Regresión Lineal: Test de Hipótesis.

José I. González Etchemaite
Aprendizaje Estadístico

7 de mayo de 2020

1. Desarrollo

1.1. Ejercicio 1

Mediante la función *ggpairs()* que calcula la correlación entre variables de a pares, se obtiene el siguiente gráfico.

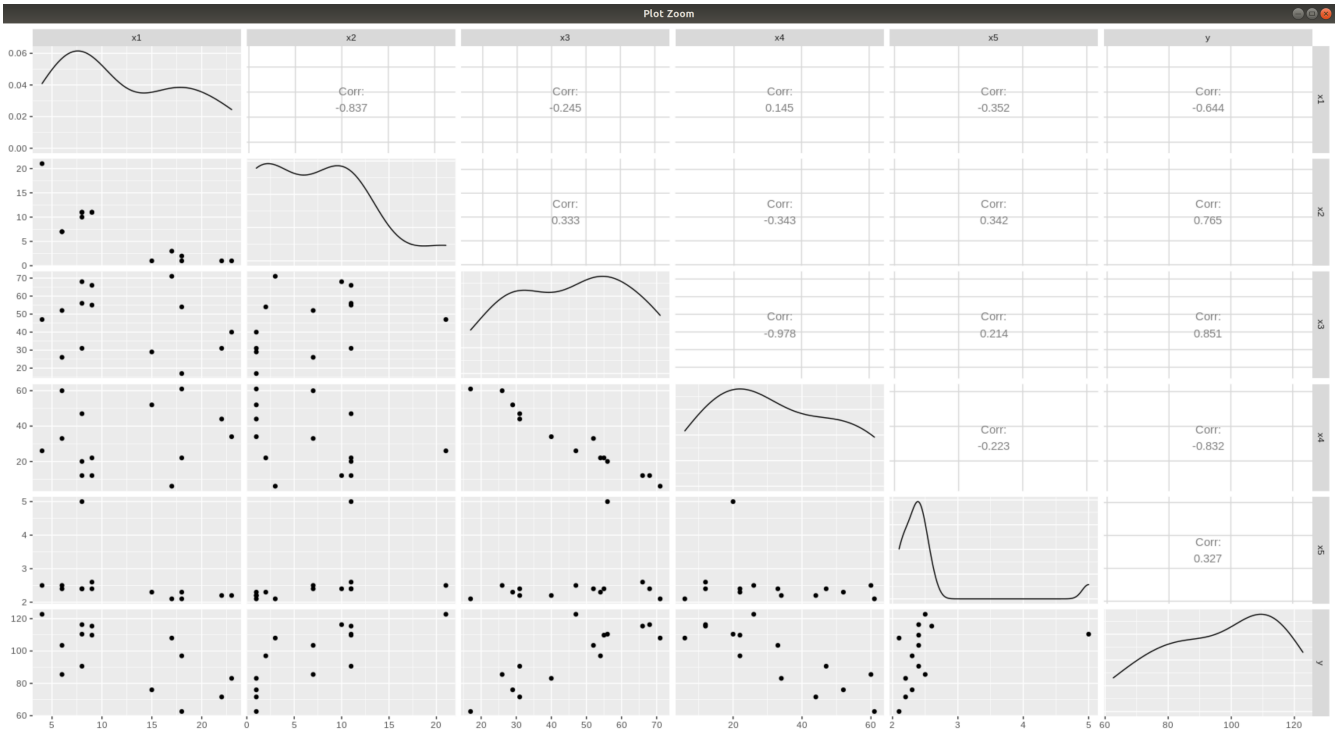

Figura 1: Correlación entre variables.

Vemos que las variables mas correlacionadas con la salida Y son, en orden descendente, x_3, x_4, x_2, x_1, x_5.

Si bien la cantidad de significación en cada variable en este análisis es subjetiva, las dos variables candidatas a ser elegidas como las mas significativas son x_3 y x_5. Como se verá luego en los resultados de las regresiones, la correlación entre variables no implica significación, ya que la disminución de una variable no significativa podría estar dando aumento a una que si lo es, y se tendría un efecto aparente de que la primera tiene impacto directo sobre el resultado (como el ejemplo en ISLR de que la venta de helados es significativa en una regresión de los ataque de tiburones en una playa).

1.2. Ejercicio 2

En la siguiente imagen se muestran los estadísticos obtenidos para la regresión lineal de Y en las variables x_i, con $i = 1, ..., 5$.

```
> summary(reg)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = cemento)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58166 -2.17473 -0.05122  1.84522  3.11955

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.6101   105.9653   0.695    0.507
x1           -0.4497     1.1312  -0.398    0.701
x2            1.2995     1.0660   1.219    0.258
x3            0.5630     1.0587   0.532    0.609
x4           -0.1704     1.0494  -0.162    0.875
x5           -0.3859     1.5221  -0.254    0.806

Residual standard error: 2.7 on 8 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.979
F-statistic: 122.2 on 5 and 8 DF,  p-value: 2.48e-07
```

Figura 2: Resultados de la regresión lineal.

En la segunda columna de la figura 2 se ve el estimador para cada parámetro de la regresión lineal, β_i, junto con el desvío estándar aproximado en la tercera. Haciendo un cociente entre ellas, se obtiene el estadístico t(cuarta columna) que se usará para testear la hipótesis de que cada uno sea cero.

Se observa que para cada variable se hallaron valores t relativamente chicos, por lo que las probabilidades relacionadas a estas (p-*valor*) son altas(quinta columna). En principio ninguna de las variables llega a ser significativamente distinta de cero. En cuanto a la regresión vemos que es significativa, teniendo un $p - valor$ asociado muy chico(~2.5e-07). En este caso se observa que ninguna de las variables es significativa por si misma, pero si lo son en conjunto.

1.3. Ejercicio 3

En la siguiente tabla se calcula en una columna adicional la suma de las cinco variables independientes x_i. Vemos que aproximadamente entre los cinco componentes se conforma 100 % del cemento. Esto se justifica desde el punto de partida del problema, donde planteamos que queremos encontrar el calor generado por una mezcla de cemento en función de distintas proporciones de sus componentes.

obs	x1	x2	x3	x4	x5	y	Suma
1	6	7	26	60	2.5	85.5	101.5
2	15	1	29	52	2.3	76.0	99.3
3	8	11	56	20	5.0	110.4	100
4	8	11	31	47	2.4	90.6	99.4
5	6	7	52	33	2.4	103.5	100.4
6	9	11	55	22	2.4	109.8	99.4
7	17	3	71	6	2.1	108.0	99.1
8	22	1	31	44	2.2	71.6	100.2
9	18	2	54	22	2.3	97.0	98.3
10	4	21	47	26	2.5	122.7	100.5
11	23	1	40	34	2.2	83.1	100.2
12	9	11	66	12	2.6	115.4	100.6
13	8	10	68	12	2.4	116.3	100.4
14	18	1	17	61	2.1	62.6	99.1

Figura 3: Tabla de variables.

La variable que se busca estimar es el calor generado durante el fraguado de la mezcla, por lo que el intercept sería el calor que se genera en ausencia de ellos, lo que en un principio no tiene mucho sentido.

Si bien uno supone que siempre se llegara al 100 % de los materiales de una forma u otra y nunca habrá ausencia de ellos en su totalidad, no se puede suponer a priori que el calor generado sea solo debido a los elementos que componen la mezcla. Esto es, el intercept podría estar captando el fenómeno del calor que se genera durante el fraguado independiente de estas cinco variables (algo así como un piso de calor), y por ello no debiera ser eliminado sin un debido análisis.

1.4. Ejercicio 4

En la tabla de la figura 4 se resumen los estadísticos obtenidos mediante la regresión lineal habiendo quitado el intercept. En esta vemos que hubo un cambio importante en las significaciones de cada parámetro estimado, así como también subió aun mas la significancia de la regresión conjunta.

Si utilizamos un nivel de confianza del 95 %, entonces entre las tres variables elegiremos x_2, x_3 y x_4, dado que las variables x_1 y x_5 no logran juntar la evidencia suficiente para suponer que sean significativas.

```
> summary(reg2)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 - 1, data = cemento)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1390 -1.9789  0.1514  1.5559  3.9401

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1  0.32652    0.17086   1.911   0.0883 .
x2  2.02517    0.20611   9.826 4.14e-06 ***
x3  1.29718    0.05993  21.646 4.52e-09 ***
x4  0.55768    0.05039  11.067 1.53e-06 ***
x5  0.35444    1.05496   0.336   0.7446
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.621 on 9 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9993
F-statistic:  3933 on 5 and 9 DF,  p-value: 9.691e-15
```

Figura 4: Regresión sin intercept.

Procedemos entonces a realizar la regresión del calor en estos tres componentes, donde obtenemos los siguientes resultados:

```
> summary(reg3)

Call:
lm(formula = y ~ x2 + x3 + x4 - 1, data = cemento)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3859 -1.5090 -0.4010  0.8653  4.4326

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x2  1.74693    0.14064   12.42 8.15e-08 ***
x3  1.39137    0.02857   48.71 3.35e-14 ***
x4  0.63361    0.02795   22.67 1.39e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.829 on 11 degrees of freedom
Multiple R-squared:  0.9993,    Adjusted R-squared:  0.9992
F-statistic:  5627 on 3 and 11 DF,  p-value: < 2.2e-16
```

Figura 5: Regresión con 3 variables (sin intercept).

Vemos finalmente que este modelo es el que mas significación presenta, tanto en los estadísticos individuales como en su conjunto, por lo que finalmente seria el elegido como modelo óptimo.