Análisis de Regresión - 3006918, semestre 02 de 2022
Trabajo No. 01

TALLER 3

Briahan Serna

Al final se dejarán todos los codigos empleados para este trabajo.

El siguiente informe continúa con el análisis basado en el ajuste de un modelo de regresión lineal múltiple haciendo uso de datos recolectados para un estudio sobre la eficacia del control de infecciones nosocomiales (infecciones adquiridas durante la hospitalización), haciendo uso de la base de datos APC 1modifm 1.csv que corresponden a una muestra aleatoria de 80 hospitales seleccionada de los 338 hospitales originales investigados, se excluirán dos observaciones que en análisis previos se detectaron como muy influenciales. La base de datos contiene información sobre las siguientes variables:

Tabla 1: Caracterización de las variables de interés

# variable	Código	Nombre	Nombre de la variable en el modelo	Descripción
1	**ID**	Número de identificación del registro	**NA**	Numérica
2	**DPERM**	Longitud de permanencia	**Y**	Longitud promedio de permanencia de todos los pacientes en el hospital (en días).
3	**EDAD**	Edad	$\boldsymbol{X_1}$	Edad promedio de los pacientes (en años).
4	**RINF**	Riesgo de infección	$\boldsymbol{X_2}$	Probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).
5	**RRC**	Razón de rutina de cultivos	$\boldsymbol{X_3}$	Razón del número de cultivos desarrollados al número de pacientes sin signos o síntomas de infección adquirida en el hospital, por 100.
6	**RRX**	Razón de rutina de rayos X del pecho	$\boldsymbol{X_4}$	Razón del número de rayos X llevados a cabo al número de pacientes sin signos o síntomas de neumonía, por 100.
7	**NCAMAS**	Número de camas	$\boldsymbol{X_5}$	Número promedio de camas en el hospital durante el periodo de estudio.
8	**AEM**	Afiliación a escuela de medicina	$\boldsymbol{A_1}$	1=SÍ, 2=NO
9	**PDP**	Censo promedio diario	$\boldsymbol{X_6}$	Número promedio de pacientes en el hospital por día durante el periodo de estudio.
10	**NENFERM**	Número de enfermeras	$\boldsymbol{X_7}$	Número promedio de tiempos completos equivalentes registrados y enfermeras de práctica licenciadas durante el periodo de estudio (número de tiempos completos + 1/2 del número de tiempo parcial).
11	**FSD**	Facilidades y servicios disponibles	$\boldsymbol{X_8}$	Porcentaje de 35 facilidades potenciales y servicios que son proporcionados por el hospital.
12	**REGION**	Región	$\boldsymbol{R_i, i = 1, 2, 3}$	Región geográfica, donde 1=NE, 2=NC, 3=S, 4=W.

Se considerará a DPERM como variable respuesta y aquellas variables que tienen asignado un $X_j, con\ j = 1, \ldots, 8$ son las predictoras del modelo que se propondrá a continuación.

1. Definición y ajuste del MRLM.

Se ajustará un modelo de regresión lineal múltiple usando la base de datos sin considerar las observaciones con ID = 47, ID = 112. En la **Tabla 1** fueron definidas las variables que serán consideradas en el modelo siendo $\boldsymbol{Y}$ la variable respuesta y $X_j, j = 1,2\ldots,8$ las variables predictoras. El modelo que se define busca explicar la longitud promedio de permanencia en días de todos los pacientes en el hospital (DPERM) a través de un conjunto de variables relacionadas con características del hospital y procedimientos llevados en él.

La ecuación del modelo es: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + E_i, E_i \sim IID\ N(0, \sigma^2)$ (1)

Tabla 2: Parámetros estimados modelo

Parámetros	Estimación	Error Std	T_0	$P(\lvert t_{69}\rvert > \lvert T_0\rvert)$
β_0	2.055201	1.793834	1.146	0.2559
β_1	0.072344	0.031344	2.308	0.0240
β_2	0.367188	0.147992	2.481	0.0155
β_3	0.011391	0.016390	0.695	0.4894
β_4	0.014971	0.008183	1.829	0.0717
β_5	-0.003927	0.004340	-0.905	0.3687
β_6	0.011631	0.005664	2.054	0.0438
β_7	-0.004784	0.002886	-1.657	0.1020
β_8	0.008357	0.014846	0.563	0.5753
$\sqrt{MSE} =$ 1.2	$R^2 = 0.4473$, $R^2_{adj} = 0.3832$,			

De tal modo que la ecuación ajustada del modelo es:

$$\hat{Y}_i = 2.055201 + 0.072344X_{i1} + 0.367188X_{i2} + 0.011391X_{i3} + 0.014971X_{i4} - 0.003927X_{i5} + 0.011631X_{i6} - 0.004784X_{i7} + 0.008357X_{i8}$$

Test Anova del modelo

Para la prueba de significancia del modelo que ha sido ajustado, se plantean las siguientes hipótesis:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_8 = 0 \ \ vs. H_1: Algún\ \beta_j \neq 0\ con\ j = 1,2\ldots,8$$

El estadístico de prueba en este test está dado por $F_0 = \frac{SSR/8}{SSE/69} = \frac{MSR}{MSE}$, $F_0 \sim f_{8,69}$ Bajo H_0 y los supuestos del MRL

Tabla 3: Tabla ANOVA del MRLM.

Fuente	Sum Sq	Df	Mean Sq	F_0	$P(f_{8,69} > F_0)$
Modelo	80.398	8	10.0498	6.9801	1.052×10^{-6}
Error	99.345	69	1.4398		

El resultado para este caso $F_0 = \frac{80.398/8}{99.345/69} = 6.9801$. La hipótesis nula se rechaza para valores grandes de F_0 o usando $VP = P(f_{8,69} > 6.9801) = 1.052 \times 10^{-6}$, se rechaza la hipótesis nula pues el valor p toma un valor muy pequeño así que hay poca probabilidad de equivocarse al rechazar H_0. Lo anterior quiere decir que la regresión es significativa, es decir, al menos una de las ocho variables predictoras contribuye de forma significativa a explicar la longitud promedio de permanencia de los pacientes en el hospital. Lo anterior no implica buen ajuste.

Coeficiente de determinación muestral R^2

Para el modelo considerado en este caso, donde la variable respuesta corresponde a la longitud de permanencia, se obtuvo un coeficiente de determinación numéricamente igual a 0.4473, el cual se define como $R^2 = \frac{SSR}{SST}$. Este resultado corresponde a la porporción de la variabilidad total observada en la variable respuesta que es explicada por el modelo que ha sido ajustado previamente. De acuerdo con esto, el 44.73% de la variabilidad observada en la permanencia promedio de los pacientes en el hospital, es explicada por el modelo de regresión múltiple que incluye ocho variables regresoras. El valor tomado por el R^2 indica que poco más de la mitad de la variabilidad observada en la variable dependiente no está siendo representada por el modelo propuesto. Además, el R^2_{adj} es una medida que se espera que esté cercana a 1 cuando hay buen ajuste y para este caso donde el valor tomado fue de 0.3832 no hay un fuerte indicio para hablar de buen ajuste.

2. Diagnósticos de multicolinealidad

Matriz de correlación variables predictoras.

Tabla 4: Matriz de correlación de las variables predictoras

	EDAD	RINF	RRC	RRX	NCAMAS	PDP	NENFERM	FSD
EDAD	1.000000	-0.084054	-0.253506	-0.045818	-0.100342	-0.110876	-0.105664	-0.058849
RINF	-0.084054	1.000000	0.571985	0.446660	0.418851	0.429037	0.460512	0.428596
RRC	-0.253506	0.571985	1.000000	0.402608	0.106515	0.086504	0.139861	0.173928
RRX	-0.045818	0.446660	0.402608	1.000000	0.036630	0.031780	0.071851	0.113300
NCAMAS	-0.100342	0.418851	0.106515	0.036630	1.000000	0.981398	0.919845	0.783623
PDP	-0.110876	0.429037	0.086504	0.031780	0.981398	1.000000	0.921138	0.782661
NENFERM	-0.105664	0.460512	0.139861	0.071851	0.919845	0.921138	1.000000	0.758830
FSD	-0.058849	0.428596	0.173928	0.113300	0.783623	0.782661	0.758830	1.000000

La matriz anterior correspende a la matriz de correlaciones muestrales entre las variables predictoras incluidas en el modelo propuesto. Su análisis pretende detectar posibles problemas de multicolinealidad entre pares de variables. Tras examinar la matriz de halla lo siguiente:

Relaciones lineales débiles o nulas entre predictoras:

La variable Edad (X_1) presenta correlaciones débiles con todas las demás variables, es decir que a partir del análisis de la matriz de correlaciones esta variable no parece estar involucrada en algún caso de multicolinealidad por pares de variables. Esto no quiere decir que no esté involucrada en un caso de multicolinealidad donde intervengan más de dos variables.

La variable Razón de rutina de cultivos (X_3) presenta correlaciones débiles con las variables Número de camas (X_5), Censo promedio diario (X_6), Número de enfermeras (X_7) y Facilidades y servicios disponibles (X_8)

La correlación entre Razón de rutina de rayos x (X_4) con Número de camas, Censo promedio diario y Número de enfermeras en cada caso no supera a 0.08 y entre Razón de rutina de rayos x y Facilidades y servicios disponibles es del orden de 0.1133.

Relaciones lineales moderadas entre predictoras:

Entre Riesgo de infección (X_2) y todas las demás predictoras (X_j, j = 3,…,8) se hallan relaciones lineales por pares moderadas pues las correlaciones muestrales oscilan entre 0.4 y 0.572.

Entre Razón de rutina de cultivos (X_3) y Razón de rutina de rayos x (X_4) también se halla una relación lineal moderada pues corr=0.403

Relaciones lineales fuertes entre predictoras:

Las correlaciones entre Facilidades y servicios disponibles (X_8) con Número de camas, Censo promedio diario y Número de enfermeras oscilan entre 0.75 y 0.784. Es decir que se hallan relaciones lineales que pueden ser consideradas fuertes entre estos pares de variables. Finalmente, los casos más severos de relaciones lineales entre pares de predictoras son Número de camas (X_5) vs. Censo promedio diario (X_6), Número de camas (X_5) vs. Número de enfermeras (X_7) y Censo promedio diario (X_6) vs. Número de enfermeras (X_7) donde las correlaciones oscilan entre 0.91 y 0.982.

Esto último tiene mucho sentido en el contexto del problema pues las variables con correlaciones más altas son aquellas que apuntan a describir un mismo aspecto que es el tamaño y la infraestructura de los hospitales.

Tras examinar la matriz de correlación de las predictoras se detectan 6 pares de variables que posiblemente están envueltas en una multicolinealidad fuerte.

Factores de inflación de varianza.

Cuando hay multicolinealidad, las varianzas de los coeficientes de regresión estimados se inflan afectando, por ejemplo, procesos de estimación. Tal inflación se cuantifica mediante los VIFs que se definen de la siguiente manera: $VIF_j = \frac{1}{1-R_j^2}, j = 1,\ldots,8$ donde R_j^2 es el coeficiente de determinación muestral de la regresión ajustada con X_j como variable respuesta vs. las demás predictoras, esto puede verse como una forma de determinar si algunas de las predictoras son explicada en términos de las demás mediante una regresión linel. El criterio para detectar multicolinealidad usando los VIFs consiste en identificar aquellos $VIF_j > 10$, en tal caso se dice que hay un caso de multicolinealidad severa. Además, puede evaluarse si $5 < VIF_j \leq 10$, en tal caso se dice que hay un caso de multicolinealidad moderada. Mediante este criterio y haciendo uso de la **Tabla 5** se detectan dos problemas de multicolinealidad fuerte en los siguientes casos:

$$VIF_5 = \frac{1}{0.03383775} = 29.552793 \; y \; VIF_6 = \frac{1}{0.03247043} = 30.797248$$

Es decir que las variables X_5 y X_6 están siendo explicadas por las demás predictoras. Además, se tiene la relación $VIF_j = c_{jj}, j = 1,\ldots,8,$ con c_{jj} el j-ésimo elemento de la diagonal principal de la matriz $(X'X)^{-1}$, por lo tanto, las varianzas de $\hat{\beta}_5$ y $\hat{\beta}_6$ se están inflando producto de las dos relaciones de multicolinealidad severa halladas. También se detecta un caso de multicolinealidad moderada pues $VIF_7 = \frac{1}{0.13774258} = 7.259919 > 5$. La varianza de $\hat{\beta}_7$ se está inflando producto de tal multicolinealidad, pero no de forma tan severa como en los casos anteriores.

Tabla 5. Factores de inflación de varianza

Variables	$1 - R_j^2$	VIF
EDAD	0.89836191	1.113137
RINF	0.45000536	2.222196
RRC	0.55740170	1.794038
RRX	0.74540716	1.341549
NCAMAS	0.03383775	29.552793
PDP	0.03247043	30.797248
NENFERM	0.13774258	7.259919
FSD	0.36125550	2.768124

Tabla 6. Valores propios de $(X'X)$ índices de condición.

Valor propio	Índice de condición
7.895044753	1.000000
0.693814211	3.373306
0.256935067	5.543265
0.045908090	13.113925
0.035310976	14.952797
0.034107873	15.214230
0.029233007	16.433897
0.006475567	34.917118
0.003170454	49.901839

Análisis de los valores propios de $(X'X)$.

Número de condición.

Se define como $k = \frac{\lambda_{max}}{\lambda_{min}}$ y sirve para detectar el grado más alto de multicolinealidad hallada en el modelo. Si $\sqrt{k} > 31$ hay multicolinealidad grave. En este caso, según la **Tabla 6** $\sqrt{k} = \sqrt{\frac{7.895044753}{0.003170454}} = 49.901839 > 31$ entonces se dice que al menos dos de las variables predictoras se encuentran envueltas en multicolinealidad severa.

Índices de condición.

Los índices de condición k_j están dados por $k_j = \frac{\lambda_{max}}{\lambda_j}$ donde λ_j es el j-ésimo valor propio de la matriz $(X'X)$ con $j = 0,\ldots,8$

El criterio para determinar problemas de multicolinealidad usando los índices de condición es:

Si $\sqrt{k_j} < 10 \; \forall j$ no hay problemas serios de multicolinealidad.

Si $10 \leq \sqrt{k_j} < 31$ para al menos un j, la multicolinealidad es moderada.

Si $\sqrt{k_j} \geq 31$ para al menos un j la multicolinealidad es severa.

En este caso, según la **Tabla 6** se detectan cuatro asociaciones lineales moderas pues $10 \leq \sqrt{k_j} < 31 \; para \; j = 3,4,5,6.$

$$\sqrt{k_3} = \sqrt{\frac{7.895044753}{0.045908090}} = 13.113925, \sqrt{k_4} = \sqrt{\frac{7.895044753}{0.035310976}} = 14.952797, \sqrt{k_5} = \sqrt{\frac{7.895044753}{0.034107873}} = 15.214230, \sqrt{k_6} = \sqrt{\frac{7.895044753}{0.029233007}} = 16.433897$$

Estos índices de condición indican que hay 4 casos de multicolinealidad moderada entre dos o más variables predictoras.

Además, se detectan dos asociaciones lineales fuertes pues $\sqrt{k_j} \geq 31 \; para \; j = 7,8.$

$$\sqrt{k_7} = \sqrt{\frac{7.895044753}{0.006475567}} = 34.917118, \sqrt{k_8} = \sqrt{\frac{7.895044753}{0.003170454}} = 49.901839$$

Estos índices de condición indican que hay dos casos de multicolinealidad severa entre dos o más variables predictoras.

Proporciones de descomposición de varianza.

El diagnóstico de multicolinealidad mediante las proporciones de descomposición de varianza π_{ij} permiten identificar cuáles variables están envueltas en algún caso de multicolinealidad. Las medidas anteriores no eran tan específicas a la hora de identificar las variables partícipes en el problema de multicolinealidad. Esta medida corresponde a la proporción que el i-ésimo valor propio λ_i aporta a la varianza de cada coeficiente estimado $\hat{\beta}_j$. El diagnóstico mediante esta medida consiste en detectar aquellos $\pi_{ij} > 0.5$ y asociados a un mismo valor propio λ_i pequeño para al menos dos coeficientes de regresión. Si se encuentra con una situación de este tipo, se dice que las variables asociadas a tales parámetros están envueltas en un problema de multicolinealidad. Se hará este análisis con los datos centrados para eliminar problemas de muticolinealidad debidos al intercepto del modelo.

Tabla 7. Proporciones de descomposición de varianza con datos centrados

Indice val propio	EDAD	RINF	RRC	RRX	NCAMAS	PDP	NENFERM	FSD
1	0.001593	0.011728	0.003725	0.002309	0.001904	0.001828	0.007616	0.017022
2	0.020717	0.047418	0.114464	0.130224	0.000793	0.000798	0.001965	0.002349
3	0.767595	0.015258	0.006869	0.070095	0.000011	0.000021	0.000008	0.001364
4	0.103200	0.059939	0.276977	0.731685	0.000111	0.000177	0.000253	0.000270
5	0.036626	0.625714	0.356154	0.028634	0.000022	0.000506	0.002865	0.283057
6	0.057943	0.205730	0.200497	0.034652	0.009235	0.006369	0.041807	0.688826
7	0.000456	0.005368	0.002760	0.001338	0.060357	0.053576	0.945296	0.007090
8	0.011871	0.028845	0.038554	0.001063	0.927567	0.936726	0.000190	0.000021

Tras inspeccionar la tabla de proporciones de descomposición de varianza solo se detecta un caso de multicolinealidad entre las variables Número de camas $(\pi_{8,5})$ y Censo promedio diario $(\pi_{8,6})$ mediante este criterio de diagnóstico. Estas variables brindan información sobre la capacidad de los hospitales así que tiene mucho sentido que exista una asociación lineal fuerte entre ellas. Una alternativa sería construir un único indicador que reuna la información relacionada con la capacidad de los hospitales, pues en todos los análisis de multicolinealidad presentados estas son las variables con problemas más severos. Los π_{ij} señalados con azul superan la

cota de 0.5 pero es necesario hallar al menos dos que lo hagan asociados a un mismo valor propio, como no es el caso, no es posible sacar una conclusión de ellos por sí solos más que la proporción que el valor propio sociado a ellos aporta a la varianza del respectivo coeficiente estimado.

3. Método de todas las posibles regresiones

Haciendo uso de la función ols_step_all_possible() en R se construye a partir del modelo definido en la **ecuación 1**, todas las posibles regresiones. En este caso, resultan 255 modelos, en la siguiente tabla se presentan los mejores tres modelos con k variables predictoras siendo k=1,2,3,4,5,6, 7 y 8, incluyendo valores de interés de cada modelo como lo son los grados de libertad de la suma de cuadrados de residuales con su respectivo número de parámetros, el coeficiente de determinación R^2 y el $R^2{}_{adj}$. En la última columna se presenta C_p el cual es una medida de sesgo en los modelos de regresión, es decir, $E[\hat{Y}_i] - \mu_{Y|x_i}$.

Tabla 8. Los tres mejores modelos de la tabla de todas las posibles regresiones para cada k=1,..7.

Modelo	p	Dfe	Predictoras	R^2	$R^2{}_{adj}$	Cp	|Cp-p|
1	2	76	X2	0,3026699	0,2934945	13,055199	10,055199
2	2	76	X6	0,1734469	0,1625712	29,187490	26,18749
3	2	76	X8	0,1591077	0.1480433	30,977614	27,977614
9	3	75	X2 X6	0,3425706	0,3250392	10,073965	6,073965
10	3	75	X2 X8	0,3352534	0,3175268	10,987453	6,987453
11	3	75	X2 X5	0,3340993	0,3163419	11,131533	7,131533
37	4	74	X1 X2 X6	0,3797523	0,3546071	7,432179	2,432179
38	4	74	X2 X4 X6	0,3709499	0,3454478	8,531079	3,531079
39	4	74	X1 X2 X5	0,3697474	0,3441967	8,681193	3,681193
93	5	73	X1 X2 X4 X6	0,4098889	0,3775541	5,669902	0,330098
94	5	73	X1 X2 X6 X7	0,4043870	0,3717507	6,356760	0,35676
95	5	73	X1 X2 X4 X5	0,3971850	0,364154	7,255872	1,255872
163	6	72	X1 X2 X4 X6 X7	0,4358390	0,3966612	4,430273	2,569727
164	6	72	X1 X2 X4 X5 X6	0,4200079	0,3797307	6,406637	0,593363
165	6	72	X1 X2 X3 X4 X6	0,4120132	0,3711808	7,404702	0,404702
219	7	71	X1 X2 X4 X5 X6 X7	0,4403450	0,3930502	5,867739	2,132261
220	7	71	X1 X2 X3 X4 X6 X7	0,4388750	0,3914560	6,051253	1,948747
221	7	71	X1 X2 X4 X6 X7 X8	0,4381606	0,3906813	6,140439	1,859561
247	8	70	X1 X2 X3 X4 X5 X6 X7	0,4447572	0,3892329	7,316917	1,683083
248	8	70	X1 X2 X4 X5 X6 X7 X8	0,4434266	0,3877692	7,483036	1,516964
249	8	70	X1 X2 X3 X4 X6 X7 X8	0,4407384	0,3848122	7,818633	1,181367
255	9	69	X1 X2 X3 X4 X5 X6 X7 X8	0,4472958	0,3832141	9,000000	1

En la **Figura 1**, se puede ver la comparación del aumento o reducción en cada una de las medidas previamente mencionadas para los mejores modelos con k=1,…, 8 variables predictoras. Los números encerrados en triángulos rojos son los modelos que de acuerdo con el criterio dado por cada estadístico es el mejor a nivel numérico. La idea es partir de un modelo que considera todo el conjunto de variables que han sido observadas, para posteriormente identificar un subconjunto de estas variables que sean potencialmente útiles en la construcción del modelo final. La selección de este subconjunto debe tener en cuenta el uso que se le dará al modelo de regresión, por ejemplo, para el caso de pronóstico se consideran estadísticos como el coeficiente de determinación y medidas de sesgo como el C_p, entre otros. Para esto se construyen todas las posibles combinaciones de este conjunto de k variables y se comparan los modelos resultantes con base a algunos estadísticos que miden la calidad del modelo. Adicionalmente, un factor de interés para el método de selección será la parsimonia del modelo, es decir, en este caso, la menor cantidad de predictoras posible. En este estudio, se tienen en consideración los siguientes criterios:

Coeficiente de determinación R^2

Este estadístico mide la proporción de variabilidad total observada en la variable respuesta que es explicada por el modelo de regresión, por ende, es de esperarse que, con base a esta medida, se busque un modelo que abarque una parte importante de esta variabilidad, es decir, de todas las posibles regresiones, el modelo candidato de acuerdo con el R^2 será aquel que tenga un valor mayor. Observando la **tabla 8**, es de interés notar que incluso con las 8 variables regresoras el R^2 no alcanza un valor superior a 0.4473, y éste al ser una medida que se infla cuando aumenta el número de predictoras, es de esperarse que el resto de los valores observados sean inferiores al del modelo completo, y continúa decreciendo a medida que se reduce el número de variables.

A pesar de que se cuenta con proporciones de variabilidad explicada bajos, en este caso se puede pensar que es importante garantizar al menos el 40% de la variabilidad sea explicada por el modelo que se seleccione, teniendo en cuenta esto y que las reducciones en esta medida parecen ser poco significativas, los modelos 93, 163 y 219 son los más apropiados de acuerdo con el criterio, los valores observados en el R^2 de estos modelos en la **tabla 8** se encuentran resaltados de color naranja. No obstante, es importante destacar que como se observa en la **figura 1**, a partir de modelos con 4 variables predictoras no hay un aumento considerable en el R^2 lo que podría indicar que no es razonable afectar la parsimonia del modelo agregando más variables que no contribuyan de forma útil al incremento de éste. Teniendo en consideración lo anterior, el modelo candidato a ser seleccionado por este criterio es el modelo número 93.

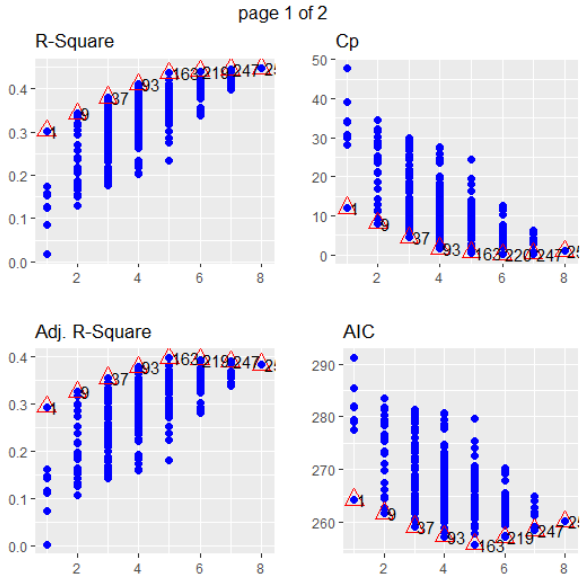

Figura 1: Gráfico de resumen de los estadísticos de interés.

Coeficiente de determinación ajustado R^2_{adj}

A pesar de que esta medida no tenga una interpretación directa como la tiene el R^2, si se espera que este valor se aproxime a la unidad cuando hay buen ajuste. Como criterio de selección, es semejante al R^2, se busca aquel modelo con mayor valor observado. De la **figura 1,** se puede identificar que hay incrementos en el R^2_{adj} al reducir los modelos hasta 5 variables predictoras, de allí en adelante los modelos con 4 o menos variables predictoras tienen R^2_{adj} inferiores en comparación al modelo completo, los modelos candidatos con R^2_{adj} más elevados son aquellos que se pueden observar en la **tabla 8** con estos valores resaltados de color azul, es decir, los modelos 93, 163, 219 y 247. Como sucedió con el criterio anterior, al no existir diferencias relevantes en estos incrementos o reducciones, se convierte en un factor importante la parsimonia del modelo, en este orden de ideas, los modelos más apropiados en base a esta medida son el 93 y el 163, con 4 y 5 variables respectivamente, de ellos se prefiere el 93.

Medida de sesgo C_p

Este estadístico mide el sesgo del modelo, siendo p el número de parámetros. Según este criterio, lo que se busca es que C_p sea pequeño, así, el modelo candidato según esta medida será aquel para el cual C_p es el más pequeño posible y que adicionalmente la diferencia absoluta entre C_p y p sea mínima. Este estadístico se calcula así: $C_p = \frac{SSE_P}{MSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)} - (n-2p)$,

donde SSE_P es la suma de cuadrados de los residuos del modelo con $p-1 > k$ variables predictoras y n el número de observaciones, para este estudio, 78. Este criterio al referirse a el sesgo del modelo $E[\hat{Y}_i] - \mu_{Y|x_i}$, hace énfasis en el ajuste del modelo.

En la **figura 1** se puede determinar fácilmente que a partir de modelos con 4 o más variables, la diferencia entre estas medidas de sesgo parece no variar de manera importante, esto se puede ver de forma más clara en los valores resaltados con color rojo en la **tabla 8**, donde la diferencia entre los modelos 93, 163 y 219 para esta medida en cuestión no cambia en más de 2 unidades. Sin embargo, observando aquellos valores que se encuentran resaltados de color púrpura, los cuales corresponden a $|C_p - p|$, se puede identificar que el modelo 93 es el que minimiza notablemente esta diferencia.

Al analizar de forma individual y general los modelos destacados según los criterios mencionados y teniendo presente el principio de parsimonia, parece ser evidente que el mejor modelo es el enumerado en la **tabla 8** como 93. La proporción de variabilidad explicada y el valor detectado para su medida de sesgo en comparación a los otros modelos que además ingresaban más variables indica que este modelo es el que mejor equilibra las medidas de calidad del ajuste con el criterio de parsimonia.

Ajuste del modelo seleccionado

El modelo que se ajustará es el modelo 93 de las 255 posibles regresiones, el cual tiene como variables regresoras la edad promedio de los pacientes, riesgo de infección, razón de rutina de rayos X del pecho y censo promedio diario, a través de las cuales se busca explicar la longitud promedio de permanencia. En la **Tabla 1** fueron definidas las variables que serán consideradas en el modelo 93 siento Y la variable respuesta y $X_j, j = 1,2,4,6$ las variables predictoras.

La ecuación del modelo es: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$ (2)

Tabla 9. Parámetros estimados del modelo 93.

Parámetros	Estimación	Error Std	T_0	$P(\lvert t_{73}\rvert > \lvert T_0\rvert)$
β_0	2.534541	1.759750	1.440	0.15406
β_1	0.065986	0.030065	2.195	0.03136
β_2	0.398305	0.125860	3.165	0.00226
β_4	0.015631	0.008096	1.931	0.05739
β_6	0.003172	0.001162	2.729	0.00795
$\sqrt{MSE}$ = 1.205 R^2 = 0.4099, R^2_{adj} = 0.3776,				

La ecuación ajustada es: $\hat{Y}_i = 2.534541 + 0.065986X_{i1} + 0.398305X_{i2} + 0.015631X_{i4} + 0.003172X_{i6}$

Test ANOVA del modelo

Para la prueba de significancia del modelo que ha sido ajustado, se contrastan las siguientes hipótesis:

$$H_0: \beta_1 = \beta_2 = \beta_4 = \beta_6 = 0 \ \ vs. H_1: Algún\ \beta_j \neq 0\ con\ j = 1,2,4,6$$

El estadístico de prueba está dado por $F_0 = \frac{SSR/4}{SSE/73} = \frac{MSR}{MSE}$, $F_0 \sim f_{4,73}$ Bajo H_0 y los supuestos del MRLM

Tabla 10. Tabla ANOVA del MRLM 93.

Fuente	Sum Sq	Df	Mean Sq	F_0	$P(f_{4,73} > F_0)$
Modelo	73.675	4	18.419	12.676	6.953×10^{-8}
Error	106.068	73	1.453		

El resultado para este caso $F_0 = \frac{73.675/4}{106.068/73} = 12.676$. La hipótesis nula se rechaza para valores grandes de F_0 o usando $VP = P(f_{4,73} > 12.676) = 6.953 \times 10^{-8}$, se rechaza la hipótesis nula pues el valor p toma un valor muy pequeño así que hay poca probabilidad de equivocarse al rechazar H_0. Lo anterior quiere decir que la regresión es significativa, es decir, al menos una de las cuatro variables predictoras contribuye de forma significativa a explicar la longitud promedio de permanencia de los pacientes en el hospital.

Coeficiente de determinación muestral R^2

La proporción de variabilidad total observada de la longitud promedio de permanencia, que es explicada por el modelo se cuantifica en este caso está dada por $R^2 = \frac{SSR}{SST} = \frac{73.675}{106.068+73.675} = 0.4099$, precisamente este estádistico fue usado como uno de los criterios de selección. Por lo tanto, se concluye que el 40.99% de la variabilidad total observada en la permanencia promedio de los pacientes en el hospital, es explicada por la regresión lineal múltiple propuesta de Y vs. X_1, X_2, X_4, X_6. Así que alrededor del 60% de la variabilidad observada en la variable respuesta no está siendo representada por el modelo. El R^2_{adj} en este caso está dado por $R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{1.453}{\frac{106.068+73.675}{77}} = 0.3775501$, como este valor no se acerca esta medida da un indicio de que no hay buen ajuste.

4. Reducción de variables mediante selección automática

Hay tres procedimientos de selección automática, que en esencia buscan seleccionar de todas predictoras propuestas inicialmente en un modelo completo, un subconjunto de ellas donde todas sean simultaneamente significativas en presencia de las demás.

Método Forward

El método consiste en agregar variables de manera consecutiva partiendo del modelo nulo, modelo sin variables predictoras. Para ello, se escoge una variable candidata a ingresar examinando de la **tabla 8** aquel modelo con mayor R^2 y k=1,2,3,4,5,6,7,8 variables según sea el interés en el paso, se comprueba la significancia de la candidata a entrar en presencia de las variables que ya se encuentran en el modelo mediante una prueba de hipótesis. El estadístico de prueba estará dado por $F_0 = \frac{\text{SSR}(X_j|X_i \text{ que ya entraron})/[dfe(MR) - dfe(MF)]}{MSE(MF)}$, donde la suma de cuadrados del numerador corresponde a la suma de cuadrados de tipo 2 de X_j, variable candidata a entrar al modelo, dfe corresponde a los grados de libertad de la suma de cuadrados de residuos del modelo. El modelo reducido (MR) es el que no considera a la candidata a entrar y el completo (MF) es el que la incluye. El estadístico de prueba bajo H_0 y los supuestos de los errores del modelo, se distribuye como una f con $[dfe(MR) - dfe(MF)]$ y $dfe(MF)$ grados de libertad, en el proceso forward, la diferencia $dfe(MR) - dfe(MF)$ siempre es igual a 1 y $dfe(MF) = $ v. F_0 puede escribirse en términos del coeficiente de determinación (R^2) del modelo reducido y completo usando las siguientes relaciones:

$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST}$ De la relación anterior se tiene que: $SSE = SST(1 - R^2)$ y $MSE = \frac{SST(1-R^2)}{dfe}$, así F_0 puede expresarse del siguiente modo: $F_0 = \frac{[SSE(MR) - SSE(MF)]/[dfe(MR) - dfe(MF)]}{MSE(MF)} = \frac{[SST(1-R^2(MR)) - SST(1-R^2(MF))]/1}{\frac{SST(1-R^2(MF))}{dfe(MF)}} = \frac{dfe(MF)[R^2(MF) - R^2(MR)]}{1-R^2(MF)}$(2)

La ecuación anterior será la utilizada durante el proceso para el cálculo de los estadísticos de prueba, el criterio de rechazo para la hipótesis nula de la prueba en cualquier caso es $F_0 > f_{0.05,1,\nu}$, de rechazarse la hipótesis nula la variable es agregada al modelo y en el siguiente paso se repite el mismo procedimiento, pero ahora buscando entre modelos con una variable más. El procedimiento finaliza cuando la variable candidata a ingresar resulta no ser significativa en presencia de las demás variables.

Paso 0: Partiendo del modelo sin predictoras $Y_i = \beta_0 + E_i, E_i \sim IID\ N(0, \sigma^2)$

Tabla 11: Resumen del procedimiento Forward

Paso 1		**Paso 2**	
Candidata a ingresar	$X_2 = \boldsymbol{RINF}$	Candidata a ingresar	$X_6 = \boldsymbol{PDP}$
Modelo completo (MF)	$Y_i = \beta_0 + \beta_2 X_{i2} + E_i, E_i \sim IID\ N(0, \sigma^2)$	Modelo completo (MF)	$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$
Modelo Reducido (MR)	$Y_i = \beta_0 + E_i, E_i \sim IID\ N(0, \sigma^2)$	Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_2 X_{i2} + E_i, E_i \sim IID\ N(0, \sigma^2)$
Hipótesis	$\boldsymbol{H_0}: \beta_2 = 0\ vs. \boldsymbol{H_1}: \beta_2 \neq 0$	Hipótesis	$\boldsymbol{H_0}: \beta_6 = 0\ vs. \boldsymbol{H_1}: \beta_6 \neq 0$
Estadístico de prueba	$F_0 = \frac{\text{SSR}(X_2)/1}{MSE(X_2)} \sim f_{1,76}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{76[0.30266992]}{1 - 0.30266992} = 32.9871$	Estadístico de prueba	$F_0 = \frac{\text{SSR}(X_6 \mid X_2)/1}{MSE(X_2, X_6)} \sim f_{1,75}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{75[0.34257063 - 0.30266992]}{1-0.34257063} = 4.5519.$
Conclusión	Como F_0 toma un valor muy grande, en particular $F_0 > f_{0.05,1,76} = 3.967$, se rechaza H_0 y se concluye que X_2 sí es significativa en el MRLS que la tiene a ella como variable predictora, por lo tanto, entra al modelo propuesto en el paso 0.	Conclusión	Se tiene que $F_0 > f_{0.05,1,75} = 3.9685$, se rechaza H_0 y se concluye que X_6 es significativa en presencia de X_2 y entra al modelo reducido.
Modelo final del paso 1	$Y_i = \beta_0 + \beta_2 X_{i2} + E_i, E_i \sim IID\ N(0, \sigma^2)$	Modelo final del paso 2	$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$
Paso 3		**Paso 4**	
Candidata a ingresar	$X_1 = \boldsymbol{EDAD}$	Candidata a ingresar	$X_4 = \boldsymbol{RRX}$
Modelo completo (MF)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$	Modelo completo (MF)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$
Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$	Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$
Hipótesis	$\boldsymbol{H_0}: \beta_1 = 0\ vs. \boldsymbol{H_1}: \beta_1 \neq 0$	Hipótesis	$\boldsymbol{H_0}: \beta_4 = 0\ vs. \boldsymbol{H_1}: \beta_4 \neq 0$

Estadístico de prueba	$F_0 = \frac{\text{SSR}(X_1\|X_2,X_6)/1}{MSE(X_1,X_2,X_6)} \sim f_{1,74}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{74[0.37975229-0.34257063]}{1-0.37975229} = 4.4360.$	Estadístico de prueba	$F_0 = \frac{\text{SSR}(X_4\|X_1,X_2,X_6)/1}{MSE(X_1,X_2,X_4,X_6)} \sim f_{1,73}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{73[0.40988891-0.37975229]}{1-0.40988891} = 3.7281.$
Conclusión	Se tiene que $F_0 > f_{0.05,1,74} = 3.970$, se rechaza H_0, así, X_1 es significativa en presencia de X_2 y X_6, por lo tanto, entra al modelo reducido.	Conclusión	Se tiene que $F_0 < f_{0.05,1,73} = 3.920$, no se rechaza H_0 y la conclusión de ello es que X_4 no es significativa en presencia de las demás variables, por lo tanto, no es razonable ingresar esta variable al modelo reducido.
Modelo final del paso 3	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$	Modelo final del paso 4	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$

El algoritmo finaliza y el modelo final del proceso es: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$

Método Backward

Este método de selección consiste en ir eliminando variables partiendo del modelo que considera todas las variables a disposición, en este caso 8 covariables. En cada paso se selecciona una variable candidata a salir del modelo buscando entre los modelos con el número de variables de interés que tenga mayor R^2_{adj}, ese modelo será el denotado como reducido (MR) y el modelo completo (MF) será el que considere también la variable candidata a salir entre sus predictoras. La decisión de excluir o no a la variable candidata se basa en una prueba de hipótesis para comprobar si la variable es significativa en presencia de las demás variables, si no lo es, la variable se excluye del modelo y en el siguiente paso se busca entre modelos con una variable menos para la selección de la candidata a salir. De ser significativa, no sale y finaliza el proceso. El estadístico de prueba se define igual que en el punto anterior, empleando la suma de cuadrados de tipo 2 y el MSE del modelo completo, pero en este caso se utiliza el equivalente en términos del R^2_{adj} haciendo uso de las siguientes relaciones: $R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{\frac{SSE}{dfe}}{MST}$, de la relación anterior se tiene que: $MSE = MST\left(1 - R^2_{adj}\right)$ y $SSE = MST\left(1 - R^2_{adj}\right)dfe$, así F_0 puede expresarse del siguiente modo:

$$F_0 = \frac{\frac{[SSE(MR) - SSE(MF)]}{[dfe(MR) - dfe(MF)]}}{MSE(MF)} = \frac{\frac{\left[MST\left(1 - R^2_{adj}(MR)\right)dfe(MR) - MST\left(1 - R^2_{adj}(MF)\right)dfe(MF)\right]}{1}}{MST\left(1 - R^2_{adj}(MF)\right)} = \frac{dfe(MR)\left(1 - R^2_{adj}(MR)\right) - dfe(MF)\left(1 - R^2_{adj}(MF)\right)}{\left(1 - R^2_{adj}(MF)\right)}$$

Donde dfe corresponde a los grados de libertad de la suma de cuadrados de residuos del modelo. El estadístico de prueba bajo H_0 y los supuestos de los errores del modelo, se distribuye como una f con $[dfe(MR) - dfe(MF)]$ y $\nu = dfe(MF)$ grados de libertad, en el proceso Backward al igual que en el forward, la diferencia $dfe(MR) - dfe(MF) = 1$ siempre será igual a 1, $F_0 \sim f_{1,\nu}$. El criterio de rechazo de la hipótesis nula es $F_0 > f_{0.05,1,\nu}$.

Paso 0: se parte del modelo definido en la **ecuación 1** con las ocho predictoras consideradas inicialmente

Tabla 12: Resumen del procedimiento Backward

Paso 1		**Paso 2**	
Candidata a salir	$X_8 =$ **FSD**	Candidata a salir	$X_3 =$ ***RRC***
Modelo completo (MF)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + E_i, E_i \sim IID\ N(0,\sigma^2)$	Modelo completo (MF)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0,\sigma^2)$
Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0,\sigma^2)$	Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0,\sigma^2)$
Hipótesis	$\boldsymbol{H_0}: \beta_8 = 0\ vs.\ \boldsymbol{H_1}: \beta_8 \neq 0$	Hipótesis	$\boldsymbol{H_0}: \beta_3 = 0\ vs.\ \boldsymbol{H_1}: \beta_3 \neq 0$
Estadístico de prueba	$F_0 = \frac{\text{SSR}(X_8\|X_1,X_2,X_3,X_4,X_5,X_6,X_7)}{MSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)} \sim f_{1,69}$ bajo y los supuestos de los errores del modelo. $F_0 = \frac{70[1-0.3892329]-69[1-0.3832141]}{1-0.3832141} = 0.3169169399.$	Estadístico de prueba	$F_0 = \frac{\frac{\text{SSR}(X_3\|X_1,X_2,X_4,X_5,X_6,X_7)}{1}}{MSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7)} \sim f_{1,70}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{71[1-0.3930502]-70[1-0.3892329]}{1-0.3892329} = 0.5562493.$
Conclusión	Como F_0 toma un valor pequeño, en particular $F_0 < f_{0.05,1,69} = 3.979807$, no se rechaza H_0 y se concluye que X_8 no es significativa en presencia de las demás variables, por lo tanto, sale del modelo propuesto en el paso 0.	Conclusión	Como F_0 toma un valor pequeño, en particular $F_0 < f_{0.05,1,70} = 3.977779$, no se rechaza H_0 y se concluye que X_3 no es significativa en presencia de las demás variables predictoras, sale del modelo final del paso 1.
Modelo final del paso 1	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0,\sigma^2)$	Modelo final del paso 2	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0,\sigma^2)$
Paso 3		**Paso 4**	
Candidata a ingresar	$X_5 =$ ***NCAMAS***	Candidata a ingresar	$X_7 =$ ***NENFERM***
Modelo completo (MF)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0,\sigma^2)$	Modelo completo (MF)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0,\sigma^2)$
Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0,\sigma^2)$	Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$
Hipótesis	$\boldsymbol{H_0}: \beta_5 = 0\ vs.\ \boldsymbol{H_1}: \beta_5 \neq 0$	Hipótesis	$\boldsymbol{H_0}: \beta_7 = 0\ vs.\ \boldsymbol{H_1}: \beta_7 \neq 0$
Estadístico de prueba	$F_0 = \frac{\text{SSR}(X_5\|X_1,X_2,X_4,X_6,X_7)/1}{MSE(X_1,X_2,X_4,X_5,X_6,X_7)} \sim f_{1,71}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{72[1-0.3966612]-71[1-0.3930502]}{1-0.3930502} = 0.5716416745.$	Estadístico de prueba	$F_0 = \frac{\text{SSR}(X_7\|X_1,X_2,X_4,X_6)/1}{MSE(X_1,X_2,X_4,X_6,X_7)} \sim f_{1,72}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{73[1-0.3775541]-72[1-0.3966612]}{1-0.3966612} = 3.3118325.$
Conclusión	Como $F_0 < f_{0.05,1,71} = 3.97581$, no se rechaza H_0 y se concluye que X_5 no es significativa en presencia de las demás variables predictoras, por lo tanto, sale del modelo final del paso 2.	Conclusión	F_0 toma un valor lo suficientemente pequeño para que $F_0 < f_{0.05,1,72} = 3.973897$, no se rechaza H_0 y se concluye que X_7 no es significativa en presencia de las demás variables predictoras, por lo tanto, sale del

			modelo final del paso 3.
Modelo final del paso 3	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + \beta_7 X_{i7} + E_i, E_i \sim IID\ N(0, \sigma^2)$	Modelo final del paso 4	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$
Paso 5		**Paso 6**	
Candidata a salir	$X_4 = \boldsymbol{RRX}$	Candidata a salir	$X_1 = \boldsymbol{EDAD}$
Modelo completo (MF)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$	Modelo completo (MF)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$
Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$.	Modelo Reducido (MR)	$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$
Hipótesis	$\boldsymbol{H_0}: \beta_4 = 0\ vs.\ \boldsymbol{H_1}: \beta_4 \neq 0$	Hipótesis	$\boldsymbol{H_0}: \beta_1 = 0\ vs.\ \boldsymbol{H_1}: \beta_1 \neq 0$
Estadístico de prueba	$F_0 = \frac{\mathrm{SSR}(X_4 \mid X_1, X_2, X_6)/1}{MSE(X_1, X_2, X_4, X_6)} \sim f_{1,73}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{74[1-0.3546071]-73[1-0.3775541]}{1-0.3775541} = 3.72807.$	Estadístico de prueba	$F_0 = \frac{\mathrm{SSR}(X_1 \mid X_2, X_6)/1}{MSE(X_1, X_2, X_6)} \sim f_{1,74}$ bajo H_0 y los supuestos de los errores. $F_0 = \frac{75[1-0.3250392]-74[1-0.3546071]}{1-0.3546071} = 4.4360.$
Conclusión	F_0 toma un valor lo suficientemente pequeño para que $F_0 < f_{0.05,1,73} = 3.972038$, no se rechaza H_0 y se concluye que X_4 no es significativa en presencia de las demás variables predictoras, por lo tanto, sale del modelo final del paso 4.	Conclusión	Como $F_0 > f_{0.05,1,74} = 3.97023$ se rechaza H_0 concluyendo que X_1 sí es significativa en presencia de las demás variables predictoras, por lo tanto, no es razonable retirar esta variable del MRLM.
Modelo final del paso 5	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$	Modelo final del paso 6	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$

El algoritmo termina y modelo final es $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$.

Método Stepwise

Paso 0: se tiene el modelo sin variables predictoras $Y_i = \beta_0 + E_i, E_i \sim IID\ N(0, \sigma^2)$

Paso 1:

Sub-paso Forward: elegir variable X_j candidata a entrar al modelo con j=1,..8.

Se selecciona aquella variable cuyo MRLS $Y\ vs.\ X_j$ presenta mayor R^2, usando la **tabla 8** se elige a X_2, que representa el riesgo de infección, como candidata a ingresar al modelo, este razonamiento es el mismo para cada sub-paso Forward del proceso.

Modelo completo (MF): $Y_i = \beta_0 + \beta_2 X_{i2} + E_i, E_i \sim IID\ N(0, \sigma^2)$; Modelo reducido (MR): $Y_i = \beta_0 + E_i, E_i \sim IID\ N(0, \sigma^2)$

Se debe verificar mediante una prueba si X_2 es significativa en el modelo completo, es decir, si es razonable ingresar esta variable al modelo nulo: $\boldsymbol{H_0}: \beta_2 = 0\ vs.\ \boldsymbol{H_1}: \beta_2 \neq 0$. Se define F_0 como el estadístico de prueba, se usarán los R^2 para calcular a F_0 como se hizo en el método forward, de la siguiente manera: $F_0 = \frac{\mathrm{SSR}(X_2)/1}{MSE(X_2)} = \frac{dfe(MF)[R^2(MF)]}{1-R^2(MF)} \sim f_{1,76}$ bajo H_0 y los supuestos de los errores, para el modelo reducido de este paso SSE=SST pues en el modelo nulo SSR=0, por lo tanto, $R^2(MR) = 1 - 1 = 0$. $F_0 = \frac{76[0.30266992]}{1-0.30266992} = 32.9871$, el criterio de rechazo de H_0 es si $F_0 > f_{0.05,1,76}$.

Como F_0 toma un valor muy grande, en particular $F_0 > f_{0.05,1,76} = 3.967$, se rechaza H_0 y se concluye que X_2 sí es significativa en el MRLS que la tiene a ella como variable predictora, por lo tanto, entra al modelo propuesto en el paso 0.

Sub-paso Backward: al ser la primera variable en ingresar al modelo no hay otras variables a las cuales se les deba verificar significancia. Modelo final del paso 1: $Y_i = \beta_0 + \beta_2 X_{i2} + E_i, E_i \sim IID\ N(0, \sigma^2)$

Paso 2

Sub-paso Forward: elegir otra variable X_j, j≠ 2 candidata a entrar al modelo. Se elige X_6, variable que representa el censo promedio diario, como candidata a entrar al modelo final del paso 1.

Modelo completo (MF): $Y_i = \beta_0 + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$; Modelo reducido (MR): $Y_i = \beta_0 + \beta_2 X_{i2} + E_i, E_i \sim IID\ N(0, \sigma^2)$

Se debe verificar mediante una prueba si X_6 es significativa en el modelo completo, es decir, si es significativa en presencia de X_2. Planteando: $\boldsymbol{H_0}: \beta_6 = 0\ vs. \boldsymbol{H_1}: \beta_6 \neq 0$ Se define el estadístico de prueba:

$F_0 = \frac{\mathrm{SSR}(X_6 \mid X_2)/1}{MSE(X_2, X_6)} \sim f_{1,75}$ bajo H_0 y los supuestos de los errores, haciendo uso de la expresión deducida en el paso anterior para F_0, el valor observado del estadístico de prueba es: $F_0 = \frac{dfe(MF)[R^2(MF) - R^2(MR)]}{1-R^2(MF)} = \frac{75[0.34257063-0.30266992]}{1-0.34257063} = 4.5519.$

Se tiene que $F_0 > f_{0.05,1,75} = 3.9685$ y empleando un criterio de rechazo de H_0 análogo al del paso anterior, se rechaza H_0 y la conclusión de ello es que X_6 es significativa en presencia de X_2, por lo tanto, entra al modelo reducido.

Sub-paso backward: se comprueba si X_2 es significativa en el modelo en presencia de la variable X_6 que acaba de ingresar al modelo.

Modelo completo (MF): $Y_i = \beta_0 + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$; Modelo reducido (MR): $Y_i = \beta_0 + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$

Se debe comprobar mediante una prueba si X_2 es significativa en el modelo completo, es decir, si es significativa en presencia de X_6.

Contrastando: $\boldsymbol{H_0}: \beta_2 = 0\ vs. \boldsymbol{H_1}: \beta_2 \neq 0$, definiendo el estadístico de prueba $F_0 = \frac{\mathrm{SSR}(X_2 \mid X_6)/1}{MSE(X_2, X_6)} \sim f_{1,75}$ bajo H_0 y los supuestos de los errores, el estadístico de prueba es numéricamente igual a: $F_0 = \frac{dfe(MF)[R^2(MF) - R^2(MR)]}{1-R^2(MF)} = \frac{75[0.3425706-0.1734469]}{1-0.3425706} = 19.2937485.$ El criterio de rechazo de H_0 es $F_0 > f_{0.05,1,75}$.

Como F_0 toma un valor grande, en particular $F_0 > f_{0.05,1,75} = 3.968471$, por ende, se rechaza H_0 concluyendo que X_2 sí es significativa en presencia de la variable predictora que entró al modelo, por lo tanto, X_2 esta variable debe permanecer en el MRLM que estima la longitud promedio de permanencia de los pacientes. Modelo final del paso 2: $Y_i = \beta_0 + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$.

Paso 3

Sub-paso Forward: elegir otra variable X_j, j≠ 2, 6 candidata a entrar al modelo.

Se selecciona X_1, variable que representa la edad promedio de los pacientes.

Modelo completo (MF): $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$;Modelo reducido (MR): $Y_i = \beta_0 + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0, \sigma^2)$

Se debe verificar mediante una prueba si X_1 es significativa en el modelo completo, es decir, si es significativa en presencia de X_2 y X_6.

$\boldsymbol{H_0}: \beta_1 = 0\ vs. \boldsymbol{H_1}: \beta_1 \neq 0$

Se define el estadístico de prueba: $F_0 = \frac{\text{SSR}(X_1|X_2,X_6)/1}{MSE(X_1,X_2,X_6)} \sim f_{1,74}$ bajo H_0 y los supuestos de los errores, el valor observado del estadístico de prueba es $F_0 = \frac{dfe(MF)[R^2(MF)-R^2(MR)]}{1-R^2(MF)} = \frac{74[0.37975229-0.34257063]}{1-0.37975229} = 4.4360$.

Se tiene que $F_0 > f_{0.05,1,74} = 3.9702$ y empleando un criterio de rechazo de H_0 análogo al de pasos previos, se rechaza H_0 y la conclusión de ello es que X_1 es significativa en presencia de X_2 y X_6, por lo tanto, entra al modelo reducido.

Sub-paso Backward: ¿Alguna de las $X_j, j = 2,6$ que ya se encontraban en el modelo dejaron de ser significativas en presencia de la nueva variable X_1? Para responder a la pregunta se realizan 2 pruebas para verificar la significancia individual de las variables mencionadas en el modelo completo. Se plantean las pruebas así: $\boldsymbol{H_0}: \beta_j = 0 \; vs. \boldsymbol{H_1}: \beta_j \neq 0 \; , j = 2,6$

se comprueba si X_j es significativa en el modelo en presencia de las demás variables.

Modelo completo (MF): $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$

Modelo reducido para j=2: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$ y MR para j=6, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i, E_i \sim IID\ N(0,\sigma^2)$

El estadístico de prueba del primero es $F_{02} = \frac{\text{SSR}(X_2|X_1,X_6)/1}{MSE(X_1,X_2,X_6)}$ y del segundo $F_{06} = \frac{\text{SSR}(X_6|X_1,X_2)/1}{MSE(X_1,X_2,X_6)}$ ambos se distribuyen $f_{1,74}$ bajo H_0 y los supuestos. Haciendo el cálculo pertinente se tiene $F_{02} = \frac{74[0.3797523-0.20440632]}{1-0,3797523} = 20.920$ y $F_{06} = \frac{74[0.3797523-0.33348960]}{1-0,3797523} = 5.519$ .El criterio de rechazo de H_0 es $F_{0j} > f_{0.05,1,74=3.97023}$. Las dos hipótesis nulas se rechazan y por ende, no se deben retirar las variables del modelo completo.

Modelo final del paso 3: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$

Paso 4

Sub-paso Forward: elegir otra variable X_j, j≠ 1, 2, 6 candidata a entrar al modelo.

X_4 es seleccionada, variable que representa la razón de rutina de rayos X del pecho.

Modelo completo (MF): $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 X_{i4} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$; Modelo reducido (MR): $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$;

Se debe comprobar mediante una prueba si X_4 es significativa en el modelo completo, es decir, si es significativa en presencia de X_1, X_2 y X_6. Se plantea $\boldsymbol{H_0}: \beta_4 = 0 \; vs. \boldsymbol{H_1}: \beta_4 \neq 0$

Se define el estadístico de prueba: $F_0 = \frac{\text{SSR}(X_4|X_1,X_2,X_6)/1}{MSE(X_1,X_2,X_4,X_6)} \sim f_{1,73}$ bajo H_0 y los supuestos de los errores, haciendo uso de la expresión en (2) para F_0, el valor observado del estadístico de prueba es $F_0 = \frac{dfe(MF)[R^2(MF)-R^2(MR)]}{1-R^2(MF)} = \frac{73[0.40988891-0.37975229]}{1-0.40988891}$ 3.7281.

Se tiene que $F_0 < f_{0.05,1,73} = 3.920$ y empleando un criterio de rechazo de H_0 análogo al de pasos previos, no se rechaza H_0 y la conclusión de ello es que X_4 no es significativa en presencia de X_1, X_2 y X_6, por lo tanto, no es razonable ingresar esta variable al modelo reducido. Como en el sub-paso anterior no ingresó una nueva variable al modelo, el sub-paso backward no se realiza y finaliza el proceso Stepwise, el modelo final es $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_6 X_{i6} + E_i, E_i \sim IID\ N(0,\sigma^2)$ (3), **modelo 37 de la tabla 8.**

Tabla 13: Resumen de R del método de selección Backward

Variables candidatas					$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$	
Mediante el uso del valor p se eliminan					X_3, X_4, X_5, X_7, X_8	
Resumen de eliminación						
Paso	Variable eliminada	R^2	R^2_{adj}	C_p	AIC	RMSE
1	X_8	0.4448	0.3892	7.3169	258.5789	1.1940
2	X_3	0.4403	0.3931	5.8677	257.1963	1.1903
4	X_5	0.4358	0.3967	4.4303	255.8218	1.1868
3	X_7	0.4099	0.3776	5.6699	257.3295	1.2054
5	X_4	0.3798	0.3546	7.4322	259.2145	1.2274

Tabla 14: Resumen de R del método de selección 1. Forward 2. Stepwise.

Variables candidatas					$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$	
Mediante el uso del valor p se selecciona					X_1, X_2, X_6	
Resumen de selección						
Paso	Variable ingresada	R^2	R^2_{adj}	C_p	AIC	RMSE
1	X_2	0.3027	0.2935	13.0552	264.3515	1.2842
2	X_6	0.3426	0.3250	10.0740	261.7556	1.2552
3	X_1	0.3798	0.3546	7.4322	259.2145	1.2274

Variables candidatas					$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$	
Mediante el uso del valor p se selecciona					X_1, X_2, X_6	
Resumen de selección/eliminación						
Paso	Variable Entró/salió	R^2	R^2_{adj}	C_p	AIC	RMSE
1	X_2-Entró	0.3027	0.2935	13.0552	264.3515	1.2842
2	X_6-Entró	0.3426	0.3250	10.0740	261.7556	1.2552
3	X_1-Entró	0.3798	0.3546	7.4322	259.2145	1.2274

Tabla 15: Resumen del modelo final de la selección automática

ANOVA						
Fuente		Sum Sq	Df	Mean Sq	F_0	$P(f_{3,74} > F_0)$
Modelo		68.258	3	22.753	15.102	0.0000
Error		111.485	74	1.507		
Total		179.743	77			
Parámetros estimados						
Parámetros	Estimación	Error Std	Error estimación	T_0	$P(\lvert t_{74}\rvert > \lvert T_0\rvert)$	Intervalo
β_0	3.471	1.722		2.015	0.047	(0.039, 6.903)
β_1	0.064	0.031	0.194	2.106	0.039	(0.003, 0.125)
β_2	0.515	0.113	0.464	4.574	0	(0.290, 0.739)
β_6	0.003	0.001	0.239	2.349	0.021	(0,0.005)
$\sqrt{MSE}$ = 1.227, R^2 = 0.380, R^2_{adj} = 0.355,						

Ajuste del modelo final

Los tres procesos de selección automática que se desarrollaron previamente como se esperaba llegaron al mismo modelo final, el cual corresponde al modelo 37 de la **tabla 8,** modelo que busca explicar la longitud promedio de permanencia de los pacientes a través del subconjunto de variables nombradas como EDAD, RINF y PDP, que corresponden a la edad promedio de los pacientes, el riesgo de infección y el censo promedio diario respectivamente.

Usando la **tabla 15** la ecuación ajustada es: $\hat{Y}_i = 3.471 + 0.064X_{i1} + 0.515X_{i2} + 0.003X_{i6}$

Coeficiente de determinación muestral R^2

La proporción de variabilidad total observada de la variable respuesta, DPERM, que es explicada por el modelo se cuantifica usando el R^2. El valor observado para este estadístico, valor que fue implementado en la selección automática para el cálculo del estadístico de prueba, es 0.3798. Por lo tanto, se dice que el 37.98% de la variabilidad total observada en la permanencia promedio de los pacientes en el hospital es explicada por la regresión lineal múltiple propuesta de Y vs. X_1, X_2, X_6 lo que indica que el modelo está dejando gran proporción de la variabilidad en la variable respuesta sin explicar. El R^2_{adj} está dado por $R^2_{adj} = 1 - \frac{MSE}{MST} = 0.3546$, como se ha mencionado en análisis previos este estadístico no tiene una interpretación como el R^2, lo que se espera es que se aproxime a 1 cuando hay buen ajuste, lo que no se observa para este modelo. Así, se puede pensar que dejar solo las variables signficativas no contribuyó mucho con la mejora de estas medidas de calidad del modelo.

5. Selección del mejor modelo

De los procedimientos observados en el punto 3 y 4, se elige el modelo 93 por encima del modelo 37 porque a pesar de que la variable adicional X_4 considerada por el modelo 93 no es significativa de manera individual se observa que su inclusión en el modelo aporta a la reducción del C_p y a alcanzar al menos un 40% en la variabilidad que logra explicar el modelo, por ello se prefiere este modelo.

Análisis de residuales del modelo 93.

En la **Figura 2** no se observa evidencia fuerte en contra del supuesto de media 0 de los errores, de nuevo se evidencia el inconveniente con la forma en que fueron tomados los datos de este estudio, aglomeración de observaciones y por ende de residuales en algún punto del rango de observación de algunas variables y muy pocos datos en otros niveles, esto dificulta juzgar con certeza el comportamiento de la varianza, en este escenario podría decirse que tampoco hay evidencia fuerte en contra del supuesto de varianza constante de los errores del modelo. En las gráficas de residuales vs. RRX y PDP se observa algo de carencia de ajuste lo cual es coherente con el R^2_{adj} observado en este modelo que toma un valor de 0.3775, es decir que este modelo pese a tener muy buenas características de ajuste respecto a todos los modelos con los que fue comparado, es un modelo que se queda corto a la hora de explicar la variable respuesta en términos de estas covariables mediante un MRLM. Se detectan posibles puntos de balanceo y observaciones atípicas.

Validación del supuesto de normalidad de los errores del modelo.

Para validar la normalidad de los errores del modelo se van a contrastar las siguientes hipótesis: $H_0: E_i \sim Normal \; vs. H_1: E_i \nsim Normal$

La **Figura 3** no muestra evidencia fuerte en contra del supuesto de normalidad de los errores del modelo, en general se observa un comportamiento bastante similar de los cuantiles teóricos de la distribución normal y los cuantiles de los residuales con un poco de desajuste en los extremos, pero en general puede pensarse que el supuesto de normalidad de los errores es razonable, además el valor p de la prueba de Shapiro-Wilks respalda tal afirmación.

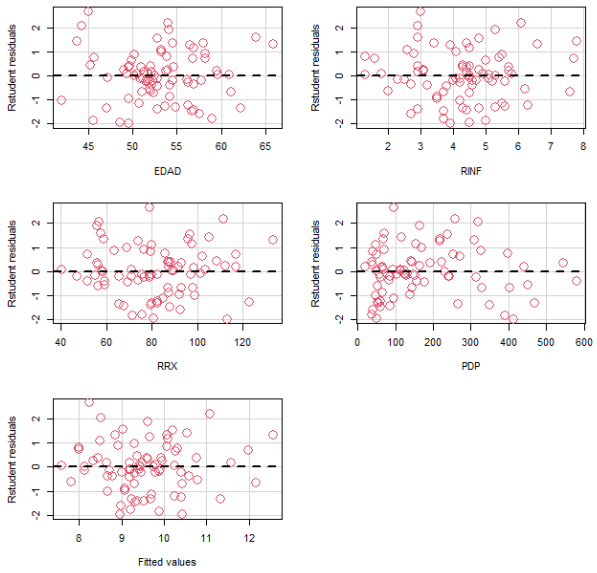

Figura 2: Residuales estudentizados vs. valores ajustados y predictoras modelo 93.

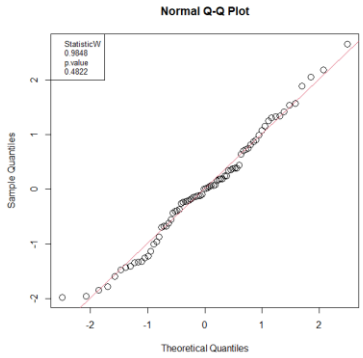

Figura 3: Gráfico de probabilidad normal con residuales estudentizados modelo 93.

PARTE II

Se quiere ajustar un MRLM para estudiar la relación entre la longitud de permanencia promedio de los pacientes en un hospital (Y) en función de la variable riesgo de infección (X_2) en presencia de las variables cualitativas Región y Afiliación a escuela de medicina. A continuación, se presentan algunos análisis descriptivos previos al ajuste del modelo de interés.

En la **figura 4** se observa que al diferenciar los pares de puntos observados (X_2, Y) según la región en la que se ubica el hospital quizás no sería tan razonable pensar en una única relación lineal para explicar la longitud de permanencia a través del riesgo de infección sin discriminar por la región, pues se observan varias relaciones lineales en el gráfico de dispersión con indicios de que estas podrían tener características diferentes según la región como variación en el intercepto (p.e en región 2 y 3) o variación en intercepto y pendiente (p.e región 1 respecto a la región 2 y 3). También se reporta una correlación general entre Y y X_2 de 0.55 pero al discriminar por región hay algunos cambios considerables que podrían ser indicio de que es necesario explicar la relación de interés mediante más de una recta pues la correlación de Y y X_2 en la región 1 y 3 es de 0.726 y 0.605 respectivamente, es decir que en estos casos la diferenciación por región mejora las correlaciones, lo cual indica una asociación lineal más fuerte de las variables de interés al eliminar el posible ruido que hagan las demás regiones en cada caso. Para la región 2 la correlación se mantiene prácticamente igual a la general y para la región 4 la correlación disminuye notablemente, tomando el valor de 0.08, esto último es coherente con lo que muestra el diagrama de dispersión, pues en la región 4 no es claro que exista tal tendencia lineal esperada en los pares (X_2, Y) observados, lo cual podría reflejarse en un problema de falta de ajuste en el modelo ajustado con las predictoras cualitativas. Otra situación que debe tenerse en cuenta es que en los diagramas de cajas de la variable predictora según la región se observa una variación en el rango observación de tal variable de región a región, es decir que la variable riesgo de infección en algunas regiónes fue observada en rangos considerablemente más pequeños respecto a otras regiones (p.e región 4 respecto a región 3), sin embargo, no se observan cambios muy significativos a nivel gráfico en las medias de la variable predictora al discriminar por región, pero si cambios en la variabilidad del riesgo de infección según la región. En cuanto a lo observado en la variable respuesta, las cajas no estan centradas alrededor del mismo valor, es decir, que aparentemente la media de la longitud de permanecia promedio cambia según la región, esto es un indicio de que es razonable pensar en diferentes relaciones lineales para explicar el fenómeno de interés, pues un MRL estima la respuesta media, sin embargo, no hay que perder de vista que en este análisis no se está involucrando el efecto de la variable AEM. Finalmente se detectan posibles puntos de balanceo.

Ahora, en la **figura 5** ignorando el efecto de la variable región y discriminando las observaciones (X_2, Y) según la variable Afiliación a escuela de medicina (AEM), no es tan claro si relación lineal entre la longitud de permanencia y el Riesgo de infección cambia según la categoría de afiliación a escuela de medicina pues no hay suficiente información para juzgarlo debido a que hay un evidente desbalance en la cantidad de observaciones que se tienen en cada nivel de la variable AEM. Sin embargo, los diagramas de cajas para la variable respuesta si presentan diferencias en su localización, es decir, que la media de la longitud de permanencia promedio (respuesta media) parece variar según la afiliación a una escuela de medicina, esto sin tener en cuenta el efecto de la región. En el diagrama de dispersión, como ya se mencionó, no hay la suficiente información para determinar si hay diferencias significativas en la relación lineal de interés según los niveles de AEM. Podría pensarse en el caso donde la relación lineal solo sufra cambios en el intercepto según el grupo, pero hay muy poca información para afirmarlo. La correlación general entre Y y X_2 es de 0.55. Al discriminar según AEM se tienen correlaciones entre Y y X_2 de 0.662 y 0.504 para hospitales afiliados y no afiliados a la escuela de medicina respectivamente. En este caso sí se observa diferencia en la media de la variable predictora según el nivel de AEM y también cambios en la variabilidad de nivel a nivel. Se detecta un posible punto de balanceo.

Tabla 16: Resumen variables de interés.

	DEPERM	RINF	AEM		REGIÓN	
Min	6.7	1.3	1 (SÍ)	11	1	18
Media	9.515	4.4	2 (NO)	67	2	22
Mediana	9.567	4.276			3	26
Max	13.95	7.8			4	12

A continuación, se presentan las frecuencias observadas en los 8 tratamientos o combinaciones posibles de los niveles de las dos variables cualitativas involucradas en el modelo que se desea ajustar.

Tabla 17: Frecuencias observadas en los tratamientos de interés.

AEM \ REGIÓN	1 (NE)	2 (NC)	3 (S)	4 (W)
1 (SÍ)	2	5	2	2
2 (NO)	16	17	24	10

En la **tabla 17** se observa gran desbalanceo en la cantidad de observaciones por tratamiento, esto es, muchas observaciones en algunas combinaciones de los niveles de las variables cualitativas en cuestión y muy pocas en otras combinaciones, en particular en aquellos tratamientos que involucran la no afiliación de los hospitales a una escuela de medicina hay muchos más datos respecto a los tratamientos donde hay afiliación. Esto podría ser un problema de la toma de datos o del diseño experimental que podría sesgar las conclusiones que se obtengan basadas en esta muestra, además tener tratamientos con 2 observaciones no es suficiente para juzgar con un buen nivel de certeza el comportamiento del fenómeno en ese caso.

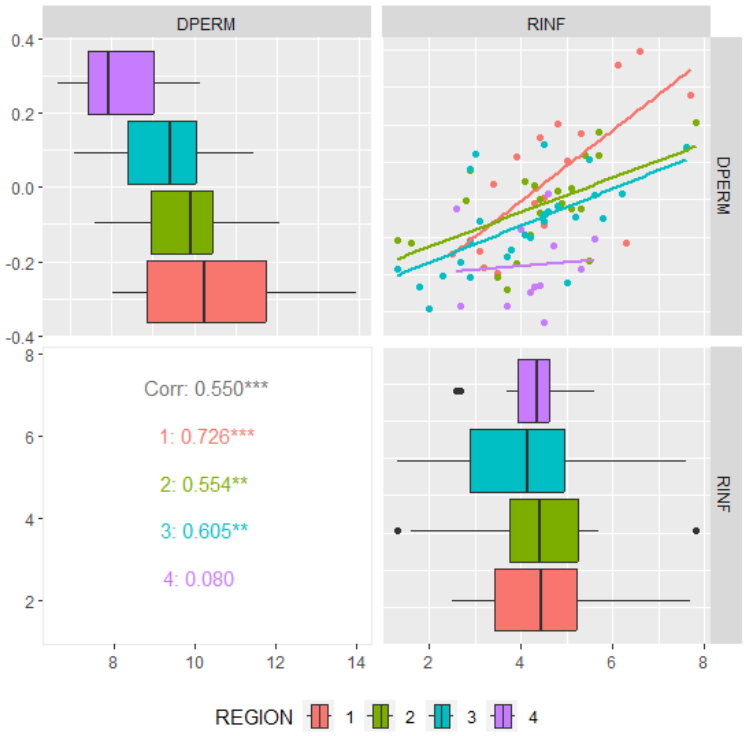

Figura 4: Diagramas de dispersión y boxplot de Y y X_2 en presencia de la variable Región.

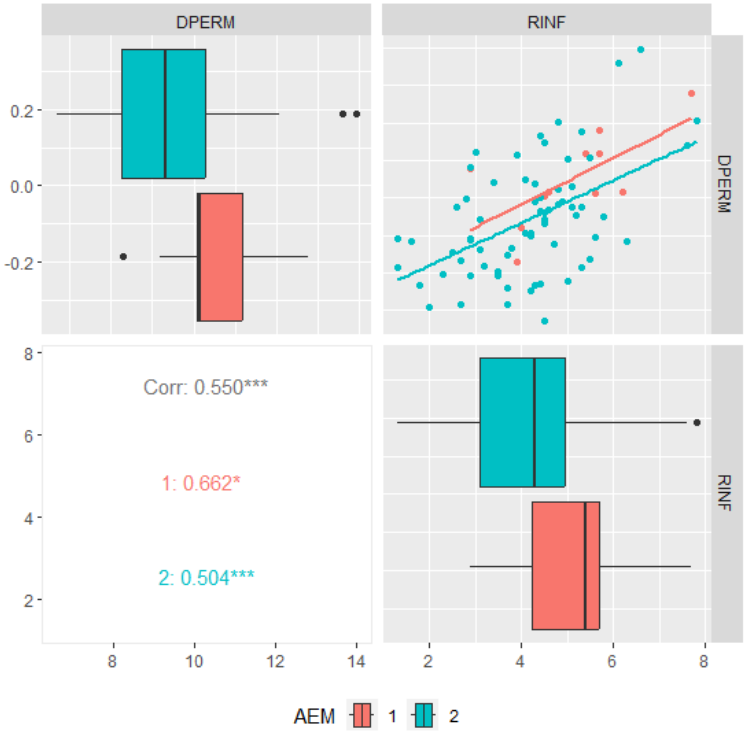

Figura 5: Diagramas de dispersión y boxplot de Y y X_2 en presencia de la variable AEM.

Definición del modelo general y ecuaciones de las ocho rectas derivadas de él.

Se quiere ajustar el siguiente modelo siendo Y la variable respuesta longitud de permanencia promedio y X_i el preditor cuantitativo riesgo de infección. Las demás variables se definen a continuación:

$$Y_i = \beta_0 + \beta_1 X_i + \alpha_1 A_{i1} + \gamma_1 R_{i1} + \gamma_2 R_{i2} + \gamma_3 R_{i3} + (\alpha\beta)_{11} X_i A_{i1} + (\beta\gamma)_{11} X_i R_{i1} + (\beta\gamma)_{12} X_i R_{i2} + (\beta\gamma)_{13} X_i R_{i3} + E_i, E_i \sim IID\ N(0, \sigma^2)\ (4)$$

Donde $A_{i1} = \begin{cases} 1\ si\ en\ la\ observación\ AEM = 1 \\ 0\ e.o.c \end{cases}$ y $R_{ij} = \begin{cases} 1\ si\ la\ observación\ i\ pertenece\ a\ la\ región\ j \\ 0\ e.o.c \end{cases}$; j=1, 2, 3

Ecuaciones para cada una de las combinaciones de los niveles de las variables categóricas:

El tratamiento de referencia corresponde a la combinación $A_{i1} = 0$ y $R_{ij} = 0;\ j = 1,2,3$; cuyo modelo asociado es $Y_i = \beta_0 + \beta_1 X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$. Este es el caso donde un hospital de la Región 4 no está afiliado a una escuela de medicina.

Tratamiento 1: $A_{i1} = 1$ y $R_{i1} = 1$; $Y_i = (\beta_0 + \alpha_1 + \gamma_1) + (\beta_1 + (\alpha\beta)_{11} + (\beta\gamma)_{11}) X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$

α_1 corresponde al incremento en el intercepto de la recta de regresión respecto al tratamiento de referencia cuando el hospital se encuentra afiliado a una escuela de medicina y γ_1 corresponde al incremento en el intercepto de la recta de regresión respecto al tratamiento de referencia cuando el hospital se ubica en la región 1. En otras palabras $(\alpha_1 + \gamma_1)$ es la diferencia que hay en la respuesta media cuando $X_i = 0$ entre un hospital de la región 4 no afiliado a una escuela de medicina y un hospital de la región 1 afiliado a una escuela de medicina.

Por su parte $(\alpha\beta)_{11}$ es el incremento en la pendiente de la recta de regresión respecto al tratamiento de referencia cuando un hospital está afiliado a una escuela de medicina y $(\beta\gamma)_{11}$ es el incremento que sufre la pendiente respecto al tratamiento de referencia cuando un hospital se ubica en la Región 1. Así $((\alpha\beta)_{11} + (\beta\gamma)_{11})$ es el incremento total que sufre la pendiente cuando un hospital de la Región 1 está afiliado a una escuela de medicina respecto a un hospital de la Región 4 que no está afiliado a una escula de medicina, es decir la diferencia en el efecto medio que tiene el riesgo de infección X_i sobre la longitud de permanencia promedio Y_i cuando se pasa del tratamiento de referencia al tratamiento con hospitales en la región 1 afiliados a la escuela de medicina.

Tratamiento 2: $A_{i1} = 0$ y $R_{i1} = 1$ $Y_i = (\beta_0 + \gamma_1) + (\beta_1 + (\beta\gamma)_{11}) X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$

Este es el caso donde un hospital ubicado en la Región 1 no está afiliado a una escuela de medicina, así que el efecto de A_1 es nulo en la ecuación. La interpetaciones de γ_1 y $(\beta\gamma)_{11}$ por sí solos fueron presentadas en el tratamiento anterior.

Tratamiento 3: $A_{i1} = 1$ y $R_{i2} = 1$ $Y_i = (\beta_0 + \alpha_1 + \gamma_2) + (\beta_1 + (\alpha\beta)_{11} + (\beta\gamma)_{12}) X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$

γ_2 corresponde al incremento en el intercepto de la recta de regresión respecto al tratamiento de referencia debido a que el hospital se ubica en la región 2 y $(\beta\gamma)_{12}$ es el incremento que sufre la pendiente respecto al tratamiento de referencia cuando un hospital se ubica en la Región 2. Así $(\alpha_1 + \gamma_2)$ es la diferencia que hay en la respuesta media cuando $X_i = 0$ entre un hospital de la región 4 no afiliado a una escuela de medicina y un hospital de la región 2 afiliado a una escuela de medicina. Por su parte $((\alpha\beta)_{11} + (\beta\gamma)_{12})$ es el incremento total que sufre la pendiente cuando un hospital de la Región 2 está afiliado a una escuela de medicina respecto a un hospital de la Región 4 que no está afiliado a una escuela de medicina, es decir la diferencia en el efecto medio que tiene X_i sobre la longitud de permanencia promedio cuando se pasa del tratamiento de referencia al tratamiento con hospitales en la región 2 afiliados a una escuela de medicina.

Tratamiento 4: $A_{i1} = 0$ y $R_{i2} = 1$ $Y_i = (\beta_0 + \gamma_2) + (\beta_1 + (\beta\gamma)_{12}) X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$

Este es el caso donde un hospital ubicado en la Región 2 no está afiliado a una escuela de medicina, así que el efecto de A_1 es nulo en la ecuación. La interpetaciones de γ_2 y $(\beta\gamma)_{12}$ por sí solos fueron presentadas en el tratamiento anterior.

Tratamiento 5: $A_{i1} = 1$ y $R_{i3} = 1$ $Y_i = (\beta_0 + \alpha_1 + \gamma_3) + (\beta_1 + (\alpha\beta)_{11} + (\beta\gamma)_{13}) X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$

γ_3 corresponde al incremento en el intercepto de la recta de regresión respecto al tratamiento de referencia debido a que el hospital se ubica en la región 3 y $(\beta\gamma)_{13}$ es el incremento que sufre la pendiente respecto al tratamiento de referencia cuando un hospital se ubica en la Región 3.De tal forma que $(\alpha_1 + \gamma_3)$ es la diferencia que hay en la respuesta media cuando $X_i = 0$ entre un hospital de la región 4 no afiliado a una escuela de medicina y un hospital de la región 3 afiliado a una escuela de medicina. Por su parte $((\alpha\beta)_{11} + (\beta\gamma)_{13})$ es el incremento total que sufre la pendiente cuando un hospital de la Región 3 está afiliado a una escuela de medicina respecto a un hospital de la Región 4 que no está afiliado a una escula de medicina, o equivalentemente la diferencia en el efecto medio que tiene X_i sobre Y_i cuando se pasa del tratamiento de referencia al tratamiento con hospitales en la región 3 afiliados a la escuela de medicina.

Tratamiento 6: $A_{i1} = 0$ y $R_{i3} = 1$ $Y_i = (\beta_0 + \gamma_3) + (\beta_1 + (\beta\gamma)_{13}) X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$

Este es el caso donde un hospital ubicado en la Región 3 no está afiliado a una escuela de medicina, así que el efecto de A_1 de nuevo es nulo en la ecuación. La interpetaciones de γ_3 y $(\beta\gamma)_{13}$ por sí solos fueron presentadas en el tratamiento anterior.

Tratamiento 7: $A_{i1} = 1$ y $R_{ij} = 0;\ j = 1,2,3.$ $Y_i = (\beta_0 + \alpha_1) + (\beta_1 + (\alpha\beta)_{11}) X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$

Este tratamiento corresponde al caso donde un hospital de la Región 4 está afiliado a una escuela de medicina, así que como se mencionó anteriormente α_1 y $(\alpha\beta)_{11}$ en este caso correponden al incremento en intercepto y pendiente respectivamente que tiene la recta en cuestión comparada con la recta del tratamiento de referencia donde un hospital ubicado en la región 4 no está afiliado a una escuela de medicina, es decir que $(\alpha\beta)_{11}$ es la diferencia en el efecto de medio de que tiene el riesgo de infección sobre la longitud de permanencia promedio de un hospital de la región 4 afiliado a la escuela de medicina respecto a un hospital de la misma región no afiliado.

Tratamiento 8: $A_{i1} = 0$ y $R_{ij} = 0;\ j = 1,2,3.$ $Y_i = \beta_0 + \beta_1 X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$

Este es el tratamiento de referencia en el cual β_0 corresponde al intercepto de la recta, es decir, a la respuesta media cuando $X_i = 0$, el hospital no está afiliado a una escuela de medicina y se ubica en la región 4. Por su parte β_1 (pendiente de la recta asociada al taratmiento de referencia) corresponde al cambio medio en la respuesta Y, cuando incrementa en una unidad el riesgo de infección dado que el hospital se encuentra ubicado en la región 4 y sin afiliación a escuela de medicina.

Ajuste del modelo.

A continuación, se presenta la tabla de parámetros estimados resultado del ajuste del MRLM propuesto

Tabla 18: Tabla de parámetros estimados MRLM con predictoras cualitativas.

Parámetros	Estimación	Error Std	T_0	$P(\lvert t_{68}\rvert > \lvert T_0\rvert)$
β_0	7.72639	1.59402	4.847	7.6×10^{-6}
β_1	0.09065	0.37030	0.245	0.8074
α_1	2.02939	1.54892	1.310	0.1945
γ_1	-1.69777	1.85087	-0.917	0.3622
γ_2	-0.14059	1.77553	-0.079	0.9371
γ_3	-0.34146	1.71516	-0.199	0.8428
$(\alpha\beta)_{11}$	-0.29212	0.29961	-0.975	0.3330
$(\beta\gamma)_{11}$	0.88155	0.42229	2.088	0.0406
$(\beta\gamma)_{12}$	0.38985	0.40808	0.955	0.3428
$(\beta\gamma)_{13}$	0.38351	0.39889	0.961	0.3397
$\sqrt{MSE} =$ 1.009	$R^2 = 0.5428$, $R^2_{adj} = 0.4823$			

Así la ecuación general ajustada del modelo es:

$$\widehat{Y}_i = 7.72639 + 0.09065X_i + 2.02939A_{i1} - 1.69777R_{i1} - 0.14059R_{i2} - 0.34146R_{i3} - 0.29212X_iA_{i1} + 0.88155X_iR_{i1} + 0.38985X_iR_{i2} + 0.38351X_iR_{i3}$$

Y las ecuaciones ajustadas de las ocho posibles combinaciones de los niveles de las variables categóricas son:

Tabla 19: Resumen modelos y ecuaciones ajustadas para cada tratamiento de Región y AEM.

Tratamiento	Modelo	Ecuación ajustada
$A_{i1} = 1$ y $R_{i1} = 1$	$Y_i = (\beta_0 + \alpha_1 + \gamma_1) + (\beta_1 + (\alpha\beta)_{11} + (\beta\gamma)_{11})X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$	$\hat{Y}_i = 8.05801 + 0.68008X_i$
$A_{i1} = 0$ y $R_{i1} = 1$	$Y_i = (\beta_0 + \gamma_1) + (\beta_1 + (\beta\gamma)_{11})X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$	$\hat{Y}_i = 6.02862 + 0.9722X_i$
$A_{i1} = 1$ y $R_{i2} = 1$	$Y_i = (\beta_0 + \alpha_1 + \gamma_2) + (\beta_1 + (\alpha\beta)_{11} + (\beta\gamma)_{12})X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$	$\hat{Y}_i = 9.61519 + 0.18838X_i$
$A_{i1} = 0$ y $R_{i2} = 1$	$Y_i = (\beta_0 + \gamma_2) + (\beta_1 + (\beta\gamma)_{12})X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$	$\hat{Y}_i = 7.5858 + 0.4805X_i$
$A_{i1} = 1$ y $R_{i3} = 1$	$Y_i = (\beta_0 + \alpha_1 + \gamma_3) + (\beta_1 + (\alpha\beta)_{11} + (\beta\gamma)_{13})X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$	$\hat{Y}_i = 9.41432 + 0.18204X_i$
$A_{i1} = 0$ y $R_{i3} = 1$	$Y_i = (\beta_0 + \gamma_3) + (\beta_1 + (\beta\gamma)_{13})X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$	$\hat{Y}_i = 7.38493 + 0.47416X_i$
$A_{i1} = 1$ y $R_{ij} = 0;\ j = 1,2,3.$	$Y_i = (\beta_0 + \alpha_1) + (\beta_1 + (\alpha\beta)_{11})X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$	$\hat{Y}_i = 9.75578 - 0.20147X_i$
$A_{i1} = 0$ y $R_{ij} = 0;\ j = 1,2,3.$	$Y_i = \beta_0 + \beta_1 X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$	$\hat{Y}_i = 7.72639 + 0.09065X_i$

Interpretaciones de los parámetros estimados.

$\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}$ se estima que la media de la longitud de permanencia promedio de los pacientes en hospitales de la región 4 que no están afiliados a escuela de medicina cuando el riesgo de infección $X_i = 0$ es de 7.7263.

$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$ se estima que la por cada incremento unitario en el riesgo de infección la media de la longitud de permanencia promedio aumenta 0.09065 para hospitales de la región 4 no afiliados a la escuela de medicina.

$\widehat{\boldsymbol{\alpha}}_{\boldsymbol{1}}$ se estima que el incremeto en los interceptos de las rectas que están asociadas a hospitales que cuentan con afiliación a una escuela de medicina es del 2.02939 respecto al intercepto del tratamiento de referencia. Este incremento es independiente de la región en la que se encuentre el hospital, solo proviene de la presencia de la característica relacionada com la afiliación.

$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{1}}$ se estima que el decremento en los interceptos de las rectas que están asociadas a hospitales que se encuentran en la región 1 es de 1.69777 respecto al intercepto del tratamiento de referencia.

$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{2}}$ se estima que el decremento en los interceptos de las rectas que están asociadas a hospitales que se encuentran en la región 2 es de 0.14059 respecto al intercepto del tratamiento de referencia.

$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{3}}$ se estima que el decremento en los interceptos de las rectas que están asociadas a hospitales que se encuentran en la región 3 es de 0.34146 respecto al intercepto del tratamiento de referencia.

Por lo tanto, los términos $(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{1}} + \widehat{\boldsymbol{\gamma}}_{\boldsymbol{j}}); \boldsymbol{j = 1,2,3}$. Representan la diferencia total estimada de los interceptos de las rectas en cuestión respecto al tratamiento de referencia cuando los hospitales de la región $\boldsymbol{j = 1,2,3}$ están afiliados a una escuela de medicina.

$\widehat{(\boldsymbol{\alpha\beta})}_{\boldsymbol{11}}$ se estima que el incremento en el efecto del riesgo de infección sobre la media de la longitud de permanencia promedio debido a la afiliación de un hospital a una escuela de medicina es de -0.29212 respecto al tratamiento de referencia.

$\widehat{(\boldsymbol{\beta\gamma})}_{\boldsymbol{11}}$ se estima que el incremento en el efecto del riesgo de infección sobre la media de la longitud de permanencia promedio debido a que un hospital se ubica en la región 1 es de 0.88155 respecto al tratamiento de referencia.

$\widehat{(\boldsymbol{\beta\gamma})}_{\boldsymbol{12}}$ se estima que el incremento en el efecto del riesgo de infección sobre la media de la longitud de permanencia promedio debido a que un hospital se ubica en la región 2 es de 0.38985 respecto al tratamiento de referencia.

$\widehat{(\boldsymbol{\beta\gamma})}_{\boldsymbol{13}}$ se estima que el incremento en el efecto del riesgo de infección sobre la media de la longitud de permanencia promedio debido a que un hospital se ubica en la región 3 es de 0.38351respecto al tratamiento de referencia.

De tal forma que los términos $(\widehat{(\boldsymbol{\alpha\beta})}_{\boldsymbol{11}} + \widehat{(\boldsymbol{\beta\gamma})}_{\boldsymbol{1j}}); \boldsymbol{j = 1,2,3}$. Representan la diferencia total estimada de las pendientes (los efectos) de las respectivas rectas, respecto al tratamiento de referencia cuando los hospitales de la región $\boldsymbol{j = 1,2,3}$ están afiliados a una escuela de medicina.En los casos donde no aparecen $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{1}}$ **y** $\widehat{(\boldsymbol{\alpha\beta})}_{\boldsymbol{11}}$ no hay afiliación de los hospitales a una escuela de medicina y los cambios en intercepto y pendiente solo se deben a la región donde se ubique el hospital.

Test ANOVA del modelo general.

Para verificar la significancia del modelo general se plantea el siguiente juego de hipótesis:

$$H_0: \beta_1 = \alpha_1 = \gamma_1 = \gamma_2 = \gamma_3 = (\alpha\beta)_{11} = (\beta\gamma)_{11} = (\beta\gamma)_{12} = (\beta\gamma)_{13} = 0\ vs.\ H_1: Al\ menos\ uno\ de\ los\ parámetros\ reportados\ en\ H_0\ es\ diferente\ de\ 0$$

El estadístico de prueba en este test está dado por $F_0 = \frac{SSR/9}{SSE/68} = \frac{MSR}{MSE}$, $F_0 \sim f_{9,68}$ Bajo H_0 y los supuestos del MRLM

Tabla 20: Tabla ANOVA del modelo general

Fuente	Sum Sq	Df	Mean Sq	F_0	$P(f_{9,68} > F_0)$
Modelo	97.564	9	10.8404	8.9702	8.778×10^{-9}
Error	82.178	68	1.2085		
Total	179.743	77	2.3343		

El valor observado del estadístico de prueba es $F_0 = 8.9702$. El criterio de rechazo dice que la hipótesis nula se rechaza para valores grandes de F_0, usando el criterio de valor p, se rechaza H_0 pues este toma un valor muy pequeño, en particular se tiene que $VP = P(f_{9,68} > 8.9702) = 8.778 \times 10^{-9}$ Lo anterior permite concluir que la regresión es significativa. Es decir, al menos una de las variables predictoras contribuye significativamente a explicar la longitud media de permanencia de los pacientes en el hospital.

Coeficiente de determinación muestral R^2

El coeficiente de determinación muestral esta dado por $R^2 = \frac{SSR}{SST}$ y es un estadístico que permite cuantificar la proporción de variabilidad total observada que logra ser explicada por el modelo propuesto. En este caso toma un valor de $R^2 = \frac{97.564}{97.564+82.178} = \frac{97.564}{179.743} = 0.5428$ y se dice que el 54.28% de la variabilidad total observada en la longitud de permanencia promedio de los pacientes en el hospital logra ser explicada por el MRLM propuesto, de tal modo que una proporción importante de la variabilidad no está siendo representada por el modelo. Este estadístico no mide la bondad del ajuste, para tal fin puede usarse el R^2_{adj} que, aunque no tiene una interpretación tan práctica como el R^2, en este caso $R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{1.2085}{2.3343} = 0.4823$, esta medida se espera cercana a 1 cuando hay buen ajuste.

Análisis de residuales.

En la **Figura 6**, en particular en la gráfica de residuales estudentizados vs. predictor cuantitativo se puede observar que, en primer lugar, los residuales no muestran evidencia fuerte en contra del supuesto de media cero de los errores del modelo ($E[E_i] = 0 \; \forall i$), pues los residuales se encuentran dispersos de manera uniforme entre -2 y 2 alrededor del cero salvo por unos pocos cuya magnitud es mayor a 2, estos son identificados como valores atípicos o outliers, sin embargo al analizar los residuales vs. los niveles de las variables cualitativas, en particular residuales vs. región si se observa evidencia en contra del supuesto de media 0 de los errores, al separar los residuales según la región, se hace más clara la tendencia de los residuos en algunos niveles a no centrarse alredor de 0. Las gráficas de residuales estudentizados vs. predictores cualitativos no muestran comportamientos en contra del supuesto del varianza constante de los errores, se puede pensar que no hay muchos cambios en su dispersión según los niveles de las variables cualitativas, sin embargo es importante resaltar que, por ejemplo, en el caso de la variable AEM hay mucha diferencia en la cantidad de observaciones registradas en sus niveles, lo que dificulta juzgar con certeza la varianza de los errores en este caso. El comportamiento de los residuales en el gráfico vs. Región podría interpretarse como evidencia de carencia de ajuste pues los residuos no están centrados alrededor del 0 y muestran a lo largo de los niveles una curvatura en forma de U.

Validación del supuesto de normalidad de los errores del modelo.

Para validar la normalidad de los errores del modelo se van a contrastar las siguientes hipótesis: $H_0: E_i \sim Normal \; vs. H_1: E_i \nsim Normal$

En la **Figura 7**, se enseña el gráfico de probabilidad normal construido con los residuales estudentizados y el resultado de la prueba de Shapiro-Wilk. Esta prueba reporta un valor p de 0.4 que conduce a no rechazar la hipótesis nula. En el gráfico de probabilidad normal en general se observa concordancia de los cuantiles teóricos de la distribución normal y los cuantiles de los residuales estudentizados, un poco de discrepancia en los extremos, sobretodo en el inferior, pero no es claro ningun tipo de tendencia con concavidad que podría no ser coherente con el supuesto de normalidad. Se observa un residual muy apartado de la recta, pero este corresponde a una observación atípica. En este caso, se puede decir que hay coherencia entre lo observado a nivel gráfico y el resultado de la prueba, por lo tanto, no hay razones claras para dudar del supuesto de normalidad de los errores y se concluye que $E_i \sim Normal$.

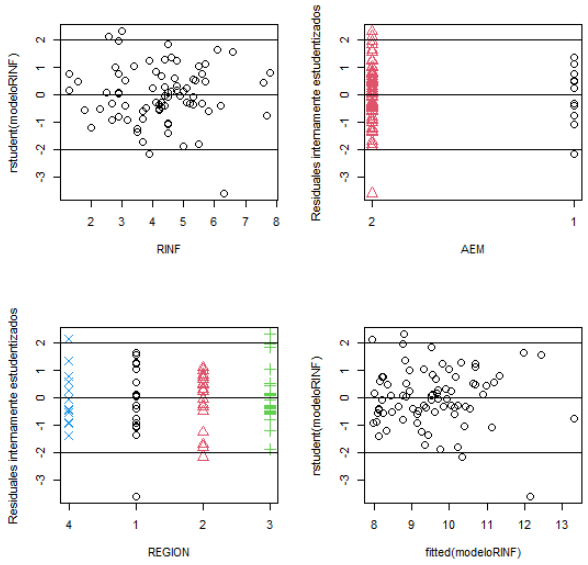

Figura 6: Gráficos de residuales estudentizados vs. valores ajustados y variables predictoras.

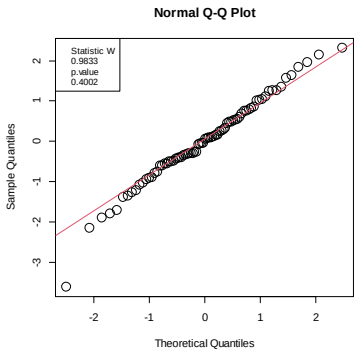

Figura 7: Gráfico de probabilidad normal con residuales estudentizados.

5. Prueba para determinar si la relación lineal de DPERM vs. RINF difiere según AEM y/o Región.

La prueba de interés es equivalente a probar si las rectas asociadas a cada combinación de los niveles de las variables cualitativas son coincidentes, es decir, determinar si el intercepto y la pendiente de las 8 rectas resultantes son iguales, por lo tanto se requiere que:

En los interceptos: $\beta_0 + \alpha_1 + \gamma_1 = \beta_0 + \gamma_1 = \beta_0 + \alpha_1 + \gamma_2 = \beta_0 + \gamma_2 = \beta_0 + \alpha_1 + \gamma_3 = \beta_0 + \gamma_3 = \beta_0 + \alpha_1 = \beta_0 \Longrightarrow \alpha_1 = \gamma_1 = \gamma_2 = \gamma_3 = 0$ (5)

En las pendientes: $\beta_1 + \alpha\beta_{11} + \beta\gamma_{11} = \beta_1 + \beta\gamma_{11} = \beta_1 + \alpha\beta_{11} + \beta\gamma_{12} = \beta_1 + \beta\gamma_{12} = \beta_1 + \alpha\beta_{11} + \beta\gamma_{13} = \beta_1 + \beta\gamma_{13} = \beta_1 + \alpha\beta_{11} = \beta_1 \Longrightarrow$
$\alpha\beta_{11} = \beta\gamma_{11} = \beta\gamma_{12} = \beta\gamma_{13} = 0$ (6)

Así, el juego de hipótesis a analizar es

$$\boldsymbol{H_0}: \alpha_1 = \gamma_1 = \gamma_2 = \gamma_3 = \alpha\beta_{11} = \beta\gamma_{11} = \beta\gamma_{12} = \beta\gamma_{13} = 0 \; vs. \boldsymbol{H_1}: Al\ menos\ uno\ de \text{ los } par\acute{a}metros \text{ en } H_0 \; es\ diferente\ de\ 0$$

Por lo tanto, el modelo reducido en este caso es MR: $Y_i = \beta_0 + \beta_1 X_i + E_i, E_i \sim IID\ N(0, \sigma^2)$ y el modelo completo (MF) corresponde al modelo de la **ecuación 4.**

El estadistico de prueba está dado por

$$F_0 = \frac{SSR(A_1,R_1,R_2,R_3,XA_1,XR_1,XR_2,XR_3|X)/8}{MSE(MF)} = \frac{[SSE(MR)-SSE(MF)]/8}{MSE(MF)} \sim f_{8,68} \; Bajo\ H_0\ y\ los\ supuestos\ de\ los\ errores\ del\ modelo.$$

Los grados de libertad de la ss extra estan dados por: $gl[SSR(A_1,R_1,R_2,R_3,XA_1,XR_1,XR_2,XR_3|X)] = gl[SSE(MR)] - gl[SSE(MF)] = 76 - 68 = 8$

Tabla 21: Test lineal general para determinar si las 8 rectas son coincidentes

Fuente	DF SSE	SSE	DF SSR parcial	SSR parcial	F_0	$P(f_{8,68} > F_0)$
MR	76	125.340				
MF	68	82.178	8	43.162	4.4644	0.000219
$\boldsymbol{H_0}: \alpha_1 = \gamma_1 = \gamma_2 = \gamma_3 = \alpha\beta_{11} = \beta\gamma_{11} = \beta\gamma_{12} = \beta\gamma_{13} = 0$						

El valor observado del estadístico de prueba es $F_0 = \frac{[125.340-82.178]/8}{82.178/68} = 4.4644$. El criterio de rechazo a un nivel de significancia α consiste en rechazar $\boldsymbol{H_0}$ cuando $F_0 > f_{\alpha,8,68}$ o usando el criterio de valor p, se rechazará $\boldsymbol{H_0}$ cuando VP sea muy pequeño, en particular menor a α. Según este último criterio se rechaza la hipótesis nula pues $VP = P(f_{8,68} > 4.4644) = 0.000219$ y se concluye que la relación lineal entre DPERM y RINF si difiere según la afiliación a escuelas de medicina y/o la región en que se ubican los hospitales.

6. Prueba para determinar si existen diferencia entre las ordenadas de las 8 rectas en cuestión en el origen.

Esta prueba es equivalente a probar si las rectas asociadas a cada combinación de los niveles de las variables cualitativas difieren en su intercepto, es decir, se quiere ver si lo relacionado en (5) es cierto, por lo tanto se proponen las siguientes hipótesis:

$$\boldsymbol{H_0}: \alpha_1 = \gamma_1 = \gamma_2 = \gamma_3 = 0 \; vs. \boldsymbol{H_1}: Al\ menos\ uno\ de \text{ los } par\acute{a}metros \text{ en } H_0 \; es\ diferente\ de\ 0.$$

Por lo tanto, el modelo reducido en este caso es MR: $Y_i = \beta_0 + \beta_1 X_i + (\alpha\beta)_{11} X_i A_{i1} + (\beta\gamma)_{11} X_i R_{i1} + (\beta\gamma)_{12} X_i R_{i2} + (\beta\gamma)_{13} X_i R_{i3} + E_i, E_i \sim IID\ N(0, \sigma^2)$ y el modelo completo (MF) corresponde al modelo de la **ecuación 4.**

El estadistico de prueba está dado por

$$F_0 = \frac{SSR(A_1,R_1,R_2,R_3|X,XA_1,XR_1,XR_2,XR_3)/4}{MSE(MF)} = \frac{[SSE(MR)-SSE(MF)]/4}{MSE(MF)} \sim f_{4,68} \; Bajo\ H_0\ y\ los\ supuestos\ de\ los\ errores\ del\ modelo.$$

Los grados de libertad de la ss extra estan dados por: $gl[SSR(A_1,R_1,R_2,R_3|X,XA_1,XR_1,XR_2,XR_3)] = gl[SSE(MR)] - gl[SSE(MF)] = 72 - 68 = 4$

En este caso el valor observado del estadístico de prueba es $F_0 = \frac{[86.257-82.178]/4}{82.178/68} = 0.8437$. H_0 se rechaza cuando $F_0 > f_{\alpha,4,68}$ a un nivel de significancia α dado o según criterio de valor p, se rechaza H_0 si VP muy pequeño, en particular menor a α. Usando el criterio de valor p no es posible rechazar H_0 pues $VP = P(f_{4,68} > 0.8437) = 0.5024$ por lo tanto la probabilidad de equivocarse al rechazar la hipótesis nula es alta y se concluye que la muestra respalda que no existen diferencias entre los interceptos de las rectas asociadas a las 8 combinaciones de los niveles de las variables cualitativas AEM y Región.

Tabla 22: Test lineal generalpara determinar si los interceptos de las 8 rectas son diferentes

Fuente	DF SSE	SSE	DF SSR parcial	SSR parcial	F_0	$P(f_{4,68} > F_0)$
MR	72	86.257				
MF	68	82.178	4	4.0783	0.8437	0.5024
$\boldsymbol{H_0}: \alpha_1 = \gamma_1 = \gamma_2 = \gamma_3 = 0$, $\boldsymbol{H_1}: Al\ menos\ uno\ de$ los $par\acute{a}metros$ en H_0 $es\ diferente\ de\ 0$.						

7. Prueba para determinar existen diferencias en las pendientes de las rectas 8 rectas en cuestión.

Esto es equivalente a probar si el efecto medio de RINF sobre DPERM no varía según las combinaciones de los niveles de AEM y Región, es decir, si sucede lo relacionado en (6). Para esto se formulan las siguientes hipótesis:

$$\boldsymbol{H_0}: (\alpha\beta)_{11} = (\beta\gamma)_{11} = (\beta\gamma)_{12} = (\beta\gamma)_{13} = 0 \; vs. \boldsymbol{H_1}: Al\ menos\ uno\ de\ estos\ par\acute{a}metros\ es\ diferente\ de\ 0$$

El modelo reducido en este caso es MR: $Y_i = \beta_0 + \beta_1 X_i + \alpha_1 A_{i1} + \gamma_1 R_{i1} + \gamma_2 R_{i2} + \gamma_3 R_{i3} + E_i, E_i \sim IID\ N(0, \sigma^2)$ y el modelo completo (MF) es el modelo definido en la ecuación 4.

El estadistico de prueba está dado por

$$F_0 = \frac{\text{SSR}(\text{XA}_1, \text{XR}_1, \text{XR}_2, \text{XR}_3|\text{X}, \text{A}_1, \text{R}_1, \text{R}_2, \text{R}_3)/4}{MSE(MF)} = \frac{[SSE(MR) - SSE(MF)]/4}{MSE(MF)} \sim f_{4,68} \; Bajo \; H_0 \; y \; los \; supuestos \; de \; los \; errores \; del \; modelo.$$

Los grados de libertad de la ss extra estan dados por: $gl[\text{SSR}(\text{XA}_1, \text{XR}_1, \text{XR}_2, \text{XR}_3|\text{X}, \text{A}_1, \text{R}_1, \text{R}_2, \text{R}_3)] = gl[SSE(MR)] - gl[SSE(MF)] = 72 - 68 = 4$

Tabla 23: Test lineal generalpara determinar si las pendientes de las 8 rectas son diferentes

Fuente	DF SSE	SSE	DF SSR parcial	SSR parcial	F_0	$P(f_{4,68} > F_0)$
MR	72	90.331				
MF	68	82.178	4	8.1529	1.6866	0.1632
$\boldsymbol{H_0}: (\alpha\beta)_{11} = (\beta\gamma)_{11} = (\beta\gamma)_{12} = (\beta\gamma)_{13} = 0$ $\boldsymbol{H_1}: Al \; menos \; uno \; de \; parámetros \; es \; diferente \; de \; 0$						

De tal forma que el valor observado del estadístico de prueba es $F_0 = \frac{[90.331 - 82.178]/4}{82.178/68}$ =1.6866. Se rechazará H_0 cuando el valor del estadístico de prueba supere al cuantil superior α de una $f_{4,68}$ o cuando el valor p de la prueba sea pequeño. En este caso, usando el criterio de valor p se tiene que $VP = P(f_{4,68} > 1.6866) = 0.1632$, así que no se rechaza la hipótesis nula y se concluye es más razonable pensar, según la información muestral, que el efecto medio del riesgo de infección sobre la longitud de me permanencia promedio es igual en todas las combinaciones de los niveles de las variables AEM y Región, es decir que las pendientes de las rectas asociadas a cada combinación no difieren.

8. Significancia de las interaciones dobles.

Significancia de la interacción doble de RINF con REGION.

Esto es equivalente a probar si la variable Región no modifica el efecto que tiene la variable RINF sobre la media de la longitud promedio de permanencia, en presencia de AEM. Para esta prueba se contrastan las siguientes hipotésis:

$$\boldsymbol{H_0}: (\beta\gamma)_{11} = (\beta\gamma)_{12} = (\beta\gamma)_{13} = 0 \; vs. \; \boldsymbol{H_1}: Al \; menos \; uno \; de \; \text{los} \; parámetros \; \text{en} \; \text{H}_0 \; es \; diferente \; de \; 0$$

El modelo reducido en este caso corresponde a MR: $Y_i = \beta_0 + \beta_1 X_i + \alpha_1 A_{i1} + \gamma_1 R_{i1} + \gamma_2 R_{i2} + \gamma_3 R_{i3} + (\alpha\beta)_{11} X_i A_{i1} + E_i, E_i \sim IID \; N(0, \sigma^2)$ y el modelo completo es el definido en la ecuación 4.

Se el estadístico de prueba $F_0 = \frac{\text{SSR}(\text{XR}_1, \text{XR}_2, \text{XR}_3|\text{X}, \text{A}_1, \text{XA}_1, \text{R}_1, \text{R}_2, \text{R}_3)/3}{MSE(MF)} = \frac{[SSE(MR) - SSE(MF)]/3}{MSE(MF)} \sim f_{3,68} \; Bajo \; H_0 \; y \; los \; supuestos \; de \; los \; errores \; del \; modelo.$

Los grados de libertad de la ss extra estan dados por: $gl[\text{SSR}(\text{XR}_1, \text{XR}_2, \text{XR}_3|\text{X}, \text{A}_1, \text{XA}_1, \text{R}_1, \text{R}_2, \text{R}_3)] = gl[SSE(MR)] - gl[SSE(MF)] = 71 - 68 = 3$

Tabla 24: Test lineal general para probar la significancia de la interacción doble de RINF con REGION .

Fuente	DF SSE	SSE	DF SSR parcial	SSR parcial	F_0	$P(f_{3,68} > F_0)$
MR	71	90.127				
MF	68	82.178	3	7.9487	2.1924	0.09687
$\boldsymbol{H_0}: (\beta\gamma)_{11} = (\beta\gamma)_{12} = (\beta\gamma)_{13} = 0$						

Tomando los resultados de la **tabla 24** el estadístico de prueba es igual a $F_0 = \frac{[90.127 - 82.178]/3}{82.178/68}$=2.1924. El criterio de rechazo de es $F_0 > f_{\alpha,4,68}$ H_0 o si el valor p de la prueba es pequeño. Usando como referencia el valor p para concluir sobre esta prueba, se tiene que $VP = P(f_{3,68} > 2.1924) = 0.09687$, así que no se rechaza la hipótesis nula y la conclusión consecuente es que, de acuerdo con la información muestral, la región en la que se ubica el hospital no modifica el efecto del riesgo de infección sobre la media de la permanencia promedio de los pacientes, en presencia de la variable que indica el estatus de afiliación a la escuela de medicina del mismo.

Significancia de la interacción doble de RINF con AEM

Otra manera de verlo, es probar si la variable AEM no modifica el efecto que tiene la variable RINF sobre la media de DPERM, en presencia de REGION. Para esta prueba se plantean las siguientes hipotésis: $\boldsymbol{H_0}: (\alpha\beta)_{11} = 0 \; vs. \; \boldsymbol{H_1}: (\alpha\beta)_{11} \neq 0$

El modelo reducido en este caso corresponde a MR: $Y_i = Y_i = \beta_0 + \beta_1 X_i + \alpha_1 A_{i1} + \gamma_1 R_{i1} + \gamma_2 R_{i2} + \gamma_3 R_{i3} + (\beta\gamma)_{11} X_i R_{i1} + (\beta\gamma)_{12} X_i R_{i2} + (\beta\gamma)_{13} X_i R_{i3} + E_i, E_i \sim IID \; N(0, \sigma^2$ y el modelo completo es el definido en la ecuación 4.

Se define el estadístico de prueba $F_0 = \frac{\text{SSR}(\text{XA}_1|\text{X}, \text{A}_1, \text{R}_1, \text{R}_2, \text{R}_3, \text{XR}_1, \text{XR}_2, \text{XR}_3)/1}{MSE(MF)} = \frac{[SSE(MR) - SSE(MF)]/1}{MSE(MF)} \sim f_{1,68} \; Bajo \; H_0 \; y \; los \; supuestos \; de \; los \; errores \; del \; modelo.$

Los grados de libertad de la ss extra estan dados por: $gl[\text{SSR}(\text{XA}_1|\text{X}, \text{A}_1, \text{R}_1, \text{R}_2, \text{R}_3, \text{XR}_1, \text{XR}_2, \text{XR}_3)] = gl[SSE(MR)] - gl[SSE(MF)] = 69 - 68 = 1$

Tabla 25: Test lineal general para probar la significancia de la interacción doble de RINF con AEM

Fuente	DF SSE	SSE	DF SSR parcial	SSR parcial	F_0	$P(f_{1,68} > F_0)$
MR	69	83.327				
MF	68	82.178	1	1.1488	0.9506	0.333
$\boldsymbol{H_0}: (\alpha\beta)_{11} = 0$						

El estadístico de prueba es numéricamente igual a $F_0 = \frac{[83.327 - 82.178]/1}{82.178/68}$=0.9506. El criterio de rechazo de es $F_0 > f_{\alpha,1,68}$ H_0 o si el valor p de la prueba es pequeño. Empleando el valor p para concluir respecto a la prueba, se tiene que $VP = P(f_{1,68} > 0.9506) = 0.333$, así que no se rechaza la hipótesis nula y se concluye que, de acuerdo con la información muestral, el estatus de afiliación a la escuela de medicina del hospital no modifica el efecto del riesgo de infección sobre la media de la permanencia promedio de los pacientes en presencia de la variable REGION.

Modelo reducido de DPERM vs. RINF en presencia de AEM y REGIÓN.

Considere ahora el siguiente modelo reducido $Y_i = \beta_0 + \beta_1 X_i + \alpha_1 A_{i1} + \gamma_1 R_{i1} + \gamma_2 R_{i2} + \gamma_3 R_{i3} + (\beta\gamma)_{11} X_i R_{i1} + E_i; \; E_i \overset{iid}{\sim} N(0, \sigma^2)$(7)

Se supone que la variable AEM afecta en la relación lineal entre DPERM vs. RINF solo en el intercepto del modelo, pero no en la pendiente, ahora, la variable REGIÓN también afecta en la relación lineal entre DPERM vs. RINF en el intercepto y en uno de sus niveles afecta la pendiente, es decir que se supone que el efecto promedio del riesgo de infección sobre la longitud de permanencia promedio cambia si el hospital está en la **REGIÓN 1**, pero no según el estado de afiliación a una escuela de medicina, en tanto que, la media general de Y cambia según los niveles de las dos variables cualitativas.

Tabla 26: Resumen ecuaciones asociadas a los diferentes tratamientos del modelo definido en (7)

Tratamiento		Modelo
Cuando se está afiliado y en la región 1	$(A_1 = 1, R_1 = 1, R_2 = 0, R_3 = 0)$	$Y_i = (\beta_0 + \alpha_1 + \gamma_1) + (\beta_1 + (\beta\gamma)_{11})X_i + E_i;\ E_i \overset{iid}{\sim} N(0, \sigma^2)$
Cuando se está afiliado y en la región 2	$(A_1 = 1, R_1 = 0, R_2 = 1, R_3 = 0)$	$Y_i = (\beta_0 + \alpha_1 + \gamma_2) + \beta_1 X_i + E_i;\ E_i \overset{iid}{\sim} N(0, \sigma^2)$
Cuando se está afiliado y en la región 3	$(A_1 = 1, R_1 = 0, R_2 = 0, R_3 = 1)$	$Y_i = (\beta_0 + \alpha_1 + \gamma_3) + \beta_1 X_i + E_i;\ E_i \overset{iid}{\sim} N(0, \sigma^2)$
Cuando se está afiliado y en la región 4	$(A_1 = 1, R_1 = 0, R_2 = 0, R_3 = 0)$	$Y_i = (\beta_0 + \alpha_1) + \beta_1 X_i + E_i;\ E_i \overset{iid}{\sim} N(0, \sigma^2)$
Cuando no está afiliado y en la región 1	$(A_1 = 0, R_1 = 1, R_2 = 0, R_3 = 0)$	$Y_i = (\beta_0 + \gamma_1) + (\beta_1 + (\beta\gamma)_{11})X_i + E_i;\ E_i \overset{iid}{\sim} N(0, \sigma^2)$
Cuando no está afiliado y en la región 2	$(A_1 = 0, R_1 = 0, R_2 = 1, R_3 = 0)$	$Y_i = (\beta_0 + \gamma_2) + \beta_1 X_i + E_i;\ E_i \overset{iid}{\sim} N(0, \sigma^2)$
Cuando no está afiliado y en la región 3	$(A_1 = 0, R_1 = 0, R_2 = 0, R_3 = 1)$	$Y_i = (\beta_0 + \gamma_3) + \beta_1 X_i + E_i;\ E_i \overset{iid}{\sim} N(0, \sigma^2)$
Cuando no está afiliado y en la región 4	$(A_1 = 0, R_1 = 0, R_2 = 0, R_3 = 0)$	$Y_i = \beta_0 + \beta_1 X_i + E_i;\ E_i \overset{iid}{\sim} N(0, \sigma^2)$

Interpretaciones:

$(\boldsymbol{\beta_0 + \alpha_1 + \gamma_1})$ es la ordenada en el origen o el intercepto de la recta cuando se está afiliado y en la ubicación NE y $[\boldsymbol{\beta_1 + (\beta\gamma)_{11}}]$ es el cambio medio de la Longitud de permanencia por unidad de cambio en la probabilidad promedio estimada de adquirir infección, es decir, la pendiente de la recta cuando se está afiliado y se ubica en la región 1.

$(\boldsymbol{\beta_0 + \alpha_1 + \gamma_2})$ es la ordenada en el origen o el intercepto de la recta cuando se está afiliado y en la ubicación NC y $\boldsymbol{\beta_1}$ es el cambio medio de la Longitud de permanencia por unidad de cambio en la probabilidad promedio estimada de adquirir infección cuando se está afiliado y se ubica en la región 2.

$(\boldsymbol{\beta_0 + \alpha_1 + \gamma_3})$ es la ordenada en el origen o el intercepto de la recta cuando se está afiliado y en la ubicación S y $\boldsymbol{\beta_1}$ es el cambio medio de la Longitud de permanencia por unidad de cambio en la probabilidad promedio estimada de adquirir infección cuando se está afiliado y se ubica en la región 3.

$(\boldsymbol{\beta_0 + \alpha_1})$ es la ordenada en el origen o el intercepto de la recta cuando se está afiliado y en la ubicación W y $\boldsymbol{\beta_1}$ es el cambio medio de la Longitud de permanencia por unidad de cambio en la probabilidad promedio estimada de adquirir infección cuando se está afiliado y se ubica en la región 4.

$(\boldsymbol{\beta_0 + \gamma_1})$ es la ordenada en el origen o el intercepto de la recta cuando no está afiliado y en la ubicación NE y $(\boldsymbol{\beta_1 + (\beta\gamma)_{11}})$ es el cambio medio de la Longitud de permanencia por unidad de cambio en la probabilidad promedio estimada de adquirir infección cuando no está afiliado y se ubica en la región 1.

$(\boldsymbol{\beta_0 + \gamma_2})$ es la ordenada en el origen o el intercepto de la recta cuando no está afiliado y en la ubicación NC, $\boldsymbol{\beta_1}$ es el cambio medio de la Longitud de permanencia por unidad de cambio en la probabilidad promedio estimada de adquirir infección cuando no está afiliado y se ubica en la región 2.

$(\boldsymbol{\beta_0 + \gamma_3})$ es la ordenada en el origen o el intercepto de la recta cuando no está afiliado y en la ubicación S, $\boldsymbol{\beta_1}$ es el cambio medio de la Longitud de permanencia por unidad de cambio en la probabilidad promedio estimada de adquirir infección cuando no está afiliado y se ubica en la región 3.

$\boldsymbol{\beta_0}$ es la ordenada en el origen o el intercepto de la recta cuando no está afiliado y en la ubicación W, $\boldsymbol{\beta_1}$ es el cambio medio de la Longitud de permanencia por unidad de cambio en la probabilidad promedio estimada de adquirir infección cuando no está afiliado y se ubica en la región 4. (tratamiento de referencia).

2. Ajuste del modelo

Tabla 27: Parámetros estimados del modelo definido en (7)

Parámetros	Estimación	Error Std	T_0	$P(\lvert t_{68}\rvert > \lvert T_0\rvert)$
β_0	6.37272	0.54244	11.748	2×10^{-16}
β_1	0.41940	0.10644	3.940	0.000188
α_1	0.57772	0.37210	1.553	0.124961
γ_1	-0.04572	1.04382	-0.044	0.965187
γ_2	1.49731	0.39210	3.819	0.000284
γ_3	1.21138	0.38218	3.170	0.002252
$(\beta\gamma)_{11}$	0.47924	0.21427	2.237	0.028453
$\sqrt{MSE} = 1.09$ $R^2 = 0.5305$, $R^2_{adj} = 0.4908$				

Ecuación ajustada: $\hat{Y}_i = 6.3727 + 0.4194X_i + 0.5777A_{i1} - 0.04572R_{i1} + 1.49731R_{i2} + 1.21138R_{i3} + 0.47924X_iR_{i1}$;

A continuación, se hará el test Anova para evaluar la significancia de la regresión, para ello se plantean las siguientes hipótesis:

$\boldsymbol{H_0}$: $\beta_1 = \alpha_1 = \gamma_1 = \gamma_2 = \gamma_3 = (\beta\gamma)_{11} = 0$ $\quad \boldsymbol{H_1}$: Algún $\beta_1, \alpha_1, \gamma_1, \gamma_2, \gamma_3, (\beta\gamma)_{11}$ es diferente de cero

La expresión del estadístico es $F_0 = \frac{SSR/6}{SSE/71} = \frac{MSR}{MSE} \sim f_{6,71}$ Bajo H_0 y los supuestos del MRLM.

Tabla 28: Tabla ANOVA del modelo definido en (7)

Fuente	Sum Sq	Df	Mean Sq	F_0	$P(f_{6,71} > F_0)$
Modelo	95.357	6	15.8928	13.372	4.448×10^{-10}
Error	84.385	71	1.1885		
Total	179.743	77	2.3343		

El valor observado del estadístico de prueba es $F_0 = \frac{15.8928}{1.1885} = 13.372$

y el valor p de la prueba es $P(f_{6,71} > F_0) = 4.448 * 10^{-10}$ por tanto se rechaza H_0 y se concluye que el modelo es significativo.

De la **tabla 28** se tiene que el valor de R^2 es 0.5305, este se calcula en este caso como $R^2 = \frac{SSR}{SST} = \frac{95.357}{179.743}$, este estadístico cuantifica la proporción de variabilidad total observada en la respuesta, que es explicada por el modelo propuesto, es decir, por la asociación lineal de Y con el conjunto de variables $X_i, A_{i1}, R_{i1}, R_{i2}, R_{i3}, X_iR_{i1}$, por lo tanto este modelo logra explicar un 53.05% de la variabilidad total

observada de la respuesta, este valor es un poco menor al del R^2 del modelo con todas las interacciones ajustado anteriormente, pero el cambio tampoco es drástico.

3. Análisis de residuales.

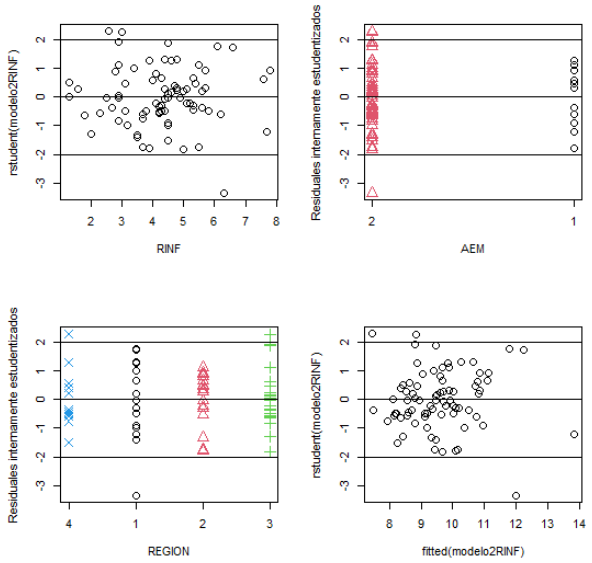

Figura 8: Residuales estudentizados vs. valores ajustados y predictoras modelo definido en (7)

De la **Figura 8** se observan algunos puntos atípicos pues la magnitud de sus residuales supera en valor absoluto la cota de 2 o 3, no es clara la detección de puntos de balanceo. En general no hay evidencia fuerte en contra del supuesto de varianza constante de los errores en las gráficas de residuales vs. valores ajustados y vs. la predictora cuantitativa. De los gráficos de de residuales vs. predictoras cualitativas se observa nuevamente lo mencionado en el análisis de la **Figura 6**, no hay suficiente información en una de las categorías de la variable AEM (muy pocos datos sobre hospitales afiliados) para juzgar con certeza el supuesto de varianza constante entre niveles de las variables categóricas. Por su parte en la gráfica de residuales vs. Región no hay evidencia fuerte en contra del supuesto de varianza constante pero sí hay un comportamiento en los residuales que conduce a pensar que no están centrados en 0, pues para algunos niveles de la variable Región los residuales tienen tendencia a localizarse por encima o por debajo de 0 y no centrase alrededor de él, también se evidencia de nuevo un poco de carencia de ajuste en esta gráfica vs. Región.

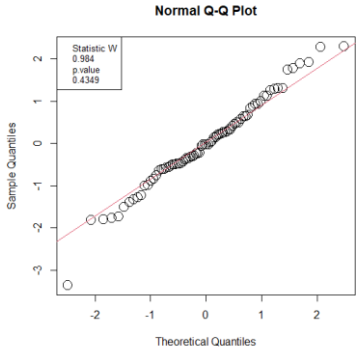

Figura 9: Normal Q-Q plot con Test de normalidad sobre residuales estudentizados.

Para validar la normalidad de los errores del modelo se van a contrastar las siguientes hipótesis:

$$H_0: E_i \sim Normal \; vs. H_1: E_i \nsim Normal$$

Se nota la presencia del dato atípico otra vez, pero a pesar de esto, se aprecia que los cuantiles observados de los residuos se ajustan a los cuantiles teóricos de la distribución normal en la mayoría de puntos, un poco de discrepancia en los extremos, sin embargo puede observarse en general un buen comportamiento en favor de la normalidad de los errores del modelo, además el valor p de la prueba de Shapiro-Wilk es de 0.4349, por consiguiente no hay evidencia contundentes en contra del supuesto de distribución normal en los errores.

4. Comparación con el modelo inicial

Evaluando lo obtenido en los ajustes de ambos modelos, se observa que no hay diferencias sustanciales entre ellos en cuanto a estadísticos que permiten juzgar la calidad del modelo, sin embargo, si hay bastante diferencia en temas de parsimonia. El modelo con los términos principales y todas las interacciones de X y las variables cualitativas presenta un R^2 de 0.5428 y un R^2_{adj} de 0.4823, por su parte el modelo con los términos principales y únicamente la interacción de X con R_1 reporta un R^2 y R^2_{adj} de 0.5305 y 0.4908 respectivamente, así que no se observan diferencias significativas en estos estadísticos de un modelo a otro. Los MSE de los modelos con ecuación (4) y (7) son respetivamente 1.2085 y 1.1885, también sin diferencias importantes. Los análisis de residuales tampoco arrojan conclusiones con diferencias notorias, así que basados en el principio de parsimonia y teniendo en cuenta que no se hallaron

términos de interacción significativos en el modelo 3 se podría pensar que el modelo con ecuación (7), donde se supone que el efecto de RINF sobre DPERM solo cambia según la variable REGION, en específico si el hospital se encuentra en la región 1, es una mejor opción.

Conclusiones

Respecto al modelo presentado en la **ecuación 1** se ha identificado que, pese a que la regresión resulta significativa, algunas medidas de la calidad del modelo como el R^2 y el R^2_{adj} conducen a dudar de la confiabilidad del modelo para realizar estimaciones y predicciones pues el modelo no está capturando una parte importante de la variabilidad observada en la respuesta y el R^2_{adj} que es una alternativa para medir la bondad del ajuste está alejado de 1, lo cual quiere decir, según esta medida, que el ajuste del modelo no es del todo satisfactorio. En análisis previos realizados a este conjunto de datos se evidenciaron problemas a nivel descriptivo con la forma en la que fueron tomados los datos y las variables que fueron medidas y se tenía la hipótesis de que el modelo tenía problemas de multicolinealidad. Esto ha sido verificado; se tiene pues, que mediante varios métodos de diagnóstico se ha detectado multicolinealidad, la más apremiante o severa es aquella que se debe a las variables que describen de alguna forma la capacidad, tamaño e infraestructura de un hospital, como el censo promedio diario, número de enfermeras, número de camas y facilidades y servicios disponibles. Estas variables, de algún modo apuntan a describir un mismo aspecto de un hospital y las correlaciones, VIFs, valores propios y proporciones de descomposición de varianza en algún punto diagnostican problemas de multicolinealidad que involucran a estas predictoras, esto afecta la calidad del modelo para hacer estimaciones y predicciones, además puede dar falsas ideas del comportamiento del fenómeno y por lo tanto es un problema que debería atenderse. En este orden de ideas una alternativa sería construir un indicador que de alguna forma reúna la información relacionada con el tamaño y capacidad de un hospital e ingresarlo como una sola predictora al modelo, esto de paso contribuiría con la reducción de dimensionalidad del problema. Otra alternativa que se usó para atacar el problema fue considerar reducir el modelo inicial en cantidad de variables y evaluar qué tan factible era dejar en él solo aquellas que aportaran información no redundante y relevante para estudiar el fenómeno de interés, esto se hizo indagando si modelos con subconjuntos de estas variables explicaban mejor o al menos de forma muy parecida la situación de interés, es decir, se realizó una depuración de variables, para quizás en este proceso no solo se reducir la dimensionalidad sino, de paso, intentar solucionar algunos problemas de multicolinealidad como resultado de la eliminación de dichas variables del modelo inicialmente propuesto. Esto arrojó como resultado un nuevo modelo con k=4 predictoras donde ya solo una de ellas hace referencia a la capacidad y tamaño de un hospital, como lo es censo promedio diario y tres variables adicionales (edad, riesgo de infección y razón de rutina de rayos x de pecho) que apuntan a describir otros aspectos que podrían explicar mejor la longitud de permanencia promedio. Sin embargo, este nuevo modelo, aunque más parsimonioso y con menos problemas de multicolinealidad, sigue sin ser del todo satisfactorio pues logra explicar solo el 40.99% de la variabilidad total observada de la respuesta y su ajuste podría pensarse muy pobre pues su R^2_{adj} = 0.3776. Muchas veces en la práctica no se alcanzan resultados tan satisfactorios haciendo uso de una metodología específica, esto debido a la calidad de los datos usados en los análisis o por ejemplo a que hay que explorar otras alternativas (otros tipos de modelos) que permitan explicar de mejor manera la situación de interés, de no ser posible esto, debe trabajarse con la mejor alternativa hallada, que en este caso podría ser el modelo descrito anteriormente.

Por otra parte, se quiso evaluar si al intentar explicar la longitud de permanencia promedio de los pacientes en un hospital mediante una regresión lineal con el riesgo de infección como predictora, tal relación lineal resultaba afectada por la región en la que se ubicaba el hospital y su estatus de afiliación a una escuela de medicina, esto implica la inclusión de predictoras cualitativas en el modelo de regresión lineal. Al realizar un análisis descriptivo incluyendo las variables cualitativas se hallan indicios de que la región y la afiliación tienen un efecto a la hora de explicar la longitud de permanencia promedio en función del riesgo de infección, sin embargo, se halla también un inconveniente relacionado con el conjunto de datos que se tiene disponible para realizar el análisis, debido a que se tienen muy pocos registros en algunas de las 8 posibles combinaciones de estos niveles (tratamientos), por ejemplo, hay casos donde solo se cuenta con 2 observaciones por tratamiento, esto dificulta el análisis pues es muy poca información para evidenciar y juzgar con certeza si hay una tendencia lineal en ese caso particular. Aún así, se ajustó un modelo que supone que tanto la media general como el efecto de RINF sobre DEPERM se ven afectados por estas variables, definiendo los hospitales de la región 4 no afiliados a una escuela de medicina como el tratamiento de referencia, se halla un modelo significativo y se halla mediante una prueba que la relación lineal de DEPERM vs. RINF sí difiere según las variables región y/o AEM, pero luego haciendo pruebas separadas para las variables indicadoras y los términos de interacción de ellas con RINF, se halla que no son significativos, esto podría deberse a algún tipo de error (tipo I o II) en las pruebas realizadas o a inconvenientes de multicolinealidad, este modelo no tiene R^2 y R^2_{adj} del todo satisfactorios.

Posteriormente, se decidió ajustar un modelo donde solo se supone que la variable REGION puede cambiar el efecto medio que tiene RINF sobre DEPERM cuando el hospital se encuentra en la región 1, es decir, solo se involucró ese término de interacción en presencia de la variable AEM, allí se halló un modelo que tampoco tiene los mejores resultados en términos de la calidad del ajuste (R^2_{adj} = 0.4908) y la variabilidad que logra explicar (53.05%), este modelo no difiere significativamente en este tipo de medidas respecto al modelo que incluía todas las interacciones, sin embargo, sí mejora en parsimonia, al tener resultados tan parecidos, se prefiere la simplicidad del segundo modelo. En el análisis de residuales ambos reportaron algo de carencia de ajuste y resultados similares para los demás supuestos. Estos resultados no tan satisfactorios en ambos modelos se deben en gran parte al problema de la cantidad de datos disponible por tratamiento que dificulta juzgar si verdaderamente es razonable pensar en un modelo lineal de este tipo para explicar el fenómeno de interés. Quizás debió pensarse en un proceso de toma de datos estratificado que tenga en cuenta la ubicación y el estatus de los hospitales para no tener un desbalanceo tan drástico en la cantidad de observaciones por tratamiento y así buscar el modelo más apropiado para estudiar longitud de permanencia promedio en función del riesgo de infección en presencia de las variables región y afiliación a la escuela de medicina.

PARTE I

3. Ajuste del modelo 93 y su ANOVA

```
#AJUSTE DEL MODELO 93

modelo93=lm(DPERM~.,data=datos2[,c(1,2,3,5,7)])
summary(modelo93)

#ANOVA DEL MODELO 93

anova(rsm(DPERM~FO(EDAD,RINF,RRX,PDP),data=datos2))
```

4. Ajuste del modelo 37

```
modelo37=lm(DPERM~.,data=datos2[c(1,2,3,7)])
summary(modelo37)
```

5. Gráficos para la validación de los supuestos para el modelo 93

```
#GRÁFICOS DE RESIDUALES ESTUDENTIZADOS EXTERNAMENTE

residualPlots(modelo93,tests=FALSE,type="rstudent",q
uadratic=FALSE,col=2,cex=2)

##GRÁFICO DE PROBABILIDAD NORMAL

test2=shapiro.test(rstudent(modelo93)) #Test de
normalidad sobre residuales estudentizados
qqnorm(rstudent(modelo93),cex=1.5)
qqline(rstudent(modelo93),col=2)
legend("topleft",legend=rbind(c("StatisticW","p.valu
e"),round(c(test2$statistic,test2$p.value),digits=4)
),cex=0.8)
shapiro.test(rstudent(modelo93))
```

PARTE II (TEMA A)

1.1. Análisis descriptivo

```
#RESUMEN VARIABLES DE INTERÉS

summary(datos3)#datos3 contiene las cuatro variables
de interés: DPERM, RINF, AEM y REGION

#FRECUENCIAS DE CADA COMBINACIÓN DE NIVELES DE AEM y
REGIÓN

table(datos3[,c(3)], datos3[,c(4)])
```

1.5 Prueba de significancia de todos los parámetros asociados a las variables indicadoras y su interacción con RINF

```
#"AEM1" "REGION1" "REGION2" "REGION3" "RINF:AEM"
"RINF:REGION1" "RINF:REGION2" "RINF:REGION3"

vectorhipotesis5=paste0(nombres[2:9],"=0")
#codificando hipótesis nula
vectorhipotesis5
linearHypothesis(modeloRINF,vectorhipotesis5)
#ejecución del test
```

1.6 Prueba de significancia términos principales

```
#"AEM1" "REGION1" "REGION2" "REGION3

vectorhipotesis6=paste0(nombres[2:5],"=0")
#codificando hipótesis nula
vectorhipotesis6
linearHypothesis(modeloRINF,vectorhipotesis6)
#ejecución del test
```

1.7 Prueba de significancia de los parámetros asociados a la interacción entre las indicadoras y la variable RINF

```
#"RINF:AEM" "RINF:REGION1" "RINF:REGION2"
"RINF:REGION3"

vectorhipotesis7=paste0(nombres[6:9],"=0")
#codificando hipótesis nula
vectorhipotesis7
linearHypothesis(modeloRINF,vectorhipotesis7)
#ejecución del test
```

1.8.1 Prueba de significancia de la interacción doble de RINF con REGION

```
#"RINF:REGION1" "RINF:REGION2" "RINF:REGION3"

vectorhipotesis81=paste0(nombres[7:9],"=0")
#codificando hipótesis nula
vectorhipotesis81
linearHypothesis(modeloRINF,vectorhipotesis81)
#ejecución del test
```

1.8.1 Prueba de significancia de la interacción doble de RINF con AEM

```
#"RINF:AEM1"

vectorhipotesis82=paste0(nombres[6],"=0")
#codificando hipótesis nula
vectorhipotesis82
linearHypothesis(modeloRINF,vectorhipotesis82)
#ejecución del test
```

2.2 ANOVA del modelo2RINF

```
#Anova del modelo "modelo2RINF"
nombres2=attr(model.matrix(modelo2RINF),"dimnames")[
2][[1]][-1]
nombres2

vectorhipotesismodelo2RINF=paste0(nombres2,"=0")
#codificando hipótesis nula del ANOVA
vectorhipotesismodelo2RINF
linearHypothesis(modelo2RINF,vectorhipotesismodelo2R
INF) #ejecución del test
```