

Aprendizaje Automático - Trabajo Práctico 3

Gonzalo Castiglione - 49138

May 16, 2012

Objetivo: Aplicar diversos métodos estadísticos para aprender a hacer inferencia a partir de datos experimentales.

1 Métodos de estadística paramétrica

1. Resultados

(a) Se calculó para cada especie, su medida y su desvío estandar:

Especie	Estimador	Largo Sépalo	Ancho Sépalo	Largo Pétalo	Ancho Pétalo
Virginica	$\hat{\mu}$	6.5880	2.9740	5.5520	2.0260
	$\hat{\sigma}$	0.6232	0.3160	0.5409	0.2692
Versicolor	$\hat{\mu}$	5.9360	2.7700	4.2600	1.3260
	$\hat{\sigma}$	0.5058	0.3075	0.4605	0.1938
Setosa	$\hat{\mu}$	5.0060	3.4280	1.4620	0.2460
	$\hat{\sigma}$	0.3454	0.3715	0.1702	0.1033

(b) Cálculo de los errores cuadráticos medios

Especie	Largo Sépalo	Ancho Sépalo	Largo Pétalo	Ancho Pétalo
Virginica	0.0081	0.0021	0.0061	0.0015
Versicolor	0.0053	0.0020	0.0044	0.0008
Setosa	0.0025	0.0029	0.0006	0.0002

(c) Intervalos de confianza para con un nivel de confianza de 0.95.

	Intervalo			
Especie	Largo Sépalo	Ancho Sépalo	Largo Pétalo	Ancho Pétalo
Virginica	6.7687 3.0657	5.7088 2.1041	6.4073 2.8823	5.3952 1.9479
Versicolor	6.0827 2.8592	4.3935 1.3822	5.7893 2.6808	4.1265 1.2698
Setosa	5.1062 3.5357	1.5114 0.2760	4.9058 3.3203	1.4126 0.2160

2. Se tienen 80 componentes, de las cuales 12 son defectuosas. Por ser este experimento una secuencia de ensayos Bernoulli, repetidos n veces a probabilidad constante, se lo puede considerar como una distribución binomial.

(a) La proporción de componentes no defectuosos de la muestra = $\bar{x}_{nd} \frac{80-12}{80} = 0.85$

i. Un estimador \hat{x} es un estimador insesgado para estimar a p si $E[\hat{x}] = p$.

$$E[\bar{x}] = E\left(\frac{\sum x_i}{n}\right) = \frac{1}{n} \sum E(x_i) = \frac{1}{n} np = p.$$

Por lo tanto este es un estimador *insesgado*.

ii. Muestra: 68 mediciones con $\{x_i, y_i\} = 1$ y 12 mediciones con $\{x_i, y_i\} = 0$.

$e_0 = (0 - 0.85)$ (para las 12 muestras defectuosas)

$e_1 = (1 - 0.85)$ (para las 68 muestras no defectuosas)

Por lo que el error cuadrático medio estaría dado por la fórmula:

$$E_{CM} = \sqrt{\frac{(1-0.85)^2 * 68 + (0-0.85)^2 * 12}{80}} = \sqrt{\frac{1.53 + 8.67}{80}} \simeq 0.35$$

(b) Sean:

x_a = total de formas de tomar los componentes que andan de a pares (sin importar orden).

x_t = total de formas de tomar el total de componentes de a pares (sin importar orden).

Proporción de sistemas que funcionan correctamente $\frac{x_a}{x_t} = \frac{\binom{80-12}{2}}{\binom{80}{2}} = \frac{2278}{3160} = 0.72$

3. Resultados obtenidos:

Variables	Valor numérico
alpha	0.05
h	1
p	0.0118
ci	[0.0314,0.2040]
df	12
sd	0.1428

Para el alpha tomado, la hipótesis nula se rechaza y los científicos están en lo cierto, hay variación entre el color de las plumas.

4. Solución

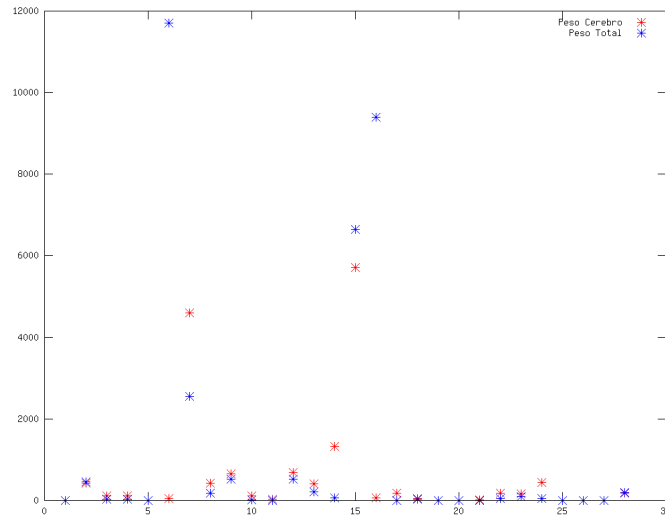


Figure 1: Peso del cerebro y peso total para cada medición en brains.txt*

(a) ¹

¹*El valor del peso del cerebro de la medición 25 no se ve en la figura ya que se aleja demasiado del resto de los valores y el ajustar los ejes solo para mostrar ese valor produce que todas las demás mediciones no puedan apreciarse correctamente.

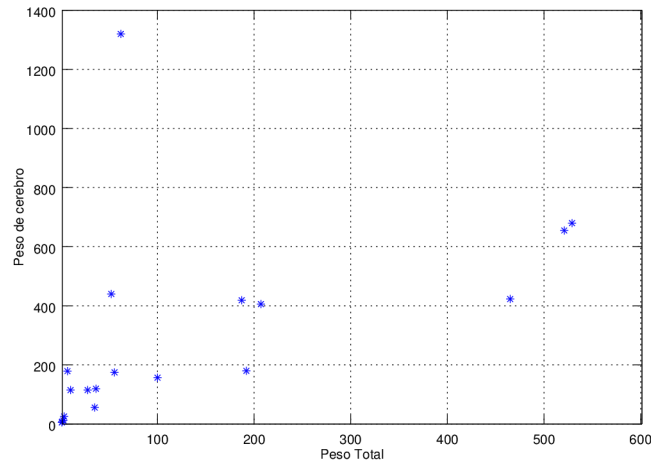


Figure 2: Peso total en Kg Vs peso del cerebro en G**

2

- i. En una observación a simple vista, se puede ver que las mediciones que se diferencian notablemente del resto son: 6, 7, 14, 15, 16 y la 25.
 - ii. En el segundo gráfico, puede verse que debido a la dispersión de los datos, si aproximamos los valores por una recta, a simple vista se ve que se va a estar cometiendo un error muy grande para la mayoría de los datos. Por lo que no existe una relación lineal.
- (b) En el gráfico a continuación puede verse que graficando el \log de ambas variables, los datos tienden a formar una línea mas definida.

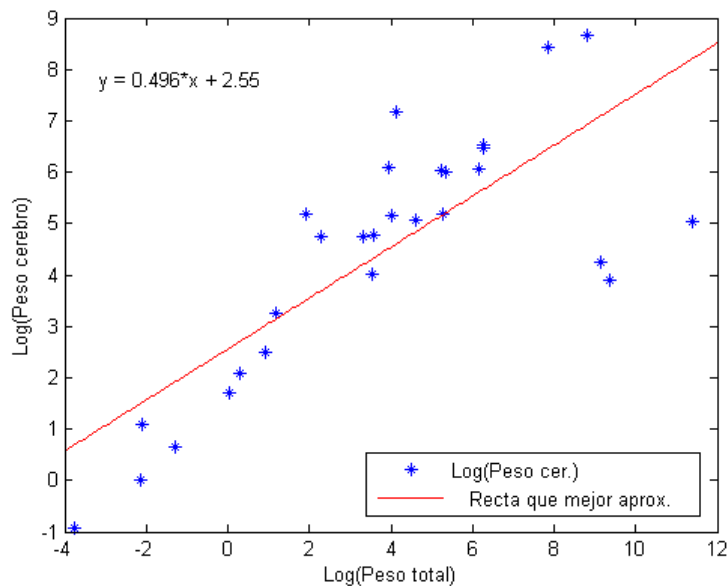


Figure 3: Logaritmo de ambas mediciones y la línea que mejor los aproxima

- (c) La recta que mejor ajusta con los ejes x e y con la función \log aplicados a ambos es: $y = a * x + b$. Con $a = 0.496$ y $b = 2.55$.
 Para obtener la recta que ajusta a los puntos x e y sin aplicar las transformación, se debe aplicar la tranformación inversa al \log , es decir $\text{pow}(10, n)$. Quedando así la curva que aproxima a los puntos como: $10^y = a * 10^x + b$. Despejando por y , se obtiene $y = \log(a * 10^x + b)$.

**Se removieron los valores para los 4 valores de x mayores a 2000 ya que ocultaban la visualización de todos los demás valores

Coeficientes de Regresion	$\sum res^2$	E_{cm}
0.496 - 2.55	60.99	2.34

- (d) En los gráficos del punto *b*, se puede observar que las mediciones que estan muy fuera del común son no solo las 14, 15 y 25 sino que también la 6 y la 16. Por lo que fueron removidos de la tabla de valores. El gráfico obtenido luego de ajustados los valores se muestra en la figura 5 del Anexo.

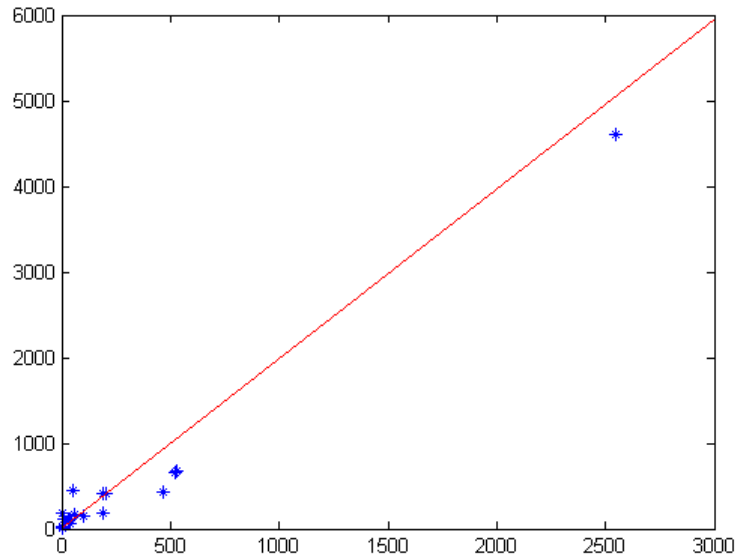


Figure 4: Logaritmo de ambas mediciones y la línea que mejor los aproxima

5. Censo poblacional de Estados Unidos 1790 - 1990

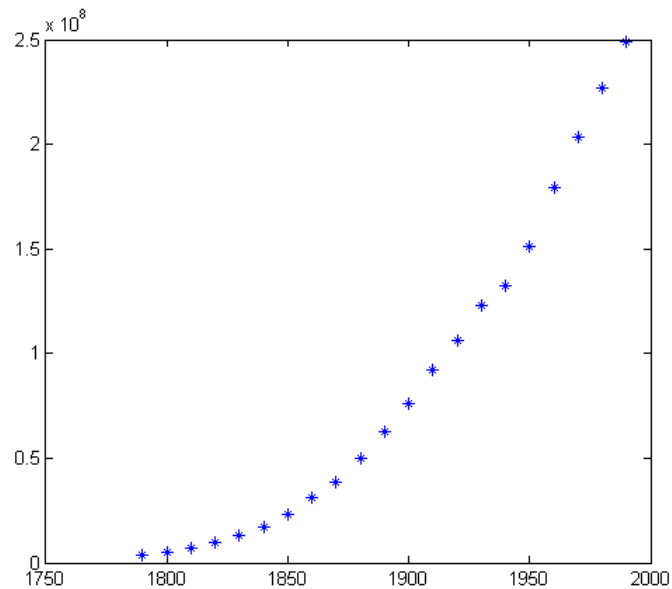


Figure 5: Cantidad de habitantes medidas por el censo por año

- (a) De la figura puede verse que existe un patrón entre las mediciones, pero no lineal.

- (b) $P_2(x) = p1 * z^2 + p2 * z + p3$
 $z = \frac{x-\mu}{\sigma} = \frac{x-1890}{62.05}$
 $p1 = 2.5049e7$
 $p2 = 7.5414e7$
 $p3 = 6.1927e7$
- (c) $P_3(x) = p1 * z^3 + p2 * z^2 + p3 * z + p4$
 $p1 = 7.7279e5$
 $p2 = 2.5049e7$
 $p3 = 7.4093e7$
 $p4 = 6.1927e7$

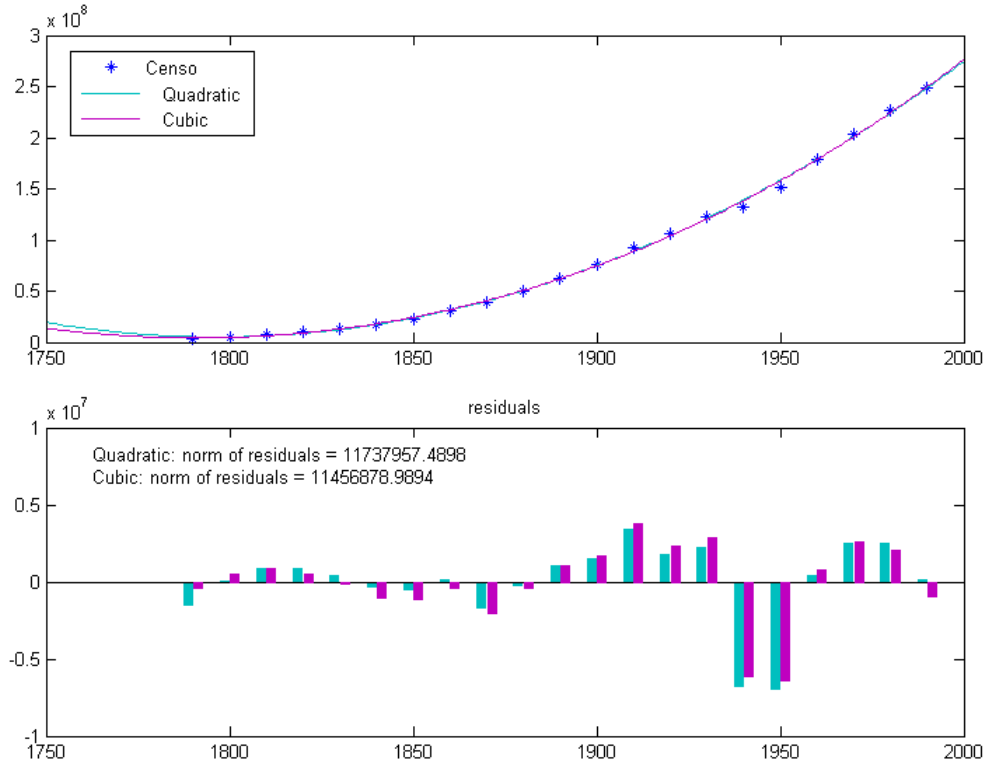


Figure 6: Cantidad de habitantes por cada año

- (d) A partir de los residuos generados por cada polinomio, puede verse que el polinomio de grado 3 es ligeramente menor, por lo que es el que mejor representaría a esta muestra.
- (e) De acuerdo a cada polinomio,
 $P_2(2000) = 2.74e8$
 $P_3(2000) = 2.76e8$
 Sabiendo que el valor real para ese año fue 281421906 ($2.8e8$). Por haberle acertado con menor grado de error, el polinomio de grado 3 es el que mejor aproxima.

6. Estimadores de maxima verosimilitud

Especie	Vector de medias [largo S., ancho S., largo P., ancho P.]	μ	σ
Virginica	[6.5880, 2.9740, 5.5520, 2.0260]	4.2850	3.4327
Versicolor	[5.9360, 2.7700, 4.2600, 1.3260]	3.5730	3.5730
Setosa	[5.0060, 3.4280, 1.4620, 0.2460]	2.5355	3.3235

Especie	Matriz de covarianza				σ
Virgínica	0.4043	0.0938	0.3033	0.0491	0.8695
	0.0938	0.1040	0.0714	0.0476	
	0.3033	0.0714	0.3046	0.0488	
	0.0491	0.0476	0.0488	0.0754	
Versicolor	0.2664	0.0852	0.1829	0.0558	0.6150
	0.0852	0.0985	0.0827	0.0412	
	0.1829	0.0827	0.2208	0.0731	
	0.0558	0.0412	0.0731	0.0391	
Setosa	0.1242	0.0992	0.0164	0.0103	0.3064
	0.0992	0.1437	0.0117	0.0093	
	0.0164	0.0117	0.0302	0.0061	
	0.0103	0.0093	0.0061	0.0111	

7. Matriz obtenida:

(a)

1	0.2286	-0.8241	-0.2454
0.2286	1	-0.1392	-0.9730
-0.8241	-0.1392	1	0.0295
-0.2454	-0.9730	0.0295	1

Cada x_{ij} representa la correlación de la cantidad de calor por ingrediente i por gramo de cemento con relación al elemento j . Cuanto mas cercano en módulo a 1 es, mas se correlacionan.

(b) selección hacia adelante

	R^2	F					R^2	F		
x_1	0.5339	12.6025	\Rightarrow	x_4x_1	0.9725	176.6270	\Rightarrow	$x_1x_4x_2$	0.9823	166.8317
x_2	0.6663	21.9606		x_4x_2	0.6801	10.6280		$x_1x_4x_3$	0.9813	157.2658
x_3	0.2859	4.4034		x_4x_3	0.9353	72.2674				
x_4	0.6745	22.7985								

8. Proporción de la varianza = $\frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_p}$ para cada una de las variables:

1 variable: 0.8660

2 variables: 0.9789

3 variables: 0.9996

4 variables: 1

Del análisis de los autovalores de la matriz de covarianza se puede ver que solamente con 2 componentes, es posible representar mas del 90% de la varianza.

2 Anexo

1. Lirios Fisher

(a) Código

```
%range toma los valores 0:50; 51:100 y 101:150
data=meas(range,:);
n=size(data,1);
uHat=mean(data);
sHat=(n-1)*std(range)/n;
```

(b) Código

```
ecm=std(data).^2;
ecm=ecm/n;
```

(c) Código

```
mu=mean(data);
sigma=std(data);
p=0.05/2;
zAlpha=tinv(p,n-1) % T student
aux=zAlpha.*sigma/sqrt(n);
interval=[mu-aux,mu+aux];
```

2. -

3. Carpinteros escapularios

```
birds=[
    ...meditions here...
];
alpha=0.05};
[h,p,ci,stats]=ttest(birds(:,2),birds(:,3),alpha);
```

4. Masa croporal vs Masa cerebral

(a) Código

```
load brains.txt;
x = 1:28;
y = brains(:,1);
z = brains(:,2);
clf;
hold on;
plot(x, y, '*b;Peso Promedio en Kg;')
plot(x, z, '*r;Peso Cerebro Promedio en G;')
print('-dpng', './TotalWeightVsBrainWeight.png')
```

(b) Código

```
load brains.txt
regstats(brains(:,1), brains(:,2),'línear')
```

5. Censo

(a) .

```
load population.txt;
x = population(:, 1);
y = population(:, 2);
plot(x,y, '*')
```

(b) .

```

load population.txt;
c2=polyfit(cdate,pop,2)
poli2=polyval(c2,cdate);

```

6. Lirios Fisher

```

meanVector=mean(data)
n=size(meanVector,2);
mu=mean(meanVector)
sigma= sum((meanVector-ones(1,n)*mu).^2)/n
% segunda parte
cm=cov(data)
sigma=cm(1,1)+cm(2,2)+cm(3,3)+cm(4,4)*(n-1)/n

```

7. Evolución de las calorías por gramo de cemento

```

[r] = corrcoef(ingredients)

```