Escalamiento de longitud unitaria

$$\mathbb{X} = \begin{bmatrix} | & | & & | \\ \mathbb{1} & x_1 & \cdots & x_K \\ | & | & & | \end{bmatrix}_{n \times p}$$

Nota normalizar es dividir entre la norma (e.g. $v = \frac{x}{\|x\|}$)

Se coge X_j y se normaliza luego de restarle la media de la columna

$$X_j^* = \frac{1}{\sqrt{\sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2}} \begin{bmatrix} x_{1j} - \bar{x}_j \\ x_{2j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix}$$

Se llama de longitud unitaria porque las columnas resultantes son de longitud 1.

i) Y (la respuesta) también se normaliza (Y^*)

ii) Se ajusta el modelo sin intercepto

$$Y_i^* = \sum_{j=1}^{K} \beta_j^* X_{ij}^*$$

¿Por qué no incluir β_0? Una superficie de regresión pasa por el punto $(\bar{Y}, \bar{X})$

Supongamos que se agrega β_0

$$Y_i^* = \beta_0 + \sum_{j=1}^{K} \beta_j^* X_{ij}^*$$ Cambiando X_{ij}^* por la media muestral se tiene que

$$Y_i^* = \beta_0 + \sum_{j=1}^{K} \beta_j^* \cdot 0 = \beta_0,$$ pero la regresión evaluada en $\bar{x}$ es igual a $\bar{Y} \Rightarrow 0 = \beta_0$

1. Responda las siguientes preguntas.

a) Suponga que se realiza escalamiento de longitud unitaria en las predictoras pero no en la variable respuesta, ¿qué unidades tienen los coeficientes de la regresión una vez esta es ajustada?

b) ¿Por qué hay problemas de multicolinealidad cuando se tienen más covariables que observaciones en los datos?

c) Si la traza de la matriz $\mathbf{X}'\mathbf{X}$ es muy grande, ¿mayor es la distancia entre el vector de parámetros estimados y el verdadero vector de parámetros?

d) Si la correlación entre las variables X_j y X_k es pequeña, ¿se puede descartar la presencia de multicolinealidad?

a) $Y = \beta_0 + \beta_1 X_1^* + \cdots + \beta_K X_K^* + \varepsilon, \ \varepsilon \sim N(0, \sigma^2 I_n)$

R// Como X_j^* es adimensional para todo j, se tiene que los coeficientes de regresión cargan con las mismas unidades de la respuesta

b) $0 \leq \text{Rango}(X) \leq \min(n, p) = n$

$0 \geq -\text{Rango}(X) \geq -n$

$p \geq p - \text{Rango}(X) \geq p - n > 0$

$(p > n)$

¿? # Variables libres

Variables libres > 0

Si uno considera $\text{Rango}(X^T X) > 0 \Rightarrow |X^T X| = 0$

$X^T X \beta = X^T Y$ este sistema { No tiene sol / tiene infinitas soluciones

c) Falso, es $(X^T X)^{-1}$

$$Tr\left(\left[X^T X\right]^{-1}\right) = \sum_{i=1}^{k} \frac{1}{\lambda_i}$$

Si $X^T X$ no es invertible $\Rightarrow \lambda_i = 0$ para algún i

$\Rightarrow \sum_{i=1}^{k} \frac{1}{\lambda_i}$ explota

d) No, porque la correlación solo mide la dependencia lineal entre dos variables

c.y. Si $X_1 = 2X_2 + 3X_3$ eso probablemente no sea detectado en $Cor(X_1, X_2)$ y $Cor(X_1, X_3)$

2) Para los VIF (factor de inflación de varianza)

$$VIF_j = \begin{cases} \leq 5 \Rightarrow \text{no hay problemas} \\ 5 < y \leq 10 \text{ hay problemas moderados} \\ > 10 \text{ hay problemas graves} \end{cases}$$

```
> car::vif(mod1)
      x1       x2
294997.4 294997.4
> car::vif(mod2)
      x1       x2
172810.7 172810.7
```

} Hay problemas graves

2)

Haciendo análisis espectral

Collinearity Diagnostics

			Variance Decomposition Proportions				
	Eigen_Value	Condition_Index	Intercept	x1	x2	x_3	x_4
1	2.8349e+00	1.000000	0.024457	0.000000	0.000000	—	—
2	1.6508e-01	4.143992	0.923034	0.000001	0.000001		
3	5.4051e-07	2290.167014	0.052509	0.999999	0.999999	0.6	0.1

x_i está en orden descendente

$\sqrt{K_J}$ índice de condición

π_{ij} proporciones de descomposición de varianza

- # Condición $\sqrt{K} = \sqrt{\lambda_{max}/\lambda_{min}} = 2290.167$ (Se saca de Condition_Index) (es el mayor)

Si $\sqrt{K} \leq 10$, todo oK (no hay problemas)
$10 < \sqrt{K} \leq 31.62$, problemas moderados
$\sqrt{K} > 31.62$, problemas graves

- índice de condición $\sqrt{K_J} = \sqrt{\lambda_{max}/\lambda_J}$

$\sqrt{K} \leq 10$, todo oK (no hay problemas) Para todo J
$10 < \sqrt{K} \leq 31.62$, problemas moderados y Para algún J
$\sqrt{K_J} > 31.62$, problemas graves

hace referencia a los coeficientes de regresión

Proporción de descomposición de varianza: para $i=3$ π_{23} y π_{33} son > 0.5 lo que indica evidencia de multicolinealidad entre X_1 y X_2

3)

k	R_sq	adj_R_sq	SSE	Cp	Variables_in_model		
1	0.017	0.016	29042245	0.101			latitude
1	0.002	0.001	29475559	32.538			longitude
1	0.000	0.000	29527416	36.420			magnitude
2	0.017	0.016	29040930	2.002	latitude	longitude	
2	0.017	0.016	29042210	2.098	latitude		magnitude
2	0.002	0.001	29472811	34.333		longitude	magnitude
3	0.017	0.015	29040902	4.000	latitude	longitude	magnitude

se mira $|Cp - p|$ ↑ pequeño

R_sq: es la columna del R_p^2 y se desea el modelo con R^2 más grande
nos quedamos con el 7 por parsimonia

adj_R_sq: R_p^2 cno busca el más alto, nos quedamos con el modelo 7

Cp: lo que se busca es que el Cp sea bajito y que a su vez $|Cp - p|$ sea mínino también (posible)

El mejor modelo según este criterio es el 5