ESTADISTICA Y ANALISIS DE DATOS

# Regresión estadística

Autor: Josemari Sarasola

sigmalitika.hirusta.io

## 0.4. Ejercicios sobre regresión

1. Se han recogido datos sobre horas de estudio semanales y calificaciones de matemáticas en un grupo de alumnos:

| Alumno | A | B | C | D | E |
|---|---|---|---|---|---|
| Horas de estudio | 2 | 3 | 4 | 6 | 8 |
| Calificación | 3.2 | 4.6 | 6 | 5.8 | 8.4 |

   *a*) Estime e interprete los parámetros de la recta de regresión que relaciona la dos variables.

   *b*) Calcule las predicciones y los errores.

   *c*) Dibuje la nube de puntos y la recta de regresión, así como un término de error.

2. A lo largo de varios dias de agosto se han recogido datos sobre temperatura máxima y ventas de helados (en euros) en un establecimiento. He aquí los datos:

| Dia | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Ventas | 124 | 145 | 202 | 196 | 154 | 176 | 188 | 167 |
| Temperatura | 25 | 29 | 42 | 39 | 28 | 37 | 38 | 33 |

   *a*) Especifique la recta de regresión minimocuadrática.

   *b*) Para mañana se ha pronosticado una temperatura máxima de 35 grados. En base a ello, haga una predicción para las ventas de helados. Sería adecuado hacer una predicción del mismo modo para una temperatura de 10 grados? Analice la fiabilidad de las predicciones.

   *c*) Dibuje e interprete el diagrama de errores.

3. En una factoría se ha realizado un experimento controlando las horas de mantenimiento y observando su efecto en la producción diaria. Estos son los datos compilados:

| Dia | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Horas de mantenimiento | 2 | 3 | 4.5 | 6 | 7 | 8 |
| Producción | 4 | 6 | 9 | 11 | 11.5 | 11.8 |

   *a*) Especifique la recta de regresión minimocuadrática y su coeficiente de determinación. Es la recta ajustada a los datos?

   *b*) Dibuje la nube de puntos y establezca la lńea que pueda ajustarse mejor a dicha nube.

   *c*) Ajuste la curva $\hat{w} = k - \frac{m}{v}$ a los datos y establezca la producción máxima para ella. Calcule el coeficiente de determinación a partir de la varianza de las predicciones.

   *d*) Ajuste la curva $\hat{w} = kv^m$ a los datos y analice si es más ajustada que la curva anterior. Nota: calcule $R^2$ a través de la varianza residual.

4. A lo largo de varios experimentos se han recogido datos sobre la probabilidad de parada de un dispositivo a diferentes temperaturas:

| Temperatura | 38 | 46 | 54 | 68 | 80 |
|---|---|---|---|---|---|
| Probabilidad de parada | 0.12 | 0.24 | 0.38 | 0.58 | 0.94 |

   *a*) Ajuste la curva $\hat{w} = ke^{mv}$ a los datos.

   *b*) Para que la probabilidad de parada sea inferior a 0.5, cuál debe ser la temperatura?

   *c*) Para qué temperatura se considera segura la parada?

   *d*) Analice la bondad de ajuste a través de la varianza residual.

5. Se han inyectado diferentes dosis de un fármaco a ratones y para cada dosis se ha calculado el porcentaje de ratones que han experimentado una mejoría::

| Dosis (mg) | 4 | 6 | 8 | 12 |
|---|---|---|---|---|
| Porcentaje de mejora | 22 % | 43 % | 69 % | 82 % |

   *a*) Haga una predicción del porcentaje de mejora para una dosis de 10 mg, a través del modelo logit.

   *b*) Calcule la dosis necesaria para un porcentaje de mejora del 99 %.

   *c*) Analice la bondad de ajuste a traves de la varianza de las predicciones.

6. En un examen de conducir colectivo, se compiló el número de horas de entrenamiento y si se aprobó el examen. He aquí los datos:

| Horas | Ha aprobado? |
|---|---|
| 4 | no-si-no-no-no-no-no-no-no-no |
| 5 | no-si-no-no-no-si-si-no-no-no |
| 8 | no-si-si-no-si-no-si-si-no-no |
| 10 | si-si-si-si-si-si-no-no-no-no |
| 14 | no-si-si-si-si-si-si-si-si-si |

   *a*) A través del modelo logit, haga una predicción sobre le número de horas necesario para aprobar con el 99 % de probabilidad.

   *b*) Idem, con una probabilidad del 99.9 % . Interprete ambos resultados en relación a las propiedades de la curva logística.

7. Se han recogido datos sobre el porcentaje de plantas que pueden considerarse comestibles con diferentes dosis de un producto fitosanitario:

| Dosis (mg) | 4 | 6 | 8 | 12 |
|---|---|---|---|---|
| Porcentaje de comestibles | 90 % | 67 % | 54 % | 12 % |

   *a*) Si se persigue un porcentaje de comestibles del 80 %, calcule para ello la dosis máxima del producto, con el modelo logit.

   *b*) Calcule el coeficiente de determinación, a través de la varianza residual.

# Plantillas de resolución

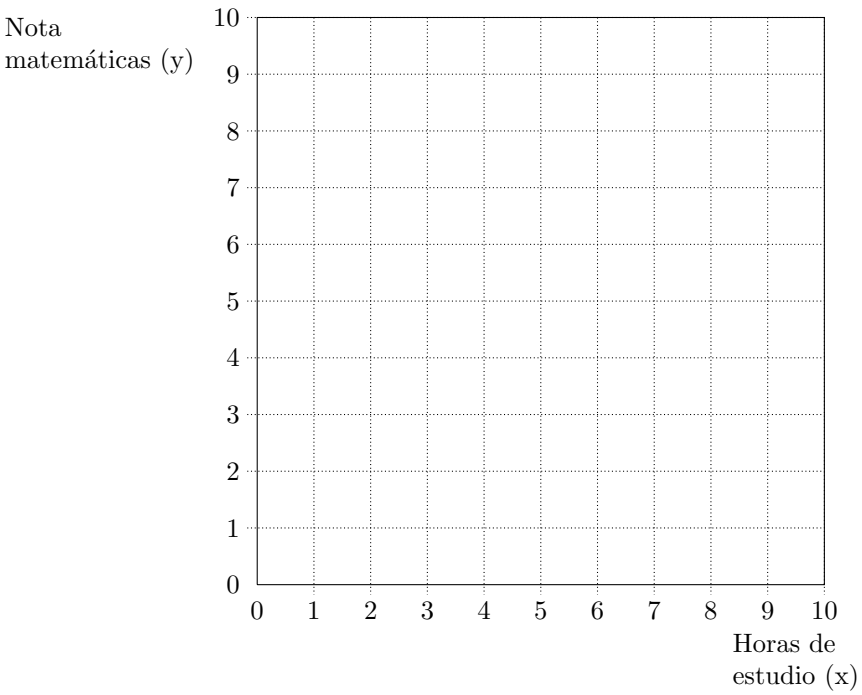
**Ejercicio 1**

| $x_i$ | $y_i$ | $x_i y_i$ | $x_i^2$ | $\hat{y} =$ + $x_i$ | $e_i = y_i - \hat{y}_i$ |
|---|---|---|---|---|---|
| 2 | 3.2 | | | | |
| 3 | 4.6 | | | | |
| 4 | 6 | | | | |
| 6 | 5.8 | | | | |
| 8 | 8.4 | | | | |
| | | | | | |

$$\overline{x} = \quad ; \quad \overline{y} =$$

$$\left.\begin{array}{l} s_{xy} = \text{———————} = \\ s_x^2 = \sqrt{\text{———————}} = \end{array}\right\} b = \frac{s_{xy}}{s_x^2} =$$

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



**Ejercicio 2**

(a)

| $x_i$ | $y_i$ | $x_i y_i$ | $x_i^2$ |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

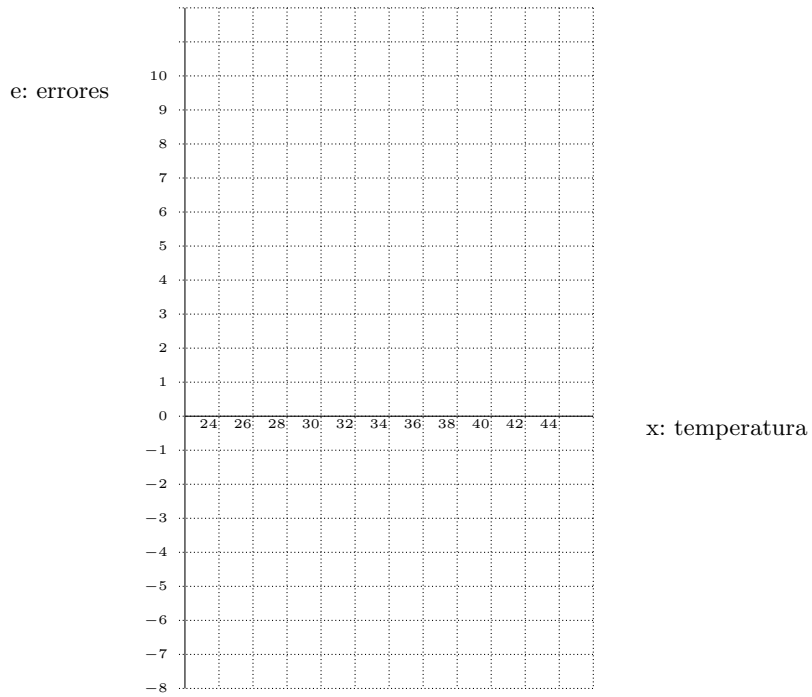
- $\overline{x} =$
- $\overline{y} =$
- $s_{xy} =$
- $s_x^2 =$
- $b = \dfrac{s_{xy}}{s_x^2} =$
- $a = \overline{y} - b\overline{x} =$

(b)

| $\hat{y} = \quad + \quad x_i$ | $e_i = y_i - \hat{y}_i$ | $y^2$ | $\hat{y}^2$ | $e^2$ |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

- $s_y^2 = \dfrac{\sum y^2}{n} - \overline{y}^2 =$
- $s_{\hat{y}}^2 = \dfrac{\sum \hat{y}^2}{n} - \overline{\hat{y}}^2 =$
- $s_e^2 = \dfrac{\sum e^2}{n} - \overline{e}^2 = \dfrac{\sum e^2}{n} =$
- $R^2 = \dfrac{s_{\hat{y}}^2}{s_y^2} =$
- $R^2 = 1 - \dfrac{s_e^2}{s_y^2} =$

(c)




**Ejercicio (3)**

(a)

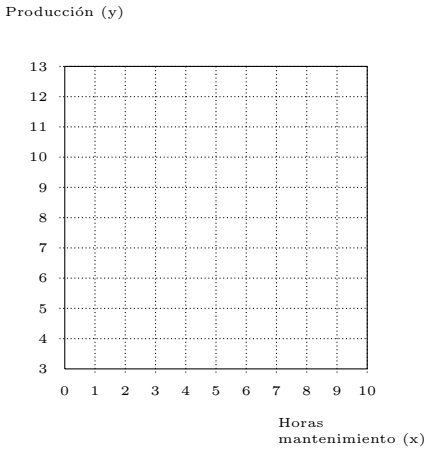
| $x_i$ | $y_i$ | $x_i y_i$ | $x_i^2$ |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

- $\overline{x} =$
- $\overline{y} =$
- $s_{xy} =$
- $s_x^2 =$
- $b = \dfrac{s_{xy}}{s_x^2} =$
- $a = \overline{y} - b\overline{x} =$

| $\hat{y} = \quad + \quad x_i$ | $y^2$ | $\hat{y}^2$ |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

- $s_y^2 = \dfrac{\sum y^2}{n} - \overline{y}^2 =$
- $s_{\hat{y}}^2 = \dfrac{\sum \hat{y}^2}{n} - \overline{\hat{y}}^2 =$
- $R^2 = \dfrac{s_{\hat{y}}^2}{s_y^2} =$

(b)



**3 (c)**

| $v_i$ | $w_i$ | $x_i =$ | $y_i =$ | $x_i y_i$ | $x_i^2$ | $\hat{y}_i =$ | $y_i^2$ | $\hat{y_i}^2$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |

- $\overline{x} =$
- $\overline{y} =$
- $s_{xy} =$
- $s_x^2 =$
- $b = \dfrac{s_{xy}}{s_x^2} =$
- $a = \overline{y} - b\overline{x} =$
- $s_y^2 = \dfrac{\sum y^2}{n} - \overline{y}^2 =$
- $s_{\hat{y}}^2 = \dfrac{\sum \hat{y}^2}{n} - \overline{\hat{y}}^2 =$
- $R^2 = \dfrac{s_{\hat{y}}^2}{s_y^2} =$

Sigmalitika - Josemari Sarasola

**3 (d)**

| $v_i$ | $w_i$ | $x_i =$ | $y_i =$ | $x_i y_i$ | $x_i^2$ | $\hat{y}_i =$ | $e_i = y_i - \hat{y}$ | $y_i^2$ | $e_i^2$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |

- $\overline{x} =$
- $\overline{y} =$
- $s_{xy} =$
- $s_x^2 =$
- $b = \dfrac{s_{xy}}{s_x^2} =$
- $a = \overline{y} - b\overline{x} =$
- $s_y^2 = \dfrac{\sum y^2}{n} - \overline{y}^2 =$
- $s_e^2 = \dfrac{\sum e^2}{n} =$
- $R^2 = 1 - \dfrac{s_e^2}{s_y^2} =$

**Ejercicio (4)**

| $v_i$ | $w_i$ | $x_i =$ | $y_i =$ | $x_i y_i$ | $x_i^2$ | $\hat{y}_i =$ | $e_i = y_i - \hat{y}$ | $y_i^2$ | $e_i^2$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |

- $\overline{x} =$
- $\overline{y} =$
- $s_{xy} =$
- $s_x^2 =$
- $b = \dfrac{s_{xy}}{s_x^2} =$
- $a = \overline{y} - b\overline{x} =$
- $s_y^2 = \dfrac{\sum y^2}{n} - \overline{y}^2 =$
- $s_e^2 = \dfrac{\sum e^2}{n} =$
- $R^2 = 1 - \dfrac{s_e^2}{s_y^2} =$

**Ejercicio (5)**

| $x_i$ | $p_i$ | $y_i = ln\frac{p_i}{1-p_i}$ | $x_i y_i$ | $x_i^2$ | $\hat{y}_i =$ | $\hat{y}_i^2$ | $y_i^2$ |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

- $\overline{x} =$
- $\overline{y} =$
- $s_{xy} =$
- $s_x^2 =$
- $b = \frac{s_{xy}}{s_x^2} =$
- $a = \overline{y} - b\overline{x} =$
- $s_y^2 = \frac{\sum y^2}{n} - \overline{y}^2 =$
- $s_{\hat{y}}^2 = \frac{\sum \hat{y}^2}{n} - \overline{\hat{y}}^2 =$
- $R^2 = \frac{s_{\hat{y}}^2}{s_y^2} =$

**Ejercicio (6)**

| $x_i$ | $n_i$ zenbatetik | $p_i$ | $y_i = ln\frac{p_i}{1-p_i}$ | $x_i y_i$ | $x_i^2$ |
|---|---|---|---|---|---|
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |

- $\overline{x} =$
- $\overline{y} =$
- $s_{xy} =$
- $s_x^2 =$
- $b = \frac{s_{xy}}{s_x^2} =$
- $a = \overline{y} - b\overline{x} =$

**Ejercicio (7)**

Como puede observarse al examinar los datos, en este caso la probabilidad disminuye a medida que se aumenta la dosis, de modo que no podemos aplicar el modelo logit tal como lo hemos planteado en un principio. Sin embargo, la solución es inmediata: no tenemos más que sustituir a la probabilidad de ser comestible por la probabilidad complementaria de ser incomestible como variable dependiente.

| $x_i$ | $p_{1i}$ | $p_{2i}$ | $y_i = ln\frac{p_{2i}}{1-p_{2i}}$ | $x_i y_i$ | $x_i^2$ | $\hat{y}_i = -4{,}02 + 0{,}50x$ | $e_i = y_i - \hat{y}_i$ | $y_i^2$ | $e_i^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 0.90 | 0.10 | -2.20 | -8.79 | 16 | -2.02 | -0.18 | 4.83 | 0.031 |
| 6 | 0.67 | 0.33 | -0.71 | -4.25 | 36 | -1.02 | 0.31 | 0.50 | 0.097 |
| 8 | 0.54 | 0.46 | -0.16 | -1.28 | 64 | -0.02 | -0,14 | 0.02 | 0.019 |
| 12 | 0.12 | 0.88 | 1.99 | 23.91 | 144 | 1.98 | 0.01 | 3.97 | 0.000 |
| 30 | | | -1.07 | 9.59 | 260 | -1.07 | 0 | 9.32 | 0.1484 |

- $\overline{x} = \dfrac{30}{4} = 7{,}5;\ \overline{y} = \dfrac{-1{,}07}{4} = -0{,}27$
- $s_{xy} = \dfrac{9{,}59}{4} - 7{,}5 \times -0{,}27 = 4{,}42$
- $s_x^2 = \dfrac{260}{4} - 7{,}5^2 = 8{,}75$
- $b = \dfrac{s_{xy}}{s_x^2} = \dfrac{4{,}42}{8{,}75} = 0{,}50;\quad a = \overline{y} - b\overline{x} = -0{,}27 - 0{,}50 \times 7{,}5 = -4{,}02 \rightarrow \ln\left(\widehat{\dfrac{p}{1-p}}\right) = -4{,}02 + 0{,}50x$
- $p_1 = 0{,}8 \rightarrow p_2 = 0{,}2 \rightarrow \ln(\dfrac{0{,}2}{1-0{,}2}) = -4{,}02 + 0{,}50 * x \rightarrow -1{,}38 = -4{,}02 + 0{,}50 * x \rightarrow x = 5{,}28$

---

- $s_y^2 = \dfrac{\sum y^2}{n} - \overline{y}^2 = \dfrac{9{,}32}{4} - (-0{,}27^2) = 2{,}25$
- $s_e^2 = \dfrac{\sum e^2}{n} = \dfrac{0{,}1484}{4} = 0{,}0371$
- $R^2 = 1 - \dfrac{s_e^2}{s_y^2} = 1 - \dfrac{0{,}371}{2{,}25} = 0{,}986 \rightarrow$ muy buen ajuste