

# MAT-266: Selección del mejor conjunto de regresores

**Felipe Osorio**

[fosorios.mat.utfsm.cl](mailto:fosorios.mat.utfsm.cl)

Departamento de Matemática, UTFSM



## Objetivo:

Se desea obtener medidas de la calidad o bondad de ajuste del modelo. En efecto, si el modelo está bien ajustado, los residuos tenderán a ser pequeños.

Consideraremos los siguientes procedimientos:

- ▶ Métodos de bondad de ajuste:  $R^2$ ,  $s^2$  y  $C_p$ .
- ▶ Criterios de información.
- ▶ Validación cruzada.
- ▶ Métodos automáticos de selección de variables.



## Bondad de ajuste y selección de modelos

Una medida de bondad de ajuste ampliamente usada es el **coeficiente de determinación**  $R^2$ , que es definido como:

$$R^2 = \{\text{cor}(\mathbf{Y}, \hat{\mathbf{Y}})\}^2.$$

En caso de modelos de regresión **con intercepto**, tenemos

$$R = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\left\{ \sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \right\}^{1/2}}.$$

Sabemos que

$$\sum_{i=1}^n e_i = 0, \quad \sum_{i=1}^n e_i \hat{Y}_i = 0.$$

Además,

$$\frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i.$$



Tenemos

$$\begin{aligned}\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})(\hat{Y}_i - \bar{Y}) \\&= \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\&= \sum_{i=1}^n e_i \hat{Y}_i - \bar{Y} \sum_{i=1}^n e_i + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\&= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.\end{aligned}$$

De ahí que

$$R = \left\{ \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \right\}^{1/2}$$



Sea

$$SYY = \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

y recordando que

$$SYY = \text{RSS} + SS_{\text{Regr}}$$

con  $\text{RSS} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$  y  $SS_{\text{Regr}} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ . Sigue que

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{SS_{\text{Regr}}}{SYY} = 1 - \frac{\text{RSS}}{SYY}.$$

$\sqrt{R^2}$  corresponde al coeficiente de correlación múltiple entre  $\mathbf{Y}$  y  $\hat{\mathbf{Y}}$ , de este modo

$$0 \leq R^2 \leq 1,$$

y un modelo será bien ajustado cuando  $R^2$  sea cercano a 1.



## Bondad de ajuste y selección de modelos

Lamentablemente,  $R^2$  no toma en cuenta la cantidad de parámetros en el modelo. Considere un modelo con  $k$  regresores, entonces podemos usar por ejemplo

$$s_k^2 = \frac{\text{RSS}_k}{n - k}.$$

Otro criterio es el  $R^2$ -ajustado, dado por

$$R_{\text{adj}}^2 = 1 - (1 - R_k^2) \left( \frac{n - 1}{n - p} \right)$$

Asumiendo un modelo con intercepto, tenemos  $R_k^2 = 1 - \text{RSS}_k / \text{SYY}$  y de ahí que

$$R_{\text{adj}}^2 = 1 - \frac{\text{RSS}_k}{\text{SYY}} \left( \frac{n - 1}{n - p} \right) = 1 - \frac{s_k^2}{\text{SYY} / (n - 1)}.$$

De este modo, modelos con máximo  $R_{\text{adj}}^2$  corresponden a modelos con mínimo  $s_k^2$ .



## Bondad de ajuste y selección de modelos

Considere  $\epsilon = Y - \mu$  y suponga un modelo con  $k$  regresores. Luego,  $\hat{\mu} = X\hat{\beta} = HY$  y sea

$$\text{ME} = \|\mu - X\hat{\beta}\|^2,$$

el error de modelo. De este modo,

$$\begin{aligned}\text{ME} &= \|\mu - HY\|^2 = \|\mu - H(\epsilon + \mu)\|^2 = \|(I - H)\mu - H\epsilon\|^2 \\ &= \|(I - H)\mu\|^2 + \|H\epsilon\|^2 = \mu^\top (I - H)\mu + \epsilon^\top H\epsilon.\end{aligned}$$

Esto nos permite obtener

$$\begin{aligned}\text{E}(\text{ME}) &= \mu^\top (I - H)\mu + \text{E}(\epsilon^\top H\epsilon) = \mu^\top (I - H)\mu + \sigma^2 \text{tr } H \\ &= \mu^\top (I - H)\mu + \sigma^2 k\end{aligned}$$

Ahora,

$$\begin{aligned}\text{E}(\text{RSS}_k) &= \text{E}\{Y^\top (I - H)Y\} = \mu^\top (I - H)\mu + (n - k)\sigma^2 \\ &= \text{E}(\text{ME}) + (n - 2k)\sigma^2.\end{aligned}$$



De este modo,

$$\frac{E(\text{ME})}{\sigma^2} = \frac{E(\text{RSS}_k)}{\sigma^2} + 2k - n.$$

Si consideramos estimar  $\sigma^2$  por  $s_p^2$ , entonces podemos usar el criterio  $C_p$  de Mallows, dado por

$$C_p = \frac{\text{RSS}_k}{s_p^2} + 2k - n,$$

como una estimación de  $E(\text{ME})/\sigma^2$ .

Si el modelo está bien ajustado, entonces  $\|(I - H)\mu\|^2 = \mu^\top (I - H)\mu$  será pequeño. Así,

$$\begin{aligned} E(C_p) &\approx \frac{E(\text{RSS}_k)}{\sigma^2} + 2k - n = \frac{\mu^\top (I - H)\mu}{\sigma^2} + n - k + 2k - n \\ &\approx k. \end{aligned}$$

Es decir, escogemos aquél modelo cuyo  $C_p$  sea más cercano a  $k$ .





La **discrepancia de Kullback-Leibler (KL)** entre las funciones de densidad  $g(x)$  y  $f(x)$  es dada por:

$$D_{\text{KL}}(g : f) = \int \log \left( \frac{g(x)}{f(x)} \right) g(x) \, dx = \mathbb{E}_G \left[ \log \left( \frac{g(x)}{f(x)} \right) \right].$$

Propiedades de la discrepancia KL (o información):

(a)  $D_{\text{KL}}(g : f) \geq 0$ .

(b)  $D_{\text{KL}}(g : f) = 0 \Leftrightarrow g(x) = f(x)$  (casi en toda parte).



## Bondad de ajuste y selección de modelos

Suponga  $Y_1, \dots, Y_n$  variables aleatorias siguiendo el modelo verdadero  $g$  y denote por  $\theta_0$  el valor verdadero de  $\theta$ . Considere que se ajusta el modelo candidato  $f(\mathbf{y}; \theta)$  maximizando

$$\ell(\theta) = \sum_{j=1}^n \log f(y_j; \theta)$$

Esto sugiere escoger aquél modelo que minimice la discrepancia  $D_{\text{KL}}(g : f_\theta)$ .

Considere la expansión en series de Taylor,

$$\begin{aligned} \log f(\mathbf{y}; \hat{\theta}) &\approx \log f(\mathbf{y}; \theta_0) + (\hat{\theta} - \theta_0)^\top \frac{\partial \log f(\mathbf{y}; \theta_0)}{\partial \theta} \\ &\quad + \frac{1}{2} (\hat{\theta} - \theta_0)^\top \frac{\partial^2 \log f(\mathbf{y}; \theta_0)}{\partial \theta \partial \theta^\top} (\hat{\theta} - \theta_0). \end{aligned}$$

Notando que  $\theta_0$  minimiza  $D_{\text{KL}}(g : f_\theta)$ , tenemos

$$\int \frac{\partial \log f(\mathbf{y}; \theta_0)}{\partial \theta} g(\mathbf{y}) \, d\mathbf{y} = \mathbf{0}$$



De ahí que

$$\begin{aligned} nD_{\text{KL}}(g : f_{\hat{\theta}}) &= n \int \log \left( \frac{g(\mathbf{y})}{f(\mathbf{y}; \hat{\theta})} \right) g(\mathbf{y}) \, d\mathbf{y} \\ &\approx nD_{\text{KL}}(g : f_{\theta_0}) + \frac{1}{2} \text{tr}\{(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^\top \mathcal{F}_g(\theta_0)\}, \end{aligned}$$

con

$$\mathcal{F}_g(\theta) = -n \int \frac{\partial^2 \log f(\mathbf{y}; \theta)}{\partial \theta \partial \theta^\top} g(\mathbf{y}) \, d\mathbf{y} = n \mathbb{E}_g \left\{ - \frac{\partial^2 \log f(\mathbf{y}; \theta)}{\partial \theta \partial \theta^\top} \right\}.$$

Sea

$$\mathcal{K}_g(\theta) = n \int \frac{\partial \log f(\mathbf{y}; \theta)}{\partial \theta} \frac{\partial \log f(\mathbf{y}; \theta)}{\partial \theta^\top} g(\mathbf{y}) \, d\mathbf{y} = n \text{Cov}_g(\mathbf{U}(\theta)).$$

*Observación:*

Si  $g(\mathbf{y}) = f(\mathbf{y}; \theta)$ , entonces  $\mathcal{F}_g(\theta) = \mathcal{K}_g(\theta) = \mathcal{F}(\theta)$ .



Cuando el modelo es mal especificado, tenemos

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N_p(\mathbf{0}, \mathcal{F}_g^{-1}(\theta_0) \mathcal{K}_g(\theta_0) \mathcal{F}_g^{-1}(\theta)).$$

De ahí que

$$n E_g \{D_{\text{KL}}(g : f_{\hat{\theta}})\} \approx n D_{\text{KL}}(g : f_{\theta_0}) + \frac{1}{2} \text{tr}\{\mathcal{F}_g^{-1}(\theta_0) \mathcal{K}_g(\theta_0)\}. \quad (1)$$

Mientras que, cuando el modelo es correcto y regular, tenemos  $\mathcal{F}_g(\theta_0) = \mathcal{K}_g(\theta_0)$  de modo que

$$\text{tr}\{\mathcal{F}_g^{-1}(\theta_0) \mathcal{K}_g(\theta_0)\} = p$$

Para estimar (1), considere

$$\ell(\hat{\theta}) = \ell(\theta_0) + \{\ell(\hat{\theta}) - \ell(\theta_0)\}.$$



De este modo

$$\begin{aligned} \mathbb{E}_g\{-\ell(\hat{\boldsymbol{\theta}})\} &= -\mathbb{E}\{\ell(\boldsymbol{\theta}_0) + \tfrac{1}{2}LR(\boldsymbol{\theta})\} \\ &\approx nD_{\text{KL}}(g : f_{\boldsymbol{\theta}_0}) - \tfrac{1}{2}\text{tr}\{\mathcal{F}_g^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}_0)\} - n \int g(\mathbf{y}) \log g(\mathbf{y}) \, d\mathbf{y}, \end{aligned}$$

con  $LR(\boldsymbol{\theta}_0) = 2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\}$  el estadístico de razón de verosimilitudes.

Un estimador de (1) es  $-\ell(\hat{\boldsymbol{\theta}}) + c$  donde  $c$  estima  $\text{tr}\{\mathcal{F}_g^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}_0)\}$ . Dos posibles elecciones de  $c$  son  $p$  y  $\text{tr}(\hat{\mathcal{J}}^{-1}\hat{\boldsymbol{\mathcal{K}}})$  con

$$\begin{aligned} \hat{\mathcal{J}} &= -\sum_{j=1}^n \frac{\partial^2 \log f(y_j; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}, \\ \hat{\boldsymbol{\mathcal{K}}} &= \sum_{j=1}^n \frac{\partial \log f(y_j; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(y_j; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^\top}. \end{aligned}$$



Lo anterior lleva a los **criterios de información de Akaike y de la red**:

$$AIC = -2\ell(\hat{\boldsymbol{\theta}}) + 2p,$$

$$NIC = -2\ell(\hat{\boldsymbol{\theta}}) + 2 \operatorname{tr}(\hat{\mathcal{J}}^{-1} \hat{\mathcal{K}}),$$

otra posibilidad es el **criterio de información de Schwarz** (bayesiano), dado por

$$SIC = -2\ell(\hat{\boldsymbol{\theta}}) + p \log n.$$



Suponga en modelo de regresión lineal con  $k$  regresores, es decir  $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$  es vector  $(k+1)$ -dimensional. Tenemos

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Además,  $\text{RSS} = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$ ,  $\hat{\sigma}^2 = \text{RSS} / n$ . De este modo,

$$\begin{aligned} AIC &= n \log 2\pi + n \log \hat{\sigma}^2 + \frac{1}{\hat{\sigma}^2} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + 2(k+1) \\ &= n \log 2\pi + n \log(\text{RSS} / n) + \frac{n}{\text{RSS}} \text{RSS} + 2(k+1) \\ &= n \log(\text{RSS} / n) + 2(k+1) + n(\log 2\pi + 1) \end{aligned}$$



Suponga  $(\tilde{\mathbf{x}}_1, \tilde{Y}_1), \dots, (\tilde{\mathbf{x}}_n, \tilde{Y}_n)$  un conjunto de datos nuevos que siguen el mismo modelo que los datos de entrenamiento  $(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$ . Podemos considerar la siguiente medida de error (de predicción)

$$\text{PE} = \frac{1}{m} \sum_{j=1}^m (\tilde{Y}_i - \tilde{\mathbf{x}}_i^\top \hat{\boldsymbol{\beta}})^2,$$

donde  $\hat{\boldsymbol{\beta}}$  es calculado usando los datos de entrenamiento.

### Objetivo:

Podemos subdividir el conjunto de entrenamiento en dos conjuntos disjuntos uno para obtener  $\hat{\boldsymbol{\beta}}$  y otro para medir el error.





Considere seleccionar un subconjunto  $D$  con  $d$  observaciones, donde usamos las  $n - d$  observaciones restantes para calcular  $PE(D)$ .

Una alternativa es repetir el proceso, seleccionando  $D_1, D_2, \dots$ , y promediar  $PE(D_1), PE(D_2), \dots$ . Este método es llamado validación cruzada.

Existen diversos procedimientos para elegir tales subconjuntos:

- ▶ Métodos exhaustivos: [leave- \$p\$ -out](#), [leave-one-out](#).
- ▶ Métodos no exhaustivos: [k-fold](#), [Monte Carlo CV](#).



Considere seleccionar un subconjunto  $D$  con  $d$  observaciones, donde usamos las  $n - d$  observaciones restantes para calcular  $PE(D)$ .

Una alternativa es repetir el proceso, seleccionando  $D_1, D_2, \dots$ , y promediar  $PE(D_1), PE(D_2), \dots$ . Este método es llamado validación cruzada.

Existen diversos procedimientos para elegir tales subconjuntos:

- ▶ Métodos exhaustivos: [leave-p-out](#), [leave-one-out](#).
- ▶ Métodos no exhaustivos: [k-fold](#), [Monte Carlo CV](#).



La versión más simple de validación cruzada es eliminar una observación a la vez<sup>1</sup> y obtener  $\hat{\beta}_{(i)}$ ,  $i = 1, \dots, n$ . Esto lleva al error de predicción leave-one-out o CV

$$\text{CV} = \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{x}_i^\top \hat{\beta}_{(i)})^2$$

### Observación:

La estadística

$$\text{PRESS} = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2 = \sum_{i=1}^n (Y_i - \mathbf{x}_i^\top \hat{\beta}_{(i)})^2$$

es conocida como **suma de cuadrados (residual) de predicción**.

---

<sup>1</sup>Seleccionar conjuntos  $D$  con  $d > 1$  es computacionalmente intenso.

Es fácil mostrar que

$$\begin{aligned} Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{(i)} &= Y_i - \mathbf{x}_i^\top \left[ \hat{\boldsymbol{\beta}} - \frac{e_i}{1 - h_{ii}} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i \right] \\ &= Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{e_i h_{ii}}{1 - h_{ii}} = \frac{e_i}{1 - h_{ii}}. \end{aligned}$$

De ahí que

$$CV = \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{1 - h_{ii}} \right)^2.$$

### Observación:

Anteriormente usamos el [criterio de validación cruzada generalizada](#) para seleccionar el parámetro de sesgo  $k$  en regresión ridge

$$V(k) = \frac{1}{n} \frac{\|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}_k\|^2}{\{\text{tr}(\mathbf{I} - \mathbf{H}(k))/n\}^2}.$$



### *Ejemplo:*

Considere datos de ventas en 15 regiones con 3 regresores  $X_1$ ,  $X_2$  y  $X_3$ , donde se obtuvo

$p$	Variables	$RSS_p$
1	—	428 144.64
2	$X_1$	88 473.00
	$X_2$	44 683.00
	$X_3$	32 483.00
3	$X_1, X_2$	43 968.00
	$X_1, X_3$	32 086.00
	$X_2, X_3$	535.00
4	$X_1, X_2, X_3$	273.00



## Bondad de ajuste y selección de modelos

Tenemos que

$$SYY = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 428144.64, \quad s^2 = \frac{RSS}{n-p} = \frac{273.00}{15-4} = 24.818$$

De este modo, los criterios de selección de modelos resultan:

Variables	$R_p^2$	$s_p^2$	$R_{adj}^2$	$C_p$	$AIC$	$SIC$
—	0.000	30581.760	0.000	17238.249	157.887	159.304
$X_1$	0.793	6805.615	0.777	3553.846	136.236	138.360
$X_2$	0.896	3437.154	0.888	1789.414	125.989	128.114
$X_3$	0.924	2498.692	0.918	1297.839	121.206	123.330
$X_1, X_2$	0.897	3664.000	0.880	1762.604	127.748	130.580
$X_1, X_3$	0.925	2673.833	0.913	1283.842	123.022	125.854
$X_2, X_3$	0.999	44.583	0.999	12.557	61.613	64.445
$X_1, X_2, X_3$	0.999	24.818	0.999	4.000	53.521	57.062



## Selección automática de variables regresoras

```
# ajuste LS para los datos de fecundidad en Suiza  
> fm <- lm(Fertility ~ ., data = swiss)  
> fm
```

```
Call:  
lm(formula = Fertility ~ ., data = swiss)
```

Coefficients:

(Intercept)	Agriculture	Examination
66.9152	-0.1721	-0.2580
Education	Catholic	Infant.Mortality
-0.8709	0.1041	1.0770



# Selección automática de variables regresoras

```
# stepwise regression
> z <- step(fm, direction = "both")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
          Infant.Mortality
```

	Df	Sum of Sq	RSS	AIC
- Examination	1	53.03	2158.1	189.86
<none>			2105.0	190.69
- Agriculture	1	307.72	2412.8	195.10
- Infant.Mortality	1	408.75	2513.8	197.03
- Catholic	1	447.71	2552.8	197.75
- Education	1	1162.56	3267.6	209.36

```
Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
```

	Df	Sum of Sq	RSS	AIC
<none>			2158.1	189.86
+ Examination	1	53.03	2105.0	190.69
- Agriculture	1	264.18	2422.2	193.29
- Infant.Mortality	1	409.81	2567.9	196.03
- Catholic	1	956.57	3114.6	205.10
- Education	1	2249.97	4408.0	221.43





# Selección automática de variables regresoras

```
# backward selection
> z <- step(fm, direction = "backward")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
          Infant.Mortality
```

	Df	Sum of Sq	RSS	AIC
- Examination	1	53.03	2158.1	189.86
<none>			2105.0	190.69
- Agriculture	1	307.72	2412.8	195.10
- Infant.Mortality	1	408.75	2513.8	197.03
- Catholic	1	447.71	2552.8	197.75
- Education	1	1162.56	3267.6	209.36

```
Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
```

	Df	Sum of Sq	RSS	AIC
<none>			2158.1	189.86
- Agriculture	1	264.18	2422.2	193.29
- Infant.Mortality	1	409.81	2567.9	196.03
- Catholic	1	956.57	3114.6	205.10
- Education	1	2249.97	4408.0	221.43



