XS-2130 Modelos de Regresión Aplicados II Sem 2020 Grupo 2

Práctica examen parcial 2.

Pregunta 1.

Un tecnólogo de alimentos cree que el éxito en producir “quequitos” (“cupcakes”) está en factores que se encuentran en distintas etapas del proceso: el refinamiento de la harina, la temperatura a la que se deja crecer la mezcla, el tamaño del molde en el que se hornea al queque completo antes de sacar los quequitos, y si al quequito se le pone gotas de limón ya terminado. La variable dependiente es una escala promedio entre 0 a 100 con que se calificó el gusto por el quequito. Los resultados del experimento en el que analiza el peso de los factores los generó a partir de un modelo de regresión. La salida de R se ofrece más abajo. A partir de ella conteste las siguientes preguntas.

a) Calcule el R cuadrado ajustado (3 ptos.)

b) Si la log-verosimilitud mulitplicada por -2, de un modelo gaussiano es equivalente a: [n log(SCRes/n)], calcule el BIC del modelo (3 ptos.)

c) ¿Cuál modelo es preferible: el modelo2 o el modelo3?, y ¿por qué? (3 ptos.)

> modelo1=lm(escala~tamanyo+temperatura+harina)

> summary(modelo1)

Call:

lm(formula = escala ~ tamanyo + temperatura + harina)

Residuals:

Min 1Q Median 3Q Max

-20.9531 -4.2708 0.5495 4.6003 14.1562

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.521 3.010 2.499 0.0142 \*

tamanyo 29.479 1.460 20.193 <2e-16 \*\*\*

temperatura 10.891 0.894 12.182 <2e-16 \*\*\*

harina 5.062 1.460 3.468 0.0008 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.152 on 92 degrees of freedom

Multiple R-squared: 0.8606, Adjusted R-squared: XXX

F-statistic: 189.4 on 3 and 92 DF, p-value: < 2.2e-16

> anova(modelo1)

Analysis of Variance Table

Response: escala

Df Sum Sq Mean Sq F value Pr(>F)

tamanyo 1 20856.5 20856.5 407.746 < 2.2e-16 \*\*\*

temperatura 1 7590.8 7590.8 148.400 < 2.2e-16 \*\*\*

harina 1 615.1 615.1 12.025 0.0008001 \*\*\*

Residuals 92 4705.9 51.2

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> modelo2=lm(escala~tamanyo+temperatura)

> modelo3=lm(escala~tamanyo+harina)

> AIC(modelo2)

[1] 665.8821

> AIC(modelo3)

[1] 746.2985

Pregunta 2. Se le da el archivo **colegAlajuela2.Rdata** en formato R que contiene un muestra de los colegios de Alajuela en el 2011 (son datos reales provenientes del Estado de la Educación, del Proyecto Estado de la Nación). Se tiene la hipótesis de que en algunos colegios la aprobación ("pasar el año") entre sétimo y décimo es un mecanismo de "coladero", por lo que los que llegan a hacer examen de Bachillerato son solo los "buenos estudiantes". Como variable dependientes se tiene el porcentaje de estudiantes matriculados en undécimo año que aprobaron bachillerato. En algunos casos este porcentaje es mayor a 100% pues hay estudiantes que se pasaron de colegio durante undécimo año y aplicaron Bachillerato. Interesa entonces analizar si los porcentajes de aprobación de cada año de la cohorte de estudiantes (o sea, sétimo año en 2007, octavo año en 2008, y así sucesivamente) predicen la aprobación del examen de Bachillerato. Además, se desea analizar si los años de existencia del colegio, el porcentaje de docentes titulados, la razón estudiantes/docentes, y si el colegio es privado o no también predicen la variable dependiente. Las variables se describen abajo. Con base en esta información, conteste las siguientes preguntas:

|  |  |
| --- | --- |
| **Variable** | **Descripción** |
| nombre | Nombre del colegio |
| porc\_bachiundec11 | Porcentaje de estudiantes matriculados en undécimo a principios de año que aprobaron bachillerato |
| p\_aprobt7\_07 | Porcentaje de aprobación total séptimo 2007 |
| p\_aprobt8\_08 | Porcentaje de aprobación total octavo 2008 |
| p\_aprobt9\_09 | Porcentaje de aprobación total noveno 2009 |
| anyosexiste | Años desde la fundación del colegio |
| provincia | 2=Alajuela, 6=Puntarenas |
| porcdocentetitulado | Porcentaje de docentes que están titulados (en puntos porcentuales) |
| razon\_estud\_docente | Razón del número de estudiantes entre el número de docentes. |
| privado | 1=Privado o semiprivado, 0=Público |

Conteste con R las siguientes preguntas.

1. Estime una ecuación de regresión en la que prediga el porcentaje de aprobación de bachillerato por las variables independientes descritas anteriormente. Presente el summary del modelo (4 ptos.)
2. Haga un gráfico de residuos estandarizados vs. leverage y, si hay valores extremos, identifique los casos con la variable nombre (Limítese como máximo a 6 casos más extremos ya sea usando leverage o residuos) (4 ptos.).
3. Busque el caso con el leverage más alto y presente los valores que tiene ese caso en las variables predictoras. Explique por qué este caso tiene el leverage más alto, comparando con el resto de los casos (4 ptos.)
4. Haga un gráfico de residuos vs. predichos y argumente si la forma de dicho gráfico sugiere o no homoscedasticidad (3 ptos.)
5. Analice el supuesto de homoscedasticidad con una prueba de Breusch-Pagan, con un =0.05, y diga si coincide este resultados con el del gráfico del punto anterior? (3 ptos.)
6. Analice el supuesto de multicolinealidad con los VIFs. ¿Qué concluye? (3 ptos)
7. Usando un procedimiento de selección "hacia atrás" de variables (stepwise backwards) basado en el criterio de información BIC, genere un modelo más parsimonioso que el estimado al principio de esta pregunta. (4 ptos.)
8. Presente un summary del modelo escogido en el punto anterior y discuta si la alta correlación entre variables independientes le estaba generando problemas de inferencia asociados a la multicolinealidad. (4 ptos.)