banuls_marc_Reg_PEC1

Marc Bañuls Tornero

9/4/2020

Contents

Problema 1

Cargamos los datos en R:

Como se dice en el ejercicio, vamos a trabajar con variables dicotómicas, por lo que separamos la variable **Sex** en dos variables: **Male** y **Female**. Para ello creamos las dos variables con todos los valores. Cabe destacar que al crear las dos variables mencionadas, debemos eliminar de los datos la variable **Sex**.

Posteriormente a cada variable le indicamos con el valor "0" O "1".

En la variable **Alcohol** tan solo indicamos como "1" a los sujetos que contienen "Alcoholic" y "0" a los sujetos que contienen "Non-alcoholic".

Ahora realizamos los modelos de regresión indicados:

(a) Entre el modelo M_1 y el modelo M_2 está claro que $\beta_2^{(2)} = -\beta_2^1$. ¿Cuál es la relación entre β_0^2 y los parámetros del modelo M_1?

Una manera de observar los distintos parámetros de cada modelo es investigando el resumen del modelo estadístico con los datos de la tabla, es decir, ajustando los modelos indicados y observando los valores de los distintos coeficientes:

```
##
## Call:
## lm(formula = Metabol ~ Gastric + Female, data = alcohol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2779 -0.6328 -0.0966  0.5783  4.5703
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3292     0.7022  -0.469  0.64270
## Gastric       1.9656     0.2674   7.352 4.24e-08 ***
## Female       -1.6174     0.5114  -3.163  0.00365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.331 on 29 degrees of freedom
## Multiple R-squared:  0.7654, Adjusted R-squared:  0.7492
## F-statistic: 47.31 on 2 and 29 DF,  p-value: 7.41e-10
```

```
##
## Call:
## lm(formula = Metabol ~ Gastric + Male, data = alcohol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2779 -0.6328 -0.0966  0.5783  4.5703
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.9466     0.5198  -3.745 0.000796 ***
## Gastric       1.9656     0.2674   7.352 4.24e-08 ***
## Male          1.6174     0.5114   3.163 0.003649 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.331 on 29 degrees of freedom
## Multiple R-squared:  0.7654, Adjusted R-squared:  0.7492
## F-statistic: 47.31 on 2 and 29 DF,  p-value: 7.41e-10
```

Observamos que $\beta_0^2 = \beta_0^1 + \beta_2^1$. Es decir, la suma del intercepto y el coeficiente de la variable **Female** del modelo 1 equivalen al intercepto del modelo 2.

(b) ¿Cuál es la diferencia en término medio de Metabol entre hombres y mujeres que tienen un mismo nivel alcohol deshidrogenasa?

Para responder la pregunta debemos observar los coeficientes de los modelos lineares generados. Concretamente, el coeficiente de la variable **Male** en el modelo M_1 o el coeficiente de la variable **Female** del modelo M_2 nos dan la respuesta al apartado (ya que al tener una correlación perfecta el resultado del coeficiente en ambos modelos es el mismo pero con el valor invertido). Por lo tanto, podemos decir que los hombres tienen un valor medio de la variable **Metabol** 1.6174 veces superior a las mujeres.

(c) De los cuatro modelos $M_i, i = 1, ..., 4$, ¿cuál es el mejor según el coeficiente de determinación? ¿y según el RMSE?

Observamos el coeficiente de determinación de los 4 modelos:

```
## [1] 0.7653987
```

```
## [1] 0.7653987
```

```
## [1] 0.7653987
```

```
## [1] 0.8736451
```

Los tres primeros modelos técnicamente son iguales debido a que las variables **Male** y **Female** aportan la misma información, así que independientemente de si ponemos una u otra u ambas, la información disponible para explicar la variación de los valores en la regresión es la misma. En cambio, el modelo M_4 tiene un mayor coeficiente de determinación (pasando de 0.76 a 0.87). Esto puede ser debido a que empezando el modelo a partir del punto 0 (al haber indicado que el intercepto es 0) el modelo puede haber explicado mejor la variación de los valores.
El RMSE es la raíz cuadrada de la media de los residuos del modelo al cuadrado, por lo que utilizamos los valores de los residuos de cada modelo para calcular su RMSE:

```
## [1] 1.267378
```

```
## [1] 1.267378
```

```
## [1] 1.267378
```

```
## [1] 1.267378
```

Como el RMSE nos indica las diferencias entre los valores predecidos por el modelo y los valores observados de la tabla y en los cuatro modelos estamos utilizando los mismos valores, los valores de RSME de los cuatro modelos son iguales, en este caso siendo el RSME de 1.267378.

(d) ¿Cuál es el rango de la matriz de diseño del modelo M_3? Resolver las ecuaciones normales para este modelo y hallar una estimación alternativa de los parámetros con la ayuda de la g-inversa de Moore-Penrose. Comprobar que los residuos son los mismos que proporciona R con la función *lm()*.

Observamos la matriz de diseño del modelo:

```
##   (Intercept) Gastric Male Female
## 1           1     1.0    0      1
## 2           1     1.6    0      1
## 3           1     1.5    0      1
## 4           1     2.2    0      1
## 5           1     1.1    0      1
## 6           1     1.2    0      1
```

Podemos ver que la matriz tiene 3 variables independientes. El intercepto, la variable **Gastric** y el sexo (las variables **Male** y **Female** son dependientes entre ellas). De igual manera se puede medir el rango de la matriz dentro del propio modelo ajustado:

```
## [1] 3
```

Por lo tanto confirmamos que la matriz de diseño es de rango 3.

Para realizar las ecuaciones normales mediante la g-inversa del Moore-Penrose, utilizamos el paquete *MASS*. Señalamos como valor respuesta la variable **Metabol** y como predictores las otras variables del modelo, y realizamos la ecuación normal:

De esta manera hemos obtenido los coeficientes mediante la g-inversa de Moore-Penrose. Ahora obtenemos los residuos:

```
##         [,1]
## 1  0.5810679
## 2 -0.5982789
## 3  0.4982789
## 4 -1.9776257
## 5 -0.1154900
## 6 -0.2120478
```

Ahora comprobamos que los residuos obtenidos son los mismos que en el ajuste en R:

```
##          1          2          3          4          5          6
##  0.5810679 -0.5982789  0.4982789 -1.9776257 -0.1154900 -0.2120478
```

Como podemos observar, los residuos calculados con la ayuda de la ginversa de Moore-Penrose son iguales que los calculados mediante el ajuste realizado con la función `lm()`.

(e) Hallar el intervalo de confianza de la función paramétrica $\beta_0^{(2)} + \beta_1^2$ en el modelo M_2.

Obtenemos los valores de x e y del modelo M2 y obtenemos las betas como en el apartado anterior.

Calculamos además la varianza de los errores σ^2 junto con los valores necesarios para ello.

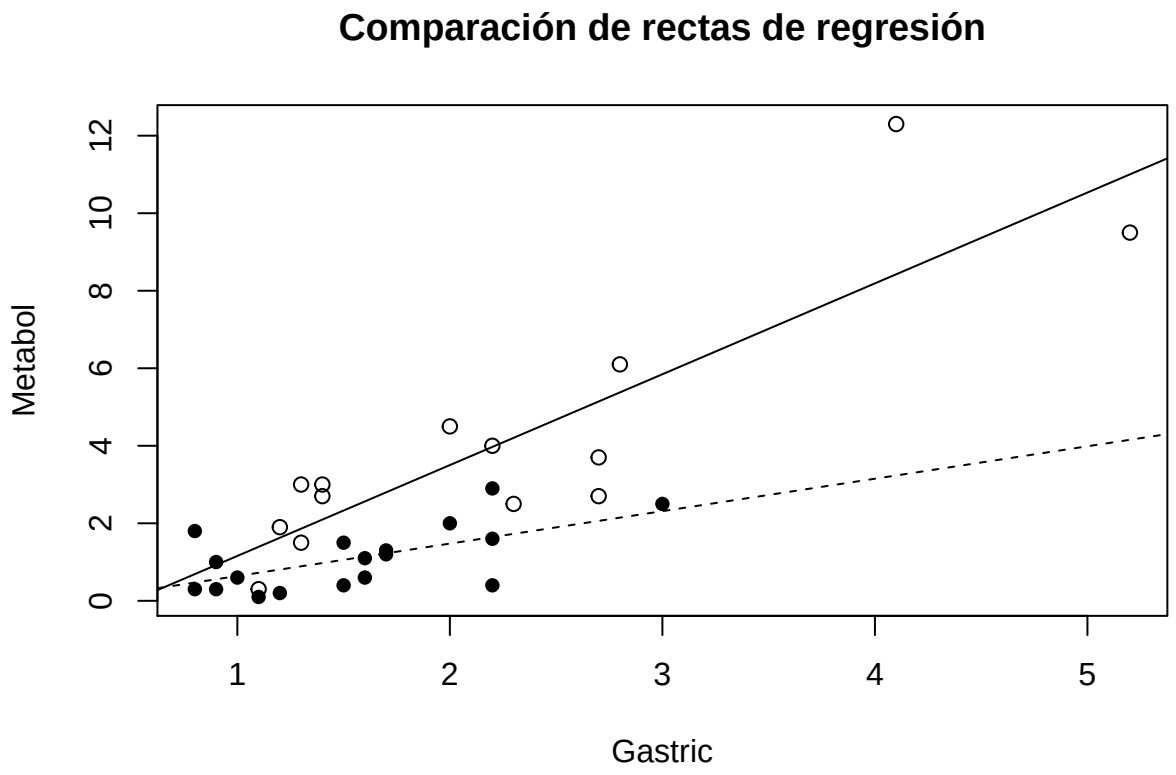
Ahora con los valores obtenidos realizamos el intervalo de confianza al 95% del modelo:

```
## [1] -0.6898185  0.7276828
```

(f) Comparar las rectas de regresión que relacionan el metabolismo Metabol con la actividad gástrica Gastric para hombres y para mujeres. ¿Son paralelas? ¿Son iguales?

Para obtener las rectas de regresión se separan los datos de la tabla alcohol por sexos mediante subsets en la función `lm()`:

Ahora podemos observar mediante un gráfico de dispersión los datos y las lineas obtenidas en cada modelo:



En el gráfico observamos que estas rectas parecen tener una intersección en el origen y la clara diferencia entre las pendientes de ambas rectas implica que las rectas no son paralelas y consecuentemente tampoco son iguales.

(g) Si consideramos el modelo completo con interacciones, ¿Podemos prescindir de todas las interacciones y también de la variable Alcoholic y quedarnos con el modelo M_2?

Para responder a la pregunta, ajustamos el modelo completo con interacciones propuesto (la variable **Alcoholic** del enunciado se llama en nuestra tabla de datos **Alcohol**, ya que en vez de crear una nueva variable con los datos dicotómicos, hemos modificado la variable inicial **Alcohol** directamente):

Ahora observamos la significación de las distintas variables y sus interacciones en el resumen del modelo ajustado:

```
##
```

```
## Call:
## lm(formula = Metabol ~ Gastric + Male + Alcohol + Gastric * Male +
##     Gastric * Alcohol + Male * Alcohol + Gastric * Male * Alcohol,
##     data = alcohol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4286 -0.6189 -0.0466  0.5150  3.6516
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.1939     0.8811  -0.220   0.8277
## Gastric                0.8407     0.5165   1.628   0.1166
## Male                  -1.4657     1.3326  -1.100   0.2823
## Alcohol                0.3004     3.9393   0.076   0.9398
## Gastric:Male           1.6734     0.6202   2.698   0.0126 *
## Gastric:Alcohol       -0.2601     2.8069  -0.093   0.9269
## Male:Alcohol           2.2517     4.3937   0.512   0.6130
## Gastric:Male:Alcohol  -1.1987     2.9978  -0.400   0.6928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.254 on 24 degrees of freedom
## Multiple R-squared:  0.8277, Adjusted R-squared:  0.7774
## F-statistic: 16.47 on 7 and 24 DF,  p-value: 9.354e-08
```

Los p-valores del test estadístico t para cada predictor o interacción entre estos nos indica si estos predictores o interacciones son significativos o no para el ajuste del modelo con los datos que tenemos (la hipótesis nula del test es que el coeficiente del predictor o la interacción es 0). En la tabla resumen observamos que tan solo la interacción entre la variable **Gastric** y la variable **Male** rechaza la hipótesis nula del t-test, implicando que esta interacción es la única que presenta una relación significativa entre esta interacción y los datos de la tabla *alcohol*.

Para saber si podemos prescindir de las interacciones y de la variable **Alcoholic** (en nuestro caso **Alcohol**) comparamos el modelo completo con el modelo M_2, el cual es precisamente el modelo del que ha anidado el modelo completo. Para comparar modelos anidados realizamos un F-test, donde se realiza un contraste de la hipótesis nula en que la variación entre estos dos modelos no es significativa:

```
## Analysis of Variance Table
##
## Model 1: Metabol ~ Gastric + Male + Alcohol + Gastric * Male + Gastric *
##     Alcohol + Male * Alcohol + Gastric * Male * Alcohol
## Model 2: Metabol ~ Gastric + Male
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     24 37.754
## 2     29 51.400 -5   -13.646 1.7349  0.165
```

El F-valor mayor a 0.05 nos indica que no hay una variación significativa entre los modelos. Por lo tanto, en caso de elegir qué modelo usar elegiríamos el modelo más simple, es decir, el modelo M_2.

Problema 2

Para poder realizar este ejercicio primero guardamos los datos del archivo senic.txt en una tabla de datos llamada **senic**:

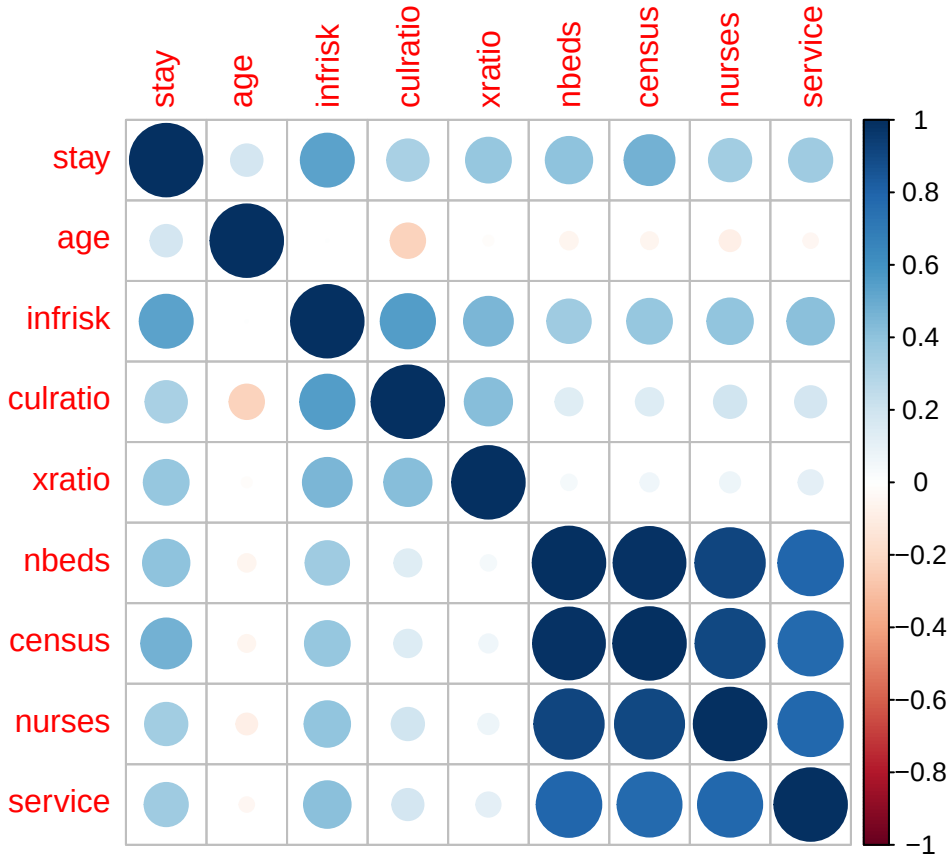
(a) Calcular la matriz de correlaciones entre las variables que sea posible. ¿Qué variables son las más correlacionadas con infrisk? Añadir algún gráfico adecuado. Dibujar los gráficos de caja (*boxplot*) para la variable infrisk primero separados según la variable medschl y después según region. ¿El riesgo de infección es igual en todas las regiones? ¿Y según la variable medschl?. Nota: Plantear las dos últimas preguntas como un contraste de modelos con la función anova().

Para realizar la matriz de correlaciones entre las variables del estudio senic utilizamos la función `cor()`. Cabe tener en cuenta que como la variable id no proporciona ningún valor significativo, descartamos dicha variable de la matriz de correlaciones.

Modificaremos la variable **region** para que sea factorial, y de esta manera se puedan interpretar los valores para cada región por separado. Por ello, para realizar la matriz de correlación, deberemos descartar esta variable (ya que una variable categórica no puede ser interpretada correctamente en una matriz de correlaciones). Además modificaremos los valores de la variable **medschl** a valores dicotómicos, es decir, el valor 1 implicará que el hospital no está afiliado a una escuela médica y el valor 0 que el hospital no está afiliado.

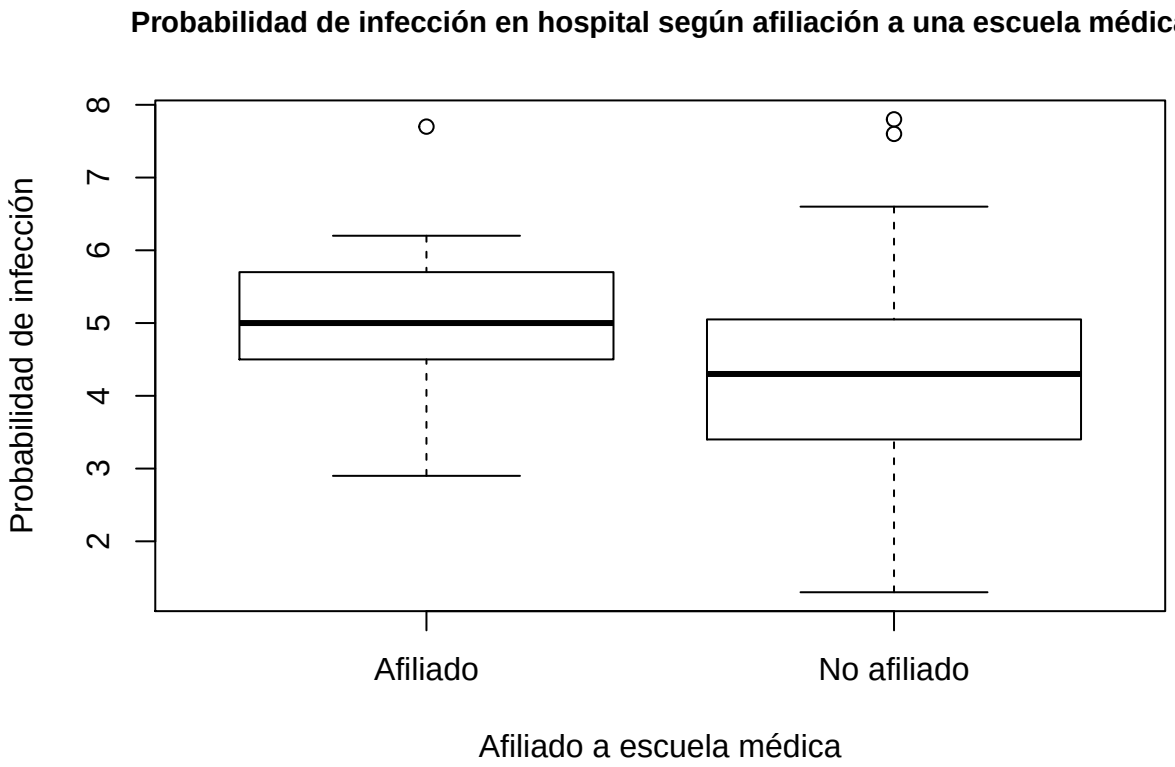
```
##          stay   age infrisk culratio xratio nbeds census nurses service
## stay     1.00  0.19    0.53     0.33   0.38  0.41   0.47   0.34    0.36
## age      0.19  1.00    0.00    -0.23  -0.02 -0.06  -0.05  -0.08   -0.04
## infrisk  0.53  0.00    1.00     0.56   0.45  0.36   0.38   0.39    0.41
## culratio 0.33 -0.23    0.56     1.00   0.42  0.14   0.14   0.20    0.19
## xratio   0.38 -0.02    0.45     0.42   1.00  0.05   0.06   0.08    0.11
## nbeds    0.41 -0.06    0.36     0.14   0.05  1.00   0.98   0.92    0.79
## census   0.47 -0.05    0.38     0.14   0.06  0.98   1.00   0.91    0.78
## nurses   0.34 -0.08    0.39     0.20   0.08  0.92   0.91   1.00    0.78
## service  0.36 -0.04    0.41     0.19   0.11  0.79   0.78   0.78    1.00
```

Las variables que tienen una mayor correlación con la variable **infrisk** son las variables **culratio**, **stay** y **xratio** con un coeficiente de correlación de 0.56, 0.53 y 0.45 respectivamente.
Para poder visualizar con un gráfico las correlaciones, podemos usar el paquete de R `corrplot`:



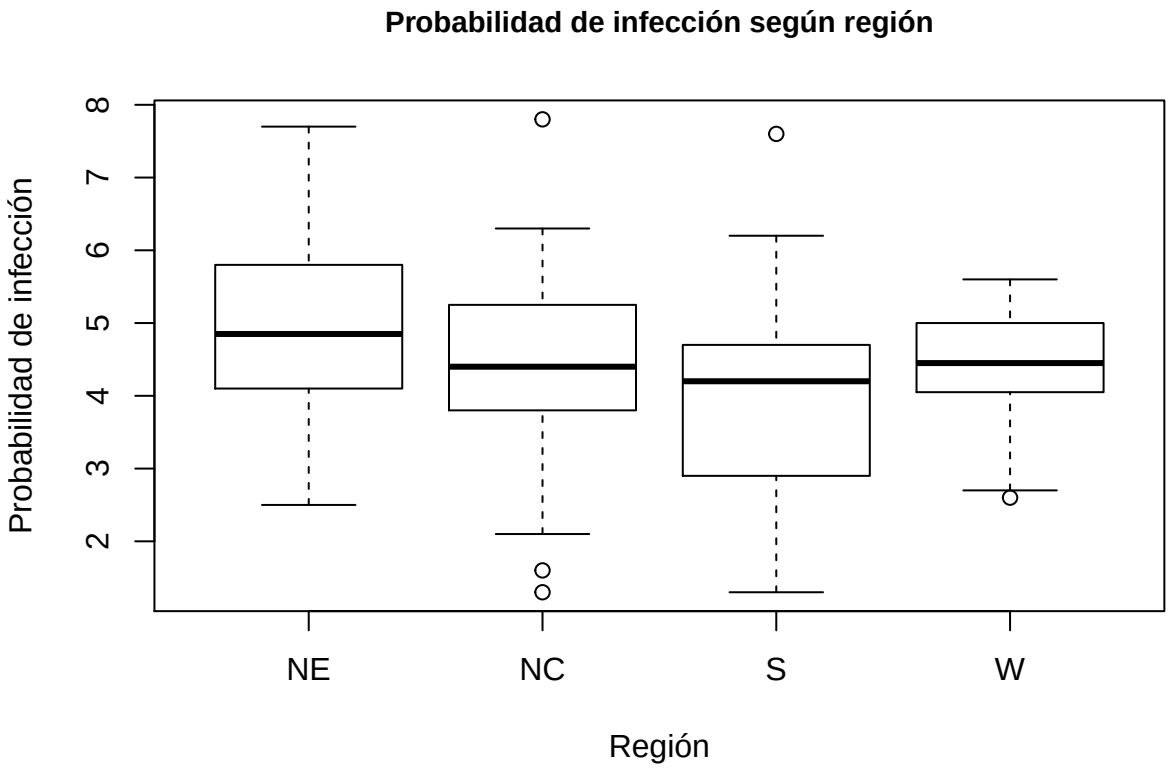
En este gráfico también se pueden observar visualmente que las variables mencionadas anteriormente son las que tienen una mayor correlación con la variable **infrisk**.

Ahora nos disponemos a visualizar en un gráfico de cajas la variable **infrisk** separados según la variable **med-**



schl:
En este gráfico de cajas se observa que los hospitales afiliados y no afiliados tienen algún outlier y los hospitales no afiliados tienen unos cuartiles 1 y 4 significativamente separados del rango de intercuantiles (IQR), donde la mayor parte de valores se encuentra. Además, observamos que las medianas entre hospitales afiliados y no afiliados son distintas, aunque no parecen encontrarse suficientemente separadas como para ser significativas.

También realizamos el respectivo boxplot de la variable **infrisk** según la variable **region**:



En este gráfico de cajas se observa que las medianas entre regiones no cambian en gran medida y que su IQR cambia levemente entre regiones, observando por ejemplo que la región S suele tener la menor probabilidad de infección.

Para saber objetivamente si hay variación del riesgo de infección según regiones o según la variable **medschl** y **region** realizaremos un modelo de regresión para cada contraste de hipótesis y posteriormente realizaremos un análisis de varianza de cada modelo utilizando la función `anova()`. De esta manera, sabremos si existe una correlación entre estas variables y **infrisk**.

Primero averiguamos la variación del riesgo de infección según la variable **medschl**

```
## Analysis of Variance Table
##
## Response: infrisk
##            Df  Sum Sq Mean Sq F value Pr(>F)
## medschl     1  10.936 10.9355  6.3737  0.013 *
## Residuals 111 190.444  1.7157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con un p-valor del F test menor a 0.05 podemos decir que hay una variación significativa entre estar o no afiliado a una escuela médica en relación a la probabilidad de infección. Esto indica que el riesgo de infección si que varía según la variable **medschl**.

Ahora analizamos si el riesgo de infección es igual en todas las regiones (variable **region**).

```
## Analysis of Variance Table
##
## Response: infrisk
```

```
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## region       3  13.997  4.6656   2.714 0.04839 *
## Residuals  109 187.383  1.7191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con un p-valor del F test menor a 0.05, también podemos afirmar que hay una variación significativa entre las distintas regiones en relación a la probabilidad de infección. De esta manera confirmamos que el riesgo de infección no es igual en todas las regiones, es decir, la variable **infrisk** varía significativamente según el valor de la variable **region**.

(b) Calcular el modelo de regresión que tiene como variable respuesta infrisk. Escribir el modelo obtenido. ¿es significativa la variable region? ¿Cómo interpretas el coeficiente de la variable medschl? ¿Y el de stay?

Calculamos el modelo que se pide:

Realizamos ahora el resumen estadístico del modelo:

```
##
## Call:
## lm(formula = infrisk ~ ., data = senic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8362 -0.4945 -0.0606  0.5284  2.5225
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.513670   1.322379  -1.901 0.060199 .
## stay                0.242403   0.070184   3.454 0.000812 ***
## age                 0.013231   0.021771   0.608 0.544745
## culratio            0.054485   0.010550   5.165 1.23e-06 ***
## xratio              0.011553   0.005264   2.195 0.030504 *
## nbeds              -0.003489   0.002677  -1.303 0.195420
## medschlNo afiliado  0.660823   0.321392   2.056 0.042375 *
## regionNC            0.425509   0.264415   1.609 0.110715
## regionS             0.366762   0.272487   1.346 0.181353
## regionW             1.149535   0.339173   3.389 0.001004 **
## census              0.003865   0.003451   1.120 0.265477
## nurses              0.001787   0.001695   1.055 0.294165
## service             0.020570   0.010059   2.045 0.043492 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9138 on 100 degrees of freedom
## Multiple R-squared:  0.5854, Adjusted R-squared:  0.5356
## F-statistic: 11.76 on 12 and 100 DF,  p-value: 1.963e-14
```

Si consideramos que la hipótesis nula es que la variable **region** no es significativa con un nivel de significación de $\alpha = 0.05$ entonces como el p-valor del t test de uno de los niveles en esta variable **region** es menor a 0.05 (concretamente, 0.001) podemos decir que almenos un factor de esta variable es significativa, es decir, es un predictor que tiene una relación lineal con la variable respuesta **infrisk**, cosa que hemos confirmado en el anterior apartado.

En la variable **medschl** se observa sólo uno de los dos niveles de esta variable (tal y como ocurre también con

la variable **region**), ya que para variables categóricas, utilizamos uno de los niveles como referencia (con su coeficiente = 0). Por lo tanto, el coeficiente mostrado en el resumen estadístico indica que cuando el hospital no está afiliado a una escuela médica, el riesgo de infección es 0.66 veces mayor que cuando el hospital sí está afiliado a una escuela médica.

La variable **stay** al ser una variable contínua nos indica que cada día de promedio de estancia que pasan los pacientes en el hospital aumenta el porcentaje del riesgo de infección en un 0.24%.

(c) Utilizar un test *F* para determinar la significación de la regresión del modelo. Escribe las hipótesis de este test e interpreta el resultado obtenido. ¿Qué predictoras son significativas al 5%? ¿Concuerdan estas variables con las del apartado (a)? ¿Cuáles son las variables más correlacionadas? ¿Y las que menos? ¿Concuerdan con los resultados del modelo de regresión?

En el resumen anterior ya se realiza un F-test donde se rechaza la hipótesis nula (no existe ninguna relación lineal entre la variable respuesta y los predictores del modelo) por tener un p-valor inferior a 0.05, por lo que se afirma que almenos un predictor contribuye significativamente al modelo. Para determinar la significación de la regresión reproducimos la tabla anova del modelo de regresión utilizando la función `aov()` y creando su respectivo resumen estadístico. En este caso tenemos una hipótesis nula para cada predictor. La hipótesis nula es que el predictor en concreto no tiene una relación lineal con la variable respuesta. Realizando el contraste de hipótesis para cada variable resolveremos qué predictores contribuyen significativamente al modelo de regresión estudiado:

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## stay          1  57.31   57.31  68.629 5.54e-13 ***
## age           1   2.08    2.08   2.485  0.11809
## culratio      1  31.46   31.46  37.673 1.69e-08 ***
## xratio        1   3.85    3.85   4.608  0.03424 *
## nbeds         1   6.52    6.52   7.804  0.00625 **
## medschl       1   1.15    1.15   1.383  0.24239
## region        3   8.91    2.97   3.557  0.01703 *
## census        1   1.34    1.34   1.600  0.20884
## nurses        1   1.79    1.79   2.139  0.14669
## service       1   3.49    3.49   4.182  0.04349 *
## Residuals   100  83.50    0.84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En el resumen estadístico observamos que las variables significativas al 5% (p-valor menor a 0.05), es decir, los predictores que contribuyen significativamente al modelo de regresión, son las variables **stay**, **culratio**, **xratio**, **nbeds**, **region**, y **service**. Estas variables concuerdan en cierta medida con las del apartado (a), ya que las variables que tienen un menor p-valor (las que con más certeza se demuestra que existe una relación lineal entre el predictor concreto y la variable respuesta) son las que mayor correlación se observa en la matriz de correlaciones. Las variables más correlacionadas que concuerdan con las correlaciones obtenidas en el apartado (a) son las variables **stay**, **culratio** y **xratio** (ya que son las más significativas con un p-valor menor). Las variables que menor correlación tienen son las que tienen un p-valor mayor a 0.05 (aceptando la hipótesis nula), que son concretamente las variables **age**, **medschl**, **census** y **nurses**. En términos generales las variables más o menos relacionadas con la variable **infrisk** obtenidas con el F-test concuerdan con el test de correlación.

Comparando los predictores más significativos del resumen realizado en el modelo de regresión (mediante un t-test) coinciden varios predictores en su significancia excepto en algunos casos. Concretamente, la variable **nbeds** no es significativa según el t-test, pero sí lo es según el F-test. De igual manera, en el F-test la variable **medschl** no es significativa, pero sí lo es según el t-test.

(d) Si amplificamos el modelo tomando únicamente las variables significativas al 5%, contrastar si se puede aceptar ese modelo simplificado frente al completo.

Ajustamos el modelo con las variables significativas al 5% mencionadas en el anterior apartado:

Ahora realizamos un test ANOVA para comparar el modelo completo con el anidado. La hipótesis nula de este contraste es que no hay diferencias significativas entre los dos modelos (la varianza entre estos dos modelos es 0).

```
## Analysis of Variance Table
##
## Model 1: infrisk ~ stay + culratio + xratio + medschl + region + service
## Model 2: infrisk ~ stay + age + culratio + xratio + nbeds + medschl +
##     region + census + nurses + service
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    104 86.349
## 2    100 83.500  4    2.8492 0.853  0.495
```

Con un nivel de significación del 5% como tenemos un p-valor del F-test mayor a 0.05, aceptamos la hipótesis nula del contraste, es decir, no hay diferencias significativas entre los dos modelos. Por ello aceptamos el modelo simplificado frente al completo, ya que de esta manera con menos predictores obtenemos un modelo con una efectividad similar (se recomienda generalmente utilizar modelos con menor cantidad de predictores si los que se eliminan o no se añaden al modelo no mejoran el modelo).

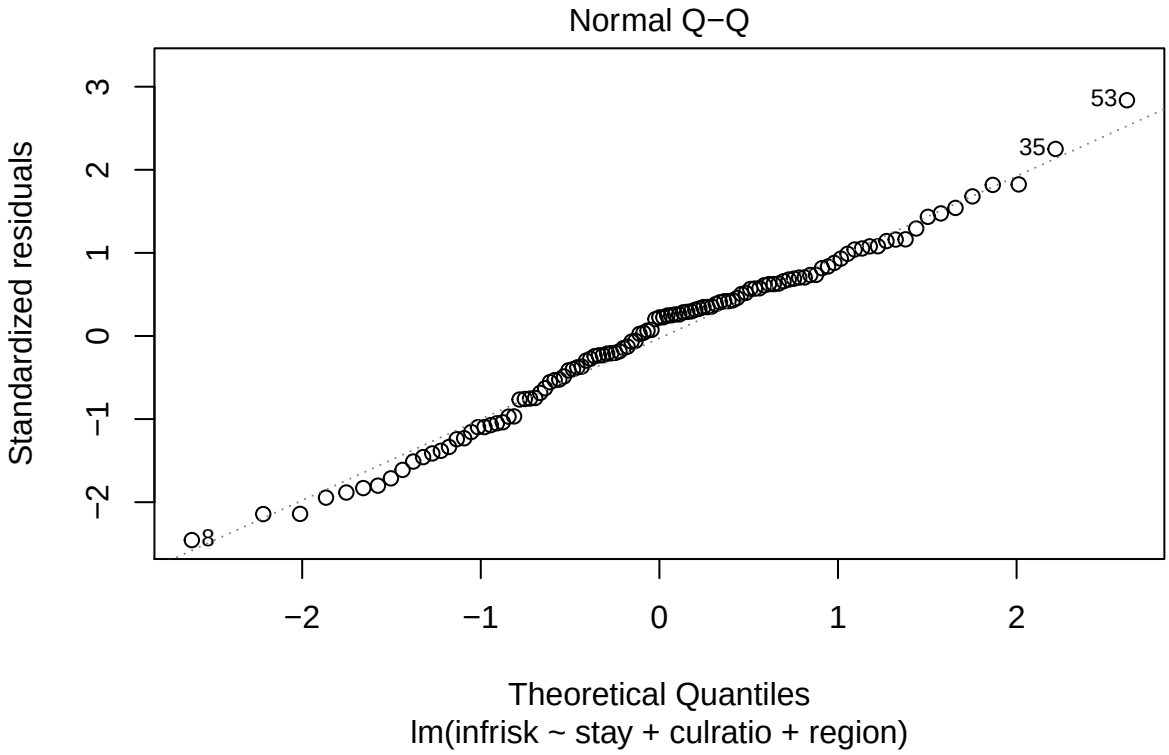
(e) Consideramos ahora el modelo con las variables: stay, culratio y region. Estudia la normalidad y la heterocedasticidad del error. ¿Hay alguna observación con un alto leverage? ¿ Y con una gran influencia? Dibujar los gráficos oportunos para explicar los resultados.

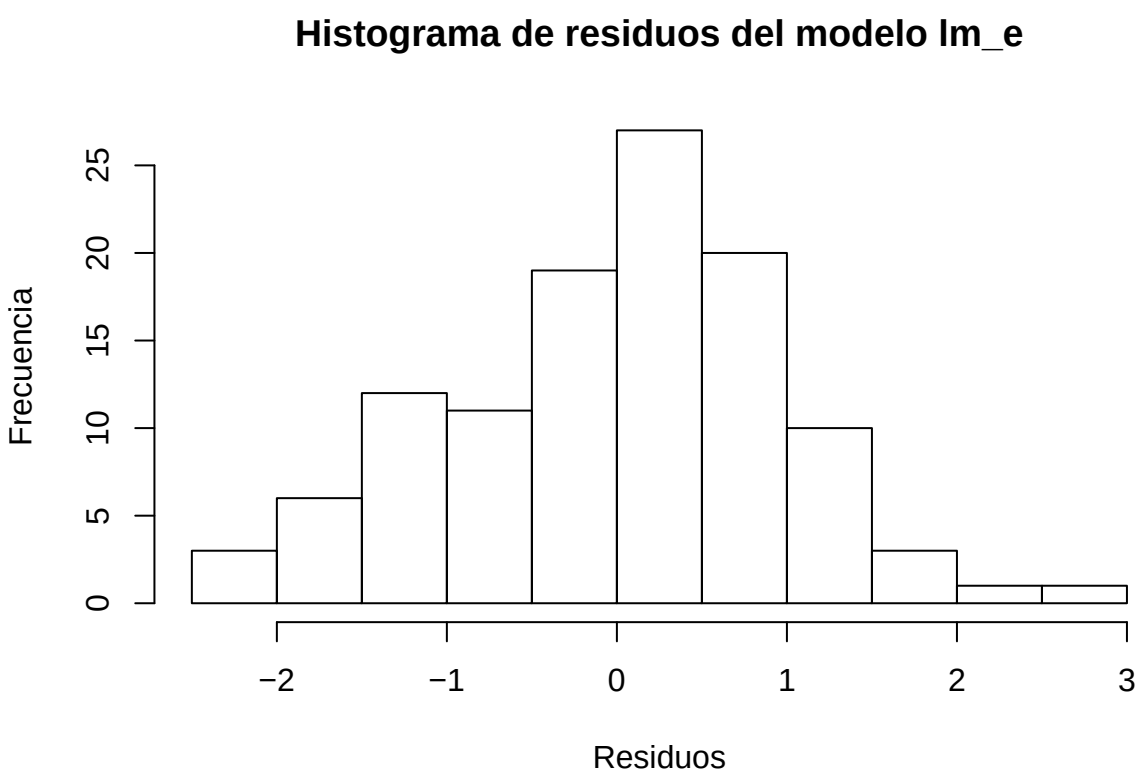
Ajustamos el modelo con las variables mencionadas.

```
##
## Call:
## lm(formula = infrisk ~ stay + culratio + region, data = senic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1084 -0.6564  0.2108  0.6047  2.6994
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.161945   0.648026  -0.250  0.80314
## stay         0.341202   0.056888   5.998 2.76e-08 ***
## culratio     0.058438   0.009729   6.007 2.65e-08 ***
## regionNC     0.320149   0.265812   1.204  0.23109
## regionS      0.199331   0.271845   0.733  0.46501
## regionW      1.030779   0.350042   2.945  0.00397 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9726 on 107 degrees of freedom
## Multiple R-squared:  0.4973, Adjusted R-squared:  0.4738
## F-statistic: 21.17 on 5 and 107 DF,  p-value: 1.147e-14
```

Para estudiar la normalidad visualmente podemos presentar los gráficos del modelo para observar si los errores siguen una distribución normal. Para ello creamos un gráfico Q-Q normal (la función `plot()` presenta

4 gráficos, pero para la normalidad necesitamos tan solo éste gráfico) y un histograma de los residuos:



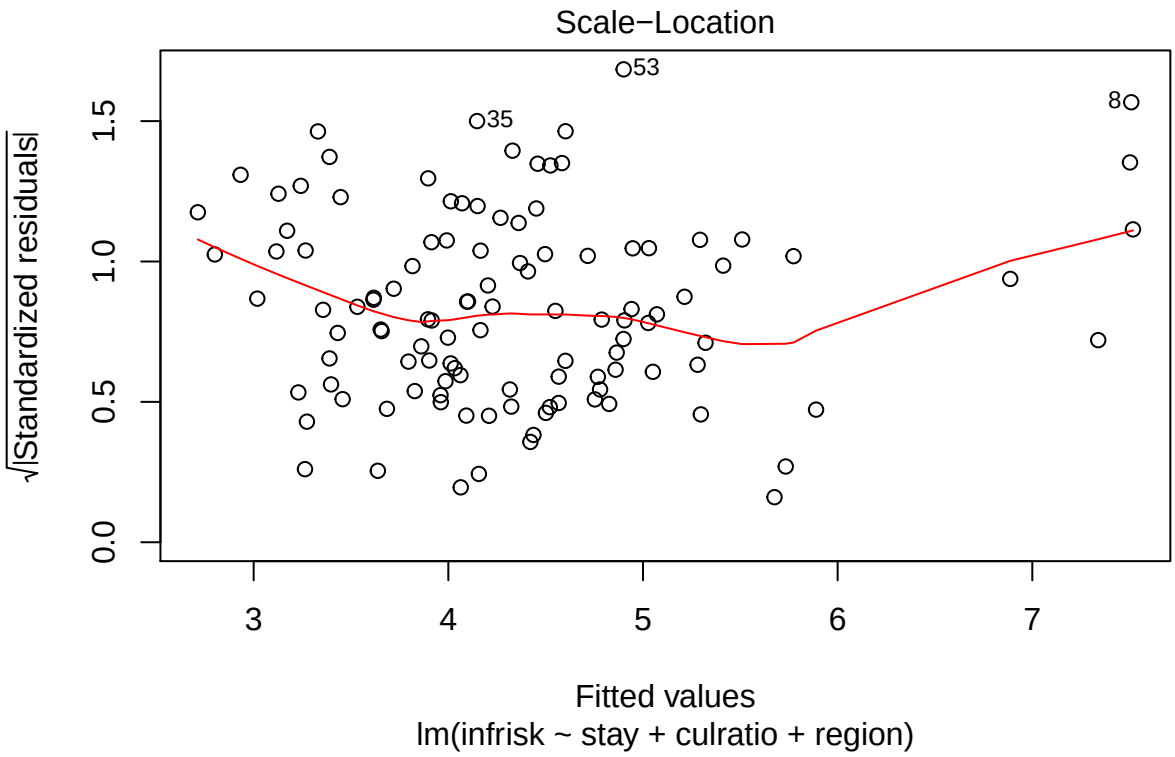


Observando el gráfico Q-Q de normalidad (Q-Q plot) se distingue una línea recta de los residuos estandarizados, indicando que es probable que los residuos se encuentren normalmente distribuidos. Además en el histograma de los residuos se intuye la forma de campana de Gauss típica de una dsitribución normal, aunque tampoco nos da la certeza de que estos residuos siguen una distribución normal. Para estudiar con más profundidad la normalidad del modelo, podemos realizar un test de normalidad de Shapiro-Wilk o de Kolgomorov-Smirnov. En nuestro caso utilizamos el test de Shapiro-Wilk, teniendo como hipótesis nula que el modelo sigue una distribución normal.

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_e$residuals
## W = 0.98587, p-value = 0.2832
```

Si utilizamos un nivel de significación del 5% y siendo el p-valor mayor a 0.05, aceptamos la hipótesis nula. Por lo tanto confirmamos que los residuos siguen una distribución normal.

Para estudiar visualmente la homocedasticidad de los residuos (que los residuos del modelo presenten una misma o similar varianza). Para visualizar si los residuos presentan o no homocedasticidad podemos observar otro gráfico del modelo creado mediante la función `plot`:

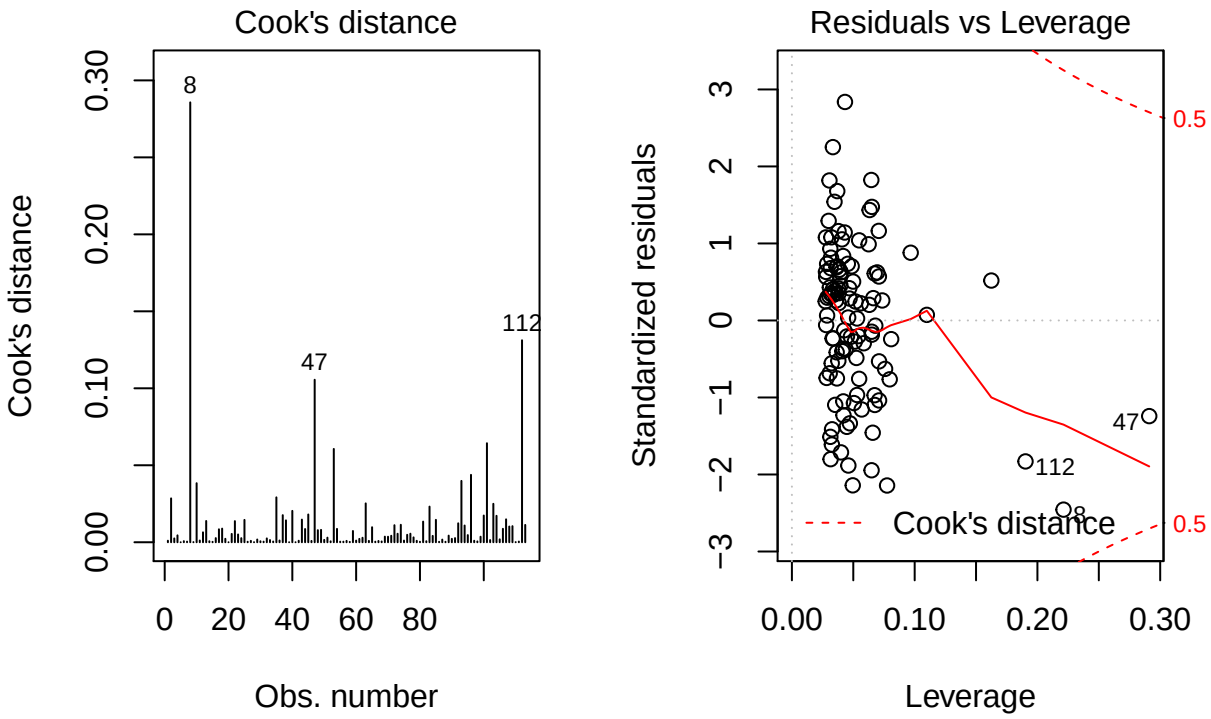


El gráfico nos presenta los valores ajustados respecto a los residuos del modelo estandarizados. Para demostrar los residuos presentan homocedasticidad, dichos residuos deberían estar distribuidos de manera aleatoria por todo el gráfico de manera similar, dando una línea de la media de los residuos estandarizados distribuidos (la línea roja) horizontal. En el caso de los residuos de nuestro modelo, se observa que no hay una perfecta homocedasticidad, aunque en este caso no se puede saber concretamente si la varianza de los residuos es los suficiente significativa para no asumir que homocedasticidad. Para ello, utilizaremos el test estadístico de Breusch-Pagan que presenta como hipótesis nula que los residuos presentan homocedasticidad con un 5% de sgnificación:

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2392009, Df = 1, p = 0.62478
```

Con un p-valor muy superior al 5% de significación, se confirma que aceptamos la hipótesis nula, asumiendo la homocedasticidad de los residuos.

Para saber si hay alguna observación con un alto leverage o una gran influencia, podemos observar otros gráficos creados mediante la función `plot()`, como por ejemplo la distancia de Cook o el gráfico de Residuals vs Lever-



age:

El gráfico de Residuals vs Leverage indica con su nombre (el número de la muestra en nuestro caso) los tres valores más extremos, que son las muestras 8, 112 y 47, que pueden verse además en el gráfico de la distancia de Cook. Para ver si estos valores con un alto leverage (llamados también outliers) tienen una alta influencia en el modelo, podemos utilizar de igual manera el gráfico de Residuals vs Leverage. Los outliers marcados anteriormente son los que pueden tener una influencia significativa en el modelo, pero para que estos outliers puedan afectar de manera significativa, deberían encontrarse por encima de la línea discontínua de la esquina superior derecha o por debajo de la línea discontínua de la esquina inferior derecha del gráfico mencionado (valor de los residuos estandarizados menor a -2 o superior a 2 con una distancia de Cook superior a 0.3 por norma general). Como en nuestro modelo los outliers no se encuentran en esas posiciones del gráfico, podemos confirmar que no tienen una influencia significativa en el ajuste del modelo de regresión lineal.

(f) Un amigo americano está a punto de entrar en un hospital. Quiere saber el intervalo de confianza al 90% para la predicción del riesgo de infección utilizando el modelo del apartado anterior. Sabe que el hospital tiene los valores de stay= 9.6 días, culratio=15.5 y region=NE.

Como tenemos el modelo lineal del apartado anterior, podemos utilizar la función `predict()` con el intervalo de confianza indicado. Introducimos primero los valores de **stay**, **culratio** y **region** en un nuevo dataframe para que tenga en cuenta los valores a partir los cuales se quiere predecir el riesgo de infección.

```
##        fit      lwr     upr
## 1 4.019383 2.370046 5.66872
```

Según la predicción sobre el modelo lineal del apartado anterior y con los datos de las variables aportados, en un intervalo del 90% de confianza, el riesgo de infección que tiene se encuentra entre el 2.37% al 5.67%, con una media del 4.02% de riesgo de infección.

(g) En la estimación el modelo del apartado (e), el coeficiente de region==NE no aparece. ¿Puedes dar alguna explicación? Observar las tres últimas columnas de la matriz de diseño del modelo ¿Cómo se codifican los 4 valores del factor region? ¿Cómo se calcula *a mano* una predicción para un hospital con region==NE con la ecuación de este modelo? Si ejecutamos la siguiente instrucción: options(contrasts=c('contr.sum','contr.poly')) y recalculamos la estimación del modelo, ¿cómo se codifican ahora los 4 valores del factor region? Ahora, ¿cómo se calcula *a mano* una predicción para un hospital con region==NE con la ecuación de este otro modelo?

Observamos el resumen estadístico, la matriz de contrastes en forma "treatment" y la matriz de diseño del modelo:

```
##
## Call:
## lm(formula = infrisk ~ stay + culratio + region, data = senic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1084 -0.6564  0.2108  0.6047  2.6994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.161945   0.648026  -0.250  0.80314
## stay         0.341202   0.056888   5.998 2.76e-08 ***
## culratio     0.058438   0.009729   6.007 2.65e-08 ***
## regionNC     0.320149   0.265812   1.204  0.23109
## regionS      0.199331   0.271845   0.733  0.46501
## regionW      1.030779   0.350042   2.945  0.00397 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9726 on 107 degrees of freedom
## Multiple R-squared:  0.4973, Adjusted R-squared:  0.4738
## F-statistic: 21.17 on 5 and 107 DF,  p-value: 1.147e-14
```

```
##   2 3 4
## 1 0 0 0
## 2 1 0 0
## 3 0 1 0
## 4 0 0 1
```

```
##   (Intercept)  stay culratio regionNC regionS regionW
## 1           1  7.13      9.0        0       0       1
## 2           1  8.82      3.8        1       0       0
## 3           1  8.34      8.1        0       1       0
## 4           1  8.95     18.9        0       0       1
## 5           1 11.20     34.5        0       0       0
## 6           1  9.76     21.9        1       0       0
```

El coeficiente de la variable **region** cuando **region==NE** no aparece porque se utiliza dicha región como coeficiente base de las regiones de la variable (pudiendose ver en la matriz de contraste, donde el primer factor tiene valor 0). Es decir, esta región tiene un coeficiente 0 y se utiliza como región de referencia para comparar con las otras regiones y obtener un coeficiente de éstas.En la matriz de diseño esto se observa de manera que la región "NE" se identifica cuando las otras tres regiones tienen como valor 0 (como se puede observar en la

muestra 5 de la matriz de diseño). Para las otras regiones, la región de la muestra se identifica como valor 1 mientras que las otras regiones se identifican como 0, para así determinar qué región es la que posee la muestra. Para calcular a mano una predicción para un hospital con **region == NE** deberemos identificar aparte los valores de **culratio** y **stay** y resolver la ecuación $infrisk = -0.16 + 0.34 * stay + 0.058 * culratio + 0$ donde hemos identificado el coeficiente de NE como 0 (como hemos comentado anteriormente).

Recalculamos ahora el modelo del apartado (e) introduciendo previamente la instrucción mencionada en el enunciado:

Realizamos el resumen estadístico del modelo recalculado junto con su matriz de diseño:

```
##
## Call:
## lm(formula = infrisk ~ stay + culratio + region, data = senic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1084 -0.6564  0.2108  0.6047  2.6994
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.225620   0.539719   0.418   0.6768
## stay         0.341202   0.056888   5.998 2.76e-08 ***
## culratio     0.058438   0.009729   6.007 2.65e-08 ***
## region1     -0.387565   0.187069  -2.072   0.0407 *
## region2     -0.067416   0.155301  -0.434   0.6651
## region3     -0.188234   0.151375  -1.243   0.2164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9726 on 107 degrees of freedom
## Multiple R-squared:  0.4973, Adjusted R-squared:  0.4738
## F-statistic: 21.17 on 5 and 107 DF,  p-value: 1.147e-14
```

Observamos ahora para saber como funcionan los parámetros de contrastes la matriz de contraste para cuatro variables, junto con la matriz del modelo para observar los ejemplos de algunas observaciones:

```
##   [,1] [,2] [,3]
## 1    1    0    0
## 2    0    1    0
## 3    0    0    1
## 4   -1   -1   -1
```

```
##   (Intercept)  stay culratio region1 region2 region3
## 1           1  7.13      9.0      -1      -1      -1
## 2           1  8.82      3.8       0       1       0
## 3           1  8.34      8.1       0       0       1
## 4           1  8.95     18.9      -1      -1      -1
## 5           1 11.20     34.5       1       0       0
## 6           1  9.76     21.9       0       1       0
```

En esta matriz de contraste se indica en cada columna los contrastes realizados. Por ejemplo en la primera columna se calcula la diferencia entre el primer y último nivel del factor que en nuestro caso entonces sería la diferencia entre NE y NW. En las siguientes columnas se realiza el mismo contraste con la diferencia entre la segunda región con la última y la tercera con la última respectivamente, indicando que siempre utilizaremos como base en este caso el último nivel del factor (NW).

En este caso los coeficientes se obtienen con las diferencias entre distintos pares de regiones. Por lo tanto, si

queremos realizar a mano la misma predicción que en el caso anterior, tenemos que realizar la ecuación con los coeficientes de manera distinta, aunque el resultado será el mismo. La predicción del modelo entonces sería $infrisk = 0.22 + 0.34 * stay + 0.058 * culratio - 0.3875 * 1$.

Para devolver las opciones a como estaban anteriormente:

(h) Consideremos ahora el modelo con 4 variables regresoras: stay, age, xratio y medschl. Alguien sugiere que el efecto de medschl sobre el riesgo de infección puede interactuar con age y con xratio. Añadir los términos de interacción apropiados al modelo de regresión, ajustar el modelo ampliado y contrastar si los términos de interacción ayudan. Usar $\alpha = 0.1$. Indicar la hipótesis nula, la alternativa, la regla de decisión y la conclusión.**

Primero ajustamos un modelo con las cuatro variables regresoras:

Ahora ajustamos el modelo ampliado con interacciones sugerido:

Para saber si los términos de interacción añadidos al modelo ampliado ayudan a mejorar el modelo, podemos realizar una comparación del modelo ampliado con el modelo más simple. Como el modelo más simple es un modelo anidado del ampliado, podemos realizar un F-test. La hipótesis nula del contraste de hipótesis es que la variación entre estos dos modelos no es significativa. Por lo tanto la hipótesis alternativa es que sí que hay una variación significativa entre estos modelos. La regla de decisión que aplicaremos la basaremos en el nivel de significación determinado en el enunciado, es decir, , con un $\alpha = 0.1$. Esto implica que si el p-valor del F-test es mayor que alfa (0.1), se aceptará la hipótesis nula, y si el p-valor es menor que alfa, entonces se rechazará la hipótesis nula. Dicho esto, ahora realizamos el F-test:

```
## Analysis of Variance Table
##
## Model 1: infrisk ~ stay + age + xratio + medschl + medschl * age + medschl *
##     xratio
## Model 2: infrisk ~ stay + age + xratio + medschl
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    106 122.05
## 2    108 127.24 -2   -5.1964 2.2566 0.1097
```

Como el p-valor del F-test es mayor al nivel de significancia estipulado (aunque por poco), aceptamos la hipótesis nula. Por lo tanto, confirmamos que con este nivel de significancia consideramos los dos modelos similares. Por ello, nos quedaremos preferiblemente con el modelo más simple, es decir, el modelo que no tiene interacciones.