Fundadores de los modelos lineales

Figura: Legendre (1805) y Gauss (1809) ambos aplicaron el método al problema de determinar, a partir de observaciones astronómicas, las órbitas de los cuerpos en torno al Sol (principalmente cometas, pero también, más tarde, los entonces recién descubiertos planetas menores)

Clase 25 de Bioestadística

Modelos lineales (2): R y variaciones

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Regresión lineal: mínimos cuadrados (repaso)

El comando `lm` (linear model)

Generalizaciones

Regresión lineal: mínimos cuadrados (repaso)

- Queremos modelar la dependencia entre dos o más variables.
- Por ejemplo, si X es una variable aleatoria muy relacionada a Y, podemos postular un modelo lineal de la forma

$$Y = aX + b,$$

- Aquí a y b son constantes que desconocemos y queremos estimar.

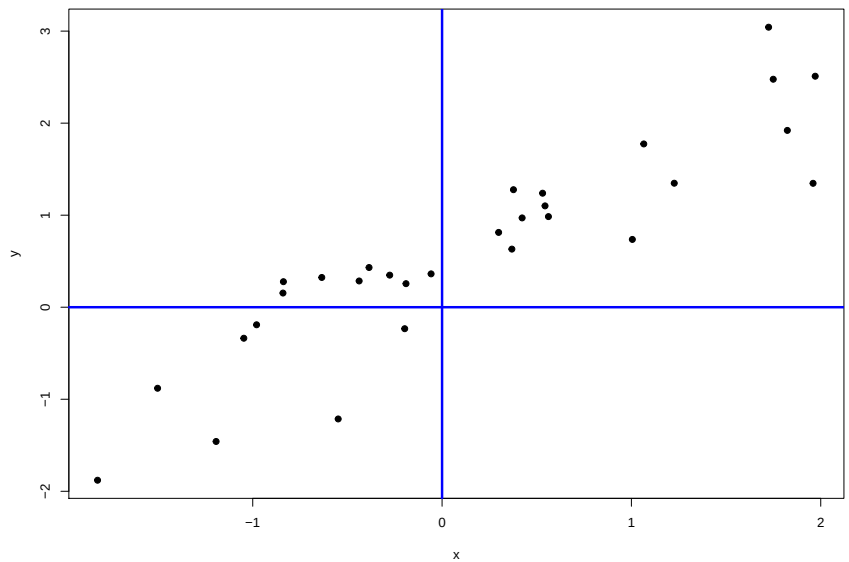
- Sin embargo, en general, el efecto de la variable X sobre la Y no sigue este modelo perfectamente lineal sino que aparece un error ε, y escribimos

$$Y = aX + b + \varepsilon$$

- El error ε es desconocido, y puede deberse
  - a errores de medición
  - Más en general, existen otras variables que no incluímos en el modelo

Aquí se muestra el modelo

$$Y = \frac{1}{2}X + 1 + \varepsilon$$



Este **es** nuestro problema

El problema estadístico

- Tenemos $(X_1, Y_1), \ldots, (X_n, Y_n)$ de pares que se asumen independientes e idénticamente distribuidos
- Suponemos que todas tienen la distribución del par (X, Y),

  $P(X_k \leq a, Y_k \leq b) = P(X \leq a, Y \leq b)$ para todo $k = 1, \ldots, n$

- Suponemos entonces que las X son determinísticas
- Suponemos que los errores ε_k son normales con media 0 y varianza σ^2.
- Las constantes a y b no se conocen.

- Nuestro modelo es

$$Y_k = aX_k + b + \varepsilon_k \quad \text{para todo } k = 1, \ldots, n.$$

- El primer problema es producir estimadores $\hat{a}_n$ de a, y $\hat{b}_n$ de b basados en dicha muestra.

- En este caso, el valor que "predice"nuestro modelo con los parámetros estimados: dado un dato X_k de la muestra estimamos la Y_k correspondiente:

$$\hat{Y}_k = \hat{a}_n X_k + \hat{b}_n.$$

- Aquí el error ε_k no aparece, porque no se conoce,

- La diferencia entre el valor observado Y_k y el valor estimado $\hat{Y}_k$ se conoce como residuo k-ésimo.

Estimación de a y b

- Buscamos la recta $y = ax + b$ que más se parezca a los datos
- Como conocemos los valores sólo en X_k, lo que queremos es que

$$Y_k \approx aX_k + b$$

- Es decir, queremos que

$$Y_k - aX_k - b \approx 0$$

- Como hicimos ya varias veces, para que los signos no compensen, pedimos

$$(Y_k - aX_k - b)^2 \approx 0$$

- En realidad, si los datos no están en una recta, no va a ser posible encontrar a y b tal que las diferencias sean cero.
- Nos resignamos con encontrar los a y b que den más cerca de cero
- Buscamos el par $(\hat{a}_n, \hat{b}_n)$ que minimice la suma de los cuadrados:

$$S(a,b) = \sum_{k=1}^{n} (Y_k - aX_k - b)^2 \tag{1}$$

- Entonces el par (a, b) va a ser un mínimo absoluto de la función

$$S(a, b)$$

- Ojo: ahora pensamos que los (X_k, Y_k) están fijos, son los datos que observamos, y a, b son las variables

- Cómo se hallan los mínimos de una función de dos variables?

- La condición necesaria de mínimo es que las derivadas parciales[1] se anulen.

- Es decir, buscamos a y b tales que

$$\frac{\partial S(a, b)}{\partial a} = \frac{\partial S(a, b)}{\partial b} = 0$$

[1] Mmm, ¿que era eso?

- Como las variables son a y b, lo que obtenemos es un sistema de dos ecuaciones lineales con dos incógnitas:

$$(\sum_{k=1}^{n} X_k Y_k) - a(\sum_{k=1}^{n} X_k^2) - b(\sum_{k=1}^{n} X_k) = 0$$

$$\sum_{k=1}^{n} Y_k - a\sum_{k=1}^{n} X_k - nb = 0$$

- Usamos las notaciones

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad \overline{Y}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$$

$$\overline{X^2}_n = \frac{1}{n}\sum_{k=1}^{n} X_k^2, \quad \overline{Y^2}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k^2 \quad \overline{XY}_n = \frac{1}{n}\sum_{k=1}^{n} X_k Y_k$$

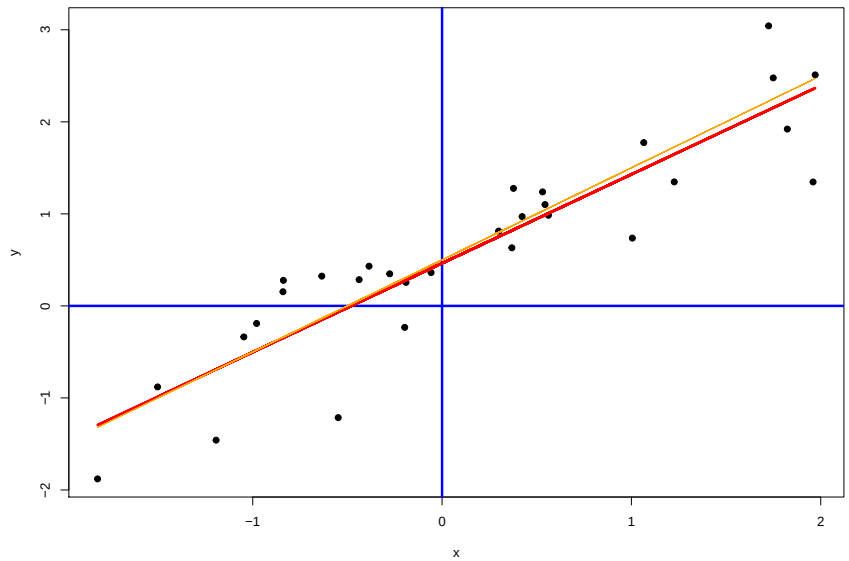
- El sistema de ecuaciones resultante de derivar queda:

$$a\overline{X^2}_n + b\overline{X}_n = \overline{XY}_n$$
$$a\overline{X}_n + b = \overline{Y}_n$$

- Despejando

$$\hat{a}_n = \frac{\overline{XY}_n - \overline{X}_n\overline{Y}_n}{\overline{X^2}_n - (\overline{X}_n)^2}$$
$$\hat{b}_n = \overline{Y}_n - \hat{a}_n\overline{X}_n$$

Veamos la solución:



- En rojo la recta estimada, con $\hat{a}_n = 0{,}97$ y $\hat{b}_n = 0{,}46$
- En naranja la recta teórica con $a = 1$ y $b = 1/2$

El comando `lm` (linear model)

- R implementa el ajuste por mínimos cuadrados con el comando `lm`
- Tenemos los vectores x e y de largo n, el modelo es

$$Y_k = aX_k + b + \varepsilon_k \qquad (k = 1, \ldots, n)$$

- El comando para calcular los coeficientes y otras características del modelo es

```
lm(formula = y ~ x)
```

- El resultado de este comando es:

```
> lm(formula = y ~ x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.5783       1.0703
```

Sumario

- El comando calcula otras cosas, que se obtienen mediante el comando `summary`:

```
> summary(lm(formula = y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.20552 -0.24126  0.08357  0.26289  1.04546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57833    0.09294   6.222 1.01e-06 ***
x            1.07031    0.08963  11.942 1.67e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5074 on 28 degrees of freedom
Multiple R-squared:  0.8359,    Adjusted R-squared:   0.83
F-statistic: 142.6 on 1 and 28 DF,  p-value: 1.67e-12
```

Veamos cada ítem

- `Residuals` son los residuos, las diferencias entre los valores observados y los calculados:

$$R_k = Y_k - (\hat{a}_n X_k + \hat{b}_k) = Y_k - \hat{Y}_k$$

- Es importante saber que los residuos no son muy grandes, y cómo se distribuyen

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.20552 -0.24126  0.08357  0.26289  1.04546
```

- `Coefficients` nos da los coeficientes:
  - `Intercept` es el intercepto, el $\hat{b}_n$
  - `x` es el coeficiente de x, es decir $\hat{a}_n$
  - `Estimate` es el valor de la estimación
  - `Std. Error` es el desvío estándar del estimador (como variable aleatoria)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57833    0.09294   6.222 1.01e-06 ***
x            1.07031    0.08963  11.942 1.67e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Además el comando `lm` hace un dos test de hipótesis preguntando si los valores de los coeficientes son cero, es decir:

$$H_0 : a = 0 \qquad \text{y} \qquad H_0 : b = 0$$

Como resultado se reporta:

- t`-value` es el valor del estadístico, y nos da el p-valor que calcula

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57833    0.09294   6.222 1.01e-06 ***
x            1.07031    0.08963  11.942 1.67e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

- Luego evalúa cuan bien ajusta el modelo:
  `Residual standard error: 0.5074 on 28 degrees of freedom`

- Primero se calcula el `SSE` que es la sum of squared errors:

$$SSE = \sum_{k=1}^{n} (R_k)^2$$

- Se descuentan la cantidad de coeficientes k menos 1: en este caso $k = 2$ y se resta el intercepto

- Se promedia entre la cantidad de datos menos $k - 1$, se toma raíz para pasar a las unidades del modelo:

$$RSE = \sqrt{\frac{SSE}{n-2}}$$

Otra forma de medir la calidad del ajuste es comparar los residuos versus la variación de las y. Es pensar:

- Si no hay modelo ($a = b = 0$) la estimación es $\bar{Y}_n$ y el error el desvío estándar de las Y

$$SSyy = \sum_{k=1}^{n} (Y_k - \bar{Y}_n)^2$$

- Si hay modelo, ese error es SSE
- Cuanto más pequeño el SSE con respecto al $SSyy$ mejor ajusta el modelo.
- Se calcula entonces el `Multiple R squared`

$$MRS = 1 - \frac{SSE}{SSyy}$$

- Cuanto más cerca de uno, mejor el ajuste

R cuadrado ajustado

- Cuantos más parámetros tenga un modelo mejor ajusta
- Para comparar distintos modelos se descuentan la cantidad de parámetros que se usan.
- Se produce el `Adjusted R squared`
- La fórmula es

$$ARS = 1 - \frac{SSE}{SSyy}\left(\frac{n-1}{n-(k+1)}\right)$$

- Aquí k es la cantidad de parámetros (para nosotros $k = 2$)

¿Son cero los coeficientes?

- El último procedimiento que hace el comando `lm` es un test que asume que todos los coeficientes (salvo el intercepto b)son cero. Es decir

$$H_0 : a = 0 \quad \text{contra} \quad H_1 : a \neq 0$$

- Observar que H_1 consiste en que algún coeficiente no es cero
- El estadístico que se calcula compara el SSE con el $SSyy$
- La idea es que si en realidad $a = b = 0$, entonces los estimadores son cercanos a cero y tenemos

$$SSE \sim SSyy$$

- El estadístico que se cacula es entonces la diferencia, relativa al tamaño de los errores:

$$F = \frac{\frac{SSyy - SSE}{k}}{\frac{SSE}{n-(k+1)}}$$

  Siempre $SSE < SSyy$, entonces $F > 0$

- F tiene distribución de Fisher[2]

- El R nos proporciona el p-valor de este test

```
F-statistic: 142.6 on 1 and 28 DF,  p-value: 1.67e-12
```

- En nuestro caso $k = 1$ (el coef a) y tenemos 28 grados de libertad

[2] Ronald Fisher (Londres, 1890 - Adelaida, Australia 1962) bioestadístico inglés

Generalizaciones: Varios regresores

El modelo que vimos es el más sencillo, pero a veces se requieren modelos más sofisticados

- El vector x se llama regresor
- Podemos tener un modelo de la forma

$$Z = aX + bY + c + \varepsilon$$

- Se ajusta con datos de la forma (Z_k, X_k, Y_k) para $k = 1, \ldots, n$
- Los procedimientos son similares
- En R se escribeustedes dicten
  `lm(formula=z~x+y)`
  si los datos son (z, x, y)

Polinomios

- Queremos ajustar un polinomio de segundo grado, por ejemplo

$$Z = aX + bX^2 + c + \varepsilon$$

- Ponemos en el modelo anterior:

$$Y = X^2$$

- Es decir, elevamos al cuadrado los datos y ajustamos

$$Z = aX + bY + c + \varepsilon$$