



PRUEBAS SELECTIVAS PARA EL INGRESO EN EL CUERPO SUPERIOR DE ACTUARIOS, ESTADÍSTICOS Y ECONOMISTAS DE LA ADMINISTRACIÓN DE LA SEGURIDAD SOCIAL

TERCER EJERCICIO: 9 de enero de 2024

ÁREA DE ESTADÍSTICA

Primer supuesto

En una región se desea estimar el precio medio anual de los seguros de salud. Para ello se han definido dos estratos dependiendo de si el municipio de residencia del asegurado tiene o no centro de salud. En el estrato 1 (municipios con centro de salud) residen 10.000 personas cuya edad media es de 50 años y en el estrato 2 (municipio sin centro de salud) residen 5.000 personas cuya edad media es de 55 años. La siguiente tabla muestra la información recopilada de una muestra aleatoria simple sin reposición en cada estrato:

Estrato 1 Municipios con centro de salud	Precio	Edad	Estrato 2 Municipios sin centro de salud	Precio	Edad
	190	31		270	34
	241	37		492	69
	398	63		388	60
	327	59		300	45
	288	47		370	57
	292	46			
	326	52			
	206	33			
	310	60			
	410	70			

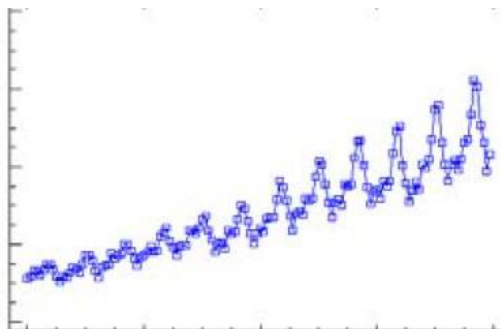
- Estimar el precio medio de los seguros de salud mediante un estimador de regresión separado utilizando como variable auxiliar la edad del asegurado.
- Calcular la varianza del precio medio de los seguros de salud de la estimación de regresión separada que utiliza como variable auxiliar la edad del asegurado.
- Determinar el tamaño muestral que hubiera sido necesario para que la estimación estratificada separada del precio medio de los seguros de salud (muestreo aleatorio sin reposición y sin tener en cuenta la variable edad) no difiriera en más de 5 del verdadero valor aplicando una afijación de mínima varianza o de Neyman suponiendo que la variabilidad de los seguros de salud es similar a la variabilidad de los seguros de vida: $S_1^2 = 2.500$ y $S_2^2 = 6.400$.



Segundo supuesto

a. Análisis e identificación gráfica:

1. En la serie temporal representada en el gráfico siguiente:



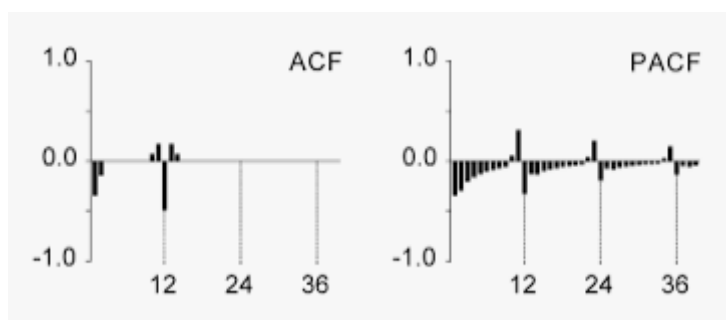
Razone sobre su estacionariedad en el contexto de los modelos ARIMA.

2. En otro proceso ARIMA donde las función teóricas de autocorrelación (FAC) y de autocorrelación parcial (FACP) son de la siguiente forma:



Indique los órdenes “p” y “q” de las correspondientes partes autorregresivas y de media móvil del modelo ARIMA (p,d,q).

3. En otro proceso ARIMA estacional se observa que:



Indique los órdenes “p”, “q”, “P” y “Q” de las correspondientes partes autorregresivas y de media móvil del modelo SARIMA (p,d,q) x (P,D,Q).



b. Suponga el siguiente proceso autorregresivo AR(2) :

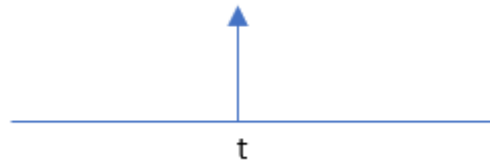
$$Y_t = 0,8 Y_{t-1} - 0,15 Y_{t-2} + a_t \quad \text{con } \sigma_a^2 = 1$$

Calcule la media del proceso, suponiendo que $E(a_t) = 0$

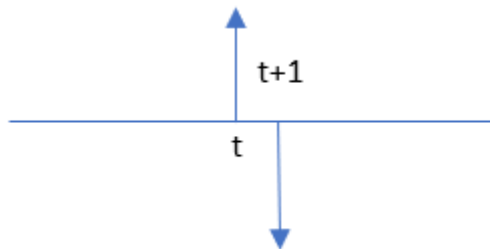
c. En el proceso AR(2) del apartado b), justifique matemáticamente si es estacionario.

d. En una serie temporal se detecta un outlier o anomalía en un momento “t”, explique el tipo de anomalía en cada uno de los siguientes casos, si la representación gráfica teórica del resultado de aplicar el operador diferencias sobre la serie es:

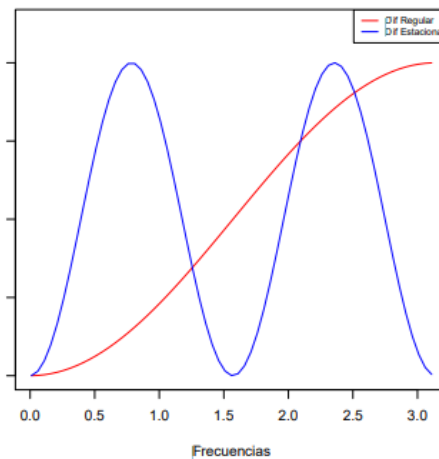
1) Caso 1:



2) Caso 2:



e. Suponga una serie temporal trimestral. En el siguiente gráfico se muestran los espectros de los operadores diferencias regular y estacional:



Explique los efectos de ambos filtros.



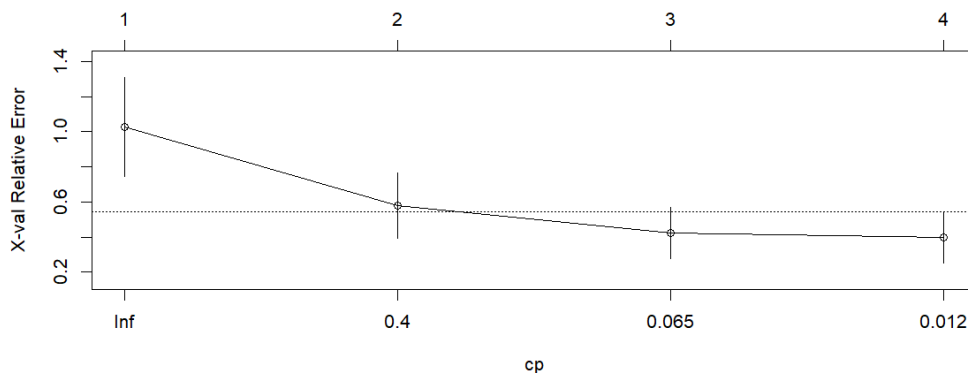
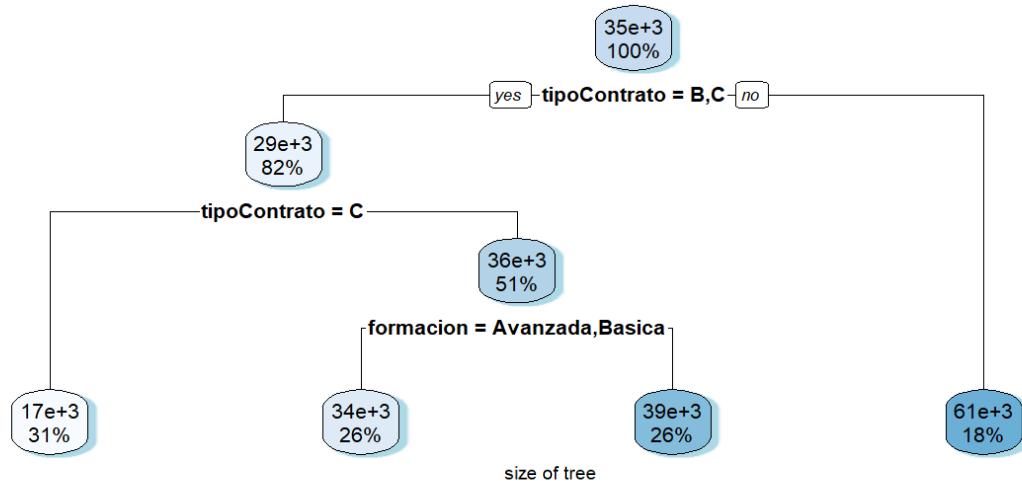
Tercer supuesto

- a. Se quiere predecir el valor del salario de un nuevo trabajador teniendo en cuenta variables auxiliares: sexo, salario, tipo de contrato, educación, siHorasExtra (si hacen o no horas extra), internacional (si trabajan o no en mercado internacional). Para ello se dispone de los datos de los 57 trabajadores de la empresa (en el escritorio de su ordenador podrá encontrar un fichero excel denominado "Datos tercer supuesto área Estadística - ejercicio 3 CSAEEE.xlsx"). Se decide ajustar un modelo de aprendizaje automático basado en árboles de regresión. Interpreta los resultados expuestos a continuación y explica cómo se podrían realizar posibles mejoras de los mismos.

```
formula = salario ~ sexo + tipoContrato + educación + siHorasExtra +  
internacional  
splitting index: Gini; method: ANOVA  
Training data set: 70%  
Test data set: 30%  
hyperparam grid top5:
```

	minsplit	maxdepth		cp	error
1		10	14	0.03341214	0.2723181
2		11	14	0.03341214	0.2737975
3		10	11	0.03341214	0.2758402
4		12	11	0.03341214	0.2759927
5		7	10	0.03341214	0.2770316

Árbol de regresión para salario



MSE: 25872.86



- b. Finalmente se quiere predecir si un trabajador realiza horas extras o no en base a su salario, sexo, tipo de contrato, educación e internacional. Se dispone de datos de 57 trabajadores de una empresa. Se ajusta un modelo de aprendizaje automático de random forest. Interpreta los resultados expuestos a continuación y explica cómo se podrían realizar posibles mejoras de los mismos u otros análisis más exhaustivos.

Call:

```
ranger(siHorasExtra ~ salario + sexo + tipoContrato + educacion +  
internacional, data = data_train, importance = "impurity")
```

Type:	Classification
Number of trees:	500
Sample size:	41
Number of independent variables:	5
Mtry:	2
Target node size:	1
Variable importance mode:	impurity
Splitrule:	gini
OOB prediction error:	36.59 %

Confusion Matrix and Statistics

	Reference
Prediction 0 1	
0 9 2	
1 2 3	

Accuracy : 0.75
95% CI : (0.4762, 0.9273)
No Information Rate : 0.6875
P-Value [Acc > NIR] : 0.4069

Kappa : 0.4182

Mcnemar's Test P-Value : 1.0000

Sensitivity : 0.6000
Specificity : 0.8182
Pos Pred Value : 0.6000
Neg Pred Value : 0.8182
Prevalence : 0.3125
Detection Rate : 0.1875
Detection Prevalence : 0.3125
Balanced Accuracy : 0.7091

'Positive' Class : 1



MINISTERIO
DE INCLUSIÓN, SEGURIDAD SOCIAL
Y MIGRACIONES

```
Impurity_importance  Variables
7.7082436            salario
3.8472330            educacion
3.0841560            tipoContrato
1.7804285            internacional
0.9374767            sexo

> data_train[, table(siHorasExtra)]
siHorasExtra
 0  1
21 20
```