Cablinski Pablo Luis

Regresión Avanzada Examen Diciembre 2024

1. a) Construya el mejor modelo lineal simple y realice el análisis diagnóstico del mismo. En caso de ser necesario transforme la variable respuesta. Concluya.
2. El mejor modelo lineal simple para explicar el salario usando las variables continuas surge de usar como predictor **educ** (años de educación) con un R2 ajustado de 0.1632 con lo cual se puede concluir que el 16% de la variabilidad de salario se puede explicar mediante los años de educación a través de la recta de regresión. El modelo resulta significativo dado el p-valor del estadístico F.
3. Del análisis diagnóstico surge que se rechazan los 3 supuestos: Normalidad, Heterocedasticidad e Independencia.
4. La gráfica muestra estructura en los residuos y la falta de normalidad

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Se hace una transformación de Box Cox obteniendo un lambda de -0.18. Al ajustar con ese lambda se sigue rechazando la normalidad de los residuos y la independencia pero se logra la homocedasticidad de los mismos de acuerdo a los tests Breusch Pagan y NCVTest

Gráfico

Descripción generada automáticamente

b) Mediante selección de variables elija el mejor modelo multivariado para explicar salario. Estudie la presencia de multicolinealidad. Si la hubiera aplique alguna metodología para evitarla.

El mejor según el R2 ajustado es el Modelo 11 : lm(formula = salario ~ educ + exper + tenure + female + married +numdep + smsa + south + west + construc + profocc + servocc, data = datos)

**R^2 ajustado: 0.418146**

Del análisis del VIF surgen que hay variables con valor VIF mayor a 5 y además hay coeficientes no significativos lo que implica presencia de multicolinealidad

Eliminando aquellos con mayor VIF y eligiendo solo aquellos coeficiente significativos del modelo se arriba a la siguiente fórmula

lm(formula = salario ~ educ + tenure + female + smsa + northcen + profocc, data = datos) con un **R^2 ajustado 0.4362**

c) Analice la pertinencia de un modelo gamlss para explicar Salario.

En este caso se sugiere la utilización de un modelo GAM ya que si bien es necesario hallar una función a cada observación utilizando la transformación de Box Cox se consigue homocedaticidad en los residuos por lo que no es necesario modelizar la varianza mediante un modelo gamlss.

d) Call:

glm(formula = salario\_grupo ~ educ + tenure + female + smsa + profocc, family = binomial, data = datos)

(Intercept) educ tenure female1 smsa1 profocc1

-5.04226193 0.21945612 0.08762021 -1.32554052 0.73570252 1.17154723

---

Null deviance: 599.14 on 525 degrees of freedom

Residual deviance: 431.25 on 520 degrees of freedom

AIC: 443.25

**Análisis de los coeficientes**

* **(Intercepto)**: -5.04226, p-value < 0.001. El intercepto es significativo, indicando que cuando todas las variables predictoras son 0, la chance de pertenecer al grupo con salario mayor o igual a 6.88 es muy baja.
* **educ**: 0.21946, p-value < 0.001. La educación tiene un efecto positivo y significativo en la probabilidad de pertenecer al grupo con salario mayor o igual a 6.88.
* **tenure**: 0.08762, p-value < 0.001. La antigüedad en el trabajo también tiene un efecto positivo y significativo.
* **female1**: -1.32554, p-value < 0.001. Ser mujer tiene un efecto negativo y significativo en la probabilidad de pertenecer al grupo con salario mayor o igual a 6.88.
* **smsa1**: 0.73570, p-value = 0.0165. Vivir en un área metropolitana tiene un efecto positivo y significativo.
* **profocc1**: 1.17155, p-value < 0.001. Tener una ocupación profesional tiene un efecto positivo y significativo.

**Prueba de bondad del ajuste**En este caso, el test de Hosmer y Lemeshow da un p valor de 0.1596, no hay evidencia significativa de que el modelo logístico no se ajuste bien a los datos. Por lo tanto es posible concluir que el modelo tiene un buen ajuste.

**Conclusión:** *El modelo logístico muestra que la educación, la antigüedad en el trabajo, vivir en un área metropolitana, y tener una ocupación profesional aumentan la probabilidad de tener un salario mayor o igual a 6.88, mientras que ser mujer disminuye esa probabilidad.*

e) A modo de ejemplo se compara el resultado obtenido mediante el modelo logístico con el equivalente utilizando un árbol de decisión con RPART de R

**Matriz de confusión del modelo logístico:**

Reference

Prediction 0 1

0 359 63

1 32 72

Accuracy : 0.8194

Sensitivity : 0.9182

Specificity : 0.5333

Pos Pred Value : 0.8507

Neg Pred Value : 0.6923

**Matriz de confusión del árbol de decisión:**

**Reference**

Prediction 0 1

0 372 73

1 19 62

Accuracy : 0.8251

Sensitivity : 0.9514

Specificity : 0.4593

Pos Pred Value : 0.8360

Neg Pred Value : 0.7654

Del análisis de las métricas se puede concluir que ambos métodos son similares en cuanto a su capacidad de clasificación, sin embargo la interpretabilidad del modelo logístico es muy superior por lo que en este caso lo hace preferible.