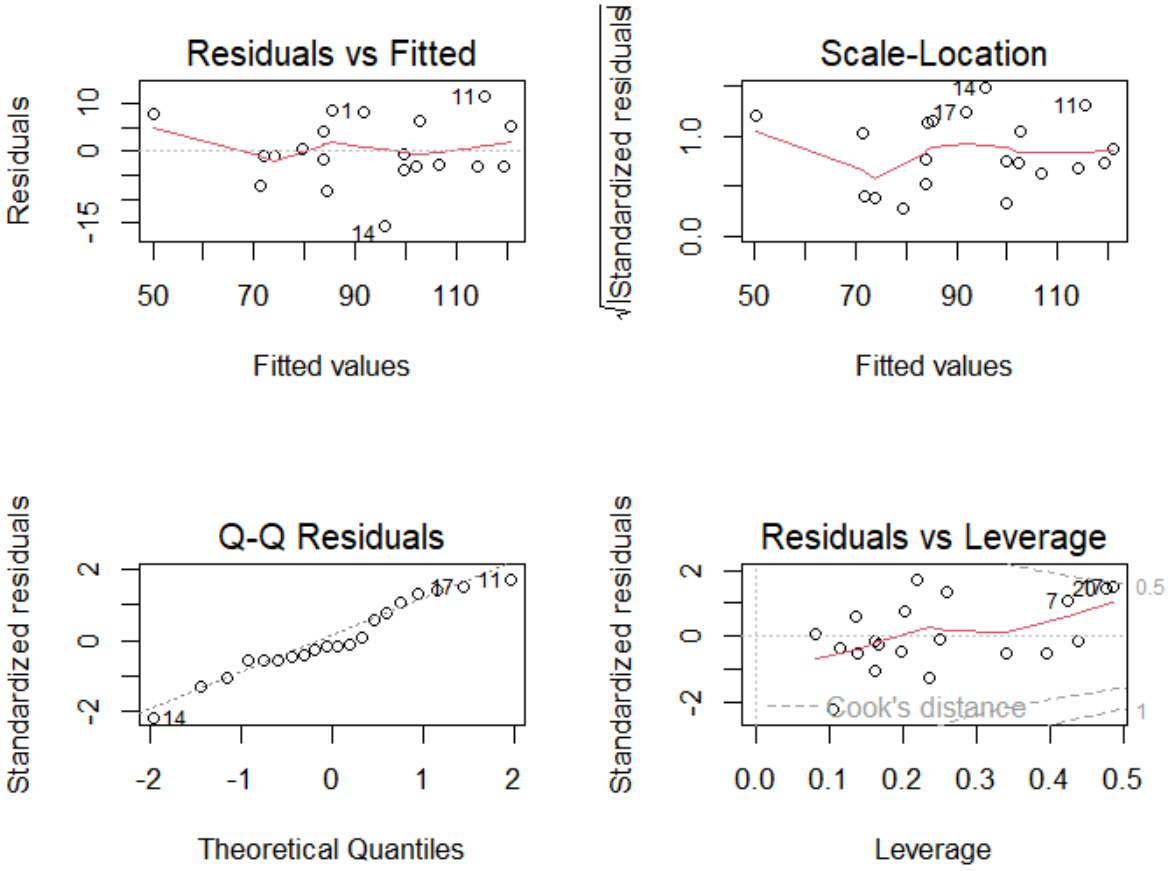
A)

Queremos estudiar la relación entre 4 variables cuantitativas y 1 variable cuantitativa. Para ello vamos a plantear modelos de regresión múltiple. Para estudiar si el modelo de regresión es lineal planteamos una regresión lineal y estudiamos como se ajustan los datos a dicho modelo. Tras plantear el modelo podemos observar las gráficas generadas por el modelo:



Observando la primera gráfica (Residuals vs Fitted) podemos ver que el modelo parece ajustarse bien a una regresión lineal. Vamos a plantear el modelo.

Primero ajustamos el modelo a través de RStudio, donde obtenemos la siguiente tabla:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -175.88440   48.61025  -3.618  0.00253 **
x1             0.49108    0.35937   1.366  0.19192
x2             0.02018    0.13987   0.144  0.88719
x3             1.30217    0.35010   3.719  0.00206 **
x4             0.83057    0.25328   3.279  0.00507 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.505 on 15 degrees of freedom
Multiple R-squared:  0.8847,    Adjusted R-squared:  0.8539
F-statistic: 28.77 on 4 and 15 DF,  p-value: 7.033e-07
```

En la primera columna podemos observar las estimaciones para los parámetros del modelo de regresión lineal asociada a cada una de las variables, pero antes de definir el modelo con dichos coeficientes debemos de comprobar las condiciones de validez:

-Normalidad: Para comprobar la normalidad realizamos el test de contras de hipótesis de Shapiro-Wilk a un nivel de significación de $\alpha=0.1$, donde H_0: los residuos siguen una distribución normal, H_1: LANC. Tras realizar el test con RStudio obtenemos un p-valor de $0.3931>0.1$ por lo tanto no tenemos motivos como para rechazar H_0, es decir, podemos suponer que los residuos siguen una distribución normal.

También podemos estudiar la normalidad remitiéndonos a la gráfica “Q-Q Residuals” que se mostró previamente, en la que se comparan los cuantiles teóricos de la distribución normal con los cuantiles muestrales de los residuos estandarizados. Cuanto más se ajusten los cuantiles de los residuos a la recta (los cuantiles teóricos) será más posible que los residuos se ajusten a una distribución normal.

-Homocedasticidad: Debemos de comprobar que todos los residuos tienen la misma varianza, para ello planteamos un contraste de hipótesis a un nivel de significación $\alpha=0.05$ en el que H_0: Todos los residuos tienen la misma varianza, H_1: LANC. Tras realizar el test de Breusch-Pagan en RStudio obtenemos un p-valor de $0.1936>0.05$, por tanto, no tenemos motivos para rechazar H_0, es decir, podemos suponer que todas las varianzas son iguales.

También podemos estudiar la homocedasticidad remitiéndonos a la gráfica anterior “Residuals vs FItted”, cuanto más cerca de la recta 0 del eje y estén distribuidos los valores mayores será la igualdad entre varianzas, en este caso podemos observar que algunos valores distan bastante de dicha recta, lo cual concuerda con un p-valor cercano al nivel de significación (0.05).

-Incorrelación: Planteamos el test de hipótesis de Durbin-Watson en RStudio el cual contrasta H_0: los residuos son incorrelados, H_1: LANC. Tomando un nivel de significación α=0.05, tras realizar el test obtenemos un p-valor de 0.2765>0.05, resultado no significativo, es decir, podemos suponer que los residuos son incorrelados.

-Linealidad: Este, a mi juicio, es la primera condición que debería comprobarse, pues en el caso de que no se cumpla el resto de contraste son innecesarios, para comprobar la linealidad de los datos se estudia la tendencia de los residuos con respecto a los valores ajustados, en la gráfica "Residuals vs Fitted" en el caso de que los datos no se ajustasen a una función lineal la parte no lineal de ese ajuste se vería reflejada en los residuos, haciendo que estos adquiriesen una tendencia no lineal. En este caso podemos observar que presentan una tendencia lineal, por lo tanto, podemos aplicar el modelo. Este estudio se realizó con menos detalle al inicio del ejercicio.

Por tanto, podemos aplicar el modelo de regresión lineal, el cual queda de la forma:

$$Y = -175.88440 + 0.49108 \cdot x_1 + 0.02018 \cdot x_2 + 1.30217 \cdot x_3 + 0.83057 \cdot x_4$$

Donde Y representa el valor de la nota de la prueba de evaluación y las variables x_i representan los valores de las notas que un individuo obtendría en las diferentes pruebas, de esta forma conociendo esas 4 notas podemos estimar la nota de la prueba de evaluación. Este modelo tiene un R^2 de 0.8847 y un R_a^2 de 0.8539.

B)

La recta de regresión lineal nos indica que existe una relación lineal entre los resultados de las 4 pruebas de selección y los resultados de la prueba de evaluación. Concretamente cuanto mayor sea la nota en cualquiera de las 4 pruebas de selección mayor será la nota de la prueba de evaluación, siendo la tercera prueba de evaluación la que mayor impacto positivo tiene en la nota de la prueba donde por cada punto que saquemos en la 3ª prueba de selección se traducirá en 1.3 puntos de más en la prueba de evaluación. Por otra parte, la prueba de selección que menor efecto tiene sobre la nota de la prueba de evaluación es la 2ª prueba, en la cual por cada punto que saquemos en dicha prueba solo obtendremos un aumento de 0.02 puntos en la prueba de evaluación. Esto nos lleva a plantearnos si podríamos quitar dicha prueba. Para un análisis más profundo realizamos a través de RStudio un análisis de la multicolinealidad, en el cual calculamos los factores de inflación de la varianza, los cuales se obtienen a partir de los coeficientes de determinación de la regresión lineal

de cada una de las variables. Si alguno de estos coeficientes es mayor que 10 entonces hay claros indicios de multicolinealidad, en este caso los respectivos factores de inflación de la varianza son:

```
      x1       x2       x3       x4
1.052861 1.401904 2.900594 2.410771
```

Como todos son < 10 no tenemos indicios de multicolinealidad, por lo tanto, ninguna de las variables regresoras está proporcionando información redundante, es decir, no hace falta quitar ninguna de las pruebas de selección.

C)

Para realizar un diagnóstico de las hipótesis teóricas planteamos el siguiente contraste:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1: LANC

Donde $\beta_i$representa el coeficiente de regresión lineal asociado a la variable x_i.

Realizamos el test y fijando un nivel de significación α=0.05 obtenemos:

```
F-statistic: 28.77 on 4 and 15 DF,  p-value: 7.033e-07
```

Por tanto, aceptamos H_1, es decir, alguno de los coeficientes es no nulo, debemos de realizar 4 contrastes de hipótesis:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

Para i=1,2,3,4

Obtenemos los siguientes resultados:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -175.88440   48.61025  -3.618  0.00253 **
x1             0.49108    0.35937   1.366  0.19192
x2             0.02018    0.13987   0.144  0.88719
x3             1.30217    0.35010   3.719  0.00206 **
x4             0.83057    0.25328   3.279  0.00507 **
```

Observamos que existen pruebas de selección cuyos coeficientes de regresión no podemos asumir que sean no nulos, por tanto, no podemos considerarlos en el modelo, vamos a eliminar dichas pruebas con el fin de obtener un modelo lineal en el

que todas las pruebas tienen un papel significativo en la predicción de la nota de la prueba de evaluación.

Para ellos vamos a plantear distintos métodos:

Backward: En este método realizamos los test parciales de todas las variables del modelo y eliminamos la variable con mayor p valor, repetimos este procedimiento hasta llegar a un modelo donde todos los test parciales tienen un resultado significativo.

En este caso la variable con mayor p valor en los test parciales es x_2 por tanto la eliminamos del modelo, obteniendo el nuevo modelo:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -176.5616    46.8793  -3.766 0.001689 **
x1             0.4995     0.3436   1.454 0.165384
x3             1.3222     0.3113   4.247 0.000615 ***
x4             0.8294     0.2453   3.381 0.003807 **
```

Podemos observar que en este modelo el coeficiente $x_1$da un p valor mayor que 0.05, por tanto, eliminamos dicha variable del modelo:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -114.9880    20.7321  -5.546 3.55e-05 ***
x3             1.2657     0.3189   3.969 0.000991 ***
x4             0.8414     0.2530   3.325 0.004006 **
```

Ya hemos llegado a un modelo en el cual todas las variables son significativas, por tanto, concluimos que las únicas pruebas de selección que nos ayudan a estimar la nota de la prueba de evaluación son la 3ª y 4ª prueba. Además, ambos presentan una relación directa, es decir, cuando aumenta la nota en las pruebas de elección aumenta la nota en la prueba de evaluación. La ecuación del modelo sería:

$$Y = -114.9880 + 1.2657 \cdot x_3 + 0.8414 \cdot x_4$$

Con un valor R^2 de 0.8693 y un valor R_a^2 de 08539, es decir, podemos explicar un 88.45% de la nota de la prueba de evaluación a través de las pruebas 3 y 4 de selección. Por último, los intervalos de confianza son:

```
                   5 %       95 %
(Intercept) -151.0537836 -78.922213
x3             0.7109798   1.820366
x4             0.4012223   1.281578
```

Modelo de selección basado en AIC: Este método se basa en buscar el modelo que tenga el menor AIC, el AIC es un coeficiente que se basa en la idea de sumar al desajuste medio del modelo el sobreajuste de este. Aplicando el modelo obtenemos:

```
Step:  AIC=82.9
y ~ x1 + x3 + x4

       Df Sum of Sq     RSS    AIC
<none>               846.09 82.898
- x1    1    111.74  957.82 83.379
- x4    1    604.63 1450.71 91.682
- x3    1    953.69 1799.78 95.994

Call:
lm(formula = y ~ x1 + x3 + x4, data = trabajo)

Coefficients:
(Intercept)           x1           x3           x4
  -176.5616       0.4995       1.3222       0.8294
```

Un resultado distinto al obtenido con el método anterior, aunque podemos observar que si eliminásemos x_1 el AIC no aumentaría mucho. Las ecuaciones del modelo serían:

$$Y = -176.5616 + 0.4995 \cdot x_1 + 1.3222 \cdot x_3 + 0.8294 \cdot x_4$$

Con un valor de 0.8845 y un valor de 08629, es decir, podemos explicar un 88.45% de la nota de la prueba de evaluación a través de las pruebas 1, 3 y 4 de selección. Por último, los intervalos de confianza son:

```
                    5 %       95 %
(Intercept) -258.4073513 -94.715791
x3             0.7786424   1.865801
x4             0.4011519   1.257595
x1            -0.1004228   1.099375
```

Concluimos que los resultados obtenidos en los apartados anteriores no son correctos, pues en ellos no hemos tenido en cuenta el nivel de significación de cada una de las variables explicativas en las variables respuestas.