Tarea 1B. Introducción a los modelos lineales generalizados

Gonzalo Pérez, César Valle y Rodrígo Jiménez

Semestre 2024-2

La solución del examen se deberá subir al classroom antes de las 11:59 PM del 24 de marzo de 2024. Todas las preguntas tienen un valor de 2 puntos. Favor de argumentar con detalle las respuestas.

NOTA 1. En caso de que se identifiquen respuestas iguales en otros examenes o alguna mala práctica, se procederá a la **anulación** de los examenes involucrados.

NOTA 2. Usar una confianza de 95 % o una significancia de .05 en los casos en donde no se requiera otro nivel de forma explícita.

NOTA 3, sobre el formato de entrega. La solución de cada pregunta se debe incluir en un **REPORTE EJECUTIVO** (pdf) que será la base de la evaluación, además se debe incluir otro archivo donde se pueda **REPLICAR TODO** resultado que se presente en el reporte ejecutivo, así como lo correspondiente al procesamiento y modelado que se realizó (R, Rmd, etc.). El reporte ejecutivo **no debe ser mayor a 4 PÁGINAS por pregunta** y deberá incluir la explicación, sustento y descripción del modelo que se usa, así como la justificación, descripción de hipótesis que se contrastan y conclusión de las pruebas de hipótesis relevantes. Este reporte deberá incluir AL MENOS una figura y una sección de resultados numéricos (puede ser una tabla); pero toda figura, tabla, resultado o extracto de código que se incluya DEBE estar descrito (explicado) y referido en el texto, de otra forma aunque se presente en el documento o se pueda generar con los scripts no se tomará en cuenta como parte de la solución. Por otra parte, los scripts deben estar comentados, al menos de grosso modo, es decir, indicando el objetivo de conjuntos de líneas de código (por ejemplo, por secciones: # búsqueda-selección de un modelo; # verificación de supuestos; # interpretación y pruebas de hipótesis, etc.).

NOTA 4. Cada pregunta contiene un conjunto de incisos que se deben tomar como guía para responder la pregunta principal del problema práctico que se presenta. Esto corresponde a lo **mínimo** que se podría hacer para obtener una solución, pero ustedes pueden explorar otras formas de modelado, así como de visualización de los datos y resultados.

1. Regresión lineal múltiple

La base de datos *Preg1B.csv* contiene información sobre 295 pacientes seleccionados de forma aleatoria. Se desea analizar si existe una asociación entre la presión arterial sistólica (bpsystol) y el índice de masa corporal (bmi), en particular, si es posible observar que tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica. Para realizar este análisis se indica que se considere el sexo (sex: 1-hombre, 2-mujer) y la edad (age) de los pacientes, pues la presión arterial sistólica podría variar de acuerdo con estos factores.

i. Ajuste el modelo de regresión lineal múltiple para $\mathrm{E}(\text{bpsystol}; \text{bmi}, \text{sex}, \text{age})$, donde las variables bmi y age entran sin modificación alguna y sin interacciones entre ellas ni con la variable sex. Indique si este modelo parece adecuado para realizar el análisis solicitado. En caso afirmativo, pase a inciso iii).

ii. En caso de que considere que el modelo en i) no es adecuado, presente un modelo que le parezca adecuado, donde de ser necesario se transformen las variables bpsystol, bmi o age. Por simplicidad no considere en el modelo interacciones entre las variables. Tanto en i) como en este inciso, revisar la linealidad de forma global y también analizando por cada covariable continua.

iii. ¿Se puede indicar que para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica? Argumente su respuesta, indicando con claridad la prueba o pruebas de hipótesis usadas y las hipótesis que se están contrastando.

iv. Para complementar la interpretación de los resultados del inciso iii), presente una gráfica resumen con la estimación puntual asociada a la relación entre bpsystol y bmi. Para esto considere sólo tres posibles edades: 30, 50 y 64, así como la diferenciación entre mujeres y hombres. Comente e interprete lo que observa en la gráfica, indicando con claridad a qué parámetro corresponde la curva/recta.

2. Modelos lineales generalizados para datos continuos

Considere los mismos datos que en la pregunta 1.

i. Explorando los diferentes modelos lineales generalizados comúnmente usados cuando la variable dependiente es continua (normal, gamma, inversa gaussiana), presente un modelo que le parezca adecuado para modelar $\mathrm{E}(\text{bpsystol}; \text{bmi}, \text{sex}, \text{age})$. Considere por simplicidad que no hay interacción entre las covariables del modelo. Deberá indicar con claridad cuál es la expresión matemática que se usa para modelar $\mathrm{E}(\text{bpsystol}; \text{bmi}, \text{sex}, \text{age})$, así como describir el procedimiento y criterio usado para seleccionar el modelo.

ii. Repita los incisos iii) y iv) de la pregunta 1 con el modelo en i).

iii. Comparando el modelo en i) con el usado en la pregunta 1, compare las conclusiones e interpretaciones que se pueden obtener e indique qué modelo prefiere usar. Argumente con claridad su respuesta, por ejemplo, debe incluir los valores de AIC o BIC, así como ventajas y desventajas en la interpretación.

3. Modelos lineales generalizados para datos binarios

La base de datos *Preg3B.csv* contiene información sobre 862 insectos que fueron expuestos a diferentes dosis (Deposit, mg) de tres insecticidas (Insecticide). La asignación a una dosis y al tipo de insecticida se realizó de forma aleatoria. Después de seis días se analizó si los insectos se habían muerto, de manera que la base de datos contiene también el número de insectos muertos (Killed) y el número total de insectos expuestos (Number) por cada dosis e insecticida. Dado que se asume que el costo de los insecticidas es el mismo, el objetivo del análisis es identificar para cada insecticida qué dosis es la mínima con la que se puede indicar que el 70 % de los insectos se muere, así como si considerando la menor de esas tres dosis se puede afirmar que un insecticida es el mejor comparado con el resto.

Notar que aquí el evento de interés es si el insecto muere o no. Además, dado que se tienen varios insectos para diferentes valores de dosis e insecticida, es posible realizar gráficas que ayudan a entender lo que se está modelando; de hecho la base de datos está en un formato agregado.

i. Presente una gráfica de dispersión en donde en el eje x se incluye la dosis del insecticida y en el eje y la proporción de insectos muertos observados para cada combinación dosis-insecticida, distinguiendo con un color el insecticida asociado. Describa lo que se observa.

ii. Ajuste modelos para datos binarios (ligas: logit, probit, cloglog) en donde incluya como covariables a Insecticide y lnD ($\text{lnD} = \ln(\text{Deposit})$), así como su interacción. Describa las expresiones del componente lineal o sistemático para cada insecticida como función de lnD. Indique si alguno de los modelos parece adecuado para realizar el análisis deseado.

iii. Ajuste modelos para datos binarios (ligas: logit, probit, cloglog) en donde adicional a las covariables incluidas en ii), también incluya a la interacción de Insecticide con lnD^2. Describa las expresiones del componente lineal o sistemático para cada insecticida como función de lnD. Indique si alguno de los modelos parece adecuado para realizar el análisis deseado y si tiene alguna ventaja la inclusión de los términos cuadráticos en el modelo.

iv. Sólo con el modelo que considere más adecuado entre los que se ajustaron en ii) y iii)

   a. presente en la misma gráfica generada en i) los resultados de la estimación puntual para el valor esperado de la variable binaria (probabilidad de que un insecto muera).
   b. calcule la dosis mínima para cada insecticida con la que se puede indicar que el 70 % de los insectos se muere.
   c. considerando la menor de las dosis encontradas en b), ¿se puede indicar que un insecticida es el mejor? Realice una prueba de hipótesis para argumentar en favor o en contra.
   d. En general ¿se puede indicar que los insecticidas A y B tienen un desempeño similar? Realice una prueba de hipótesis para argumentar en favor o en contra.

4. Modelos lineales generalizados para datos de conteos

La base de datos *Preg4.csv* contiene información sobre el número de casos de cáncer de pulmón (Cases) registrados entre 1968 y 1971 en cuatro cuidades de Dinamarca (City). En estos casos se registró también la edad de los pacientes (Age, variable categorizada en 5 grupos). El interés del análisis es estudiar si se puede indicar que a mayor edad existe mayor incidencia de cáncer de pulmón.

Notemos que para realizar el análisis la variable de conteos Cases depende de forma inherente de la población de la ciudad (Pop), pues entre más grande la ciudad es mayor el número de casos que se pueden observar; de manera que el estudio se debe enfocar en las tasas de incidencia.

i. Presente una gráfica de dispersión en donde en el eje x se incluyan los grupos de edad (ordenados de menor edad a mayor) y en el eje y la tasa de incidencia (Cases/Pop) por cada cruce Age-City, distinguiendo con un color la Ciudad. Describa lo que se observa.

ii. Como un primer modelo considere la distribución Poisson con liga logarítmica y las covariables Age y City, así como su interacción. Dado que las dos covariables son categóricas, este modelo con interacciones tiene muchos parámetros y es deseable trabajar con uno más simple. Para esto considere un segundo modelo donde sólo se usa como covariable a Age. Realice una prueba de hipótesis para argumentar si es posible considerar el segundo modelo [recuerde que dado que los modelos son anidados, podría usar la función $anova(mod1, mod2, test = "Chisq")$, también puede usar multcomp, pero hay muchos parámetros y podría ser tedioso]. Complemente su decisión con lo que se observa en la gráfica en i) y con medidas como AIC o BIC.

iii. Considerando el modelo seleccionado en ii), ajuste un modelo binomial negativo. Compare ambos modelos e indique cuál podría ser adecuado para realizar el análisis deseado. Con el modelo seleccionado, calcule intervalos de confianza simultáneos de las tasas de incidencia para cada grupo de edad, incluya estos en la gráfica presentada en i). Comente los resultados, en particular si se puede indicar que a mayor edad existe mayor incidencia de cáncer de pulmón.

iv. **Opcional (1 punto)** Los incisos anteriores usaron a la variable Age como categórica, sin embargo, eso dificulta un poco la interpretación, además de que por su naturaleza esa variable se podría haber registrado sin categorizar. Con los datos actuales, una aproximación sería usar el punto medio de cada intervalo de edad que define las categorías de Age y usar la variable resultante como una variable continua, llámela Ageprima. Ajuste modelos usando la distribución Poisson y Binomial Negativa con la covariable Ageprima, también considere la opción de incluir a Ageprima^2. Entre esos 4 modelos indique cuál podría ser adecuado para realizar el análisis. Con ese modelo indique si a mayor edad existe mayor incidencia de cáncer de pulmón, por ejemplo, analizando si la función es creciente considerando que el intervalo de edad que es de interés es entre 40 y 74 años. Presente una gráfica que complemente su análisis.

5. Modelos lineales generalizados para datos categóricos

La base de datos *Preg5.csv* contiene información sobre el nivel de satisfacción (Sat) de un conjunto de individuos que rentan una vivienda. El interés es identificar si entre los factores que definen este nivel están: el tipo de vivienda (Type), la percepción sobre su influencia en las decisiones sobre el mantenimiento de la vivienda (Infl) y el contacto que tienen con el resto de inquilinos (Cont).

i. En este caso todas las covariables son categóricas, así que se puede obtener una gráfica que describa las frecuencias relativas para los tres niveles de satisfacción considerando cada cruce Type-Infl-Cont. Presente esta gráfica y comente lo que se observa.

ii. Ajuste un modelo logístico multinomial considerando todas las posibles interacciones entre Type, Infl y Cont. Este modelo tiene demasiados parámetros y es de interés buscar si es posible considerar un modelo más simple. Ajuste un modelo logístico multinomial que incluya a las tres covariables Type, Infl y Cont, pero sin considerar interacciones. Realice una prueba de hipótesis para argumentar si es plausible considerar el modelo más simple. Complemente esto con medidas como AIC o BIC.

   Nota computacional. Para los modelos ajustados en vglm no es posible usar directamente el paquete multcomp, pero está disponible una función anova() que permite comparar dos modelos, siempre que estos estén anidados, por ejemplo: $anova(mod1, mod2, test = "LRT", type = "I")$ o equivalentemente $lrtest(mod1, mod2)$.

iii. Considerando las covariables del modelo ajustado en ii) y notando que la variable Sat puede ser considerada como ordinal, ajuste un modelo logístico acumulativo (cumulative logit) sin considerar el supuesto de proporcionalidad. También ajuste un modelo que asuma el supuesto de proporcionalidad y dado que este último está anidado en el primero, realice una prueba de hipótesis para analizar si es plausible asumir este modelo más sencillo. Complemente su decisión con medidas como AIC o BIC.

iv. Usando el AIC o BIC, seleccione sólo un modelo entre los ajustados en ii) y iii). Con ese modelo, trate de interpretar los resultados. Por ejemplo, puede apoyarse de una gráfica como la realizada en i), pero donde se presenten las probabilidades estimadas para cada nivel de satisfacción. Por simplicidad sólo analice el efecto que se observa al considerar la variable Infl, cuando se asume que la persona renta una vivienda tipo Apartment y tiene un nivel de contacto con otros inquilinos como High.