

GEM

Alfonso Camarillo Núñez

12 de marzo de 2024

Data Analytics con R

Docente Manuel Juipa

Resolver el caso de retención de empleados de la empresa STORE24 haciendo un modelo de MINIMOS CUADRADOS ORDINARIOS o MINIMOS CUADRADOS PONDERADOS.

Analizando el problema queremos ver qué relación hay entre la utilidad y la antigüedad por lo que realizamos un análisis lineal para observar esta relación en conjunto con las demás variables las cuales tienen la información en el archivo “store24_data.xls”.

Por lo que esperamos tener un modelo en el cual podamos analizar la utilidad en base a la antigüedad de los gerentes y empleados además de las variables con las que contamos, pero al hacer directamente el modelo1:

Profit ~ Sales + MTenure + CTenure + Pop + Comp + Visibility + PedCount + Res + Hours24 + CrewSkill + MgrSkill + ServQual

Da los siguientes coeficientes

Coefficients:							
(Intercept)	Sales	MTenure	CTenure	Pop	Comp	Visibility	PedCount
-9.506e+04	2.161e-01	1.569e+02	2.661e+02	5.920e-01	-1.285e+04	3.610e+03	9.570e+03
Res	Hours24	CrewSkill	MgrSkill	ServQual			
3.149e+04	2.849e+03	-1.900e+03	1.983e+04	6.082e+01			

Donde vemos que hay parámetros que son significativos para las utilidades como lo es la antigüedad de los gerentes o las ventas como se esperaba, pero otros no son significativos como la antigüedad del personal, a pesar de que el modelo explica con r^2 ajustado de 0.8681.

Residuals:				
Min	1Q	Median	3Q	Max
-81468	-20403	618	17325	77033

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-9.506e+04	6.168e+04	-1.541	0.1283
Sales	2.161e-01	2.176e-02	9.931	1.95e-14 ***
MTenure	1.569e+02	9.171e+01	1.710	0.0922 .
CTenure	2.661e+02	2.580e+02	1.031	0.3063
Pop	5.920e-01	9.163e-01	0.646	0.5206
Comp	-1.285e+04	3.602e+03	-3.569	0.0007 ***
Visibility	3.610e+03	5.401e+03	0.668	0.5064
PedCount	9.570e+03	5.771e+03	1.658	0.1023
Res	3.149e+04	2.333e+04	1.350	0.1819
Hours24	2.849e+03	1.308e+04	0.218	0.8283
CrewSkill	-1.900e+03	1.078e+04	-0.176	0.8606
MgrSkill	1.983e+04	1.080e+04	1.836	0.0711 .
ServQual	6.082e+01	3.526e+02	0.172	0.8636

Signif. codes: 0 '***' 0.001 '**' 0.01 '.' 0.05 ' ' 0.1 ' ' 1

Residual standard error: 32470 on 62 degrees of freedom
Multiple R-squared: 0.8895, Adjusted R-squared: 0.8681
F-statistic: 41.58 on 12 and 62 DF, p-value: < 2.2e-16

Al probar los supuestos con un orden 1 vemos que, si se cumplen que los errores son homocedasticos, no tienen autocorrelación y tienen normalidad.

<pre>> #probamos los supuestos</pre>				
<pre>> test_ols(m= prueba1, order = 1)</pre>				
# A tibble: 3 x 4				
Assumption	Test	Statistic	P-value	
<chr>	<chr>	<dbl>	<dbl>	<dbl>
1 Homoskedasticity	studentized Breusch-Pagan test	14.8	0.251	
2 No serial correlation	Breusch-Godfrey test for serial correlation of order up to 1	0.319	0.572	
3 Normality	Jarque Bera Test	1.56	0.457	

Al probar los supuestos con un orden 2 vemos que no cumple con el supuesto de correlación.

```
> #probamos 1so supuestos
> test_ols(m= prueba, order = 2)
# A tibble: 3 x 4
  Assumption      Test      Statistic 'P-value'
  <chr>          <chr>          <dbl>    <dbl>
1 Homoskedasticity studentized Breusch-Pagan test 14.8     0.251
2 No serial correlation Breusch-Godfrey test for serial correlation of order up to 2 7.06     0.029
3 Normality Jarque Bera Test 1.56     0.457
```

Por lo que proponemos un segundo modelo ya pensando en usar el margen de utilidad bruta de ventas el cual es una métrica en base a la utilidad entre las ventas, con este modelo2 probamos, pero tampoco nos ayuda con el problema de la correlación.

(Profit/Sales) ~ Sales + (MTenure/Sales) + (CTenure/Sales) + (Pop/Sales) + (Comp/Sales) + (Visibility/Sales) + (PedCount/Sales) + (Res/Sales) + (Hours24/Sales) + (CrewSkill/Sales) + (MgrSkill/Sales) + (ServQual/Sales)

```
> #probamos 1so supuestos
> test_ols(m= prueba2, order = 2)
# A tibble: 3 x 4
  Assumption      Test      Statistic 'P-value'
  <chr>          <chr>          <dbl>    <dbl>
1 Homoskedasticity studentized Breusch-Pagan test 16.2     0.849
2 No serial correlation Breusch-Godfrey test for serial correlation of order up to 2 11.4     0.003
3 Normality Jarque Bera Test 2.77     0.251
```

Por lo que una nueva prueba seria la segmentación con grupos de 25 datos en cada uno para tener la misma cantidad en cada grupo estos ordenados de menor a mayores utilidades, para poder visualizar el efecto en base a las que tienen más utilidades de las que tienen menos.

Al evaluar los modelos planteados en se nota que hay un problema de correlación a pesar de probar de forma segmentada con los dos modelos.

Modelo 1

```
[[1]]
# A tibble: 3 x 4
  Assumption      Test      Statistic 'P-value'
  <chr>          <chr>          <dbl>    <dbl>
1 Homoskedasticity studentized Breusch-Pagan test 4.67     0.946
2 No serial correlation Breusch-Godfrey test for serial correlation of order up to 1 7.82     0.005
3 Normality Jarque Bera Test 0.152     0.927

[[2]]
# A tibble: 3 x 4
  Assumption      Test      Statistic 'P-value'
  <chr>          <chr>          <dbl>    <dbl>
1 Homoskedasticity studentized Breusch-Pagan test 13.3     0.351
2 No serial correlation Breusch-Godfrey test for serial correlation of order up to 1 1.95     0.162
3 Normality Jarque Bera Test 0.63     0.73

[[3]]
# A tibble: 3 x 4
  Assumption      Test      Statistic 'P-value'
  <chr>          <chr>          <dbl>    <dbl>
1 Homoskedasticity studentized Breusch-Pagan test 4.68     0.968
2 No serial correlation Breusch-Godfrey test for serial correlation of order up to 1 0.669     0.413
3 Normality Jarque Bera Test 12.2     0.002
```

Modelo2

```

[[1]]
# A tibble: 3 × 4
  Assumption      Test      Statistic 'P-value'
  <chr>          <chr>      <dbl>    <dbl>
1 Homoskedasticity studentized Breusch-Pagan test 18.3      0.627
2 No serial correlation Breusch-Godfrey test for serial correlation of order up to 1 11.5      0.001
3 Normality Jarque Bera Test 2.57      0.277

[[2]]
# A tibble: 3 × 4
  Assumption      Test      Statistic 'P-value'
  <chr>          <chr>      <dbl>    <dbl>
1 Homoskedasticity studentized Breusch-Pagan test 21.7      0.479
2 No serial correlation Breusch-Godfrey test for serial correlation of order up to 1 21.2      0
3 Normality Jarque Bera Test 0.776      0.678

[[3]]
# A tibble: 3 × 4
  Assumption      Test      Statistic 'P-value'
  <chr>          <chr>      <dbl>    <dbl>
1 Homoskedasticity studentized Breusch-Pagan test 22.2      0.51
2 No serial correlation Breusch-Godfrey test for serial correlation of order up to 1 25      0
3 Normality Jarque Bera Test 1.92      0.382

```

Después de revisar los modelos propuestos consideramos quedarnos con el primero porque al cumple con los supuestos de normalidad, homocedasticidad y correlación y podría usarse para estimar las utilidades en relación con la antigüedad de los trabajadores y sus habilidades, siendo la de los gerentes la que más significancia tiene, comparándolo con la antigüedad de los demás trabajadores.

Por lo que entenderíamos de ese modelo que al incrementar en una unidad la utilidad se incrementa en $1.569e+02$ la antigüedad del gerente si las demás variables quedan constantes.