \cdot 72 + 2^2 \cdot 69 + 3^2 \cdot 69 + 4^2 \cdot 54)/300 - 2.11^2 = 1.6579.$$

Desviación típica del número de compras por semana: $s_y = \sqrt{s_y^2} = \sqrt{1.6579} = 1.29$.

c) Distribución del número de tarjetas de crédito:

| # tarjetas de crédito | $n_i$ |
|---|---|
| 1 | 122 |
| 2 | 107 |
| 3 | 81 |
| Total | 300 |

Número más frecuente de tarjetas de crédito (moda): 1.

d) Distribución del número de compras semanales pagadas con tarjetas de crédito que realizan las personas que poseen tres tarjetas:

| Núm. compras por semana \| num. tarjetas=3 | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| $fr_{y_j\|x=3}$ | 0.037 | 0.111 | 0.222 | 0.296 | 0.296 | 1 |

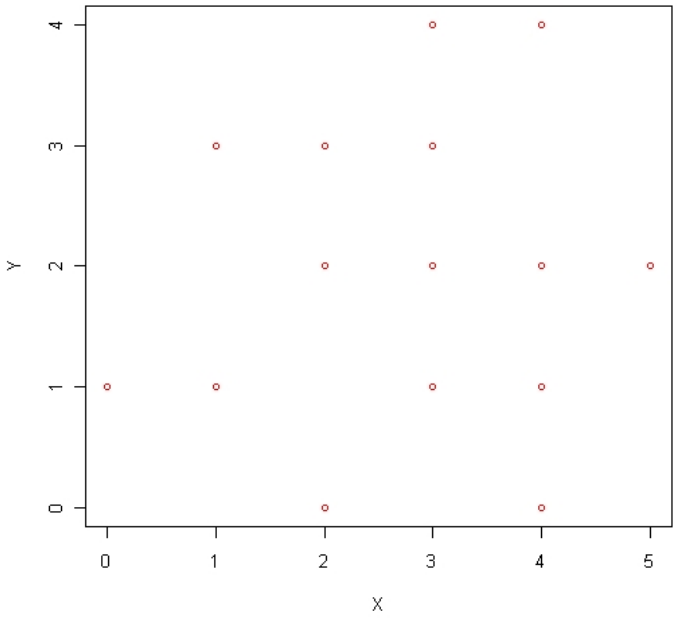
Media de esta distribución:

$$\overline{y}|x = 3 = \sum_{j=1}^{5} y_j \cdot fr_{y_j|x=3} = 0 \cdot 0.037 + 1 \cdot 0.111 + 2 \cdot 0.222 + 3 \cdot 0.296 + 4 \cdot 0.296 = 2.627.$$

4. a) Tabla de doble entrada (distribución conjunta de frecuencias):

| X: \ Y: | 0 | 1 | 2 | 3 | 4 | D. marg. de X |
|---|---|---|---|---|---|---|
| 0 | 0 | 4 | 0 | 0 | 0 | 4 |
| 1 | 0 | 3 | 0 | 4 | 0 | 7 |
| 2 | 2 | 0 | 9 | 3 | 0 | 14 |
| 3 | 0 | 6 | 12 | 5 | 2 | 25 |
| 4 | 2 | 7 | 15 | 0 | 1 | 25 |
| 5 | 0 | 0 | 5 | 0 | 0 | 5 |
| D. marg. de Y | 4 | 20 | 41 | 12 | 3 | 80 |

Diagrama de dispersión:




b) Tanto la covarianza como el coeficiente de correlación han de ser positivos ya que las dos variables parecen tener una relación creciente. Además, sobre el valor del coeficiente de correlación, podemos decir que no estará próximo a 1, ya que la relación lineal entre las dos variables no parece muy fuerte.

c)

$$r_{(x,y)} = \frac{Cov(s,y)}{s_x \cdot s_y} \qquad Cov(x,y) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{xy}\right)$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{80}(0\cdot 4 + 1\cdot 7 + 2\cdot 14 + 3\cdot 25 + 4\cdot 25 + 5\cdot 5) = 2.9375.$$

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{1}{80}(0\cdot 4 + 1\cdot 20 + 2\cdot 41 + 3\cdot 12 + 4\cdot 3) = 1.875.$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{80}(0^2\cdot 4 + 1^2\cdot 7 + 2^2\cdot 14 + 3^2\cdot 25 + 4^2\cdot 25 + 5^2\cdot 5) - 2.9375^2 = 1.5336.$$

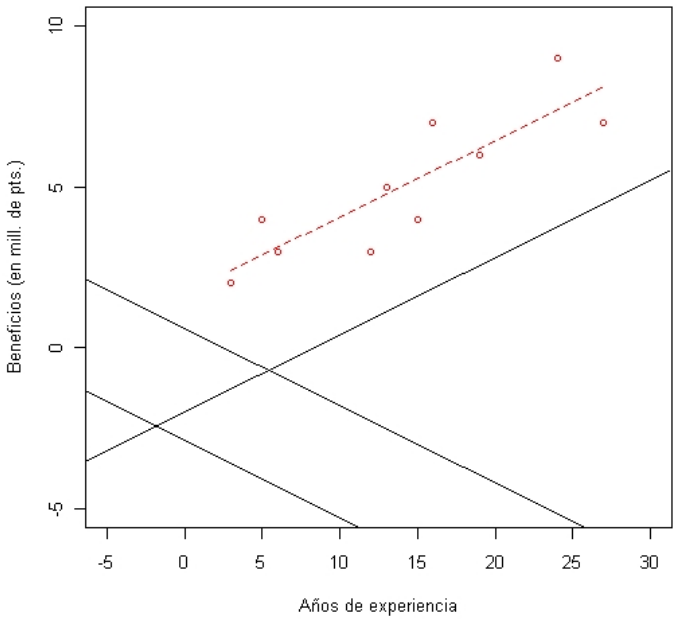
$$s_y^2 = \frac{1}{n}\sum_{j=1}^{n} y_j^2 - \overline{y}^2 = \frac{1}{80}(0^2\cdot 4 + 1^2\cdot 20 + 2^2\cdot 41 + 3^2\cdot 12 + 4^2\cdot 3) - 1.875^2 = 0.7344.$$

$$Cov(x,y) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i \cdot y_i - n\cdot\overline{x}\cdot\overline{y}\right) = \frac{1}{79}(0\cdot 0\cdot 0 + 0\cdot 1\cdot 4 + \cdots + 5\cdot 4\cdot 0 - 80\cdot 2.9375\cdot 1.5336) = 0.0174.$$

$$r_{(x,y)} = \frac{Cov(x,y)}{s_x \cdot s_y} = \frac{0.0174}{\sqrt{1.5336} \cdot \sqrt{0.7344}} = 0.0162.$$

Como habíamos predicho, obtenemos valores positivos para la covarianza y el coeficiente de correlación. El valor del coeficiente de correlación es muy cercano a cero, lo que indica que prácticamente no hay relación lineal entre estas dos variables.

5. a) y b) Diagrama de dispersión y rectas:



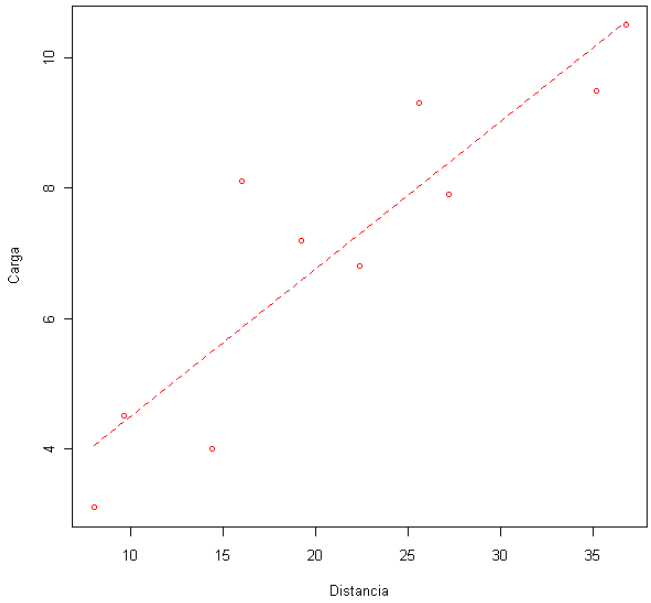

c) La recta de regresión parece ser $y = 1.66 + 0.24x$, es decir, la que aparece punteada en el gráfico anterior.

d) Coeficiente de correlación (los cálculos se harían como en el ejercicio 4): $r_{(x,y)} = 0.8587$.

e) La relación linea es positiva, es decir, a mayores valores de $x$, mayores valores de $y$, ya que el coeficiente de correlación es positivo. Además, como toma un valor alto (próximo a 1) podemos decir que la relación lineal es fuerte.

f) Recta de regresión de los años de experiencia ($x$) en función de los beneficios ($y$): $x = c + dy$ donde $d = \dfrac{Cov(x,y)}{s_y^2} = 3.0909$ y $c = \overline{x} - d\overline{y} = -1.4545$. La interpretación de la pendiente sería que un aumento de 1 millón de pesetas en los beneficios, se corresponde con un aumento de 3.0909 años en la experiencia de la empresa. (Parece que claro, que para este par de variables, la variable independiente debería ser los años de experiencia y la independiente los beneficios y no al revés).
La ordenada en el origen se interpretaría como los años de experiencia para una empresa que no obtuviese beneficios (0 millones). Evidentemente, el valor obtenido (-1.4545 años) no tiene sentido en este caso ya que 0 no está dentro del rango de valores de la variable beneficios utilizados para predecir la recta de regresión.

6. a) Diagrama de dispersión:



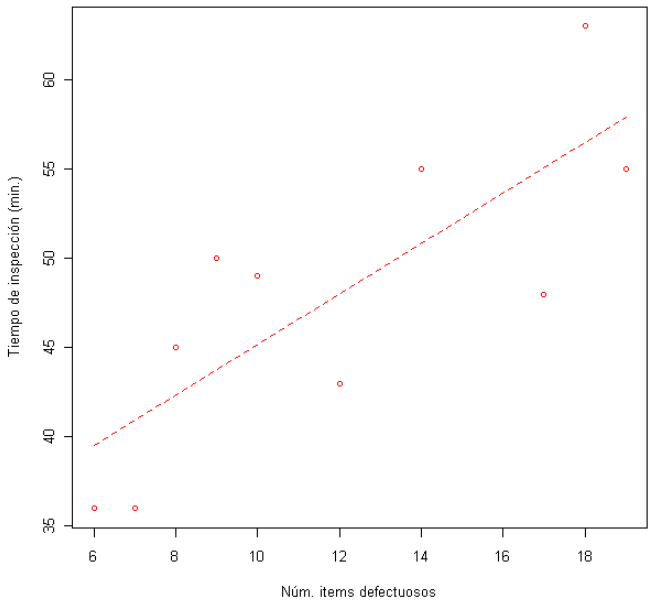

b) Recta de regresión por mínimos cuadrados: $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$.
La ecuación de la recta es: $y = 2.2405 + 0.2261x$.

7. a) Recta de regresión por mínimos cuadrados: $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$.
La ecuación de la recta es: $y = 7.9531 + 0.4408x$.

b) Coeficiente de correlación (se calcula como en el ejercicio 4): $r_{(x,y)} = 0.9848$.

c) El coeficiente de determinación es $r^2_{(x,y)} = 0.9699$, es decir, casi el 97 % de la variabilidad del tiempo de espera queda **explicada** por su dependencia lineal del número de pasajeros que llegan. Esto es, la relación lineal entre ambas variables es muy fuerte.

8. a) Diagrama de dispersión:



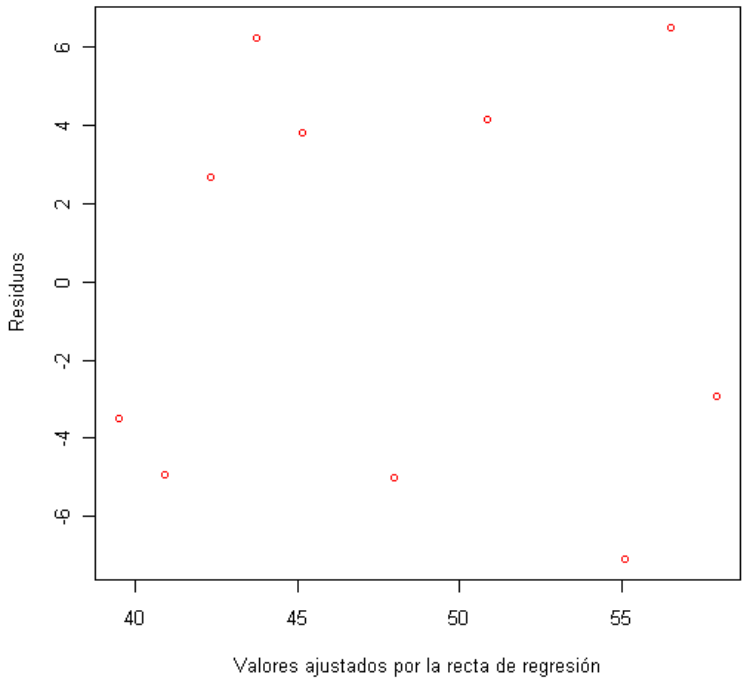
b) Recta de regresión por mínimos cuadrados: $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$.

La ecuación de la recta es: $y = 31 + 1.4167x$.

c) Residuos:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # items def. ($x_i$) | 17 | 9 | 12 | 7 | 8 | 10 | 14 | 18 | 19 | 6 |
| t. inspección ($y_i$) | 48 | 50 | 43 | 36 | 45 | 49 | 55 | 63 | 55 | 36 |
| $\hat{y}_i = 31 + 1.4167x_i$ | 55.08 | 43.75 | 48.00 | 40.92 | 42.33 | 45.17 | 50.83 | 56.50 | 57.92 | 39.50 |
| Res. ($r_i = y_i - \hat{y}_i$) | 7.08 | 6.25 | -5.00 | -4.92 | 2.67 | 3.83 | 4.17 | 6.50 | -2.92 | -3.50 |

d) Gráfico de los residuos $r_i$ frente a los valores predichos $\hat{y}_i$:




Los residuos se reparten de forma aleatoria en torno a la linea horizontal $y = 0$, y por tanto podemos decir que el ajuste de la recta de regresión es bueno.

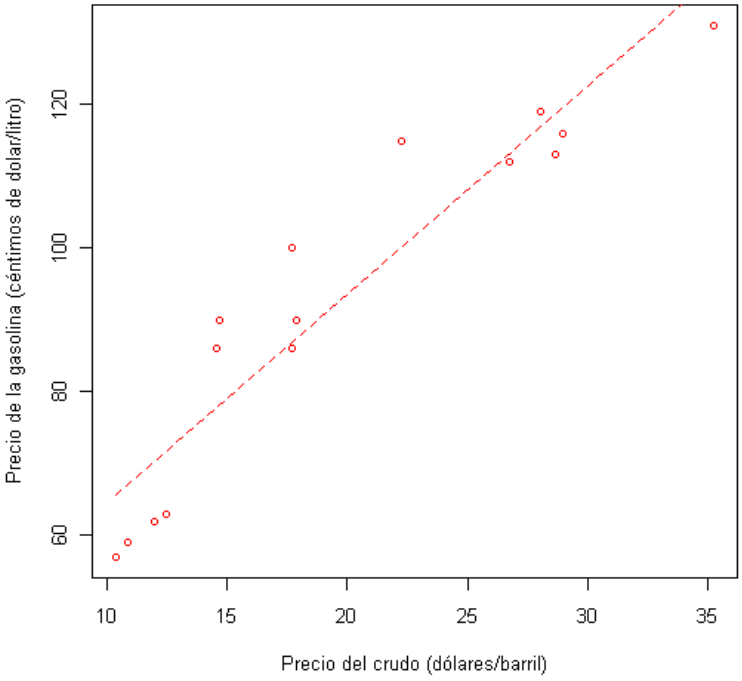
e) Coeficiente de determinación (cálculos como en el ejercicio 4): $r^2_{(x,y)} = 0.6299$. Esto quiere decir que el 70 % de la variabilidad del tiempo de inspección viene explicada por su dependencia lineal del número de items defectuosos.

9. El coeficiente de correlación es (cálculos como en el ejercicio 4): $r_{(x,y)} = 0.7911$ . La relación lineal entre estas dos variables es positiva, es decir, a mayor tamaño de la familia mayor es el consumo de detergentes, ya que $r_{(x,y)}$ es positivo. Además, podemos decir que la relación lineal es fuerte ya que el valor del coeficiente de correlación es bastante alto (próximo a 1).

10. El coeficiente de correlación es (cálculos como en el ejercicio 4): $r_{(x,y)} = 0.7607$ . La relación lineal entre estas dos variables es positiva, es decir, a mayor número de inventarios mayor es el porcentaje de ventas de estas compañías, ya que $r_{(x,y)}$ es positivo. Además, podemos decir que la relación lineal es fuerte ya que el valor del coeficiente de correlación es bastante alto (próximo a 1).

11. a) Recta de regresión para el precio de la gasolina ($y$) en función del precio del crudo ($x$): $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$. La ecuación de la recta es: $y = 35.51 + 2.91x$.

    b) Diagrama de dispersión y la recta ajustada en el apartado anterior:

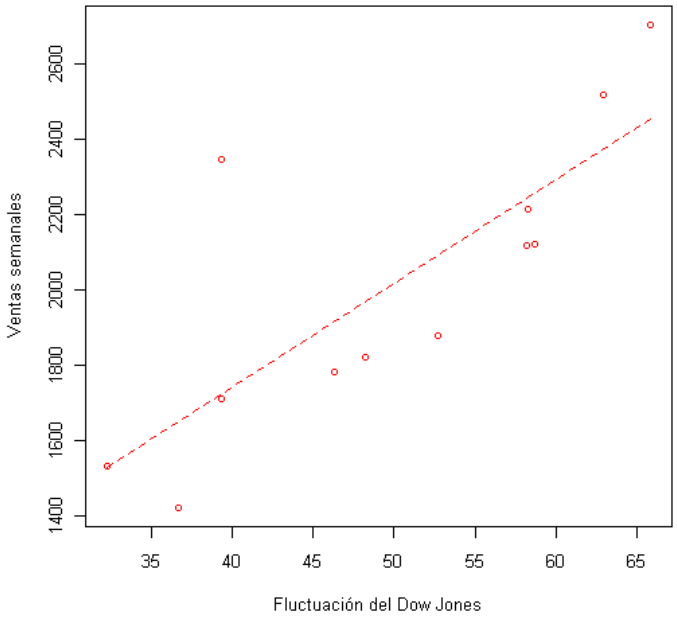



    c) Si el precio del crudo cae a los 15\$, el precio estimado del litro de gasolina será $y(15) = 35.51 + 2.91 \cdot 15 = 79.16$ céntimos de dólar.

    d) No tiene sentido hacerse la pregunta anterior para un precio del crudo de 0 dólares, ya que 0 no está dentro del rango de valores de $x$ utilizados para calcular la recta de regresión.

    e) Tampoco se puede emplear la recta de regresión obtenida en el apartado a) para predecir a futuro el precio del crudo a partir del precio de la gasolina, porque la relación a futuro entre los dos precios puede cambiar y dejar de tener el comportamiento descrito por la recta de regresión.

12. a) Recta de regresión para las ventas semanales ($y$) en función de la fluctuación del *Dow Jones* ($x$): $y = a + bx$ donde $b = \dfrac{Cov(x,y)}{s_x^2}$ y $a = \overline{y} - b\overline{x}$. La ecuación de la recta es: $y = 640.98 + 27.53x$.
    Diagrama de dispersión y recta de regresión:




b) Parece haber cierta relación entre las dos variables, es decir, a mayores fluctuaciones en el *Dow Jones* se observan mayores ventas. En ese sentido se corroboraría la sospecha del dueño de la tienda. Sin embargo, podemos observar que el ajuste de la recta de regresión no es muy bueno. Se aprecia un dato atípico que "desplaza"la recta del centro de la nube. Y aún eliminando ese dato atípico, el resto de puntos tampoco parece seguir una tendencia lineal.

c) No necesariamente, ya que correlación no implica causalidad. En este caso, no parece razonable pensar que mayores fluctuaciones en el *Dow Jones* "provoquenün aumento en las ventas. Lo que puede ocurrir es que haya variables subyacentes que tengan a la vez relación con las fluctuaciones del Dow Jones y las ventas de la tienda, y que hagan que cuando las primeras suban, las segundas suban también.