XS-2130 Modelos de Regresión Aplicados II Sem 2022 Grupo 2

Examen Parcial 1.

Se entrega el jueves 15 de setiembre a más tardar 11:50 am, al correo electrónico [Gilbert.brenes.camacho@gmail.com](mailto:Gilbert.brenes.camacho@gmail.com). Pueden enviar este mismo documento de Word con su nombre en el nombre del archivo o bien un documento de pdf generado con Markdown. En ambos documentos DEBE APARECER EL ENUNCIADO DE CADA PREGUNTA CON SU RESPECTIVO PUNTAJE. Al final del documento, debería venir copia del código usado para resolver el examen. Tiene 170 minutos para terminar el examen.

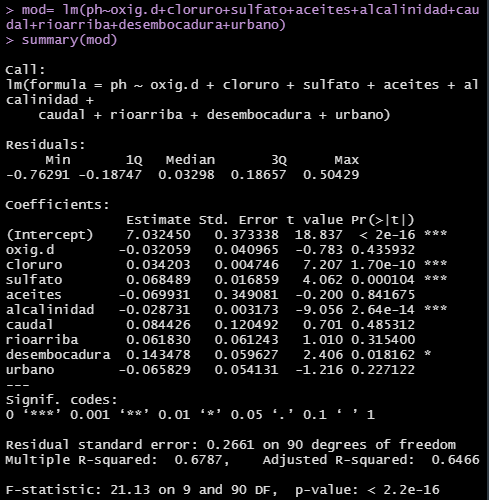
REPITO: EN LAS RESPUESTAS DEL EXAMEN TIENEN QUE APARECER NECESARIAMENTE LOS TEXTOS DE LAS PREGUNTAS (CON SU RESPECTIVO PUNTAJE) Y LOS RESULTADOS GENERADOS EN LA CONSOLA DE R-STUDIO.

Puntaje del examen: 50 ptos.

1. Preguntas de cálculos y teóricas a partir de procedimientos en cómputo
2. (Basado en García, Lorena, Parra y Pincay-Pilay, 2019). Un conjunto de ingenieros ambientales y estadísticos desean analizar los mejores predictores del PH (grado de acidez) en un río en Ecuador. Para ello, escogen 100 puntos al azar del río, y extraen muestras de agua, para analizar sus componentes. La base se encuentra en el archivo rios.Rdata. Abajo se presentan las variables de la base de datos:

|  |  |
| --- | --- |
| id | Número de identificación del punto de recolección |
| ph | Grado de acidez del agua, medido en “unidades logarítmicas” |
| oxig.d | Oxígeno disuelto, en mg por litro |
| cloruro | Cloruro, en mg por litro |
| sulfato | Sulfato, en mg por litro |
| aceites | Cantidad de aceites, en mg por litro |
| alcalinidad | Alcalinidad (cantidad de CaCO3), en mg por litro |
| caudal | Caudal del río, en metros cúbicos por segundo |
| rioarriba | 1=Recolección en río arriba, 0=Recolección en río abajo |
| desembocadora | 1=Recolección cerca de desembocadura, 0=Recolección lejos de desembocadura |
| urbano | 1=Recolección cerca de zona urbana, 0=Recolección en zona rural |
| distrito | Numeración de distritos en los que está ubicado el punto de recolección: Distrito 1, distrito 2, distrito 3 y distrito 4. |

1.A. Estime con R un modelo lineal en que se prediga el valor del pH en función del oxígeno disuelto, el cloruro, el sulfato, la cantidad de aceites, la alcalinidad, el caudal, si el punto de recolección está río arriba o río abajo, si el punto de recolección está cerca o no de la desembocadura y si el punto de recolección está cerca de zona urbana o cerca de zona rural. En otras palabras, usar todas las variables independientes **excepto distrito y, evidentemente, excepto id**. Con base en las salidas de R, conteste las siguientes preguntas:



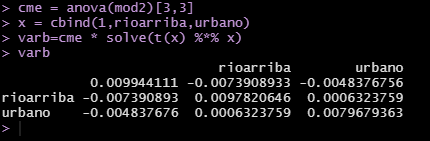
1. Independientemente del valor del pvalue (probabilidad asociada a la prueba), interprete el coeficiente de regresión para la variable *cloruro*. (3 ptos)

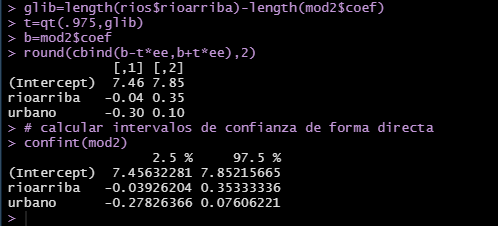
**Por cada aumento de un mg por litro de la variable cloruro, el ph aumenta en promedio 0,0342 unidades logarítmicas, manteniendo constantes las otras variables**

1. Independientemente del valor del pvalue (probabilidad asociada a la prueba), interprete el coeficiente de regresión para la variable *rioarriba*. (3 ptos)

**Cuando la variable rioarriba es igual a uno el ph aumentan en promedio 0,0618 mantenidndo las demás variables constantes**

1. Debido a que no hay interacciones, la diferencia en el PH entre puntos de recolección urbanos río arriba por un lado y en puntos de recolección rurales río abajo por el otro lado sería la suma de los coeficientes de las variables urbano y rioarriba. Con un procedimiento paso a paso (o sea, “a pie” calculando cada componente), calcule un intervalo de confianza al 95% para la suma de las estimaciones de esos dos coeficientes (Nota: Necesita usar la matriz de variancia-covariancia de los betas). (6 ptos.)





Los intervalos para rioarriba=[-0.04, 0,35]

Los intervalos para urbano= [-0.30, 0.10]

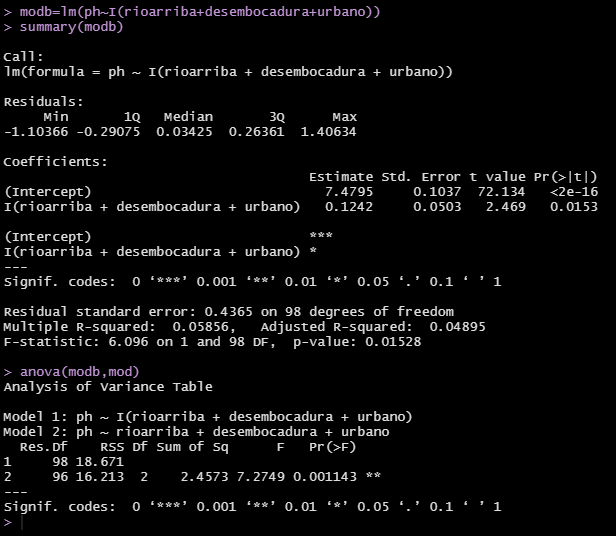
1. Al 5% de significancia, pruebe la hipótesis nula de que los coeficientes poblacionales de las variables rioarriba, desembocadura y urbano son iguales entre sí (6 ptos).

**H0: b1=b2=b3**

**H1: algún bi es diferente**

**Como Pr(>F)< alfa se rechaza**

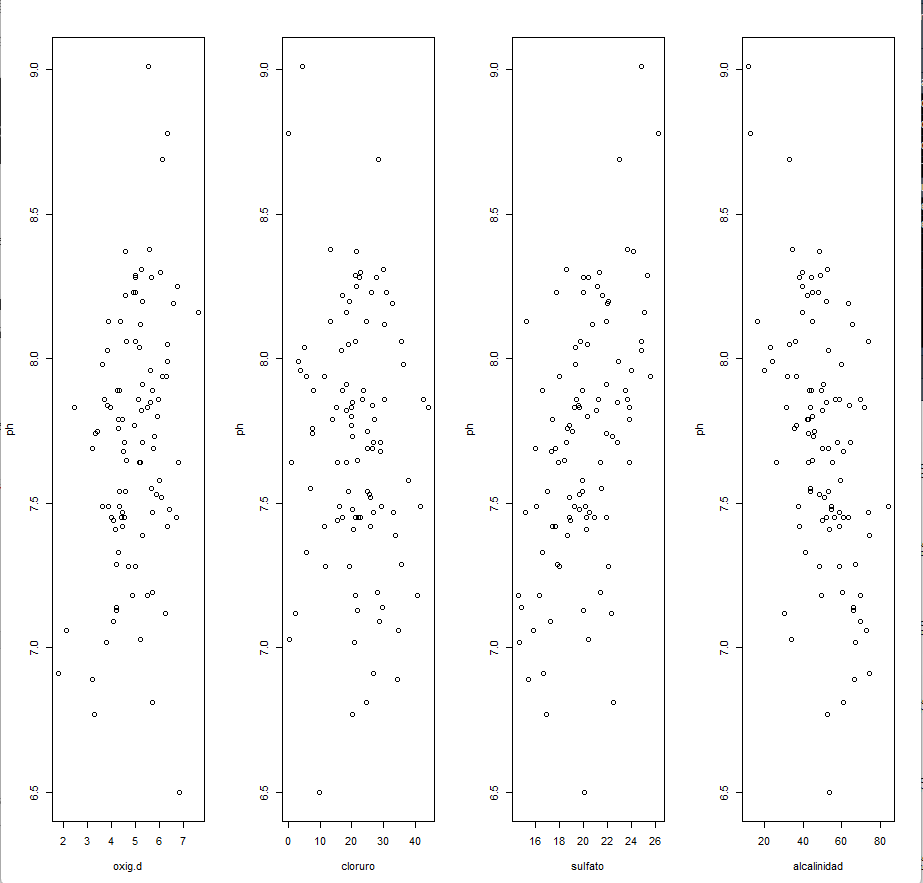
**Hay suficiente evidencia estadística para rechazar la hipótesis nula de que b1=b2=b3 con una significancia de 5%**

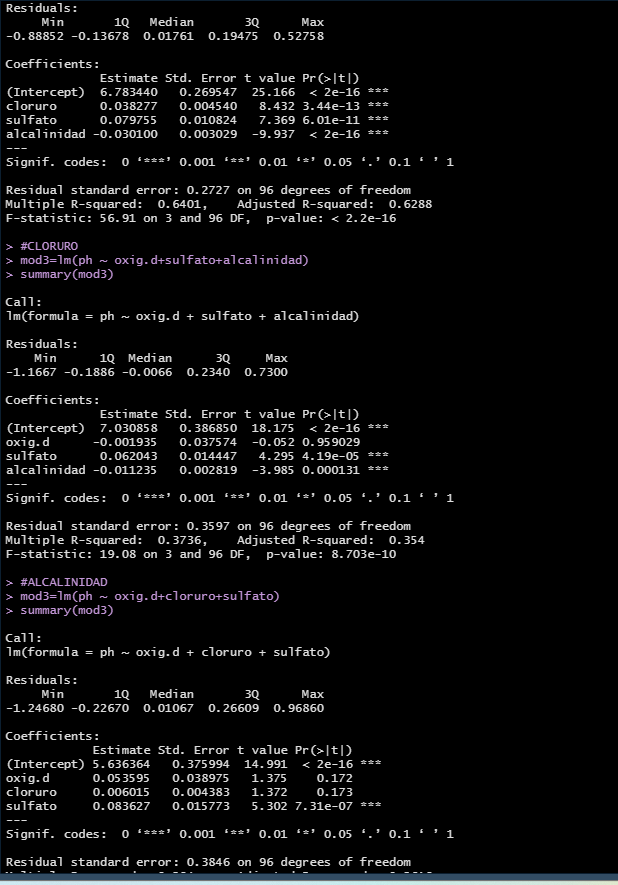


1.B. Estime con R un modelo lineal en que se prediga el valor del pH en función únicamente del oxígeno disuelto, el cloruro, el sulfato y la alcalinidad. Con base en este modelo, conteste las siguientes preguntas:

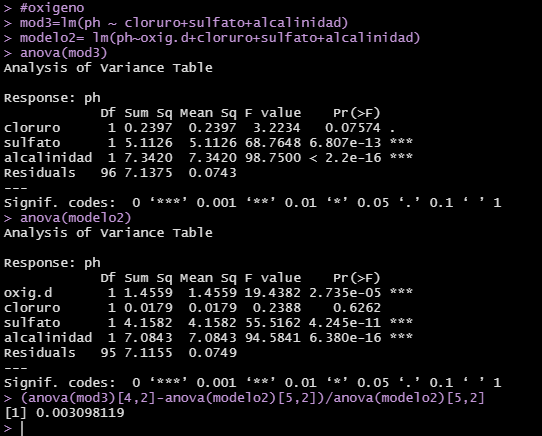
1. Cree una matriz de 4 gráficos de dispersión, con la variable ph en el eje de las ordenadas, y cada una de las 4 variables cuantitativas (oxigeno.d, cloruro, sulfato y alcalinidad), y señale cuáles son las dos variables predictoras con las que está más clara y fuerte la asociación lineal con el ph (6 ptos).

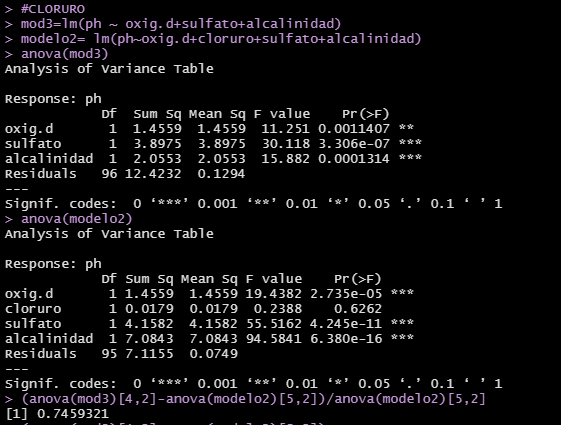
**Las variables más claras para predecir el ph son sulfato y alcalinidad.**

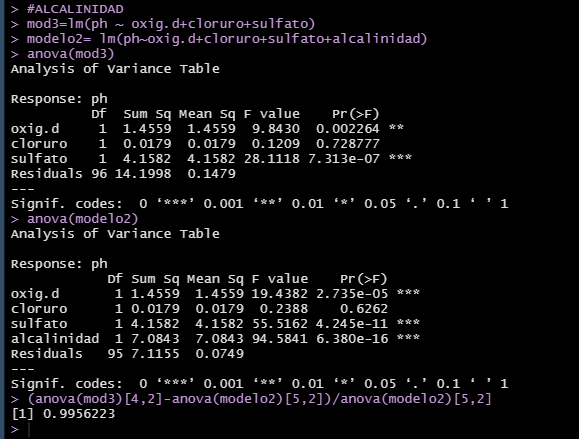


1. Calcule el coeficiente de determinación parcial de cada predictor condicional a la incorporación de los otros 3 predictores, y concluya cuál es el predictor que hace reducir menos la Suma de Cuadrados del Error del modelo más simple (6 ptos.) 

El predictor que hace reducir menos la suma de cuadrados del error es el oxigeno.







1.C. Estime con R un modelo lineal en que se prediga el valor del pH en función únicamente del caudal, el sulfato, el distrito y la interacción entre el sulfato y distrito. Con base en este modelo, conteste las siguientes preguntas:

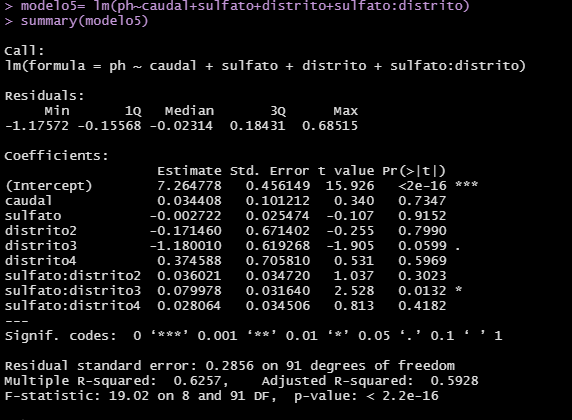
1. Haga un summary del modelo con interacciones que se le pide en el encabezado de la pregunta 1.C, y con base en él, escriba las ecuaciones implícitas para cada distrito (independientemente del pvalue). (6 ptos.)

7,26+0,03\*caudal-0,003\*sulfato-0,17\*distrito2-1,18\* distrito3+0,37\*distrito4+0,04\*distr2\*sulfato+0,08\*distr3\*sulf+0,03\*distr4\*sulf

Distrito2=7,13+0,03\*caudal-0,037\*sulfato

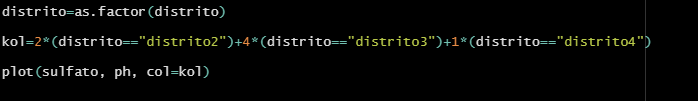
Distrito3=6,08+0,03\*caudal-0,077\*sulfato

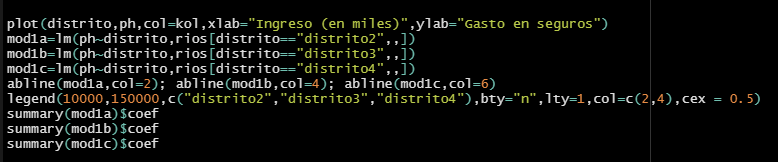
Distrito4= 7,64+0,03\*caudal-0,027\*sulfato

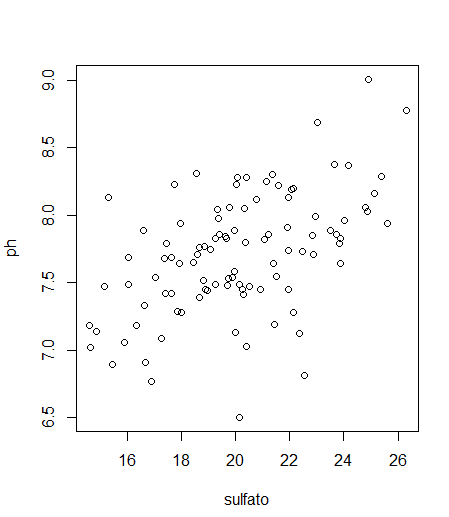


1. Fijando el valor del caudal en 1 metro cúbico por segundo, diseñe un gráfico de dispersión entre sulfato y ph, poniéndole distintos colores para los pares ordenados correspondientes a cada distrito, y superponga una recta de regresión condicional a cada distrito con el color correspondiente sobre cada nube de puntos. Cada una de las rectas de regresión se estima dentro de cada submuestra definida por cada distrito. Por esta razón, puede usar las ecuaciones estimadas en el punto anterior (Nota: Sería un gráfico parecido al hecho con el gasto en seguros vs el ingreso, para cada nivel de educación). Interprete si Ud cree que puede suponer paralelismo, justificando su interpretación. (4 ptos.)

**Con las ecuaciones calculadas anteriormente podría asumir cierto paralelismo en los distritos dos y cuatro porque su pendiente es muy similar.**







No tengo ni idea de lo que estaba haciendo :´(

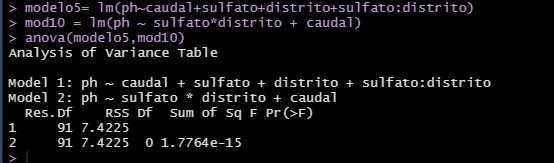
1. Al 5% de significancia, pruebe si existe interacción entre sulfato y distrito en la predicción del ph, incluyendo al caudal en el modelo (4 ptos.)

Ho: beta2=0 y beta3=0

H1= beta2<>0 y beta3<>0

Como Pr(>F) es menor a alfa se rechaza

Hay suficiente evidencia estadística para rechazar ho de que existe la interacción entre el sulfato y el modelo con una significancia de 5%



1. Preguntas teóricas.
2. Explique brevemente cuál es la diferencia entre normalidad de la variable dependiente y el supuesto del modelo lineal denominado normalidad condicional de la variable dependiente. (3 ptos).

La normalidad cuando está dada en función de otra variable puede hacer que cambie un poco el comportamiento y distribución de la variable dependiente ya que la está condicionado dado una o varias características.

1. Defina qué es una verosimilitud y qué significa que la log-verosimilitud de un modelo estimado A sea más cercana a cero que una log-verosimilitud de otro modelo estimado B (3 ptos.)

La verosimilitud es la maximización de la función de distribución. Y que sea más cercano significa que el modelo puede explicar mejor la realidad por medio de las variables.

Sinceramente no tengo ni idea jajaj.