Biología Sanitaria Examen Enero 2017

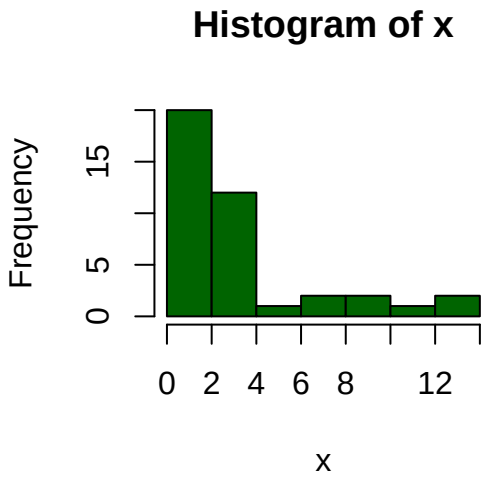
Equipo docente

12 de enero de 2017

Cuestión voluntaria 1

Supón que dispones de una muestra de una variable cuantitativa en la que la mediana es mucho menor que la media. Dibuja, de forma aproximada, un histograma que corresponda a esos datos, y explica porqué tiene esa forma.

La media, que es una medida numérica, es sensible a la presencia de unos pocos valores anormalmente (comparados con el resto) grandes o pequeños



Por ejemplo, para el diagrama anterior, la media vale 3.1690249 y la mediana 2.0087763

Cuestión voluntaria 2

Dispones de dos muestras pareadas de una variable cuantitativa. Se quiero probar que la media en la muestra 1 es mayor que en la muestra 2. Escribe las hipótesis nula y alternativa que definen el contraste. Fijamos

$$H_0 : \mu_1 \leq \mu_2, \qquad H_1 : \mu_1 > \mu_2$$

Así, sólo supondremos que $\mu_1 > \mu_2$ cuando haya una evidencia muestral muy fuerte a su favor. En caso de haber elegido fijas las hipótesis nula y alternativa al revés, nuestra comprobación se limitaría a "a la vista de nuestra muestra, no hay evidencia de que $\mu_1 < \mu_2$".

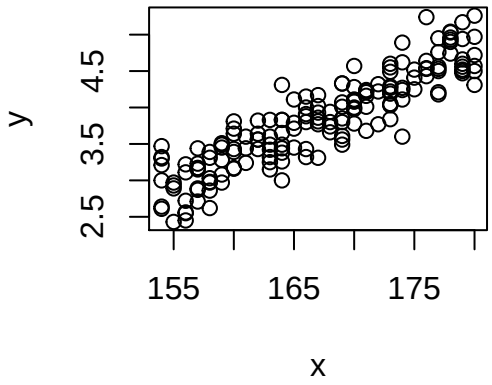
Pregunta regresión

Se dispone de los siguientes datos para estudiar la asociación entre el volumen máximo expirado en el primer segundo de una expiración forzada (FEV1) y la talla medida en centímetros de un grupo de 170 adolescentes de edades comprendidas entre los 14 y los 18 años

```
##
## Call:
## lm(formula = y ~ x)
```

```
##
## Coefficients:
## (Intercept)            x
##    -8.65153      0.07473
```

Aunque no se pide explícitamente, **siempre** hay que representar los datos, para cercionarse de que tiene sentido aproximarlos por una recta.

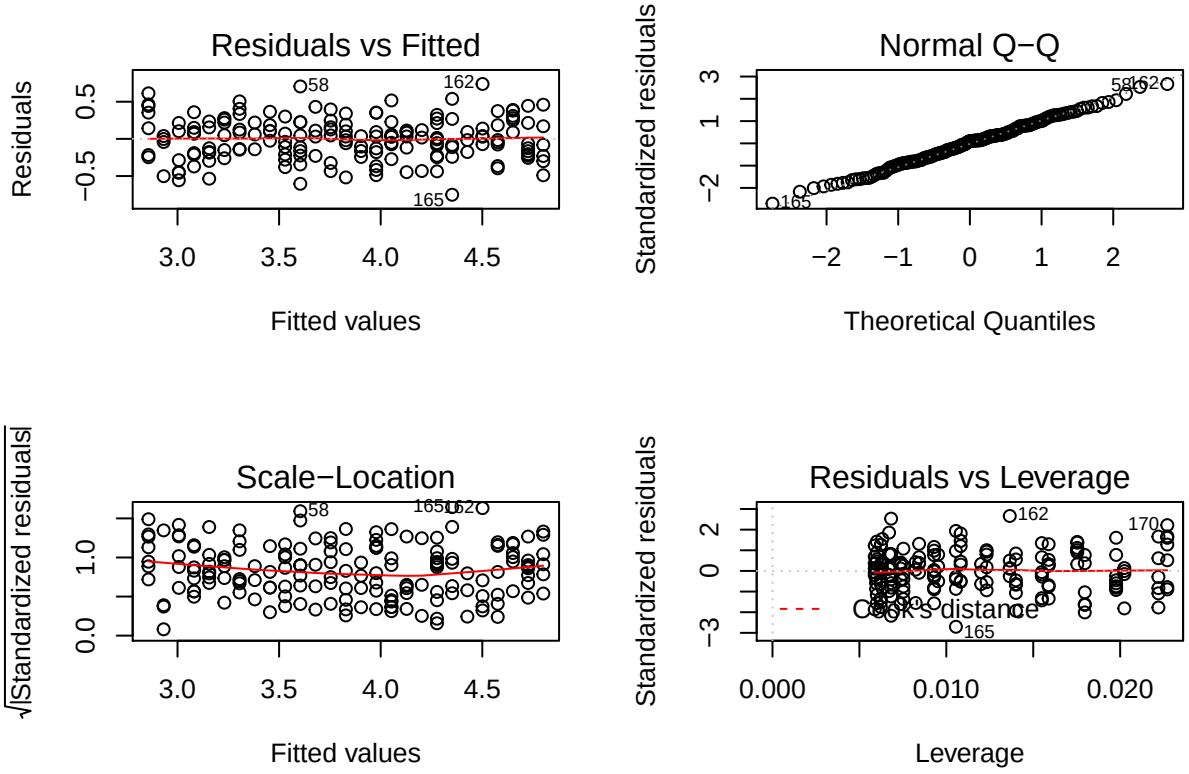


Calcula los coeficientes de la recta de regresión

```
## (Intercept)           x
## -8.65153345  0.07473125
```

¿Se cumplen las condiciones para aplicar la regresión lineal?

Si usas métodos gráficos, lo más adecuado es visualizar los residuos (homocedasticidad) y los residuos estandarizados (normalidad), como en las gráficas de la primera fila



Por supuesto, hay que interpretar tanto diagrama de residuos frente a valores observados (arriba, izquierda) como el QQ-plot (arriba, derecha). Puedes encontrar la interpretación de estos diagramas en las páginas

438-450 del libro, o en el tutorial 12 (página 12).

Otra alternativa pasa por usar la librería `gvlma` para obtener el p-valor correspondiente a los contrastes sobre el sesgo (skewness), la kurtosis y la homocedasticidad. Las hipótesis nulas de esos contrastes se formulan, esencialmente, como $H0$: el sesgo es como el de una normal, $H0$: la kurtosis es como la de una normal, $H0$: hay homocedasticidad.Los correspondientes p-valores son

```
## [1] 0.9557
```

```
## [1] 0.5096
```

```
## [1] 0.8387
```

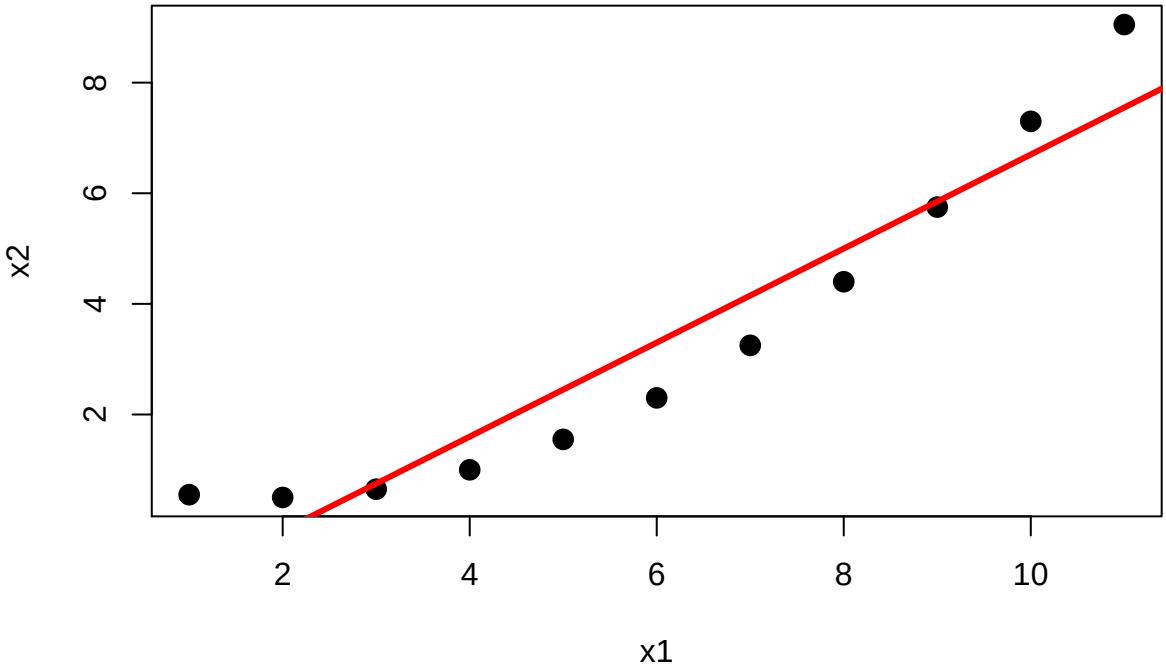
lo que nos lleva a no tener evidencia para rechazar las hipótesis de normalidad y homocedasticidad.

Calcula e interpreta los coeficientes de correlación y de determinación

```
## [1] 0.9015422
```

```
## [1] 0.8127783
```

- **Coeficiente correlación** vale 0.9015422. Su signo coincide con el de la pendiente de la recta. Por tanto, en este caso la relación es directa (al aumentar la variable explicativa, aumenta la respuesta). Los puntos de la nube parecen sugerir una relación línea y eso es compatible con el valor alto de r. La interpretación recíproca (como r es próximo a 1 los puntos tienden a estar sobre una recta) es **erronea**. Mira la siguiente nube de puntos,



que tiene forma de (¡es una!) parábola, a pesar de que su coeficiente de correlación lineal es 0.9500321 superior al obtenido en nuestro problema.

- **Coeficiente determinación** vale 0.8127783. Es el cociente entre las variabilidades explicadas por el modelo y total. Cuanto más próxima es a 1, más *explicativo* es el modelo, en el sentido de que no es preciso el concurso de más variables (que no hemos tenido en cuenta) para explicar satisfactoriamente cómo varían de forma conjunta las variables explicativa y respuesta.

A la vista del modelo, si la estatura de individuos difiere en 10cm (y está dentro del rango analizado), estima la diferencia en los valores de su FEV1.

Hay que calcular el intervalo de confianza de la pendiente

```r
confint(lmXY, level=0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept) -9.5653730 -7.73769386
## x            0.0692683  0.08019421
```

y multiplicar sus extremos por 10,

```
## [1] 0.692683
```

```
## [1] 0.8019421
```

Pregunta Anova

Se quiere evaluar la eficacia de distintas dosis de un fármaco contra la hipertensión arterial, comparándola con la de una dieta sin sal. Para ello se seleccionan al azar 400 hipertensos y se distribuyen aleatoriamente en 4 grupos, todos del mismo tamaño. Al primero de ellos no se le suministra ningún tratamiento, al segundo una dieta, al tercero el fármaco a una dosis (baja) y al cuarto el mismo fármaco a otra dosis (alta). Las presiones arteriales sistólicas de los 400 sujetos al finalizar los tratamientos son:

Decide si hay diferencias significativas en la respuesta a los distintos tratamientos y, en caso afirmativo, ordenalos según su eficacia. En particular, nos interesa saber si hay diferencias significativas entre administrar dosis bajas y altas del medicamento No olvides, como mínimo,

- Leer detenidamente el enunciado ¡no te dejes nada!
- Escribir las hipótesis nula y alternativas adecuadas
- Comprobar si se verifican las condiciones teóricas necesarias para aplicar la técnica estadística que decidas utilizar

La hipótesis nula es

$$H_0 : \mu_1 = \cdots = \mu_4$$

y la alternativa H_1: "al menos una media es diferente del resto".

Un error muy frecuente es establecer como hipótesis alternativa "todas las medias son distintas", que no es lo contrario de "H_0: todas las medias son iguales".

Contraste ANOVA

```r
datos.lm = lm(datos$Respuesta ~ datos$Tratamiento)
anova(datos.lm)
```

```
## Analysis of Variance Table
##
## Response: datos$Respuesta
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## datos$Tratamiento   3  23063  7687.7  45.278 < 2.2e-16 ***
## Residuals         396  67237   169.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

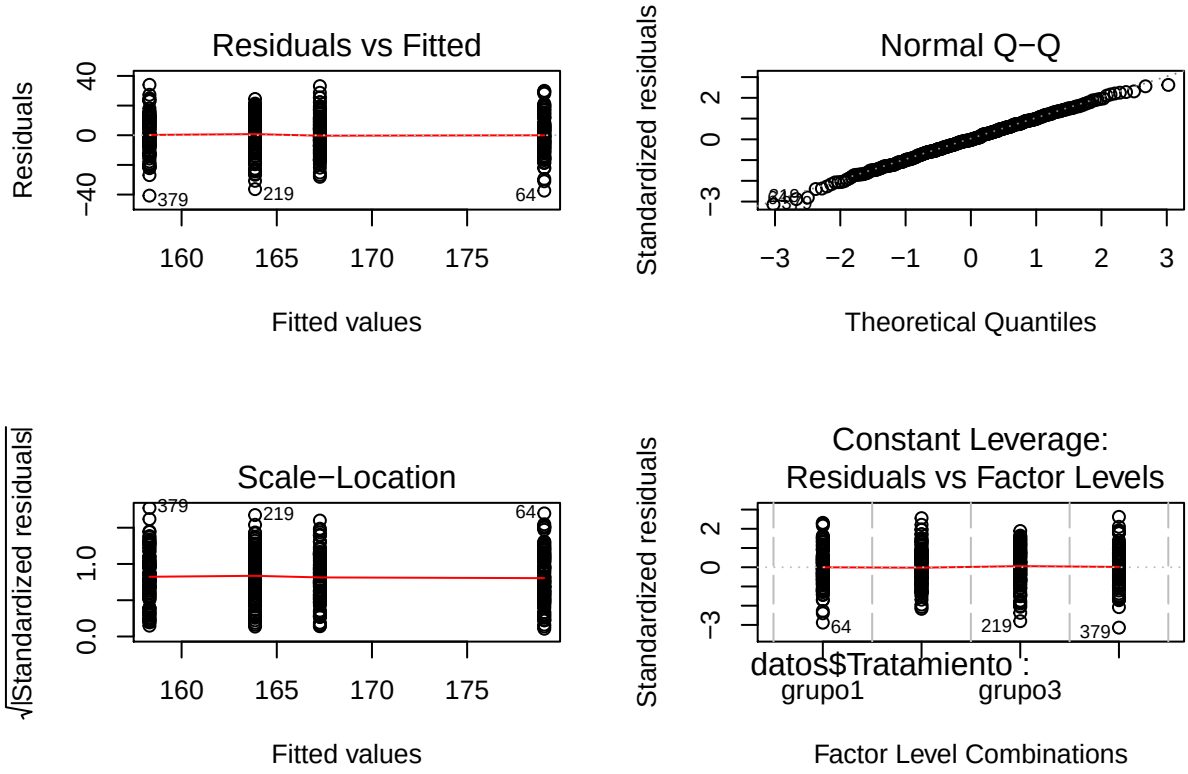
El p-valor es muy pequeño (menor que $2.2 \r{u} 10^{-16}$), por lo que el contraste ANOVA es significativo: alguno de los tratamientos es más efectivo que los demás. Para calcular el p-valor exacto del contraste, hay que calcular $P(F>45.278) $, donde F sigue la distribución de una F de Fisher-Snedecor con 3 y 396 grados de libertad. En concreto

```r
pf(45.278, df1 = 3, df2=396, lower.tail = F)
```

```
## [1] 3.534691e-25
```

Condiciones ANOVA

Ver los comentarios hechos en el apartado correspondiente a la recta de regresión.



Ordenación de medias

Una vez que el contraste ANOVA es significativo hay que ordenas las medias.

Podemos usar la tabla de Bonferroni. El ajuste de Bonferroni reparte el error de tipo I entre todas las comparaciones 2 a 2 de modo que la probabilidad de cometer error de tipo I en al menos una comparación se mantiene dentro del nivel de significación prefijado.

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  datos$Respuesta and datos$Tratamiento
##
##        grupo1  grupo2  grupo3
## grupo2 5.3e-09 -       -
## grupo3 2.3e-13 0.378   -
## grupo4 < 2e-16 1.1e-05 0.019
##
## P value adjustment method: bonferroni
```

El siguiente diagrama incorpora un código de letras que sintetiza las comparaciones dos a dos de las medias:

- La media del grupo1 (letra c) es significativamente mayor que las demás
- Las medias de los grupos 2 y 3 (letra b) son iguales entre sí, y significativamente menores (respectivamente, mayores) que la del grupo1 (respectivamente, que la del grupo4).
  - La media del grupo 4 (letra a) es significativamente menor que el resto de grupos.