Taller 1 Regresión polinómica y por segmentos

Andrés Felipe Palomino - David Stiven Rojas
Códigos:1922297-1924615
Universidad del Valle

29 de marzo de 2023

Universidad del Valle

Ejercicio 1: Los datos ChemReact.csv contiene información de un experimento químico. Se tiene como objetivo determinar las condiciones de temperatura (x_1) y tiempo (x_2) que proporcionan un alto rendimiento (y). Para ello, se propone un modelo cuadrático completo:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}^2\beta_3 + x_{2i}^2\beta_4 + x_{1i}x_{2i}\beta_5 + \varepsilon_i$$

Donde ε_i $N(0, \sigma^2)$ y $cov(\varepsilon_i, \varepsilon_k) = 0$

A) Ajuste el modelo y verifique el cumplimiento de los supuestos del modelo. En caso de que no se cumplan los supuestos, realice transformaciones sobre la variable respuesta para corregirlo

	Time	Temp	Yield	Temp.c	Time.c
Sin Centrar	5888.79	3703.33	1253.70	5310.88	3028.54
Centrado	1.00	1.00	1.01	1.01	1.00

Tabla 1: VIF

Antes de evaluar los supuestos del modelo verificamos la multicolinealidad a través del indicador VIF, obtuvimos altos valores que evidenciamos en la Tabla 1, indicándonos problemas graves de multicolinealidad, por lo cual centramos el modelo, obteniendo una corrección de este problema y a continuación realizamos la validación de sus supuestos utilizando las tres metodologías vistas en clase dónde lo explicaremos a continuación.

Modelo/ Prueba	Shapiro Wilk(p-value)	Breusch Pagan(p-value)
Modelo Centrado	0.006628	0.558
Modelo Box-Cox	0.005993	0.6521
Modelo MCP	0.0002527	No Aplica

Tabla 2: Resultados de pruebas de hipótesis para la validación de supuestos

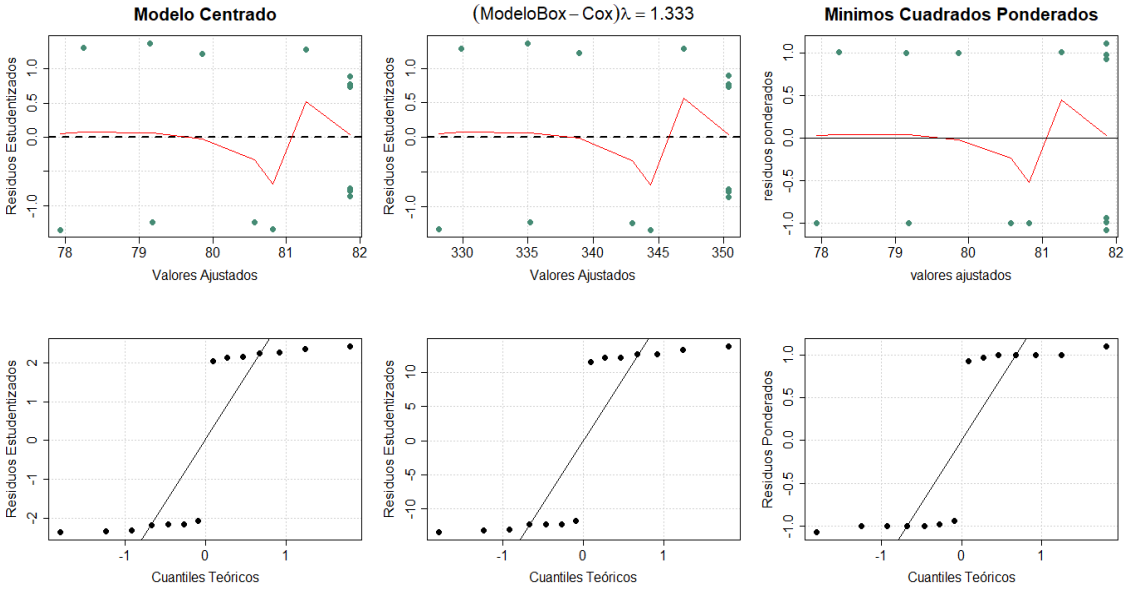

Figura 1: Validación de supuestos general para las diferentes transformaciones realizadas al modelo.

En la Figura 1 se presenta el gráfico del modelo centrado, modelo utilizando transformación de Box-Cox y la estimación por MCP, se observa que en los 3 modelos no se presentan problemas heterocedasticidad, sin embargo, con la ayuda del qqplot y la prueba de Shapiro-Wilk se afirma que con una significancia de 0.05 rechazamos la hipótesis nula para los 3 modelos, por lo cual no se cumple el supuesto de normalidad de los errores. Estas afirmaciones podemos evidenciarlas en la Tabla 2. Dado el incumplimiento del supuesto de normalidad de los errores en todos los modelos, más que no hubo una manera de arreglarlo con las transformaciones conocidas, se procede a trabajar con el modelo centrado.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	81.8662	1.2053	67.92	0.0000
Temp.c	0.1155	0.2088	0.55	0.5951
Time.c	0.1865	0.2088	0.89	0.3978
I(Time.c^2)	-0.0523	0.0435	-1.20	0.2631
I(Temp.c^2)	-0.0373	0.0435	-0.86	0.4155
Temp.c:Time.c	0.0050	0.0590	0.08	0.9346

Tabla 3: Resumen del modelo centrado.

También contamos un valor-p para el estadístico F de 0.683, por lo cual concluimos que no se rechaza la hipótesis nula y **todos** los coeficientes del modelo son significativamente iguales a 0. Aun así, por motivos de pedagogía procederemos a realizar los demás puntos a tratar, aun sabiendo que no se encuentra ninguna significancia importante en el modelo, y en general es mejor estimar la variable de interés con su promedio aritmético.

B) ¿Cómo afectan las covariables a la variable respuesta? (esto lo puede hacer gráficamente y/o interpretando los coeficientes)

La interpretación de los coeficientes es complicada debido a que si x_1 cambia en α unidades, al mantener x_2 constante, el valor esperado cambia en $(\beta_1\alpha + \beta_3\alpha^2) + 2\beta_3\alpha x_1 + \beta_5\alpha x_2$, por ende se realiza un gráfico 3D que visualice el valor esperado de "y" en el modelo ajustado centrado.

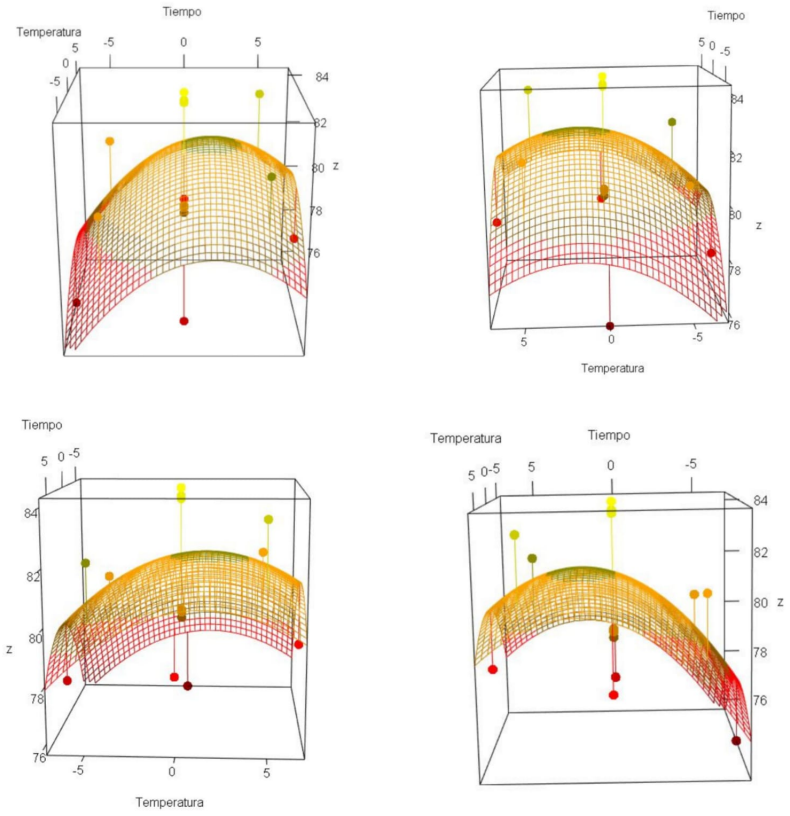

Figura 2: Curva generada por el modelo polinómico.

En la Figura 2 se observa como cambia el rendimiento medio, a medida que interactúan las demás covariables, a manera de ejemplo, si nos mantenemos a una temperatura de 0 (175), a medida que avanza el tiempo, el rendimiento incrementa hasta que llega a un rendimiento máximo, y luego este empieza a disminuir, teniendo en cuenta la no extrapolación en el proceso. Cabe aclarar que la curva generada cuenta con un vértice, en este se encuentra el rendimiento máximo.

C) Evalúe si el efecto interacción es significativo. ¿Qué implicaciones tiene esto en la relación de las covariables con la variable respuesta? Para evaluar si el efecto de interacción es significativo se procede a realizar la prueba ANOVA, el primero modelo con la interacción y el segundo modelo sin la interacción.

Tabla 4: ANOVA para comparación entre modelo completo y modelo reducido (no interacción)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	8	69.73				
2	9	69.79	-1	-0.06	0.01	0.9346

En la Tabla 4 se evidencia que al realizar la prueba de hipótesis $H_0 : \beta_5 = 0$ vs $H_1 : B_5 \neq 0$ se obtiene un valor P de 0.93, por lo tanto, el efecto de la interacción no es significativo. Esto implica que trabajaremos con el modelo sin la interacción.
D) Encuentre la combinación de temperatura y tiempo que maximiza el rendimiento. Calcule e interprete un intervalo de predicción del rendimiento en este punto.
Para realizar este punto debemos derivar parcialmente respecto a cada covariable e igualar a cero.

$$\frac{\partial E(Y)}{\partial} = \frac{\partial}{\partial x}(\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}^2\beta_3 + x_{2i}^2\beta_4)$$

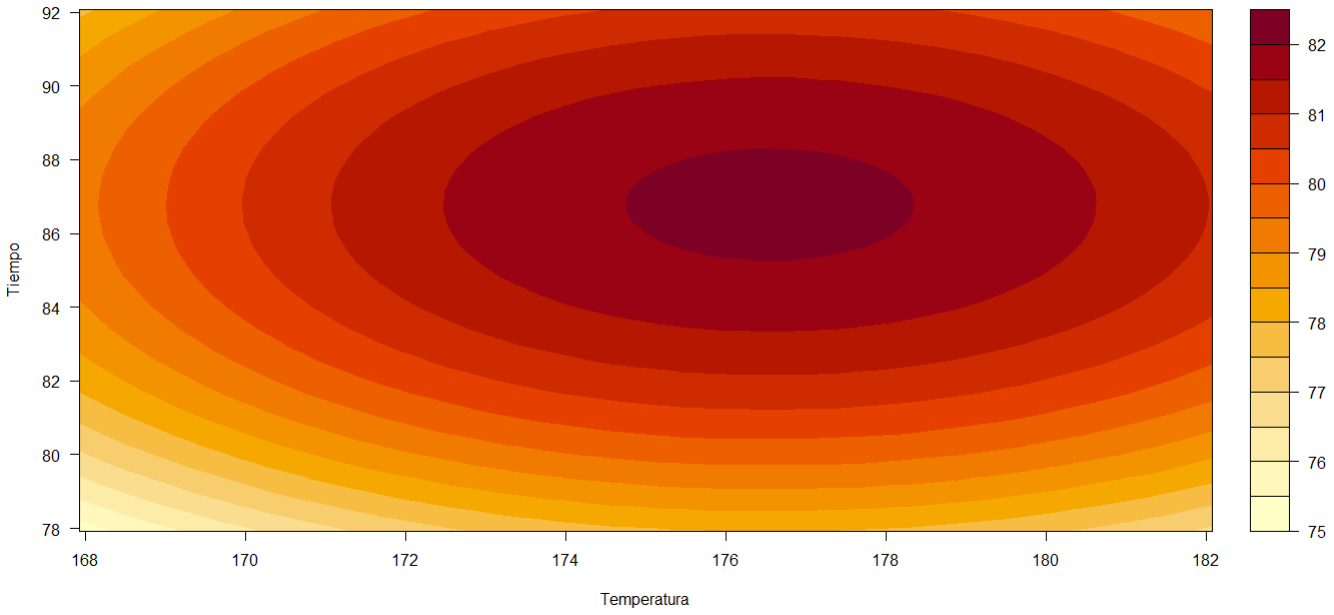

Figura 3: Gráfico de contornos.

Donde, al reemplazar los valores obtenidos en el modelo y derivar parcialmente en el software matemático Wolfram Alpha encontramos que el vector que maximiza esta combinación después de haberle sumado el promedio de cada covariable, dado que estamos considerando el modelo centrado es (176.54, 86.78), encontramos que el valor máximo cuando la temperatura y tiempo dados en el vector anterior sean esos, se espera que en promedio máximo del rendimiento en el experimento químico sea 82.12 con un intervalo de confianza del 95 % de [79.59,84.65]. Valor que podemos evidenciar en la Figura 2 o en el siguiente gráfico de contorno. Aun así, como mencionamos inicialmente, este máximo no tiene una significancia estadística.

Ejercicio 2: Considere los datos oldfaith de la libreria alr4 sobre las erupciones del gieiser Old Faithful durante octubre de 1980. Las variables observadas son: la duración en segundos de la erupción actual (x), y el intervalo de tiempo en minutos hasta la próxima erupción (y). La relación entre las variables es aproximadamente lineal, sin embargo, se cree que hay un cambio de pendiente. Por esta razón se propone el siguiente modelo:

$$y_i = \beta_0 + x_i\beta_1 + (x_i - 180)\beta_2 + \varepsilon_i$$

Donde:

$$(x_i - 180) = \begin{cases} 0 & \text{si } x_i \leq t \\ (x_i - 180) & \text{si } x_i > t \end{cases}$$

A) Ajuste el modelo e interprete los coeficientes.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.3582	2.9595	8.57	0.0000
Duration	0.2455	0.0223	11.02	0.0000
Duration.x	-0.1149	0.0362	-3.17	0.0017

Tabla 5: Coeficientes del modelo por segmenetos.

También contamos con un R^2 de 0.8101 que a comparación del modelo sin el cambio de pendiente tiene un leve aumento, puesto que este era de 0.8029. Dado los valores de la Tabla 5 podemos concluir que por cada aumento en un segundo que dure la erupción actual el tiempo medio en minutos hasta la próxima erupción aumenta en 0.1769 si se esta en valores por debajo a los 3, también que cuando sobre pase estos 180 segundos el valor de aumento seguirá siendo positivo, pero tendrá un cambio de -0.1149, quedando de 0.1306, lo cual nos indica que el aumento en la variable dependiente será menor por encima de este punto de corte.

B) ¿El cambio de pendiente es significativo?
Para esto realizaremos un ANOVA con el modelo sin el cambio y otro con este. Así obtenemos la siguiente tabla.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	268	9659.42				
2	267	9308.97	1	350.45	10.05	0.0017

Tabla 6: ANOVA entre los modelos para considerar el cambio de pendiente.

En la Tabla 6 en el valor-p del estadístico podemos observar que el cambio en la pendiente tiene un aporte significativo dentro del modelo y es diferente de cero, por lo cual no debemos omitirlo.

C) Compare el modelo anterior con un modelo lineal simple. ¿Cuál proporciona mejor ajuste?

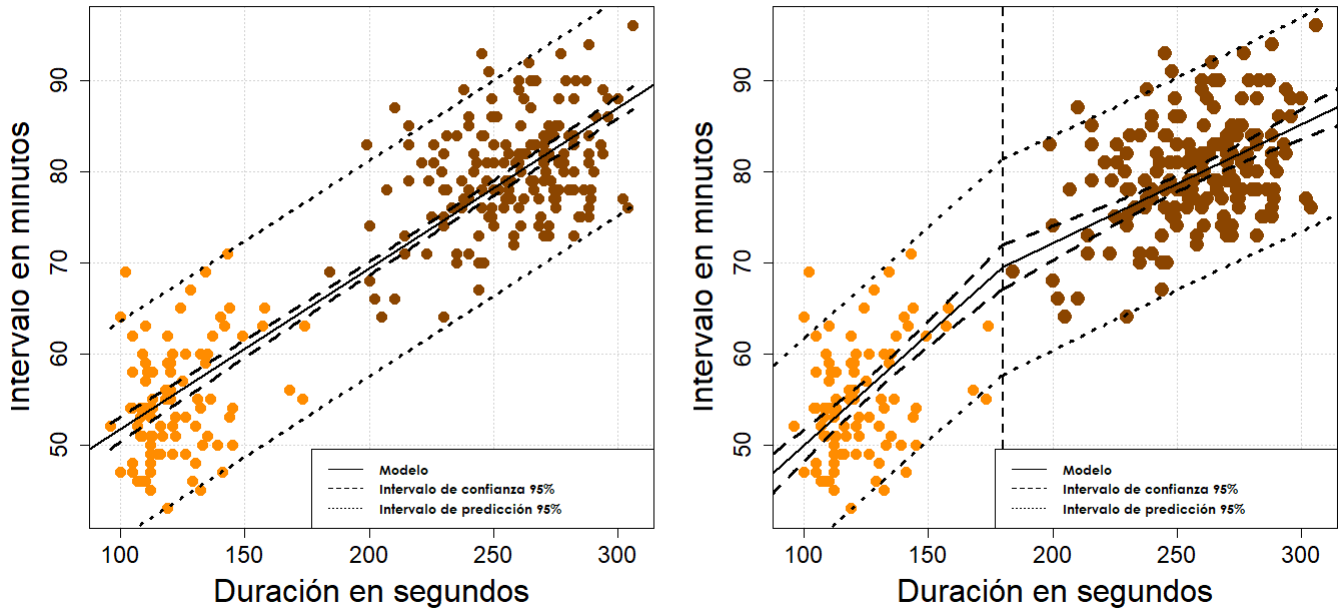

Figura 4: Modelo lineal simple y por segmentos.

Al comparar los R^2 del modelo encontramos que el modelo por segmentos presenta un valor mayor como mencionamos anteriormente, junto con que la prueba ANOVA nos dice que el aporte de este cambio de pendiente es significativo, adicionalmente a esto realizamos el cálculo del BIC y encontramos que el modelo con menor valor de este criterio de información fue el modelo segmentado, con un valor de 1744. Por lo cual seleccionamos este modelo.