

Clase auxiliar # 7

Selección de modelos (método LASSO y test de Ramsey) y omisión de variables relevantes

Método LASSO: este método busca minimizar la suma de los residuos al cuadrado (como ya se vio antes), pero penalizada por un factor que permite controlar la cantidad de variables.

$$\min_{\hat{\beta}} \sum_{n=1}^N (Y_n - \hat{\beta}_0 - \sum_{k=1}^K \hat{\beta}_k X_{nk})^2 + \lambda \sum_{k=1}^K |\hat{\beta}_k|$$

Test Reset de Ramsey:

1. Especificar $Y = X\beta + U$ y estimar $\hat{\beta}$ e $\hat{Y}$.
2. A partir de la estimación $\hat{Y}$, crear las variables $\hat{Y}^p$ para $p \in \{1, \dots, P\}$
3. Regresionar $Y = X\beta + \gamma_2 \hat{Y}^2 + \dots + \gamma_P \hat{Y}^P + V$
4. Testear $H_0 : \gamma_2 = \dots = \gamma_P = 0$ (restricciones múltiples).

1. Método LASSO

Tras fracasar en la Liga Kalos, nuestro héroe Ash Ketchum de Pueblo Paleta, ha decidido dedicarse a su verdadera vocación: la política. Afortunadamente y contra todo pronóstico, Ash ha logrado la elección como presidente de su país. Los analistas políticos buscan una explicación y lo contratan a usted para seleccionar aquellas variables que **no** deberían considerar en su análisis.

Para llevar a cabo esta tarea, se le ha entregado la base de datos almacenada en el archivo `auxiliar5.xlsx` (almacenado en Material Docente de U-Cursos). Dicha tiene información de distintas ciudades y pueblos, como el promedio de escolaridad, desempleo, ingreso, nivel de los pokemones de la zona, un índice de ideología política, la popularidad de Pikachu y una serie de “variables basura”.

1. Elija un valor de λ y ejecute el método LASSO.
2. ¿Qué variables descartaría en primera instancia? ¿Por qué?
3. Discuta: ¿cuál es la importancia del parámetro λ en el problema de optimización?

Comente:

4. “El método LASSO sirve para seleccionar variables o forma funcional correcta dentro de un modelo, incluso si hay un número mayor de variables que observaciones”.
5. “No es posible aproximar un modelo cuya forma funcional no sea polinómica mediante el método LASSO”.

2. Test Reset de Ramsey

Utilise la base de datos Auxiliar 07 Test Reset Ramsey BD.dta y considere el modelo $y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + u_i$.

1. ¿Qué modelo plantearía para realizar el test Reset de Ramsey con potencias $p = 2, 3$?
2. Plantee un test de hipótesis (hipótesis nula, hipótesis alternativa, estadístico y región de rechazo) para dicho test.
3. Realice dicho test en *Stata*. ¿Qué puede concluir?

```
. reg log_sal esc exp exp2
```

Source	SS	df	MS	Number of obs	=	97.926
Model	15100,4243	3	5033,47478	F(3, 97922)	=	11036,67
Residual	44659,1283	97.922	,456068384	Prob > F	=	0,0000
Total	59759,5527	97.925	,610258388	R-squared	=	0,2527
				Adj R-squared	=	0,2527
				Root MSE	=	,67533

log_sal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
esc	,114268	,0006729	169,82	0,000	,1129492 ,1155869
exp	,0115993	,0004689	24,74	0,000	,0106803 ,0125183
exp2	-,000063	8,17e-06	-7,71	0,000	-,000079 -,000047
_cons	7,38576	,0110628	667,62	0,000	7,364077 7,407443

3. Omisión de variables relevantes

1. Demuestre que al omitir variables relevantes $\hat{\beta}_{MCO}$ es sesgado.
2. Explique a qué se deben las diferencias de $\hat{\beta}_{esc}$ y $\hat{\hat{V}}(\hat{\beta}_{esc})$ entre las regresiones de las tablas que se muestran a continuación.

log_sal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
esc	,0974079	,0005637	172,80	0,000	,0963031 ,0985128
_cons	7,818565	,0068344	1144,01	0,000	7,80517 7,831961

log_sal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
esc	,114268	,0006729	169,82	0,000	,1129492 ,1155869
exp	,0115993	,0004689	24,74	0,000	,0106803 ,0125183
exp2	-,000063	8,17e-06	-7,71	0,000	-,000079 -,000047
_cons	7,38576	,0110628	667,62	0,000	7,364077 7,407443